MAP/REDUCE
IMDB DATA ANALYSIS

Abstract

In this project, we implemented a Map/Reduce program to find the number of movies with genre combinations of Comedy, Romance; Action, Thriller; and Adventure, Sci-Fi for the time periods [2000-2006], [2007-2012], and [2014-2020] using IMDB dataset. In the second part of the project, we write SQL queries to find the top 5 and bottom 5 movies of the one of the time periods and all the genre combinations. We also find the query plan using the 'EXPLAIN PLAN' command.

Contents

Overall Status

TASK 1:

- With the task pdf provided, we got logic about where to start and how to proceed with the installation of hadoop single node cluster.
- After completing the installation, we then went for executing basic WordCount program.
- The hadoop daemons were started by typing below command, and this started three nodes viz. namenode, datanode and secondary namenode.
  start-dfs.sh start-yarn.sh
- Hadoop uses HDFS file system. Hence, we first had to decode the file system of Hadoop.
- This is how we loaded input,
  - **Hadoop -dfs copyFromLocal imdb.txt**
- To execute the project, we executed the following operation.
  - **bin/hadoop com.sun.tools.javac.main imdb.java**
  - **jar cf imdb.jar imdb*.class**
  - **hadoop jar project/imdb.jar project.imdb /imdb/input /imdb/output**
- First we had to figure out whether or not with one mapper or output will be enough for the project to be implemented. And we found out that partitioning the data rightly will solve our problem. Hence we created three class – TokenizerMapper & one reduce was enough to simplify the process.
- With the above mentioned executions, we implemented the mentioned tasks

Analysis Results

TASK 1:

We have attached a separate "Analysis" report in the current directory. We thought keeping here would make this report lengthy. And hence went on to attach separately.

Thus, we have compiled a detailed report of all the three disjoint periods as well as complete data histogram chart from 2000 – 2020 year.

File Description

No new files were created in this project. All files used were provided in the assignment attachment.

Logical Errors

TASK 1:

- We had a lot of difficulties while installing ubuntu. And we had to install all the dependencies in order for Hadoop to work efficiently.
- ArrayBound errors were encountered as we didn't knew the existence of "\N" in the string on every line.
- Counting the genre's as given led us to wrong version of output. And, later we understood that we were to find those genres in the subarray and then we had to match that line with counting.
- Also, we had some wrong counting in the mapper done because we were partitioning data with nested ifs and this led to some bad logic which gave a faulty of version of count to the reducer.
- Giving wrong argument for the output led us to empty data problem even when the Hadoop ran successfully.