

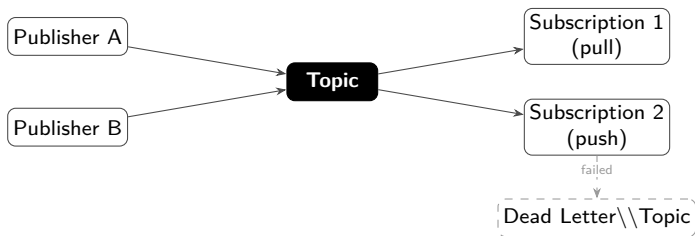


Data Ingestion & Processing

GCP Professional Data Engineer | Pub/Sub, Dataflow, Dataproc & Processing Patterns

Pub/Sub: Messaging & Streaming Ingestion

Cloud Pub/Sub is a fully managed, real-time messaging service for decoupling producers and consumers. It guarantees at-least-once delivery, scales automatically, and supports both push and pull delivery modes.



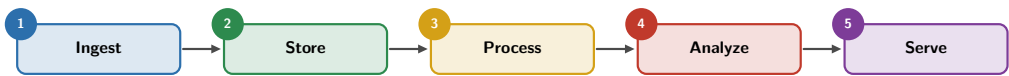
Pub/Sub Architecture

Pull subscription	Consumer polls for messages. Ack within deadline (10-600s). Best for: Dataflow, batch consumers, variable processing time.
Push subscription	Pub/Sub sends HTTP POST to endpoint. Best for: Cloud Run, Cloud Functions, webhooks. Automatic retry with back-off.
BigQuery subscription	Direct write from Pub/Sub to BigQuery. No Dataflow needed. Schema auto-detection or manual. Best for simple streaming to warehouse.
Cloud Storage subscription	Write messages as files to GCS. Configurable batch size and duration. Good for data lake ingestion.
Ordering	Enable ordering key for ordered delivery within a key. Messages with same key go to same subscriber.
Exactly-once delivery	Available with pull subscriptions. Prevents redelivery of acked messages. Use for financial/transactional data.
Dead letter topic	Messages that fail after max delivery attempts are forwarded to dead letter topic for investigation.
Message retention	Default 7 days (configurable 10 min to 31 days). Replay: seek to timestamp or snapshot for reprocessing.

7 days

Default Pub/Sub retention

Configurable from 10 minutes to 31 days



Data Pipeline Stages

Pub/Sub
Async messaging

Dataflow
Beam runner

Dataproc
Spark/Hadoop

Cloud Composer
Airflow DAGs

Dataflow: Unified Batch & Stream Processing

Cloud Dataflow is a fully managed service for Apache Beam pipelines. It handles both batch and streaming data with the same programming model. Autoscaling, exactly-once processing, and tight GCP integration.

Apache Beam model	PCollection (data), PTransform (operation), Pipeline (DAG of transforms), Runner (Dataflow, Spark, Flink).
Windowing	Fixed (tumbling), Sliding (overlapping), Session (activity-based gaps), Global (all elements). Defines grouping for aggregation.
Triggers	When to emit window results: AfterWatermark (event time), AfterProcessingTime, AfterCount, Repeatedly. Combine for complex timing.
Watermarks	Track event time progress. Estimate when all data for a window has arrived. Late data handled by allowed lateness.
Exactly-once	Streaming: exactly-once per-record semantics with checkpointing. Batch: inherently exactly-once.

Key Beam Transforms

Transform	Purpose	Example
ParDo	Element-wise processing	Parse, filter, enrich records
GroupByKey	Group by key	Aggregate per user/session

CoGroupByKey	Join two PCollections	Enrich stream with side input
Combine	Reduce/aggregate	Sum, mean, count per window
Flatten	Merge PCollections	Union multiple sources
Partition	Split into N outputs	Route by category

- Dataflow templates: reusable, parameterized pipelines (Google-provided and custom)
- Flex templates: containerized pipelines for custom dependencies
- Dataflow Prime: enhanced autoscaling with right-fitting worker sizes
- Streaming Engine: offload state/shuffling from workers to managed service (faster, cheaper)

Dataproc: Managed Spark/Hadoop

Cluster types	Standard (1 master), High Availability (3 masters), Single Node (dev/test). Preemptible workers for cost savings.
Dataproc Serverless	Run Spark batch jobs without managing clusters. Auto-scaling, pay per use. Best for ad-hoc Spark SQL/PySpark.
Dataproc on GKE	Run Spark on existing GKE clusters. Share K8s resources. Good for mixed workloads (Spark + other containers).
Component Gateway	Secure access to Spark UI, Jupyter, Zeppelin without SSH tunnel or firewall rules.
Initialization actions	Scripts that run on cluster startup. Install custom packages, configure Hadoop. Store in GCS.
Ephemeral clusters	Create cluster, run job, delete cluster. Use Cloud Composer or Workflows to orchestrate. Cost-efficient for batch.

Dataflow vs Dataproc

Criteria	Dataflow	Dataproc
Best for	New pipelines, streaming	Existing Spark/Hadoop jobs
Model	Apache Beam	Spark, Hadoop, Hive, Pig
Management	Fully serverless	Managed clusters (or serverless)
Streaming	Native, exactly-once	Spark Structured Streaming
Cost model	Per vCPU-hour + data	Per cluster (VMs)
Ecosystem	Beam connectors	Spark/Hadoop ecosystem

Data Fusion & Other Processing

Cloud Data Fusion	Visual ETL/ELT tool (based on CDAP). Drag-and-drop pipeline builder. Code-free transforms. Runs on Dataproc under the hood.
Dataprep (Trifacta)	Interactive data wrangling. Visual exploration, profiling, cleaning. Generates Dataflow jobs. Good for data analysts.

Dataflow streaming capability ●●●●●

Dataproc Spark ecosystem ●●●●●

Data Fusion ease of use ●●●●●

Datastream real-time CDC ●●●●●

ETL vs ELT

Aspect	ETL	ELT
Transform location	Before loading	After loading (in warehouse)
Best tool	Dataflow, Data Fusion	BigQuery, dbt/Dataform
Use case	Complex transforms, streaming	SQL transforms, BigQuery-centric
Scalability	Scales with pipeline	Scales with warehouse (BQ slots)

Official resources

⚠️ Check the current exam guide

Google Cloud is updating its exams to the products announced at Next '26 (certification hub, checked 2026-10-05). The Dataproc documentation is now titled **Managed Service for Apache Spark** and Vertex AI's is **Gemini Enterprise Agent Platform**: these sheets keep the older names; read the current exam guide for the wording it uses.

EXAM GUIDE 2 links

Professional Data Engineer exam guide

Google Cloud · [services.google.com/fh/files/misc/profession...ata_engineer_exam_guide_english.pdf](#)
domains and topics

FREE

Professional Data Engineer certification page

Google Cloud · [cloud.google.com/learn/certification/data-engineer](#)
format, price, sample questions link, renewal

FREE

VIDEO 1 link

Professional Data Engineer learning path

Google Skills · [skills.google/paths/16](#)
Google Skills (ex Cloud Skills Boost)

FREE

DOCS 4 links

BigQuery documentation

Google Cloud · [cloud.google.com/bigquery/docs](#)

FREE

Dataflow documentation

Google Cloud · [cloud.google.com/dataflow/docs](#)

FREE

Pub/Sub documentation

Google Cloud · [cloud.google.com/pubsub/docs](#)

FREE

Managed Service for Apache Spark (ex Dataproc)

Google Cloud · [docs.cloud.google.com/managed-spark/docs](#)
renamed in 2026

FREE

CHEATSHEET 2 links

gcloud CLI cheat sheet

Google Cloud · [cloud.google.com/sdk/docs/cheatsheet](#)
the commands on one page

FREE

bq command-line tool reference

Google Cloud · [cloud.google.com/bigquery/docs/reference/bq-cli-reference](#)

FREE

9 official links · 9 free, 0 paid · checked 2026-10-05