AN ANALYTICAL APPROACH TO ANALYZE WEB DATA

Kirti Bhatia[1], Lalit[2]

[1]*HOD, Department of Computer Science, SKITM Bahadurgarh, Haryana, India*
bhatia.kirti.it@gmail.com
[2]*M Tech 4th sem, SKITM Bahadurgarh, Haryana, India*
lalit.dang3@gmail.com

Abstract

The basis of this dissertation is the rise of "big data" and the use of analytics to mine that data. Big data is basically used to analyze the large amount of data. Companies have been storing and analyzing large volumes of data. In the recent times, data warehousing was mostly used for this purpose which was invented in the early 1990s. Big Data warehouses means terabytes that is data stored in the warehouses was terabytes in storage. But now it is petabytes, and the data is growing at a high speed. The growth of data is because the companies and organizations store and analyze the greater levels of transaction details, as well as Web data and machine-generated data, to achieve a better understanding of customer behavior and market behavior. Big data helps to achieve this goal. A large number of companies are using Big Data. It seems in coming years the growth of big data concept must be higher.

Keywords: *Web Data, Big Data, Large Data, Growth of Data,*

----------***----------

1. INTRODUCTION

This dissertation is in the area of data mining in the large organizations to keep pace with the desire to store and analyze ever larger volumes of structured data. The vendors of RDBMS have given a number of special platforms that provide higher levels of price improvement and also a higher level of performance as compared to general purpose Relational DBMS. The platforms used for this purpose are available in a variety of shapes and sizes, from software only databases [1]. A large number of persons who make survey or we can say a large number of survey respondents said that they have an analytical platform that is applicable to this description.

As we know, new technologies are coming also have come to handle the complex data. These technology can handle a large volumes of this complex data which also includes the Web Data.

By web data, I mean the social media data like Facebook data. As we all use Facebook that is have an account on Facebook. A large number of pictures are uploaded to the Facebook server the content on Facebook server is very large and this type of content is social media content. New technologies also handle or overlook the machine generated data or the sensors and data like GPS that is Global Positioning System. These type of data is big data. By Big Data means very big data.

I have discuss in above paragraph that big data or large volume of data is handled or we can also say analyzed by the new technologies. Now what are these new technologies? Let's discuss them in brief. A large number of companies, which includes Internet and also social media companies, are using new open source frameworks such as Hadoop and MapReduce to analyze, store and process large volumes of data which can be structured data also unstructured data, can be Web data or any typical data. Which also includes the data in batch jobs that run on a plenty of database servers.

The term BIG DATA is used to address the data sets that are too large and too complex than the traditional data that the traditional processing applications are not enough to handle or process this large and complex data. Traditional processing applications cannot process such large volume of data because of some issues. These issues include analysis, capture, search, sharing storage transfer, querying, and information privacy. These are the some challenges which comes while processing such large volume data or complex data by traditional processing system. Big Data refers to the use of some analytics to extract value from the complex data or it refers to other methods like this to extract the value or to summarize the complex and large data. These analytical methods can also be used to summarize the web data. The main advantage of Big Data is its accuracy which leads to more confidence in decision making. If decisions are better they can result in greater performance, greater efficiency, increase productivity and also the cost effective that is can help in the cost reduction also can help in the reduction of risks.

Big data includes very large volume data sets and very complex datasets with the sizes which cannot be processed by commonly used software tools to process and analyze the traditional data within a small limit of size and tolerable time period [1]. In concept of Big data size is a constantly increasing ranging from gigabytes to terabytes and now to petabytes. This requires a new set of techniques with the extended power to process the large volume data.

In an article published in IEEE Access Journal the writer of the article defined the big data definitions into three types: Attribute Definition, Comparative Definition and Architectural Definition. The author also showed a big-data map that shows the key features of big data. By analysis of data sets, we can find new techniques to find market trends also find ways to prevent serious diseases and so on [15].

1.1 Hadoop: Big Data Overview

Hadoop is key frame work in the analysis of big data. It is an open-source framework that allows to store and process large volume data, structured and unstructured complex data [3]. This framework is designed to analyze very large data and also can be used to scale up from single servers to thousands of servers, with each server has capacity to perform local computation and has its local storage. With the invention of new technologies, new devices and new and advance communication means like Facebook, Twitter and WhatsApp, the data is growing in volume at a very high speed. Data produced by only social networking sites is the amount of data that if we bind up the data into form of disks it might fill an entire football ground. This shows that social networking sites are producing very large amount of data. The data produced by these sites was in gigabyte at the start of 21st century. But it grew at a very large speed. Then in terabytes and now in petabytes. In coming years this data is going to be very large due to advent of new technologies and new communication devices. To process this large data we were requiring a framework. Then Hadoop came into picture which is open source. This framework is used to process and analyze the big data. With the help of Hadoop we can run applications with a large number of nodes with large number data which may be in terabytes. This allows the very fast transfer of data between nodes. It is also very reliable framework which allows continue operating if even any node fails [4]. It also minimize the risk and chance of system failure even if a good number of nodes get failed.

1.2 Mapreduce: Heart of Big Data

MapReduce is the heart of Big Data. MapReduce is also a framework which is used to write applications to process and analyze a large amounts of data in parallel on big servers in a reliable manner. It is programming method or programming technique which allows for large scalability across a large number of servers, may be hundreds or even thousands. The MapReduce concept is very simple to understand. For those who are familiar with large scale-out data processing solutions, this concept is simpler. For new commerce it is somewhat difficult to understand because it is not like traditional system. MapReduce actually refers to two different tasks which is performed by Hadoop [2]. The first task is the Map task, which takes input as a big data in the form of data sets and converts them into another set of data, in which elements are broken down into further lower elements called tuples. Tuples are the key/value pairs. The Second task is the Reduce task. Reduce task takes the output from a Map task as input and combines those datasets tuples into further smaller tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

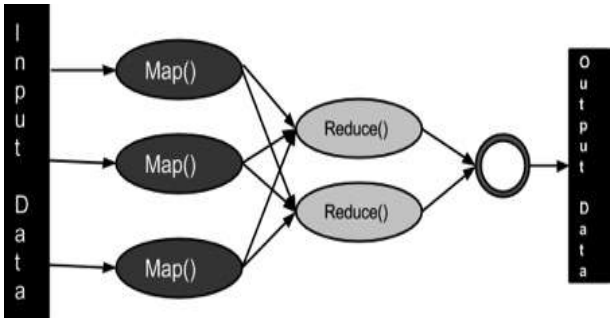
The main advantage of MapReduce is its accuracy and also it is easier to process data over a large number of computing nodes. In the MapReduce model, the data processing attributes are called mappers and reducers. MapReduce decompose the data processing applications into Mappers and Reducers. This Decomposition of application into mappers and reducers is sometimes nontrivial that it may have some variables or terms that are not equal to zero or that are not equal to identity. But, once we had written an application using MapReduce model that is we had decomposed the application into Mapper and Reducer then we can process or scale the application to run over a large number, may be hundreds, thousands or might be tens of thousands of machines in a cluster by changing only few configurations. This is the main feature of MapReduce model which enables many programmers to use the MapReduce model. Generally, MapReduce model is based on sending the computer to the place where the actual data resides [4].

The execution of MapReduce program consist of three stages which are: Map stage, Shuffle stage and Reduce stage.

Map stage: The Map stage or Mapper's task is to process the big data which is taken as input to mapper. This input data is in the form of datasets or file or directory which is stored in the Hadoop file system (HDFS). The input datasets are passed to the Mapper function line by line. The Mapper processes the data and decompose them into smaller elements or creates several small chunks of data [15].

Reduce stage: The Reduce stage or Reducer's task is to process the data which is output from the Mapper. After processing the data which comes from Mapper, Reducer produces a new set of output, which is then stored in the HDFS [15].

In this process of MapReduce, Hadoop sends the Map and Reduce tasks to the respective servers in the cluster. The Hadoop framework manages all the processing and details of data-passing such as issuing tasks, verifying task completion, and transfer the data around the cluster between the nodes. Most of the computation takes place on nodes with data on local disks due to which the traffic on the network gets reduced. After these tasks are completed, the framework collects the data and reduces it to form the required appropriate result, and then sends it back to the Hadoop server [16].



2. BIG DATA SIGNIFICANCE

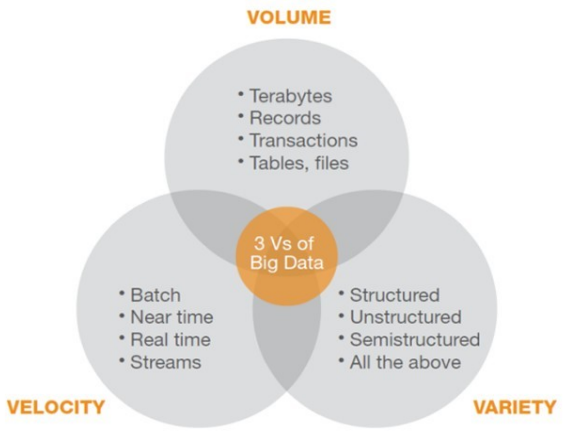
Big Data is growing at a regular interval at a very high speed. It is becoming so large and complex that it is difficult to process, store, manage, share and analyze within current computational powers or with the traditional processing applications. Big data is very important field. It is becoming the growing field. Almost all big organizations and social media companies are lying on Big Data. There are four key characteristics of big data which make them so important and significant. These characteristics are Volume, Velocity and Variety. These are called 3vs of Big Data.

Volume: Volume is the most important characteristic of Big Data. As we all know, Web Data is growing in volume. With the help of Big Data frameworks available in market like Hadoop and MapReduce, we can manage, store process and analyses such large volume of data easily [11].

Velocity: By the advent of Big Data it is data streaming at a high speed is now possible. This type of framework is real-time. For example we can easily stream video on YouTube. Also the larger volume of Twitter data ensured high velocity of even at 140 characters per tweet.

Variety: This characteristic refers to the fact that Big Data is capable of processing a variety of data which includes Structured, Unstructured, Semi-Structured, Text, Images and other media and so on. With the help of Big Data technology, we are now capable of analyzing and processing these variety of data.

Finally, social media sites like Facebook, Twitter and LinkedIn would not exist without the big data.



3. PROPOSED WORK

This dissertation is about analyzing the data of YouTube. This analysis is performed using the Hadoop and MapReduce frameworks. The YouTube data is open source and publically available to the users. This YouTube data set is described below under the Data Set Description section. Using this dataset we will perform some Analysis and will find some results of this analysis like what are the top rated videos on YouTube also who uploaded the most number of videos on the YouTube. By using, MapReduce it is possible to analyze the web data. The Analysis is described below.

DATA SET DESCRIPTION

The data set or data table for YouTube consist of following information.
1: A Unique Video id.
2: Details about Uploader of the video.
3: Time gap between date of uploading of the video and YouTube establishment date.
4: Category to which video belongs to.
5: Length in time of the video.
6: Number of views till now.
7: Rating.
8: Number of ratings.
9: Number of comments.
10: Ids of related videos.

Problem Statement 1
Our motive is to find out the top 5 categories having maximum number of videos uploaded.

SOURCE CODE

Now by using Mapper code we will get the video category as key and final int value as values which will be passed to the next stage which is shuffle and sort phase and then this value is sent to the reducer phase where the aggregation of the values is performed.

Mapper Code

```
public class TopFiveCategories {
public static class Map_1 extends
Mapper<LongWritable, Text, Text,
IntWritable>
{
private Text cat = new Text();
private final static IntWritable one_1 = new IntWritable(1);
public void map1(LongWritable key, Text value, Context
context )
throws IOException, InterruptedException
{
String one_line = value.toString();
String str[]=one_line.split("\t");
if(str.length > 5){
cat.set(str[3]);
}
context.write(cat, First);
}
}
```

Explanation of the above Mapper code:
In **line 1** we are taking a class by name TopFiveCategories.
In **line 2** we are extending the Mapper default class with arguments KeyIN, ValueIn, KeyOut, ValueOut
In **line 3** we are declaring a private Text variable 'cat' that is category which will store the category of videos on YouTube.
In **line 4** we are declaring a variable which is private final static IntWritable variable one_1 which will be a constant value for every value.
In **line 5** we are overriding the map_1 method .
In **line 7** we are storing the line in a string variable one_line.
In **line 8** the line is splitted using comma "," delimiter and value is stored in a array of string so that all the columns in a row are stored in the string array.
In **line 9** if block is used to check whether the string array of length is greater than 6 which means it will enter into the if block and execute the code to eliminate the Exceptions.
In **line 10** we are storing the cat that is category.
In **line 12** we are writing the key and .Which will be the output of the Mapper's Map_1 method.

Reducer Code

```
public static class ReduceCat extends Reducer<Text, IntWritable,Text,IntWritable>
{
public void reducecat(Text key, Iterable<IntWritable> values,Context context throws IOException, InterruptedException
{
int sumcat = 0;
for (IntWritable val : values)
{
sumcat += val.get();
}
context.write(key, new IntWritable(sumcat));
}
}
```

While coming to the Reducer code:
line 1 a class Reducercat which is extended the default Reducer class with arguments KeyIn , ValueIn, same as the outputs of the mapper class and , ValueOut that is used to store final outputs of our MapReduce program.
In **line 2** Reduce method is overridden which is run each time for every key.
In **line 3** An integer variable sumcat is declared which is used to store the sum of all the values for each key.
In **line 4** A Loop which is foreach is taken. It will run each time for the values inside the Iterable values. Which are coming from the second phase that is shuffle and sort phase after the mapper phase.
In **line 5** Value for sumcat that is "sum" is calculated and stored.
In **line 7** sumcat is obtained as value to next context.

How to Execute

hadoop jar topfivecategories.jar /youtubedata.txt /topfivecategories_out

Here '**hadoop'** is a command and jar specifies that we are running java type of application and **topfivecategories.jar** is the jar file which is created and consists of the above source code.

How to view output

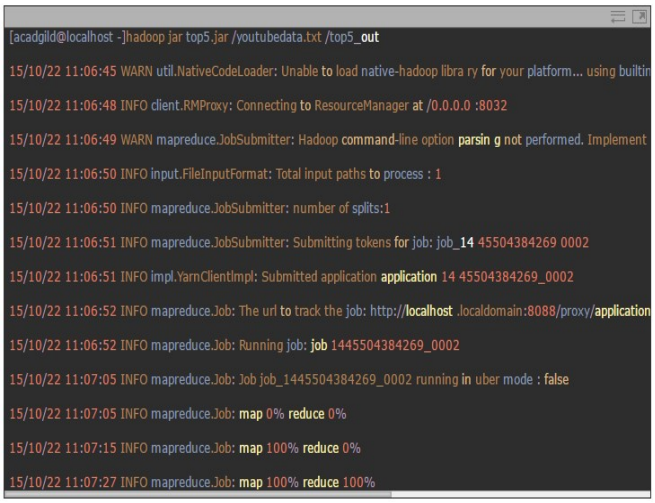
hadoop fs-cat/topfivecategories_out/part-r-00000 | sort –n –k2 –r | head –n5

Here **'hadoop'** is a command. And **dfs** is related to the File System of Hadoop (HDFS) which is used to perform some operations on Hadoop Distributed file System. The **– cat** command is used to view the contents of a file and **topfivecategories/part-r-00000** is the file where final output is stored.

This Part file is created by Text Input Format which consist of the actual final output. This Part file is created by default. Then, **sort –n –k2 –r | head –n5** is the main command which shows the top 5 categories with maximum number of videos uploaded. Also we can mention the secondary sort instead of this command.

Sort is used to sort the data, **–n** stands for numerically that is sorting will be numerically,
–k2 stands for second column, **–r** stands for recursive operation, **–n5** stands for the first 5 values after sorting.

Output



Problem Statement 2
In this problem statement we will find the top 10 best rated videos in YouTube.

SOURCE CODE

Now by using Mapper code we will get the video id as key and rating as final Int value as values which will be passed to the next stage which is shuffle and sort phase and then this value is sent to the reducer phase where the aggregation of the values is performed.

Mapper Code

```
1. public class Top_Video_Rating {
2. public static class Map_1 extends
   Mapper<LongWritable, Text, Text,
3. FloatWritable> {
4. private Text top_video_name = new Text();
5. private  FloatWritable top_rating = new
   FloatWritable();
6. public void map_1(LongWritable key, Text
   value, Context context )
7. throws IOException, InterruptedException {
8. String top_line = value.toString();
9.          If(top_line.length()>0) {
10.          String str_1[]=top_lineline.split("\t");
11.               top_video_name.set(str_1[0]);
12.               if(str_1[6].matches("\\d+.+")){
13.               float f=Float.parseFloat(str_1[6]);
14.               top_rating.set(f);
15.         }
16.         }
17. context.write(top_video_name, top_rating);
18.         }
19.         }
20.         }
```

Explanation of the above code
In **line 1** we are creating a class with name Top_Video_rating
In **line 2** Map_1 is extended by Mapper default class. Which have the arguments keyIn, ValueIn, KeyOut and ValueOut.
In **line 4** a variable is declared which is a private Text variable top_video_name which is used to store the video name in the encrypted format.
In **line 5** a variable is declared which is private FloatWritable with name top_rating which is used to store the rating of the video.
In **line 6** map_1 method is overridden which is run one time for every line.
In **line 8** a string variable top_line is declared which is used to store the line.
In **line 9** top_line is splitted by using comma *","* delimiter and values are stored in a string array.
In **line 10** a if block is declared to check weather the string array length greater than 7 or less than 7.
In **line 11** video name is stored in the variable top_video_name declared in 2nd line
In **line 13** numeric data is converting into float type by using type cast method
In **line 14** the rating of the video is stored in the variable top_rating.
In **line 17** we are writing the key and .Which will be the output of the Mapper's Map_1 method.

Reducer Code

```
public static class Top_Reduce extends
Reducer<Text, FloatWritable,Text,
FloatWritable> {
public void Top_Reduce(Text key,
Iterable<FloatWritable> values,Context context)
throws IOException, InterruptedException
{
        float top_sum = 0;
        Int l=0;
        for (FloatWritable top_val : values) {
            l+=1;
           top_sum += top_val.get();
}
  Top_sum=top_sum/l;
  context.write(key, new
  FloatWritable(top_sum));
}
}
```

In Reducer code

line 1 a class Top_Reduce which is extended the default Reducer class with arguments KeyIn , ValueIn, same as the outputs of the mapper class and , ValueOut that is used to store final outputs of our MapReduce program.
In **line 2** Reduce method is overridden which is run each time for every key.
In **line 4** An integer variable top_sum is declared which is used to store the sum of all the values for each key.
In **line 5 another variable is** as *"l"* which is incremented every time. This variable is increamented as many values are there for the key.
In **line 6** A Loop which is foreach is taken. It will run each time for the values inside the Iterable values. Which are coming from the second phase that is shuffle and sort phase after the mapper phase.
In **line 8** Value for top_sum that is "sum" is calculated and stored.
In **line 10** An average of the obtained topsum is performed.

How to Execute

hadoop jar video_rating.jar /youtubedata.txt /videorating_out
Explanation for this statement is same as in Problem statement 1.

How to Execute

hadoop jar topratingvideos.jar /youtubedata.txt / topratingvideos out
Here **'hadoop'** is a command and jar specifies that we are running java type of application and **topratingvideos.jar** is

the jar file which is created and consists of the above source code.

How to view output

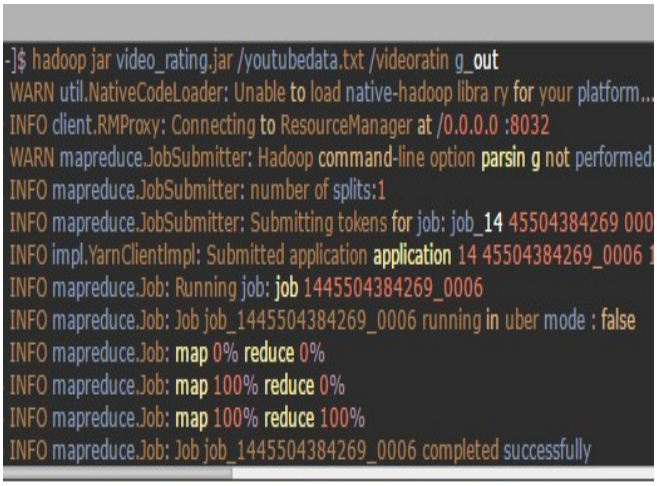
hadoop fs-cat/ topratingvideos_out/part-r-00000 | sort –n –k2 –r | head –n5

Here **'hadoop'** is a command. And **dfs** is related to the File System of Hadoop (HDFS) which is used to perform some operations on Hadoop Distributed file System. The **– cat** command is used to view the contents of a file and topratingvideos **/part-r-00000** is the file where final output is stored.

This Part file is created by Text Input Format which consist of the actual final output. This Part file is created by default. Then, **sort –n –k2 –r | head –n5** is the main command which shows the top 5 categories with maximum number of videos uploaded. Also we can mention the secondary sort instead of this command.

Sort is used to sort the data, **–n** stands for numerically that is sorting will be numerically,
–k2 stands for second column, **–r** stands for recursive operation, **–n5** stands for the first 5 values after sorting.

Output

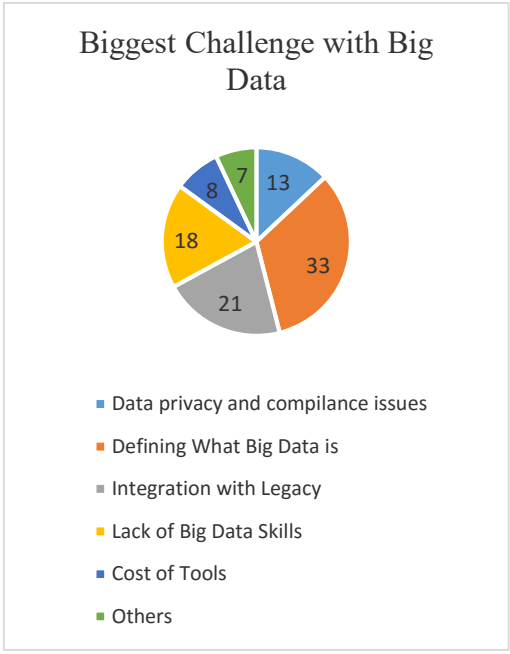


4. BIG DATA CHALLNGES

Understanding and Utilizing Big Data – The main challenge in using the Big Data approach is to understand the data that is available to process and use. It is a very challenging task to understand the data available in Information Technology Industries. This type of analysis is very important for any organization and it is performed on regular basis in order to keep pace with the market and to develop the organization [17].

Privacy, Security, and Regulatory Considerations – We all know the volume and complexity of data is becoming more complex. Keeping this in mind, it is very difficult to obtain a reliable content or data. Also it is very difficult to prevent the data from the threats [10]. Various Security mechanism are available today. But implementing those mechanism may increase the cost of a company.

The Need for IT, Data Analyst, and Management Resources – To analyze the big data efficient human resources with proper skill sets are required. The challenge in Big Data described in first point that is understanding the big data. To understand and process the large volume of data, People with proper skill sets are required. Some of the Challenges are highlighted below in the pie chart.



5. CONCLUSION

A large volume of data availability of Big Data, have produced a very good method for data analysis. The availability of big data and its framework has enabled us to analyze the data at a very large scale. The main advantage of Big Data is its accuracy which leads to more confidence in decision making. If decisions are better they can result in greater performance, greater efficiency, increase productivity and also the cost effective that is can help in the cost reduction also can help in the reduction of risks. Also there are a lot of challenges in implementing the big data

approach. The biggest challenge does not seem to be technology itself as it is growing at a very high speed. But the main challenge is weather we have proper resources, proper skill sets and proper dedicated hardware to implement this. If we have proper resources another challenges are also there like there are many legal like data privacy and integrity, cyber security which we need to resolve. With the help of Big Data we can run applications with a large number of nodes with large number data which may be in terabytes. This allows the very fast transfer of data between nodes. It is also very reliable framework which allows continue operating if even any node fails [4]. It also minimize the risk and chance of system failure even if a good number of nodes get failed.

Now a days, a number of companies have implemented this Big Data approach. They are working on it and with a great pace. Companies like Facebook which have a large volume of data uses this approach.

Finally, social media sites like Facebook, Twitter and LinkedIn would not exist without the big data.

ACKNOWLEDGEMENT

I would like to thank my guide Ms. Kirti Bhatia for her indispensible ideas and continuous support ,encouragement ,advice and understanding me through my difficult times and keeping up my enthusiasm, encouraging me and for showing great interest in my thesis work, this work could not finished without his valuable comments and inspiring guidance.

REFERENCES

[1] H.Herodotou,H.Lim,G.Luo,N.Borisov,L.Dong,F.B.Ceti n,and S. Babu. Starfish: A Selftuning System for Big Data Analytics. In CIDR, pages 261–272, 2011.

[2] Shweta Pandey, Vrinda Tokekar. Prominence of MapReduce in BIG DATA Processing. In Fourth International Conference on Communication Systems and Network Technologies, IEEE, pages 555-560 , 2014.

[3] Hadoop. http://hadoop.apache.org/.

[4] Hadoop MapReduce Tutorial. http://hadoop.apache.org/common/docs/r0.20.2/ mapred_tutorial.html

[5] T. Nykiel, M. Potamias, C. Mishra, G. Kollios, and N. Koudas. MRShare: Sharing Across Multiple Queries in MapReduce. PVLDB,

[6] http://www.drdobbs.com/parallel/indexing-and-searching-on-a-hadoop-distr/226300241

[7] http://www.pgs-soft.com/harnessing-big-data/

[8] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009) .

[9] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011).

[10] Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013) .

[11] EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012) .

[12] He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208 (2011) .

[13] Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272 (2011).

[14] Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28 (2012) .

[15] Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011) .

[16] Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011) .

[17] Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 (2011) .

[18] TechAmerica: Demystifying Big Data: A Practical Guide to Transforming the Business of Government.