Project name :- data pipeline   
Objective :- get a raw data and transform it in a cleaned data so that it can be visualize in a power BI

we downloaded a dataset from Kaggle of OOT platform apps in csv file of 5 gb data   
put it in a s3 bucket through EMR (downloaded it in emr and uploaded it in s3

then we cleaned the data by using pyspark .  
if the same type of data comes again then we need to cleaned it again in pyspark   
Hence we wrote a pyspark script in glue job which can clean the same type of data   
(command to clean the data )

stopwords , null ,replace , change the data type , split , column drops   
we converted the csv file into the parquet bcz it consumes less data and can be queried in athena then the cleaned data is uploaded to other s3 bucket

then the cleaned data is uploaded to a s3 bucket from the glue job   
and the crawler takes the data from the s3 bucket by using a trigger

crawler will define the schema of the data so the data can be queried in the athena   
and then by using ODBC connector we connected the data from athena to power BI   
In power BI we made separate reports for each app and created two pages for IMDB and rotten tomatos

Insights for the power BI are :- user can decide the which movie he has to see by applying various filter on the power BI .   
genere , rating , movie release year , director , cast , timming , movei/series

If the user want to see a movie he can also check it in the IMDB or rotten tomatos page which contain all the movies and the series irrespective to the application.