# Big Data in Cloud Computing: Features and Issues

Pedro Caldeira Neves[1,2], Bradley Schmerl[1], Javier Cámara[1] and Jorge Bernardino[1,2,3]
[1]*Carnegie Mellon University, Institute for Software Research, Pittsburgh, PA 15213, U.S.A*
[2]*ISEC – Superior Institute of Engineering of Coimbra, Polytechnic of Coimbra, 3030-190 Coimbra, Portugal*
[3]*CISUC – Centre of Informatics and Systems of the University of Coimbra, FCTUC, University of Coimbra, 3030-290 Coimbra, Portugal*

Keywords: Big Data, Cloud Computing, Big Data Issues.

Abstract: The term big data arose under the explosive increase of global data as a technology that is able to store and process big and varied volumes of data, providing both enterprises and science with deep insights over its clients/experiments. Cloud computing provides a reliable, fault-tolerant, available and scalable environment to harbour big data distributed management systems. Within the context of this paper we present an overview of both technologies and cases of success when integrating big data and cloud frameworks. Although big data solves much of our current problems it still presents some gaps and issues that raise concern and need improvement. Security, privacy, scalability, data governance policies, data heterogeneity, disaster recovery mechanisms, and other challenges are yet to be addressed. Other concerns are related to cloud computing and its ability to deal with exabytes of information or address exaflop computing efficiently. This paper presents an overview of both cloud and big data technologies describing the current issues with these technologies.

## 1 INTRODUCTION

In recent years, there has been an increasing demand to store and process more and more data, in domains such as finance, science, and government. Systems that support big data, and host them using cloud computing, have been developed and used successfully (Hashem et al., 2014) .

Whereas big data is responsible for storing and processing data, cloud provides a reliable, fault-tolerant, available and scalable environment so that big data systems can perform (Hashem et al., 2014). Big data, and in particular big data analytics, are viewed by both business and scientific areas as a way to correlate data, find patterns and predict new trends. Therefore there is a huge interest in leveraging these two technologies, as they can provide businesses with a competitive advantage, and science with ways to aggregate and summarize data from experiments such as those performed at the Large Hadron Collider (LHC).

To be able to fulfil the current requirements, big data systems must be available, fault tolerant, scalable and elastic.

In this paper we describe both cloud computing and big data systems, focusing on the issues yet to be addressed. We particularly discuss security concerns when hiring a big data vendor: data privacy, data governance, and data heterogeneity; disaster recovery techniques; cloud data uploading methods; and how cloud computing speed and scalability poses a problem regarding exaflop computing.

Despite some issues yet to be improved, we present two examples that show how cloud computing and big data can work well together.

Our contributions to the current state-of-the-art is done by providing an overview over the issues to improve or have yet to be addressed in both technologies.

The remainder of this paper is organized as follows: Section 2 provides a general overview of big data and cloud computing; Section 3 discusses and presents two examples that show how big data and cloud computing work well together and especially how hiring a big data vendor may be a good choice so that organizations can avoid IT worries; Section 4 discusses the several issues to address in cloud computing and big data systems; and Section 5 presents the discussion, conclusions and future work.



Neves, P., Schmerl, B., Cámara, J. and Bernardino, J.
Big Data in Cloud Computing: Features and Issues.
DOI: 10.5220/0005846303070314
In *Proceedings of the International Conference on Internet of Things and Big Data (IoTBD 2016)*, pages 307-314
ISBN: 978-989-758-183-0

## 2 BIG DATA & CLOUD COMPUTING

The concept of big data became a major force of innovation across both academics and corporations. The paradigm is viewed as an effort to understand and get proper insights from big datasets (big data analytics), providing summarized information over huge data loads. As such, this paradigm is regarded by corporations as a tool to understand their clients, to get closer to them, find patterns and predict trends. Furthermore, big data is viewed by scientists as a mean to store and process huge scientific datasets. This concept is a hot topic and is expected to continue to grow in popularity in the coming years.

Although big data is mostly associated with the storage of huge loads of data it also concerns ways to process and extract knowledge from it (Hashem et al., 2014). The five different aspects used to describe big data (commonly referred to as the five "V"s) are Volume, Variety, Velocity, Value and Veracity (Sakr and Gaber, 2014):

- *Volume* describes the size of datasets that a big data system deals with. Processing and storing big volumes of data is rather difficult, since it concerns: scalability so that the system can grow; availability, which guarantees access to data and ways to perform operations over it; and bandwidth and performance.
- *Variety* concerns the different types of data from various sources that big data frameworks have to deal with.
- *Velocity* concerns the different rates at which data streams may get in or out the system and provides an abstraction layer so that big data systems can store data independently of the incoming or outgoing rate.
- *Value* concerns the true value of data (i.e., the potential value of the data regarding the information they contain). Huge amounts of data are worthless unless they provide value.
- *Veracity* refers to the trustworthiness of the data, addressing data confidentiality, integrity, and availability. Organizations need to ensure that data as well as the analyses performed on the data are correct.

Cloud computing is another paradigm which promises theoretically unlimited on-demand services to its users. Cloud's ability to virtualize resources allows abstracting hardware, requiring little interaction with cloud service providers and enabling users to access terabytes of storage, high processing power, and high availability in a *pay-as-you-go* model (González-Martínez et al., 2015). Moreover, it transfers cost and responsibilities from the user to the cloud provider, boosting small enterprises to which getting started in the IT business represents a large endeavour, since the initial IT setup takes a big effort as the company has to consider the total cost of ownership (TCO), including hardware expenses, software licenses, IT personnel and infrastructure maintenance. Cloud computing provides an easy way to get resources on a pay-as-you-go model, offering scalability and availability, meaning that companies can easily negotiate resources with the cloud provider as required. Cloud providers usually offer three different basic services: Infrastructure as a Service (IaaS); Platform as a Service (PaaS); and Software as a Service (SaaS):

- *IaaS* delivers infrastructure, which means storage, processing power, and virtual machines. The cloud provider satisfies the needs of the client by virtualizing resources according to the service level agreements (SLAs);
- *PaaS* is built atop of IaaS and allows users to deploy cloud applications created using the programming and run-time environments supported by the provider. It is at this level that big data DBMS are implemented;
- *SaaS* is one of the most known cloud models and consists of applications running directly in the cloud provider;

These three basic services are closely related: SaaS is developed over PaaS and ultimately PaaS is built atop of IaaS.

From the general cloud services other services such as Database as a Service (DBaaS) (Oracle, 2012), BigData as a Service (BDaaS) and Analytics as a Service (AaaS) arose.

Since the cloud virtualizes resources in an on-demand fashion, it is the most suitable and compliant framework for big data processing, which through hardware virtualization creates a high processing power environment for big data.

## 3 BIG DATA IN THE CLOUD

Storing and processing big volumes of data requires scalability, fault tolerance and availability. Cloud computing delivers all these through hardware virtualization. Thus, big data and cloud computing are two compatible concepts as cloud enables big data to be available, scalable and fault tolerant.

Business regard big data as a valuable business opportunity. As such, several new companies such as Cloudera, Hortonworks, Teradata and many others, have started to focus on delivering Big Data as a Service (BDaaS) or DataBase as a Service (DBaaS). Companies such as Google, IBM, Amazon and Microsoft also provide ways for consumers to consume big data on demand. Next, we present two examples, Nokia and RedBus, which discuss the successful use of big data within cloud environments.

### 3.1 Nokia

Nokia was one of the first companies to understand the advantage of big data in cloud environments (Cloudera, 2012). Several years ago, the company used individual DBMSs to accommodate each application requirement. However, realizing the advantages of integrating data into one application, the company decided to migrate to Hadoop-based systems, integrating data within the same domain, leveraging the use of analytics algorithms to get proper insights over its clients. As Hadoop uses commodity hardware, the cost per terabyte of storage was cheaper than a traditional RDBMS (Cloudera, 2012).

Since Cloudera Distributed Hadoop (CDH) bundles the most popular open source projects in the Apache Hadoop stack into a single, integrated package, with stable and reliable releases, it embodies a great opportunity for implementing Hadoop infrastructures and transferring IT and technical concerns onto the vendors' specialized teams. Nokia regarded Big Data as a Service (BDaaS) as an advantage and trusted Cloudera to deploy a Hadoop environment that copes with its requirements in a short time frame. Hadoop, and in particular CDH, strongly helped Nokia to fulfil their needs (Cloudera, 2012).

### 3.2 RedBus

RedBus is the largest company in India specialized in online bus ticket and hotel booking. This company wanted to implement a powerful data analysis tool to gain insights over its bus booking service (Kumar, 2006). Its datasets could easily stretch up to 2 terabytes in size. The application would have to be able to analyse booking and inventory data across hundreds of bus operators serving more than 10.000 routes. Furthermore, the company needed to avoid setting up and maintaining a complex in-house infrastructure.

At first, RedBus considered implementing in-house clusters of Hadoop servers to process data. However they soon realized it would take too much time to set up such a solution and that it would require specialized IT teams to maintain such infrastructure. The company then regarded Google bigQuery as the perfect match for their needs, allowing them to:

- Know how many times consumers tried to find an available seat but were unable to do it due bus overload;
- Examine decreases in bookings;
- Quickly identify server problems by analysing data related to server activity;

Moving towards big data brought RedBus business advantages. Google bigQuery armed RedBus with real-time data analysis capabilities at 20% of the cost of maintaining a complex Hadoop infrastructure (Kumar, 2006).

As supported by Nokia and RedBus examples, switching towards big data enables organizations to gain competitive advantage. Additionally, BDaaS provided by big data vendors allows companies to leave the technical details for big data vendors and focus on their core business needs.

## 4 BIG DATA ISSUES

Although big data solves many current problems regarding high volumes of data, it is a constantly changing area that is always in development and that still poses some issues. In this section we present some of the issues not yet addressed by big data and cloud computing.

As the amount of data grows at a rapid rate, keeping all data is physically cost-ineffective. Therefore, corporations must be able to create policies to define the life cycle and the expiration date of data (data governance). Moreover, they should define who accesses and with what purpose clients' data is accessed. As data moves to the cloud, security and privacy become a concern that is the subject of broad research.

Big data DBMSs typically deal with lots of data from several sources (variety), and as such heterogeneity is also a problem that is currently under study. Other issues currently being investigated are disaster recovery, how to easily upload data onto the cloud, and Exaflop computing.

Within this section we provide an overview over these problems.

## 4.1 Security

Cloud computing and big data security is a current and critical research topic (Popović & Hocenski, 2015). This problem becomes an issue to corporations when considering uploading data onto the cloud. Questions such as who is the real owner of the data, where is the data, who has access to it and what kind of permissions they have are hard to describe. Corporations that are planning to do business with a cloud provider should be aware and ask the following questions:

a) *Who is the Real Owner of the Data and Who has Access to it?*
   The cloud provider's clients pay for a service and upload their data onto the cloud. However, to which one of the two stakeholders does data really belong? Moreover, can the provider use the client's data? What level of access has to it and with what purposes can use it? Can the cloud provider benefit from that data?
   In fact, IT teams responsible for maintaining the client's data must have access to data clusters. Therefore, it is in the client's best interest to grant restricted access to data to minimize data access and guarantee that only authorized personal access its data for a valid reason.
   These questions seem easy to respond to, although they should be well clarified before hiring a service. Most security issues usually come from inside of the organizations, so it is reasonable that companies analyse all data access policies before closing a contract with a cloud provider.
b) *Where is the Data?*
   Sensitive data that is considered legal in one country may be illegal in another country, therefore, for the sake of the client, there should be an agreement upon the location of data, as its data may be considered illegal in some countries and lead to prosecution.

The problems to these questions are based upon agreements (Service Level Agreements – SLAs), however, these must be carefully checked in order to fully understand the roles of each stakeholder and what policies do the SLAs cover and not cover concerning the organization's data. This is typically something that must be well negotiated.

Concerning limiting data accesses, (Tu et al., 2013) and (Popa et al., 2011) came up with an effective way to encrypt data and run analytical queries over encrypted data. This way, data access is no longer a problem since both data and queries are encrypted. Nevertheless, encryption comes with a cost, which often means higher query processing times.

## 4.2 Privacy

The harvesting of data and the use of analytical tools to mine information raises several privacy concerns. Ensuring data security and protecting privacy has become extremely difficult as information is spread and replicated around the globe. Analytics often mine users' sensitive information such as their medical records, energy consumption, online activity, supermarket records etc. This information is exposed to scrutiny, raising concerns about profiling, discrimination, exclusion and loss of control (Tene and Polonetsky, 2012). Traditionally, organizations used various methods of de-identification (anonymization or encryption of data) to distance data from real identities. Although, in recent years it was proved that even when data is anonymized, it can still be re-identified and attributed to specific individuals (Tene and Polonetsky, 2012). A way to solve this problem was to treat all data as personally identifiable and subject to a regulatory framework. Although, doing so might discourage organizations from using de-identification methods and, therefore, increase privacy and security risks of accessing data.

Privacy and data protection laws are premised on individual control over information and on principles such as data and purpose minimization and limitation. Nevertheless, it is not clear that minimizing information collection is always a practical approach to privacy. Nowadays, the privacy approaches when processing activities seem to be based on user consent and on the data that individuals deliberately provide.

Privacy is undoubtedly an issue that needs further improvement as systems store huge quantities of personal information every day.

## 4.3 Heterogeneity

Big data concerns big volumes of data but also different velocities (i.e., data comes at different rates depending on its source output rate and network latency) and great variety. The latter comprehends very large and heterogeneous volumes of data coming from several autonomous sources. Variety is one of the "*major aspects of big data characterization*" (Majhi and Shial, 2015) which is triggered by the belief that storing all kinds of data may be beneficial to both science and business.

Data comes to big data DBMS at different velocities and formats from various sources. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also result in diverse data representations (Wu et al., 2014). Dealing with

such a wide variety of data and different velocity rates is a hard task that Big Data systems must handle. This task is aggravated by the fact that new types of file are constantly being created without any kind of standardization. Though, providing a consistent and general way to represent and explore complex and evolving relationships from this data still poses a challenge.

## 4.4 Data Governance

The belief that storage is cheap, and its cost is likely to decline further, is true regarding hardware prices. However, a big data DBMS does also concern other expenses such as infrastructure maintenance, energy, and software licenses (Tallon, 2013). All these expenses combined comprise the total cost of ownership (TCO), which is estimated to be seven times higher than the hardware acquisition costs.

Regarding that the TCO increases in direct proportion to the growth of big data, this growth must be strictly controlled. Recall that the "Value" (one of big data Vs) stands to ensure that only valuable data is stored, since huge amounts of data are useless if they comprise no value.

Data Governance came to address this problem by creating policies that define for how long data is viable. The concept consists of practices and organizational polices that describe how data should be managed through its useful economic life cycle. These practices comprise three different categories:

1. Structural practices identify key IT and non-IT decision makers and their respective roles and responsibilities regarding data ownership, value analysis and cost management (Morgan Kaufmann, 2013).
2. Operational practices consist of the way data governance policies are applied. Typically, these policies span a variety of actions such as data migration, data retention, access rights, cost allocation and backup and recovery (Tallon, 2013).
3. Relational practices formally describe the links of the CIO, business managers and data users in terms of knowledge sharing, value analysis, education, training and strategic IT planning.

Data Governance is a general term that applies to organizations with huge datasets, which defines policies to retain valuable data as well as to manage data accesses throughout its life cycle. It is an issue to address carefully. If governance policies are not enforced, it is most likely that they are not followed. Although, there are limits to how much value data governance can bring, as beyond a certain point stricter data governance can have counterproductive effects.

## 4.5 Disaster Recovery

Data is a very valuable business and losing data will certainly result in losing value. In case of emergency or hazardous accidents such as earthquakes, floods and fires, data losses need to be minimal. To fulfil this requirement, in case of any incident, data must be quickly available with minimal downtime and loss. However, although this is a very important issue, the research in this particular area is relatively low (Subashini and Kavitha, 2011), (Wood et al., 2010), (Chang, 2015).

For big corporations it is imperative to define a disaster recovery plan – as part of the data governance plan – that not only relies on backups to reset data but also in a set of procedures that allow quick replacement of the lost servers (Chang, 2015).

From a technical perspective, the work described in (Chang, 2015) presents a good methodology, proposing a "*multi-purpose approach, which allows data to be restored to multiple sites with multiple methods*", ensuring a recovery percentage of almost 100%. The study also states that usually, data recovery methods use what they call a "single-basket approach", which means there is only one destination from which to secure the restored data.

As the loss of data will potentially result in the loss of money, it is important to be able to respond efficiently to hazardous incidents. Successfully deploying big data DBMSs in the cloud and keeping it always available and fault-tolerant may strongly depend on disaster recovery mechanisms.

## 4.6 Other Problems

The current state of the art of cloud computing, big data, and big data platforms in particular, prompts some other concerns. Within this section we discuss data transference onto the cloud; Exaflop computing, which presents a major concern nowadays; and scalability and elasticity issues in cloud computing and big data:

***a)*** **Transferring Data onto a Cloud** is a very slow process and corporations often choose to physically send hard drives to the data centres so that data can be uploaded. However, this is neither the most practical nor the safest solution to upload data onto the cloud. Through the years there has been an effort to improve and create efficient data uploading algorithms to minimize upload times and provide a secure way to transfer data onto the cloud (Zhang et

al., 2013), however, this process still remains a major bottleneck.

***b)*** **Exaflop Computing** (Geller, 2011), (Schilling, 2014) is one of today's problems that is subject of many discussions. Today's supercomputers and clouds can deal with petabyte data sets, however, dealing with exabyte size datasets still raises lots of concerns, since high performance and high bandwidth is required to transfer and process such huge volumes of data over the network. Cloud computing may not be the answer, as it is believed to be slower than supercomputers since it is restrained by the existent bandwidth and latency. High performance computers (HPC) are the most promising solutions, however the annual cost of such a computer is tremendous. Furthermore, there are several problems in designing exaflop HPCs, especially regarding efficient power consumption. Here, solutions tend to be more GPU based instead of CPU based. There are also problems related to the high degree of parallelism needed among hundred thousands of CPUs.

Analysing Exabyte datasets requires the improvement of big data and analytics which poses another problem yet to resolve.

c) **Scalability and Elasticity** in cloud computing and in particular regarding big data management systems is a theme that needs further research as the current systems hardly handle data peaks automatically. Most of the time, scalability is triggered manually rather than automatically and the state-of-the-art of automatic scalable systems shows that most algorithms are reactive or proactive and frequently explore scalability from the perspective of better performance. However, a proper scalable system would allow both manual and automatic reactive and proactive scalability based on several dimensions such as security, workload rebalance (i.e.: the need to rebalance workload) and redundancy (which would enable fault tolerance and availability). Moreover, current data rebalance algorithms are based on histogram building and load equalization (Mahesh et al., 2014). The latter ensures an even load distribution to each server. However, building histograms from each server's load is time and resource expensive and further research is being conducted on this field to improve these algorithms.

## 4.7 Research Challenges

As discussed in Section 3, cloud and big data technologies work very well together. Even though the partnership between these two technologies have been established, both still pose some challenges.

Table 1 summarizes the issues of big data and cloud computing nowadays. The first column specifies the existing issues whereas the second describes the existing solutions and the remaining present the advantages and disadvantages of each solution.

Concerning the existing problems, we define some of the possible advances in the next few years:

- Security and Privacy can be resolved using data encryption. However, a new generation of systems must ensure that data is accessed quickly and that encryption does not affect processing times so badly;
- Big Data variety can be addressed by using data standardization. This, we believe, is the next step to minimize the impact of heterogeneity;
- Data governance and data recovery plans are difficult to manage and implement, but as Big Data become a de facto technology, companies are starting to understand the need of such plans.;
- New and secure QoS (quality of service) based data uploading mechanisms may be the answer to ease data uploading onto the cloud;
- Exaflop computing is a major challenge that involves governments funding and which is in its best interest. The best solutions so far use HPCs and GPUs;
- Scalability and elasticity techniques exist and are broadly used by several Big Data vendors such as Amazon and Microsoft. The major concern relies upon developing fully automatic reactive and proactive systems that are capable of dealing with load requirements automatically.

# 5 CONCLUSIONS

With data increasing on a daily base, big data systems and in particular, analytic tools, have become a major force of innovation that provides a way to store, process and get information over petabyte datasets. Cloud environments strongly leverage big data solutions by providing fault-tolerant, scalable and available environments to big data systems.

Although big data systems are powerful systems that enable both enterprises and science to get insights over data, there are some concerns that need further investigation. Additional effort must be employed in developing security mechanisms and standardizing data types. Another crucial element of Big Data is scalability, which in commercial techniques are mostly manual, instead of automatic.

Table 1: Big data issues.

| Issues | Existent solutions | Advantages | Disadvantages |
|---|---|---|---|
| Security | Based on SLAs and data Encryption | Data is encrypted | Querying encrypted data is time-consuming |
| Privacy | -De-identification<br>-User consent | Provides a reasonable privacy or transfers responsibility to the user | It was proved that most de-identification mechanisms can be reverse engineered |
| Heterogeneity | One of the big data systems' characteristics is the ability to deal with different data coming at different velocities | The major types of data are covered up | It is difficult to handle such variety of data and such different velocities |
| Data Governance | Data governance documents | -Specify the way data is handled;<br>-Specify data access policies;<br>-Role specification;<br>-Specify data life cycle | -The data life cycle is not easy to define;<br>-Enforcing data governance policies so much can lead to counterproductive effects |
| Disaster recovery | Recovery plans | Specify the data recovery locations and procedures | Normally there is only one destination from which to secure data |
| Data Uploading | -Send HDDs to the cloud provider<br>-Upload data through the Internet | Physically sending the data to the cloud provider is quicker than uploading data but it is much more unsecure | Physically sending data to the cloud provider is dangerous as HDDs can suffer damage from the trip.<br>- Uploading data through the network is time-consuming and, without encryption, can be insecure |
| High Data processing (Exabyte datasets) | -Cloud computing<br>-HPCs | Cloud computing is not so cost expensive as HPCs but HPCs are believed to handle Exabyte datasets much better | HPCs are very much expensive ant its total cost over a year is hard to maintain. On the other hand, cloud is believed that cannot cope with the requirements for such huge datasets |
| Scalability | Scalability exists at the three levels in the cloud stack. At the Platform level there is: horizontal (Sharding) and vertical scalability | Scalability allows the system to grow on demand | Scalability is mainly manual and is very much static. Most big data systems must be elastic to cope with data changes |
| Elasticity | There are several elasticity techniques such as Live Migration, Replication and Resizing | Elasticity brings the system the capability of accommodating data peaks | Most load variations assessments are manually made, instead of automatized |

Further research must be employed to tackle this problem. Regarding this particular area, we are planning to use adaptable mechanisms in order to develop a solution for implementing elasticity at several dimensions of big data systems running on cloud environments. The goal is to investigate the mechanisms that adaptable software can use to trigger scalability at different levels in the cloud stack. Thus, accommodating data peaks in an automatic and reactive way.

Within this paper we provide an overview of big data in cloud environments, highlighting its advantages and showing that both technologies work very well together but also presenting the challenges faced by the two technologies.

## ACKNOWLEDGEMENTS

This research is supported by Early Bird project funding, CMU Portugal, Applying Stitch to Modern Distributed Key-Value Stores and was hosted by Carnegie Mellon University under the program for CMU-Portugal undergraduate internships

## REFERENCES

Chang, V., 2015. Towards a big data system disaster recovery in a Private cloud. *Ad Hoc Networks*, 000, pp.1–18.

cloudera, 2012. Case Study Nokia: Using big data to Bridge the Virtual & Physical Worlds.

Geller, T., 2011. Supercomputing's exaflop target. *Communications of the ACM*, 54(8), p.16.

González-Martínez, J. a. et al., 2015. cloud computing and education: A state-of-the-art survey. *Computers & Education*, 80, pp.132–151.

Hashem, I.A.T. et al., 2014. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, pp.98–115.

Kumar, P., 2006. Travel Agency Masters big data with Google bigQuery.

Mahesh, A. et al., 2014. Distributed File System For Load Rebalancing In cloud Computing. , 2, pp.15–20.

Majhi, S.K. & Shial, G., 2015. Challenges in big data cloud Computing And Future Research Prospects: A Review. *The Smart Computing Review*, 5(4), pp.340–345.

Morgan Kaufmann, B., 2013. Chapter 5 – data governance for big data analytics: considerations for data policies and processes, in: D. Loshin (Ed.), big data Analytics. , pp.pp. 39–48.

Oracle, 2012. Database as a Service ( DBaaS ) using Enterprise Manager 12c.

Popa, R.A., Zeldovich, N. & Balakrishnan, H., 2011. CryptDB : A Practical Encrypted Relational DBMS. *Design*, pp.1–13.

Popović, K. & Hocenski, Z., 2015. cloud computing security issues and challenges. , (January), pp.344–349.

Sakr, S. & Gaber, M.M., 2014. *Large Scale and big data: Processing and Management* Auerbach, ed.,

Schilling, D.R., 2014. Exaflop Computing Will Save the World ... If We Can Afford It - Industry Tap. Available at: http://www.industrytap.com/exaflop-computing-will-save-world-can-afford/15485 [Accessed May 26, 2015].

Subashini, S. & Kavitha, V., 2011. A survey on security issues in service delivery models of cloud computing. *Journal of Network and Computer Applications*, 34(1), pp.1–11.

Tallon, P.P., 2013. Corporate governance of big data: perspectives on value, risk, and cost. *Computer 46*, pp.pp. 32–38.

Tene, O. & Polonetsky, J., 2012. Privacy in the Age of big data.

Tu, S. et al., 2013. Processing analytical queries over encrypted data. *Proceedings of the VLDB Endowment*, 6(5), pp.289–300.

Wood, T. et al., 2010. Disaster recovery as a cloud service: Economic benefits & deployment challenges. *2nd USENIX Workshop on Hot Topics in cloud Computing*, pp.1–7.

Wu, X. et al., 2014. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1), pp.97–107.

Zhang, L. et al., 2013. Moving big data to the cloud. *INFOCOM, 2013 Proceedings IEEE*, pp.405–409