# Building an ETL Pipeline using Azure Data Services

## DESCRIPTION

Use the data analytics stack to build a data pipeline using Data Factory, Databricks and Synapse.

## Problem Statement:

As a Data Engineer, you’ve been asked to access the services that can help with ETL of data in the cloud data storage to enable analytics through Synapse. In this POC, we will be collecting the data from SQL Database using ADF and the transformed data will be the source for databricks to run complex transformations and once data is analysed using Databricks, it is synced into synapse analytics data warehouse as historical dataset for enabling various analytics.

## Domain: Analytics

## Steps for building ETL pipeline :

In this project, perform the following steps:

- Create a Resource Group.
- Create a Storage account.
- Create an Azure SQL Database.
- Create a data factory.
- Configure Databricks cluster
- Create Synapse analytics Data Warehouse.
- Use the different Azure data factory tools to build a pipeline (SQL Database-> Copy-> ADLS Gen 2 -> Transform using Databricks -> Copy to Synapse DW).
- Use Databricks notebook for mounting ADLS Gen 2 storage, transforming the data (clean, join, filter, aggregate, pivot) and persist result to ADLS.
- Schedule and Monitor the pipeline and activity runs.

## Questions that need to be answered/Evaluation steps while building the ETL Pipeline :

## Task 1: Create a dataflow with the following requirement:

1. Create a data stream named CleaningGenreRomance and perform data cleansing on the Genre column using Derived Column and case expression.  (While collecting data it was observed that some genres have spelling mistakes like romance, Romence for Romance, comedy, Comdy for Comedy.)

2. Create a data stream named CountMoviesBasedOnGenre that can calculate number of films for each genre and store it as a separate dataset in ADLS under folder name “solution/genreCount”

3. Create a new stream named  JoinMovieCountWithCleanData. Perform join operation on CountMoviesBasedOnGenre with CleaningGenreRomance stream and store the same in the Azure SQL Database.

## Task 2: Create the following activity pipeline

1. Get the clean data from Azure SQL DB. Create an activity that can copy the data from SQLDB to ADLS Gen2.

2. Create an activity that can use Azure Databricks to read the data from the ADLS Gen2 and perform rank operation on the Genre column. Ensure this activity gets activated only after the data is stored in ADLS from SQL DB. The result of Databricks must be stored in the ADLS.

3. Create a final activity that will read the output of previous activity in ADLS and store the same in Synapse.

## Data Dictionary :

| Column Name | Data Type | Description |
|---|---|---|
| Film | String | Name of the Film |
| Genre | String | Type/category  of the Film |
| Lead Studio | String | Name of the Production Company |
| Audience score % | integer | score of the audience for movie |
| Profitability | String | profit |
| Rotten Tomatoes % | Integer | ratings of rotten tomatoes |
| Worldwide Gross | float | collections of the fim |
| Year | integer | release year of the film |

## Dataset :

Find the dataset here