Project - Analyzing IMDB Datasets

**Overview:**

This project involves provisioning a Spark Cluster on AWS Elastic MapReduce (EMR) to analyze IMDB datasets obtained from Kaggle ([LINK](https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset)). The analysis is performed using Jupyter Notebook, and the output is captured in the Project2\_Analysis.ipynb file. The primary goal is to showcase the ability to configure and provision infrastructure using AWS EMR, connecting it to a Jupyter Notebook, and conducting data analysis tasks with PySpark.

**Datasets:**

The relevant datasets are as follows:

* **Actors Dataset** (s3://cis9760-lecture9-movieanalysis/name.basics.tsv):
  + Contains information about actors.
* **Genres Dataset** (s3://cis9760-lecture9-movieanalysis/title.basics.tsv):
  + Provides details about movie genres.
* **Movie Actors Dataset** (s3://cis9760-lecture9-movieanalysis/title.principals.tsv):
  + Includes information about actors associated with movies.
* **Movie Ratings Dataset** (s3://cis9760-lecture9-movieanalysis/title.ratings.tsv):
  + Contains ratings and other details for movies.

**Technology Leveraged:**

* **AWS EMR (Elastic MapReduce):** Utilized to provision and manage the Spark cluster.
* **S3 Bucket:** IMDB data is read directly from a publicly available S3 bucket (cis9760-lecture9-movieanalysis)
* **IAM:** To assign permissions to the Elastic MapReduce user created for this task.
* **PySpark:** Employed for performing data analysis tasks using the DataFrame API & Spark SQL.
* **Jupyter Notebook:** Used as the interface for running analysis queries on the Spark cluster.

**Project Steps:**

1. **AWS EMR Cluster Provisioning and Notebook Steps:** The cluster is configured to read data from the specified S3 bucket. Screenshots of the EMR cluster configuration are available below, demonstrating the setup. Screenshots of the Notebook configuration within the AWS EMR environment are also provided. The setup includes establishing a connection between Jupyter Notebook and the Spark cluster, ensuring seamless data access from the S3 bucket.

**Cluster Configuration Screenshots-**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

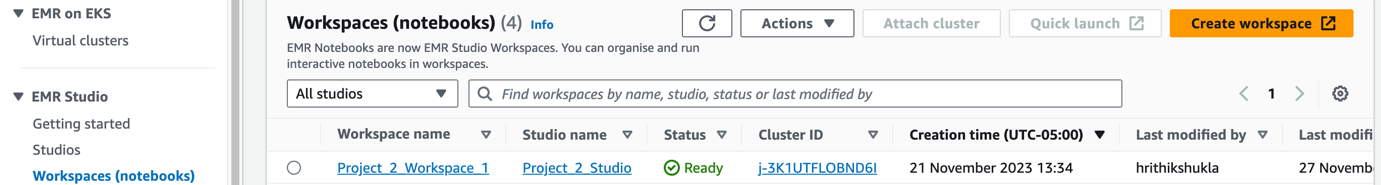
Description automatically generated

**Studio-**

A screenshot of a computer

Description automatically generated

**Notebook Configuration-**



A screenshot of a computer

Description automatically generated

1. **Data Analysis:** The 'Project2\_Analysis.ipynb' file contains a structured analysis broken into five parts:

Part I: Installation and Initial Setup

* Dependencies such as pandas and matplotlib are imported.
* The IMDB dataset is loaded as PySpark DataFrames.

Part II: Analyzing Genres

* Denormalization of genres associated with each title is performed.
* A basic analysis is conducted on the result.

Part III: Analyzing Job Categories

* Top job categories in the dataset are identified.

Part IV: Answering Questions

This part involves answering specific questions:

* Find all the "movies" featuring "Johnny Depp" and "Helena Bonham Carter"
* Find all the "movies" featuring "Brad Pitt" after 2010
* What is the number of "movies" "acted" by "Zendaya" per year
* What are the "movies" with an average rating greater than "9.7" and released in "2019"

Part V: Additional Analysis

This part involves answering certain additional questions:

* What are the top 10 highest-rated “movies” (based on Average Ratings and Number of Votes) that were released after 2015?
* What are the top 5 genres for “short” films by the average ratings?
* Which genres have the highest count of top-rated movies, defined as those within the top 25th percentile of ratings?
* Among the actors and actresses in the dataset, which 20 film professionals have the highest number of movies, and what is the count of films for each of them? Visualize the results.
* What patterns emerge from the evolution of movie ratings over decades?

The notebook provides code snippets and outputs for each analysis, offering insights into the IMDB dataset based on the posed questions.