Data Ingestion and Processing pipeline on GCP

Problem Statement

With the evolution of Cloud Platforms, nowadays most of the companies are shifting their data platforms on cloud. While AWS and Azure are already in the market for a while now, but with the release of Google Cloud Platform, many companies are planning to include GCP in their data migration journey as it provides much more flexibility in terms of compute, security and rich machine learning capabilities. This brings the need to design an effective data pipeline utilizing GCP capabilities for batch and real time loads for data driven products and companies.

Solution Approach

Data Platform is the core of any Data product and setting it in a right way is a key for its viability. The solution design will harness the power of services offered by Google to build an end to end highly scalable ingestion and processing pipeline. The pipeline design will cater the needs for both real time and batch loads.

Services Used/Tech Stack

Cloud Composer/Airflow - Scheduling and Orchestration

Google Cloud Storage(GCS) - Storage and Archival

Pub-Sub - Real time Ingestion

Dataflow - Scaling the compute based on Throughput

BigQuery – Warehouse for storing structural data

GcsFuse/Google SDK – Data transfer to GCS

Data Studio – Visualization of Data

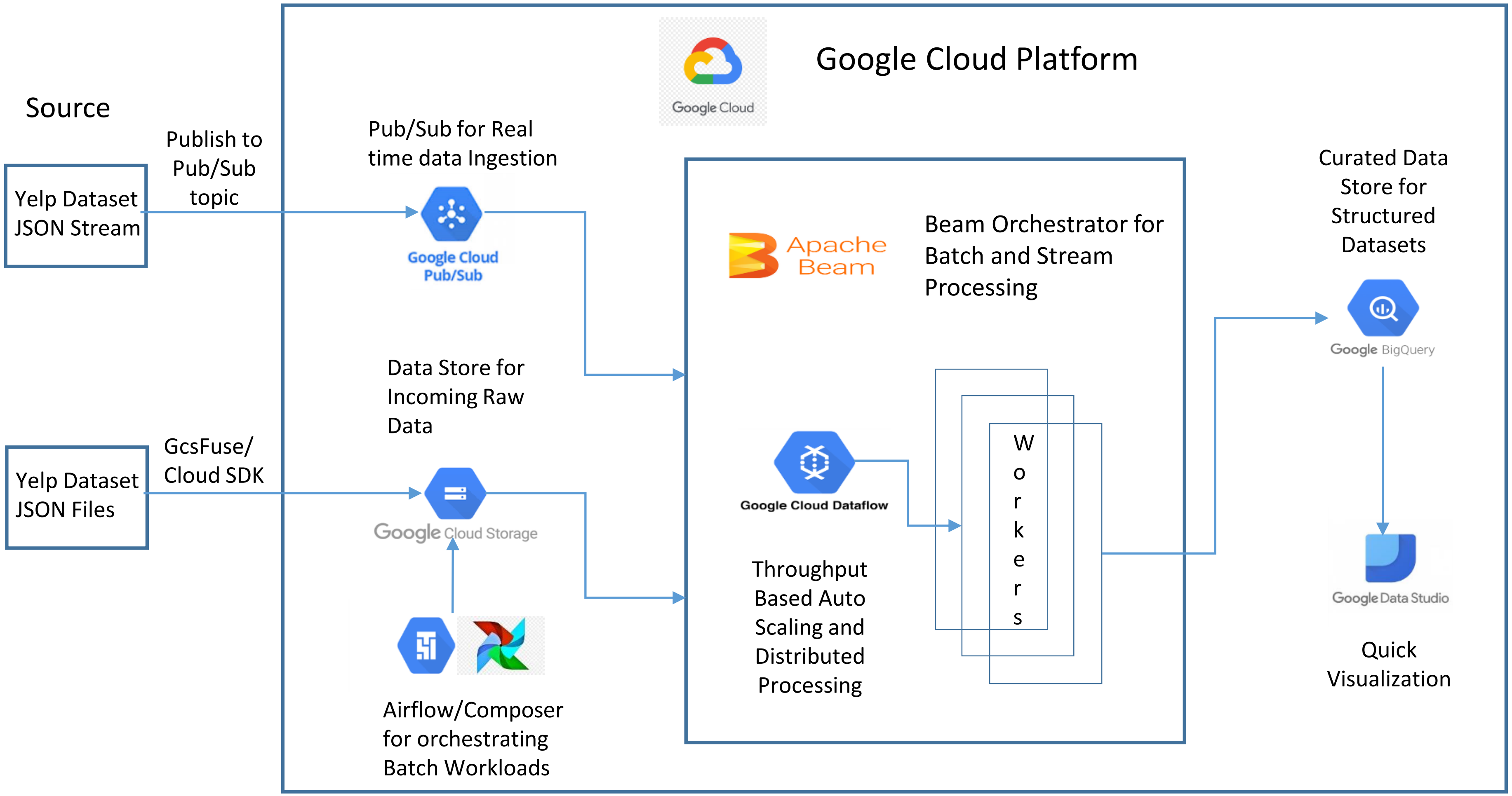
Apache Beam – Orchestrating Dataflow Jobs

Python 3.7 – Programming Needs

DataSet

Yelp Dataset will be used to prove the efficacy of this approach. Yelp provide dataset for academic and research purposes which can be utilized without any restrictions.

Solution Design



Source
Yelp Dataset JSON Stream
Publish to Pub/Sub topic
Yelp Dataset JSON Files
GcsFuse/ Cloud SDK
Google Cloud
Google Cloud Platform
Pub/Sub for Real time data Ingestion
Google Cloud Pub/Sub
Data Store for Incoming Raw Data
Google Cloud Storage
Airflow/Composer for orchestrating Batch Workloads
Apache Beam
Beam Orchestrator for Batch and Stream Processing
Google Cloud Dataflow
Throughput Based Auto Scaling and Distributed Processing
Workers
Curated Data Store for Structured Datasets
Google BigQuery
Google Data Studio
Quick Visualization