ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Web History Analysis Using MapReduce Framework

Aatmaja Kulkarni[1], Gayatri Shiras[1], Samrudhhi Deshmukh[1], Ketaki B. Naik[2]

Students, Dept of Information Technology, Bharti Vidyapeeth's College of Engineering For Women

Pune, India[1]

Associate Professor, Department of Information Technology, Bharti Vidyapeeth's College of Engineering For Women,

Pune, Savitribai Phule Pune University, Pune India.[2]

ABSTRACT: Big Data is a term for datasets that are so large or complex that traditional data processing applications are inadequate. Big data analytics is the process of examining large data sets containing a variety of data types, in an effort to uncover hidden patterns, unknown correlations and other useful information. Hadoop, the software framework for computing large amount of data. The project includes following modules of Hadoop: HDFS (Hadoop Distributed File System) and Hadoop Map Reduce. Proposed system used to find out most searched data after firing a query by user by analyzing web browsing history of different people using Hadoop framework. System will predict the most trending topic for constraints like city, university, food, personality on basis of the surfing history of a search engine with respect to different people. The proposed architecture can be used to make estimation of trending results of given constraint for future user. Mostly different analytics method or filter methods was used before, for getting predictions. Map reduces runs parallel that's why we get output fast and also handle large dataset so scalability is really very good than any other technique.

KEYWORDS: Big Data, Hadoop, MapReduce, HDFS, Search engine dataset

I. INTRODUCTION

The big data is data of different forms obtained from different sources [1]. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. The same amount was created in every two days in 2011, and in every Five minutes in 2015. This rate is still growing enormously. This data is generated from banking transaction, social networking [2], SMS, records of government, study material and much more [3].Data is stored on vast scale database to perform computing and get the required result. [4] The result may be used for predicting trend in the society, correcting sales and marketing strategy and even in educational context to study trends for elective allocations. There are many frameworks to perform operation on big data. One of them is Apache Hadoop framework [5]. This paper explores the application of Hadoop to analyze trends in a dataset. This paper proposes a key-value representation for analysis of dataset through Hadoop .Core technology of Hadoop framework is MapReduce. MapReduce is two stages process. In first stage, it maps the record in key value pairs. Then in second stage it reduces the same key-value pairs to one set. The Hadoop stores these key-value pairs in Hadoop distributed file system. The system will work on history of search engine. It will take the browsing history of a search engine with respect to different people as an input and predict the most trending topic on basis of constraints like city, university, food, personality.

II. HADOOP AND MAPREDUCE

The two main components of Hadoop are HDFS and MapReduce. Hadoop Distributed File System (HDFS) is used for reliable data storage. MapReduce is a framework which writes an application for processing large amounts of both unstructured and structured data in parallel across a cluster of machines, in a fault-tolerant and reliable manner. There are three major categories of machine roles in a Hadoop deployment. They are Name Node, Data Node and Client machines.

A. Name Node

Name Node works as the master node which stores the file system metadata i.e. keeps the track of which file and blocks are to be stored on which Data Node. All information are stored in RAM.

B. Data Node

The Data Node works as the slave on which the actual data resides. To indicate its presence in the system. In order to keep the replication high and to rebalance the data, the Data Nodes interact with one another moves and copies the data around. The Data Node is responsible for serving the read and write request for the client. The daemon named as Task Tracker runs on the Data Node which is responsible for executing the individual tasks assigned by the JobTracker.

C. HDFS Client

Client machines have Hadoop installed with all the cluster settings, but are neither a Master nor a Slave. Instead, the role of the Client machine is to load data into the cluster, submit Map Reduce jobs describing how that data should be processed and then retrieve or view the results of the job when it's finished. Client can read, write and delete files and also perform the operations to create and delete directories with contacting to Name Node.

III. PROBLEM STATEMENT

To find out most searched data after firing a query by user by analyzing web browsing history of different people using Hadoop framework.

III. MOTIVATION

Big data analytics is the process of examining large data sets containing a variety of data types. Big data can be analyzed with the software tools commonly used as part of advance analytics disciplines such as predictive analysis. We are looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tools such as MapReduce. These technology form the software framework that supports the processing of large database [6].

"WEB HISTORY ANALYSIS FOR DISPLAYING TOP SEARCHED RESULTS USING HADOOP FRAMEWORK" provides process of examining this large amount of different data types, or big data, in an effort to uncover hidden patterns, unknown correlations and other useful information. It also provides most relative results or trending topics which are asked by many users from data using users web browsing history.

IV. LITERATURE SURVEY

1]Title:- Big Data representation for Grade Analysis through Hadoop Framework[7]

Functionality:-

The grade estimation system is built on MapReduce Architecture and Hadoop based framework. The proposed architecture grading of student can be used to make predication. It can be used for analysis of various attributes of the Hadoop framework over the cloud environment. The paper has clearly deliberated the data distribution and the respective key-value pairs at each stage of Hadoop architecture.

Advantages:-

- Programmer only has to design map and reduce function
- Scalability: Can handle large data sets
- Easy to use.

Disadvantages:-

- Programming model is very restrictive

2] Title:- Web users Browsing Behaviour Prediction by implementing Support Vector machines in MapReduce using Cloud Based Hadoop[8]

Functionality:-
According to the observations, Support vector machine and many other machines learning algorithms do not fit when the source of input data is too large. The parallel support vector machines for web page prediction based on MapReduce programming model, which runs on Hadoop framework. It removes scalability problem of present SVM algorithm. From the experiments, we have improved the pre-processing time by comparing the results with non-Hadoop based approach and Hadoop-based approach. In this research work, one vs. one approach is used.

Advantages:-

- Reduces the user's browsing access time and avoids the visit of unnecessary pages to ease network traffic.

Disadvantages:-

- Joins of multiple datasets are tricky and slow.

3] Title: - Analysis of the Patients' Treatment Process in a Hospital in Thailand using Fuzzy Mining Algorithms[9]
Functionality:-
The main goal was to investigate and analyze the behaviour of patients within the given period of time. To do this Fuzzy Mining algorithms supported by ProM and Disco Fluxicon open-source applications was used. Considering the resulting fuzzy models/graphs, we understood that the "Acute upper respiratory infection, unspecified" allocated the highest most significant activity type to itself (in both of the fuzzy models created by ProM and Disco).

Advantages:-

- Flexible and convenient user interface.
- Easy computation.

Disadvantages:-

- Time consuming,.
- Hard to develop model

4] Title: - Social Network Analysis on SinaWeibo Based on K-Means Algorithm[10]
Functionality:-
In this research, study on users of SinaWeibo has been carried out, 98 user's information has been collected and stored in database. The K-means algorithm has been used to cluster the data. Furthermore, an improved K-means algorithm has been designed to find the optimal value of k. using the improved K-means algorithm; the 98 users were classified into 3 types, which were "Youth", "Technical enthusiast" and "Loving life". The classification method and the results could help enterprise to do targeted marketing to different types of users. This research provides a novel idea and a practical solution that classify users by analyse the hobbies and interests of their social circle, and proposed an improved K-means algorithm which can determined the number of clusters automatically.

Advantages:-

- Simple implementation with high efficiency and scalability.
- Works on well on big data sets.

Disadvantages:-

- Sensitive to data sets.
- Unstable result.
-

5]Title: - Analyzing students 'data using a classification technique based on genetic algorithm and fuzzy Logic[11]
Functionality:-
This paper examines applications of genetic algorithm and fuzzy logic for finding and optimal classification technique. The two parameters namely entropy and gini index are used for evaluating a classification technique. Genetic algorithm is employed which selects an optimal value of the function that considered these two parameters. rules. Thus by

running genetic algorithm for specified number of generation and optimal solution of classification technique can be determined.

Advantages:-

- Gives best result for overlapped data set and comparatively better then k-means algorithm.

Disadvantages:-

- In fuzzy algorithms we have to calculate Euclidean distance measures which can unequally weight underlying factors.

6] Title: - Unstructured Data Analysis on Big Data using Map Reduce[12]

Functionality:-

In this research work, the unstructured data is structured and processed by using MapReduce technique and the automatic prediction of user's taste is done through collaborative filtering. Map reduce is the most efficient technique for processing large volume of data and the application of collaborative filtering and sentiment analysis provides recommendation generation for any number of data provided as input. This MapReduce job can also be implemented in distributed mode in which we can use an N number of slaves for a single master. For lodging the huge data sets, Apache HBASE database support can be used. A pre-history, cache table can also be used for generating recommendations for a single user.

Advantages:-

- Stores and retrieves the data in an efficient way that would scale to very large scales.
- Supports the structure of big data as it is a parallel performing platform.

Disadvantages:-

- Large amount of unstructured data needs structural arrangement for processing the data.

V. SYSTEM OVERVIEW

A MapReduce is a programming paradigm. We are using it for computation of large datasets. A standard MapReduce process computes terabytes or petabytes of data. So MapReduce basically splits the huge data into chunks. We here mainly are considering three classes in this algorithm. Mapper Class, Reducer Class and the Driver Class.

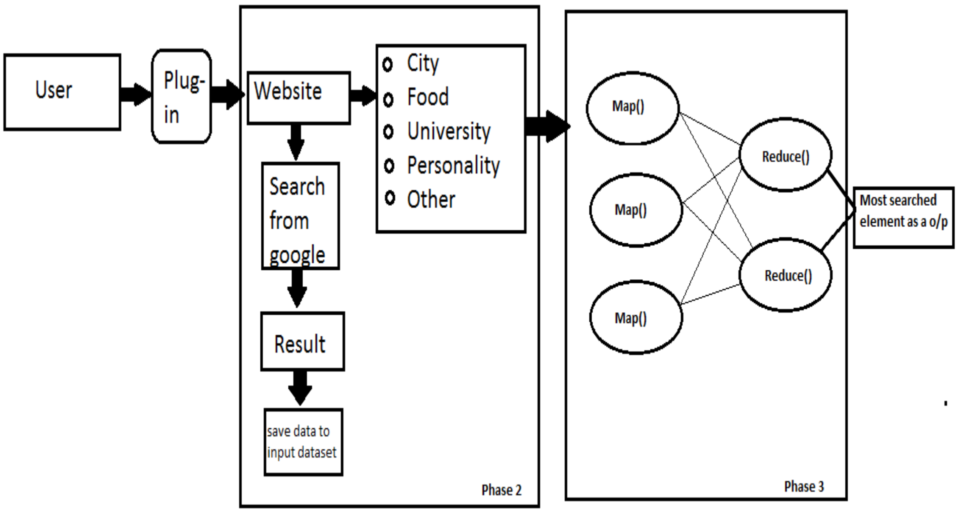

Fig: General block diagram for web search

DOI: 10.15680/IJIRCCE.2017. 05032321

The dataset in the form of text file format will be given as an input to the mapper class. The datasets will get split into the Chunks. The dataset will split word wise after the word wise splitting the output of Mapper class will be given as an input to the reducer class, Reducer function will actually combine the value elements of paired intermediate dataset having the same key. Each of the Map and Reduce steps are performed in parallel on pairs of data members. Thereby the program is segmented into 2 distinct and well defined stages namely Map and Reduce.
The Map stage involves execution of a function on given dataset in the form of (key, value) and generates the intermediate dataset. The generated intermediate dataset is then organized for the implementation of reduce operation. So here data transfer takes place between Map and Reduce functions. The reduce Function compiles all the dataset. Then the driver class will finally display the output including the count.

VI. PROPOSED MODEL

System will predict the most trending topic for constraints like city, university, food, personality on basis of the surfing history of a search engine with respect to different people. The browsing history will get stored on HDFS. Here assume entering search history as input. MapReduce will assign on dataset in a distributed environment. Dataset will compute at slave node level. Each system will return result. The JobTracker will retrieve the result information. The top five trending result in each constraint is the predicated or suggested for future user.

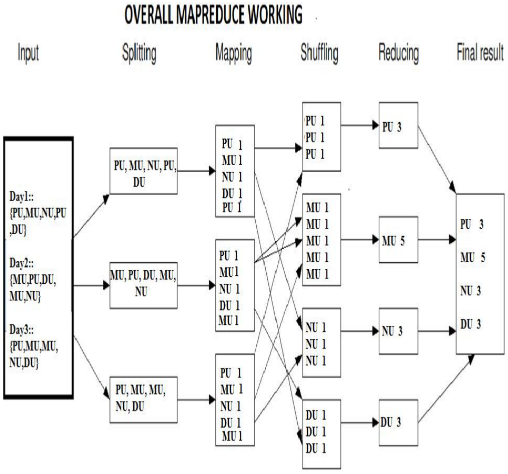

Fig: Example for web search

The proposed architecture can be used to make estimation of trending results of given constraint for future user. It can be used for analysis of various attributes of the Hadoop framework over the cloud environment.

VII. CONCLUSION

In this paper, we are using ‘Map Reduce’ algorithm to get result from big data. Mostly different analytics method or filter methods was used before, for getting predictions. Map reduces runs parallel that’s why we get output fast and also handle large dataset so scalability is really very good than any other technique. The system is built on MapReduce Architecture and Hadoop based framework. The proposed architecture can be used to make predication. It can be used for analysis of various attributes of the Hadoop framework over the cloud environment. The paper has clearly deliberated the data distribution and the respective key-value pairs at each stage of Hadoop architecture.

REFERRENCES

1. Sharma, S., Tim, U. S., Wong, J., Gadia, S., & Sharma, S. (2014).A Brief Review on Leading Big Data Models. Data Science Journal, 13(0), 138-157.
2. NoSQL models for handling Big Data: a brief review., International Journal of Business Information Systems, Inderscience, 2015
3. Sagiroglu, Seref, and DuyguSinanc. "Big data: A review." Collaboration Technologies and Systems (CTS), 2013 International Conference on.IEEE, 2013.
4. Chaudhuri, Surajit. "How different is big data?." Data Engineering (ICDE), 2012 IEEE 28th International Conference on.IEEE, 2012.
5. Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data" http://www.gartner.com/newsroom/id/1731916
6. http://www.tutorialspoint.com/articles/advantages-of-hadoop-mapreduce-programming
7. Big Data representation for Grade Analysis through Hadoop Framework. Hadoop official site May 2015, http://hadoop.apache.org/core
8. Web users Browsing Behaviour Prediction by Implementing Support Vector Machines in MapReduce using Cloud Based Hadoop. Hadoop official site May 2015, http://hadoop.apache.org/core/
9. Analysis of the Patients' Treatment Process in a Hospital in Thailand using Fuzzy Mining Algorithms, IEEE paper 2016.
10. Social Network Analysis on SinaWeibo Based on K-means algorithm, IEEE paper 2016.
11. Analyzing students 'data using a classification technique based on genetic algorithm and fuzzy logic
12. http://www.sciencedirect.com/science/article/pii/S1877050915005165