Proceedings of
National Conference on Innovative Trends in Computer Science Engineering (ITCSE-2015)
held at BRCMCET , Bahal on 4th April 2015

Research Analysis of Big Data Issue & Cloud Computing

[1]Nishi Midha , [2]Dr. Rajender Bathla,
[1]M.Tech Scholar , [2]Assistant Professor
[1,2]HCTM College Kaithal
[1]midha.nishi9@gmail.com , [2]dr.bathla@gmail.com

Abstract-- Recently, Big Data has attracted a lot of attention from academia, industry as well as government. It is a very challenging research area. Big Data is term defining collection of large and complex data sets that are difficult to process using conventional data processing tools. Every day, we create trillions of data all over the world. These data is coming from social networking sites, scientific experiments, mobile conversations, sensor networks and various other sources. We require new tools and techniques to organize, manage, store, process and analyze Big Data. This paper systematically presents various dynamic research issues related to Big Data analytics and Cloud.

Keywords:- **Big Data, MapReduce, Hadoop, Analyticscloud computing, cloud-computing.**

1.INTRODUCTION

What is Big Data? Many Researchers and organizations have tried to define Big Data in different ways. Gartner defines Big Data are high-volume, high-velocity and high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization [1].Big Data is defined as the representation of the progress of the human cognitive processes, usually includes data sets with sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time [2].

Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools [3].Scientists break down Big Data into many dimensions: Volume, Velocity, Variety, Veracity and Value [4, 5]

1. Volume – The amount of data is at very large scale. The amount of information being collected is so huge that modern database management tools are unable to handle it and therefore become obsolete.
2. Velocity – We are producing data at an exponential rate. It is growing continuously in terabytes and petabytes.
3. Variety – We are creating data in all forms -unstructured, semi structured and structured data. This data is heterogeneous is nature. Most of our existing tools work over homogenous data, now we require new tools and techniques which can handle such a large scale heterogeneous data.
4. Veracity-The data we are generating is uncertain in nature. It is hard to know which information is accurate and which is out of date.
5. Value-The data we are working with is valuable for society or not. IBM estimates that every day 2.5 quintillion bytes of data are created ,out of which 90% of the data in the world today has been created in the last two years .This data comes from sensors used to gather climate information, posts to social media sites, digital pictures and videos uploaded on internet, purchase transaction records, and cell phone conversation.

Fig: 1 Type of Big Data

All this data is Big Data [6]. The International Data Corporation (IDC) study predicts that the world will generate 50 times the amount of information and 75 times the number of information containers by 2020 while IT personnel to manage it will grow less than 1.5 times. The unstructured information such as files, email and video will constitute 90% of all data created over the next decade [7]. The 2011 digital universe study: extracting values from chaos says that the digital universe is 1.8 trillion gigabytes in size and stored in 500 quadrillion files and its size gets more than double in every two years time frame. If we compare the digital universe with our physical universe then it's nearly as many bits of information in the digital universe as stars in our physical universe [8]. According to Intel, 90% of the data today was created in the last two years, and the growth continues. It is estimated that the amount of data generated until 2012 is 2.7 zettabytes and it is expected to grow 3 times larger than that until 2015 [9]. According to cnet, the initial presidential debate between U.S. president barack obama and former governor mitt Romney on october 4, 2012, generated more than 10 million tweets, making it the most tweeted political event in U.S. history. website, usage across the world. It is estimated that 60 hours of video is uploaded every minute on YouTube and over 4 billion YouTube videos are viewed everyday [12]. Another example is Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 .All these examples show that enormous amount of data is generated everyday on internet by users only. All this data is Big Data and it is exploding at exponential rate.

In order to tackle the Big Data challenges, many governments, organizations and academic institutions come forward to take initiates in this direction. Recently, the US government announced a Big Data Research and Development initiative , to develop and improve the tools and techniques needed to access, organize, and analyze Big Data and to use Big Data for scientific discovery,

environmental and biomedical research, education, and national security [13] .Such a federal initiative has resulted in a number of winning projects to investigate the foundations for Big Data management (led by theUniversity of Washington), analytical approaches for genomics based massive data computation (led by BrownUniversity), large scale machine learning techniques for high-dimensional datasets which may be as large as 500,000 dimensions (led by Carnegie Mellon University), social analytics for large-scale scientific literatures (led by Rutgers University), and several others. These projects seek to develop methods, algorithms, frameworks, and research infrastructures which allow us to bring the massive amounts of data down to a human manageable and interpretable scale. Other countries such as the National Natural Science Foundation of China (NSFC) are also catching up with national
grants on Big Data research [14].

II. BIG DATA ANALYTICS TOOLS

There are varieties of applications and tools developed by various organizations to process and analyze Big Data. The Big Data analysis applications support parallelism with the help of computing clusters. These computing clusters are collection of hardware connected by ethernet cables. The following are major applications in the area of Big Data analytics.

A)MapReduce

MapReduce is a programming model for computations on massive amounts of data and an execution framework for largescale data processing on clusters of commodity servers. It was originally developed by Google and built on well-known principles in parallel and distributed processing [15]. MapReduce program consists of two functions –Map function and Reduce function. MapReduce computation executes as follows 1. Each Map function is converted to key-value pairs based on input data. The input to map function is tuple or document. The way key-value pairs are produced from the input data is determined by the code written by the user for the Map function 2. The key-value pairs from each Map task are collected by a master controller and sorted by key. The keys are divided among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task. 3. The Reduce tasks work on one key at a time, and combine all the values associated with that key in some way. The manner of combination of values is determined by the code written by the user for the Reduce function. MapReduce has two major advantages: the MapReduce model hide details related to the data storage, distribution, replication, load balancing and so on. Furthermore, it is so simple that programmers only specify two functions, which are map function and reduce function, for performing the processing of the Big Data. MapReduce has received a lot of attentions in many fields, including data mining, information retrieval, image retrieval, machine learning, and pattern recognition.

B)Hadoop

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop was inspired by Google's MapReduce Programming paradigm [16]. Hadoop is a highly scalable compute and storage platform. But on the other hand, Hadoop is also time consuming and storage-consuming. The storage requirement of Hadoop is extraordinarily high because it can generate a large amount of intermediate data. To reduce the requirement on the storage capacity,Hadoop often compresses data before storing it. Hadoop takes a primary approach to a single big workload, mapping it into smaller workloads. These smaller workloads are then merged to obtain the end result. Hadoop handles this workload by assigning a large cluster of inexpensive nodes built with commodity hardware. Hadoop also has a distributed, cluster file system that scales to store massive amounts of data, which is typically required in these workloads. Hadoop has a variety of node types within each Hadoop cluster; these include DataNodes, NameNodes, and EdgeNodes. The explanations are as follows:

a)NameNode: The NameNode is the central location for information about the file system deployed in a Hadoop environment. An environment can have one or two NameNodes, configured to provide minimal redundancy between the NameNodes. The NameNode is contacted by clients of the Hadoop Distributed File System (HDFS) to locate information within the file system and provide updates for data they have added, moved, manipulated, or deleted.

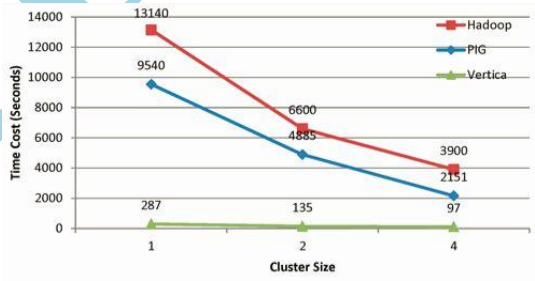

Figure 2. Hadoop Time Cost Efficiency

b)DataNode: DataNodes make up the majority of the servers contained in a Hadoop environment. Common Hadoop environments will have more than one DataNode, and oftentimes they will number in the hundreds based on capacity and performance needs. The DataNode serves two functions: It contains a portion of the data in the HDFS and it acts as a compute platform for running jobs, some of which will utilize the local data within the HDFS.

c)EdgeNode: The EdgeNode is the access point for the external applications, tools, and users that need to utilize the Hadoop environment. The EdgeNode sits between the Hadoop cluster and the corporate network to provide access control, policy enforcement, logging, and gateway services to the Hadoop environment. A typical Hadoop environment will have a minimum of one EdgeNode and more based on performance needs.

C)IBM InfoSphere BigInsights

It is an Apache Hadoop based solution to manage and analyze massive volumes of the structured and unstructured data. It is built on an open source Apache Hadoop with IBM

big Sheet and has a variety of performance, reliability, security and administrative features.

III.RESEARCH DONE IN BIG DATA ANALYTICS

Big Data analytics is a hot research area today. There are several research papers published to tackle Big Data problems efficiently. Sachchidanand Singh et al. explained the concept, characteristics & need of Big Data and different offerings available in the market to explore unstructured large data [17]. Changqing ji et al. discussed the scope of Big Data processing in cloud computing environment [18]. Dan Garlasu et al. proposed an architecture for managing and processing Big Data using grid technologies [19]. Tyson Condie et al. discussed the machine learning computational models for Big Data [20]. Xindong Wu et al. presented a HACE theorem that characterizes the features of the Big Data revolution and proposed a Big Data processing model from the data mining perspective [21]. Dr. Sun-Yuan Kung proposed cost-effective design on kernel-based machine learning and classification for Big Data learning applications [22]. Kapil Bakshi explored the approaches to analyze unstructured data like imagery, sensors, telemetry, video, documents, log files, and email data files [23].Xiaoyan Gu et al. investigated energy efficient architecture for Big Data application [24].Chansup Byun et al. brought together the Big Data and Big compute by combining Hadoop clusters and MPI clusters [25]. These are few developments in the area of Big Data. Big Data analytics is a new research area and there is lot of scope of research in this area. Preliminary Research has been started but still lot to be done in future.

IV.RESEARCH SCOPE IN BIG DATA ANALYTICS

Many researchers have suggested that commercial DBMSs are not suitable for processing extremely large scale data. They are suggesting new Big Data base management system which must be cost effective and scalable. The use of parallelization techniques and algorithms is the key to achieve better scalability and performance for processing Big Data. Big Data is a new challenge for academia and industry. Researchers are defining new theories, methods and technologies for Big Data management and analysis. Advancing Machine learning, data mining and statistical techniques for processing of Big Data are key to transforming Big Data into actionable knowledge. Current data base management systems are unable to store the increasing flood of Big Data. There is a need of hierarchical storage architecture to handle the challenge of storing the Big Data. Existing data processing algorithms are excellent at processing homogeneous and static data. But today data is continuously generating from various resources. This data is heterogeneous and dynamic in nature. New scalable data processing algorithms are required to process such data. While processing a query in Big Data, speed is major criteria .In such case, indexing is a optimal choice for complex query processing. Parallelization and divide and conquer are good algorithmic solutions to handle Big Data effectively. Organizations are reducing their cost by using online Big Data applications. This strategy is profitable to organizations but producing new security threats. Security of Big Data is prime concern for researchers as well as industry. Security in Big Data is mainly in the form of how to process data mining without exposing sensitive information of users. As Big Data is dynamic in nature hence producing new challenges for researcher to create algorithms to handle such situations. The main challenges of Big Data are data variety, volume, analytical workload complexity and agility. Many organizations are struggling to deal with the increasing volumes of data. In order to solve this problem, the organizations need to reduce the amount of data being stored and exploit new storage techniques which can further improve performance and storage utilization. The IT professionals and students looking to build a career and skills in Big Data & Apache Hadoop can take advantage of IBM's BigDataUniversity.com website where users can learn the basics of Hadoop, stream computing and open-source software development [26].

V BIG-DATA TECHNOLOGY: SENSE, AND ANALYZE.

The rising importance of big-data computing stems from advances in many different technologies:

A)Sensors: Digital data are being generated by many different sources, including digital imagers (telescopes, video cameras, MRI machines), chemical and biological sensors (microarrays, environmental monitors), and even the millions of individuals and organizations generating web pages.

B)Computer networks: Data from the many different sources can be collected into massive data sets via localized sensor networks, as well as the Internet.

C)Data storage: Advances in magnetic disk technology have dramatically decreased the cost of storing data. For example, a one-terabyte disk drive, holding one trillion bytes of data, costs around $100. As a reference, it is estimated that if all of the text in all of the books in the Library of Congress could be converted to digital form, it would add up to only around 20 terabytes.

D)Cluster computer systems: A new form of computer systems, consisting of thousands of "nodes," each having several processors and disks, connected by high-speed local-area networks, has become the chosen hardware configuration for data-intensive computing systems. These clusters provide both the storage capacity for large data sets, and the computing power to organize the data, to analyze it, and to respond to queries about the data from remote users. Compared with traditional high-performance computing (e.g., supercomputers), where the focus is on maximizing the raw computing power of a system, cluster computers are designed to maximize the reliability and efficiency with which they can manage and analyze very large data sets. The "trick" is in the software algorithms – cluster computer systems are composed of huge numbers of cheap commodity hardware parts, with scalability, reliability, and programmability achieved by new software paradigms.

E)Cloud computing facilities: The rise of large data centers and cluster computers has created a new business model, where businesses and individuals can *rent* storage and computing capacity, rather than making the large capital investments needed to construct and provision large-scale

computer installations. For example, Amazon Web Services (AWS) provides both network-accessible storage priced by the gigabyte-month and computing cycles priced by the CPU-hour. Just as few organizations operate their own power plants, we can foresee an era where data storage and computing become utilities that is ubiquitously available.

F)**Data analysis algorithms:**

The enormous volumes of data require automated or semi automated analysis – techniques to detect patterns, identify anomalies, and extract knowledge. Again, the "trick" is in the software algorithms - new forms of computation, combining statistical analysis, optimization, and artificial intelligence, are able to construct statistical models from large collections of data and to infer how the system should respond to new data. For example, Netflix uses machine learning in its recommendation system, predicting the interests of a customer by comparing her movie viewing history to a statistical model generated from the collective viewing habits of millions of other customers.

VI BIG DATA AND THE CLOUD

The term big data is derived from the fact that the datasets involved are so large that typical database systems are not able to store and analyze the datasets. The datasets are large because the data are unstructured data and are from many new sources, including e-mail, social media etc. The characteristics of big data present data storage and data analysis challenges to businesses. Analyzing big data is done using a programming paradigm called MapReduce. The MapReduce paradigm requires that huge amounts of data be analyzed. The mapping is done concurrently by each separate NAS device; the mapping requires parallel processing. The parallel processing needs of MapReduce are costly, and require the configuration noted previously for storage. The processing needs can be met by cloud-service providers. Big Data is a data analysis methodology enabled by recent advances in technologies and architecture which support high-velocity data capture, storage, and analysis. However, big data entails a huge commitment of hardware and processing resources, making adoption costs of big data technology prohibitive to small and medium sized businesses. Cloud computing offers the promise of big data implementation to small and medium sized businesses. Data sources extend beyond the traditional corporate database to include email, mobile device output, sensor-generated data, and social media output Big Data requires huge amounts of storage space. While the price of storage continues to decline, the resource needed to leverage big data still poses financial difficulties for small to medium sized businesses. A typical big data storage and analysis infrastructure will be based on clustered network-attached storage (NAS) [14]. Clustered NAS infrastructure requires configuration of several NAS "pods" with each NAS "pod" comprising of several storage devices connected to an NAS device [14]. The series of NAS devices are then interconnected to allow massive sharing and searching of data [14]. Data storage using cloud computing is a viable option for small to medium sized businesses considering the use of Big Data analytic techniques. Cloud computing is on-demand network access to computing resources which are often provided by an outside entity and require little management effort by the business. A number of architectures and deployment models exists for cloud computing, and these architectures and models can be used with other technologies and design approaches [15]. Owners of small to medium sized businesses who are unable to afford adoption of clustered NAS technology can consider a number of cloud computing models to meet their big data needs. Small to medium sized business owners need to consider the correct cloud computing in order to remain both competitive and profitable.

A)TYPES OF CLOUDS

In the cloud deployment model, networking, platform, storage, and software infrastructure are provided as services that scale up or down depending on the demand. The Cloud Computing model has three types of clouds model which are – the public cloud, the private cloud, and the hybrid cloud.

Fig: 3.Types of Cloud

a)Public cloud

A public cloud is the pay- as-you-go services available to the general public. In this configuration, a business does not own the core technology resources and services but outsource these to service providers. Public cloud is also considered to be an external cloud. Public cloud describes cloud computing in the traditional mainstream sense, whereby resources are dynamically provisioned on a fine-grained, self-service basis over the Internet, via web applications/web services, from an off-site third-party provider who shares resources and bills on a fine-grained utility computing basis. It is typically based on a pay-per-use model, similar to a prepaid electricity metering system, whose flexibility caters for spikes in demand for cloud optimization. Public clouds are less secure than the other cloud models because it places an additional burden of ensuring all applications and data accessed on the public cloud are not subjected to malicious attacks.

b)Private cloud

A private cloud is internal data center of a business that is not available to the general public but uses cloud structure. In this configuration, resources and services are owned by the business, with the services accessible within the business through the intranet and since the technology is owned and operated by the business, this type of cloud is more expensive than a public cloud. It is also more secure and because of its specified internal exposure, only the organization and designated stakeholders may have access to operate on a specific Private cloud. A private cloud is an internal cloud residing inside the company's firewall and managed by the company.

c)Hybrid cloud

Hybrid cloud is a combination of both public and private cloud, when a company uses a hybrid cloud; it uses a public cloud for some tasks and a private cloud for other tasks. In this model, a company uses the public cloud to expedite extra tasks that cannot be easily run in the company's data center or on its private cloud [1]. A hybrid cloud allows a company to maintain critical, confidential data and information within it firewall while leveraging the public cloud for non-confidential data. The private cloud portion of the hybrid cloud is accessed by company employees, both in the company and on the go, and is maintained by the internal technology group. The private cloud part of the hybrid cloud is also accessed by the company employees but is maintained by external service providers. Each portion of the hybrid cloud can connect to the other portion.

d)Community cloud

Community cloud is a private cloud that is shared by several customers with similar security concerns and the same data and applications sensitivity.

VII. CLOUD COMPUTING CHALLENGES

Cloud computing is associated with numerous challenges and the major challenges that prevent Cloud Computing from being adopted are as follows:

A)**Security**

Security issue plays the most important role in hindering Cloud computing acceptance. Security issues such as data loss, phishing, botnet poses serious threats to organization's data and software. For example, hackers can use Cloud to organize botnet as Cloud often provides more reliable infrastructure services at a relatively cheaper price for them to start an attack.

B)Costing Model

Cloud consumers must consider the tradeoffs amongst computation, communication, and integration. While migrating to the Cloud can significantly reduce the infrastructure cost, it does raise the cost of data communication, i.e. the cost of transferring an organization's data to and from the public and community Cloud and the cost per unit of computing resource used is likely to be higher. This problem is particularly prominent if the consumer uses the hybrid cloud deployment model where the organization's data is distributed amongst a number of public/private (in-house IT infrastructure)/community clouds.

C)**Charging Model**

The elastic resource pool has made the cost analysis a lot more complicated than regular data centers, which often calculates their cost based on consumptions of static computing. Moreover, an instantiated virtual machine has become the unit of cost analysis rather than the underlying physical server. For SaaS cloud providers, the cost of developing multitenancy within their offering can be very substantial. These include: re-design and redevelopment of the software that was originally used for single-tenancy, cost of providing new features that allow for intensive customization, performance and security enhancement for concurrent user access, and dealing with complexities induced by the above changes.

D)**Service Level Agreement (SLA)**

Since cloud consumers do not have control over the underlying computing resources, they do need to ensure the quality, availability, reliability, and performance of these resources when consumers have migrated their core business functions onto their entrusted cloud. It is therefore vital for consumers to obtain guarantees from providers on service delivery. Typically, these are provided through Service Level Agreements (SLAs) negotiated between the providers and consumers. The very first issue is the definition of SLA specifications in such a way that has an appropriate level of granularity, namely the tradeoffs between expressiveness and complicatedness, so that they can cover most of the consumer expectations and is relatively simple to be weighted, verified, evaluated, and enforced by the resource allocation mechanism on the cloud.

VIII.BENEFITS OF CLOUD COMPUTING

There are lots of benefits in using Cloud computing to render or access computing resources. Presently a lot of people use Cloud computing without even knowing what it means. For example, Gmail, Yahoo mail, YouTube, and Skype users…are all in the Cloud. Increasingly companies and organizations are becoming aware of the huge benefits that Cloud computing provides. Some of these benefits include:

A)Flexibility and storage: With Cloud computing Files are stored in the "Cloud". This allows for development in the organization because workers no longer have to worry about the storage of documents. Also, workers can access office files from wherever and whenever. Workers can also work together virtually even when they are not at the same place at the same time. Various documents can be viewed simultaneously provided Internet connection is available.

B)**Time saving:** Alongside easy collaboration, Cloud computing also aids the easy access to information. Easy access in this context could be seen in how fast it is to access Gmail, Yahoo mail, mailboxes in general. It is fast and easy in contrast to the time it would take to download and install software.

C)**Reduced Cost:** Cloud computing puts a stop to the illegal reproduction and distribution of software. Some software on the Cloud is free. For example, most SaaS solutions have a pay-as-you-go pricing model instead of a large up-front investment. Such pricing models allow end users to pay only for what they use thus freeing up resources such as time and money for other more important (core) business activities. Cloud computing is therefore cheaper and less labor intensive for companies. There is no need to buy and install expensive software. There is no need to acquire, track and manage software licenses.

IX CONCLUSION

The paper is a systematic study of various issues of Big Data analytics. Big Data is a very challenging research area. Data is too big to process using conventional tool of data processing. Academia and industry has to work together to design and develop new tools and technologies which effectively handle the processing of Big Data. Big Data is an emerging trend and there is immediate need of new machine learning and data mining techniques to analyze massive

amount of data in near future. Despite the benefits enumerated, it is surprising that not many companies and organizations are rushing to leverage the advantages of Cloud computing, especially in developing countries because the benefits of cloud computing are tempered by two major concerns – security and loss of control. Although Big data and Cloud computing is a new phenomenon which is set to revolutionize caution must be exercised in the way we use the Internet. There are many new technologies emerging at a rapid rate, each with improvements in making living much easier for users. However, there is a need for a cost-performance trade off while deliberating on what type of cloud service to adopt. If the data being processed is considered mission critical to the company, the more expensive private cloud, implemented in-house, would provide a more secured environment with the company keeping the mission critical data in-house.

REFERENCES

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G…Zaharia, M. (2010, April). A view of cloud computing. Communications of the ACM, 53(4), 50-58. DOI: 10.1145/1721654.1721672.

[2] Rouse, M. (2010b, August). Infrastructure as a Service. Retrieved from http://searchcloudcomputing.techtarget.com/definition /Infrastructure-as-a -Service-IaaS

[3] Cisco. (2009). Infrastructure as a Service: Accelerating time to profitable new revenue streams. Retrieved from http://www.cisco.com/en/US/solutions/collateral/ns341/ns991/ns995 /IaaS_BDM_WP.pdf

[4] Salesforce.com. (2012). The end of software: Building and running applications in the cloud. Retrieved from http://www.salesforce.com/paas/

[5] Cole, B. (2012). Looking at business size, budget when choosing between SaaS and hosted ERP. E-guide: Evaluating SaaS vs. on premise for ERP systems. Retrieved from http://docs.media.bitpipe.com/io_10x/io_104515/item_548729/SAP_sManERP_IO%23104515_EGuide_061212.pdf

[6] Rouse, M. (2010a, August). Software as a service. Retrieved from http://searchcloudcomputing .techtarget.com/definition/Software-as-a-Service

[7] Rouse, M. (2007, December). Hardware as a service. Retrieved from http://searchitchannel .techtarget.com/definition/Hardware-as-a-Service-in-managed-services

[8] IOS Press. (2011). Guidelines on security and privacy in public cloud computing. Journal of EGovernance, 34 149-151. DOI: 10.3233/GOV-2011-0271

[9] Chansup Byun, William Arcand, David Bestor, Bill Bergeron, Matthew Hubbell, Jeremy Kepner, Andrew McCabe,Peter Michaleas, Julie Mullen, David O'Gwynn, Andrew Prout, Albert Reuther, Antonio Rosa, Charles Yee," Driving Big Data With Big Compute"

[10] "Big Data: science in the petabyte era," *Nature 455 (7209):1*, 2008

[11] Douglas and Laney, "The importance of 'Big Data': A definition" ,2008

[12] http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/Federation.html

[13] http://int3.de/res/GfsMapReduce/GfsAndMapReduce.pdf

[14] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011

[15] The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. *National Public Radio*, Nov. 30, 2011. http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data.