Escalation the Power of Big Data

Sujata A. Pardeshi
Department of Computer Science & Engineering
Sanjay Ghodawat Group of Institution's, Atigre
Kolhapur, India
sujatapardeshi@rediffmail.com

Pooja K. Akulwar
Department of Computer Science & Engineering
Sanjay Ghodawat Group of Institution's,
Kolhapur, India
poojaakulwar13@gmail.

Abstract – Today's life is the fast food life, everyone is in hurry up to get what they want as a superiority within short span of time and all such needs are easily available through Internet means now a day. The use of e-commerce, social media, mobile commerce, devices like sensors used by the users which leads to enormous data generation and it changes in proportion to user population. The data is a central point of Economical World & required to manage and manipulate with emerging trends. According to IDC Digital Universe Study done in 2011, 130 exabytes of data were created and stored. In 2005 this grew to 1,227 exabytes in 2010 and is projected to grow to 7,910 exabytes in 2015. The explosion of such larger data results into catastrophe situation in front of organizations since they are managing the large data sets and its operational structure with the traditional data management tools and techniques. But these traditional ways are facing limitations in organizing such escalation of data. So to elaborate in the world of today's Economical System, digging up the Big Data Technological phenomena is the greatest challenge and upcoming opportunity.

Index Terms – Big Data, Big Data Infrastructure, Digital Data

I. INTRODUCTION

The term coined by Roger Magoulas from O'Reilly Media in 2005[5], refers to a wide range of large data sets almost impossible to manage, analyze and processed using traditional database management tools due to their size as well as because of their complexity. In 2004, Wall Mart claimed that to have the largest data warehouse with 500 terabyte of storage. In 2009, eBay storage amounted to 8 petabytes. Two years later, the Yahoo Warehouse totaled 170 petabytes. Due to rise in digitization, enterprises from various verticals have been generating voluminous digital data and there is need of capturing trillions of bytes of information about the stake holders. The rapid growth of data constitutes the technological phenomenon called as "Big Data". According to International Data Corporation Digital Universe Study done in 2011 [1] [2] [3], 130 exabytes of data were created and stored. In 2005 this grew to 1,227 exabyte and is projected to grow to 7,910 exabytes in 2015[6].

In today's scenario with the explosion of sensors, smart devices as well as social networking verities of complex data are generated per minute through the digital customers such as e-commerce, social media and mobile commerce used and through which structured, semi structured and unstructured data generated and this growth increases per second. The structured data is grouped into relational schema, semi-structured is self describing and contains the tags or other markers to enforce hierarchies of records and fields with the dataset and unstructured data contains formats which can't be easily indexed into relational schemas for analysis or querying [15].

Big Data is used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using on-hand database management tools or traditional data processing applications. The most of application scenarios are looking forward to data that moves too fast or it exceeds current processing capacity of existing Data Warehouses.

Existing studies [1],[2],[3],[6],[7],[8],[10],[11],[12] show that it is challenging job to organize, acquire and analyze the growing digital data that changes with respect to usage and utilization of Internet things and population digital devices. The key point in escalating the supremacy of Big Data lies in the use of its awareness and starting the journey towards its Infrastructure to become a big with Big Data.

The rest of this paper is organized as follows: In Section II, we introduce the background study of Big Data, & its characteristics. In Section III, the Big Data Infrastructure, Tools and Techniques shows the way in use of this advancement of Big Science. The section IIII specifies the study on the usage and applications of Big Data. The section V looking forward of the challenges and big opportunities to become a big with Big Data and section VI gives the conclusion on the proposed study.

II. BIG DATA & ITS WORLD

According to McKinsey [4] Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, store, manage and analyse. There is no explicit definition of how big a dataset should be in order to be considered Big Data. New technology has to be in place to manage this Big Data phenomenon. IDC defines Big Data technologies as a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. According to O'Reilly, "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing

database architectures. To gain value from these data, there must be an alternative way to process it."[5]

A. Definition 1:

"Big Data refers to the massive amounts of data consist of structured, unstructured data that collect over time & are difficult to analyze and handle using common database management tools. Big Data includes business transactions, e-mail messages, photos, surveillance videos and activity, scientific data from sensors can reach in huge proportions over time".

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few [6].

B. Characteristics of Big Data

Figure 1 focus on 3 V's that describes the characteristics of Big Data technology. Volume is a huge amount of data that are likely to have mere gigabytes or terabytes to exabytes of data. Data volume will continue to grow, regardless of the organisation's size. Data can come from a variety of sources typically both internal and external to an organisation and in a variety of types. The velocity of data is defined as a frequency its generation. Conventional understanding of velocity typically considers how quickly the data arrives and is stored, and how quickly it can be retrieved. The characteristics of Big Data are structured data that are grouped into relational schema, semi-structured is self describing and contains the tags or other markers to enforce hierarchies of records and fields with the dataset and unstructured data contains formats which can't be easily indexed into relational schemas for analysis or querying [15].

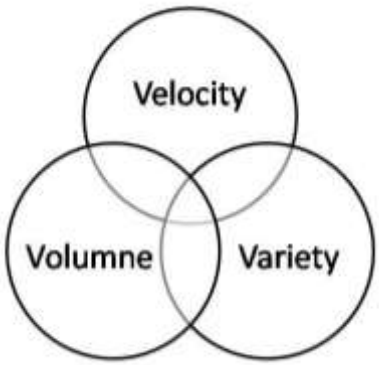

Figure1: The 3V's of Big Data

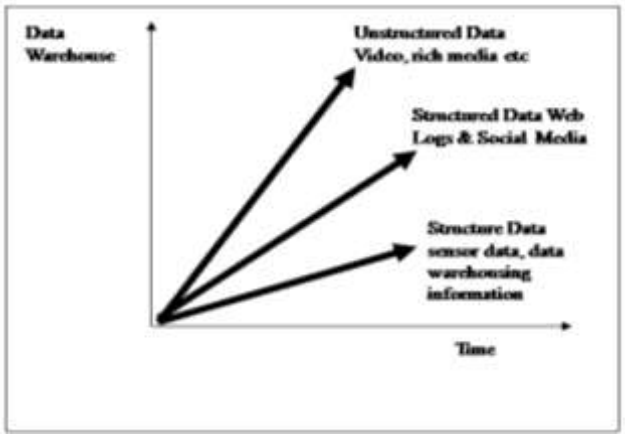

Figure2: The characteristics of Big Data

C. Significant of Big Data

As we know Big Data is used to organize, process & analyze & acquisition of large amount of data set. it can be used in finance & business where enormous amount stock exchange, banking, online and onsite purchasing data flows through Internet things every day and then data are captured, stored for Inventory Monitoring, Customer Behaviour & Market Intelligence. It can also be seen in the life science where big sets of data such Genome Sequencing , Clinical & Patient Data are analyzed & used in breakthrough in Science & Research.

III. BIG DATA INFRASTRUCURE – OPEN PLATFORMS

A. Investment in Big Data Infrastructure

For handling Big Data Hadoop technology is used as it handles ample amount of unstructured data. Today Big Data Infrastructure is affordable since it is available as open source platforms. According to IDC, data will grow 50 times by 2020 and there is need to make investment in handling such digital data and there is need of experts with efficient skill. In terms on investment, according to IDC, the global market size of Hadoop projects in 2011 was US$77 million. The market is expected to grow almost nine fold to US$682.5 million by 2015[10]. 15-18% of skilled experts are there to handle Big Data Infrastructure and there is need of more % of experts in today and future. Big Data Infrastructure stands to provide solutions and vendors such as IBM and EMC have plant launch training programs for their Hadoop based programs [22]. These experts are expected to understand and study about what strategies to be planed to utilize the Big Data Infrastructure with respect to planned budget and investment defined structure, hence these experts should have expatiation in such required and demanding Infrastructure.

B. Open Platforms

The Big Data having torrent of data streams with lots of variety, the infrastructure required to support the data organization, acquisition in both capturing and running complex or short queries in distributed environment always have dynamic data structure As with data warehousing, web

stores & IT platforms have to have use of Big Data Infrastructure which consist of components and it depends on: 1) Data Organization 2) Data Acquisition 3) Data Analysis.

Data Organization – In classical data warehouse organizing data is called as data integration and Big Data is able to process and manipulate data in original storage location. Hadoop is the new technology allows voluminous data to be organized and processed as data is stored at original data storage clusters. Hadoop Distributed File System is a long term storage system for a web logs and the web logs are turned into browsing behavior by running MapReduce programs on the clusters & generated the aggregated results on the same cluster.

Data Analysis – The infrastructure to analyze the data must support deeper analytics such as statistical analysis & data mining on a wider variety of data types stored in a diverse system and would deliver in a faster response time.

Hadoop is almost synonymous with the term "Big Data" in the industry and is popular for handling huge volumes of unstructured data. The Hadoop Distributed File System enables highly scalable, redundant data storage and processing environment that can be used to execute different types of large-scale computing projects. For large volume structured data processing, enterprises use analytical databases such as Greenplum Database and Teradata Aster Data Systems [14]. Many of these appliances offer connectors or plug-ins for integration with Hadoop systems.

1) Hadoop

Hadoop, formally called *Apache Hadoop*, is an Apache Software Foundation project and open source software platform for scalable, distributed computing. Hadoop can provide fast and reliable analysis of both structured data and unstructured data. Given its capabilities to handle large data sets, it's often associated with the phrase *big data*. Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data, Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits [15].

2) MapReduce

Most enterprise data management tools are designed to make simple queries run quickly. Typically, the data is indexed so that only small portions of the data need to examine in order to answer a query. This solution, however does not work for data that cannot be indexed, namely in semi-structured form or unstructured form. To answer a query in this case, all the data to be examined. Hadoop uses the MapReduce [9] technique to carry out this exhaustive analysis quickly, and it is data processing algorithm that uses a parallel programming implementation. It is a programming paradigm that involves distributing a task across multiple nodes running a "map" function. The map function takes the problem, splits it into sub-parts and sends them to different machines so that all sub-parts can run concurrently. The results from the parallel map functions are collected and distributed to a set of servers running "reduce" functions, which then takes the results from the sub-parts and re-combines them to get the single answer.

3) HDFS – Hadoop Distributed File System

The HDFS is a fault tolerant storage system that can store huge amounts of information, scale up incrementally and survive storage failure without failure without losing data. Hadoop clusters are built within expensive computers. If one computer fails, the cluster can continue to operate without losing data or interrupting work by simply re-distributing the work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking files into small blocks and storing duplicated copies of them across the pool of nodes. HDFS offers the two key advantages: Firstly, HDFS requires no special hardware as it can be built from one common hardware. Secondly, it enables an efficient technique of data processing in the form of MapReduce.

4) The Hadoop Eco-System

Hadoop refers to a collection of other software projects that uses the MapReduce and HDFS framework. A tool designed for transferring bulk data between Hadoop and structured data stores such as relational databases [17]. The following table briefly describes some of the tools of HDFS frame work

Technology	Use
HBase	A key value pair DBMS that runs on HDFS
Hive	A system of functions that support data summarization and ad hoc query of the Hadoop MapReduce result set used for data warehousing
Pig	High level language for managing data flow and application execution in the Hadoop environment
Mahout	Machine learning system implemented on Hadoop
Zookeeper	Centralised service for maintaining configuration information, naming, providing distributed synchronisation and group services
Sqoop	A tool designed for transferring bulk data between Hadoop and structured data stores such as relational databases

Table 1: Big Data Open Source Technology Types

5) NoSQL Database Management System

NoSQL database management systems (DBMSs) are available as open source software and designed for use in high data volume applications in clustered environments. They often do not have fixed schema and are non-relational, unlike the traditional SQL database management system (also known as RDMS) in many data warehouses today. Because they do not adhere to a fixed schema, NoSQL DBMS permit more flexible

usage, allowing high-speed access to semi-structured and unstructured data. However, SQL interfaces are also increasingly being used alongside the MapReduce programming paradigm [7].

NoSQL databases are designed to be able to scale out on commodity hardware to manage the exploding data and transaction volumes. The result is that the cost per gigabyte or transactions per second for NoSQL can be many times less than the cost for RDBMS, allowing more data storage and processing at a lower price point. However, it is important to recognize that the NoSQL database can realistically focus on two of the three properties of Consistency, Availability and Partition Tolerance (CAP Theorem). NoSQL databases need partition tolerance in order to scale properly, so it is very likely they will have to sacrifice either availability or consistency.

Semi-structured and unstructured data sets are the two fastest growing data types in the digital universe. Analysis of these two data types will not be possible with traditional database management systems. Hadoop HDFS and MapReduce enable the analysis of these two data types, giving organizations the opportunity to extract insights from bigger datasets within a reasonable amount of processing time. Hadoop MapReduce's parallel processing capability has increased the speed of extraction and transformation of data. Hadoop MapReduce can be used as a data integration tool by reducing large amounts of data to its representative form which can then be stored in the data warehouse. At the current stage of development, Hadoop is not meant to be a replacement for scale-up storage and is designed more for batch processing rather than for interactive applications.

C. Cloud Services & Big Data

In Early days the users of Big Data needs to billions of investment budget for deploying their applications with Big Data Infrastructure and it is not affordable for organization having millions of investment with respect to their annual budget. Cloud services with the ability to ingest, store and analyse data have been available for some time and they enable organisations to overcome the challenges associated with Big Data. The Cloud Service providers such as Infrastructure-as-a-Service (IaaS) providers such as Amazon Web Services and Rackspace, for test and development, and analysis of existing datasets offer data storage and data back-up in a cost-effective manner and facilitated to test and development, and analysis of existing datasets. They deliver a low-cost and reliable environment that gives organisations the computing resources to store their structured and unstructured data as per needs. At the Software-as-a-Service (SaaS) level, embedded analytics engines help to analyse the data stored on the cloud, then analytics output can be provides to then end user through a graphical interface. However, the development of queries and integration to the data source on the cloud are prerequisites that organisations need to undertake before the usability can be delivered. Cloud services are useful for organisations looking to test and develop new processes and applications.

IV. BIG DATA AND APPLICATIONS

The organizations to survive in a highly competitive environment they need the application to handle and manage the ample amount of data; the traditional technology has certain limitations. And the supremacy of Big Data prepares the users to face the challenges to create meaningful insights that will be translated to new economic value.

A. Big Data & Business World

Organizations can leverage Big Data technologies to create highly specific customer segmentations at granular level with respect to their behaviour and to customise products and services that meet their needs. Such functionality may be well-known in the field of marketing and retailing used in defining the business forecasting strategies [13]. Sophisticated analytics with automated algorithms can provide valuable insights that would otherwise remain hidden. These insights can then be used to minimize decision risks and improve data driven decision making. In some cases, decisions may not be completely automated but only augmented by the analysis of huge data sets using Big Data techniques rather than small data sets and samples that individual decision makers can handle and understand. Innovating new business models, products and services: Using emerging Big Data technologies, companies can enhance and create new products and services.

Business such as retailers, enterprisers, health care marketing , cross selling etc have been spending lots if investment on Big Data Infrastructure and Services & will be reached nearly by $10Million in 2013 and in 2016 $20Billion [12]. The business leaders are interested in such investment to harness the power of Big Data, the small scale businesses are attracted in driving their business the Big Data infrastructure – Hadoop technology that requires billions of investment with respect to budget and investment. But it's become difficult for small business to cope with such huge amount of investment and hence they turn to Big Data Infrastructure and Services provided by third party Cloud Services. Cloud Services provide Small Business Enterprisers an affordable and easy way to grow their existing business & uncover new opportunities, hence there no barrier between Small Business Enterprisers and use of Big Data Advancements. The some providers like Infrastructure as a service (IaaS) such as Amazon, Microsoft, Google, Go Grid, Rack space, Slices hot etc. provides all the services on "Pay-as-you-use" basis. It is safe to say that there are solutions that allow a business to perform simple data analytics on a terabytes of data for as low as $100. Leveraging the benefits of Cloud Services offers a quick, low cost and easy business analytics and solutions that help the users in different areas.

B. Big Data & Public Sector

The government agencies also harness the power of Big Data in the same way it does other organisations. Big Data brings the potential to transform the work of government agencies by helping them to operate more efficiently, create more transparency and make more informed decisions. The various government agencies accumulate large data set over the years

and offer new opportunities which can use Big Data technologies to extract insights and keep track of citizens' specific needs. In turn, these insights could then be used to improve government services by performing data analytics that uses data mining to provide actionable, forward-looking intelligence to make improvement across government agencies [8] & offers a range of new capabilities that includes operational improvements in the areas of citizen service provision and tax fraud detection, policy development, etc.

C. Big Data & Education

In the education sector, learners are creating information at the same time as they are consuming knowledge. Students are facing with increasingly demanding curricula where they are no longer expected to regurgitate facts from hard memorising but are required to learn the subjects with deep understanding. At the same time, there is high expectation from stake holders regarding the personal attention and monitoring during teaching learning session. The challenge for the educators is to be able to have a clear profile of each of the students under their charge. Creating a profile for each of the students would require disparate sets of information so that it is possible for them to make study about which teaching methodology will be applicable and this involves the opportunity for Big Data analytics.

As more emphasis is being placed on adaptive, personal learning and to make it flexible as per the learners, there is the need to mine unstructured data such as student interactions and any form of student generated content. Learning analytics, such as Social Networks Adapting Pedagogical Practice (SNAPP), can be deployed to analyse this data. SNAPP is a software tool that allows users to visualise the network of interactions resulting from discussion forum posts and replies. The visualisation provides an opportunity for teachers to rapidly identify patterns of user behaviour at any stage of course progression and it provides quick identification of the levels of learner engagement and network density emerging from any online learning activities. From there, disengaged and low performing students can be identified.

V. CHALLENGES & OPPORTUNITY WITH BIG DATA

Figure 3: A Big……Opprotunity to become a Big with Big Data

A. Challenges

1. Data generated from the Internet things through sensors and devices embedded in the physical world and connected by networks to computing resources, is another trend in driving the growth of big data, it is big challenge to organize it with Big Data by understanding its Infrastructure [12].
2. The penetration of social networks is growing and frequent use of smart phones and sharing of information among the users rapidly increase, since social media and other online resources are used by the users leads to challenges to manipulate such digital data [12] [21].
3. There is a lack of experts in understanding & effective use of the analytical tools of Big Data skills. There are lots of technical engineers who can build distributed systems, and works with voluminous amount of captured data sets, and because of lacking in skills about what to do with the data sets and how it is useful? In BI Industry today, the challenge is to provide the faster and easier way to access their existing knowledge rather than reaching out into the distance and discovering new knowledge. The people with pure data analysis and knowledge discovery skills are much harder to find. These are people, who can make a real and significant impact on an organizations bottom line, and help to solve some of the R&D to find out better insights. Data Geeks are the people to be revered in the future and hopefully & there is need of experts called as Data Scientists & this need will grow rapidly.

B. Opportunity

1. Big Data is becoming fast track technology in solving the functional problems and it includes enabling the target marketing, improving marketing spend and enhancing campaign. It supports for supply chain management for retailers with expansive item catalogs and analytics to improve supplier negotiation. [15]. In Indian organizations are seeing greater opportunity to become big with big data. A new survey by Informatica Corporation states that 72% of Indian organizations are now considering, planning or running 'Big Data' projects, with organizations viewing the trend as an opportunity rather than an IT challenge [23].
2. Aiming to expand its Big Data analytics portfolio, Hewlett-Packard has launched the big data analytics platform 'HAVEn', that leverages HP's analytics software, hardware and services to enable organizations to gain better insight into their data and deliver real-time outcomes and designed to work with next generation of big data-ready analytics applications and solutions" [4].
3. In education [24], the opportunities for Big Data analytics can be exemplified by Civitas Learning, a digital education platform that uses predictive analytics to help guide educational decision making. Civitas takes data such as demographic, behavioural and academic data provided by partner institutions, and anonymises and combines them. The data is then analysed to identify trends and provide insights, such as identifying the

courses and tracks that are most beneficial, suitable as per the student profiles and the instructional approaches that tend to be most effective at ensuring good educational outcomes. These insights are translated into real-time recommendations for students, instructors, and administrators through a customised platform.

4. Cloudera was founded in October 2008 to deliver the first enterprise-class implementation of Apache Hadoop & help to deploy an Enterprise Data Hub solve data management problems powered by Apache Hadoop.
5. IDC predicts spending of more than $14 billion on big data technologies and services or 30% growth year-over-year", as demand for big data analytics skills continue to outstrip supply." The cloud will play a bigger role with IDC predicting a race to develop cloud-based platforms capable of streaming data in real time. There will be increased use by enterprises of externally-sourced data and applications and "data brokers will proliferate." IDC predicts explosive growth in big data analytics services, with the number of providers to triple in three years. 2014 spending on these services will exceed $4.5 billion, growing by 21%.[22]

VI. CONCLUSION

In this paper, we have proposed how the rapid growth of data constitutes advancements in technological world that have given rise to an ecosystem of software and hardware products and enabling users to analyze the data to produce new insights. This technological phenomenon is the "Big Data" that provides the Infrastructure which easily handles & manipulate the explosion of digital data assets and upcoming healthy and wealthy property of the organization to survive them in the competitive environment. Today's global economy takes everybody to run on the highly economical track and hence for organization needs to understand its running speed. According to George Washington Carver, "Start where you are, with what you have. Make something of it & never be satisfied" and according to him, it is need to find what we have and start immediately with to run with global economical world. And therefore it is an opportunity to become a member of Big Data Family to start & tolerate the challenges of it and grab the chance to become a big with Big Data.

ACKNOWLEDGMENT

We would like to thanks to Dr. Aniruddha D. Joshi, Dr. V. A. Raikar Advisory Member, Prof. S. M. Ingale, the In-charge Director, and Prof. A. S. Kamble H.O.D. of Computer Science & Engineering Department, Sanjay Ghodawat Group of Institutions, Atigre, Shivaji University for their valuable guidance and motivation for this work

REFERENCES

[1] IDC. The Digital Universe Study: Extracting Value from Chaos, 2011.

[2] IDC Internet consumer traffic analysis, 2010.

[3] IDC. Worldwide Big Data Technology and Services 2012-2015 Forecast. (for the purpose of data growth)

[4] A HAVEN Platform - http://www8.hp.com/us/en/software-solutions/big-data-platform-haven.html

[5] Ovum. What is Big Data: The End Game http://ovum.com/research/what-is-big-data-the-end-game

[6] IBM. Data growth and standards. http://www.ibm.com

[7] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters http://static.usenix.org

[8] Mike Loukides. What is data science?: http://radar.oreilly.com/2010/06/what-is-data-science.html

[9] Carl W. Olofson, Dan Vesset. Worldwide Hadoop – MapReduce Ecosystem Software 2012-2016 Forecast.: http://www.idc.com

[10] Whatsthebigdata.com. A Very Short History of Data Science http://whatsthebigdata.com/2012/04/26/a-very-short-history-of-data-science. [analytics]

[11] Mike Loukides. What is data science? http://radar.oreilly.com/2010/06/what-is-data-science.html.

[12] James Manyika, et al. Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.

[13] Srinivasan. N, Rajeeb Nayar "Harnesing the Power of Big Data Big Opportunity for Retailers to win Customers", publishes on www.infosys.com.

[14] Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future" SIGKDD Explorations Volume 14, Issue 2.

[15] Apache Hadoop, http://hadoop.apache.org.

[16] Lelia Voinea, Alina Filip, "Analyzing the Main Changes in New Consumer Buying Behavior during Economic Crisis", International Journal of Economic Practices and Theories, Vol. 1, No. 1, 2011 (July).

[17] Douglas Laney, Lisa Kart. Emerging Role of the Data Scientist and the Art of Data Science. http://www.gartner.com.

[18] Rob Addy. Emerging Technology Analysis: Predictive Support Services. http://www.gartner.com

[19] David White. Picture this: Self-Service BI through Data Discovery & Visualization. http://aberdeen.com/Aberdeen-Library/7729/AI-business-intelligence-analytics.aspx

[20] Massimo Pezzini. Net IT Out: In-Memory Computing — Thinking the Unthinkable Applications. http://agendabuilder.gartner.com.

[21] Information Management. Mobile Business Intelligence for Intelligent Businesses. http://www.information-management.com

[22] IDC. http://www.idc.com

[23] UNICOM is organizing a series of Conference UNICOM Big Data & Cloud on Big Data called "India Big data Week 2014", http://www.bigdatainnovation.org

[24] Dawson, S. (2010). 'Seeing' the learning community: An exploration of the development of a resource for monitoring online student networking. British Journal of Educational Technology, 41 (5), 736-752.

.