Project 2 - Analyzing IMDB Datasets

For this project, you will be tasked with provisioning a Spark Cluster on AWS EMR for loading and running some analysis on IMDB’s datasets from [**Kaggle**](https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset). You will run your analysis via Jupyter Notebook and the expected output artifact is **Project2\_Analysis.ipynb** file.

[Requirements 2](#_Toc131699857)

[Artifacts 2](#_Toc131699858)

[Notebook File 2](#_Toc131699859)

[README 3](#_Toc131699860)

[S3 Bucket 4](#_Toc131699861)

[Submission 4](#_Toc131699862)

[Assignment 4](#_Toc131699863)

[Part I: Installation and Initial Setup 4](#_Toc131699864)

[Part II: Analyzing Genres 5](#_Toc131699865)

[Part III: Analyzing Job Categories 5](#_Toc131699866)

[Part IV: Answering Questions 5](#_Toc131699867)

[Extra credit: Analysis of Your Choice (1.5 pts) 5](#_Toc131699868)

[Extra credit: Answering a Question of Your Choice (1.5 pts) 5](#_Toc131699869)

# 

# 

# 

# 

# Requirements

This project is very simple: you are to 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.

In doing so, you are demonstrating your ability to configure and provision infrastructure using the AWS Elastic Map Reduce ecosystem. Also, you are demonstrating your understanding of how to leverage transformations and actions (as per the Spark terminology) with **PySpark** in performing basic data analysis tasks on information sources that are too large to manage in memory.

# Artifacts

You are to submit a zip file with your project work content (as seen below) inside. Expected Zip file structure:

**project02**

**+-- Project2\_Analysis.ipynb**

**+-- Project2\_Analysis.pdf**

**+-- assets**

**+-- +-- cluster\_configuration.png**

**+-- +-- notebook\_configuration.png**

**+-- README.md**

Note: I’m ok with the images submitted as **jpegs** and README submitted as **pdf** as well.

Note2: If you cannot download Project2\_Analysis.ipynb as pdf, you can press Ctrl + P on your keyboard and save it as a PDF.

## Notebook File

The **ipynb** file that contains your analysis and **the outputs of the code you wrote** to arrive at your results. This is very important as this is the sole method of validation that you actually ran an EMR cluster successfully.

You **must** name your Notebook file **Project2\_Analysis.ipynb**. **PLEASE PLACE THIS IN YOUR PROJECT ROOT FOLDER**.

## README

The **README**, in markdown or pdf, should contain a brief blurb describing the project and the technology leveraged to conduct your analysis. This ought to be brief and informational, in case folks in the future want to recreate your results.

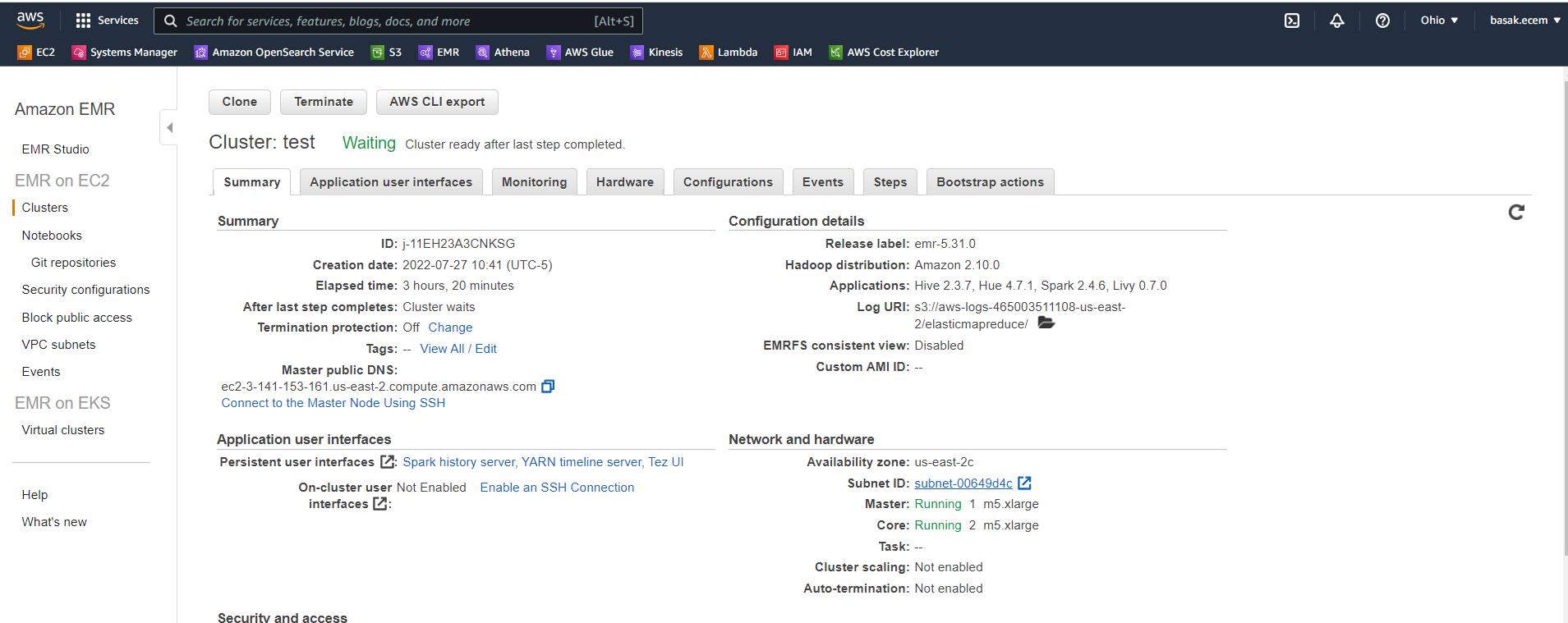
**ALSO**, your README must contain screenshots of your EMR cluster configuration and Notebook configuration. Here are mine, shared below as reference

**Example markdown code:**

**![cluster\_iamge](assets/cluster\_configuration.png)**

PS: if you wanted to “test” your readme, you can download a readme viewer like this [one](https://typora.io/)

**Cluster Configuration**



**Notebook Configuration**

Graphical user interface, text, application, email

Description automatically generated

## S3 Bucket

You **must** read IMDB data from my publicly available **S3** bucket. Your **Project2\_Analysis.ipynb** file must demonstrate that the data is being read from **S3** - this is largely as simple loading your DataFrame like so:

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

Note: The path to each dataset is given in Project2\_Analysis.ipynb file.

## Submission

Please submit your zip folder to Blackboard before the deadline.

# Assignment

The actual analysis is broken into four parts - three which are guided and one that is freeform. I have published an **Project2\_Analysis.ipynb** demonstrating this project. Note that the output of the code written is provided as a means to give you **structure** as you write your analysis.

## Part I: Installation and Initial Setup

In this portion, you will import the necessary dependencies (**pandas** and **matplotlib**) and load your dataset as a pyspark dataframe.

## Part II: Analyzing Genres

For this part, you will take a stab at denormalizing the genres that are associated with each title (there may be more than one, presented as a string of comma-separated identifiers) and then run some basic analysis on the result.

## Part III: Analyzing Job Categories

For this next part, you will attempt to get the top job categories in the dataset.

## Part IV: Answering Questions

For this final part, you will answer the following questions:

* Find all the movies acted by both Johnny Depp and Helena Bonham Carter
* Find all the movies acted by 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?

## Extra credit: Analysis of Your Choice (1.5 pts)

Try and analyze some interesting dimension to this data. You should specify the name of the analysis in your **Project2\_Analysis.ipynb**. This part similar to Part 2 and Part 3.

* You must create at least one visual.

## Extra credit: Answering a Question of Your Choice (1.5 pts)

Come up with a question and get an answer like in Part 4. The question of your choice should not be similar to the ones given in Part 4. Try to find some interesting questions using different fields and arguments. You should specify the question in **Project2\_Analysis.ipynb**.

* You must join at least two datasets.

This project is due **May 5th, MIDNIGHT.**

If you need, you can request an extension; however, you will lose the opportunity of extra credits.

|  |  |
| --- | --- |
| **RUBRIC** | |
| **Overall** | **5** |
| Parent folder is named “**project02**” and is exposed when unzipped | 1 |
| **README** exists under the **project02** folder and describes the project | 1 |
| **README** contains screenshots of your EMR cluster configuration AND Notebook configuration | 1 |
| The “**assets**” folder exists and contains **cluster\_configuration.png** and **notebook\_configuration.png** | 1 |
| **Project2\_Analysis.ipynb** and **Project2\_Analysis.pdf** exist under project02 folder | 1 |
| **Part I: Installation and Initial Setup** | **7** |
| Any necessary packages are loaded into the environment (pandas, matplotlib, etc) | 1 |
| All four datasets are loaded from the S3 bucket and saved as Spark DF | 2 |
| Overview: The number of rows and columns in each Spark DF is displayed | 2 |
| The remaining cells are filled properly. | 2 |
| **Part 2 - Analyzing Genres** | **11** |
| Association table | 3 |
| Total unique genres | 2 |
| Average rating / genre | 3 |
| A horizontal bar chart of top genres | 3 |
| **Part 3 – Analyzing Job Categories** | **5** |
| Total unique job categories | 2 |
| Top job categories | 1 |
| A bar chart of top job categories | 2 |
| **Part 4 – Answer to the following questions:** | **12** |
| Find all the movies acted by both Johnny Depp and Helena Bonham Carter | 3 |
| Find all the movies acted by Brad Pitt after 2010 | 3 |
| What is the number of movies acted by Zendaya per year? | 3 |
| What are the movies by an average rating greater than 9.7 and released in 2019? | 3 |