IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 5, May 2018

Big Data Analytics on Cab DataSet Using Hadoop

Vivek Sachapara[1], Hrishikesh Shinde[2], Abhishek Puri[3,] Shraddha Aggrawal[4] Prof. Sachin Wandre[5]

Department of Computer Engineering, Sinhgad Institutes, Pune, Maharastra, India[1,2,3,4,5]

Abstract: A tremendous amount of data is generated with the increasing use of the internet, as a result, new fields like big data handling, IOT, Machine learning have evolved all of them aiming to improve the standard of human living. This data is apprehended and have become certainly manageable for first hand properties of data driver analysis which can be recycled for an improvement of folks living in urban zones and even maximize the business profits. One such area is CAB or Taxi mode of transportation .large amount of data is generated by cab companies, earlier data generated by taxi commission was physically analyzed by a various analyst to find superlative practices but with the exponential increase ranging to almost in GBs, to perform analysis manually was next to impossible. Hence came the tools like big data to our aid. Big data can effortlessly help us to analyze thousands of Gbs in a fraction of seconds. This data can be analyzed to for purposes like to avoid traffics, judge the peak hours, help frequent customers with beneficial offers etc. This analysis can also be used by Indian government authorities to aid public transportation. In this paper detailed analyze of Hadoop map reduce and spark is made using a different parameter.

Keywords: Big Data, Hadoop , Map Reduce, Spark.

I. INTRODUCTION

Since the time all walks of human lives have been digitized, a large amount of data is being generated by the users. This data generated in gbs although poses a problem of Big data. But the study of this data is extremely useful in business development ultimately focusing on the raising profits, improving their facilities given to their customers by understanding the customer better and acting upon the feedback received from their customer, and even understanding the needs of the employees by evaluating their performance etc. Ultimately analysis of this big data can be used to study or reveal the patterns that can be used to form the basis of decisions that can improve the business strategies. In our research work, we have collected the data generated by the cab companies and analyzed the same using different tools and drawn insights on which tools work better in different situations. We have collected the template data set from"OLA CAB COMPANY" on which we have done the analysis using two parameters ie. location and time using two different tools Apache Map reduce hive and Spark.
This paper is organized in '5' sections.

A. Hadoop

Hadoop is a collection of software utilities that facilitate using a network of many computers to solve problems of massive amounts of data and computation[1]. It provides different frameworks of software for distributed kind of storage and processing of Large amount of data i.e. Big Data using the MapReduce Programming.It is originally designed for Computer-Clusters built from commodity Hardwarestill it has also found use on clusters of higher-end hardware.All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and it should be automatically handled by the software framework. Hadoop Distributed File System (HDFS) is a core storage part of Apache Hadoop, and a processing part which is a MapReduce programming model.The Hadoop framework is mostly written in the Java programming language, with some native code in C and command line utilities written as shell scripts.

B. Spark

Apache Spark is an Open-source cluster computing framework. Originally developed at the University of California. Apache Spark has as its architectural foundation the resilient distributed dataset, a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way[2].Spark and its resilient distributed dataset was developed in 2012 in response to limitations in the MapReduce cluster computing paradigm. Apache Spark requires a distributed storage system and cluster manager. For cluster management, Spark supports standaloneHadoop YARN. For distributed storage, Spark can interface with a wide variety, including Hadoop Distributed File System, Cassandra, OpenStack Swift, Amazon S3, Kudu, MapR File System (MapR-FS)or a custom solution can be implemented. MapReduce read the input data from disk, then map a function across the data, reduce the results of the map, and store reduction results on disk. Spark's resilient distributed dataset function as a working set for distributed programs that offers a deliberately restricted form of distributed shared memory.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 5, May 2018

C. Map Reduce

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster[3]. The model is a specialization of the split then apply then combine strategy for data analysis. It is inspired by the map function and reduce functions commonly used in functional programming. MapReduce libraries have been written in many programming languages as per the requirements, with different levels of optimization. A popular open-source implementation that has support for distributed shuffles is part of Apache Hadoop.MapReduce allows for distributed processing of the map and reduction operations. Maps can be performed in parallel, provided that each mapping operation is independent of the others in practice, this is limited by the number of independent data sources and the number of CPUs near each source.

II. RELATED WORK

In this paper complete experiment analysis was made on NYC taxi Data using tools like HadoopMapreduce , Hive and conclusion was made based on different dataset[4] . A comparative analysis was made of hadoop and spak engine based on factors like data handling capacity, processing speed[5]. Analysis was mades based on total CPU utilization in which Map reduce more cpu than Spark. In this paper hadoop analysis was done on YouTube data using hadoop on Amazon Web Service and accordingly result was generated[6].

III.PROBLEM DEFINATION

During this course of study about Cab data sets consequence are entirely centred on certain analysis. These problem statements are elucidated clearly one after the other in the sequence.

A. Problem Definition I – Analysis on Individual

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

1) Driver with most distance travelled: This analyse is done on an individual person who has cover most of distance will be figured out.

Map Reduce Function:

Mapper:
map(licence,Distance_trip)=Emit(inter_licence,inter_Distance_trip)

Reducer:
Reduce(inter_licence, inter_distance_trip)=Emit(licence,Max_distance_trip)

2) *Driver with most fare collected:* This analysis will be done individually where we the drive collected that collected most of fare will be figured.

Map Reduce Function:

Mapper:
map(Hack_license, Total_amount)=Emit(inter_Hack_license, inter_Total_amount)

Reducer:
Reduce(inter_ Hack_license, inter_Total_amount)=Emit(Hack_license, Max_Total_amount)

3) *Driver with most time travelled:* Analysis of Drivers performance based on the time, the individual analysis is done by calculating the time spent by each driver driving the cab in seconds.

Map Reduce Fucntion:

Mapper:
map(inter_ Hack_license, inter_Trip_time)=Emit(inter_Hack_license, inter_Trip_time)

Reducer:
Reduce(inter_ Hack_license, inter_Trip_time)=Emit(Hack_license, Max_Trip_time)

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 5, May 2018

4) *Driver with most efficiency based on customer feedback:* Analysis of Drivers performance based on feedback received by the customers, the individual analysis is done by evaluating the customer rating

Map Reduce Function:

Mapper:
map(Hack_license, Trip_distance / Trip_time)=Emit(inter_Hack_license, inter_Trip_distance / inter_ Trip_time)

Reducer:
Reduce(inter_ Hack_license, inter_Trip_distance / inter_ Trip_time)=Emit(Hack_license, Min_effeciency)

B. Problem Definition II– Analysis on Region

Analysis on Region The problem definition II consists of analysis on a region.

1) *Maximum pick up location*: Frequently visited pick up locations: Buzzing or In demand pick up locations are analyzed by storing the alttitude ie pickup_latitude and pickup_longitude.

Map Reduce Function:

Mapper:
map(Pickup_location)=Emit(inter_Pickup_location,1)

Reducer:
Reduce(inter_Pickup_location, 1)=Emit(Pickup_location, Sum)

2) *Maximum pick up location*: Frequently visited drop off locations: Buzzing or In demand drop off locations are analyzed by storing the alttitude ie dropoff_latitude and dropoff_longitude.

Map Reduce Function:

Mapper:
map(Dropoff_location)=Emit(inter_ Dropoff_location,1)

Reducer:
Reduce(inter_ Dropoff_location, 1)=Emit(Dropoff_location, Sum)

IV. EXPERIMENT ANALYSIS

Based on the following fuctions used experiment was carried out and following result was generated which is shown in Figure 1.

Problem Definitation	Time In Seconds	
	Map Reduce	Spark
Problem Definitation 1.1	16	21
Problem Definitation 1.2	13	14
Problem Definitation 1.3	35	37
Problem Definitation 1.4	24	26
Problem Definitation 2.1	8	7
Problem Definitation 2.2	14	12

Table 1

V. CONCLUSION

In this paper a comparison is drawn between two tools ie. hadoop map reduce and spark. We analysed the Cabdataset using both these tools on the basis of various individual spatial and temporal parameters and results are drawn. Basically as the size of the data grew Spark showed efficient performance by evaluating the same query in way less amount of time as compared to MapReduce.Here the same saved plays a great role when real time analysis comes into picture due to large amount of data ie in Gbs or Pbs is involved and query execution time effects the organizations largely. Ultimately we conclude that this project is used by various data analysts, machine learning engineers to forecast taxi demand factors and how this demand can be used expected to grow and change.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 5, May 2018

FUTURE SCOPE

In this paper currently few queries like most pick up locations and drop off locations and driver with most distance travelled is analyzed. But in future many queries can analyzed using this model like predicting trips between railway station and airports etc.

ACKNOWLEDGEMENT

The special vote of thanks to **Mr S.N. Wandre** and the Department of Computer Engineering SIT for their incessant support.

REFERENCES

[1] https://en.wikipedia.org/wiki/Apache_Hadoop
[2] https://en.wikipedia.org/wiki/Apache_Spark
[3] https://en.wikipedia.org/wiki/MapReduce.
[4] Umang Patel "NYC Taxi Trip and Fare Data Analytics using BigData",University of BridgePort,USA
[5] Akaash Vishal Hajrika,G Jagdeesh,Eeti Jain, "Performance comparision of hadoop and spark engine",International conference on I-SMAC,I-SMAC 2017
[6] Jaime Raigoza,Vijay Parmar," Research the Data Analysis and Processing Comparison between MapReduce and Spark",International Conference on Computational Science and Computational Intelligence,2016
[7] PrathyushaRani Merla, Yiheng Liang, "Data Analysis using Hadoop MapReduce Environment", IEEE International Conference on Big Data (BIGDATA),2017
[8] Ahmed Qasim Mohammed, Rajesh Bharati," An Efficient Technique to Improve Resources Utilization for Hadoop MapReduce in Heterogeneous system" International Conference on Intelligent Communication and Computational Techniques (ICCT) 2017.
[9] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar," A Review Paper on Big Data and Hadoop" , international Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.