DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5414

Volume 9, No. 1, January-February 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

MAPREDUCE: INSIGHT ANALYSIS OF BIG DATA VIA PARALLEL DATA PROCESSING USING JAVA PROGRAMMING, HIVE AND APACHE PIG

Dr. Ujjwal Agarwal
Lecturer (I.T.), Salalah College of Technology,
Salalah, Sultanate of Oman

Abstract: Digitaldata which come from different sources like office, school, hospital, social media or machine generated data. Apache Hadoop is a software framework to store and process this enormous amount of data. Hadoop is using HDFS and MapReduce to store and process this huge volume of data. MapReduce is a programming model initiated by Google which can be written in different programming languages like Java, Python and Ruby. The main objective of this paper is to describe the concepts of MapReduce and showing the operation by using Java Program, Apache Pig and Hive.Hive and Apache Pig working on top layer of Hadoop ecosystem and provide the level of abstraction to run the MapReduce jobs. In this research paper we write MapReduce program in Java to find anagram words from input files, group them together and save the result in output file.At the end we perform the operation in HiveQL (Hive query language) and Pig Latin Script and showing the backend process in MapReduce job.

Index Terms— Big Data, Hadoop, MapReduce, Hive. Apache Pig

I. INTRODUCTION

1.1 Big Data: In today's world data is everywhere, either in shop, office, hospital, colleges, universities, mobile application, websites, digital devices, etc.Technologies are changing with extremely high rate and producing a huge amount of data, which comes in a variety of form along with a high velocity. Big data is a growing term that describes any huge amount of structured , semi structured and un-structured data that has the to be used for analysis [1]. There are three sources of big data:

a) Organization generated data
b) People generated data
c) Machine generated data

Organization generated data: is the most structured data because all the data is stored in different databases either in MS ACCESS, Oracle, MS SQL etc. in form of tables, rows and column. To search any information in database is an easy task as the data already in particular shape.

People generated data: is in semi-structured or un-structured form like Facebook messages, twitters tweet, Google search query, pictures on Instagram,

Picasa. Social media is main platform where people will generate terabyte/petabyte of data every day.

Machine generated data: This is the major source of big data about 85% - 90% of the total data is generated by machine. This data is generated by various real time sensors, camera, mobile; devices like sleeping monitor, heart beat monitoring devices. The widespread availability of smart devices like smart phone, smart cars, smart homes, smart camera, now sea, oceans, and sky also connected to this devises and all together producing a huge amount of data every minute, hours and days. In big data the quantity or volume of data is very high, every minute total amount of data is generated may be in terabyte or petabyte. This data is generated at a very high velocity for example every minute Facebook generated about 350GB of data, 72 hours video uploaded on YouTube. Data can be Batch, near time, real time or stream data.

As data is coming from various sources therefore the structure of big data varies it can be structure, semi-structure or un-structure.

1.2 Hadoop: Apache Hadoop is an open-source framework which is used to store large amount of data using HDFS and process this data by using the MapReduceprogramming model. It consists of computer clusters built from commodity hardware. Hadoop is designed in such a way that always it assumes hardware failures are very common and it should be automatically handled by the framework. [2]. Hadoop is a major platform for storing and analysis of big data. The first release of Hadoop was launched on 10th December 2011 and the first stable version (2.7.3) came into existence on 25th August 2016. The base component of Hadoop is HDFS (Hadoop Distributed File System). HDFS works on Name Node and Data Node. HDFS works on Master-slave architecture as it has one name node and multiple data node. Name Node is the controller/master of the system. Name node spreads data to data node. It stores the metadata of all the files in the HDFS. This metadata includes name, location of each block, block size and file permission. The data that we store in Hadoop is store in different clusters across the nodes. By default the replication factor of any data is 3 on data node.

1.3 Hadoop Ecosystem: Hadoop is a software framework which is used for storage and analysis of Big Data. Hadoop Ecosystem tied up with many different applications some for structured data like sqoop and hive, some for semi-structured and un-structured. In the core Hadoop distributed file system which responsible to store all the data in different racks. Above that MapReduce will work, which provide the parallel data processing and above all different application is working which provides the level of abstraction like Pig, Hive or Sqoop etc.

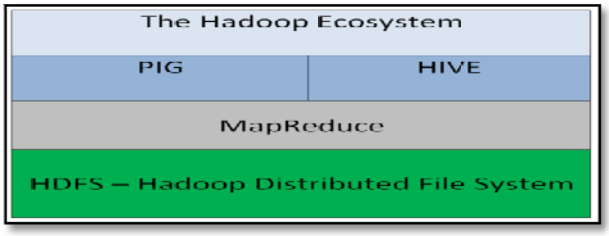

Figure 1.1 Hadoop Ecosystem

II MAPREDUCE FRAMEWORK

MapReduce [1, 2] is a programming model for data processing. This model is very simple but still too complex. Hadoop can run MapReduce program written in different language, like Java, Python and Ruby. The term MapReduce has two different operations that Apache Hadoop program performs. The first step is MAP and another step is REDUCE, but between map and reduce we have two another steps called sort and shuffle. The first step is our Map which takes the group of data or files and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

The next step is sort and shuffle, which takes the output of Map step and sort and shuffle the result, based on (key, value) pairs.

The last sand final step is reducing step,reduce job takes the output from sort and shuffle as input and combines values from Shuffling phase and returns a single output value. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.[3], [4] [5], [6]

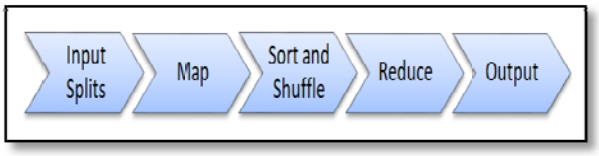

Figure 2.1 Architecture of MapReduce

We can divide MapReduce task as two different categories called:-

a) MapReduce with single reduce task

b) MapReduce with multiple reduce task

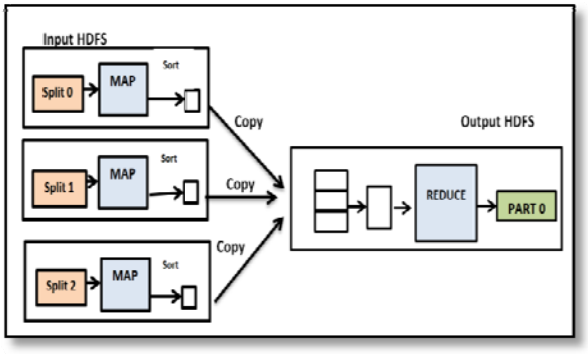

Figure 2.2 MapReduce Single Reduce Task

By default number of reducers is set to 1, means always only one reducer task will be executed as shown below.

Many reducers can run in parallel, as they are working independently. The number of reducers is decided by the user.We can change the number of reducer by using the following commands:

```
Job job = new Job(conf);
job.setNumReduceTasks(3); // 3 Reducers
                    or
Configuration conf = new Configuration();
conf.set("mapreduce.job.reduces", "5"); //5 Reducer
```

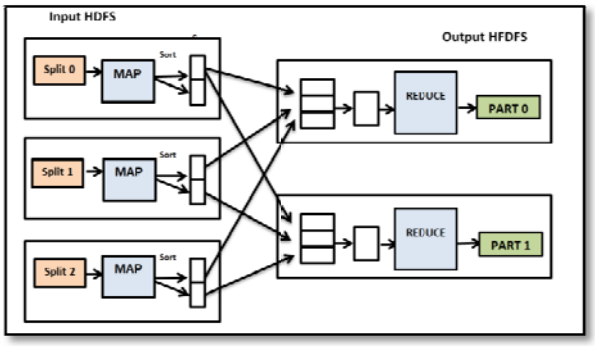

Figure 2.3 MapReduce Multiple Reduce Task

MapReduce Framework: The MapReduce framework working on <Key,Value> pair. Firstly the Map operation takes the input job as a pair of <Key, Value> and produce the output again in the form <key, Value> which will be the input for Reduce job.

The key and the value classes should be in serialized manner by the framework and hence, need to implement the Writable interface. Additionally, the key classes have to implement the Writable-Comparable interface to facilitate sorting by the framework. Input and Output types of a MapReduce job: (Input) <k1, v1> -> map -><k2, v2>-> reduce -><k3, v3> (Output).

	Input Value	Output Value
Map Job	<key1, value1>	List <key2, value2>
Reduce	<key2, List (value2) >	List (Key3, value3>

Table 2.1

III MAPREDUCE IMPLEMENTATION

There are different ways to execute MapReduce jobs:-

A) The first way to execute by using Java MapReduce program, these programs can be used for structured, semi-structured and un-structured data like .csv, .json, or DBMS files. A MapReduce program usually consists of the three different classes:

a) Mapper Class
b) Reducer Class
c) Driver Class

B) Apache Pig is working on the top layer of Hadoop ecosystem and it is used to analysis of structure and semi-structure data and used to process the scripting approach for MapReduce. [7]

C) The Hive Query Language (HiveQL or HQL) for MapReduce is used for analysis of structured data which in the form of row and columns. [8]

3.1. To implement the MapReduce program in Java we are taking the example of an anagram. Anagram is a word or set of character formed by rearranging the letters of a different word or characters, For example, the world “hadoop” can be rearranged into "adhpoo".Input files contain n number of varying string, we need to identify anagrams words and group them together.For solving this problem by using MapReduce we are usingthree classes:-

a) AnagramMapper
b) AnagramReducer
c) AnagramDriver

AnagramMapper Class(SortKeyMapper): This will read input file line by line.It will create a token from line and sort each toke to form a key with original value.In this way Hadoop will create a group of similar words.For example: aba and baa will get sorted as aab. So reducer will receive a key as aab with group {aba,baa}

```
public void map(Text key, Text value, Context context)
throwsIOException, InterruptedException
{
System.out.println("In mapper");
String keystring = key.toString();
StringTokenizerst = newtringTokenizer(keystring," ");
String sorted = null ;
while(st.countTokens()!=0)
{
String s1 = newString();
s1=(String) st.nextElement();
char[] chars =s1.toLowerCase().toCharArray();
Arrays.sort(chars);
sorted = new String(chars);
context.write(new Text(sorted), new Text(s1));
}
}
```

AnagramReducerClass: It will read each member of group and combine them into single string separated by tab.

```
publicvoid reduce(Text
key,Iterable<Text>values,Contextcontext)
throwsIOException,InterruptedException
{
StringBufferstr=newStringBuffer();
String finalkey =values.iterator().next().toString();
for (Text t : values)
{
str=str.append(t.toString()+" ");
}
context.write(new Text(finalkey), newext(str.toString()));
}
```

AnagramDriver Class: The driver program controls the execution of program by calling both the classes first it will call the AnagramMapper class then it will call AnagramReducer class and finally write the results in new file.

```
public static void main(String[] args) throws IOException,
ClassNotFoundException, InterruptedException{

jobjob = new Job();
job.setJarByClass(AnagramDriver.class);
job.setJobName("Ana");

FileInputFormat.addInputPath(job, new
Path("C:\\Users\\ujjwal\\eclipse-
workspace\\Anagram1\\hadoop\\anagram.txt"));

FileOutputFormat.setOutputPath(job, new
Path("C:\\Users\\ujjwal\\eclipse-
workspace\\Anagram1\\hadoop\\out_ana"));

job.setInputFormatClass(KeyValueTextInputFormat.class);
job.setMapperClass(AnagramMapper.class);
job.setReducerClass(AnagramReducer.class);
job.setMapOutputKeyClass(Text.class);
job.setMapOutputValueClass(Text.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);

System.exit(job.waitForCompletion(true)?0:1);
}
```

3.2: Hive: Hive is data warehouse tool which is used to process and analysis of structured data in the form of tables and databases. Hive is working on top layer of Hadoop ecosystem. Hive has mainly three basic functions:-

a) Data summarization
b) Query
c) Analysis of data

Hive supports query language called HiveQL, which translate SQL like query into the MapReduce jobs. Once we execute the query it will pass to Job Tracker and then Job Tracker pass the job to Task Tracker and finally the Map and Reduce task will execute and fetch the data from HDFS.

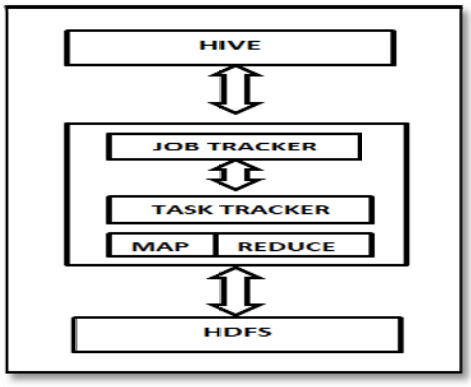

Figure 3.2.1Architecture of Hive

To exhibit the working structure of MapReduce in Hive first we take the databasecalled “employee” and inside the employee database we have table called “emp”. Our main objective is to fine the total number of rows in the “emp table”. We will execute the following HiveQL query:-

```
Hive > use employee;
Hive >Select Count (*) from emp;
```

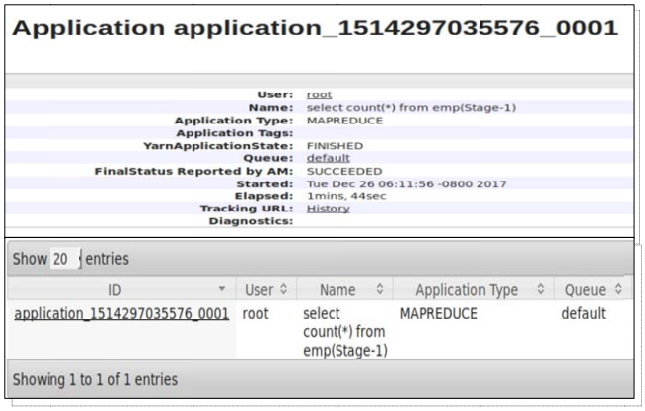

Figure 3.2.2 MapReduce job for HiveQL query

To execute this query, Hive performs the MapReduce job as shown in 3.2.2 and the total time required is 1 min and 22 second.

3.3 Apache Pig: Apache Pig is an abstraction over MapReduce. Pig is used for analysis of large or distributed data set without using writing the program of MapReduce.We can perform all the data manipulation operations in Hadoop using Apache Pig by using scripting language called Pig Latin. Every pig program has three parts:-

a) Loading
b) Transforming
c) Dumping of the data.

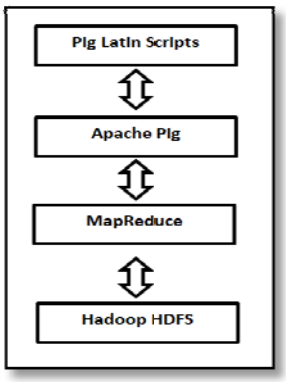

Figure 3.3.1 Architecture of Apache Pig

The following command we will execute to display the MapReduce execution on Apache Pig:-

$ pig –version

$ pig –x mapreduce

Grant>A=LOAD 'hdfs://localhost:9000/pig_data/ student.txt' USING PigStorage(',') as (id:int, name:chararray , city:chararray);/*Load File */

Using dump method, the processed data is displayed on the standard output.

Grant >Dump A;

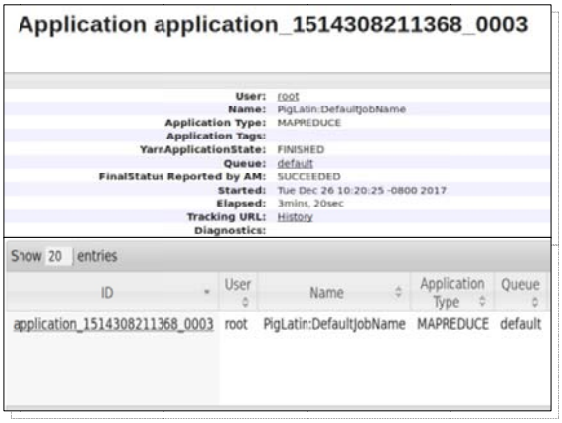

Figure 3.3.2 MapReduce job for Pig Latin Script

To execute this Pig Script,Pigperforms the MapReduce job as shown in 3.2.2 and the total time required is 3 min and 20 second.

IV CONCLUSION

This paper exploits an overview ofbigdata, Hadoop ecosystem and deeply explained MapReduce architecture. MapReduce programming model used to deal with huge volume of data by using parallel data processing. MapReduce break the job in two groups first Map job will execute then Reduce job will perform. MapReduce provides this feature in Hadoop ecosystem so that data can be break down as per <key,value> given by the programmer. By default there is one reducer but we can change the number of reducer as per need. We solve anagram problem by MapReduce job in Java programming language and further we execute the HiveQL and Pig Latin Script to execute the query and showing that at backend the whole query is executed by MapReduce job. Hive and Apache Pig working on top layer of Hadoop ecosystem and provide the level of abstraction, as user no need to write the code of MapReduce but just he write the SQL like script in hive its HiveQL and for Pig its Pig Latin Script, that automatically converted into MapReduce Jobs.

V REFERENCE

[1] Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1,January 2010, pp 72-77.
[2]Jefry Dean and Sanjay hemwat,.MapReduce: Simplified data processing on large clusters, Communications of the ACM, Volume 51 pp. 107–113, 2008
[3] M. Dhavapriya, N. Yasodha, Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table, (IJCST) – Volume 4 Issue 1, Jan - Feb 2016
[4] Dr. Urmila R. Pol, Big Data Analysis Using Hadoop Mapreduce, American Journal of Engineering Research (AJER), Volume-5, Issue-6, pp-146-151
[5]AbdelrahmanElsayed, Osama Ismail, and Mohamed E. El-Sharkawi, MapReduce: State-of-the-Art and Research Directions, International Journal of Computer and Electrical Engineering, Vol. 6, No. 1, February 2014
[6]Shafali Agarwal, Map Reduce: A Survey Paper on Recent Expansion, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 8, 2015

[7] Pig Latin: A Not-So-Foreign Language for Data Processing, Christopher Olston∗ Yahoo! Research Benjamin Reed † Yahoo! Research UtkarshSrivastava ‡ Yahoo! Research

[8] Hive – A Petabyte Scale Data Warehouse Using Hadoop, AshishThusoo, JoydeepSenSarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy, Facebook Data Infrastructure Team

VI PROFILE

Authors Profile

Dr. Ujjwal Agarwal completed his PhD(Computer Science), M.Phil. (Computer Science) and MSc(Information Technology). He is currently working in Salalah College of Technology, Salalah, Sultanate of Oman as a Lecturer (IT) alogwith he is member in many International Journals. He has published more than 12 research papers in reputed international journals and conferences. His main research work focuses on Big Data Analytics, Machine Learning using Python and R Programming language. He has more then 12 years of teaching experience and 6 years of Research Experience.