Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.S1, 2019, pp.1-3

DOI: https://doi.org/10.51983/ajcst-2019.8.S1.2017

Social Media Analysis through Big Data Using Map Reduce Algorithm

S. Lingeswari
Assistant Professor, Department of Computer Science,
P.K.R. Arts College for Women, Tamil Nadu, India
E-Mail: lingesmsc@gmail.com
(Received 2 November 2018; Revised 5 December 2018; Accepted 2 January 2019; Available online 12 January 2019)

***Abstract* - Few years back the Internet usage was very low when compared now-a-days. It has become a very important part in our day to day life. Billions of people are using social media and social networking every day all over the world. Such a huge number of people generate a large number of data which have become a quite difficult to manage. Here solving these types of problem by using a term called Big Data. It refers to the huge number of datasets. Data may be structured, unstructured or semi structured. Big data is defined by three Vs such as Volume, Velocity and Variety. Big Data use an algorithm known as Map Reduce algorithm. Large number of datasets is very difficult to manage. This problem has been solved using Map Reduce algorithm. In this paper, we focus to analyze social media through big data using Map Reduce algorithm.**
Keywords: **Big Data, Hadoop, Map Reduce, Social media**

I. INTRODUCTION

Big data is a collection of large datasets. It cannot be processed by using traditional computing software. Data can be stored, accessed and processed in the form of fixed format is termed as structured data. Data stored in a relational database management system is one of example of structured data. Consider example as employee table. Table consists of number of fields such as emp-id, emp-name, gender, dept, salary is format of structured Data. Data with unknown form or the structure is classified as unstructured data. Example of unstructured data is a heterogeneous data source such as satellite images, scientific data, sensor data, images, video, audio, etc. Semi-structured Data is information that does not reside in a relational database, but it does have some organizational properties that make it easier to analyze.

A. Three Vs in Big Data

Big Data [1] is defined by three Vs namely volume, velocity and variety.
Volume: It refers to the amount of data is generated. This data can be low density, high volume, structures/ unstructured data with unknown value. This unknown value is converted into useful one using technologies like Hadoop. The data can range from terabytes to penta bytes. Velocity: It refers to the rate at which the data is generated. The data is received at an unprecedented speed and is acted upon in a timely manner. It also requires real-time evaluation and action in case of the internet of things applications. Variety: It refers to different formats of data. It may be structured, unstructured or semi-structured. The can be audio, text, video, sensor data. In this additional processing is required to derive the meaning of data and also to support the metadata. In addition to these three Vs of data, following Vs are also defined in big data.

Value: Each form of data has some value which needs to be discovered. These are certain qualitative and quantitative techniques to derive meaning the data. For deriving value from data, certain new discoveries and techniques are required. Variability: Another dimension for big data is the variability of data. It means the flow of data may be high or low. There are challenges in managing this flow of data.

II. HADOOP

Hadoop [7] is open source software for reliable, scalable, distributed computing. This software allows for the distributed processing of datasets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machine. Each offering local computation and storage. It is developed by using Java language. Hapoop consists of a storage part known as Hadoop distributed file system (HDFS) and a processing part which is a Map Reduce programming model. It is used for processing and generating big datasets with a parallel computing, distributed computing algorithm on a cluster. A Map Reduce program creates a procedure (method) and a reduce method. Procedure which performs sorting and filtering the data in the big data set. Reduce method, which performs summary operations such as counting, finding sum and average the records.

A. Modules of Hadoop Framework

In a fig -1 shows modules of Hadoop framework. Hadoop common: It contains libraries and utilities needed by other Hadoop modules. Hadoop Distributed File Systems (HDFS): A distributed file system that stores data on commodity machines, providing very high aggregate bandwidth across the clusters.

1. *Hadoop Yarn:* It is a platform responsible for managing computing resources in cluster and using them for scheduling user's applications.
2. *Hadoop Map Reduce:* An implantation of map reduce programming model for large scale data processing.

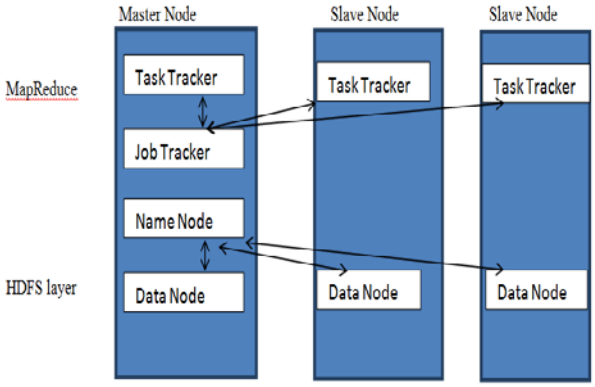

Fig. 1 Hadoop Framework

III. MAP REDUCE ALGORITHM

Traditional techniques cannot be suitable for processing a huge amount of data. Google solved the problem using an algorithm called Map Reduce [2]. Map Reduce divides a task into small parts and assigns them to many computers. Later the results are collected at one place and integrated to form the result dataset. This algorithm contains two important tasks, namely Map and Reduce. The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples. The Reduce task takes the output from the Map as an input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.

A. Phases of Map Reduce Algorithm

1. Input Phase: Here we have a record reader that translates each record in an input files and sends the parsed data to the mapper in the form of key-value pairs.
2. Map phase: Map is a user-defined function, which takes a series of key-values pairs and processes each one of them to generate zero or more key-value pairs.
3. Intermediate Keys: The key-value pairs generated by the mapper are known as intermediate keys.
4. Combiner: A combiner is a type of local reducer that groups similar data from the map phase into identifiable sets. It takes intermediate keys from the mapper as input and applies a user-defined code to aggregate the values in a small scope of one mapper. It is not a part of the main Map Reduce algorithm, it is optional.

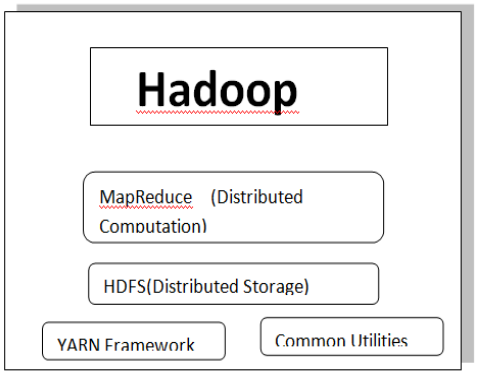

Fig.2 Map Reduce algorithm

5. Shuffle and Sort: The Reducer task starts with the shuffle and sort step. It downloads the grouped key-value pairs onto the local machine, where the Reducer is running. The individual key-value pairs are sorted by key into a larger data list.
6. Reducer: The Reducer takes the grouped key-value paired data as input and runs a Reducer function on each one of them. Here, the data can be aggregated, filtered and combined in a number of ways and it requires a wide range of processing. Once the execution is over, it gives zero or more key-value pairs to the final step.
7. Output phase: In the output phase, we have an output format that translates the final key-value pairs from the reducer function and writes them onto a file using a record writer.

IV. SOCIAL MEDIA IN BIG DATA

Social media [7] is computer based technology which facilitates the sharing of ideas, thoughts and information through the building of virtual network and communities. Social media is an internet based and users are allowed to access the information quickly. Information includes personal details, documents, videos, audios and photos. Users engage with social media via computer, tablet, and smart phone via web based software or web applications often utilizing it for message. Social media originated as a way to interact with friends, family and business advertisement. The advantage of social media is the ability to connect and share information with many people simultaneously. Some of the most popular social media websites are Face book (2.167 billion users), YouTube (1.5 billions of users), what's app (1.3 billions of users), Face book Messenger (1.3 billion of users) and Instagram (800 million of users).

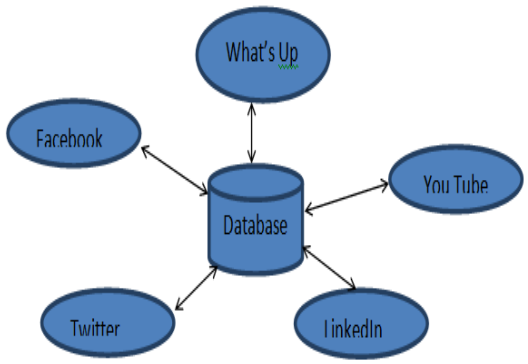

Fig. 3 Social Media

Social media is a form of electronic media through users can create online communities to share ideas, personal information, images, videos, audios and message. Statics shows that five hundred and above terabytes of new data gets ingested into the database of social media site facebook, every day. This data is mainly generated in terms of photos, videos, message exchange, putting comments,

etc. Ten terabytes of data need for every day in Google classroom. Managing these types of databases by using traditional method is very difficult. This problem has been solved using Map Reduce algorithm in Big Data.

V. CONCLUSION

A large amount of data cannot be processed using traditional data processing software. Here we use a new term called as Big Data. Big data is a collection of large datasets. Big Data use a new tool called Hadoop. It is open source software for reliable, scalable, distributed computing. This software allows for the distributed processing of datasets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machine. Each offers local computation and storage. Hadoop use an algorithm known as Map Reduce algorithm. In this paper, analyze social media through Big Data using Map Reduce algorithm. Map Reduce divides a task into small parts and assigns them to many computers. Later the results are collected at one place and integrated to form the final result dataset.

REFERENCES

[1] Blokdijk Gearad, *Big Data Analytics – Simple steps to Win, Insights and Opportunities for Maxing out Success*, Complete publishing.
[2] DT Editorial Services, *Big Data, Black: Covers Hadoop 2, Map Reduce, Hive, Yarn, Pig, R and Data Visualization*, Paperback, 2016.
[3] David Loshin, *Big Data Analytics*, Morgon Kaufmann.
[4] Judith Hurwitz, *Big Data for Dummies*, Wiley Publishers.
[5] Dirk DeRoos, Paul C. Zikopoulos, Roman B.Melnyk, Phd., Bruce Rrown, Rafael Coss, *Hadoop for Dummies*, Wiley Publishers.
[6] Robert D. Schneider, "*Hadoop for Dummies*, John Wiley & Sons publishers, Canada, 2012.
[7] Social media Website, [Online] Available at: https://smallbiz trends.com/2016/05/popular-social-media-sites.html.