How to bring data from SAP to GCP

A screenshot of a video game

Description automatically generated

Cloud Data Fusion

Cloud Data Fusion is a fully managed, cloud-native, enterprise data integration service for quickly building and managing data pipelines. Business users, developers, and data scientists can easily and reliably build scalable data integration solutions to cleanse, prepare, blend, transfer, and transform data without having to wrestle with infrastructure.

A blue and white logo

Description automatically generated

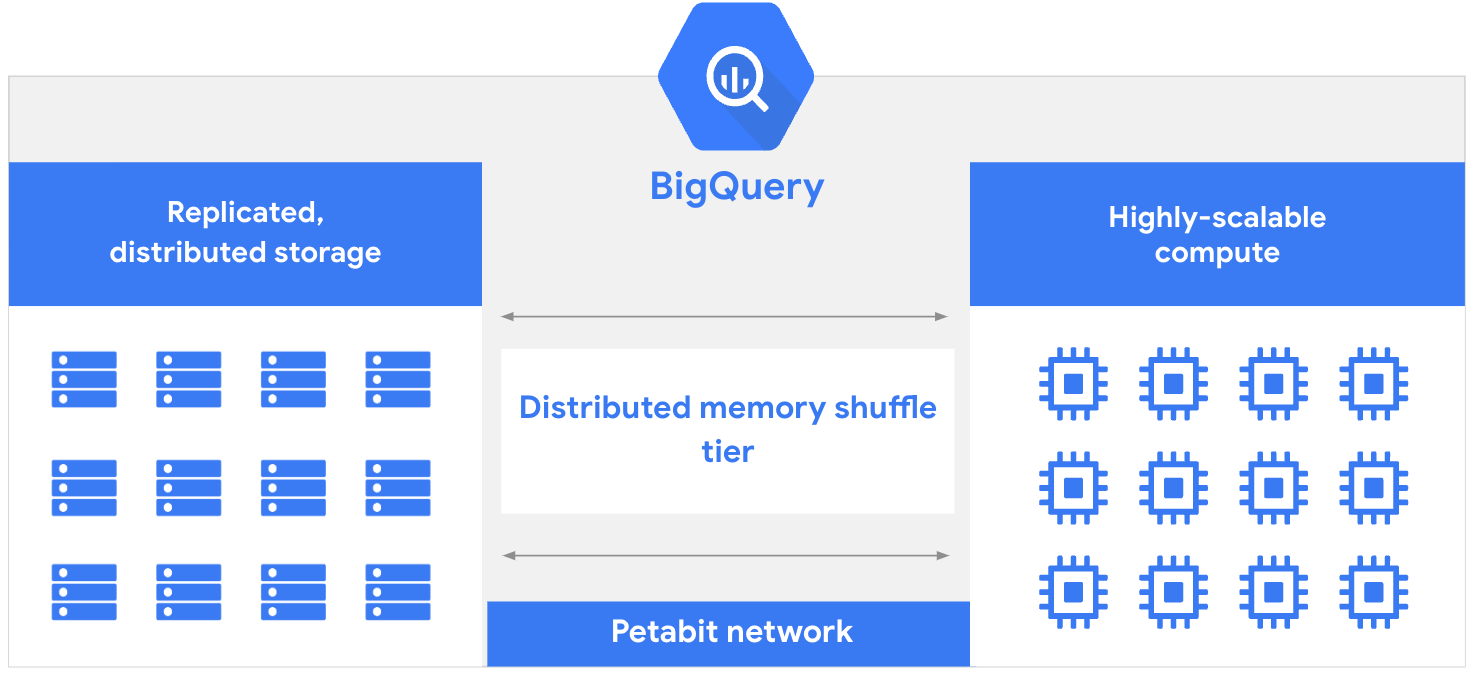
Cloud Data Fusion can Wrangle, Cleanse and Load the data coming from the database (SAP) and load it to BigQuery.

BigQuery is a fully managed, AI-ready data platform that helps you manage and analyze your data with built-in features like machine learning, search, geospatial analysis, and business intelligence. BigQuery's serverless architecture lets you use languages like SQL and Python to answer your organization's biggest questions with zero infrastructure management.

BigQuery provides a uniform way to work with both structured and unstructured data and supports open table formats like Apache Iceberg, Delta, and Hudi. BigQuery streaming supports continuous data ingestion and analysis while BigQuery's scalable, distributed analysis engine lets you query terabytes in seconds and petabytes in minutes.

BigQuery's architecture consists of two parts: a storage layer that ingests, stores, and optimizes data and a compute layer that provides analytics capabilities. These compute and storage layers efficiently operate independently of each other thanks to Google's petabit-scale network that enables the necessary communication between them.

Legacy databases usually have to share resources between read and write operations and analytical operations. This can result in resource conflicts and can slow queries while data is written to or read from storage. Shared resource pools can become further strained when resources are required for database management tasks such as assigning or revoking permissions. BigQuery's separation of compute and storage layers lets each layer dynamically allocate resources without impacting the performance or availability of the other.



FROM SAP TO GCP

To integrate with SAP Data Sources, CDF supports 4 plugins:

* Table Batch Source
* OData
* ODP
* SLT replication

Pipelines:

A screenshot of a computer

Description automatically generated

It’s possible to consume the so called Accelerators: pre-packaged and standard pipelines -> based on OData technology.

Once the data is in BigQuery we can build our AI Based Applications, with VertexAI for example.