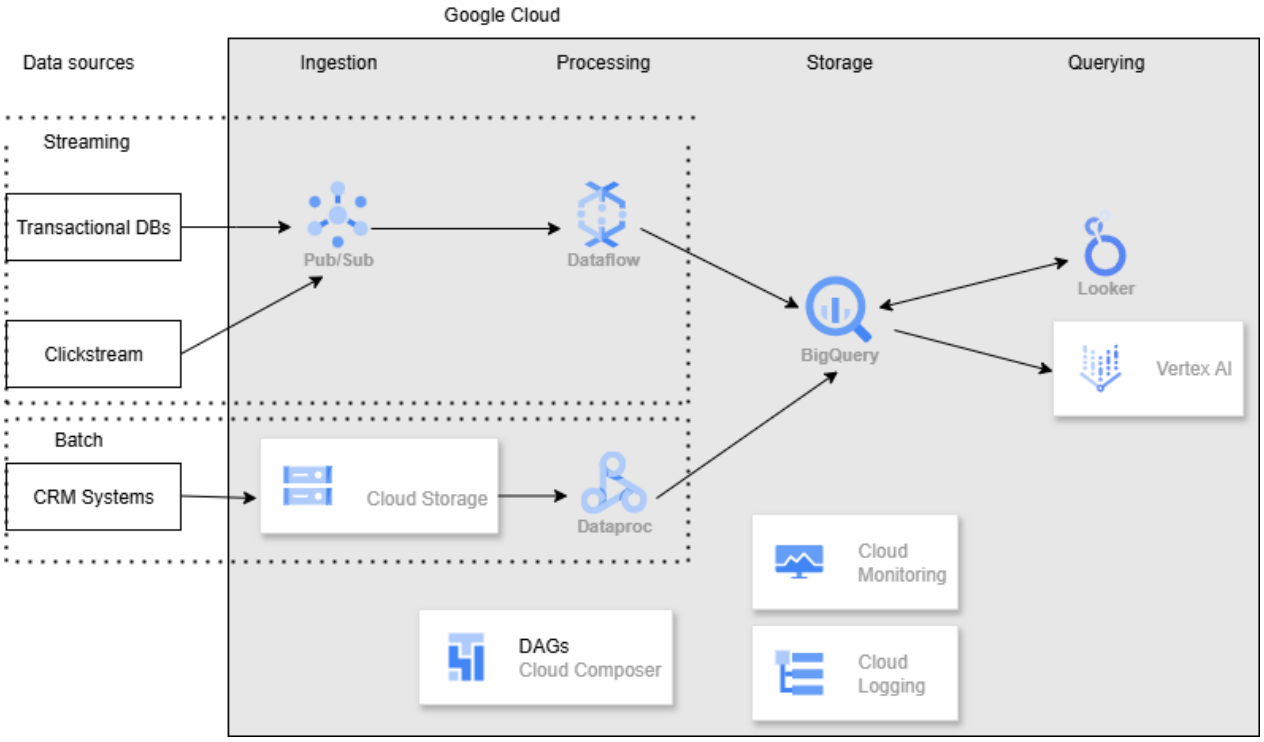

The diagram above shows high-level architecture for an e-commerce platform. It includes three main types of data sources:

- Transactional data from OLTP systems (orders, inventory, user profile data, payments);
- Clickstream data (webpage views, clicks, session time, time spent on page);
- CRM data (marketing campaigns, interactions with customer support).

Streaming is used for transactional data to reflect real-time changes to a user profile, basket, or payments. Clickstream data can be near-real time and can be fed to ML recommendation models to display products and services similar to what the user is browsing. Batch processing can be used for CRM data to plan future marketing campaigns.

A combination of PubSub and Dataflow is used to ingest and process/transform streaming data. Cloud Composer DAGs are used to schedule and trigger Dataflow pipelines and data movement to BigQuery.

CRM batch data is pulled via API daily. The extracted files land in Cloud Storage buckets, which is scheduled and managed by Cloud Composer. Dataproc is used to clean and transform the data, which is later written to BigQuery. Dataflow can be used instead of Dataproc as both services can handle large batch processing.

Data analysts and data scientists are granted viewer access to BigQuery datasets and tables, while Looker is used for further data analysis and visualization, and Vertex AI enables ML engineers to explore the data and build machine learning models.

Cloud Monitoring and Logging can be enabled for ingestion and processing stages, and logs can be stored in Cloud Storage for further analysis. When data lands in BigQuery, the event can also be written to a separate monitoring table (timestamps, number of records, file names, etc.)

Scalability, cost-efficiency, and security

PubSub can handle millions of messages per second with low latency. It ensures message delivery at least once and the price is based on message delivery volume. Data is encrypted in transit and at rest, and IAM ensure controlled access to topics.

Dataflow can autoscale up and down based on data volume, which ensures no overprovisioning of resources. Security is ensured via assigned worker roles and permissions.

Dataproc can also autoscale, uses IAM assigned controls and can run within private networks.

Cloud Composer offers scalability through managed Apache Airflow on GKE. It ensures cost-efficiency by autoscaling DAG execution and reducing manual overhead, while maintaining security with VPC support, IAM roles, and encrypted metadata storage.

BigQuery pricing is comprised of storage and compute costs, which can be reduced by adding partitioning and clustering, querying only the data you need, limiting the number of columns, and using less computationally heavy functions. BigQuery allows for fine-grained access control (row/table level), data masking, and column-level security.

Cloud Storage can handle raw data and support various file formats. It offers tiered pricing based on access frequency. It provides managed IAM access control per bucket and object.