Big Data in Cloud Computing: A taxonomy of risks

Holmes E. Miller
Department of Accounting, Business and Economics, Muhlenberg College
Allentown, Pennsylvania USA 18104

Abstract:
Introduction. *Technological advances enable storing, processing, analyzing, retrieving, and reporting large amounts of data residing on servers accessible by the Internet, referred to here as big data in cloud computing. As with many technological advances, this new environment creates new risks.*
Method. *A grid, coupling the uses of information with various levels of aggregation of information, informs a scenario-based analysis of risks that are discussed in three categories: Data sources, cloud computing, and big data.*
Analysis. *Data source risks are analyzed in four categories: Data quality, data privacy, data security, and data integration. Cloud computing risks are analyzed in these four categories plus vendor risk. Finally big data risks are analyzed.*
Results. *Given the above structure, over 40 separate risks are articulated and discussed. These risks can inform risk analyses performed by organizations that have migrated or are migrating to big data and cloud computing environments.*
Conclusions. *Given these risks, a sample five-step process for risk management is proposed. The overall conclusion is that risks in big data and cloud computing environments may be managed in ways beneficial for all affected parties, including those generating the data, those using the data, and all others profiting from it.*

INTRODUCTION

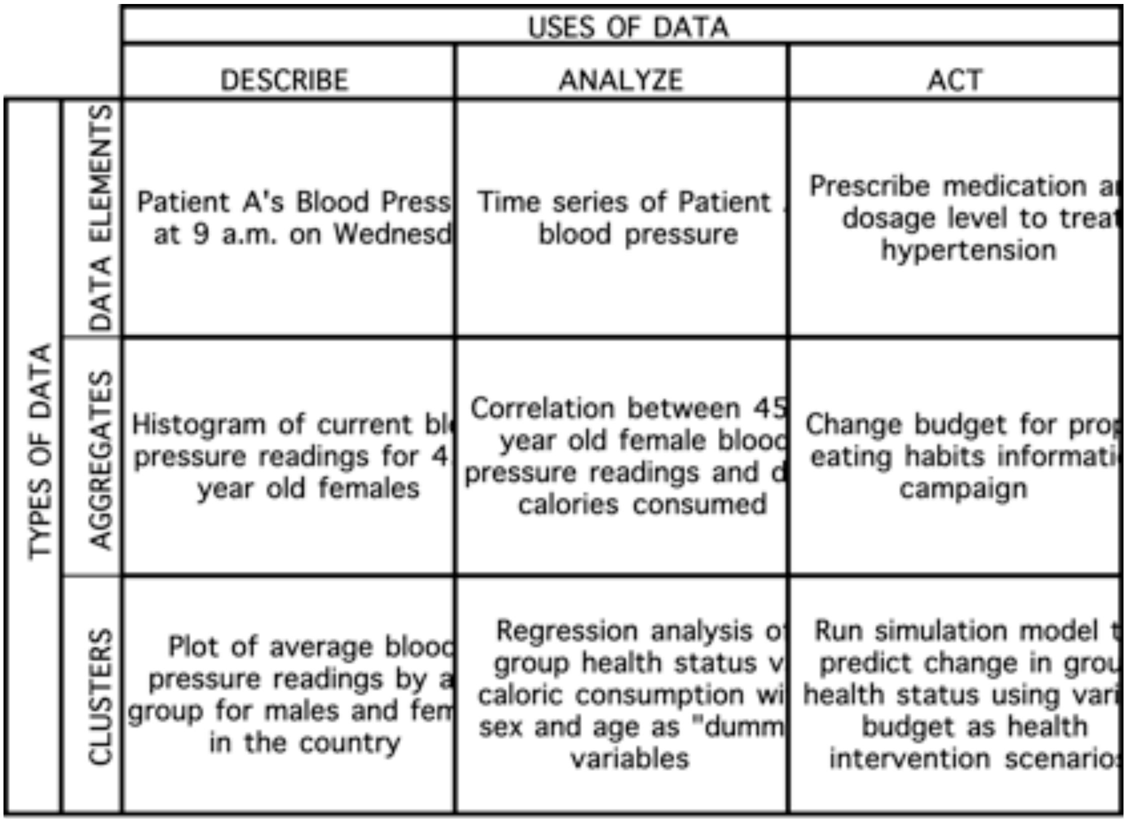
Cloud computing and big data are complementary approaches to storing, processing, analyzing, accessing, and reporting information. Cloud computing provides services (e.g., infrastructure, platforms, applications, and data) that run on servers accessible through the Internet rather than residing on the desktop. Cloud computing allows data to be uploaded and accessed from the desktop, from mobile devices such as tablet computers and cell phones, and from appliances such as vending machines and automobiles. Although perhaps conceptually nothing new vis-à-vis timesharing (Ellison 2008), cloud computing attributes including scale, accessibility, cost, and timeliness have created new capabilities that previously did not exist. These capabilities include company-centric uses of information where customer data can be aggregated to support company needs (Prabhaker 2000), and in the near future, customer-centric uses where *smart products* and data aggregation may be used to empower individuals to tailor purchases to their specific needs without resorting to intermediaries (Searls 2012). Whereas prior marketing approaches involved *pushing* advertisements onto customers, big data and cloud computing approaches facilitate customers *pulling* information from vendors to better satisfy the customer's individually defined and articulated needs.

The term *big data* refers to analytical technologies that have existed for years but can now be applied faster, on a greater scale, and are accessible to more users. Kusnetsky (2010, para. 3) defines big data as, 'the tools, processes and procedures allowing an organization to create, manipulate, and manage very large data sets and storage facilities.' While big data operations can be processed locally, as organizations migrate to the cloud, so will their corporate data. Moreover cloud-based architectures will become more important as individual entities (i.e., both appliances and people) generate continuous data streams that can be collected, stored, processed, analyzed and reported.

In this paper we will present and discuss risks and opportunities of big data applications in cloud computing environments (BD/CC) and present some ideas for managing these risks.

BIG DATA AND CLOUD COMPUTING OVERVIEW

To inform the discussion below, Figure 1 presents a grid (using a health care example) that indicates how data can be classified, aggregated, and used:



		USES OF DATA		
		DESCRIBE	ANALYZE	ACT
TYPES OF DATA	DATA ELEMENTS	Patient A's Blood Press at 9 a.m. on Wednesd	Time series of Patient blood pressure	Prescribe medication a dosage level to treat hypertension
	AGGREGATES	Histogram of current bl pressure readings for 4 year old females	Correlation between 45 year old female blood pressure readings and d calories consumed	Change budget for pro eating habits informati campaign
	CLUSTERS	Plot of average blood pressure readings by a group for males and fem in the country	Regression analysis o group health status v caloric consumption wi sex and age as "dumm variables	Run simulation model t predict change in grou health status using vari budget as health intervention scenario

FIGURE 1: TYPES AND USES OF DATA IN A HEALTH CARE CONTEX

The three columns present three processes associated with information: Describing the information; analyzing the information to draw meaning from it; and using the analysis to change the underlying environment. For example, when a patient has back pain *descriptive* attributes are collected. Some of these attributes might include the patient's identity, age, weight, and blood pressure; a description of the back pain; the patient's description of events that might have caused the back pain; and medications currently taken.

Analysis questions use this descriptive and other historical information, coupled with the physician's knowledge of back pain and the patient, to arrive at a diagnosis and a proscribed treatment. Further methods of analysis might include using correlation and extrapolation methods to explore the relationship between the medicine and exercise prescribed and changes in health status. Furthermore, these methods can be used to support cost benefit analysis or to forecast improvements in the state of a patient's health using different combinations of medicine and exercise regimes.

Data generated from *Actions* might include a record of minutes spent each day performing exercises to combat back pain, a record of taking the prescribed medications, and perhaps in conjunction with future *descriptive* data, changes in health status. The above elements concern *one patient*. Similar examples would apply to hierarchical groups of patients that then can form *aggregates* of individuals and *clusters* of those aggregates, as defined by a subset of the attributes.

The data environment for this example existed well before the era of cloud computing and big data and arguably, even before the era of electronic computing. Patient X's data was stored in the physician's office, most likely in paper-based files. Although access to the data was immediate, the data was meaningful only after the files were properly updated. Not immediately available, or most likely ever available, was information about patients in other physician's offices at other geographic locations that could be analyzed in concert with patient X. These analytical operations were limited by the available computing capabilities of the hardware and software and how data was processed to meet volume, timeliness, cost, and data requirements. The inability to access, analyze, report, and act on this information constituted lost opportunity in terms of data management and decision-making.

Cloud computing environments allow users to employ many disparate data sets to ask and answer questions regarding data generated in many different geographic locations and temporal states. This allows users to ask and answer many questions that can lead to insights heretofore unknown. Before cloud computing only questions about patient X's back pain might be addressed. Cloud computing technologies and big data can use information about back pain for all patients throughout a country or even between countries to examine the effectiveness of policies and treatment regimes across wider patient cohorts. Data storage in the cloud not only creates the ability to answer the same questions faster and with more veracity, but also opens up the possibility to review and analyze information that was previously inaccessible. The results from the data analysis are constrained only by the ingenuity in applying analytical methods and cloud computing technologies.

Discussion of Big Data in Cloud Computing Environments

In discussing big data pathologies Jacobs (2009) states that 'it is easier to get data in than out' (para. 19) and that the 'pathologies of big data are primarily those of analysis.' (para. 10) In the subsequent discussion he indicated that what makes big data *big* is the repeated observation over time and/or space, an attribute that favours cloud computing. Data may be generated in a network of millions or even billions of nodes. Data stored in the cloud can be analyzed and reported in the aggregate, or can be distributed back to the individual nodes themselves in a granular format. Regarding the latter option, McKendrick (2010) indicates that the centre of gravity in data management will continue to shift to end users, facilitated by downloadable apps, more intuitive search techniques and video game-like interfaces. This creates a scenario of end users both generating data to the cloud and processing data retrieved from the cloud. This creates many analytical possibilities and as we will see, risks.

Some factors facilitating the uses of big data in cloud environments include:

1. *No upper limit on the amount of information stored.* Big data storage amounts now are in the petabytes (10^{15} bytes) rather than terabytes or gigabytes. Indeed, given historical growth, one can surmise in the future terms such as exabytes (10^{18} bytes), zettabytes (10^{21} bytes), and yottabytes (10^{24} bytes) eventually will enter the conversation. Rather than being restricted by the size of a single storage media, cloud users have the ability to increase scale by adding and connecting servers and other storage media through networking and telecommunications technologies.

2. *Ability to access, integrate and analyze data from multiple sources and databases whose owners are geographically and even temporally disbursed.* Scalability potential creates the capability to collect and analyze more granulated data. Examples of this include: individual Google searches; scientific data generated in experiments in particle physics and genomics; changes in location using connected smart phones and other GPS enabled devices; ongoing status of equipment and other devices such as cars, home appliances, and medical equipment; and aggregating in real time opinions such as those enunciated using social media like Twitter and Facebook. This capability creates new worlds of potential applications and also creates new risks.

3. *The elimination of intermediaries – owners and users of the data can be individuals, not only large organizations.* Until recently, intermediaries in data use were required on two fronts. First, large data centres stored data and staff served as gatekeepers regarding how data was collected, analyzed, and disseminated. This led to an organization-centric view of information. Second, models and software used to analyze and report data often required a level of sophistication and cost that inhibited democratization of information. BD/CC environments, while not intermediary free, provide more freedom for an individual user to generate, store, access, analyze, and use information.

4. *Immediate access to personalized results using distributed input and output mechanisms.* Eliminating intermediaries through standard evolutionary processes leads to more creative applications and new insights. Because of the granular capabilities inherent in big data, many of these applications and insights are personalized, which allows individuals to receive specific information tailored to specific needs. Tailored information in turn, leads to users asking more questions and demanding even faster and more granular answers. And so the cycle continues.

5. *The capability to change the paradigm from 'model and analysis' to 'correlation and extrapolation'.* Chris Anderson, editor of *Wired* magazine, hypothesizes *the end of hypotheses*. He argues that (Anderson 2008) 'faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.' (para. 10) Instead, Anderson posits that correlation is enough. 'We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.' (para. 13)

While Anderson's views are not without detractors (Conway 2008, Morton 2008) large datasets when coupled with more powerful and accessible analytical methods can lead to democratization of information and also to new ways of creating and discovering information. One key new capability is that using correlation methods on large datasets, while recognizing that correlation does not imply causality, creates the capability to discover new, previously hidden insights.

OPPORTUNITIES FOR USING BIG DATA IN CLOUD COMPUTING

We will discuss opportunities for big data in cloud computing environments in three categories that correspond to types of data given in Figure 1: Data elements; data aggregates; and clusters of data aggregates.

Data Elements

Being able to capture and link existing data elements on larger, timelier, and more accessible scales creates new opportunities. On example of this is data captured by remote sensors. Bollier (2010, 3) points out regarding remote sensors, that 'remote sensors are generating new streams of digital data from telescopes, video cameras, traffic monitors, magnetic resonance imaging machines, and biological and chemical sensors monitoring the environment.' Personal data streams also may be generated by mobile phones, laptops, Websites, and mobile phone enabled GPS devices. A good example of aggregating data is discussed by Bryant *et al*. (2008) with regard to how Wal-Mart contracted with Hewlett Packard to store 4 petabytes of data in a data warehouse that encompass 267 million daily transactions made at 6000 stores, worldwide. Machine learning was then applied to this data to help Wal-Mart exercise various pricing and advertising strategies to better manage its inventories and supply chain. This example illustrates how one may not only describe the underlying entities of data sets but may also apply various analytical tools to analyze, correlate and extrapolate information to instigate future actions.

Aggregates

Big data in cloud environments can help organizations describe, analyze, and act on data aggregates. Whereas an individual product (e.g., shampoo) might be analyzed by evaluating the effectiveness of an advertising campaign, so might aggregates of all beauty products be analyzed regarding how pricing differences and store placement affects sales. The promise of BD/CC is to create *new aggregates* that provide new information that may yield newer and richer insights.

Bollier (2011) reports several interesting examples of how aggregation, combined with analysis and visualization techniques, yields surprising new insights. Google Earth maps sampled 8510 cows in 308 herds around the world and discovered two thirds of them aligned their bodies with magnetic north in the Earth's magnetic field. A second big data Google example is the Google Flu Trends service, where by tracking flu-related search terms, the service identified possible flu outbreaks in Atlanta a week or two before actual flu cases are reported by the Center for Disease Control.

These examples involve data elements that existed but either were unreported or if reported, neither aggregated nor analyzed. As Bollier points out, one of the benefits of big data in cloud environments is *right now* data. Hal Varian of Google has used the term *now casting,* using right now data coupled with analytical methods to create various innovative uses (Bollier 2010, 20). Google has many real time information streams but real time information also is available through monitoring of appliances, cell phones, cameras, and any other device that can report its attributes on an ongoing basis.

Clusters

Clusters of aggregated information can be used to analyze and predict conditions for the clusters or their components. Several examples of how big data in cloud environments might be used for clusters appear with Chris Anderson's *Wired* article (Anderson 2008) where clusters of worldwide agricultural information are analyzed and used to forecast cluster conditions once this is disaggregated back to the individual field level.

Predicting airfare prices is a second example. Anderson (2008) discusses how Farecast (now incorporated in Bing) uses a prediction model to track information on 175 billion fares originating at 79 US airports. The database is used to forecast when airline prices are going to change and can be used to develop ticket-buying strategies such as the day of the week to purchase, the length of stay, etc.

RISKS OF USING BIG DATA IN CLOUD COMPUTING

As we come to depend more on data, lack of access to that data becomes more costly; as we offload more decisions to analytical methods, improper use of these methods or inherent weaknesses in the methods may yield decisions that are wrong and costly. As more data can be integrated to help society, data can also be integrated in more harmful ways. Three obvious domains for BD/CC are: data sources, the cloud, and big data. These will be used to frame the discussion of BD/CC risks.

Data Sources

Data source risks include data quality risks and risks relating to how the data is defined, stored, and generated prior to transmission to the cloud. These will be discussed in terms of data quality, data security, data privacy, and data integration.

Data quality may be defined in many dimensions and is tied to often evolving user requirements. Acceptable quality for one user might be unacceptable for another; acceptable quality in 2010 might be unacceptable in 2013. A corollary to poor data quality is the risk that arises from using this same poor data to draw conclusions and make decisions. Information used for stated purposes might also create unintended consequences that result in real losses.

Dimensions of data quality include (Miller 1996):

1. Relevance: Does the data address its user's needs?
2. Accuracy: Does the data reflect the underlying reality? Is the level of precision consistent with the user's demands?
3. Timeliness: Is the data current relative to user demands?

4. Completeness: Does the level of completeness correspond with user demands? Data can be incomplete or even too complete!
5. Coherence: How well does the data *hang together*? Do irrelevant details, confusing measures or ambiguous format make it incoherent?
6. Format: How is the data presented to the user? Is the context appropriate?
7. Accessibility: Can the data be obtained when needed?
8. Compatibility: Is the data compatible in format and definition with other data with which it is being used?
9. Security: Is the data physically and logically secure?
10. Validity: Does the data satisfy appropriate standards related to other dimensions such as accuracy, timeliness, completeness and security?

Errors along one or more quality dimensions create problems for downstream users leading to risks similar to those encountered in current IT environments. In BD/CC environments several new data quality related risks appear. Data that is transferred to the cloud to be analyzed later may violate accuracy, timeliness, completeness, format, and accessibility quality dimensions. Some examples include real time status information about assets such as appliances and vehicles (Bughin 2010 refers to this as the 'internet of things'). Poor data quality casts a shadow on conclusions drawn from analysis, precludes analysis in the first place, and creates an environment with unintended consequences for users.

The *Economist* (Anonymous 2008) discusses such an environment concerning the U. S. legal system regarding the Completeness quality dimension. In a lawsuit concerning Horizon Blue Cross a client sued the insurance company over a claim for their anorexic daughter. The company asked to see all online postings, messages, etc. from the daughter's Facebook and MySpace pages. Beyond privacy issues (the daughter's lawyer objected on those grounds and lost), this example of e-discovery illustrates a *new normal*, i.e., the ability to review the equivalent of millions of paper pages. This increases the costs of the legal system and places certain populations (the less-wealthy) at a disadvantage. In the *Economist* article cited above, Justice Stephen Breyer of the U.S. Supreme court is quoted as saying, 'you're going to drive out of the litigation system a lot of people who ought to be there,' indicating that 'justice is determined by wealth, not be the merits of the case.' (para. 9)

Data security risks become magnified when data sources become ubiquitous. When data origination is distributed across personal devices such mobile phones, tablets, and appliances, data can be modified or stolen by direct access or apps containing computer viruses. Mobile device security often is less robust than in more controlled environments, which themselves have weaknesses.

Even in cloud computing environments, some data (including passwords) still resides on remote devices before being uploaded. Systems relying on this information can be compromised by denial of service attacks on the devices themselves. Such acts could become particularly costly in health care or transportation, where actions could physically affect users and cause financial losses or even deaths. Cell phone companies are aware of the threat and are taking action to protect information currently stored on phones. Ante (2010) discusses an application for Android-based devices that facilitates physical security and blocks malicious apps. One area of concern is mobile banking operations where phones are used to access personal information located in the cloud for transactions. The immediate risk of intercepting data, e.g., unencrypted or poorly encrypted passwords or content, is obvious, as are risks from subsequent analysis of changed data resulting in compromised decisions.

Data privacy issues involve interlopers intercepting and disclosing data either stored on mobile devices or appliances, or while being uploaded to or from the cloud. Although privacy breaches are more severe when an entire cloud-based database is compromised, aggregates of individual and seemingly insignificant data elements also may lead to significant privacy disclosures. Weitzner et al. (2006) recognized this problem and proposed a technology infrastructure to increase accountability and transparency for information use on the Web. As an example of how data elements might be integrated, consider mobile phones equipped with GPS location tracking devices. Currently user location information from mobile phones can be used to facilitate location-based marketing efforts (Hopkins and Turner 2012). This same information also could be used to facilitate criminal activity. Using GPS enabled location-based technologies, intercepted transmissions or a rouge app (e.g., an application that is placed on a device that purports to do one thing while surreptitiously doing another) could collect and transmit information to a third party which, beyond violating privacy standards, could alert criminals that individuals are away from home (Europol 2011). One can conjure many scenarios where location information can be used for nefarious ends, for example, disclosing the location of investment banker working on a merger to support insider trading; aiding a criminal bent on kidnapping a business leader; or a enabling a terrorist plotting an assassination (see Brandon 2012 for a Website that can pinpoint a mobile phone user's location by entering a mobile phone number). Sensing devices are another example of how mobile phones and appliances can generate data to be used by BD/CC applications. Shilton (2009) discusses the potential for abuse of privacy in this setting. One suggestion is that software developers implement data-protection standards for collecting personal information from sensing technology. Such standards may be necessary because it will be harder for users to voluntarily adhere to good data security practices as computing power and data generation becomes distributed among a mass audience.

Data Integration is the final risk involving data sources. Compatibility is a dimension of the quality of individual data elements, and incompatibility among the data elements may preclude otherwise promising analysis opportunities, or lead to spurious results. Cafarella et al. (2011) discusses integrating data across multiple databases on the Web and the ensuing challenges. When data that does not conform to any standardized data design this poses significant problems for traditional data management techniques. As with the above discussion regarding privacy, data elements generated from user-controlled devices adds more complexity to this problem. Table 1 summarizes Data Source Risks.

Table 1: DATA SOURCE RISKS

Data Quality
Fail to meet user requirements along specific data quality dimensions
Cast shadow re: conclusions from subsequent data analysis
Foster unintended consequences
Data overload responding to information requests

Data Security
Allow direct access to devices containing data
Facilitate access via apps or downloaded viruses
Denial of service on devices (e.g., health care applications)
Accessing online payment systems

Data Privacy
Releasing copied/stolen data on mobile devices or appliances
Creating data aggregates that compromise privacy
Using confidential information for financial gain
Eavesdropping on sensing or mobile devices

Data Integration
Lost opportunities caused by incompatible data elements
Spurious results caused by incompatible data elements
Data management problems

The Cloud

Many *data quality* risks in the cloud are similar to those discussed above. The difference is now the data resides in the cloud rather than on user equipment. Changes in data may occur through transmission from user to cloud, from cloud to user, or while in storage. Examples might affect quality dimensions such as accuracy, completeness, format, validity, timeliness, and accessibility. Intercepted data may be changed, erased, or copied. Modifying data in the cloud database could facilitate monetary crimes (e.g., using a mobile payment system); erased data could impact infrastructure control systems; and data that has been disclosed could compromise user privacy. Modified data would be particularly important for big data applications where large amounts of data from appliances are transferred up to the cloud, analyzed, and decisions for action are then sent back to the user appliances, such as traffic controls, or individuals administering medications.

Data obsolescence risk occurs when data in the cloud becomes stale. Applications such as mobile phone location systems that seek to pitch advertising messages to groups of customers based on their location and past purchase histories would be affected. Data redefinition occurs when data is redefined in a way that unintentionally affects data quality. Mullich (2011) quotes a Verizon executive as saying that enterprise level platforms that manage and synchronize data transmitted from the cloud to mobile devices will enable workers to access data ‘when and where they want.’ This creates challenges, especially when data once in sync becomes unsynchronized and creates a false picture of reality that distorts future actions.

Data security risks in the cloud appear in many media stories. In some cases the cloud arguably increases security. Cloud computing has enabled more robust solutions to the challenges of disaster recovery planning, with cost and accessibility advantages (Linthicum 2010). Backing up cloud data is more economical than replicating individual systems; when data from many firms are aggregated cloud architectures clearly provide security advantages. Centralization of information, however, creates two risks. First, either through a failure of the cloud servers or through the telecommunications systems linking cloud to customer, data may become unavailable bringing down a user organization. An example of this occurred when services provided by Amazon, a leading cloud vendor, were interrupted for two days affecting thousands of corporate customers (Lohr 2011). Events like this should motivate companies to rethink what applications are put on the cloud, especially those that are mission-critical. As BD/CC applications become more prevalent, and as companies have few alternatives other than migrating to the cloud, putting mission-critical applications on the cloud may no longer be a choice but a necessity, thus resulting in increased risk exposure.

A second risk stems from the ability to access cloud data, either by hacking into databases, intercepting telecommunications messages (as discussed above), or by sloppy vendor controls enabling one company to access a competitor's data in a cloud environment. Dodge (2011) maintains that the 'hacking in the cloud' risk is overblown. He is more concerned with denial of service risks, either intentional or unintentional. These risks are increasingly important as mission critical applications migrate to the cloud. Clayton (2011) discusses solutions that involve mixed applications, e.g., storing some locally and others on the cloud. Quoting one executive, Clayton says, 'Where the cloud becomes really interesting is with new things where you have the option of doing it on your own server in-house or going with the cloud. Then you're much more likely to move those to the cloud than traditional IT.' (para. 13) *New things* describe evolving big data applications and the challenges to control risk. Another risk is failure to properly coordinate cloud and locally stored applications.

Data privacy issues in the cloud result from information that is transmitted to the cloud from mobile devices or from other input channels, being hacked, stolen, or intercepted after being uploaded to the cloud. One concern for big data applications is privacy breaches for medical information. Big data applications can take many disparate data elements stored in the cloud and analyze them in ways that could present previously hidden individual profiles. These could affect employment, insurance, and psychological wellbeing. Osterhaus (2010) discusses these risks and concludes that the need for privacy trumps the need for access, and that cloud vendors will need to ensure HIPAA regulations (in the U.S., the Health Insurance Portability and Accountability Act) are met. This conclusion could apply to all cloud privacy risks; there is no reason why privacy in the cloud need be less secure than privacy in traditional environments. Indeed, the cloud environment as compared to storage in paper files in physicians' offices arguably is inherently *more* secure. The new risk to be controlled is the big data related risk stemming from integration and processing of large amounts of medical data and personal data. These are challenges, not barriers, to cloud security.

Data Integration in cloud environments includes customer information regarding health profiles, financial profiles, location information, information from social networking databases, and integrating data from infrastructure appliances. With regard to big data and integration leading to personalization Bollier (2010, 23), quoting Kim Taipale, points out that, 'the benefits of personalization tend to accrue to businesses but the harms are inflicted on disbursed and unorganized individuals.' A second risk is integrating data in ways that seek to achieve an objective but because of poor data quality or incompatible databases, miss opportunities. Examples of this occur when the data elements one is seeking to integrate exist in different formats, adhere to different standards, or have different timeliness or accuracy attributes.

Vendor risk is the final cloud risk. Since vendors control the cloud environment, all of the risks discussed above relate to vendor operations. Two other specific vendor risks involve how the vendor conducts business and *if* the vendor is able to conduct business. How the vendor conducts business includes the vendor's policies for data security, disaster recovery planning, transaction authentication and many other details. Often vendor policies exceed those and are more secure than policies in non-cloud environments. There are cases, however, where vendor policies and procedures are less robust than a client's. Currently, if the client is aware of vendor deficiencies, the client can process the application locally. This may be impossible for BD/CC applications. If a client uses a private cloud for critical applications and a public cloud for others, another risk surfaces: Blending public and private clouds. One solution is developing security infrastructures spanning both environments (Scheier 2009). Clayton (2011) discusses these along with the importance of vendors' understanding corporate and industry policies and regulatory requirements. The cloud environment is inherently harder to customize to each organization and regulatory policy violations may be more likely, especially when processing big data.

A final and more immediate vendor risk is the possibility that the vendor goes out of business. While this is less likely for vendors such as Amazon, Apple, and Google, smaller cloud vendors are now in the marketplace and may be more likely to exit the market. Organizations are vulnerable when they entrust their business operations to such a vendor. Managing this risk is not unlike vendor risk management procedures for entrusting mission-critical software development and operations to smaller vendors, or outsourcing mission critical operations. If customers view this form of vendor risk as unacceptable, larger vendors may crowd smaller cloud vendors out of the marketplace.

Table 2 summarizes Cloud Computing Risks.

Table 2: CLOUD COMPUTING RISKS

Data Quality
Transmission and storage breaches reducing data quality
Modifying data used for actions (financial, traffic signals, medication doses)
Lost marketing opportunities
Challenges in synchronizing data

Data Security
More robust disaster recovery plans reduce risk
Inaccessibility to the cloud
Increased access risk for mission critical systems
Hackers gaining cloud access
Denial or service attacks
Failure to coordinate cloud and local applications

Data Privacy
Releasing copied/stolen data (financial, medical)
Violating government or industry privacy regulations
Increased legal exposure

Data Integration
Lost opportunities caused by incompatible databases in cloud
Spurious results caused by incompatible databases in cloud
Data management problems

Vendor Risk
Poor vendor management of information-based risk
Lax vendor policies and procedures re: data
Problems blending vendor and client applications
Vendors who are unaware of regulatory requirements or industry policies
Vendors going out of business

Big Data

How information is collected, processed, reported, and acted on in a way that it adds value to its customer and does not adversely affect other information stakeholders is a BD/CC risk (for more on who owns customer information, see Hagel et al. 1994, Harison 2010, and Prabhaker 2000). For well-defined problems risks of big data involve dimensions that can be categorized under data quality, security, privacy, and integration, as discussed above. For example, a power company using big data to collect data from sensors on electrical devices to turn some of the devices off to optimize its electrical grid, is using big data to do something new but not something different in kind from using mathematical methods to optimize operational networks, something done for decades. In these situations the unique big data risk is being able to work faster and work on a larger scale. Using a sport analogy, the risk is similar to that faced by an amateur who, when promoted to play against professionals, realizes that the game now is faster, more complex, and more challenging.

When analytical methods are used in new ways the challenges and risks increase. One example is using big data to make inferences in situations where in the past, making data-based inferences was impractical (such as using correlations to extrapolate). While correlation analysis is a standard analytical technique, using it for new applications on a large scale may become problematic. Bollier (2010) quotes a classic *Wall Street Journal* article, 'My TIVO Thinks I'm Gay' where, based on the author's viewing habits, the system kept recommending gay-themed films. To remedy this, the customer began viewing war movies, and the system responded by recommending documentaries about the Third Reich! While these examples may be comical, one sees the risk of using BD/CC for categorizing tastes or other personal attributes. There is certainly a risk for privacy and if actions are taken, even in non-human environments, may create liabilities and risks of doing the wrong thing.

This leads to perhaps the most fundamental big data risk, which is exacerbated in cloud environments. The risk is not 'Are the results accurate?' but is 'Has big data addressed the right problem and if so, addressed it in the right way?' BD/CC relies heavily on integrating and analyzing massive amounts of data, and human intervention often is absent. The methodology *reports* what the data analysis indicates rather than is a methodology based on an underlying theory that *informs* data collection, analysis, and reporting. This risk is reflected in Anderson's (2008) article discussed above.

Big data applications often analyze information in a *near domain*, i.e., one in which the data is readily available. For example, when Google uses searches for flu symptoms to predict flu outbreaks (Guth 2008) the results are timelier than health reports from the U.S. Center for Disease Control (Bollier 2010). However, using Internet searches rather than a scientific model focuses on a related attribute (victim searches for information) rather than a root cause. A risk of rampant use of big data to develop inferential relationships in cloud environments, while appealing because the nexus of the availability of large amounts of data and processing power to generate seemingly useful results, is that users may think something is true based on past history even though future data may indicate otherwise. Over reliance on history creates the potential for mischaracterization of data, violations of privacy, poor decisions, and solving the wrong problems in a way failed generals fight the last war. Over reliance on history also risks repeating the classic error of treating correlation as causation.

When human links, characterized by grounding analysis in theories are severed, technologies may run amok (Postman 1993). Moreover, modeling assumptions that disenfranchise certain groups, e.g., using credit-scoring models (Anonymous 2011), may be embedded in those technologies. W. Edwards Deming, a pioneer in the quality movement, held that (Deming 1994, 103), 'Without theory, experience has no meaning, and without theory, one has no questions to ask. Hence, without theory, there is no learning. Theory is a window into the world. Theory leads to prediction, and without theory, experience and examples teach nothing' Rampant use of correlations and other inferences obtained only from processing massive amounts of data without human intervention, and removed from any underlying theory, may lead to a data cycle of collect-analyze-act which may appear valid, but which in fact is seriously flawed.

Table 3 summarizes Big Data Risks.

Table 3: BIG DATA RISKS

Fail to meet user requirements due to inability to process data
Inappropriate use of analytical methods
Applications that adversely affect stakeholders
Uncertainty regarding who owns customer information
Addressing the wrong problem
Focusing on the "near domain" and ignoring the real problem
Mischaracterizing data resulting in privacy violations
Mischaracterizing data resulting poor decisions
Disenfranchising groups by ignoring theory and over-relying on data

CONCLUSION

BD/CC represents the confluence of two technologies that promise to create new capabilities for creating more value for customers and businesses at reduced costs. Its promise is hard to dismiss. Indeed, as BD/CC environments become commonplace, more and more success stories will lead to more and more applications. These applications will create a plethora of new products and marketing opportunities, which may fundamentally change the relationship between organizations and their customers.

Five steps, informed by Reinhold et al. (2011) may be used to manage big data risks in cloud computing environments. They are:

1. Implement a process to continually review, assess, and understand BD/CC risks in an organization's local and global environments;
2. View the BD/CC risk management process and technology improvements strategically rather than tactically, and act accordingly;
3. Ensure all underlying data is consistent, well-defined, and understood by managers and other stakeholders;
4. Ensure data is of high quality along all relevant quality dimensions, and that data governance responsibilities are well-defined;
5. Continually invest in improvements in both cloud computing infrastructure and analytical big data capabilities, and ensure these capabilities are congruent with the strategic needs of point 1.

A key tenet of risk management is that risk can never be eliminated but once understood, can be managed. What is required is ongoing understanding of the evolving customer and technical environments, understanding of the evolving risk environment, and the commitment to categorize quantify and ensure that the risk exposure is consistent with the organization's underlying objectives. A final challenge for big data is while it is possible for organizations to manage their risk, who will manage the risk exposure of the individuals generating and using the data? Incentives should be created to make investing in risk management mutually beneficial both for those generating the data, those using the data, and all others profiting from it.

About the Author

Holmes E. Miller is Professor of Business in the Department of Accounting, Business, and Economics at Muhlenberg College in Allentown, Pennsylvania (USA). He has a Ph.D. in Management Science from Northwestern University in Evanston, Illinois. Prior to coming to Muhlenberg he taught at Rensselaer Polytechnic Institute in Troy, New York and worked in the chemical and financial services industries. He can be reached at homiller@muhlenberg.edu.

Citations and References

Anderson, C. (2008). The end of theory: will the data deluge makes the scientific method obsolete?. *Wired*, (June 30), Retrieved 1 November 2012 from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Anonymous (2008). The big data dump. *The Economist*, Retrieved 26 November 2012 from http://www.economist.com/node/12010377

Anonymous (2011). Credit scoring. Electronic Privacy Information Centre, Retrieved 17 June 2012 from http://epic.org/privacy/creditscoring/

Ante, S. (2010). Fortifying phones from attackers; AT&T hires PhDs for security lab; Verizon Wireless teams with start-up on data-security app. *Wall Street Journal* (Online). (Dec 23), Retrieved 1 November from 2012 from http://online.wsj.com/article/SB10001424052748704774604576035960449272404.html

Bollier, D. (2010). The promise and peril of big data. *The Aspen Institute*, Retrieved 10 June 2012 from http://web.resourceshelf.com/go/resourceblog/57852

Brandon, J. (2012). Website can find your exact location with your phone number. Retrieved 8 November 2012 from http://www.foxnews.com/tech/2012/02/05/website-can-find-with-your-phone-number/#ixzz2AplRtG00

Bryant, R., R. Katz, and E. Lazowska (2008); Big Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. *Computing Community Consortium*. Retrieved 8 November 2012 from http://www.cra.org/ccc/initiatives.

Bughin, J., M. Chui, and J. Manyika (2010). Clouds, big data and smart assets: ten tech-enabled business trends to watch. *McKinsey Quarterly*, **4**, 26-43.

Cafarella, M., A. Halevy and J. Madhavan (2011). Structured data on the web. *Communications of the ACM*, **54**(2), 72-79.

Clayton, N. (2011). Get of my cloud. *Wall Street Journal*, (February 15), Retrieved 8 November 2012 from http://online.wsj.com/article/SB10001424052748704739504576067823202021708.html.

Conway, D. (2008), The hubris of 'The end of theory'. Retrieved 17 June 2012 from http://www.drewconway.com/zia/?p=209.

Deming, W.E. (1994). The New Economics for Industry, Government, Education, 2nd Edition, MIT Centre for Advanced Educational Services: Cambridge, MA.

Dodge, J. (2011). Is credit card hack damage to cloud reputation overblown? CIO, Retrieved 17 June 2012 from http://www.enterprisecioforum.com/en/blogs/jdodge/credit-card-hack-damage-cloud-reputation-overblown.

Ellison, L. (2008). What the hell is cloud computing. Retrieved 5 June 2012 from http://www.youtube.com/watch?v=0FacYAI6DY0.

Europol, (2011). Threat Assessment: Internet facilitated organized crime iOCTA - Europol. Retrieved 8 November 2012 from https://www.europol.europa.eu/content/publication/iocta-threat-assessment-internet-facilitated-organised-crime-1455

Guth, R. (2008). Sniffly surfing: Google unveils flu-bug tracker. *Wall Street Journal - Eastern Edition*, **252** (114) D1, D16

Hagel, J, and Lansing, J. (1994). Who owns the customer? *McKinsey Quarterly*, (4), 63-75,

Harison, E. (2010). Who owns enterprise information? Data ownership rights in Europe and the U.S.. *Information & Management*, **47** (2), 102-108

Hopkins, J. and Turner, J. (2012). Go Mobile: location-based marketing, apps, mobile optimized ad campaigns, 2d codes and other mobile strategies to grow your business. John Wiley & Sons.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, **52** (8), Retrieved 8 November 2012 from http://cacm.acm.org/magazines/2009/8/34493-the-pathologies- of-big-data/fulltext,

Kusnetsky, D. (2010). What is big data?. *ZDNet,* Retrieved 16 February 2012 from http://www.zdnet.com/blog/virtualization/what-is-big-data/1708

Linthicum, D. (2010). Leveraging cloud computing for business continuity. *Disaster Recovery Journal*, **23** (3), 29-30.

Lohr, S. (2011). Amazon's trouble raises cloud computing doubts. *New York Times*, (April 23), B1.

McKendrick, J. (2010). Big Data for the year ahead: 10 predictions. Retrieved 17 December 2012 from http://www.zdnet.com/blog/service-oriented/big-data-for-the-year-ahead-10-predictions/6258

Miller, H. (1996). The multiple dimensions of information quality, *Information Systems Management,* **13** (2), 79-82.

Morton, O. (2008). The end of theory. *Edge*, Retrieved 1 November 2012 from http://www.edge.org/discourse/the_end_of_theory.html

Mullich, J. (2011). Mid-year IT trends: reassessing your IT game plan. *Wall Street Journal (Special Advertising Section),* (June 22); B4.

Osterhaus, L. (2010). Cloud computing and health information. *University of Iowa*, Retrieved 22 June 2012 from http://ir.uiowa.edu/bsides/19/

Postman, N. (1993). Technopoly: The surrender of culture of technology. Vintage Books: New York.

Prabhaker, P. (2000); Who owns the online consumer? *Journal of Consumer Marketing*, **17** (2/3), 158-169

Reinhold, B., J. Goherty, and D. Higgins (2011). Rethink risk, rethink technology. **103** (4), *ABA Banking Journal*, 26-30

Scheier, R. (2009). Busting the nine myths of cloud computing. Retrieved 20 November 2012 from http://www.infoworld.com/d/cloud-computing/busting-nine-myths-cloud-computing-260

Searls, D. (2012). The Customer as a God. Retrieved November 20 2012 from http://online.wsj.com/article/SB100008723963904448732045775353525210921 54.html

Shilton, K. (2009). Four billion little brothers? Privacy, mobile phones, and ubiquitous data collection. *Communications of the ACM*; **52** (11), 48-53.

Weitzner, D. J.; Abelson, H.; Berners-Lee, T.; Hanson, C.; Hendler, J.; Kagal, L.; McGuinness, D. L.; Sussman, G. J.; Waterman, K. K. (2006). Transparent Accountable Data Mining: New Strategies for Privacy Protection; Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory; Retrieved 8 November 2012 from http://hdl.handle.net/1721.1/30972.