Data Ingestion and Processing Pipeline on GCP Project Overview

What is Data Ingestion?

Data Ingestion is defined as the transportation of data from various assorted sources to the storage medium where it can be thoroughly accessed and analyzed by any organization. The storage medium acts as a destination which is typically the data warehouse, database, data mart or any document store. The data can come from various sources such as RDBMS and other different types of databases like S3 buckets, CSVs files etc.

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 Data ingestion and processing pipeline on Google cloud platform with real-time streaming and batch loads. .Yelp dataset, which is used for academics and research purposes is used. We first create a service account on GCP followed by downloading Google Cloud SDK(Software developer kit). Then, Python software and all other dependencies are downloaded and connected to the GCP account for further processes. Then, the Yelp dataset is downloaded in JSON format, is connected to Cloud SDK following connections to Cloud storage which is then connected with Cloud Composer and Yelp dataset JSON stream is published to PubSub topic. Cloud composer and PubSub outputs are Apache Beam and connecting to Google Dataflow. Google BigQuery receives the structured data from workers. Finally., the data is passed to Google Data studio for visualization.

Usage of Dataset:

Here we are going to use Yelp data in JSON format in the following ways:

- Yelp dataset File: In Yelp dataset File, JSON file is connected to Cloud storage Fuse or Cloud SDK to the Google cloud storage which stores the incoming raw data followed by connections to Google Cloud Composer or Airflow to the Google cloud storage for scheduling and orchestration to batch workloads.

- Yelp dataset Stream: In Yelp dataset Stream, JSON Streams are published to Google PubSub topic for real-time data ingestion followed by connections to Apache beam for further processing.

Key Takeaways

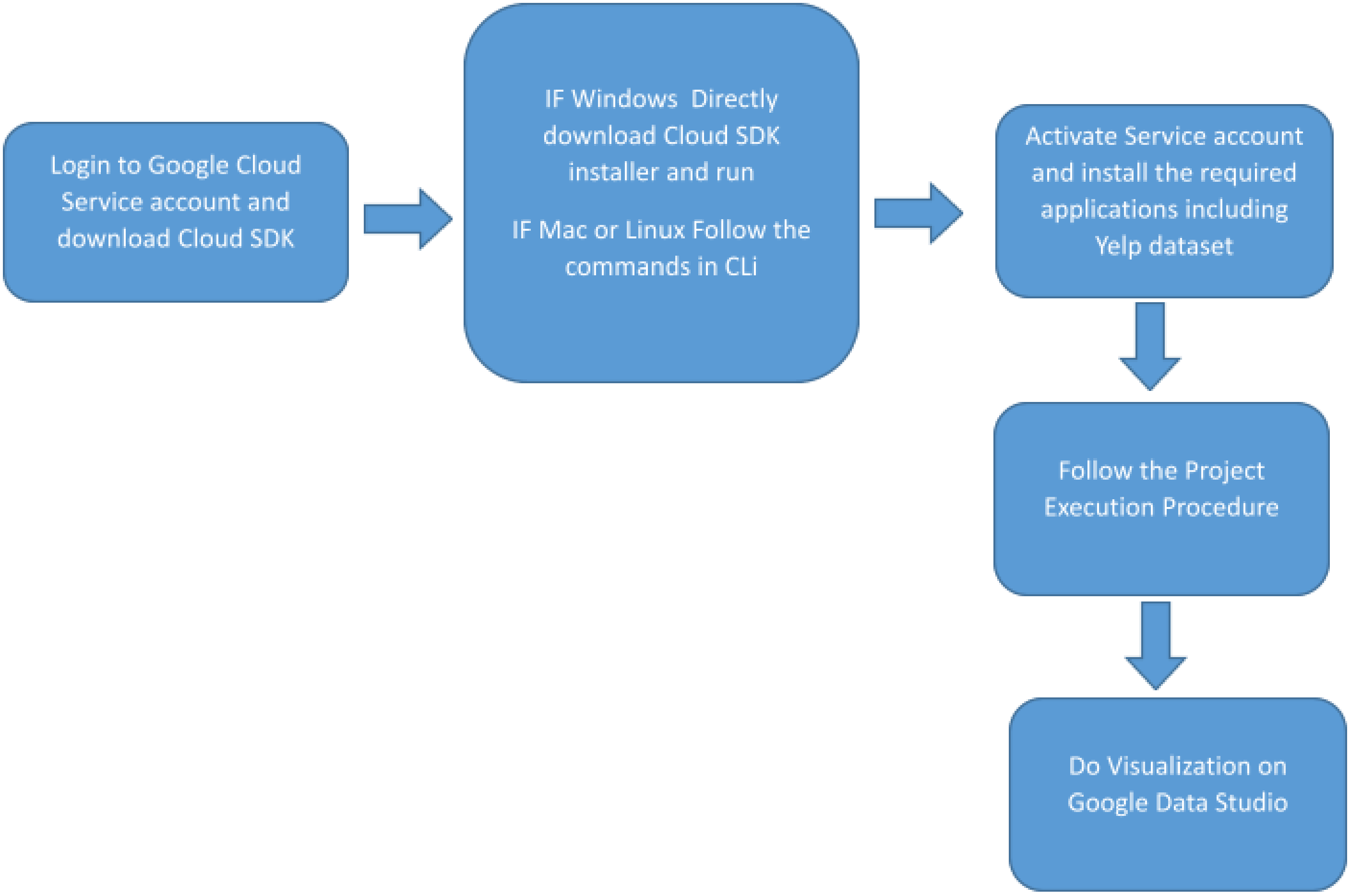
- Understanding the project and how to use Google Cloud Storage.
- Understanding the basics of data pipelines, data ingestion and their application
- Visualizing the complete Architecture of the system
- Introduction to Google Cloud SDK
- Usage of Google cloud SDK and connecting it to the service account
- Exploring Yelp dataset and Using its JSON stream and JSON file
- Creating Google Cloud Bucket

- Understanding PubSub and using it for data ingestion by creating topics
- Understanding Apache Beam and using it for executing data processing pipelines
- Creating data flow stream jobs and understanding it
- Creating data flow batch jobs and understanding it
- Understanding Cloud Composer/Airflow and using it for orchestrating batch workloads
- Creating Cloud Composer Instance and using for scheduling a job
- Tracking Dag runs and other configurations in Cloud Composer
- Integrating PubSub, Cloud Composer/Airflow to Apache beam
- Integrating Google Cloud Data flow and Google BigQuery
- Understanding Google BigQuery and using it as a data warehouse.
- Integrating BigQuery and Data Studio
- Displaying live stream results using Google Data Studio

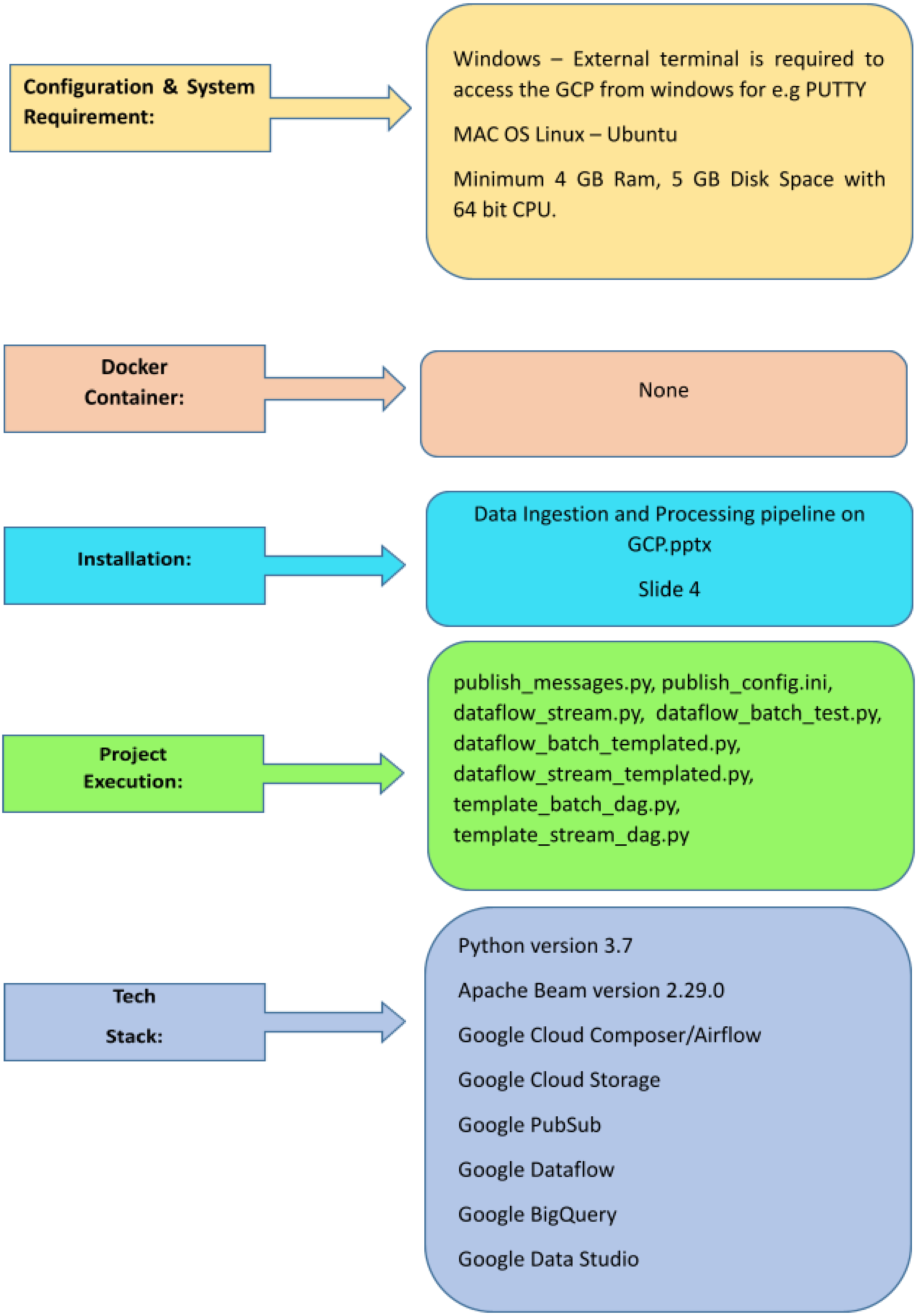
Data Analysis:

- From the given website, the Yelp dataset is downloaded in JSON format. The Yelp JSON file is connected to Google SDK or GcsFuse for transfer of data to Google cloud storage which is connected to Google Cloud composer/Airflow for scheduling and orchestration of batch workloads.
- Yelp dataset JSON streams are published to Google PubSub which is used for real-time ingestion or streaming datasets.
- Data pipeline is created by apache beam which receives the real-time data from Google PubSub and the data from Google cloud storage as inputs which are followed by creating Google dataflow stream job and batch job scaling the compute based on throughput.
- Apache beam orchestrates stream and batch jobs following the output of Google Dataflow to workers.
- Google BigQuery acts as a Data warehouse storing structured data which receives the input from workers and queries the data.
- Finally data is visualized using different graphs and table definitions in Google Data Studio.

Project Workflow:


Login to Google Cloud Service account and download Cloud SDK
IF Windows Directly download Cloud SDK installer and run
IF Mac or Linux Follow the commands in CLi
Activate Service account and install the required applications including Yelp dataset
Follow the Project Execution Procedure
Do Visualization on Google Data Studio

Folder Structure:



Configuration & System Requirement:

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

MAC OS Linux – Ubuntu

Minimum 4 GB Ram, 5 GB Disk Space with 64 bit CPU.

Docker Container:

None

Installation:

Data Ingestion and Processing pipeline on GCP.pptx

Slide 4

Project Execution:

publish_messages.py, publish_config.ini, dataflow_stream.py, dataflow_batch_test.py, dataflow_batch_templated.py, dataflow_stream_templated.py, template_batch_dag.py, template_stream_dag.py

Tech Stack:

Python version 3.7

Apache Beam version 2.29.0

Google Cloud Composer/Airflow

Google Cloud Storage

Google PubSub

Google Dataflow

Google BigQuery

Google Data Studio