Analyzing Yelp Dataset with Spark & Parquet Format on Azure Databricks

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:

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. Right from extracting or capturing data using various tools, storing raw data, cleaning, validating data, transforming data into query worthy format, visualisation of KPIs including Orchestration of the above process is data pipeline.

What is the Agenda of the project?

The Agenda of the project involves Analyzing Yelp Dataset with Spark & Parquet Format on Azure Databricks. We first download the Yelp dataset from the Yelp website and understand the problem. Then a solution architecture is designed which defines ingestion of data, preparation of data and publishing it on Databricks. Then subscription is set up for using Microsoft Azure and categorisation of resources are done into a resource group. A standard storage account is set up to store all the data required for Analyzing Yelp Dataset with spark & Parquet format on Azure Databricks. Creation of containers in a standard storage account and uploading of the Yelp dataset in it. Creating an Azure data factory, a copy data pipeline and starting link storage for standard storage account in Azure data factory. Copying data from Azure storage to Azure data lake storage using a copy data pipeline in Azure data factory. It is followed by the conversion of Yelp dataset from JSON to Parquet file format and JSON to Delta format. Then performing partitioning, repartitioning and coalesce on the dataset in Databricks. Performing data analysis on the repartitioned dataset and finally deducing the recommendations.

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:

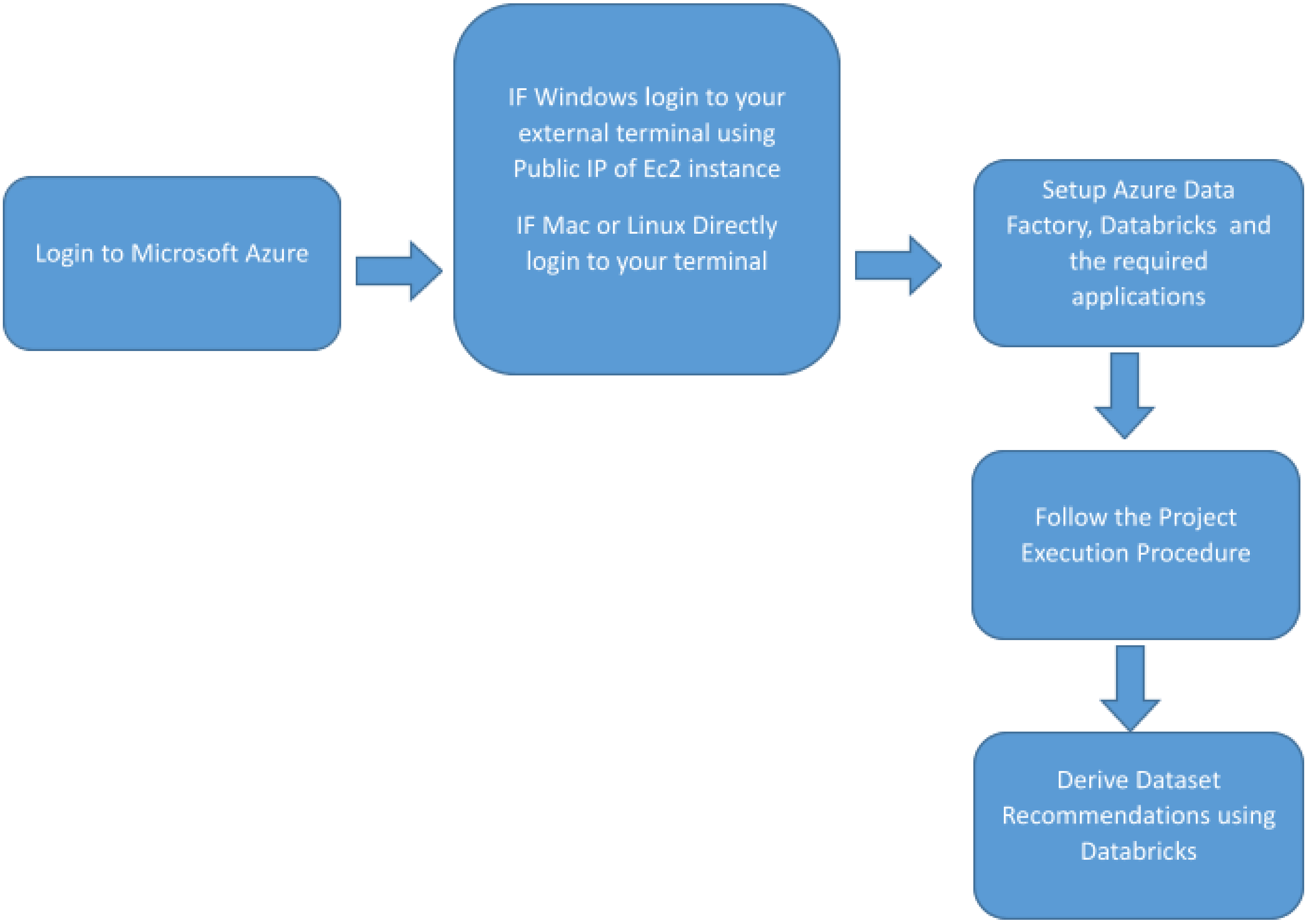
- From the Yelp website, the academic dataset is downloaded containing business, checkin, review, tips and users.

- 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.

Project Workflow:



Login to Microsoft Azure

IF Windows login to your external terminal using Public IP of Ec2 instance

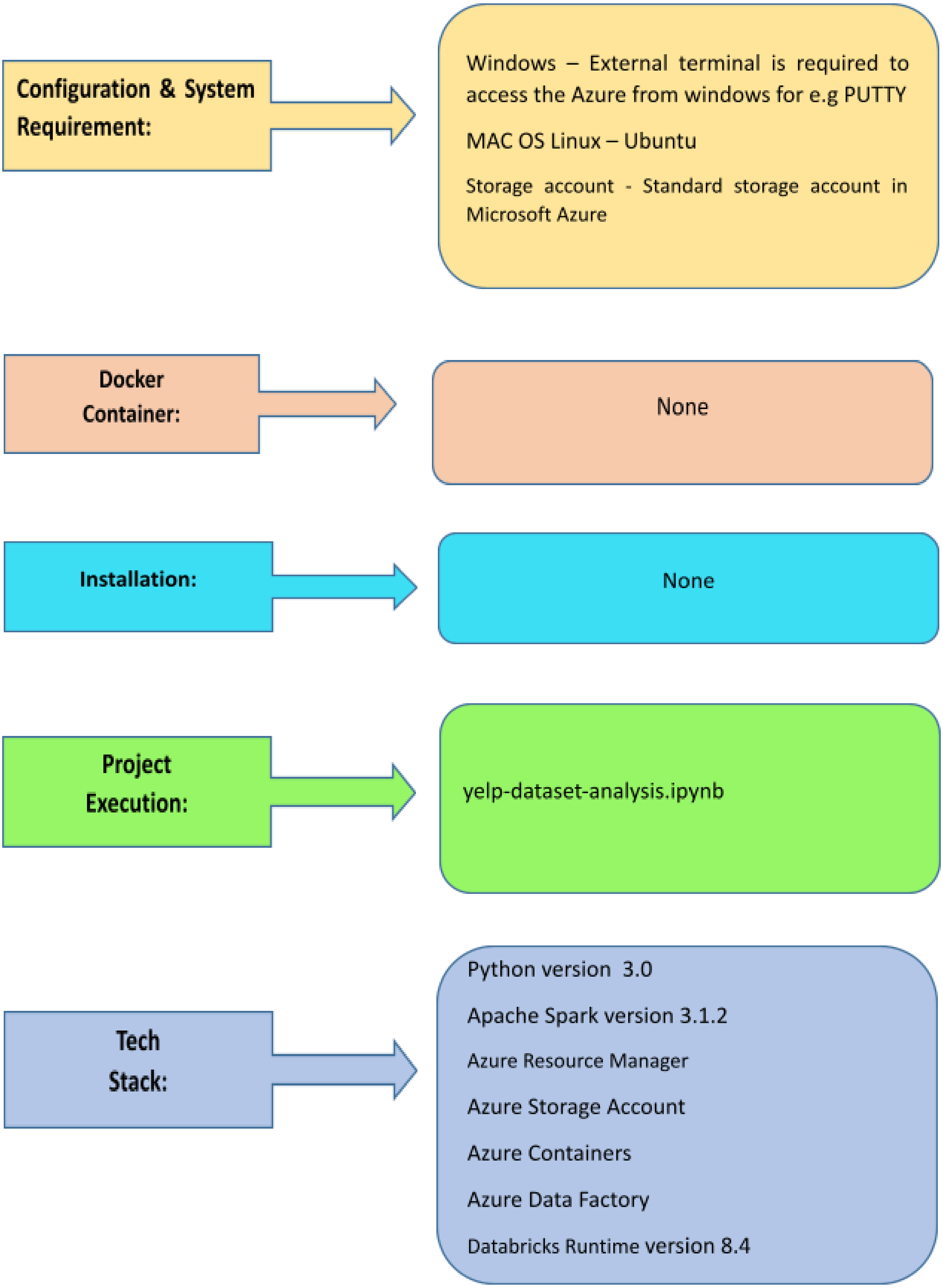
IF Mac or Linux Directly login to your terminal

Setup Azure Data Factory, Databricks and the required applications

Follow the Project Execution Procedure

Derive Dataset Recommendations using Databricks

Folder Structure:



Configuration & System Requirement:

Windows – External terminal is required to access the Azure from windows for e.g PUTTY

MAC OS Linux – Ubuntu

Storage account - Standard storage account in Microsoft Azure

Docker Container:

None

Installation:

None

Project Execution:

yelp-dataset-analysis.ipynb

Tech Stack:

Python version 3.0

Apache Spark version 3.1.2

Azure Resource Manager

Azure Storage Account

Azure Containers

Azure Data Factory

Databricks Runtime version 8.4