ISSN: 2321-7782 (Online)
Volume 2, Issue 9, September 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Big Data: Moving Forward with Emerging Technology and Challenges

Vibha Shukla[1]
Computer Science Department
NIET Gr. Noida
Uttar Pradesh – India

Pawan Kumar Dubey[2]
Electronics and communication Department
NIET Gr. Noida
Uttar Pradesh – India

Abstract: ***The amount of available data has exploded in the past few years because of new social behaviour, social transformation as well as vast spread of social system. Big data become very important driver for innovation and growth. This paper provides a brief review of several papers, articles to know about the background of big data; explains the emerging technologies and what existing data and technologies should be augmented with big data technology enablers. We also address the challenges and opportunities rising from the use of big data.***

Keywords: Big Data, 3Vs, Traditional Data, Technologies, Enterprise Initatives.

I. INTRODUCTION

The amount of data that is generated and stored is increasing rapidly, even exponentially. It may be doubling every two years, according to one estimate (IDC 2011). New advanced analytics techniques and technologies allow practitioners to connect and interrogate datasets. Big data describes innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte- or larger- sized datasets with high-velocity and diverse structures that conventional data management methods are incapable of handling.

Traditional data management and analysis systems are based on the relational database management system (RDBMS). It is apparent that the traditional RDBMSs could not handle the huge volume and heterogeneity of big data. For solutions of permanent storage and management of large-scale disordered datasets, distributed file systems and NoSQL (Not Only SQL) databases are good choices. Traditional Data analysis approaches are: Cluster Analysis, Factor Analysis, Correlation Analysis, Regression Analysis, A/B Testing, Statistical Analysis and Data Mining Algorithms. Big data analytic methods Big Data analysis approaches are: Bloom Filter, Hashing, Index, Triel & Parallel Computing.

With the amount of data expanding, technology challenges, organization limitations, and privacy/trust concerns - among other obstacles – organization must approach analyzing the data with an array of new and emerging technologies .While the amount of large datasets is drastically rising, it also brings about many challenging problems demanding prompt solutions. Big is characterized by the 3 Vs. The three Vs of Big Data are: Variety, Volume and Velocity.

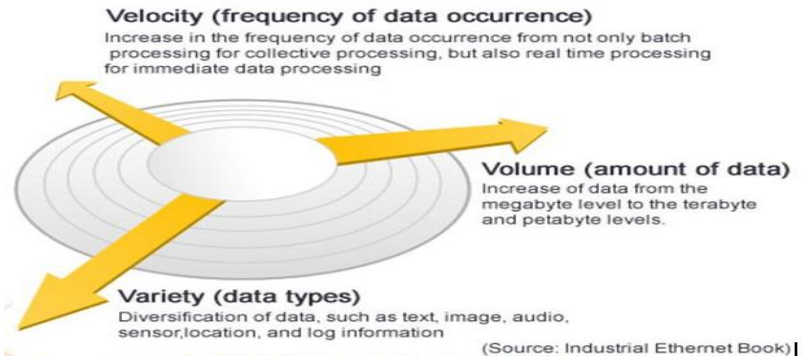

Fig:1 Model Expressing Big Data in 3 dimensions

II. LITERATURE REVIEW

John Gantz and David Reinsel , in 2011 , said that The growth of the digital universe continues to outpace the growth of storage capacity. But keep in mind that a gigabyte of stored content can generate a petabyte or more of transient data that we typically don't store (e.g., digital TV signals we watch but don't record, voice calls that are made digital in the network backbone for the duration of a call). So, like our physical universe, the digital universe is something to behold — 1.8 trillion gigabytes in 500 quadrillion "files" — and more than doubling every two years. That's nearly as many bits of information in the digital universe as stars in our physical universe

Mango DB white paper states that companies should consider 5 critical dimensions to make the right choice for their applications and their businesses. These are - Data Model, Query Model, Consistency Model, APIs & Commercial Support and Community Strength. Moniruzzaman, A. B. M.& Hossain, Syed Akhter presented a paper to motivate - classification, characteristics and evaluation of NoSQL databases in Big Data Analytics. This report is intended to help users, especially to the organizations to obtain an independent understanding of the strengths and weaknesses of various NoSQL database approaches to supporting applications that process huge volumes of data.

In 2012, Danah boyd & Kate Crawford presented that the era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Six Provocations given them for Big Data are: 1. Big Data changes the definition of knowledge 2. Claims to objectivity and accuracy are misleading 3. Bigger data are not always better data 4. Taken out of context, Big Data loses its meaning 5. Just because it is accessible does not make it ethical 6. Limited access to Big Data creates new digital divides.

Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared disk big data analytics with Apache Hadoop" Big data analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns. Big data analytics refers to the Mapreduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model [2]. In this the performance of SF-CFS is compared with the HDFS using the SWIM by the facebook job traces .SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns.

In December 2012, Hsinchun Chen , Roger H & Veda C said ,"Now, in this era of Big Data, even while BI&A 2.0 is still maturing, we find ourselves poised at the brink of BI&A 3.0, with all the attendant uncertainty that new and potentially revolutionary technologies bring". In a paper submitted in April 2013, Renu Kanwar, Prakriti Trivedi & Kuldeep Singh claimed that NoSQL is the solution for use cases where ACID is not the major concern and uses BASE instead which works up on eventual consistency.

Real Time Literature Review about the Big data According to 2013, facebook has 1.11 billion people active accounts from which 751 million using facebook from a mobile. Another example is flicker having feature of Unlimited photo uploads (50MB per photo), Unlimited video uploads (90 seconds max, 500MB per video), the ability to show HD Video, Unlimited storage, Unlimited bandwidth. Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily. In Feb 2014, Wenliang Huang, Zhen Chen, Wenyu Dong, Hang Li, Bin Cao, and Junwei Cao presented a comparison of HBase and Oracle, "Compared with Oracle database, our HBase shows very consistent performance, and the peak insertion rate reaches approximately 100 000 records per second".

III. TRADITIONAL DATA VERSUS BIG DATA ANALYTICS

Traditional Analytics analyzes on the known data terrain that too the data that is well understood. Most of the data warehouses have a elaborate ETL processes and database constraints, which means the data that is loaded inside a data warehouse is well under stood, cleansed and in line with the business metadata. The biggest advantages of the Big Data is it is

targeted at unstructured data outside of traditional means of capturing the data. Which means there is no guarantee that the incoming data is well formed and clean and devoid of any errors. This makes it more challenging but at the same time it gives a scope for much more insight into the data.

Traditional Analytics is built on top of the relational data model, relationships between the subjects of interests have been created inside the system and the analysis is done based on them. In typical world, it is very difficult to establish relationship between all the information in a formal way, and hence unstructured data in the form images, videos, Mobile generated information, RFID etc... Have to be considered in big data analytics. Most of the big data analytics database is based out columnar databases.

Traditional analytics is batch oriented and we need to wait for nightly ETL and transformation jobs to complete before the required insight is obtained. Big Data Analytics is aimed at near real time analysis of the data using the support of the software meant for it.Parallelism in a traditional analytics system is achieved through costly hardware like MPP (Massively Parallel Processing) systems and / or SMP systems. While there are appliances in the market for the Big Data Analytics, this can also be achieved through commodity hardware and new generation of analytical software like Hadoop or other Analytical databases.

	Traditional Data Analytics	Big Data Analytics
1.	Environment suitable for structured data only.Usual unit of volume is megabyte/gigabyte.	Environment suitable for any data – structured, semi-/multi- /unstructured data.Usual unit of volume is terabyte or petabyte
2.	Technology used is Relational database (data to function model).	.Tecnolgy used is Hadoop framework (function to data model).
3.	Batch processing (offline) of "historical," static data	Stream processing (online) of (near) real time, live data
4.	No open source	Open source
5.	Individual manufacturer/ developer can work independently.	Nobody works alone; all related parties must.

Table:1

IV. BIG DATA EMERGING TECHNIQUES AND TECHNOLOGY

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data. There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies.

A. NoSQL databases

There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

B. MapReduce

This is a programming paragigm that allows for massive jobs execution scalability against thousand of servers. Any MapReduce impletation consists of two tasks:

The "Map" task, where an input dataset is converted onto a different set of value/key pairs, or tuples.

The "Reduce"task, where several of the "Map " are combined task to form a reduced set of tuples(Hence the name).

C. Hadoop

Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

D. Hive

Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.

E. PIG

PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

F. PLATFORA

Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases. PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

G. HDFS

HDFS is a block-structured distributed file system that holds the large amount of Big Data. In the HDFS the data is stored in blocks that are known as chunks. HDFS is client-server architecture comprises of NameNode and many DataNodes.The name node stores the metadata for the Name Node. Name Nodes keeps track of the state of the DataNodes. NameNode is also responsible for the file system operations etc. When Name Node fails the Hadoop doesn’t support automatic recovery, but the configuration of secondary nod is possible

H. HPCC

HPCC is a open source computing platform and provide the services for management of big data workflow. HPCC’ data model is defined by the user. HPCC system is designed to manage the most complex and data-intensive analytical problems. HPCC system is a single platform, a single architecture and a single programming language used for the data processing.HPCC system is based on Enterprise control language that is declarative, on-procedural programming language HPCC system was built to analyze the large volume data for the purpose of solving complex problem.

I. STORM

Storm is probably the open-sourced tool the most used for streams processing. Storm is a distributed real-time computation system that is fault-tolerant and guarantees data processing. Storm makes it easy to reliably process unbounded streams of data, doing for real time processing what Hadoop did for batch processing. Storm is very simple and was designed from the ground up to be usable with any programming language.

The rampant growth of connected technologies creates even more acceleration in Big Data trends – it is a "data tsunami". Smart cloud technology will emerge to respond to the needs of the enterprise and the consumer there are the some Existing data and technology initiatives that should be augmented with Big Data technology enablers to make it more beneficial: Below table shows briefly Enterprise initiatives and their scenario.

Enterprise Initiative	Scenarios
CRM Platform	• Explore hosting of CRM platform in the cloud • Extend CRM platform by integrating social media technology to enhance customer view. • Integrate operational analytics with CRM modules to aid sales interaction.
Customer Portal	• Enhance use of interfaces for internal customers to allow for seamless integration with external interfaces. • Incorporate all customer data into one portal so the consumer experience can be further customized.
Enterprise Data Management	• Explore alternative technologies for data integration across channels (i.e. Hadoop, MapReduce). • Leverage technology to integrate structured enterprise data with unstructured data from social media and other sources.
Risk Management	• New analytical tools and processes can enable enhanced risk management models and proactive fraud and loss prediction. • Information security initiatives will also be necessary to protect data especially in light of trends such as Bring Your Own Device

Table:2

V. MAJOR CHALLENGES

Big data's researchers, developers and users are experiencing many unknowns during the journey of discovery and application. Organisations big data users specifically mainly face four big challenges: talent, leadership, technology capacity, and budget.

A. Talent- By the McKinsey Global Institute (2011) that, by 2018, the United States alone may face a 50 to 60 percent gap between supply and the requisite demand of deep analytic talent and data managers, respectively. While "there is a storm approaching on the Big Data talent front," more than 60 percent of the survey respondents in the *Big Data Executive Survey 2013* found that it is very difficult or even impossible to find or hire data scientists for their organizations (NewVantage Partners, 2013). We should implement formal education through institutions of higher education. Another solution is training from inside. An innovative solution is crowd sourcing, where the needed services, ideas, or content are obtained by soliciting contributions from a large pool of experts, especially from an online community, rather than from traditional employees or suppliers.

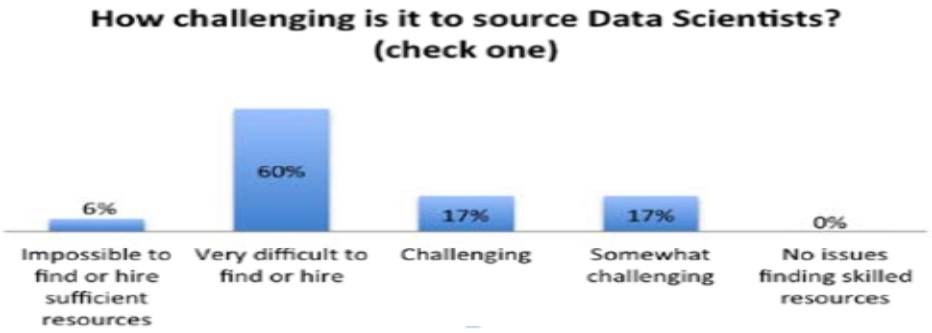

Fig2: Big Data Executive Survey 2013, Summary Report, NVP

B. Leadership- Leadership on big data comes with a few challenges. One challenge is taking initiative on big data. When the question, "Where is your agency with big data?" was asked to 151 Federal agency IT professionals in the *Big Data Gap* survey (NetApp, 2012), less than 10 percent of respondents reported that they have the infrastructure to successfully leverage big data, while 60 percent of respondents from civil agencies are just learning about it, indicating the urgency of big data awareness. Further, respondents reported that it will take agencies at least 3 years on average to take full advantage of big data. There should be Clear responsibility is the to the leadership in the implementation of big data because ambiguity about responsibilities could cause confusion in practice.

C. Capacity And Budget- To take the advantage of big data technologies, practical capacity may hold them from moving ahead. about 40 percent of senior managers or business leaders have adequate ability to use data and analytics to improve and transform the business. Although big data analytics have great potential to bring value to organizations, traditional servers in organizations' data centers are not designed to process big data. Analytics servers, and in some cases, high-performance computing servers and applications will be needed, which requires new IT investment. Solution-

D. Technology And Privacy -Technical challenges also exist for big data analytics. technical challenges also exist for big data analytics. Privacy of data is another huge concern, and one that increases in the context of Big Data. Organization must establish effective policies, procedures, responsibility and processes for big data analytics and use,and incorporate privacy and security controls into the related processes before actually putting them into use.

There are many additional challenging research problems. For Example, Velocity, volume, and variety of data show no signs of slowing which will require new technology solutions. And Most of the enterprise implementations are in pilot stages. Processing big data is also a major challenge. A critical issue is whether or not an analytic process scales as the data set increases by orders of magnitude.

VI. CONCLUSION

We have entered an era of Big Data. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to massage and process data. Big data is not formally and structurally defined. The big data technology is still in its infancy. The analysis of big data is confronted with many challenges, but the current research is early stage. This paper is a collaborative research effort to begin examining the traditional view of data analytics and big data analytics, introduces new techniques and technologies to handle big data. We have identified some major challenges regarding big data which big data users specifically face. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data. Our future research will concentrate on developing a more complete understanding of challenges associated with big data.

References

1. Gantz, J. and E. Reinsel. 2011. "Extracting Value from Chaos", IDC's Digital Universe Study, sponsored by EMC.
2. Hopkins, B. "Big opportunities in big data". Forrest Research, Inc. , May 18, 2011 .
3. McKinsey Global Institute (2011), "Big Data: The Next Frontier for Innovation, Competition and Productivity".
4. Gartner (2012a), "The Importance of 'Big Data': A Definition" ,October,2013.
5. Gantz, J. and E. Reinsel. 2011. "Extracting Value from Chaos", IDC's Digital Universe Study, sponsored by EMC.
6. Mango DB, "Top 5 considerations when evaluating NoSQL Databases", White Paper.
7. Danah boyd & Kate Crawford , "CRITICAL QUESTIONS FOR BIG DATA",Routledge, Information, Communication & Society Vol. 15, No. 5, June 2012, pp. 662–679 ISSN 1369-118 (2012).
8. Hsinchun Chen , Roger H. L , Veda C., "BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT", MIS Quarterly Vol. 36 No. 4/December 2012,Eller College of Management, University of Arizona, Tucson, AZ 85721 U.S.A.(2012).
9. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. , "Shared disk big data analytics with Apache Hadoop" , Dec.,2012.
10. Stephen Kaisler,Frank Armour Big Data: Issues and Challenges Moving Forward, 46th Hawaii International Conference on System Sciences, 2013.
11. The Economist. 2010. "Data, Data Everywhere",(online edition, February 28) http://www.economist.com/node/15557443

AUTHOR(S) PROFILE

Vibha Shukla received Btech degree in Information Technology with HONOURS and Pursuing M.Tech degrees in Software Engineering from Noida Institute of Engineering & Technology in 2011 and 2014, respectively. Her interest is in research field and she is very good in research work. Her interest area is big data and software.

Pawan Kumar Dubey received Btech degree in Electronics and Communication. And Mtech degree in Telecommunication from Noida Institute of Engineering and Technology. He is faculty in Sharda Group of Institutions. He is very good in research work.