Big Data and Cloud Computing: Challenges and Opportunities

Sheetal Singh
Department of Computer Science
Ramanujan College, University of Delhi

Vipin Kumar Rathi
School of Computer and Systems Sciences
Jawharlal Nehru University, New Delhi

Bhawna Chaudhary
School of Computer and Systems Sciences
Jawharlal Nehru University, New Delhi

Abstract - A brief survey of Big Data and Cloud Computing has been presented in this paper. These two fields have gained tremendous momentum in the recent years and have attracted attention of several researchers. The paper describes the concepts and characteristics of Big Data along with a number of tools like HADOOP, Apache Spark, HPCC for managing Big Data. It also presents an overview of Cloud Computing and the services provided by the Cloud (SaaS, PaaS, IaaS). We have also explored several cloud platforms which offers storage, power and infrastructure such as Google Cloud Platform, Microsoft Azure, Amazon Web Services and so on. Some challenges and issues related to the fields of Big Data and Cloud Computing have also been highlighted.

Keywords – Big data, Cloud Computing, Distributed file system, Platforms,Services

I. INTRODUCTION

The continuous growth of computational resources generates an enormous amount of data every day. Data is available in abundance but it is difficult to extract useful information from such Big Data. The prominent social networking Website, Facebook, serves 0.57 trillion page hits in a month, stores 3000 million new pictures every month and handles 25 billion segments of content [13]. The new Paradigms such as Cloud computing, Grid computing, Distributed Systems have access to huge amounts of computational power by accumulation of resources and they offer a single system image view. Cloud computing becomes a solid platform for performing large scale complex computing. The most fundamental intent of these computing technologies is to provide a mechanism or solution for handling Big Data.

According to Gartner's Hype Cycle for Emerging Technologies, Big Data and Cloud Computing have been identified as the recent emerging technologies [14]. Cloud Computing is a new paradigm which provides infrastructure for computing and processing of all types of data resources. The technologies based on cloud have been adopted to deal with the large amounts of data.

The Internet of things is all set to bring the revolution in the information industry as it provides the interconnectivity of physical objects and allows them to exchange the data with the other connected devices. According to [13], the number of connected devices in 2014 was 3. 7 billion and this number is estimated to reach 25 billion till 2020. These interconnected devices, growing exponentially in number, are also giving rise to new types of data in large volumes. Big Data here becomes extremely important to convert this data into information.

We are creating 2.5 quintillion bytes of data everyday by using handheld devices, Internet of things, Clouds, mobile networks, social networks, online machine to machine communication [29]. The concept of Big Data has been proposed to store and manage, visualize and analyze such enormous volumes of data generated quickly per day. Cloud Computing provides a platform to the users that is accessible and flexible for storage and processing of such Big Data applications.

This paper is organized as follows: Section 2 gives an overview of Big Data and Cloud Computing Technologies. In Section 3, we discuss some emerging trends in the field of Big Data and Cloud Computing. Section 4 describes the tools and platforms for managing Big Data with cloud technologies. Section 5 mentions Challenges and Opportunities of these emerging technologies. Section 6 concludes this paper.

II. Overview of Big Data and Cloud Computing Technologies

A. Cloud Computing:

Cloud computing is a new archetype for enabling pervasive and favorable network access to a shared pool of resources for computing where different services such as servers, storage, network applications and data center fabric are delivered to customers devices and systems over the Internet and other computing services which can be promptly provided with minimum effort and service provider interaction [1]. In Cloud Computing, the word *Cloud* stands for *Internet*. Hence, it is a technology in which delivery of computing resources takes place over the Internet. Cloud Computing focuses on increasing the power of computation to an extent that a millions of instructions can be executed per second.

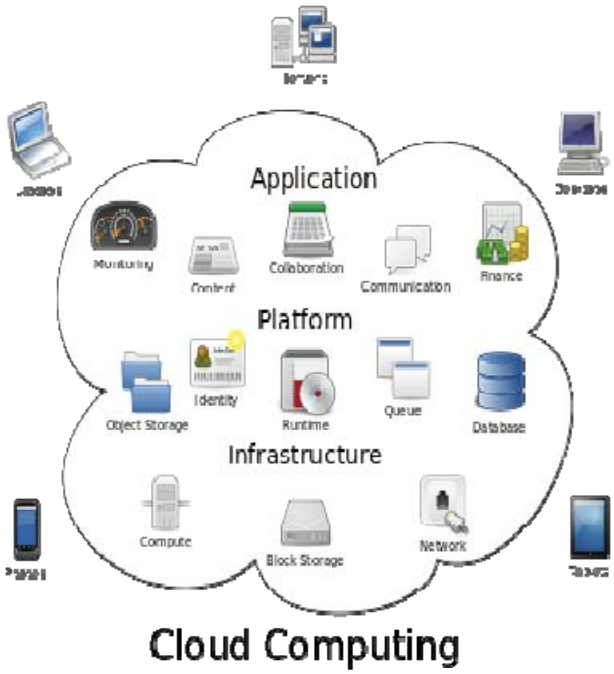

Figure 1. Cloud Computing environment[28]

The swift development of computing technologies and the success of internet, the resources have become much more powerful and are available ubiquitously. Cloud computing is one model which has enabled the availability of the resources like CPU, storage etc as general services that can be rented and released by the users according to their demand through internet [2].

The essential characteristics for a cloud computing model as defined by [1] [19] are:

- **On-demand self service:** The users who are accessing the Cloud services can interact with the Cloud to perform several tasks like building and deploying applications, managing and scheduling. Computing capabilities including processing power, storage, virtual machines etc are available to the user as and when needed without the requirement for human association.

- **Broader network access:** Cloud computing services provide accessibility to the solutions for business management techniques via smartphones, tablets, laptops, and office computers. These gadgets can be used via a simple online access point at any place. Broad network access incorporates public clouds, private clouds, or a hybrid deployment.
- **Resource Pooling:** Cloud administration suppliers create a shared pool of resources that is accessible by everyone and then these resources are available for use by a large number of customers through that pool. It uses multi-tenancy in which the resources are allocated and de-allocated to the users on demand, dynamically.
- **Rapid elasticity:** The allocation of resources is elastic that changes instantly and quickly according to demand. Users can scale out the resources by releasing them back to the Cloud that is not needed anymore. Hence, they can gain more resources by again scaling in their resource requirement.
- **Measured service:** The resource usage is monitored by measuring CPU hours, storage space usage, bandwidth usage, etc. The above said metrics is applied to all the clouds, but each cloud provides services according to their organizational policy with different levels of abstraction, providing an alternate to the administration.

The role of the cloud provider is to manage the cloud and resources available in cloud, rent the resources from one cloud to another to serve the end users. The emerging progression of cloud computing has made the Multi National firms like Google, Rackspace , Amazon , Microsoft etc to provide cloud platforms to gain benefit from this new paradigm.

A cloud can collaborate with its customers in a mixed bag of courses, through capacities called services. Three major types of service models have emerged across the web [21].

Software as a service (SaaS) model also known as On-demand service provides the capability to the user to access the software that is offered as a service over the web through applications running on the cloud infrastructure. The users need not to know about how the system, network, storage or individual applications. Google Apps , Microsoft Office 365 is a prominent example of SaaS [20].

Platform as a Service (PaaS) is a platform for developers to create and write their own Software as a service that is applications and associated services onto the cloud infrastructure without installation or downloads. Users need to keep control on the sent applications and the environment arrangements facilitating the applications. Google App Engine, Salesforce.com is based on the PaaS model.

Infrastructure as a service (IaaS) model gives infrastructure as demand by users including storage, hardware, processing, servers and networking components. The client has full control on storage, operating systems, and deployed applications. Some examples of IaaS are OpenStack, Amazon Web Services, Eucalyptus, Cloud Stack and Open Nebula are prominent examples that use the IaaS model of cloud computing [21].

Cloud computing, an Internet cloud, is being used to make cost efficient computing resources. It can be either distributed or a centralized computing Resource; most of the time cloud is built with Virtualized resources but it is also possible to built cloud using Physical resources. Cloud Computing is also known as utility Computing. Cloud can be divided in to two types' public cloud or private cloud; *public cloud* is accessible to the general users in pay-as – you- go fashion. The *private cloud* is the one which is internal to an organization and not available to the general users.

B. BIG DATA:

Big Data signifies the data sets that are enormously large and complex. The traditional data processing applications are insufficient to deal with the Big Data. The term "Big Data [8]" is believed to be generated from the website research. Describing Big Data as just great volume of data is simply not enough, neither justified. The data represents a great variety; and it can be processed in different ways, depending on the analysis. According to Gartner: "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"[22].

The three terms that signify Big Data [8] [3] are :

Volume refers to the ever growing data expanding beyond terabytes (for example transaction data, sensor data etc).

Variety refers to the data that is collected from heterogeneous sources such as machines, sensors etc making it tough to manage (for example emails, audio-visuals, text documents etc.)

Velocity refers to the pace of the generation of new data and also the fact that how fast it must be processed. In fact the data can become obsolete in a very short time.

Variability and *Veracity* are the two other dimensions that are of certain concern regarding Big Data. *Variability* defines the inconsistency of data at different intervals of time, thereby hampering the reliability and effective management of data; whereas *Veracity* refers to the traits of the data that is being captured. The analysis of the data depends on the data captured.

Figure 2. Big Data Characteristics[29]

With the rapid development of networking technologies and the emerging trend of Cloud Computing, Big Data are now amplifying in all the domains of engineering technologies. A large amount of data is produced per day. A survey identified that 90 percent of the total data in the world were produced in the last 2 years [5]. As an example to consider, more than 10 million tweets were recorded within the first two hours of the presidential debate held between US President, Barack Obama and the Governor, Mitt Romney in 2012 [6]. Confident decision making depends on the accuracy of Big Data which in turn leads to greater operational efficiency, reduced costs and reduced risks. The unprecedented volumes of data require an emphatic data analysis and prediction platform to achieve quick response and real-time classification for aforesaid Big Data.

III. EMERGING TRENDS IN BIG DATA AND CLOUD COMPUTING

In the recent years, context awareness has widely manifested its importance in achieving optimized management of resources, systems and services in many application domains. The urge to store, manage, and handle the ever increasing amounts of data is being felt. Another issue is concerned with the integration of multiple data sources in an automated way to aggregate and store such heterogeneous and large amounts of data for conducting analysis on the combined data set. In particular, the Internet of Things (IoT) has given rise to new types of data, for instance

data emerging from the collection of sensor data and the control of actuators. [15] Have analyzed the challenges and the requirements of Big Data coming from different Smart City applications. They proposed a solution that uses Big Data technologies for redesigning an IoT context aware application for the exploitation of pervasive environment.

The security issues for Cloud Computing and Big Data have been discussed by [9]. They proposed several possible solutions for the issues related to cloud computing security using Hadoop environment such as File and Network Encryption, Logging, Nodes Authentication, Layered Framework for Assuring Cloud etc. A lot of work has been done in the field of Big Data and cloud computing. For instance, a processing Model was proposed by [3] for Mining of Big Data. They proposed the HACE theorem for characterizing the features of Big Data as "The Heterogeneous data and the Autonomous sources with decentralized control seek to explore Complex and Evolving relationships among data".

A detailed analysis between Big Data and cloud computing has been done by [10]. They discussed on the security issues and challenges on the cloud computing types and the service delivery types. He discussed various challenges like Security, Cost, Service Level Agreements etc and also briefly described the benefits of both. The relationship between cloud and Big Data has been studied by [11]. They have discussed various pros and cons of cloud and Big Data. They also discussed about Apache Hadoop, a tool for managing structured or unstructured Big Data.

[12] Proposed a solution to the problem of limited usage of Cloud by the businessmen due to the problem of moving their data (measuring in tens of terabytes) to and fro of the cloud. **Aspera** has been proposed as a solution to solve the technical issues of the WAN as well as the cloud Input Output bottlenecks, thereby delivering incomparable performance for transferring Big Data in and out of the cloud.

IV. BIG DATA MANAGEMENT TOOLS

A.HADOOP

HADOOP is a framework launched by Apache providing a distributed environment that enables storage and processing of Big Data. It is spread across clusters of computers using simple programming models. It is a java based programming framework that uses a Master/Slave structure. Hadoop is a platform where a large number of data sets are processed over a cluster of servers and applications can be run on systems with a huge number of nodes [16]. It provides fast data transfer rates and since it is a distributed file system, the system does not break down even in the case of failure of a number of nodes. The Hadoop framework is scalable, cost effective, fault tolerant and flexible. Prominent MNCs' like IBM, Google, Yahoo etc have been using Hadoop for supporting their applications consisting of large volumes of data. The two main Sub projects of Hadoop are – Map/Reduce and Hadoop Distributed File System (HDFS).

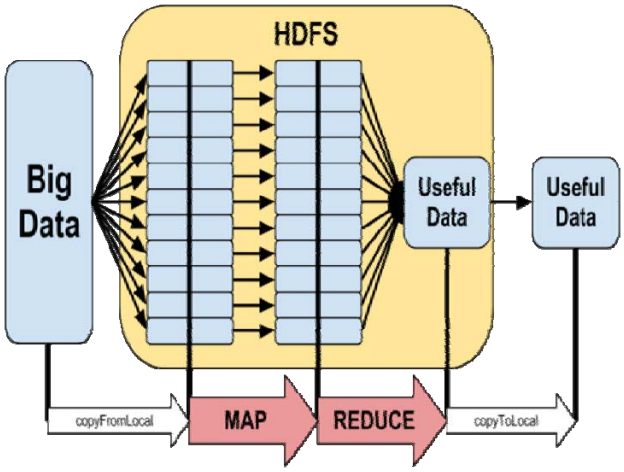

Figure 3. HADOOP Encironment

HDFS:

Hadoop Distributed File System is a block-structured system especially designed to hold large amounts of data, in a reliable and scalable manner that is easy to operate. It is an open source file system that processes large volumes of data in a distributed computing environment. The blocks are called as chunks and the default size of a block is 64 MB. HDFS is based on a client-server architecture which is comprised of many nodes. Metadata is stored in memory; hence HDFS offers a very high speed in terms of operations per second.

HDFS being written in java offers great portability. It also offers reliability since all the files are replicated on more than one node. Default number of replicas is three. HDFS was developed on the fact that moving the computation is easier than moving the data especially when the data is Big [17].

MAP REDUCE:

Hadoop MapReduce is a software programming framework having clusters containing thousands of nodes connected parallelly for processing of large data sets measuring in terabytes in a reliable and fault tolerant manner [18]. The *map tasks* processes the data sets, usually split into independent blocks, in a parallel manner. The framework then sorts the maps and they are sent as input to *reduce tasks*. The framework also handles scheduling, and monitors the execution of tasks. Basically Hadoop MapReduce framework is meant majorly for batch processing.

The MapReduce framework also works on the Master/Slave architecture. The master node is responsible for the scheduling of jobs, monitoring of jobs and re-execution of the failed tasks. The responsibility of the slave nodes is to execute the tasks as directed by the master node.

B. APACHE SPARK

Apache Spark is distributed cluster computing system to speed up the data analytics and it is open source. It's in-memory primitives make it a better framework that provides a much faster performance for certain applications. It is based on a general execution model which allows the user programs to load the data into a cluster's memory thereby helping in in-memory computing and optimization.

Spark is based on two things; first one is a cluster manager and the second is a distributed storage system[23]. Hadoop Yarn and Apache Mesos are supported Spark clusters and HDFS, Cassandra, and Amazon S3 are some of the supported Spark distributed storage systems. Spark has an application Programming Interface that consists of several parallel collections facilitating the use of the functional programming language Scala. Some of the main features that enhance the performance of Spark and make it a perfect framework for Big Data applications are:

- Fast processing speed as it executes a batch of jobs upto100times faster than MapReduce due to reduced number of the read/writes from disc.
- Scalability upto 8000 nodes[23]
- In memory caching of the datasets for interactive data analysis
- Spark Streaming feature provides real time method with higher level library for stream processing
- Supports Structured query processing through Spark SQL

C. HPCC

High performance Computing Cluster framework is a massive parallel-processing computing platform and it is open source also. It is implemented on the commodity computing clusters which provide higher performance for solving Big Data problems. It has two different processing clusters. The *Thor Processing Cluster* is a data refinery that processes large volumes of heterogeneous data. It is responsible for extracting, transforming and loading processed raw data. It also supports high performance structured queries and data warehouse applications by creating keyed data and indexes. The Thor cluster is much similar to the Hadoop MapReduce platform in its environment and file system. The *Roxie Processing Cluster* is a parallel data processing system that works as a rapid data delivery engine.

It uses a distributed indexed file system for providing parallel processing of queries in an optimized file system environment. It supports thousands of simultaneous queries and users of online applications. It competes with the Hadoop's HBase in producing near to real time predictable query latencies.Both the clusters of HPCC use ECL programming language for its applications.

V. CLOUD COMPUTING PLATFORMS

A. AWS (Amazon Web Services)

AWS is a cloud services platform which is a collection of remote computing services. It offers web services like compute power, storage space, content delivery, and other functionalities. AWS is used by a number of organizations for deploying their services and applications in a cost effective manner. AWS cloud platform is flexible, Scalable and Reliable. It is a self service platform that operates from 11 geographical regions. The most well known services areAmazonEC2and Amazon S3[24].

Amazon Elastic Compute Cloud (EC2) is a Dashboard web service to launch a virtual server ,Known as an Amazon EC2 instance, instances are called as Amazon Machine Images(AMI) and to resize compute capacity in the cloud. Through EC2 Dashboard the users can take computers(instances) on rent , virtually instead of physical machines on the cloud and run their applications. A user can create and terminate as many instances of virtual machines as needed. It gives the user a complete control of the computing resources. Amazon provides economic and time efficient resources to the cloud users.

Amazon Simple Storage Service (Amazon S3) is a web service which provides online storage service for storing and retrieving any amount of data at any place and at anytime. The storage space is known as Amazon S3 bucket. It provides highly scalable, reliable, fast, efficient and secure infrastructure. According to [25], Amazon S3 provides the object-oriented storage service for users. User can access the objects by SOAP. S3 provides 99.99999999 percent durability.

B.Microsoft Azure

Azure is a cloud computing platform that allows its users to store data and control the applications directly written onto the virtual machines on its next-gen data center. It provides both the platform and infrastructure services to the users. Azure has a set of integrated tools and managed services facilitating the rapid creation of mobile and web apps [26]. Azure has one of the broadest range of supporting operating Systems, programming languages, tools, devices and databases. Specific Software Development Kits are provided by Microsoft for languages like Java, Python, .Net, etc. Azure supports both relational and NoSQL databases.

Azure provides an application programming interface which allows the user to interact with the services provided on the cloud. Its specialized operating system consists of a fabric layer that manages computing and storage resources. Virtual services are provided by the Windows server 2008 that has a customized version of Hyper V. The Platform as a Service environment of Azure helps in creating scalable and reliable applications. Microsoft Azure is the most widespread cloud platform running worldwide across 19 regions. Machine learning and Stream Analytics services of Azure are making intelligent business worldwide.

C. Google Cloud Platform

Google Cloud Platform is Google's own infrastructure. It is a set of flexible cloud-based services that allow a user to create, store, compute and process from simple to complex applications [27]. It includes servers, physical networks and software's for higher efficiency. Google cloud consists of several components:

- Google App Engine
- Cloud Data Storage
- Compute Engine

- Big Data Query

Google Cloud was made available for users in 2013 with the added features of load balancing, extended support for operating systems, faster disk persistence and the live migration of Virtual Machines. Google provides faster network access because it offers a vast online storage. Google cloud allows the user to easily develop and deploy mobile apps, create online games, as it is flexible for both IOS and Android operating systems. It can handle large sized files including videos, high resolution pictures etc.

VI. CHALLENGES AND OPPORTUNITIES

As Big Data has offered enormous appreciation to the organizations with terabytes and petabytes of data, traditional infrastructures are not up to the challenge. The most elementary bottleneck for Big Data applications is to analyze the large volumes of data and extract beneficial information or knowledge for future actions. Other Challenges include searching, sharing, storage, transfer, analysis, visualization and information privacy. The large volumes of data and the level of details needed in accessing it; that too with a high speed is another hurdle. Data security is another big risk when Big Data carries credit/debit card data, personal and other sensitive information. A more flexible design is required as Big Data comes in all shapes, colors and sizes.

Despite of the flexibility and efficiency of the cloud, a number of Security challenges have been identified in cloud computing environments. Data loss and data breaching is the most common security issue related to the cloud. A malicious hacker can have access to a target's data and he/she could use side-channel timing to extract personal cryptographic keys which are being used by other virtual machines. Also the insecure interfaces and weak API's can expose a user's confidentiality, integrity, availability and liability. Denial of Service and Cloud Abuse are another issues related to security in cloud computing. To maintain scalability, the infrastructure, platforms, and applications are shared by the Cloud Service Providers to deliver their services, which in turn gives rise to another security threat of shared vulnerabilities.

Cloud computing has reached a maturity level that leads it into a productive phase thereby offering a number of opportunities in the various field of business and entrepreneurship. Big Data analytics has been the hottest sector in the Silicon Valley for the past five years. The Internet of Things has also become a buzz nowadays; and the most interesting thing about The Internet of Things is all the data, which is itself a part of Big Data.

VII. CONCLUSION

In recent years, Cloud Computing and Big Data have gained a lot of attention. In this paper, we studied about the concepts and features of the cloud and Big Data. The Cloud provides three types of services i.e. Infrastructure as a Service ,Software as a Service and Platform as a Service and. We also discussed number of leading Cloud platforms such as Google Cloud Platform, Microsoft Azure and Amazon Web Services. We took an insight into Big Data and the characteristics that comprise Big Data. A number of tools have been introduced in the market for managing and analyzing Big Data. We discussed a number of tools like Hadoop's HDFS and MapReduce, HPCC and Apache's Spark for dealing with Big Data and Big Data Analytics. The challenges and opportunities were identified highlighting the pros and cons of the Big Data and Cloud Computing.

REFERENCES

[1] Peter Mell, Timothy Grance, The NIST Definition of Cloud Computing, September 2011

[2] Qi Zhang, Lu Cheng, Raouf Boutaba, Cloud computing: state-of-the-art and research challenges, 20 April 2010

[3] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, Data Mining with Big Data, January 2014
[4] A. Rajaraman and J. Ullman, Mining of Massive Data Sets, Cambridge Univ. Press, 2011.
[5] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.
[6] "Twitter Blog, Dispatch from the Denver Debate," http://blog.twitter.com/2012/10/dispatch-from-denver-debate.html,Oct. 2012.
[7] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloud computing reference architecture: Cloud service management perspective.". Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.
[8] A, Katal, Wazid M, and Goudar R.H. "Big Data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[9] Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri, Security Issues Associated With Big Data In Cloud Computing, May 2014
[10] OlukunleA Iyanda, Big Data and Current Cloud Computing Issues and Challenges, June 2014
[11] Hirdesh Shivhare, Nishchol Mishra, Jitendra Agarwal, Sanjeev Sharma, Cloud Computing and Big Data, Nov 13-15
[12] Enabling cloud computing & storage for Big Data applications with on-demand, high-speed transport , Taking Big Data to the Cloud, http://www.cloud.asperasoft.com,
[13] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, *Big Data Processing in Cloud Computing Environments*, 2012 IEEE
[14] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The eucalyptus open-source cloud-computing system," in *Cluster Computing and the Grid, 2009. CCGRID'09. 9th IEEE/ACM International Symposium on*. IEEE, 2009, pp. 124–131.
[15] Alba Amato, Beniamino Di Martino, Salvatore Venticinque, Big Data Processing for Pervasive Environment in Cloud Computing, 2014, IEEE
[16] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.
[17] Apache Hadoop Documentation, http://hadoop.apache.org/
[18] http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
[19] Sasikala, P. "Research challenges and potential green technological applications in cloud computing." International Journal of Cloud Computing, 2(1), pp. 1-19, 2013.
[20] Armbrust, Michael, et al. "A view of cloud computing." Communications of the ACM, 53(4), pp. 50-58, 2010.
[21] Deepak Puthal, B. P. S. Sahoo, Sambit Mishra, and Satyabrata Swain, Cloud Computing Features, Issues and Challenges: A Big Picture, 2015 IEEE
[22] Douglas and Laney, "The importance of 'Big Data': A definition,"2008.
[23] "Apache Spark FAQ". *apache.org*. Apache Software Foundation. Retrieved5 December 2014.
[24] https://d36cz9buwru1tt.cloudfront.net/AWS_Overview.pdf
[25] - Amazon S3 - Two Trillion Objects, 1.1 Million Requests / Second
[26] http://azure.microsoft.com/en-in/overview/what-is-azure/
[27] "Google Cloud Platform". cloud.google.com. Retrieved 2014-04-05
[28] Figure 1(image source: https://en.wikipedia.org/wiki/Cloud_computing#/media/File:Cloud_computing.svg)
[29] Figure 2(image source: http://bigdata.iexpertify.com)
[30] Figure 3 (img src: http://www.glennklockwood.com/data-intensive/hadoop/overview.html)