# Google Cloud Platform (GCP) - Big Data & Analytics

Efficiently capture, process, and analyze data with Google Cloud data analytics products. Focus on finding insights rather than managing infrastructure. Combine cloud-native services with open source tools as needed, both in batch and stream mode providing big data & analytics solutions.

## Features

* Fully-managed big data (petabyte-scale) SQL data warehouse
* Managed SQL and highly-scalable managed NoSql databases
* Process structured and unstructured data
* Fully-managed global-scale transactionally-consistent SQL database
* Google Chronicle security analytics platform (big data security solution)
* Fully managed alternatives to Hadoop
* Process data in batch or in streams
* Easy-to-use report creation tools
* Build on Google's data processing knowhow
* Easy to use visualisation tools including Data Studio and Looker

## Benefits

* Focus on analytics not operational support
* Data analytics that grows with you, however much you need
* GCP is highly scalable
* Highly available
* High performance
* Industry standard interfaces
* Easy-to-use stream processing tools
* Easily create and share reports
* Benefit from Google's data processing innovation via Netpremacy resale
* Interactive analytics tools which enable faster time-to-insight

**Google Big Data Services**

GCP offers a wide variety of big data services you can use to manage and analyze your data, including:

**Google Cloud BigQuery**

BigQuery lets you store and query datasets holding massive amounts of data. The service uses a table structure, supports SQL, and integrates seamlessly with all GCP services. You can use BigQuery for both batch processing and streaming. This service is ideal for offline analytics and interactive querying.

**Google Cloud Dataflow**

Dataflow offers serverless batch and stream processing. You can create your own management and analysis pipelines, and Dataflow will automatically manage your resources. The service can integrate with GCP services like BigQuery and third-party solutions like Apache Spark.

**Google Cloud Dataproc**

Dataproc lets you integrate your open source stack and streamline your process with automation. This is a fully managed service that can help you query and stream your data, using resources like Apache Hadoop in the GCP cloud. You can integrate Dataproc with other GCP services like Bigtable.

**Google Cloud Pub/Sub**

Pub/Sub is an asynchronous messaging service that manages the communication between different applications. Pub/Sub is typically used for stream analytics pipelines. You can integrate Pub/Sub with systems on or off GCP, and perform general event data ingestion and actions related to distribution patterns.

**Google Cloud Composer**

Composer is a fully-managed cloud-based workflow orchestration service based on Apache Airflow. You can use Composer to manage data processing across several platforms and create your own hybrid environment. Composer lets you define the process using Python. The service then automates processing jobs, like ETL.

**Google Cloud Data Fusion**

Data Fusion is a fully-managed data integration service that enables stakeholders of various skill levels to prepare, transfer, and transform data. Data Fusion lets you create code-free ETL/ELT data pipelines using a point-and-click visual interface. Data Fusion is an open source project that provides the portability needed to work with hybrid and multicloud integrations.

**Google Cloud Bigtable**

Bigtable is a fully-managed NoSQL database service built to provide high performance for big data workloads. Bigtable runs on a low-latency storage stack, supports the open-source HBase API, and is available globally. The service is ideal for time-series, financial, marketing, graph data, and IoT. It powers core Google services, including Analytics, Search, Gmail, and Maps.

**Google Cloud Data Catalog**

Data Catalog offers data discovery capabilities you can use to capture business and technical metadata. To easily locate data assets, you can use schematized tags and build a customized catalog. To protect your data, the service uses access-level controls.

To classify sensitive information, the service integrates with Google Cloud Data Loss Prevention.

**Architecture for Large Scale Big Data Processing on Google Cloud**

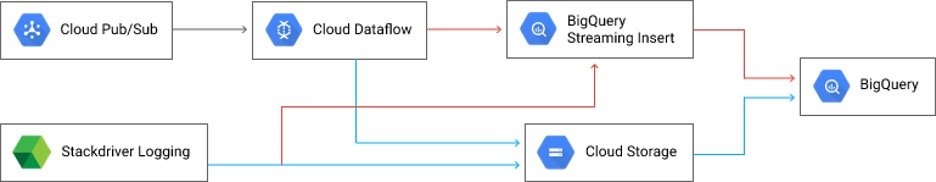
Google provides a [reference architecture](https://cloud.google.com/solutions/architecture/optimized-large-scale-analytics-ingestion) for large-scale analytics on Google Cloud, with more than 100,000 events per second or over 100 MB streamed per second. The architecture is based on Google BigQuery.

Google recommends building a big data architecture with hot paths and cold paths. A **hot path** is a data stream that requires near-real-time processing, while a **cold path** is a data stream that can be processed after a short delay.

This has several advantages, including:

* Ability to store logs for all events without exceeding quotas
* Reduced costs, because only some events need to be handled as streaming inserts (which are more expensive)

The following diagram illustrates the architecture.



Data originates from two possible sources—analytics events published to Cloud Pub/Sub, and logs from Google Stackdriver Logging. Data is divided into two paths:

* The hot path (red arrows) feeds into BigQuery using a streaming insert, to enable continuous data flow
* The cold path (blue arrows) feeds into Google Cloud Storage and from there, loaded in batches to BigQuery.

**GCP Big Data Best Practices**

Here are a few best practices that will help you make more of key Google Cloud big data services, such as Cloud Pub/Sub and Google BigQuery.

**Data Ingestion and Collection**

Ingesting data is a commonly overlooked part of big data projects. There are several options for ingesting data on Google Cloud:

* **Using APIs on the data provider**—pull data from APIs at scale using Compute Engine instances (virtual machines) or Kubernetes
* **Real time streaming**—best with Cloud Pub/Sub
* **Large volume of data on-premises**—most suitable for Google transfer appliance or GCP Online Transfer, depending on volume
* **Large volume of data on other cloud providers**—use Cloud Storage Transfer Service

**Streaming Insert**

If you need to stream and process data in near-real time, you’ll need to use streaming inserts. A streaming insert writes data to BigQuery and queries it without requiring a load job, which can incur a delay. You can perform a streaming insert on a BigQuery table using the Cloud SDK or Google Dataflow.

Note that it takes a few seconds for streaming data to become available for querying. After data is ingested using streaming insert, it takes up to 90 minutes for it to be available for operations like copy and export.

**Use Nested Tables**

You can nest records within tables to create efficiencies in Google BigQuery. For example, if you are processing invoices, the individual lines inside the invoice can be stored as an inner table. The outer table can contain data about the invoice as a whole (for example, the total invoice amount).

This way, if you only need to process data about invoices, and not individual invoice lines, you can run a query only on the outer table to save costs and improve performance. Google only accessed items in the inner table when the query explicitly refers to them.

**Big Data Resource Management**

In many big data projects, you will need to grant access to certain resources to members of your team, other teams, partners or customers. Google Cloud Platform uses the concept of “resource containers”. A container is a grouping of GCP resources that can be dedicated to a specific organization or project.

It is best to define a project for each big data model or dataset. Bring all the relevant resources, including storage, compute, and analytics or machine learning components, into the project container. This will allow you to more easily manage permissions, billing and security.

**How Does Google BigQuery Work?**

BigQuery runs on a serverless architecture that separates storage and compute and lets you scale each resource independently, on demand. The service lets you easily analyze data using Standard SQL.

When using BigQuery, you can run compute resources as needed and significantly cut down on overall costs. You also do not need to perform database operations or system engineering tasks, because BigQuery manages this layer.

**What is BigQuery BI Engine?**

BI Engine performs fast in-memory analyses stored in BigQuery. BI Engine offers sub-second query response time and with high concurrency.

You can integrate BI Engine with tools like Google Data Studio and accelerate your data exploration and analysis jobs. Once integrated, you can use Data Studio to create interactive dashboards and reports without compromising scale, performance, or security.