A Review on Big Data and Security Challenges

Er. Achharpreet Bhalla and Er. Gagandeep Bhalla
(Assistant Professors, Computer science and Engineering, PIT GTB Garh, Moga, India)

Abstract- Data is currently one of the most essential distinctions for companies in each and every field. The rapidly increasing growth of volume of data has created a new challenge. So, these challenges therefore, solved through the innovation of a new standards of big data. Thirty years ago, one gigabyte of data could amount to a big data problem and therefore needs special purpose computing resources should be needed. Now today, Terabytes of data can be rapidly or easily transmitted, processed and stored and used for customer oriented devices for the better service. The analysis of big data is a multifaceted aspire the stir of Mathematics, computer science and statistics. Data analytic defines how to manage the complete data lifecycle to carry out the whole big data concept. Big data also defines the core components of Hadoop system that are responsible for large amount of data processing in better way. Big data is not related only to data volume or variety of data but also add big data privacy and security. The main objective of this paper is to explore the overall view of big data aspects, security challenges and issues and various related tools combined with this. As a result, this paper gives a way to enhance big data at different stage of work. It is almost difficult to define deeply and detailed research but is a big picture of the main reason related to big data approach along with principle solutions.

Keywords- *Big data introduction, Big data characteristic, Data Analytics and lifecycle, Hadoop system and privacy and security challenges*

I. INTRODUCTION

Big data is an area related to the analysis, processing and storage of huge collection of data that is fetched from different sources related to data. Big data need come into existence when the analysis, processing and storage technologies and techniques are not properly worked then the better decisions of big data introduces to sort out these difficulties that are related to data issues. Big data defines different requirements like merging of different data sets, processing large amount of unbalanced data and combining hidden data in proper time slots. Big data analysis is important for business people and researches to make much better decisions that were not attained previously. Analysis of big data reduces frauds, unauthorized access and helps to enhance scientific researches and field development. At the time of development of computers the large amount of data is not able to store in computer because of limited capacity of storage. So when networking comes into existence then with internet technology massive data storage capacity has been increased. After that new techniques come into existence like image processing, cloud computing, data mining, and another algorithms, concepts and advanced methods and also advancement in social media that arises the need for large storage capacity of data because of large size of data it is considered as "Big Data". The Existing software's for big data applications like Hadoops and Google Maps that reduces framework in intermediate data. The Hadoop components define proper framework of data management. Big data provides a framework for Hadoops, Cloud computing and distributed framework which incorporates large scale data analysis. Big data lead to wide range of benefits such as scientific discoveries, fault and fraud detection, improved decision making, correct predictions, searching of new markets, increase optimization etc.

II. IDENTIFYING BIG DATA CHARACTERISTICS

Big data characteristics used to help differentiate data as "Big "from other datasets. Basically five data characteristic shown below:

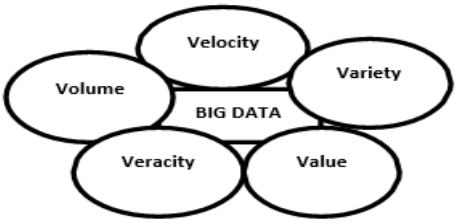

Fig.1: 5v's of big data

A. Volume

Volume is a huge amount of data. The data resources that is responsible for producing large volume of data that are-online transaction, research experiment, sensor and social media etc.

B. Velocity

Velocity related to high speed of data. There is huge and rapidly flow of data that represents how fast the data is produced and processed to meet the requirements.

C. Variety

It related to the nature about the data that defines what type of data it is. It includes structure of data.

D. Veracity

Veracity refers to the quality and consistency that approaches to

data which is available sometime get messy and difficult to control.

E. Value

Value refers to the usefulness of the data. Data in itself has no use but it needs to be converted into valuable and important information.

III. DATA ANALYTICS

Data analytic is a greater term includes data analysis. Data analytic include the management of the complete data lifecycle which includes collection, cleansing, and organizing, storing, analyzing data. Data analytic has developed methods that allow data analysis to occur through the use of highly scalable distributed technology and framework that are used for analyzing large volume of data from variety of resources. The four categories of analytics that represents the different results that is:

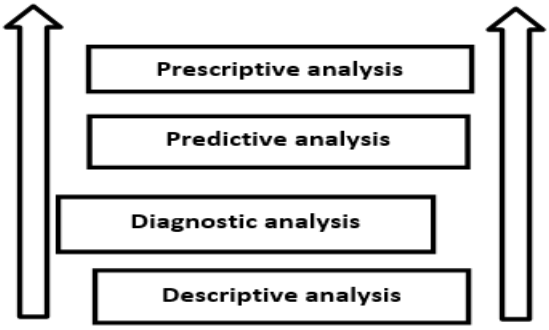

Fig.2: Data analytics

A. Descriptive analytics

It includes the number of processes executed and find out how many processes are executed within giving time slots and descriptive analysis of data and causes behind the events and behaviors.

B. Diagnostic analytics

Diagnostic analytics main focus to find out the convictions that produced in the previous question and the events to identify what information is matches to the sensation and tells why something has produced. It collects data from various resources and storing it in exact format.

C. Predictive analytics

Predictive analytics defines the outcomes of events and find out structure, exceptions arises in past and recently used data and used to evaluate the risks.

D. Prescriptive analytics

This analytic technique tells us what action should be carried out when numerous results are calculated. This technique used a huge amount of internal and external data results and defines which course of actions is best and when these actions should be taken.

IV. BIG DATA ANALYTICS LIFECYCLE

When analysis of big data then lifecycle organize and managed the events and actions related to big data analysis. Lifecycle is divided into different categories:-

A. Evaluation

Evaluation part describes a clear understanding of the validation, motivation and scope of carrying out the analysis of data. It helps in decision making and understands how business resources should be managed properly.

B. Identification

This data identification part defines the data sets used for the analysis of project and their used resources. Identifying large variety of data sources may enhance the chances to find out the hidden information.

C. Acquisition and Filtering

During this stage data is collected from last explained and then acquire data is considered to automated filtering for the vanishes of unused, invalid and corrupted data.

D. Extraction

The data execution stage is related to extracting different data and changes it into a format that used for data analysis.

E. Validation and cleansing

Sometime invalid data gives false results because of unstructured data analysis. So, the validation and cleansing used to develop validation rules and vanishes any invalid data.

F. Aggregation and Representations

Sometime data may be spread across multiple datasets and requiring the data sets is joined together. Then data aggregation and representation technique is used to combining multiple datasets together to define common view. It is a time and effort rigorous.

G. Data analysis

Data analysis stage carried out the actual analysis task that is repeated until exact patterns or results inuncovered. This technique finds out the anomalies to exact statics and mathematical model of data.

H. Visualization and Utilization of analysis of results

Data visualization stage is used techniques and tools to communicate the analysis outcomes for better interpretation by users. Analysis results being made available to user for better decision making. It determines how and where processed data can be further grasped.

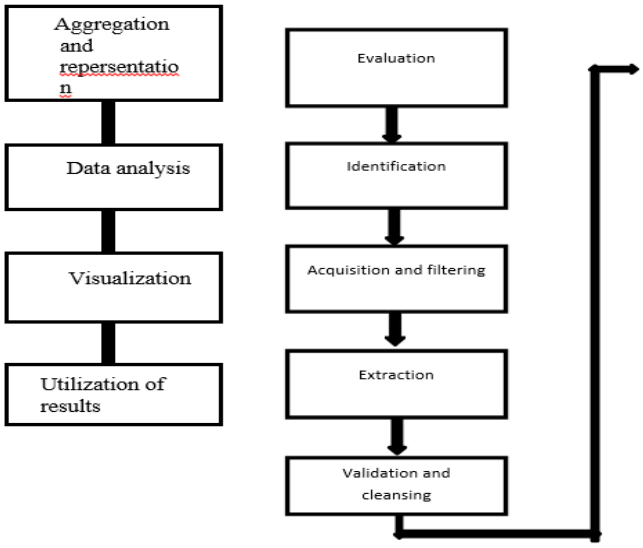

Fig.3: Data analytic lifecycle

V. HADOOP SYSTEM

Hadoop is a programming frame work used for large data processing on distributed environments. Hadoop is called Apache Hadoop. The various core components of Hadoop system is given below:

A. HDFS

HDFS is a distributed block structure basically used for big data management.The data in HDFS is strored in blocks. These blocks are called chunks.It is dividede into master-slave nodes. The master slave monitors the whole data thet are processed by slave nodes and slave node is comprises of actual data and file processing. Master node is called name node and slave node is called data node. The combination of both name node and data node is known as cluster.

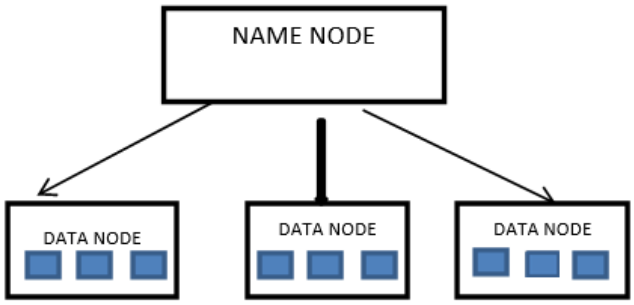

Fig.4: HDFS Architecture

B. Map Reduce

Map Reduce is the heart of Hadoops system. It is an important paradigm technique that allows huge scalability across large number of Hadoops clusters. Map Reduce provide parallel processing of data and useful for batch processing to store Zeta byte data for processing. It takes set of data and converts it into another dataset where the each and every element broken into key value pairs and then it reduce the task into single set of elements with single key value pair.

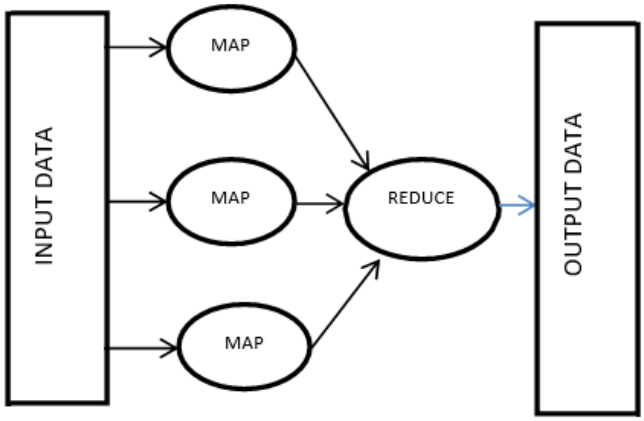

Fig.5: Map Reduce Architecture

C. D

YARN provides different data processing engines like batch processing, stream processing and graph processing to run and process data that is stored in HDFS. For large processing of data it is very useful because it provide better scheduling and management of data. YARN has application manager, node managers and resource managers that are useful for managing computer resources as well. YARN has high scalability, compatibility and multi tenancy.

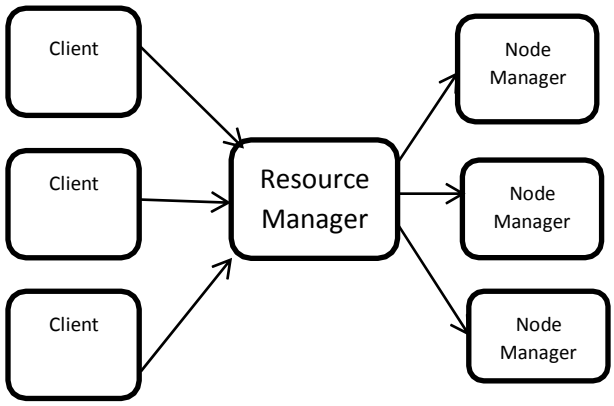

Fig.6 :YARN Architecture

VI. BIG DATA SECURITY CHALLENGES

Regular checking of data to detect security Is not sufficient. For this purpose, you need full time protection of data so; various techniques are using for security purposes that are:

A. Protecting transaction data

Data is stored in storage medium and continuously transfers of data between different locations. The size of data is also being

increased so the new challenges are being introduced for big data storage management.

B. Validation and Filtration of data

Storage and processing of data is done with the help of input data, which is provide by end points. Therefore, it should be necessary to use authentic devices for data security.

C. Protecting data in real time

Huge amount of data is generated so the maintenance of large amount of data is difficult. However, it is necessary to perform security checks in real time manner.

D. Protecting access control method and Encryption

Most of the data storage devices are unfortified so, it is mandatory to encrypt the access control methods for better data security.

E. Data Provenance

To assort data, it is important to be aware of its origins to discern the data accurate, authentic, validate and control access could be gained.

F. Granular access control and auditing

Granular access control of big data by NoSQL database and Hadoops distributed file system for great authentication task and regular auditing helps to identify any illegal attack.

VII. CONCLUSION

This paper is conjuring up an image of big data characteristics, big data analytic techniques, lifecycle of big data and security challenges. The main approach of this paper was an evolution of big data analytics. The main intend of this paper on tools and techniques and architectures and security challenges that perform exact data analysis and transfer of large volume of data. Hadoop system is an open source software that is used for large amount of data storage and processing and different components like HDFS, MAP-REDUCE, YARN produces data storage management techniques that transforms operational, financial and commercial problems into human interpretation constraints by discrete data sets. Another main concern of this paper on different methods used to save the data from different security issues.

VIII. REFRENCES

[1]. P-Joseph Charles and Carol, "Big Data security an Overview", International Research Journal of Engineering and Technology, vol 05, issue 2, Feb 2018.

[2]. Zahir Irani, "Critical analysis of big data challenges analytic methods", Journal of business research, 10 August 2018.

[3]. Snehalata Funde, Laxman Bharate, "Survey of Big Data Security", ISSN , vol.5, issue.2, 2018.

[4]. Dr. S. Vijayarani and Ms. S. Sharmila. "Research in big data-an overview", Informatics Engineering an International Journal, vol.4, no.3, Sep 2016.

[5]. Althaf Rehman.Sk and Sai Rajesh, "Challenging tools on research Issue in big data analytics", IJEDR, vol.6, issue.1, 2018.

[6]. Nada Elgendy and Ahmed Elragal "Big Data Analytics: A Literature Review Paper" Springer's International Publishing Switzerland 2014.

[7]. Sammidha Mukhrjee and Ravi Shaw "Big Data- Concepts, Applications, Challenges and Future Scope" IJARCCE, vol 5,Issue 2, February 2016.

[8]. Mashooque A. Memon and Muneer A. Kartio "Big Data Analytics and its Applications" AETiC, vol.1, No.1, 2017.

[9]. Xiaolong Jin and Benjamin W. Wah "Significance and Challenges of Big Data Research"vol.2, Issue 2,June 2015.

[10]. H.V jagdish "Big Data and Science: Myths and Reality" vol.2, Issue 2,June 2015.