GCP Data Engineer Certification Notes

- Some data can be directly loaded into data warehouses using an extract and load approach, while others might be transformed before being uploaded into the data warehouse.
- Building, deploying, and operating effective flexible data pipelines for all the stages of data processing is a primary expectation from you as a Professional Data Engineer.
- EL, ETL, or ELT and choose the right Google Cloud tools for the job.
- Cloud Dataflow, Cloud Dataproc, Cloud Pub/Sub, and Cloud Composer
- Data pipelines are modeled as Directed acyclic graphs (DAGs).
- Cycles are not allowed in data pipelines, and for that reason the graphs that model data pipelines are directed acyclic graphs.

Data Pipeline Stages

The four types of stages in a data pipeline are as follows

- Ingestion
- Transformation
- Storage
- Analysis

Ingestion:

- Ingestion (see Figure 3.3) is the process of bringing data into the GCP environment. This can occur in either batch or streaming mode.
- In batch mode, data sets made up of one or more files are copied to GCP. Often these files will be copied to Cloud Storage first. There are several ways to get data into Cloud Storage, including gsutil copying, Transfer Service, and Transfer Appliance
- Streaming ingestion receives data in increments, typically a single record or small batches of records, that continuously flow into an ingestion endpoint, typically a Cloud Pub/Sub topic

Transformation:

- In GCP, Cloud Dataflow and Cloud Dataproc are often used for transformation stages of both batch and streaming data.
- Cloud Dataprep is used for interactive review and preparation of data for analysis. Cloud Datafusion can be used for the same purpose, and it is more popular with enterprise customers

Storage:

- BigQuery can treat Cloud Storage data as external tables and query them. Cloud Dataproc can use Cloud Storage as HDFS-compatible storage.
- BigQuery is an analytical database that uses a columnar storage model that is highly efficient for data warehousing and analytic use cases.
- Bigtable is a low-latency, wide-column NoSQL database used for time-series, IoT, and other high-volume write applications. Bigtable also supports the HBase API, making it a

good storage option when migrating an on-premises HBase database on Hadoop (see Figure 3.5).

Analysis

- Data in BigQuery, for example, is analyzed using SQL. BigQuery ML is a feature of the product that allows SQL developers to build machine learning models in BigQuery using SQL.
- Data Studio is a GCP service used for interactive reporting tool for building reports and exploring data that is structured as dimensional models. Cloud Datalab is an interactive workbook based on the open source Jupyter Notebooks. Datalab is used for data exploration, machine learning, data science, and visualization.
- Large-scale machine learning models can be built using Spark machine learning libraries running on Cloud Dataproc while accessing data in Bigtable using the HBase interface

---

Cloud Dataflow, Cloud Dataproc, Cloud Dataprep, Cloud Pub/Sub, Cloud Composer, Cloud Datafusion, Transfer Service, Transfer Appliance, BigQuery, Bigtable, Data Studio