Issues and Challenges in Convergence of Big Data, Cloud and Data Science

Neha Mathur
Dept of Computer Science and Engg,
Jai Narain Vyas University,
Jodhpur, Rajasthan, India

Rajesh Purohit, PhD
Dept of Computer Science and Engg,
Jai Narain Vyas University,
Jodhpur, Rajasthan, India

ABSTRACT
Big data, Cloud Computing and Data Science are currently trending in organizations across the globe. Big Data refers to technologies and techniques that involve data that is massive, heterogeneous and fast-changing for conventional technologies, skills and infra-structure to address efficiently. Cloud Computing is a paradigm that provides dynamically scalable and virtualized resource as a service over the Internet. The need to store, process, and analyze large amounts of data is making enterprise customers adopt cloud computing at scale. Cloud enables users to perform advanced analytics with big data. Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis. It is the umbrella of techniques used when trying to extract insights and information from data. Big Data Analytics the science of examining big data with the purpose of drawing conclusions and inferences. It is a subset of data science. Big data analytics is unimaginable without cloud in the current scenario. This paper discusses the convergence of big data, cloud and data science. It also identifies various issues in Big Data, Cloud, Data Science and their convergence.

Keywords
Big Data, Cloud, Data Science, Convergence, Issues

1. INTRODUCTION
IT industry is in a time of transition. Data is growing at an unconstrained and exponential rate. Organizations and people are now becoming more data centric. Researchers, academicians and business leaders believe that the world has entered the big data era.

Big data is the data that one is not able to process using pre-exist technology, method and theory. Big data is about any attribute that challenges constraints of a system's capability or business need. Put in simple words, big data is the ability to use the data that one has built up in the past to inform and improve what one is going to do in the future. Data gets more accurate and precise as time passes and more data about the data is created, calculated or inferred. Big Data tries to find meaning out of data that is constantly changing and to find relationships between data created. Understanding this interconnectedness and being able to harvest the information in Big Data unlocks the real value of big data. Big data provides opportunities but also poses challenges for businesses. Big data must be processed and analyzed in a timely manner in order to extract value from it. The results of the analysis need to be available in such a way as to that they can influence business decisions.

The need to store, process, and analyze big data is making enterprise customers adopt cloud computing at scale. Cloud is large-scale distributed computing paradigm, providing storage, networking, computing and other resources to its users as a measured service, whenever and wherever needed. The reduced cost, rapid elasticity, efficient management, and availability of cloud appeals users.

When organizations want to gain value and profitability from Big Data, Data Science and Big Data Analytics come into picture. Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis. In simple terms, it is the umbrella of techniques used to extract insights and information from data. Big Data analytics is the process of collecting, organizing and analyzing large sets of data (Big Data) to discover useful information. Businesses have to know how to move. Big Data analytics help identify the right moves to make. Big Data Analytics is a part of Data Science.

This paper analyzes the convergence of big data, cloud computing and data science. The convergence of these three technologies is changing the IT and business world significantly. This paper also analyzes the relationship between Big Data, cloud and data science. The analysis of big data is confronted with many challenges. This paper discusses the issues in big data analytics through cloud computing. The issues are categorized in three categories: i) Issue in Big Data, ii) Issues in Cloud and iii) Issues unresolved after convergence of Big Data and Cloud. For successful big data application in cloud, these challenges need to be addressed and resolved.

The rest of the paper is organized as follows. Section II describes the relationship between big data, cloud and data science. Section III presents the convergence of big data, cloud and data science. Section IV presents the issues in big data analytics through cloud computing. Section V provides a summary of the work.

2. RELATIONSHIP BETWEEN BIG DATA, CLOUD AND DATA SCIENCE
The simple block diagram figure 2 depicts the relationship between big data, cloud and data science. Big Data due to its huge size, high velocity and vivid variety needs a platform capable of storing, processing and managing it. Cloud efficiently provides the required platform for storage, processing and management of big data. So, Big Data resides on the cloud. Big data is like crude oil — it needs filtering and refining to unlock its value and make it usable. To gain value and derive insights from big data, analytics needs to be performed on it. Data Science and big data analytics are involved in this process. The tools and algorithms of data science and big data analytics reside on the cloud. So, big data analytics is performed on the big data residing in cloud through cloud as a platform.

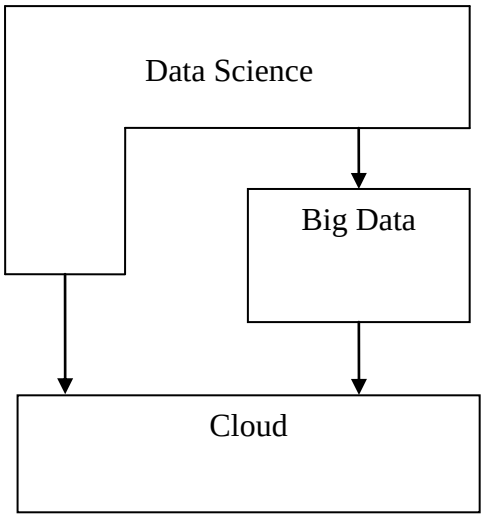

Fig 1: Relationship between Big Data, Cloud and Data Science

3. CONVERGENCE OF BIG DATA, CLOUD AND DATA SCIENCE

In solving any business use case, the following questions need to be answered –Where, How, What.

Cloud, Big Data and Data Science will create magic triangle where any business can harvest huge amount of insights, values and efficiency. Figure 1 depicts the magic triangle formed by cloud, big data and data science.

Cloud will answer the question – Where. It will cater to the question where to store data, where to process data. Hence, Cloud will provide the infrastructure for storage and processing Big Data.

Big Data will answer the question- How. Big data is responsible for answering the questions how to handle huge amount of data, how to store huge amount of data, how to process huge amount of data, how to do anything (clean, analyze etc.) with huge amount of data.

Data Science will answer the question – What. Data Science enables Predictive Analytics (What will happen) and also Prescriptive Analytics (What should happen).

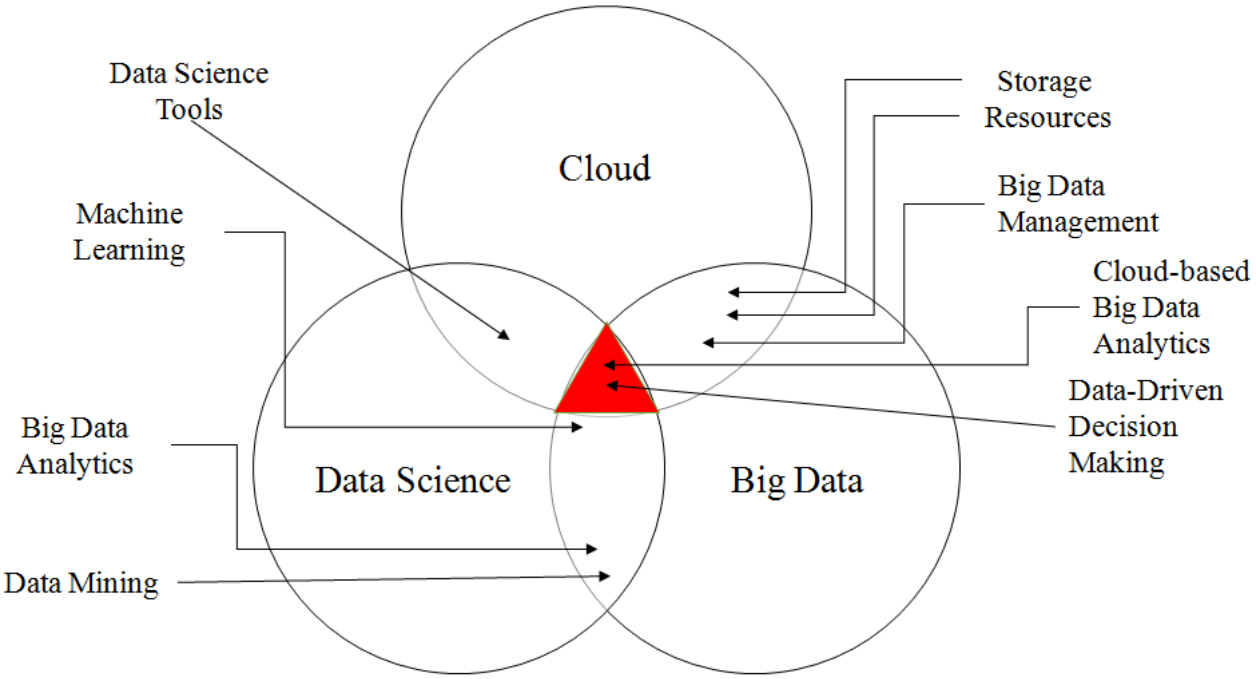

Fig 2: Convergence of Big Data, Cloud and Data Science

3.1 Big Data and Cloud

Big Data and cloud computing go hand-in-hand. The reasons for this complementary relationship are stated below.

Both Cloud and Big Data reduce the cost of ownership and deliver value to the enterprise. Forrester has defined big data as ―Technologies and techniques that make capturing value from data at an extreme scale economical. [1] Cloud computing enables fast analytics at a much lower cost than before. So, companies of all sizes can derive more value from their data than ever before. This, in turn drives companies to acquire and store even more data, creating more need for processing power and driving a virtuous circle.

Cloud provides everything as a service by hiding the complexity and challenges involving in building a scalable elastic self-service application. The same is the requirement for Big Data Processing. Hadoop hides the complexity of the large scale distributed processing from the end user in a similar way. Thus, the prime reason for the mass adoption of Big Data and Cloud is the simplification provided by Cloud and Big data.

Cloud computing and big data are complementary, forming a dialectical relationship. Cloud computing is a trend in technology development, while big data is an inevitable phenomenon of the rapid development of a modern information society. To solve big data problems, modern means and Cloud computing technologies are needed [2]. The breakthrough of big data technologies can make Cloud computing and the Internet of Things' technologies land on the ground and be promoted and applied in in-depth ways. It can be summarized that: IoT is the King, Big Data is the queen and Cloud is their palace.

The distributed storage technology of Cloud computing allows effective management of Big Data. The parallel computing capacity of cloud computing can improve the efficiency of acquiring and analyzing Big Data.

Big Data should be actionable data. The collection and analysis of data is worthless without appropriate action. Cloud computing due to its flexibility, makes the collection, analysis and dissemination of results and actions easier.

Moving it to the cloud makes sense because big data needs a lot of space. Many organizations use the cloud for their big data analytics. Cost savings in hardware and processing and the ability to experiment with big data technology are the benefits of implementing big data technology through cloud computing. Several models of cloud computing services are available to the businesses to consider, with each model having trade-offs between the benefit of cost savings and the concerns data security and loss of control.

3.2 Big Data and Data Science

Data science is the secret sauce for an organization to leverage Big Data to gain value and profitability. When machine learning and data mining tools are applied to big data, they lead to big data analytics.

Big Data + Data Science = Big Data Analytics

Big Data = High Volume, High Variety, High Velocity Data

Data Science = Maths + Programming + Statistics +Domain expertise+ Tools

3.3 Data Science and Cloud

Analysis and storage are the two important challenges for organizations whether large or small. The amount of Big Data generated has accelerated tremendously. One of the top priorities of organization where Cloud comes into picture is storing big data in an economic and secure manner. This has given rise to the trend of hiring skilled data analysts, data engineers and above all data scientists. A data scientist must possess skills like analysis, statistics and programming. A data scientist is expected to work on newer platforms in which the organization stores data.

Data science and cloud computing essentially go hand in hand. A Data Scientist typically analyses different types of data that are stored in the Cloud. With the increase in Big Data, organizations are increasingly storing large sets of data online and there is a need for Data Scientists. Given that the storage is now much cheaper, and the open source platforms and tools are available for data scientists, cloud is the key.

3.4 Cloud-based Big Data Analytics

Cloud + Big Data + Data Science = Cloud-Based Big Data Analytics

The intersection of big data, cloud and data science leads to cloud-based big data analytics and data-driven decision making. Data-driven decision making (DDD) refers to the practice of basing decisions on the analysis of data rather than purely on intuition. Businesses can harvest huge amount of insights, values and efficiency using DDD.

4. ISSUES IN BIG DATA ANALYTICS THROUGH CLOUD

The sharply increasing data deluge in the big data era brings huge challenges. To solve big data problems, modern means and Cloud computing technologies are needed. The challenges in cloud based big data analytics are classified in three broad categories. Figure 3 depicts the categorization of issues in big data analytics through cloud.

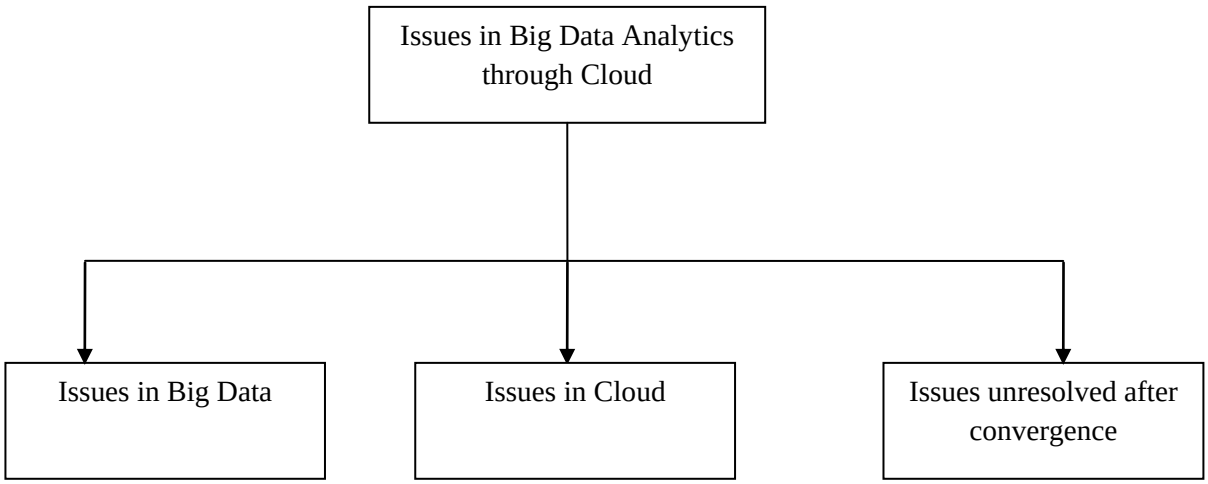

Fig 3: Categorization of Issues

4.1 Issues in Big Data

Big data brings not only opportunities but also challenges. Suitable solutions must be applied to overcome these challenges and gain valuable insights from big data.

4.1.1 Heterogeneity (Root Level)

The variety of data produced by the multiple sources like sensors, smart devices, and social media in raw, structured, semi-structured, unstructured and rich-media formats is

complicating the storage, processing and management of data. Certain levels of heterogeneity prevail in type, structure, semantics, organization, granularity, and accessibility of the datasets. Due to variety and heterogeneity, data representation is a significant challenge.

4.1.2 Storage Issues

The capability of existing storage technologies to store and manage data is restricted by the rapid growth of data. Over the past few years, traditional storage systems have been utilized to store data through structured RDBMS. However, almost every storage system has its own limitations and is inapplicable to the storage and management of big data. To store and manage large dataset, there is a need of a storage architecture that can be accessed in a highly efficient manner while achieving availability and reliability is required to store and manage large datasets.

4.1.3 Data Transportation Issue

It would take longer to transmit the data from a collection or storage point to a processing point than it would to actually process it. Current disk technology limits are about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. Even if an Exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Access to that data would overwhelm current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an Exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained [3].

4.1.4 Data Management Issues

Two things, big data and data management constitute big data management. Also, how the two work together to achieve business and technology goals falls under BDM. Big data is about high volume, high velocity data that comes from multiple sources and is multi-structured (structured, semi-structures, unstructured). Data management involves the collection and storage of data, plus its processing and delivery. Data management covers a number of data disciplines, including data warehousing, data integration, data quality, data governance, content management, event processing, database administration, and so on. Traditional data and new big data can differ in terms of content, structure, and intended use, and each category has many variations within it. Management will be the most difficult problem to address with big data. Resolving issues of access, metadata, utilization, updating, governance, and reference have proven to be major problem areas. There is no perfect big data management solution yet.

4.1.5 Processing Issue

Extensive parallel processing and new analytics algorithms will be required for effective processing of exabytes of data in order to provide timely and actionable information. Speed is a significant demand, while processing a query in big data However, this may cost a lot of time because it cannot traverse all the related data in the whole database in a short time. In this case, index will be an optimal choice. A desirable solution will be the combination of appropriate index for big data and up-to-date preprocessing technology.

4.1.6 Data Security Issue

At present most big data service providers or owners, because of their limited capacity, could not effectively maintain and analyse huge datasets. They must rely on professionals or tools to analyze the data. This increases the potential safety risks. Only when proper preventive measures are taken to protect the sensitive data, to ensure its safety, the analysis of big data should be delivered to a third party for processing.

4.1.7 Scalability

Managing large and rapidly increasing volumes of data is a challenging issue. Scalability has three aspects: Data Volume, Hardware Size and Concurrency. Data volume is increasing at an exponential rate. The size of hardware required for big data usage is quite large and increasing with the volume requirements. The analytical system of big data must support present and future datasets. The analytical algorithm must be able to process increasingly expanding and more complex datasets. A Distributed and scalable architecture required to store, process and analyse big data. High level of concurrency required. Extensive parallel computing is needed to handle data.

4.1.8 Timeliness/Response Time

Speed is the flip side of size. The larger the data set to be processed, the longer it will take to analyze. The challenge is having tight response time limits.

4.2 Issues in Cloud

Because users are still skeptical about the authenticity of the cloud, the current adoption of cloud computing is associated with numerous challenges. Based on a survey conducted by IDC in 2008, [4] the major challenges that prevent Cloud Computing from being adopted are recognized by organizations are described below.

4.2.1 Security

The security issue has played the most important role in hindering Cloud computing acceptance. Cloud computing and storage solutions provide users and enterprises with various capabilities to store and process their data in third-party data centers. Organizations use the Cloud in a variety of different service models (SaaS, PaaS, and IaaS) and deployment models (Private, Public, Hybrid, and Community). There are a number of security concerns associated with cloud computing [4]. These issues fall into two broad categories:

1. security issues faced by cloud providers (organizations providing software-, platform-, or infrastructure-as-a-service via the cloud) and
2. security issues faced by their customers (companies or organizations who host applications or store data on the cloud).

4.2.2 Costing Model

The cost of transferring an organization's data between the public and community Cloud and the cost per unit of computing resource used is likely to be higher. The problem is particularly eminent if the consumer uses the hybrid cloud deployment model where the organization's data is distributed amongst a number of public/private/community clouds. In addition, the cost incurred for data integration can be very high as different clouds often use proprietary protocols and interfaces.

4.2.3 Charging Model

A sound charging model needs to incorporate cost based on consumptions of static computing as well as VM associated items such as software licenses, usage of virtual network, node and hypervisor management overhead etc. A critical and viable charging model for Software as a Service provider is

important for the profitability and acceptability of cloud providers.

4.2.4 Service Level Agreements (SLA)

It is important for consumers to obtain guarantees from providers on service delivery which are provided through Service Level Agreements (SLAs) between the providers and consumers. The agreement should be specified in such a way that has an appropriate trade-off between expressiveness and complicatedness, can cover most of the consumer expectations and is relatively simple to be verified, evaluated and enforced by the resource allocation mechanism on the cloud.

4.2.5 Cloud Interoperability

Each cloud has its own way on how clients, applications and users interact with the cloud. Proprietary clouds APIs make it difficult to integrate cloud services with an organization's own existing legacy. This severely disrupts the cloud ecosystems development. The interoperability should be prevalent between both amongst different clouds and the connection between a cloud and an organization's local systems.

4.3 Issues unresolved after convergence of big data with cloud

Research on big data in the cloud remains in its early stages, although cloud computing has been broadly accepted by many organizations. Several existing issues have not been fully addressed. New challenges continue to emerge from applications by organization. Due to convergence of big data and cloud, most issues of big data and cloud are resolved. Some remain unresolved. The following issue remained unresolved after big data and cloud convergence.

4.3.1 Heterogeneity (Leaf Level)

Interoperability can mean different things to different people. One can mean the ability of applications to move from one environment to the next, and for the applications to work exactly the same in both places. Another might mean applications running in different clouds being able to share information, which might require having a common set of interfaces. To others, cloud interoperability refers to the ability of customers to use the same management tools, server images and other software with a variety of cloud computing providers and platforms. The essence of the problem, though, is that each vendor's cloud environment supports one or more operating systems and databases. Each cloud contains hypervisors, processes, security, a storage model, a networking model, a cloud API, licensing models and more. When big data analytics is performed on the cloud platform, Cloud Interoperability is a concern. It makes adopting and migrating applications and data to the cloud difficult. Challenges here include vendor lock-in, technology lock-in, and licensing related issues [5].

4.3.1 Data Integrity

A key aspect of big data security is integrity. Integrity means that data can be modified only by authorized parties or the data owner to prevent misuse. Users get the opportunity to store and manage their data in cloud data centers because of the rapid increase of cloud-based applications. Data integrity must be ensured in such applications. One of the main challenges to be addressed is safeguarding the correctness of user data in the cloud.

4.3.2 Data Quality

Earlier, data processing was performed on clean datasets from well-known and limited sources. Therefore, the results were accurate. With the advent of big data, data originate from multiple sources; not all of these sources are verifiable. Because data are often collected from different sources, poor data quality has become a critical issue for many cloud providers. For example, huge amounts of data are generated from smart-phones, where inconsistent data formats can be produced as a result of heterogeneous sources. The data quality problem is usually defined as "any difficulty encountered along one or more quality dimensions that render data completely or largely unfit for use". Hence, obtaining data of high-quality from vast collections of data sources is a challenge.

4.3.3 Privacy

Privacy concerns continue to hinder users who out- source their private data into the cloud storage. This concern has become serious with the development of big data mining and analytics, which require personal information to produce relevant results, such as personalized and location-based services. Information on individuals is exposed to scrutiny, a condition that gives rise to concerns on profiling, stealing, and loss of control. Currently, encryption is utilized by most researchers to ensure data privacy in the cloud.

4.3.4 Governance

Data governance embodies the exercise of control and authority over data-related rules of law, transparency, and accountabilities of individuals and information systems to achieve business objectives. The key issues of big data in cloud governance pertain to applications that consume massive amounts of data streamed from external sources. Therefore, a clear and acceptable data policy with regard to the type of data that need to be stored, how quickly an individual need to access the data, and how to access the data must be defined. Big data governance involves leveraging information by aligning the objectives of multiple functions, such as telecommunication carriers having access to vast troves of customer information in the form of call detail records and marketing seeking to monetize this information by selling it to third parties. Big data provides significant opportunities to service providers by making information more valuable. However, policies, principles, and frameworks that strike stability between risk and value in the face of increasing data size and deliver better and faster data management technology can create huge challenges. Cloud governance recommends the use of various policies together with different models of constraints that limit access to underlying resources [5].

4.3.5 Legal/Regulatory Issues

Different countries have different laws and regulations to achieve data privacy and protection. In several countries, monitoring of company staff communications is not allowed. However, electronic monitoring is permitted under special circumstances. Therefore, the question is whether such laws and regulations offer adequate protection for individuals' data while enjoying the many benefits of big data in the society at large.

5. SUMMARY

This paper described the convergence of big data, cloud and data science. A relationship between big data, cloud and data science is established. The paper also identified the various issues in big data analytics through cloud computing. The findings of the study can be summed up as follows. Cloud is a facilitator of big data analytics. Cloud and big data converge to become a competitive advantage. More data results into

more accurate analysis. More accurate analysis results into better decision making. Better decisions result into better operational efficiencies, cost reductions and reduced risk. To realize the full potential of big data, the challenges posed by it must be addressed. Cloud platform can handle the challenge of storing, processing, transporting, managing big data. Proper measures should be taken for data security and privacy. Skilled humans (Data Scientists) need to be involved. For better utilization of Cloud, the challenges posed by it need to be handled. Security, service level agreements, costing and charging model and cloud interoperability are the key challenges of cloud. Big data analytics on cloud platform faces many challenges. Heterogeneity, privacy, data integrity, data quality, governance and legal issues are the major challenges. For successful big data analytics on cloud, all these issues must be resolved. Both big data and cloud continue to evolve. As these technologies advance, many challenges get resolved on the way. Data driven decisions can be made by businesses who manage to handle these issues. They gain competitive advantages and increase their profitability.

6. REFERENCES

[1] H. Kisker, "Big Data Meets Cloud", August 15, 2012, http://blogs.forrester.com/holger_kisker/12-08-15-big_data_meets_cloud [Last Accessed On: 2-Jan-2017]

[2] W. Tian and Y. Zhao "Big Data technologies and Cloud Computing", Optimized Cloud Resource Management and Scheduling, DOI: http://dx.doi.org/10.1016/B978-0-12-801476-9.00002-1, 2015 Elsevier Inc.

[3] S. Kaisler et al, " Big Data: Issues and Challenges Moving Forward ", in 2013 46th Hawaii International Conference on System Sciences, 2012 IEEE, DOI 10.1109/HICSS.2013.645

[4] Gens, F. "New IDC IT Cloud Services Survey: Top Benefits And Challenges". IDC eXchange., 2009. [Online]. [Last Accessed:17 December 2016].

[5] I.A.T. Hashem et al.,"The Rise of "Big Data" On Cloud Computing: Review and Open Research Issues", Information Systems, vol 47, pp 98–115, 2016