International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 4 Issue 12, December-2015

Literature Review: Big Data Tools and Techniques and its Opportunities

Meenakshi

ABSTRACT

Huge changes are happening in Cloud Computing, Big Data, Communication technology and Internet of things in past few years. New challenges come with adapting of latest technologies. 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. Views of various researchers, Big Data tools and techniques required for storage, management and analytic and its growth and suspected challenges in various domains is discussed in this paper.

I. INTRODUCTION

During the past few years, Several changes are happening in the IT field especially in the area of Big data, Cloud computing, mobility and Internet of things(IoT), which creates a new platform for the enterprises to enter into new business. Vast amount of data produced significantly in the past two years due to internet and social media penetration. Today's 90% of the data created within last two years of time, Everyday's data generation exceeds 2.5 quintillion bytes of data. This enormous speed of data growth is due to the services and users who are producing vast amounts of data.

Big Data becomes very important in this context which is making possible to turn into this amount of data in information, decision making, knowledge and, ultimately insights. Gartner defines Big Data as velocity, high volume 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].

3Vs are 3 categories of Big data:

- Volume (data size and number)
- Velocity (the rate at which data are generated or need to the processed)
- Variety (different types / different forms of content) [7].

As Big Data are very large and complex so only with traditional approaches it is difficult to process and analyze the data. Effective data management for Big Data sets is not possible with traditional RDBMS (Relational database management systems). It is very difficult to extract the information in a proper and required manner due to the size of Big Data.

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 analyses and visualize various formats of data in different industries and various domains.

In this paper, Section III deals about the technical facts of Big Data and the main aim of this paper is to provide technical aspects of Big Data and Big Data challenges in Internet of things (IoT).

.

II. CHALLENGES

The main challenges in the business environment are to combine structured and unstructured formats from various applications and projects into single format and it is being handled by Big Data very well and it also provides valuable insights.

For any organization to move forward, it is very important to impact right technology at right time. 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. So this is the right time to move towards Big Data technology.

III. RELATED WORK

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

[1]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., Also, focuses on the improvement the performance of the old services and offer new services in an open and dynamic environment and from IoT point of view and various opportunities analyzed using new categories.

[2] Hierarchical distributed architecture for IoT is proposed in this paper and Special focus given to Analytics and challenges of Big Data. The next level of the cloud computing is Fog computing and it uses common resources. Fog computing uses a Smart Traffic Light System and Wind form systems. The outcome of the use case studies discussed various attributes between Fog and Cloud. Primary aim of this paper is how cloud computing can be extended into Fog.

[3] Developing materialized view architecture as a smart service for Smart city is focused in this study. 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. To meet the challenge, Big Data is required. Processing everything with raw data will take enormous time. This paper proposed architecture named Materialized View as a Service.

[4] Necessity of Big Data and Big Data techniques which is required to process huge amount of data and to discover insights is discussed in this study. For implementing of Mapreducer Model, Hadoop is a open source platform which is used. 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.

[5] Huge distributed, structured and unstructured data can't be handled by traditional database management itself. Big Data plays a role in solving the issues of handling huge, complicated and dynamic data. It is supported by Hadoop and NoSQL to eradicate these problems. Various technologies associated to MapReduce discussed in this paper.

IV. BIG DATA TOOLS

Actually Big data is not a new concept. It is only differentiated by its size, how complicated it is and its fast growth. To handle the challenges it required new tools. 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. Applications like Google, Facebook, Amazon, Twitter, Linked in required latest database management technologies to handle dynamic and complicated datasets. These companies initiated NoSQL which are essential for the Enterprises to handle huge dataset generated through Cloud computing, IoT, Big Data and Big Users.

NoSQL has following key properties[10].

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

Open source big data are mostly available in market. Few important Big Data tools briefly explained below:

A. Big Data Analysis Platforms [26]

1) ***Hadoop And MapReduce :*** One of the popularly used Big Data tool and it 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: Gridgain is an alternative of Mapreduce which also supports HDFS. For fast analysis of real time data using in –memory processing it is used.

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

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

B. Database Ecosystem

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.

V. Big Data Oppurtunities

Even though the Big Data boom started few years ago, the opportunities are growing as the speed of data keeps growing. 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. 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.

C. Automation

In the IoT environment current emerging trend is collection of sensor data. The immediate need is to store, manage and analyze the increasing data which comes via IoT.

D. Manufacturing Industries

Big Data tools will help the manufacturing industries to Store , Retrieve and analyse the data.

E. Defence

In arms race information is an important treasure. Data received from aircrafts, satellites, and messages from various devices are important in the military technology.

F. Smart City

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

CONCLUSION

We have discussed about various survey papers related to Cloud Computing, IoT, Smart city, Hadoop and Map Reduce in this paper. The literature brings out the conclusion that importance of Big Data and the requirements of change and adoption to latest technologies is important. Requirement of cultural and technological change to adopt the new technology is the biggest challenges in front of all the enterprises are the. Valuable insights will be derived from available traditional data also. Initiatives to be taken by organizational leaders to understand and move towards 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
[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,