International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

Big Data Analysis with R Programming and RHadoop

U. Prathibha[1], M. Thillainayaki[2], A. Jenneth[3]
[1,2,3]Assistant Professor, [1]Department of CS, [2]Department of CA, [3]Department of IT
Karpagam Academy of Higher Education
Coimbatore, Tamil Nadu, India

ABSTRACT
Big data is a technology to access huge data sets, have high Velocity, high Volume and high Variety and complex structure with the difficulties of management, analyzing, storing and processing. The paper focuses on extraction of data efficiently in big data tools using R programming techniques and how to manage the data and the components that are useful in handling big data. Data can be classified as public, confidential and sensitive. This paper proposes the big data applications with the Hadoop Distributed Framework for storing huge data in cloud in a highly efficient manner. This paper describes the tools and techniques of R which is integrated with Big data tools for the parallel processing and statistical method. Using RHadoop data tools helps organization to resolve the scalability, issues and solve their predictive analysis with high performance by using Map reducing Framework.

Keywords: *Big data, R, RHadoop, Map Reduce*

I. INTRODUCTION:

Big data is not only containing data, it also contains various tools, techniques and frameworks. Data that has extra-large Volume, comes from Variety of sources, Variety of formats and comes at us with a great Velocity is normally referred to as Big Data.

Variety – Different type of data including text, audio, video, click streams, log files, and more which can be structured, semi-structure or unstructured.

Volume - Hundreds of terabytes and petabytes of information.

Velocity – Speed of data to be analyzed in real time to maximize the data's business value.

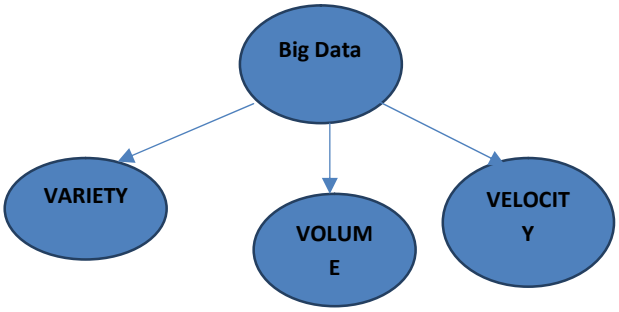

Fig 1: Big Data

R: R. is an open source language that uses data modelling, handling, statistics, prediction, time analysis and data visualization. The R language uses your computer's RAM, the RAM of your machine is very large, and the larger data you can work for R. We have more than 4000 different packages created by various scholars as per the requirement. The latest version of R will be R 3.0.2, initially R is not used as a large data analysis language due to its memory limit problems. Gradually, some libraries such as R, ffbase, Rodbc, rmr2 and Rhdfs were available to handle large data. Rmr2 and rhdfs use Hadoop power in order to handle great data effectively.

II. LITERATURE REVIEW

In [1], Anju Gahlawat 2014 explained that the company would solve their performance-efficient analyzes in resolving greater performance and complexity by using R and Hadoop with Integrating. In [2] Anshul Jatain and Amit Ranjan2017 described the larger diagram that includes the tools and techniques that can load, extract and disseminate different data when performing complex coordinate power to perform complex changes and analyzes. In [3] Harish D, Anusha M.S and *et.al* 2015 describe the process of applying some tools and techniques to analyze, visualize and predict the future trend of the

Research and also implemented the RHadoop, It is complete set where we can process our data efficiently, perform some meaningful analysis. It is one of the approach to developing algorithms that have been explicitly parallelized to run within Hadoop. In [4] Raissa Uskenbayeva a, Abu Kuandykova and et.al., 2015 proposed the integration of Hadoop-based data and R, which is popular for processing statistical information. Hadoop database contains libraries, Distributed File System (HDFS), and resource management platform. It can implement a version of the Map Reduce programming model for processing large-scale data and it allows us to integrate various data sources at any level, by setting arbitrary links between circuit elements, constraints and operations. In [5] Ross Ihaka and Robert Gentleman,1996 explained and developed the R programming and implemented in the area of probability, Computational Efficiency, Memory management and scoping. In [6] Shubham S. Deshmukh, Harshal Joshi and *et.al.,* 2017 implemented the twitter data analysis and visualization in R platform. It mainlyfocuses on real-time analysis rather than historic datasets. Twitter API allow for collecting the sentiments information in the form of positive score, negative score or neutral. After it decided to build our back-end on top of Hadoop platform which includes Hadoop HDFS as distributed file system and Map-reduce as distributed computation.

III. BIG DATA AND HADOOP

Big Data is a term used for a collection of data sets so large and complex that it is difficult to process using traditional applications/tools. It is the data exceeding Terabytes in size. Because of the variety of data that it encompasses, big data always brings a number of challenges relating to its volume and complexity. A recent survey says that 80% of the data created in the world are unstructured.

HADOOP

Hadoop is a software framework which stores huge amount of data and process it.

Scalable: It can store reliably and process petabytes.

Economical: It distributes the data and processing across clusters of commonly available computers (in thousands).

Efficient: By distributing the data, it can process in parallel on the nodes where the data is located.

Reliable: It automatically maintains multiple copies of data and automatically redeploys computing tasks based on failures.

The data can be managed with Hadoop to distribute the data and duplicates chunk of each data file across several nodes. Locally available resource is used to process, parallel process Handles failover smartly and automatically.

Features of Hadoop

It is optimized to handle massive quantities of various types of data. It Shared Nothing Architecture. Hadoop replicates data across multiple computers. It provides High throughput with low latency. It complements both OLTP and OLAP. It is not good when work is not parallelized. It is not good for processing small files because it stores a huge amount of data.

1. HDFS Daemons

Daemons mean “Background process”.
Name node
Data Node
Secondary name node

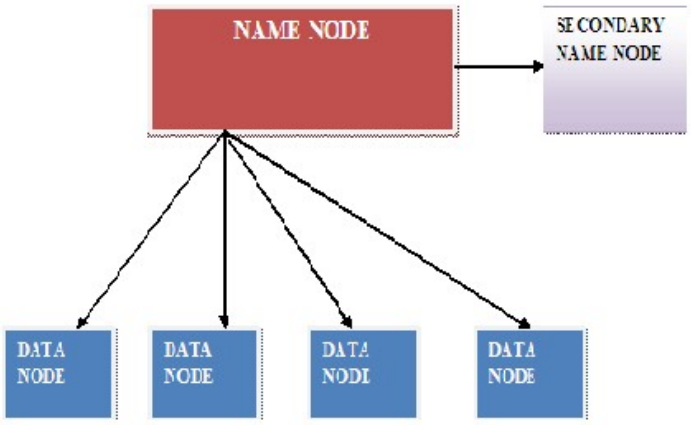

Fig 2: HDFS Daemons

A. HDFS Daemons – Name node(NN)

It is the ‘master’ machine. It controls all the meta data for the cluster. Eg - what blocks make up a file, and what data nodes those blocks are stored on. HDFS breaks large data into smaller pieces called Blocks. Default block size is 64MB.NN uses RACKID identify. Rack is a collection of data nodes within cluster. NN keeps tracks of blocks of a file as it is placed on various Data nodes. NN manages file related operations such as read, write, create and

delete. Its main job is managing the File System Namespace.

B. File System Namespace

File system namespace refers a collection of files in cluster. It includes mapping of blocks to file, file properties and it is stored in a file called FS Image. HDFS supports a traditional hierarchical file organization. A user or an application can create directories and store files inside these directories. The file system namespace hierarchy is similar to most other existing file systems; one can create and remove files, move a file from one directory to another, or rename a file. The Name Node maintains the file system namespace. Any change to the file system namespace or its properties is recorded by the Name Node. An application can specify the number of replicas of a file that should be maintained by HDFS. The number of copies of a file is called the replication factor of that file. This information is stored by the Name Node. HDFS stores multiple data nodes per cluster. It stores each block of HDFS data in a separate file. It performs a Read/Write operation to communicate with Name node and Data node.

2. Map Reduce Programming

Map Reduce Programming is a software frame work. Map Reduce Programming helps you to process massive amounts of data in parallel. It provides a Key- value pair, Job Tracker (master) /Cluster, Task Tracker (slave)/Node, Job Configuration: Application and Job parameters, Interaction between Job tracker and task tracker

Input: Text file

Driver class: Job configuration details

Mapper class: Overrides Map function based on the problem statement

Reducer class: Overrides Reduce function based on the problem statement.

The map task is done by Mapper class and the reduce task is done by reducer class. Input data set is split into multiple pieces of data. The Framework creates several master and slave processes. There are several map tasks works simultaneously. Map workers uses partitioner function to divides the data into regions. Once map is completed reducer work begins. Mapper class tokenizing the given input and after sorts it. In next step reducing process reduces the matching pairs and produces the perfect output.

IV. R AND RHADOOP

R PROGRAMMING

R is the programming language and environment commonly used in statistical computation, data analysis, and scientific research. This is one of the most popular languages for retrieving, analyzing, displaying, and retrieving data by statisticians, data analysts, researchers and vendors. It has become popular in recent years, due to its expression syntax and easy-to-use interface. Gradually, some libraries such as R, ffbase, Rodbc, rmr2 and Rhdfs were available to handle large data. Rmr2 and rhdfs use Hadoop power in order to handle great data effectively. Map Reduce is a computational model that divides the map function into subsystems, and then the main value that results in the final release is to use the pair to seal and capture the approach. The following example is a number of objects and aggregation of a particular category

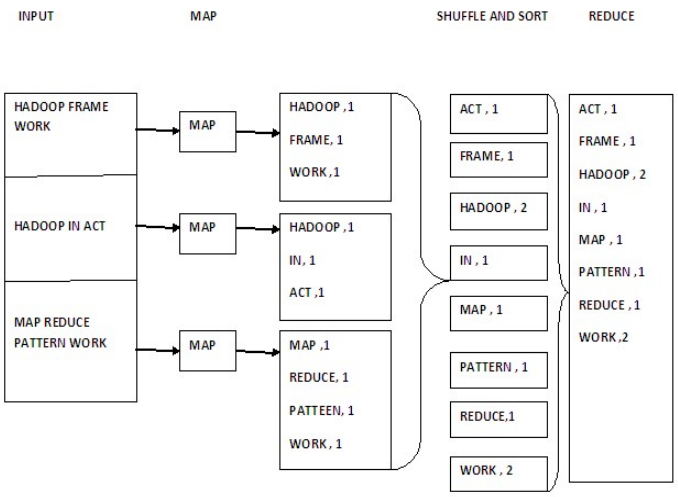

Fig 3 Map Reduce Work

R AND STREAMING

R and streaming streaming is a technology integrated into the hotspot sharing, and writes decisions to release jobs that are made from Standard Input and Works of Mapper or Reuser, as users stop operating with the script or operating system to enable Map / Users. This can be done using Streaming from R, and / or Reduced with the script in R. In this application, the customer side integration with R does not have users use the HAdoop command line to bring articles referring to the string jobs marker and reducing the scripts. Implemented muscles with a line of lines and R scripts are reduced. [4]

```
& ${HADOOP_HOME}/bin/Hadoop jar ${HADOOP_HOME}/contrib/streaming/*.jar \ -inputformat
org.apache.hadoop.mapred.TextInputFormat \ -input input_data.txt \
-output output \
-mapper /home/tst/src/map.R \
-reducer /home/tst/src/reduce.R \
-file /home/tst/src/map.R \
-file /home/tst/src/reduce.R
```

Fig 4 An Example of Map Reduce task with R and Hadoop integrated by streaming

RHADOOP ARCHITECTURE

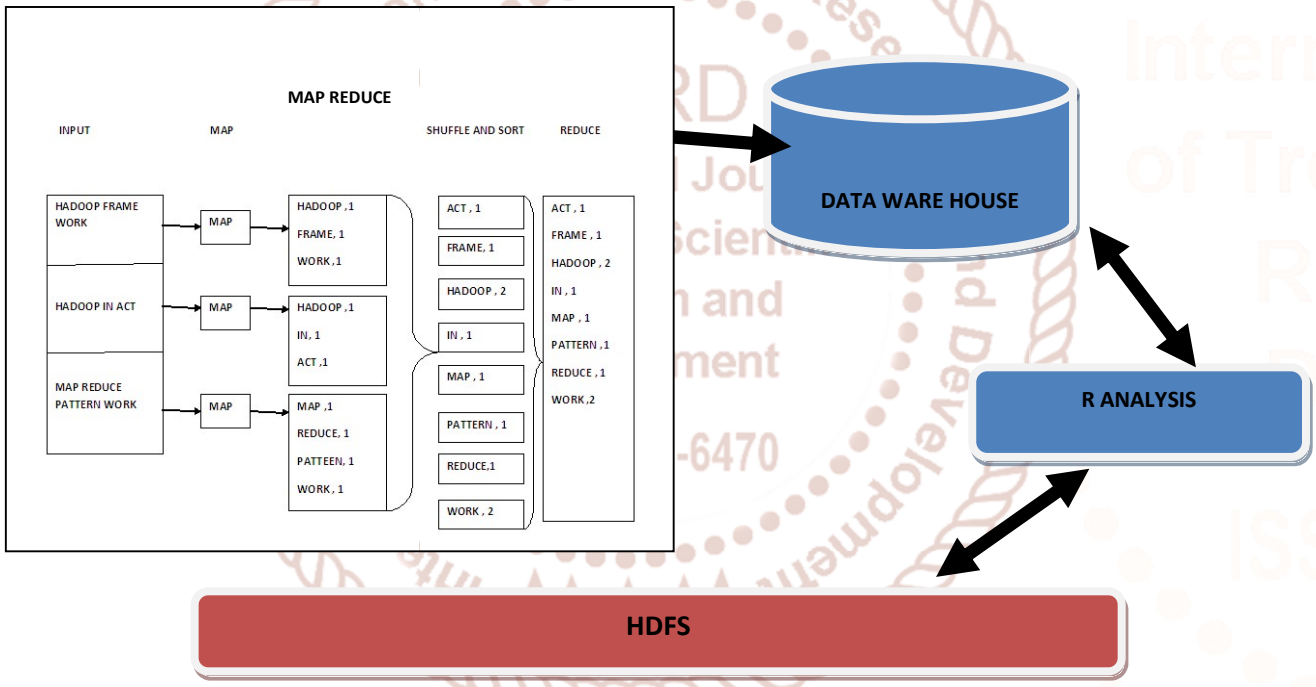

Fig 5 RHadoop Architecture

RHadoop is a set of three R packages: rmr, rhdfs and rhbase.rmr Rcpp, RJSONIO, Pittos, Gears, Functional, stringr, plyr, reshape2. The RhDfs rJava package is required. You must install these packages before installing Rmr and rhdfs. RMR RHR provides RAB and RHBs in HBase database management from HBase database management R. Rmr2 - rmr2 Hadoop's function provides us with the RADA rhdfs - rhdfs supply of Hadoop Map Reduce function. Eg file management for HDFS with rhbase - rhbase We have the R [1] with the HBase partition database Map Reduce structure is a hypothetical nerve system. Map Reduce divides and approach, which runs parallel. As a researcher, you can do this in many dimensions. First we can use our analytics to transform our work into smaller analyzes. We process our data there, reducing our small data packages and rewrite our output back to HDFS or Hbase. Based on the calculated results, we can draw the feelings of users using R.

V. CONCLUSION

RHadoop is a complete package that can make our data efficient and some useful analyzes. We have

reviewed the design and architecture of Hadoop Map Reduce architecture. Specifically, our analysis focuses on data processing. The big data results in saying that the new buzz term and Hadoop Map Reduce is the best tool that data mining and its distribution, column-based information, HBase uses its basic storage HDFS, and the best tool that provides support for the support system. It combines strong data analytics and visualization features with large data capabilities supporting Hadoop, so it certainly is worth a closer look at the RHadoop features. There are packages to connect with RR, which are important components of the Hadoop ecosystem with Map Reduce, HDFS, and HBase. In future we can activate RHadoop with the Big data protection for encryption and encryption process

REFERENCES

1. Anju Gahlawat "Big Data Analysis using R and Hadoop" International Journal of Computer Applications, Volume 108 – No 12, December 2014
2. Anshul Jatain and Amit Ranjan "A Review Study on Big Data Analysis Using R Studio" IJCSMC, Vol. 6, Issue. 6, June 2017
3. Harish D, Anusha M. S and *et.al.,* "Big Data Analysis Using RHadoop ", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Issue 4, Volume 2, April 2015
4. RaissaUskenbayeva a, AbuKuandykova and et.al., "Integrating of data using the Hadoopand R" The 12th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2015)
5. Ross Ihaka and Robert Gentleman " R : A Language for Data Analysis and Graphics" Journal of Computational and Graphical Statistics, Volume 5, Number 3, 1996
6. Shubham S. Deshmukh, Harshal Joshi and *et.al.,* "Twitter Data Analysis using R", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 6, Issue 4, April 2017
7. http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages
8. http://www.analytics-tools.com/2012/04/r-basicsintroduction-to-r-analytics.html
9. http://blog.revolutionanalytics.com/
10. http://www.r-bloggers.com/handling-large-datasetsin-r/
11. http://www.analytics-tools.com/2012/04/r-basicsintroduction-to-r-analytics.htm