A Survey on Big Data Analytics

Pallavi Dubey, Prof. Manaswini Panigrahi

Department of Computer Science and Engineering, IES Group of Institute, Bhopal, India

Abstract - We are living in highly digital universe in which demands are very high so to fulfill these demands Organizations, Institution, Machines generates that highly proliferative data. This data is known as big data. Big data is the term which defines three characteristics volume, velocity, and veracity. Most of the data are generated in the form of unstructured and semi-structured form. Data streams generate from these devices are the challenges for tradidational approaches for data management and processing on these data. The volume of data and heterogeneity of data with high generation speed makes it difficult for present computing approaches to manage big data. Thus to resolve these problems big data tools are used to handle these data for storing and processing.

Keyword: Big Data, Big Data Analytic Tools.

I. INTRODUCTION

Big data [1] analysis is one of the major challenges of our era. Referring the huge amount of data, big data is growing exponential rate. The limits to what can be done are often times due to how much data can be processed in a given time slots. Big data is usually transformed by 3V's i.e. volume, velocity and veracity. Volume refers to the huge amount of data in TB, PB and so on which are generated from social media, organizations and many other institutes. Velocity refers to the Frequency of data generated or frequency of data delivery while Veracity refers to inherent unpredictability of some data which requires for the analysis of big data to gain reliable prediction [2]. Big data technologies describe a new generation of technologies and architectures, which are designed for extracting value from large volumes of a wide variety of the data by enabling the high-velocity, discovery and analysis [3].The growth of the Big data depends on the increase in the storage capabilities, increase in processing power and availability of data. Data are classified into three categories they are structured, unstructured and semi-structured form. The term Structured Data generally refers to data that has a defined their format .Example of structured data which include numbers, dates, and groups of words and numbers called string which are generated from sensor data, web log data, and financial data and so on. Structured data generally resides in a relational database in the form of tables which consist of rows and columns; it is also known as relational data. This type of data can be easily mapped into pre-defined fields' .While unstructured data is not relational and doesn't fit into these pre-defined data models. The third type of data is semi-structured data, it is information that doesn't reside in relational database but that does have some organizational properties that make it easier to analyze such as

XML document or log files and so on [4]. Big data refers to the larger datasets that are difficult to analyze by the traditional tools .Big data can consists both structured and unstructured data but many organization estimates that 90 percent of big data is unstructured data.

As we known to that there are many problem are arise to handle big data such as storing due to most of the data in the form of unstructured or semi-structured and processing these data. Thus to overcome these problem frameworks are used for storing and processing big data is popularly known as Hadoop. In this paper we discuss about Applications of big data, problem arises in big data and how to handle these problems using Big data analytics frameworks.

II. PROBLEM IN HANDLING BIG DATA

Heterogeneity, scalability, complexity, and securities problems with Big Data hamper the progress at all stages of the process that can generate value from the data. Much data today is not natively in structured format, such as tweets and blogs are structured pieces of text, while video and images are structured for storage and display, but not reliable for semantic content and search, transformation of such data content into a structured format for later analysis is a major challenge [5]. The value of data that increases when it can be linked with other data, thus data integration is a major creator of the value. Now a day, most of the data is directly generated in digital format, thus we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. There are many challenges such as data analysis, organization, retrieval, and modeling. Big Data analysis is a bottleneck in many applications, due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed[6].Some other semantic challenge for to extract the meaning of the information from massive volumes of unstructured data are discussed below.

A. *High volume of processing using low power consumed by the digital processing architecture:*

Power that needs to process the data as well as power that used to required the cooling of processing system need to be considered together in designing such systems.

B. Designing scalable data storages that provides efficient data mining.

Storing any kind of the data is useless or unless they cannot be retrieved and extract the knowledge efficiently.

C. Data security and privacy

This is about managing access control to the big data.

D. Approximate results

Due to the volume and the velocity of big data, approximate results would be order of magnitude faster compared to traditional query execution. It need be decided where and when to use approximation where it will not harm the accuracy of the results or decision that emerged based on approximated results.

E. Data exploration to enable deep analytics

Certainly, new technologies are required to analyse large volumes of data faster with efficient resource and power consumptions.

F. Enterprise data enrichment with web and social media

The big data has more relationships among themselves that ever in history. This is also related to the linked data and social media. Real value of big data will be merged once when we are able to preserver and identify the relationships between data.

G. Query optimization

In order to harvest the knowledge hidden in big data optimized query processing is an essential step. Optimization need to consider in different perspectives such as energy consumption, memory required, processing time and storage requirement, etc. Parallel processing would be the key in query processing in cloud environments.

III. BIG DATA ANALYTIC TOOLS

The need for efficient and scalable outcome which support partial component failures and provide data consistency. The Big Data analysis tools which are used for the efficient and precise data analysis and management for these data. Following are the tools which are used to handle these problems which are discuss below.

3.1 Apache Hadoop: Apache Hadoop [7] is an open source framework for the storing and processing the large datasets using clusters of commodity hardware. Hadoop is designed to scale up the nodes. The various components of a Hadoop are shown in Figure 1. The Hadoop framework contains two major components: they are Hadoop Distributed File System anf Hadoop YARN. Hadoop Distributed File System (HDFS) [8] is a distributed file system that is used to store data across cluster of commodity machines which provides high availability of data and fault tolerance. Hadoop YARN is a resource management which schedules the jobs across the cluster.

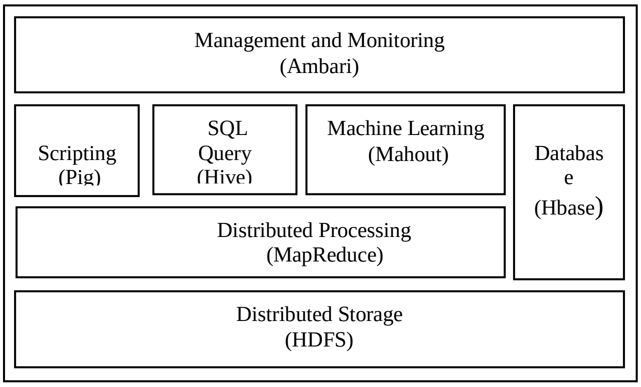

Figure 1: Apache Hadoop Ecosystem

A. MapReduce:

The programming model used in Hadoop is MapReduce [9]. It was proposed at Google by Dean and Ghemawat. MapReduce is the basic data processing unit used in Hadoop framework which breaks the entire task into two parts, known as mappers and reducers.

B. Apache Pig :

Apache Pig [10] is a SQL like environment developed at Yahoo, it used by many organizations such as Yahoo, Twitter, and LinkedIn etc.

C. Hives:

Hive is another MapReduce wrapper developed by Facebook. These have two wrappers provide a better environment for processing and make the code development simpler thus the programmers do not have to deal with the complexities of MapReduce coding. Hive converts HiveQL queries into MapReduce jobs for execution on a cluster.

D. Mohout:

Mahout is an open source machine learning library built on Hadoop which provide distributed analytics capabilities. Mahout provides a wide range of data mining techniques including effective filtering, classification and clustering algorithms.

E. Hbase:

HBase is a distributed database that resides on top of the Hadoop Distributed File System, which providing real-time read/write random-access to large datasets. Hive defines a SQL-like query language which is called as HiveQL, that reduce the complexity of writing MapReduce jobs from the users.

3.2 Spark: Spark is a next generation paradigm for big data processing It was developed at Berkeley at the University of California . It is an alternative of Hadoop framework. The main aim is to designed to overcome the disk I/O limitations and improve the performance of the systems. Spark is designed to support in-memory processing .The main benefit of keeping everything in memory is the ability to perform iterative computation at very fast speeds. Spark permit programmer to write applications in Java, Python and to build parallel application designed to take full advantage of a distributed environment along with supporting map and reduce operation. Spark also supports the SQL queries, data streaming, and complex analytics such as machine learning and graph algorithm. Spark runs on existing Hadoop clusters and is compatible with HDFS, HBase and any Hadoop storage system, the users can combine all these capabilities into a single workflow while accessing and processing data in the Hadoop environments. The major feature of Spark is that it allows the data to be cached in memory, thus eliminating the Hadoop's disk overhead limitation for iterative tasks. The components of Spark Ecosystem are discussed in Figure 2.

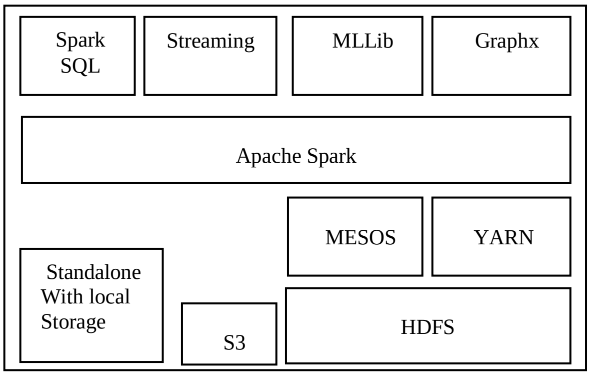

Figure. 2 Spark Ecosystems

A. Spark SQL :

Spark SQL is a component on top of Spark framework that introduces a new data abstraction called Data Frames, which provides support for structured and semi-structured data both. Spark SQL provides domain-specific languages to manipulate Data Frames in Java, Python.

B. Spark Streaming:

Spark Streaming component of Spark Core provides faster scheduling capabilities to perform streaming analytic to convert data into sub batches and perform RDD(Resilient Distributed Datasets) transformation on these batches of data.

C. Machine Learning Library:

Spark MLLib is a distributed machine learning framework on the top of Spark.

D. Graphx:

Graphx is a distributed graph processing framework on the top of Spark Core.It proveides an API for expressing graph computation .It also provides an optimized runtime for the abstraction.

Apache Hadoop and Spark is both popular framework in the big data analytics. Apache Spark is an improvement on the original Hadoop MapReduce component of the Hadoop framework used for big data analysis. There are many advantages in Apache Spark as it provides benefits in interactive data interrogation on in memory data set and also in multi-pass iterative machine learning algorithms. We elaborate a detailed discussion on Spark Hadoop comparisons in Table 1.

Table.1 Comparison Between Hadoop And Spark

Parameter	Description
Faster	Spark execute batch processing jobs ,thus it can more faster than the Hadoop framework. Spark divides data into sub-batches. MapReduce does not leverage the memory of the Hadoop cluster to the maximum. While Spark use the concept of RDDs in which to save data on memory and preserve it on the disc.. Thus the general execution engine of Spark is much faster than Hadoop MapReduce with the use of memory.
Easy Management	In Spark framework it is possible to perform batch processing and machine learning in the same cluster in a same time while in Hadoop framework not possible. In Spark it is possible to control different kinds of workloads, so that if there is an interaction between various workloads in the same process it can easily managed and secure such workloads which come as a limitation with MapReduce
Caching	Spark ensures lower latency computations by using cache the partial results across its memory of distributed workers while MapReduce are disk oriented completely.
Recovery	RDD is the main abstraction in the Spark. It allows recovery of failed nodes by re-computation. It also support more similar recovery pattern to Hadoop which used to way of check pointing, and to reduce the dependencies of an RDD.
Failure Tolerance	Spark has retries per job and speculative execution such as MapReduce in Hadoop. If the process crashes in the middle of execution for any reason in MapReduce, it could be continue where it left off, whereas Spark will have to start processing from the beginning.

IV. APPLICATIONS OF BIG DATA

Big data refers the ability to collect and analyze the huge amounts of data which are generating in the real world .Now a day's every aspect of lives are affected by the big data. However, there are some areas where big data is already making a real difference today [11].The most widespread use as well as the highest benefits of big data discussed as follows.

A. Understanding and Targeting the Customers

It is one of the biggest and most famous areas of big data. Big data is used for the better understanding the customers and their behaviors and preferences. Companies are expanding their traditional data sets with the social media data, and browser logs and text analytics and sensor data for to complete the customer criteria. The main objective is to create predictive models.

B. Understanding and Optimizing Business Processes

Big data is also used to optimize the business processes. Retailers are able to optimize their stock based on the predictions which are generated from social media data, web search and weather forecasts. Thus the business process used a lot of big data analytics is to applied on the chain or delivery route optimization. By the use of big data geographic positioning and radio frequency identification are also used to track goods or delivery vehicles or to optimize the routes by integrating live traffic data, etc.

C. *Personal Quantification and Performance Optimization*

Big data is not only used for companies and governments but also for all of us individually. Now we can benefit from the data generated from devices such as smart watches or smart bracelets. Examples are taking the up band from Jawbone; the armband collects data on our calorie consumption, activity levels, and our sleep patterns. Thus it gives individuals rich insights; the real value is in analyzing from the collective data. In above case, the companies' now collecting earlier year's data of worth of sleep data every night. Analyzing such volumes of data will bring entirely new insights that it can feedback to individual users. The other area where we have many benefit from big data analytics. Most online sites are also apply big data tools and algorithms to find us the most appropriate matches.

D. Improving Healthcare and Public Health by using Big Data

The computing power of big data analytics enables us to decode entire DNA in a minutes and will allow us to find new cures and better understand and predict disease patterns. Big data techniques are already being used to monitor the babies premature babies. By recording and analyzing every heart beat and breathing pattern of every baby, thus the unit was able to develop algorithms that can predict infections at an hours before any physical symptoms appear. The team can intervene early and save fragile babies in an environment where every hour counts. Big data analytics allow us to monitoring and predict the developments of the epidemics and about disease. Integrating data for the medical records from the social media analytics which enables us to monitor outbreaks in real-time.

E. Improving Sports Performance

Most of the sports have embraced big data analytics. There are many tools which is used in sport such as IBM Slam Tracker tool for tennis tournaments sport. Big data used for video analytics that record the performance of every player in many games such as football or baseball game, and sensor technology also used in sports equipment such as basket ball's or golf clubs that allows us to get feedback from the smart phones on our game and how to improve it. Many sports teams also track athletes outside of the sporting environment using smart technology to track nutrition and sleep, by the social media conversations to monitor emotional wellbeing.

F. Improving the Science and Research field

Science and research is being transformed by the new possibilities of big data. For example, the Swiss nuclear physics laboratory which is the worlds largest and the most powerful lab for partial accelerator. Experiments to unlock the secrets of the universe, how it started and works which generated a huge amounts of data. The CERN data center has number of processors to analyze it's about 30 to 35 petabytes of data. However, it uses the computing powers of thousands of computers distributed across number of data centers generated by worldwide which to analyze the data. Such computing powers can provide maximum advantages to transform many s areas of science and research.

G. Optimizing Machine and Device Performance

Big data analytics help machines and devices become smarter. For example big data tools are used to operate Google's driving car. The GPS as the powerful computers and sensors for the safely drive on the road without the intervention of the human beings[12]. Big data tools are also used to optimize energy using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouses.

H. Improving Security and Law Enforcement.

Big data is applied for the improving security and enabling law enforcement. I know the awareness of the revelations that is the National Security Agency by using big data analytics to foil the terrorist plots. Others use of big data techniques to detect and prevent cyber attacks. Police forces use big data tools to catch criminals and to predict criminal activity and credit card companies use it to detect fraudulent transactions by big data .

I. Improving and Optimizing Cities and Countries

Big data is used to improve many aspects of our cities and also our countries. For example, it allows the cities to optimize their traffic flows based on real time traffic information from social media and weather data[13]. A number of cities are piloting big data analytics with the aim of changing themselves into Smart Cities, where as the transport infrastructure and utility processes are all joined up. A bus would wait for a delayed train and where traffic signals are predict from traffic volumes and operate or minimize jams.

J. Financial Trading

My last type of big data application from the financial trading. High-Frequency Trading is an area where the big data used to finds a lot of things now a days. Big data algorithms are useful to make trading decisions. Today, the majority of equity trading are replaced by data algorithms that increasingly take into account signals from social media networks and websites which make buy and sell decisions in split seconds.

V. CONCLUSION

This paper presents the fundamental concepts of the Big Data, which include the increase in data, the progressive demand for hard disks, and the role of Big Data in the current environment of enterprise and technology area. Thus to enhance the efficiency of data management, a data-life cycle that uses the technologies and terminologies of Big Data. The stages in the data life cycle which include collection, filtering, analysis, storage, retrieval, and discovery. All the steps convert raw data into the data which has significant aspect for the management of data in significant manner. Many organizations are often face troubles with respect to the creating, managing, and manipulating the rapid influx of information from the large datasets. The increase in data volume, thus data sources have increased in terms of size and variety. Data are also generated in different formats such as structured ,unstructured or semi-structured form, which affect data analysis, management, and storage. This variation in the data is accompanied by complexity and the development of additional means of data acquisition. The extraction of valuable data from large dataset influx of information is a critical issue in the field of s Big Data. Qualifying and validating of the items in Big Data are impractical; hence, many new approaches must be developed in this feild. From a security perspective, the major concerns of Big Data such as privacy, integrity, availability, and confidentiality with respect to outsources of data. Large amounts of data are stored in cloud platforms. However, customers cannot physically check the outsourced data. Thus, data integrity is imperil. Thus the lack of data support which caused by remote access and the lack of information regarding storage, spoil the data integrit. Big Data involves large systems, profits, and challenges. Therefore, additional research is necessary to improve the efficiency of integrity evaluation online, as well as the display, analysis, and storage of Big Data.

REFERENCES

[1] Seoung-hun Park,Young-guk Ha, “Large Imbalance Data Classification Based on MapReduce for Traffic Accident Predication” , IEEE International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing,2014,pp 45-49.

[2] Dong-Hee Shin , “Demystifying big data: Anatomy of big data”,Sung kyun kwan University ,90327 International Hall, 53 Myungryun-dong3-ga, Jongro-gu, Seoul, 2015, pp. 110-745, Republicof Korea.

[3] Samuel FossoWamba a,b,n, ShahriarAkter c, AndrewEdwards d, GeoffreyChopin e, Denis Gnanzou,“How ‘big data’ can make big impact:Findings from asystematic review and alongitudinal casestudy” Int. J.ProductionEconomics,2015, vol. 165,pp 234–246

[4] Philip Carter, Associate Vice President of IDC Asia Pacific,“Big Data Analytics: Future Architectures,Skills and Roadmaps for the CIO”,2011.

[5] J., Han, F., & Liu, H. , “ Challenges of big data analysis”, National ScienceReview,2014, vol. 1(2),pp. 293–314.

[6] E. Ackerman and E. Guizzo(2011), "5 technologies that will shape the web," Spectrum, IEEE, vol. 48, no. 6, pp. 40-45.

[7] Labrinidis, A., & Jagadish, H. V. , “Challenges and opportunities with big data”,Proceedings of the VLDB Endowment,2012, vol. 5(12), pp. 2032–2033,.

[8] http://hadoop.apache.org/ website Hadoop.

[9] Y. Demchenko, C. Ngo, and P. Membrey, “Architecture Framework and Components for the Big Data Ecosystem,” Journal of System and Network Engineering,2013, pp. 1–31.

[10] Dean J, Ghemawat S , “ MapReduce: simplified data processing on large clusters”,2008, vol. 51(1), pp.107-113

[11] Olston C, Reed B, Srivastava U, Kumar R, Tomkins A , “Pig latin: a not-so-foreign language for data processing”, In: Proceedings of the ACM SIGMOD international conference on Management of Data. ACM,2008, pp 1099-1110

[12] Chen,H.,Chiang,R.,&Storey,V.,“Business intelligence and analytics:From big data to big impact”, MIS Quarterly, 2012, vol.36(4), pp.1165–1188.

[13] Lizhe Wang, Senior Member, IEEE, Yan Ma, Member, IEEE, Albert Y. Zomaya, Fellow, IEEE, Rajiv Ranjan, Member, IEEE, and Dan Chen, “A Parallel File System with Application-Aware Data Layout Policies for Massive Remote

Sensing Image Processing in Digital Earth”, IEEE Transactions on Parallel and Distributed Systems, 2015, vol.. 26, pp. 6.

[14] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang, Fellow, IEEE, “Traffic Flow Predsiction With Big Data: A Deep Learning Approach”, IEEE Transactions on Intelligent Transportation Systems, 2015, vol. 16, NO. 2, pp 865.