# **Assignment – 11**

# **S3 and Athena**

**Due Date:** Feb 19th, 2023

Description:

We have two datasets: netflix\_list.csv and imdb\_title\_ratings.tsv. Upload these files into your S3 bucket (encrypted bucket) and create athena tables on top of these S3 files. Query the data on those tables and verify if it is working as expected and then run the below data analysis queries.

Data analysis queries:

* Number of movies in the dataset
* Number of tv shows in the dataset
* Top 5 runtimes in the dataset
* Top 5 movies w.r.t imdb ratings
* Top 5 tv shows runtimes (multiple episodes with runtime)
* Top 5 movies w.r.t imdb ratings whose numOfVotes is greater than 1000
* Number of titles in each origin\_country
* Average imdb ratings of all titles that released in each startYear whose numOfVotes must be greater than 100
* Using window functions, extract any data insight
* Using window functions, extract any data insight

Outcomes:

* AWS CLI commands which you used to copy data into S3
* Athena tables DDL’s - .sql file
* Data analysis SQL’s - .sql file

Submissions:

* Create a folder with your name under “Staffing Training 2023/Submissions”
* Share above folder with me so I can access your submissions
* Create another folder with the assignment name “Assignment 11 - S3 and Athena”
* Copy all the above-created files into that folder
* Update the Assignment Submission Tracker sheet in the Assignments folder