Table of Contents

Big-Table (Analytical Workload)

- Ideal solution for storing ***time-series data.***
- Ideal solution to provide low latency and high throughput data-processing option with analytics.
- ***HBase managed service*** alternative on google cloud.
- It's regional and not relational data service.
- To get ***good*** performance it's essential to ***design a schema*** that makes it possible to distribute ***reads and writes evenly*** across each table.
- ***Storage*** for clusters ***cannot*** be updated.

Big-Query (Analytical Workload)

Three types of resources available in BigQuery is ***organizations, projects and datasets.***

- It can export Avro data natively to Cloud Storage.
- Provides ***99.9%*** SLA.
- It does ***not facilitate*** direct data load from ***cloud SQL***.
- Access ***can only*** be controlled on ***Datasets and Views***.
- Ideal stack to handle ***IoT*** Data.
- Caching is in ***report settings.***
- Column types ***cannot*** be changed. *[Error in update operation]*
- Dataset ***location cannot be changed*** once it is created.
- It charges only for ***Storage, Queries and Streaming inserts***.
- ***Loading and Exporting*** are ***free*** operations.
    - Provides ***two metrics*** for slots : ***Slot Allocated*** and ***Slot Available***

Data Flow

- Allows access to create and work on dataflow pipeline.
- Denies the access to view data maintaining privacy.
- Always a ***pull*** end-point with cloud Pub/Sub.
- DataFlow pipeline can be stopped using ***Drain Option***. It would stop new processing but allow existing processing to complete.
- ***PCollections*** is cloud DataFlow pipeline. Cloud runner enables pipeline to scale production.
- It helps in ***ordering the data*** received from cloud pub/sub.
- It allows ***updates*** to an ***existing pipeline***.

DataLab

- Cloud Data lab provides a powerful interactive, scalable tool on Google Cloud with the ability to analyze, visualize data.

DataProc

- DataProc has a ***BigQuery connector library*** which allows it directly interface with BigQuery. (BigQuery connector library)
- It handles spark and Hadoop jobs. ***(Spark and high-memory machines needs standard modes).***

DataPrep

- Ability to ***detect, clean and transform data*** through graphical interface ***without any programming knowledge.***
- Helps in visually exploring, cleaning and preparing structured and unstructured data for analysis.

- Automatically identifies ***data anomalies***.
- It can be used to handle ***schema changes*** by Data Analysts ***without any programming knowledge***, but through an easy to use GUI.

Cloud Composer

- Single interface to ***manage*** and ***monitor*** the jobs.
- It allows us to focus on ***authorizing, Scheduling, Monitoring*** as opposed to provisioning resources.
- Helps in completion of ***interdependent jobs***.
- Help ***create*** workflows that connect ***data, processing, and services*** across clouds, giving you a ***unified data environment***.

Cloud Datastore (Transactional workload)

- Highly scalable ***NoSQL database*** for web and mobile applications.
- Provides ***transactional data service***.
- Fully managed with ***NoOps*** required.

Cloud Spanner (Transactional workload)

- Horizontal scaling.
- Low latency.

Cloud Storage

- It provides ***long-term archival*** option.
- Does ***not*** provide ***SQL*** interface.
- Qualities:
    - Direct access
    - HDFS compatibility
    - Interoperability

- Data accessibility
- High data availability
- No Storage management overhead ***(No routine maintenance***)
- Quick startup

Transfer Appliance

- ***Huge data*** can be transferred in time and cost-effective way.
- ***Rehydration*** of data.
- One-way one-time migration.
- Most cost-effective.

Pub/Sub

- Helps in handling streaming data but does not handle the ***ordering*** of the same.
- Provides elastic and scalable ingestions.

Imports and exports of data

1) Cloud SQL ➡ Big-Query
    - Export to Cloud storage, then import to Big-Query.
2) Collect data via IoT. Process store analyze data in real-time.

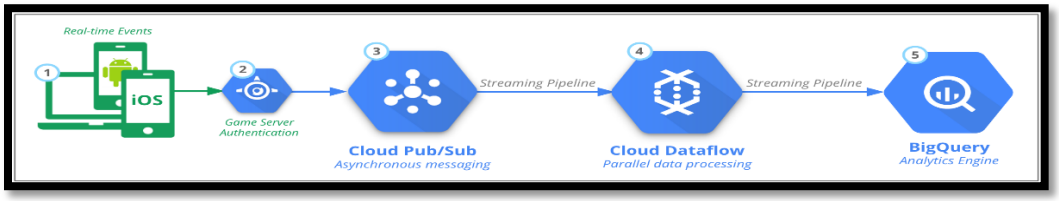


Figure 1System design

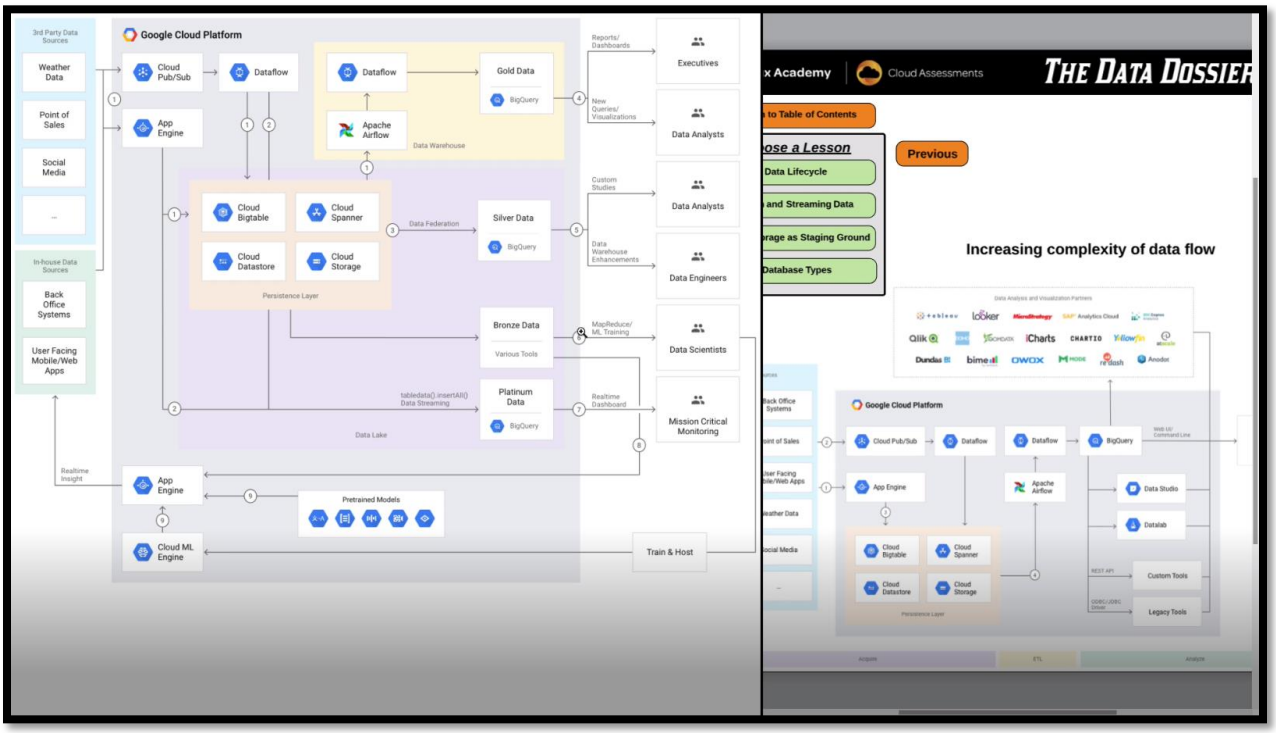
3) Avro format data transfer from BigQuery to Cloud Storage (using web console) Big-Query ➡ Cloud Storage
    - Export table to BigQuery and then provide Cloud Storage location to export to.

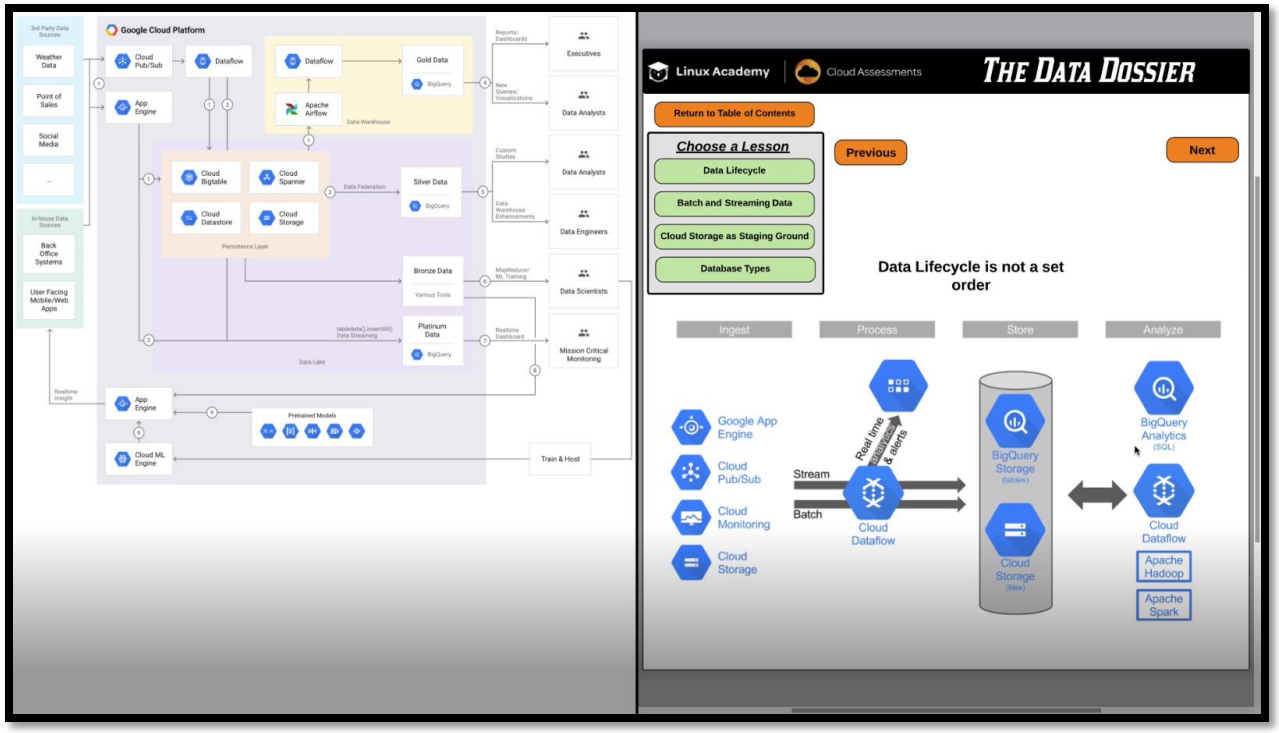
Best Practices

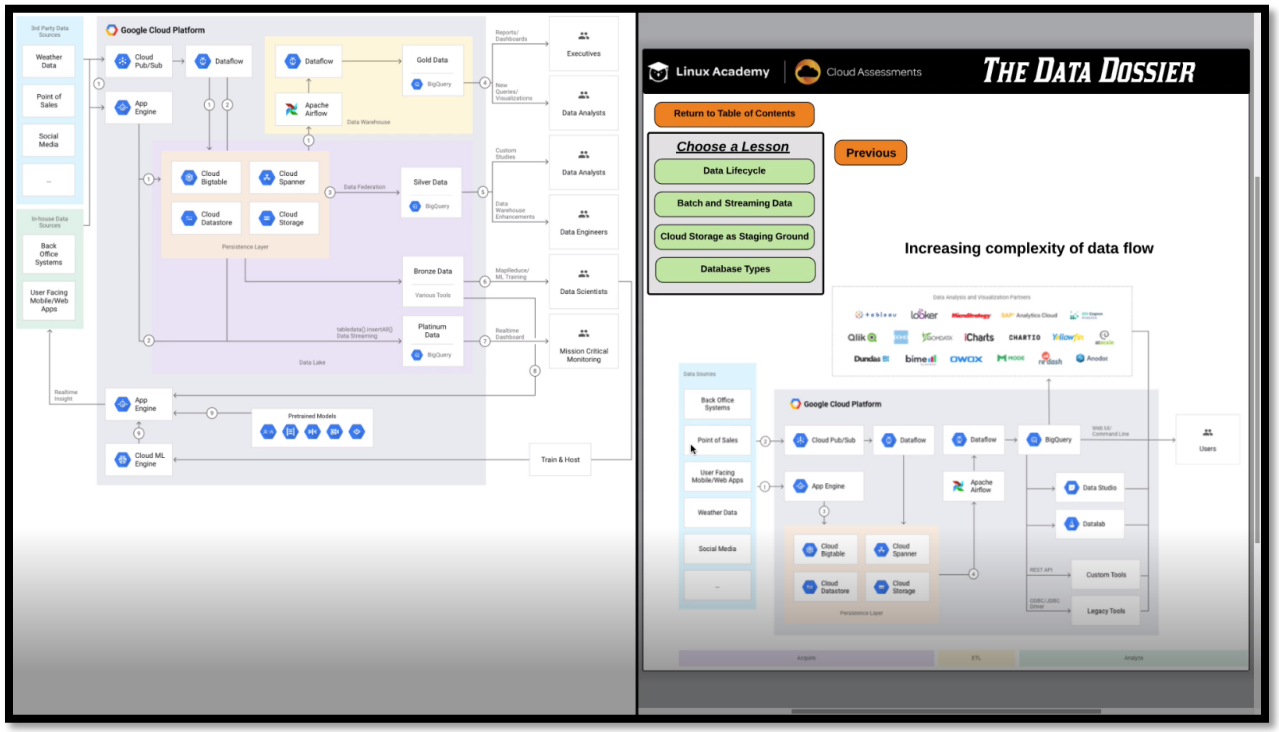
A. Best way to limit and expose number of columns and access is to create a View. (BigQuery).

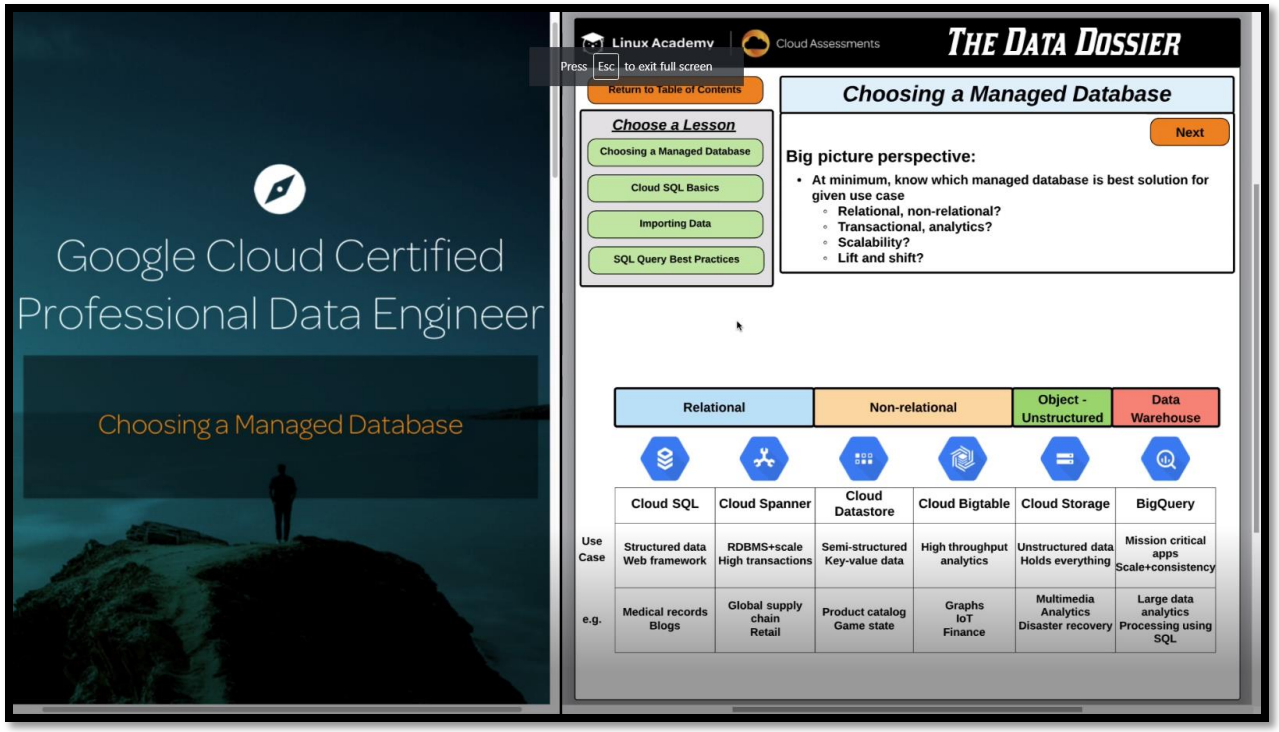
B. Table name should include a * for the wildcard and it must be enclosed in back-tick characters. (Ex: `bigquery-public-data.noaa_gsod.gsod*`).

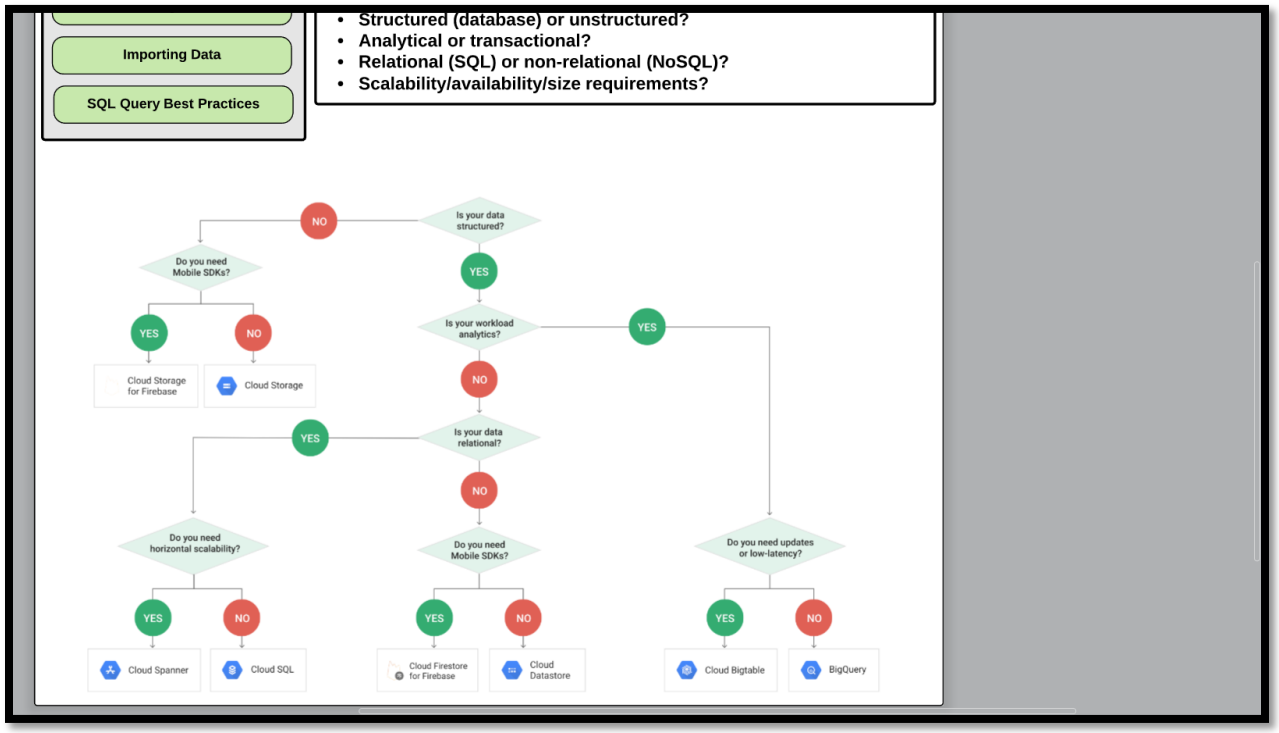
C. When we require to reuse Hadoop jobs with minimizing the infrastructure management with the ability to store data in a durable external storage, ***Dataproc with Cloud Storage*** would be an ideal solution.




Google Cloud Platform
Weather Data
Point of Sales
Social Media
Back Office Systems
User Facing Mobile/Web Apps
Cloud Pub/Sub
Dataflow
App Engine
Cloud Bigtable
Cloud Spanner
Cloud Datastore
Cloud Storage
Gold Data
Apache Airflow
Silver Data
Bronze Data
Platinum Data
Pretrained Models
Cloud ML Engine
Train & Host
Executives
Data Analysts
Data Engineers
Data Scientists
Mission Critical Monitoring
Linux Academy
Cloud Assessments
THE DATA DOSSIER
Return to Table of Contents
Choose a Lesson
Data Lifecycle
Batch and Streaming Data
Cloud Storage as Staging Ground
Database Types
Previous
Next
Data Lifecycle is not a set order
Ingest
Process
Store
Analyze
Google App Engine
Cloud Pub/Sub
Cloud Monitoring
Cloud Storage
Stream
Batch
Real time analytics & alerts
Cloud Dataflow
BigQuery Storage (tables)
Cloud Storage (files)
BigQuery Analytics (SQL)
Cloud Dataflow
Apache Hadoop
Apache Spark


Google Cloud Platform
Weather Data
Point of Sales
Social Media
Back Office Systems
User Facing Mobile/Web Apps
Cloud Pub/Sub
Dataflow
App Engine
Cloud Bigtable
Cloud Spanner
Cloud Datastore
Cloud Storage
Gold Data
Apache Airflow
Silver Data
Bronze Data
Platinum Data
Pretrained Models
Cloud ML Engine
Train & Host
Executives
Data Analysts
Data Engineers
Data Scientists
Mission Critical Monitoring
Linux Academy
Cloud Assessments
THE DATA DOSSIER
Return to Table of Contents
Choose a Lesson
Data Lifecycle
Batch and Streaming Data
Cloud Storage as Staging Ground
Database Types
Previous
Increasing complexity of data flow
Data Analysis and Visualization Partners
Data Sources
Back Office Systems
Point of Sales
User Facing Mobile/Web Apps
Weather Data
Social Media
Google Cloud Platform
Cloud Pub/Sub
Dataflow
Dataflow
BigQuery
App Engine
Apache Airflow
Data Studio
Datalab
Cloud Bigtable
Cloud Spanner
Cloud Datastore
Cloud Storage
Custom Tools
Legacy Tools
Users


Google Cloud Certified
Professional Data Engineer
Choosing a Managed Database
Linux Academy
Cloud Assessments
THE DATA DOSSIER
Press Esc to exit full screen
Return to Table of Contents
Choose a Lesson
Choosing a Managed Database
Cloud SQL Basics
Importing Data
SQL Query Best Practices
Choosing a Managed Database
Next
Big picture perspective:
At minimum, know which managed database is best solution for given use case
Relational, non-relational?
Transactional, analytics?
Scalability?
Lift and shift?
Relational
Non-relational
Object - Unstructured
Data Warehouse
Cloud SQL
Cloud Spanner
Cloud Datastore
Cloud Bigtable
Cloud Storage
BigQuery
Use Case
Structured data Web framework
RDBMS+scale High transactions
Semi-structured Key-value data
High throughput analytics
Unstructured data Holds everything
Mission critical apps Scale+consistency
e.g.
Medical records Blogs
Global supply chain Retail
Product catalog Game state
Graphs IoT Finance
Multimedia Analytics Disaster recovery
Large data analytics Processing using SQL


Importing Data
SQL Query Best Practices
Structured (database) or unstructured?
Analytical or transactional?
Relational (SQL) or non-relational (NoSQL)?
Scalability/availability/size requirements?
Is your data structured?
NO
YES
Do you need Mobile SDKs?
YES
NO
Cloud Storage for Firebase
Cloud Storage
Is your workload analytics?
YES
NO
Is your data relational?
YES
NO
Do you need horizontal scalability?
YES
NO
Cloud Spanner
Cloud SQL
Do you need Mobile SDKs?
YES
NO
Cloud Firestore for Firebase
Cloud Datastore
Do you need updates or low-latency?
YES
NO
Cloud Bigtable
BigQuery