Analytical Study on Big Data
Dr. Vivek Raich, Pankaj Maurya
Department of Computer Science, Govt. Holkar Science College, Indore, Madhya Pradesh, India

Abstract— in the time of the Information Technology, the big data store is going on. Due to which, Huge amounts of data are available for decision makers, and this has resulted in the progress of information technology and its wide growth in many areas of business, engineering, medical, and scientific studies. Big data means that the size which is bigger in size, but there are several types, which are not easy to handle, technology is required to handle it. Due to continuous increase in the data in this way, it is important to study and manage these datasets by adjusting the requirements so that the necessary information can be obtained.

The aim of this paper is to analyze some of the analytic methods and tools. Which can be applied to large data. In addition, the application of Big Data has been analyzed, using the Decision Maker working on big data and using enlightened information for different applications.

Keywords— Big data, data mining, data warehousing, Hadoop, Analytics, Decision making.

I. INTRODUCTION

In the era of Information Technology, the data is being created day by day, which is to produce and store data in the amount of the Society and Organizations. A place where every detail about a human, organization or Company, every deal or transaction performed, or every aspect which can be documented and record are lost directly after use.

Human beings now create 2.5 quintillion bytes of data per day [1]. The rate of data creation has increased so much that 90% of the data in the world today has been created in the last 8 years alone. This acceleration in the production of information has created a need for new technologies to analyze massive data sets. The urgency for collaborative research on Big Data topics is underscored by the U.S. federal government’s recent $200 million funding initiative to support Big Data research.[2]

The variety, size, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.[3]

The contribution of this paper is to provide an analysis of the available literature on big data analytics.[1] This document describes how the incorporation of Big Data is changing security analytics by providing new tools and opportunities for leveraging large quantities of data. Briefly discusses tools used in Big Data analytics, methods, and technologies.[3]

II. BIG DATA ANALYTICS

- The way toward examining and mining Big Data – can deliver operational and business learning at an uncommon scale and specificity. The need to examine and use slant information gathered by organizations is one of the primary drivers for Big Data examination instruments. They are datasets whose size is past the capacity of generally utilized programming devices and capacity frameworks to catch, store, manage, and in addition process the information inside a middle of the road passed time.

- Big information sizes are continually expanding, right now extending from a couple of dozen terabytes (TB) to numerous petabytes (PB) of information in a solitary informational collection. Thus, a portion of the troubles identified with huge information incorporate catch, stockpiling, look, sharing, investigation, and picturing. Today, ventures are investigating substantial volumes of exceptionally point by point data by point information in order to find realities they didn't know previously.

- Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage.[3]

- Big Data technologies can be divided into two groups: batch processing, which are analytics on data at rest, and stream processing, which are analytics on data in motion. Real-time processing does not always need to reside in memory, and new interactive analyses of large-scale data sets through new technologies. [2]
Hadoop is one of the most popular technologies for batch processing. The Hadoop framework provides developers with the Hadoop Distributed File System for storing large files and the MapReduce programming model, which is tailored for frequently occurring large-scale data processing problems that can be distributed and parallelized.[3]

Several tools can help analysts create complex queries and run machine learning algorithms on top of Hadoop. New frameworks such as Spark4 were designed to improve the efficiency of data mining and machine learning algorithms that repeatedly reuse a working set of data, thus improving the efficiency of advanced data analytics algorithms. There are also several databases designed specifically for efficient storage and query of Big Data, including Cassandra, CouchDB, Greenplum Database, HBase, MongoDB, and Vertica.[3]

1. Data Privacy:
The preservation of privacy largely relies on technological limitations on the ability to extract, analyze, and correlate potentially sensitive data sets. However, advances in Big Data analytics provide tools to extract and utilize this data, making violations of privacy easier. As a result, along with developing Big Data tools, it is necessary to create safeguards to prevent abuse. [3]

The scope of this paper is on how Big Data can improve information security best practices.

2. Big Data Characteristics:
Big data refers to large-scale data architectures and facilitates tools addressing new requirements in handling data volume, velocity, and variability. Traditional databases (data warehousing) assume data are organized in rows and columns and employ data-cleansing methods on the data, while the data volumes grow over a time period and often lack on handling such large-scale data processing. Traditional data base/warehousing systems were designed to address smaller volumes of structure data, with the predictable updates and consistent data structure, which mostly operate on single server and lead to operational expenses with the increased data volume. However, big data comes in a variety of diverse formats with both batch and stream processing in several areas such as geospatial data, 3D data, audio and video, structured data, unstructured text including log files, sensor data, and social media. Below, we discuss the properties of traditional database (data warehousing) and big data.[4]

Big Data can be simply defined by explaining the 3V’s – volume, velocity and variety which are the driving dimensions of Big Data quantification. introduced the famous 3 V’s concept.

a. Volume: This essentially concerns the large quantities of data that is generated continuously. Initially storing such data was problematic because of high storage costs. However with decreasing storage costs, this problem has been kept somewhat at bay as of now. However this is only a temporary solution and better technology needs to be developed. Smartphones, E-Commerce and social networking websites are examples where massive amounts of data are being generated. This data can be easily distinguishes between structured data, unstructured data and semi-structured data.

b. Velocity: In what now seems like the pre-historic times, data was processed in batches. However this technique is only feasible when the incoming data rate is slower than the batch processing rate and the delay is much of a hindrance.

c. Variety: Documents to databases to excel tables to pictures and videos and audios in hundreds of formats, data is now losing structure. Structure can no longer be imposed like before for the analysis of data. Data generated can be of any type- structures, semi-structured or unstructured. The conventional form of data is structured data. For example text. Unstructured data can be generated from social networking sites, sensors and satellites.[1]

III. BIG DATA ANALYTICS TOOLS AND METHODS
To develop the technology and analyze the data being used in organizations, faster and more efficient methods have become necessary. From which we can get the correct decision, but still we have not been able to work on it completely. Advanced technology will be needed to manage and process Big Data. Traditional Data Management, Analysis Techniques and Infrastructures will have to change with this type of data set.

1. Tools:
Key-value stores: Key-value pair (KVP) tables are used to provide persistence management form any NoSQL technologies. The concept is that the table has two columns – one is the key; the other is the value. The value could be a single value or a data block containing many values; the format of which is determined by program code. KVP tables may use indexing and have tables or sparse arrays to provide rapid retrieval and insertion capability, depending on the need for fast lookup, fast insertion, or efficient storage. KVP tables are best applied to simple data structures and on the Hadoop MapReduce environment. [4]
Document-oriented database: A document-oriented database is a database designed for storing, retrieving, and managing document-oriented or semi-structured data. The central concept of a document-oriented database is the notion of a document where the contents within the document are encapsulated or encoded in some standard format such as JavaScript object notation, binary JavaScript object notation, or XML. [4]

Column family/big table database: Instead of storing key-values individually, they are grouped to create the composite data, each column containing the corresponding row number as key and the data as value. This type of storage is useful for streaming data such as web logs, time series data coming from several devices, sensors, and so on. [4]

Graph database: A graph database uses graph structures similar to nodes, edges, and properties for data storing and semantic query on the data. In a graph database, every entity contains direct pointers to its adjacent element, and index lookups are not required. A graph database is useful when large-scale multi-level relationship traversals are common and desirable for processing complex many-to-many connections such as social networks. A graph may be captured by a table store, which supports recursive joins such as big table and Cassandra. [4]

2. Programming models:

Programming models like data intensive, stream computing, batch processing.[4]

Batch and stream processing: Today developers are analyzing Terabytes and Petabytes of data in the Hadoop Ecosystem. Many projects are helping to speed up this innovation. All of these project are rely on two aspects. They are:[5]

- Batch Processing
- Stream Processing

Batch Processing [5]: Batch processing is where the processing happens of blocks of data that have already been stored over a period of time. For example processing all the transaction that have been performed by a major financial firm in a week. This data contains millions of records for a day that can be stored as a file or record etc. This particular file will undergo processing at the end of the day for various analysis that firm wants to do. Obviously it will take large amount of time for that file to be processed. That would be what Batch Processing is:

Hadoop MapReduce is the best framework for processing data in batches. The following figure gives you detailed explanation how Hadoop processing data using MapReduce.

![Figure 1 Processing data Using MapReduce](image)

Batch processing works well in situations where you don’t need real-time analytics results, and when it is more important to process large volumes of information than it is to get fast analytics results.

Stream Processing [5]: Stream processing is a golden key if you want analytics results in real time. By building data streams, you can feed data into analytics tools as soon as it is generated and get instant analytics results using platforms like Spark Streaming. Apache Storm is a stream processing framework. I would recommend WSO2 Stream Processor (WSO2 SP), the one I have helped built. It can ingest data from Kafka, HTTP requests, message brokers. You can query data stream using a “Streaming SQL” language. WSO2 SP is open source under apache license. With just two commodity servers it can provide high availability and can handle 100K+ TPS throughput. It can scale up to millions of TPS on top of Kafka. The Business Rules Manager of WSO2 SP allows you to define templates and generate business rules from them for different scenarios with common requirements.

Stream processing is useful for tasks like fraud detection. If you stream-process transaction data, you can detect anomalies that signal fraud in real time, then stop fraudulent transactions before they are completed.[5]

![Figure 2 : Real time processing in spark](image)
IV. BIG DATA APPLICATION

Due to the development of technology development, the information is increasing day by day. Which can be used in the development of the organization. Decisions can be taken on data sets with the help of Technologies.

Big Data is slowly becoming universal. Every field of business, health or general living standards now can implement big data analytics. To put simply, Big Data is a field which can be used in any zone whatsoever given that this large quantity of data can be harnessed to one’s advantage. The major applications of Big Data have been listed below.[1]

- **The Third Eye- Data Visualization**
  Organizations worldwide are slowly and perpetually recognizing the importance of big data analytics. From predicting customer purchasing behavior patterns to influencing them to make purchases to detecting fraud and misuse which until very recently used to be an incomprehensible task for most companies big data analytics is a one-stop solution. Business experts should have the opportunity to question and interpret data according to their business requirements irrespective of the complexity and volume of the data. In order to achieve this requirement, data scientists need to efficiently visualize and present this data in a comprehensible manner. Giants like Google, Facebook, Twitter, EBay, Wal-Mart etc., adopted data visualization to ease complexity of handling data. Data visualization has shown immense positive outcomes in such business organizations. Implementing data analytics and data visualization, enterprises can finally begin to tap into the immense potential that Big data possesses and ensure greater return on investments and business stability.[1]

- **Integration- An exigency of the 21st century**
  Integrating digital capabilities in decision-making of an organization is transforming enterprises. By transforming the processes, such companies are developing agility, flexibility and precision that enables new growth. Gartner described the confluence of mobile devices, social networks, cloud services and big data analytics as the as nexus of forces. Using social and mobile technologies to alter the way people connect and interact with the organizations and incorporating big data analytics in this process is proving to be a boon for organizations implementing it. Using this concept, enterprises are finding ways to leverage the data better either to increase revenues or to cut costs even if most of it is still focused on customer-centric outcomes. Such customer-centric objectives may still be the primary concern of most companies, a gradual shift to integrating big data technologies into the background operations and internal processes.[1]

- **Big Data in Healthcare:**
  Healthcare is one of those arenas in which Big Data ought to have the maximum social impact. Right from the diagnosis of potential health hazards in an individual to complex medical research, big data is present in all aspects of it. [1]

- **Big Data and the World of Finance:**
  Big Data can be a very useful tool in analyzing the incredibly complex stock market moves and aid in making global financial decisions. For example, intelligent and extensive analysis of the big data available on Google Trends can aid in forecasting the stock market.[1]

- **Big Data and the Food Industry:**
  The impact of Big Data on the food industry is increasing exponentially. Be it for tracking the quality of products or presenting recommendations to the customer or developing marketing strategies for better customer experience, the presence of Big Data analytics on the food industry is slowly becoming ubiquitous.[1]

V. CONCLUSION

This literature survey discusses on Big Data, Big Data is an evolving area, there is a lot of work to be done in big data right now. In the Information Age we are currently living, the vast varieties of high velocity data are being produced daily, and internal patterns and hidden knowledge patterns are kept inside those which should be removed and used. Therefore, large data analytics can be applied to advanced analytical techniques on big data, and to find hidden knowledge and valuable things.

The study was reviewed to analyze large data analytics concepts. They are being constantly researched and they are also being used in decision making. Consequently, big data was discussed, as well as its characteristics and importance. Moreover, some of the big data analytics tools and methods in particular were examined. Thus, this study provided a detailed description of large data collection and management as well as large data analytics processing. Some of the different advanced data analytics techniques were discussed. Apart from this, the details of various Big Data applications were explained in detail so that the distribution maker.
REFERENCES


[8] G. Sabarmathi, Dr. R. Chinnaiyan, Dr. V. Ilang, Big Data Analytics Research Opportunities and Challenges-October 2016


