Big Data Paradigm- Analysis, Application, and Challenges

U. Z. EDOSIO
School of Engineering, Design and Technology
University of Bradford

***Abstract*- This era unlike any other is faced with explosive growth in the sizes of data generated. Data generation underwent a sort of renaissance, driven primarily by the ubiquity of the internet and ever cheaper computing power, forming an internet economy which in turn fed an explosive growth in the size of data generated globally. Recent studies have shown that about 2.5 Exabyte of data is generated daily, and researchers forecast an exponential growth in the near future.**

This has led to a paradigm shift as businesses and governments no longer view data as the byproduct of their activities, but as their biggest asset providing key insights to the needs of their stakeholders as well as their effectiveness in meeting those needs. Their biggest challenge however is how to make sense of the deluge of data captured.

This paper presents an overview of the unique features that differentiate big data from traditional datasets. In addition, we discuss the use of Hadoop and MapReduce algorithm, in analyzing big data. We further discuss the current application of Big Data. Finally, we discuss the current challenges facing the paradigm and propose possibilities of its analysis in the future.

Keywords: Big Data, Large Dataset, Hadoop, Data Analysis

I. Introduction

Over the last two decades of digitization, the ability of the world to generate and exchange information across networks has increased from 0.3 Exabyte in 1986 (20 % digitized) to 65 Exabyte's in 2007 (99.9 % digitized) [1].

In 2012 alone, Google recorded a total of 2,000,000 searches in one minute, Facebook users generated over 700,000 contents, and over 100,000 tweets are generated per minute on Twitter.

In addition, data is being generated (per second) through: telecommunication, CCTV surveillance cameras, and the "Internet of things" [2].

However, the "Big Data" paradigm is not merely about the increasing volume of data but how to derive business insight from this data. With the advent of Big data technologies like Hadoop and existing models like Clustering algorithm and MapReduce there is much promises for effectively analyzing big data sets. Notwithstanding, there are a lot of challenges associated with Big Data analysis, such as: Noise accumulation and highly probabilistic outputs; Data privacy issues and need for very expensive infrastructure to manage Big data [3]. Fig 1 Illustrates the exponential growth of data between 1986 and 2007.

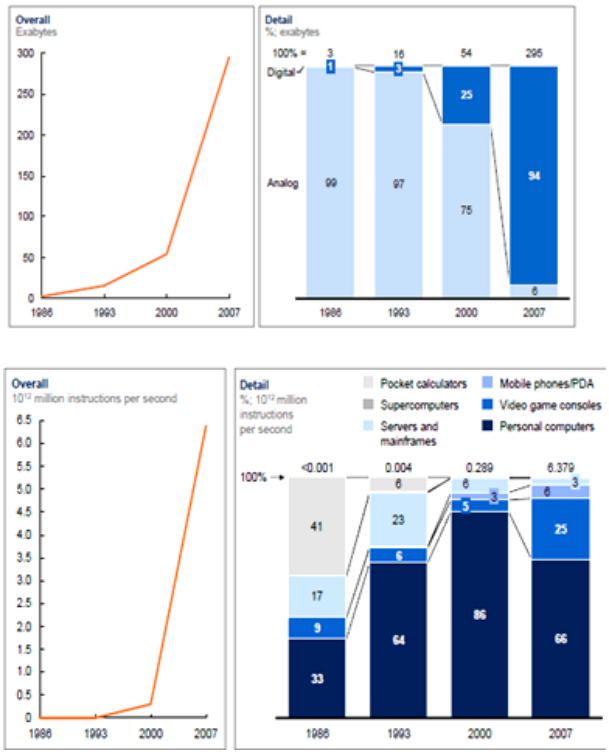

Fig.1. Data growth between 1986 and 2007[1]

II. Definition of Big Data

Currently, there is no single unified definition of "Big Data", various researchers define the term either based on analytical approaches, or its characteristics. For instance; according to Leadership Council for Information Advantage, Big Data is summation of infinite datasets (comprising of mainly unstructured data) [2]. This definition focuses solely on the size/volume of data. Volume is only one feature of Big Data, and this does not uniquely differentiate Big Data from any other dataset. On the other hand, some researchers define Big Data in terms of 3 characteristics, volume, velocity and variety also known as the 3 V [2]'s.–Variety depicts its heterogeneous nature (comprising of both structured and unstructured datasets), velocity represent the pace to which data is acquired, and volume illustrates the size of data(usually in Exabyte , Petabytes and Terabytes). This is a holistic definition of the term Big Data; as it encapsulates key features that uniquely identify Big Data and it opposes the common notion that Big Data is merely about data

size[2][4]. Fig. 2 illustrates the "3 V" characteristics of Big Data, and also highlights how these characteristics uniquely identify Big Data.

Recently, some researchers have proposed another V –"Veracity" as a characteristic of Big Data. Veracity deals with the accuracy and authenticity of Big Data. In this paper we will focus solely on the volume, velocity and velocity as this is more widely accepted characteristics [5].

Analysis in Big Data refers to the process of making sense of data captured [4]. In order terms it can be seen as systematic interpretation of Big Data sets to provide insights or business intelligence which can foster business intelligence[6][7].

III. Characteristics of Big Data

This section explains each of the 3v characteristics of Big Data.

A. *Big Data Volume*

According to [8], in 2012, it was estimated that about 2.5 Exabyte of data was created. Researchers have further forecasted that, these estimates will double every 40 months.

Various key events and trends have contributed tremendously to the continuous raise in the volume of data. Some of these trends include:

1) *Social Media*: There has been significant growth in the amount of data generated from social media sites .For instance, an average Facebook user creates over 90 contents in a month. Also, each day there are about 35 million status updates on Facebook.This is just a small picture considering that there are hundreds of social media application fostering users interaction and content sharing daily [9][10]. Fig. 3 provides more insight on the amount of data generated from social media sites in 60 seconds.

2) *Growth of Transactional Databases*: Businesses are aggressively capturing customer related information, in order to analyze consumer behavior patterns [9].

3) *Increase in Multimedia Content*: Currently multimedia data accounts for more than half of all internet traffic. According to the Internet Data Corporation the number of multimedia content grew by 70% in the year 2013[9].

B. *Big Data Variety*

This characteristic is also referred to as the Heterogeneity of Big Data. Big Data is usually obtained from diverse sources, with different data type, such as: DNA sequences, Google searchers, Facebook messages, traffic information, weather forecast amongst others.

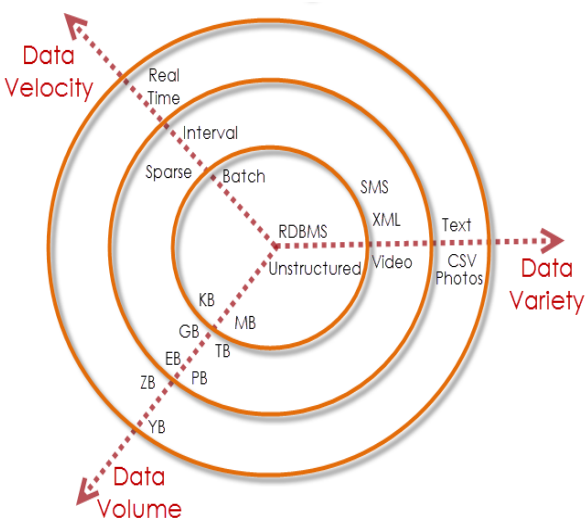

Fig. 2. Characteristics of Big Data **[7]**

Specifically these data can be generically categorized into structured, unstructured, semi-structured, and mixed [2]. This data is obtained from wide variety of sample sizes (cutting across age, geographical location, gender, religion, academic backgrounds).

C. *Big Data Velocity*

This refers to the speed at which data is generated [2]. Data can be captured either real time, in batches or at intervals. For instance web analytics sensors usually capture number of clicks on a website, (by utilizing specialized programming functions which listening to click event) on real time basis. For every click, the web sensor analytic is updated immediately (real time). However, in some cases data is captured in batches, an instance of this is- Bank daily transaction data- this is reviewed in batches at the end of each day.

Big Data analytic systems unlike traditional database management systems have the capacity of analyzing data in real time to provide necessary insight immediately.

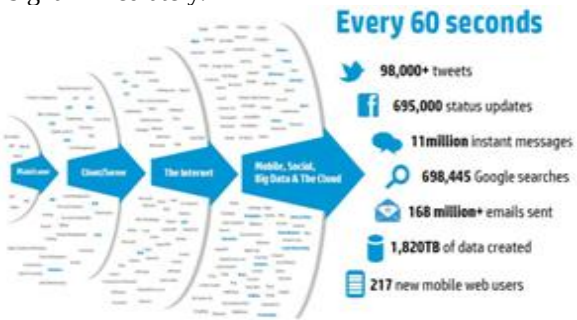

Fig. 3. Sources of Big Data [11]

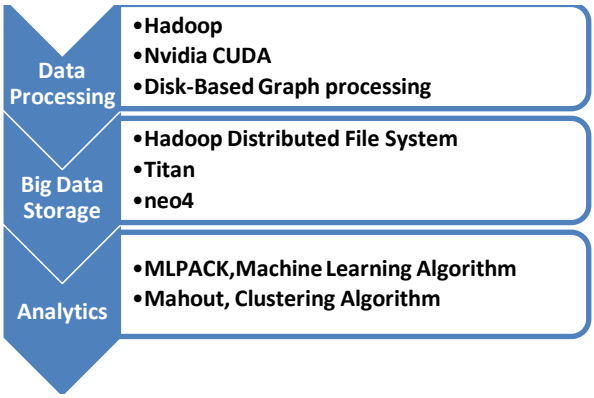

Fig. 4: Technologies for Big Data Analytics **[12]**

IV. Technologies for Analysising and Managing Big Data

Due to the 3v properties mentioned above, it is impossible to process, store and analyze Big Data using traditional relational database (RDBMS). However, there are myriads of individual technologies and libraries which provide an overall Big Data analytics framework (when combined together). Fig 4 (above) highlights each of these technologies and how they fit into the overall framework. However, this report focus only on the Hadoop architecture

A. Hadoop Architecture

Hadoop is an open source software built by Apache. The software provides a platform for managing large dataset, with 3V characteristics. In order words it allows for effective and efficient management of Big Data.

Hadoop consists of a data management system, referred to as the Hadoop distributed storage file system (HDFS) - this is a distributed file system that processes and stores Big Data sets.

The HDFS is responsible for dividing Big Data (both structured and unstructured data) into smaller data blocks. After which the HDFS performs the following on the data chunks:

1) *Creates replicas of the smaller data*: HDFS has an inbuilt fault tolerance function, which creates replicas of data, and distributes it cross various data blocks (this good incase of disaster recovery) [3].

2) *Distribute smaller data sets to slave nodes*: Each of these smaller data blocks are spread across the distributed system, to slave nodes.

3) *MapReduce:* using a special functions/task referred to Map() and Reduce() respectively, the HDFS can effectively process this data, for effective analysis [13].

B. MapReduce in Hadoop

MapReduce is a programming model integrated into Hadoop, which allows for parallel processing of individual data blocks or chunks in the HDFS[13]. This model comprises a Map() task - which is responsible for breaking data chunks into more granular forms for easy analysis. While the Reduce() task is responsible for fetching the outputs of the map task and aggregating it into a more concise format, which is easy to understand. A function called Maplogik connect enables storage of final output from the Reduce task to the HDFS[13] [11]. All tasks on a single node are managed by a Task Tracker, and all Task Trackers in a cluster are managed by a single Job Tracker. Fig. 5. presents a diagram illustrating mapping of data using the Map task, and Reduce function to obtain a more concise result.

For Example:

MapReduce can be used in a scenario where one needs to obtain the number of word count in a document. The pseudo-code is illustrated in fig. 6.:

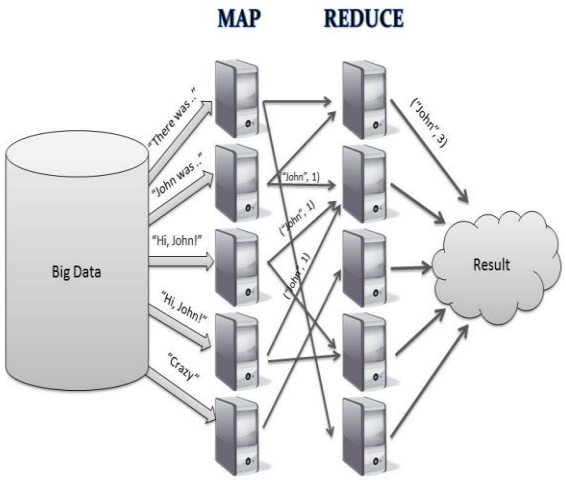

Fig. 5: Illustration of Map and Reduce task in Hadoop [14]

```
map(String key, String value):
  // key: document name
  // value: document contents
  for each word w in value:
    EmitIntermediate(w, "1");

reduce(String key, Iterator values):
  // key: a word
  // values: a list of counts
  int result = 0;
  for each v in values:
    result += ParseInt(v);
  Emit(AsString(result));
```

Fig. 6: MapReduce Algorithm [14]

V. Applications of Big Data

A. Business Intelligence

Based on a survey carried out by MIT Sloan Management Research, in collaboration with IBM, on over 3000 top management employees, it was discovered that top performing organizations adopt Big Data analytics five times more in their activities and strategies than any other [17]. Many organizations are increasingly adopting analytics in order to make more informed decisions and manage organizational risks effectively.

With the help of analytical tools like Hive, JAQL and Pig integrated with Hadoop HDFS, business can visualize (pictorially and diagrammatically) key insight from Big Data [6].

For example: Many organizations currently analyze unstructured data such as social feeds, tweets, chats, alongside structured data such as stocks ,exchange rates, trade derivative, transactional data in order to gain understanding and insight on consumers perception on their brands. By Analyzing these voluminous data, in real time, organizations can make calculated strategic decisions, in advertisement and marketing, portfolio, risk management amongst others[17].

B. Predictive Analytics in Ecommerce, Health Sector

The major principle applied in predictive analysis is studying of words accumulated from various sources, such as tweets, RSS feed(Big Data).

In 2008, Google was able to successful predict trends in swine flu outbreak based on analysis of searches only. This prediction was about two weeks earlier than the U.S. Center of Disease Control. With such vast sample size of real time data available, manual statistical models are quickly becoming a thing of the past[6].

Majority of the online stores are taking advantage of words tracking, to provide smart advertisements to customers. For example Amazon and E-bay analyze user online searches, by depositing cookies on browsers in order to study consumers' habit on the web. Based on results from this analysis, they can provide customized offers, discounts and advertisements based. Predictive analysis promises to provide competitive advantages to businesses by preempting "what if" scenarios [19].

C. E-Government and E-Politics

Online campaign was first used in the United States of America in 2008. These campaigns brought about great success and participation by the general public. Politicians used the web to publicize their policies, events, discussions, and donations. These campaigns were rich in multimedia contents and interactive with the general public [6].

As e-government is gaining more grounds, politicians are taking advantage of Big Data analytics, together with business intelligence tools to analysis voter's perception. They are also democratizing campaigns based on different audiences. This is made possible by adopting clustering algorithm and HDFS system to cluster different set of audiences. Other applications of Big Data in politics include opinion mining, social networking analytics.

VI. Challenges of Big Data Analysis

A. Data Privacy

Privacy is one of the major challenges facing Big Data paradigm. Majority of people are concerned about how their personal information is utilized. Big Data analytics tends to infringe on our daily lives taking advantage of the ubiquity of the internet. For instance, Big Data analysis may involve studying consumer's social interactions, shopping patterns, location tracking, communication and even innocuous activities like power usage at one's home. A commutation of all these data can uniquely identify an individual and serve as what is known as a "digital DNA"[3].

There major underlying data privacy challenges that have plagued the Big Data Analysis paradigm. These issues are not new to just Big Data but almost all ICT innovations.

It is involves issues pertaining to confidentiality (who owns data generated?, who owns the results from the analytic of the data?), integrity(Who vouches the accuracy of the data?, who would take responsibility of false positive data analysis?) , interoperability (who stipulates the standards for data exchange?)and availability [6] .

B. Noise Accumulation

Due to the heterogeneous nature of Big Data, statistician and scientist predict high noise accumulation in the Big Data sets. This feature is peculiar to datasets which are varied and voluminous from wide sample sources[3][20].

In the case of Big Data, majority of the data acquired are mainly based on estimations, incomplete and probabilistic in nature. For instance, imagine analysis of consumer behavior; however in this instance the study is based on a public system in a library. The data generated from such analysis will be highly dispersed as different individuals use that computer in their own fashion. Imagine that a Big Data sample set contained probably 50% of such noisy data. The resulting data will be incoherent. Hence, results generated from Big Data are not always truthful , since data is generated is highly prone to error.

In response to this challenge scientist are adapting machine learning to analyze Big Data. However this is impractical as the machine learning algorithms expect only homogenous input data[20].

C. Cost of Infrastructure

Over the past decades the data generation has out grown computer resources. Many computer manufacturers are constantly producing systems with higher processing powers and Hard disk space [20]. Currently, the traditional hard disk storage system is been replaced with solid disk drive state.

However, in the case of Big data due to the volume and the velocity of data generated per time, a more stable computer processor will be need in futures. Also with the advent of a new paradigm referred to as the "internet of things" there will be more data generated into the internet (data generated from objects and human, or object to object interaction). This would require large data ware houses.

These resources are very expensive to install and manage. Many organizations may find it too expensive to acquire and may choose to stick with the traditional means of data analysis.

VII. Conclusion

Although, the analytic framework for Big Data is not 100% accurate many organizations, are applying the technology regardless. According to Mckinsey in [1], majority of the top five business organizations in the USA claim to yielding tremendous growth.

Currently, statisticians and computer scientist are still searching for better statistical models, and algorithms to fine tune the noisy results and produce more precise and accurate insight from big data analysis.

However, I recommend there is need for more urgent research on stable hard ware systems and computational algorithms to manage and produce insights at optimum. As data growth is a going concern.

References

[1] J. Manyika, et al., "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Jun. 2011.

[2] M. D. Assuncoa, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data Computing and Clouds: Challenges, Solutions, and Future Directions," *arXiv*, vol. 1, no. 1, pp. 1-39, Dec. 2013.

[3] J. Fan, F. Han, and H. Liu, "Challenges of Big Data Analysis," *ResearchGate*, vol. 1, no. 1, pp. 1-38, Oct. 2013.

[4] P. Russom, "Big Data Analytics," *The Data Warehousing Institute*, vol. 4, no. 1, pp. 1-36, 2011.

[5] V. López, S. d. Río, and J. M. H. Benítez, "Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data," *Fuzzy Sets and Systems*, vol. 1, no. 1, pp. 1-47, Feb. 2014.

[6] M. Hilbert, "Big Data for Development: A Systematic Review of Promises and Challenges ," *United Nations Economic Commission for Latin America and the Caribbean (UN ECLAC)* , vol. 1, no. 1, pp. 1-36, Jan. 2013.

[7] J. Hurt. (2012, Jul.) The Three Vs Of Big Data As Applied To Conferences. [Online]. http://jeffhurtblog.com/2012/07/20/three-vs-of-big-data-as-applied-conferences/

[8] A. McAfee and E. Brynjolfsson, "Big data: the management revolution," *Harvard business review*, vol. 90, no. 10, pp. 60-68, Oct. 2012.

[9] J. Wielki, "Implementation of the Big Data concept in organizations – possibilities, impediments and challenge," in *FEDCSIS*, 2013, pp. 985-985.

[10] H. Keshavarz, W. H. Hassan, S. Komaki, and N. Ohshima, "Big Data Management: Project and Open Issue," *Malaysia-Japan International Institute of Technology*, vol. 1, no. 1, pp. 1-8, 2012.

[11] P. Rajesh and Y. M. Latha, "HADOOP the Ultimate Solution for BIG DATA," *International Journal of Computer Trends and Technology (IJCTT)* , vol. 4, no. 4, pp. 550-552, Apr. 2013.

[12] S. Chalmers, C. Bothorel, and R. CLEMENTE, "Big Data-State of the Art," Telecom Bretagne, Institut Mines-Telecom Thesis, Feb. 2013. [Online]. https://www.researchgate.net/publication/258868555_Big_Data_-_State_of_the_Art?ev=srch_pub

[13] H. P. Fung, "Using Big Data Analytics in Information Technology (IT) Service Delivery ," *Internet Technologies and Applications Research*, vol. 1, no. 1, pp. 6-10, Jan. 2013.

[14] J. Dean and S. Ghemawat, "MapReduce: Simplied Data Processing on Large Clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107-113, 2008.

[15] F. C. P. Muhtaroglu, G. T. TUBITAK-BILGEM, S. Demir, M. Obali, and C. Girgin, "Business Model Canvas Perspective on Big Data Applications," in *Big Data, 2013 IEEE International Conference*, Silicon Valley, CA, 2013, pp. 32-37.

[16] "Real-Time Urban Traffic State Estimation with A-GPS Mobile Phones as Probes," *Journal of Transportation Technologies*, vol. 2, no. 1, pp. 22-31, Jan. 2012.

[17] S. LaValle, et al., "Big Data, Analytics and the Path From Insights to Value," MIT Sloan Management Review Survey, 2011.

[18] F. D'Amuri and J. Marcucci, ""Google it!" Forecasting the US unemployment rate with a Google job search index," *Fondazione Eni Enrico Mattei Working Papers*, vol. 1, no. 1, pp. 1-54, Apr. 2010.

[19] A. Mosavi and A. Vaezipour, "Developing Effective Tools for Predictive Analytics and Informed Decisions," University of Tallinn, Technical Report, 2013.

[20] A. Labrinidis and H. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032-2033, 2012.

[21] "Big Data Computing and Clouds: Challenges, Solutions, and Future Directions".

[22] A. Rabkin, "How Hadoop Clusters Break," *IEEE*, vol. 30, no. 4, pp. 88-94, Jul. 2013.

[23] Y. M. ESSA, G. ATTIYA, and A. EL-SAYEd. (2013, Feb.) Mobile Agent based A New Framework for Improving Big Data Analysis . [Online]. http://www.researchgate.net/publication/258194086_Mobile_Agent_based_New_Framework_for_Improving_Big_Data_Analysis/file/60b7d52cbb2bef0bc4.pdf