Azure-Databricks-project-on-Yelp-Dataset

Author: Santhi Maddipudi

GitHub: https://github.com/santhimaddipudi/Azure-Databricks-project-on-Yelp-Dataset

Analyse Yelp Dataset with Spark & Parquet Format on Azure Databricks

In this Databricks Azure project, you will use Spark & Parquet file formats to analyse the Yelp reviews dataset. As part of this you will deploy Azure data factory, data pipelines and visualise the analysis.

What is Dataset Analysis?

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

Data Pipeline

A data pipeline is a technique for transferring data from one system to another. The data may or may not be updated, and it may be handled in real-time (or streaming) rather than in batches. The data pipeline encompasses everything from harvesting or acquiring data using various methods to storing raw data, cleaning, validating, and transforming data into a query-worthy format, displaying KPIs, and managing the above process.

Business Overview

Azure Databricks is a data analytics tool tailored for the Microsoft Azure cloud services platform. In massive data pipelines, raw and structured data is imported into Azure in batches via Azure Data Factory or streamed near real-time via Apache Kafka, Event Hub, or IoT Hub. This data is stored in a data lake for long-term sustained storage, either in Azure Blob Storage or Azure Data Lake Storage. Azure Databricks is utilized to read data from different data sources and transform it into breakthrough insights using Spark as part of the analytics workflow.

Yelp is a community review site and an American multinational firm based in San Francisco, California. It publishes crowd-sourced reviews of local businesses as well as the online reservation service Yelp Reservations. Yelp has made a portion of their data available in order to launch a new activity called the Yelp Dataset Challenge, which allows

anyone to do research or analysis to find what insights are buried in their data. Due to the bulk of the data, this project only selects a subset of Yelp data in a zip file named 'dataset.zip,' which comprises three JSON files, including 'business.json', which provides business data such as location data, attributes, and categories.

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

- **Conversion**: During the conversion process, the Yelp academic dataset JSON file is converted to Parquet format and further Parquet format is converted to the Delta format for further data analysis in Databricks.
- **Transformation and Load**: During the transformation and load process, the uploaded dataset in Spark is read into Spark data frames. And dataset is finally analyzed in Databricks into Spark and further recommendations are deduced.

Data Analysis:

- From the Yelp website, the academic dataset is downloaded containing business, checkin, review, tips and users.
- The resource manager is created in Azure to categorise the resources required followed by Storage account for storing data required and the Creation of containers for uploading the dataset.
- The pipeline is created to copy the data from Azure storage to Azure data lake storage in the Azure data factory.
- The Databricks workspace and cluster is created, accessed and configured Azure data lake storage from databricks.
- The conversion process is done by converting the Yelp academics data file from JSON format to Parquet format and further converting it to Delta format for smooth analysis.
- In the transformation and load process, the uploaded dataset in Spark is read into Spark data frames.
- Finally, data is analyzed into Spark in Databricks deducing recommendations and data are visualized using bar charts.

NOTE:

- The Container in Azure is created with the name “yelpcontainer” for uploading the dataset.

- The Yelp dataset files are uploaded in the Container in Azure.

Approach

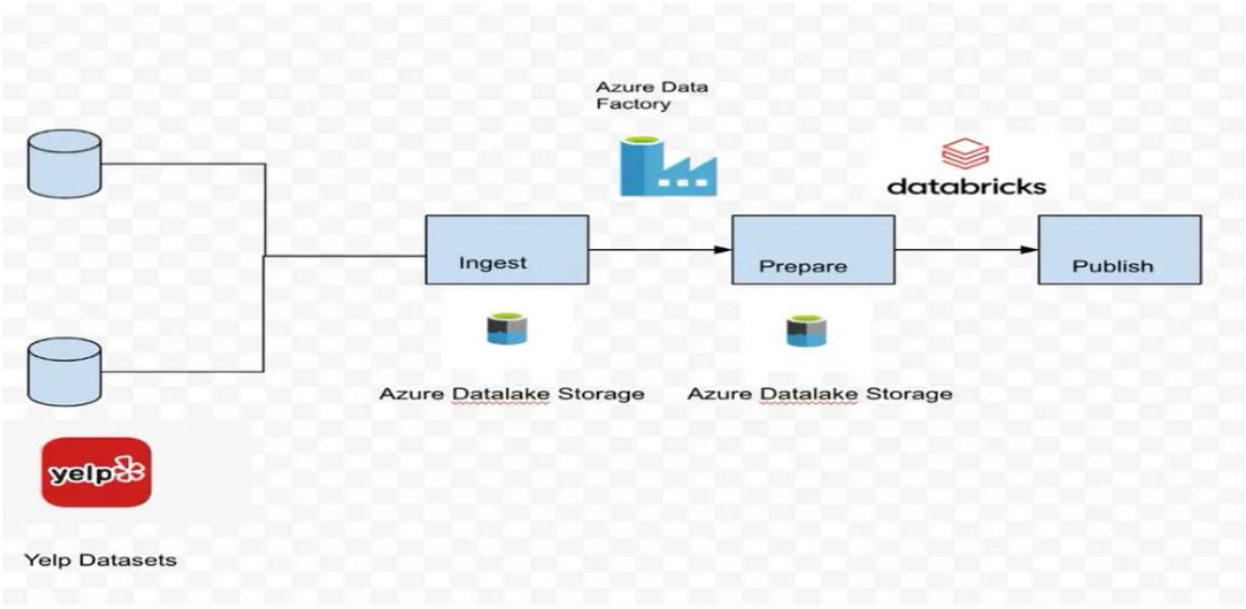
- Read yelp datasets in ADLS and convert JSON to parquet for better performance.
- Convert JSON to Delta Format.
- Total records in each dataset.
- Partition tip dataset tip by a date column.
- repartition() vs coalesce()
- Find the top 3 users based on their total number of reviews.
- Find the top 10 users with the most fans
- Analyse the top 10 categories by a number of reviews.
- Analyse top businesses which have over 1000 reviews.
- Analyse Business Data: Number of restaurants per state.
- Analyze the top 3 restaurants in each state.
- List the top restaurants in a state by the number of reviews.
- Numbers of restaurants in Arizona state per city.
- Broadcast Join: restaurants as per review ratings in Pheonix city.
- Most rated Italian restaurant in Pheonix.

Tech Stack

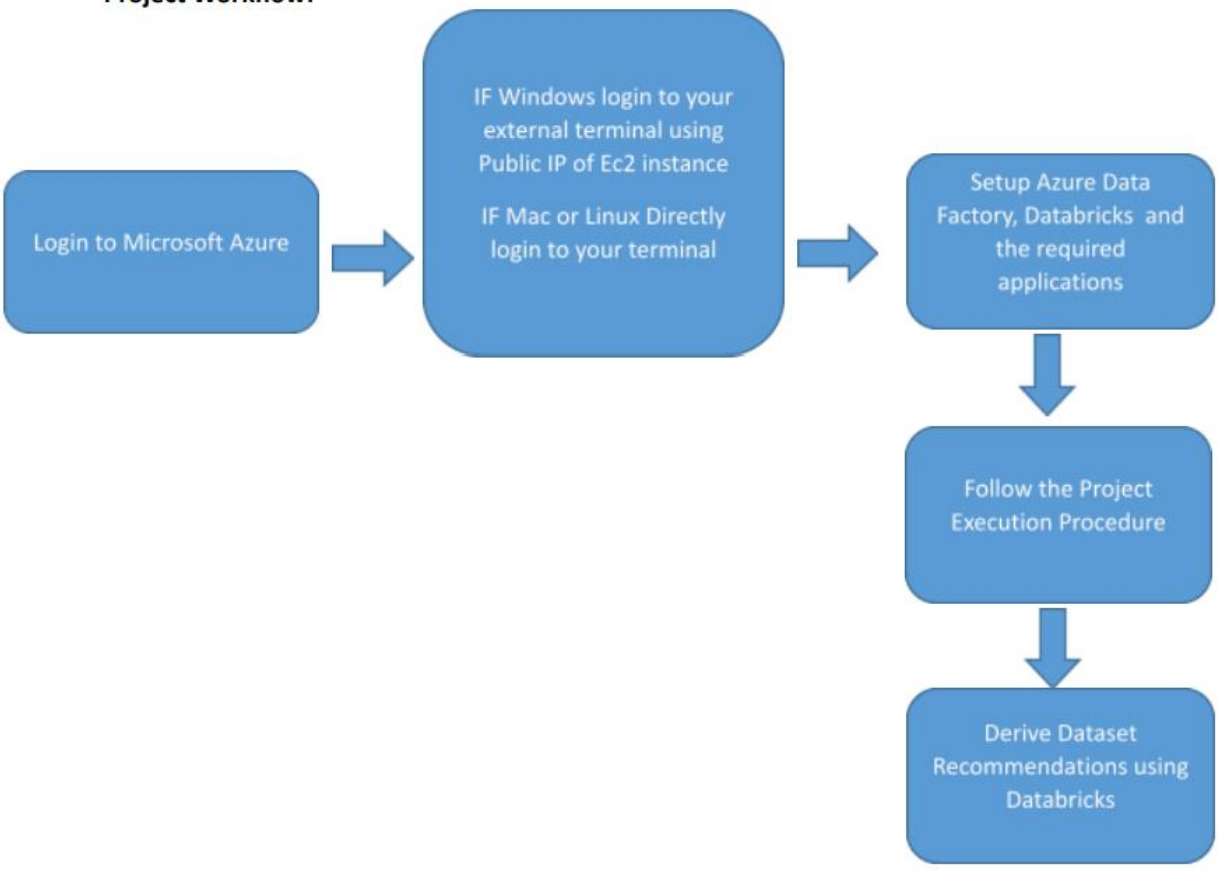
➔ Language: Python3

➔ Services: Azure Data factory, Azure Databricks, ADLS

Architecture



Project Workflow:



Folder Structure: