ASSIGNMENT-3

Part B: Assignments based on the Hadoop

Aim:

Integrate R/Python and Hadoop and perform the following operations on forest fire dataset
1) Text mining in RHadoop
2) Data analysis using the Map Reduce in Rhadoop
3) Data mining in Hive

Introduction

R+ Hadoop:

• R will not load all data (Big Data) into machine memory. So, Hadoop can be chosen to load the data as Big Data.
• Not all algorithms work across Hadoop, and the algorithms are, in general, not R algorithms. Despite this, analytics with R have several issues related to large data.
• In order to analyze the dataset, R loads it into the memory, and if the dataset is large, it will fail with exceptions such as "cannot allocate vector of size x".
• Hence, in order to process large datasets, the processing power of R can be vastly magnified by combining it with the power of a Hadoop cluster.

RHadoop;

• If we think about a combined RHadoop system, R will take care of data analysis operations with the preliminary functions, such as data loading, exploration, analysis, and visualization, and
• Hadoop will take care of parallel data storage as well as computation power against distributed data.

Learning RHadoop:

• RHadoop is a great open source software framework of R for performing data analytics with the Hadoop platform via R functions.
• RHadoop has been developed by Revolution Analytics, which is the leading commercial provider of software and services based on the open source R project for statistical computing.
• The Rhadoop project has three different R packages: rhdfs , rmr , and rhbase .
• rhdfs : This is an R package for providing all Hadoop HDFS access to R. All distributed files can be managed with R functions.
• rmr : This is an R package for providing Hadoop MapReduce interfaces to R. With the help of this package, the Mapper and Reducer can easily be developed.
• rhbase : This is an R package for handling data at HBase distributed database through R.

Ways to Link R and Hadoop:

• Three ways to link R and Hadoop are as follows:
– RHIPE
– RHadoop
– Hadoop Streaming

RHIPE:

RHIPE stands for R and Hadoop Integrated Programming Environment.

• As mentioned on http://www.datadr.org/ , it means "in a moment" in Greek and is a merger of R and Hadoop.

• It was first developed by Saptarshi Guha for his PhD thesis in the Department of Statistics at Purdue University in 2012.

• Currently this is carried out by the Department of Statistics team at Purdue University and other active Google discussion groups.

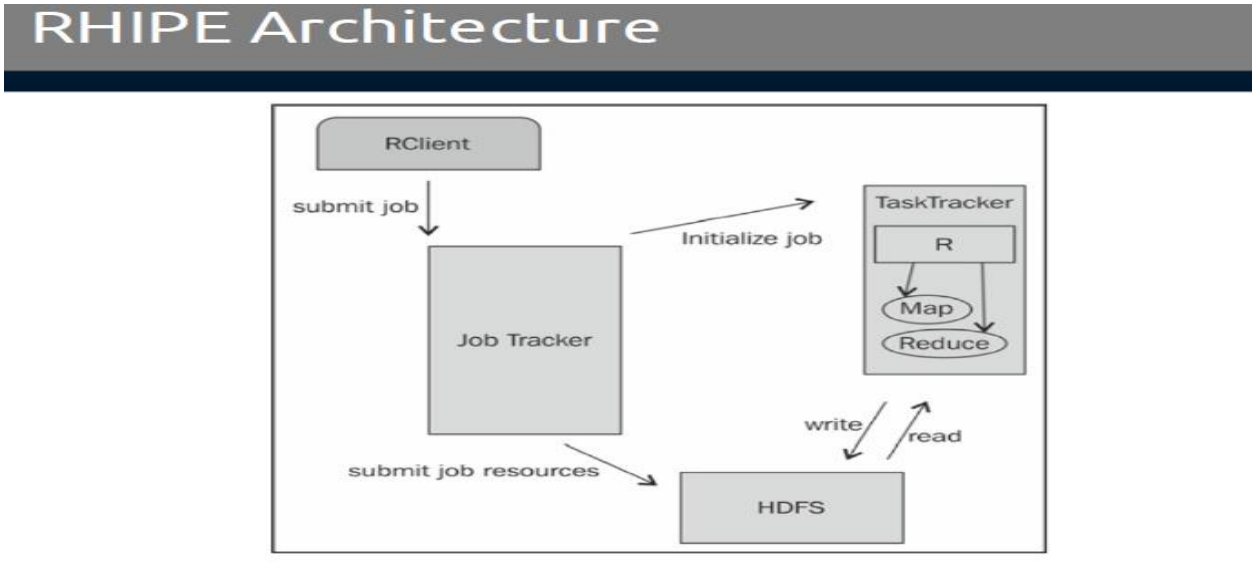
• The RHIPE package uses the Divide and Recombine technique to perform data analytics over Big Data.

• In this technique, data is divided into subsets, computation is performed over those subsets by specific R analytics operations, and the output is combined. RHIPE has mainly been designed to accomplish two goals that are as follows:

– Allowing you to perform in-depth analysis of large as well as small data.

– Allowing users to perform the analytics operations within R using a lower-level language. RHIPE is designed with several functions that help perform Hadoop Distribute File System (HDFS) as well as MapReduce operations using a simple R console.

• RHIPE is a lower-level interface as compared to HDFS and MapReduce operation. Use the latest supported version of RHIPE which is 0.73.1 as **Rhipe_0.73.1-2.tar.gz .**



RHIPE COMPONENTS:

• RClient: RClient is an R application that calls the JobTracker to execute the job with an indication of several MapReduce job resources such as Mapper, Reducer, input format, output format, input file, output file, and other several parameters that can handle the MapReduce jobs with RClient.

• JobTracker: A JobTracker is the master node of the Hadoop MapReduce operations for initializing and monitoring the MapReduce jobs over the Hadoop cluster.

• TaskTracker: TaskTracker is a slave node in the Hadoop cluster. It executes the MapReduce jobs as per the orders given by JobTracker, retrieve the input data chunks, and run R-specific Mapper and Reducer over it. Finally, the output will be written on the HDFS directory.

• HDFS: HDFS is a filesystem distributed over Hadoop clusters with several data nodes. It provides data services for various data operations.

RHADOOP:

RHadoop is a collection of three R packages for providing large data operations with an R environment.

• It was developed by Revolution Analytics, which is the leading commercial provider of software based on R.

• RHadoop is available with three main R packages: rhdfs , rmr , and rhbase . Each of them offers different Hadoop features.

RHDFS:

• rhdfs is an R interface for providing the HDFS usability from the R console.

• As Hadoop MapReduce programs write their output on HDFS, it is very easy to access them by calling the rhdfs methods.

• The R programmer can easily perform read and write operations on distributed data files.

• Basically, rhdfs package calls the HDFS API in backend to operate data sources stored on HDFS.

RMR:

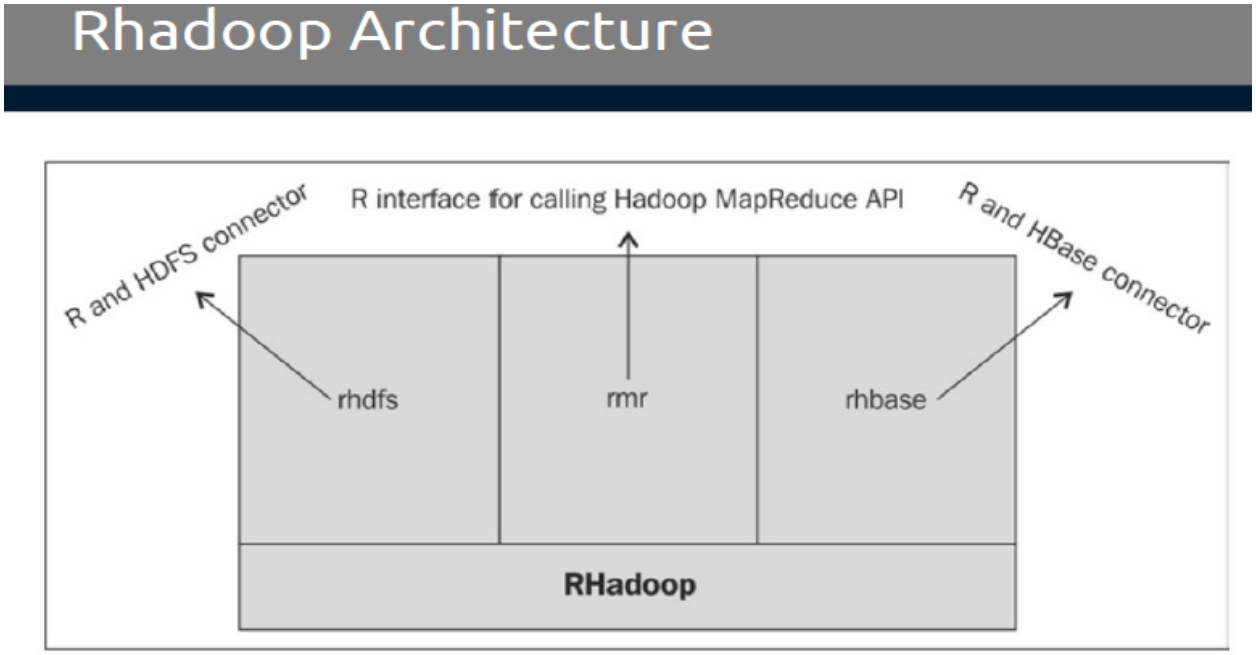
• rmr is an R interface for providing Hadoop MapReduce facility inside the R environment.

• So, the R programmer needs to just divide their application logic into the map and reduce phases and submit it with the rmr methods.

• After that, rmr calls the Hadoop streaming MapReduce API with several job parameters as input directory, output directory, mapper, reducer, and so on, to perform the R MapReduce job over Hadoop cluster.

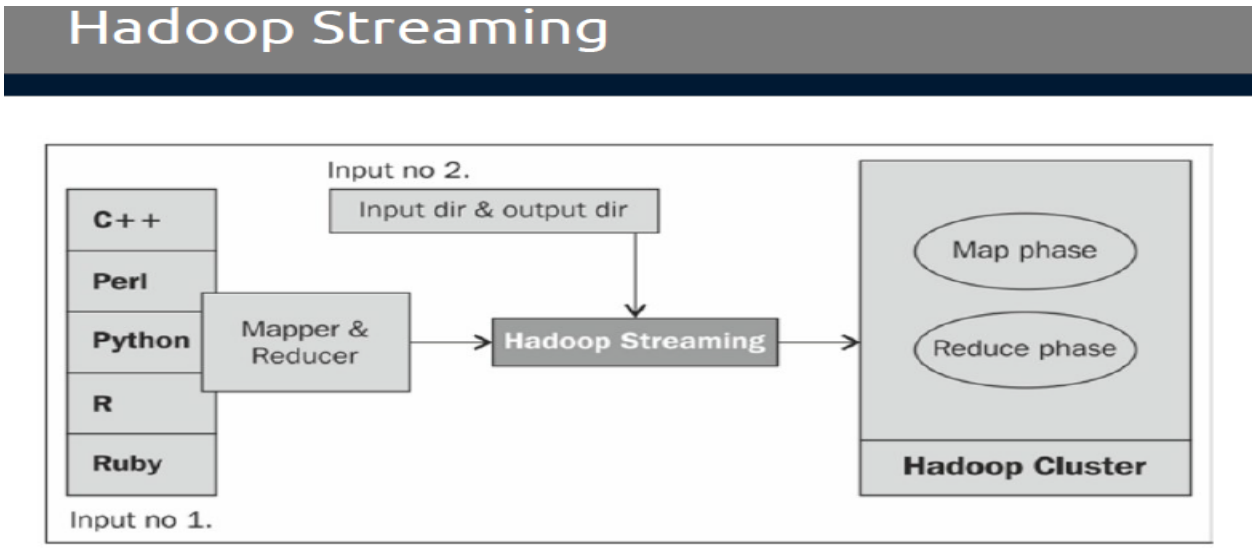
R HBASE:

• rhbase is an R interface for operating the Hadoop HBase data source stored at the distributed network via a Thrift server.
• The rhbase package is designed with several methods for initialization and read/write and table manipulation operations.



HADOOP STREAMING:

• Hadoop streaming is a Hadoop utility for running the Hadoop MapReduce job with executable scripts such as Mapper and Reducer.
• This is similar to the pipe operation in Linux.
• With this, the text input file is printed on stream (stdin), which is provided as an input to Mapper and the output (stdout) of Mapper is provided as an input to Reducer; finally, Reducer writes the output to the HDFS directory.
• The main advantage of the Hadoop streaming utility is that it allows Java as well as non-Java programmed MapReduce jobs to be executed over Hadoop clusters.
• Also, it takes care of the progress of running MapReduce jobs.
• The Hadoop streaming supports the Perl, Python, PHP, R, and C++ programming languages.
• To run an application written in other programming languages, the developer just needs to translate the application logic into the Mapper and Reducer sections with the key and value output elements.



Conclusion: Thus we have acquired the brief knowledge of (1. Text mining in RHadoop 2. Data analysis using the Map Reduce in Rhadoop 3. Data mining in Hive) with **R Language .**