Uber Data Analysis using Map Reduce

P. Devika, Y. Prasanna , P. Swetha ,G. Akhilesh Babu

Abstract: Map Reduce has become of the foremost often used framework for processing of giant quantity of knowledge hold on in Hadoop cluster. It is used for multi processing of giant quantity of knowledge speedily. Firstly, it had been designed by google to produce the correspondence and cut back the fault tolerance of knowledge. We are using Uber Data for analyzing the vehicle with most popular trips. As mapreduce is used to process huge amounts of data, we are using mapreducing model to analyze uber data and give insights about the most used vehicle, number of trips it has covered. The main objective of this project is to investigate no of trips so as to produce data for the company to take care of the records and helps to company in creating huge information for long run endeavor.

I. INTRODUCTION

The main objective of Uber Data Analysis is to find the days on which each basement has more trips and the days on which each basement has more no of active vehicles. In this project we will be performing analysis on Uber dataset using MapReduce. In that case, if use Hadoop cluster it gives the fastest results without any error problems and also the server doesn't go down. This can be to supply quickest results and conjointly to boost the standard of social control choices. For example, if there is more information that containing base number, date, active vehicles and trips, the data not containing any of the above that can be removed or necessary changes can also be done for future endeavors.

In today's generation huge information, vast amounts of structured and unstructured Uber information are being created daily. Massive data is difficult to work with and desires massively parallel package running on large amount of Mapreduce could be a programming model that simplifies distributed applications that handle massive information. Increasing digitization of student records means that prognosticative analytics is anticipated to rework teaching and become a key tool in learning additional regarding students. Prognosticative analytics could be a method during which information collected regarding the scholar, generally attending, subjects taken, assessment is employed to know learning patterns, establish talent gaps, predict performance and establish learning opportunities. The effective feature choice technique is needed to research the economical classification algorithmic rule. Handling big data, a novel partition mechanism in MapReduce is also required. This is the main part of process during a hadoop system because it supplies the logic of method. In another words, mapreduce can be a coding system framework that helps in writing applications that processes big information sets victimization distributed and parallel algorithms at intervals Hadoop setting.

Mapreduce program consists of, Map() and Reduce() those are two unit functions. Map performs actions like grouping, filtering and sorting. If it reduce perform aggregates and summarizes the result created by map perform. The output generated by the map perform might be a key value combine(key,value) that acts as a result of the input for reduce perform.

A. BigData-5V's:

The term “big knowledge” is outlined as data that becomes thus giant that it can not be processed exploitation standard strategies. the dimensions of the knowledge which might be thought-about to be huge knowledge could be a perpetually varied issue and newer tools area unit incessantly being developed to handle this huge data. so as to create sense out of this overwhelming quantity of knowledge it's typically softened exploitation 5 V's: Velocity, Volume, Value, Variety, and Veracity

II. LITERATURE SURVEY

Before mapReduce, doing this type of calculation would are troublesome. Currently programmers will handle this type issues with an ease. The advanced algorithms have been coded by the knowledge scientists for frameworks so that it becomes easy to use for the programmers. They don't want the department of PhD scientists to develop a whole complex framework. As mapreduce will work on network which provides a straightforward analysis.

MapReduce is gaining ground chop-chop as a result of the Apache Hadoop and Spark parallel computing systems lets programmers use mapReduce to run models over giant distributed sets of knowledge and use advanced applied math and machine learning techniques to try to predictions, realize patterns, uncover correlations, etc.

Manuscript published on November 30, 2019.
* Correspondence Author

P. Devika*, Department of Computer Science and Engineering, MLR Institute of Technology, Dundigal, Hyderabad, India.
Y. Prasanna , Department of Computer Science and Engineering, MLR Institute of Technology, Dundigal, Hyderabad, India.
P. Swetha , Department of Computer Science and Engineering, MLR Institute of Technology, Dundigal, Hyderabad, India.
G. Akhilesh Babu, Department of Computer Science and Engineering, MLR Institute of Technology, Dundigal, Hyderabad, India.

In the existing system we are using DBMS to store data and PHP to process the data.But DBMS does not support huge amounts of data.This is the major drawback of traditional systems as it cannot meet the requirements of growing data.So to overcome these drawbacks we are moving to new system where data is stored in HDFS and processing is done by Mapreduce.

A.Proposed System

MapReduce could be a recent programming model that simplifies distributed applications that handle massive information. For implementing the thought of MapReduce, it's to divide the information and perform the information mining method. Then the collective results created. With the quick emergence of leading edge technologies, ancient information management solutions square measure inadequate to catch up with them. These technological changes square measure golf shot pressure on the adoption of “big information.” to understand why massive information is far higher than RDBMS for information analytics we've to understand the benefits of mistreatment massive information for analytics.

III. REQUIREMENTS

A. Hardware Requirements:

- Hard Disk of size 1 - 4 TB
- 8GB RAM
- 64 Bit OS

B. Software Requirements:

- Eclipse IDE -MARS
- JAVA –JDK-1.7
- WinScp
- Putty
- Hadoop 2.x-hadoop version

IV. SYSTEM DESIGN

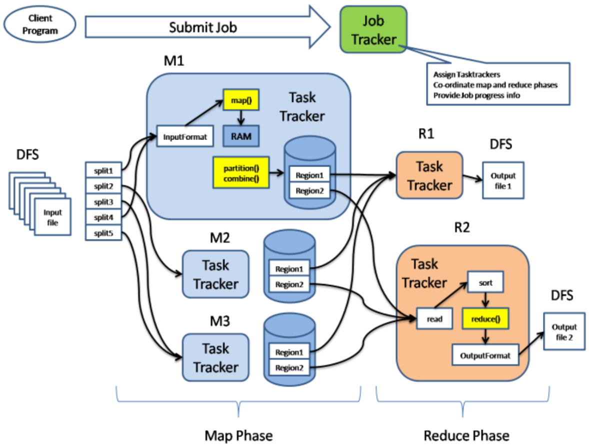
A. System Architecture

MapReduce has became of the foremost often used framework for processing of giant quantity of knowledge hold on in Hadoop cluster. It is used for multiprocessing of giant quantity of knowledge speedily. Firstly, it had been designed by google to produce the correspondence and cut back the fault tolerance of knowledge.

MR processes the info within the type of key price pairs. we are able to select the key price pairs supported our alternative. We need to use the key price pairs for Mapreduce as our schema isn't static. After we have static schema we are able to use columns for analysing the info Map cut back API can furnish the next choices like process, multiprocessing of giant amounts of knowledge and high accessibility.

The Map cut back work flow can undergoes totally different phases that stores the lead to hdfs with replications at the top. Job Trackers can do the work of checking all the Map cut back jobs that ar acting on the Hadoop Cluster.

The Job huntsman can play an important role in planning jobs and it'll keeps the track of each map and cut back jobs. The task hunter can do the particular map and cut back jobs. Map cut back design principally consists of 2 process stages. 1st one is that the map stage and thus the opposite is cut back stage. Between these 2 stages there's an extra stage referred to as intermediate stage that will the work of taking the input from the mappers and doing the tasks like shuffle, sort, mix etc.



Mapper Phase

a) Mapper Phase

In this phase, the mapper get the input values from the record reader it sends the key value pairs to the mapper as it can understand only them.

The input received by the mapper will be splited into some other key value pairs so that based on each key they go to the specified reducer which depends on the partitioning constarint. The out put of this phase is also key value pairs that is indermediate key value pairs.

b) Intermediate Process

This is the phase which comes in between the map and reduce phases. In this phase many operations are done based on the results required.

In this phase the same key values from different mappers will get into one mapper i.e., here the operations like shuffle, sort etc are done.

It uses the Round Robin algorithm to write the intermediate key values pairs into the local disk so that the condition is obtained correctly.

a) Reducer Phase

This is the second stage of the Map Reduce data flow. In this phase it receives the input from the partitioner and combiner. After performing them the intermediate values are taken as the input for this phase.

The reducers logic will begin with the operations performed by the mapper. It produces the output files like part files which contains the actual output of the analysed data.

Each time when we run a job it shows the number of reducers needed for the our job to jet executed. In the mapreduce configuration file we need to change the permissions so that the partitioning tasks can be performed.

As it is a parallel processing job when one mapper completes it's tasks and sit then the task tracker will assign the other task to it from the mapper which has many tasks to do. This will produce the output which is optimized

By using this we can increase the performance and reduce the processing time.

V. IMPLEMENTATION

Step: 1

Create a new directory with name Uber in your cluster.

EdgeNode@$ Hadoop fs mkdir Uber;

Step: 2

Write your mapreduce Program in Eclipse.

Uber.java

Step: 3

Create a Jar file Uber.jar

Step: 4

Copy the jar file, input file to local edge node using winscp.

Uber.jar, Uberdata.csv

Step: 5

Login into your cluster using putty.

Step: 6

Copy the input file from local to hadoop cluster.EdgeNode@$hadoopfs–copyFromLocal.Uberdata.csv /user/sai/

Step: 7

Run the MapReduce Program EdgeNode@$ hadoop fs jar Uber.jar /user/sai/Uberdata.csv Uber

Step: 8

Result can be seen through command interface as

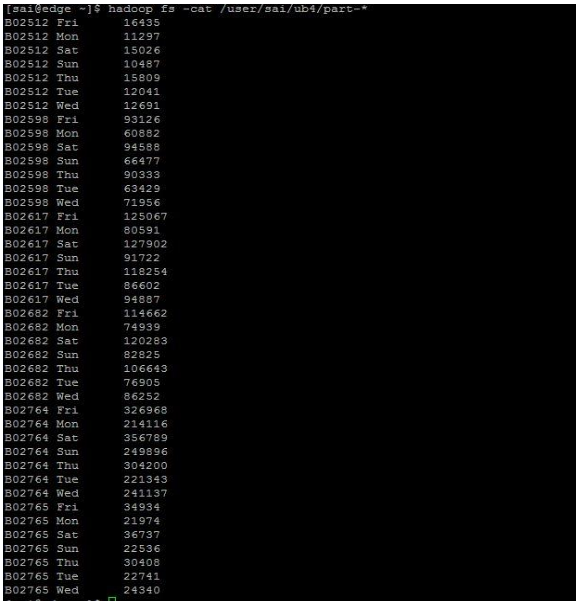
EdgeNode@$ hadoop fs –cat user/sai/Uber/part-*

Or

You can use GUI for results

By using the address of the cluster: 172.16.103.68:8888

VI. RESULT



```
[sai@edge ~]$ hadoop fs -cat /user/sai/ub4/part-*
B02512 Fri      16435
B02512 Mon      11297
B02512 Sat      15026
B02512 Sun      10487
B02512 Thu      15809
B02512 Tue      12041
B02512 Wed      12691
B02598 Fri      93126
B02598 Mon      60882
B02598 Sat      94588
B02598 Sun      66477
B02598 Thu      90333
B02598 Tue      63429
B02598 Wed      71956
B02617 Fri      125067
B02617 Mon      80591
B02617 Sat      127902
B02617 Sun      91722
B02617 Thu      118254
B02617 Tue      86602
B02617 Wed      94887
B02682 Fri      114662
B02682 Mon      74939
B02682 Sat      120283
B02682 Sun      82825
B02682 Thu      106643
B02682 Tue      76905
B02682 Wed      86252
B02764 Fri      326968
B02764 Mon      214116
B02764 Sat      356789
B02764 Sun      249896
B02764 Thu      304200
B02764 Tue      221343
B02764 Wed      241137
B02765 Fri      34934
B02765 Mon      21974
B02765 Sat      36737
B02765 Sun      22536
B02765 Thu      30408
B02765 Tue      22741
B02765 Wed      24340
```

VII. CONCLUSION

Storing the data and processing the data is the hectic task now-a-days in various businesses. As data is increasing tremendously we do not have appropriate ways to store and process this data. Traditional methods are not capable of handling this data. In this scenario, Hadoop comes into picture. HDFS solves the problem of storing and MapReduce can be used to process data. In our project we are using MapReduce framework to analyze student data. There are many ways to store and process the data, but these days' storage and handling of data has become a major issue, instead of using traditional databases we are moving towards a new Solution for dealing with large amount of data that is HDFS in Hadoop. The main objective of our project is to analyze student data in order to provide information for the education system to maintain records and help in making faster and useful decisions for the benefit of the students. We have taken a data set of students which contains student roll number and credits in a csv format. Next steps are as follows. We obtain results from moving the data from local to Hadoop and execute the MapReduce program by converting it in to a jar file.

REFERENCES

1. Sai Prasad K., Chandra Sekhar Reddy N., Rama B." Analyzing and predicting academic performance of students using data mining techniques" In Journal of Advanced Research in Dynamical and Control Systems 2018
2. Prasanna Y.,Kumar R.A., Sobharani, Reddy S.N., "A study on techniques of facial recognition and IOT cam vulnerabilities: A survey"In International Journal of Engineering and Technology(UAE) 2018.
3. Varalakshmi V., Anvesh E., Sandeep P." Analysis of raw and ro treated water â€““a case study of medchal district, telangana state, India" In International Journal of Civil Engineering and Technology 2017
4. https://acadgild.com/blog/mapreduce-use-case-uber-data-analysis hadoop.apache.org/
5. https://www.owler.com/reports/acadgild/acadgild-blog-mapreduce-use-case---uber-data-analy/1470309362454
6. https://ieeexplore.ieee.org/document/8389665
7. https://github.com/vickyg12/Uber-MapReduce-Data-analysis.

Retrieval Number: D7111118419/2019©BEIESP
DOI:10.35940/ijrte.D7111.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)