Big Data And Hadoop: A Review Paper

Article · January 2015

CITATIONS
14

READS
13,554

1 author:

Rahul Beakta
Baddi University of Emerging Sciences and Technologies
1 PUBLICATION **14** CITATIONS

SEE PROFILE

Big Data And Hadoop: A Review Paper

Rahul Beakta
CSE Deptt., Baddi University of Emerging Sciences & Technology, Baddi, India
rahulbeakta93@gmail.com

***Abstract*— In this world of information the term BIG DATA has emerged with new opportunities and challenges to deal with the massive amount of data. BIG DATA has earned a place of great importance and is becoming the choice for new researches. To find the useful information from massive amount of data to organizations, we need to analyze the data. Mastery of data analysis is required to get the information from unstructured data on the web in the form of texts, images, videos or social media posts. This paper presents an overview on Big Data, Advantages and its scope for the future research. Big Data present opportunities as well as challenges to the researchers. An overview on opportunities to healthcare, technology etc. is given. This paper gives an introduction to Hadoop and its components. This paper also concentrates on application of Big Data in Data Mining.**

Keywords*— *big data; Hadoop; Map Reduce; HDFS; data mining.

I. INTRODUCTION

BIG DATA is a vague topic and there is no exact definition which is followed by everyone. Data that has extra-large Volume, comes from Variety of sources, Variety of formats and comes at us with a great Velocity is normally refer to as Big Data. Big data can be structured, unstructured or semi-structured, which is not processed by the conventional data management methods. Data can be generated on web in various forms like texts, images or videos or social media posts. In order to process these large amount of data in an inexpensive and efficient way, parallelism is used [1].

There are four characteristics for big data. They are Volume, Velocity, Variety and Veracity.

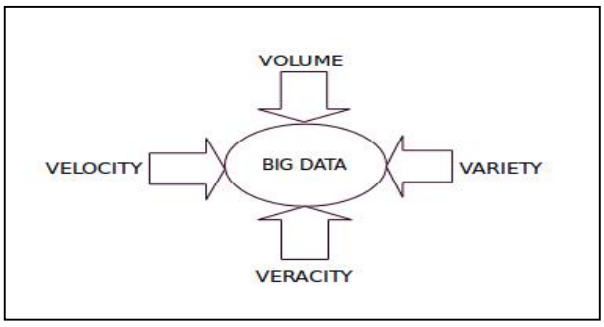

Fig. 1. 4 V's of BIG DATA.

Volume means scale of data or large amount of data generated in every second. Machine generated data are examples for these characteristics. Nowadays data volume is increasing from gigabytes to petabytes [2]. 40 Zettabytes of data will be created by 2020 which is 300 times from 2005 [3]. Second characteristic of Big Data is velocity and it means analysis of streaming data.

Velocity is the speed at which data is generated and processed. For example social media posts [2].

Variety is another important characteristic of big data. It refers to the type of data. Data may be in different forms such as Text, numerical, images, audio, video, social media data [2]. On twitter 400 million tweets are sent per day and there are 200 million active users on it [3].

Veracity means uncertainty or accuracy of data. Data is uncertain due to the inconsistency and incompleteness [2].

II. CHALLENGES AND OPPORTUNITIES

There are 800 million web pages on Internet giving information about Big Data. Big Data is the next big thing after Cloud [11]. Big data comes with a lot of opportunity to deal in health, education, earth, and businesses but to deal with the data having large volume using traditional models becomes very difficult. So we need to look on big data challenges and design some computing models for efficient analysis of data [13].

A. Challenges with Big Data: [12]

1) Heterogeneity and Incompleteness:

If we want to analyze the data, it should be structured but when we deal with the Big Data, data may be structured or unstructured as well. Heterogeneity is the big challenge in data Analysis and analysts need to cope with it. Consider an example of patient in Hospital. We will make each record for each medical test. And we will also make a record for hospital stay. This will be different for all patients. This design is not well structured. So managing with the Heterogeneous and incomplete is required. A good data analysis should be applied to this.

2) Scale:

As the name says Big Data is having large size of data sets. Managing with large data sets is a big problem from decades. Earlier, this problem was solved by the processors getting faster but now data volumes are becoming huge and processors are static.

World is moving towards the Cloud technology, due to this shift data is generated in a very high rate. This high rate of increasing data is becoming a challenging problem to the data analysts.

Hard disks are used to store the Data. They are slower I/O performance. But now Hard Disks are replaced by the solid state drives and other technologies. These are not in slower rate like Hard disks, so new storage system should be designed.

3) Timeliness:

0 Another challenge with size is speed. If the data sets are large in size, longer the time it will take to

analyze it. Any system which deals effectively with the size is likely to perform well in term of speed. There are cases when we need the analysis results immediately. For example, If there is any fraud transaction, It should be analyzed before the transaction is completed. So some new system should be designed to meet this challenge in data analysis.

4) Privacy:

Privacy of data is another big problem with big data. In some countries there are strict laws regarding the data privacy, for example in USA there are strict laws for health records, but for others it is less forceful. For example in social media we cannot get the private posts of users for sentiment analysis.

5) Human Collaborations:

In spite of the advanced computational models, there are many patterns that a computer cannot detect. A new method of harnessing human ingenuity to solve problem is crowd-sourcing. Wikipedia is the best example. We are reliable on the information given by the strangers, however most of the time they are correct. But there can be other people with other motives as well as like providing false information. We need technological model to cope with this. As humans, we can look the review of book and find that some are positive and some are negative and come up with a decision to whether buy or not. We need systems to be that intelligent to decide.

B. Opportunities to Big Data: [14]

Now this is Data Revolution time. Big Data is giving so many opportunities to business organizations to grow their business to higher profit level. Not only in technology but big data is playing an important role in every field like health, economics, banking, and corporates as well as in government.

1) Technology:

Almost every top organization like Facebook, IBM, yahoo have adopted Big Data and are investing on big data. Facebook handles 50 Billion photos of users. Every month Google handles 100 billion searches. From these stats we can say that there are a lot of opportunities on internet, social media.

2) Government:

Big data can be used to handle the problems faced by the government. Obama government announced big data research and development initiative in 2012. Big data analysis played an important role of BJP winning the elections in 2014 and Indian government is applying big data analysis in Indian electorate

3) Healthcare:

According to IBM Big data for Healthcare, 80% of medical data is unstructured. Healthcare organizations are adapting big data technology to get the complete information about a patient. To improve the healthcare and low down the cost big data analysis are required and certain technology should be adapted.

4) Science and Research:

Big data is a latest topic of research. Many researchers are working on big data. There are so many papers being published on big data. NASA center for climate simulation stores 32 petabytes of observations [15].

5) Media:

Media is using big data for the promotions and selling of products by targeting the interest of the user on internet. For example social media posts, data analysts get the number of posts and then analyze the interest of user. It can also be done by getting the positive or negative reviews on the social media.

III. HADOOP FRAMEWORK

Hadoop is open source software used to process the Big Data. It is very popular used by organizations/researchers to analyze the Big Data. Hadoop is influenced by Google's architecture, Google File System and MapReduce. Hadoop processes the large data sets in a distributed computing environment. An Apache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and other components like Apache Hive, Base and Zookeeper [1]

A. Hadoop consists of two main components:

1) Storage: The Hadoop Distributed File System (HDFS): It is a distributed file system which provides fault tolerance and designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS can store data across thousands of servers. HDFS has master/slave architecture [5]. Files added to HDFS are split into fixed-size blocks. Block size is configurable, but defaults to 64 megabytes.

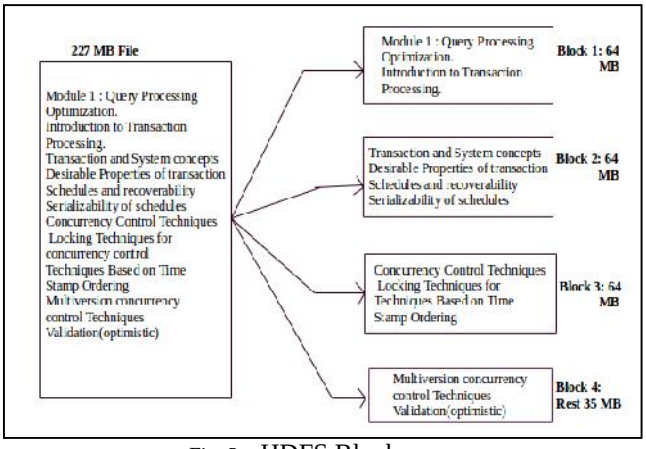

Fig. 2. HDFS Blocks.

2) *Processing: MapReduce* [*4*]: It is a programming model introduced by Google in 2004 for easily writing applications which processes large amount of data in parallel on large clusters of hardware in fault tolerant manner. This operates on huge data set, splits the problem and data sets and run it in parallel.

Two functions in MapReduce are as following:

a) Map – The Map function always runs first typically used to filter, transform, or parse the data. The output from Map becomes the input to Reduce.

b) **Reduce** – The Reduce function is optional normally used to summarize data from the Map function.

IV. APPLICATIONS IN DATA MINING

Big Data is very useful for Business Organizations as well as to the researchers to observe the data patterns in big data sets. Extracting useful information from large amount of big data is called as Data Mining. There is huge amount of data on Internet in form of text, numbers, social media posts, images and videos. 40 Zettabytes of data will be created by 2020 which is 300 times from 2005 [3]. To analyze this data to get useful information for security, health, education etc., we need to introduce new data mining system which is effective. There are many Data mining techniques which can be used with big data, some of them are:

A. Classification Analysis:

It is a systematic process for obtaining important information about data and metadata. Classification can also be used to cluster the data.

B. Cluster Analysis:

It is the process to identify data sets that are similar to each other. This is done to get the similarities and differences within the data. For example clusters of customers having similar preferences can be targeted on social medial [6].

C. Evolution Analysis:

It is also called as genetic data mining mainly used to mine data from DNA sequences. But can be used in Banking, to predict the Stock exchange by previous years' time series Data [7].

D. Outlier Analysis:

Some observations, identifications of items are done which do not make a pattern in a Data Set. In medical and banking problems this is used.

V. LITERATURE REVIEWS

Anupam Jain, Rakhi N K and Ganesh Bagler studied Indian Recipes and discovered that the presence of certain spices makes a meal much less likely to contain ingredients with flavors in common. Jain and others chose an online website TarlaDalaa.com and downloaded more than 2500 recipes for their research. 194 different ingredients were found in these recipes. Then they studied Network of links between these recipes. They found that Indian cuisine is characterized by strong negative food pairing that even higher than any before. According to them, "Our study reveals that spices occupy a unique position in the ingredient composition of Indian cuisine and play a major role in defining its characteristic profile". "Our study could potentially lead to methods for creating novel Indian signature recipes, healthy recipe alterations and recipe recommender systems," conclude Jain and mates [8,9].

Vidyasagar S. D did a survey on Big Data and Hadoop system and found that organizations need to process and handle petabytes of Data sets in efficient and inexpensive manner. According to him if there is any node failure then we can lose some information. Hadoop is an Efficient, reliable, Open Source Apache License. Hadoop is used to deal with large data sets. Author explained its need, uses and application. Now days, Hadoop is playing an important role in Big Data. Vidyasagar S.D concluded that "Hadoop is designed to run on cheap commodity hardware, it automatically handles data replication and node failure, it does the hard work – you can focus on processing data, Cost Saving and efficient and reliable data processing" [10].

VI. CONCLUSION

In this review paper, an overview is provided on Big Data, Hadoop and applications in Data Mining. 4 V's of Big Data has been discussed. An overview to big data challenges is given and various opportunities and applications of big data has been discussed. This paper describes the Hadoop Framework and its components HDFS and Map reduce. The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. Hadoop plays an important role in Big Data. This paper also focuses on current researches in Data Mining and some literature reviews have also been studied.

REFERENCES

[1] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

[2] SMITHA T, V. Suresh Kumar "Application of Big Data in Data Mining" in International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 7, July 2013).

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

[4] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" in International Journal of Advance Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

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

[6] Smitha.T, Dr.V.Sundaram, "Classification Rules by Decision Tree for disease prediction" International journal for computer Application, (IJCA) vol 43, 8, No-8, April 2012 edition. ISSN0975-8887; pp- 35-37

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

[8] Anupam Jain, Rakhi N K and Ganesh Bagler, arxiv.org/abs/1502.03815 Spices Form The Basis Of Food Pairing In Indian Cuisine.

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

[10] 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.

[11] BIG DATA: Challenges and opportunities, Infosys Lab Briefings, Vol 11 No 1, 2013.

[12] Divyakant Agrawal, Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the United States.

[13] Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challenges and Solutions in International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[14] Big Data, Wikipedia, http://en.wikipedia.org/wiki/Big_data

Webster, Phil. "Supercomputing the Climate: NASA's Big Data Mission". *CSC World*. Computer Sciences Corporation. Retrieved 2013-01-18.