Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Artificial Intelligence

APPLANT

Dr.Keskin

24 May 2021, Delft, The Netherlands

Email: keepgoing@applant.nl

CONTENT

Abstract.

Cloud computing and big data analytics are, without a doubt, two of the most important technologies to enter the mainstream IT industry in recent years. Surprisingly, the two technologies are coming together to deliver powerful results and benefits for businesses. Cloud computing is already changing the way IT services are provided by so called cloud companies and how businesses and users interact with IT resources.

Big Data is a data analysis methodology enables by recent advances in information and communications technology.

But big data analysis requires a huge amount of computing resources making adoption costs of big data technology is not affordable for many small to medium enterprises.

In this paper, we outline the the benefits and challenges involved in deploying big data analytics through cloud computing. We argue that cloud computing can support the storage and computing requirements of big data analytics. We discuss how the consolidation of these two dominant technologies can enhance the process of big data mining enabling businesses to improve decision-making processes. We also highlight the issues and risks that should be addressed when using a so called CLaaS, cloud-based service model.

Search and review under responsibility of DR.KESKIN

Keywords: Cloud Computing, Big Data Analytics, Cloud Analytics, Security, Privacy, Artificial Intelligence, AWS, AaaS, CLaaS

1. INTRODUCTION

Artificial intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing (NLP), speech recognition and machine vision.

Organizations are being compelled to capture, understand and harness their data to support decision-making in order to improve business. The main purpose is to support better and faster business decision-making operations.

* Corresponding author.
E-mail address: ellaburcu@applant.nl

In an ever-changing business world, many companies now face growing pressure to develop and ramp up their artificial intelligence efforts quickly and at a low cost in order to remain competitive.

Recently emerged cloud computing is changing the way IT services are provided by companies and how businesses and users interact with IT resources. It represents a paradigm shift that introduces flexible service models that companies can subscribe on a pay-as-you-use model.

The data in the world is growing exponentially. Big data is an evolving term that describes any huge amount of structured, semi -structured and unstructured data that has the potential to be mined for information. Big data is data that exceeds the processing capacity of traditional databases. The data is too big to be processed by a single machine. The evolving field of big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. Big data technology has become possible with the latest developments in computer technology as well as algorithms and approaches developed to handle big data.

In this paper, our aim is to investigate the impacts of cloud computing and big data on businesses and analyse the benefits and challenges it brings to enterprises. First, we overview the concepts, issues technology of cloud computing and big data separately. We then present a framework that combines these two technologies to form an ideal platform for e-commerce. We discuss the role of big data in enhancing the main functional areas of B2B and e-commerce such as customer management, marketing, payments, supply chain and management.

2. RELATED WORK

Cloud computing popularity has prompted several academic and industry initiatives to explore capabilities and enhancements in cloud computing. The value proposition of cloud computing in comparison with on premise investments is one of the key research areas. There are several initiatives to specifically address the security issues and challenges in cloud computing.

There have been several academic initiatives investigating e-business model aspects of cloud computing. Aydin [2] discusses research of E-Commerce Based on Cloud Computing. Dan and Roger [4] compared various cloud offerings such as Google App Engine, Amazon EC2, and Microsoft Azure to provide guidance on cost, application performance (and limitations) for different deployment scenarios.

Agarwal et al [1] present various methods for handling the problems of big data analysis through Map Reduce framework over Hadoop Distributed File System (HDFS). In this paper, Map Reduce techniques have been implemented for Big Data analysis using HDFS. Yadav et al [20] present an overview of architecture and algorithms used in large data sets. These algorithms define various structures and methods implemented to handle Big Data and this paper lists various tools that were developed for analysing them. It also describes about the various security issues, application and trends followed by a large data set.

Fan and Bifet [6] present an overview of big data mining outlining its current status, controversy, and forecast to the future. This paper also covers various interesting and state-of-the-art topics on Big Data mining. Sharma and Navdeti [15] discuss about the big data security at the environment level along with the probing of built in protections. It also presents some security issues that we are dealing with today and proposes security solutions and commercially accessible techniques to address the same. The paper also covers all the security solutions to secure the Hadoop ecosystem. They also provide an overview on big data, its importance in our live and some technologies to handle big data. Jassena and David [8] discuss issues, challenges and solutions of big data mining. Padgavankar and Gupta [14] provide detail analysis

of the challenges involved in big data storage and propose some solutions to handle them. Jayasree [9] provides an overview of big data technologies such as MapReduce and Hadoop and compares with traditional data mining techniques. Zulkernine et al [22] presents a conceptual architecture for a cloud-based analytics as a service (CLaaS).

3. THE CLOUD COMPUTING PARADIGM

3.1 What is Cloud Computing?

Many researchers have defined cloud computing differently. One mostly accepted definition is given by the United States Institute of Standards (NIST). Per the NIST definition [11],

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, five service models, and four deployment models".

3.2 Cloud Computing Characteristics

Cloud computing has five essential characteristics. They are on-demand capabilities, broad network access, resource pooling, rapid elasticity and measured service. These are the characteristics that distinguish it from other computing paradigms.

On-demand Capabilities: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops and workstations).

Resource Pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned per consumer demand.

Rapid elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand.

Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth and active user accounts).

3.3 Cloud Deployment Models

Cloud deployment models are grouped broadly into four models: private cloud, public cloud, community cloud and hybrid cloud. ***Private cloud*** is the most secure way to utilize cloud computing. The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises. ***Community cloud*** is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns. It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises. ***Public cloud*** is provisioned for

open use by the public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider. ***Hybrid cloud*** is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability [2, 3].

3.4 Cloud Service Delivery Models

Cloud-based services are grouped broadly into four models: Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). Software as a Service (SaaS) is a model that provides the user with access to already developer applications that are running in the cloud. The access is achieved by cloud clients and the cloud users do not manage the infrastructure where the application resides, eliminating with this the way the need to install and run the application on the cloud user's own computers. Platform as a Service (PaaS): is a model that delivers to the user development environment services where the user can develop and run in-house built applications. The services might include an operating system, a programming language execution environment, databases and web servers. Infrastructure as a Service (IaaS) is a model that provides the user with virtual infrastructure, for example servers and data storage space. Virtualization plays a major role in this mode, by allowing IaaS-cloud providers to supply resources on-demand extracting them from their large pools installed in data centres. Data as a Service (***DaaS***) is a model in which, data is readily accessible through a Cloud-based platform. Simply put, DaaS is a new way of accessing business-critical data within an existing data centre. Figure 1 illustrates the general cloud computing architecture.

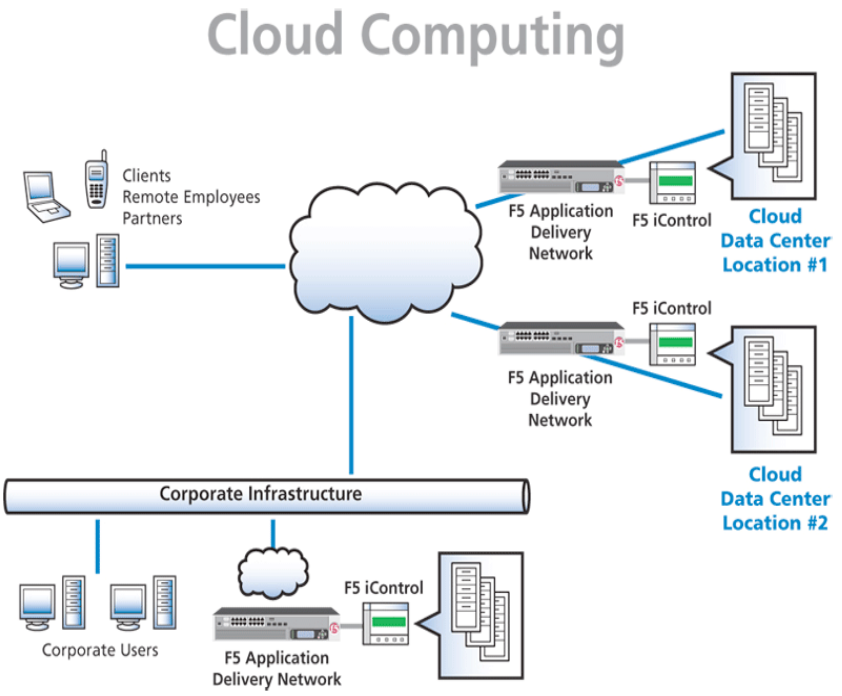

Figure 1: A Typical Cloud Computing Architecture

3.5 Cloud Computing Benefits

Cost Efficiency - This is the biggest advantage of cloud computing, achieved by the elimination of the investment in stand-alone software or servers. By leveraging cloud's capabilities, companies can save on licensing fees and at the same time eliminate overhead charges such as the cost of data storage, software updates, management etc. Renting your infrastructure can make good financial sense. The pay as you go (PAYG) model is especially

Continuous availability - Public clouds offer services that are available wherever the end user might be located. This approach enables easy access to information and accommodates the needs of users in different time zones and geographic locations. As a side benefit, collaboration booms since it is now easier than ever to access, view and modify shared documents and files. Moreover, service uptime is in most cases guaranteed, providing in that way continuous availability of resources. The various cloud vendors typically use multiple servers for maximum redundancy. In case of system failure, alternative instances are automatically spawned on other machines.

Scalability and Elasticity - Scalability is a built-in feature for cloud deployments. Cloud instances are deployed automatically only when needed and thus, you pay only for the applications and data storage you need. Hand in hand, also comes elasticity, since clouds can be scaled to meet your changing IT system demands.

Fast deployment and ease of integration - A cloud-based application can be up and running with just a few hours rather than weeks or months and without spending a large sum of money in advance. This is one of the key benefits of cloud. On the same aspect, the introduction of a new user in the system happens instantaneously, eliminating waiting periods.

Resiliency and Redundancy- A cloud deployment is usually built on a robust architecture thus providing resiliency and redundancy to its users. The cloud offers automatic failover between hardware platforms out of the box, while disaster recovery services are also often included.

Increased Storage Capacity- The cloud can accommodate and store much more data compared to a personal computer and in a way, offers almost unlimited storage capacity. It eliminates worries about running out of storage space and at the same time it spares businesses the need to upgrade their computer hardware, further reducing the overall IT cost

4. BIG DATA ANALYTICS

4.1 What is "Big Data"?

Big Data is the term for a collection of data sets so large and complex that it becomes difficult to process using conventional data mining techniques and tools. The overall goal of the big data analytics is to extract useful information from a huge data set and transform it into an understandable structure for further use. The major processes of big data include capture, curation, storage, search, sharing, transfer, analysis, and visualisation [12, 13, and 19].

Recently the importance of this field has attracted enormous attention because it gives businesses useful information and better insight of both structured and unstructured data, which may lead to better-informed decision-making [10, 16, 18]. In a business context, big data analytics is the process of examining "big data" sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information [7, 21]. Today's advances in technology combined with

the recent developments in data analytics algorithms and approaches have made it possible for organisations to take advantage big data analytics. Some of the major issues in applying big data analytics successfully include data quality, storage, visualization and processing [14].

Some business examples of big data are social media content, mobile phone details, transactional data, health records, financial documents, Internet of things and weather information.

4.2 Big Data Technologies

In order to support big data analytics, a computing platform should meet the following 3 criteria, so called 3 Vs as illustrated in Figure 2.

Variety: The platform supports wide variety of data and enables enterprises to manage this data as is in its original format, and with extensive transformation tools to convert it to other desired formats.

Velocity: The platform can handle data at any velocity, either low-latency streams, such as sensor or stock data, or large volumes of batch data.

Volume: The platform can handle huge volumes of at-rest or streaming data.

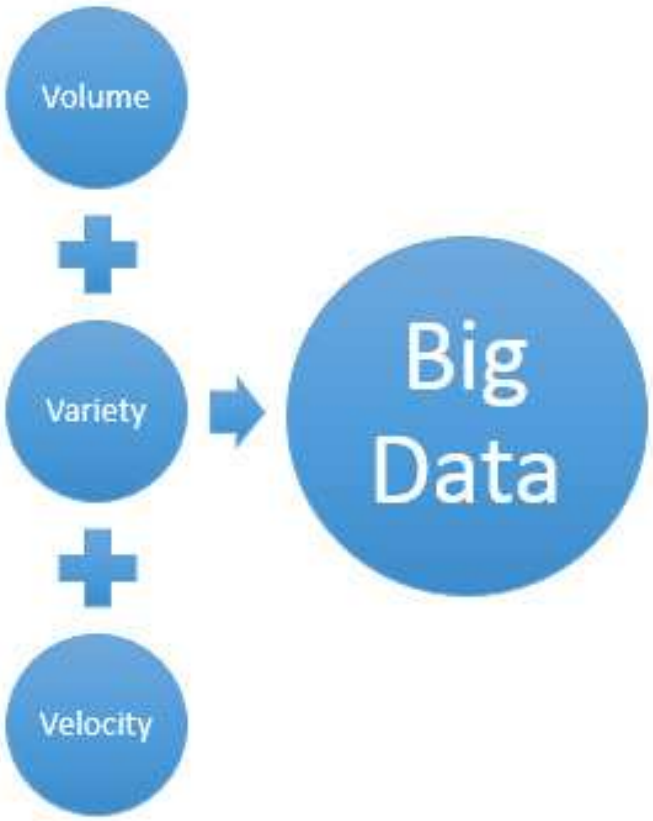

Figure 2. 3-V Criteria of Big Data

Traditional data mining involves finding interesting patterns from datasets whereas big data analytics involves large scale storage and processing of huge data sets. Traditionally Hadoop and MapReduce are two of the popular technologies for big data analytics [9].

More tools and technologies are becoming available for big data processing. Examples include Amazon's Redshift hosted BI data warehouse, Google's BigQuery data analytics service, IBM's Bluemix cloud platform and Amazon's Kinesis data processing service. The future state of big data will be a hybrid of on-premises and cloud, Alternatives to traditional SQL-based relational databases, called NoSQL (**Not Only SQL**) databases, are rapidly gaining popularity as tools for use in specific kinds of big data analytic applications

4.3 Big Data Benefits

The fact that the valuable enterprise data will reside outside the corporate firewall raises serious concerns. Some of the most common challenges are discussed below

Cost reduction - Big data technologies like Hadoop and cloud-based analytics can provide substantial cost advantages. While comparisons between big data technology and traditional architectures (data warehouses and marts) are difficult because of differences in functionality, a price comparison alone can suggest order-of-magnitude improvements. Rather than processing and storing vast quantities of new data in a data warehouse, for example, companies are using Hadoop clusters for that purpose, and moving data to enterprise warehouses as needed for production analytical applications.

Faster, better decision making - Analytics has always involved attempts to improve decision making, and big data doesn't change that. Following the Big data analytics really makes the business managers good decision makers. Large organizations are seeking both faster and better decisions with big data, and they're finding them. Driven by the speed of Hadoop and in-memory analytics, several companies are focused on speeding up existing decisions.

New products and services - Perhaps the most interesting use of big data analytics is to create new products and services for customers. Online companies have done this for a decade or so, but now predominantly offline firms are doing it too.

Product recommendation - It is obviously very clear that the adoption of big data and analytics have proved to be a very powerful strategy for online businesses. The influence of the huge data of the customers on the business is turning to be very significant and economic tool for strengthening a business. Storing and working on huge data has been always a challenge for any trade. Big data has constructed the road for managing such huge data making business much simpler and profitable.

Fraud Detection - High-performance analytics is not just another technology fad. It represents a revolutionary change in the way organizations harness data. With new distributed computing options like in-memory processing on commodity hardware, businesses can have access to a flexible and scalable real-time big data analytics solution at a reasonable cost. This is sure to change the way insurance companies manage big data across their business – especially in detecting fraud

5. DEPLOYING BIG DATA ANALYTICS IN THE CLOUD

Cloud-based big data analytics is a service model in which elements of the big data analytics process are provided through a public or private cloud [18, 20]. It uses a range of analytical tools and techniques to help businesses extract information from massive data and present it in a way that is easily categorised and readily available via a web browser. Such cloud-based data analytics applications and services are typically offered under a subscription-based or utility (pay-per-use) pricing model. This service model is called Cloud Analytics as a Service (**CLAaaS**) [22]. In this model, analytics is readily accessible through a cloud computing platform. Such cloud-based data analytics service will enable businesses to automate processes on an anytime, anywhere basis. Examples of such cloud-based analytics products and services include hosted data warehouses, software-as-a-service business intelligence (SaaS BI) and cloud-based social media analytics. Data stored in a cloud-based database can help businesses with their decision-making processes.

With cloud-based big data, analysts have not only more data to work with, but also the processing power to handle large numbers of records with many attributes. This has the ability to increase predictability. The combination of big data and cloud computing also lets analysts explore new behavioural data such as websites visited or location on a daily basis.

5.1 Major Benefits for Business Organisations

On-demand self-service: As the name describes, organisations can expand the storage or service at a click of the button without any human help. Organisations will can establish big data infrastructure as quickly as possible.

Data and Information over the net: Information is available over the network and can be accessed anytime through the net by different devices such as laptop, mobile, ipads etc

Resource pooling: Provider resources are grouped and used efficiently by multi-tenant model. Resources include storage, memory, VMs etc

Rapid elasticity: Resources (both hardware & software) can be increased or decreased efficiently and effectively in quick span of time. Customers can purchase the resources for any quantity and at any time.

Cost effective: Resource usage can be monitored and would be charged on the basis of usage. This system is very transparent which makes the provider and the user more comfortable to adopt it. Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.

5.2 Big Data and Cloud Computing Challenges

The fact that the valuable enterprise data will reside outside the corporate firewall raises serious concerns. Some of the most common challenges are discussed below [4, 9, and 17]:

Data Storage - Storing and analysing large volumes of data that is crucial for a company to work requires a vast and complex hardware infrastructure. With the continuous growth of data, data storage device is becoming increasingly more important, and many cloud companies pursue big capacity of storage to be competitive.

Data Quality - Accuracy and timely availability of data is crucial for decision-making. Big data is only helpful when an information management process is implemented to guarantee data quality.

Security and Privacy - Security is one of the major concerns with big data. To make more sense from the big data, organizations would need to start integrating parts of their sensitive data into the bigger

data. To do this, companies would need to start establishing security policies which are self-configurable: these policies must leverage existing trust relationships, and promote data and resource sharing within the organizations, while ensuring that data analytics are optimized and not limited because of such policies. **Hacking and various attacks** to cloud infrastructure would affect multiple clients even if only one site is attacked. These risks can be mitigated by using security applications, encrypted file systems, data loss software, and buying security hardware to track unusual behaviour across servers.

Service Delivery and Billing It is difficult to assess the costs involved due to the on-demand nature of the services. Budgeting and assessment of the cost will be very difficult unless the provider has some good and comparable benchmarks to offer. The service-level agreements (SLAs) of the provider are not adequate to guarantee the availability and scalability. Businesses will be reluctant to switch to cloud without a strong service quality guarantee.

Interoperability and Portability Businesses should have the leverage of migrating in and out of the cloud and switching providers whenever they want, and there should be no lock-in period. Cloud computing services should have the capability to integrate smoothly with the on premise IT.

Reliability and Availability Cloud providers still lack round-the-clock service; this results in frequent outages. It is important to monitor the service being provided using internal or third-party tools. It is vital to have plans to supervise usage, SLAs, performance, robustness, and business dependency of these services [3].

Performance and Bandwidth Cost Businesses can save money on hardware but they must spend more for the bandwidth. This can be a low cost for smaller applications but can be significantly high for the data-intensive applications. Delivering intensive and complex data over the network requires sufficient bandwidth

All these challenges should not be considered as road blocks in the pursuit of cloud computing. It is rather important to consider these issues and the possible ways out before adopting the technology.

6 CONCLUSIONS

Businesses have long used data analytics to help direct their strategy to maximise profits and support their decision-making processes. Today it is widely accepted that cloud computing and big data technologies are two dominant technologies that will shape up the business world. Cloud is no longer just a buzzword – it's a fact-of-life affecting every facet of the technology industry. Big data technologies provided through cloud computing will allow businesses to make proactive, knowledge-driven decisions as it allows them to have future trends and behaviours predicted. Businesses will be able to store their data remotely and access data and services from anywhere and anytime. Further, cloud-based data analytics provides the infrastructure that companies would otherwise have to build up themselves from scratch.

Alongside data analytics, cloud computing is also capable of keeping businesses stay competitive by providing many benefits such as cost effectiveness, resource pooling, on-demand service, rapid elasticity, and ease of management. Despite these benefits, there are some challenges and drawbacks, particularly in relation to privacy and security. Before investing in cloud-based big data analytics, an organisation needs to fully grasp the extent of what's involved. Investing in cloud analytics can be profitable for an organization but proper planning is essential to ensure that all phases of analytics elements are covered.

REFERENCE LIST

[1] Agarwal, D., Das, S. and Abbadi, A. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM 978-1-4503-0528-0/11/0003. Retrieved from: http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pd

[2] Aydin, N. (2015). Cloud Computing for E-Commerce, Journal of Mobile Computing and Application. Vloume 2, Issue, 1, pp 27-31.

[3] Barthelus, L. (2010). adopting cloud computing within the healthcare industry: opportunity or risk? Online Journal of Applied Knowledge Management, Volume 4, Issue 1.

[4] Dan, S and Roger, C. (2010). Privacy and consumer risks in cloud computing, Computer Law and Security Review, Vol 26, pp: 391-397.

[5] Fan, J., Han, F. & Liu, H., 2013. Challenges of Big Data Analysis. ResearchGate, 1(1), pp.1-38.

[6] Ilieva G, Yankova T. and Klisarova, S. (2015). Big Data Based System Model of Electronic Commerce, Trakia Journal of Sciences, Vol. 13, Suppl. 1, pp 407-413.

[7] Fan, W. and Bifet, A. "Mining Big Data, Current Status and Forecast to the Future". SIGKDD, Explorations, 14(2), pp:1-5.

[8] Jaseena K.U and David J.M. (2014). ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING, Natarajan Meghanathan et al. (Eds) : NeTCoM, CSIT, GRAPH-HOC, SPTM – 2014, pp. 131–140.

[9] Jayasree, M. (2013). Data Mining: Exploring Big Data Using Hadoop and MapReduce, International Journal of Engineering Science Research, IJESR, Volume 04, Issue 01.

[10] Khan, I., Naqvi, S.K. Alam, M. Rizvi, S.N.A. (2015). Data model for Big Data in cloud environment. Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference. pp. 582 -585

[11] Mell, P, and Grance, T. (2011). The NIST Definition of Cloud Computing, National Institute of Standards and Technology Special Publication.

[12] Neaga, I. and Hao, Y. (2014). A Holistic Analysis of Cloud Based Big Data Mining. International Journal of Knowledge, Innovation and Entrepreneurship. Volume 2 No. 2, 2014, pp. 56—64. Retrieved from: http://ijkie.org/IJKIE_December2014_IRINA&HAO.pdf

[13] NESSI. (2012). Big Data: A New World of Opportunities. Retrieved from:

http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf

[14] Padgavankar M.H and Gupta, S.R. (2014). Big Data Storage and Challenges, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2218-2223.

[15] Sharma, P, Navdeti, C. (2014). "Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", IJCSIT, Vol 5(2), 2126-2131

[16] Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis : Challenges and Solutions", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV

[17] Rachana, R.C and Guruprasad, H.S (2014). Emergency Security Issues and Challenges in Cloud Computing, International Journal of Engineering Science and Innovative Technology (IJESIT) , Vol 2, Issue 2, pp 485-490.

[18] Pareek,A and Gupta, M. (2012). Review of Data Mining Techniques in Cloud Computing Database, International Journal of Advanced Computer Research Volume-02, Number-02, Issue-01.

[19] Shukla, D.P., Patel, S.B, and Sen, A.K. (20114). A Literature Review in Health Informatics Using Data Mining Techniques, International Journal of Software and Hardware Research in Engineering, Volume 2, Issue 2, pp.123-129.

[20] Talia, D. (2013). Clouds for Scalable Big Data Analytics. Published by IEEE Computer Society.

[21] Yadav, C. Wang, S. and Kumar M. (2013). "Algorithm and Approaches to handle large Data- A Survey", IJCSN, Vol 2, Issue 3, ISSN: 2277-5420.

[22] Zulkernine, F. Bauer, M. and Aboulnaga,A.(2013). Towards Cloud-based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud, 2013 IEEE International Congress on Big Data