ANALYSIS OF IMDB DATASET

In this project, I provisioned a Spark Cluster on AWS EMR, connect it to a Jupyter Notebook. I then ran a series of queries and performed data analysis on IMDB's datasets from Kaggle with PySpark to answer the following nine questions:

1、What are all the "movies" featuring "Johnny Depp" and "Helena Bonham Carter"?
2、What are all the "movies" featuring "Brad Pitt" after 2010?
3、What is the number of "movies" "acted" by "Zendaya" per year?
4、What are the "movies" by average rating greater than "9.7" and released in "2019"?
5、What are the Top 5 "movies" by numvotes greater than 1.5 million and average rating greater than 8.5?
6、How many movies have been released per year since 2000?
7、What is the average number of votes per year since 2000?
8、What is the average rating of votes per year since 2000?
9、What are the top 20 good movies in 2019?

The four datasets used in this project come from Kaggle and they have been uploaded to an S3 bucket:

- s3://cis9760-lecture9-movieanalysis/name.basics.tsv ---> (actors)
- s3://cis9760-lecture9-movieanalysis/title.basics.tsv ---> (basics)
- s3://cis9760-lecture9-movieanalysis/title.principals.tsv ---> (principals)
- s3://cis9760-lecture9-movieanalysis/title.ratings.tsv ---> (ratings)

File Structure

project02

+-- Project2_Analysis.ipynb

+-- Project2_Analysis.pdf

+-- assets

+-- +-- cluster_configuration.png

+-- +-- notebook_configuration.png

+-- README

Table Structure

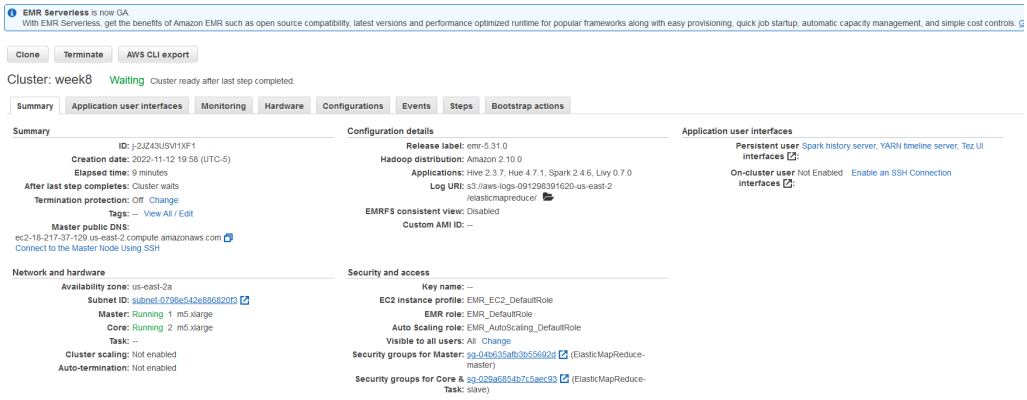
actors
nconst
primaryName
birthYear
deathYear
PrimaryProfession
known for titles

basics
tconst
titleType
PrimaryTitle
OriginalTitle
isAdult(boolean)
startYear
endYear
runtimeMinutes
genres

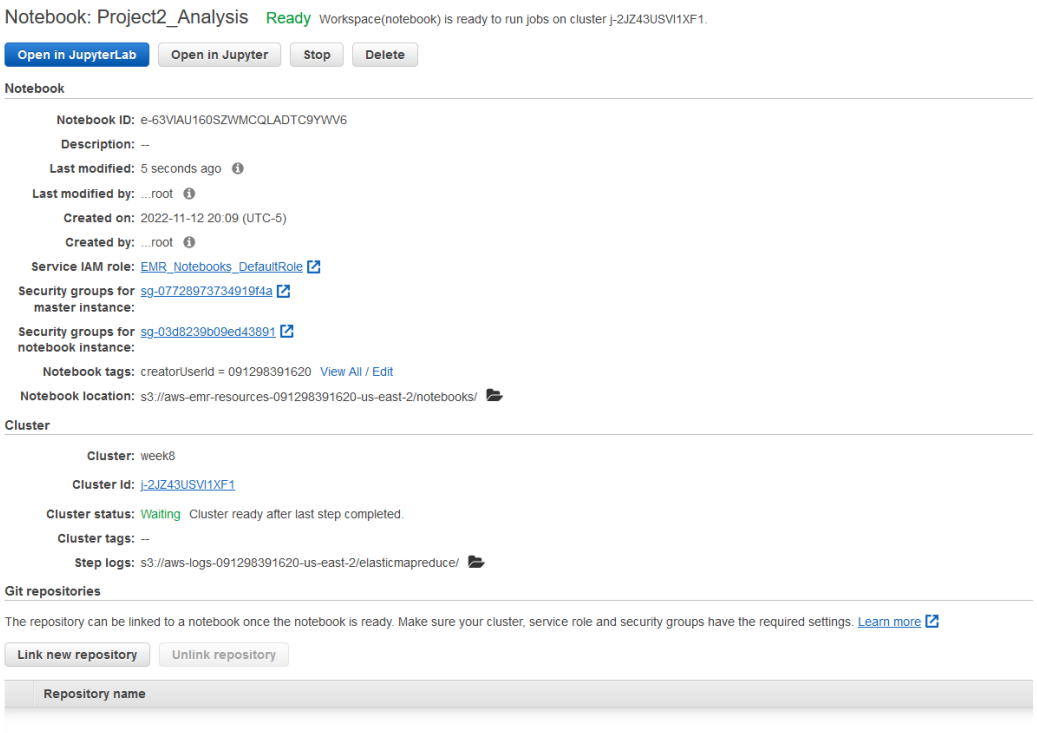
principals
tconst
ordering
nconst
category
job
characters

ratings
tconst
averageRating
numVotes

Cluster Configuration



Notebook Configuration



Technology Used

Read dataset from publicly available S3 bucket

Create a Cluster on AWS EMR

Connect the cluster with Jupyter Notebook

Perform data analysis by using PySpark

Analysis

PART 1 - Installation and Initial Setup

Imported the necessary dependencies (pandas and matplotlib) and loaded dataset as a pyspark dataframe.

PART 2 - Analyzing Genres

Created association table and performed some basic analysis about top genres by movies

PART 3 - Analyzing Job Categories

Analyzed top job categories in the dataset

PART 4 - Answer to the below nine questions

1、What are all the "movies" featuring "Johnny Depp" and "Helena Bonham Carter"?

```
+---------------------------------------------+
|PrimaryTitle                                 |
+---------------------------------------------+
|Dark Shadows                                 |
|Sweeney Todd: The Demon Barber of Fleet Street|
|Alice Through the Looking Glass              |
|Alice in Wonderland                          |
|Charlie and the Chocolate Factory            |
|Corpse Bride                                 |
+---------------------------------------------+
```

2、What are all the "movies" featuring "Brad Pitt" after 2010?

primaryTitle	startYear
Babylon	2021
Kajillionaire	2020
Irresistible	2020
Ad Astra	2019
Once Upon a Time ... in Hollywood	2019
The King	2019
Vice	2018
War Machine	2017
Voyage of Time: Life's Journey	2016
Allied	2016
By the Sea	2015
Hitting the Apex	2015
The Big Short	2015
Fury	2014
Kick-Ass 2	2013
World War Z	2013
12 Years a Slave	2013
Killing Them Softly	2012
The Tree of Life	2011
Moneyball	2011

3、What is the number of "movies" "acted" by "Zendaya" per year?

startYear	count
2020	1
2018	2
2017	1

4、What are the "movies" by average rating greater than "9.7" and released in "2019"?

PrimaryTitle	averageRating
Our Scripted Life	10.0
The Twilight Zone: A 60th Anniversary Celebration	10.0
Bu Can Var Oldugu Sürece	10.0
L'Enfant Terrible	10.0
Kirket	10.0
A Grunt's Life	10.0
A Medicine for the Mind	10.0
Love in Kilnerry	10.0
The Butcher Baronet	10.0
Square One	9.8
Time and motion	9.8
Kamen Rider Zi-O: Over Quartzer	9.8
Randhawa	9.8
From Shock to Awe	9.8
Gini Helida Kathe	9.8
We Shall Not Die Now	9.8
Puritan: All of Life to The Glory of God	9.9
Superhombre	9.9
The Cardinal	9.9

5、What are the Top 5 "movies" by numvotes greater than 1.5 million and average rating greater than 8.5?

PrimaryTitle	numvotes	averageRating
The Shawshank Redemption	2159745	9.3
The Dark Knight	2134602	9.0
Inception	1892958	8.8
Fight Club	1725444	8.8
Pulp Fiction	1695159	8.9

6、How many movies have been released per year since 2000?

startYear	Total
2000	2880
2001	3079
2002	3276
2003	3490
2004	3912
2005	4385
2006	4862
2007	5080
2008	5727
2009	6504
2010	6769
2011	7336
2012	7660
2013	7977
2014	8382
2015	8534
2016	8777
2017	8893
2018	8182
2019	5371

7、What is the average number of votes per year since 2000?

startYear	avg_num_of_votes
2000	6299.577083333334
2001	7027.099058135758
2002	6685.843101343101
2003	6429.0555873925505
2004	6878.109151329243
2005	5559.932269099202
2006	5706.264294529001
2007	5800.278937007874
2008	5443.445433909551
2009	4614.460793357934
2010	4596.446890234894
2011	4548.010496183206
2012	4353.268276762402
2013	4566.240315908236
2014	4210.141016463851
2015	3163.5385516756505
2016	3158.0945653412327
2017	2631.9877431687846
2018	2327.9615008555365
2019	1646.7909141686837

8、What is the average rating of votes per year since 2000?

startYear	avg_rating
2000	6.034722218869461
2001	6.093894123566464
2002	6.135195361534463
2003	6.133094557240221
2004	6.215388550074554
2005	6.22159635264898
2006	6.176593996060331
2007	6.220413389238786
2008	6.2001571516435305
2009	6.209578724951468
2010	6.226384990217767
2011	6.264462923616853
2012	6.274138384746198
2013	6.255120974070031
2014	6.280505848425836
2015	6.234614484645959
2016	6.300284838260843
2017	6.341515799146743
2018	6.251723297313675
2019	6.4932787226822715

9、What are the top 20 good movies in 2019?
(I define good by the movie's average rating is greater than 2019 overall average rating and the movie's number of votes is greater than 2019 overall average number of votes)

PrimaryTitle	averageRating	numvotes
Love in Kilnerry	10.0	2360
Zana	9.4	3932
Mosul	9.1	2643
Little Baby	9.0	3987
Kaithi	8.9	3076
Saand Ki Aankh	8.9	1960
Joker	8.8	466912
Asuran	8.8	3918
The Blue Elephant 2	8.8	3819
The Irishman	8.7	8992
Jersey	8.7	3991
Parasite	8.6	73962
Kumbalangi Nights	8.6	4138
Agent Sai Sriniva...	8.6	3084
Tell No One	8.6	2325
Avengers: Endgame	8.5	602740
Klaus	8.5	4490
Uri: The Surgical...	8.4	35278
Super Deluxe	8.4	4535
Nerkonda Paarvai	8.4	4405

only showing top 20 rows