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 :

- How do you integrate SQL DB,ADLS Gen 2 to Databricks using ADF Pipeline
    - Get the clean data from dataflow and perform Rank the column based on genre using Azure notebook.

Dataflow activity may include,

- Clean the data and derive a new column for the genre ‘Romance’
- Calculate number of films for each genre
- Join the newly generated dataset with the existing one and remove the duplicates
- Sink the output to Synapse for analytics

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