Naseem Dabiran

CIS 9760 – Big Data Technologies

Analyzing IMDB Datasets

Background information about the project:

The goal of this project was to utilize AWS ElasticMapReduce (EMR) to create an Apache Spark cluster to perform queries using IMDB data provided. Apache Spark is a computing engine and set of libraries used to perform big data analytics tasks much faster than EMR.

After logging into AWS, we searched for EMR to create a cluster and went to the Advanced Options in Create Cluster. The software configuration should be set to emr-5.31.0. Spark 2.4.6 and Livy 0.7.0 must be selected as well. EC2 m5.xlarge instance types should be automatically created but do verify the instances are made correctly. Figure 1 following image is an example of a successfully created cluster.

A Jupyter Notebook was configured and connected to the created cluster. Figure 2 is an example of a successfully created Jupyter Notebook. The preset Python kernel of the Jupyter Notebook was then changed to PySpark kernel. The syntax and data structures of Python and PySpark are similar, such as DataFrames which are used heavily in this project, but PySpark allows us to interact with the Spark core.

First, Python libraries pandas and matplotlib were installed and imported into the notebook for future use. 4 datasets in a publicly available Amazon S3 bucket that contained data from IMDB (an online database for movies and television shows) were then loaded into the Jupyter notebook as Spark DataFrames. Exploratory data analysis was performed on the datasets to get an overview of the data in each DataFrame. PySpark functions was used to answer the questions provided in the original Project2\_Analysis.ipnyb, such as querying the data to find the movies featuring both Johnny Depp and Helena Bonham Carter.

Graphical user interface, text, application, email

Description automatically generated

Fig 1. CLUSTER CONFIGURATION

Graphical user interface, text, application, email

Description automatically generated

Fig 2. NOTEBOOK CONFIGURATION