Movie Recommendation System using Spark in Azure

What is Dataset Analysis?
Dataset Analysis is defined as manipulating or processing unstructured data or raw data to draw valuable insights and conclusions that will help derive critical decisions that will add some business value. The dataset analysis process is followed by organizing the dataset, transforming the dataset, visualizing the dataset, and finally modeling the dataset to derive predictions for solving the business problems, making informed decisions, and effectively planning for the future.

Data Pipeline:
Data pipeline involves extracting or capturing data using various tools, storing raw data, cleaning, validating data, transforming data into a query-worthy format, visualizing KPIs, and orchestration of the above process. It refers to a system for moving data from one system to another. The data may or may not be transformed, and it may be processed in real-time (or streaming) instead of batches.

What is the Agenda of the project?
The Agenda of the project involves deriving Movie Recommendations using Python and Spark on Microsoft Azure. We first understand the problem and download the Movielens dataset from the grouplens website. Then a subscription is set up for using Microsoft Azure, and categorization of resources is done into a resource group. A standard storage account is a setup to store all the data required for serving movie recommendations using Python and Spark on Azure, followed by creating a standard storage blob account in the same resource group. Firstly, we make containers in a standard storage account and standard storage blob account and upload the movielens zip file dataset in its standard storage blob account. Then we create an Azure data factory, a copy data pipeline, and start link storage for standard blob storage account in the Azure data factory. We are copying data from Azure blob storage to Azure data lake storage using a copy data pipeline in the Azure data factory. It is followed by creating the databricks workspace, cluster on databricks, and accessing Azure data lake storage from databricks. We are creating mount points and extracting the zip file to get CSV files. Finally, we upload files into databricks, read the datasets into Spark dataframes in databricks, and analyze the dataset to get the movie recommendations.

Usage of Dataset:
Here we are going to use Movielens data in the following ways:

- Extraction: During the extraction process, the Movielens data zip file is extracted to get the CSV files out of it in two ways: the Databricks local file system(DFS) and the Azure data factory(ADF) copy pipeline.

- Transformation and Load**:** During the transformation and load process, the uploaded dataset in Spark is read into Spark dataframes. Data tags are also read into Spark in

Databricks, and output is displayed through Bar chart. And dataset is finally analyzed in Databricks into Spark, and movies are recommended.

Data Analysis:

- From the grouplens website, data is downloaded containing names of movies, ratings given to the movies, links of the movies, and tags assigned to the movies.

- Resource manager is created in Azure to categorize the resources required, followed by a Storage account.
- The Copy Data pipeline is created to copy the data from Azure blob storage to Azure data lake storage in the Azure data factory.

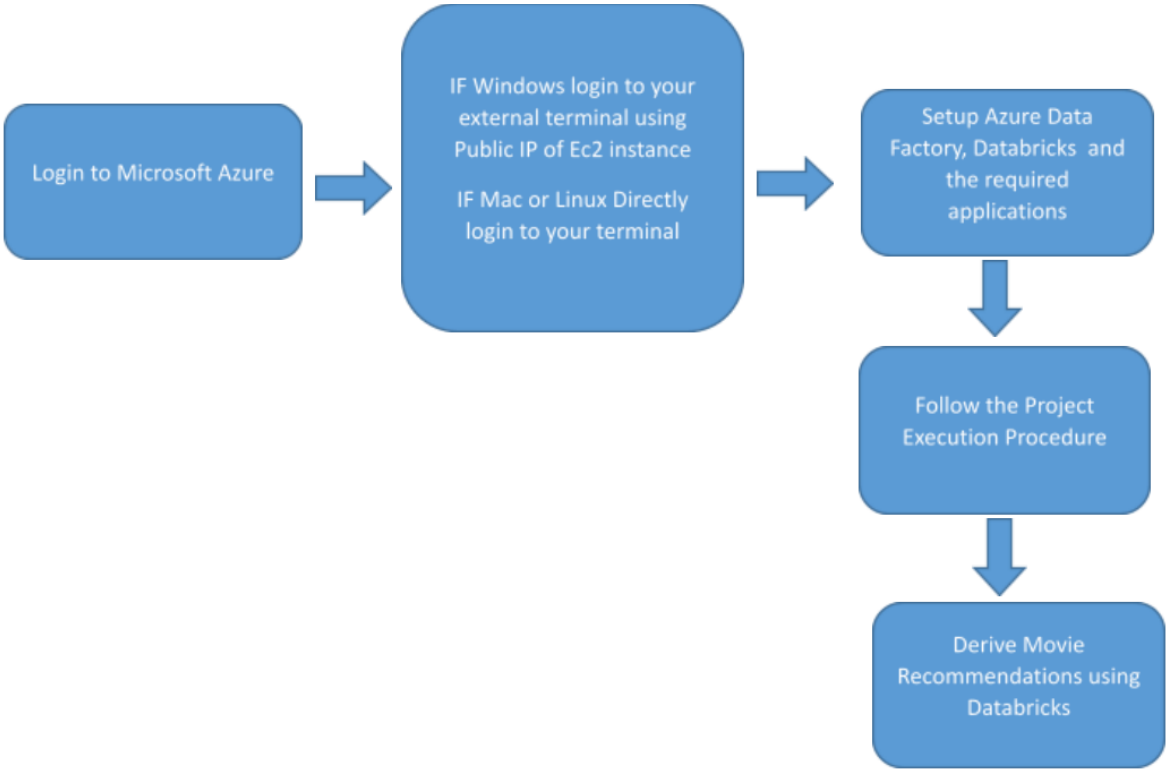
- The Databricks workspace and cluster are created and accessed Azure data lake storage from databricks followed by the creation of Mount pairs.

- The extraction process is done by extracting the Movielens data zip file to get the CSV files out of it using the Databricks file system(DFS) and using the Azure data factory(ADF).

- In the transformation and load process, the uploaded dataset in Spark is read into Spark dataframes. Data tags are read into Spark in Databricks.

- Finally, data is analyzed into Spark in Databricks using mount points, and data is visualized using bar charts.

Project Workflow:



Folder Structure: