

GOOGLE CLOUD DATAFLOW



- Cloud Dataflow is a fully-managed service for transforming and enriching data in real time and batch modes
- Cloud dataflow provides a serverless architecture that can be used to process very large batch data sets, or high volume live streams of data in parallel
- It enables developers to set up processing pipelines for integrating, preparing and analyzing large data sets
- It is based on the Apache Beam open source project
- Tightly integrated with Cloud Pub/Sub, Big query and cloud ML

GOOGLE CLOUD BIGQUERY



- BigQuery is a fully-managed enterprise data warehouse, that helps to manage and analyze data with built-in features like machine learning, geospatial analysis, and business intelligence
- BigQuery's serverless architecture uses SQL queries to answer organization's biggest questions with zero infrastructure management
- The query engine is capable of running SQL queries on terabytes of data in a matter of seconds, and petabytes in only minutes
- BigQuery relies on two systems unique to Google, the Colossus File System and Jupiter networking, to ensure that data can be queried quickly no matter where it physically resides in the compute cluster
- BigQuery can process the data from sources such as GCP cloud storage, GCP Bigtable etc.,

GOOGLE CLOUD DATAPREP



- Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning
- Dataprep is serverless and works at any scale, there is no infrastructure to deploy or manage
- Fast exploration and anomaly detection
- Dataprep uses a proprietary inference algorithm to interpret the data transformation intent of a user's data selection
- Comprehensive connectivity
- Schedule and automate your data preparation jobs by chaining them together in sequential and conditional order