Project 2 - Analyzing IMDB Datasets

I worked with provision a Spark cluster on AWS EMR, connect it to a Jupyter Notebook and then run a series of queries (in python with DataFrame API or Spark SQL) that answer a few simple questions about the IMDB Data available.

**Cluster Configuration:**

Graphical user interface, text, application, email

Description automatically generated

**Notebook Configuration:**

Graphical user interface, text, application, email

Description automatically generated

**S3 Bucket**

actors = spark.read.csv('s3://cis9760-lecture9-movieanalysis/name.basics.tsv', sep=r'\t', header=True)

genres = spark.read.csv('s3://cis9760-lecture9-movieanalysis/title.basics.tsv', sep=r'\t', header=True)

movie\_actors = spark.read.csv('s3://cis9760-lecture9-movieanalysis/title.principals.tsv', sep=r'\t', header=True)

movie\_ratings = spark.read.csv('s3://cis9760-lecture9-movieanalysis/title.ratings.tsv', sep=r'\t', header=True)

# **Artifacts**

Text, letter

Description automatically generated