Conference Title

The Second International Conference on Data Mining, Internet Computing, and Big Data (BigData2015)

Conference Dates

June 29 - July 1, 2015

Conference Venue

University of Mauritius, Reduit, Mauritius

ISBN

978-1-941968-13-0 ©2015 SDIWC

Published by

The Society of Digital Information and Wireless Communications (SDIWC)
Wilmington, New Castle, DE 19801, USA
www.sdiwc.net

TABLE OF CONTENTS

Machine Learning Techniques for Building a Large Scale Production Ready Classifier

Arthi Venkataraman
Wipro Technologies
72, Electronics City, Bangalore
arthi.venkat@wipro.com

ABSTRACT

This paper brings out the various techniques we have followed to build a production ready scalable classifier system to classify the tickets raised by employees of an organization. The end users raise the tickets in Natural language which is then automatically classified by the classifier. This is a practical applied research paper in the area of machine learning. We have applied different machine learning techniques like active learning for improving the accuracy of the prediction and have used clustering for handling the data issues found in the training data. The approach we used for the core classifier combined the results of multiple machine learning algorithms using suitable scoring techniques.

Use of this system has given more than 50% improvement in the tickets re-assignment index and more than 80% accuracy has been achieved in correctly identifying the classes for the tickets. The system is able to perform at scale, has response times well within the expectations and handles the peak load.

Key takeaways from this paper include:

- How to build live production ready classifier system
- How to overcome the data related challenges while building such a system
- Solution architecture for the classifier system
- Deployment architecture for the classifier system
- Being prepared for the kind of post deployment challenges one can face for such a system

Benefits of building such a system include Improved Productivity, improved End user experience and quick turnaround time.

KEYWORDS

Natural Language Processing, Machine Learning, Active Learning, Classification, Clustering, Name Entity Recognition, Committee-based approach, Cluster than Label, Shallow semantic parsing, Ontology, Cognitive Process Automation

1 INTRODUCTION

Across the world there is a significant pressure to drive costs down. Employees want better working conditions and better pay while at same time customers want better service and quicker responses. One answer to this is Cognitive Process Automation (CPA). CPA can be used to automate different business processes so that they can be performed consistently and quickly with reduced manual effort.

Employees face a lot of issues across different functions in the organization like Payroll, Infrastructure, Facilities, Applications, etc. When employees raise a ticket they are traditionally asked to manually select a category for their ticket from a hierarchical menu driven interface. This leads to a lot of wrong selection of choice by end user. This in turn causes tickets to be raised in wrong bucket and delays the resolutions due to re-assignments.

ISBN: 978-1-941968-13-0 ©2015 SDIWC

We have built a system which accepts a Natural language text input from end user and automatically classifies the ticket in the right category. The technical steps needed to build such a system include cleaning the data, natural language processing, machine learning both supervised and un-supervised, working with large data and continuous learning. In addition system should work in near real time.

This paper brings out the challenges we faced, the techniques we used to overcome these challenges, results obtained and some of the future work we want to do. This paper falls in the category of Industry applied research. We have researched the different techniques for our problem areas. Techniques were customized and improvised for our use case to build the system. Key focus areas of this paper are the Data Cleaning and Augmentation, Building the Classifier and Deployment.

2 BACKGROUND

Incidents in the organization are logged using different incident logging mechanisms. All of them have a traditional User interface which users navigate based on their understanding of the problem statement. There was no natural language interface. There were many in-accuracies in the incident logging process since it was dependent on the users understanding of the issue and whether they chose the right system, right screen and right item in screen to log the problem. Users would navigate, select the classification, enter their problem and submit. The issue with this was that users frequently did not know under which category to raise the ticket and hence raised the same in the wrong category. This leads to re-assignment of tickets where the ticket landed in the wrong bucket and hence was continuously passed around before it landed in the correct bucket to finally get resolved. There was more than 35% re-assignment index.

A huge volume of tickets are raised across functions every day. Hence a 35% re-assignment was in effect compounding the load of an already loaded support team. A system had to be designed which understands the natural language description entered by end user to automatically log the incident in the correct category.

3 TECHNICAL CHALLENGES

There were many challenges to address while building this solution.

Input Data related challenges

- Classification to happen on unstructured text (Unstructured data)
- The input data had 3000 + classes. (Data with large number of classes)
- Ticket data was unclean with unwanted texts including names, dates, etc. (Unclean Data)
- There were also a significant number of tickets assigned to wrong classes. (Wrongly Labelled data)
- Ticket data was imbalanced with some classes having large number of tickets and others having too few tickets. (Imbalanced training data)

Deployment Related challenges.

- System was to be designed for a live use of hundred thousand users
- System has to be designed for 1000+ simultaneous users
- Response has to be under .5 sec for the system to be usable
- System was to be accessible across the globe both within and outside the organization network for all the employees

We will delve deeper into how the data related challenges were handled in a dedicated section for the same. The deployment related challenges are addressed as part of the Solution approach section.

4 PRIOR RESEARCH

There is a lot of research addressing each of the challenges brought out in the previous section. Few of the key ones are highlighted in this section.

Sanjoy (2008) handles large training data with minimal label using active learning approaches. Ullman (Chapter 12) details many algorithms in the “Large Scale Machine Learning” chapter. Each of these algorithms has potential application for the current use case. Gunter (2010) discusses how classifications returned by multiple classifiers can be combined using novel boosting techniques to improve the accuracy of classification. Zhu (2009) discusses how semi supervised techniques can be used to handle combinations of labelled and un-labelled data to build classifiers. Haibo discusses how synthetic sampling techniques can be used to handle imbalanced classes.

Handling large number of classes in machine learning is a separate focused research area. Active learning has also been used frequently for handling large number of classes. Jain (2009) has found a probabilistic k-nearest neighbors technique to efficiently handle very large number of classes. Sotoris (2006) brings out the different techniques in handling imbalanced data. These cover techniques of random under sampling, oversampling, etc. There are also techniques which handle this issue at the machine learning model level itself. (C4.5 Tree, One class classification, is some of the possible algorithms.).

Classification of service tickets by correlating different data sources as well as contextual services has been done in Gargi (2014). In Julia (2015) classification of data with noisy labels has been done using two class classification and application of inter class. In Weizhang (2014) active learning of labels for training data using LSI subspace signature is discussed. In Andreas (2015) an ensemble of classifiers is used. The aim of this is to be able to classify using data which is represented in different ways. In Bruce (1996) it was found that an ensemble of classifiers performed much better than individual classifiers

5 OVERALL SOLUTION APPROACH

Solution had to address the different challenges listed above. Figure 1 represents the main high level modules of the solution:

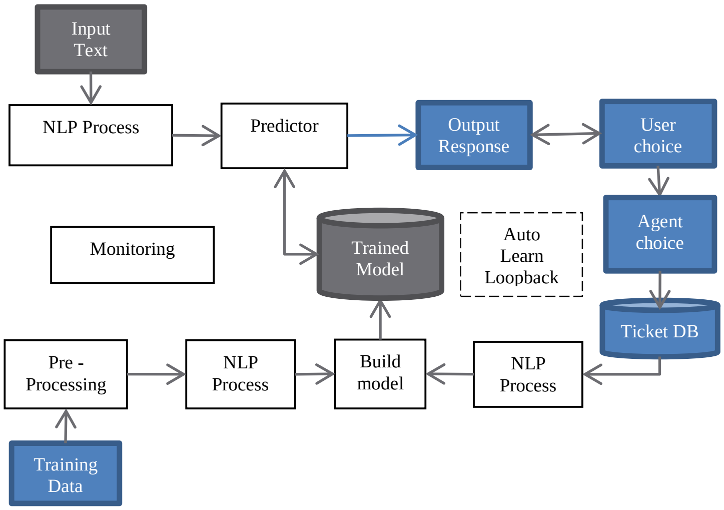


Figure 1: Solution Diagram

5.1 Pre-Processing

This step is done to ensure data is suitable for building accurate models. This is an offline activity and happens each time additional data is available to train the system. Different machine learning techniques are applied based on the understanding of the data as to what is applicable.

The data is analyzed to understand its characteristics. If it is found to have less clean training data but a larger unclean or unlabeled data set, then techniques like cluster then label as well as identification of most unclean data for manual labelling is followed. If it is found to have imbalanced classes then random oversampling techniques will be applied. If there are too many classes then cluster followed by auto labelling techniques will be applied to reduce the number of classes. All these techniques are discussed in the Handling Data Related challenges section.

5.2 NLP Process

This block accepts unstructured text. The unstructured text is tokenized. Stop words are removed. Special named entity recognition is run to remove irrelevant words like user names and contact details. Then the important features are extracted. Features can include bi-grams / tri-grams / n-grams, presence of certain important terms, etc. A vectorization is done on features. This is followed by compression to lower dimensionality.

5.3 Build Model

This block accepts the suitably processed features from the NLP process block.

Different algorithms are available as part of the Model tool box. A one-time configuration of the set of chosen algorithms for the class of input data is done. Once the algorithms are chosen the different parameters for the algorithms is also tuned. This happens one time initially. This is repeated regularly when a significant change in data composition is foreseen. Every time a new data set arrives from the NLP processing block the models are re-built as part of this. Different models are built as output of this block.

There are two sources of input data for this block. One is the historical training data. The other is the feedback from the usage of the system which comes back from live data. The historical data goes through the Pre-processing block as well as NLP process block. The live data goes through only the NLP process block.

The composition of this module is described in detail under the Implementation approach section.

5.4 Trained model

This is the model store. The trained models are stored inside this store. The models in the store are updated whenever the build model block is triggered with new inputs.

5.5 Prediction

Every time a new text arrives it is passed through the NLP block. The output of this is then fed to the prediction module. Prediction module predicts the correct classes for the given input. It predicts using the models stored in the trained model store. The prediction model is responsible for running the different ensemble of models, combining results from different models and scoring the same. This combining and scoring can be done in different ways. It is described in detail under the Building the Classifier system.

5.6 Automatic Learning Loop Back

This module is responsible for ensuring the accuracy of the system increases with use. This is done by implementing a feedback mechanism. The usage of the choices provided by the system is tracked to provide feedback. Whenever a ticket is closed in a class which is different from the originating class it is fed back as training data for the closure class. The ticket database is monitored regularly for such cases. These tickets are then fed through the NLP process block and then to the Build model module to re-build the models.

5.7 Monitoring Process

There is a need to continuously monitor the different processes. Monitoring will be done to check if the different blocks are up and running. The response time for a prediction is tracked. If there is a degradation of performance below a certain limit suitable corrective actions will be taken. Any process which is not responsive for a pre-configured amount of time will be re-started. Suitable logging of such incidents will be done. Root cause analysis will be done for all such cases to ensure the issue will not re-occur in future.

6 HANDLING THE DATA RELATED CHALLENGES

This section brings out how we can handle the different data related challenges. It brings out under what premise the approach will hold good.

6.1 Handling Unclean Data

Premise

- Training data is not clean with unwanted text including names, dates, and unwanted words.

Philosophy

- Good understanding of the data and what is causing it to be un-clean
- Creation of pattern recognizers / machine learning models to identify entities to tag and remove tokens causing issues
- Named Entity Recognition techniques used to identify dates and proper names. These were then filtered from the text.
- A domain specific stop word list to be created. This will be used to further filter the non-needed text. Concept graphs were extracted with focus only on the nouns. This will again be used to weed out unwanted text.

Targeted End Result

Clean corpus with all the unwanted data removed.

High Level Approach

Named Entity Recognition Approach (NER)

- Train classifier for named entity recognition
- Tag the data using the NER
- Remove the offending data
  - This processing has to be done on both the testing and training sets.

Stop Word List Approach

- A domain specific stop word list was created.
- This was used to further filter the unwanted text.

6.2 Increasing Available Clean Training corpus

Premise

- Difficult to manually get an accurately labelled complete corpus
- Have a small clean base corpus.
- Have a larger wrongly labelled / un-labelled data

Philosophy

Active Learning Approach to build a training corpus from un-labelled data.

- Use the small clean corpus to build a bigger corpus using the Active Learning approach
- Rather than clean complete corpus, find instances which are most un-clean
- Manually curate this un-clean set and label them
- Higher benefits since we are targeting tickets which are most un-clean.

Targeted End Result

To get a larger corpus for training the machine learning algorithm

Alternate High Level Approaches

- Identify the most un-clean instances for manually labelling
  - Train classifier with core set of clean tickets
  - For each unlabeled instance
  - Measure prediction uncertainty
  - Choose instances with highest uncertainty
  - Manually validate and label these instances.
- Self-Training
  - A classifier is built using smaller corpus. This is used to

    - predict the labels of the un-labelled data
- Label using 1 nearest neighbor
    - Find the nearest labelled neighbor and label using this
- Cluster and then Label
    - Cluster the tickets
    - Build a classifier for each cluster using labelled instances in that cluster
    - Classify all un-labelled instances in that cluster.

Detailed Approach to Identify the Most Unclean Data for Re-labelling

Voting approach:

- Get the uncertainty of classification for an instance based on the disagreements amongst members of the voting group
    - Members of the voting group are the set of algorithms selected for voting
- Tickets ranked on descending order with tickets with maximum disagreement will be at top of list
- Top x% (say 20%) of tickets will be chosen for further cleaning
- Most successfully used method
- A classifier is built using smaller corpus. This is used to predict the labels of the un-labelled data

Detailed Approach for Increasing the Training Data Set Using Cluster-then-Label

We selected a clustering algorithm A (k-means) and a supervised classifier algorithm L. (Support Vector Machine)

Input is a set of training points some of which are labelled and the rest are not.

- All the input training data points are clustered using the k-means algorithm.
- For each resulting cluster, get all the labeled instances in this cluster (Say S)
- If S is non-empty, learn a supervised predictor from S: fS = L(S).
- Apply fS to all unlabeled instances in this cluster.
- If S is empty, use the predictor f trained from all labeled data.

The output will be the labelled data based on the supervised classifier L.

6.3 Technique for Handling Wrongly Labelled Data

Premise

- There is a list of labelled training data which are known to have been wrongly labelled

Philosophy

- Use the wrongly labelled data to build a classifier to identify further wrongly labelled data in training corpus

End Result

Have a more accurately labelled corpus for training the machine learning algorithm

Detailed Approach

- Build a classifier to identify mislabeled data.
- Collect mislabeled data and use the same to train this classifier. This assumes there is some kind of pattern to the mislabeling.
- This could be used to predict other data if they are mislabeled.
- Re-label manually all mislabeled data

6.4 Technique for Handling Large Number of Classes

Premise

- Data set has too many classes
- Many of the classes were highly correlated. There is a possibility that many of these labels are very close to each other
- Is there a way closely related labels can be identified and merged automatically

Philosophy

- Group similar labels using different techniques

Targeted End Result

Have a smaller number of classes. Training data set will be grouped based on the smaller set of classes.

Different Approaches

There are many available approaches in literature. These include reducing the classes by calculating the similarity of the classes, Reduce the number of labels by finding the correlations between labels using generative models, etc. We went with the approach of clustering followed by auto labelling. We have detailed the approach in the following section.

Detailed Approach – Cluster Followed By Auto Labelling

- Cluster the entire training set
- Number of clusters will be set to the target number of classes we are looking for
- Post clustering identify the labels of original data set which fall into a cluster
- Choose the label with maximum data in a cluster as the name of the cluster
- To improve quality an audit can be done of the clustering to ensure there is indeed a logical grouping of similar labels

6.5 Technique for Handling Imbalanced Classes

The issue is that traditional classifiers seeking an accurate performance over a full range of instances are not suitable to deal with imbalanced learning tasks, since they tend to classify all the data into the majority class. We need to ensure classes with less data are also suitably represented. One positive point of the domain of our interest was typically the classes with less data are correspondingly the less important class. Hence a lower accuracy on the classes with less data can be tolerated.

Premise

- Data set has variation in the number of instances available for each class
- Some classes have lot of data and others have very less

Philosophy

- Choose a representative data set so that
  - Accuracy of classifications are good
  - Classes with lots of data do not overwhelm the classes with very data

Targeted End Result

Good classification accuracy. Results of classification are usable.

Different Approaches

There are many available approaches in literature. These include:

- Adaptive synthetic sampling to generate a balanced sample
- Oversample the classes with less data to ensure comparative data across classes
- Random under sampling of classes with more data

Detailed Approach – Oversample the Classes with Less Data

- Process the training data
- For classes with less data randomly pick data from these classes till number of tickets in these classes in equivalent to the other classes

7 BUILDING THE MODEL

Once the data pre-processing is over we will come to the actual task of building the classifier models. Algorithm should have following characteristics:

- Is a supervised machine learning algorithm
- Handles text classification
- Can handle the size of the training set
    - Many algorithms will satisfy 1 and 2 above however when used for training they never completed the training cycle.
- Prediction should be in near real time

7.1 Choosing the algorithms

An iterative process is followed to shortlist the correct algorithms. Different algorithms perform well with different kinds of data. Based on benchmarking of different algorithms the best performing algorithms are selected

- Parameters for selection will vary from domain to domain
- Key Parameters which we considered in order of importance are :
    - F- Score, Precision, Re-call
        - Without good benchmarks in this system would give wrong classifications
    - Model Prediction Time
        - Near Real time system needed this
    - Model building Time
        - Certain algorithms did not complete the training cycle and hence rejected.

After every cycle the above parameters are verified. If there is a scope for improvement in these parameters then another cycle is performed till we get the required level of performance. If for any algorithm we are unable to reach the desired level of performance and there is no scope for improvement then the algorithm is rejected.

At the end of this process the set of algorithms which are relevant are selected. These algorithms are enabled as part of the Build Model block. Whenever new training data arrives it is passed through the Build Model block. The output of this block is a set of trained models which are stored in the Trained Model store.

8 PREDICTING USING THE CLASSIFIER

In live use whatever incidents are entered by user are passed through NLP Process block. The output of this is then fed to the predictor. The predictor predicts the output classes using the models in the trained model store. The key responsibilities of the predictor are

- Predict using the models in the Trained model store
- Combine the results of multiple models in the store
- Score the results
- Share the predicted results with the score to the end user

8.1 Combining and Scoring the Results Using a Committee

This section elaborates the approach followed to combine results of multiple classifiers and score the results. The models built by the shortlisted algorithms are considered members of a committee. Each of the models predicts possible classes for the given input. It is possible for a model to return more than one class also. The different unique responses can be combined. The scoring can be based on the number of occurrences of a class across models. A choice which is predicted by all models will be given maximum weightage.

It is also possible to use the individual probability scores for occurrence of a class as predicted by a model. The challenge here would be to design a way to combine probabilities given by different models.

Scoring of results can also follow a weighted combination approach. It is possible to assign more weight for choices given by certain models of the ensemble based on their historical prediction accuracy.

9 SCALABLE DEPLOYMENT WITH NEAR REAL TIME RESPONSE

The system should be designed to handle the stated number of live simultaneous calls. It should be able to respond for prediction requests within the near real time performance constraints.

The system should be horizontally scalable. There will be multiple instances of the prediction system each residing on one machine. To understand the number of instances a benchmarking is to be done on an instance of the prediction solution to evaluate what is the peak load capacity for an instance beyond which the response time starts degrading. Based on the peak simultaneous load to be supported, the number of instances to be provisioned will be calculated. The load balancer will be configured to balance the requests across the different instances based on the performance and load on each of the instances. An auto recovery script will be created for each instance which will ensure that if the performance drops for any instance then the instance will automatically re-start.

System is to be accessible across the globe both within and outside the organization network for all employees. Hence services need to be deployed across special machines which are accessible outside the company network. Based on origin of request if request is outside the company network, load balancer will ensure that it will be serviced from these machines. Please see Figure 2 for the Production deployment diagram.

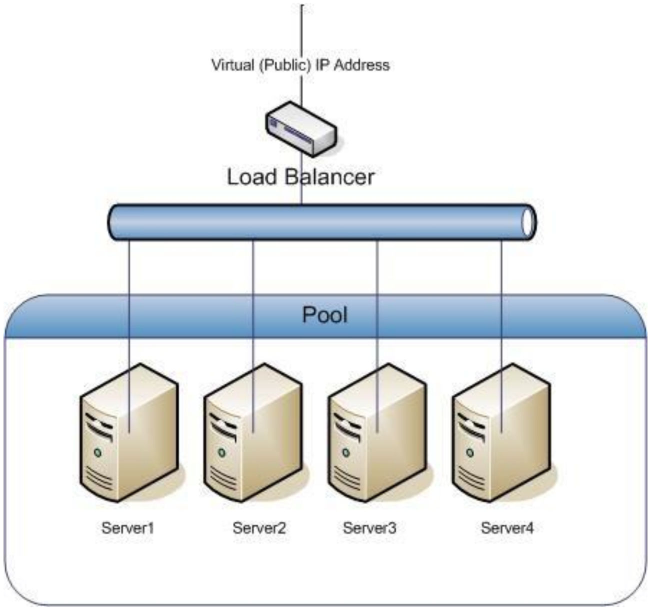


Figure 2: Production Deployment Diagram

10 Implementation Approach

We applied the different techniques as detailed in the previous sections. Detailed below are the different stages we followed in the actual implementation of this system.

10.1 Stage 1 – Build Initial Classifier Model

The overall requirements for the solution were clear and detailed out. We were to develop a classifier system which was to be able to accurately classify every ticket to the correct category. We received historical data dump of incident tickets with their closure category. Size of data was in range of half a million incidents and number of classes was in excess of 2500.

A classifier model was built which can handle text and the given scale of data. Logistic Regression was chosen for the initial classification step. Initial accuracy checks showed us that we were below 30% on overall accuracy.

10.2 Stage 2 – Data Cleanup

A detailed function wise data analysis was done. Data issues with each of the function were unearthed. All the data challenges as discussed in the Handling Data Challenges section were present in different degrees for the different functions data.

This step was done in close co-ordination with the associated function teams. Based on data analysis the applicable technique for the data relevant to the function was done.

To identify the data challenges a quick audit of 5% of random sample tickets was done for each of the functions data. Following categories of unwanted words were identified and removed

- People names
- Phone numbers
- Repetitive sentences across incidents which add no value to identify a ticket as they occur across different ticket. (For e.g. Please close this issue at earliest)

During the audit, a check was also done on the data labelling accuracy. If there was a significant mismatch, then that data set was marked as candidate for the pre-processing for wrongly labelled data.

We used Pivot table of excel sheet to understand what is the number of classes and the split of data among the classes. This test revealed if

- Number of tickets per class is very less
- Number of classes is too large
- Data is imbalanced with few classes having too many training tickets

Based on the results the corresponding techniques that were identified in the Handling Data Challenges were applied.

Actions taken for the different issues are summarized in the table 1.

Table 1: Techniques followed to solve data challenges

ID	Data Challenge	Technique followed
1	Unclean data	Named Entity recognition tagging and filtering out the tagged data Stop word removal
2	Wrongly labelled data	Active learning
3	Large number of classes	Clustering followed by Auto Labelling
4	Imbalanced data	Random oversampling for some classes based on given accuracy.
5	Too less data	Active learning

10.3 Stage 3 – Building the Classifier Model

Once data cleansing step was done we focused our efforts in building a more accurate classifier. A list of feasible algorithms which were applicable to our context was short listed. Some of the algorithms which we considered include:

- Naïve Bayes
- Support Vector Machine
- Ridge Classifier
- Multi-Layer Perceptron
- K-nearest neighbors
- Random Forest classifier, etc.

Algorithm chosen should be able to accurately discriminate between close classes.

There were issues such as if user specified Printer was out of paper it had to classified to printer issues under facilities management. However if printer has hardware issues it had to be classifier under printer hardware problem categories. If there was a network issue which was in turn causing the printer not being accessible then it had to classify to the network category. If user just specified Printer issue then system would have return all related categories.

Similarly if problem stated is My XYZ application is slow then it should be placed under XYZ application. If problem stated is machine is slow them should be placed under desktop / laptop issues.

For each of the short listed algorithms a detailed analysis was done on all the listed parameters F- Score, Precision, Re-call, Model building Time and Prediction Time. We also took the precision measurements at individual function levels. Our conclusion was no one algorithm performed equally well for all functions. Also certain algorithms like the Random forest took very long time to complete the training process. Algorithms which took more time were dropped from consideration.

10.4 Stage 4 – Building the Predictor

The predictor block was developed to use the models trained by the Model Training block to predict the class using the NLP processed incident ticket. The scoring component of the predictor block combines the results of the multiple models and scores them. We faced many challenges in designing the scoring and resulting combination approach. Each of the individual models gave multiple results with probabilities. We had to combine the probability scores returned by the different algorithms. In the end we combined the unique responses across the models. The scoring was based on the number of occurrences of a class across models. A choice which is predicted by all models was given maximum weightage.

Once this combination was done, the system was tested with a hidden test set. The overall accuracy was calculated. If the accuracy was below a threshold then the training process was repeated by asking additional data from the respective function teams or by increasing the data pre-processing efforts for these classes.

This was followed by a few iterations of the following steps till acceptable accuracy was reached.

1. Obtain more training data wherever possible
2. Apply the Data cleaning approaches
3. Perform Re-Training
4. Perform Accuracy validation

This whole process was repeated till a satisfactory accuracy was obtained across the system. The system is now ready to be used for classification.

In our case the ensemble of Support Vector Machine and Ridge Classification along with some in house techniques gave maximum performance.

10.5 Stage 5 – Production Rollout

While we had the classifier running at good accuracy in isolation we had to plan for its integration with the company wide incident logging portal. At this stage we also needed to take care that the classifier will be able to handle the load of all the company users and respond in near real time.

We created REST services of the classifier interface. It was deployed in a standard Windows 7 desktop with 8 GB RAM. A front end user interface was created as part of the employee’s incident logging portal. Employees could enter their issues in free text in this user interface. On submitting the incident a call was made to our deployed REST service. The service in turn responded with the set of predicted classes with scores.

To enable the handling of the load multiple such REST services were deployed across different machines. A load balancer was used. All requests from the employee portal hit the load balancer. The load balancer automatically re-directed the request to the correct back end service based on its knowledge of the services being up and the load on each of the services. The deployment diagram followed the Figure 2. The number of instances required was calculated based on load.

10.6 Stage 6 – Continuous Learning

Post deployment there were various teething issues with the system. This was more to do with the large scale infrastructure. Once this was resolved the next issue was accuracy of the system. Post roll out the system was performing at an average accuracy of around 70% which was not acceptable. This is when the continuous learning system was brought in. A list of all tickets where final closure class was not same as initial logging class was generated. These tickets were fed back to the system for learning. For extremely poorly performing functions we took a step back and repeated the Get Additional data, Data Cleanup and Re-train step to improve the accuracies.

10.7 Stage 7 –Monitoring

Post deployment there were various issues faced by employees. These included

- Complaints that the system never gave any classification
- Complaints that the new system gave wrong classification

Due to lack of upfront monitoring we were unable to pin-point where the issue was as there were multiple points of failures. For the first kind of issue where the user complaint was that the system never gave any classification the failures could be because

- Issue with front end
- Front end was fine but on submitting it was making a call to a wrong service
- Network issue
- Issue with different backend models

For the second kind of issue where users complained that they never got back any response it could be because

- User specified the issue badly
- Users specified the issue correctly but there was a genuine problem with the classification accuracy for the user stated problem

Hence there was a strong need for monitoring. To enable monitoring we enabled Logging at different levels. At end user interface level user raised ticket description, system returned classes, Time for system to respond were logged. At the classification level the ticket description, the classes returned by the system as well as response time were logged. Between these two logs we could easily pinpoint the cause of the different issues.

A continuous monitoring dash board was built to monitor parameters at different granularity levels. These included Number of tickets raised in a day, Number of tickets re-assigned, Number of Tickets open, Function wise view of metrics, etc. A script was also written which ensured that if the performance of the system was below some threshold a restart was done.

11 RESULTS

We will track the journey we went through before we could get a reasonably stable system. There were multiple releases made before stabilization. It will cover the issues faced in each release and improvements made.

11.1 Internal Release

(Baseline Internal Release / Not released for production)

Initially when we directly used our designed classifier system the maximum accuracy which we could get was ~ 60%. In many classes the accuracy was less than 30%. Since the solution had to support classes from different functions the quality of data and challenges with respect with data was different with respect to each of the functions. A detailed function wise analysis of data was done. In parallel for every function a study was undertaken as to the best data level techniques which can be applied for that kind of data. A benchmarking was also done for the most suitable algorithm for that data. Hence the correct combination of the data processing techniques and the machine learning algorithm was selected. After multiple iterations for each of the different functions a significant improvement in accuracy was seen. In testing case we could get accuracy in excess of 80%. A sample metric is attached for a representative sub-function in Table 2.

On analyzing the metrics we see that, post the different machine learning steps data has become more balanced. The number of classes has also significantly got reduced. This metric was felt good enough and we made our first production release.

11.2 First Production Release

Within a day of our release we start getting complaints about the system. When we made our release we had not planned for any logging or monitoring. There was no way for a user to give us feedback directly. Whatever feedback we got was through informal channels. These included company social groups and escalation channels where employees directly complained to the top management. Adding to the issue was when a complaint came we could not pinpoint the issue due to lack of logging.

An immediate action which we undertook was to enable logging. We also debugged issues statically. On analyzing the complaint data we could spot the trend that user outside the company network were regularly unable to reach the service. On debugging this we found that calls were being made to a service which was not accessible outside the company network. This was fixed and a release was made.

11.2 Second Production Release

At this stage we had logging enabled in the system. There was still lot of complaints which were coming to us through the informal channels.

Two top complaints related to the system were

- System does not respond
- System does not give the correct classification

On monitoring the logs we figured that system did not respond because the network was slow at times. To handle this default classes were returned to users from front end if back end did not respond within a certain time.

Further analysis of the logs brought out the challenge of wrong configuration in front end and load balancer. All calls were by default routed to only one back end service. As a result the response was slow. This issue was fixed.

For the System not giving correct classes we could narrow the issue to two cases. The first was that the end user was not entering proper information to describe their issue. They were using meaningless statements like I have an issue or generic statements like Laptop issue. To resolve this feedback was given in user interface to properly describe their problem.

There were quite a few cases where the description was correct and the system did not give a proper response. We had to figure a way to handle this.

11.3 Third Production Release

By this time users were well educated and all the basic issues with the system were sorted out. There were still many complaints from end users that they were not getting a correct classification. However till date we did not have a way to track the accuracy of the system. We created and deployed the monitoring system. Using this we could get a view of the classification accuracy and performance of the system across the different functions. Using this metric we went back to each of the functions to collectively improve the classification accuracy of the system.

Table 2 : Pre & post data clean-up metric

For 1 subset of services	
Initial Data composition	
Training Data	100000
Number of Classes	500
Average Tickets per class	200
Percent of classes contributing to 80% of Tickets	5%
Overall Accuracy	55%
Post cleanup data composition	
Training Data	100000
Number of Classes	100
Average Tickets per class	1000
Percent of classes contributing to 80% of Tickets	75%
Overall Accuracy	80%

One finding of deploying the monitoring solution was that the accuracy numbers in live use was much less than the accuracy which we got in our internal testing. This could be attributed to the multiple dependencies in a live deployment scenario. There can be issues with the

- End user interface going down
- Network going down
- End user's system having issues
- End user not knowing how to describe his problem

11.4 Further Production Releases

Post the monitoring system release, multiple regular releases were made to improve the accuracy of the system. Fixes include

- Domain specific keywords to improve classification accuracy
- Additional data
- Class consolidation
- More rigorous automated and manual cleaning of the training data

11.5 Current State

Use of this system has given > 50% reduction in transfer index as compared to manual process. There is a continuous improvement in classification accuracy. At outset many classes had less than < 30% accuracy. Through the application of the different techniques as mentioned in previous section > 80% accuracy has been achieved across all areas. In some classes we have seen an accuracy of > 95%.

12 CONCLUSION

It is a daunting task to create a working prediction system with good level of accuracy and performance for a large scale live system while supporting diverse types of predictions. Text classification in real life is a challenging issue. It requires a deep understanding of the data and kind of data cleansing algorithms which are required to handling these data types. It requires a close coordination with the different data sources to understand the data better so that the appropriate data techniques can be short listed, applied and the result verified.

Key challenges are due to the data which brings in issues like volume, wrong mapping, insufficient data, and too many close classes. As a result a direct supervised classification approach will not work. Specific techniques would need to be used for cleaning the data, handling the wrongly mapped data, handling too many classes and handling the issue of too less data. A combination of techniques have to be used for cleaning including domain specific stop words and domain specific NER's to identify unwanted data. Clustering techniques have to be used to reduce the number of classes. Concept extraction techniques have to be used to enhance the quality of classification.

It requires a good understanding of the different possible machine learning algorithms which can be applied. The system should be designed in a flexible manner so that the appropriate algorithms can be easily applied or removed. A continuous learning of the system is required so that it can learn from use and be flexible to understand new classes which are not there in the initial training set.

A reasonably good accuracy has been obtained using the mentioned techniques and the system is being used actively by all the employees of the company. To further improve the accuracy additional techniques would need to be used including building a domain specific ontology and domain specific relationship / key concept extractor.

Building such a system has given the needed productivity improvements since the percentage of re-assigned tickets has gone down by greater than 50%. There is an improved End user experience since users no longer need to bother about navigating to the right selection and can just type their problem right away. There is a quicker turnaround time due to tickets falling to correct bucket and being allocated to the correct agent.

13 FUTURE WORK

Different approaches will have to be explored to ensure accurate classification even in face of the data issues as listed in paper. These would include semantic techniques and shallow semantic parsing to extract the relations. A domain specific ontology could be used to ensure only relevant and clean tickets will be used for training. Alternate techniques will have to be explored for the data cleaning as well as the supervised prediction approach. The self-learning loop has to be automated by finding a solution to work around users under specifying the problem and agents closing tickets under wrong closure category. There is also a need for improving the monitoring tools for this system.

REFERENCES

[1] "Hierarchical Sampling for Active Learning" by Sanjoy Dasgupta, Daniel Hsu. Proceeding
ICML '08 Proceedings of the 25th international conference on Machine learning
Pages 208-215

[2] Chap 12 of book "Large Scale Machine Learning" by Ullman

[3] "New Boosting Algorithms for Classification Problems with Large Number of Classes Applied to a Handwritten Word Recognition Task" by Simon Gunter and Horst Bunke. Multiple Classifier Systems. doi:10.1007/3-540-44938-8_33

[4] Introduction to Semi-Supervised Learning by Xiaojin Zhu and Andrew B. Goldberg From the SYNTHESIS LECTURES ON ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING #6, 2009

[5] Adaptive Synthetic Sampling Approach for Imbalanced Learning by Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li.

[6] Active Learning for Large Multi-class Problems by Prateek Jain (Univ of Texas) and Ashish Kapoor (Microsoft Research). IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2009

[7] Handling imbalanced data – A review by Sotiris Kotsiantis and et al. GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006

[8] Towards Auto-remediation in Services Delivery: Context-Based Classification of Noisy and Unstructured Tickets by Gargi Dasgupta, Tapan Kumar Nayak, Arjun R. Akula, Shivali Agarwal, Shripad J. Nadgowda. Service-Oriented Computing - 12th International Conference, ICSOC 2014, Paris, France, November 3-6, 2014.

Proceedings. Volume 8831 of Lecture Notes in Computer Science, pages 478-485, Springer, 2014

[9] Data Algorithms for Processing and Analysis of Unstructured Text Documents by Artem Borodkin, Evgeny Lisin, Wadim Strielkowski Applied Mathematical Sciences, Vol. 8, 2014, no. 25, 1213 - 1222

[10] Classification of Historical Notary Acts with Noisy Labels by Julia Efremova, Alejandro Montes García, Toon Calders in Advances in Information Retrieval. Lecture Notes in Computer Science Volume 9022, 2015, pp 49-54

[11] Classifying Unstructured Text Using Structured Training Instances and an Ensemble of Classifiers Andreas Lianos, Yanyan Yang. Journal of Intelligent Learning Systems and Applications, 2015, 7, 58-73 Published Online May 2015 in SciRes. http://www.scirp.org/journal/jilsa http://dx.doi.org/10.4236/jilsa.2015.7200

[12] Feature Selection for Effective Text Classification using Semantic Information by Rahul Jain and Nitin Pise, International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 10, March 2015 18

[13] Active learning for text classification: Using the LSI Subspace Signature Model. Weizhang Zhu and Allen Data Science and Advanced Analytics (DSAA), 2014 International Conference

[14] Combining Classifiers for Text Categorization. Leah.S.Larkey and W.Bruce.Croft. SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. Pages 289-297

Predicting the Price of Second-hand Cars using Artificial Neural Networks

Saamiyah Peerun, Nushrah Henna Chummun and Sameerchand Pudaruth
University of Mauritius
Reduit, Mauritius

ABSTRACT

The number of cars on Mauritian roads has been rising consistently by 5% during the last decade. In 2014, 173 954 cars were registered at the National Transport Authority. Thus, one Mauritian in every six owns a car, most of which are second hand reconditioned cars and used cars. The aim of this study is to assess whether it is possible to predict the price of second-hand cars using artificial neural networks. Thus, data for 200 cars from different sources was gathered and fed to four different machine learning algorithms. We found that support vector machine regression produced slightly better results than using a neural network or linear regression. However, some of the predicted values are quite far away from the actual prices, especially for higher priced cars. Thus, more investigations with a larger data set are required and more experimentation with different network type and structures is still required in order to obtain better predictions.

KEYWORDS
Car price prediction, neural network, linear regression, support vector regression.

1 INTRODUCTION

According to the data obtained from the National Transport Authority (2014), there has been an increase of 254% in the number of cars from 2003 (68, 524) to 2014 (173, 954), as shown in Figure 1. We can thus infer that the sale of second-hand imported (reconditioned) cars and second-hand used cars has eventually increase given that new cars represent only a very small percentage of the total number of cars sold each year. Most individuals in Mauritius who buy new cars also want to know about the resale value of their cars after some years so that they can sell it in the used car market.

Price prediction of second-hand cars depends on numerous factors. The most important ones are manufacturing year, make, model, mileage, horsepower and country of origin. Some other factors are type and amount of fuel per usage, the type of braking system, its acceleration, the interior style, its physical state, volume of cylinders (measured in cubic centimeters), size of the car, number of doors, weight of the car, consumer reviews, paint colour and type, transmission type, whether it is a sports car, sound system, cosmic wheels, power steering, air conditioner, GPS navigator, safety index etc. In the Mauritian context, there are some special factors that are also usually considered such as who were the previous owners and whether the car has had any serious accidents.

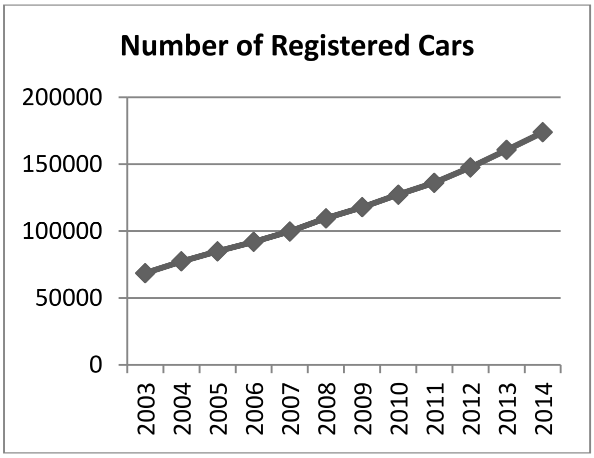


Figure 1. Number of registered cars from 2003-2014

Thus, predicting the price of second-hand cars is a very laudable enterprise. In this paper, we will assess whether neural networks can be used to accurately predict the price of second-hand cars. The results will also be compared with other methods like linear regression and support vector regression.

This paper proceeds as follows. In Section II, various works on neural networks and price prediction have been summarised. The methodology and data collection are described in section III. Section IV presents the results for price prediction of second-hand cars. Finally, we end the paper with a conclusion and some ideas towards future works.

2 RELATED WORKS

Predicting the price of second-hand cards has not received much attention from academia despite its huge importance for the society. Bharambe and Dharmadhikari (2015) used artificial neural networks (ANN) to analyse the stock market and predict market behaviour. They claimed that their proposed approach is more accurate than existing ones by 25%.

Pudaruth (2014) used four different supervised machine learning techniques namely kNN (k-Nearest Neighbour), Naïve Bayes, linear regression and decision trees to predict the price of second-hand cars. The best result was obtained using kNN which had a mean error of 27000 rupees.

Jassbi *et al.* (2011) used two different neural networks and regression methods to predict the thickness of paint coatings on cars. The error for the final thickness of the paint was found to be 2/99 microns for neural networks and 17/86 for regression. Ahangar *et al.* (2010) also compared the use of neural networks with linear regression in order to predict the stock prices of companies in Iran. They also found that neural networks had superior performance both in terms of accuracy and speed compared to linear regression.

Listiani (2009) used support vector machines (SVM) to predict the price of leased cars. They showed that SVM performed better than simple linear regression and multivariate regression. Iseri and Karlik (2009) used neural networks to predict the price of automobiles and achieved a mean square error of 8% compared with 14.4% for regression.

Yeo (2009) used neural networks to predict the retention rate for policy holders of automobile insurance. The neural network was able to predict which customers were likely to renew their policy and which ones would terminate soon. Doganis *et al.* (2006) used artificial neural networks and genetic algorithm in order to predict the sales of fresh milk with an accuracy of 95.4%. Rose (2003) used neural networks to predict the production of cars for different manufacturers.

Thus, we have seen that neural networks have been used successfully for predicting the price of various commodities. Our objective, therefore, in this work, is to use neural networks in a new application, i.e., that of predicting the price of second-hand cars.

3 METHODOLOGY

In order to carry out this study, data have been obtained from different car websites and from the *small adverts* sections found in daily newspapers like L'Express and Le Defi. The data was collected in less than one month interval (i.e. in the month of August in 2014) because like other goods, the price of cars also changes with time. Two hundred records were collected.

The data comprises of different features for second-hand cars such as the year (YEAR) in which it was manufactured, the make (MAKE), engine capacity (ENGINE) measured in cubic centimetres, paint (PAINT) type (normal or metallic), transmission (T/N) type (manual or automatic), mileage (MILEAGE) (number of kilometres the car has been driven) and its price (PRICE) in Mauritian rupees.

Table 1. Snapshot of the car dataset

YEAR	MAKE	ENGINE	PAINT	T/N	MILEAGE	PRICE
2008	Toyota	1400	1	1	70000	315000
2008	Toyota	1400	1	1	70000	315000

2008	Ford	1400	1	0	75000	300000
2005	Mazda	1400	1	1	80000	290000
2009	Citroen	1400	1	0	13000	285000
2005	Mazda	1400	1	1	80000	285000
2006	Renault	1400	0	0	151000	260000
2005	Opel	1400	0	0	125000	230000
2006	Toyota	1400	1	1	75000	230000
2004	Ford	1400	0	0	204000	230000
2004	Toyota	1400	1	0	105000	225000

Table 1 shows eleven records selected from our dataset of 200 records. The range for the year attribute was 2000-2012. A total of fifteen make was studied. Chevrolet and Peugeot had only 3 instances while Toyota has 63 instances. The smallest horsepower in the dataset was 900 and the highest one was 2900. For paint type, 0 stood for normal paint while 1 stood for metallic. For transmission type, a value of 0 means manual transmission while a value of 1 means automatic transmission. The lowest mileage recorded was 2000 km and the highest one was 275,000 km while the price range was 110000 to 685000.

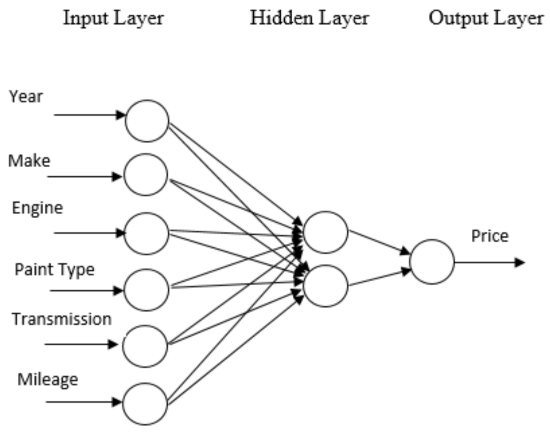


Figure 2. Neural Network Architecture

A neural network with six inputs and one hidden layer containing two nodes was used to predict the price of second-hand cars.

4 EXPERIMENTS AND RESULTS

A large number of experiments have been conducted in order to find the best network structure and the best parameters for the neural network. We found that a neural network with 1 hidden layer and 2 nodes produced the smallest mean absolute error among various neural network structures that were experimented with. However, we found that Support Vector Regression and a multi-layer perceptron with back-propagation produced slightly better predictions than linear regression while the k-Nearest Neighbour algorithm had the worst accuracy among these four approaches. All experiments were performed with a cross-validation value of 10 folds. The results are summarised in Table 2 below.

Table 2. Mean Absolute Errors

Machine Learning Algorithm	**Mean Absolute Error (Rupees)**
Support Vector Regression	30605
Linear Regression	30828
k-Nearest Neighbour (kNN)	42240
MLP, 500 cycles, learning rate = 0.05	30746

Pudaruth (2014) used only 97 records, 3 make and only 3 features and obtained a mean absolute error of 27000 with kNN (*k*=1). However, in this work, we have used 200 records with 6 inputs and experimented with more complex approaches. Although, the mean absolute error was slightly higher in our experiments, the value of 30605 can be considered to be a satisfactory outcome as the mean price of the cars was found to be Rs 311586, which is less than 10%. The actual price values obtained from the different sources have been assumed to reflect the true value of the cars but we should point out that these values are estimated by car owners who often do not have much experience in the car business.

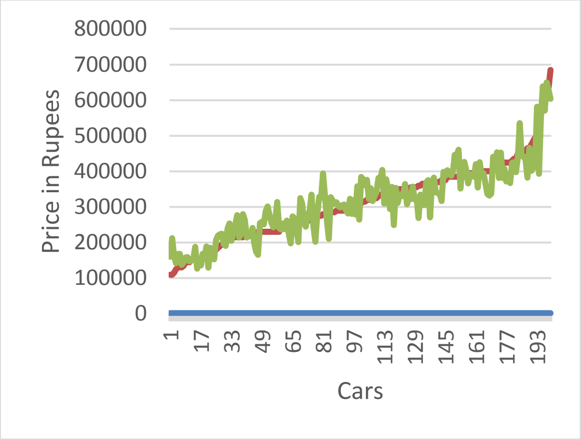


Figure 3. Actual Price v/s Predicted Price using MLP

Figure 3 above shows the variation in actual price (red line) against the variation in the predicted predicted price (green line) using a multi-layer perceptron. The graph shows that as the price gets higher, the deviation from the actual price also increases by a small amount. Nevertheless, the graph also shows that the predictions are also fairly accurate and can relied upon in many cases.

5 CONCLUSION

The aim of this paper was to predict the price of second-hand reconditioned and second-hand used cars in Mauritius. The car market has been increasing steadily by around 5% for the last ten years, showing the high demand for cars by the Mauritian population. There are hundreds of car websites in Mauritius but none of them provide such a facility to predict the price of used cars based on their attributes. Our dataset of 200 records was used with the cross-validation technique with ten folds. The car make, year manufactured, paint type, transmission type, engine capacity and mileage have been used to predict the price of second-hand cars using four different machine learning algorithms. The average residual value was reasonably low for all four approaches. Thus, we conclude that predicting the price of second-hand cars is a very risky enterprise but which is feasible. This system will be very useful to car dealers and car owners who need to assess the value of their cars. In the future, we intend to collect more data and more features and to use a larger variety of machine learning algorithms to do the prediction.

REFERENCES

[1] NATIONAL TRANSPORT AUTHORITY. 2015. Available at: http://nta.govmu.org/English/Statistics/Pages/Archives.aspx. [Accessed 24 April 2015].

[2] Bharambe, M. M. P., and Dharmadhikari, S. C. (2015) "Stock Market Analysis Based on Artificial Neural Network with Big data". *Fourth Post Graduate Conference, 24-25th March 2015, Pune, India.*

[3] Pudaruth, S. (2014) "Predicting the Price of Used Cars using Machine Learning Techniques". *International Journal of Information & Computation Technology,* Vol. 4, No. 7, pp.753-764.

[4] Jassibi, J., Alborzi, M. and Ghoreshi, F. (2011) "Car Paint Thickness Control using Artificial Neural Network and Regression Method". *Journal of Industrial Engineering International*, Vol. 7, No. 14, pp. 1-6, November 2010

[5] Ahangar, R. G., Mahmood and Y., Hassen P.M. (2010) "The Comparison of Methods, Artificial Neural Network with Linear Regression using Specific Variables for Prediction Stock Prices in Tehran Stock Exchange". *International Journal of Computer Science and Information Security*, Vol.7, No. 2, pp. 38-46.

[6] Listiani, M. (2009) "*Support Vector Regression Analysis for Price Prediction in a Car Leasing Application"*. Thesis (MSc). Hamburg University of Technology.

[7] Iseri, A. and Karlik, B. (2009) "An Artificial Neural Network Approach on Automobile Pricing". *Expert Systems with Application: ScienceDirect Journal of Informatics*, Vol. 36, pp. 155-2160, March 2009.

[8] Yeo, C. A. (2009) "Neural Networks for Automobile Insurance Pricing". *Encyclopedia of Information Science and Technology, 2nd Edition*, pp. 2794-2800, Australia.

[9] Doganis, P., Alexandridis, A., Patrinos, P. and Sarimveis, H. (2006) "Time Series Sales Forecasting for Short Shelf-life Food Products Based on Artificial Neural Networks and Evolutionary Computing". *Journal of Food Engineering*, Vol. 75, pp. 196–204.

[10] Rose, D. (2003) “Predicting Car Production using a Neural Network Technical Paper- Vetronics (In-house)*”*. Thesis, U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC).

[11] LEXPRESS.MU ONLINE. 2014. [Online] Available at: http://www.lexpress.mu/ [Accessed 23 September 2014].

[12] LE DEFI MEDIA GROUP. 2014. [Online] Available at: http://www.defimedia.info/ [Accessed 23 September 2014].

[13] He, Q. (1999) “*Neural Network and its Application in IR”*. Thesis (BSc). University of Illinois.

[14] Cheng, B. and Titterington, D. M. (1994). “Neural Networks: A Review from a Statistical Perspective”. *Statistical Science*, Vol. 9, pp. 2-54.

[15] Anyaeche, C. O. (2013). “Predicting Performance Measures using Linear Regression and Neural Network: A Comparison”. *African Journal of Engineering Research*, Vol. 1, No. 3, pp. 84-89.

SOREST, A Novel Framework Combining SOAP and REST for Implementing Web Services

Roopesh Kevin Sungkur
Computer Science and Engineering Dept
University of Mauritius, Reduit, Mauritius

Sachin Daiboo
Computer Science and Engineering Dept
University of Mauritius, Reduit, Mauritius

ABSTRACT

The two leading technologies being used to implement Web services are SOAP and REST. SOAP-based services have been widely used for interfacing software applications within and across boundaries of organisations. However, lately the idea behind REST has been extensively investigated for creating Web services and it is becoming a challenge to the adoption of SOAP as the technology of choice for implementing Web services. Therefore, it has become imperative to investigate and compare the effectiveness of the two types of implementations in terms of their relative performance. This will help developers and researches in making the right choice of technology for implementing the required functionalities through web services. Also, by analysing the two approaches, it can be investigated whether and how the best features of the two can be combined to provide a more useful and improved way of implementing web services. Finally, a novel approach to implementing web services, named SOREST has been proposed which improved the performance of SOAP-based web services while at the same time making use of the benefits of both SOAP and REST.

KEYWORDS

Web Services; Web Services Architecture; SOAP; REST; Performance Testing.

1. INTRODUCTION

With the rapid advancement of internet technology, the World Wide Web is becoming more and more an area for communication between applications, over and above the actual human-to-machine interactions. One of the reasons for this change is the advent of Web Services technology. Web services have become popular in the last few years mainly because they allow interoperability between distributed and disparate systems, independent of the underlying technology, be it the hardware, operating system or the programming language used.

With businesses operating in an increasingly competitive environment, there is an ever-growing demand to deliver effective and efficient services within and across enterprise boundaries over the internet. Consequently, organisations need to use and interact with business partners often operating on heterogeneous systems. The integration costs can have high financial impact. This is where web services can help by allowing corporations to find and seamlessly integrate the services provided by other businesses within their processes as well as deliver them at lower costs to consumers and other companies [1, 2].

The benefits of web services technology (and SOA generally) have led to extensive research and development into the design and implementation of web services, resulting in several standards and specifications. Two approaches have emerged as industry standards: using SOAP protocol and REST architectural design. While some major companies like PayPal offer SOAP-based web services, others such as Twitter, Flickr and Yahoo offer RESTful web services. Others, on the other hand, provide both SOAP-based and RESTful web services such as Google and Amazon [3, 4].

Both implementations have the same common goals of allowing platform independent communication and providing powerful web services. Each technology approach also has its relative advantages and disadvantages as well as its uses. This has resulted in a challenge for software engineers as to which approach is better and more efficient to be applied for their respective needs [5].

2. LITERATURE REVIEW

A. *Web Services Architecture and SOA*

The Web Services Architecture reflects the SOA approach and has the same objectives as the latter which are to allow software components and business applications to be available through standardised interfaces, within and across networks [6, 7]. The Web Service Architecture basically consists of 3 roles: Service Provider, Service Requester and Service Broker; and 3 basic operations: publish, bind and find. These are depicted in figure 1 below.

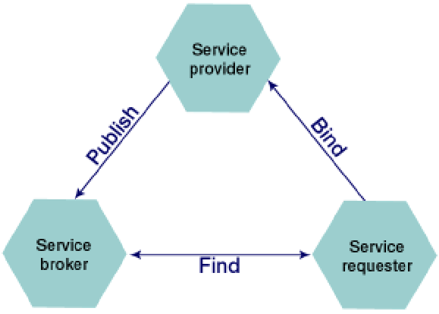


Fig 1: Web Services Architecture [8]

A **service provider** offers services and *publishes* them to a **service broker**. A **service requester** needs a service and requests for or *finds* same using the service broker. The requested service is then *bound* to the service requester [8].

B. *SOAP as a Web Services Technology*

As the need for flexibility and interoperability increased, two main methods of implementing web services emerged as industry standards, one based on the SOAP protocol and the other based on REST architectural design. The W3C specifications describe an implementation of web services based on the SOAP protocol [9]. There are four main technologies that are used in this approach:

- XML

XML has been designed to describe data and provide more flexible and adjustable ways of representing information, thereby improving the functionality of the Web. XML is called extensible as unlike HTML, it does not have a fixed format. XML is rather a meta-language that allows design of our own customised markup languages. Thus, our own elements, attributes and structures can be used to describe any data. Because of this flexibility, XML helps to easily represent and exchange complex data over the internet in a meaningful manner. XML is text-based and hence, platform independent [10, 11].

- SOAP

SOAP is a lightweight protocol designed for the exchange of information in a decentralised and distributed environment. It is no more used as an acronym. SOAP is based on XML and hence, allows the definition of an extensible messaging framework which enables messages to be exchanged between diverse parties, independent of the programming language or platform [12]. SOAP messages can be exchanged over different underlying protocols such as HTTP, SMTP and FTP. SOAP has an error structure as well which allows faults to be gracefully handled [13].

- Anatomy of a SOAP message

A SOAP message consists of an envelope which is XML based. The envelope has an optional header but a mandatory body. The header can consist of different header blocks which contain the metadata for the message while the body defines the data within the message. The header blocks can be used by intermediary nodes during the SOAP transmission while the message body is to be used by the final recipient of the message. The SOAP body can also contain a Fault element which indicates error messages [14, 15].

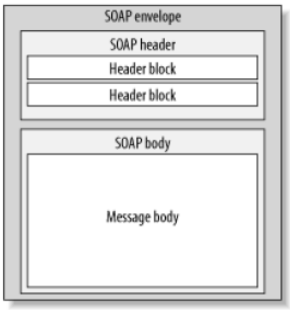


Figure 2: Soap Message Structure

C. *Serialisation and Deserialisation*

Serialisation and Deserialisation are important phases in the processing of SOAP request and response as they are based on XML. Serialisation is the process of transforming an object into a stream of data so that it is easier transmitted over a network. Deserialisation is the reverse process whereby the original object is reconstructed from the stream of data [16, 17].

The stream of data must be in an understandable format for both ends of the communication channel so that it can be reconstructed and serialised at each end [18].

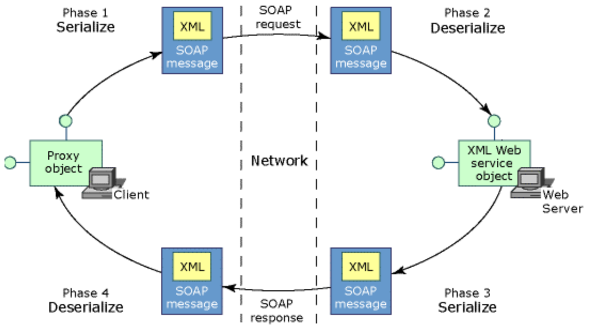


Figure 3: Serialisation and Deserialisation [18]

Figure 3 above shows the lifecycle of SOAP messages between client and server. The SOAP requests are serialised at the client's end before transmitted to the server. At the server's end, the serialised stream is deserialised, that is, reconstructed before being processed in order to extract the name of the web service and the input parameters, if any. After processing the request, the SOAP response is serialised again and sent to client. At client's end, the message is deserialised in order for the response to be processed [19, 20].

D. REST as a Web Services Technology

REST uses the concept of accessing and manipulating representations of resources and makes use of the power of Internet protocols such as HTTP to obtain representations of resources in different states.

The following terms need to be understood to be able to grasp the idea behind RESTful Web services [21]:

- Resource

A resource is generally anything on the Web which can be identified by a URI .

- URI

A URI identifies a resource, for example, using the "http:" URI scheme. An example of URI can be: 'http://example.com:8042/over/there?name=ferret#nose'.

- Representation

A representation of a resource is not the resource itself but a description of the resource in the form of metadata.

In the REST architecture, every resource is identified by a unique URI

E. Related Works

I. Portal Framework by Mulligan and Gracanin [22]

The aim of the work described by Mulligan and Gracanin [22] is to evaluate the potential effectiveness of using SOAP protocol and REST architectural design in achieving the backend requirements for a data transmission component.

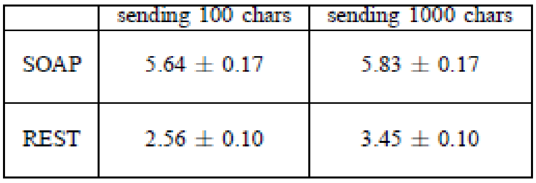
A data transmission component based on SOA was developed to maintain communication between the clients and servers. It is the component with which we are most interested. This component relays the CRUD commands from different clients over the Internet to manipulate the profiles on the servers. Hence, each service of the SOA would represent a CRUD operation being done on any of the four profiles on the server: application, device, user and user account.

The two objectives of using SOA for this component are firstly effectiveness, that is, to rapidly exchange requests between clients and servers in order not to hamper performance of the client applications and secondly, scalability, that is, the component should be able to deal with disparate number of clients per server.

From this work, it is found that it is more straightforward to implement RESTful services than SOAP-based services. A WSDL file has to be created in the latter implementation and then published. However, there is no need of service discovery in case of REST. Only the URI of the resource is needed to allow operation on it. Similarly, any HTTP client library can be used to implement REST as compared to SOAP where specialised SOAP clients are needed.

II. GA Implementation by Castillo et al [23]

Another related work has been done in [23]. A high-level comparison was done of SOAP and REST. In order to test the efficiency of both approaches to implement web services, two experiments, one with a client-server model and one with a master-slave based GA were carried out. The 2 models were implemented on Perl language, the SOAP model used the SOAP::Lite module and the REST model used the Perl Dancer module. These were chosen due to their stability. Also, servers implemented using these modules are easy to develop and deploy using the available computing infrastructure in the laboratory of the department where these experiments were done. For the first experiment a classic client-server model was developed where the client can send and receive text strings. To be able to analyse the load, different lengths of string were used. Thus, this experiment would determine how the amount of data affects the running time. Strings of length 100 and 1000 characters were used and each experiment carried out 50 times. The function 'gettimeofday' was used to obtain good precision. The results are shown below:

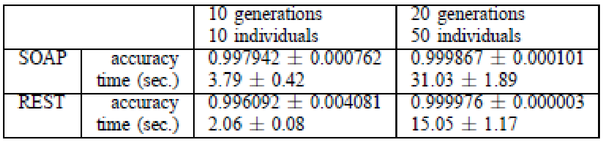


	sending 100 chars	sending 1000 chars
SOAP	5.64 ± 0.17	5.83 ± 0.17
REST	2.56 ± 0.10	3.45 ± 0.10

Figure 4: Response time as string length increases [23]

It can be seen that the SOAP implementation takes more time to send and receive the strings as compared to the REST implementation. However, there is not much difference between the time taken for string length 100 and 1000 characters for both SOAP and REST.

For the second experiment, the master-slave model was used to implement a distributed genetic algorithm. The distributed GA could not be implemented on REST technology as asynchronous processing and invocation is not supported by. Hence, the master-slave model has been used to implement the distributed GA instead. The results for this experiment are as follows:



		10 generations 10 individuals	20 generations 50 individuals
SOAP	accuracy	0.997942 ± 0.000762	0.999867 ± 0.000101
	time (sec.)	3.79 ± 0.42	31.03 ± 1.89
REST	accuracy	0.996092 ± 0.004081	0.999976 ± 0.000003
	time (sec.)	2.06 ± 0.08	15.05 ± 1.17

Figure 5: Response time as population increases [23]

Again, the REST implementation is better than the SOAP one for both configurations of load.

In both experiments, REST did better in terms of load and communication time. This can be attributed to the fact the SOAP uses XML which is verbose while for REST, no additional XML information is sent [23].

Critical Appraisal

From the results of the experiments, it is found that SOAP-based web services are more heavy-weight and take greater time to respond as compared to RESTful ones. It is the SOAP envelope and SOAP headers that increase the size of the SOAP messages. The greater the packet size, the higher the response time. The response time also increases for the SOAP implementation as in this work, XML is used for communication such that the time taken to parse the SOAP messages is higher.

Moreover, the response time for 100 and 1000 characters for both SOAP and REST messages are rather the same. This shows that load and hence, response time when strings are sent does not increase drastically when the number of characters in the string increases.

On the other hand, REST could not be used to implement a distributed GA as asynchronous processing is not supported by it. Hence, SOAP is more appropriate for asynchronous processing as compared to REST. Thus, although REST is better than SOAP in terms of latency and load, each approach has its own advantages and uses.

III. REST2SOAP

After extensive research, a framework, REST2SOAP presented in [24] was found which converts RESTful web services into SOAP web services. These can be used into BPEL service composition as BPEL support for

RESTful web services is still underway. BPEL is the standard for service composition for business processes, whereby a service is provided by combining or using the functionality provided by other web services [25, 26]. Figure 9 below provides an overview of the system architecture.

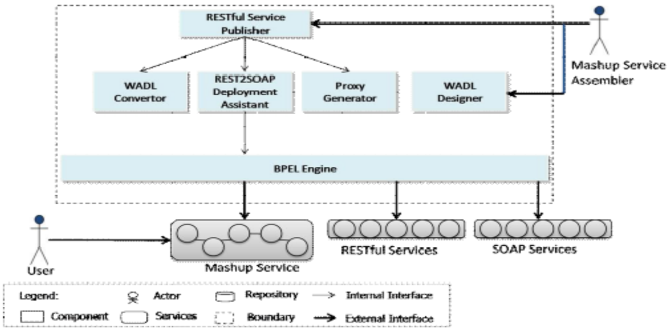


Figure 6: REST2SOAP System Architecture [24]

In this work, RESTful web service is transformed to SOAP service. However, REST2SOAP does not make use of the best features of SOAP and REST to implement an improved approach. Nevertheless, we can make use of the idea behind it to achieve our end.

3. METHODOLOGY AND PROPOSED SOLUTION

In the above section, several works have been analysed where the performance of SOAP and REST implementation of web services have been compared. In this research, the performance of these two web service technologies when applied to a DBMS system implemented using client-server architecture will be investigated. This will help to identify which technology is more appropriate to implement DBMS systems.

A. Proposed Performance Testing and Optimisation Process

The following performance testing and optimisation process based on [26] will be adopted:

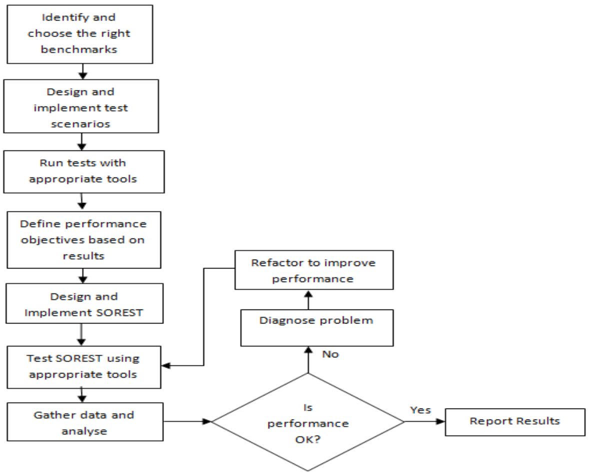

Identify and choose the right benchmarks
Design and implement test scenarios
Run tests with appropriate tools
Define performance objectives based on results
Design and Implement SOREST
Test SOREST using appropriate tools
Gather data and analyse
Is performance OK?
Yes
Report Results
No
Diagnose problem
Refactor to improve performance

Fig 7: Performance comparison plan

From the different works analysed in the Literature Review Section, different performance benchmarks used for comparing SOAP and REST based web services have been identified.

The right benchmarks which are relevant in assessing the performance of each type of web service have to be chosen. The following have been chosen to perform the performance comparison:

- Response time

In SOA, applications often need to access web services in remote environment. The performance of web services is therefore measured by how quickly the server responds to the requests from the clients. The response time determines the network latency. It also determines the network QoS provided to clients. An increase in response time induces a decrease in QoS.

- Packet size

The network over which web services are invoked is restricted by the network bandwidth. The latter is determined by the amount of the data sent and received. Hence, an important factor to measure is the packet size of the service request and response; the larger the size of the message, the greater the bandwidth utilisation. The size also affects the response time of the service.

The packet size and response time will be measured for each CRUD operation for each technology. The same tests will be performed 5 times and the average value derived.

- Load

The scalability of servers to handle load is important as well. As demand for a service increases, the processing time can increase as opposed to service suddenly stopped from being provided. Therefore, the response time will be measured as the number of requests increases. This will also determine the behaviour of the web service when load increases and thus, which type of service is better scalable.

The response time will be measured as the number of GET requests increases with each web service technology.

- Message Complexity

The response time and packet size will also be measured when the size of the request and response are increased so as to assess impact on them when message size increases.

- Database complexity

In order to determine how the web services are affected when transactions are done on large databases as in real-life situations, the size of the database will be increased to 10000 records as well.

For the latter two benchmarks, average response time and packet size will be derived over 5 transactions when GET requests are used.

Moreover, the benchmarks will be performed from different client/server perspectives in order to obtain better comparison data depicting real-life situations:

1. client on the server computer itself
2. client and server setup as LAN
3. client outside the LAN, via the Internet (WAN)

B. *Overview of Proposed System*

In order to compare the performance, we will need to develop two DBMS, one based on SOAP and one on REST. The primary aim of our project is performance comparison. Therefore, the system to be developed should at least be able to perform the CRUD operations. A tourist information system for Airports of Mauritius Limited (AML) will be developed. A conceptual design of the system to be developed is shown below:

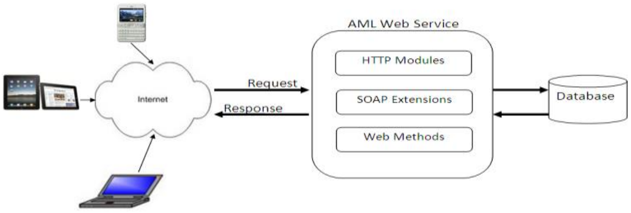


Figure 8: Conceptual design of the AML Web Service

The AML web services will be implemented with SOAP and REST. Tourists and people in general will be able to access them through different devices such as laptop, iPad and mobile phones.

C. *Objectives*

One of the main objectives of this project is devise an innovative concept to combine SOAP and REST web services. The outcome of this combination will be called SOREST.

D. *Rationale*

From the critical analysis, it was concluded that the overall performance of RESTful web services was better than that for SOAP based ones. However, both have their relative advantages and disadvantages. A larger number of functionalities, in addition to the CRUD operations, can be performed by SOAP web services. These include asynchronous operations which are not realisable with REST architecture. Also, SOAP is an effective communication protocol in relation to implementing SOA between disparate enterprise servers and applications. Thus, many web services particularly enterprise systems are already based on the SOAP protocol. Moreover, the WS* [27] specifications provide greater extensibility to the SOAP protocol such as security and reliable messaging as compared to REST. On the other hand, REST provides better response time due to the smaller size of its packets in both its request and response. It is also simpler to implement. Therefore, looking into a compromise between SOAP protocol and REST architectural design in implementing web services, which take the advantages of both techniques to provide an improved web service, is worthwhile.

Whenever we speak about SOAP and REST, we see mostly the differences between these two. Let us now look at the similarities between them. Both use XML for exchanging messages and HTTP for packet transmission. Also, both displays requested data based on collected input parameters. These similarities can be used to merge the two technologies so that the benefits provided by their differences can be put to advantage to produce a better approach to implementing web services. Furthermore, the idea behind combining SOAP and REST will provide future direction for developers wanting to merge the best features of the two.

E. Merging of SOAP and REST

The main requirement for merging the 2 technologies is to take advantage of the benefits provided by both. The added payload of SOAP messages is due to the mandatory SOAP envelope and optional headers added to the request and response which are not incorporated in REST messages. However, the packet size of SOAP request, especially for GET/POST operation, which is the most commonly used method on the web, is smaller than for SOAP response. Furthermore, using SOAP request we can take advantage of the extensibility of the SOAP protocol such as security and compression of messages. As for REST, the input parameters are sent in clear via the URL which can lead to security breaches.

The response time is greater for SOAP messages as SOAP is based on XML and since the size of the response is larger, the time taken for serialisation and deserialisation is greater. On the other hand, the response time for REST, even if based on XML is smaller than SOAP. Also, other content types apart from XML can be returned by REST. Therefore, we can return a REST response. Security features such as ‘https’ can be used for securing the latter.

Moreover, since SOAP web services are well established and are used for implementing web services in enterprise applications, existing clients will not have to break their existing connection with SOAP web services. Their web reference will almost remain the same and they will continue interacting with a SOAP web service. It is the SOAP service that will perform the necessary processing to be able to return a REST response to the requesting client afterwards. Hence, it has been decided to send a SOAP request but return a REST response to the calling SOAP client. This will in theory induce a reduction in the overall packet size during transmission as well as response time as compared to SOAP web services while enjoying the benefits of REST.

F. Software Methodology Used

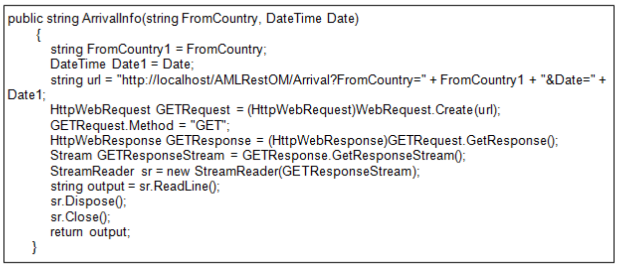
The creation of SOA needs a methodology that can quickly react and respond to changes. The Agile methodology is the most appropriate methodology that can be adopted to develop system like SOAs. In simple terms, Agile methodology means breaking down a picture into small size puzzles and when the right time arrives, reassembling the puzzle to finally form the big picture. The agile methodology will be able to produce launchable modules at the end of each tested stage to ensure that errors are identified and corrected in the development phase itself. Also, changes can be made easily without writing the whole program again which will reduce overheads and allow easy upgrading of programs. Also SOREST will be implemented in an evolving fashion meaning that changes will be done as and when required to ensure that the non-functional requirements are achieved, hence the appropriateness of this methodology

G. Implementation of SOREST

An evolving approach in the implementation of SOREST has been used until the performance improvements (decrease in packet size and response time) could be achieved.

1. A SOAP web service, similar to the AML SOAP web service, was created that contains a web method to allow client to get details of a particular plane arrival.
2. The method ‘ArrivalInfo’ is called when the SOAP web service is invoked. The former is modified as compared to how it was implemented in the SOAP project as now it will call a REST web service

instead. The new ‘ArrivalInfo’ class is now as follows:



```
public string ArrivalInfo(string FromCountry, DateTime Date)
    {
        string FromCountry1 = FromCountry;
        DateTime Date1 = Date;
        string url = "http://localhost/AMLRestOM/Arrival?FromCountry=" + FromCountry1 + "&Date=" +
Date1;
        HttpWebRequest GETRequest = (HttpWebRequest)WebRequest.Create(url);
        GETRequest.Method = "GET";
        HttpWebResponse GETResponse = (HttpWebResponse)GETRequest.GetResponse();
        Stream GETResponseStream = GETResponse.GetResponseStream();
        StreamReader sr = new StreamReader(GETResponseStream);
        string output = sr.ReadLine();
        sr.Dispose();
        sr.Close();
        return output;
    }
```

Figure 9: ‘ArrivalInfo’ method

4. RESULTS AND INTERPRETATION

F. *Performance Testing of SOAP and REST Implementation*

The most important part of our testing is to compare the performance of the SOAP-based and RESTful web services. There are several testing tools available for performing such tests. After examining the advantages and disadvantages of some of the tools, the SoapUI tool has been selected for several reasons. The main advantage with SoapUI is that it can be used to test both SOAP and REST web services. Moreover, it can be used to easily create and execute automated functional, regression and load tests. SoapUI is a free testing tool for web services and SOA. It also provides an easy to use GUI . The four methods in table below will be the prime target to compare both types of web services. As per the benchmark set, both web services will be compared in terms of average response time, average packet size and load.

Table 1: Equivalent Methods in SOAP and REST

Scenario	**SOAP**	**REST**
1	ArrivalInfo()	GetArrivalReq()
2	updateArrival()	PUTArrivalReq()
3	SaveArrival()	POSTArrivalReq ()
4	DeleteArrival()	DELETEArrival Req()

G. *Response Time of SOAP and REST*

Each method was invoked 5 times using both the SOAP and the REST web services. The average response time (RT) in milliseconds and number of bytes (B) returned were calculated. The database size and input criteria for testing were same for both implementations:

Table 2: Average RT

	Get Arrival / GET Method	Update Arrival / PUT method	Save Arrival / POST method	Delete Arrival / DELETE method
SOAP	213.2ms	201.8ms	218.6ms	190.2ms
REST	126.6ms	172.2ms	154.2ms	153ms

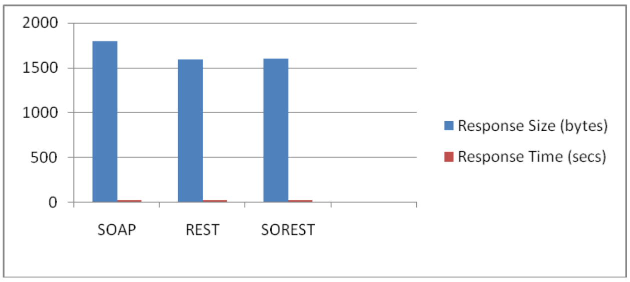


Figure 10: Response Time (SOAP and Rest)

The graph clearly demonstrates that using the REST style architecture for web services, the response time is better compared to SOAP web services.

H. *Packet Size of SOAP and REST*

The average packet size returned (bytes) by both web services was also compared and summarized in table 11.

Table 3: Packet Size (SOAP against Rest)

	Get Arrival / GET Method	Update Arrival / PUT method	Save Arrival / POST method	Delete Arrival / DELETE method
SOAP	533.8 bytes	297.4 bytes	295.4 bytes	296.4 bytes
REST	343.6 bytes	51.8 bytes	36.2 bytes	24.2 bytes

The below graph depicts the above figures obtained:

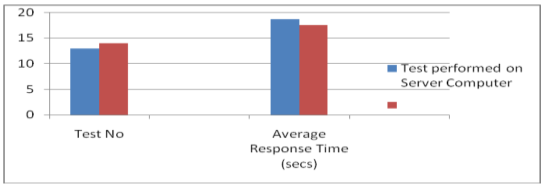


Figure 11: Packet Size (SOAP vs Rest)

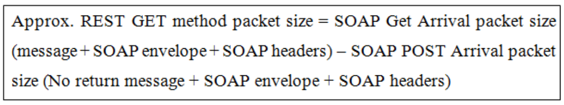
In this test case also, the average packet size of REST is smaller than that for SOAP.

I. Evaluation of Response Time and Packet Size

Compared to the other methods, the packet size of the GET method in both types of web service was larger. This is simply because using the GET method, more data were being returned to the client. For the same type of GET request, REST returned a packet size smaller than its SOAP counterpart. Apart from HTTP headers, REST does not add any overhead to the request/response message whereas, SOAP uses XML tags to create a SOAP envelope to enclose each request/response message which causes this.

The packet size and response time of the other SOAP requests are also larger compared to their REST counterparts. Interesting to note here that using the SOAP web service, the packet size for the Update, Save and Delete arrival methods were almost the same. The reason is that when performing these operations, no data is returned. The packet size returned is in fact the size of the SOAP envelope and a few SOAP or HTTP headers.

If from the number of bytes returned for a retrieval method using SOAP, we subtract the number of bytes returned by one of the other method used in SOAP, we get an answer that is close to the number of bytes returned by a GET method using the REST architecture:



Approx. REST GET method packet size = SOAP Get Arrival packet size (message + SOAP envelope + SOAP headers) – SOAP POST Arrival packet size (No return message + SOAP envelope + SOAP headers)

Figure 12: REST packet size conclusion

This calculation demonstrates that REST does not add any payload to the request/response messages which affects the performance of the web service compared when using SOAP.

J. Load Testing

SoapUI can also generate load tests in order to compare the performance of web services. This facility has been used to test both the SOAP and the REST web services. The values that will be analysed are the average time taken to process the request and the number of bytes processed per second. These will be recorded as the number of users sending retrieval requests is increased. Load Test No. 1 on GET method using 10 virtual users and every second each batch of request was sent for 1 minute. The results for each approach are as follows:

Table 4: Load Test legends

Option	Description
min	Shortest Time Taken
max	Longest Time Taken
avg	Average Time Taken (ms)
last	Time Taken for Last Test
cnt	No. of times test has been executed
tps	Transactions per second
bytes	Number of bytes processed
bps	Number of bytes processed per second
err	Number of assertion error
rat	Percentage failed request

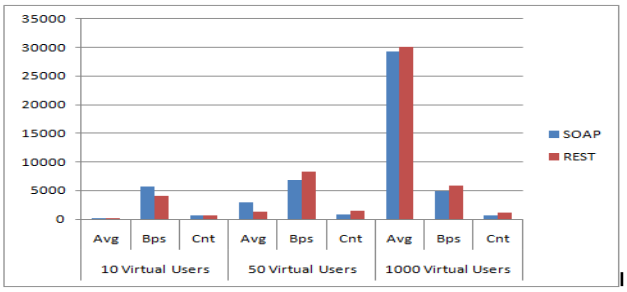


Figure 13: Evaluation of Load Testing

The REST web service performed better than the SOAP web service based on the results obtained in figure above. Although the average time taken to process a request and the number of bytes processed per second using REST was higher when the number of users increased, at the same time the number of times the request was executed (cnt) almost doubled than for the SOAP web service. This has a direct implication on the performance of the two types of web services. The REST web service is able to execute twice the number of requests as that for the SOAP web service for the same type of request over same time period. Thus, the network latency, network bandwidth utilisation and scalability of REST web services are better than for SOAP-based ones.

K. PERFORMANCE TESTING and EVALUATION OF SOREST

SOAPUI cannot be used to test the SOREST web service as it used for testing SOAP and REST services but SOREST is a combination of both. Instead, Nikhil's web development helper [27, 28] was used to perform front end testing as it is a freely available tool and allows tracing of HTTP messages [29].

I. Test Plan

The following test plan was devised so as to compare the performance of SOREST with SOAP and REST. The same test data was used. Also, the tests were performed when a style sheet was applied to the response to show the results in XML.

Moreover, to test for message and database complexity the size of the request will be increased ('FromCountry' - 30 characters) and response ('Day' – 30 characters) as well as increasing the database to store 10000 records. Five GET tests were performed in a LAN setup and the average RT and packet size compared. Only the responses without applying the style sheet will be tracked.

II. Evaluation of SOREST

The average response size and time is calculated and summarized in the table below.

Table 5: Average Response Size and Time to compare SOREST

	Test No		Average Response Size (bytes)	Averag e Response Time (secs)
Test performed on Server Computer	1	SOAP (style sheet)	8969	1.42
	2	REST (style sheet)	7804	0.54
	3	SOREST (style sheet)	7813	1.15
	4	SOAP	1714	1.22
	5	REST	1510	0.39
	6	SOREST	1519	0.54
LAN Testing (client/server) (192.168.1.12)	7	SOAP (style sheet)	8969	17.47
	8	REST (style sheet)	7804	14.16
	9	SOREST (style sheet)	7813	17.02
	10	SOAP	1714	15.92
	11	REST	1510	14.55
	12	SOREST	1519	15.02
WAN Testing (197.224.115.117)	13	SOAP (style sheet)	8969	18.76
	14	REST (style sheet)	7804	17.53
	15	SOREST (style sheet)	7813	17.88
	16	SOAP	1714	17.45
	17	REST	1510	13.26
	18	SOREST	1519	16.43

For all the different types of testing performed (on server, LAN, WAN), SOREST average response size was almost similar to REST average response as both returns the response in XML only. Compared to SOAP, SOREST average response size decreased considerably as now no additional soap envelope is being returned which is present and hence, adding additional payload to the average response size of SOAP web services.

The average response time of REST is the best as the request/response is only performed using XML. SOREST average response time is somewhere between the average response times of REST and SOAP. The reason is simply that SOREST sends requests using SOAP messages and return response in XML only.

Average response size and time for the complexity tests are summarised in table 10:

Table 6: Average Response Size and Time Complexity Test via LAN

		SOAP	REST	SOREST
Response Size (bytes)		1797	1593	1602
Response Time (secs)		17.43	14.75	16.57

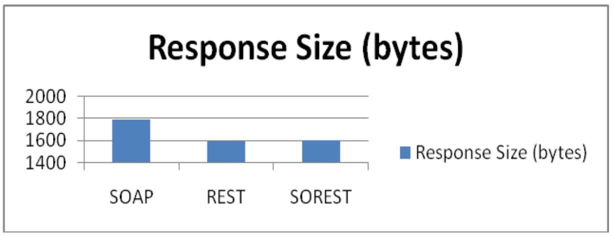


Figure 14: Average Response Size –Complexity Test

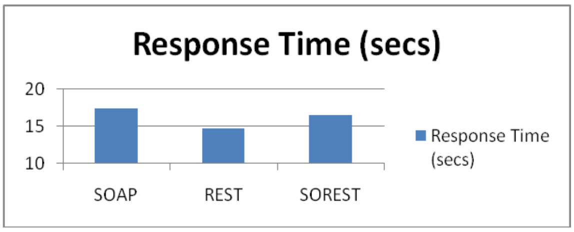


Figure 15: Average Response Time –Complexity Test

Although the complexity of request and response as well as database size were increased, the average response size and average response time of SOREST was better than for SOAP web services although they were greater as compared to REST.

We can conclude that SOREST is working as expected for the GET messages. It has improved the response time and response size compared to those for SOAP web services but of course it cannot outperform REST web service. Hence SOREST, the combination of SOAP request and REST response, will definitely improve a client experience while using a SOREST web service.

L. General Guidelines for Creating a SOREST Web Service

Guidelines are provided below to enable developers to implement SOREST:

1. Create a normal asp.net SOAP web service but make the web service one way.
2. The service should include the URL of the calling client.
3. Allow client to add web reference to the web service to able to invoke the web methods.
4. When client initiates requests, use soap extension methods at server side to capture SOAP request before the deserialise stage from client to server.
5. Extract parameters from the SOAP request.
6. Construct a REST request and send the request to a REST web service.
7. Return only the response received from the REST web server to the calling SOAP client.

5. DISCUSSION AND CONCLUSION

All the main objectives initially set have been successfully achieved as it has been demonstrated during the design, implementation and testing phase. An airport web service system was implemented using both the SOAP protocol and using the REST design architecture. Consumer applications were also created to use and test the web service implemented. Similar methods were implemented in both types of web services to provide a good benchmark of results. The relative performance of the web services have been well compared using the benchmarks identified.

From the test results, it was seen that RESTful web services were better than SOAP-based web services in terms of latency, bandwidth utilized and overall system resiliency. Moreover, the overhead caused by the SOAP envelop was clearly demonstrated in this project. However, by examining related works in this area, we also came to the conclusion that even if REST was better than SOAP, each approach to implementing web services has their own uses. Keeping this in mind, a new technique, SOREST, has been successfully devised in which a SOAP web service was merged with a REST web service in order to improve on performance of SOAP-based web services while using the benefits of both approaches. This is a new idea which we have successfully implemented and tested.

A complex system was not implemented in this project as the main aim of the project was to compare the two types of web services. If a complex system was to be implemented, it would have taken too much time and less time would have been spent in doing a comprehensive analysis and comparison of the performance of SOAP, REST and SOREST. However, additional worthwhile features could have been implemented but which are considered as future work. These consist of:

1. Security measures such as digital signature or certificates could have been implemented to ensure information is exchanged securely over the web.

2. Caching techniques could have been used, for example, using GET methods with ETags headers help in efficient caching .

3. The use of wireless technologies can be incorporated in the web service to allow passengers to have real time information, for example, about the length of a queue in a gate for boarding.

4. Compression techniques, such as ZIP algorithm to compress and decompress SOAP/XML messages could be used which would improve the performance of SOAP and SOREST based web services.

5. Extend the SOREST approach to the other CRUD operations and optimise SOREST codes.

6. Create a framework for SOREST so that it is easily adaptable for implementing web services in general.

7. Make the implementation more flexible by using dynamic binding instead of static binding for example, for extracting the input parameters, method name and creating URL of REST Web service and returning other content types for implementation of SOREST.

REFERENCES

[1] S. Graham, “Introduction,” in *Building Web Services with Java^{TM}: Making sense of XML, SOAP, WSDL and UDDI*, Indianapolis, IN 46290, Sams Publishing

[2] P.A. Castillo et al, “SOAP vs REST: Comparing a master-slave GA implementation”, Granada Univ., Dept of Architecture and Computer Technology, May 2011

[3] A. Widmann, “Why Implementing Web Services using Representational State Transfer (REST) is better than SOAP”, 2006, pp 1-4

[4] M. Champion and D. Hollander. (2004, February 10). *Web Services Glossary* [Online]. Available: http://www.w3.org/TR/ws-gloss/ [Accessed 01 May 2015]

[5] The Spiritus Temporis Web Ring Community. (2005). *Web Services, Advantages of Web Services* [Online]. Available: http://www.spiritus-temporis.com/web-service/advantages-of-web-services.html [Accessed 01 May 2015]

[6] L. F. Cabrera (2004, October). *An Introduction to the Web Services Architecture and Its Specifications* [Online]. Available at http://msdn.microsoft.com/en us/library/ms996441.aspx. version 2.0. [Accessed 01 May 2015]

[7] E. Cerami, “Introduction,” in *Web Services Essentials*, O'Reilly Media, Inc., 2002, ch. 1, pp 1-40.

[8] K. Gottschalk. (2000, September 6). *Web Services Architecture Overview* [Online]. Available: http://www.ibm.com/developerworks/webservices/library/w-ovr/ [Accessed 02 May 2015]

[9] D. Booth (2004, February 11). *Web Services Architecture* [Online]. Available: http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/ [Accessed 01 May 2015]

[10] J. Zhuk, Integration-Ready Architecture and Design, Cambridge, U.K: Univ. Press, 2004.

[11] Sun Microsystems, The Java EE5 Tutorial, U.S.A.: Sun Microsystems Inc, 2008.

[12] G. Martin et al. (2007, April 27). *SOAP Version 1.2 Part 1: Messaging Framework (2nd Ed)* [Online]. Available: http://www.w3.org/TR/soap12-part1/

[13] J.R. Erenkrantz, “Web Services: SOAP, UDDI, and Semantic Web”, Univ. California, Irvine, May, 2004

[14] Oracle. (2010). *Overview of SOAP* [Online]. Available: http://java.sun.com/developer/technicalArticles/xml/webservices/ [Accessed 01 May 2015]

[15] W3schools.com. (2011). *SOAP Envelope Element* [Online]. Available: http://www.w3schools.com/soap/soap_envelope.asp [Accessed 01 May 2015]

[16] W3schools.com. (2011). *SOAP Example* [Online]. Available: http://www.w3schools.com/SOAP/soap_example.asp [Accessed 01 May 2015]

[17] B. Hartman et al, "Web Services", in *Mastering Web Services Security*, Indianapolis, Wiley and Sons, 2003, ch. 2, pp. 48

[18] Microsoft. (2011). *XML Web Services Basics* [Online]. Available: http://msdn.microsoft.com/en-us/library/ms996507.aspx [Accessed 01 May 2015]

[19] Anshumas. (2009, Feb 11). *Serialization and De-serialization* [Online]. Available: http://www.codeproject.com/KB/aspnet/Serialization.aspx [Accessed 01 May 2015]

[20] Microsoft. (2011). *Altering the SOAP message using SOAP extensions* [Online]. Available: http://msdn.microsoft.com/en-us/library/esw638yk(v=vs.71).aspx [Accessed 04 May 2014]

[21] K. Scribner and S. Seely, "RESTful Systems: Back to the Future" in *Effective REST services via .NET, For .NET framework 3.5*, Addison-Wesley, 2009, ch. 1, 2, pp. 1-39

[22] G. Mulligan and D.Gracanin, "A comparison of Soap and REST implementations of a service based interaction independence middleware framework", *Proc. IEEE Winter Simulation Conference*, pp. 1423-1432, 2009

[23] P.A. Castillo et al, "SOAP vs REST: Comparing a master-slave GA implementation", Granada Univ., Dept of Architecture and Computer Technology, May 2011

[24] Y. Penf et al., "REST2SOAP: a framework to Integrate SOAP Services and RESTful Services," *IEEE Int. SOCA*, pp 1-4, Feb 2010.

[25] *Web Services Composition*, STI.Innsbruck, 2010.

[26] M. Mulugeta, "Performance Testing and Optimization in Web-Service Based Applications," presented at Blackboard developers Conf.

[27] Microsoft. (2007, October). *Web Services Security Specifications Index Page* [Online]. Available: http://msdn.microsoft.com/en-us/library/ms951273.aspx [Accessed 01 May 2015]

[28] Informer Tech. Inc. (2011). *Web Development Helper* [Online]. Available: http://web-development-helper.software.informer.com/ [Accessed 01 May 2015]

[29] N. Kothari. (2008). *Web Development Helper* [Online]. Available: http://projects.nikhilk.net/WebDevHelper [Accessed 04 May 2014]

Evaluating Cloud Computing Management Challenges for Non-Expert Clients

Karim Mualla *(Author)*
School of Energy, Geoscience, Infrastructure and Society
Heriot-Watt University
Edinburgh, United Kingdom
unizoul@gmail.com

David Jenkins *(co-Author)*
School of Energy, Geoscience, Infrastructure and Society
Heriot-Watt University
Edinburgh, United Kingdom
d.p.jenkins@hw.ac.uk

ABSTRACT

Non-expert managers in various types of organizations usually find it challenging to decide whether to adopt cloud computing services, which are internet-based, instead of the conventional in-house technologies, which are physically owned and controlled in-house. This paper evaluates various cloud computing management challenges that were selected and ranked by a number of non-expert managers through a decision-making survey. The outcome argues that the Urgent Support Availability aspect is selected as the most worrying factor amongst the majority of non-expert managers taking into account the associated costs, and future demand changes. In addition, multiple decision-making considerations are identified and evaluated from a non-expert management perspective in relation to performance, network reliability, integration, quality of service, and actual business benefits.

KEYWORDS

Cloud Computing, Challenges, Decision Making, ICT Management, Non-expert.

1. INTRODUCTION

A basic definition of cloud computing for non-expert clients is the use of the Internet for the tasks performed on computers. The Cloud here represents the Internet. Virtualization of processing power, storage, and networking applications via cloud computing platforms allows organizations to operate heavy demand computing resources off-premises. While this approach reduces in-house costs and energy use, recent case-studies have highlighted complexities in the decision-making process associated with implementing various models of cloud computing. This complexity is due to the rapid evolvement of these technologies without standardization of approach by today's top providers. In addition, the difficulty of understanding and predicting ICT demand growth in organizations has caused managers not to take advantage appropriately of the cloud cost-saving factor.

Non-expert managers look at cloud computing as the process of taking the services and tasks performed by computers and bringing them to the web. To a large extent this is correct. However, cloud computing technologies offer a wider range of processing, networking, and storage capabilities which assist organizations in performing many heavy or small ICT tasks at the cloud provider's datacenters. This is mostly achieved through on-demand, remotely-controlled, scalable, and pay-as-you-go approaches. In many cases, conventional in-house methods were observed not as cost-effective as current cloud-based services.

However, several challenges were observed in the long-run due to management aspects which can be summarized as follows: [1]

- Improper analysis of the organization's actual ICT requirements
- Unclear contracts with the cloud provider
- Unreliable internet performance which affects the entire cloud service-delivery process
- Security considerations and data integrity issues regarding the methods in which the cloud providers handle access to resources, store data, and secure the virtualized server infrastructure.

The above aspects can result in additional expenses and management complexities in the long-term as will be discussed in this paper. The different models and techniques of cloud computing deployments and services have a significant impact on the decision-making process in any organization's ICT environment. This paper discusses various management challenges of cloud computing, which were voted as most relevant by a number of non-expert managers through a risk-analysis survey. The outcome is analyzed against present decision-making considerations regarding the adoption of different types of cloud computing services. The study highlights non-expert clients as key users of the outcomes from this project given the diverse nature of the operational objectives in today's organizations.

The structure of this paper is divided as follows: Section 2 will introduce briefly a background on cloud computing from the perspective of non-expert clients. In Section 3, a brief literature review will be discussed regarding cloud computing challenges and end-user potential risks. Section 4 will evaluate the client-cloud computing management challenges regarding the three cloud service delivery models (IaaS, PaaS, and SaaS). In Section 5, the selected client-cloud computing management challenges will be analyzed through a decision-making risk-analysis survey which targeted 54 non-expert management-level users. Following this, the paper will analyze data results and evaluate the outcome of the survey in relation to previously stated decision-making aspects. At the end, conclusions and future works are listed.

2. BACKGROUND

Whether ICT clients realize it or not, cloud computing services are being used on a daily basis and for a long period of time. For example, internet email accounts, social networks, GPS locations, and numerous other forms of online data storage and sharing are constantly being accessed by millions of users worldwide [2]. These services are supplied by ICT providers that own virtualized datacenters for end-users to access through the Internet. In general terms, cloud-computing is a ubiquitous platform which provides on-demand ICT services through either the public Internet, or other privately-managed and secure tunneling networks like Virtual Private Networks (VPN) [3]. The Cloud concept came to life mainly because of the growing ICT requirements in almost each industry, which were not being fulfilled through previous models due to costly services and complex management procedures. However, multiple tradeoffs and challenges have risen as a result of the rapid evolvement of these technologies, while other challenges have remained from previous ICT models.

Several cloud-computing scientists and organizations have identified different characteristics, service-delivery models, architectural types, and legal aspects of a system necessary to support cloud-computing. According to the NIST definition of cloud-computing concepts, five essential characteristics were necessary: On-Demand Self Service, Broad Network Access, Resource Pooling, Rapid Elasticity, and Measured Services [4]. In addition, experts from The Cloud Security Alliance have identified a sixth cloud characteristic and named it Multi

Tenacity [5]. Furthermore, another client-cloud computing characteristic was widely discussed by many organizations is the Economy of Scale [6], which indicates the distributed manner of computing access and sharing of resources across the cloud. This characteristic is significant to this paper given the security considerations needed for non-expert clients to evaluate before signing contracts with the cloud provider.

Cloud computing hosting models were divided into four interrelated models as follows:

- Public: Cloud providers offer a full range of computing services via online means, which enables organizations to outsource the entire ICT infrastructure into the cloud.
- Private: Organizations operate either on-site, exclusively managed, or via a third-party outsourced cloud, or a combination of both.
- Community: Multiple organizations with similar operational goals and security policies, share the same virtual ICT services and platform, which can be managed by one of the above, a third-party, or a combination of all.
- Hybrid: Often the most preferable cloud deployment method for end-users, as it ensures additional management flexibilities regarding security, risk elimination, information systems portability, and better standardization. The hybrid solution offers a mixture of various sub-components from previous deployment approaches. In particular, this model irrespectively combines the technical and nontechnical aspects from Private, Public and Community models [7].

Moreover, client-cloud computing potential benefits extend beyond obtaining cost reductions and management flexibility. On this note, multiple energy saving characteristics were pointed out by academics and service providers given that ICT virtualization can have a significant potential for eliminating plugged-in equipment, thus minimizing associated electricity consumption, space and management. These Green characteristics are summarized as follows: [8] [9].

- Dynamic Provisioning: The ability to reduce unwanted cloud computing components through better matching of server capacity with actual clients' demand.
- Multi-Tenancy: The ability to normalize and flatten unmeasured peak loads by serving large numbers of clients on a shared hosting infrastructure.
- Server Utilization: The ability to operate servers at higher utilization rates via virtualization techniques.
- Data Center Efficiency: The ability to use advanced datacenter features which reduce the overall power loss through improved methods of power conditioning, air cooling, and other methods.

3. LITERATURE REVIEW

According to Carrenza and HP, upgrading an existing ICT system for three consecutive years is more costly than the system itself. This was studied on applying intensive cloud solutions for several large organizations across the United Kingdom [10]. As a result, these providers identified several vital security aspects related to the concept of virtualization in which non-expert users must thoroughly understand before outsourcing their critical business applications onto the cloud. On this ground, several reliability concerns were raised by clients after adopting software, platform, or infrastructure cloud services for at least a 1-year lifecycle.

One report argued that a slower pace of virtual ICT adoption is currently spreading across large organizations, simultaneously with the rapid evolution of cloud techniques [11]. These risks were argued to range from technical, management, all the way to legal aspects of ICT employment. Furthermore, numerous standards for specific industries were argued to be missing regarding optimizing the way in which cloud services are disparately purchased, supported, and governed.

One of the major issues in standardizing cloud computing is the large range of different purchase standards and technical definitions. Currently, these were estimated to reach nearly 160 different definitions around the world [12].These standards began developing in 1999 when Salesforce introduced the first online-based application [13].

A number of definitions and standards of cloud computing were published by top ICT providers such as Cisco, Microsoft and IBM. Many academics and papers stated that these cloud standards are developed inaccurately given that ICT providers usually tend to market their services in order to increase sales against other competitors [14]. This causes decision-making challenges for non-expert clients as will be discussed in the next section.

Other general assumptions were arguing for outsourcing of non-core ICT capacity into a third-party provider that owns the infrastructure. However, numerous growth-limiting barriers were explored concerning knowledge sharing and data breach risks [15]. Adopting a fully outsourced cloud computing solution is currently considered an unfavorable decision by most non-expert managers given the uncertainty of private data whereabouts and many other considerations related to less control over owned resources [16].

Other concerns regarding credibility and authenticity of cloud services were observed among managers from different organizations. According to a survey by the IDC Enterprise Panel in 2009, the following barriers were identified and rated depending on the level of concern in contrast to the acceptance percentages attained from purchasing on-demand cloud benefits (Figure 1) [17].

4. EVALUATION OF CLIENT-CLOUD MANAGEMENT CHALLENGES

With regard to the three primary service layers of cloud computing (SaaS, IaaS, and PaaS), the following table was constructed to evaluate the reliability and security challenges from the perspective of non-expert clients and in relation to each cloud layer separately (Table 1).

Many ICT providers are currently analyzing the adoption patterns in which their clients are turning towards cloud computing [18]. Their objective behind this analysis is essentially to identify the key concerns and challenges regarding why many clients are still reluctant to adopt cloud computing services for businesses and heavy ICT tasks. The following section presents a risk-analysis survey which addresses the point of view of 54 non-expert decision-makers from different specialties.

5. SURVEY-BASED EVALUATION FOR NON-EXPERT MANAGERS

This paper conducted a risk-analysis survey which targeted 54 non-expert management-level personnel form different organizations and companies. The purpose was to collect data on cloud computing tradeoffs and management risks by following the viewpoint of non-expert decision makers across different types of industries. This survey includes a single rating-scale question which offers 5 multi-choices as available answers.

The Likert approach was specifically selected for this survey given the nature of opposing opinions between different non-expert managers [19]. This was identified by this study from observing different ICT management aspects, such as the degree of concern towards the utilization of novel technologies among managers with medium-level technical background. The survey attempts to reflect the diverse attitude of these managers towards ICT

Table 1. Evaluation of Client-Cloud Computing Management Challenges regarding each Service-Layer

Cloud Model	Description	Example	Evaluations of Challenges
SaaS (Software as a Service)	Users access applications via network-hosted infrastructure like the Internet or VPN (e.g. Gmail).	Gmail, Blogger, Cisco WebEx, Flicker, Windows Live Meeting, Windows Office Live	Given that SaaS is mostly offered free of charge, or accompanied as an additional service with larger paid solution, Software is not installed on the users' servers or personal PCs. Therefore, access can occur strictly on on-demand. As a result, only confined functionalities, selected configuration, service availability issues, and limited control of programs -to underlying ICT developments- are provided by the provider following the SaaS approach.
IaaS (Platform as a Service)	Clients develop software via a fully network-hosted platform	Force.com (development platform), GoGrid, Facebook Developers	Underlying cloud solutions, in addition to several dependencies like storage, network, servers and operating systems, are not controlled by the service-requester. However, more control is available than the SaaS model, as the main IT environment in the PaaS approach is considered *Closed* or *Contained* [20]. Nevertheless, availability restrictions are still considered a trade-off for managers against the traditional physically-owned ICT infrastructure.
PaaS (Infrastructure as a Service)	The Cloud provider rents out hardware, software, networking bandwidth, processing power, or data storage via virtual, on-demand accessing policies	Amazon EC2, IBM Cloudworks, Windows Azure	Even though non-expert managers have, to some degree, the ability to control, deploy, and run user-created programs such as operating systems, privately developed software, networking components, however, the underlying cloud solution, is again, primarily managed by the cloud provider. Therefore, security access of information, user-group permissions and other administrative dependencies are all identified as management concerns.

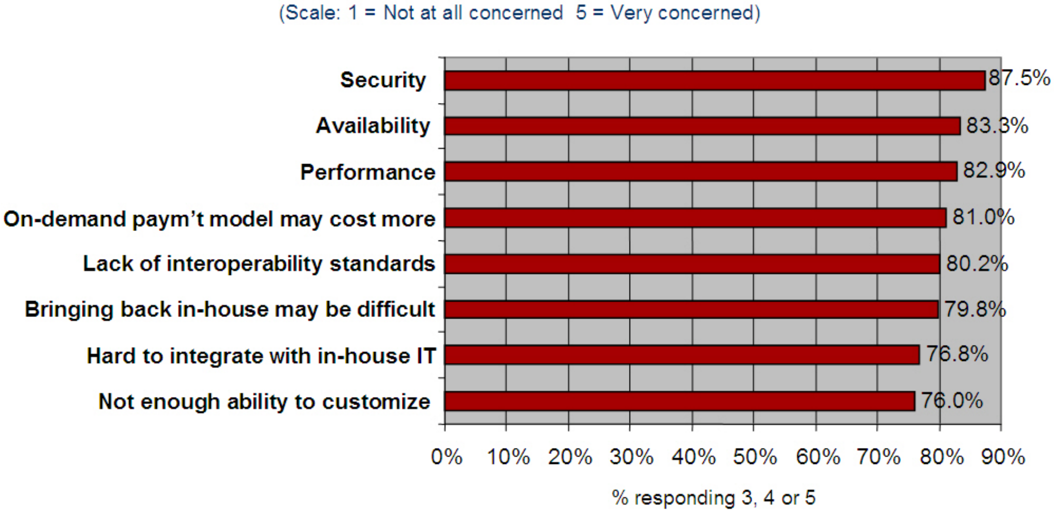


Figure 1. A 2009 Survey on Cloud Computing Management Concerns

budget acceptance, sustainability readiness, change management, and other organizational aspects [21]. The rating-scale survey was conducted via the popular online Survey provider: Survey-Monkey.
Furthermore, this survey was not structured to target a specific audience from a particular industry given that each relates to a different industry, hence, is subject to dissimilar ICT requirements.

Firstly, we asked 54 decision-makers to select the most relevant 12 risk statements of cloud computing adoption risks and challenges in their opinion. These were picked from a bigger pool of ICT risk statements and cloud adoption challenges which were listed from previous surveys and literature. We then asked the interviewees to rank those categories according to their organization's priority by selecting one of five multi-choices which reflect five levels of concern. Table 2 presents the answers and the calculated percentages of the total ranking regarding each risk category.

In addition, Figure 2 presents the completed survey findings via a bar chart, which was generated via Microsoft Excel from data inputs of Table 2.

It can be observed that the *Urgent Support Availability* aspect was classified as the most worrying factor among non-expert managers. This was demonstrated with a 4.13 average rate out of 54 participants in contrast to the rest of the statements. The *Government Hosting Regulations* came as the lowest concern and only received an average of 2.24. The two price-associated factors: *Unpredictable Costs in the Future* and *The 'on-demand' payment method of cloud-computing might actually cost more than the traditional approach*; both came at positions 3 and 4 in order. In addition, the *security* risk category landed as the second most worrying aspect following *the rapid delivery of unpredictable maintenance*.

As mentioned earlier, given that the science of cloud computing is evolving at a faster pace than most of the other services provided by various industries, it is important to identify the patterns and changes in collecting data results when performing similar surveys across time. This risk-analysis survey was intended to illustrate a relatively different viewpoint of the earlier cloud computing surveys. For instance, the IDC survey in 2009 which was discussed in the literature review section previously has covered slightly different risk categories of cloud computing. The concluded of the IDC survey results have shown obvious differences in answers in comparison to this paper's survey. For example, both *Security* and *Availability* aspects have received the highest ranking in terms of end-users' concerns. On the other hand, this paper identified the *Support* and *Unpredictable Future Costs* aspects as the highest worrying factors among managers. Moreover, while most surveys addresses operational and administrative issues of cloud computing regarding the control and access of resources, this survey has restricted the range of audience to management-level users with only a medium or low technical background.

It can be concluded from the previous survey that most non-expert managers have similar concerns when it comes to unforeseeable long-term costs, contract management issues, performance difficulties, and integration with conventional systems. This was concluded as result of reaching out to different companies which involved various ICT processes and applications. In theory, each concern was addressed based on current ICT limitations observed by these managers within their organizations. On this account, a future research work can be suggested at this point regarding the development of an automated filtering and comparison rule, which compares each of the previous cloud risk statements against the *Urgent Support Availability.* This can potentially support previous findings by highlighting the *Unpredictable Maintenance Delivery* as the most worrying aspect of different organizations' existing cloud solutions.

Table 2. Risk-Analysis Survey: Results in contrast to Cloud-Risks Statements

Level of Concern *Cloud-Risk Category*	**Not worried at all**	**Slightly worried**	**I don't mind**	**I am more worried**	**Extremely worried**	**Total Number of Participants**	**Average Rating**
Government hosting regulations	**46.30%** 25	**12.96%** 7	**18.52%** 10	**14.81%** 8	**7.41%** 4	54	2.24
Difficulties in going back to old hosting methods after using Cloud-computing	**28.85%** 15	**28.85%** 15	**26.92%** 14	**15.38%** 8	**0%** 0	52	2.29
Unknown hosting locations	**35.19%** 19	**16.67%** 9	**27.78%** 15	**16.67%** 9	**3.70%** 2	54	2.37
Integration difficulties between the Cloud and existing systems (which are already supplied by multiple vendors)	**15.38%** 8	**28.85%** 15	**26.92%** 14	**28.85%** 15	**0%** 0	52	2.69
A complete service shutdown (e.g. as a result of an internet breakdown)	**3.92%** 2	**43.14%** 22	**17.65%** 9	**15.69%** 8	**19.61%** 10	51	3.04
Contract management issues	**7.55%** 4	**32.08%** 17	**15.09%** 8	**32.08%** 17	**13.21%** 7	53	3.11
Performance issues (as everything is delivered over a network)	**1.85%** 1	**33.33%** 18	**9.26%** 5	**50%** 27	**5.56%** 3	54	3.24
Control over resources	**1.89%** 1	**28.30%** 15	**9.43%** 5	**47.17%** 25	**13.21%** 7	53	3.42
The 'on-demand' payment method of Cloud-computing might cost more than the traditional approach	**5.66%** 3	**13.21%** 7	**3.77%** 2	**56.60%** 30	**20.75%** 11	53	3.74
Unpredictable costs in the future	**3.70%** 2	**14.81%** 8	**3.70%** 2	**55.56%** 30	**22.22%** 12	54	3.78
Security (Data, access, permissions, sharing, etc.)	**1.85%** 1	**16.67%** 9	**5.56%** 3	**46.30%** 25	**29.63%** 16	54	3.85
Urgent support availability	**0%** 0	**7.41%** 4	**7.41%** 4	**50%** 27	**35.19%** 19	54	4.13

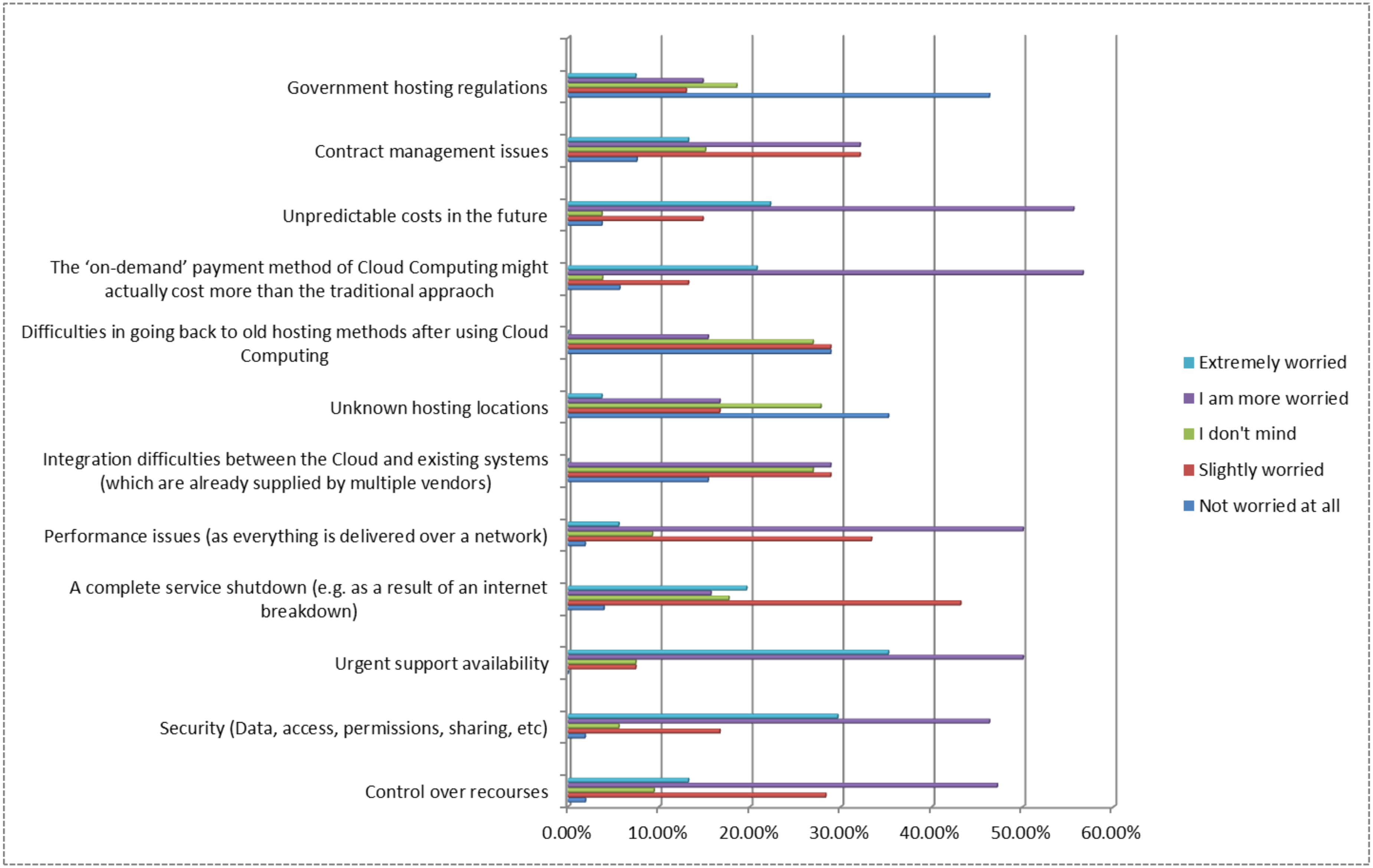


Figure 2. Survey Analysis: Microsoft Excel Representation of End-Users' Inputs

5.1 Summary of Results

The main objective of the risk-analysis survey earlier is to evaluate the level of concern of non-expert ICT clients towards cloud computing management and deployment. In conclusion, the previous collected data can be summarized in the following categories in relation to the highlighted cloud computing decision-making challenges:

Security and Privacy: According to IBM, the data security and privacy concerns rank top on almost all types of surveys [22]. On this account, cloud computing introduces another level of risk because essential services are often outsourced to a third party, which makes the management process harder to maintain data integrity and privacy, support data and service availability, and demonstrate compliance.

Actual Business Benefits: Most of today's non-expert managers are not convinced of the potential cost benefit of cloud computing. According to Netflex, in some heavy-scaling demand cases, cloud-computing can be more costly than the conventional ICT approaches [23]. This can be determined by managers in-house through a thorough identification of the organization's ICT requirements before adopting any models of cloud computing.

Managers' main concern is to realize the investment requisites to full potential, which adds value by making the cloud computing services part of their mainstream ICT portfolio. IBM argued that the return on investment (ROI) on utilizing cloud resources must be accomplished and verified by comparing certain management metrics of traditional ICT with cloud computing services. As a result, this comparison will illustrate savings on future costs, which can lead to revenue, reduction in management effort and time, compliance, and better workload assessment.

Support and Service Quality: As can be viewed in the previous survey, Service quality is one of the biggest factors that non-expert managers highlighted as a challenge against outsourcing their ICT environments and business applications onto the cloud. On this ground, if the Service Level Agreements (SLAs) provided by the cloud providers are not sufficient to guarantee the requirements for running applications on the cloud, especially related to the availability, performance and scalability, then in most cases, these non-expert users need to ensure their contracts states that the provider will cover business loss for the amount of time the service was unavailable. This is essential to any organization to take into consideration most of today's cloud contracts which usually include a limited guarantee on service quality assurances. As a result, managers are reluctant to outsource their critical business infrastructure to the service providers' cloud datacenters.

Integration: Most organizations own legacy systems which require integration with specific types of cloud computing systems if these companies decided to outsource part of their applications onto the cloud. These applications usually have complex integration requirements to interact with other cloud or in-house systems. Non-expert mangers often sense that it is often more challenging in terms of cost, effort, and time to complete any needed integration with the cloud-based systems. Therefore, in many cases these managers would rather upgrade and invest more on existing in-house technologies. On this note, a proper evaluation of the cloud contract with the provider must be thoroughly examined given that most organizations have a major requirement to integrate cloud applications with the rest of the company's systems in a quick, easily-managed, and cost-efficient manner.

Performance: Most of today's cloud business applications require intensive bandwidth and a reliable internet connection whether delivered via software, platform or infrastructure cloud solutions. Cloud computing providers usually inform clients before signing any contracts that the performance of delivering complex services through the cloud is going to be unpredictable if the network bandwidth is not reliable and adequate. Therefore, as pointed out earlier, it has been observed that the majority of non-expert managers prefer to hold off any cloud outsourcing until an improved bandwidth with lower costs is made available in their organizations.

6. CONCLUSION AND FUTURE WORK

Constructing long-term, sustainable, and cost-efficient strategies for any cloud deployment depends on the thorough identification of required services in-house and off -premises. This study points out that most of today’s heavy-burdened organizations are outsourcing these services to costly independent suppliers which causes contract limitations, unnecessary management efforts, additional costs, and complexities. These efforts are better employed by managers to enhance core competencies in their companies to maximize growth and attract new business opportunities. On this ground, this paper evaluated various management challenges of cloud computing, which were voted as most relevant by a number of non-expert managers through a risk-analysis survey. The outcome was analyzed against decision-making considerations for adopting different types of cloud computing services. The study highlighted non-expert clients as key users of the outcomes from this project given the diverse nature of the operational objectives in today’s organizations.

Future work is structured to investigate and compare each cloud computing risk category identified by this paper’s survey with the actual cost, environmental, and management benefits obtained when cloud computing services are utilized. This analysis will highlight real-life examples of management issues and barriers experienced in those organizations, which as a result will allow decision-makers to measure the actual levels of management feasibility and efficiency from adopting cloud computing services against costs and other game changing factors in their ICT infrastructure.

REFERENCES

[1] M. Almorsy, J. Grundy, A. Ibrahim, “CollaborationBased Cloud Computing Security Management Framework”, 2011.

[2] J. Rubner, “When the Sky’s the limit”. Collective Intelligence: Cloud Computing, 2011.

[3] R. Bernnat, W. Zink, N. Bieber, J. Strach, “Standardizing the Cloud: A Call to Action”. Booz and Company Inc, 2012.

[4] P. Mell, T. Grance, “The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology”, 2011.

[5] G. Brunette. R. Mogull, (2009). “Security Guidance for Critical Areas of Focus in Cloud Computing V2.1”. CSA, Cloud Security Alliance.

[6] R. Buyya. J. Broberg. A. Goscinski, “Cloud Commuting: Principles and Paradigms”. Wiley Press, New York, USA, 2011.

[7] BizCloud Corp. “Defining Cloud Deployment Models”, http://bizcloudnetwork.com/defining-cloud-deployment-models, 2010

[8] C. Mines, “4 Reasons Why Cloud Computing is Also a Green Solution”, GreenBiz, Web: http://www.greenbiz.com/blog/2011/07/27/4-reasons-why-cloud-computing-also-green-solution?page=0%2C1, 2011.

[9] A. Kofmehl, A. Levine, G. Falco, K. Schmidt, “Energy-Smart Buildings: Demonstrating how information technology can cut energy use and costs of real estate portfolios”. Accenture Corporation, 2011.

[10] Carrenza Website, Enterprise Cloud Computing for Intelligent Buildings: Service Overview Manual for Tenants, 2015.

[11] R. Bernnat, W. Zink, N. Bieber, J. Strach, “Standardizing the Cloud: A Call to Action”. Booz and Company Inc, 2012.

[12] A. Mohamed, “A History of Cloud Computing”. Web Link: http://www.computerweekly.com, 2010.

[13] J. Brodkin, “Gartner: Seven Cloud-Computing Security Risks”, InfoWorld Corporation, 2009.

[14] L. LaManna, “Top 9 Challenges in Cloud Computing”. SAP - Business innovation. Web: http://blogs.sap.com/innovation/cloud-computing/top-9-challenges-in-cloud-computing-that-are-slowing-its-adoption-011918, 2012.

[15] S.O.Kuyoro, F. Ibikunle, O. Awodele, “Cloud Computing Security Issues and Challenges”. International Journal of Computer Networks (IJCN), Volume (3), Issue (5), 2011.

[16] C. Gong, J. Liu, Q. Zhang, H. Chen, Z. Gong, “The Characteristics of Cloud Computing”. Department of

Computer Sciences, National University of Defense Technology, Changsha, China, 2010.

[17] F. Gens, “New IDC Cloud Services Survey: Top Benefits and Challenges”. IDC Exchange, 2009.

[18] H. Lim, C. Babu, S. Shivnath, J. Chase, S. Parekh, “Automated Control in Cloud Computing: Challenges and Opportunities”. ACDC, Barcelona, Spain, 2009.

[19] R. Johns, “Likert Items and Scales”. University of Strathclyde. 2010.

[20] S. Yarwood, “Intelligent Buildings: Understanding and managing the security risks”. IET: The Institution of Engineering and Technology, 2012.

[21] S. Reza, A. Adel, O. Justice, “Cloud Computing from SMES Perspective: a Survey-based Investigation”, Journal of Information Technology Management Volume XXIV, Number 1, 2013.

[22] S. Sreekanth, “Top 5 Challenges to Cloud Computing”, IBM, Web: https://www.ibm.com/developerworks/community/blogs/c2028fdc-41fe-4493-8257-33a59069fa04/entry/top_5_challenges_to_cloud_computing4?lang=en, 2011.

[23] G. Orzell, J. Becker, “Auto Scaling in the Amazon Cloud”. Netflix Inc, the Netflix Tech Blog. 2012.

Exploring the Evolutionary Change in Bollywood Lyrics over the Last Two Decades

Amina Abdul Shakoor
Department of Computer Science and Engineering
University of Mauritius
Reduit, Moka
amina.abdul@umail.uom.ac.mu

Waffiqah Bibi Sahebdin
Department of Computer Science and Engineering
University of Mauritius
Reduit, Moka
waffiqah.sahebdin@umail.uom.ac.mu

Sameerchand Pudaruth
Department of Ocean Engineering & ICT
Faculty of Ocean Studies
University of Mauritius
Reduit, Moka
s.pudaruth@uom.ac.mu

ABSTRACT

Bollywood songs have experienced a considerable change in terms of lyrics over the past decades. Long ago, bollywood songs contained mostly Hindi words, but nowadays lyricists use lots of English words to express themselves. Till date, we are not aware of any systematic quantitative analysis of how it has changed and how the evolution took place over the years. In this paper, we analysed the evolution of bollywood songs' lyrics over the past two decades. We study how the number of words and foreign words evolved over the past 20 years. The top 20 words used mostly while composing songs and the most occurring character are identified. Our dataset is composed of 300 bollywood lyrics. Based on the dataset, we show that the number of words and foreign words over the past two decades has been increasing. Our study reveals that the word "hai" (is) and the foreign word "baby" are the most widely used while writing the lyrics and the letter "a" has the highest frequency.

KEYWORDS

Data mining, Bollywood lyrics, Analysis, Change, Decade.

1 INTRODUCTION

Hindi film music, also known as "Bollywood" music, is one of the most popular forms of music in the world today [1]. The word "Bollywood" is actually a play in Hollywood, with the "B" derived from Bombay which is now known as Mumbai, the centre of film world. Bollywood forms part of one of the largest film production industry in India and has significantly influenced the Indian culture [2]. According to an estimate, around 14 million of India's population enjoys bollywood movies on a daily basis [3]. Bollywood has an amazing fan base and an incredible mass appeal. Bollywood has a great impact on the common lives of Indians week in and week out. In 1913, Dadasaheb Phalke, the father of Indian Cinema, released the first ever Indian film, Raja Harishchandra (Truthful King Harishchandra) [2]. In the 1930s, more than 200 films per year were being produced by the industry. Ardeshir Irani's Alam Ara (The Ornament of the World), released in 1931, was the first Indian sound film and it turned out to be a major commercial success. Unlike Hollywood, the initial growth in the industry was slow.

Post India's independence, according to film historians of Hindi cinema, the "Golden Age" started in the late 1940s until 1960s. This period marks the production of the Hindi films which received most criticism. Reference could be made to Guru Dutt's films Pyaasa (Thirsty) (1957) and Kaagaz Ke Phool (Paper Flowers) (1959) and Raj Kapoor's films Awaara (Tramp) (1951) and Shree 420 (Mr 420) (1955) [2].

The release of Mughal-e-Azam (The Emperor of the Mughals) in 1960 allowed Hindi cinema to move a step further [4]. The film was directed by K Asif. Featuring Dilip Kumar and Madhubala in lead roles, the film was a major hit and broke box office records in India. It became the biggest money making Bollywood film in its run at the box office. This film set the trend of romantic movies all over India.

A global growth in Bollywood's popularity was observed in 2000. This led the filmmaking of the nation to new levels due to story lines which tend to be more innovative and the technical advances in areas like animation and special effects [2]. The top 10 grossing films at the Indian box office in 2000 are Kaho Naa... Pyaar Hai (translation: Say...You Love me), Mohabbatein (Love Stories), Mission Kashmir, Josh (Frenzy), Kya Kehna (What to Say), Har Dil Jo Pyaar Karega (Every Heart that Loves), Dulhan Hum Le Jayenge (I will take the bride), Badal (Cloud), Pukar (Call) and Kurukshetra [5].

Another aspect related to Bollywood is the Filmfare Awards ceremony which is one of the most renowned film events in India. The first Filmfare Awards was held in 1954, and the best films of the year 1953 were awarded. Some of the Film Awards ceremonies which are held to honor outstanding performances are National Film Awards (1954), Filmfare Awards (1954), Screen Awards (1995) and Stardust Awards (2003) [2]. Apart from being held in India, Film Awards such as International Indian Film Academy Awards (2000) and Zee Cine Awards (1998) are held overseas.

Bombay Talkies, released on 3rd May 2013 is an Indian anthology film which consists of four short films namely Ajeeb Dastaan Hai Yeh (Strange story it is), Star (Star), Sheila Ki Jawaani (The prime youth of Sheila) and Murabba (Fruit Preserve). These short films were directed by India's four top directors- Karan Johar, Dibakar Banerjee, Zoya Akhtar and Anurag Kashyap respectively. Starring Bollywood legend Amitabh Bachan and actors like Amir Khan, Katrina Kaif and Rani Mukherji, the four films focus mainly on issues such as problems faced by a married couple, an actor's struggle and a young boy's dream about a movie star [6]. The film ends with a celebration song entitled "Apna Bombay Talkies". It contains a huge number of celebrities such as Shah Rukh Khan, Amir Khan, Madhuri Dixit, etc. defining the magic of cinema. The film release date coincided with 100th year of Indian cinema and the beginning of a new era in modern cinema. In addition, this movie was screened at the 2013 Cannes Film Festival [6].

TIMES Celebex is a monthly rating index of the Bollywood star. A score, known as the 'T Score' is calculated based on the performance and popularity of each star [7]. For instance, the top 5 ratings of actors in January 2015 are as follows: Akshay Kumar (T Score=34.0), Shah Rukh Khan (T Score=27.0), Amitabh Bachan (T Score=26.0), Arjun Kapoor (T Score=25.0), Salman Khan (T Score=21.0) [7].

Song lyrics from Bollywood movies are written by different people. Song lyrics are often targeting love; whether in happy or sad moments. They usually invoke family, mother, and self-sacrifice. In old movies, lyricists have made use of the poetic vocabulary of court Urdu [1]. Some of lyricists known for their contribution in Hindi Cinema are Gulzar, Anand Bakshi, Javed Akhtar and Mehboob. The best lyricist was first awarded in 1958. The first lyricist to be honored with this prestigious award is Shailendra for "Yeh Mera Deewanapan Hain" from the movie Yahudi (Jew) [8]. Gulzar holds the record for most awards in Best Lyricist category as he has won 11 awards from 28 nominations. Anand Bakshi has been nominated 40 times and won 4 awards. In addition, Javed Akhtar is the only one to have received the most nominations (5) in a single year (2005), where he won all of them.

The Mumbai based Hindi language film industry produces more than 800 movies every year [1]. Almost all of the Bollywood movies feature several songs that are very famous in India. They are often referred to as the heartline of Indian popular culture [1]. Actually, Bollywood songs are one of the items which are searched mostly on the Web from India. This music has influenced lives and cultures globally over decades. However, as far as we know, till date there has been no computational analysis of bollywood songs, except for a recent study that classifies Hindi songs according to moods. In this paper, we present a computational analysis of bollywood songs over a period of two decades to study the evolution.

This paper proceeds as follows. Section 2 gives an overview of works related to lyrics and songs. Section 3 describes the methodology being used. Section 4 illustrates how the lyrics corpus was created and it also involves some statistics of the corpus. Section 5 gives a summary of the results revealed from the experiments carried out. Section 6 focuses on the difficulties encountered while working and section 7 lists some interesting directions for future work.

2 RELATED WORKS

Computational analysis of bollywood songs has not received much attention by researchers. However, several studies on Music Information Retrieval (MIR), use of melodic scales in bollywood music and imagery in contemporary poetry, computational analysis of basque song collections, chorus features in popular song and corpus linguistic analysis of 50 years of Bob Dylan lyrics have been carried out in the recent decades.

A considerable amount of work has been done on the music mood classification based on audio, lyrics, social tags and all together. Automatic methods are required to classify music according to moods. This has been achieved by building a system for classifying moods of Hindi songs which makes use of audio related features such as rhythm, timber and intensity [8]. Music Information Retrieval Evaluation eXchange (MIREX) has been used during the experiment. In addition, the decision tree classifier has been employed for classification purposes and an average accuracy of 51.56% has been achieved [8].

Furthermore, the evolution of Bollywood music with respect to the use of melodic scales has been the topic under study. Through an analysis, various cultural influences on Bollywood music and its composers over the years were identified and quantified [1]. From the analysis, some striking facts were revealed about the scale usage patterns that were useful in formulating interesting conjectures which can be verified statistically.

Elements of poetic craft which include imagery, diction, emotive language, and sound devices were examined thoroughly for a quantitative analysis of style and affects in texts. From the experiment, it was observed that the frequency of concrete objects was one of the major features that could be used in detecting high quality poetry [9]. Therefore, it was concluded that imagism has an influence on contemporary professional poetry.

In recent years, cultural heritage and advances in music informatics methods has lead to a renewed interest of analysing folk song. In order to manage and understand large corpora, the ability to classify music content with respect to different properties of songs such as place name, dance type, tune family, tonality, and social function, is important. The essential elements of Basque music has been identified and analysed. A study of the evolution and origins of Basques melodies was carried out. These studies have helped in automatic classifications of songs [10].

The historical and scientific study of audio features in music, specifically the chorus, has been computationally analysed in pop songs. Choruses are known to be more prominent, more attractive and more memorable than other sections in a song, yet choruses were detected by application based on identifying the most repeated section in a song [11]. A compilation of a list of robust and interpretable features and a modelling of their influence on the ‘chorusness’ of a collection of song sections have been computationally analysed by cognitive research. This has been done through the unsupervised learning of a probabilistic graphical model. The study has shown that timbre and timbre variety are more strongly related to chorus qualities than harmony and absolute pitch height. A regression and a classification experiment have been performed to quantify these relations [11].

In 2013, the lyrics of Bob Dylan’s songs have been grouped together by D.Schmidtke, a student of PhD on the Cognitive Science of Language program at McMaster University, to form a mini corpus for the study of corpus linguistic. He put them into bins based on the years that they were released [12]. The corpus covers 50 years of lyrical material. Daniel wanted to see if there were any trends regarding Bob Dylan's lexical diversity over time. His study proved that there was no considerable change in the singer's vocabulary. He also proved that Bob Dylan's singular versus plural pronoun usage does not significantly increase with time. Daniel also studied the relationship between the word 'love' and the use of past or present tense verb. His findings suggest that Dylan prefers to talk less about love in the past tense and more about the concept in the present. It would be interesting to find out whether the preference to refer to love in the present tense over the past tense is consistent with pop music in general [12].

To the best of our knowledge, no work has been carried out on exploring the evolutionary change in Bollywood lyrics.

3 METHODOLOGY

In our present work, a Java program has been implemented for the computational analysis of bollywood songs over two decades.

First and Foremost, a corpus of bollywood lyrics consisting of both old and new songs has been created. This has been done by storing each lyric in a text file. The number of lyrics stored amounts to 300. The corpus has been divided into two parts. One corpus comprises of lyrics from 1995 to 2005 while the other consists of lyrics from 2006 to 2015. A program is written to process the text files and compute some statistics in each. For implementing the Java programs, Eclipse has been used.

Furthermore, an analysis is carried out thoroughly to find the top 20 most used words over the last two decades. In addition, character analysis is performed for the entire corpus.

The lyrics are written in Hindi Language, but we make use of its transliteration to compute the statistics. Below is an example of song lyrics "Mast Magan" in Hindi along with its transliteration [13].

इश्क़ कि धूनी रोज़ जलाये
Ishk Ki Dhooni Roz Jalaye

उठता धुँआ तो कैसे छुपाये
Uthta Dhuaa Toh Kaisein Chupaaye
ओ अंखिया करे जी हुज़ूरी
O Ankhiya Karey Jee Hajoori
मांगे है तेरी मंजूरी
Maange Hai Terii Manzoori
कजरा स्याही दिन रंग जाये
Kajra Syahi Din Rang Jaaye
तेरी कस्तूरी रैन जगाये
Teri Kasturi Rainn Jagaye

4 DATASET

There exists no database or corpus of Bollywood song lyrics. However, there are many websites where Bollywood lyrics are available. These websites have been used for the creation of lyrics corpus and reference can be made to websites such as lyricsmasti.com, hindilyrics.net, glamsham.com and lyricstaal.com. These websites also contain relevant information about the lyrics such as lyricist, movie title, composer and date of release which is very important for exploring the evolutionary change over the last two decades.

In the present task, a standard data set has been used for the computational analysis task. The lyrics used in the experiment are collected from Indian Hindi music websites. We decided to select a subset of songs for each decade. The selection was based mostly on the songs available on the websites. The final dataset of songs, that was thus created, has 300 songs composed between 1995 and 2015.

Each lyric will have the details as follows: title, release year, number of songs, number of words, number of unique words, number of foreign words, and number of unique foreign words. Based on these data, the average number of words and foreign words for each year is computed.

5 EXPERIMENTS AND RESULTS

After carrying out the experiment to study the change in number of words and foreign words over the last 20 years, the following statistics in Table 1 have been obtained.

Table 1. Statistics

Year	Number of Songs	Number of Words	Number of Foreign Words	Average Number of Words	Average Number of Foreign Words
1995	12	2536	599	211	50
1996	6	1360	335	227	56
1997	15	3675	752	251	50
1998	17	3785	769	223	45
1999	13	2913	556	224	43
2000	9	1929	460	214	51
2001	8	2179	441	272	55
2002	9	1959	401	218	45
2003	7	1551	390	222	56
2004	10	3041	826	304	83
2005	15	3570	865	238	58
2006	11	2434	671	221	61
2007	12	2243	465	187	39
2008	18	4649	1460	234	81
2009	11	2734	1003	249	89
2010	12	3610	993	301	83
2011	13	3466	967	251	74
2012	18	4119	1183	229	66
2013	30	7322	2032	244	68
2014	42	9768	2822	233	67
2015	13	2746	518	211	40

The average number of words over the first decade and second decade has been represented graphically as follows:

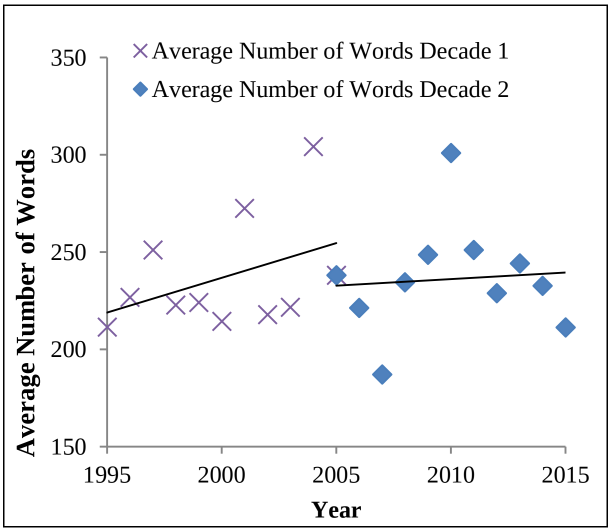


Fig. 1. Graph of Average Number of Words against Year for decade 1 (1995-2005) and decade 2 (2005-2015)

The graph represents the average number of words in songs that have been analysed over the past two decades. The linear graph for each decade shows an increase in the number of words in the Bollywood songs.

The average number of foreign words over the first decade and second decade has been represented graphically as follows:

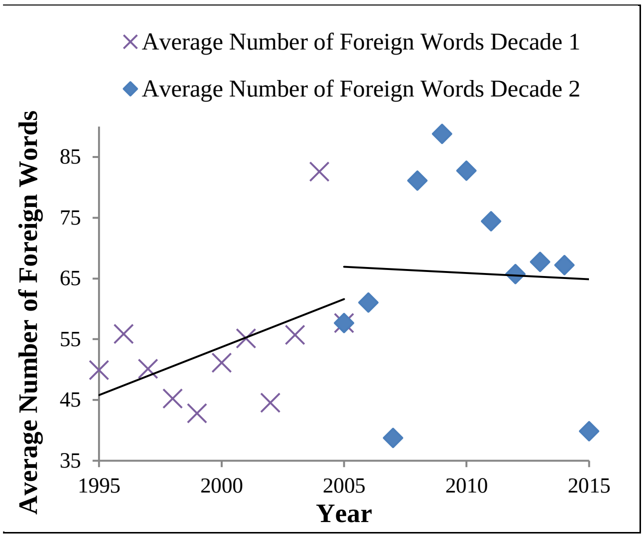


Fig. 2. Graph of Average Number of Foreign Words against Year for decade 1 (1995-2005) and decade 2 (2005-2015)

From the above graph, it can be clearly deduced that the number of foreign words found in Hindi songs has increased considerably over the years. Compared to the songs analysed over the two decades, songs in the second decade contains more foreign words than that of the first decade.

After carrying out the experiment to find the top 20 most used Hindi words while writing lyrics over the last 2 decades (1995-2015), the following results were obtained:

Table 2: Top 20 hindi words

Words	Number of occurrences	Percentage Occurrence (%)
hai	2959	18.86
mein	1115	7.11
na	942	6.00
tu	920	5.86
ho	869	5.54
dil	849	5.41

main	800	5.10
se	725	4.62
yeh	721	4.60
ke	704	4.42
ki	693	4.42
to	582	3.71
tere	561	3.58
mere	497	3.17
hain	492	3.14
hi	471	3.00
kya	466	2.97
de	458	2.92
re	432	2.75
jo	431	2.75

A graphical representation of the top 20 Hindi words used by lyricists over the last two decades is illustrated below.

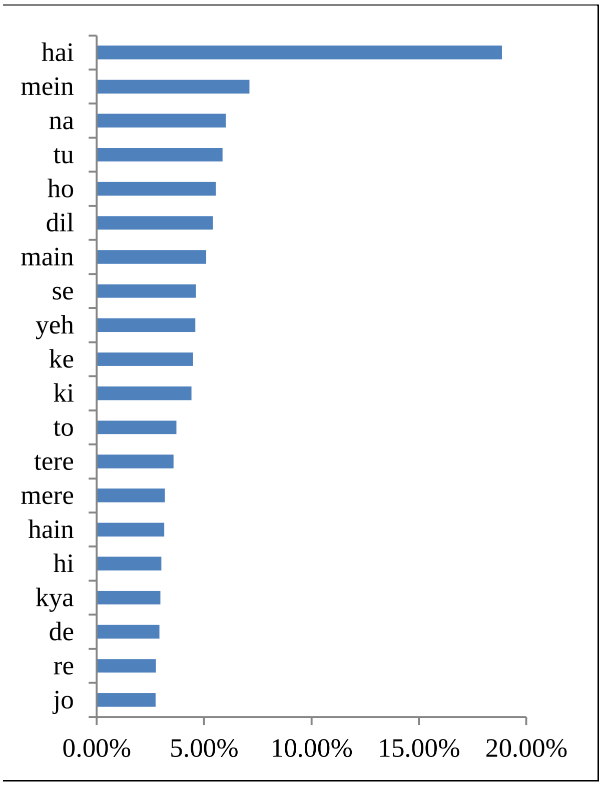


Fig. 3. Relative frequencies of top 20 Hindi words in lyrics

From the graph, it can be observed that the most commonly used verb is "be" (hai, ho, hain) and the preposition most widely used is "mein" (in). "Tu" (you) is the most used pronoun by lyricists. "Dil" is the most widely used noun while composing lyrics.

After carrying out the experiment to find the top 20 most used foreign words while writing lyrics over the last 2 decades (1995-2015), the following results were obtained:

Table 3: Top 20 foreign words

Words	Number of occurrences	Percentage Occurrence (%)
baby	110	11.20
love	108	11.0
dance	85	8.66
girl	69	7.03
twist	54	5.50
shake	49	4.99
darling	49	4.99
rock	43	4.38
floor	42	4.28
welcome	41	4.18
party	40	4.07
know	40	4.07
right	39	3.97
all	38	3.87
sunny	36	3.67
feel	34	3.46
now	31	3.16
lucky	29	2.95
touch	23	2.34
naughty	22	2.24

A graphical representation of the top 20 words used by lyricists over the last two decades is illustrated below.

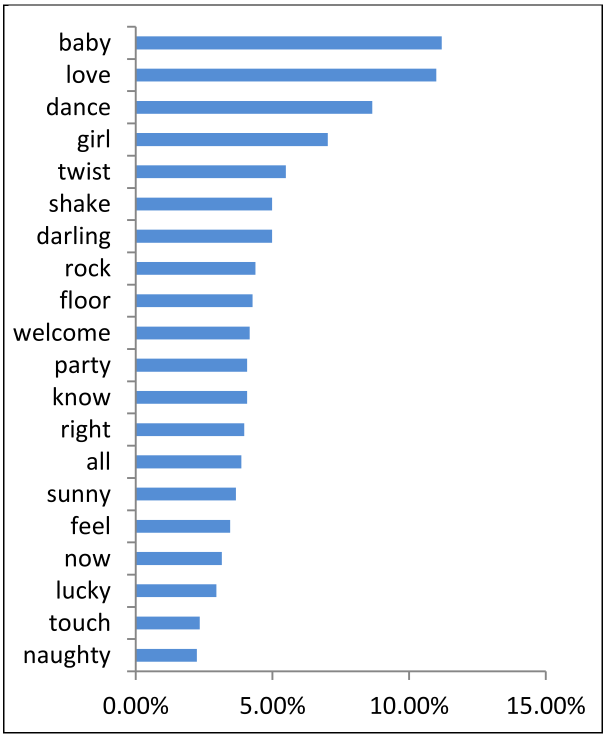


Fig. 4. Relative frequencies of top 20 foreign words in

From the graph, it can be noticed that the most commonly used foreign word is baby, followed close by love. Other popular words are dance, girl, rock, party, etc. The relative frequencies of letters in the English Language were retrieved from Pavel Mička's website [14]. This set of data was then compared with the relative frequencies of letters in the lyrics. After carrying out the analysis, the following results were obtained:

Table 4: Letter Analysis in English Language and Hindi

Letters	Frequency in Lyrics (%)	Frequency in English Language (%)
a	22.16%	8.17
h	8.95%	1.49
e	8.42%	2.78
i	8.17%	4.25
n	6.92%	12.7
o	5.12%	2.23
r	4.84%	2.02
t	4.46%	6.09
k	3.47%	6.97
m	3.43%	0.15
s	3.19%	0.77
d	3.10%	4.03
u	2.95%	2.41
y	2.89%	6.75
j	2.21%	7.51
b	1.90%	1.93
l	1.57%	0.10
g	1.57%	5.99
p	1.56%	6.33
c	0.89%	9.06
w	0.76%	2.76
f	0.45%	0.98
z	0.33%	2.36
q	0.22%	0.15
v	0.19%	1.97
x	0.00%	0.07

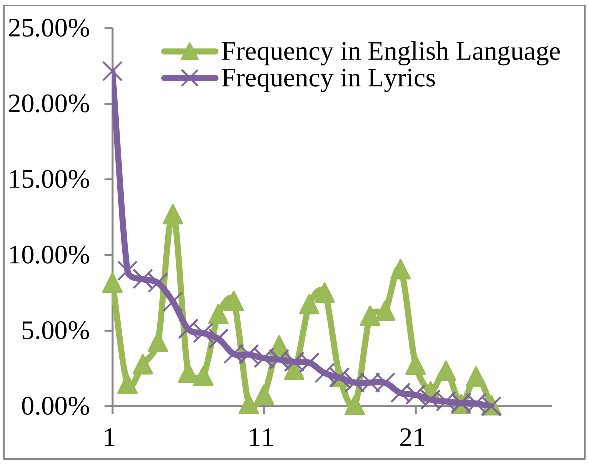


Fig. 4. Letter Frequency

Figure 4 shows the difference in frequencies between characters in the English language and in Hindi songs. There is very little correlation between these two systems.

In English Language, from most to least common: n c a j k y p t g i d e w u z o r v b h f s m q l x while in Hindi lyrics, from most to least common: a h e i n o r t k m s d u y j b l g p c w f z q v x. It can therefore be deduced that the lyricists have been making use of the letter “a” mostly over the last two decades.

6 DIFFICULTIES

One of the main difficulties we faced is the fact that some English words and the Hindi (transliterated) words are written in the same way. For instance, the word “main” (I) in Hindi and “main” (major) in English have been considered to be common during computation. Furthermore, Hindi words in the lyrics are often written in different ways. For example, “pyaar” (love) has been written as “pyar” (love) and “pyarr” (love), “bekaraar” (restless) has been written as “beqaraar” (restless).

7 CONCLUSIONS

Compared to the first decade analysed (1995-2005), the second decade (2006-2015) contains more foreign words. We may conclude that as from 2006 and up till now, the Hindi songs has experienced a change in the linguistic features. According to us, one of the main factors that contribute to this change is because of globalisation. India is more exposed to the outer world than it was in the 1990s and thus English language is now commonly spoken in India for communication. For a change, lyricists have chosen to include some English words in their songs since Indians now understand this foreign language. Our computational analysis of the Bollywood songs clearly supports this fact. Over the last two decades, the word “hai” meaning “is” has been used mostly by lyricists. After carrying out character analysis, it has been seen that the frequency of the character “a” is the highest.

One direction for future work is to analyse other musical cultures worldwide as well as regional film industries in India and extend the current dataset to be able to include more Bollywood songs. Another direction is that we can possibly work on Mauritian Sega to analyse the evolution over the years.

REFERENCES

[1] M. Choudhury, R. Bhagwan, and K. Bali. “The Use of Melodic Scales in Bollywood Music: An Empirical Study,” Microsoft Research Lab India, International Society for Music Information Retrieval, 2013.

[2] Wikipedia. (2015), *Bollywood*. [online] Available at: http://en.wikipedia.org/wiki/Bollywood [Accessed 9 Apr. 2015].

[3] N. Mishra, (2014) *Bollywood’s influence on Indian Tourism Industry*. [online] Available at: http://www.travelguru.com/travel-blog/bollywoods-influence-on-indian-tourism-industry [Accessed 14 Apr. 2015].

[4] Bollywood Tourism, (2009) *History of Bollywood*. [online] Available at: http://www.bollywoodtourism.com/bollywood-history [Accessed 13 Apr. 2015].

[5] Wikipedia, (2015). *List of Bollywood films of 2000*. [online] Available at: http://en.wikipedia.orf/wiki/List of Bollywood films of 2000 [Accessed 16 Apr. 2015].

[6] Wikipedia, (2015). *Bombay Talkies (film)*. [online] Available at: http://en.wikipedia.org/wiki/Bombay_Talkies_(film) [Accessed 12 Apr. 2015].

[7] Timescelebex.com, (2015). Top 10 *Bollywood Actors & Actresses – Times Celebex*. [online] Available at: http://timescelebex.com/ [Accessed 9 Apr. 2015].

[8] B. Patra, D. Das, and S. Bandyopadhyay, “Automatic Music Mood Classification of Hindi Songs,” [Proceedings of the 3rd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2013), IJCNLP 2013, pages 24–28, Nagoya, Japan, 2013]

[9] J. Kao and D. Jurafsky. “A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry,” Stanford University, Stanford, CA 94305, USA, 2015

[10] D. Conklin and C. Anagnostopoulou, “Comparative Pattern Analysis of Cretan Folk Song,” Journal of New Music Research, 40(2), pp.119-125, 2011

[11] J. Balen, J. Burgoyne, F. Wiering and R. Veltkamp “AN ANALYSIS OF CHORUS FEATURES IN POPULAR SONG”, Utrecht University, Department of Information and Computing Sciences, Universiteit van Amsterdam, Institute for Logic, Language and Computation, International Society for Music Information Retrieval, 2013.

[12] D. Schmidtke, (2013). *Time Out Of Mind: A corpus linguistic analysis of 50 years of Bob Dylan lyrics*. [online] Available at: https://danschmidtke.wordpress.com/2013/05/24/time-out-of-mind-a-corpus-linguistic-analysis-of-50-years-worth-of-bob-dylan-lyrics/ [Accessed 4 Apr. 2015].

[13] Music My Life, (2014). 2 States (2014) - Mast Magan|मस्त मगन |Arijit Singh |Chinmayi Sripada. [online] Available at: http://musicmylifelyrics.blogspot.com/2014/05/2-states-2014-mast-magan-arijit-singh.html#.VTYWYSGqqko [Accessed 17 Apr. 2015]

[14] Algortihmy.net, (2008). Letter frequency (English). [online] Available at: http://en.algoritmy.net/article/40379/Letter-frequency-English [Accessed 25 Apr. 2015]

The lyrics have been taken from the following websites:

1. LyricsMasti, (2015). Welcome to LyricsMasti! for Hindi Movie Songs Lyrics. [online] Available at: http://www.lyricsmasti.com [Accessed 13 Apr. 2015].
2. Glamsham.com, (2015). Movie Songs Lyrics: glamsham.com. [online] Available at: http://www.glamsham.com/music/lyrics [Accessed 13 Apr. 2015].
3. Hindilyrics.net, (2015). Hindi Lyrics: Lyrics of Hindi Songs. [online] Available at: http://www.hindilyrics.net [Accessed 13 Apr. 2015].
4. Lyricstaal, (2015). Lyricstaal | Indian Songs Lyrics. [online] Available at: http://lyricstaal.com/ [Accessed 13 Apr. 2015]

Cyber security: Threats, Vulnerabilities and Countermeasures - A Perspective on the State of Affairs in Mauritius

Tikshnayah Nelliah Maistry, Nomesh Ramkurrun, Mageshwaree Cootignan and Pierre Clarel Catherine
School of Innovative Technologies and Engineering, University of Technology, Mauritius
La Tour Koenig, Pointe-aux-Sables, Republic of Mauritius
tmaistry@umail.utm.ac.mu, nramkurrun@umail.utm.ac.mu, mcootignan@umail.utm.ac.mu and ccatherine@umail.utm.ac.mu

ABSTRACT

Recent incident analysis from CERT-MU has found that there have been an increase in cybercrime activities including unauthorised access, electronic fraud, identity theft, denial of service, spamming and fake accounts. This paper describes the most common global vulnerabilities and threats and also provide an overview of countermeasures such as encryption, back-tracing and the use of common security standards and protocols such as ISO 127K. In addition, the security issues of emerging technologies such as IPv6, Internet of Things, and Cloud Computing are investigated. The system and infrastructure of the Mauritian Cybersecurity framework is also reviewed and recommendations are provided to build a strong and resilient cyber security framework in the country. This is particularly important given that Mauritius is embarking on the ambitious project of deploying smart cities with the integration of all these emerging technologies. Security must indeed be at the centre of all considerations in this endeavour.

KEYWORDS

Cyber security, network, attack, vulnerability, threats, Internet, IPv6, IoT

1 INTRODUCTION

In April 2011, around 100 million users were attacked worldwide through a distributed denial of service (DDoS) attack. The latter allowed the attackers to access the victims’ PlayStation consoles and accounts. The company had to pay a £250,000 fine to the British regulators as it was unable to prevent its users’ data such as account names, customer addresses and bank account numbers from leaking to the internet [1]. Stories such as that one are not uncommon in the current world, largely dominated by the pervasive use of ICT. This effect is accentuated with new technologies that makes it possible for electronic devices to now be connected to the “cyberspace”. From mobile phones to video camera surveillance systems, from pacemakers to their remote monitoring systems, the “cyberspace” has become the main source of information and means of communication to most, if not all infrastructures. For this reason, cyber security has become a major priority of governments and corporations worldwide [2]. Cyber security enables the protection of assets such as building infrastructures, servers, desktop computers, information data and even lives. The objective is to protect data both at rest and during transmission [3].

However threats are becoming more aggressive with new attack methods and levels of sophistications. A correspondingly huge number of threatening issues were found on the web in 2014, and such trend is expected to continue this year and for years to come. New versions of malware and vulnerabilities suggest that the war to provide adequate level of security must be on a full-time basis [4].

The current work starts by providing an overview of common vulnerabilities and threats in section 2. Section 3 reviews the countermeasures available while section 4 provides an analysis of the growing sophistication of the cybersecurity landscape. Section 5 details some aspects of emerging technologies and the additional security concerns they raise. Sections 6 – 8 reviews the present security framework of Mauritius and provides recommendations to enhance the latter. We finally conclude with section 10.

2 VULNERABILITIES & THREATS

A vulnerability is a hole or a weakness in an application and consists of three elements: a system susceptibility or flaw, attacker access to the flaw, and attacker capability to exploit the flaw. A flaw can happen during coding, compilation or implementation of a software, also known as a bug, and allows an attacker to indirectly harm the stakeholders of the software. Stakeholders are entities that depend on the software as well as the software owner and users [5]. Most vulnerabilities, if not all, may allow a hacker to take advantage of the system by initiating an attack.

High security breaches can result in major drawbacks. In most cases, it is usually seen as the responsibility of IT managers who ignore IT systems vulnerabilities and bugs. IT security audit provides a way to reduce risks of attack, often by outsourcing the security of the system to a certified independent IT security professional or firm. The protection of the company's IT environment can then be managed properly without much input from IT managers. Penetration tests can be apply on a regular basis by performing security stress reverse test which look for and identify blind spots or the level of weaknesses and provide protection to those spots to avoid catastrophic security. Those tests are usually performed by the companies funded white hat hackers [6]. The most common software and web vulnerabilities are described in the next two subsections.

2.1 Most Common Software Vulnerabilities [7]

Some of the most common software vulnerabilities are:

a) *Injection*: Injections flaws, such as SQL (Structured Query Language), OS (Operating System) and LDAP (Lightweight Directory Access Protocol) injections happen when suspicious or infected data is unknowingly received by a user as part of request or order. The attacker's malicious data can fool the user into running unwanted requests or command codes which will enable unsanctioned data access to the attacker.

b) *Broken Authentication and Session Management*: Usually functions such as authentication and session management are wrongly implemented, providing a way for attackers to compromise credentials such as passwords and logins, keys, or session tokens or assuming another registered person's identity.

2.2 Most Recent Web Vulnerabilities & Attack

Some of the most recent web vulnerabilities and their corresponding attacks are next described:

a) "FREAK" [4]

A newly discovered vulnerability in the SSL (Secure Layer Security) and TLS (Transport Layer) cryptographic protocols could allow interception and decryption of communications of infected clients to and from servers. It facilitates man-in-the-middle (MITM) attacks against secure connections where there are unpatched or outdated version or OpenSSL this is a threat to the server which receive RSA_EXPORT cipher suites from the client with the outdated version. Further potential attacks can be launched to attack the website and steal password as well as private information.

b) "Heartbleed" [8]

Another vulnerability was recently found in the cryptographic software library of OpenSSL that may provide the attackers with access to session data such as private keys, passwords or credit card details inside a web's server memory during an encrypted session. This could provide open access to an entire encrypted data communication.

c) “ShellShock” or “Bash Bug”

ShellShock uses web servers via which attackers would use a Common Gateway Interface (CGI) which is commonly used for producing web content, to insert malware command code and infect operating system such as Linux distros, UNIX and even Mac OS X. [9].

d) Compromised Sites

Some web attack toolkits are designed to be used in the cloud, as software as a service (SaaS)

Web attack toolkits perform scans on the victims’ computers, looking for vulnerable plug-ins in order to launch the most effective attack.

Attackers are also able to control how the exploits are administered such as enabling the attacks only if a cookie has been set by the initial compromised website thereby preserving the malicious code from the prying eyes of search engines and security researchers [10].

2.3 Type of Attacks

A cyber-attack is a digital attack to a website, computer systems (individual or networked computers) or it can be any connected electronic systems these days. The attacks may compromise the confidentiality, integrity or availability of the computer or information stored on it as well as the stability of a system. Cyber-attacks take many forms, including [11]:

a) Gaining, or attempting to gain, unauthorized access to a computer system or its data.
b) Denial of service (DoS) attacks by overloading and taking down of a whole web sites. There are multiple ways to execute a DoS attack. Some of the different forms of execution include:
    a. Teardrop - sending random size data packets over the network
    b. Buffer Overflow – the server is flooded with an excess of bogus data
    c. Smurf - tricking computers to reply to a fake request, causing much traffic

2.4 Targeted Attacks

Targeted attacks are designed to target and infiltrate specific organisations globally and steal data or cause some serious damage in any particular way required by the attacker. Some of these sophisticated attacks are designed for the purpose of cyber espionage.

2.4.1 Common Attack Groups and Platforms

Some notorious groups and platforms of attacks are presented:

a) Waterbug

Waterbug is a cyberespionage group that uses sophisticated malware such as Trojan. Wipbot and Trojan.Turla to systematically target government-related infrastructures in a range of countries [12]. One example of their attacks is Heartbleed which resulted in the loss of 4.5 million healthcare records [13] [14].

b) The Regin Platform

Regin is a cyber-attack platform that can monitor GSM networks as well as some other networks. The malware can attack by collecting key logs, making screen shots, stealing files from the system, and extracting emails from Microsoft email servers and information from network data flow. As GSM infrastructures are computer controlled, attackers can compromise the controller of the Base Station, which turn control the network and launch attacks via SMS and call interception or interruption. The Regin Platform targets specific groups such as telecom operators and research institutions [15].

c) Dragonfly

Dragonfly is an ongoing cyber espionage campaign, since 2011 that has the energy sector in Europe and US as target. Their goal is to steal information and is capable of sabotage along the way. The main targets are electricity infrastructure and generation, industrial equipment providers, and petroleum pipeline operators [16].

3 COUNTERMEASURES

Depending on the kind of attacks and their resources, defenders of an information system can use different variety of countermeasures [17]. These are next described.

3.1 Type of Countermeasures

a) Training

Employees of an organization should be aware of the kinds of attacks that is likely to occur in their organization and what they should do in such cases. Proper operating procedures should be adopted to protect key attack targets such as passwords. Several studies have indicated that education is more effective than any other countermeasure for protecting information systems since for most jobs, specific knowledge of information-systems security is not a prior requirement.

b) Legal Responses

Laws prohibit all mentioned attacks. But most attackers do not fear about getting caught, as it is they are usually very difficult to track and the laws prevailing cannot be easily applied, usually because of cross-border activities of the cyber criminals. Laws should however be effective against recidivist attackers.

c) Patches

It is important to fix flaws or bugs in software as soon as they are discovered, since flaws are typically exploited very soon after discovery and attacks are made within days. Manufacturers provide "patches", "security updates", or "service packs" to fix flaws, in the form of updated versions of software.

d) Backups

Since many attacks destroy data or programs, making copies ("backups") of digital information is essential to recover from attacks and to differentiate from tampered and genuine data. Backups should to be done for any critical information, and should to be stored some distance away from the systems which are vulnerable to attacks. A backup can be an entire duplicate computer system, server room or data centre.

e) Access Controls

Automated access controls are important for the cyberspace. Access controls for computers are generally managed by passwords that must be supplied to log on and use resources. Controls can be set for individuals or for groups of people, and they can apply separately to reading, writing, or execution of resources, or to the ability to extend those privileges to other users. Access controls for networks are enforced by "firewalls", dedicated computers on a local-area network that restrict traffic to and from the network according to simple rules on such features as origin and communications protocol. Unfortunately, access controls are vulnerable to many attacks mentioned above, and will not generally protect against attacks by insiders like technical staff.

f) Encryption

Encryption hides data in some form that cannot easily be read and a "key" must be supplied to decode it when required. Any attempts to modify encrypted data will result in undecipherability, so one can tell if encrypted messages or programs have been modified. Strong and virtually unbreakable methods of encryption have been developed with "public-key cryptography", and the software for it is available for free download from a number of Web sites. Encryption methods can also be used for "authentication" or to provide digital "signatures" on documents to prove who wrote them and when. Encryption has been touted as a solution to many security problems, but is usually overrated. Some of the recent encryption cipher includes AES and TKIP.

g) Intrusion Detection and Computer Forensics

Logging records the events on a computer system or network. This can generate enormous amounts of data. As such, intrusion-detection systems (IDSs) can be set up to check and record the particular events that could indicate an attack, alerting system administrators when matters become serious. Signature checking is provided by standalone virus and worm application such as Norton Internet Security, Norton 360, McAfree, AVG, and Microsoft Security Essentials. Those anti-virus software will examine every file on a computer system to eliminate the “known” threats.

h) Intrusion Prevention Systems
Most of the methods mentioned so far just react to attacks. The alternative is an "active network defence system", which in its simpler forms is called also known as intrusion-prevention system. This includes simple things like turning off the Internet connection or logging out a user when they become sufficiently suspicious as judged by an intrusion-detection system. It can also include forms of limiting damage such as denying the user certain resources, downgrading their priority, or delaying them.

i) Backtracing
Backtracing is a form of active network defence that tries to find where an external attack is coming from so as to stop it more easily.

3.2 Security Standards and Protocols [18]

There are a number of international guidance undertaken by Integrated Computer Solutions (ICS) vendors to provide organisations and infrastructure, both in the private and public sectors locally with the best security services, standards and certifications possible. These are next depicted.

3.2.1 International Electrotechnical Commission (IEC) 62351

IEC is a standard developed for handling the security objectives include authentication of data transfer through digital signatures, ensuring only authenticated access, prevention of eavesdropping, prevention of playback and spoofing, and intrusion detection.

3.2.2 Institute of Electrical and Electronics Engineers (IEEE) 1686

Technical Standard for Security for Intelligent Electronic Devices.

3.2.3 Instrumentation, Systems, and Automation Society (ISA) 99 / IEC 62443

The ISA99 Committee establishes standards, recommended practices, technical reports, and related information that will define procedures for implementing electronically secure manufacturing and control systems and security practices and assessing electronic security performance.

3.2.4 International Organization for Standardization (ISO) 27K

The ISO 27000 series of standards have been specifically reserved by ISO for information security matters.

3.2.5 NERC CIP

NERC CIP are standards that provide a cyber-security framework for the identification and protection of Critical Cyber Assets to support US, Canada and parts of Mexico’s electrical grid operators.

3.2.6 NIST 800-53

Recommended Security Controls for Federal Information Systems and Organizations.

4 CYBER SECURITY CHALLENGES

A report from the World Economic Forum released in January 2014 examines the need for new approaches to increase resilience against cyber-attacks and suggests that the failure to effectively secure cyberspace could result in an aggregate impact of approximately US$3 trillion by 2020 [19].

Many of the challenges of cyber security are also challenges of privacy and data protection [20]. Cyber security will always need new and better systems with all the emerging technologies. The goal is to provide the best and strongest security available by keeping the data as private and secure as possible. Below are some emerging challenges that need to be taken under consideration while amending standards:

4.1 Complexity of the Connected Network

As newer and bigger network are implemented for an increased connectivity of mobile devices and "always on" services, the cyberspace has become more complex and challenging to secure.

4.2 Growing Sophistication of Threats

With the increased load on devices, along came high end threats and professional hacking and cybercrime which make cybercrime activities much more difficult to counter and track. Now that the cyberspace has a much larger footprint across, in account of huge scale of information and data flow across, it has become clear that we are more exposed than ever without some "unbreakable security". The fact that companies almost never disclose security breaches for competitive and reputational purpose, it is even more difficult to work on and find solutions to those breaches on a global scale.

4.3 Threats moving to the Mobile Devices

The global population will be exceeded by the number of smart phones in use. Mobile devices may contain anything one need for their daily life routine from valuable personal data, office work. People usually carry their mobile devices with them almost all the time to keep in touch with their social world, access their bank accounts, surf the web, and monitor their health among many other activities. Organisations are always trying to reach to potential consumers on their everyday electronic devices. But this only give more possibility for new vulnerabilities for cyber threats. Malware can easily be distributed via app stores which may appear safe can also provide infected apps that can easily make their way to mobile devices [21]. Moreover, there is an increased risk of data being intercepted when connected to free public Wi-Fi [22]. "Tap-and-pay" functions can be compromised by malware and affecting local transactions [23]. An ICSPA study concluded that mobile malware is a key emerging threat in the cyberspace. Although it has been argued that cyber criminals are building better malware specifically designed for mobile devices, actual infection rates on devices are low for now since the distribution of malware to mobile devices has not yet been perfected. This can be expected to change in the near future [24]. Data protection will become more critical to mobile devices, and faced with this coming wave of threats to mobile devices, app developers, mobile industries and companies have a much greater responsibility in ensuring the safety of the mobile platforms.

4.4 The "Big Data" Paradox

"Big Data" can be characterised as a huge amount of data storage accumulated from both traditional sources, and progressively, new accumulation areas such as the web, sensors information, time, and locations extracted from social networks [25]. Results outcome from big data are often considered as answers to any issues that need solving [26]. But this brings up two different issues from cyber security point of view, firstly how secure is the data in Big Data context. Associations that choose to use the huge data available may present new potential security vulnerabilities and open many ways for malicious input [27]. As more networks and devices are connected, the risk of potential malware infecting devices increases. Big Data is definitely set to be one of the major breakthroughs of the recent years but it will certainly be the target of attackers out there. Hence the need for a proactive inclusion of security measures for Big Data.

5 EMERGING TECHNOLOGIES & SECURITY

Gartner predicted that the security technology and services market would reach $67.2 billion in 2013, up 8.7 percent from 61.8 billion in 2012. The market is expected to grow to more than $86 billion by 2016. The growth is partly due to interest in a new set of emerging security technologies and a return of more capable defences that address mobile security, authentication weaknesses and threats to data in the cloud [28]. These technologies are next described along with some of the associated vulnerabilities as well as the countermeasures that can be undertaken.

5.1 IPv6

IPv6 has been introduced because IPv4 addresses were predicted to run out in the near future, which is the case with the rapidly increasing growth in IT and technology. This is where IPv6 solved the issues with its huge number of address space. But due to some security concerns many organisations have slacked off their transfer to IPv6 as the organisations realised that the security analysis and products for IPv6 might be inadequate in spite of the fact that the network infrastructure is ready to support IPv6 transport. IPv6 cannot be implemented without considering the security aspect. IPv6 security vulnerabilities currently exist and as the popularity of the IPv6 increases, the number of threats also is increasing.

5.1.1 Vulnerabilities Concerns with IPv6

a) Tracking the Identity of the User

For IPv6-based dial-up connections, the user is assigned a 64-bit prefix after the connection is made through router discovery and stateless address auto-configuration. If the interface identifier is always based on the EUI-64 address, it is possible to identify the traffic of a specific node regardless of the prefix, making it easy to track a specific user and their use of the Internet.

b) IPv6 Address Spoofing (MAC Address Spoofing)

Since IPv6 address depends on MAC address which in a sense the MAC address is a networked device's unique code name, used to identify devices from each other in a network. The MAC address is hard-coded on a network interface controller (NIC) and cannot be changed. But with the right tools one might want to virtually change the MAC address of an NIC, also known as MAC Spoofing, for many reasons:

1. To get past MAC address filtering on a router.
2. Sniffing other connections on the network.
3. To keep their burned or blacklisted MAC address out of IDS and security logs.
4. To pull off a denial of service attack. Therefore, many attackers change their MAC addresses in different operating systems either manually or by specific tools. Unfortunately, this is privacy risk, because anyone who has your MAC address also has your IP address.

5.1.2 Multiple Addresses Vulnerability

IPv6 assigns multiple addresses to an interface which challenges the filtering rules in the firewalls and access control lists [29]. In such cases, a firewall will need to learn all the addresses dynamically and the filtering rules will need to be automatically generate-able using sophisticated policy rule sets. And such capabilities are not available.

5.1.3 Long Address Space

Port scanning is one of the most common techniques in use today. Port scanning allows "black-hats" to listen to specific services (ports) that could be associated to well-known vulnerabilities. In IPv6 networks, IPv6 subnets use 64 bits for allocating host addresses. Scanning such a large address space (264) is not absolutely impossible [30].

5.1.4 Fragmentation Security Vulnerability

Fragmentation is the process of dissecting an IP packet into smaller packets to be easily carried across a data network that cannot transmit large packets, fragmentation is never performed by the intermediary routers but by the end nodes themselves. So, only the end hosts are allowed to create and reassemble fragments. This process may allow attackers to either hide their attacks or attack a node [31] [32].

Attackers can create fragments in such a way as to exploit weaknesses in the method an end host uses to defragment. Examples of this would be overlapping fragments, where there is an overlap in the offset and out-of-order fragments where the fragments' IDs do not match correctly with the data. Another type of fragment attack involves an attacker sending an incomplete set of fragments to force the receiving node to wait for the final fragment in the set. Fragmentation attacks also involve nested, where the IPv6 packet has multiple fragmentation headers. Fragmentation attacks are typically used by hackers with tools such as Whisker, Frag-router, Teardrop, and Bonk [33].

5.2 IPv6 Vulnerabilities Control

There are a number of steps to be taken so as to minimise the threats to IPv6. Below are a few guideline that needs to be taken under consideration:

1. Use standard, non-obvious static addresses for critical systems,
2. Ensure adequate filtering capabilities for IPv6,
3. Filter internal-use IPv6 addresses at border routers,
4. Block all IPv6 traffic on IPv4-only networks,
5. Filter unnecessary services at the firewall,
6. Develop a granular ICMPv6 filtering policy and filter all unnecessary ICMP message types,
7. Maintain host and application security with a consistent security policy for both IPv4 and IPv6,
8. Use IPsec to authenticate and provide confidentiality to assets,
9. Document the procedures for last-hop trace back and,
10. Pay close attention to the security aspects of transition mechanisms.

5.3 IPSec

IPSec, is a framework of open standards from Internet Engineering Task Force (IETF) that provides secure communications in a network by providing cryptographically-based security for IPv4 and IPv6. It offers different security administrations at the IP layer and in this way, offers security to the network layer and higher layers of the stack. The security administrations include: instance, access control, connectionless integrity, data origin authentication, protection against replays (a form of partial sequence integrity), confidentiality (encryption), and limited traffic flow confidentiality [34].

5.4 Cloud Computing

Cloud computing enables application software to be operated utilizing internet-enabled devices. Clouds can be categorized as public, private, and hybrid. Cloud computing relies on sharing of resources to achieve coherence and economies of scale, homogeneous to a utility such as the electricity grid over a network. At the substructure of cloud computing is the broader concept of converged infrastructure and apportion accommodations. Cloud computing, or in simpler shorthand just "the cloud", additionally works on maximizing the efficacy of the shared resources. Cloud resources are conventionally not only shared by multiple users but are additionally dynamically reallocated per demand. This can work for allocating resources to users [35].

5.4.1 Security of the Cloud

Cloud computing security or simply cloud security is a developing sub-area of Computer security and network security. It refers to a wide set of policies, technologies, and controls conveyed to secure information, applications, and the related framework of cloud computing. Cloud security architecture is powerful only if the right protective usage are set up. A proficient cloud security infrastructure will need to address the issues that will emerge with security administration and security controls. These controls are set up to protect any shortcomings in the framework and decrease the impact of an attack [36].

5.5 IoT [37]

The vision of the internet of things (IoT) is to manage objects around us with their own unique identifiers such as IP addresses and MAC addresses. IoT will consist of billions of devices that can sense, communicate and take actions. The rise of IPv6 also contributes to the uptake of the Internet of Things [38]. Fundamentally IoT will give the capacity of transferring information to more than a network without the need of human-to-human or human-to-PC association, and will rely mostly on Machine to Machine (M2M) communications. The definite size of the Internet of Things is difficult to gauge, as its development is gigantic and will very soon surpass whatever remains of the Internet in size (number of nodes) and will keep developing at a fast rate, maybe even in trillions of devices. The best development potential will probably come from embedded, low-power, wireless devices and systems that were not previously IP-enabled, opening up the era of the Wireless Embedded Internet.

6 OVERVIEW OF CYBER SECURITY INFRASTRUCTURE IN MAURITIUS

Mauritius is one of the few African countries that has legislations, laws and infrastructures that govern cyber security and help against related cybercrime activities. Mauritius along with many other countries have been working with ITU, and in collaboration with key partners, such as the International Multilateral Partnership Against Cyber Threats (IMPACT), to facilitate the development of cybersecurity capabilities, including the establishment of national CIRTs [40]. The security system in Mauritius is quite robust and redundant but there still some issues to be taken under considerations. In this section, the security framework of Mauritius is reviewed.

6.1 ICT Laws against Cyber Crime

There are a couple of ICT Laws that help govern the Mauritian Cyberspace [41];

6.1.1 Computer Misuse and Cyber-Crime Act 2003

The Act was enacted to provide for repression of criminal activities perpetrated through computer systems and to help reduce cybercrimes impact on the society and discourage to give a strong signal to would-be perpetrators. Cyber criminals may face severe penalties and fine for examples if found guilty for unauthorised access to restricted data and with intent to commit offences.

6.1.2 Data Protection Act 2004

The goal of this Bill is to accommodate the security of the protection privileges of people in perspective of the improvements in the procedures used to capture, transmit, and control, record or store information identifying with people. Protection laws refers to the privileges of people or associations to limit the utilization of information

6.1.3 The Information and Communication Technologies Act 2001

It provides the appropriate legal framework to facilitate electronic transactions and communications and give a new orientation to the traditional way of doing business by fostering the conduct of transactions by electronic means. It also provides for the legal recognition and regulation of electronic records, electronic signatures and their security. This allows devices to authenticate authorised access only.

6.2 Information Security Management System (ISMS) [42]

The ISMS under the Ministry of Technology, Communication and Innovation (MTCI) is a set of policies concerned with information security management or IT related risks. It is a system designed to establish, implement, operate, monitor, review, maintain, and improve information security. ISMS can be implemented as a specific information system that deals with a particular business area, or it can be implemented as an all-encompassing system involving the whole organization. The IT Security Unit facilitates the implementation of ISMS in Ministries and Departments by providing training to officers in information risk management, providing advice to their Information Security Forums, reviewing and auditing the ISMSs.

6.2.1 Standards provided by the Mauritius Standards Bureau.

- MS ISO/IEC 27001 - The Information Security Management System

This standard addresses the topic of information security management and provides a framework to initiate, implement, maintain and manage information security within an organisation.

- MS ISO/IEC 27002 - Code of Practice for Information Security Management

This is a standard code of practice which contains guidelines to be followed to set up and implement the ISMS. It can be regarded as a comprehensive catalogue of good security things to do.

6.2.2 ISMS Implementations

There are around 10 sites of the Government that are presently implementing their ISMSs. Two of which have already been certified at the national level as of March, 2015 [43]:

- Cane Planters and Millers Arbitration and Control Board

The Control and Arbitration Department has established and applied an Information Security Management System to protection the information it receives, deal with, saves and communicates while supplying arbitration and control services to Cane Planters and Millers.

- Passport and Immigration Office (PIO)

The passport and immigration office has established and applied an information security management system to protect the information it receives, deals with, or communicates while supplying its services for the issuance of passports and travel documents, immigration activities at the head office and offices at SSR International Airport and the Aurelie Perrine Passenger Terminal.

6.3 Central Informatics Bureau

The Central Informatics Bureau operates under the aegis of the MTCI. Its main function is to promote e-Governance through the provision of project management, consultancy and advisory services to Ministries and Departments for the successful implementation of e-government projects and on ICT matters.

6.4 National Computer Board (NCB)

The NCB comprises of many subsections and teams that work to help bring down cybercrime activities and bring awareness to the public. Some of those subsections are the GOC and CERT-MU.

6.4.1 Mauritius National Computer Security Incident Response Team (CERT-MU)

CERT-MU was set up by National Computer Board in May 2008 to handle and coordinate information security issues at the National level, as well as the managing of Information Security risks, such as Information Security breaches and incidents. CERT-MU Services includes [44]:

- Incident Response and Coordination
- Vulnerability Scanning and Assessment
- Assistance in the Implementation of ISO27001
- Information Security Audits
- Information Security Awareness
- Technology Workshops
- Capacity Building
- Presentations
- Issuance of Security Alerts
- Advisories & Vulnerability Notes
- Virus Alerts
- Information Security News

Basically the main objective is to provide to provide information and assistance to its constituents in implementing proactive measures to reduce the risks of information security incidents as well as responding to such incidents as and when they occur.

There some entities such as information infrastructure providers, system administrators both in the public and private sectors and the common public that benefit from security threats warning and as well as guidance on how handle those security risks [45].

6.4.2 Incident Statistics Reported

Incident statistics reports from CERT-MU shows an increased in certain cybercrime activities in Mauritius. While some activities were controlled, some others emerged. The following activities were increased from 2013 to 2014: unauthorised access by 3%, electronic fraud by 1%, identity theft by 3%, DoS attack by 1%, spamming by 2%, website defacement by 1%, offensive video by 1%, fake account by 22% [46] [44].

6.4.3 Cyber Security Mauritius Website

The website is under CERT-MU and helps instruct and improve the attention to the overall population on the technological and social issues confronting web clients, especially on the risks online. The Cyber Security Mauritius expects to give data to distinctive focused on gatherings, such as children, parents, home users and organisations.

6.4.4 The Government Online Centre

The Government online centre is the gateway to access applications online through one of its sub-portals, the Citizen Portal. 66 e-Services are online for the Citizen of Mauritius to interact with the Government anytime, anywhere and in real time. Operational since May 2005, the GOC is overseen by NCB and situated in the Ebene Cyber Tower. GOC takes a proactive, complete approach by giving its clients the business' most intense security devices and methods, designed, assembled and kept up particularly for big business class application hosting. From secure server manufactures and intrusion protection frameworks to a physically secure data centre and observed system, the GOC takes a multi-layered way to deal with keeping all application hosting operations dependable and secure.

7 IMPROVING SYSTEMS & SECURITY IN MAURITIUS CYBERSPACE

As new technologies and threats emerge, the need for system security upgrades are needed to counter the newer threats and continue protect those systems.

7.1 R&D (Research & Development)

Mauritius rely solely on other international cybersecurity organisation and research. Like other leadings countries, Mauritius does not have any R&D unit when it comes to fighting new cyber threats. As new technologies such as Smart Cities, IPv6, Cloud and IoT are emerging in Mauritius, a new infrastructure should be dedicated to the whole purpose of finding new methods of fighting cybercrime activities in collaboration with other stakeholders in Mauritius. This R&D unit should be specific for the system requirements of Mauritius.

7.2 Improved Governance Structure [48]

A new governance structure was proposed by the Ministry of ICT to ensure security development. The goal of the new structure is set in the National Strategy for Cybersecurity, 2014-2019, (see figure 1):

- Securing Mauritian Cyberspace

The aim is to set a cyber-defence strategy that prevents cyber-attacks by improving capabilities, defining roles and developing

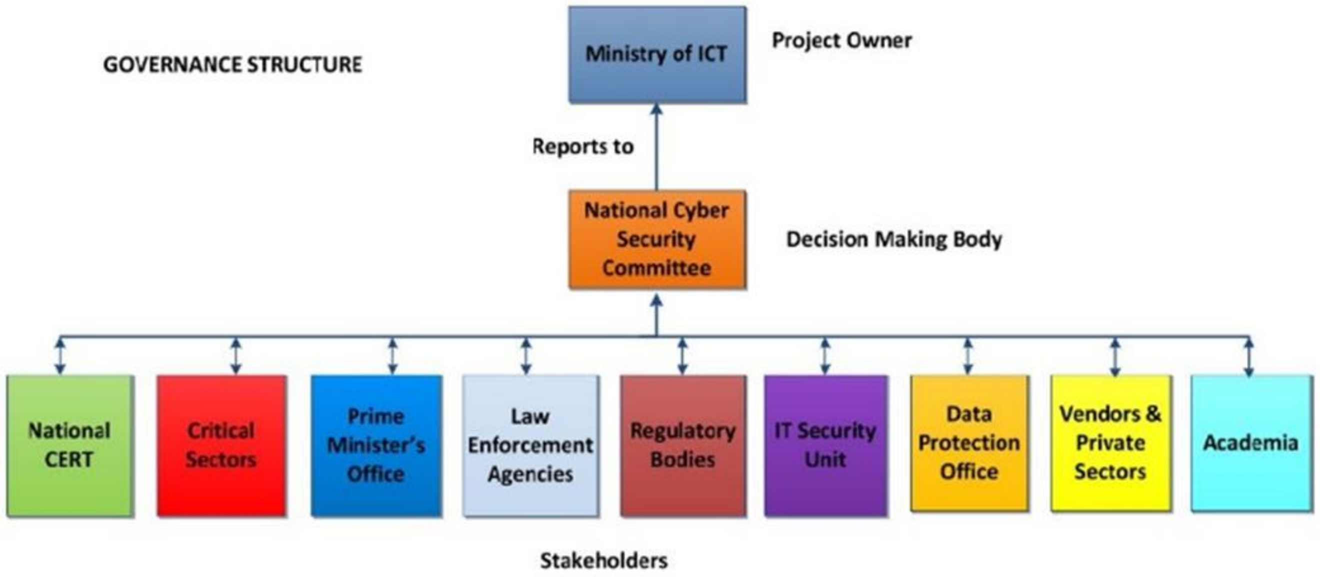


Figure 1: Governance Structure proposed by the National Computer Board

necessary responses for the public and private sector.

By creating a sense of awareness of the vulnerabilities, threats and the ability to react faster to prevent intrusions in our cyberspace. More laws should be enforced to investigate and prosecute cybercrimes.

- Enhancing resilience to Cyber Attacks

Efforts should be intensified to protect our critical network infrastructure in order to improve our resilience to cyber-attacks and to support various national projects and to maintain economic stability. Disruptions in the critical infrastructure can affect negatively the ability of a nation to perform effectively. The objective is to identify any disruptions in the cyberspace and to try to minimize their effects.

- Developing a collaboration between various authorities

There should be an active collaboration between the stakeholders whose aim is to achieve a nation free from cyber threats. Critical IT infrastructure and systems will be identified and will be protected. Cyber vulnerabilities assessments will be carried out and the Cyber Defence will be advanced.

- Improving Cyber Expertise

Cyber Security professionals and capacity building to deter and defend cyber threats should be encouraged. In this modern society, the level of competence of the professionals should be raised and to do so, a cadre of skilled cyber professionals should be promoted.

7.3 Encourage Active Monitoring System

Rather than relying only on passive security techniques or reactive defensive measures, organisations in Mauritius can better benefit from an active autonomous defence system that will automatically detect and eliminate the threats. Proactive procedures that enable the framework to identify attacks at an opportune time, and quickly react to neutralize the danger is becoming increasingly important.

7.4 Centralisation of Cybersecurity Infrastructure

The centralisation of the Government databases at a single location would represent an important step in the right direction. Currently, many of these databases contain fragmented information. As such, centralising these information systems will reduce confusion and provide only the most updated version of the information required. In addition, the security aspect would be greatly simplified. Such structure however does not preclude the building of secondary sites that replicates the data for back-up and recovery purposes.

8 CONCLUSION

Computer security endeavours to guarantee the secrecy, trustworthiness, and accessibility of computing systems and their segments. Three essential parts of a computing systems are liable to attacks: hardware, software, and data. These three, and the communications among them, are prone to computer security vulnerabilities. As Mauritius is interested in making the Information Technology Sector its main pillar of the economy by introducing Smart Cities, Internet of Things and Cloud computing among other innovative technologies, the need for better cyber security infrastructures is becoming increasingly important. Although cyber-attacks are generally the work of individual hackers, challenge gatherings, or criminals looking for unlawful monetary benefit, some likewise represent a risk to national security. To meet the security challenges of tomorrow, we must enhance our capacity to identify, research, and react to cyber-attacks today. For the case of Mauritius, the establishment of a National Cyber Security Committee is considered crucial. That committee would then take up the challenge of coping with the changing technology landscape and all the security issues this will involve.

9 REFERENCES

[1] R. Williams, “The biggest ever cyber attacks and security breaches,” Telegraph Media Group Limited, 22 May 2014. [Online]. Available: http://www.telegraph.co.uk/technology/internet-security/10848707/The-biggest-ever-cyber-attacks-and-security-breaches.html. [Accessed May 2015].

[2] R. Deibert, “Distributed Security as Cyber Strategy:Outlining a Comprehensive Approach for Canada in Cyberspace,” 2012.

[3] S. Musa, “The Evolllution,” 2013. [Online]. Available: http://www.evolllution.com/opinions/cybersecurity-understanding-online-threat/.

[4] S. R. T. Symantec, “FREAK vulnerability can leave encrypted communications open to attack | Symantec Connect,” Symantec Corporation, 4 March 2015. [Online]. Available: http://www.symantec.com/connect/blogs/freak-vulnerability-can-leave-encrypted-communications-open-attack. [Accessed May 2015].

[5] O. W. A. S. P. OWASP, “Open Web Application Security Project,” 5 April 2014. [Online]. Available: https://www.owasp.org/index.php/Category:Vulnerability.

[6] S. Bavisi, Computer and Information Security Handbook, Morgan Kaufmann Publications, Elsevier Inc, 2009.

[7] O. W. A. S. P. (OWASP), “OWASP,” Open Web Application Security Project (OWASP), 2013. [Online]. Available: https://www.owasp.org/index.php/Top_10_2013-Top_10.

[8] D. O'Brien, “Heartbleed Bug Poses Serious Threat to Unpacthed Servers | Symantec Connect,” Symantec Corporation, 4 April 2014. [Online]. Available: http://www.symantec.com/connect/blogs/heartbleed-bug-poses-serious-threat-unpatched-servers. [Accessed 2015].

[9] T. Symantec Security Response, “ShellShock: All you need to know about the Bash Bug vulnerability | Symantec Connect,” Symantec Corporation, 25 September 2014. [Online]. Available: http://www.symantec.com/connect/blogs/shellshock-all-you-need-know-about-bash-bug-vulnerability. [Accessed May 2015].

[10] C. Symantec, “Internet Security Threat Report, Volume 20,” Symantec Corporation, 2015.

[11] B. M. a. R. R. H. K. L. Vince Farhat, “Cyber Attacks:Prevention and Proactive Responses,” Practical Law Publishing Limited and Practical Law Company, Inc., 2011.

[12] C. Symantec, “Internet Security Threat Report, Volume 20,” 2015.

[13] N. BBC, “BBC News - US hospital hack 'exploited Heartbleed flaw',” BBC News -, 20

August 2014. [Online]. Available: http://www.bbc.com/news/technology-28867113. [Accessed May 2015].

[14] S. R. T. Symantec, “Torjan.Turla | Symantec,” Symantec Corporation, 2014. [Online]. Available: http://www.symantec.com/security_response/writeup.jsp?docid=2014-011316-1921-99.

[15] L. Kaspersky, “The Regin Platform | What is Regin? | Virus Definition,” 2015.

[16] A. Forzieri, “Secure and Protect your Critical Infrastructure,” in *Computer Security Day 2014* , 2014.

[17] C. P. Pfleeger and S. L. Pfleeger, Security in Computing, vol. 4, 2007.

[18] A. Forzieri, “Secure and Protect your Critical Infrastructure,” in *Computer Security Day 2014,* Mauriitius, 2014.

[19] McKinsey, “Risk and Responsibility in a Hyperconnected World,” World Economic Forum, 2014.

[20] O. o. t. P. Commissioner, “Privacy and Cyber Security,” Office of the Privacy Commissioner of Canada, 2014.

[21] A. S. B. a. M. Lockheed Martin, “International Cyber Security Protection Alliance,” 2013.

[22] I. C. S. P. Alliance, “Study of the Impact of Cyber Crime on Businesses in Canada,” 2013.

[23] M. Labs, “McAfee 2013 Threats Predictions Report,” 2013.

[24] A. Gonsalves, “Windows malware finds its way to Android | Computerworld,” 16 August 2013. [Online]. Available: blogs.computerworld.com/mobile-security/22662/windows-malware-finds-its-way-android?source=CTWNLE_nlt_security_2013-08-19.. [Accessed May 2015].

[25] C. f. I. P. Leadership, ““Big Data and Analytics: Seeking Foundations for Effective Privacy Guidance,”,” 2013.

[26] E. Schmidt and D. Wagner, Interviewees, *Why Data Analytics Is the Future of Everything.* [Interview]. 21 November 2013.

[27] P. Wood, “How to tackle big data from a security point of view,” 5 September 2013. [Online]. Available: http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view. [Accessed May 2015].

[28] R. Westervelt, “10 Emerging Security Technologies Gaining Interest, Adoption,” CRN, 17 June 2013. [Online]. Available: http://www.crn.com/slide-shows/security/240156647/10-emerging-security-technologies-gaining-interest-adoption.htm. [Accessed May 2015].

[29] A. S. A. Choudhary, “Securing IPv6 Network Infrastructure: A New Security Model,” in *IEEE Conference*, USA, 2010.

[30] P. Risztics, “Will IPv6 Bring Better Security?,” in *30th Euromicro Conference,* 2004.

[31] D. G. J. B. a. Y. P. K. Merike, “IPv6 Security Technology North,” 2006.

[32] M. H. Warfield, “Security Implications of IPv6 Whitepaper,” 2003.

[33] S. Hogg and E. Vyncke, “IPv6 Security,” Cisco Press, USA, 2009.

[34] K. Das, “IPv6.COM - IPv6 and IPSec - Securing the Next Generation Internet,” IPv6.com, 2008. [Online]. Available: http://ipv6.com/articles/security/IPsec.htm. [Accessed May 2015].

[35] H. Qusay, “"Demystifying Cloud Computing",” *The Journal of Defense Software Engineering,* pp. 16-21, 2011.

[36] a. R. D. V. Krutz Ronald L., “"Cloud Computing Security Architecture." Cloud Security: A Comprehensive Guide to Secure Cloud Computing,” Indianapolis, 2010.

[37] S. P. Pawar, “Smart City with Internet of Things (Sensor networks) and Big Dat,” 2015. [Online]. Available: http://www.academia.edu/5276488/Smart_City_with_Internet_of_Things_Sensor_networks_and_Big_Data. [Accessed 2015].

[38] *Internet of Things.* [Film]. USA: TEDxCIT, 2014.

[39] D. Maughan, “DHS S&T Showcase New Cybersecurity Technologies | Homeland Security,” DHS, 23 August 2013. [Online]. Available: http://www.dhs.gov/blog/2013/08/23/dhs-st-showcase-new-cybersecurity-technologies. [Accessed May 2015].

[40] ITU, “ITU Regional Cybersecurity Forum for Africa and Arab States,” ITU, Tunis, Tunisia, 2009.

[41] I. a. C. T. A. o. Mauritius, “ICT Authority | ICT Laws,” ICTA, 11 May 2015. [Online]. Available: https://www.icta.mu/laws/ict_laws.htm. [Accessed May 2015].

[42] .. Mauritius Standards Bureau, “Mauritius Standards Bureau - Information Security Management System,” MSB, [Online]. Available:

http://msb.intnet.mu/English/AboutUs/Pages/Structure.aspx. [Accessed 2015].

[43] .. Mauritius Standards Bureau, “Mauritius Standards Bureau - Certification Schemes - Information Management System,” 2015. [Online]. Available: http://msb.intnet.mu/English/Certification%20Schemes/Pages/Information-Security-Management.aspx. [Accessed June 2015].

[44] R. Sachindra, “The Current Legal Framework on Cybercrime in Mauritius,” CERT-MU, 2014.

[45] MTCI, “Mauritian National Computer Security Incident Response Team > About CERT-MU > Charter & Mission,” 2015. [Online]. Available: http://cert-mu.govmu.org/English/About_CERT-MU/Pages/Charter--Mission.aspx. [Accessed May 2015].

[46] K. Usmani, “An Enhanced Framework For Incident Handling,” CERT-MU, Port Louis, 2014.

[47] NCB, “National Computer Board | Government Onlince Centre,” 2015. [Online]. Available: http://www.ncb.mu/English/EPowering%20Public/Pages/Government-Online-Centre.aspx. [Accessed May 2015].

[48] MTCI, “National Cyber Security Strategy,” Port Louis, 2014-2019.

[49] ICTA, “The ICT Sector in Mauritius,” ICTA, Port Louis, 2004.

[50] Y. E. Gelogo, R. D. Caytiles and B. Park, “Threats and Security Analysis for Enhanced,” *International Journal of Control and Automation*, vol. 4, no. 4, pp. 179-184, 2011.

[51] T. Narten, E. Nordmark, W. Simpson and H. Soliman, “Neighbor Discovery for IP version 6 (IPv6),” 2007.

[52] N. M. CERT-MU, “Cybersecurity Mauritius Home,” NCB, [Online]. Available: http://cybersecurity.ncb.mu/English/Pages/default.aspx. [Accessed May 2015].

[53] MTCI, “Ministry of Technology, Communication and Innovation - ISMS Implementation,” 2015. [Online]. Available: http://mtci.govmu.org/English/Pages/IT%20Security/ISMS-Implementation.aspx. [Accessed May 2015].

The Establishment of Smart Cities in Mauritius: Requirements, Challenges and Opportunities

Muhammad Ridwan Sahadut, Mohammad Hashim Bundhoo and Pierre Clarel Catherine
University of Technology
La Tour Koenig, Pointe aux Sables, Mauritius
msahadut@umail.utm.ac.mu, hbundhoo@umail.utm.ac.mu, ccatherine@umail.utm.ac.mu

ABSTRACT

54% of the world’s population live in cities nowadays and this figure is expected to reach some 66% by 2050 as 60 million people are added to the cities every year. With such phenomenon, it will become a major challenge in the future to provide quality services to the residents due to resource constraints. In its 2015 budget, the government of Mauritius proposed “13 mega projects” including 8 smart cities and 5 techno-parks as a means to boost the economy and drive the competiveness of Mauritius in both the regional and international arena. The government is indeed promoting smart cities in terms of national benefits: promoting the business sector, improving of living standards of its citizens, as well as making a judicious use of resources. However Mauritius does not need to start this venture from scratch. There are examples of smart cities around the world where many smart ideas have been implemented from which the country can learn and adapt. In this paper we will look at essential elements that form a smart city and how they have been implemented in existing smart cities and furthermore how can we use these ideas to build smart cities in Mauritius, or in other countries ready to make the leap.

KEYWORDS

Smart city, Innovation, Efficiency, Sustainability, Reliability, Internet of things.

1 INTRODUCTION

If we look up the word “smart” in the dictionary, we will come across “intelligent”. And indeed, intelligence at all levels including government, policy making institutions, private sector, academia among others is needed.
According to Deloitte [1], a city can be defined as smart when investments in human and social capital, traditional and modern information and communication technology (ICT) communication infrastructure fuel sustainable economic development and a high quality of life, with a wise management of natural resources, through participatory action and engagement. This definition sums well the interaction of the various stakeholders required to make it work.

Smart cities are essential if the world is to respond effectively to the critical challenges it faces. The number of people living in cities is increasing every year; by 2030 it is estimated that 60% of the human population will be living in cities and 66% by 2050 [1]. The concept of a Smart City goes way beyond the transactional relationships between citizens and service providers. It encourages the citizen to become a more active and participative member of the community, for example by providing feedback on the quality of services or the state of roads and the built environment, adopting a more sustainable and healthy lifestyle, volunteering for social activities or supporting minority groups. Furthermore, citizens need a monthly income and “Smart Cities” are often attractive locations to live, work and visit.

The development of a ‘smart city’ consists of three layers [2]:

1. The physical layer, incorporating human capabilities and knowledge-intensive activities.

2. The institutional layer that incorporates proper institutional mechanisms for social cooperation towards knowledge and innovation development. (More specifically it involves institutions and mechanisms for information diffusion, transfer of technology, cooperative new product development).

3. The digital infrastructure layer that incorporates a range of ICT infrastructure, tools, applications and content in support of both individual and collective action.

These multiple elements of a smart city makes the implementation a challenge that cannot be resolved by a few stakeholders only. It will take all the components of the city to decide what the smart city should look like. As Deloitte points out [1], the primary objective of the smart city is to enhance the quality of lives of its citizens. While cities across the world will likely share common traits, there are also some specificities that should be considered.

2 CRITICAL FACTORS FOR GOING "SMART"

According to [3], a smart city is based on 3 aspects, namely: the communication means (network infrastructure and technology), the process (networking of various actors), and the goal (public involvement or others). Smart cities are also commonly defined by 5 critical success factors which are deployment of broadband communication infrastructure, effective education and training of local labor force, policies and programs that promote digital democracy, innovative capacity and marketing (see figure 1):

1. Deployment of broadband communication infrastructure is used as an evaluation of the local capacity for digital communication. The evaluation takes into consideration both the city's local vision and the affordability of costs incurred by the users.
2. Effective education and training of local force, to increase the rate of adoption of new technology infrastructures. This increases the capacity of the workforce to perform knowledge-intensive activities and enhances knowledge processes.
3. Policies and programmes that promote digital democracy, which are then applied across the society to benefit from the broadband revolution.
4. Innovative capacity, which assesses the level of creation of an innovation friendly environment. This provides the ability to attract highly creative people and businesses.
5. Marketing of the smart city as an advantageous place for living, working and setting up businesses. This increases the city's potential to attract talented employment and investments.

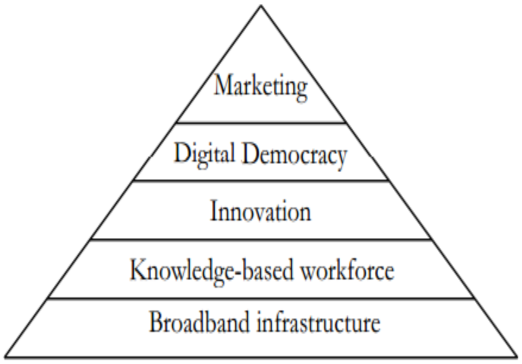


Figure 1: Critical success factors for cities 'going smart' [3].

3 DIGITAL DIMENSIONS OF SMART CITIES

The smart city can offer citizens and businesses with a range of tools and applications. These applications can be classified into 6 respective groups [3] (see figure 2):

- E-information: providing a vast sea of information to a wide range of audiences.
- E-business: potential exploitation of opportunities and adoption of new business strategies offered.
- E-marketing: supports a range of e-marketing possibilities for a city
- E-Government: provide more effectively services, businesses, and governmental institutes.
- E-innovation: refers to the potential for e-corporation as on-line development of new products.
- E-participation: strengthening active participation in the process of decision making.

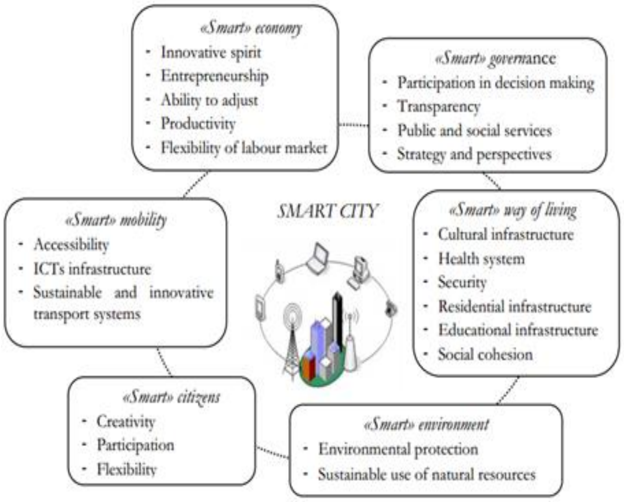


Figure 2: Dimensions of smart city development [3]

The use of digital technologies is to enhance the performance and well-being of citizens as well as reducing cost and resource consumption. In order to engage more effectively and actively with the citizens, a smart city should be able to respond faster to city and global challenges. It has been suggested that smart cities use information technologies to make a more efficient use of physical infrastructure such as intelligent lighting and smart building controls to support a strong and healthy economy. Energy-consuming city lighting can become sustainable with the use of renewable energy sources. Smart energy in solar panels, fuels cells and wind turbines can further be used to produce electricity and generate power. Such initiatives will boost the "green" into our lives. In addition, the cities of tomorrow must not only be improved in terms of ergonomics and technological comfort, but they also must be disaster-proof. The recent earthquake in Nepal and the earlier Tsunamis hitting the Indian Ocean are testimony to this need.

4 THE KEY PILLARS FOR CITIES

Cities all share some common traits since they all strive to achieve three objectives, presented here as the city sustainability pillars. The first is economic sustainability, i.e. a dynamic, productive city with numerous business opportunities that generates wealth. This requires on one hand high productivity of the private sector and on the other hand a healthy and well-financed public service. The second is social sustainability, guaranteeing access to all citizens to basic services and avoiding social exclusion. The third is environmental sustainability, guaranteeing environmental services and a healthy living environment [4]. An additional challenge, closely linked to economic sustainability would be financial sustainability so that the objectives of the city are achieved based on a financially sound plan, ensuring that costs are fully covered and the city is not at risk of insolvency. While it may be difficult to properly assess the ROI of such investments, it would be wise to lay ahead structures that will contribute to job and wealth creation of these new smart cities.

4.1 Economic Sustainability

The development of smart cities and the fullest adoption of innovations by city inhabitants, require an understanding of the economic fabric of the city and the market for smart solutions. Understanding the market allows for the development of new approaches to infrastructure financing, as well as influencing the citizen's behavior through these approaches. For cities requiring public private partnerships (PPPs) and systems of cost recovery using user charges, this knowledge is of paramount importance [4].

Smart city services contribute to the economic sustainability and the resilience of cities to economic shocks, as those generate a new level of economic diversification.

Economic sustainability is also closely linked to financial sustainability, particularly in the wake of the financial crisis. Many cities have seen their access to capital curtailed and their credit rating deteriorate, while financial institutions have restricted access to credit [4].

Nevertheless, investing in the city structures of the future can be done using novel financial models, which monetize savings and use them to finance the reimbursement of capital expenditures. In addition, the cities of the future are expected to have much more decentralized energy services and supply provision systems thereby creating new economic activities. Most importantly, financing models must be based on solid cost benefit analysis, including wider socio-economic benefits where necessary.

4.2 Social Sustainability

City authorities have a key interest to ensure social inclusion, which starts with a basic level of services for all citizens. In a smart city, it is important to take into account the risks of alienating important groups of citizens. This may happen because smart services are limited to richer areas of the city, or because user charges make many important services unaffordable for certain parts of the population. All models of development of cities must ensure that public transport, water, sanitation, electricity, and telecommunications are affordable and accessible to all population groups [4].

In particular, smart city infrastructures or services will need to respond to the following questions:

- How can it be guaranteed that basic city services remain affordable?
- Who is paying for the services? Are the users that can afford them the right target group?
- Can the new services and infrastructures be understood and used by all citizens targeted?
- Are the social and cultural values of the citizens taken into account?

Ideally, smart city projects should be carried out only if they help cities to meet their needs, with a quantifiable added value facilitated by technology integration, usability or cost reductions [4].

4.3 Environmental Sustainability

Environmental concerns are also growing in cities. Three challenges arise. The first is on resource limitations, such as water scarcity and quality. The second is on quality of life and health. The third is on risk management and resilience to environmental shocks (disasters caused by climate change) [4].

One of the first steps to address environmental sustainability is to increase resource efficiency in all domains, such as energy efficiency in buildings and networks, fuel efficiency in transport, water efficiency and new methods to transform waste to energy. Technology is not the only aspect required for sustainability, but is an important and necessary one. Efficiency gains may need significant investments, and the integration of different technologies can be complex [4].

Resilience and risk management also need to be integrated in city planning, based on estimated future risks. The smart city is essential and possibly our best bet to move towards the required environmental sustainability [4].

5. EXAMPLES ILLUSTRATING THE DIFFERENT CHALLENGES & ASPECTS OF EXISTING SMART CITIES

This section illustrates the different challenges that cities from the different parts of the world face. The speed of urbanization, levels of social inequality, infrastructure needs are highly different and complex. However these examples can serve as valuable references for building the smart cities of Mauritius, or elsewhere.

5.1 Barcelona

With the aim of becoming a good example for Smart Cities, Barcelona is working to merge urban planning, ecology, and information technology to make sure that the benefits of technology reach every people and improve the lives of its citizens. Barcelona's transformational approach follows a long-term vision based on building productive, human-scale neighborhoods within a hyper-connected, high-speed and zero-emission metropolis [5].

Barcelona Smart City proves smart in terms of the efficiency of its processes and the quality of life of its city residents. For example, 50 percent of Barcelona's lighting power is controlled remotely. With the long-term impacts of these measures, it is clear that the Barcelona Smart City has already and will continue to serve as a best practice for other global cities seeking to employ the best available technologies to develop durable sustainability initiatives in the best interest of their citizens [5]. The initiatives, failures and successes of Barcelona can indeed teach everyone valuable lessons regarding the deployment of smart cities.

5.2 Vienna

Smart City Vienna is a longstanding initiative by the city of Vienna to improve the design, development and perception of the federal capital. Vienna looks at a cross-section of the city, masking all areas of life, work and leisure activities in equal measure, and includes everything from infrastructure, energy and mobility to all aspects of urban development [6].

The city has set itself the task of systematically and uninterrupted modernize the city in order to reduce energy consumption and emissions without having to forego any of the advantages of mobility. Vienna has also accomplished bold smart-city targets and tracked their progress to reach them, with programs like the Smart Energy Vision 2050, Roadmap 2020, and Action Plan 2012-2015. Vienna's planners are incorporating stakeholder consultation processes into building and executing carbon reduction, transportation and land-use planning changes in the hopes of making the city a major European player in smart city technologies [8]. Such consultation processes with all classes of stakeholders represent an important activity that ensures the smart city meets the expectations of its users.

5.3 Santander

The Spanish city is embedded with more than 12,500 sensors to help the government operate as efficiently as possible. It is an adaptation of the way Europe thinks about cities [9]. Santander was chosen four years ago to become Europe's test bed for a sensor-based smart city. Since 2010, 12,500 sensors have been placed in and around the city's downtown district, where they quantify everything from the level of trash in containers, to the number of parking spaces available, to the size of crowds on the sidewalks. Furthermore, sensors on vehicles such as police cars and taxicabs measure air pollution levels and traffic conditions [9].

Dubbed SmartSantander, the project, which received an $11 million grant from the EU, started when Luis Muñoz, an engineering professor at the local University of Cantabria, and his development team of 20 technicians, researchers and programmers.

5.4 Birmingham

The affirmation statement of Birmingham's Smart City Commission sets out the huge ambition for Birmingham that will define the technological solutions and partnerships to meet the city's immediate and future challenges using smart city concepts.

Birmingham, like many cities, faces outstanding challenges. They need to move to a low carbon economy and adapt to climate change. Furthermore, the national and financial situation has intensify many of the inequalities of the city; some areas are extremely deprived; wages are lower, more people are unemployed, and the health outcome of residents are poorer in these places [10].

The aim of Birmingham was therefore to embed a capability for smart and sustainable re-invention of the way the city runs and in the way new businesses are created to deliver a step-change in Birmingham's economic growth, well-being and prosperity [10].

Birmingham City Council and Virgin Media Business thus initiated free public Wi-Fi across the city center, giving residents, shoppers and tourists access to unlimited data services on the go. Access points – roughly the size of a shoe-box – have been mounted on lamp-posts and other street furniture across the city center and connected back into Virgin Media Business nationwide fiber optic network to ensure unlimited, super-fast connectivity [11].

The citizens are now able to do everything from streaming videos, to staying in touch with friends on Twitter or Instagram, to looking up the latest bus timetables via Wi-Fi. With no usage limitation, people are also capable of downloading content or browsing the web as long as they need [11].

6. ESSENTIAL ELEMENTS IDENTIFIED IN MAJOR SMART CITIES.

In this section, some of the salient features of the smart cities depicted in the previous section are pointed out and summarised.

6.1 Smart Transportation

A smart bus network is a desired feature of a smart city, particularly one that is easy to understand, intuitive, faster and better connected, so travellers can save time and move around the city in a simpler and more sustainable manner. What is more, bringing on board improvements to technology ensures the system is managed more efficiently: "right-of-way" traffic lights, transfer points, in-bus and bus stop information, smart management to improve speed, frequency and service provision across the city, as well as the optimisation of resources based on people needs [12].

Barcelona has therefore decided on a New Bus Network based on vertical, horizontal and diagonal routes, where users of public transport are informed and benefit from the improvements that have been made [12].

The city of Santander is also opening up its data so that programmers can create apps that help citizens find bus arrival times or let tourists find out who is performing at concert halls simply by pointing their mobile phones at a bus stop or building. The collecting of data through numerous sensors could lead to significant improvements in how city infrastructure is used and lead to a better understanding of urban issues [9].

6.2 Broadband

The city of Seoul launched Broadband City/Broadband Metropolis, where fiber optic backbones were established in the city and enabled the interconnection of households and local enterprises to ultra-high speed networks [13]. Connections to the backbone were established with fiber optic channels. Other cities that can be classified in this category are Beijing (China), Antwerp (Belgium), Helsinki, Amsterdam and Geneva. Antwerp and Amsterdam collaborated and interconnected their broadband networks. Geneva MAN (Metropolitan Area Network) was a pilot project that was envisioned in 1998 and installed in Geneva city in 2003 [13]. Combined with 100MBps a small urban area, it offered VoIP and TV services.

6.3 Parking Management

Santander introduced the parking management system coined Mobypark that provides drivers with a rapid solution to the issue of where to park and helps city officials reduce congestion and air pollution. Information about parking is displayed on special panels located at main intersections in the city, so anyone who is heading downtown will have an estimation of how many spaces are currently available and where they are situated [9].

Mobypark brings up the idea of Amsterdam's Smart parking. Parking a car in big cities is indeed becoming a complex process. Drivers spend on average 20 minutes while looking for a parking spot. This increases CO_2 emissions and wasting the time of users.

With Mobypark, private parking lots, public parking garages, hotels, and hospitals all make their unoccupied parking spots available for drivers. They offer their available parking spots on a platform where it is possible to check real time availability and book these parking spots ahead. As a result, drivers spend less time searching for a single spot and reduce CO_2 emissions. Mobypark also ensures that one may easily rent a parking place for several days of a private individual, hotel or another institution. The service of Mobypark consists of a website and an app (Android and iOS).

6.4 Smart Waste Management

In Finland, the Enevo One solution uses a system of smart sensors and cloud computing to cut waste collection costs, while saving both time and money [14]. Wireless sensors are used to measure and forecast the amount of waste and recycling containers. The facility combines fill-level forecasts with a set of collection parameters order to calculate the most low cost collection plan. This latter can be accessed by the driver through a tablet [14].

In general, Intelligent and efficient waste logistics consists of the following [15]:

- An embedded system and communication networks.
- Real-time information on the fill level of containers.
- Monitored truck routes.
- Remote and dynamic configuration of the routes of vehicles on the road.
- Use of centralized real-time control and analysis applications.

This service is already available in the European market, but could be easily adapted for other countries.

6.5 Some Other Elements Identified from an Exemplary Barcelona

Smart Lighting: Barcelona came up with a master plan in 2012 that includes projects to control street level lighting in addition to transitioning 50 streets and a total of 1,155 lampposts to LED technology [5].

Smart Energy: From smart grid projects the same city has also developed a program to achieve greater energy efficiency and has currently deployed more than 19,500 smart meters in the Olympic Villa [5].

Smart Water: Barcelona is well on its way to set a program that includes remote irrigation control for the City's green spaces. Thus far, 77 fountains are controlled remotely.

District heating and Cooling: Two networks provide hot water in 64 buildings spanning an area of 21km [5].

Smart Transportation: The City unfolded a master plan in 2012 to make public transport better by deploying orthogonal bus lines, five of which began running in October 2012 [5].

Zero Emissions Mobility: As part of an extensive plan to encourage the use of electric vehicles, Barcelona is setting electric charging stations, as well as electric vehicle fleets and car rentals. To date, the city has more than 500 hybrid taxis, 294 public electric vehicles, 262 recharging points, 130 electric motorbikes and an estimated 400 private electric vehicles on its streets [5].

Open Government: The City has developed a program to make municipal government activities more open to its citizens, starting with the deployment of 44 “citizen’s attention” kiosks and the launch of an Open Data portal in 2010 [5].

As can be noted, Barcelona has a lot to show to the world in terms of its bold initiatives to move the smart city agenda forward.

7. IMPORTANT THINGS TO BE WATCHED

In this section, we give a brief overview of some of the technologies and techniques that may need consideration in smart cities endeavors. We also provide a glimpse of some countries actively pursuing the establishment of smart cities and how we can learn from them.

7.1 Internet of Things

The Internet of Things (IoT) involves technology in which objects, animals or people are provided with unique identifiers and the ability to transfer data over a network without requiring human intervention. IoT has evolved from the merging of wireless technologies, micro-electromechanical systems (MEMS) and the Internet [16].

So far, the Internet of Things has been most closely been associated with machine-to-machine (M2M) communication in manufacturing and power, oil and gas utilities. Products built with M2M communication capabilities are often referred to as being smart [16]. Examples are smart label, smart meter and smart grids.

IoT, with its arrays of sensors could be embedded everywhere in the cities. Data analytics could then be used to extract and generate valuable information from sensor data and helping the various systems to become more intelligent.

7.2 IPv6

IPv6’s huge increase in address space is an important factor in the development of the Internet of Things [16].

Building a smart city requires connectivity of thousand of nodes talking seamlessly with one another and transferring important data; thus, extreme reliability and scalability are essential [17].

Some models of connectivity that can be find in a smart city are smart metering communication, street lighting control and traffic lights control.

Typically, these networks can spread several miles of distance while serving thousands of nodes per edge router or data concentrator [17].

IPv6 offers a highly scalable address scheme which is quite enough to address the needs of any present and future communicating device still allowing it to have many addresses.

In fact, one could say that IPv6 enables IoT, which in turn enable smart cities.

7.3 Co-creation

Co-creation is a business strategy focusing on customer experience and interactive relationships. Co-creation allows and encourages a more active involvement from the customer to create a value rich experience [18].

With co-creation, the participants such as granting customers, suppliers or the general population all contribute to the design of the final solution. Through a series of well-grounded steps, stakeholders of the project are invited to contribute, evaluate, and refine ideas and concepts [19].

We believe this may be a very important factor in any city that wishes to enhance its services. Indeed a city cannot become smart if some of its stakeholders do not have their say. Co-creation will ensure that the city meets the actual needs of the citizens and is not just an adaptation of a solution that worked somewhere else.

7.4 Cloud Computing

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. These services are broadly divided into three categories: Infrastructure-as-a-Service, Platform-as-a-Service and Software-as-a-Service [20].

Smart cities are a relatively new concept and their basic goal is the efficient use of natural resources as water, electricity, air quality, waste management among many services to the citizens. One such smart city in particular, Dublin, supports cloud computing as a natural resource [21]. The city policy-makers argue that cloud computing was quickly adopted by companies. In addition, city university initiated new curricula around the theme of cloud computing.

Another city, Guadalajara in Mexico followed a similar journey. This city has been recognized as model city to create smart guidelines established by National Digital Strategy announced by the Mexican government to promote the adoption and development of Information Technology and Communication [22]. Digital services that will be implemented in the city are:

- Type of Service
- Cloud Data storage
- Crowd sourcing traffic information
- Cloud based creative software for educational users
- Intelligent urban security systems
- Smart parking
- Remote global education opportunities
- Advanced consumer analytics

These services, in general, need an elastic, flexible and economic infrastructure to support workflows that demand a high performance computing resources. Hence a remote cloud service is used to classify the processes foot-prints and schedules.

7.5 Countries to Watch

Some countries that could easily be added to a watch list in the area of smart cities are India, China and Japan. They are bringing massive developments indeed.

India is building 100 smart cities across the country. The city of 2 million people, commonly known as Vizag, geared up to become one of India's first smart cities. Its officials admit that much more must be done to help it cope with increasingly extreme weather - a challenge being recognized in many fast-growing metropolises worldwide [23]. India has many inspiring ideas against natural disasters from which other countries can learn. These 100 smart cities are thus a global laboratory where many of the smart city technologies deployed could be replicated.

China's urbanization and Chinese cities' eagerness to "go smart" could mean great business opportunities for enterprises. More than 260 prefecture-level Chinese cities are building digital geographic systems to provide better services to citizens. The Internet of Things (IoT) is becoming a major new industry in China. IT services provider Digital China Holdings mentioned that it aims to have 100 billion yuan ($15.8 billion) in sales revenue during China's 12th Five-Year Plan 2011 to 2015 [24]. As china rushes into the smart city competition, a lot can be learnt about their smart city deployments for existing or legacy cities.

In Japan, the Ministry of Economy, Trade and Industry (METI) has been investing in Smart City projects since 2010. The promotion of smart energy initiatives is now one of the goals established by the Fourth Energy Strategic Plan, released in April 2014. Smart Cities are a booming market in Japan. The growing economic importance of cities and the necessity of addressing environmental issues have brought Japan to develop local solutions. The smart cities of this country will definitely provide interesting case studies as new methods are deployed to cater for early warning systems for earthquakes and other calamities often affecting the region.

On another note, European and Japanese companies are also contemplating the benefits from advanced cooperation on Smart City development and management. The two economies wish to improve their models and intensify their competition at a global scale [25].

7. MAURITIUS SMART CITIES

The government of Mauritius has recently launched in its 2015 budget ‘thirteen mega projects’. These projects include bringing the concept of smart city to eight different regions across the country along with another five high-tech manufacturing and information-based industrial zone, also coined as ‘technopoles’ [26]. The main aim to this development is to promote and ease the Mauritian lifestyle, create jobs and bring a boost to the construction sector. These cities will be designed to be environmental friendly and will produce their electricity while making an efficient use of water and smart modern transportation to ease the crisis of traffic congestion the country currently faces.

The regions that were selected for these projects are: The Omnicane airport city in the south-east, St Félix Village in the south, The Médine Integrated Park in the west, Roches Noires in the north-east, The Azuri Phase 2 project in the north, The Terra project in the north, The Highlands City in the center, The Richeterre Project in the vicinity of Port Louis. The 5 ‘Technopoles’ will be located at Highlands, Rose Belle, Flacq, Rivière du Rempart and Bambous. These projects will exploit roughly up to 7,000 acres of land and will required an investment of approximately 120 billion Mauritian rupees (about 3.5 billion USD), to be essentially drawn from private and foreign investments [26].

The Omnicane Mont Tresor Airport City will extend over 400 hectares of land near the airport of Mauritius. It aims at the developing the movement of airport activities with the African continent and the rest of the world while pressing on green energy. The CEO of Omnicane sees it also as an opportunity to launch a bioethanol distillery and a carbon burn out project. The city will consist of an industrial and technological activities park and is conceptualized as a hub of services including hotel and leisure activities [27]. The St Felix Village project is an ecovillage project which will provide office and housing facilities for foreigners. The Médine Integrated Park is a three phase project and the infrastructures will be mainly of IT projects, a University and a Hospital. Extending over 800 acres of land will be the Roche Noires project which has not been finalized however. The Azuri Phase 2 project will be a residential project which consists of 132 accommodation, leisure areas, gymnasiums and shopping centers. The Terra project in the north will consist of the culture of sugar cane along with business parks and residential portions. The Richeterre project will extend over 500 acres of land and will consist mainly of industrial zones and business parks. The Highlands City project is called for as a massive project that will help to decongest the city of Port-Louis, and become the new administrative center of the country. Port-Louis will then be transformed into a regional bunkering hub.

To support these smart cities, a new submarine cable linking Mauritius to the rest of the world will be installed [28]. Provision for a full broadband connectivity and FTTH is expected within the next 3 years while Free Wi-Fi Hotspots will be increased from 15 to 350 [28].

8. ESSENTIAL ELEMENTS OF MAURITIUS SMART CITIES

Many of the projects announced in the budget 2015 are not yet concrete. As such, it is important that we consider some of the priorities the country is currently facing when developing the concept of smart cities around the island. In this section, we provide some of the elements that needs particular attention.

8.1 Traffic Management

Providing modern transportation system and reducing traffic congestion across the island is a must. Traffic congestion indeed costs the country around 4 billion rupees per year, according to [35]. Furthermore, buses are always crammed in the morning with bus attendants often having difficulties to make their way through and attending each and every passenger. This is very time consuming. Another issue arises with money change which is not always available when customers tend notes. This issue could be resolved with the introduction of NFC (near field communication devices) compatible readers in buses. People would simply to tag their cards when they enter buses and tag again when getting off. The cost would be deducted from their cards automatically depending on their journey. This will reduce cash transactions and also bring a boost to one-man operation system. Wifi availability in buses could also help people to stay connected. Applications can be designed and be available to citizens, giving information about bus routes, planning and even keep track of buses they need to get onto. These applications could also be designed to indicate availability of sitting seats as per number of people that tagged their NFC cards in specific buses.

Parking facilities are also very lacking across the country. A Smart parking a system that would help drivers to quickly find parking spots is very much required. Our cities should manage their parking spaces more efficiently. Information could be sent via smart phones and via electronic street panels. This would reduce traffic density as well as decrease CO_2 emission.

In Moscow, Fastprk, a private company implemented a system where gateways send information via the internet to a database in real time. The parking occupancy in the area is instantly reported to users via apps and illuminated panels in the street. The central control gets real time analytics about the parking space per areas and times of the day. When connected to the payment method system, the authority can even identify non-paying cars with the use of a tablet application. The system relies on embedded sensors in each parking bay in the street. When a car is parked on the sensor, it is detected and the sensor relays that information [29].

Bringing efficient solutions to our present bus system as well as to the problem of parking are considered key to a good traffic management system. It is therefore hoped that some of the ideas provided here will be adopted for the traffic system in Mauritius.

8.2 Smart Waste Management Systems

Until now collecting waste has been done using static routes and schedules where containers are collected once or twice every week regardless if they are full or not. It is not only time consuming and of high cost but traffic congestion is also created with the trucks stopping at all times with workers getting down, emptying each and every can in the trucks manually. By turning these static routes into dynamic smart plans, which are based on real data from the field [14] waste management can be efficiently operated.

Trashcans could be connected to the internet via embedded sensors monitoring the amount of wastes. Software will then calculate how much trash per region in a daily basis and smart plans will be automatically generated. Even wastes producers like offices, residents and institutions can track how much the amount of waste they generate.

Trashcans could be placed in specific places in the vicinity of the yard or office places so that trucks can have easy access to them. This will allow the truck driver to pick up the containers with the help of automated machine fitted in the trucks without the help of several workers. This could well have a positive impact on the environmental sustainability of our cities and the country in general.

8.3 Smart Energy Use

The Central Electricity Board (CEB) is the main organization responsible for the transmission, distribution and supply to the electricity to the population. 40% of energy is produced by CEB and the rest is obtained from independent power producers mainly sugar estates. Energy is produced by thermal means (80%), water, and bagasse (both accounting for the remaining 20%) from small sugar estates [30]. With the smart city projects, electricity demand should be managed.

Smart grids could help to provide additional electricity to meet the rising demand, increase reliability and quality of power supplies, increase energy efficiency, while even reducing the carbon emissions. Smart grids have the capacity to help balance electrical consumption with supply, as well as the potential to combine new technologies to enable energy storage devices and the large-scale use of electric vehicles. A smart grid will also provide greater control over energy costs and a more reliable energy supply for consumers. Environmental benefits of a smarter grid include integration of more renewable power sources, and reduced CO2 emissions and other pollutants [31].

The use of smart meters should also be introduced along with smart grids. A smart meter is an electronic device that records consumption of electric energy in intervals of an hour or less and communicates that information at least daily back to the utility for monitoring and billing. Smart meters enable two-way communication between the meter and the central system. Smart meters can gather data for remote reporting. This advanced metering infrastructure (AMI) is different from traditional automatic meter reading (AMR) in that it enables two-way communications with the meter [32]. This will record easily the consumption of electricity and facilitate billing information.

The forthcoming smart cities should look forward at producing most (or even all) of their energy. With the help of photovoltaic cells, green energy can be obtained from the sun. Photovoltaic cells basically convert visible light to direct current. Large sets of PV cells can be connected together to form solar modules, arrays, or panels which can produce sources for utility power.

We also have a vast ocean and it can be used to produce two types of energy: thermal energy from the heat of the sun, and mechanical energy from tides and waves. With the appropriate techniques, a certain amount of energy can be produced for specific areas so as to reduce the load from the CEB.

8.4 Smart Water

Water management is essential to any smart city initiative. Currently, Mauritian citizens face frequent water cuts. There is also the problem of water leakage problems across the country. It is indeed estimated that about 50% of the treated water that goes in the piping system is lost even before reaching the customers' premises.

Facilities such as sensors could therefore be placed in water treatment processes so that water authority is capable of monitoring the turbidity, salinity, conductivity, pH and chlorine levels in the water at every moment. Alerts can also be received through mobile phones or any smart devices if any issue arises and can be solved in near-real-time.

As an example, IBM has its own solution to the above set of problems. Indeed, IBM intelligent water is a software that was developed to manage pressure, detect leaks, reduce water consumption, mitigate sewer overflow, and better manage their water infrastructure, assets and operations. It uses advanced data management and technology to monitor the flow of water and also smart metering for easy billing information. IBM Intelligent Water [33]:

- Includes the Intelligent Operations for Water component, which provides extensive visibility and situational awareness spanning water and wastewater operations. This helps improve decision-making, enhance efficiency and reduce risk.
- Includes the Water Efficiency Analytic component, which helps mitigate non-revenue water through pressure optimization and pipe failure prediction.
- Turns data from smart meters into opportunities for recapturing revenue and detecting fraud.
- Delivers insights from big data and smart devices to help operators improve irrigation, flood management and sewer overflows.
- Takes advantage of flexible deployment options by offering multiple deployment models.

The IBM software or its equivalent could therefore be applied to the Mauritian cities, along with the required set of sensors to help monitor the water usage and quality in the country.

8.5 Other Essential Elements

Smart Defence against flooding: Mauritius is currently facing a problem concerning floods. We have experienced it severely on in March 2013 where 13 persons lost their live, mostly due to poor infrastructure planning and drainage systems. Full consideration should therefore be given to appropriate countermeasures. Singapore, for example, has set up a smart defense against flooding. Local authorities should build defense systems so as to prevent floods when storm events occur or when heavy rain hits [34]. Sustainable Urban Drainage systems could be built that replicates the environment’s ability to drain water away at its natural rate prior to development. This is achieved through the storage of water in tanks, collecting water for other uses – such as flushing toilets – and holding some water in ponds or lakes. The water can then be discharged back into the environment at its natural rate and flooding from this flash run-off can be prevented [34].

Smart billboards: It is customary to see a couple of workers working on billboards in Mauritius, to change the adverts. This is time very consuming. With the help of smart billboards, this will be much easier. The adverts can be easily uploaded each week without much the hassle of sending workers across the country to do the job. Energy efficient photovoltaic billboards can also be used to save energy.

Meteorological Services: meteorological facilities can directly accessed from our mobile phones. However, services like weather forecasting requires region setting to operate properly. Unfortunately, these information are not currently sourced from our own meteorological station but are provisioned elsewhere. It would be better if Mauritius could provide this national information on its own.

CONCLUSION

Smart cities in Mauritius have many requirements and have to face several challenges. The thirteen mega-projects presented in the national budget 2015 were designed to ease lifestyle, decrease unemployment and boost the construction sector. The present paper shed light on a few directions we should check out as we venture into country-wide deployment of smart cities.

In particular, it is believed that there is a need stakeholders to sit together and decide of what smart cities in the country should look like. While we can take stock of the experiences from other cities such as Barcelona or Santander or from countries such as India or China, Mauritius will need to devise its own deployment plan, and this would be best achieved through co-creation processes, where all stakeholders express their opinion so that the smart cities actually meet the needs of all its citizens.

Technologies such as IPv6 and Internet of Things, Data analytics along with cloud computing will need to be thoroughly investigated. These technologies are the essential IT ingredients required to build the platform for smart cities. Training is therefore very important, and it is believed that academia will also need to restructure its role so as to equip the future workforce with the right skills to tackle the many challenges that these cities will pose.

It will be interesting if the factors considered in this paper are taken up in the upcoming National Innovation Framework report 2015-2020, scheduled for release in July 2015 by the Ministry of Technology, Communication and Innovation. It is finally hoped that Mauritius can serve as a valuable example for the African continent in its endeavor to go smart.

REFERENCES

[1] Caragliu, A., Del Bo, C., & Nijkamp, P. (2009). Smart cities in Europe. In Proceedings of the 3rd Central European Conference in Regional Science (Košice, Slovak Republic, Oct 7--9). Available at http://www.cers.tuke.sk/cers2009/PDF/01_03_Nijkamp.pdf.AMETIC, “2012 Smart Cities”, Foro Tic para la Sostenibilidad, September, 2013.

[2] Komninos N, The architecture of intelligent cities; Integrating human, collective, and artificial intelligence to enhance knowledge and innovation. 2nd International Conference on Intelligent Environments, Institution of Engineering and Technology, Athens, 2006.

[3] Bell et al. 2008; Passerini and Wu 2008, “Intelligent cities: towards interactive and global innovation environments”, International Journal of Innovation and Regional Development, vol. 1, n° 4, pp. 337–355. Available at ICF website at http://www.netcom journal.com /volumes /articles

[4] Thwink.org - Finding and Resolving the Root Causes of the Sustainability Problem. 2015. Thwink.org - Finding and Resolving the Root Causes of the Sustainability Problem. [ONLINE] Available at: http://www.thwink.org/. [Accessed 12 June 2015].

[5] Barcelona: Urban Platform. 2015. Barcelona: Urban Platform. [ONLINE] Available at: http://cityclimateleadershipawards.com/2014-project-barcelona-urban-platform/. [Accessed 12 June 2015].

[6] UNDP et al., 2013. http://www.undp.org/content/dam/undp/library/Environment%20and%20Energy/Climate%20Strategies/UNDP%20Derisking%20Renewable%20Energy%20Investment%20%20Executive%20Summary%20(April%202013).pdf

[7] Rathaus, "Smart City Wien – Ready for the future!," 2014. Home | Smart City Wien. 2015. Home | Smart City Wien. [ONLINE] Available at: https://smartcity.wien.gv.at/site/en/home-en/

[8] Global Smart City • Top Ten Smart Cities (Globally). 2015. Global Smart City • Top Ten Smart Cities (Globally). [ONLINE] Available at:http://globalsmartcity.tumblr.com/post/28841016437/toptensmartcities. [Accessed 12 June 2015].

[9] T. Newcombe, "Governing," MAY 2014. [Online]. Available: http://www.governing.com/topics/urban/gov-santander-spain-smart-city.html.

[10] Birmingham Smart City - Birmingham City Council. 2015. Birmingham Smart City - Birmingham City Council. [ONLINE] Available at:http://www.birmingham.gov.uk/smartcity. [Accessed 12 June 2015].

[11] Caragliu, A., Del Bo, C., & Nijkamp, P. (2009). Birmingham Free wifi | Digital Birmingham. 2015. Birmingham Free wifi | Digital Birmingham. [ONLINE] Available at:http://digitalbirmingham.co.uk/project/birmingham-free-wifi/.

[12] "BCN Smart City," Barcelona.[Online]. New bus network | Projects | Barcelona Smart City | Barcelona City Council | Barcelona Smart City. 2015. New bus network | Projects | Barcelona Smart City | Barcelona City Council | Barcelona Smart City. [ONLINE] Available at:http://smartcity.bcn.cat/en/new-bus-network.html

[13] L. A. a. P. Fitsilis, "using classification and roadmapping techniques for smart city viability," in Greece. Using Classification and Roadmapping techniques for Smart City viability s realization by Academic Conferences International

[14] M. München, "Smart Logistics Solution For Waste Collection Launched at IFAT," 2014.[Online]. Available: http://www.enevo.com/news/smart-logistics-solution-waste-collection-launched-ifat/

[15] 2015. Smart waste management | Urbiotica "SMART WASTE MANAGEMENT".[Online]. Available: http://www.urbiotica.com/en/smart-solutions/intelligent-waste-management/

[16] I. Wigmore, "Internet of Things (IoT)," 2014. [Online]. Definition from WhatIs.com Available: http://whatis.techtarget.com/definition/Internet-of-Things

[17] " Smart Cities Connectivity Through IPv6|Freescale. 2015. Smart Cities Connectivity Through IPv6|Freescale.”[Online]. Available:

http://www.freescale.com/webapp/sps/site/application.jsp?code=APLESCCIPV6

[18] What is co-creation? definition and meaning. 2015"Your Digital Experts Digital Marketing– Internet Strategy".[Online} Available: http://www.businessdictionary.com/definition/co-creation.html

[19] S. Benson, "VisionCritical," 21 October 2013. [Online]. Available: https://www.visioncritical.com/cocreation-101/.

[20] M. Rouse, "TechTarget," February 2015. [Online]. Available: http://searchcloudcomputing.techtarget.com/definition/cloud-computing.

[21] P. Howlin, Silicon Republic | Technology, science and start-up news. 2015. Silicon Republic | Technology, science and start-up news."SiliconRepublic," 19 july 2012

[22] Client Validation. 2015. Client Validation. "Mexico Presidentia de la Republica," 30 01 2014. [Online]. Available: http://www.presidencia.gob.mx/objetivos-de-la-estrategia-digital-nacional.

[23] N. Bhalla, Business Insider. 2015. India is building 100 'smart cities' across the country "India is building 100 'smart cities' across the country," Reuters, New Delhi, 2015.

[24] "China Daily Information Co (CDIC)," 2013. [Online]. Available: http://www.chinadaily.com.cn/business/China-smart-city.html.

[25] P. Clarisse, "SMART CITIES IN JAPAN," in An Assessment on the Potential for EU-Japan Cooperation and Business Development, Tokyo, 2014.

[26] S. Lutchmeenaraidoo, "Mauritius at cross road Budget 2015," 2015. [Online]. Available: http://www.investmauritius.com/budget2015.

[27] Omnicane unveils ambitious Mon Trésor Airport City project | AfricaMoney. 2015"AfricaMoney," [Online]. Available: http://africamoney.info/omnicane-unveils-ambitious-mon-tresor-airport-city-project/.

[28] Inf, "Greekscribes," 29 march 2015. [Online]. Available: http://www.geekscribes.net/.

[29] Smart parking technology | Smart city sensors | FastPrk. 2015"Fastpark the easiest way to park," Fastprk, [Online]. Available: http://www.fastprk.com/.

[30] Central Electricity Board Web Site. 2015"Central Electricity Board," CEB, [Online]. Available: http://ceb.intnet.mu/.

[31] "Powered and Productivity for a Better World ABB," [Online]. Available: http://new.abb.com/smartgrids/what-is-a-smart-grid.

[32] D. Kathan, "Federal Energy Regulatory Commission," 2008. [Online]. Available: http://www.ferc.gov/legal/staff-reports/12-08-demand-response.pdf.

[33] IBM, "IBM Water Managment," [Online]. Available: http://www-03.ibm.com/software/products/en/intelligentwater.

[34] A. Langley, "localgov.co.uk," Morgan Sindall Professional Services, 2009. [Online]. Available: http://www.localgov.co.uk/A-smart-defence-against-flooding/26945

[35] C. Dorsamy, and C. Puchooa. Alleviating Traffic Congestion along the M1 corridor - An economic perspective. The Journal of the Institution of Engineers Mauritius, pp.27-34, 2013.

Towards Adaptive Analytics on Big Data Sources

Verena Kantere and Maxim Filatov
University of Geneva
Geneva, Switzerland
{verena.kantere, maxim.filatov}@unige.ch

ABSTRACT

The analysis of Big Data is a core and critical task in multifarious domains of science and industry. Such analysis needs to be performed on a range of data stores, both traditional and modern, on data sources that are heterogeneous in their schemas and formats, and on a diversity of query engines. The users that need to perform such data analysis may have several roles, like, business analysts, engineers, end-users etc. Therefore a system for Big Data analytics should enable the expression of analytics tasks in an abstract manner, adaptable to the user role, interest and expertise. We propose a novel workflow model that enables such users to define in an abstract manner the application logic of the analysis of diverse Big Data. The model focuses on the separation of task dependencies from task functionality. Our motivation and applications derive from real use cases of the telecommunication domain.

KEYWORDS

Big data, analytics, workflow management

1 INTRODUCTION

The analysis of Big Data is a core and critical task in multifarious domains of science and industry. Such analysis needs to be performed on a range of data stores, both traditional and modern, on data sources that are heterogeneous in their schemas and formats, and on a diversity of query engines.

The users that need to perform such data analysis may have several roles, like, business analysts, engineers, end-users, scientists etc. Users with different roles may need different aspects of information deduced from the data. Therefore, the various users need to perform a variety of tasks, like simple or complex data queries, data mining, algorithmic processing, text retrieval, data annotation, etc. Moreover, they may need to perform such tasks in different scheduling schemes, for example short or long-running queries in combinations with a one-time or a continuous output. Finally, the users may differ in their expertise with respect to their data management skills, as well as on their interest in implementation specifics. Thus, a system for Big Data analytics should enable the expression of simple tasks, as well as combinations of tasks, in a manner that describes the application logic of the tasks and is adaptable to the user role, interest and expertise.

To fulfill the above requirements we propose a novel workflow model for the expression of analytics tasks on Big Data. The proposed model allows for the expression of the application logic while abstracting the execution details of tasks and the details on the data formats and sources. The model enables the separation of task dependencies from task functionality, as well as the adaptation of the level of description of execution semantics, i.e. the execution rationale. In this way, the model can be easily, i.e. intuitively and in a straightforward manner, used by any type of user with any level of data management expertise and interest in the implementation. Therefore, using the proposed model, a user is not only able to express a variety of application logics for Big Data analytics, but also to set her degree of control on the execution of the workflow. This means that the model enables the user to express specific execution semantics for parts of the workflow and leave the execution semantics of other parts abstract. The latter are decided by the analytics system at the processing time of the workflow.

Motivating applications This work is part of the ASAP[1] research project that develops a dynamic open-source execution framework for scalable data analytics. Distributed data processing is the core application that motivates our research in the ASAP project. Storing and analyzing large volumes and variety of data efficiently is the cornerstone use case. A goal of the project ASAP is to design and develop an analytics platform for telecommunication analytics.

Data sources for cellular networks bring a new quantity and quality to the analysis of mobility data. We are interested in the analysis of GSM/UMTS data in application scenarios that focus on handling, interpretation and analysis of cellular data. ith a provable and feasible protection of sensitive information when the Big data approach is used. One scenario is mining mobility of telecommunication customers to improve portfolio services according to the legal context: In the new ecosystems the right to the protection of the private sphere must coexist with the right to access to knowledge and to services as a common good. Through the analysis of cellular data (voice and SMS) and the correlation with information on the customers, it is possible to identify real tourist flows, linking and counting the unique customer making or receiving calls in states/provinces/cities other than that of residence. In this way the traffic would be automatically 'sorted', easily traceable and then analyzed. Such an application is important from the marketing point of view because it will be possible, starting from the cellular data for each customer (and from her trajectories), to address a Social Network Analysis (SNA) taking into account also the legal constraints coming from the privacy rules. The analysis of such data can be useful for various situations. Examples are the reconstruction and optimization of the transportation of things and people, the behaviour analysis of people flows for the promotion of tourism, etc.

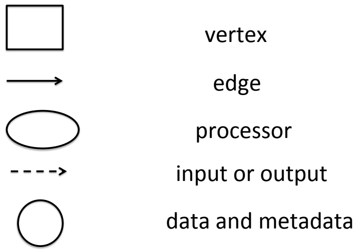


Figure 1. Notation of figures with workflows

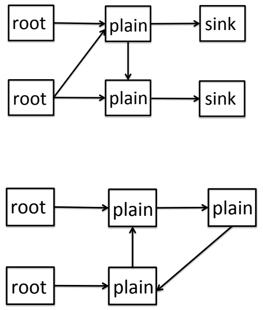


Figure 2. Workflow examples

2 WORKFLOW DEFINITION

The goal of the workflow is to enable the expression of the logical definition of user applications, which include *data processing*, i.e. *data accessing* and *computation*, as well as *dependencies* between instances of data processing. Computation may refer to algebraic computation or to more elaborate, algorithmic computation. The workflow models such applications as a graph. The vertices in the graph represent application logic and the edges represent the flow of data. Application logic includes (a) the analysis of data, and (b) the

[1] http://www.asap-fp7.eu

modification of data. Edges are directed and connect the vertices that produce and consume data. The rationale for adopting a graph model for the definition of a workflow is that the latter can enable the expression of application logic in an intuitive manner.
There are three types of vertices in a workflow, namely *root* vertices, *sink* vertices and *plain* vertices. The root vertices have only outgoing edges and they represent entry points of the application logic. Figure 1 explains the notation in all the figures that represent workflows and workflow parts. We require that each workflow has at least one root vertex. The sink vertices have only incoming edges and they represent final points of the application logic. We do not require that each workflow has one or more sink vertices. The vertices that are not of type root or sink, are plain vertices, which means that they have both incoming and outgoing edges. For applications that include many phases of data modifications or analysis, we expect that most vertices in respective workflows are plain, as they represent points in the application logic where data are both produced and consumed. Workflows that do not have sink vertices are those that express an application logic of continuous execution. It is easy to see that workflows without sink vertices are graphs with cycles. Figure 2 shows a workflow with two root and two sink vertices. and a workflow with no sink vertices, and, therefore, a cycle. Nevertheless, a workflow may comprise both acyclic sub-graphs and sub-graphs with cycles. A trivial case of such a workflow is one that expresses the logic of continuous querying that also outputs processed data, e.g. some final results to be archived. The formal definition of a workflow is the following:

Definition 1. . A workflow is a directed graph G = (V,E) where $V = V_r \cup V_s \cup V_p$ is a set that consists of three sets of vertices, the root V_r, sink V_s and plain V_p vertices. The three sets do not overlap, i.e. $V_r \cap V_s \cap V_p = \varnothing$, and there should be at least one root vertex, i.e. $V_r \neq \varnothing$. Also, $E = \{E_1, \ldots, E_m\}$ is a set of edges. An edge $E \in E$ is an ordered pair of vertices, i.e. $E = \{(V_i, V_j) \mid V_i, V_j \in V\}$.

Vertices and edges of workflows have properties. The properties of a vertex are related to tasks of the application part represented by this vertex, as well as corresponding metadata. The properties of an edge are related to data flow, and respective metadata, represented by this edge.

3 VERTICES

Each vertex in a workflow represents one or more tasks of data processing. Each task $\mathcal{T}$ is a set of *inputs*, *outputs* and a *processor*. An input inputs data to a processor; the latter represents the core of the data processing of the task, and, furthermore, an output outputs data generated from the processor. Therefore, inputs and outputs are related to descriptions of data and respective metadata. Figure 3 shows task examples: two tasks that have a shared input and one output each, and a task with two inputs and two outputs.

Definition 2. A vertex $V \in V$ corresponds to non-empty set of tasks $T \neq \varnothing$ such that each task $T \in T$ is a set of inputs I, outputs O and a processor P, i.e. T = {I,O,P}. Each input $I \in I$ and output $O \in O$ is a pair of data D and metadata descriptors M , i.e. $I = (D_I, M_I)$ and $O = (D_O, M_O)$.

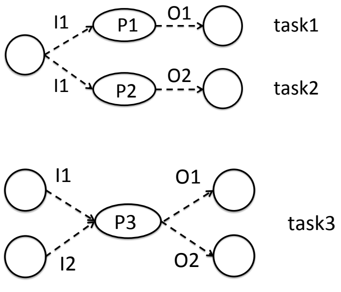


Figure 3. Task examples

As defined, a vertex may represent one or multiple tasks of the application logic. These tasks may share or not inputs, but they do not share processors and outputs. The inputs and outputs of the tasks of a vertex can be related to incoming and outgoing edges of this vertex, but they do not identify with edges: inputs and outputs represent consumption and production of data, respectively, and edges represent flow of data. Similarly, vertices do not identify with processors. This semantic differentiation is necessary in order to allow the management of the dependencies in the workflow through graph manipulation, separately from the management of data processing and computation in the workflow. Hence, it is easy to see that a vertex of any type, root, sink, or plain, may consist of tasks with non-empty sets of inputs and outputs, since the latter do not imply the existence of incoming or outgoing edges, respectively. The incoming and outgoing edges of vertices are related in a 1-1 fashion with inputs and outputs, respectively, of vertices. Therefore, if E_I and E_O are the sets of incoming and outgoing edges, respectively of a vertex V, and T is the corresponding set of tasks, then $| E_I | \subset |\cup_{\mathrm{T},I}|$, $\forall \mathrm{T} \in T$ and $|E_O| \subset |\cup_{\mathrm{T},O}|$,$\forall \mathrm{T} \in T$.

Figure 4 shows an example of a vertex with two tasks (it shows the detailed representation and the simplified representation where the cycles that represent data are omitted). The tasks share input I1 and each has one output, O1 and O2 that each are an input to an edge outgoing from this vertex. The input I1 is the output of an edge incoming to this vertex. Also, one of the tasks has one more output, O3, which is an additional input, I3, of the other task. The input/output O3 / I3 is not related with any edge, meaning that these data are not input to tasks that correspond to any dependent vertex. Figure 5 shows a vertex with one task, which creates a histogram of the input data. The task outputs the histogram itself and a set of additional statistics. These two outputs are separated and are input to two different edges to feed two different tasks. The histogram is further processed and the statistics are logged.

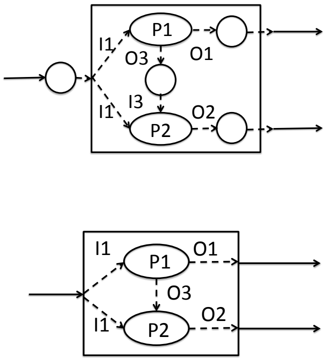


Figure 4. A vertex with multiple tasks

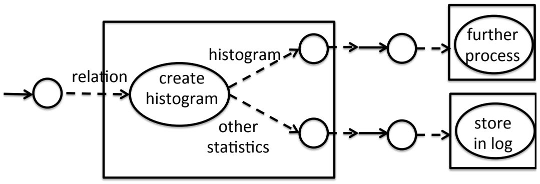


Figure 5. A vertex with a task with two outputs

A vertex needs to correspond to at least one task, but it can also correspond to more than one task of the application. Such tasks may or may not adhere to any sort of relation, e.g. concerning associations or similarities of their inputs, outputs or processors. Nevertheless, the reason why the proposed model allows the definition of vertices with multiple tasks, is to enable the user to express such associations or similarities. Therefore, the definition of a vertex in a way that it consists of multiple tasks, enables the definition of workflows that are intuitive with respect to the rationale of the application logic.

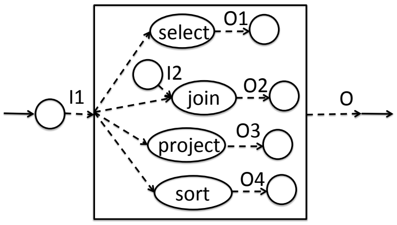


Figure 6. A vertex that represents a SQL query

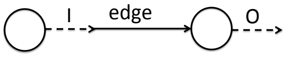


Figure 7. Edge with the respective input and output

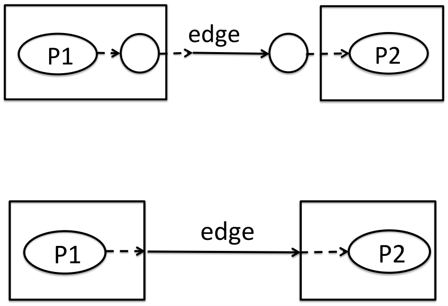


Figure 8. Edge connecting two vertices with one task each

Figure 6 shows a vertex that represents a SQL query. The vertex includes separate tasks for different parts of the SQL query. All the tasks share the input data, and one of them, the task that represents a *join*, has an additional input. Each task has an output. Note that the output data of this vertex, which is the input to the outgoing edge, is actually the output data after executing the whole group of tasks represented by this vertex. Notably, the output of the vertex can be any of the O1, O2, O3, O4 depending on the execution plan of this group of tasks. Allowing the user to define vertices with multiple tasks, enables her to represent in a vertex part of the application logic that she considers to be, conceptually, one unified complex task, without requiring her to define at the same time the way that this complex task should be executed, i.e. the execution semantics of it.

The processors in tasks realise the application logic, which is, as mentioned earlier, the analysis or the modification of data. Section 6 gives details on the concept of the processor and discusses proposed instantiations.

4 EDGES

Each edge in a workflow corresponds to a pair of an input I and an output O of the same data D. As mentioned, the I and the O of an edge correspond in a 1-1 fashion to an I and an I, respectively of a task. The data D are accompanied by metadata M, which can be different for the input and the output of the same edge. Figure 7 depicts the input and the output of an edge. Figure 8 shows an edge that connect two vertices, with one task each. The output of one task becomes the input of the other, via the dependency created by the edge connecting the two vertices. (The figure shows the detailed and the simplified representation of this example, where the cycles representing the data are omitted.) Formally:

Definition 3. An edge $E = (V_i, V_j)$, $V_i, V_j \in V$ in the workflow corresponds to a pair of an input and an output (I, O). Input I is a pair of data D and some metadata M_I, i.e. $I = (D, M_I)$, and output O is a pair of data D and some metadata M_O, i.e. $O = (D, M_O)$. Input I is equivalent with an output O' of a task that corresponds to vertex V_i, i.e. $\exists\ T \in V_i.\mathit{T}$, $\exists\ O' \in V_i.T.O$ such that $I \equiv O'$. Also, output O is equivalent with an input I' of a task that corresponds to vertex V_j, i.e. $\exists\ T \in V_j.\mathit{T}$, $\exists\ I' \in V_i.T.I$ such that $O \equiv I'$.

Hence, an edge defines the flow of data from one vertex to another according to some metadata that describe production and consumption information for these data. The production and consumption information can be the same or different and are related to (a) the data flow (b) the data persistence (c) the data reliability. Other types of metadata may be added in future work. In general, such metadata can be any information that plays a role in determining the execution plan of the workflow. Specifically, the metadata types are:

Flow. Metadata concerning the data flow pattern. The values of this type may be the following:

- Batch: The data flow in batches. Concerning input data, this means that they are consumed in batches (e.g. all data need to be available and accessed for processing to begin), and concerning output data, this means that they are available in batches, (e.g. produced data are stored in memory but they cannot be accessed until they are dumped in permanent storage).
- Stream: The data flow as a stream. Concerning input data, this means that they are consumed in a streaming manner (i.e. data are processed as soon as they are available as an input), and concerning output data, this means that they are available in a streaming manner, i.e. as soon as they are produced (e.g. either from memory or from permanent storage).

Persistence. Metadata concerning the data persistence in storage. The values of this type may be the following:

- Persistent: The data are stored in permanent storage. Therefore, concerning output data, they remain available after a task completes, and concerning input data, they remain available during and after they are used as input.
- Volatile: The data are not stored in permanent storage or they are stored only for a limited time. For example, concerning output data, they may not remain available after a task completes, and concerning input data, they may not remain available after they are used as input.

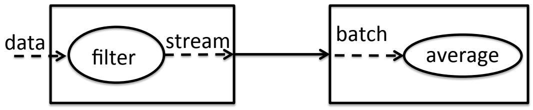


Figure 9. Example of an edge with different flow metadata of input and output data

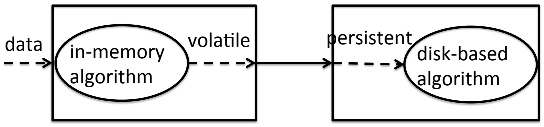


Figure 10. Example of an edge with different persistence metadata of input and output data

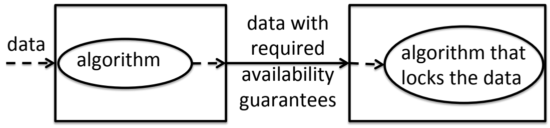

data
algorithm
data with required availability guarantees
algorithm that locks the data

Figure 11. Example of an edge with different availability metadata of input and output data

Reliability. Metadata concerning the correctness or completeness of data. The values may be the following:

- Reliable: The data are considered to be correct and complete[2].
- Non-reliable: The data are considered to be either incorrect and/or incomplete.

The above values of the metadata types may, even further, have properties, especially related to quantification. For example, the flow values may be characterised by size (e.g. the size of the batches or streams or the range of their varying sizes) or therate of their availability in time; the persistence values may be characterised by the lifetime of data or required

[2] Correctness and completeness of data may have application-specific semantics.

guarantees for availability; and the reliability values may be characterised by probability estimations.

As mentioned, the metadata of the input and the output pair of an edge can be different. Thus, the values for the same metadata type for the input and the output of an edge can be different. For example, concerning the data flow, the input value can be ’batch’, while the output value is ’stream’. Figure 8 shows an edge that connects a vertex with a task that filters some data, with a vertex with a task that computes the average of the filtered data. The filtering task outputs data in a streaming manner, but the task that computes the average takes as a input a batch of data. Therefore, even though the input and output data‘ of the connecting edge is the same, the respective metadata concerning the flow of data differ.

Concerning the data persistence, for example, the output value can be persistent with a lifetime of 1 sec, while the input value can be ’volatile’, meaning no lifetime at all after consumption. Figure 10 shows an edge that connects a vertex with an in-memory algorithmic task, with a vertex with a disk-based algorithmic task. The in-memory algorithm outputs data that are stored in the main memory and not in the disk, thus they are volatile; whereas the disk-based algorithm reads these data from the disk. Therefore, the persistence metadata of the input and output of the connecting edge are different.

Figure 11 shows another case of different metadata with respect to data persistence. The input and output data of the edge connecting two vertices are both persistent, i.e. stored in the disk, but their availability requirements differ. The algorithm in the left vertex outputs one copy of some data; the algorithm in the right, i.e. the dependent, vertex, locks this copy in order to process the data; yet, these data should be always available for reading by other tasks.

Furthermore, concerning data reliability, for example, the input and the output data may have different guarantees. A task may output data with some reliability guarantee, e.g. data can be considered correct with a 0.5 probability, whereas the task thattakes this data as a input may require that they can be considered correct with a 0.8 probability.

It is interesting to note that combinations of the metadata values for the input and output data of an edge create different execution plans.

5 DATA

The data D of inputs and outputs of edges, as well as of inputs and outputs of tasks consists of information on the data source where these data reside, as well as information on the data *unit*. Formally:

Definition 4. The data D of an input I = (D, M_I) or an output O = (D, M_O) is a set D = {S, u, *A*}, where S is the data source, u is the basic data unit and *A* includes additional information. The data source is a pair S = (n, t) of the name n and the type t of the source. The unit u takes values from a constraint domain D, which includes the names of the basic units for known types of data sources.

The type t of a data source can be one of the well known ones, e.g. ‘relational’, ‘rdf’, ‘xml’, ‘key-value’ etc. The unit for each type is unique and pre-specified; e.g. the unit of the relational type is the ’tuple’, the unit for the ’rdf’ type is the ’triple’ and the unit for the ’key-value’ type is the ’pair’. Moreover, data may include the description of additional information, such as relation and attributes names, as well as schema information (e.g. primary and foreign key constraints) and information on the respective processing engine, (e.g. engines of NoSQL databases, relational DBMSs like MySQL, etc).

6 PROCESSORS

The tasks included in vertices take as input data and metadata, process the data using a processor and output some data and metadata. Each processor can have an abstract definition

and several implementations, i.e. one or more implementations per platform. For example a processor that implements a ‘join’ for two data inputs, has an abstract definition, and can be implemented for a relational DBMS and a NoSQL database. In order for a processor to be used on a specific platform, it is required that this processor is implemented for the specific platform. The same holds for processors that perform more complex operations, such as algorithmic computation. A processor definition includes restrictions on the type and number of inputs and specifies the number and type of outputs. Defined and implemented processors form a library from which a user can select processors to describe tasks. Users can define their own processors and should provide respective implementations, in which input and output data can be in the form of raw bytes/records/key-value pairs etc.
In the following we give examples of the definition of basic processors, namely the *select*, *calc* and *join*:

$O(select, I) = \{r \mid r \in I \wedge SelectPredicate(r)\}$
$O(calc, I) = \{r \cup \{attr{:}vaue\} \mid r \in I \wedge value := CalcExpression(r)\}$
$O(join, I_1, I_2) = \{t \cup s \mid t \in I_1 \wedge s \in I_2 \wedge joinPredicate\ (t \cup s)\}$

The input and output data of a processor are accompanied by metadata that describe their type, format and other characteristics. The metadata defined for each processor have a generic tree format (JSON, XML etc). In order to allow for extensibility, the first levels of the meta-data tree are predefined; yet, users can add their ad-hoc subtrees to define their customized processors.The generic metadata tree for data definitions is the following:

```
{<input | output>: {
"constrains": {
"data_info": {
"attributes": {
<attr1..n>: {
"type": <type>}},
"engine": {
"DB": <db_meta>}}
}}
```

The generic meta-data tree for processors is:

```
{<operator_name>: {
"constrains": {
<input1..n>
<output1..m>
"op_specification": {
"algorithm": <alg_meta>}}
}}
```

<alg_meta> for considered processors:

```
{"select": { "select_condition":
 <select_predicate>
}}
```

All rows for which the <select_predicate> is ‘True’ are returned by the processor. Furthermore:

```
<select predicate> ::={<field name>
<comparison> <value>}
<comparison> ::= ['>' | '<' | '>=' |
'<=' | ' ==' | '! =']
```

```
{"calc": [{
"calc_attr": <calc_attr>,
"calc_value": <calc_expression>
}]}
```

The *calc* processor produces data with new attribute <calc_attr> and a value calculated by <calc_expression>.

```
<calc expression> ::= <attr name>
<action operator> [<attr name> |
<value>]. <action operator> ::= [['+'
| '-' | '*' | '/'] | ['concat' |
'substring'] | ['u' | 'n' | '\']]
```

The <value> and <action operator> depend of the attribute type.

```
{"join": { "join_condition":
<attr_name | attrs equality>
}}
```

The *join* processor has a minimum 2 inputs. The following is an example of a *join* instantiation:

```
{"joinOp": {
```

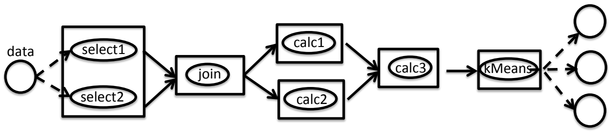
```
"constrains": {
"input1": {
"data_info": {
"attributes":["$attr1","$attr2"]}},
"input2": {
"data_info": {
"attributes":["$attr1","$attr2"]}},
"output1": {
"data_info": {
"attributes":["$attr1","$attr2"]}},
"op_specification": {
"algorithm": {
"join": {
"join_condition":
  "input1.$attr1=input2.$attr1"}}}}
}}
```

7 EXAMPLE USE CASES

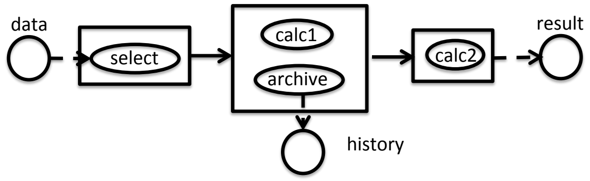
Telecommunication analytics applications may include many possible workflows, on-line queries, off-line long-running computations and ad-hoc analytics. This section describes two examples of the proposed workflow model applied on real use cases from the ASAP research project.

7.1 Traffic Jam Detection

An important use case of telecommunications analytics is the detection, and, further, the prediction of traffic jams. The telecommunication data are anonymised at regular times. The use case involves the processing of the anonymized location data for clustering along time and space. Figure 12 (a) depicts the respective workflow. Data with respect to location and time predicates are selected and joined. Then parallel processing, *calc1* and *calc2*, calculate ranges of space coordinates and time periods, respectively. These are further processed to caclulate speed, *calc3*. The speed values are input to an implementation of the k-Means algorithm that performs clustering according to a pre-defined set of criteria. Criteria may include specific area, the cut-off speed, or other parameters of the clustering algorithm. The results of the clustering algorithm are stored to several databases (relational and graph databases). The computed and stored clusters can be queried to discover traffic that occurs with regularity (detecting transport system bottlenecks) or without any regularity (anomalies, car accidents etc).



(a) Workflow for the detection of traffic jams



(b) Workflow for the detection of peaks

Figure 12. Workflows for example use cases

7.2 Peak Detection

Another important use case is the detection of peaks in the telecommunication traffic. This use case focuses on a dataset named Call Detail Records (CDR), which stores records of calls and is anonymised. CDR contains only recent data, e.g. the data of thelast day. Figure 12 (b) depicts the respective workflow. The use case involves processing of the anonymized CDR data by first selecting a spatial region and a temporal period (*select*). For this region and period, the number of calls is calculated (*calc1*). Data and calculations from CDR are archived (*archive*) in other storage (*history*). After calls are count, the application proceeds with algorithmic processing that detects peaks (*calc2*). The objective of this processing is to detect peaks in load, according to a set of criteria. Criteria may include the minimum size of a region and/or period, the cut-off distance, or other parameters for selecting regions and periods. These parameters should be adjustable by the analytics engineer, marketing expert, etc., who uses the peak analysis results. The results of this workflow are added to a database (relational or graph DBMS) that contains peaks detected in previous data. The database of peaks can then be queried by a user to discover clusters of calls that occur with regularity e.g., every week, discover clusters of calls that occur without any regularity, or similar queries based on the pre-computed peak data.

8 RELATED WORK

Workflow management systems have emerged in order to provide easy-to-use specification of tasks that have to be performed during data analysis. An essential problem that these systems need to solve is the combination of various tools for workflow execution and optimization over multiple engines into a single research analysis / system. The field of workflow management is a relatively new field of research, but there are already some promising results.

One of the oldest research projects to deal with the general problem of querying multiple heterogeneous data sources is Artemis [1]. Artemis uses ontologies and metadata, and integrates metadata in terms of semantics. The project identifies the problem of continuous metadata integration. The proposed workflow model enables the creation and execution of associative tasks that process and integrate intermediate results.

The system HFMS [2] builds on top of previous work on multi-engine execution optimization [3]. Their study is more focused on optimization and execution across multiple engines. The design of flows in HFMS is agnostic to a physical implementation. Datasets need not be bound to a data store, and operators need not be bound to an execution engine. HFMS handles flows as DAGs (i.e. Directed Acyclic Graphs) encoded in xLM, a proprietary language for expressing data flows. Alternatively, flows may be written in a declarative language (e.g., SQL, Pig, Hive) and imported. characteristics like memory budget), and so on. The proposed workflow model aims at modularity of workflow manipulation, expressibility of application logic, and adaptability to the user interests and role, goals that are out of the scope of the HFMS flow model definition.

Pegasus [4] is another workflow management system that allows users to easily express multi-step computational tasks. The workflows are formalizedso that tasks are represented as nodes and task dependencies as edges. The workflow description is separated from the description of the execution environment [5] and [6]. This allows the system to perform optimizations at ‘compile time’ and/or at ‘runtime’. A drawback of this approach is that the executing workflow be different than what the user anticipated when she submitted the workflow. As a result, in Pegasus a lot of effort is devoted toward developing a monitoring and debugging system that can connect the two different workflow representations in a way that makes sense to the user. The proposed workflow model overcomes such problems, by separating the definition of the dependancies and the processing tasks in the application logic. In this way, the user controls the detail of execution semantics she describes.

Taverna [7] is an open source domain-independent workflow management system, which includes a suite of tools used to design and execute scientific workflows. Research in [8] is focused on the issue of the analysis of data from heterogeneous and ‘incompatible’ sources. While Taverna includes tools for the composition and enactment of bioinformatics workflows, the composition of workflows is done through a graphical user interface and does not provide sophisticated methods for their efficient execution.

Apache Tez [9] is an extensible framework for building high performance batch and interactive data processing applications. Tez models data processing as a DAG. The task design is based on inputs and outputs that exist in pairs; outputs generate ‘DataMovementEvent(s)’, which are processed by inputs. The key difference from the proposed model is that in the latter inputs and outputs of dependent tasks are not connected and do not have to exist in pairs. Our model dictates that inputs and outputs of edges, and not of tasks, represent the dependencies between tasks and realise the relation between their inputs/outputs. In this way, tasks are inherently independent from each other, allowing modular manipulation of tasks and task groups, as well as separate manipulation of task execution and dependencies in the application logic.

Finally, other projects have also dealt with the problem of workflow definition and execution. The Stratosphere project [10] tackles the challenge of executing workflows with the PACT programming model, based on the Nephele execution engine [11]. This approach introduces the notion of workflows in cloud-based systems, but the solution is not mature enough to give the necessary efficiency or full-fledged capabilities of adaptive execution.

9 CONCLUSIONS

This paper proposes a novel workflow model for the expression of analytics tasks on Big Data. The model enables the separation of task dependencies from task functionality. Using the proposed model, a user is able to express a variety of application logics and to set her degree of control on the execution of the workflow. Finally, we depict the proposed workflow model on specific use cases from the telecommunication and web analytics domains. Ongoing and future work focuses on defining methods for the manipulation of the workflow structure in order to optimise its execution semantics.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n^{o} 619706 ASAP.

REFERENCES

[1] R. Tuchinda, S. Thakkar, Y. Gil and E. Deelman. Artemis: Integrating scientific data on the grid. In IAAA, 2004.

[2] A. Simitsis, K. Wilkinson, U. Dayal, and Meichun Hsu. Hfms: Managing the lifecycle and complexity of hybrid analytic data flows. In ICDE, pages 1174–1185, 2013.

[3] A. Simitsis, K. Wilkinson, M. Castellanos, and U. Dayal. Optimizing analytic data flows for multiple execution engines. In SIGMOD, 2012.

[4] Pegasus. http://pegasus.isi.edu/.

[5] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny and K. Wenger. Pegasus: a workflow management system for science automation. Future Generation Computer Systems, 2014.

[6] M. Malawski, G. Juve, E. Deelman, and J. Nabrzyski. Cost- and deadline-constrained provisioning for scientific workflow ensembles in iaas clouds. In SC, 2012.

[7] T. Oinn, M. Addis, J. Ferris, D. Marvin,T. Carver, M. R. Pocock, and A. Wipat. Taverna: A tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20, 2004.

[8] K. Wolstencroft, R. Haines, D. Fellows, A. R. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M. P. Balcazar V., S. Sufi, and C. A. Goble. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic Acids Research, 41(Webserver-Issue):557–561, 2013

[9] Apache tez. http://hortonworks.com/hadoop/tez/.

[10] A. Alexandrov, R. Bergmann, S. Ewen, J. –C. Freytag, F. Hueske, A. Heise, O. Kao, M. Leich, U. Leser, V. Markl, F. Naumann, M. Peters, A. Rheinla nder, M. J. Sax, S. Schelter, M. Hoger, K. Tzoumas, and D. Warneke. The stratosphere platform for big data analytics. VLDBJ, 23(6):939–964, December 2014.

[11] D. Battre, S. Ewen, F. Hueske, O. Kao, V. Markl, and D. Warneke. A programming model and execution framework for web-scale analytical processing. In SoCC, 2010.