World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 5.218

A STATE OF THE ART REVIEW ON DATA MINING WITH BIG DATA

M. Sharmila*

Assistant Professor, Department of Information Technology, Aditya Engineering College (Autonomous), ADB Road, Surampalem, East Godavari (DT), Andhra Pradesh, India.

Article Received on 01/05/2018 **Article Revised on 22/05/2018** **Article Accepted on 12/06/2018**

***Corresponding Author**
M. Sharmila
Assistant Professor, Department of Information Technology, Aditya Engineering College (Autonomous), ADB Road, Surampalem, East Godavari (DT), Andhra Pradesh, India.

ABSTRACT

Data Mining indulged with Big data in numerous and autonomous sources concerned with large-volume, complex information and growing data sets are the emerging impact in present scenario. Big data are now quickly developing in all domains like biological, physical and biomedical sciences with the rapid growth of data storage, data collection capacity and the networking. Big Data doesn't fits into a well formatted tables rather than it has many new unstructured and structured data that tends to be much challenging factor to get processed easily. Owing to the fast development of these data, results are needed to be studied as well provided so as to handle and mine knowledge and value from these datasets. Additionally, to gain valuable insights from varied and quickly changing data, decision makers need to be ranging from day-to-day transactions to customer interactions in addition to social network data. These values are offered using big data analytics that is the application of progressive analytics method on big data. The goal of this paper is to examine some of the diverse analytics techniques as well as tools that are applied to process and manipulate the big data.

KEYWORDS: Big Data, Data mining, Hadoop, HDFS, MapReduce.

INTRODUCTION

Big data is referred to as data sets or else mixtures of data sets whose size (volume), complexity (variability), and rate of growth (velocity) makes them problematic to be captured, managed, processed or analyzed through conventional equipment and tools, for instance relational databases as well as desktop statistics otherwise visualization packages, surrounded by the time necessary to create them useful. Whereas the size is used to determine whether a particular data set is deliberated big data is not firmly well-defined and carry on to change over time, furthermost analysts and practitioners currently denote to data sets from 30-50 terabytes that is 1012 or 1000 gigabytes per terabyte to multiple petabytes that is 1015 or 1000 terabytes per petabyte as big data.[1]

There is no unbreakable and fast instruction about exactly what size database requirements to be in command for the data inside of it to be deliberated big. As an alternative, what classically describes big data is the basic for new methods and tools so as to be able to process it. With the intention of uses of big data needs program that span many physical and virtual machines functioning together performing in order to process all of the data in a reasonable span of time.[2]

Big data can be stowed, acquired, processed in addition to analyze in a lot of ways. All big data source has diverse characteristics such as the frequency, volume, velocity, veracity and type of the data. Once big data is processed and stowed, extra dimensions come into play, including governance, security as well as policies. Selecting architecture and also building a suitable big data solution is challenging because so many factors have to be considered.[3] The Big data architecture as well as patterns sequence grants a structured and pattern-based method to simplify the task of defining complete big data architecture. Since it is significant to evaluate whether a business scenarios is a big data problem.[3]

BIG Data – Growth and Size Facts (*MGI Estimates)

- There are five billion mobile phones use in 2010.
- There are thirty billion pieces of content shared on Facebook every single month.
- There is a forty percent projected growth in global data generated per year versus five percent growth in global IT spending.
- US Library of Congress in April 2011 collects data of 235 terabytes.

In the United States 15 out of 17 major business sectors have more data stored for each company that the US Library of Congress.[1]

Structured Versus Unstructured Data

In categorizing big data, TCS (Tata Consultancy Services Limited) regarded the report that how much companies data was structured versus unstructured and how much was produced internally versus externally.[4]

- 51% of data is structured
- 21% of data is semi-structured
- 27% of data is unstructured

A much advanced than anticipated percentage of data was not structured - either unstructured or else semi-structured addition to a little less than a quarter of the data was external.

Layers of Big Data

Three model layers are[9]:

- *Physical layer* – it have dissimilar data types like logs, video, business tables, audio and so on.
- *Modeling Data Layer* – abstract data model is developed to handle physical data.
- *Computing modeling layer* – application layer to develop in addition to retrieve information for business value.

In physical layer to store the data, abstract layer is used. In global devices large volume of data with dissimilar formats are stored.[9] Big data model offers a visual method to manage data resources in addition to generate basic data architecture so as to optimize data reuse as well as condense computing costs. In order to build a data model with distinct physical data these three models were used. That is, the application is capable to access data by the data model instead of accessing the physical data which makes application flexible and data manageable. To build big data model produce a data depends on data type, read-write requirement, relationship, data storage and so on. Additional, modeling applications uphold these models, so as to data models are capable to display and store the modern data.

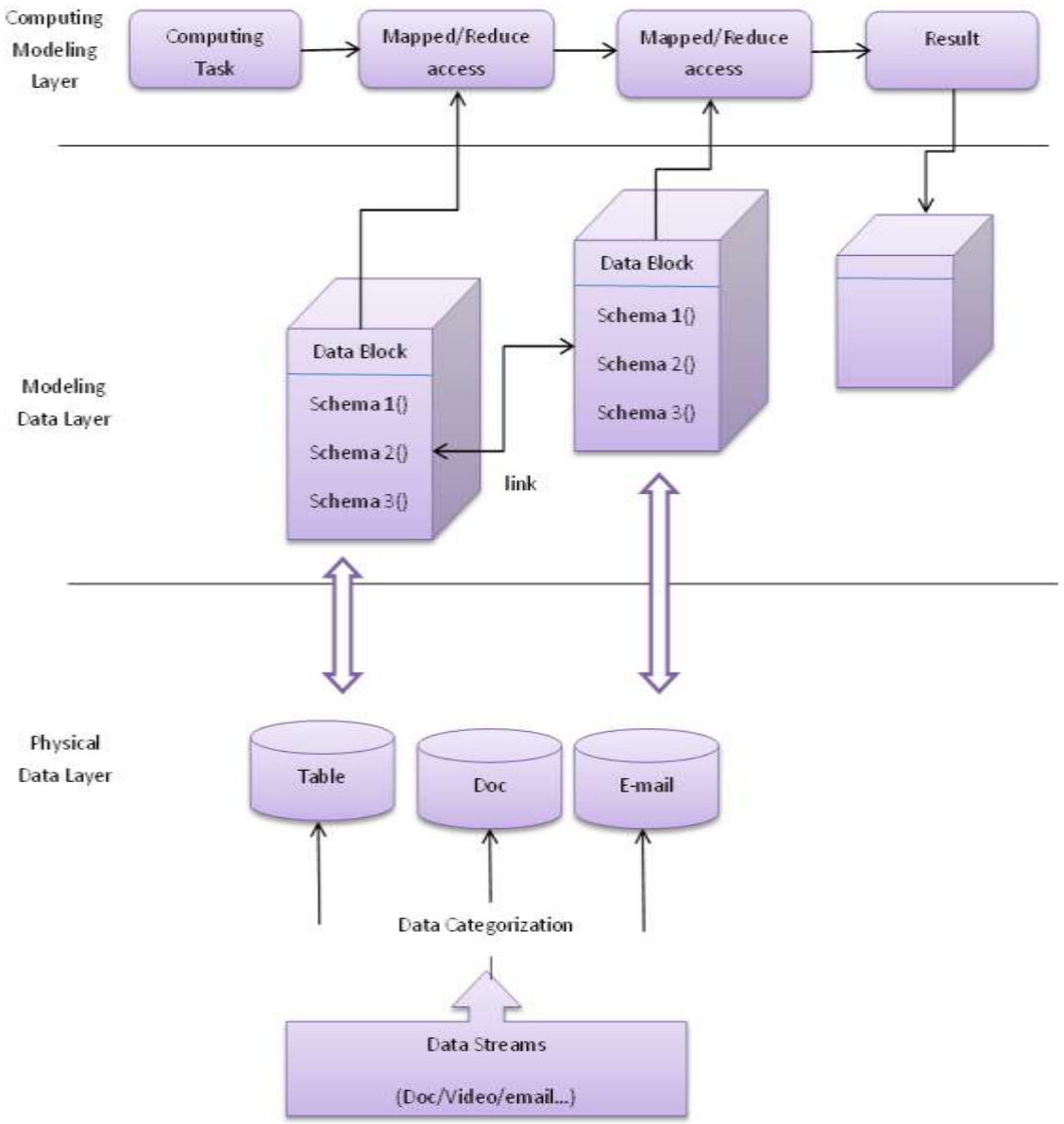

Fig. 1: Layers of Big Data.

Big Data Characteristics

Big data analysis has two perspectives[6]:

- **Decision-oriented**

Decision-oriented analysis is more similar to traditional business intelligence. Discriminating subsets and representations of larger data sources are used to apply the outcomes to the process of making business decisions. Definitely these decisions might outcome in certain kind of action or process change; however the purpose of the analysis is to augment decision making.

- **Action-oriented**

Action-oriented analysis is mainly used for rapid response, action is required when a pattern emerges or exact kinds of data are detected. Take the advantage of big data through analysis then causing proactive or else reactive behavior alterations offer great potential for early adopters. By creating analysis applications, we can find and utilize big data that can hold the key to extracting value faster rather than later. It is more operative to construct these custom applications from scratch or else via leveraging platforms and/or components.

The additional characteristics of big data analysis that make it dissimilar from traditional kinds of analysis sideways from the three Vs of volume, velocity, and variety.[6]

- Programmatic.
- Data driven.
- Use a lot of attributes.
- Iterative.
- Quick to get the calculate cycles needed via leveraging a cloud-based Infrastructure as a Service.[6]

Effective uses of big data exist in the following areas

- In order to develop use information technology (IT) logs IT troubleshooting and security breach detection, speed, effectiveness, future occurrence prevention.
- Usage of voluminous historical call center material more rapidly, with the intention to improve customer interaction and satisfaction.[1]
- Usage of social media contented with the purpose of better and more quickly understand customer sentiment around your customers and then improve products, services, and customer interaction.
- Fraud detection plus prevention in any industry which procedures financial transactions on-line, for example shopping, banking, investing, insurance, health care claims, etc.,
- Usage of financial market transaction information to more rapidly evaluates risk and takings corrective action.

6 V'S of Big Data

1. **Volume.** Big data suggests huge volumes of data. It is used for an employee to created data. Now a day's data is generated by machines, networks plus human interaction on

systems such as social media, the volume of data to be analyzed is massive.[7] Inderpal states the volume of data is not as much the problem like extra V's like veracity.

2. **Variety.** Variety states that the several sources and kinds of data both structured as well unstructured. To store data from sources alike databases, spreadsheets. Currently data arises in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. These varieties of unstructured data generate problems for storage, mining and analyzing data.[7]
3. **Velocity.** Big Data Velocity contracts with the step upon which data flows in from sources resembling business processes, machines, networks and human interaction with things such as social media sites, mobile devices, etc. The flow of data is massive also continuous. These real-time data can aids researchers and businesses make valuable decisions which offer strategic competitive advantages and ROI.[7] Inderpal[7] proposed that sampling data can help deal with problems similar to volume and velocity.
4. **Veracity.** Big Data Veracity denotes the biases, noise then abnormality in data. The data is being stored and then mined meaningful to the problem being examined.[7] Inderpal[7] states sense veracity in data analysis is the biggest challenge when compared to things such as volume and velocity.
5. **Validity.** Big data veracity is the major problem of validity and meaning is the data correct and accurate for the envisioned use. Obviously valid data is major key to making the right decisions.
6. **Volatility.** Big data volatility denotes in what way long data is valid and just how long it should be stored. Real time data is needed to determine which point is data no longer relevant to the current analysis.

BIG DATA CHALLENGES

- ***Understanding and utilizing big data*** - It is an intimidating assignment in most industries as well companies that deals with big data to understand the data that is obtainable to be used, by determining the best use of that data depends on the company, industry, strategy, and tactics. Then these kinds of analyses essential to be performed on a continuing basis as the data landscape modifying at an ever-increasing rate and as executives improve more appetite for analytics depends on all obtainable information.[1]
- ***New, Complex, and Continuously Emerging Technologies*** – From the time when the technology that is obligatory with the purpose of utilize big data is new to many organizations, it is necessary for these organizations to study about these new tools at an ever-accelerating pace and potentially engage with dissimilar technology providers and

partners. With all technology, companies entering into the world of big data will essential to balance the business requirements associated with big data with the associated costs of incoming into and residual betrothed in big data capture, storage, processing, and analysis.

- ***Cloud Based Solutions*** – A new class of business software applications is developed whereby company data is managed plus stored in data centers around the world. Whereas these solutions range from ERP, CRM, Document Management, Data Warehouses and Business Intelligence to many others, the general problem remains the safe keeping as well management of confidential company data. The keys often provide companies marvelous flexibility and cost savings opportunities associated to more traditional on evidence solutions but then again it increases a new dimension associated to data security and the general management of an enterprise's Big Data paradigm.
- ***Privacy, Security, and Regulatory Considerations*** - The volume as well complexity of big data is challenging for maximum partnerships to achieve a consistent grasp on the content of the data and to capture and secure it sufficiently, consequently that confidential or private business plus customer data are not retrieved by or disclosed to unauthorized parties. The data privacy breach cost is enormous.
- ***Archiving and Disposal of Big Data*** – From the time when big data will miss its value to present decision-making over time, subsequently it is voluminous plus varied in content and structure, it is essential to utilize new tools, technologies, and methods to archive and delete big data, without foregoing the efficiency of using big data for current business needs.[1]
- ***The Need for IT, Data Analyst, and Management Resources*** – It is projected that there is a requirement for approximately 140,000 to 190,000 additional workers with "deep analytical" knowledge and 1.5 million more data-literate managers, also retrained or hired. Consequently, it is possible that any firm that undertakes big data inventiveness will prerequisite to either reeducate existing people or else engage new people in order for their initiative to be successful.

DIFFERENT TECHNIQUES AND TOOLS TO MANIPULATE BIG DATA

Big Data supports us to generate new growth opportunities and completely new groupings of companies, like that aggregate and analyze industry data. The companies that sit in the middle of enormous information flows wherever data around products and services, buyers and suppliers, consumer preferences and determine can be captured plus analyzed. Forward-

thinking leaders across sectors[5] should begin aggressively to build their organizations in Big Data capabilities. The following passage will outline some of the techniques and tools that are playing a vital role in processing the Big Data.

APACHE HADOOP

Apache Hadoop software library is a framework which permits for the distributed processing of big data sets across group of computers by using simple programming models. It is designed to scale up from particular servers to thousands of machines every machine providing local storage and computation. Somewhat than rely on hardware to deliver high-availability, at the application layer library is intended to detect and handle failures, consequently intending a highly-available facility on uppermost of a group of computer, each of which is disposed to failure.[8]

Layers of Hadoop

Hadoop has two main layers namely[10]:

1. Processing or Computation layer (MapReduce).
2. Storage layer (Hadoop Distributed File System).

The project contains these modules

- ***HadoopMapReduce*:** In HadoopMapReduce, for parallel processing of large data sets a YARN based system is used.
- ***Hadoop Distributed File System (HDFS):*** it offers high-throughput access to application data. HDFS is based on GFS (Google File System)
- ***Hadoop YARN:*** An outline for job scheduling in addition to cluster resource management.
- ***Hadoopcommon:*** support the additional Hadoop modules.

HDFS ARCHITECTURE

Name Node – It is the commodity hardware contains GNU or Linux operating system and software of name node. Software runs on the hardware. The name node act as the master server and the tasks are:

- Manages the file system namespace.
- Regulates clients access to files.[10]
- It performs file system operations like renaming, closing, and opening files and directories.

Data Node – It is the commodity hardware contains GNU or Linux operating system and software of data node. For each node in group there is a data node. They manage the data storage.

- As per client request, Data nodes execute read-write operations on the file systems.
- According to the instructions of the name node, they implement operations like block creation, deletion, and replication.

Block –The data are stored in the file of HDFS. In a file system, the files are separated into one or more segments and stored in separate data nodes and they are known as blocks. The default size of the block is 64 MB, it may possibly be increase as per the necessity to change in HDFS configuration.

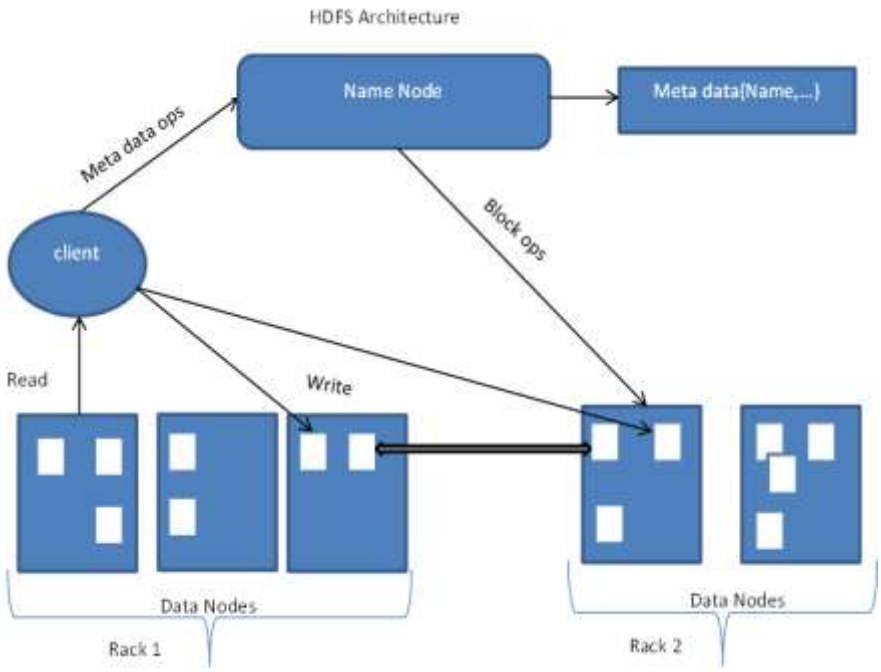

Fig. 2: Hadoop Architecture.

ADDITIONAL HADOOP RELATED TOOLS

- ***Ambari*:** Ambari is a web-based tool for provisioning, managing, and monitoring. Apache Hadoop clusters that support for Hadoop HDFS, Hive, HCatalog, ZooKeeper, Oozie, Hadoop MapReduce, Pig, HBase and Sqoop.[8] Similarly Ambari also provides a dashboard for inspecting group health like heat maps as well as capability to view MapReduce, Pig along with Hive applications visually along with features to diagnose their presentation characteristics in a user-friendly manner.

- ***Avro:*** It is a data serialization system and relies on *schemas*. When Avro data is read, the schema used when writing is always present. This permits each data to be written with no per-value overheads, making serialization both fast and small and also facilitates use with dynamic, scripting languages, since data, together with its schema, is fully self-describing.
- ***Cassandra:*** It is a scalable multi-master database using no single points of failure.
- ***Chukwa:*** It is a data collection system for managing large distributed systems.
- ***HBase:*** It is a scalable, distributed database which supports structured data storage for large tables.
- ***Hive:*** It is a data warehouse infrastructure which offers data summarization as well as ad hoc querying.
- ***Mahout:*** It is a data mining library, scalable machine learning.
- ***Pig:*** It is a high-level data-flow language and also execution framework for parallel computation.
- ***Spark:*** It is a fast plus overall compute engine for Hadoop data. Spark offers a simple addition to expressive programming model which supports an extensive range of applications such as ETL, machine learning, stream processing, and graph computation.
- ***Tez:*** It is a generalized data-flow programming framework, constructed on HadoopYARN that delivers a powerful then flexible engine to execute an arbitrary DAG of tasks to progression data for together batch and interactive use-cases.[8] A Tez is adopted by Hive, Pig and additional outlines in the Hadoop ecosystem, additional to by other commercial software to substitute HadoopMapReduce as the underlying execution engine example ETL tools.
- ***ZooKeeper:*** It is a high-performance coordination facility for distributed applications and also a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

MAPREDUCE

In Apache Hadoop, MapReduce is helpful for batch processing on terabytes or petabytes of data stored[11] Some of the MapReduce benefits are:

- **Scalability:** Petabytes of data stored in HDFS on one cluster is processed by MapReduce.
- ***Simplicity:*** Developers can write applications in Java, C++ or Python and it is easy to run MapReduce jobs.

- ***Speed:*** MapReduce can take problems that used to take days to resolve and resolve them in hours or minutes are parallel processing.
- ***Recovery:*** MapReduce takes care of failures. A machine with one replica of the data is out of stock, if another machine has a replica of the same key or value pair that can be used to resolve the same sub-task. The JobTracker remains track of it all.
- ***Minimal data motion:*** MapReduce shifts compute processes to the data on HDFS and not the other way around. The data resides when the processing tasks can happen on the physical node. This significantly diminishes the network I/O patterns plus contributes to Hadoop's processing speed.[11]

CONCLUSION

Data mining is an analytical way designed to explore enormous relevant and valid data. Big data is the term for a group of complex data sets where vast amount of semi structured, heterogeneous and unstructured data are repeatedly created at extraordinary scale. The usage of Big Data is fetching a critical way for leading firms to outperform their peers. In several industries, conventional competitors and new entrants similar will leverage data-driven approaches to innovate, compete, and capture value. The limitation of current data mining methods reveals big data, resulted in a sequence of innovative challenges related to big data mining. Big data mining needs high performance computing platforms that execute systematic designs to release the complete power of the big data. Big data mining is an auspicious area, despite of the partial work done on extracting till now more work is essential to overcome its challenges associated to heterogeneity, scalability, speed, accuracy, trust, provenance, and privacy. Besides the sheer scale of Big Data, the real-time plus high-frequency nature of the data are also significant. Correspondingly, the high frequency of data allows users to test theories in near real-time and to attain a level never before conceivable. Henceforth this art review paper delivers an outline of big data platform for managing and processing big data as well as few features of its techniques. This in turn will be greatly useful for the readers to enrich their knowledge in the developing trends of big data and further innovations for big data mining in all domains.

REFERENCES

1. Available at: http://www.navint.com/images/Big.Data.pdf.
2. Available at: https://opensource.com/resources/big-data.
3. Available at: http://www.ibm.com/developerworks/library/bd-archpatterns1/.

4. Availableat:http://www.webopedia.com/quick_ref/important-big-data-facts-for-it-professionals.html.
5. Availableat:http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/.
6. Availableat:http://www.dummies.com/how-to/content/characteristics-of-big-data-analysis.html.
7. Available at: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.
8. Available at: http://hadoop.apache.org/.
9. Jinbao Zhu, Allen Wang, "Data modeling for big data" Available at: www.ca.com/us/~/media/files/articles/ca-technology-exchange/data-modelling-for-big-data-zhu-wang.aspx.
10. "HADOOP big data analysis framework" tutorialspoint SIMPLYEASYLEARNING, Available at: www.tutorialspoint.com/hadoop/hadoop_tutorial.pdf.
11. Available at: http://hortonworks.com/hadoop/mapreduce.