Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# Big Data – Literature Survey

Sabitha M.S[1], Dr.S.Vijayalakshmi[2], R.M.Rathikaa Sre[3]
[1]*Research & Development Centre, Bharathiar University, Coimbatore,India*
[2]*Assistant Professor, Thiagarajar College of Engineering, Madurai, India*
[3]*B.E., Electronics & Communication I year, Mepco Schlenk Engineering College, Sivakasi,India*

***Abstract*— In the past few years, tremendous changes are happening in Cloud Computing, Big Data, Communication technology and Internet of things. Shift to the latest technology is envisaging new upcoming challenges. Big Data is becoming a vital transformation for the enterprises and scientific community. IoT, social websites, medical science, automated and smart devices will fuel the explosion of data for the near future. This transformation provides an opportunity to find the insights to make the businesses more agile and to answer the questions which were considered beyond our reach before. The core purpose of this paper is to discuss the views of various researchers, Big Data tools and techniques required for storage, management and analytic and its growth and suspected challenges in various domains.**

## I. INTRODUCTION

During the past few years, several changes are happening in the IT field especially in the area of Cloud computing, Big Data, mobility and Internet of things. It creates a new platform for the enterprises to penetrate into new business. Due to internet and social media penetration vast amount of data produced significantly in the past two years. Everyday's data generation exceeds 2.5 quintillion bytes of data. Today's 90% of the data created within last two years of time. The enormous speed of data growth is due to the services and users producing vast amounts of data. Internet of Things will be an important trigger for database growth for the near future. Estimated number internet connected devices in the year 2020 will be 16 to 50 billion . Machine and users need to collaborate in intelligent way to generate data and the management of these data would be the Big Data challenge. This is going to be the big challenge for large data streams that we receive from everyday devices and finding useful and meaningful hidden information from the large stream and it is very hard to detect the behavioral patterns out of it. Higher level of decision making and prediction capabilities of the applications and services are very important to get the full benefits from the context aware data intensive applications and services and make the valuable or important information transparent and available at a much higher frequency.

In this context, Big Data becomes very important, making possible to turn into this amount of data in information, knowledge, decision making and, ultimately, insights. A view of what are the Big Data has been exposed to Gartner that defines Big Data as high volume, velocity and variety information requires innovative way of information processing and to derive enhanced insights through data analytical tools, automation of process and effective decision making. Also it demands for the cost effective solution." [6]. In fact the Big Data doesn't mean only the volume .Big Data characterized as 3Vs – Volume (data size and number), Velocity (the rate at which data are generated or need to the processed) and Variety (different types / different forms of content).

The volume of Big Data is expanding beyond terabytes into peta bytes and even exabytes (1 million TB). Varity refers to the data from different types of sources like sensors, devices, machines and unknown things. It means different data types, data formats, structured, semi structured and unstructured data. Speed of the production of data and to process to the data to generate valuable insights referred as Velocity. In fact the life of data can be very short and this may become obsolete after some time. So efficient usage of Big Data analytics results will bring good useful insights from high volume, variety of data. [7]. Sometimes quality of the data is a concern area because it fetches data from different applications for making decisions. Not all the data captured from various devices are useful for making decisions. These are just information and the information has to be converted into knowledge for decision making.

Big Data are very large and complex that it is really tough and only with traditional approaches it is very difficult to process and analyze the data. Effective data management for Big Data sets is not possible with traditional RDBMS (Relational database management systems). Due to the size of Big Data it is very difficult to extract the information in a proper and required manner. Bringing insights from the large amount of data is very much useful. In fact the raw input is the data that is processed into information. The data has no meaning when it is individual. But volume of data will provide some meaningful output and it provide trends and patterns. The converted knowledge will be the used for further analysis. The combination of knowledge and experience is

Wisdom [9]. Conversion from raw data to valuable information is a challenge. To handle Big Data, Innovation in latest technologies and application of different techniques will help the individuals and organizations to collect the data, various analysis and visualize various formats of data in different industries and various domains. The aim of this chapter is to provide technical aspects of Big Data and Big Data challenges in Internet of things (IoT).

In this paper, Section III deals about the technical facts of Big Data.

## II. MOTIVATION

Big challenge in the business environment is combining structured and unstructured formats from various applications and projects into single format. Big Data handles this kind of data and provides valuable insights. Data analytics plays a major role in today's dynamic business environment. Traditional method of batch analytics is useful only to understand historical analysis. It provides the trend over a period. But real-time analytics will provide the insights of the live data.

Right technology at right time will impact much for any organisation to move forward. In the world, 90% of the data created in the last couple of years alone and every day we create 2.5 quintillion bytes of data. This is the right time to move towards Big Data technology which can handle huge amount of data. But at the same time, which can produce useful insights from the data.

The increasing data produced by the Internet of Things (IoT) will fuel the explosion of data for the near future. The combination of Internet and emerging technologies such as wireless / embedded sensor, communication technology and smart objects should to be capable to understand and react to their environment. Big Data is a right choice to address the global infrastructure created by Internet of Things.

## III. LITERATURE REVIEW

Literature review done in IoT, Cloud Computing, Smart City and Hadoop and MapReduce.

*A. Big Data Platforms For The Internet Of Things, Radu Ioan Ciobanu, Valentin Cristea, Ciprian Dobre And Florin Pop.2014,Springer*

This paper focuses on how Big Data could change the research direction in the business model by providing services along with products. Technology shift generate more data through various applications like wireless sensors, smart devices, social media etc., This paper focuses on the improvement the performance of the old services and offer new services in an open and dynamic environment. Also discusses the expected challenges and upcoming trends in the context aware environments for the Internet of Things. IoT aims to integrate and collect the information from smart objects of various domains. IoT infrastructure is best suited for integration, collection, processing, transmission and delivery of context information. It combines context model with event based organisation of services. This paper insisting on IoT is the backbone for the development of many applications which includes people, things, mobility and governance. Opportunistic networks facilitate the mobile communication when things are unable to establish the communication or it is offloaded to handle with large throughputs. The exchange of data is for the users in a closed place and the mobility happens through the short range transmission protocols. To handle the dynamic environment, the mobile devices are working in store-carry-and forward paradigm. The contact acts as the opportunities for the data to move to the destination. In this network data distribution happens through publish/subscribe model. This paper discussed from IoT point of view and various opportunities analyzed using new categories.

*B. Fog Computing: A Platform For Internet Of Things And Analytics, Flavio Bonomi, Rodolfo Milito, Preethi Natarajan And Jiang Zhu,2014,Springer*

This paper proposed a hierarchical distributed architecture for IoT. Fog computing proposes a new breed of applications and services to have a productive interaction between existing cloud and Fog. Special focus given to Analytics and challenges of Big Data. Fog computing is the next level of the cloud computing and it uses common resources. A Smart Traffic Light System (STLS) and Wind form systems were taken as the use cases in Fog computing. The outcome of the use case studies discussed various attributes between Fog and Cloud. Requirement and necessity of Fog deeply analysed and discussed in the use cases. Relevance of Fog emphasized for IoT and Big Data. Technical requirement of Fog also discussed under the Fog architecture. Primary aim of this paper is how cloud computing can be extended into Fog.

C. *Using Materialized View As A Service Of Scallop4sc For Smart City Application Services Shintaro Yamamoto, Shinsuke Matsumoto,Sachio Saiki, And Masahide Nakamura Kobe University, 1-1 Rokkodai-Cho, Nada-Ku, Kobe, Hyogo 657-8501, Japan*

This paper focuses on developing materialized view architecture as a smart service for Smart city. In a Smart city environment, houses and infrastructure which are connected to smart city produces huge amount of data and it is a big challenge to process the data and to get the required information from available data. Big Data is required to meet the challenge. Processing everything with raw data will take enormous time. This paper proposed architecture named Materialized View as a Service. Within the abstract cloud service it encapsulates all the transactions related to materialized views. The architecture designed in data platform Scallop4SC. The architecture designed in MapReduce on Hadoop and HBase KVS. It takes care of the design and implementation part of house logs. So finally the MVaaS converts the house logs into application specific materialised view. Also the effectiveness of the system demonstrated in three case studies.

D. *Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared Disk Big Data Analytics With Apache Hadoop"*

This paper discusses the necessity of Big Data and Big Data techniques which is required to process huge amount of data and to discover insights. Hadoop is a open source platform used for implementing Mapreducer Model. The performance of VERITAS Storage Foundation Cluster File System ( SF CFS) is compared with Hadoop distributed file system (HDFS) for shared data Big Data analytics. Analytics with clustered file system is best suited for this proposed model.

E. *"Towards MapReduce Performance Optimization: A Look Into The Optimization Techniques In Apache Hadoop For Big Data Analytics" Kudakwashe Zvarevashe1, Dr. A Vinaya Babu*

Traditional database management system can't handle huge distributed, structured and unstructed data. Big Data plays a role in solving the issues of handling huge, complicated and dynamic data. Hadoop and NoSQL databases supported to eradicate these problems. Various technologies associated to MapReduce discussed in this paper. Difference research problems related to the improvement of the MapReduce problem discussed.

## IV. BIGDATA TECHNOLOGY

Big Data is not actually a new concept. Enterprises are having high volume of databases and data warehouses for many years. The difference is its size, how complicated it is and its speedier growth. It require new tools to handle the challenges. Traditional RDBMS is not sufficient to handle Big Data . It requires efficient and effective technology to process huge volume of data in an efficient manner. Modern technologies and latest cloud based applications required to overcome the limitations of traditional RDBMS. Facebook, Twitter, Google, Amazon , Linked in required latest database management technologies to handle dynamic and complicated datasets. NoSQL was initiated by these companies. NoSQL are essential for the Enterprises to handle huge dataset generated through Cloud computing, IoT, Big Data and Big Users.

NoSQL does not use SQL as a querying language but it is a type of database management system in a distributed architecture. This is not another RDBMS. NoSQL has following key properties[10].

Higher scalability
Ability of partitioning and distribution of data
Simplified protocols and interfaces
Query capabilities are low
Eventual consistency rather ACID property
Efficient storage management through distributed indexing
Dynamic addition of new attributes to the records.

Most of the Big Data tools available in the market are open source. Few important Big Data tools briefly explained below

*A. Big Data Analysis Platforms And Tools [26]*

*1) Hadoop And MapReduce :* This is one of the popularly used Big Data tool. Hadoop MapReduce is a Big Data programming model used for writing applications to process very huge amount of data in parallel on various clusters of commodity hardware in a reliable and fault tolerant manner. The scheduling, monitoring and re-execution of the failed tasks taken care by the master and the slave execute the tasks as per the direction of the master[100].

*2) Gridgain:* This is an alternative of Mapreduce and this also supports HDFS. This is used for fast analysis of real time data using in –memory processing.

*3) Hpcc:* It's expansion is High performance computing cluster. Both paid version and open source is available.

*4) Storm:* It works in many programming languages and owned by Twitter. It works under Linux operating system.

*B. Data Bases / Warhouses*

1) *Apache Cassandra:* This is another open source distributed database management system developed by Facebook. This is a high performance, scalability and high availability software. It has a good built in Cache.
2) *Apache HBase:* designed to run on the top of HDFS(Hadoop Distributed File system). It provides real time access to Hadoop and it provides distributed and scalable data set. It is modeled after Google's BigTable and it used Java for programming.
3) *MongoDB [18]:* MapReduce uses this for batch processing. It provides Query by field, Range and regular expression searches. It follows master slave model and the duplicate data is useful during hardware failure.
4) *Neo4j [21]:* It is a graph database model. Its speed is thousand times higher than Traditional DBMS. It works under REST interface or Java API.
5) *Apache CouchDB [17]:* It performs MapReduce queries thourgh JavaScript. It provides synchronization even in Smart Objects.
6) *Terrastore [26]:* This works in all the operating system. It is highly scalable and consistent.
7) *FlockDB[26]:* It is a graph oriented database and works in all Operting system
8) *RIAK [16]:* is another open source distributed key-value data store. It works with map/reduce, HTTP, REST and JSON.
9) *Hypertable [19]:* This is designed after Bigtable. It runs on the top of HDFS, GlusterFS, or the Kosmos File System (KFS). Its own querying language is HQL (Hypertable querying language)
10) Hive: Like Hypertable it uses its own querying language called HiveQL . This runs in all operating system. Hive is the Hadoop based data warehouse.

*C. Business Intelligence*

Several business intelligence tools available in the market for Big Data analytics. It provides insights from various data collected from various sources. Talend, Jaspersoft, Jedox, Pentaho, SpagoBI, Knime, BIRT are some of the popular BI tools used for Big Data.

*D. Data Mining*

The primary aim of the data mining it to derive the required information in an understandable format from the available data set. RapidMiner/RapidAnalytics, Mahout, Orange, Weka, jHepWork, KEEL, SPMF, Rattle are popularly available Data Mining tools.

*E. File Systems*

*1) Gluster:* is a file storage system for objects and large datasets. This can be used beyong the limitation of HDFS.

*2) HDFS :* The default file system of Hadoop is the Hadoop distributed file system . This java based file system is reliable and scalable. Useful for large datasets.

*F. Programming Lanugages*

*1) Apache Pig [24]:* Pig is a high level scripting language used generating MapReduce programs using Hadoop. Pig Latin is the textual language available in the language layer of Pig.

*2) R:* R Programming is very popular nowadays and it is a graphial oriented programming tool similar to S.

*3) ECL:* is the short form of Enterprise Control Language. It is a high level programming language.

## V. BIG DATA POTENTIALS

Eventhough the Big Data boom started few years ago, the opportunities are growing as the speed of data keeps growing. A global survey[28] conducted by McKinsey on Big Data to understand the innovation, competition, and productivity. The survey covered Healthcare, Public sector, Retail, Manufacturing, Telecommunications. In depth study carried out with the help of existing literature reviews and various interview with industry executives. The research conducted in Economics and management. The research focused on Productivity, Competitiveness and growth .The evolution of global financial market and the economic impact of technology. Following key domains will have the great opportunities.

### *A. Marketing*

Big Data automatically will not lead to better marketing. The deeper and richer insights derived from Big Data drives the success and to read the pulse of the customers. Proper analytics leads to the prediction of tomorrow's requirements by today's purchase.

### *B. Healthcare*

Large amount of data required for better analysis. Critical insights derived from clinical data will provide good care to the patients. Clinics can play a better role due to the availability of transparent and largely available information. Best practices to be deployed to meet the challenges and complete rethinking and change in IT structure required at the time of deployment. A big revolution is happening in the genetic field. New research direction arrived by Genome project. Big Data plays significant role in data storage, retrieval, sequence analysis and visualization.

### *C. Social Media*

Many companies are interested to understand the e-commerce transactions and social media postings to understand the public interest. Get valuable insights from the flooded data is today's challenge.

### *D. Automation*

Current emerging trend is collection of sensor data in the IoT environment. The immediate need is to store, manage and analyse the increasing data which comes via IoT.

### *E. Manufacturing Industries*

Manufacturing and IoT are interrelated because many companies are having automated machines in the production environment which generates more data . Big Data tools will help the manufacturing industries to Store , Retrieve and analyse the data.

### *F. Defence*

Information is an important treasure in arms race. Data received from satellites, aircraft and messages from various devices are important in the military tech.

### *G. Smart City*

Smart city is going to change our living environment and infrastructure. This will bring the IoT into reality by embedding advanced technology and data driven methods. Big companies like IBM and Cisco are working seriously to make it real.

## VI. BIG DATA CHALLENGES

The success of Big Data in the enterprises requires biggest cultural and technological change. Enterprise wise strategy required to derive the business value by integrating the available traditional data. Biggest challenges in Big Data analysis are Heterogeneity and Incompleteness, Scalability, Timeliness and security of the data. Privacy is one of the major concerns for the outsourced data. Policies to be deployed and rule violators to be identified to avoid the misuse of data.Data integrity are a challenge for the data available in cloud platform. Organisational leaders should take the initiative to understand and move torwards the Big Data. Skilled people required for the shift to Big Data. It requires people in the area of system analysis, domain knowledge, data analytics, database management and software developers.

Large number of open source technologies available in the market for Big Data. Few are discussed in the previous section. Selection of right tool is also a challenge. During selection of tools, the fitment with the existing traditional database to be

considered to provide valuable insights.

## VII. CONCLUSION

In this paper we have discussed about various survey papers related to Internet of Things, Cloud computing, Smart city, Hadoop and MapReduce. The outcome of the literature brings out the importance of Big Data and the requirements of change and adoption to latest technologies. Big Data databases ensure better performance than traditional RBDMS in various use cases. There are many open source software available in the market. But the choice of selecting best Big Data tool is a challenge for the programmers for developing efficient scalable application. Clear analysis required before selecting the tools from developer and users point of view. Most of the Big Data tools available in the market are open source. Various Big Data Analysis Platforms and Tools Data bases / warhouses , Business Intelligence, Data Mining, File systems and Programming languages were discussed under Big Data technologies. Opportunities available in various domains discussed under Big Data Potentials.

The biggest challenges in front of all the enterprises are the requirement of cultural and technological change to adopt the new technology. Valuable insights will be derived from available traditional data also. Organisational leaders should take the initiative to understand and move torwards the Big Data. Because it involves changes in all levels. Future research problems will promise the benefits of Big Data.

## REFERENCES

[1] Radu-Ioan,Ciobanu,Valentin Cristea, Ciprian Dobre and Florin Pop, Big Data Platforms for the Internet of Things,2014,Springer

[2] Flavio Bonomi, Rodolfo Milito, Preethi Natarajan and Jiang Zhu,Fog Computing: A Platform for Internet of Things and Analytics, springer (2014)

[3] Shintaro Yamamoto, Shinsuke Matsumoto,Sachio Saiki, and Masahide Nakamura Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan, Using Materialized View as a Service of Scallop4SC for Smart City Application Services (2014)

[4] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. "Shared disk big data analytics with Apache Hadoop" (18-22 Dec. 2012)

[5] Kudakwashe Zvarevashe1, Dr. A Vinaya Babu, Towards MapReduce Performance Optimization: A Look into the Optimization Techniques in Apache Hadoop for BigData Analytics (2014)

[6] Gartner: Hype cycle for big data, 2012. Technical report (2012)

[7] IBM, Zikopoulos, P., Eaton, C.:Understanding BigData: Analytics for Enterprise Class Hadoop and Streaming Data. 1st edn. McGraw-Hill Osborne Media, New York (2011)

[8] Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., Tufano, P.: Analytics: The realworld use of big data. IBM Institute for Business Value—executive report, IBM Institute for Business Value (2012)

[9] Evans, D.: The internet of things—howthe next evolution of the internet is changing everything. Technical report (2011)

[10] Cattell, R.: Scalable sql and nosql data stores. Technical report (2012)

[11] Apache: Hadoop (2014) (Online 20 Oct 2015)

[12] Jo Foley, M.: Microsoft drops dryad; puts its big-data bets on hadoop. Technical report (2011)

[13] Locatelli, O.: Extending nosql to handle relations in a scalable way models and evaluation framework (2012012)

[14] Robinson, I., Webber, J., Eifrem, E.: Graph Databases. O'Reilly Media, Incorporated (2013)

[15] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., Vogels,W.: Dynamo: amazon's highly available key-value store. SIGOPS Oper. Syst. Rev. 41, 205–220 (2007) Big Data Management Systems for the Exploitation 89

[16] Riak: Riak (Online Oct 2015)

[17] Apache: Couchdb (Online; Oct 2015)

[18] MongoDB: Mongodb (Online; Oct 2015)

[19] Hypertable: Hypertable (Online; Oct 2015)

[20] Rabl, T., Gómez-Villamor, S., Sadoghi, M., Muntés-Mulero, V., Jacobsen, H.A., Mankovskii, S.: Solving big data challenges for enterprise application performance management. Proc. VLDB Endow. 5, 1724–1735 (2012)

[21] Neo Technology, I.: Neo4j, the world's leading graph database. (Online;Oct 2015)

[22] Amato, A., DiMartino, B., Venticinque, S.: Semantically augmented exploitation of pervasive environments by intelligent agents. In: ISPA, pp. 807–814. (2012)

[23] Jing Zhang, "A Distributed Cache for Hadoop File Distribution system in Real time Cloud Services ", 2012 ACM/IEEE 13th International Conference on Grid Computing.

[24] Pig.apachi.org (online Oct 2015).

[25] http://www.concurrentinc.com/2014/05/cascading-3-0-adds-support-for-wide-range-of-computational-frameworks-and-data-fabrics/

[26] http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.html

Volume 3 Issue XII, December 2015
ISSN: 2321-9653


[27] http://www.thesojo.net/key-domains-with-opportunities-in-big-data/

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