ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 9, Issue 1, July 2019

A Review- Analysis of Big data Using Hadoop and Analytical Tools

Poonam Harode, N K Gupta

***Abstract*— We live in on-demand, on-command Digital universe with data generated by Institutions, Individuals and Machines at a very high rate. This data is categories as "Big Data" due to its large Volume, Variety and Velocity. Most of this data is unstructured, structured or semi structured and it is heterogeneous in nature. The volume and the heterogeneity of data with the speed it is generated, makes it difficult for the present computing infrastructure to manage Big Data. Traditional data management, warehousing and analysis systems fall short of tools to analyze this data. Map Reduce is a Minimization technique which makes use of file indexing with mapping, sorting, shuffling and finally reducing. Map Reduce techniques have been studied in this paper which is implemented for Big Data analysis.**

***Keywords*-- Hadoop, bigdata, mapreduce, HDFS.**

I. INTRODUCTION

Big data is a structured and unstructured data video, audio, picture and information emails etc. it is very large amount of data provided by social site and daily activities of social media like news and news channels or new technology, television, mobile, and computers and industries all are big data [8]. That's we can say it is more than thousands of information storage for the growth of the industries. Know if we have information or history of previous data than it's very easy for the next new changes for the industries or business. Today's competitive world in this time industries and business are growing very fastly by the help of the storage of the previous data which is known as big data. Big data is very hard to process and analysis the data easily. But with the help of HADOOP [5] data is easily to process and analysis the data easily. Big data is a different-different collection of complex data sets. Big data [10] is produced by different kinds of sources like television, mobile and other sources like industries Data records. It is three characteristics of big data:

l. volume
2. Velocity
3. Verity.

HADOOP

The Apache Hadoop [6] project develops open-source software for scalable, reliable, distributed computing. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using thousands of computational independent computers and large amount (terabytes, pet bytes) of data. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. Apache Hadoop is good choice for twitter analysis as it works for distributed huge data. Apache Hadoop is an open source framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the Map Reduce algorithm, where the data is processed in parallel on different clusters nodes. In short, Hadoop framework is able enough to develop applications able of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data. Hadoop Map Reduce is a software framework [8] for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

HADOOP ECOSYSTEMS

Hadoop system could be a framework of varied forms of advanced and evolving tools and parts that have skilful advantage in finding issues. a number of the weather could also be terribly dissimilar from one another in terms of their architecture; but, what keeps all along underneath one roof is that all of them derive their functionalities from the quantifiability and power of Hadoop. Hadoop system is alienated in four completely different layers: information storage, processing, information access, information management. All the parts of the Hadoop system, as express entities area unit evident. The holistic read of Hadoop design provides prominence to Hadoop common, Hadoop YARN, Hadoop Distributed File Systems (HDFS) and Hadoop Map Reduce of the Hadoop system. Hadoop common provides all Java libraries, utilities, OS level abstraction, necessary Java files and script to run Hadoop, whereas Hadoop YARN could be a framework for job programming and cluster resource management. HDFS in Hadoop design provides high turnout access to application information and Hadoop Map Reduce provides YARN primarily based data processing of huge information sets.

II. LITERATURE REVIEW

In [6] the contemporary world, Data analysis is a challenge in the era of varied inters- disciplines though there is a specialization in the respective disciplines. In other words, effective data analytics helps in analyzing the data of any business system. But it is the big data which helps and axial rates the process of analysis of data paving way for a success of any business intelligence system. With the expansion of the industry, the data of the industry also expands. Then, it is increasingly difficult to handle huge amount of data that gets generated no matter what's the business is like, range of fields from social media to finance, flight data, environment and health. Big Data can be used to assess risk in the insurance industry and to track reactions to products in real time. Big Data is also used to monitor things

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 9, Issue 1, July 2019

as diverse as wave movements, flight data, traffic data, financial transactions, health and crime. The challenge of Big Data is how to use it to create something that is value to the user. How can it be gathered, stored, processed and analyzed it to turn the raw data information to support decision making. In this paper Big Data is depicted in a form of case study for Airline data based on hive tools.

In [1] the research work is carried out using Apache pig and hadoop on a crime dataset. It describes the large volume of data yielded from multiple sources and termed it as voluminous data. Crime and crime related datasets with ever growing population has raised to a higher extent and is a attention seeking subject to government for taking strict measures by prevailing law and procedure. Big data analytics using pig and hadoop has been applied on this crime dataset with the idea behind it as the optimal improvement for analysing some trends that needs to figure out, so among the citizens of the country there could be a feel of security and safety. Also it could help the government to furnish law and procedure and welfare among the people of the country. Analysis results shows the total number of crimes occurred in every state, crimes that took place against women, type of crime and from year 2000 to 2014 the total number of crimes that took place. Experimental setup was pseudo distributed mode of hadoop and it was concluded that scripting language Pig Latin has fewer lines of code as compared to map reduce program but the execution time increases in pig as compared to map reduce. In [2], the author describes that Big data analytics has attracted intense interest from all academia and industry recently for its attempt to extract knowledge, information and wisdom form big data. Big data and cloud computing, two of the most important trends that are defining the new emerging analytical tools. Big data analytical capabilities using cloud delivery models could ease adoption for many industry, and most important thinking to cost saving, it could simplify useful insights that could providing them with different kinds of competitive advantage. Many companies to provide online Big Data analytical tools some of the top most companies like Amazon Big data Analytics Platform ,HIVE web based Interface, SAP Big data Analytics, IBM Info Sphere Big Insights, TERADATA Big Data Analytics, 1010data Big Data Platform, Cloud era Big Data Solution etc. Those companies analyze huge amount of data with help of different type of tools and also provide easy or simple user interface for analyzing data.

In [3], Information technology gives utmost importance to processing of data. Some pet bytes of data is not sufficient for storing large amount of data. Large volume of unstructured and structured data that gets created from various sources such as Emails, web logs, social media like Twitter, Facebook etc. The major obstacles with processing Big Data include capturing, storing, searching, sharing and analysis. Hadoop enables to explore complex data. It is an open source framework written in Java which supports parallel and distributed data processing and is used for reliable storage of data. With the help of big data analytics, many enterprises are able to improve customer retention, help with product development and gain competitive advantage, speed and reduce complexity. E-commerce companies study traffic on web sites or navigation patterns to determine probable views, interests and dislikes of a person or a group as a whole depending on the previous purchases. In this paper, they compare some typically used data analytic tools.

III. PROPOSED WORK

For analyzing these large and complex data required a power tool, we are using hadoop [6] which is a open source implementation of map reduce, a powerful tool designed for deep analysis and transformation of very large data.

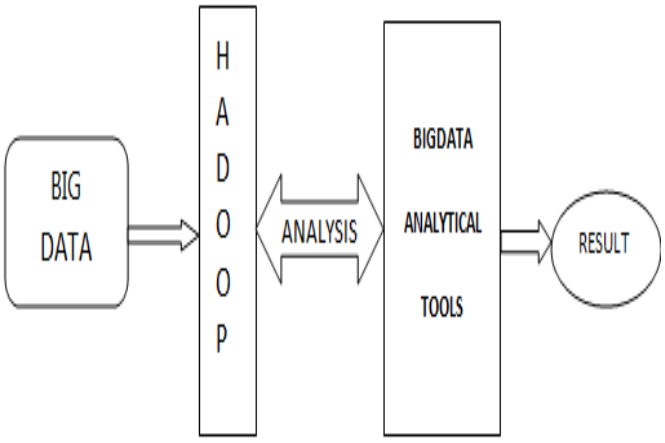

Fig 1. Workflow Diagram

This paper we design algorithm for handling the problems raised by the larger data volume and the dynamic data characteristics for finding and performing operation on social media data sets. For analysing first we used standard platform as hadoop on single node ubuntu machine to solve the challenges of big data through Map Reduce framework [12] where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling,After that we can use big data analytical tools to refine such huge data using big data analytical tools.

IV. PROPOSED METHODOLOGY

Our Steps or Algorithm Steps will follow:

1. First we can collect the big data and store it into HDFS.
2. After storing into HDFS (Hadoop Distributed File System), which is very reliable for storing such huge amount of data. After storing data into HDFS, we can pre-process the data using hadoop.
3. we can start analyzing such huge amount of data using big data analytical tools.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 9, Issue 1, July 2019

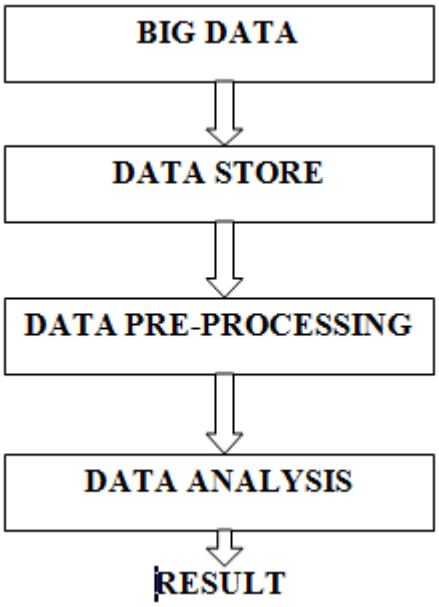

Fig 2. Analysis Steps

V. CONCLUSION

On analysing complete scenario regarding the analysis of big data we say that using traditional analytical tool we cannot perform analysis on such huge and complex data , so we uses a new powerful tool which is designed for deep analysis called hadoop and also integrate with its ecosystems. And using these various ecosystems we can easily analyse the large and complex data.

REFERENCES

[1] Arushi Jain, Vishal Bhatnagar, "Crime Data Analysis Using Pig with Hadoop" in International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA, in ELSEVIER 2015.

[2] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, "Big Data and Advanced Analytics Tools", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[3] Mrunal Sogodekar, Shikha Pandey, Isha Tupkari, Amit Manekar, "BIG DATA ANALYTICS: HADOOP AND TOOLS", in 978-1-5090-2730-9/16, 2016 IEEE.

[4] Manjunath T. N., Srividhya, "Analysis of Airport Data using Hadoop-Hive: A Case Study" in International Journal of Computer Applications (0975 – 8887) National Conference on "Recent Trends in Information Technology" (NCRTIT-2016).

[5] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.

[6] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/.

[7] Sagiroglu, S., & Sinanc, D, "Big data: A review", IEEE International Conference on Collaboration Technologies and Systems (CTS), 2013, pp 42-47.