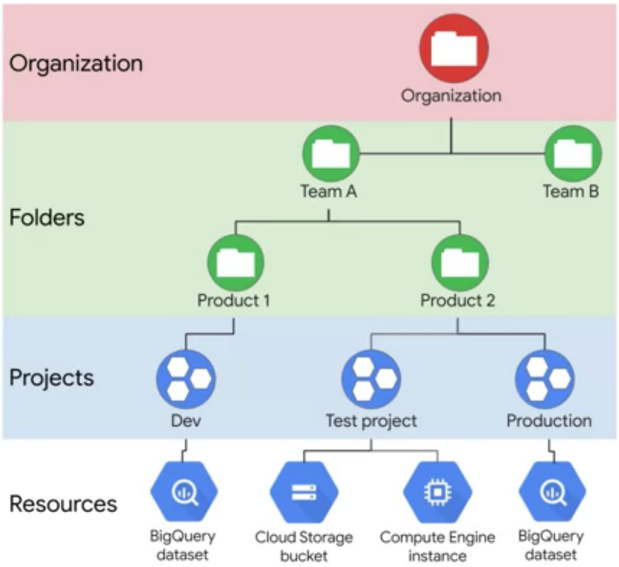
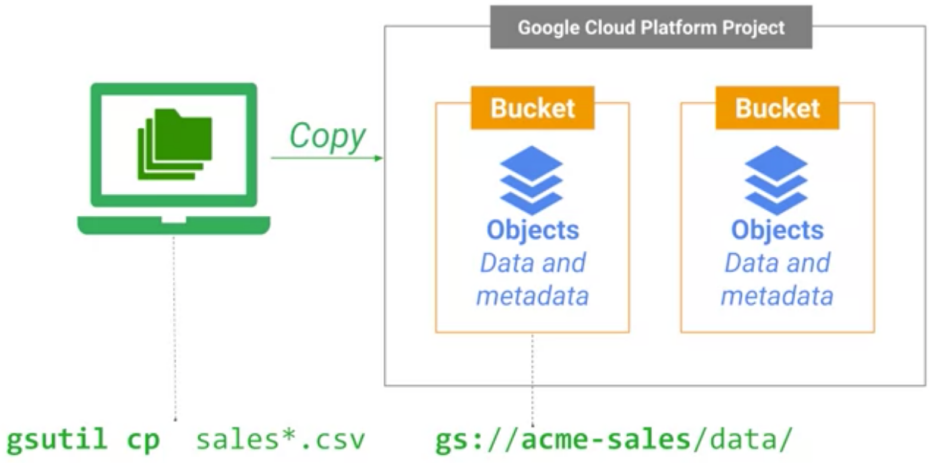


Google Cloud Platform resource hierarchy

Using gsutil to copy existing data into Cloud Storage

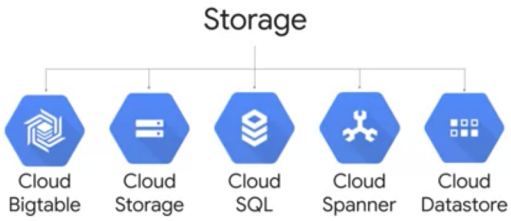


Spotlight: BigQuery granular control over data access

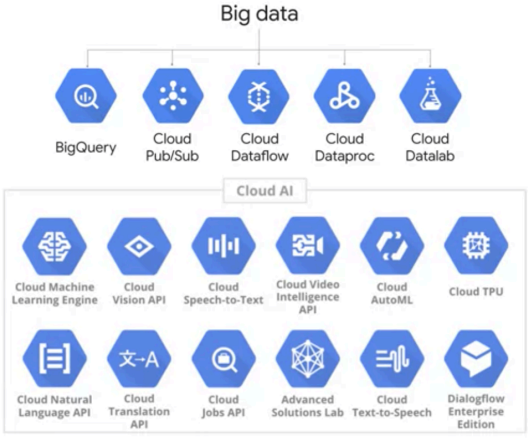
- BigQuery table data encrypted with keys (and those keys are also encrypted)
- Monitor and flag queries for anomalous behavior
- Limit data access with authorized views

GCP offers a range of services

Or use a managed service

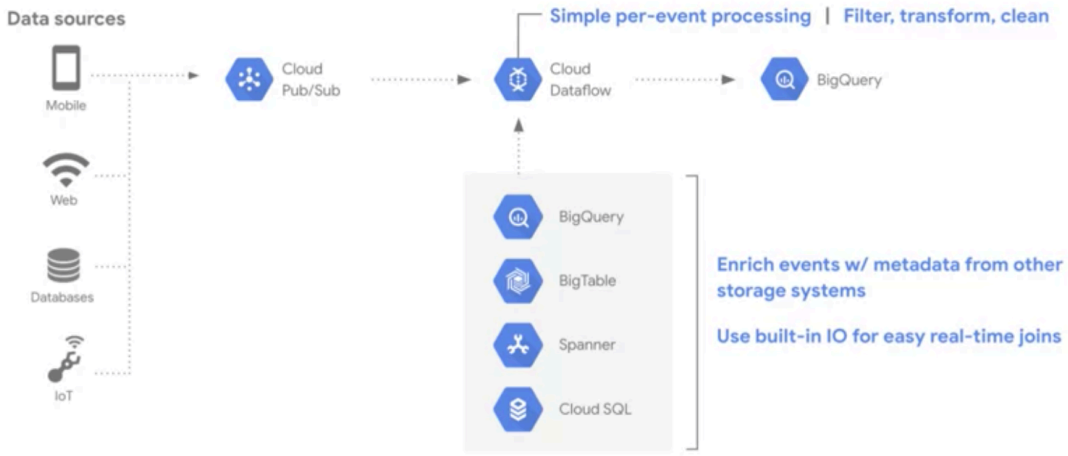


GCP offers a range of services



The suite of big data products on Google Cloud Platform

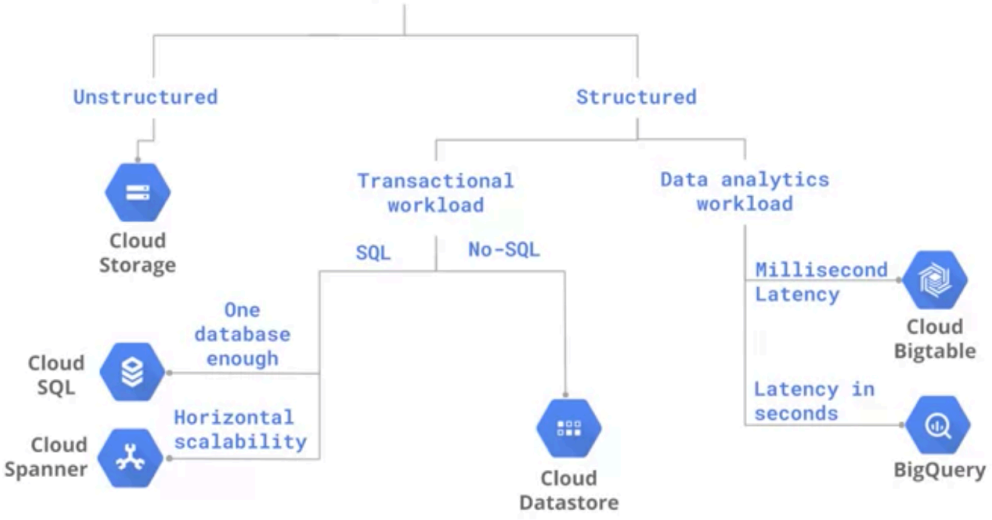
GO-JEK architecture review



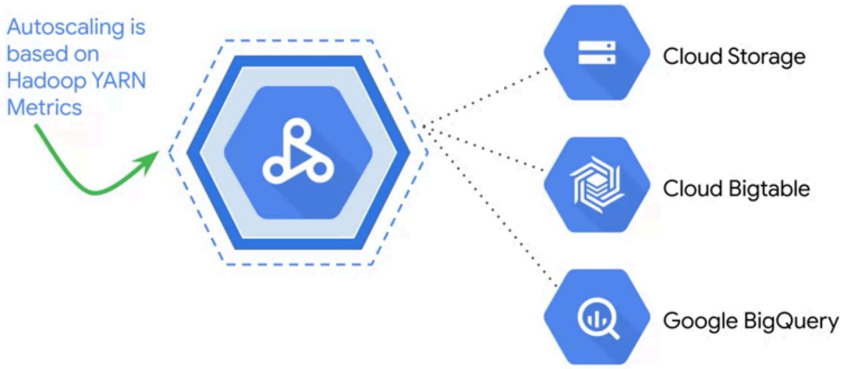
Choose your solutions based on access pattern

	Cloud Storage	Cloud SQL	Datastore	Bigtable	BigQuery
Capacity	Petabytes +	Gigabytes	Terabytes	Petabytes	Petabytes
Access metaphor	Like files in a file system	Relational database	Persistent Hashmap	Key-value(s), HBase API	Data warehouse
Read	Have to copy to local disk	SELECT rows	filter objects on property	scan rows	SELECT rows
Write	One file	INSERT row	put object	put row	Batch/stream
Update granularity	An object (a "file")	Field	Attribute	Row	Field
Usage	Store blobs	No-ops SQL database on the cloud	Structured data from AppEngine apps	No-ops, high throughput, scalable, flattened data	Interactive SQL* querying fully managed warehouse

If your data is....



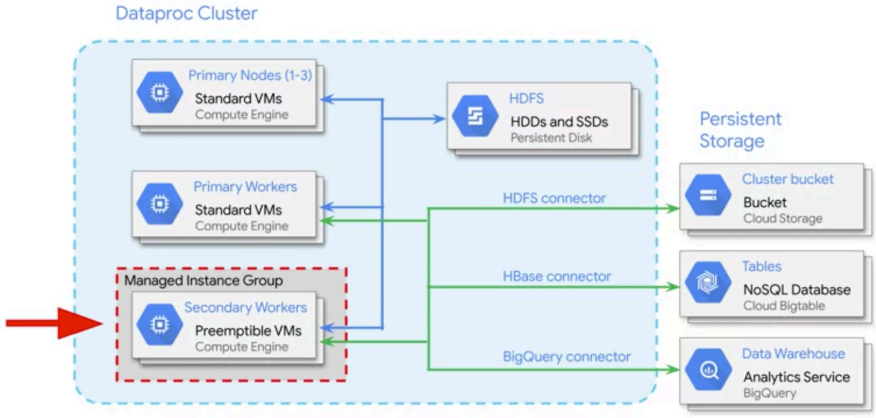
Cloud Dataproc Autoscaling (alpha) provides flexible capability



Dataproc

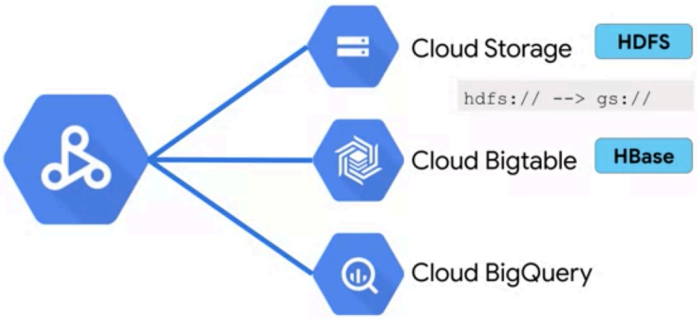
- Hadoop without cluster management
- Lift-and-shift existing Hadoop workloads
- Connect with Cloud Storage to separate compute and storage
- Re-size clusters effortlessly. Preemptible VMs for cost savings

Utilize PVMs to significantly reduce costs for fault-tolerant workloads



Off-cluster storage is the gateway to efficiency

More utilization options are available if persistent data is stored off-cluster

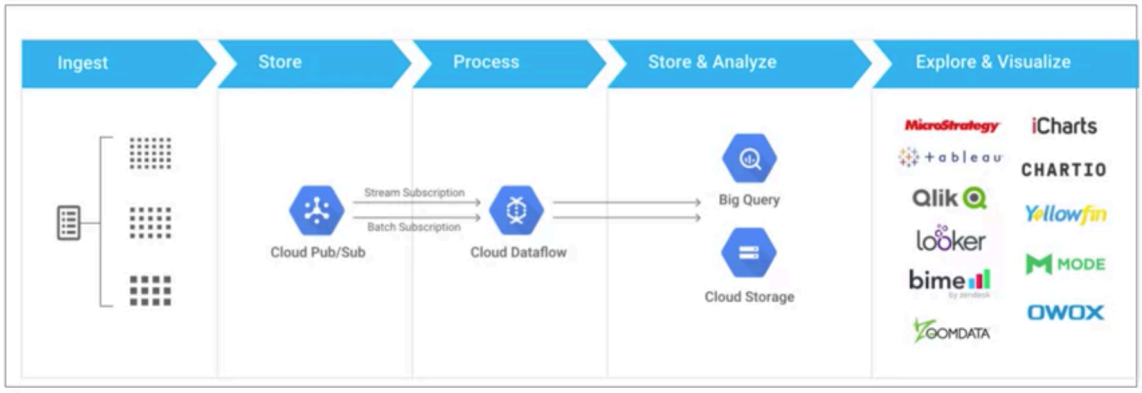


Pub/Sub offers reliable, real-time messaging

Distributed Messaging with Pub/Sub

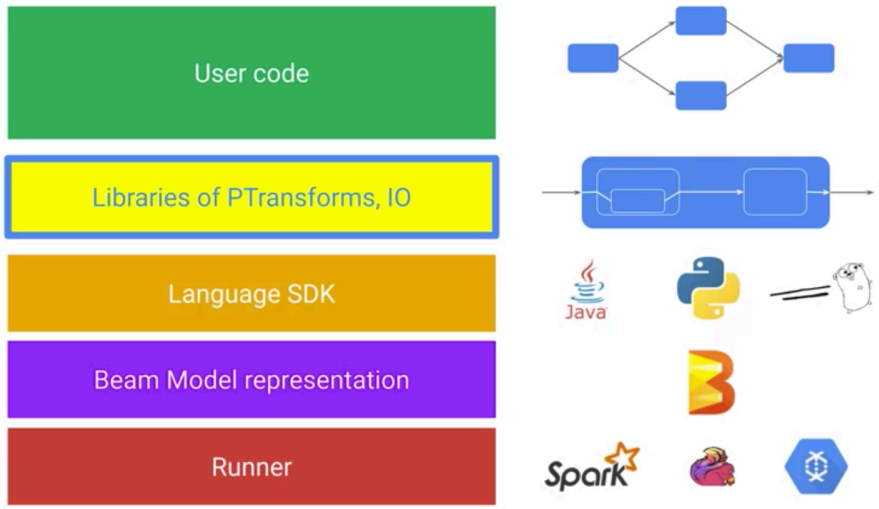
- At-least-once delivery
- No provisioning, auto-everything
- Open APIs
- Global by default
- End-to-end encryption

Google Cloud Serverless Big Data Pipeline

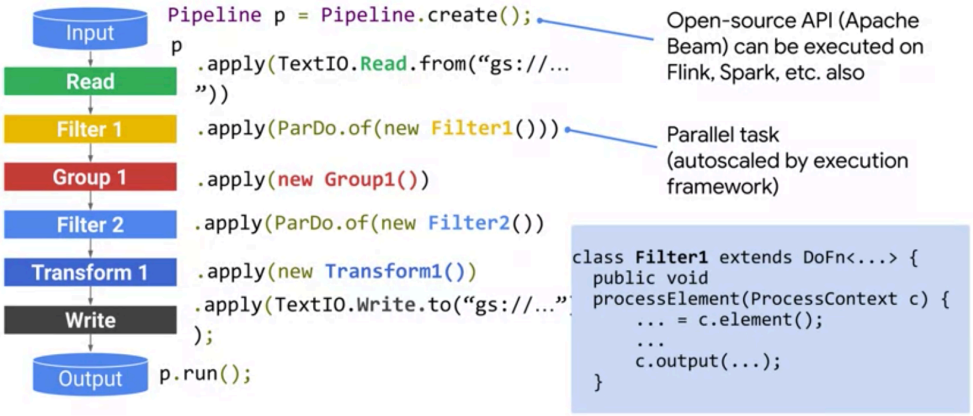


Why Apache Beam?

- **Unified** - Use a single programming model for both batch and streaming use cases
- **Portable** - Execute pipelines on multiple execution environments
- **Extensible** - Write and share new SDKs, IO connectors, and transformation libraries



Dataflow offers NoOps data pipelines



Cloud Dataflow

- Serverless, fully managed data processing
- Unified batch and streaming processing + autoscale
- Open source programming model using  beam
- Intelligently scales to millions of QPS