I J C T A, 9(17) 2016, pp. 8437-8444

Research Issues and Challenges of Big Data

K. Radha[*] and B. Thirumala Rao[*]

ABSTRACT

Many of the Organizations are generating massive volumes of data ranging from Terabytes to Petabyte through different sources such as Social media sites (Face book, Flipkart, Quickr, etc). To handle such a large amount of data Organizations are using analytical tools with respect to the interaction between cloud and big data. This paper presents big data issues and research directions towards the recent research work of processing of big data in cloud computing.

Keywords: Cloud Computing, Big Data, Heterogeneity, Security, Data Accuracy, Internet of Things

I. INTRODUCTION

Big Data keeps on growing, both in volume and complexity. Unstructured data are generated from non-traditional sources [1]. An enterprise faces the challenges and opportunities for storing and analyzing Big Data. Big Data Characteristics are:

Volume: Volume refers to that, massive amount of data is generated for every second in terms of Zettabyte or Brontobyte. It refers to the amount of data that is measured in terabytes today, but will likely to be measured in Petabytes, Exabytes, and even Zetta bytes in the future. Volume is one of the research challenges in the subsequent sections. Volume is an immediate challenge of big data it requires scalable data storage. Massive data will be Discussed in the section 3 .Big Data issues and challenges.

Velocity: Velocity refers to that, speed of the newly generated data such as social media messages. In-memory analytics analyzed the data at the time of the generation before storing that data into databases. It includes frequency of data generated, including batch, real-time, or streams. Decision Making should be fast in a timely manner to handle the unexpected changes in the enterprise environment for the successful operational execution. Big Data analytics are to transform magnitude quantities of raw data into metadata for the analysis purpose. Velocity is a significant challenge to analyze and process magnitude of the data.

Variety: It Refers to that,sources of data and can be classified as structured, unstructured, or semi-structured data web pages, web logs, emails, etc. Health data consists of pictures, test reports, medical repositories, and doctor's prescriptions. **Veracity**: Data quality is required to find the efficient decision making. Each ' V' varies even with the same datasets.Volume, variety ,velocity includes data web logs, RFID data, social data,web data, search data,video,e-commerce.Big Data uses the technologies such as massively parallel processing databases(MPP), distributed databases, File system, Cloud computing platform, etc. Its tools are NoSQL, Hadoop and CouchDB. Big Data areas are Financial Services, Health Care, Retail and Consumer Products [1].

II. STATISTICAL REPORTS ON MOBILE APPSTORE AND DIGITAL UNIVERSE DATA GROWTH

The mobile app store revenue growth is expected to grow enormously as depicted in the Fig 1. In the year 2011, the number of Paid-for Downloads is 7,139 and In-app Purchases Downloads is 7124. In the year

[*] Department of Computer Science and Engineering, KL University, Guntur, Andhra Pradesh, India, *E-mail: radha.klu13@gmail.com; thirumail@yahoo.com*

2012, the number of Paid-for Downloads is 15,375$millions and In-app Purchases Downloads is 2111 $millions.

The rest of the paper is organized as follows. Proposed embedding and extraction algorithms are explained in section II. Experimental results are presented in section III. Concluding remarks are given in section IV.

In the year 2013, number of Paid-for Downloads is 20240 $millions, In-app Purchases Downloads are 4591 $millions. Hence, for every year Number of Downloads is increasing in huge Volume and Velocity. From 2011-2016, In- App purchase Downloads are increased range from 7124$millions to 23,771 $millions downloads.

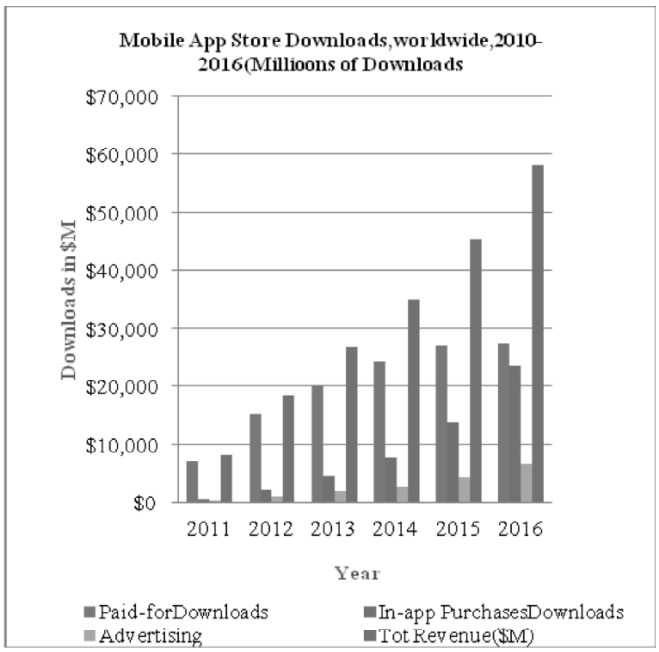

Figure 1: Mobile App Store Revenue Worldwide 2011-2017($M)

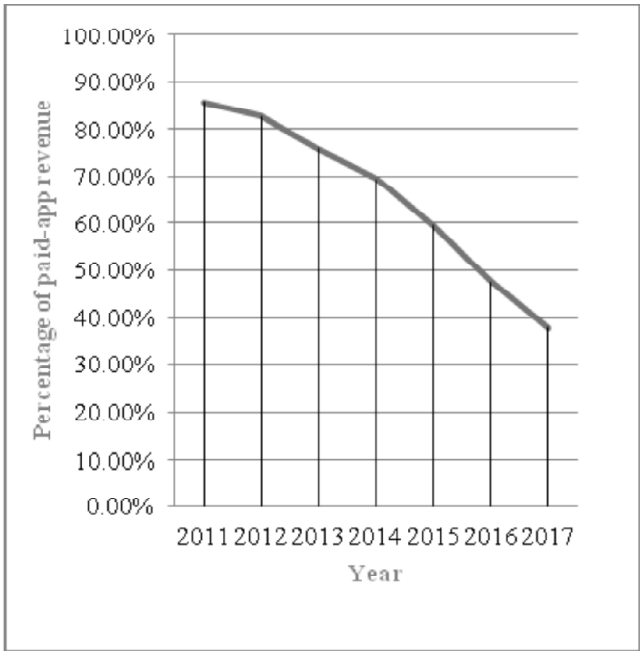

Figure 2: Percentage of Paid app revenues worldwide from 2011-2017

From 2011 to 2016, Paid-for Downloads are increasing in huge volume and velocity compared to In-App Purchase Downloads. The Percentage of Paid-for Downloads is more when compared to In-App Purchases downloads as shown in Fig. 1 . In-app purchase will be half of the total revenue in the year 2017. In the year 2012, 6.65 billion paid apps are downloaded via mobile devices. In the year 2016 paid for downloads will meet 13.49 billion. Most of the mobile apps are free to download some app developers produce revenue via paid for downloads which contains premium apps with in-app purchases as well as mobile advertising. For the paid-for downloads revenue will go to the distributor and remaining will goes to the developer. Because of rapid growth of mobile devices are increasing. App stores will receive the payment from app developers via sign-up account payments based on number of app sales. Most of the apps can be freely downloaded but the paid versions are unlocked for the extra features and also lock up the advertisements. Google play store for Android applications are having present features are 67% free apps and 33 % paid apps[36]. From 2011-2017, percentage of paid-for app revenues depicted in the Fig 2. In the year 2013, 75.9% of app revenues were produced via paid downloads. In the year 2017, Gartner estimated that paid- for app revenues will decrease to 37.8 %. This will occur in the growth of in-app purchase revenues[35]. Mobile network operators will face an explosion in the mobile data and voice traffic. Fig 3. The estimates for the total traffic growth from 2011-2016 .This is a major technical, operational, and financial challenge for mobile telecom operators all over the world. Several factors contribute to the explosion of wireless data traffic. In 2011, there were 6 billion mobile subscriptions; this number is expected to

double by 2020. The dynamic growth of data will upgrading the wireless media adopts the Fourth generation networks. Wire line networks will also face an increase in the amount of data they will carry. According to GreenTouchTM estimates, fixed Internet IP traffic in 2020 will be eight times the level in 2011. However, an increase in IP traffic is less likely to impact wire line networks than the anticipated impact on wireless networks[36].

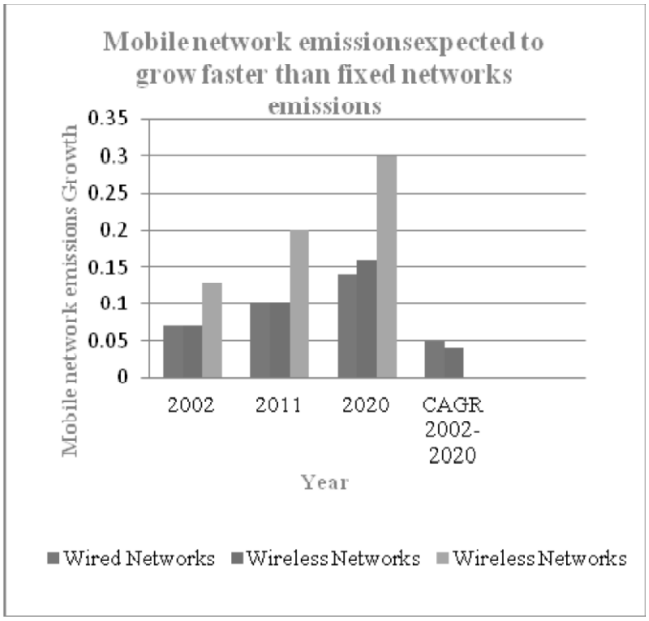

Figure 3: Mobile network emissions Growth from 2002-2020

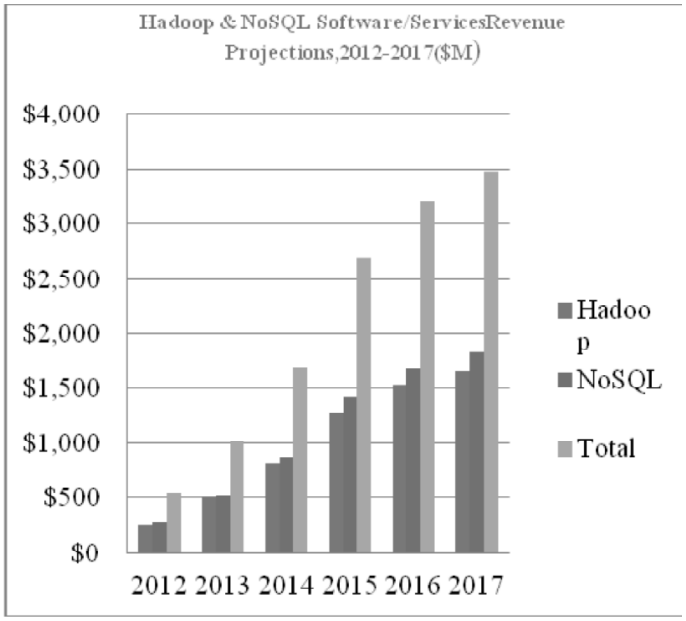

Figure 4: Hadoop & NoSQL Software/Services Revenue Projections, 2012-2017($M)

Revenue processions' of Hadoop and NoSQL services till 2017 are increasing gradually as depicted in the Fig.4. NoSQL suppliers are delivering commercial versions of open source database. Hadoop and NoSQL is using at Bank of America between the financial service industries. NoSQL market is growing massively. NoSQL massive scale transactional workloads such as performance, privacy and scalability. In 2017 the Hadoop market is forecasted that $13.95 Billion [35].

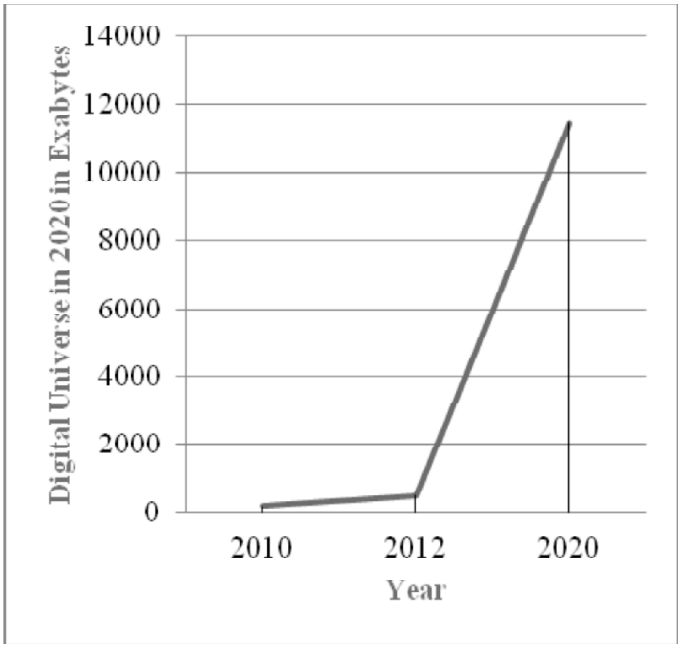

Figure 5: Digital Universe by 2020 (in Exabytes)

In a Digital Universe generated 171 Exabyte of data. In 2012, it is generated 491 Exabyte of data. By 2020 The Digital Universe Data will be 11,453 Exabyte as shown in the Fig 5. Hence, for every year, volume and velocity increases if the data growth is increasing rapidly [37].

III. BIG DATA ISSUES AND CHALLENGES

3.1. Data Quality and Data Accuracy

Accurate data is initial requirement for the efficient information systems. Decision making is not possible with the inaccurate data. Many of the enterprises are not having the knowledge on inaccurate data in their systems. Organizations are failing in improving the efficiency of data. To correct the problems of inaccurate data, organizations monitor the data collection through statistics. Data can be used in organizations, Government sectors, Educational Institutions, etc. Organizations are using more data to make important decisions. If the complexity of information systems increases inaccuracy of data also increases. Many of the organizations are suffering tremendously with inaccurate data in the company databases so that their data quality is managed strictly. The problems are as follows, systems are having low acceptance testing and less data creation process etc.Poor Quality data is due to transaction rework costs, latencies in providing data to the decision-makers, losing customers through poor service. To improve the data quality organizations must make the operational and business changes [28]. This provides the flexibility to the organization for more paybacks overtime. Data quality assurance Technology elements are availability of experts, methodologies and software tools. This allows the IT industry to use the emerging technology to transfer the knowledge between individuals and organizations . Software tools to manage the data quality are metadata repositories. Data profiling is an analytical tool .It develops the architecture and data quality. Data profiling is used for assessing the data quality. Data profiling contains two procedures to examine the data,namely First, discovery, second, assertive testing . Data supervising tool can be used as an enterprise-oriented and database-oriented. Data refining tools are used to examine the data for finding the errors and to fix the problem. In-Database Analytics are used for efficient decision-making.

3.2. Speed of Data and Massive Data

IDC forecasted that in the year 2020, 40% of the information in the digital world will meet the cloud computing, Yet, only as much as about 15% of data will be maintained in a cloud [9]. In the year 2020 digital world data growth will be 130 Exabyte to 40,000 Exabyte [15], [16]. By 2020, 40 Zettabyte of data will be generated. Organizations can't make the decisions efficiently with massively generated data and also finding and analyzing of massive data is a tedious task. If the Speed of data is increased, Data accuracy, Data quality can't be maintained properly. Big Data analysis enables to get the relevant data in less time for making decisions [11]-[13]. Data accuracy [6], speed of data and data quality are suffering from the problem of Decision-making with magnitudes of data and also unable to find, analyze, retrieve, model and process the massive data. As Shown in Fig1.Market is increasing in adoption of the mobile devices and inbuilt apps and functionalities are major in continuous growth of unstructured data. As shown in the Fig 1. Volume and Velocity of the Paid –for downloads is more compared to in-App Purchase downloads. In the year 2011 paid-for downloads are 7139 $millions, In-App purchase downloads are 7124 $millions, such that from 2011 to 2016 year, In-App purchase Downloads is 712$millions to 23,771$millions and Paid-for Downloads is 7139$millions to 27664$millions downloads $millions. Hence Volume and velocity of the Downloads will be high compared to In-app purchase downloads with paid-for downloads. Developers do not trust the information which is extracted from big data. Magnitudes of data can't be put on a graph for analysis. Social networking sites are responsible for 12 Terabytes data daily. There are 30 billion pieces of content is sharing on Facebook every day. By 2020 one third of all data will be stored in the cloud and users will create 40 Zetta bytes of data [25]. How to compute and extract massive quantities and how to access scientific resources and making an efficient use and it is tackled in various dimensions such as the creation

of data, access, interoperability, etc.W-wide ,scientific data is increasing rapidly and interoperability is very tedious to achieve. Organizational data are collected rapidly, while collecting the data, delay of minute causes a variation in analysis of output. By 2020 IT departments will be looking after 10×more servers, 50×more data, 75× more files [7].

3.3. Data Discovery, Diversity of Data

Web sites are generating Terabytes of data daily in various formats such as structured and unstructured data. Unstructured data cannot be indexed in relational table for analysis and querying. Unstructured data can't be analyzed. Unstructured data is unable to process at large scale [14], [17], [18], [19]. For the newly developed information users will not have the prior intimation [6], [8]. Search results are a problem, because both structured and unstructured data come from various sources. Understanding the search patterns is difficult. Gartner predicted that more than 80% of data is in an unstructured format, generated unstructured data will be double for every 3 months. By 2015 Digital universe data growth will be 8 Zettabyte and it will doubles for very two years [20].Unstructured data are not organized which includes bitmap images, text and other types and unstructured data is unable to locate. There are various reasons are there to achieve the integration. In various situations, there are different on-going efforts are integrating the structured data with other structured data to combine the unstructured data with other unstructured data.

3.4. Distributed Storage

Conventional storage approaches are expensive, inefficient, which results in inflexible storage infrastructure and difficult to manage at large scale data. Due to a variety of data, decision–making is not possible. Latency and data size are the problems in distributed storage [24], [25]. The Big data storage platform delivers cost-effectively and organizations data storage demand is growing rapidly. Today40-60% of additional data is generated each year, predicting a maximum amount of storage is no longer possible. If range of data is increased mean time between failures falls (MTBF) [1]. Present trends of database management systems are unable to fulfill the necessities of big data, increasing velocity of storage capability is lower than the data. Hierarchical storage architecture needs to be designed. Earlier algorithms are unable to store the data from real world, because of diversity of big data. Re-organizing the data in big data management is a challenge. Virtual server technology creates this problem due to less communication between Application, Server and Storage administrators.

3.5. Performance

If the complication of organizational systems is growing,the need for supervising and interpretation of systems also raise . Performance and capacity of big data depend on Volume, Variety, and Velocity [30]. Access latencies, lack of speed in accessing the data while in-memory, latency of network, and time to access the storage devices have performance problems. Big Data analytics provides new problems or issues for the performance of applications to monitor the systems. Performance and storage capacity challenges are derived from the Big Data system features such as Volume, Variety, Velocity, Veracity. Data can be accessed from memory at a rapid speed, storage devices will produce capacity problems. Movement of Petabytes of data over the network in one data center to another datacenter and to many data centers require highest bandwidth and less- network of latency infrastructure for good communication between the compute nodes. Hadoop Distributed File System writes data for thrice for replication to avoid the loss of data. It was designed for just a bunch of disks (JBOD) not for the organizational systems. Organizational disk systems are avoided due to heavy data traffic. Replication creates load on the memory devices and I/O devices .This causes heavy load on the network while systems are trying to control the data traffic.

3.6. Security

Security and privacy issues are amplified by volume, variety, velocity, veracity in massive large clusters of cloud infrastructures, multifariousness of data sources and diversity of data, data streams accomplishment [6]. Every day 2.5 quintillion bytes of content or information is produced. Today 90% of the data in the world has been created [3], [22], [26], [30] from sensors, social media, digital images and video etc**.** Input validation and validation of the input data source difficult for filtering. Anonymous data is allowed into big data that must be authenticated and should not distributed to unauthenticated recipients [23], [25], [2]. Through the online big data applications, many industries are reducing their IT budget. Even though, security and privacy are the problems for the storage of big data and processing of big data. Magnitude use of anonymous services infrastructures produces tedious operations. The range of data and applications are growing exponenentially and brings issues of dynamic data supervising and protecting the data.

Privacy is based on an immobile collection of data while data always changing progressively includes patterns of data, change in addition to new data. Hence it is a problem to implement protecting the data in this complicated situation and legitimate and governing issues require some awareness [32]. Users are accessing the information from mobiles devices, PCs make the efficient decisions on the real-time data to store the data and analyze the data for intelligence. Massive devices and users generating the data to create unbelievable volume, velocity, variety.An Enterprise needs to protect their data and utilizes the real-time insight from big data. In previous days, data controlling is easier. If the secrets are kept in the company network to make sure that data is kept safely a strong firewall should be placed. To meet the present security challenges faced by enterprises a new model should be developed. Businesses need the ability to secure the data. Monitor the data that should not be accessed by third-party. Utilize the intelligence to stop the attacks before the data are exfiltrate. Collect the information from the cloud, virtual, network devices, servers, database ,desktops. Data is theft in the following ways today. Firstly, Hackers find the way to enter into the network. Second, they install an agent to gather the information after discovering the information; third, they will exfiltrate the information out of the network. With these five ways information will be hacked from the organizations. Security Management in the big data is categorized into 3 ways, such as Analytics and Visualization, threat Intelligence and Scale-out Infrastructure. Scale–out infrastructure is responsible for changing the IT environment threats.

3.7. Scalability

The size of the big data is the challenge. Redundant Array of Independent Disks approach does not provide the performance and durableness deals with magnitude of the data. Storing the data from memory to disk causes difficulty, this causes the processing delay. There are some approaches to the problem of capacity and performance, such as Usage of copyrighted networks, construct the networks based on traffic, Apply data locality through ETL (Extract, Transform, Load) and Analyze the data where it is stored. Performance and capacity challenges are affected by the variety of data types such as structured, semi-structured and unstructured. Structured data have great automation value. Structured data can be efficiently stored, organized and searched. Data is growing daily, with this growth of data there are many problems. New data is added to old data,maintain the both old and new data without deleting anything. Older data also archived for legally. Structured data can be stored and retrieved easily than the unstructured data. Adding structure to extremely unstructured data and heterogeneous data improves the total process of storing retrieval process efficiently. Hence Scalability is achieved by making the unstructured data to structure. Use Human-Computer Interaction patterns (HCI) to achieve the Scalability [6].

3.8. Heterogeneity

Unprecedented data are generated by the social media and internet of things. Human beings are regularly connecting to the internet than before. Growth of Data trends in social networking sites is

increasing rapidly in human lives. The problem with sense-making is that making sense of the observed data [31].

IV. SUMMARY

As the Value of Data is increasing rapidly, Cloud services are able to store and analyze the organizational data to face the challenges with the big data. Big Data enables enterprises to achieve differentiation by reducing cost, through the efficient plan it can improve the efficiency. Big Data challenges and issues are summarized in the Table 1. Organizations are forecasting and understanding the customer behavior .Big data analytics enable to store and query massive data sets. Global data will reach 40 Zettabyte by 2020[31], [34]. The "things" of the real world will coherently interconnect into the virtual world by permitting to connect anywhere and anytime.

Table 1
Summary of Issues and Challenges of Big Data

SN	*Challenges*	*Technologies*	*HDFS*	*Big Table*	*Map Reduce*	*HBase*
1	Accuracy of data	Predictive analytics	Y			
2	Massive data	Master data management				
3	Speed of data	Saas based Business analytics	Y	Y		Y
4	Diversity of data	Data federation			Y	
6	Data quality	Predictive analytics (statistical process control-based models)		Y		
7	Data discovery	Predictive analytics	Y		Y	
8	performance	Graph analytics (optimal path analysis)	Y			
9	Distributed storage	Big Data analytics		Y	Y	
10	Context validation	Graph analytics	Y	Y		

REFERENCES

[1] "Big Data is the Future of Healthcare, Cognizant 20- 20 Insights"

[2] "Big Data A New World of Opportunities", NESSI White Paper, December 2012

[3] "Top Ten Big Data Security and Privacy Challenges", November 2012,Cloud Security Alliance.

[4] Jeff Kelly, David Floyer ,"The Industrial Internet and Big Data Analytics Opportunities and Challenges

[5] "A Guide to the Internet of Things Info graphic", Intel

[6] Ashraf Gaffar, Eman Monir Darwish, Abdessamad Tridane," Structuring Heterogeneous Big Data for Scalability and Accuracy".

[7] Olivier Monnier,"A smarter grid with Internet of Things", white paper.

[8] Big Data 101: "Unstructured Data Analytics", Intel IT Center.

[9] Lipika Dey, "Internet of Things, Big Data Analytics and Cloud Computing".

[10] Frank Gens, "TOP 10 PREDICTIONS,IDC Predictions 2012: Competing for 2020".

[11] "Big data Spectrum",Infosys.

[12] "Big Data",Economist Intelligence Unit,SAS

[13] Gartner, "IBM See Big Market for Big Data ", Data Center Knowledge, Intel

[14] Carl W.Olofson,Dan Vesset,"Big Data: Trends, Strategies", and SAP Technology, August 2012.

[15] Integrate For Insight,Oracle white Paper.

[16] Dr. Thomas Hill ,"The Big Data Revolution And How to Extract Value from Big Data"

[17] "Unstructured Data Doubles Every 3 Months, Intelligent Document Management",infopreserve, Intelligent Document Management.

[18] Ashraf Gaffar, Abdessamad Tridane, Eman Monir Darwish "The Failure and Success of Unstructured Data Reuse a Pragmatic Analysis"

[19] Ashraf Gaffar, Naouel Moha, "Semantics of a Pattern System"

[20] "Challenges of Integrated Structured and unstructured Data", January 2002

[21] Rapa Kudva,Som Mittal,"Big data :Next Big Thing",NASSCOM

[22] "Big Data for Development: Challenges & Opportunities", Global Pulse, May 2012.

[23] The Five Must-Haves of Big Data Storage,redhat White Paper.

[24] Fay Chang, Jeffrey Dean, Sanjay Ghemawat et.al."Big Table – A Distributed Storage System For Data"

[25] Tilmann Rabl, Mohammad Sadoghi, HansArno ,Jacobsen,"Solving Big Data Challenges for Enterprise Application Performance Management"

[26] "Addressing Big Data Security Challenges": The Right Tools for Smart Protection, Trend micro white paper September 2012.

[27] Jayavardhana Gubbia, Rajkumar Buyya, Slaven Marusic a, Marimuthu Palaniswami, "Internet of Things (IoT)": A vision, architectural elements, and future directions: Future Generation Computer Systems.

[28] Ralph Kimball ,"Newly Emerging Best Practices for Big Data, New Trends and Best Practices for Big data", A Kimball Group White Paper.

[29] Jack E.olson,Data Quality: The Accuracy Dimension.

[30] Dave Jewell, Arindam Hazra,et. Al ,Performance and Capacity Implications for Big Data,IBM Red paper.

[31] Vivek K. Singh, Mingyan Gao, and Ramesh Jain,"Situation Recogntition:An Evolving Problem for Heterogeneous Dynamic Big Multimedia Data".

[32] Changqing Ji ,Wenming Qiu, Yu Li ," Big Data Processing in Cloud Computing Environments", 2012 International Symposium on Pervasive Systems, Algorithms and Networks

[33] Lynn Anderson ,"Harness the Power of Big Data", HP Information Optimization Press Conference December 4, 2012

[34] Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, Michele Zorzi, "Internet of Things for Smart Cities".

[35] John A. , Mike Berners-Lee,"GeSI SMARTer 2020: The Role of ICT in Driving a Sustainable Future",December 2012 Global e-Sustainability Initiative aisbl and The Boston Consulting Group, Inc.

[36] http://wikibon.org/wiki/v/HadoopNoSQL_Software_and_Services_Market_Forecast_2012-2017

[37] http://www.statista.com/statistics/271652/worldwide-revenues-from-mobile-apps

[38] "digital universe growth from 2010- 2020". http://www.emc.com/about/news/press/2012/20121211-01.html