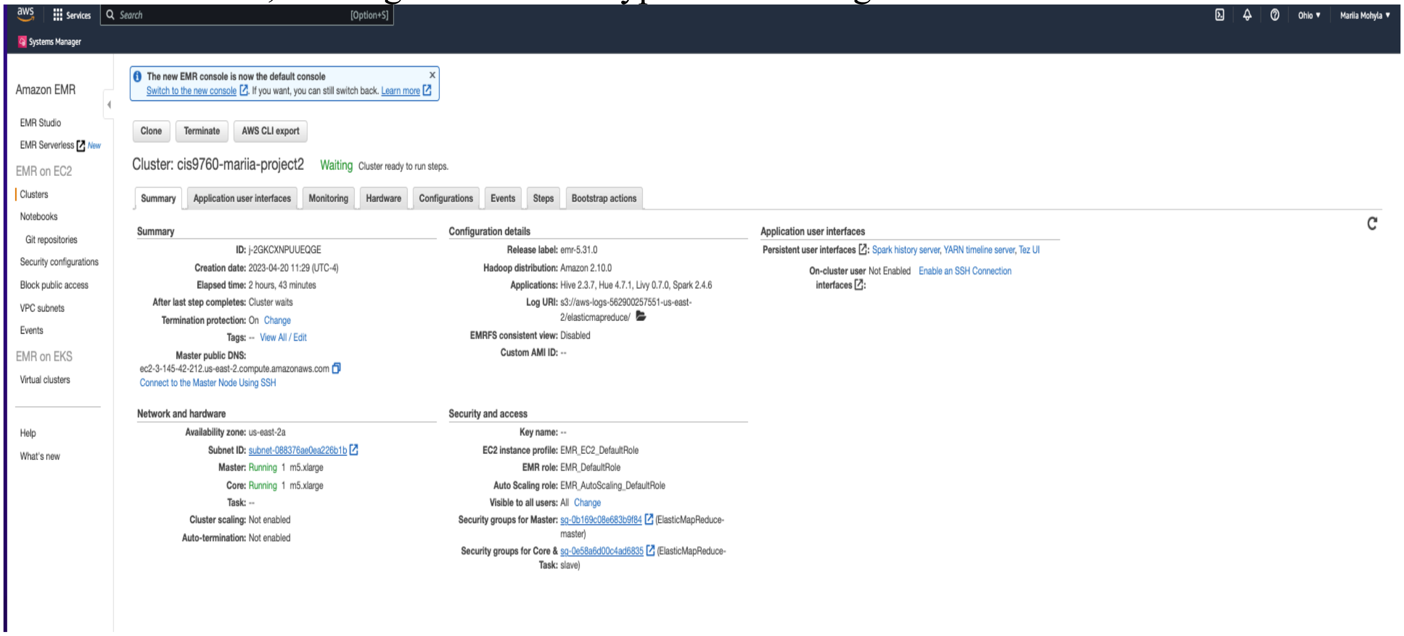
The goal of this project is to provision a Spark infrastructure on AWS EMR ecosystem to load and run some exploratory data analysis using PySpark on IMDB's dataset from Kaggle located in publicly available S3 storage.

STEP 1:

The project starts with setting up an EMR ecosystem to provision a Spark cluster of machines to run jobs of big data. For the project purposes, I chose 1 master node and 1 core node. Also, I configured instance types as m5.xlarge.



The new EMR console is now the default console
Switch to the new console. If you want, you can still switch back. Learn more

Cluster: cis9760-mariia-project2 Waiting Cluster ready to run steps.

Summary

ID: j-2GKCXNPUUEQGE
Creation date: 2023-04-20 11:29 (UTC-4)
Elapsed time: 2 hours, 43 minutes
After last step completes: Cluster waits
Termination protection: On Change
Tags: -- View All / Edit
Master public DNS: ec2-3-145-42-212.us-east-2.compute.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.31.0
Hadoop distribution: Amazon 2.10.0
Applications: Hive 2.3.7, Hue 4.7.1, Livy 0.7.0, Spark 2.4.6
Log URI: s3://aws-logs-562900257551-us-east-2/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: Spark history server, YARN timeline server, Tez UI
On-cluster user interfaces: Not Enabled Enable an SSH Connection

Network and hardware

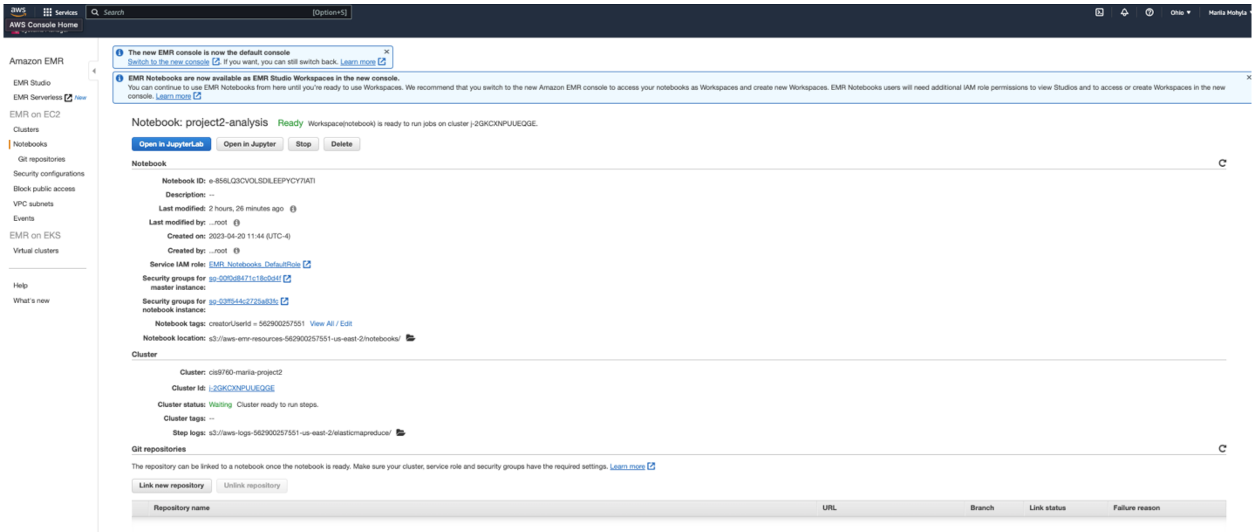
Availability zone: us-east-2a
Subnet ID: subnet-088376ae0ea226b1b
Master: Running 1 m5.xlarge
Core: Running 1 m5.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

Key name: --
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0b169c08e683b9f84 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0e58a6d00c4ad6835 (ElasticMapReduce-slave)

STEP 2:

Then, I configured Jupyter Notebook to run on that cluster and enable the data analysis job. You may notice the configuration of the Jupyter Notebook below.



The new EMR console is now the default console
Switch to the new console. If you want, you can still switch back. Learn more

EMR Notebooks are now available as EMR Studio Workspaces in the new console.
You can continue to use EMR Notebooks from here until you're ready to use Workspaces. We recommend that you switch to the new Amazon EMR console to access your notebooks as Workspaces and create new Workspaces. EMR Notebooks users will need additional IAM role permissions to view Studios and to access or create Workspaces in the new console. Learn more

Notebook: project2-analysis Ready Workspace(notebook) is ready to run jobs on cluster j-2GKCXNPUUEQGE.

Notebook

Notebook ID: e-856LQ3CVOLSDILEEPYCY7IATI
Description: --
Last modified: 2 hours, 26 minutes ago
Last modified by: ...root
Created on: 2023-04-20 11:44 (UTC-4)
Created by: ...root
Service IAM role: EMR_Notebooks_DefaultRole
Security groups for master instance: sg-00f0d8471c18c0d4f
Security groups for notebook instance: sg-03ff544c2725a83fc
Notebook tags: creatorUserId = 562900257551 View All / Edit
Notebook location: s3://aws-emr-resources-562900257551-us-east-2/notebooks/

Cluster

Cluster: cis9760-mariia-project2
Cluster Id: j-2GKCXNPUUEQGE
Cluster status: Waiting Cluster ready to run steps.
Cluster tags: --
Step logs: s3://aws-logs-562900257551-us-east-2/elasticmapreduce/

Git repositories

The repository can be linked to a notebook once the notebook is ready. Make sure your cluster, service role and security groups have the required settings. Learn more

Repository name	URL	Branch	Link status	Failure reason

STEP 3:
I performed exploratory data analysis in my Jupyter Notebook by using PySpark library which allowed leveraging Dataframe API in Python and SQL queries to explore interesting facts about the IMDB dataset. The analysis consisted of four parts.

Part I: I started with installing some required libraries such as pandas and matplotlib and a general overview of the tables.
Part II: In the second part of the project, I discerned distinct movie genres and calculated the average rating per genre. Finally, I have plotted my findings. Short movies have the highest average rating while horror movies have the lowest.
Part III: Part 3 is related to the jobs in the movie industry. To be more specific, I built a bar chart with the top 5 job categories. Actor, actress, self, writer, and director are the top 5 jobs in the movie industry.
Part IV: Part 4 includes some additional questions that conclude my analysis.