An Outlook on Big Data and Big Data Analytics

Swapnil Lokhande
Student of Master of Technology in (Computer Science)
International Institute of Information Technology, Bhubaneswar, India.

Nilay Khare, PhD
Associate Professor Computer Science Department, Maulana Azad National Institute of Technology, Bhopal, India.

ABSTRACT
The rapid growth in the field of technology in the last 10 years has generated a large amount of data which is structured, semi structured, as well as unstructured in nature. This large amount of data has been generated by the excessive use of the Internet, the growing popularity of the social networking sites like Facebook, Twitter, Quora, Linkedin etc and there are many more reasons behind this abrupt growth of data. The emerging trend of E-commerce in last few years has fastened the rate of growth of the data. The data has been growing rapidly in Volume, Variety, Velocity and Veracity. This emerging trend coined the term Big Data. The Big Data is often unstructured and real-time analysis is required to analyze and process the data. This evaluation calls for the new system architecture for the storage, transmission, analysis and processing of the data. In this paper we present a literature survey of the Big Data and Big Data analytics. The paper is divided into various sections which comprises of the definition of Big Data and its growing importance, Big Data challenges and Big Data Analytics. We also present a detailed survey of analytics platform like Apache Hadoop. Finally we delineate the evaluation and research direction for the Big Data System.

Keywords
Big Data, Big Data Analytics, stream processing, batch processing, Apache Hadoop.

1. INTRODUCTION
In the era of Information Technology, the Internet provides a large space to add information every day. The internet launched in 1984 firstly linked 1,000 hosts at university and corporate labs. As stated by The Incredible Growth of Web Usage infographic from WhoIsHostingThis.com, it took 15 years for the internet to connect to 50 million users in 1998. Eleven years later, in 2009, there were 1 billion internet users around the world. Three years later, it doubled to over 2.1 billion users, and by 2013, 39% of the world's population is connected through the internet (2.7 billion people) [1].

Google was founded in September 1998, at that time it served ten thousand search queries per day (by the end of 2006 that same amount of queries would be served in a single second). In September 1999, one year after being launched, Google already started answering 3.5 million search queries daily [2].

Now a days social networking sites are emerging speedily. Every minute Facebook users share nearly 2.5 million chunks of content, Twitter users tweet nearly 300,000 times, Instagram users post nearly 220,000 new photos, YouTube users upload 72 hours of new video content, Apple users download nearly 50,000 apps, Email users send over 200 million messages, Amazon generates over $80,000 in online sales [1].

An IDC report predicts that, from 2005 to 2020, the global data volume will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, representing a double growth every two years [3].

This scenario leads to the term "Big Data". The term Big Data is used to delineate a mammoth volume of both structured as well as unstructured data and the abundance of the data make it even more difficult to process it using traditional database and software techniques. As it can be seen from the trend that there is a huge potential associated with Big Data, it attracted the researchers from diverse domain.

In this paper we have presented a literature survey on Big Data Analytics and different outlooks on its use. In section II we define the term Big Data and the history associated with it. Section III delineate about the big data paradigm. Then section IV gives an insight about the importance of Big Data Analytics and different areas in which it can be productive. The different areas in which there is an application of big data analytics is covered in section V. Section VI delve in for the challenges associated with the Big Data Analytics. Section VII introduces Apache Hadoop, which is the current foundation of the Big Data Analytics software. The brief inference with the exhortation for future studies is stated in section VIII.

2. BIG DATA CONCEPT
Big Data is a loosely defined term used to describe datasets which are immense and intricate that it became awkward to work with using standard statistical software. The rise of digital and mobile communication has made the world become more connected, networked, and traceable and has typically lead to the availability of such large scale datasets [4]. The big data can be viewed in different ways. One such view is showcased in 2011 Mckinsey's report [5]. It states that- "Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don't define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). By this definition we can infer that the big data is not subjective to the amount of data only, for a data to be referred to as big data, other aspects of the data are also need to be considered.

IBM highlighted these aspects of the data in its report which need to be delved:

Volume: refers to the quantity of data gathered by a company. This data must be used further to obtain important knowledge.

Velocity: refers to the time in which Big Data can be processed. Some activities are very important and need

immediate response, that is why fast processing maximizes efficiency.

Variety: Refers to the type of data that Big Data can comprise. This data can be structured as well as unstructured.

Veracity: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future [6].

The above mentioned definition of big data provides tools for to the comparison between traditional data and big data. This comparison is summarized in Table 1.

First, the sheer volume of datasets is a critical factor for discriminating between big data and traditional data. Second, big data comes in three flavors: structured, semi-structured and unstructured. Traditional data are typically structured and can thus be easily tagged and stored. However, the vast majority of today's data, from sources such as Facebook, Twitter, YouTube and other user-generated content, are unstructured. Third, the velocity of big data means that datasets must be analyzed at a rate that matches the speed of data production.

Table 1- Comparison between Traditional data and Big Data.

Characteristics	Traditional Data	Big Data
Volume	GB	Constantly updated
Structure	Structured	Semi-structured or unstructured
Rate of generation	Per hour, day	More rapid
Data Source	Centralized	Fully Distributed
Data Storage	RDBMS	HDFS, NoSQL
Access	Interactive	Batch or near real-time

Finally, by exploiting a variety of mining methods to analyze big datasets, significant value can be derived from a huge volume of data with a low value density in the form of deep insight or commercial benefits [7].

3. BIG DATA PARADIGM

According to processing time requirement, big data analytics can be categorized into two paradigms.

Streaming Processing- The objective is to extract actionable intelligence as streaming analytics, and to react to operational exceptions through real-time alerts and automated actions in order to correct or avert the problem. Data are typically unstructured log records and sensor events, with each record including a timestamp indicating the exact time of data creation or arrival [8].

Stream processing is designed to analyze and act on real-time streaming data, using "continuous queries" (i.e. SQL-type queries that operate over time and buffer windows). Essential to stream processing is Streaming Analytics, or the ability to continuously calculate mathematical or statistical analytics on the fly within the stream. Stream processing solutions are designed to handle high volume data in real time with a scalable, highly available and fault tolerant architecture. This enables analysis of data in motion [9].

Storm [10] and Kafka [11] are the future of stream processing, and they are already in use at a number of high-profile companies including Groupon, Alibaba, The Weather Channel, and many more.

Batch Processing- In the batch-processing paradigm, data are first stored and then analyzed. MapReduce has become the dominant batch-processing model. The core idea of MapReduce is that data are first divided into small chunks. Next, these chunks are processed in parallel and in a distributed manner to generate intermediate results. The final result is derived by aggregating all the intermediate results [12].

Big data platform can use alternative processing paradigms, though the difference between the two can be discern by the architectural distinctions in the associated platform.

4. IMPORTANCE OF BIG DATA ANALYTICS

The importance of big data is in the potential to improve efficiency in the use of large data, storing and managing the bulk of unstructured data and processing the large data in real-time. If the big data is implemented and used accordingly, companies can efficiently improve their business and can get a better view on their business leading to efficiency in different areas like sales, improving the manufactured product and can also analyze the trend of the market which helps the company to develop the business strategy accordingly. Following are the ways [5] in which big data can offer potential transformation to create value and have suggestions for how the companies have to be designed, organized and managed.

Creating Transparency

Making the big data easily available to admissible stakeholders in a timely manner can create immense value.

Enabling experimentation to discover needs, expose variability, and improve performance

As the companies create and store more transactional data in digital form, organizations can collect more accurate and detailed performance data (in real or near real time) on everything from product inventories to personnel sick days. Big data allows organizations to create highly specific segmentations and to tailor products and services precisely to meet those needs.

Replacing/supporting human decision making with automated algorithms

Sophisticated analytics can substantially improve decision making, minimize risks, and excavate valuable insights that would otherwise remain obscure.

Innovating new business models, products, and services

The emergence of real-time location data has created an entirely new set of location-based services from navigation to pricing property and casualty insurance based on where, and how, people drive their cars.

5. APPLICATIONS OF BIG DATA ANALYTICS

The importance of big data analytics is growing in many fields. There is no question that organizations are dealing with an expanding data that is either too voluminous or too unstructured to be managed and analyzed through traditional means. Among its rapidly growing sources are the clickstream data from the Web, social media content (tweets, blogs, Facebook wall postings, etc.) and video data from retail and other settings and from video entertainment. But big data also enclose everything from call center voice data to genomic and proteomic data from biological research and medicine [13].

The efficient implementation of big data analytics in the following areas:

- ***Government***

The implementation and possession of Big Data within governmental processes is valuable and provide efficiencies in terms of cost, productivity, and innovation.

United States of America

a. In 2012, the Obama administration announced the Big Data Research and Development Initiative, to inspect how big data could be made useful to address important issues faced by the government [14]. The initiative is composed of 84 different big data programs spread across six departments [15].
b. Big data analysis played a tremendous role in Barack Obama's successful 2012 re-election campaign [16].

India

a. Big data analysis was, in parts, responsible for a highly successful win of the BJP and its allies in Indian General Election 2014 [17].

- ***Manufacturing***
  Based on TCS 2013 Global Trend Study, big data analysis provides great benefits in the improvements in supply planning and product quality for manufacturing [18].
- ***Media***
  Big Data and the Internet of Things work in conjunction. From a media point of view, data is the key imitative of device inter connectivity and allows accurate targeting [19].
- ***Technology***
  a. eBay.com uses two data warehouses at 7.5 petabytes(PB) and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising.
  b. Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. Amazon uses Linux as the core technology that keeps it running and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 terabyte (TB), 18.5 TB, and 24.7 TB [20].
- ***Private Sector***
  a. ***Retail***
  Walmart executes more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress [21].
  b. ***Real estate***
  Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day [22].

There are many more fields in which big data analytics has been successfully implemented.

6. CHALLENGES OF BIG DATA ANALYTICS

Designing and exploiting the big data system is an intricate task. As suggested by one of its definition, big data is beyond the potential of the current hardware and software platform. In our paper we have classified the challenges [7], [23] associated with big data in three categories: data acquisition and management, data analytics and system issues.

- Challenges associated with data acquisition and management:
  a. ***Redundancy Reduction and Data Compression-*** A large amount of redundant data is present in raw datasets. Redundancy reduction and data compression without sacrificing potential value are efficient ways to lessen overall system overhead and increase the overall system throughput.
  b. ***Privacy and security-*** These are also important challenges for Big Data. Data present is abundant and intricate, thus it is very difficult for a company to sort this data on privacy levels and apply the according security.
- Challenges associated with data analyst:
  a. ***Understanding of Big Data-*** In order to determine the best strategy for a company it is essential that the data that you are counting on must be properly analyzed.
  b. ***Approximate Analytics-*** As datasets are growing and the real-time requirement becomes prominent, analysis of the entire dataset is not feasible. One way to inquisitively solve this problem is to provide approximate results, such as by means of an approximation query.
- Challenges associated with large scale data systems:
  a. ***Scalability-*** Big data analytics systems must be capable of scaling up and down according to the size of the complex dataset.
  b. ***New Technologies-*** Considering the fact that Big Data is new to the organizations, it is necessary for these organizations to learn the new developed technologies. This is an important aspect which is going to bring competitive advantage to the organization.

There are many more challenges which the organizations come across in order to efficiently manage, analyze and process big data.

7. BIG DATA ANALYTICS SOFTWARE

Apache Hadoop [24] is a fast-growing big-data processing platform defined as "an open source software project that enables the distributed processing of large data sets across clusters of commodity servers". The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. Hadoop has many advantages, and the following features make Hadoop particularly suitable for big data management and analysis:

- ***Scalability***: Hadoop permits hardware infrastructure to be scaled up and down without changing data formats. The system will automatically redistribute data and computation jobs to accommodate hardware changes.
- ***Cost Efficiency***: Hadoop comes up with massively parallel computation to commodity servers, leading to a considerable decrease in cost per terabyte of storage, which makes immensely parallel computation reasonable for the emerging volume of big data.
- ***Flexibility***: Hadoop is schema-less and able to immerse any type of data from any number of sources. Moreover, different types of data from multiple sources can be append and aggregated in Hadoop for insight analysis.
- ***Fault tolerance***: Missing data and computation failures occur usually in big data analytics. Hadoop can retrieve the data and computation failures caused by node breakdown or network congestion.

7.1 Apache Hadoop Architecture

The Apache Hadoop architecture (shown in fig. 1) [7] is a massive computing framework consisting of several modules, including HDFS, Hadoop MapReduce, HBase, and ZooKeeper**.**

Hadoop HDFS and HBase are responsible for data storage. HDFS is a distributed file system developed to run on commodity hardware that references the GFS design. An HDFS cluster consists of a single NameNode that manages the file system metadata, and collections of DataNodes that store the actual data. A file is split into one or more blocks, and these blocks are stored in a set of DataNodes. Apache HBase is a column-oriented store modeled after Google's Bigtable. HBase can serve both as the input and the output for MapReduce jobs run in Hadoop and may be accessed through Java API, REST, Avor or Thrift APIs.

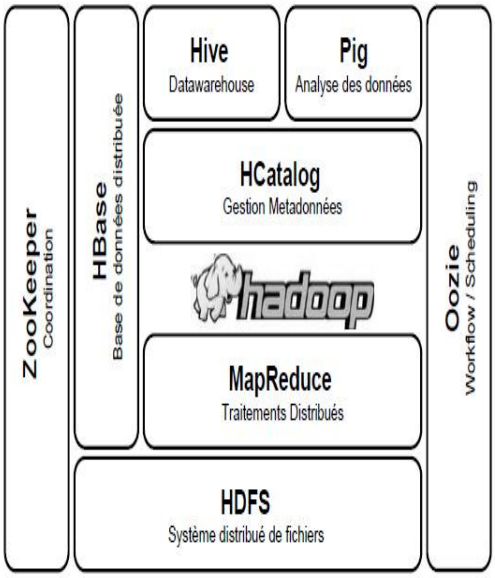

Figure 1- The Apache Hadoop Architecture

Hadoop MapReduce is the computation core for massive data analysis and is also modeled after Google's MapReduce. The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster node. The master is responsible for scheduling jobs for the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master. The MapReduce framework and HDFS run on the same set of nodes, which allows tasks to be scheduled on the nodes in which data are already present.

Pig Latin and Hive are two SQL-like high-level declarative languages that express large data set analysis tasks in MapReduce programs. Pig Latin is suitable for data flow tasks and can produce sequences of MapReduce programs, whereas Hive facilitates easy data summarization and ad hoc queries.

Zookeeper is a centralized service for maintaining configuration, naming, providing distributed synchronization, and providing group services. Whereas Oozie is a workflow scheduler system to manage Apache Hadoop jobs. Oozie workflow jobs are Directed Acyclic Graphs (DAG) of actions.

8. CONCLUSION AND FUTURE RESEARCH

This literature survey delineates the concept of Big data and Big data analytics and helps to get acquainted with the new concept and different technologies which are available through which the organizations can develop a platform for the efficient analysis of their business. This paper covers the detailed study of the concept, how the term "Big Data" coined, importance of big data, challenges faced in the analysis of big data and the intricate architecture of Apache Hadoop.

In future there would be a great demand and requirement of real-time analysis. Our survey suggests that data would be required to analyze and process in accordance with the rate of its generation. Storm is a big data analytics software which can be used for real-time analysis. The efficient use of Strom could leads to the rapid analysis of fast growing data. Further researches can be done in the field of weather forecasting, stock exchange market and analysis of the trend on Twitter where there is an ample scope of real-time analysis.

9. REFERENCES

[1] Susan Gunelius, "The Data Explosion in 2014 Minute by Minute– Infographic, July 2014. Available : http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/.

[2] Available: http://www.internetlivestats.com/google-search-statistics/#trend.

[3] J. Gantz and D. Reinsel, ``The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," in Proc. IDC iView, IDC Anal. Future, 2012.

[4] Chris Snijders, Uwe Matzat1, Ulf-Dietrich Reips, ""Big Data": Big Gaps of Knowledge in the Field of Internet Science", International Journal of Internet Science 2012, 7 (1), 1–5, ISSN 1662-5544.

[5] 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, May 2011.

[6] P. Zikipoulos, T. Deutsch, D. Deroos, Harness the Power of Big Data, 2012. Available: http://www.ibmbigdatahub.com/blog/harness-power-big-data-book-excerpt

[7] HAN HU, YONGGANG WEN, TAT-SENG CHUA, AND XUELONG LI, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", volume 2, 2014.

[8] Available: http://www.sqlstream.com/stream-processing/.

[9] "Real-Time Stream Processing as Game Changer in a Big Data World with Hadoop and Data Warehouse".

[10] Available: http://storm-project.net/.

[11] Available: http://kafka.apache.org/design.html.

[12] J. Dean and S. Ghemawat, ``Mapreduce: Simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1, pp. 107-113, 2008.

[13] Available: http://sloanreview.mit.edu/article/how-big-data-is-different/.

[14] Kalil, Tom. "Big Data is a Big Deal". White House. Retrieved 26 September 2012.

[15] Executive Office of the President (March 2012). "Big Data Across the Federal Government". White House.

[16] Lampitt, Andrew. "The real story of how big data analytics helped Obama win". Infoworld.

[17] "News: Live Mint". Are Indian companies making enough sense of Big Data?. Live Mint - http://www.livemint.com/.

[18] "Manufacturing: Big Data Benefits and Challenges". TCS Big Data Study. Mumbai, India: Tata Consultancy Services Limited.

[19] Couldry, Nick; Turow, Joseph (2014). "Advertising, Big Data, and the Clearance of the Public Realm: Marketers' New Approaches to the Content Subsidy". International Journal of Communication.

[20] Layton, Julia. "Amazon Technology". Money.howstuffworks.com.

[21] "Data, data everywhere". The Economist, 25 February 2010.

[22] Wingfield, Nick (2013-03-12). "Predicting Commutes More Accurately for Would-Be Home Buyers - NYTimes.com".

[23] Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU, "Perspectives on Big Data and Big Data Analytics", Database Systems Journal vol. III, no. 4/2012

[24] What Is Apache Hadoop? Available: https://hadoop.apache.org/.