Twitter data analysis using MapReduce

Sandeep Iddalgave[1], T.Deepika Reddy[2], S.Kunal Reddy[3], P.Jayanth[4]
[1]*Assistant Professor,* [2,3,4]*Student*
Department of InformationTechnology, MLR Institute of Technology, Hyderabad, India.

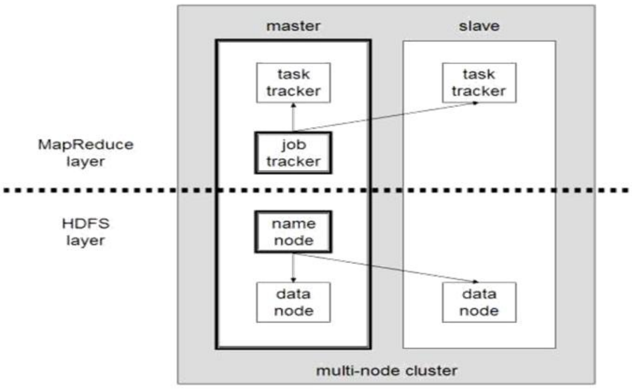
Abstract **-** Twitter is one of the largest social networking site present on the social media platform due to its high popularity and easy availability of data it has become one of the largest data hub that contains various kinds of data and in order to prevent data loss and make it more efficient and cost effective we need to conduct an analysis of data, we use Hadoop in order to have a better study on data. Using Mapreduce we will analyze the number of people discussing about a particular matter.

Keywords - big data,hadoop,mapreduce

I. INTRODUCTION

A. Big information is a term for informational indexes that are so vast or complex that customary information preparing application programming is insufficient to manage them. Enormous information challenges incorporate catching information, information stockpiling, information examination, look, sharing, exchange, perception, questioning, refreshing and data security. _ Enormous information is an issue articulation so here we will utilize HADOOP and its ecosystems, for getting crude information from the twitter.

Hadoop is an open source, Java-based programming structure that backings the preparing and capacity of to a great degree huge informational indexes in a conveyed processing condition. It is a piece of the Apache venture supported by the Apache Software Foundation.



B. Hadoop MapReduce (Hadoop Map/Reduce) is a product structure for appropriated preparing of extensive informational indexes on process groups of ware equipment. It is a sub-undertaking of the Apache Hadoop project. The structure deals with booking assignments, observing them and re-executing any fizzled undertakings.

i) Generally MapReduce worldview depends on sending the PC to where the information dwells!

ii) MapReduce program executes in three phases, specifically outline, rearrange arrange, and lessen organize.

Map organize - The guide or mapper's activity is to process the info information. By and large the info information is as record or registry and is put away in the Hadoop document framework (HDFS). The info document is passed to the mapper work line by line. The mapper forms the information and makes a few little lumps of information.

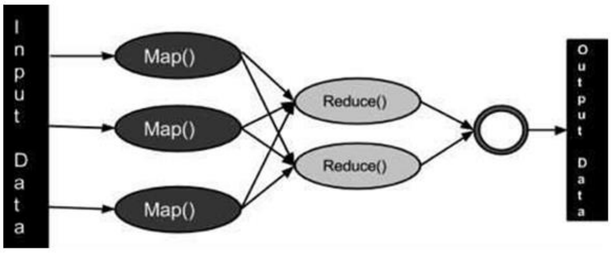
Reduce arrange - This stage is the mix of the Shufflestage and the Reduce organize. The Reducer's activity is to process the information that originates from the mapper.

In the wake of handling, it creates another arrangement of yield, which will be put away in the HDFS.

i) During a MapReduce work, Hadoop sends the Map and Reduce undertakings to the proper servers in the group.

ii) The system deals with every one of the points of interest of information passing, for example, issuing undertakings, checking assignment fruition, and replicating information around the bunch between the hubs.

iii) After fulfillment of the given assignments, the group gathers and lessens the information to shape a proper outcome, and sends it back to the Hadoop server.



II. METHODOLOGY

A. HDFS - Hadoop File System was produced utilizing appropriated record framework plan. It is keep running on item equipment. Not at all like other conveyed frameworks, HDFS is profoundly blame tolerant and composed utilizing minimal effort equipment. HDFS holds substantial measure of information and gives less demanding access. To store such tremendous information, the records are put away over various machines.

These documents are put away in repetitive design to safeguard the framework from conceivable information misfortunes if there should be an occurrence of disappointment. HDFS likewise makes applications accessible to parallel handling.

Features of HDFS:

i) It is appropriate for the dispersed stockpiling and preparing.

ii) Hadoop furnishes a summon interface to cooperate with HDFS.
iii) The worked in servers of namenode and datanode help clients to effectively check the status of group.
iv) Streaming access to document framework information.
v) HDFS gives document authorizations and confirmation. HDFS takes after the ace slave engineering and it has the accompanying components.

Namenode: The namenode is the product equipment that contains the GNU/Linux working framework and the namenode programming. It is a product that can be keep running on item equipment.

The framework having the name node goes about as the ace server and it does the accompanying assignments-

i) Manages the document framework namespace.
ii) Regulates customer's entrance to records.
iii) It likewise executes record framework activities, for example, renaming, shutting, and opening documents and registries.

Datanode: The datanode is an item equipment having the GNU/Linux working framework _ and datanode programming. For each hub (Commodity equipment/System) in a bunch, there will be a datanode. These hubs deal with the information stockpiling of their framework. Datanodes perform read-compose activities on the record frameworks, according to customer ask for .They likewise perform tasks, for example, piece creation, cancellation, and replication as indicated by the guidelines of the namenode. Square Generally the client information is put away in the documents of HDFS.

The record in a document framework will be separated into at least one fragments or potentially put away in singular information hubs. These record sections are called as pieces. At the end of the day, the base measure of information that HDFS can read or compose is known as a Block. The default piece estimate is 64MB, however it can be expanded according to the need to change in HDFS design.

Objectives of HDFS:

Fault recognition and recuperation - Since HDFS incorporates a substantial number of item equipment, disappointment of segments is visit. Subsequently HDFS ought to have systems for snappy and programmed blame location and recuperation.

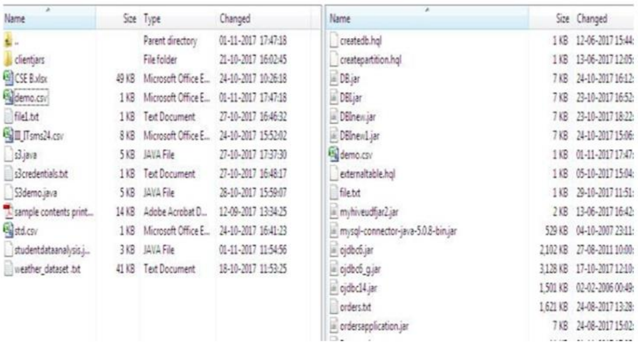
Huge datasets - HDFS ought to have several hubs for each group to deal with the applications having gigantic datasets.

Hardware at information - An asked for errand should be possible productively, when the calculation happens close to the information. Particularly where gigantic datasets are included, it lessens the system activity and builds the throughput.

There are numerous more charges in "$HADOOP_HOME/container/hadoop fs" than are exhibited here, in spite of the fact that these fundamental tasks will kick you off. Running ./canister/hadoop dfs with no extra contentions will list every one of the summons that can be keep running with the FsShell framework. Moreover, $HADOOP_HOME/receptacle/hadoop fs -enable commandName to will show a short utilization synopsis for the activity being referred to, in the event that you are trapped. Every other record and way names allude to the items inside HDFS.

B. IMPLEMENTATION

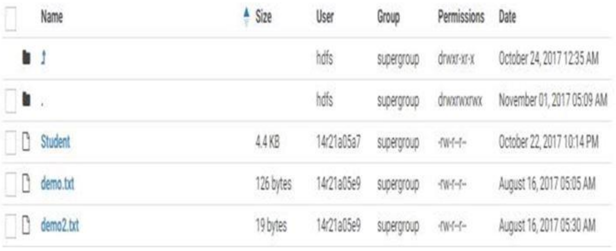
i) First we downloaded and introduced winscp and putty in our working framework.
ii) Then we associated with the server with the client name and secret word gave to us.
iii) Then collected some datasets regarding landslides and stored the dataset in the form of csv file. We stored that dataset in our folder in winscp by dragging it to the local file.



Name	Size	Type	Changed
..		Parent directory	01-11-2017 17:47:18
clientjars		File folder	21-10-2017 16:02:45
CSE B.xlsx	49 KB	Microsoft Office E...	24-10-2017 10:26:18
demo.csv	1 KB	Microsoft Office E...	01-11-2017 17:47:18
file1.txt	1 KB	Text Document	27-10-2017 16:46:32
III_ITsms24.csv	8 KB	Microsoft Office E...	24-10-2017 15:52:02
s3.java	5 KB	JAVA File	27-10-2017 17:37:30
s3credentials.txt	1 KB	Text Document	27-10-2017 16:48:17
S3demo.java	5 KB	JAVA File	28-10-2017 15:59:07
sample contents print...	14 KB	Adobe Acrobat D...	12-09-2017 13:34:25
std.csv	1 KB	Microsoft Office E...	24-10-2017 16:41:23
studentdataanalysis.j...	3 KB	JAVA File	01-11-2017 11:54:56
weather_dataset.txt	41 KB	Text Document	18-10-2017 11:53:25

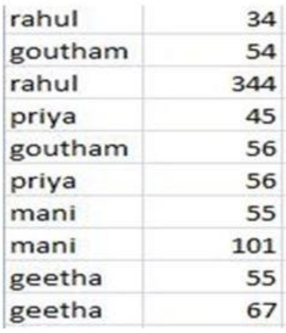
Name	Size	Changed
createdb.hql	1 KB	12-06-2017 15:44:
createpartition.hql	1 KB	13-06-2017 12:05:
DB.jar	7 KB	24-10-2017 16:12:
DB1.jar	7 KB	23-10-2017 16:52:
DB1new.jar	7 KB	23-10-2017 18:22:
DB1new1.jar	7 KB	24-10-2017 15:06:
demo.csv	1 KB	01-11-2017 17:47:
externaltable.hql	1 KB	05-10-2017 15:04:
file.txt	1 KB	29-10-2017 11:51:
myhiveudfjar2.jar	2 KB	13-06-2017 16:42:
mysql-connector-java-5.0.8-bin.jar	529 KB	04-10-2007 23:11:
ojdbc6.jar	2,102 KB	27-08-2011 10:00:
ojdbc6_g.jar	3,128 KB	17-10-2017 12:10:
ojdbc14.jar	1,501 KB	02-02-2006 00:49:
orders.txt	1,621 KB	24-08-2017 13:28:
ordersapplication.jar	7 KB	24-08-2017 15:02:

iv) And implemented some data sets to upload the dataset into the cluster like _
v) We used Hadoop fs –copyFromLocal demo.csv •With the help of jar files we are writing a code in ellicipse.
vi) Datasets is store in jar files like" hadoop jar Retweet.jar/datasets/demo.csv tweetout888.
vii) Hadoop fs –cat tweetout888/part-* .by using these command we will get an output

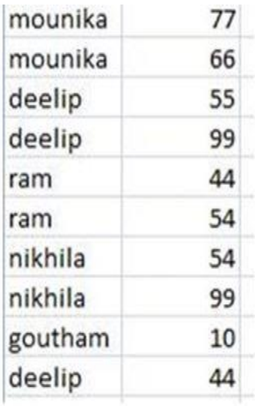


	Name	Size	User	Group	Permissions	Date
	↥		hdfs	supergroup	drwxr-xr-x	October 24, 2017 12:35 AM
□	.		hdfs	supergroup	drwxrwxrwx	November 01, 2017 05:09 AM
□	Student	4.4 KB	14r21a05a7	supergroup	-rw-r--r--	October 22, 2017 10:14 PM
□	demo.txt	126 bytes	14r21a05e9	supergroup	-rw-r--r--	August 16, 2017 05:05 AM
□	demo2.txt	19 bytes	14r21a05e9	supergroup	-rw-r--r--	August 16, 2017 05:30 AM

Data sets:



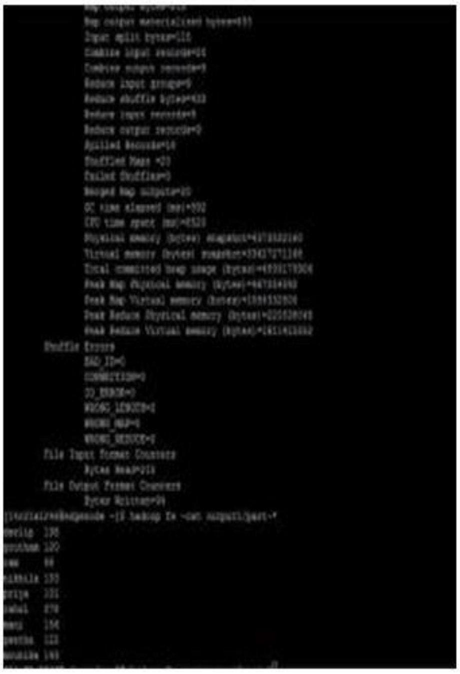
rahul	34
goutham	54
rahul	344
priya	45
goutham	56
priya	56
mani	55
mani	101
geetha	55
geetha	67



mounika	77
mounika	66
deelip	55
deelip	99
ram	44
ram	54
nikhila	54
nikhila	99
goutham	10
deelip	44

Retweetability - Retweet in twitter is the agreement action to a specific tweet, as in some cases the user passes information to his/her audiences to express their opinion on a particular tweet. The mechanism of retweetability plays a prominent role in information diffusion. retweet and number of mentions are related to the same network-topology. Additional work was done by where the reweetability was studied by deploying two different features the Content based on MapReduce.

III. RESULT



```
                Map output bytes=[illegible]
                Map output materialized bytes=[illegible]
                Input split bytes=[illegible]
                Combine input records=[illegible]
                Combine output records=[illegible]
                Reduce input groups=[illegible]
                Reduce shuffle bytes=[illegible]
                Reduce input records=[illegible]
                Reduce output records=[illegible]
                Spilled Records=[illegible]
                Shuffled Maps =[illegible]
                Failed Shuffles=0
                Merged Map outputs=[illegible]
                GC time elapsed (ms)=[illegible]
                CPU time spent (ms)=[illegible]
                Physical memory (bytes) snapshot=[illegible]
                Virtual memory (bytes) snapshot=[illegible]
                Total committed heap usage (bytes)=[illegible]
                Peak Map Physical memory (bytes)=[illegible]
                Peak Map Virtual memory (bytes)=[illegible]
                Peak Reduce Physical memory (bytes)=[illegible]
                Peak Reduce Virtual memory (bytes)=[illegible]
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=[illegible]
        File Output Format Counters
                Bytes Written=[illegible]
[illegible] -]$ hadoop fs -cat output1/part-*
deelip  [illegible]
goutham [illegible]
ram     [illegible]
nikhila [illegible]
priya   [illegible]
[illegible]
[illegible]
[illegible]
mounika [illegible]
```

IV. CONCLUSION

The sheer amount and the different types of data on twitter and the public nature of tweets have allowed exploiting twitter information in data analysis.

First, by measuring the life cycle of a specific topic by measuring the number of tweets over a period of time, and second by measuring the sentiment of users . Our aim is to enhance the analysis of twitter data how many members are retweet on a paticular tweet,we will be count and add those numbers. _ It is proposed to stream real time live tweets from twitter using Twitter API, and thelarge volume of data makes the application suitable for Big Data Analytics.

A method topredict or deduct the location of a tweet based on the tweet's information and the user'sinformation should be found in the future.

V. REFERENCES

[1]. Divya Sehgal; Ambuj Kumar Agarwal "Sentiment analysis of big data applications using Twitter Data with the help of HADOOP framework" 2016 International Conference System Modeling & Advancement in Research Trends (SMART) Year: 2016.

[2]. Twitter. From https://en.wikipedia.org/wiki/Twitter

[3]. https://www.kaggle.com/

[4]. http://www.ijcee.org/vol8/931-IT015.pdf.

[5]. https://en.wikipedia.org/wiki/Big_data

[6]. https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.html

[7]. https://hadoop.apache.org/docs/r2.6.3/hadoop-project-dist/hadoop-common/FileSystemShell.html