**READ ME FILE**

**IMDB Dataset Data Analysis**

In this project, we will be analyzing the IMDB datasets from Kaggle. We will use Amazon Warehouse Services such as S3, EMR to hold the data and Jupyter Notebook to run pyspark functions to carry out the analysis.

The steps include:

* We create a bucket on s3 to house the IMDB datasets from Kaggle
* Create an EMR cluster on AWS to facilitate Jupyter notebook
* We create a new notenbook to run the analysis
* Next, change the kernel to Pyspark
* Now, install the required packages
* Finally, run multiple analysis on the datasets

**Cluster Configuration**

**Graphical user interface, text, application, email

Description automatically generated**

**Notebook Configuration**

Graphical user interface, text, application

Description automatically generated