Data Analysis Using Map/Reduce

Submitted by

Team 2

Yash Avlani (1001670008)

Harshit Modi (1001662262)

Under Supervision of: Prof. Sharma Chakravarthy and GTA, Mr. Abhishek Santra

Department of Computer Science,

University of Texas at Arlington

Acknowledgements

My sincere thanks to **Prof. Sharma Chakravarthy,** Department of Computer Science & Engineering, University of Texas at Arlington, who gave us great lecture on Cloud Computing and Map/Reduce and gave us opportunity to implement project on Hadoop Cluster. Apart from this we thank Prof. Sharma for make us understand the other DBMS topic in depth.

We are also grateful to the **GTA, Mr. Abhishek Santra** for solving all our queries as soon as possible on Blackboard and in person. For his kind assistance and cooperation during the development of the project, we are delighted to have such a supportive GTA.

Overview:

This project implements a basic data analysis using Hadoop Cluster. The given .txt file with huge data will make us follow the method to produce a productive analysis in the real word. Here we have to analyse the IMDB dataset and generate <key, value> pair indicating the the data of year and genre. This is how the mapper function arrange the data and it will be reduced by adding the count of particular genre in specific year. So, this is how it merges together these values to produce possibly smaller set of values.

Overall Status & File Description:

This project is successfully completed. A mapper and reducer are implemented to derive the expected output.

The major components of our implementations and their details are as follows:

- Installation and setup of Hadoop 2.9.1
- Successful implementation of Mapper and Reducer
- Successful deployment and tested by taking small files.

Configurations:

- Hadoop 2.9.1
- OS : Ubuntu 16.04
- JAVA 10.0.2

Setup:

- After going through the project details and pdf given I got the basic idea about where to begin with and how to setup installation of hadoop single node cluster. We used the link given in project material to follow the installation process.
- After successfully completing the installation we made sure that we were able to solve basic WordCount problem using the java file provided.
- Starting and Stopping Services.
    - start-dfs.sh start-yarn.sh
    - stop-dfs.sh stop-yarn.sh
    - Namenode format : hadoop namenode -format

HDFS File System:

- HDFS file system is what Hadoop uses for the mechanism it provides for map/reduce.
- Input can be loaded to HDFS file system by following command.
    - Hadoop -dfs copyFromLocal IMDBTitles.txt

Generate jar and Run Map/Reduce:

- Jar file will allow the directory to access file
    - *bin/hadoop com.sun.tools.javac.Main QueryIMDB.java*
    - *jar cf project3.jar QueryIMDB*.class*
    - *hadoop jar project3.jar QueryIMDB inputhd outputhd*

Get output from HDFS to local:

- The output would be added to the HDFS directory
    - *dfs -copyToLocal /user/hadoop/output titlesOutput*

Where we encountered difficulty

The major difficulty for us was at installation. Even after completing the **installation** we were not able to connect to the **namenode** after restarting the system. It took us more than 2 hours to connect to the namenode. Again the same query came up with **datanode** but we solved it easily. The simple word count problem took us many attempts to be done. As while reaching to the **hadoop directory** and local host we faced many issues. Apart from this .jar file made some trouble for us. Analysing the output was a big task indeed.

Division of Labor

We have completed the project in a group of 2 members. As the skeleton code provided was with hadoop single node cluster and there was depth documentation available, though it took us time to get familiar with some components of the projects. We also spent significant time in making documentation on some methods and flows of project for our future reference, which helped us in further implementation. After completing the big task of install and setup hadoop, it didn't take much time for us to run the WordCount problem and the basic project implementation as well. The analysis task was really challenging as we had never dealt with analysing such data in past. Overall it took **5-6 sessions** to completing whole project.

Total time spent: 36 hrs approx

Analysis Result

1. Taking results from the bar graph from year 2015 - 2019 will make us enable to conclude that : DRAMA and COMEDY are the leading genre for each year. It can be said that due to hard and stressful life people seems more directed towards COMEDY as a result movie makers have produced movie on this genres. Here DRAMA can be "COMEDY-DRAMA" as well. So, by this graphs the current trends can be shown to the producers. On the other hand, genres like MUSICAL, WESTERN, ADULT, NEWS caught least number of producers eyes. As, such genres have very limited and specific audience. Moreover, SHORT genre which was consistently on third position from year 2015-2018 suddenly slept to position 6, which shows producers decreasing interest in creating SHORT.

2.

Genre	AVERAGE of Movies Count
Western	59.46
War	139.93
Musical	295.00
Biography	661.44
History	684.07
Sci-Fi	698.31
Thriller	853.29
Fantasy	908.01
Mystery	1,030.67
Horror	1,045.06
Sport	1,196.75
Adventure	1,337.40
Game-Show	1,879.75
Crime	1,937.08
Adult	2,203.40
Animation	2,260.24
Music	2,284.98
Family	2,720.82
Reality-TV	3,721.69
Romance	3,772.85
News	4,524.42
Action	4,682.82
Documentary	5,218.60
Talk-Show	6,587.38
Short	6,822.65
Drama	10,047.24
Comedy	10,237.99

This table show the average number of movies count in a year from year 2000 to 2150. If we analyse the overall scenario of movie creation over the 150 years history, still COMEDY tops the chart. And WESTERN is having least number of productions. CRIME, ADULT, ANIMATION were at the middle on an average. It can be said that investing on COMEDY being blindfold would not really make lose.

3. By exploring the table of number of movies released from year 2000 to 2018 it can be said that, Starting from 2000 with the 103172 by the time in year 2016-17 the number of movies releases were almost for times than the former year. But, sudden downfall can be seen in 2018, as there can be interest of people decreased in movies, or some major sports tournament held which was diverting people towards sports or else the producers can be on strike that year due to some clash or issue with theater.

Logical errors

1. Connecting to **NAMENODE** after restarting the system.

   This problem was very critical as one of us engaged setting up the hadoop cluster for very long time and after successfully finishing up the installation process we needed to do a lot work to connect to namenode. We didn't know that the established connection breaks after restarting the system. Somehow we got the solution that formatting namenode can solve out issues.

2. **Permission Denied** when accessing java file.

   While applying the java file for the given input of word count we got an error of permission denied about accessing the class TokenizerMapper. But after some R&D we came up with the solution that the file should be located inside the hadoop directory to be accessed. As the created **hadoopusr** will have the permission to use all the data inside hadoop directory and it won't be accessible by any other user.

3. **Analysis** phase of the project.

   This was the most different task for both of us as none of us had tried our hands on data analysis in past. Moreover, what to be put and what can be useful results as a real time example was the challenging task for us. The GTA Abhishek's suggestion guided us to the right direction in this phase of project. The describing the derived graphs and charts and opt out the best possible results from it took really long for us.