International Journal of Science, Technology and Management (IJSTM) ISSN (online): 2321-774X

Volume 9, Issue 1, 2022

BIG DATA MINING THROUGH HADOOP

Tinky Singh

tinkysingh11@gmail.com

M.Tech Scholar, Department of Computer Science & Engineering, BRCM CET, Bahal, (Haryana)

Amit Ranjan

tmitrnjn87@gmail.com

Assistant Professor, Department of Computer Science & Engineering, BRCM CET, Bahal, (Haryana)

ABSTRACT

The capacity of human brain is about 2.5 petabytes, which is also the estimated size of walmart databases that handle 1 million transactions a day. The amount of data that gets created every day will only continue to increase. The rate of data generation is so fast that there will be a need to implement cost effective and easy data storage and retrieval mechanisms. For such large amount of data, the term used BIG DATA, emerged with new opportunities and challenges. By analyzing the data, we are able to get useful information for the organizations. Big Data can be structured through HADOOP which is an open source software project. This paper is on Big Data, Advantages and Data Mining of Big Data. A coin has two faces, similarly Big Data also has, one is opportunities and other is challenges to the researchers. An overview of opportunities in different field like healthcare, technology etc is given. The paper consist basic understanding of Big Data and usefulness for organization. Also, include Hadoop Architecture and its basic functionalities. This contains descriptions on the HDFS and Map Reduce framework. This paper also has application of Big Data in Data Mining.

Keywords: Big Data, Hadoop, Data Mining

INTRODUCTION

BIG DATA is a large capacity of data sets structured, semi-structured or unstructured that is being generated from different sources at startling rate. Key facilitators for the growth of BIG DATA are increasing processing power, increasing storage volume and availability of data. It is thus necessary to develop techniques for easy storage and retrieval. Data can be induced on web in several forms like texts, images, audios, videos or social media posts. For such big amount of data to process in economical and efficient way, parallelism is used. Velocity, Volume, Variety and Veracity are the main properties of BIG DATA.

Volume

Data is being generated at a startling rate. The total capacity of data being produced makes the issue of data processing a tedious task. The storage and retrieval of data collected from multiple sources has become easier with the help of technologies such as Hadoop.

Velocity

Velocity means speed of data coming from different sources. This characteristic is not being restricted to the speed of incoming data but also speed at which data pass and accumulated.

Variety

Data is being produced from different sources, including stock exchange data, social media data and black box data. The data can presume various forms-text, audio, numerical, video etc. On twitter 500 million tweets are sent per day and there are 330 million active users on it.

Veracity

Veracity means unpredictability or precision of data. Data is obscure due to the inadequacy.

CHALLENGES AND OPPORTUNITIES

Approx 850 million web pages on internet provide information of Big Data. Big Data comes with a lot of good fortune to deal in health, education, earth and businesses but to deal with the data having large capacity using old models become laborious. So, it's essential to look into challenges and work out on models for storage and retrieval of data.

Challenges

1) **No uniformity and in completeness:** If we would like to look at the information, it ought to be organized however once we deal with the massive information, information could also be organized or world organization organized further. No uniformity is that the massive challenge in information Examination and predictors have to be compelled to manage with it.
2) **Scale:** As the name pronounces massive information has massive size of information sets. Handling with massive information sets is a massive drawback from decades. Earlier, this drawback was resolved by the computers obtaining previous however now information volumes have become monumental and processors are static. Universe is moving on the thanks to the Cloud technology, thanks to this variation information is made terribly very high rate. This high rate of growing information is becoming a difficult drawback to the information authorities
3) **Appropriateness:** Further challenge with size is speed. If the information gatherings are massive in size, longer the time it'll go for examine it. Any methodology that deals with efficiency with the scale is probably going to accomplish well in term of speed.
4) **Secrecy:** Secrecy {of information of knowledge|} is another massive drawback with massive data. In some countries there are strict acts regarding the data secrecy, as an example in USA there are strict acts for fitness records, except for others it's less powerful.
5) **Human Cooperation:** In malice of the advanced procedure replicas, there are several styles that a computer cannot notice. A replacement technique of harnessing human cleverness to unravel drawback is crowd-tracking. Wikipedia is that the best example. We have a tendency to be dependable on the data given by the strange person, however most of the time they're correct. However, there will be others with different functions further as like providing false info. We'd like technical model to handle with this. As humans, we will look the appraisal of book and realize that some are positive and a few are negative and are available up with a conclusion to whether obtain or not. We'd like ways to be that intelligent to determine

Opportunities

The technology like Big Data is having a very important a part of every field like political economy, health, Education, banking, and corporate as well as in government.

1) **Technology:** Nearly every high business like Facebook, IBM, and yahoo have accepted huge information and are exploiting on huge information. Facebook manages fifty Billion photos of shoppers. Every month Google manages a hundred billion quests.
2) **Government:** Huge information is wont to grip the difficulties handled by the govt. Obama government exposed huge information analysis and growth creative

thinking in 2012. Big data exploration vie a very important a part of BJP winning the elections in 2014 and Indian government is placed on huge information analysis in Indian body.

3) **Care:** Rendering to IBM huge information for Healthcare, eightieth of medical aid information is shapeless. Healthcare organizations are familiarizing huge information technology to urge the total info a few patient. To increase the care and low down the charge huge information examination are necessary and sure technology ought to be improved.
4) **Science and Research:** Huge information could be a current topic of analysis. Varied scientists are engaged on huge data.
5) **Media:** Media is exploitation huge information for the advertisings and selling of merchandise by targeting the interest of the user on net. For instance social media posts, information predictors get the quantity of posts and then examine the eye of user. It also can be complete by obtaining the positive or negative evaluations on the social media.

HADOOP FRAMEWORK

Hadoop is free software used to process the Big Data. It is very prominent used by administrations/researchers to analyze the Big Data. Hadoop is swayed by Google's structural design; Google file system and map reduce. Hadoop processes the large data sets in a divided computing environment. An apache Hadoop ecosystem comprised of Hadoop Kernel, Map Reduce, HDFS and other components like Apache Hive, Base and Zookeeper.

Hadoop has two main units:

1] HDFS (Hadoop Distributed file system)

It's a divided file system which has capacity of tolerating fault and modeled to run on commodity hardware. HDFS provides high work rate to application data and is fit for applications that have large data sets. HDFS can keep data across thousands of servers. HDFS has master/slave architectonics. Files added to HDFS are divided into fixed size section. Section size is configurable.

2] Map Reduce

It is a programming model launched by Google in 2004 for simply writing application which processes huge amount of data in parallel on large collection of hardware in defect tolerant manner. This work on hug data sets, divide the problem and data sets and run it in parallel.

Two basis in map reduce are as following:

- **Map:** The map function work as to filter, transform or parse the data. The result from map becomes the input to reduce.
- **Reduce:** The reduce function is used to condense data from the map function.

IMPLEMENTATION IN DATA MINING

Data mining is finding of valuable information from raw data. There are many technologies used for data mining – Presto, Rapid miner, Elastic search etc.

Presto

It is a distributed SQL query engine that's helpful in operating analytic queries against a variety of data sources e.g. Cassandra, Hadoop, MySQL and MongoDB. One of the advantages of presto is that it allows users to query data from multiple sources through one query.

Rapid miner

This is a centralized software package for mining data and running foresight analytics. Users can enter huge volumes of raw data e.g. databases and text for immediate and intelligent analysis. The advantages of rapid miner are user friendliness and affordability.

Elastic search

It is a whole text search and analytics engine that allows users to keep, search and analyze large data volumes in near real time. The advantages of elastic search are rapid search and the potential to filter large data sets.

These are the types of analysis done after mining:

Ordering Analysis

It is a sorted process for finding important information about data. Ordering can also be used to bundle the data.

Bundle Analysis

It is the process to identify datasets that are alike to each other. This is done to get the resemblance and contrast within the data. For example, clusters of customers having alike preferences can be selected on social media.

Evolution Analysis

It is also called as hereditary data mining used to mine data from DNA sequencing, but can be used in banking to observe the stock exchange by previous years' time series data.

Outlier Analysis

Some inspection, recognition of units is done which do not make a form in a data set. In medical and banking problems this is helpful.

LITERATURE REVIEWS

1] Big Data stands for set of numerical data generated by the use of new technologies for different purposes. In this paper we have studied Big Data features and discussed the challenges of Big Data quality might affect induced by Big Data. Big Data analytics is a rapid growing technology.

2] In recent years' data are produced at a rapid pace. Analyzing theses data is difficult for a general man. The author studied the various research issues, challenges and tools used to analyze these Big Data. From this studied, it's acknowledge that every Big Data platform has its individual attention. Some of them are planned for batch processing whereas some are best at real time analytic.

3] In this research paper, we have studied the innovative topic of Big Data, which has recently gained lots of interest due to its perceived unprecedented chances and profits.

4] We have entered in generation of Big Data. The paper consists the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data. The paper emphasized on Big Data processing problems. These technical challenges must be handled for fast and efficient processing of Big Data. Then, the challenges of handling big data are analyzed, followed by the limitations of using the traditional big data processing approach. The paper represents Hadoop which is free software used for processing of Big Data.

5] The concept of Big Data is as prominent as its meaning is uncertain. With this article we have clarified the necessary features of Big Data, mainly: Information, Technology, Methods and Impact. For each of them the paper put forward an exploration of the main research areas, bringing understandable examples across different areas.

CONCLUSION

In this review paper, an overview is delivered on Big Data, Hadoop and applications in processing. 4 V's of Big Data info has been mentioned. The paper has challenges and opportunities of Big Data. This paper consists the Hadoop Framework and its mechanisms HDFS and Map reduce. The paper further has implementation of data mining and types of analysis.

FUTURE SCOPE

This research can extend to different sectors- Health, Insurance, Energy and Traffic System etc.

REFERENCES

[1] Abdulbaset Salem Albaour and Yousof Abdulrahman Aburawe, Big Data: Review Paper, volume 7, issue 1, 2021, IJARIIE-ISSN (O)-2395-4396.

[2] D.P. Acharjya and Kauser Ahmed P, A Survey on Big Data Analytics: Challenges, Open Research Issues And Tools, IJACSA, volume 7, no. 2, 2016.

[3] Nada Elgendy and Ahmed Elragal, Big Data Analytics: A Literature Review Paper, ICDM 2014.

[4] IBM Big Data analytics HUB, www.ibmbigdatahub.com/infographic/four-vs-big-data

[5] Apache Hadoop Project, http://hadoop.apache.org/, 2013.

[6] Mucherino A. Petraq papajorgji P.M.Paradalos 1998. A survey of data mining techniques allied to agriculture CRPIT.3 (3): 555560.

[7] MIT Technology Review, http://www.technologyreview.com/view/535451/data-mining-indian-recipesreveals-new-food-pairing-phenomenon/.

[8] Vidyasagar S. D, A Study on "Role of Hadoop in Information Technology era", GRA - GLOBAL RESEARCH ANALYSIS, Volume: 2 | Issue: 2 | Feb 2013 • ISSN No 2277 –8160.