



# The Data Dossier

## Choose a Section

[Introduction](#)[Case Studies](#)[Foundational Concepts](#)[Cloud SQL](#)[Cloud Datastore](#)[Cloud Bigtable](#)[Cloud Spanner](#)[Real Time Messaging with Cloud Pub/Sub](#)[Data Pipelines with Cloud Dataflow](#)[Cloud Dataproc](#)[BigQuery](#)[Machine Learning Concepts](#)[Cloud ML Engine](#)[Pre-trained ML API's](#)[Cloud Datalab](#)[Cloud Dataprep](#)[Data Studio](#)[Additional Study Resources](#)

[Return to Table of Contents](#)

## **Choose a Lesson**

[What is a Data Engineer?](#)[Exam and Course Overview](#)

[Return to Table of Contents](#)

## Choose a Lesson

[What is a Data Engineer?](#)[Exam and Course Overview](#)

## *What is a Data Engineer?*

### Google's definition:

A Professional Data Engineer enables data-driven decision making by collecting, transforming, and visualizing data. The Data Engineer designs, builds, maintains, and troubleshoots data processing systems with a particular emphasis on the security, reliability, fault-tolerance, scalability, fidelity, and efficiency of such systems.

The Data Engineer also analyzes data to gain insight into business outcomes, builds statistical models to support decision-making, and creates machine learning models to automate and simplify key business processes.

### What does this include?

- Build data structures and databases:
  - Cloud SQL, Bigtable
- Design data processing systems:
  - Dataproc, Pub/Sub, Dataflow
- Analyze data and enable machine learning:
  - BigQuery, Tensorflow, Cloud ML Engine, ML API's
- Match business requirements with best practices
- Visualize data ("make it look pretty"):
  - Data Studio
- Make it secure and reliable

### Super-simple definition:

Collect, store, manage, transform, and present data to make it useful.

[Return to Table of Contents](#)

## Choose a Lesson

[What is a Data Engineer?](#)[Exam and Course Overview](#)

## *Exam and Course Overview*

[Next](#)

### **Exam format:**

- **50 questions**
- **120 minutes (2 hours)**
- **Case study + individual questions**
- **Mixture of high level, conceptual, and detailed questions:**
  - **How to convert from HDFS to GCS**
  - **Proper Bigtable schema**
- **Compared to the architect exam it is more focused and more detailed:**
  - **Architect exam = 'Mile wide/inch deep'**
  - **Data Engineer exam = 'Half mile wide, 3 inches deep'**

### **Course Focus:**

- **Very broad range of topics**
- **Depth will roughly match exam:**
  - **Plus hands-on examples**

[Return to Table of Contents](#)

## Choose a Lesson

[What is a Data Engineer?](#)[Exam and Course Overview](#)

## *Exam and Course Overview*

[Previous](#)

### **Exam topics:**

- Building data representations
- Data pipelines
- Data processing infrastructure
- Database options - differences between each
- Schema/queries
- Analyzing data
- Machine learning
- Working with business users/requirements
- Data cleansing
- Visualizing data
- Security
- Monitoring pipelines

### **Google Cloud services covered:**

- Cloud Storage
- Compute Engine
- Dataproc
- Bigtable
- Datastore
- Cloud SQL
- Cloud Spanner
- BigQuery
- Tensorflow
- ML Engine
- Managed ML API's - Translate, Speech, Vision, etc.
- Pub/Sub
- Dataflow
- Data Studio
- Dataprep
- Datalab

[Return to Table of Contents](#)

## **Choose a Lesson**

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

## *Flowlogistic Case Study*

[Next](#)

### Link:

<https://cloud.google.com/certification/guides/data-engineer/casestudy-flowlogistic>

### Main themes:

- Transition existing infrastructure to cloud
- Reproduce existing workload ("lift and shift"):
  - First step into cloud transition

### Primary cloud objectives:

- Use proprietary inventory-tracking system:
  - Many IoT devices - high amount of real-time (streaming) data
  - Apache Kafka stack unable to handle data ingest volume
  - Interact with both SQL and NoSQL databases
  - Map to Pub/Sub - Dataflow:
    - Global, scalable
- Hadoop analytics in the cloud:
  - Dataproc - managed Hadoop
  - Different data types
  - Apply analytics/machine learning

### Other technical considerations:

- Emphasis on data ingest:
  - Streaming and batch
- Migrate existing workload to managed services:
  - SQL - Cloud SQL:
    - Cloud Spanner if over 10TB and global availability needed
  - Cassandra - NoSQL (wide-column store) - Bigtable
  - Kafka - Pub/Sub, Dataflow, BigQuery
- Store data in a 'data lake':
  - Further transition once in the cloud
  - Storage = Cloud Storage, Bigtable, BigQuery
  - Migrate from Hadoop File System (HDFS)

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

## Flowlogistic Case Study

[Previous](#)[Next](#)

### Inventory Tracking Data Flow



Tracking Devices



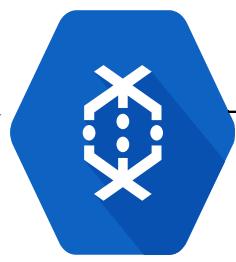
Tracking Devices



Tracking Devices



Tracking Devices



Inventory tracking

Metadata - tracking messages

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X

Pub/Sub is used for streaming (real-time) data ingest. Allows asynchronous (many-to-many) messaging via published and subscribed messages.

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

Cloud Dataflow is a data processing pipeline, transforming both stream and batch data.



[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X

Cloud SQL is a fully managed MySQL and PostgreSQL database. It is a perfect transition step for migrating SQL workloads.

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X  
Cloud Bigtable is a managed, massively scalable non-relational/NoSQL database based on HBase.

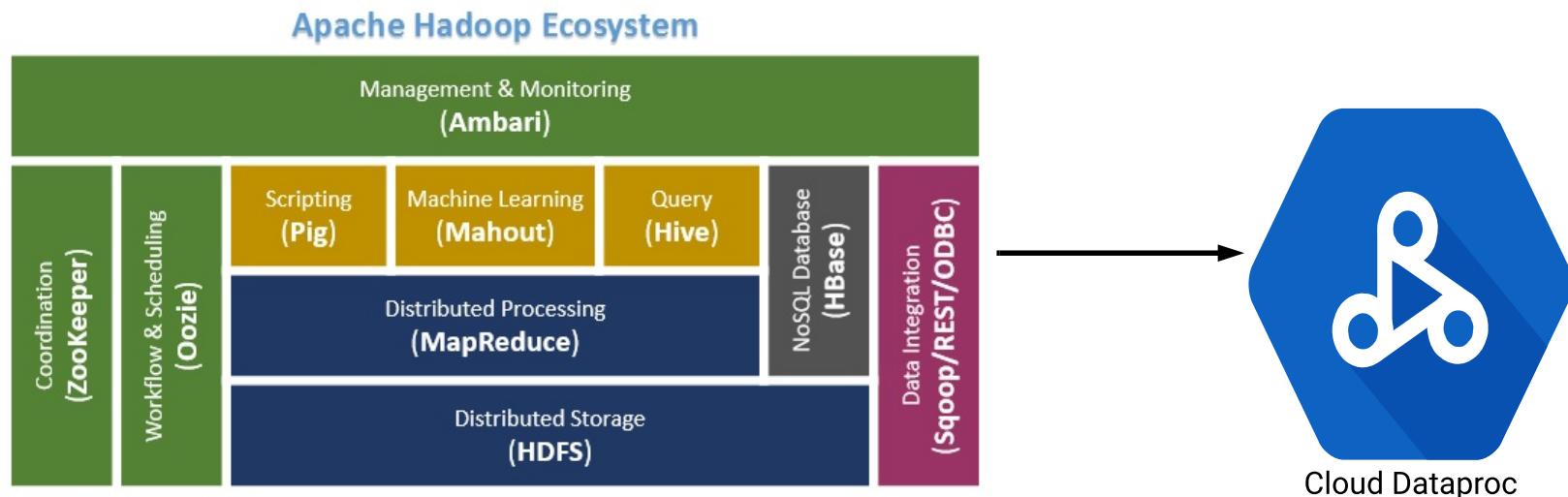
[Return to Table of Contents](#)

## Flowlogistic Case Study

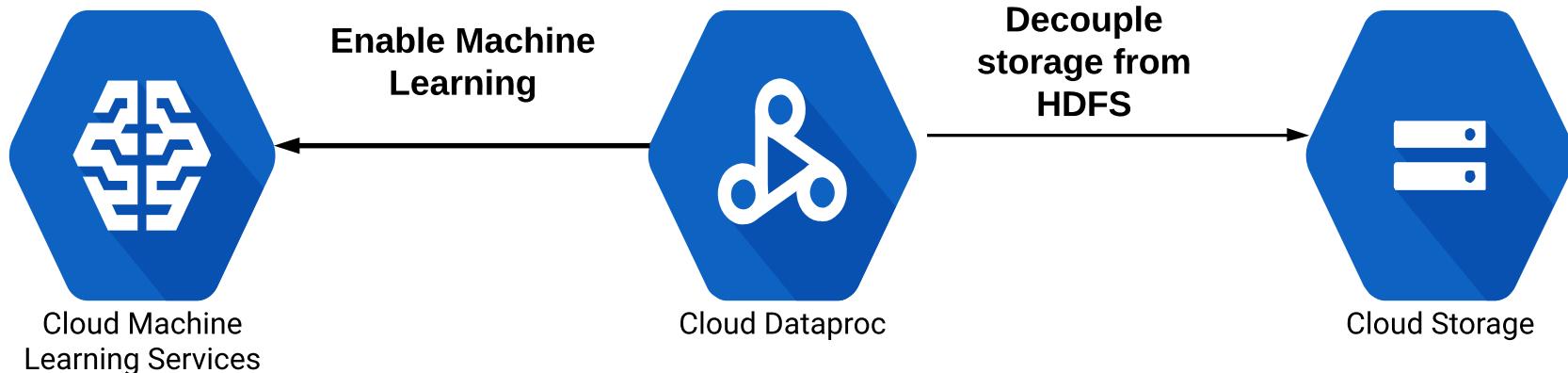
### Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)[Previous](#)

### Phase 1: Initial migration of existing Hadoop analytics



### Phase 2: Integrate other Google Cloud Services



[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X

Cloud Dataproc offers fully managed Apache, Hadoop, and Spark cluster management. It integrates easily with other GCP services.

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X

Managed machine learning service  
for predictive analytics.

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X

Decoupling storage from the Dataproc cluster allows for destroying the cluster when the job is complete as well as widely available, high-performance storage.

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

## **Case Study Overview**

- Exam has 2 possible case studies
- Exam case studies available from Google's training site:  
<https://cloud.google.com/certification/guides/data-engineer>
- Different 'themes' to each case study = insight to possible exam questions
- Very good idea to study case studies in advance!
- Case study format:
  - Company Overview
  - Company Background
  - Solution Concept – current goal
  - Existing Technical Environment – where they are now
  - Requirements – boundaries and measures of success
  - C-level statements – what management cares about

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

## *MJTelco Case Study*

[Next](#)

### Link:

<https://cloud.google.com/certification/guides/data-engineer/casestudy-mjtelco>

### Main themes:

- No legacy infrastructure - fresh approach
- Global data ingest

### Primary Cloud Objectives:

- Accept massive data ingest and processing on a global scale:
  - Need no-ops environment
  - Cloud Pub/Sub accepts input from many hosts, globally
- Use machine learning to improve their topology models

### Other technical considerations:

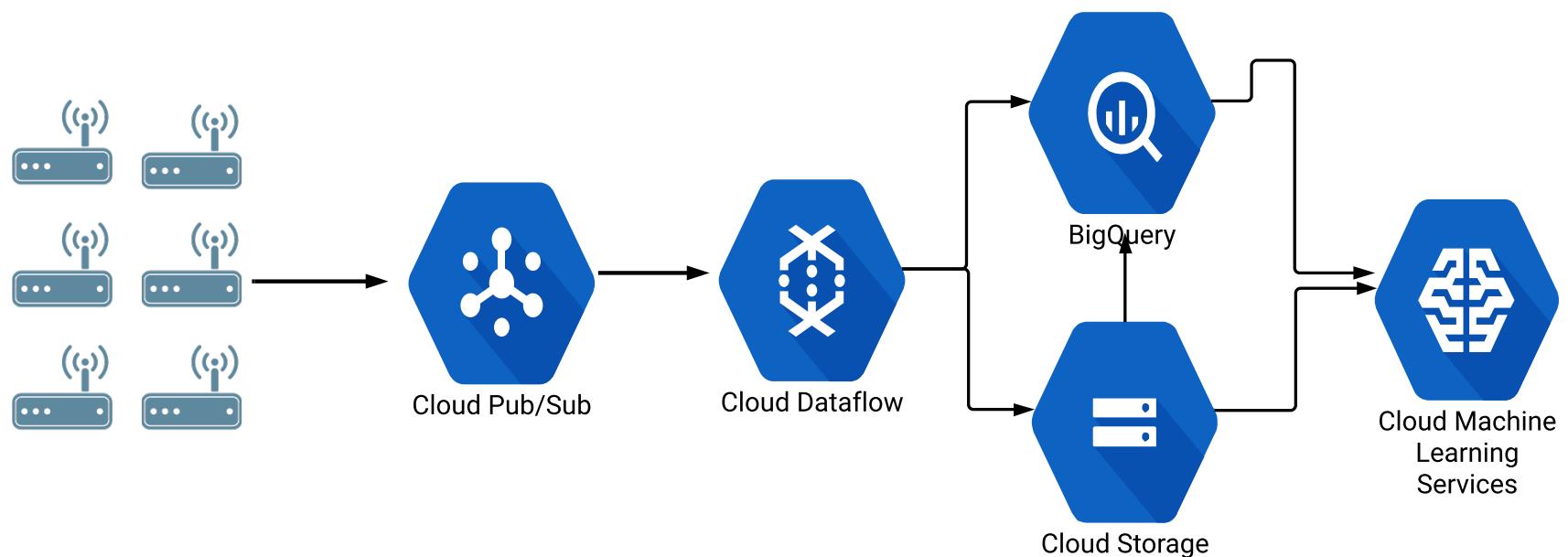
- Isolated environments:
  - Use separate projects
- Granting access to data:
  - Use IAM roles
- Analyze up to 2 years worth of telemetry data:
  - Store in Cloud Storage or BigQuery

[Return to Table of Contents](#)

### Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)[Previous](#)

### Data Flow Model



[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X

Cloud Storage provides globally available, long-term, high-performance storage for all data types.

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X

Cloud Dataflow is a data processing pipeline, transforming both stream and batch data.

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X

BigQuery is a no-ops data warehouse used for massively scalable analytics.

[Return to Table of Contents](#)

## Choose a Lesson

[Case Study Overview](#)[Flowlogistic](#)[MJTelco](#)

X  
Managed machine learning service  
for predictive analytics.

[Return to Table of Contents](#)

## **Choose a Lesson**

[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)

[Return to Table of Contents](#)

## Choose a Lesson

[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)

## *Data Lifecycle*

[Next](#)

- Think of data as a tangible object to be collected, stored, and processed
- Lifecycle from initial collection to final visualization
- Needs to be familiar with the lifecycle steps, what GCP services are associated with each step, and how they connect together
- Data Lifecycle steps:
  - Ingest - Pull in the raw data:
    - Streaming/real-time data from devices
    - On-premises batch data
    - Application logs
    - Mobile-app user events and analytics
  - Store - data needs to be stored in a format and location that is both reliable and accessible
  - Process and analyze - Where the magic happens. Transform data from raw format to actionable information
  - Explore and visualize - "Make it look pretty"
    - The final stage is to convert the results of the analysis into a format that is easy to draw insights from and to share with colleagues and peers

[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 and Associated Services

Ingest	Store	Process & Analyze	Explore & Visualize
 App Engine	 Cloud Storage	 Cloud Dataflow	 Cloud Datalab
 Compute Engine	 Cloud SQL	 Cloud Dataproc	 Google Data Studio
 Kubernetes Engine	 Cloud Datastore	 BigQuery	 Google Sheets
 Cloud Pub/Sub	 Cloud Bigtable	 Cloud ML	
 Stackdriver Logging	 BigQuery	 Cloud Vision API	
 Cloud Transfer Service	 Cloud Storage for Firebase	 Cloud Speech API	
 Transfer Appliance	 Cloud Firestore	 Translate API	
	 Cloud Spanner	 Cloud Natural Language API	
		 Cloud Dataprep	
		 Cloud Video Intelligence API	

[Return to Table of Contents](#)**Choose a Lesson**[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)[Previous](#)[Next](#)

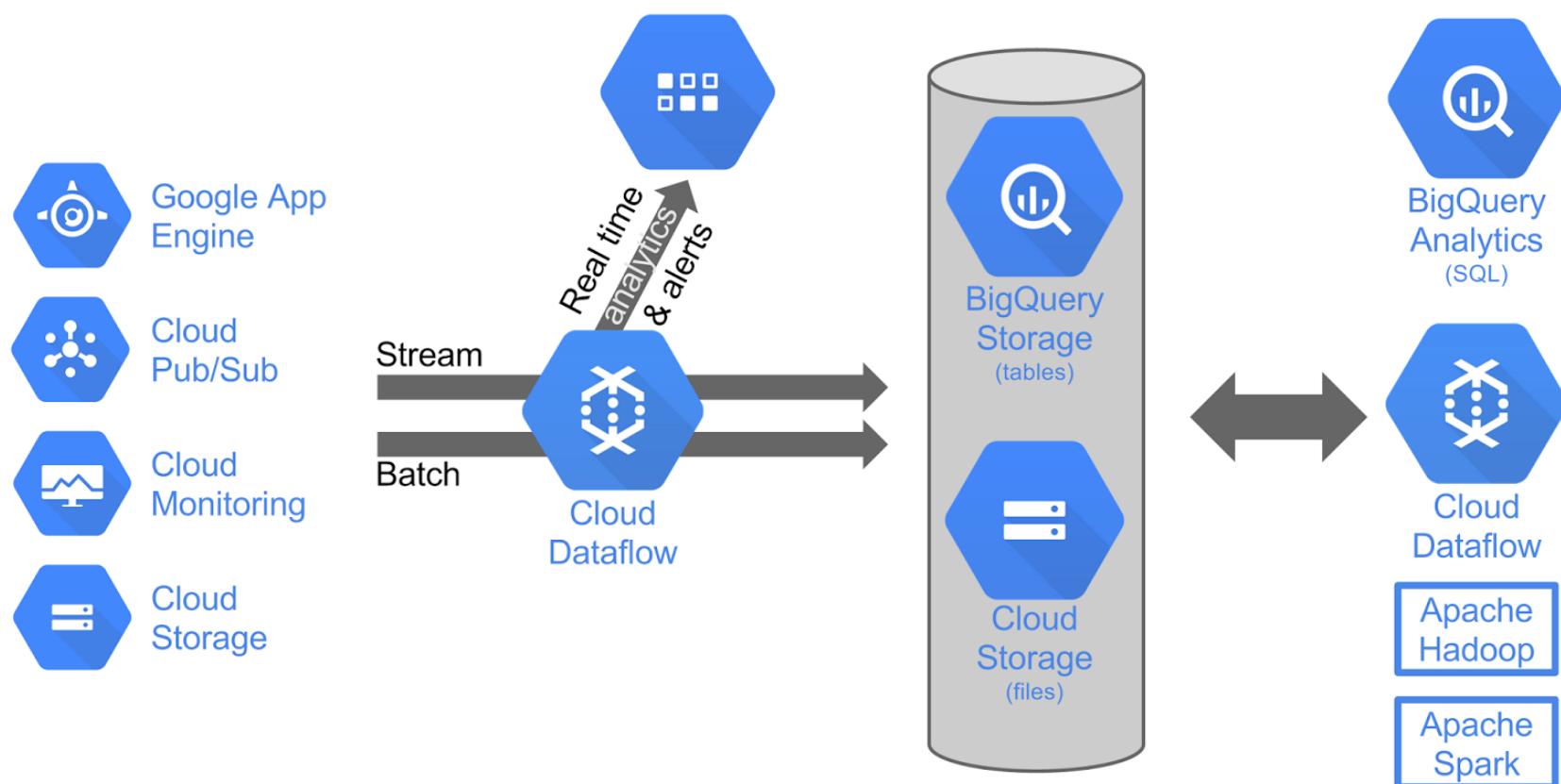
## Data Lifecycle is not a Set Order

Ingest

Process

Store

Analyze

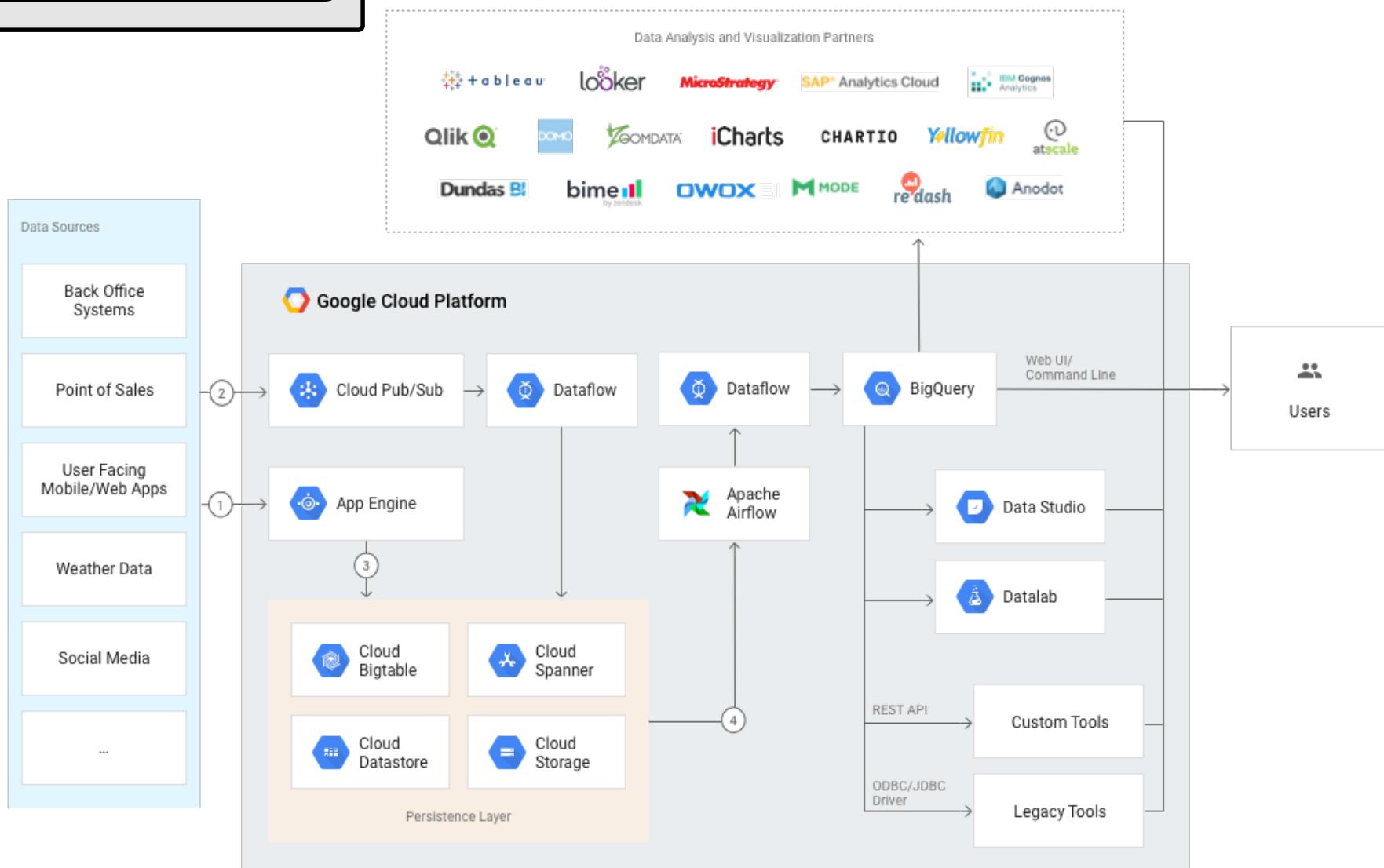


[Return to Table of Contents](#)

## Choose a Lesson

[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)[Previous](#)

## Increasing Complexity of Data Flow



[Return to Table of Contents](#)

## Choose a Lesson

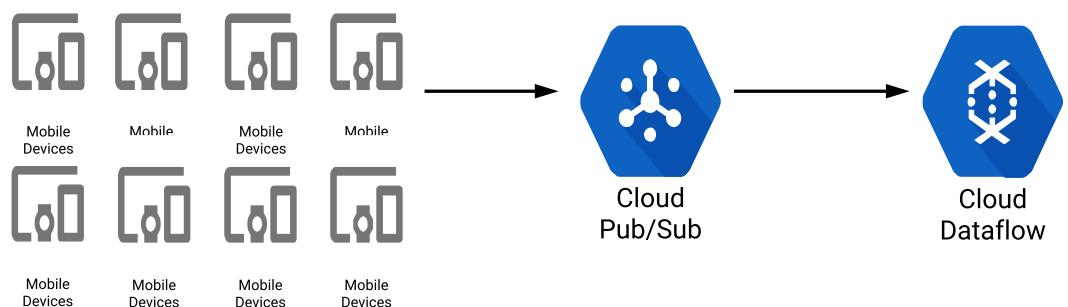
[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)

## *Streaming and Batch Data*

### Data Lifecycle = Data Ingest

#### Streaming (or real-time) data:

- Generated and transmitted continuously by many data sources
- Thousands of data inputs, sent simultaneously, in small sizes (KB)
- Commonly used for telemetry - collecting data from a high number of geographically dispersed devices as it's generated
- Examples:
  - Sensors in transportation vehicles - detecting performance and potential issues
  - Financial institution tracks stock market changes
- Data is processed in small pieces as it comes in
- Requires low latency
- Typically paired with Pub/Sub for the streaming data ingest and Dataflow for real-time processing

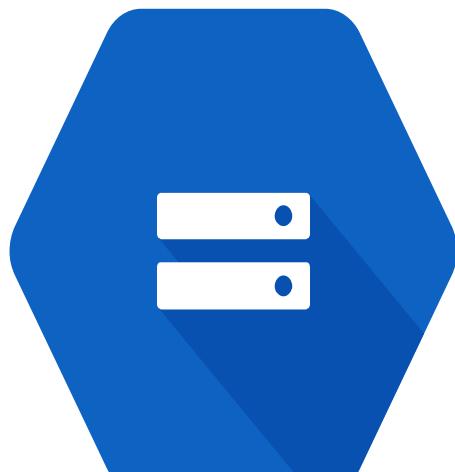


#### Batch (or bulk) data:

- Large sets of data that 'pool' up over time
- Transferring from a small number of sources (usually 1)
- Examples:
  - On-premise database migration to GCP
  - Importing legacy data into Cloud Storage
  - Importing large datasets for machine learning analysis
  - `gsutil cp [storage_location] gs://[BUCKET]` is an example of batch data import
- Low latency is not as important
- Often stored in storage services such as cloud storage, CloudSQL, BigQuery, etc.

[Return to Table of Contents](#)

## Choose a Lesson

[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)

Cloud Storage

## *Cloud Storage as Staging Ground*

[Next](#)

### **Storage 'swiss army knife':**

- **GCS holds all data types:**
  - All database transfer types, raw data, any format
- **Globally available:**
  - Multi-regional buckets provide fast access across regions
  - Regional buckets provide fast access for single regions
  - Edge caching for increased performance
- **Durable and reliable:**
  - Versioning and redundancy
- **Lower cost than persistent disk**
- **Control access:**
  - Project, bucket, or object level
  - Useful for ingest, transform, and publish workflows
  - Option for Public read access

### **Data Engineering perspective:**

- **Migrating existing workloads:**
  - Migrate databases/data into Cloud Storage for import
- **Common first step of data lifecycle - get data to GCS**
- **Staging area for analysis/processing/machine learning import:**
  - 'Data lake'

[Return to Table of Contents](#)

## Getting data in and out of Cloud Storage

### Choose a Lesson

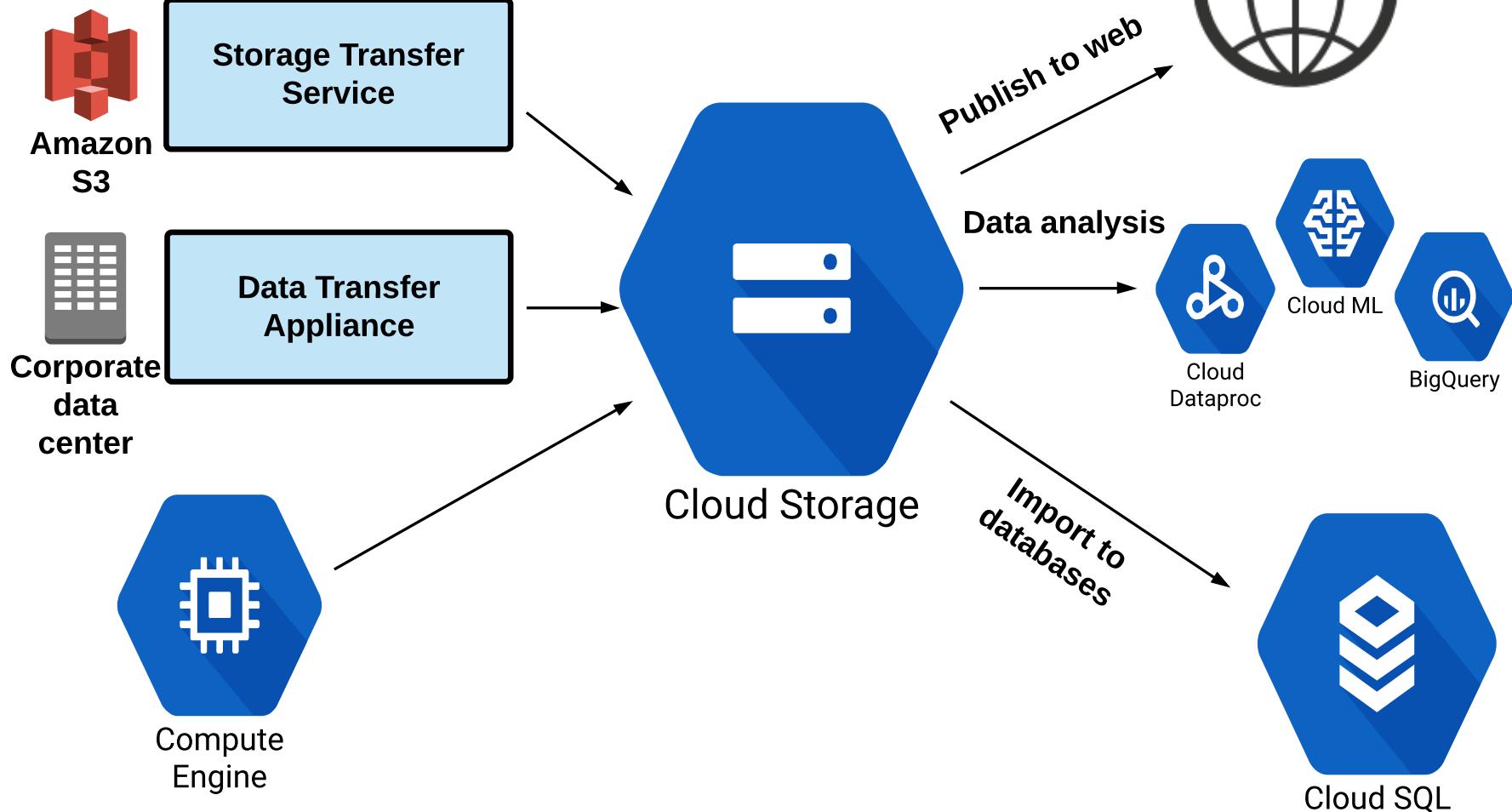
[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)[Previous](#)

### Storage Transfer Service - S3, GCS, HTTP --> GCS:

- One time transfer, periodic sync

### Data Transfer Appliance - physically shipped appliance:

- Load up to 1 petabyte, ship to GCP, loaded into bucket
- gsutil, JSON API - "gsutil cp ..."



[Return to Table of Contents](#)

## Choose a Lesson

[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)[Next](#)

## Database Types

### Two primary database types:

- Relational/SQL
- Non-relational/NoSQL

### Relational (SQL) database:

- SQL = Structured Query Language
- Structured and standardized:
  - Tables - rows and columns
- Data integrity
- High Consistency
- ACID compliance:
  - Atomicity, Consistency, Isolation, Durability
- Examples:
  - MySQL, Microsoft SQL Server, Oracle, PostgreSQL
- Applications:
  - Accounting systems, inventory
- Pros:
  - Standardized, consistent, reliable, data integrity
- Cons:
  - Poor scaling, not as fast performing, not good for semi-structured data
- "Consistency and reliability over performance"

[Return to Table of Contents](#)

## Choose a Lesson

[Data Lifecycle](#)[Batch and Streaming Data](#)[Cloud Storage as Staging Ground](#)[Database Types](#)

## Database Types

[Previous](#)

### Non-relational (NoSQL) Database:

- Non-structured (no table)
- Different standards - key/value, wide table
- Some have ACID compliance (Datastore)
- Examples:
  - Redis, MongoDB, Cassandra, HBase, Bigtable, RavenDB
- Application:
  - Internet of Things (IoT), user profiles, high-speed analytics
- Pros:
  - Scalable, high-performance, not structure-limited
- Cons:
  - Eventual consistency, data integrity
- "Performance over consistency"

### Exam expectations:

- Understand descriptions between database types
- Know which database version matches which description
- Example:
  - "Need database with high throughput, ACID compliance not necessary, choose three possible options"

[Return to Table of Contents](#)

## **Choose a Lesson**

[Choosing a Managed Database](#)[Cloud SQL Basics](#)[Importing Data](#)[SQL Query Best Practices](#)

[Return to Table of Contents](#)Choose a Lesson[Choosing a Managed Database](#)[Cloud SQL Basics](#)[Importing Data](#)[SQL Query Best Practices](#)[Next](#)

## *Choosing a Managed Database*

### Big picture perspective:

- At minimum, know which managed database is the best solution for any 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 frameworkRDBMS+scale  
High transactionsSemi-structured  
Key-value dataHigh throughput  
analyticsUnstructured data  
Holds everythingMission critical  
apps  
Scale+consistency

e.g.

Medical records  
BlogsGlobal supply  
chain  
RetailProduct catalog  
Game stateGraphs  
IoT  
FinanceMultimedia  
Analytics  
Disaster recoveryLarge data  
analytics  
Processing using  
SQL

[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

[Previous](#)

### Decision tree criteria:

- Structured (database) or unstructured?
- Analytical or transactional?
- Relational (SQL) or non-relational (NoSQL)?
- Scalability/availability/size requirements?



[Return to Table of Contents](#)

## Choose a Lesson

[Choosing a Managed Database](#)[Cloud SQL Basics](#)[Importing Data](#)[SQL Query Best Practices](#)[Scaling](#)[High Availability](#)[Database Backups](#)[Software Patches](#)[Database Installs](#)[OS Patches](#)[OS Installation](#)[Server Maintenance](#)[Physical Server](#)[Power-Network-Cooling](#)

## *Cloud SQL Basics*

### What is Cloud SQL?

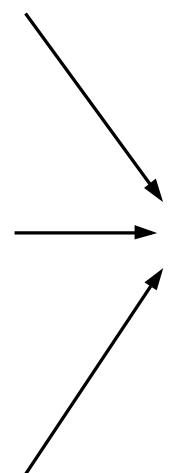
- Direct lift and shift of traditional MySQL/PostgreSQL workloads with the maintenance stack managed for you

### What is managed?

- OS installation/management
- Database installation/management
- Backups
- Scaling - disk space
- Availability:
  - Failover
  - Read replicas
- Monitoring
- Authorize network connections/proxy/use SSL

### Limitations:

- Read replicas limited to the same region as the master:
  - Limited global availability
- Max disk size of 10 TB
- If > 10 TB is needed, or global availability in RDBMS, use Spanner



[Return to Table of Contents](#)

## Choose a Lesson

[Choosing a Managed Database](#)[Cloud SQL Basics](#)[Importing Data](#)[SQL Query Best Practices](#)

## *Importing Data*

### Importing data into Cloud SQL:

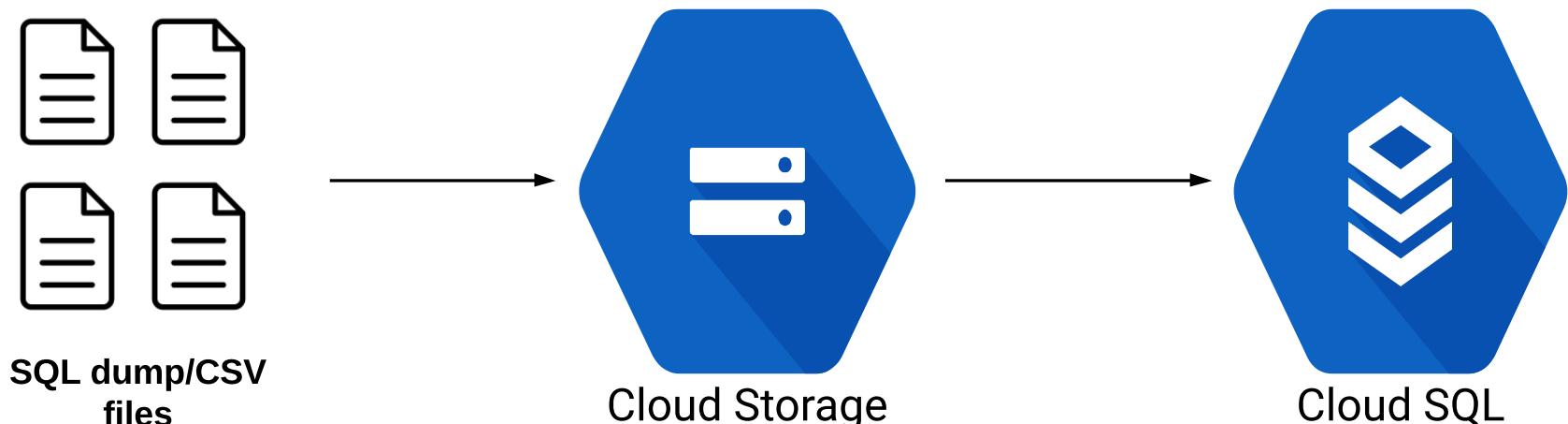
- Cloud Storage as a staging ground
- SQL dump/CSV file format

### Export/Import process:

- Export SQL dump/CSV file:
  - SQL dump file cannot contain triggers, views, stored procedures
- Get dump/CSV file into Cloud Storage
- Import from Cloud Storage into Cloud SQL instance

### Best Practices:

- Use correct flags for dump file ('--flag\_name'):
  - Databases, hex-blob, skip-triggers, set-gtid-purged=OFF, ignore-table
- Compress data to reduce costs:
  - Cloud SQL can import compressed .gz files
- Use InnoDB for Second Generation instances



[Return to Table of Contents](#)

## Choose a Lesson

[Choosing a Managed Database](#)[Cloud SQL Basics](#)[Importing Data](#)[SQL Query Best Practices](#)

## ***SQL Query Best Practices***

### **General SQL efficiency best practices:**

- More, smaller tables better than fewer, large tables:
  - Normalization of tables
- Define your SELECT fields instead of using SELECT \*:
  - SELECT \* acts as a 'select all'
- When joining tables, use INNER JOIN instead of WHERE:
  - WHERE creates more variable combinations = more work

[Return to Table of Contents](#)

## **Choose a Lesson**

[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)

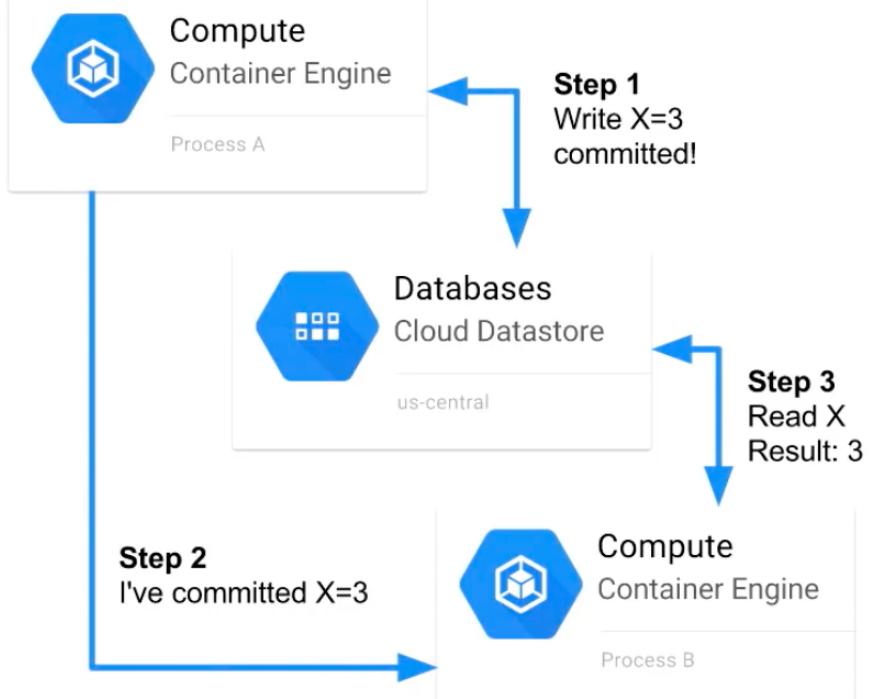
[Return to Table of Contents](#)

### Choose a Lesson

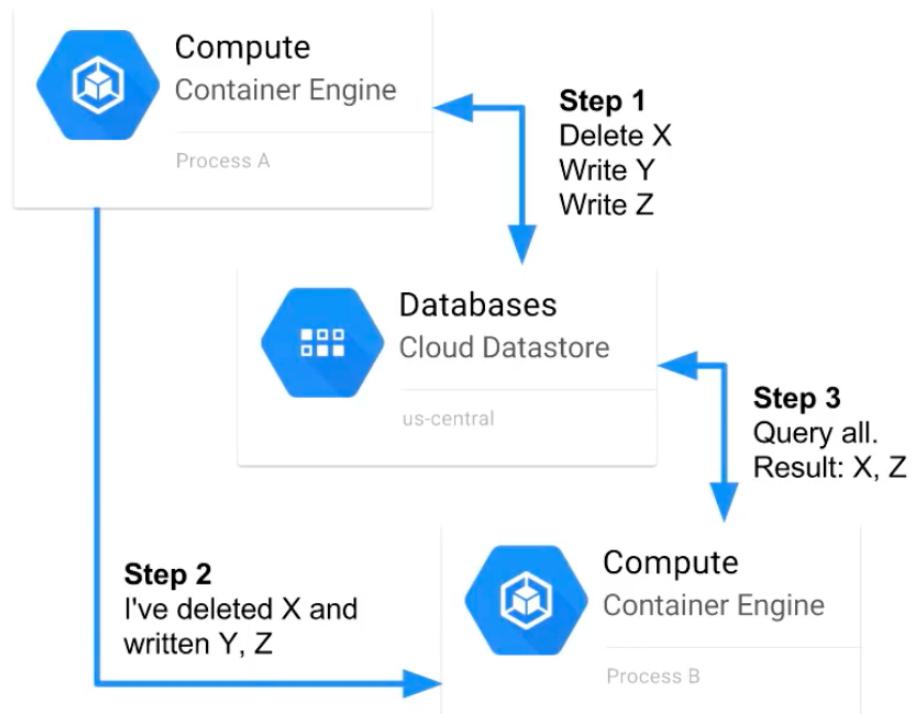
[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)[Previous](#)

### Strong

## Data Consistency



### Eventual



[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)

## Data Consistency

[Next](#)

### What is data consistency in queries?

- "How up to date are these results?"
- "Does the order matter?"
- **Strongly consistent** = Parallel processes see changes in same order:
  - Query is guaranteed up to date but may take longer to complete
- **Eventually consistent** = Parallel process can see changes out of order, will eventually see accurate end state:
  - Faster query, but may \*sometimes\* return stale results
- **Performance vs. accuracy**
- **Ancestor query/key-value operations** = strong
- **Global queries/projections** = eventual

### Use cases:

- **Strong** - financial transaction:
  - Make deposit -- check balance
- **Eventual** - census population:
  - Order not as important, as long as you get eventual result

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)

## *Queries and Indexing*

[Previous](#)

### Danger - Exploding Indexes!

- Default - create an entry for every possible combination of property values
- Results in higher storage and degraded performance
- Solutions:
  - Use a custom index.yaml file to narrow index scope
  - Do not index properties that don't need indexing

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)[Next](#)

## *Queries and Indexing*

### Query:

- Retrieve an entity from Datastore that meets a set of conditions
- Query includes:
  - Entity kind
  - Filters
  - Sort order
- Query methods:
  - Programmatic
  - Web console
  - Google Query Language (GQL)

### Indexing:

- Queries gets results from indexes:
  - Contain entity keys specified by index properties
  - Updated to reflect changes
  - Correct query results available with no additional computation needed

### Index types:

- Built-in - default option:
  - Allows single property queries
- Composite - specified with an index configuration file (index.yaml):
  - gcloud datastore create-indexes index.yaml

```
indexes:
- kind: Task
  properties:
    - name: tags
    - name: created
- kind: Task
  properties:
    - name: collaborators
    - name: created
```

[Return to Table of Contents](#)

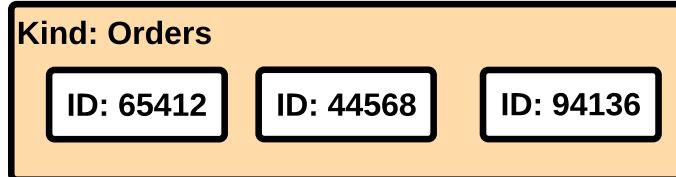
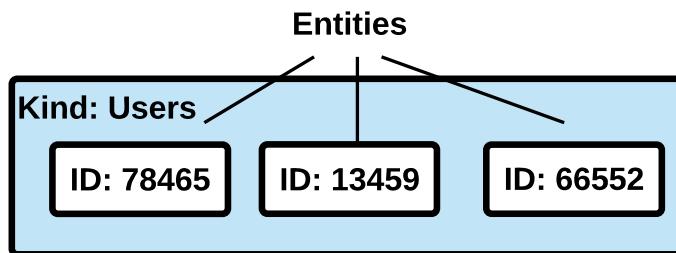
### Choose a Lesson

[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)

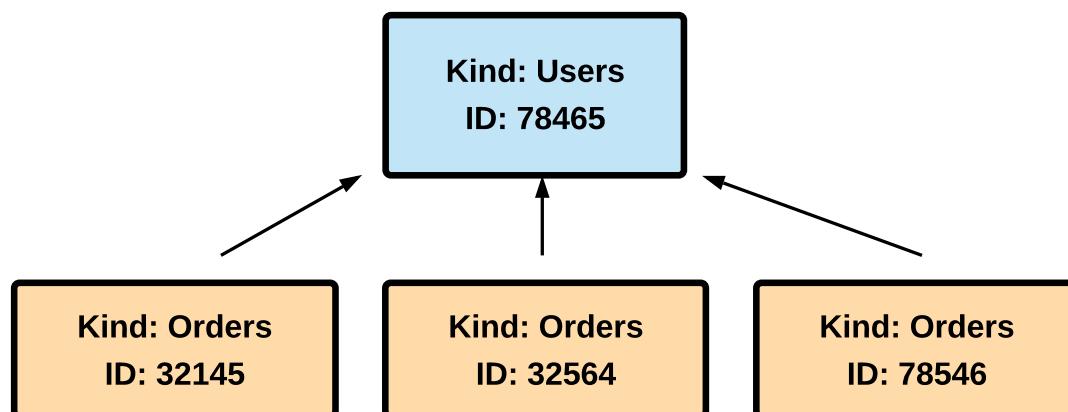
## *Data Organization*

[Previous](#)

### Simple Collections of Entities



### Hierarchies (Entity Groups)



[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)

## Data Organization

[Next](#)

### Short version:

- Entities grouped by kind (category)
- Entities can be hierarchical (nested)
- Each entity has one or more properties
- Properties have a value assigned

Concept	Relational Database	Datastore
Category of object	Table	Kind
Single Object	Row	Entity
Individual data for an object	Field	Property
Unique ID for an object	Primary key	Key

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)

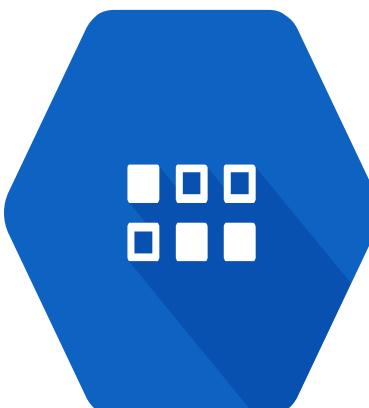
## **Cloud Datastore Overview**

[Previous](#)

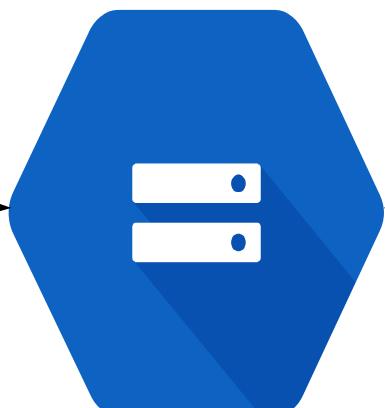
### Other important facts:

- Single Datastore database per project
- Multi-regional for wide access, single region for lower latency and for single location
- Datastore is a transactional database
- Bigtable is an analytical database
- IAM roles:
  - Primitive and predefined
  - Owner, user, viewer, import/export admin, index admin

**Backup/Export/Import/Analyze**  
**Managed export/import service**



Cloud Datastore



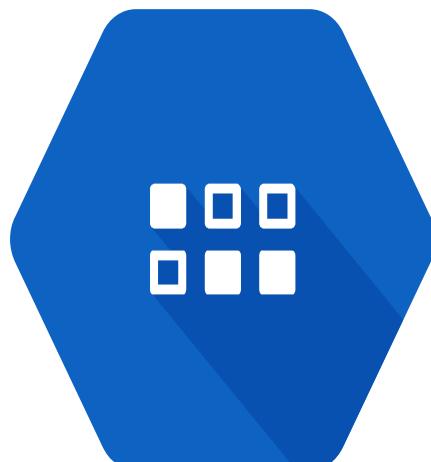
Cloud Storage



BigQuery

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Datastore Overview](#)[Data Organization](#)[Queries and Indexing](#)[Data Consistency](#)

Cloud Datastore

## **Cloud Datastore Overview**

[Next](#)

### What is Cloud Datastore?

- **No Ops:**
  - No provisioning of instances, compute, storage, etc.
  - Compute layer is abstracted away
- **Highly scalable:**
  - Multi-region access available
  - Sharding/replication handled automatically
- **NoSQL/non-relational database:**
  - Flexible structure/relationship between objects

### Use Datastore for:

- Applications that need highly available structured data, at scale
- Product catalogs - real-time inventory
- User profiles - mobile apps
- Game save states
- ACID transactions - e.g., transferring funds between accounts

### Do not use Datastore for:

- **Analytics (full SQL semantics):**
  - Use BigQuery/Cloud Spanner
- **Extreme scale (10M+ read/writes per second):**
  - Use Bigtable
- **Don't need ACID transactions/data not highly structured:**
  - Use Bigtable
- **Lift and shift (existing MySQL):**
  - Use Cloud SQL
- **Near zero latency (sub-10ms):**
  - Use in-memory database (Redis)

[Return to Table of Contents](#)

## **Choose a Lesson**

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

[Return to Table of Contents](#)

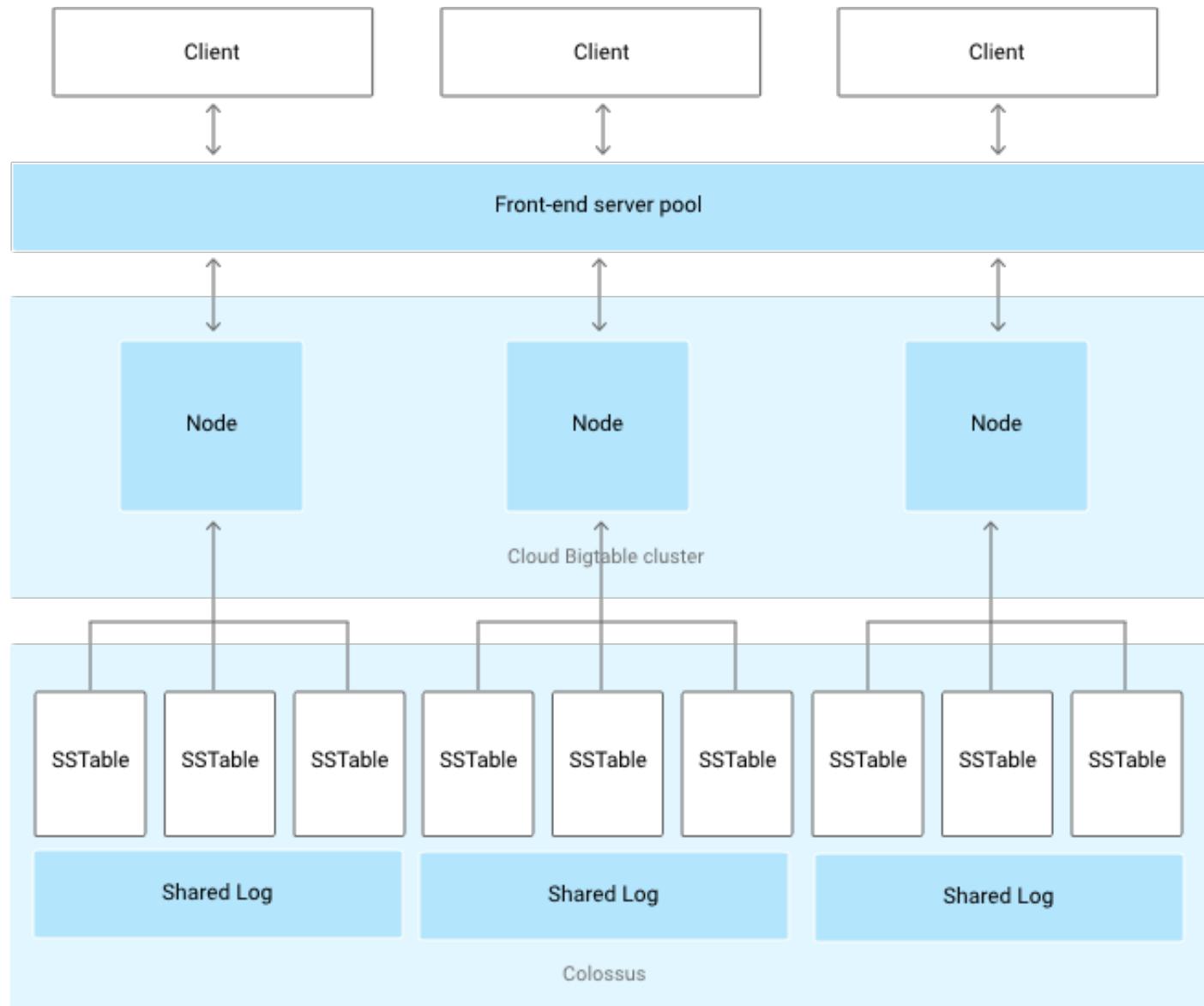
## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

## Cloud Bigtable Overview

[Previous](#)

### Cloud Bigtable Infrastructure



[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

Cloud Bigtable

## *Cloud Bigtable Overview*

[Next](#)

### What is Cloud Bigtable?

- High performance, massively scalable NoSQL database
- Ideal for large analytical workloads

### History of Bigtable

- Considered one of the originators of NoSQL industry
- Developed by Google in 2004
  - Existing database solutions were too slow
  - Needed realtime access to petabytes of data
- Powers Gmail, YouTube, Google Maps, and others

### What is it used for?

- High throughput analytics
- Huge datasets

### Use cases

- Financial data - stock prices
- IoT data
- Marketing data - purchase histories

### Access Control

- Project wide or instance level
- Read/Write/Manage

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

Cloud Bigtable

## *Instance Configuration*

### Instance basics

[Next](#)

- Not no-ops
  - Must configure nodes
- Entire Bigtable project called 'instance'
  - All nodes and clusters
- Nodes grouped into clusters
  - 1 or more clusters per instance
- Auto-scaling storage
- Instance types
  - Development - low cost, single node
    - No replication
  - Production - 3+ nodes per cluster
    - Replication available, throughput guarantee

### Replication and Changes

- Synchronize data between clusters
  - One additional cluster, total
  - (Beta) available cross-region
- Resizing
  - Add and remove nodes and clusters with no downtime
- Changing disk type (e.g. HDD to SSD) requires new instance

### Interacting with Bigtable

- Command line - cbt tool or HBase shell
  - cbt tool is simpler and preferred option

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

Cloud Bigtable

[Previous](#)

## *Instance Configuration*

- Install the cbt command in Google SDK
  - sudo gcloud components update
  - gcloud components install cbt
- Configure cbt to use your project and instance via .cbtrc file'
  - echo -e "project = [PROJECT\_ID]\ninstance = [INSTANCE\_ID]" > ~/.cbtrc
- Create table
  - cbt createtable my-table
- List table
  - cbt ls
- Add column family
  - cbt createfamily my-table cf1
- List column family
  - cbt ls my-table
- Add value to row 1, using column family cf1 and column qualifier c1
  - cbt set my-table r1 cf1:c1=test-value
- Delete table (if not deleting instance)
  - cbt deletetable my-table
- Read the contents of your table
  - cbt read my-table

Get help with cbt command using 'cbt --help'

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

## **Data Organization**

### **Data Organization**

- One big table (hence the name Bigtable)
- Table can be thousands of columns/billions of rows
- Table is sharded across tablets

### **Table components**

- Row Key
  - First column
- Columns grouped into column families

	Column-Family-1		Column-Family-2	
Row Key	Column-Qualifier-1	Column-Qualifier-2	Column-Qualifier-1	Column-Qualifier-2
r1	r1, cf1:cq1	r1, cf1:cq2	r1, cf1:cq1	r1, cf1:cq2
r2	r2, cf1:cq1	r2, cf1:cq2	r2, cf1:cq1	r2, cf1:cq2

### **Indexing and Queries**

- Only the row key is indexed
- Schema design is necessary for efficient queries!
- Field promotion - move fields from column data to row key

Row key	Column data
BATTERY#Corrie#20150301124501001	METRIC:PERCENTAGE:98
BATTERY#Corrie#20150301124501003	METRIC:PERCENTAGE:96
BATTERY#Jo#20150301124501002	METRIC:PERCENTAGE:54
BATTERY#Sam#20150301124501004	METRIC:PERCENTAGE:43
BATTERY#Sam#20150301124501005	METRIC:PERCENTAGE:38

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

Front-end server pool serves client requests to nodes X

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

X

Nodes handle cluster requests. It acts as the compute for processing requests.

No data is stored on the node except for metadata to direct requests to the correct tablet

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

X  
Bigtable's table is sharded into block of rows, called tablets.  
Tablets are stored on Colossus, Google's file system, in SStable format  
Storage is separate from the compute nodes, though each tablet is associated with a node.  
As a result, replication and recovery of node data is very fast, as only metadata/pointers need to be updated.

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Bigtable Overview](#)[Instance Configuration](#)[Data Organization](#)[Schema Design](#)

### Row Key

memusage+user+timestamp

20-mattu-201805082048

## ***Schema Design***

### Schema Design

- Per table - Row key is only indexed item
- Keep all entity info in a single row
- Related entities should be in adjacent rows
  - More efficient reads
- Tables are sparse - empty columns take no space

### Schema efficiency

- Well defined row keys = less work
  - Multiple values in row key
- Row key (or prefix) should be sufficient for search
- Goal = spread load over multiple nodes
  - All on one node = 'hotspotting'

### Row key best practices

- Good row keys = distributed load
  - Reverse domain names (com.linuxacademy.support)
  - String identifiers (mattu)
  - Timestamps (reverse, NOT at front/or only identifier)
- Poor row keys = hotspotting
  - Domain names (support.linuxacademy.com)
  - Sequential ID's
  - Timestamps alone/at front

[Return to Table of Contents](#)

## **Choose a Lesson**

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

## **Cloud Spanner Overview**

[Next](#)

### **What is Cloud Spanner?**

- Fully managed, highly scalable/available, relational database
- Similar architecture to Bigtable
- "NewSQL"

### **What is it used for?**

- Mission critical, relational databases that need strong transactional consistency (ACID compliance)
- Wide scale availability
- Higher workloads than Cloud SQL can support
- Standard SQL format (ANSI 2011)

### **Horizontal vs. vertical scaling**

- Vertical = more compute on single instance (CPU/RAM)
- Horizontal = more instances (nodes) sharing the load

### **Compared to Cloud SQL**

- Cloud SQL = Cloud incarnation of *on-premises* MySQL database
- Spanner = designed from the ground up for the cloud
- Spanner is not a 'drop in' replacement for MySQL
  - Not MySQL/PostgreSQL compatible
  - Work required to migrate
  - However, when making transition, don't need to choose between consistency and scalability

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

## **Cloud Spanner Overview**

[Previous](#)[Next](#)

**Transactional Consistency vs. Scalability**  
**Why not both?**

	Cloud Spanner	Traditional Relational	Traditional Non-relational
Schema	Yes	Yes	No
SQL	Yes	Yes	No
Consistency	Strong	Strong	Eventual
Availability	High	Failover	High
Scalability	Horizontal	Vertical	Horizontal
Replication	Automatic	Configurable	Configurable

**Primary purpose of Cloud Spanner:**  
**No compromises relational database**

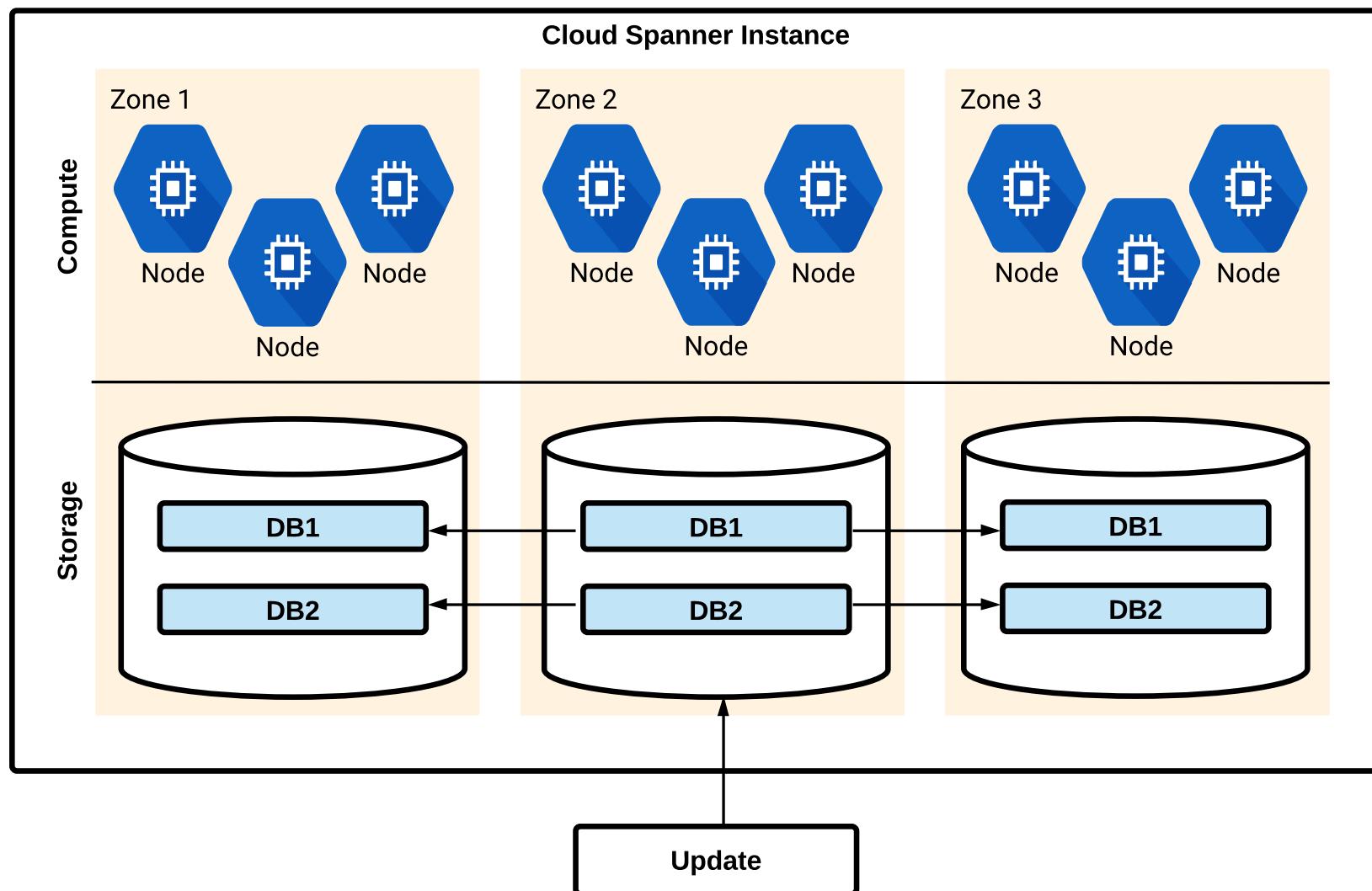
[Return to Table of Contents](#)

## Cloud Spanner Overview

### Choose a Lesson

[Cloud Spanner Overview](#)[Previous](#)[Next](#)[Data Organization and Schema](#)

### Cloud Spanner Architecture (similar to Bigtable)



[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

## ***Cloud Spanner Overview***

[Previous](#)

### **Identity and Access Management (IAM)**

- Project, Instance, or Database level
- roles/spanner.
- Admin - Full access to all Spanner resources
- Database Admin - Create/edit/delete databases, grant access to databases
- Database Reader - read/execute database/schema
- Viewer - view instances and databases
  - Cannot modify or read from database

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

X

Nodes handle computation for queries, similar to that of Bigtable.

Each node serves up to 2 TB of storage.

More nodes = more CPU/RAM = increased throughput

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

X

Storage is replicated across zones (and regions, where applicable).  
Like Bigtable, storage is separate from computing nodes

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

X

Whenever an update is made to a database in one zone/region, it is automatically replicated across zones/regions.

Automatic synchronous replication

- When data is written, you know it has been written
- Any reads guarantee data accuracy

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

## **Data Organization and Schema**

### Organization

[Next](#)

- RDBMS = tables
- Supports SQL joins, queries, etc
- Same SQL dialect as BigQuery
- Tables are handled differently
  - Parent/child tables
  - Interleave Data Layout

### Typical Relational Database

**Two sets of related data = Two tables**

SingerId	SingerName
1	Beatles
2	U2
3	Pink Floyd

SingerId	AlbumId	AlbumName
1	1	Help!
1	2	Abbey Road
3	1	The Wall

### Spanner Interleave Tables

Singers(1)	"Marc"	"Richards"	<Bytes>		
Albums(1, 1)				"Total Junk"	
Albums(1, 2)				"Go, Go, Go"	
Songs(1, 2, 1)					"42"
Songs(1, 2, 2)					"Nothing Is The Same"
Singers(2)	"Catalina"	"Smith"	<Bytes>		
Albums(2, 1)				"Green"	
Songs(2, 1, 1)					"Let's Get Back Together"
Songs(2, 1, 2)					"Starting Again"
Songs(2, 1, 3)					"I Knew You Were Magic"
Albums(2, 2)				"Forever Hold Your Peace"	
Albums(2, 3)				"Terrified"	
Songs(2, 3, 1)					"Fight Story"

[Return to Table of Contents](#)

## Choose a Lesson

[Cloud Spanner Overview](#)[Data Organization and Schema](#)

## ***Data Organization and Schema***

[Previous](#)

### **Primary keys and Schema**

- How to tell which child tables to store with which parent tables
- Usually a natural fit
  - 'Customer ID'
  - 'Invoice ID'
- Avoid hotspotting
  - No sequential numbers
  - No timestamps (also sequential)
    - Use descending order if timestamps required

[Return to Table of Contents](#)

## **Choose a Lesson**

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

[Return to Table of Contents](#)

### Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

Traffic sensors



Credit Card Transactions



Mobile Gaming

### Examples

#### Fast action is often necessary

- Must quickly collect data, gain insights, and take action
- Sending to storage can add latency
- Credit card fraud
- Predict highway traffic

[Next](#)

## *Streaming Data Challenges*

### What is streaming data?

[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

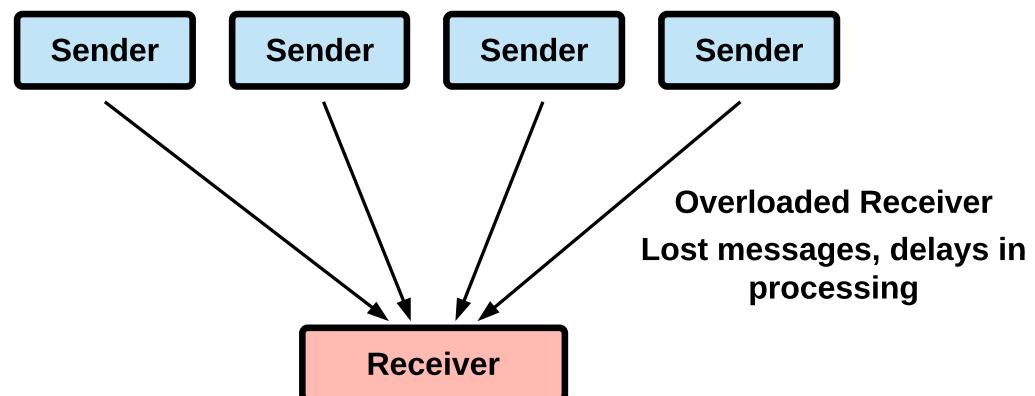
## *Streaming Data Challenges*

[Previous](#)

### Tight vs loosely coupled systems

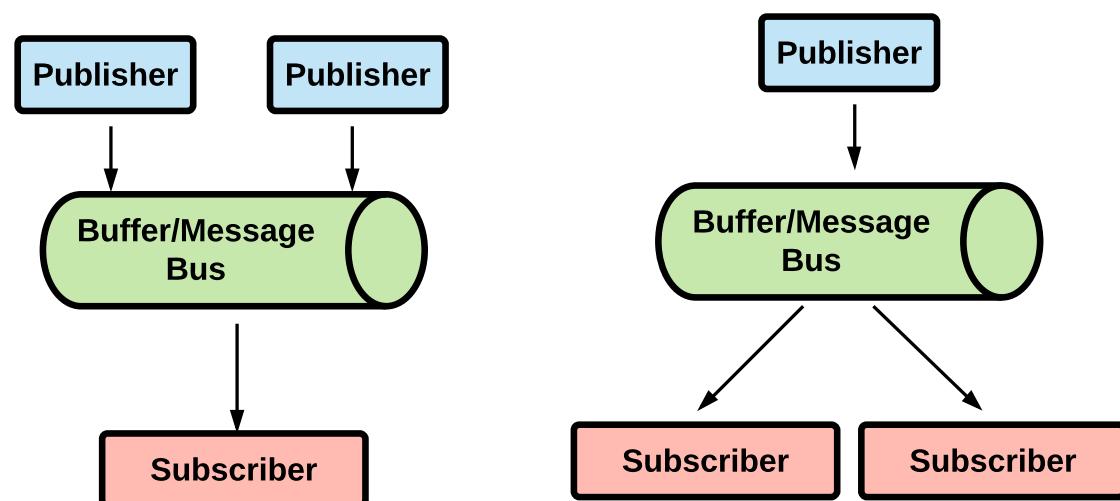
- Tightly (direct) coupled systems more likely to fail
- Loosely coupled systems with 'buffer' scale with better fault tolerance

#### Tightly coupled system



#### Loosely coupled system

- Fault tolerance
- Scalability
- Message queuing



[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

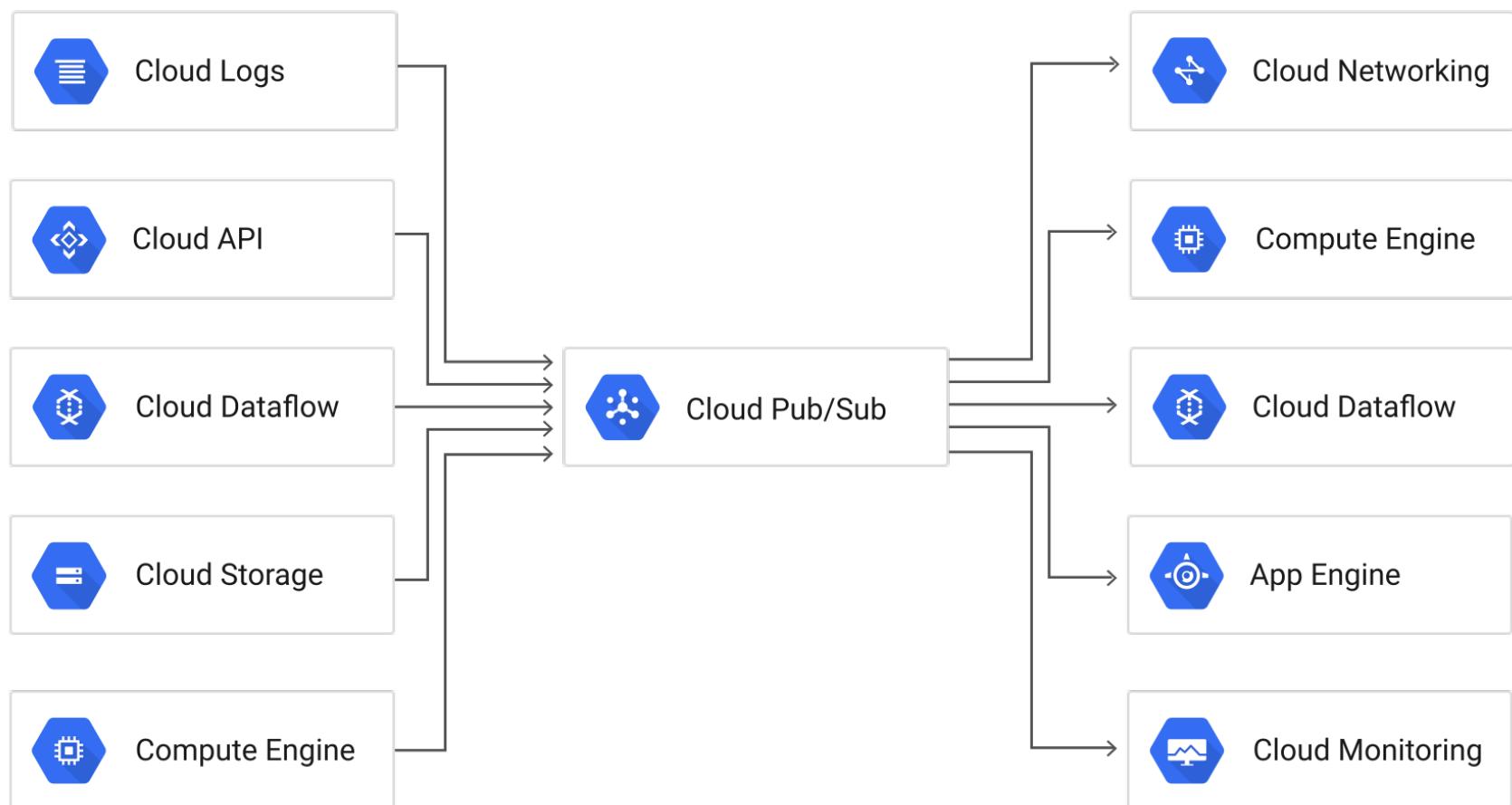
## Cloud Pub/Sub Overview

[Next](#)

### What is Cloud Pub/Sub?

- Global-scale messaging buffer/coupler
- No-ops, global availability, auto-scaling
- Decouples senders and receivers
- Streaming data ingest
  - Also connects other data pipeline services
- Equivalent to Apache Kafka (open source)
- Guaranteed at-least-once delivery

Asynchronous messaging - many to many (or any other combination)



[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

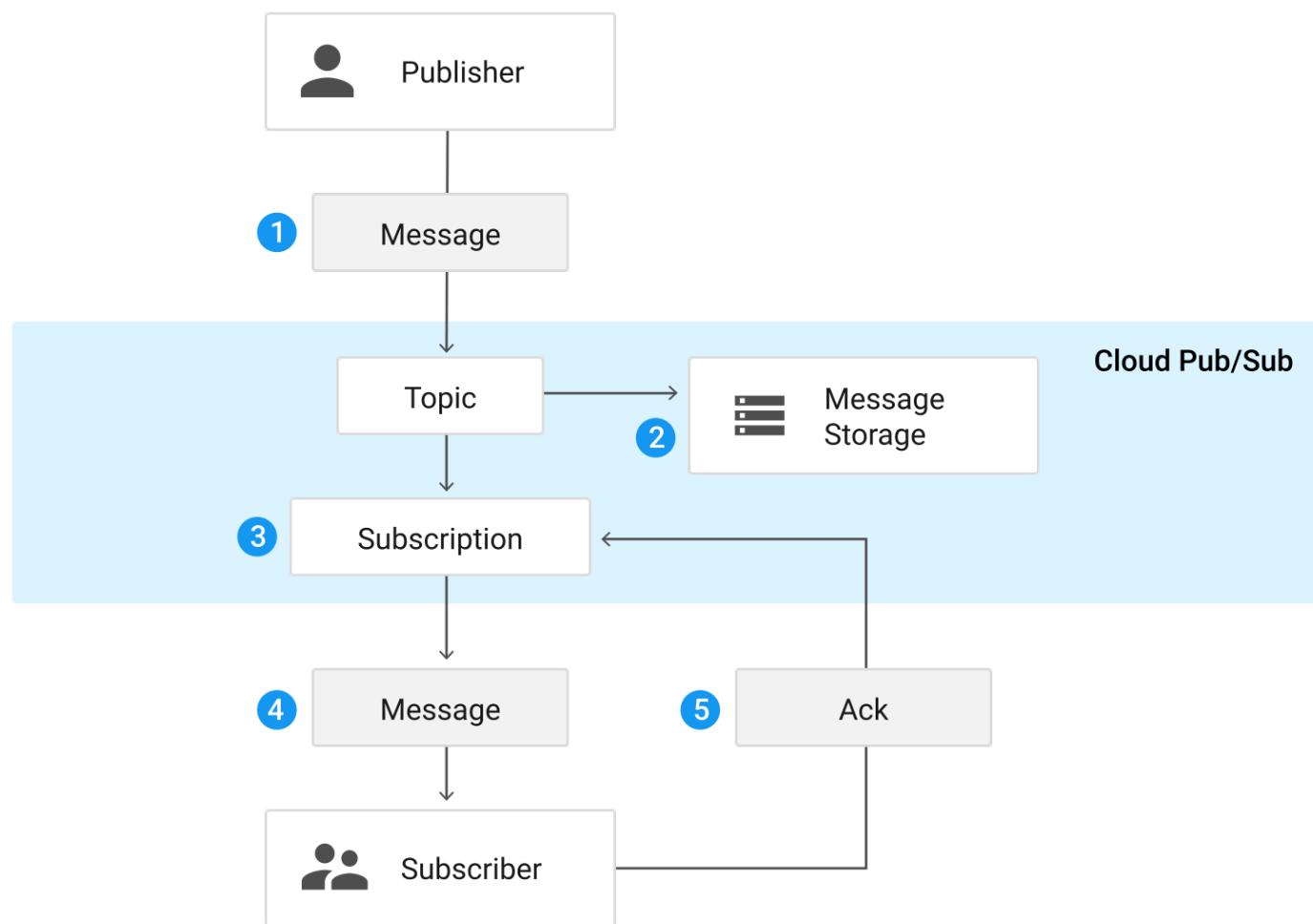
## Cloud Pub/Sub Overview

[Previous](#)[Next](#)

### How it works - terminology

- Topics, Messages, Publishers, Subscribers, Message Store

Click numbers for process steps



[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

Publisher application creates a *topic* in the Cloud Pub/Sub service and sends *messages* to the topic. A message contains a payload and optional *attributes* that describe the payload content.



[Return to Table of Contents](#)

## **Choose a Lesson**

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

X

Messages are stored in a *message store* until they are delivered and acknowledged by subscribers.

[Return to Table of Contents](#)

## **Choose a Lesson**

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)**X**

Pub/Sub forwards messages from a topic to all subscribers, individually.

Messages can be either pushed by Pub/Sub to subscribers, or pulled by subscribers from Pub/Sub.

[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

Subscriber receives pending messages from its subscription and acknowledges each one to the Pub/Sub service.

[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

After message is acknowledged by the subscriber, it is removed from the subscription's queue of messages.



[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

## *Cloud Pub/Sub Overview*

[Previous](#)[Next](#)

### Push and Pull

- Pub/Sub can either push messages to subscribers, or subscribers can pull messages from Pub/Sub
- Push = lower latency, more real-time
- Push subscribers must be Webhook endpoints that accept POST over HTTPS
- Pull ideal for large volume of messages - batch delivery

### IAM

- Control access at project, topic, or subscription level
- Admin, Editor, Publisher, Subscriber
- Service accounts are best practice

### Pricing

- Data volume used per month (per GB)

### Out of order messaging

- Messages may arrive from multiple sources out of order
- Pub/Sub does not care about message ordering
- Dataflow is where out of order messages are processed/resolved
- Can add message attributes to help with ordering

Monthly data	Price Per GB
First 10 GB	\$0.00
Next 50 TB	\$0.06
Next 100 TB	\$0.05
Beyond 150 TB	\$0.04

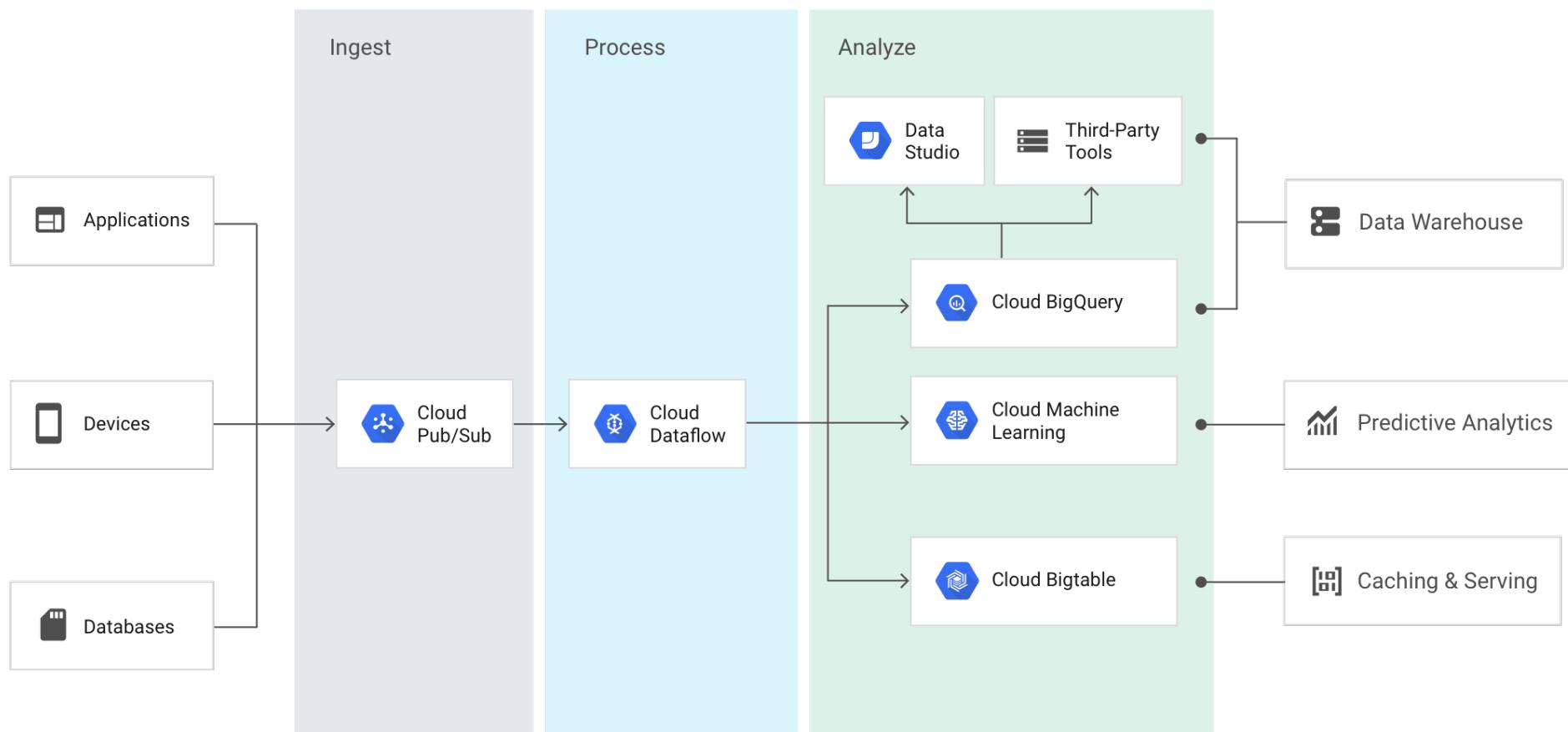
[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)[Previous](#)

## Cloud Pub/Sub Overview

### Big picture - data lifecycle for streaming data ingest



[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)[Next](#)

## ***Pub/Sub Hands On***

### **The steps**

- Create a topic
- Create a subscription
- Publish messages
- Retrieve messages

### **Simple topic/subscription/publish via gcloud**

#### **Create a topic called 'my-topic'**

- `gcloud pubsub topics create my-topic`

#### **Create subscription to topic 'my-topic'**

- `gcloud pubsub subscriptions create --topic my-topic mySub1`

#### **Publish a message to your topic**

- `gcloud pubsub topics publish my-topic --message "hello"`

#### **Retrieve message with your subscription, acknowledge receipt and remove message from queue**

- `gcloud pubsub subscriptions pull --auto-ack mySub1`

#### **Cancel subscription**

- `gcloud pubsub subscriptions delete mySub1`

[Return to Table of Contents](#)

## Choose a Lesson

[Streaming Data Challenges](#)[Cloud Pub/Sub Overview](#)[Pub/Sub Hands On](#)

## *Pub/Sub Hands On*

[Previous](#)

### Traffic data exercise

- Clone GitHub
- Copy data points
- Simulate traffic data
- Pull messages

Clone GitHub data to Cloud Shell (or other SDK environment), and browse to publish folder

- cd ~
- git clone https://github.com/linuxacademy/googledataengineer
- cd ~/googledataengineer/courses/streaming/publish

Create a topic called 'sandiego'

- gcloud pubsub topics create sandiego

Create subscription to topic 'sandiego'

- gcloud pubsub subscriptions create --topic sandiego mySub1

Run script to download sensor data

- ./download\_data.sh

May need to authenticate shell to ensure we have the right permissions

- gcloud auth application-default login

View script info

- vim ./send\_sensor\_data.py or use viewer of your choice

Run python script to simulate one hour of data per minute

- ./send\_sensor\_data.py --speedFactor=60 --project=YOUR-PROJECT-ID

If you receive error: google.cloud.pubsub can not be found OR 'ImportError: No module named iterator', run below pip command to install components then try again

- sudo pip install -U google-cloud-pubsub

Open new Cloud Shell tab (using + symbol)

Pull message using subscription mySub1

- gcloud pubsub subscriptions pull --auto-ack mySub1

Create a new subscription and pull messages with it

- gcloud pubsub subscriptions create --topic sandiego mySub2
- gcloud pubsub subscriptions pull --auto-ack mySub2

[Return to Table of Contents](#)

## **Choose a Lesson**

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

[Return to Table of Contents](#)

## Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

## Data Processing Challenges

### What is Data Processing?

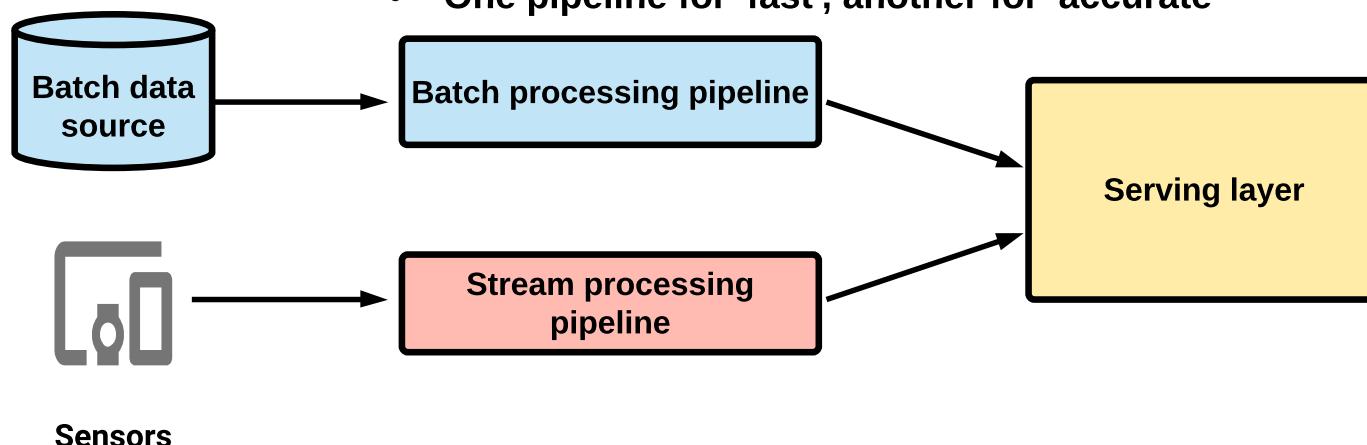
[Next](#)

- Read Data (Input)
- Transform it to be relevant - Extract, Transform, and Load (ETL)
- Create output



### Challenge: Streaming and Batch data pipelines:

- Until recently, separate pipelines are required for each
- Difficult to compare recent and historical data
- One pipeline for 'fast', another for 'accurate'



### Why both?

- Credit card monitoring
- Compare streaming transactions to historical batch data to detect fraud

[Return to Table of Contents](#)

## Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)[Previous](#)

## *Data Processing Challenges*

### Challenge: Complex element processing:

- Element = single data input
- One at a time element ingest from single source = easy
- Combining elements (aggregation) = hard
- Processing data from different sources, streaming, and out of order (composite) = REALLY hard

### Solution: Apache Beam + Cloud Dataflow



beam +



Cloud Dataflow

[Return to Table of Contents](#)

## Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

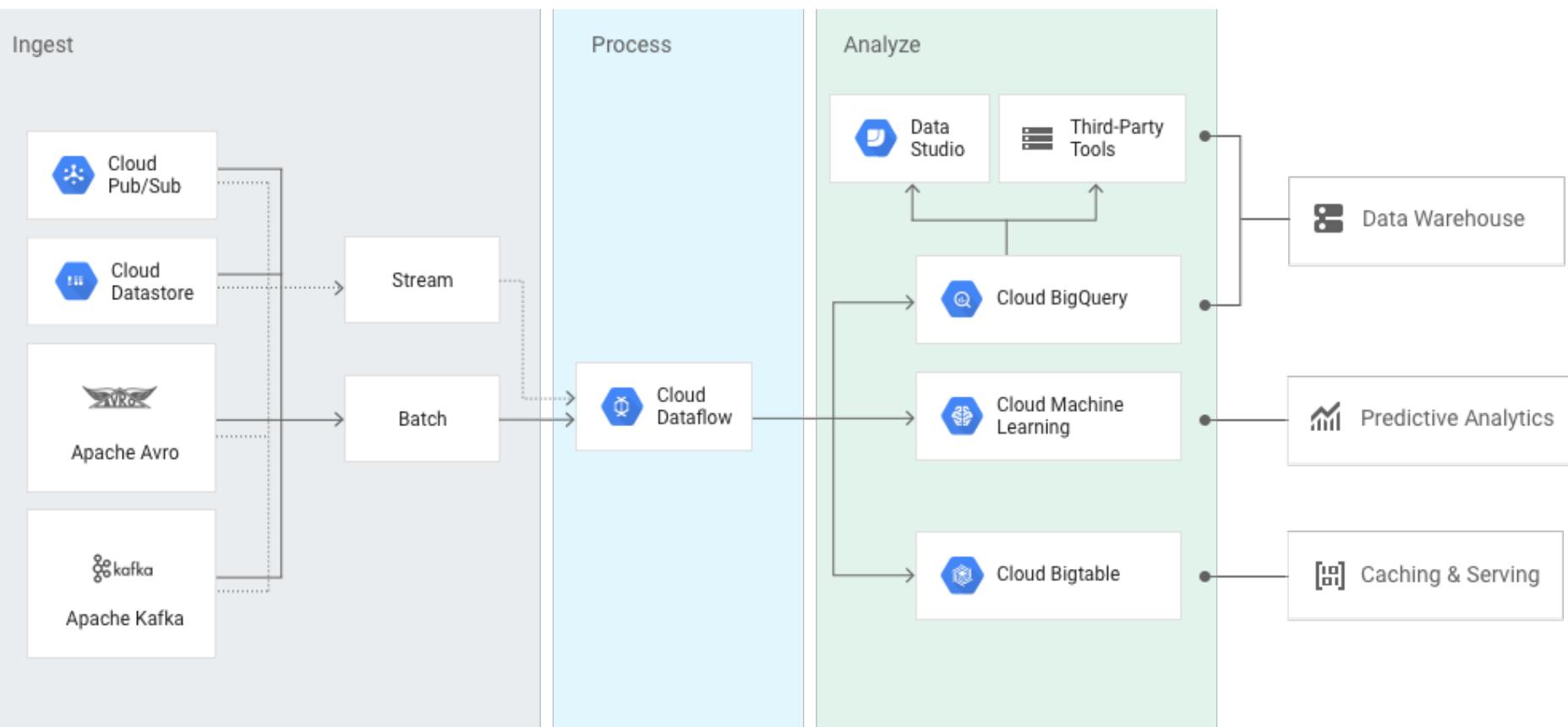
## Cloud Dataflow Overview

[Next](#)

### What is it?

- Auto-scaling, No-Ops, Stream, and Batch Processing
- Built on Apache Beam:
  - Documentation refers to Apache Beam site
  - Configuration is 100% code-based
- Integrates with other tools (GCP and external):
  - Natively - Pub/Sub, BigQuery, Cloud ML Engine
  - Connectors - Bigtable, Apache Kafka
- Pipelines are regional-based

### Big Picture - Data Transformation



[Return to Table of Contents](#)

## Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

## *Cloud Dataflow Overview*

[Previous](#)[Next](#)

### IAM:

- Project-level only - all pipelines in the project (or none)
- Pipeline data access separate from pipeline access
- Dataflow Admin - Full pipeline access plus machine type/storage bucket config access
- Dataflow Developer - Full pipeline access, no machine type/storage bucket access
- Dataflow Viewer - view permissions only
- Dataflow Worker - Specifically for service accounts

### Dataflow vs Dataproc?

### Beam vs. Hadoop/Spark?

#### Dataproc:

- Familiar tools/packages
- Employee skill sets
- Existing pipelines

#### Dataflow:

- Less Overhead
- Unified batch and stream processing
- Pipeline portability across Dataflow, Spark, and Flink as runtimes

WORKLOADS	CLOUD DATAPROC	CLOUD DATAFLOW
Stream processing (ETL)		X
Batch processing (ETL)	X	X
Iterative processing and notebooks	X	
Machine learning with Spark ML	X	
Preprocessing for machine learning		X (with Cloud ML Engine)

[Return to Table of Contents](#)

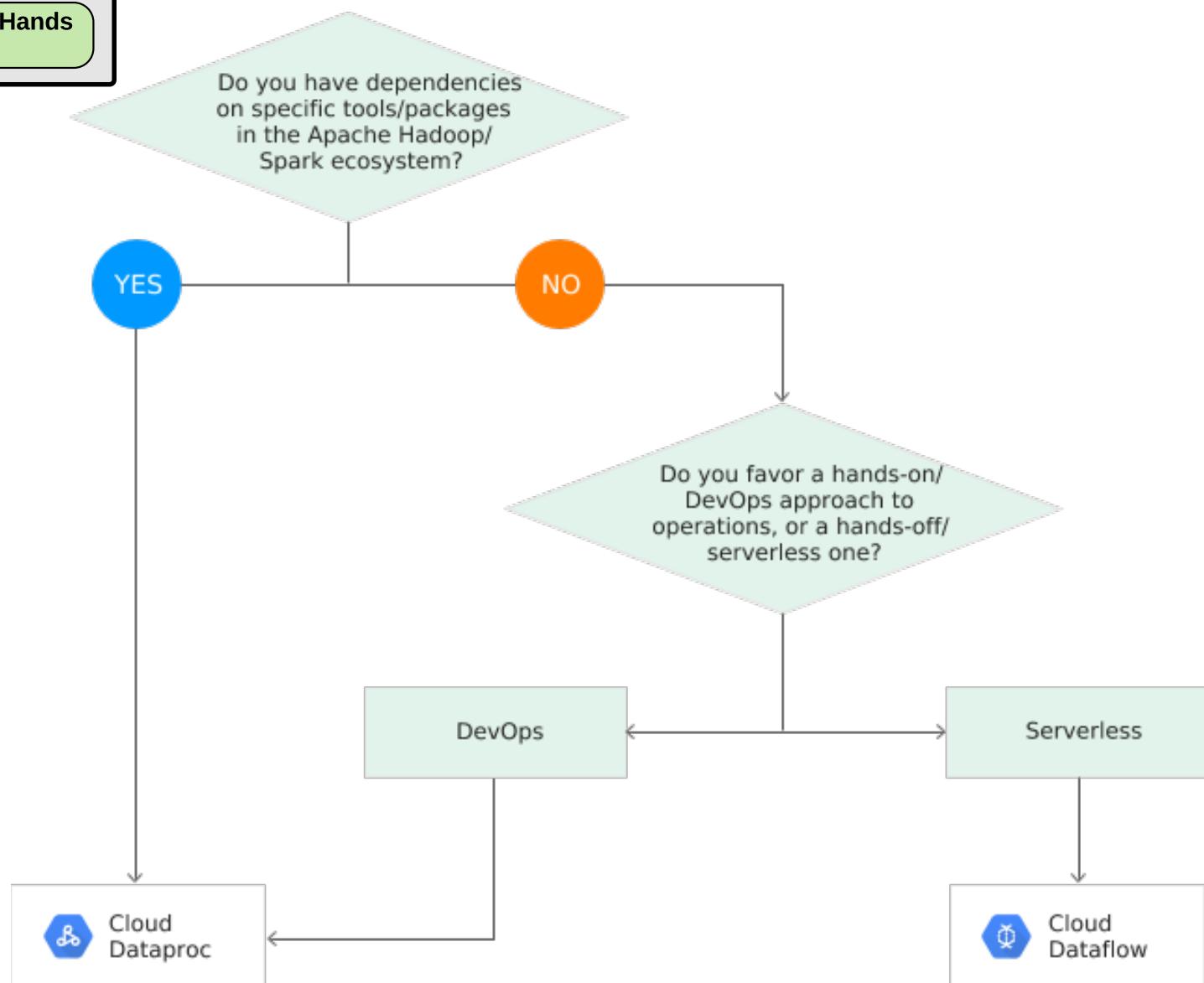
## Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

## Cloud Dataflow Overview

[Previous](#)

### Dataflow vs. Dataproc decision tree



[Return to Table of Contents](#)

## Key Concepts

### Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

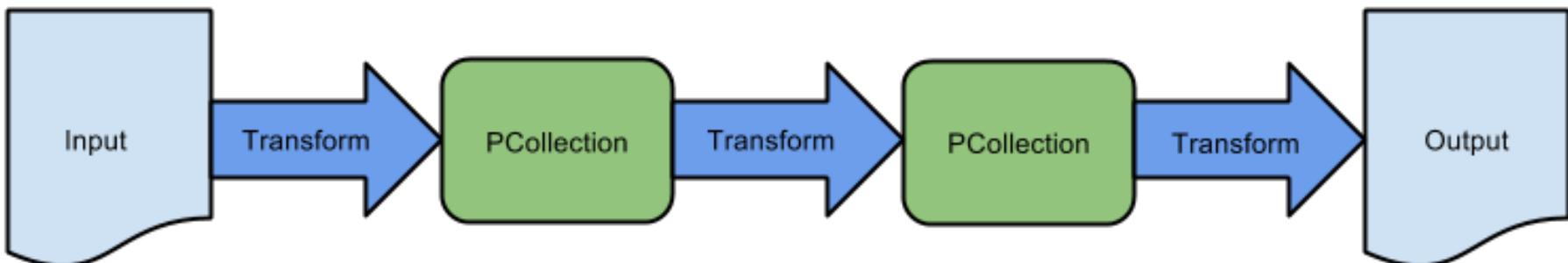
### Course/exam perspective:

[Next](#)

- Dataflow is very code-heavy
- Exam does not go deep into coding questions
- Some key concepts/terminology will be tested

### Key terms:

- Element - single entry of data (e.g., table row)
- PCollection - Distributed data set, data input and output
- Transform - Data processing operation (or step) in pipeline:
  - Uses programming conditionals (for/while loops, etc.)
- ParDo - Type of transform applied to individual elements:
  - Filter out/extract elements from a large group of data



**PCollection and ParDo in example Java code.**

**One step in a multi-step transformation process.**

```

PCollection<LaneInfo> currentConditions = p // 
    .apply("GetMessages", PubsubIO.readStrings().fromTopic(topic)) // 
    .apply("ExtractData", ParDo.of(new DoFn<String, LaneInfo>() {
        @ProcessElement
        public void processElement(ProcessContext c) throws Exception {
            String line = c.element();
            c.output(LaneInfo.newLineInfo(line));
        }
    }));
  
```

[Return to Table of Contents](#)

## Choose a Lesson

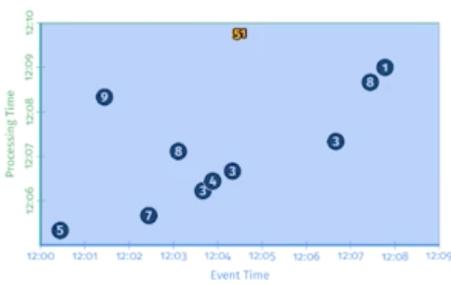
[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

## **Cloud Dataflow Overview**

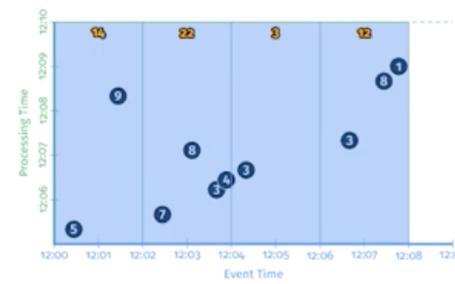
[Previous](#)

### Dealing with late/out of order data:

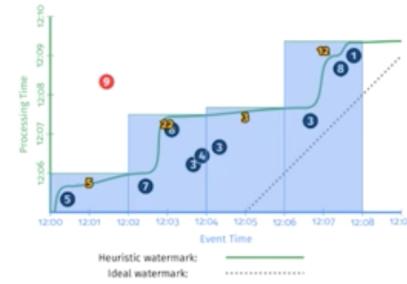
- Latency is to be expected (network latency, processing time, etc.)
- Pub/Sub does not care about late data, that is resolved in Dataflow
- Resolved with Windows, Watermarks, and Triggers
- Windows = logically divides element groups by time span
- Watermarks = 'timestamp':
  - Event time = when data was generated
  - Processing time = when data processed anywhere in the processing pipeline
  - Can use Pub/Sub-provided watermark or source-generated
- Trigger = determine when results in window are emitted (submitted as complete):
  - Allow late-arriving data in allowed time window to re-aggregate previously submitted results
  - Timestamps, element count, combinations of both



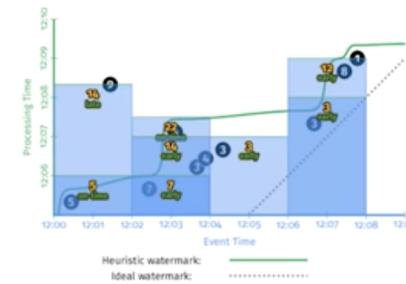
**1**  
**Classic  
Batch**



**2**  
**Windowed  
Batch**



**3**  
**Streaming  
w/Discarding**



**4**  
**Streaming  
+ Accumulation**

[Return to Table of Contents](#)

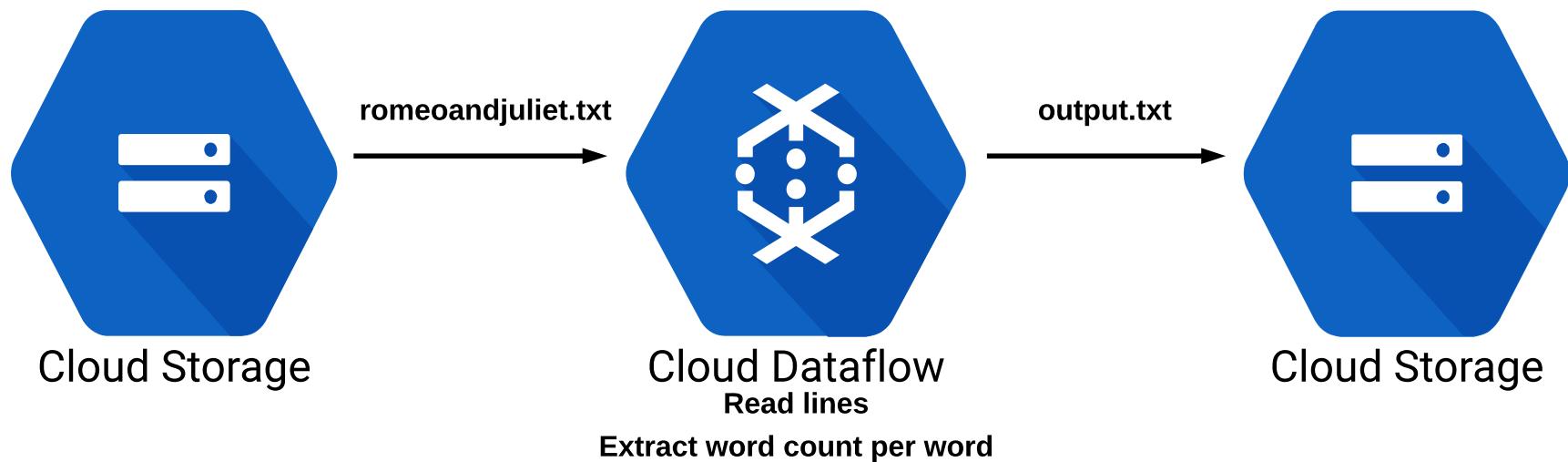
### Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

# The Data Dossier

## *Template Hands On*

- Google-provided templates
- Simple word count extraction



[Return to Table of Contents](#)

## Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)[Next](#)

## ***Streaming Ingest Pipeline Hands On***

- Take San Diego traffic data
- Ingest through Pub/Sub
- Process with Dataflow
- Analyze results with BigQuery
- First: Enable Dataflow API from API's and Services



Traffic data

Data ingest  
Published streaming sensor data



Topic: sandiego

Subscription pulls messages



Cloud Dataflow  
Transform data  
to calculate  
average speed.

Output to  
BigQuery.



BigQuery

[Return to Table of Contents](#)

## Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

## *Streaming Ingest Pipeline Hands On*

[Previous](#)[Next](#)

### Quick command line setup (Cloud Shell)

- Create BigQuery dataset for processing pipeline output:
  - bq mk --dataset \$DEVSHELL\_PROJECT\_ID:demos
- Create Cloud Storage bucket for Dataflow staging:
  - gsutil mb gs://\$DEVSHELL\_PROJECT\_ID
- Create Pub/Sub topic and stream data:
  - cd ~/googledataengineer/courses/streaming/publish
  - gcloud pubsub topics create sandiego
  - ./download\_data.sh
  - sudo pip install -U google-cloud-pubsub
  - ./send\_sensor\_data.py --speedFactor=60  
--project=\$DEVSHELL\_PROJECT\_ID

### Open a new Cloud Shell tab:

- Execute Dataflow pipeline for calculating average speed:
  - cd ~/googledataengineer/courses/streaming/process/sandiego
  - ./run\_oncloud.sh \$DEVSHELL\_PROJECT\_ID \$DEVSHELL\_PROJECT\_ID AverageSpeeds
- Error resolution:
  - Pub/Sub permission denied, re-authenticate
    - gcloud auth application-default login
  - Dataflow workflow failed - enable Dataflow API

[Return to Table of Contents](#)

## Choose a Lesson

[Data Processing Challenges](#)[Cloud Dataflow Overview](#)[Key Concepts](#)[Template Hands On](#)[Streaming Ingest Pipeline Hands On](#)

## *Streaming Ingest Pipeline Hands On*

[Previous](#)

### **View results in BigQuery:**

- List first 100 rows:
  - `SELECT * FROM [<PROJECTID>:demos.average_speeds]  
ORDER BY timestamp DESC LIMIT 100`
- Show last update to table:
  - `SELECT MAX(timestamp) FROM  
[<PROJECTID>:demos.average_speeds]`
- Look at results from the last minute:
  - `SELECT * FROM  
[<PROJECTID>:demos.average_speeds@-60000] ORDER BY  
timestamp DESC`

### **Shut down pipeline:**

- Drain - finishing processing buffered jobs before shutting down
- Cancel - full stop, cancels existing buffered jobs

[Return to Table of Contents](#)

## **Choose a Lesson**

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

[Return to Table of Contents](#)

## Choose a Lesson

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

## Managed Hadoop/Spark Stack

Custom Code

Monitoring/Health

Dev Integration

Manual Scaling

Job Submission

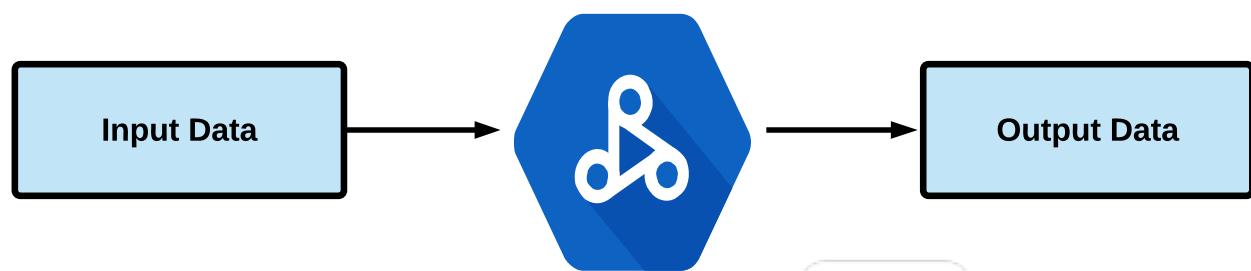
Google Cloud Connectivity

Deployment

Creation

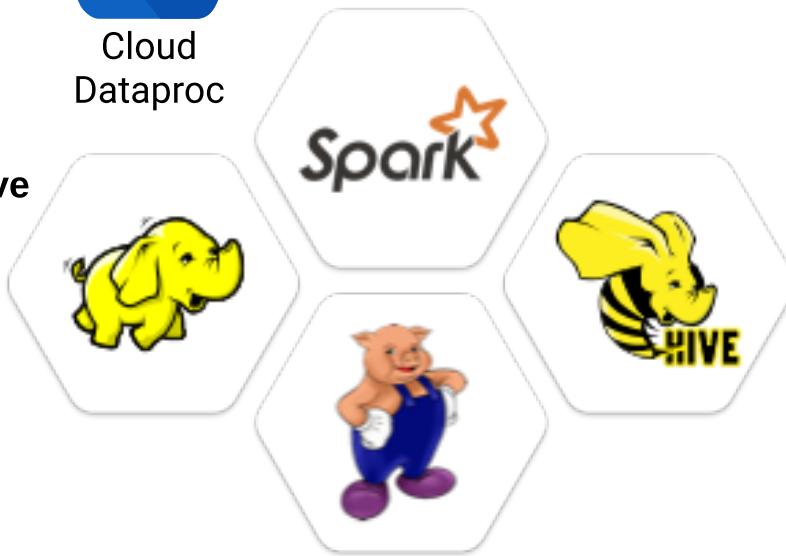
## Dataproc Overview

### What is Cloud Dataproc?

[Next](#)

#### Hadoop ecosystem:

- Hadoop, Spark, Pig, Hive
- Lift and shift to GCP



#### Dataproc facts:

- On-demand, managed Hadoop and Spark clusters
- Managed, but not no-ops:
  - Must configure cluster, not auto-scaling
  - Greatly reduces administrative overhead
- Integrates with other Google Cloud services:
  - Separate data from the cluster - save costs
- Familiar Hadoop/Spark ecosystem environment:
  - Easy to move existing projects
- Based on Apache Bigtop distribution:
  - Hadoop, Spark, Hive, Pig
- HDFS available (but maybe not optimal)
- Other ecosystem tools can be installed as well via initialization actions

[Return to Table of Contents](#)

## Choose a Lesson

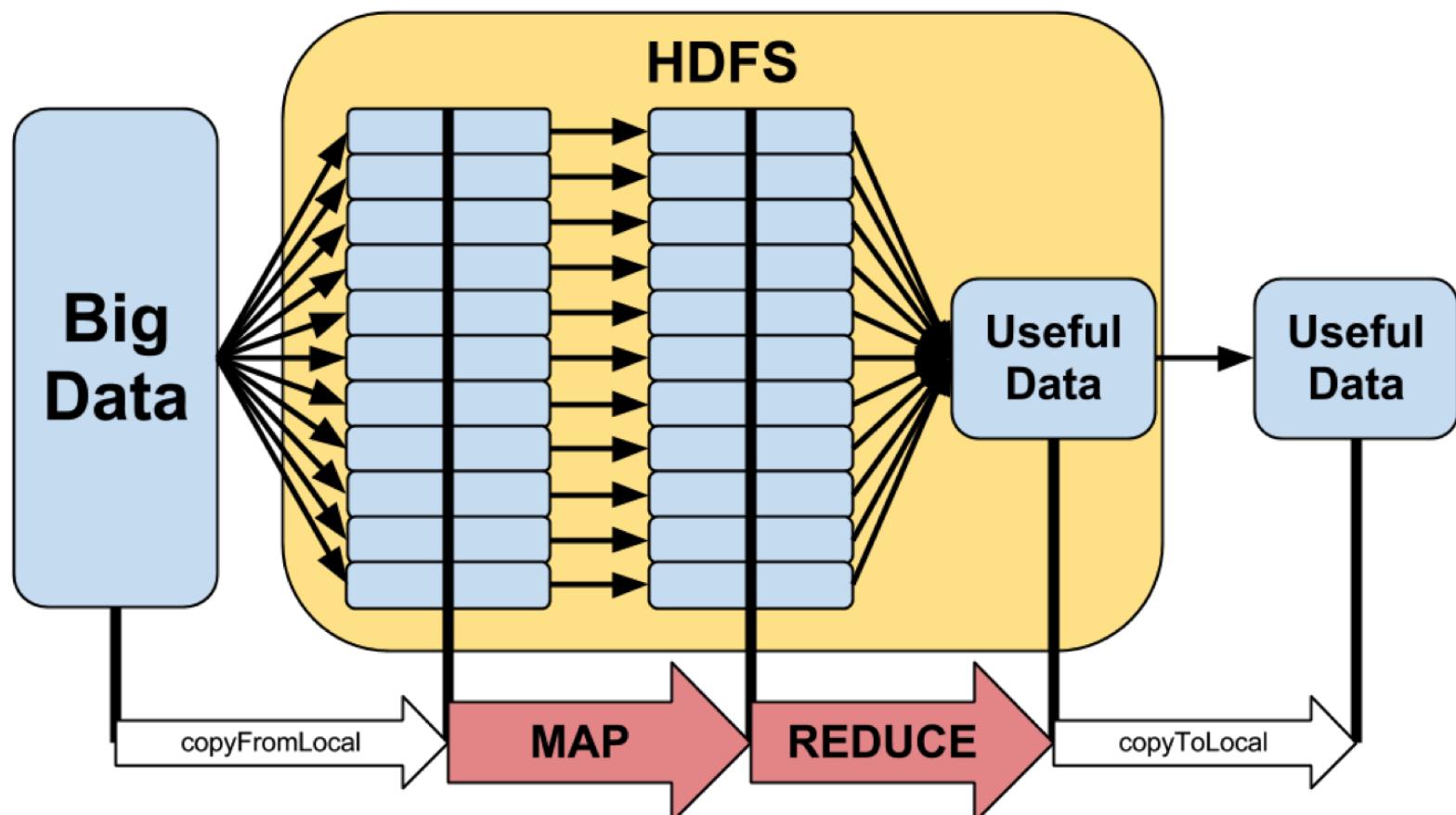
[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

## Dataproc Overview

[Previous](#)[Next](#)

### What is MapReduce?

- Simple definition:
  - Take big data, distribute it to many workers (map)
  - Combine results of many pieces (reduce)
- Distributed/parallel computing



[Return to Table of Contents](#)

## Choose a Lesson

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

## Dataproc Overview

[Previous](#)[Next](#)

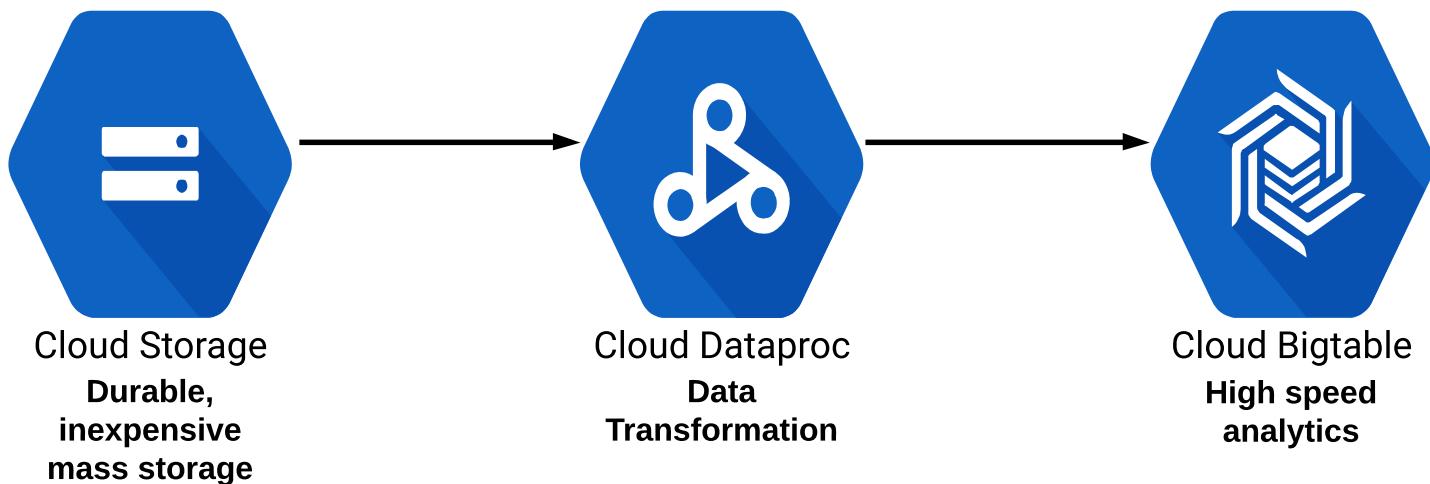
### Pricing:

- Standard Compute Engine machine type pricing + managed Dataproc premium
- Premium = \$0.01 per vCPU core/hour

Machine type	Virtual CPUs	Memory	Dataproc
n1-highcpu-2	2	1.80GB	\$0.020
n1-highcpu-4	4	3.60GB	\$0.040
n1-highcpu-8	8	7.20GB	\$0.080
n1-highcpu-16	16	14.40GB	\$0.160
n1-highcpu-32	32	28.80GB	\$0.320
n1-highcpu-64	64	57.60GB	\$0.640

### Data Lifecycle Scenario

#### Data Ingest, Transformation, and Analysis



[Return to Table of Contents](#)

### Choose a Lesson

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

# The Data Dossier

## Dataproc Overview

[Previous](#)

### Identity and Access Management (IAM):

- Project level only (primitive and predefined roles)
- Cloud Dataproc Editor, Viewer, Worker
- Editor - Full access to create/delete/edit clusters/jobs/workflows
- Viewer - View access only
- Worker - Assigned to service accounts:
  - Read/write GCS, write to Cloud Logging

[Return to Table of Contents](#)

## Choose a Lesson

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

## Configure Dataproc Cluster

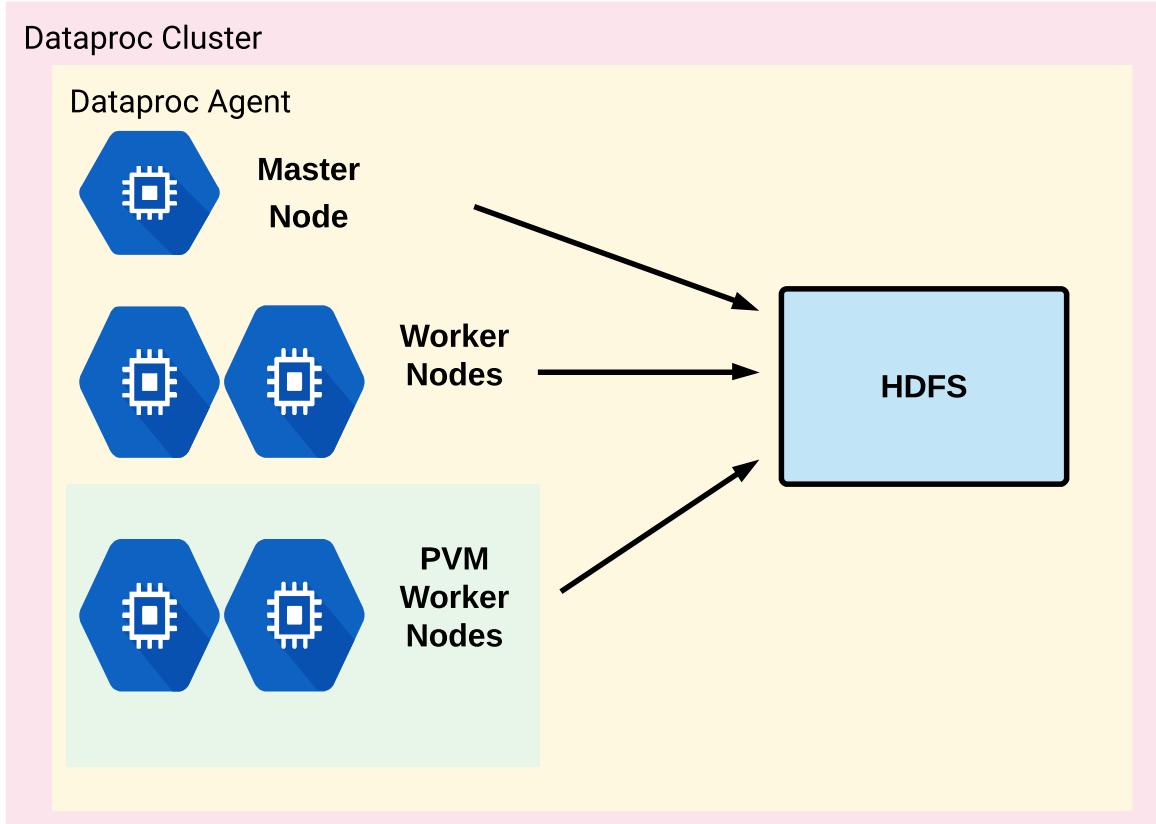
[Next](#)

### Create cluster:

- `gcloud dataproc clusters create [cluster_name] --zone [zone_name]`
- **Configure master node, worker nodes:**
  - Master contains YARN resource manager
  - YARN = Yet Another Resource Negotiator

### Updating clusters:

- Can only change # workers/preemptible VM's/labels/toggle graceful decommission
- Automatically reshards data for you
- `gcloud dataproc clusters update [cluster_name] --num-workers [#] --num-preemptible-workers [#]`



[Return to Table of Contents](#)

## Choose a Lesson

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)[Previous](#)

## Configure Dataproc Cluster

### Preemptible VM's on Dataproc:

- Excellent low-cost worker nodes
- Dataproc manages the entire leave/join process:
  - No need to configure startup/shutdown scripts
  - Just add PVM's...and that's it
- No assigned disks for HDFS (only disk for caching)
- Want a mix of standard + PVM worker nodes

### Access your cluster:

- SSH into master - same as any compute engine instance
- gcloud compute ssh [master\_node\_name]

### Access via web - 2 options:

- Open firewall ports to your network (8088, 9870)
- Use SOCKS proxy - does not expose firewall ports

### SOCKS proxy configuration:

- SSH to master to enable port forwarding:
  - gcloud compute ssh *master-host-name* --project=*project-id* --zone=*master-host-zone* -- -D 1080 -N
- Open new terminal window - launch web browser with parameters (varies by OS/browser):
  - "/Applications/Google Chrome.app/Contents/MacOS/Google Chrome" --proxy-server="socks5://localhost:1080" --host-resolver-rules="MAP \* 0.0.0.0 , EXCLUDE localhost" --user-data-dir=/tmp/cluster1-m
- Browse to [http://\[master\]:port](http://[master]:port):
  - 8088 - Hadoop
  - 9870 - HDFS

### Using Cloud Shell (must use for each port):

- gcloud compute ssh *master-host-name* --project=*project-id* --zone *master-host-zone* -- -4 -N -L *port1:master-host-name:port2*
- Use Web Preview to choose port (8088/9870)

[Return to Table of Contents](#)

## Choose a Lesson

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

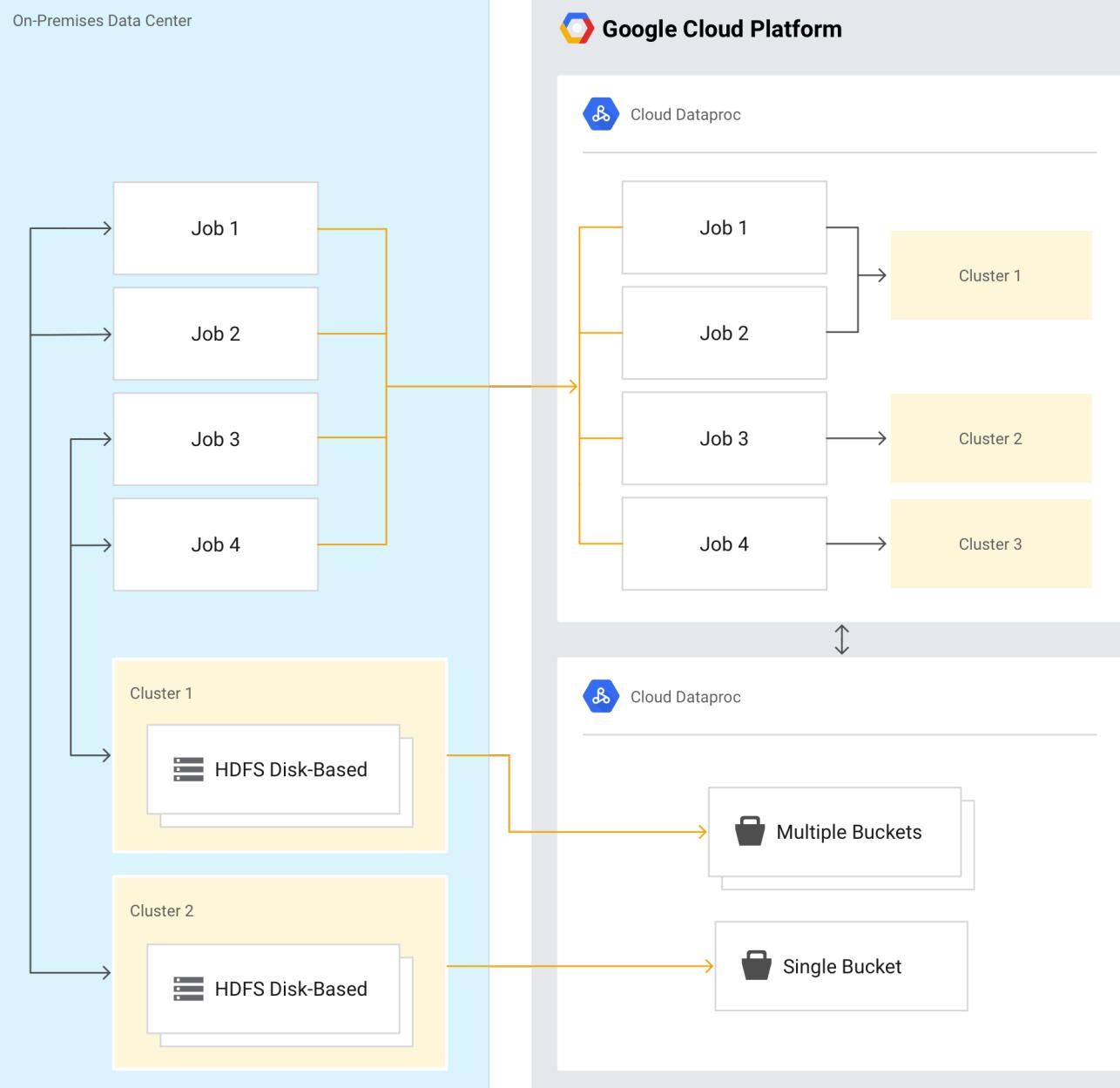
## Migrating and Optimizing for Google Cloud

### Migrating to Cloud Dataproc

[Next](#)

#### What are we moving/optimizing?

- Data (from HDFS)
- Jobs (pointing to Google Cloud locations)
- Treating clusters as ephemeral (temporary) rather than permanent entities



Install Cloud Storage connector to connect to GCS (Google Cloud Storage).

[Return to Table of Contents](#)

## Choose a Lesson

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

## *Migrating and Optimizing for Google Cloud*

[Previous](#)[Next](#)

### Migration Best Practices:

- Move data first (generally Cloud Storage buckets):
  - Possible exceptions:
    - Apache HBase data to Bigtable
    - Apache Impala to BigQuery
    - Can still choose to move to GCS if Bigtable/BQ features not needed
- Small-scale experimentation (proof of concept):
  - Use a subset of data to test
- Think of it in terms of ephemeral clusters
- Use GCP tools to optimize and save costs

[Return to Table of Contents](#)

## Migrating and Optimizing for Google Cloud

### Choose a Lesson

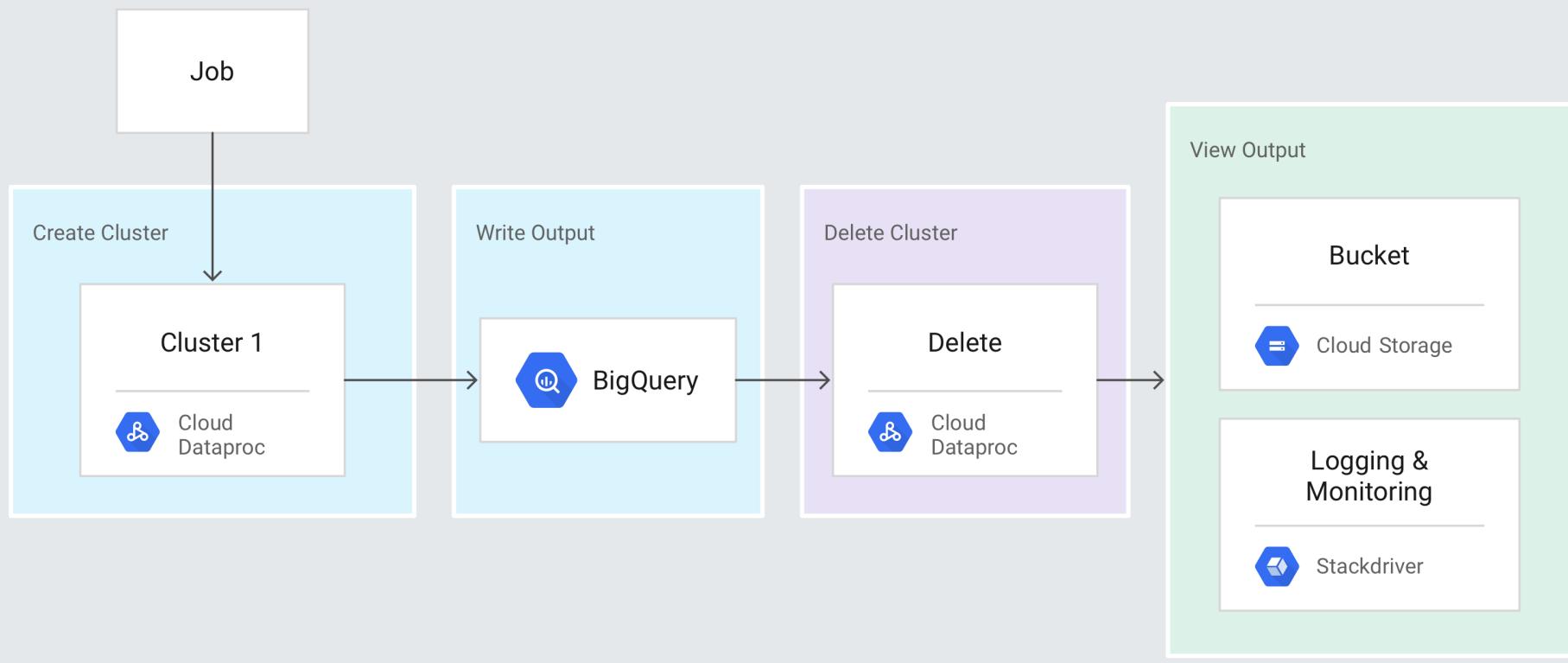
[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)[Previous](#)[Next](#)

### Optimize for the Cloud ("Lift and Leverage")

Separate storage and compute (cluster):

- Save on costs:
  - No need to keep clusters to keep/access data
- Simplify workloads:
  - No shaping workloads to fit hardware
  - Simplify storage capacity
- HDFS --> Google Cloud Storage
- Hive --> BigQuery
- HBase --> Bigtable

### Google Cloud Platform



[Return to Table of Contents](#)

## Choose a Lesson

[Dataproc Overview](#)[Configure Dataproc Cluster and Submit Job](#)[Migrating and Optimizing for Google Cloud](#)

## *Migrating and Optimizing for Google Cloud*

[Previous](#)

### Converting from HDFS to Google Cloud Storage:

#### 1. Copy data to GCS:

- Install connector or copy manually

#### 2. Update file prefix in scripts:

- From `hdfs://` to `gs://`

#### 3. Use Dataproc, and run against/output to GCS

The end goal should be to eventually move toward a cloud-native and serverless architecture (Dataflow, BigQuery, etc.).

[Return to Table of Contents](#)

## **Choose a Lesson**

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)[Next](#)

## BigQuery Overview

### What is BigQuery?

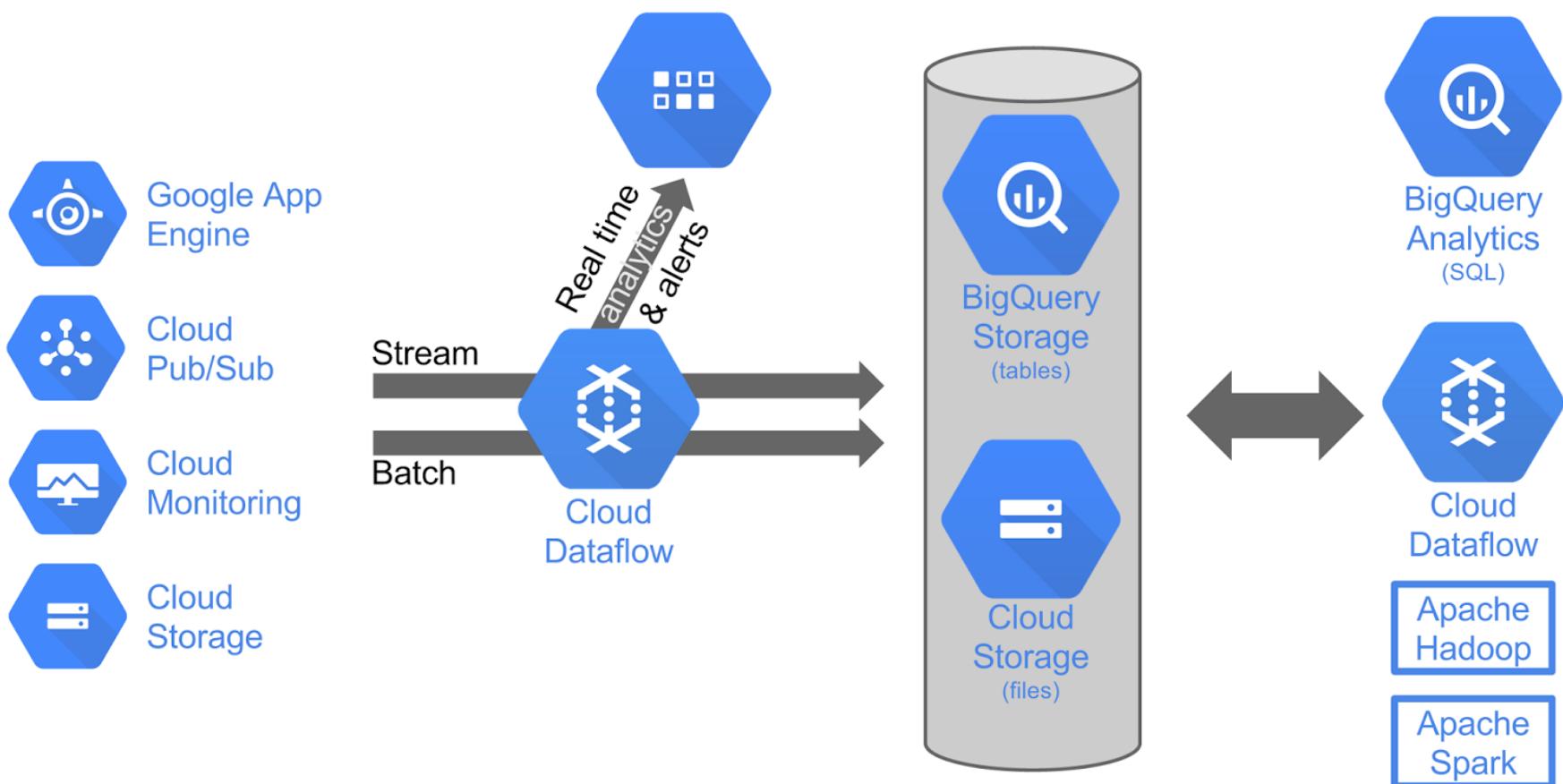
- Fully Managed Data warehousing
  - Near-real time analysis of petabyte scale databases
- Serverless (no-ops)
- Auto-scaling to petabyte range
- Both storage and analysis
- Accepts batch and streaming loads
- Locations = multi-regional (US, EU), Regional (asia-northeast1)
- Replicated, durable
- Interact primarily with standard SQL (also Legacy SQL)
  - [SQL Primer course](#)

Ingest

Process

Store

Analyze



[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## BigQuery Overview

[Previous](#)[Next](#)

### How BigQuery works

- Part of the "3rd wave" of cloud computing
  - Google Big Data Stack 2.0
- Focus on serverless compute, real time insights, machine learning...
  - ...instead of data placement, cluster configuration
  - No managing of infrastructure, nodes, clusters, etc

2002      2004      2005      2006      2008      2010      2012      2014      2015



Google Papers



GFS      MapReduce      BigTable      Dremel      PubSub      Flume Java      Millwheel      Dataflow      TensorFlow



Cloud Dataproc



BigQuery



Pub/Sub



Cloud Dataflow

Google Big Data Stack 1.0

Google Big Data Stack 2.0

[Return to Table of Contents](#)

## Choose a Lesson

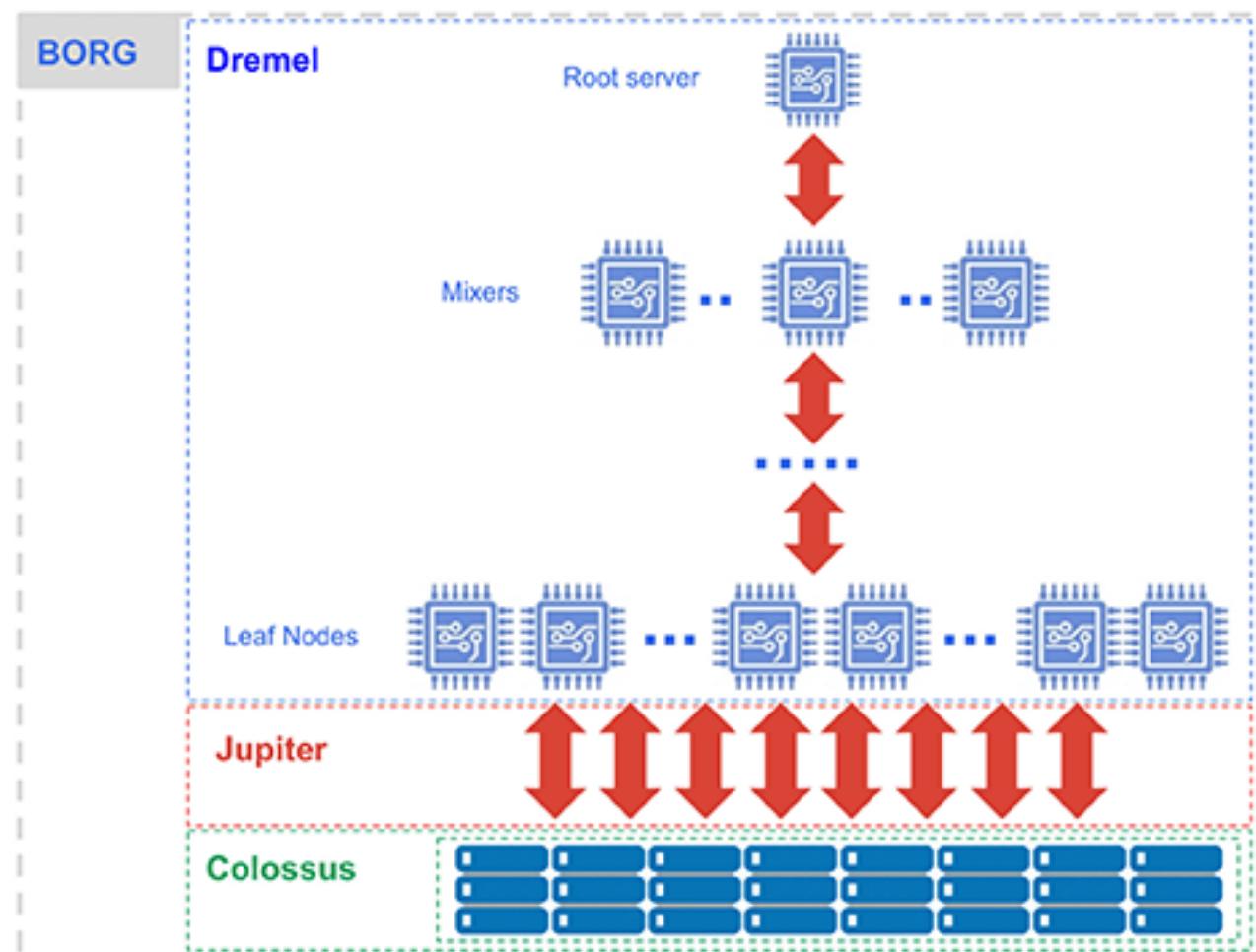
[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## BigQuery Overview

[Previous](#)[Next](#)

### How BigQuery works (cont)

- Jobs (queries) can scale up to thousands of CPU's across many nodes, but the process is completely invisible to end user
- Storage and compute are separated, connected by petabit network



[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## ***BigQuery Overview***

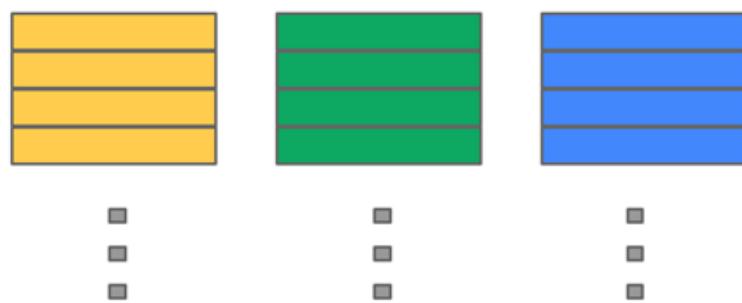
[Previous](#)[Next](#)

### **How BigQuery works (cont)**

- **Columnar data store**
  - Separates records into column values, stores each value on different storage volume
  - Traditional RDBMS stores whole record on one volume
  - Extremely fast read performance, poor write (update) performance - BigQuery does not update existing records
  - Not transactional



Record Oriented Storage



Column Oriented Storage

[Return to Table of Contents](#)

## Choose a Lesson

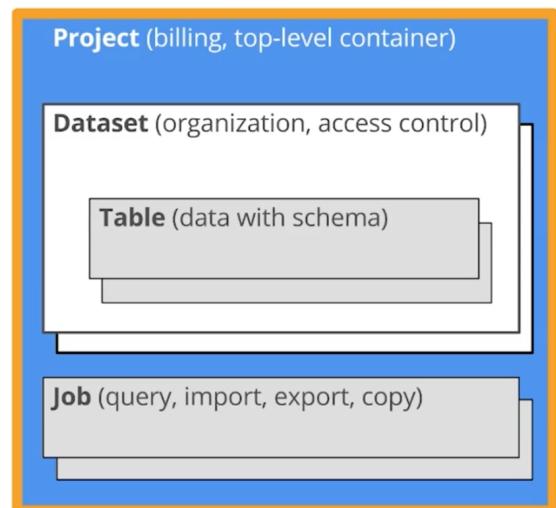
[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## *BigQuery Overview*

[Previous](#)[Next](#)

### BigQuery structure

- **Dataset** - contains tables/views
- **Table** = collection of columns
- **Job** = long running action/query



### Identity and Access Management (IAM)

- Control by project, dataset, view
- Cannot control at table level
  - But can control by views via datasets as alternative (virtual table defined by SQL query)
- Predefined roles - BigQuery...
  - Admin - full access
  - Data Owner - full dataset access
  - Data Editor - edit dataset tables
  - Data Viewer - view datasets and tables
  - Job User - run jobs
  - User - run queries and create datasets (but not tables)
- [Roles comparison matrix](#)
- Sharing datasets
  - Make public with All Authenticated Users

[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## *BigQuery Overview*

[Previous](#)

### Pricing

- Storage, Queries, Streaming insert
- Storage = \$0.02/GB/mo (first 10GB/mo free)
  - Long term storage (not edited for 90 days) = \$0.01/GB/mo
- Queries = \$5/TB (first TB/mo free)
- Streaming = \$0.01/200 MB
- Pay as you go, with high end flat-rate query pricing
- Flat rate - starts at \$40K per month with 2000 slots

[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## *Interacting with BigQuery*

### Interaction methods

[Next](#)

- Web UI
- Command line (bq commands)
  - bq query --arguments 'QUERY'
- Programmatic (REST API, client libraries)
- Interact via queries

### Querying tables

- FROM `project.dataset.table` (Standard SQL)
- FROM [project:dataset.table] (Legacy SQL)

### Searching multiple tables with wildcards

Query across multiple, similarly named tables

- FROM `project.dataset.table\_prefix\*`

Filter further in WHERE clause

- AND \_TABLE\_SUFFIX BETWEEN 'table003' and 'table050'

### Advanced SQL queries are allowed

- JOINS, sub queries, CONCAT

[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)[Previous](#)

## Views

- Virtual table defined by query
- 'Querying a query'
- Contains data only from query that contains view
- Useful for limiting table data to others

## Cached queries

- Queries cost money
- Previous queries are cached to avoid charges if ran again
- command line to disable cached results
  - `bq query --no_use_cache '(QUERY)'`
- Caching is per user only

## User Defined Functions (UDF)

- Combine SQL code with JavaScript/SQL functions
- Combine SQL queries with programming logic
- Allow much more complex operations (loops, complex conditionals)
- WebUI only usable with Legacy SQL

[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## Data formats:

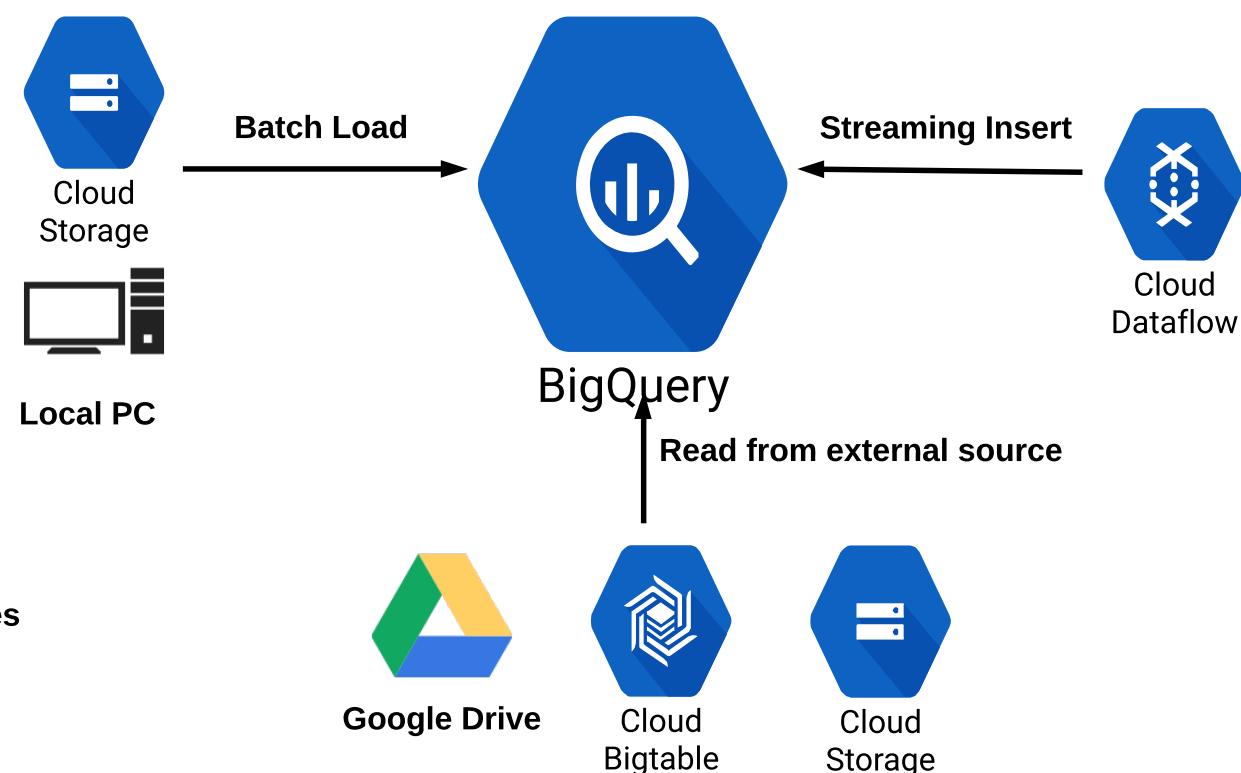
### Load

- CSV
- JSON (Newline delimited)
- Avro - best for compressed files
- Parquet
- Datastore backups

### Read

- CSV
- JSON (Newline delimited)
- Avro
- Parquet

## Loading and reading sources

[Next](#)

## Why use external sources?

- Load and clean data in one pass from external, then write to BigQuery
- Small amount of frequently changing data to join to other tables

## Loading data with command line

- `bq load --source_format=[format] [dataset].[table] [source_path] [schema]`
- Can load multiple files with command line (not WebUI)

[Return to Table of Contents](#)

## Choose a Lesson

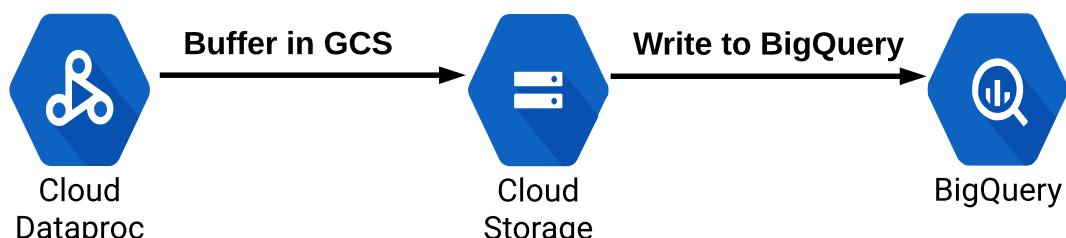
[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## Load and Export Data

[Previous](#)

### Connecting to/from other Google Cloud services

- Dataproc - Use BigQuery connector (installed by default), job uses Cloud Storage for staging



### Exporting tables

- Can only export to Cloud Storage
- Can copy table to another BigQuery dataset
- Export formats: CSV, JSON, Avro
- Can export multiple tables with command line
- Can only export up to 1GB per file, but can split into multiple files with wildcards
- Command line
  - `bq extract 'projectid:dataset.table' gs://bucket_name/folder/object_name`
  - Can drop 'project' if exporting from same project
  - Default is CSV, specify other format with `--destination_format`
  - `--destination_format=NEWLINE_DELIMITED_JSON`

### BigQuery Transfer Service

- Import data to BigQuery from other Google advertising SaaS applications
- Google AdWords
- DoubleClick
- YouTube reports

[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## ***Optimize for Performance and Costs***

### Performance and costs are complementary

[Next](#)

- Less work = faster query = less costs
- What is 'work'?
  - I/O - how many bytes read?
  - Shuffle - how much passed to next stage
  - How many bytes written?
  - CPU work in functions

### General best practices

- Avoid using **SELECT \***
- Denormalize data when possible
  - Grouping data into single table
  - Often with nested/repeated data
  - Good for read performance, not for write (transactional) performance
- Filter early and big with **WHERE** clause
- Do biggest joins first, and filter pre-JOIN
- **LIMIT** does not affect cost
- Partition data by date
  - Partition by ingest time
  - Parition by specified data columns

[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

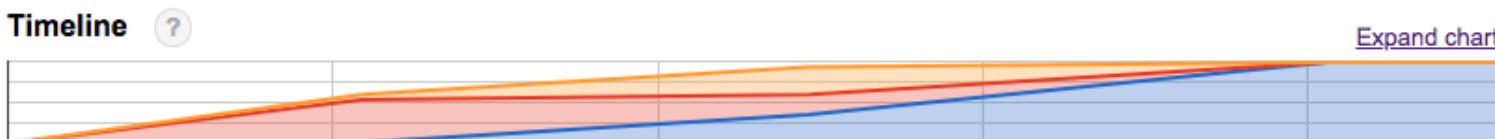
## Optimize for Performance and Costs

[Next](#)

### Monitoring query performance

- Understand color codes
- Understand 'skew' in difference between average and max time

#### Timeline



#### Execution Plan

	Stage timing					Parallel Inputs	Rows	
		Wait	Read	Compute	Write		Input	Output
▶ S00: Input	✓	█	███████	███	███	105	122 M	5.34 K (78.2 KB)
▶ S02: Coalesce	✓	██	██	██	██	100	5.34 K	5.34 K (78.2 KB)
▶ S03: Join+	✓	██	██	██	██	74	22.1 M	37 (925 B)
▶ S04: Aggregate+	✓	██	██	██	██	17	37	19 (646 B)
▶ S05: Output	✓	█	██	██	██	1	19	19 (532 B)

[Return to Table of Contents](#)

## Choose a Lesson

[BigQuery Overview](#)[Interacting with BigQuery](#)[Load and Export Data](#)[Optimize for Performance and Costs](#)[Streaming Insert Example](#)

## *Streaming Insert Example*

### Quick setup

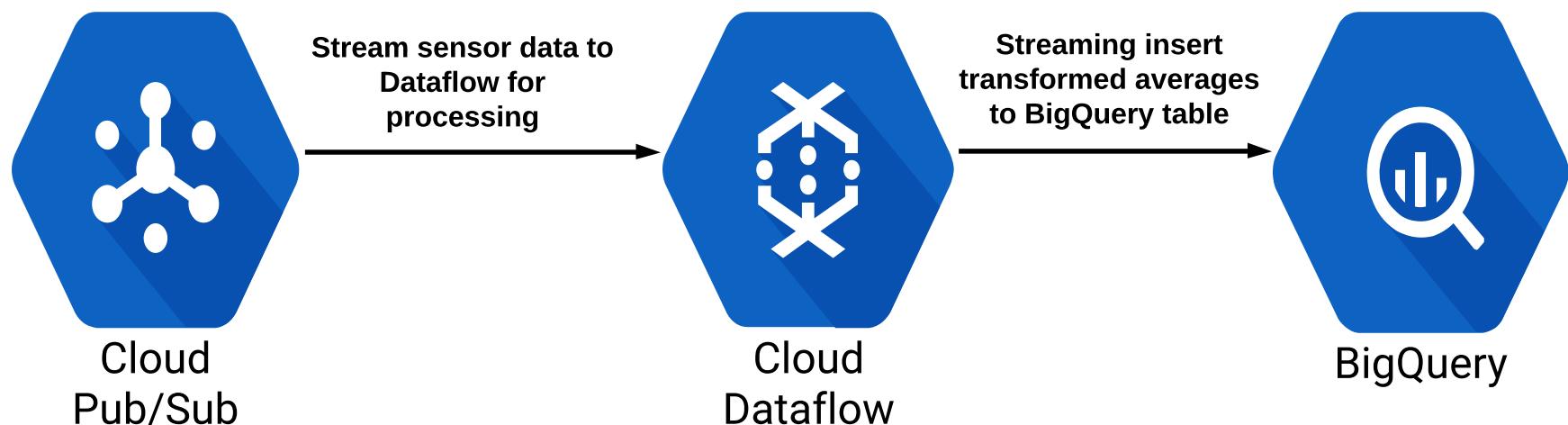
```
cd
```

```
gsutil cp -r gs://gcp-course-exercise-scripts/data-engineer/* .  
bash streaming-insert.sh
```

### Clean up

```
bash streaming-cleanup.sh
```

Manually stop Dataflow job



[Return to Table of Contents](#)

## **Choose a Lesson**

[What is Machine Learning?](#)[Working with Neural Networks](#)

# The Data Dossier

[Return to Table of Contents](#)

## Choose a Lesson

[What is Machine Learning?](#)

[Working with Neural Networks](#)

## What is Machine Learning?

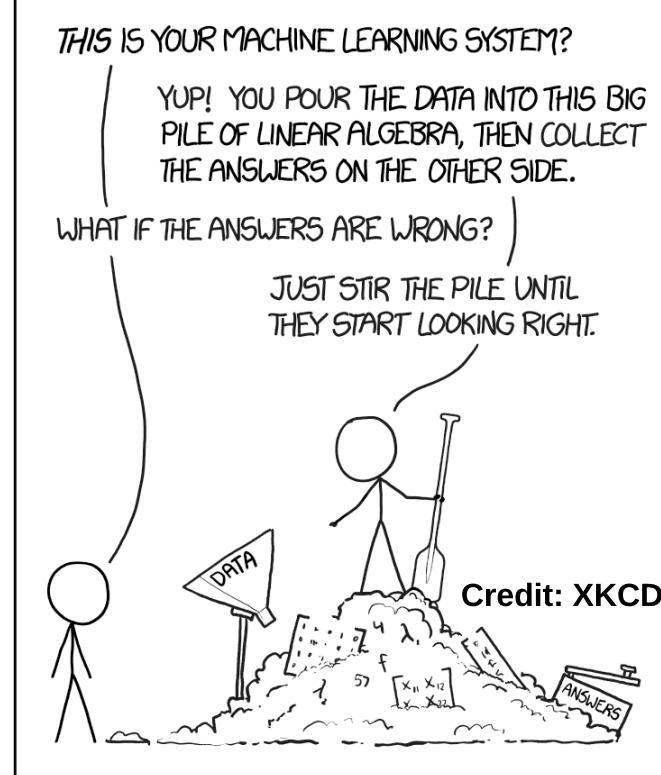
Popular view of machine learning...

[Next](#)

DATA →

→ MAGIC!

For Data Engineer:  
Know the training and inference  
stages of ML



## So what is machine learning?

Process of combining inputs to produce useful predictions on never-before-seen data

Makes a machine learn from data to make predictions on future data, instead of programming every scenario



New, unlabeled  
image



"I have never seen  
this image before,  
but I'm pretty sure  
that this is a cat!"

[Return to Table of Contents](#)

## Choose a Lesson

[What is Machine Learning?](#)
[Working with Neural Networks](#)

## *What is Machine Learning?*

[Previous](#)
[Next](#)

### How it works

- Train a model with examples
- Example = input + label
- Training = adjust model to learn relationship between features and label - minimize error:
  - Optimize weights and biases (parameters) to different input features
- Feature = input variable(s)
- Inference = apply trained model to unlabeled examples
- Separate test and training data ensures model is generalized for additional data:
  - Otherwise, leads to overfitting (only models to training data, not new data)

Input + Label



"Cat"

Train on many examples  
Training dataset

Everything is numbers!



Train on ML model

"I think this is a cat"

Predict with trained model



Match labels by adjusting weights to input features

Array RGB									
Page 3 - blue intensity values									
0.342	0.647	0.515	0.816	...	0.689	0.703	0.118	0.884	...
0.111	0.300	0.205	0.526	...	0.538	0.538	0.653	0.925	...
0.523	0.428	0.712	0.929	...	0.314	0.268	0.159	0.101	...
0.214	0.604	0.918	0.344	...	0.553	0.633	0.528	0.493	...
0.100	0.121	0.173	0.126	...	0.441	0.465	0.512	0.512	...
0.204	0.175	...	...	...	0.298	0.401	0.421	0.398	...
0.760	0.531	...	...	...	0.912	0.713	...	...	...
0.997	0.910	...	...	...	0.219	0.328	...	...	...
0.995	0.726	...	...	...	0.128	0.133	...	...	...

Page 1 - red intensity values

Page 2 - green intensity values

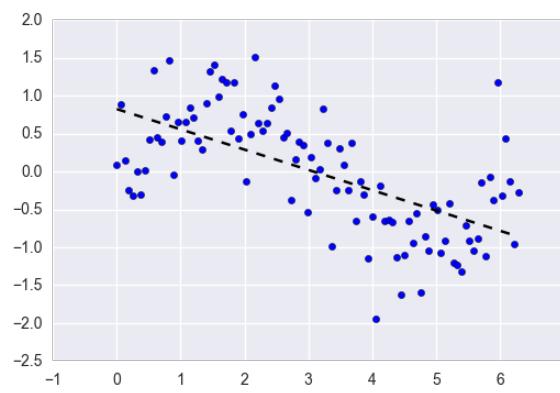
n-dimensional arrays called 'tensor', hence TensorFlow

[Return to Table of Contents](#)

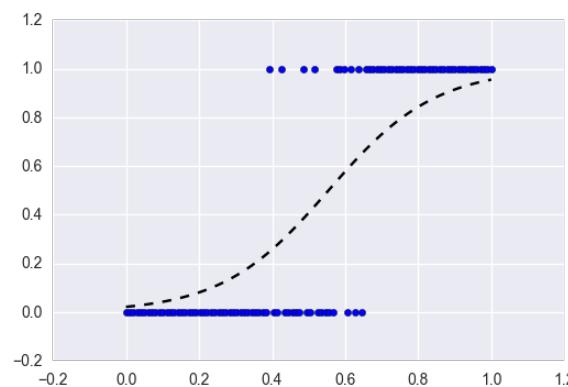
## Choose a Lesson

[What is Machine Learning?](#)[Working with Neural Networks](#)

### Regression



### Classification



## What is Machine Learning?

[Previous](#)

### Learning types

- Supervised learning
  - Apply labels to data ("cat", "spam")
  - Regression - Continuous, numeric variables:
    - Predict stock price, student test scores
  - Classification - categorical variables:
    - yes/no, decision tree
    - "is this email spam?" "is this picture a cat?"
  - Same types for dataset columns:
    - continuous (regression) and categorical (classification)
    - income, birth year = continuous
    - gender, country = categorical
- Unsupervised learning
  - Clustering - finding patterns
  - Not labeled or categorized
  - "Given the location of a purchase, what is the likely amount purchased?"
  - Heavily tied to statistics
- Reinforcement Learning
  - Use positive/negative reinforcement to complete a task
    - Complete a maze, learn chess

[Return to Table of Contents](#)

## Choose a Lesson

[What is Machine Learning?](#)[Working with Neural Networks](#)

## **Working with Neural Networks**

Hands on learning tool

[playground.tensorflow.org](http://playground.tensorflow.org)

[Next](#)

### Key terminology

- Neural network - model composed of layers, consisting of connected units (neurons):
  - Learns from training datasets
- Neuron - node, combines input values and creates one output value
- Input - what you feed into a neuron (e.g. cat pic)
- Feature - input variable used to make predictions
  - Detecting email spam (subject, key words, sender address)
  - Identify animals (ears, eyes, colors, shapes)
- Hidden layer - set of neurons operating from same input set
- Feature engineering - deciding which features to use in a model
- Epoch - single pass through training dataset
  - Speed up training by training on a subset of data vs. all data

### Making Adjustments with Parameters

- Weights - multiplication of input values
- Bias - value of output given a weight of 0
- ML adjusts these parameters automatically
- Parameters = variables adjusted by training with data

$$w_1x_1 + w_2x_2 > b$$

[Return to Table of Contents](#)

## Choose a Lesson

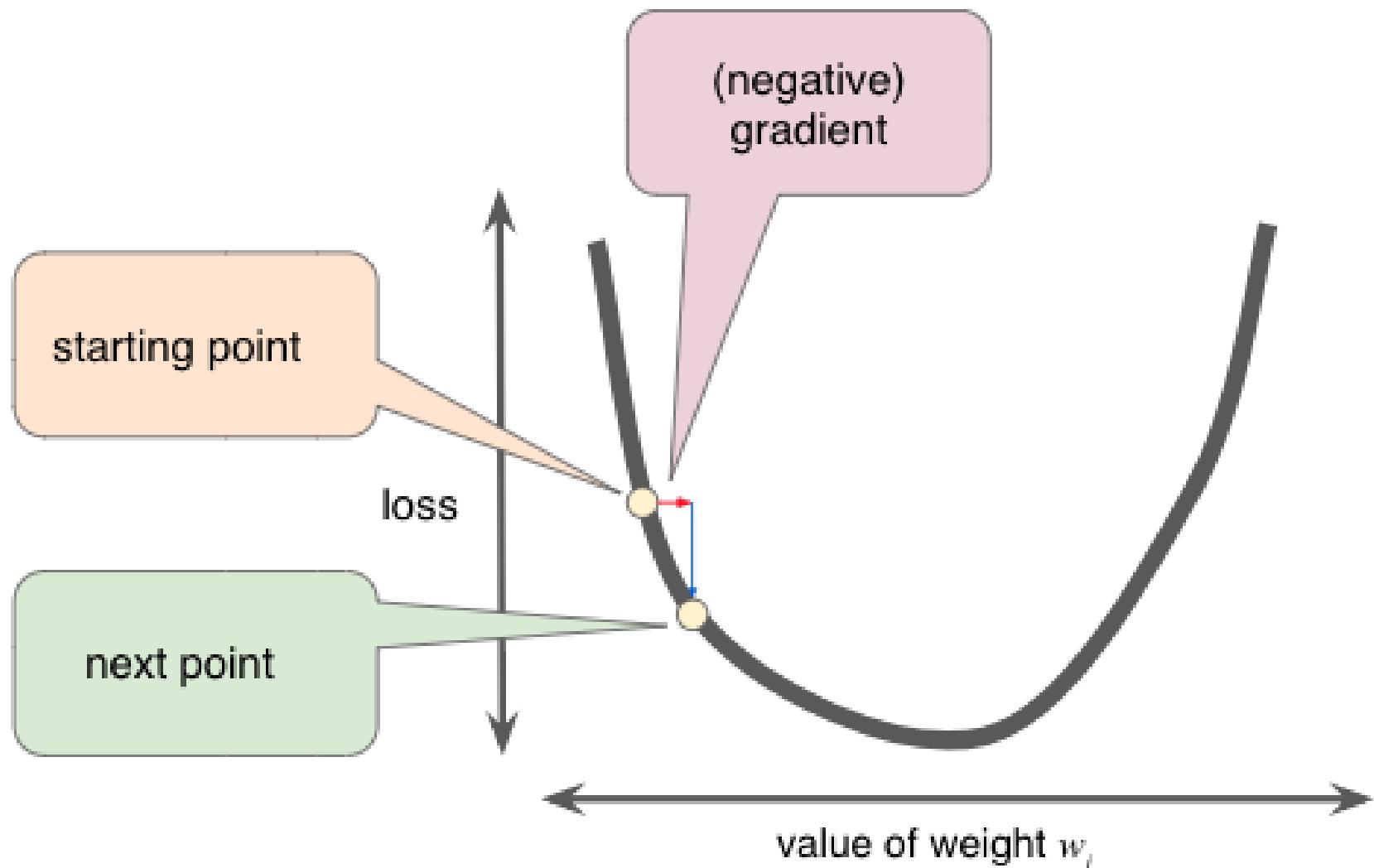
[What is Machine Learning?](#)[Working with Neural Networks](#)

## Working with Neural Networks

[Previous](#)[Next](#)

### Rate of adjustments with Learning Rate

- Magnitude of adjustments of weights and biases
- Hyperparameter = variables about the training process itself:
  - Also includes hidden layers
  - Not related to training data
- Gradient descent - technique to minimize loss (error rate)
- Challenge is to find the correct learning rate:
  - Too small - takes forever
  - Too large - overshoots



[Return to Table of Contents](#)

## Choose a Lesson

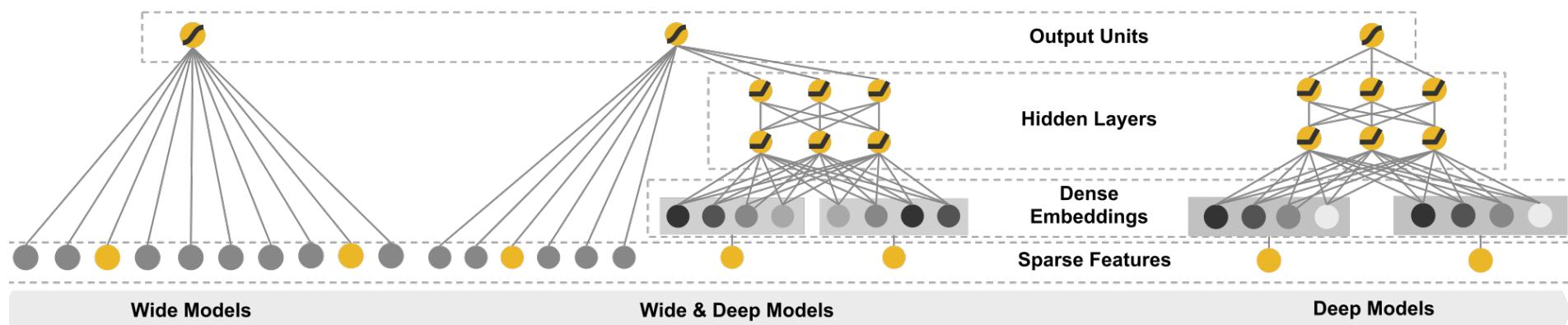
[What is Machine Learning?](#)[Working with Neural Networks](#)

## *Working with Neural Networks*

[Previous](#)

### Deep and wide neural networks

- **Wide - memorization:**
  - Many features
- **Deep - generalization:**
  - Many hidden layers
- **Deep and wide = both:**
  - Good for recommendation engines



[Return to Table of Contents](#)

## **Choose a Lesson**

[ML Engine Overview](#)[ML Engine Hands On](#)

[Return to Table of Contents](#)

## Choose a Lesson

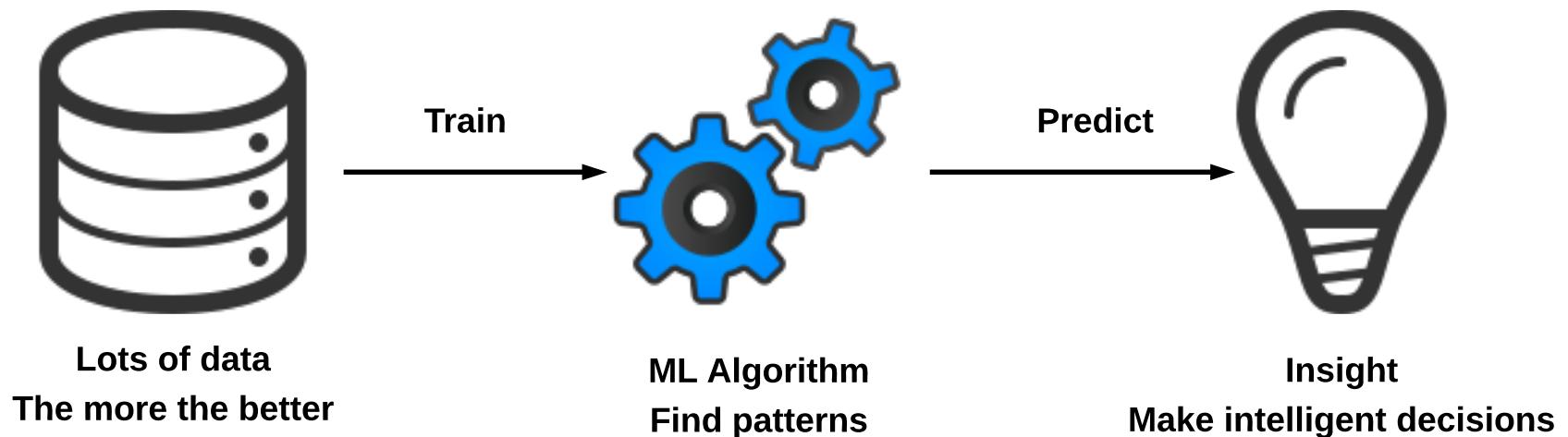
[ML Engine Overview](#)[ML Engine Hands On](#)

## *ML Engine Overview*

### Machine Learning - in a nutshell

[Next](#)

- Algorithm that is able to learn from data



[Return to Table of Contents](#)

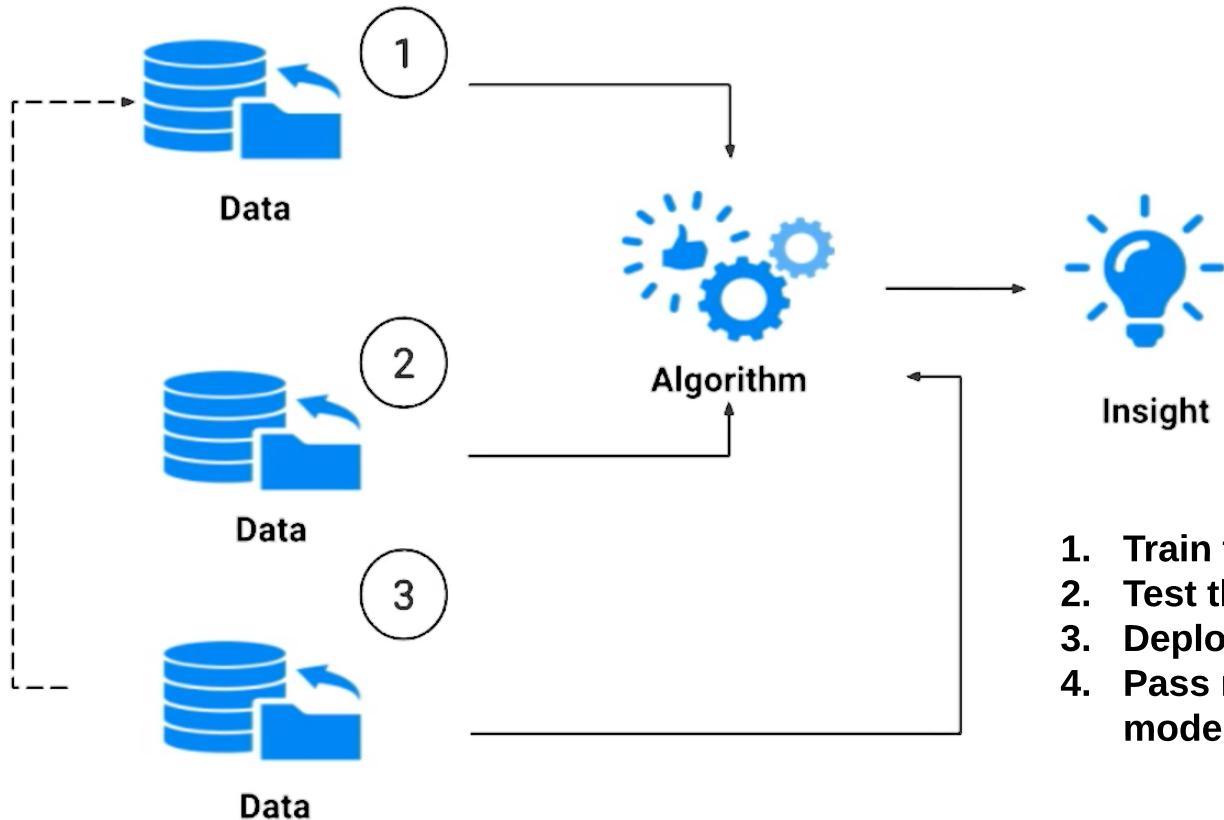
## Choose a Lesson

[ML Engine Overview](#)[ML Engine Hands On](#)

## ML Engine Overview

[Previous](#)[Next](#)

### Machine Learning in production



1. Train the model
2. Test the model
3. Deploy the model
4. Pass new data back to train the model (keep it fresh)

Notice a theme?  
Machine Learning needs Data!  
Big Data makes ML possible.

[Return to Table of Contents](#)

## Choose a Lesson

[ML Engine Overview](#)[ML Engine Hands On](#)

## *ML Engine Overview*

[Previous](#)[Next](#)

### Machine learning on Google Cloud

- Tensorflow
- Cloud ML Engine
- Pre-built ML API's



### TensorFlow

- Software library for high performance numerical computation
- Released as open source by Google in 2015
- Often default ML library of choice
- Pre-processing, feature creation, model training
- "I want to work with all of the detailed pieces."

### ML Researcher



- Fully managed Tensorflow platform
- Distributed training and prediction
- Scales to tens of CPU's/GPU's/TPU's
- Hyperparameter tuning with Hypertune
- Automate the "annoying bits" of machine learning
- "I want to train my own model, but automate it."

### Data Engineer/Scientist



- Pre-built machine learning models
- "Make Google do it"



### App Developer

[Return to Table of Contents](#)

## Choose a Lesson

[ML Engine Overview](#)[ML Engine Hands On](#)

## *ML Engine Overview*

[Previous](#)[Next](#)

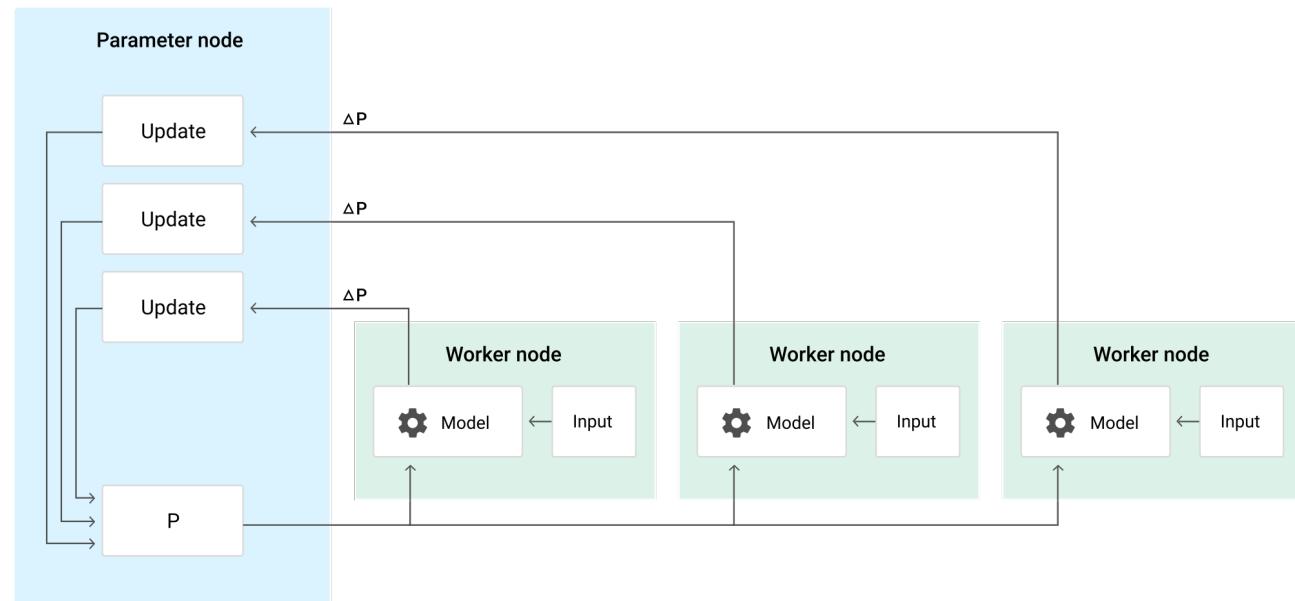
### How ML Engine works

Prepare trainer and data for the cloud:

- Write training application in Tensorflow
- Python is language of choice
- Run training model on local machine

Train your model with Cloud ML Engine:

- Training service allocates resources by specification (cluster of resources)
- Master - manages other nodes
- Workers - works on portion of training job
- Parameter servers - coordinate shared model state between workers
- Package model and submit job
- Package application and dependencies



Get Predictions - two types:

- Online:
  - High rate of requests with minimal latency
  - Give job data in JSON request string, predictions returned in its response message
- Batch:
  - Get inference (predictions) on large collections of data with minimal job duration
  - Input and output in Cloud Storage

[Return to Table of Contents](#)

## Choose a Lesson

[ML Engine Overview](#)[ML Engine Hands On](#)

## *ML Engine Overview*

[Previous](#)[Next](#)

### Key terminology

- **Model** - logical container individual solutions to a problem:
  - Can deploy multiple versions
  - e.g. Sale price of houses given data on previous sales
- **Version** - instance of model:
  - e.g. version 1/2/3 of how to predict above sale prices
- **Job** - interactions with Cloud ML Engine:
  - Train models
    - Command = 'submit job train model' on Cloud ML Engine
  - Deploy trained models
    - Command = 'submit job deploy trained model' on Cloud ML Engine
  - 'Failed' jobs can be monitored for troubleshooting

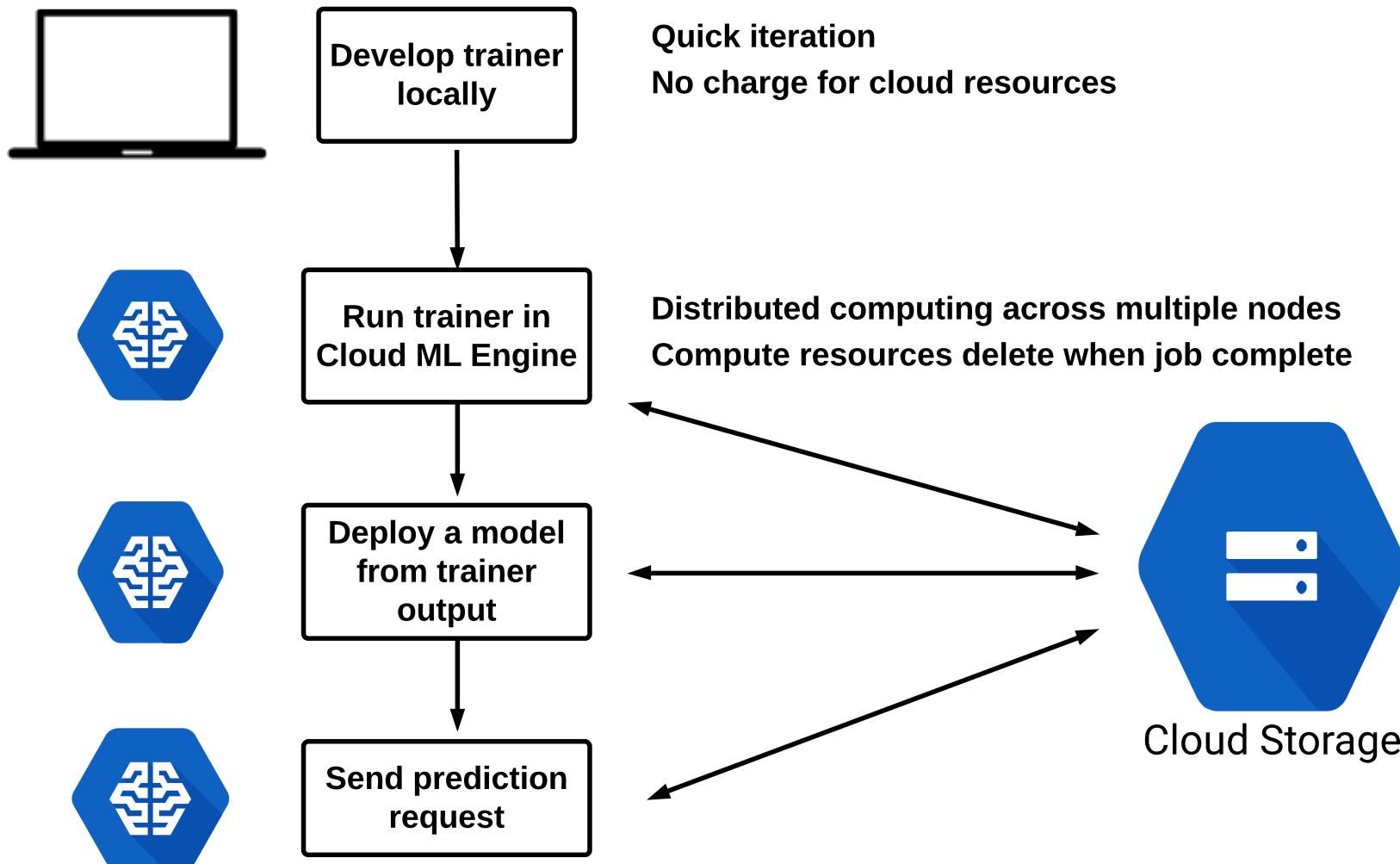
[Return to Table of Contents](#)

## ML Engine Overview

### Choose a Lesson

[ML Engine Overview](#)[ML Engine Hands On](#)[Previous](#)[Next](#)

### Typical process



[Return to Table of Contents](#)

## Choose a Lesson

[ML Engine Overview](#)[ML Engine Hands On](#)

## ML Engine Overview

[Previous](#)[Next](#)

### Must-know info

- Currently supports Tensorflow, scikit-learn, and XGBoost frameworks
- Keras currently in beta
- Note: This list is subject to change over time

### IAM roles:

- Project and Models:
  - Admin - Full control
  - Developer - Create training/prediction jobs, models/versions, send prediction requests
  - Viewer - Read-only access to above
- Models only:
  - Model Owner:
    - Full access to model and versions
  - Model User:
    - Read models and use for prediction
    - Easy to share specific models

### Using BigQuery for data source:

- Can read directly from BigQuery via training application
- Recommended to pre-process into Cloud Storage
- Using gcloud commands only works with Cloud Storage



BigQuery



Cloud Storage



Cloud Machine Learning Services

[Return to Table of Contents](#)

## Choose a Lesson

[ML Engine Overview](#)[ML Engine Hands On](#)

## *ML Engine Overview*

[Previous](#)[Next](#)

### Machine scale tiers and pricing

- **BASIC** - single worker instance
- **STANDARD\_1** - 1 master, 4 workers, 3 parameter servers
- **PREMIUM\_1** - 1 master, 19 workers, 11 parameter servers
- **BASIC\_GPU** - 1 worker with GPU
- **CUSTOM**

#### GPU/TPU:

- Much faster processing performance

#### Pricing:

- Priced per hour
- Higher cost for TPU/GPU's

Training - Predefined scale tiers - price per hour		Training - Machine types - price per hour		Batch prediction - price per node hour.	Online prediction - price per node hour.
<b>BASIC</b>	\$0.2774	<b>standard</b>	\$0.2774	\$0.09262	\$0.3
<b>STANDARD_1</b>	\$2.9025	<b>large_model</b>	\$0.6915		
<b>PREMIUM_1</b>	\$24.1683	<b>complex_model_s</b>	\$0.4141		
<b>BASIC_GPU</b>	\$1.2118	<b>complex_model_m</b>	\$0.8281		
<b>BASIC_TPU (Beta)</b>	\$9.7674	<b>complex_model_l</b>	\$1.6562		
<b>CUSTOM</b>	If you select CUSTOM as your scale tier, you have control over the number and type of virtual machines used for your training job. See the table of machine types.	<b>standard_gpu</b>	\$1.2118		
		<b>complex_model_m_gpu</b>	\$3.7376		
		<b>complex_model_l_gpu</b>	\$7.4752		
		<b>standard_p100 (Beta)</b>	\$2.6864		
		<b>complex_model_m_p100 (Beta)</b>	\$9.636		
		<b>cloud_tpu (Beta)</b>	\$9.4900		

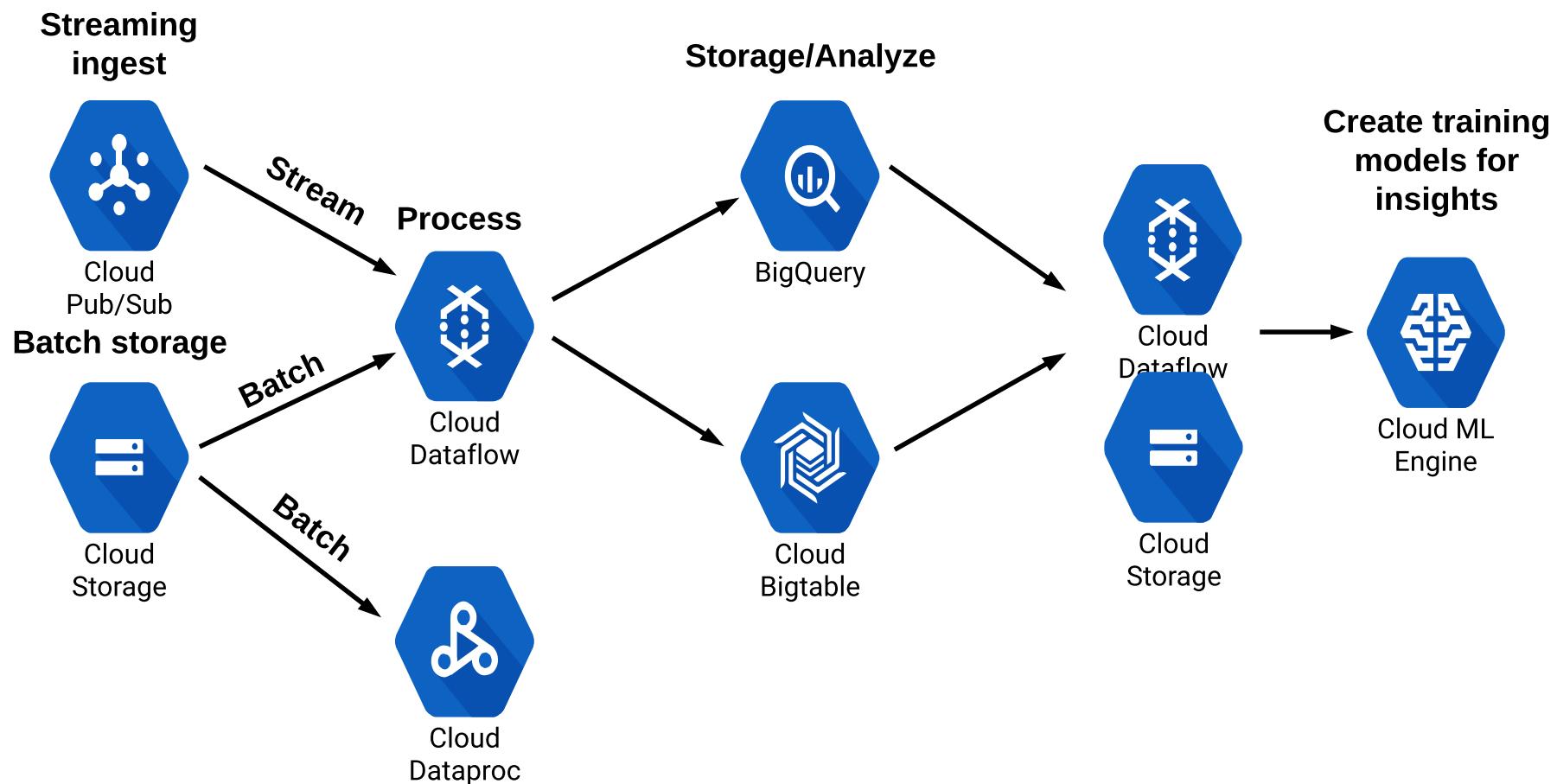
[Return to Table of Contents](#)

## ML Engine Overview

### Choose a Lesson

[ML Engine Overview](#)[ML Engine Hands On](#)[Previous](#)

### Big picture



[Return to Table of Contents](#)

## Choose a Lesson

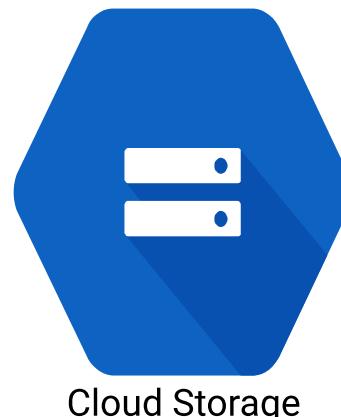
[ML Engine Overview](#)[ML Engine Hands On](#)

## *ML Engine Hands On*

### What we are doing:

- Working with pre-packaged training model:
  - Focusing on the Cloud ML aspect, not TF
- Heavy command line/gcloud focus - using Cloud Shell
- Submit training job locally using Cloud ML commands
- Submit training job on Cloud ML Engine, both single and distributed
- Deploy trained model, and submit predictions

```
gcloud ml-engine jobs submit training $JOB_NAME \
--package-path $TRAINER_PACKAGE_PATH \
--module-name $MAIN_TRAINER_MODULE \
--job-dir $JOB_DIR \
--region $REGION \
--config config.yaml \
-- \
--user_first_arg=first_arg_value \
--user_second_arg=second_arg_value \
--train-files $TRAIN_DATA \
--eval-files $EVAL_DATA
```



Cloud Storage

### Instructions for hands on:

Download scripts to Cloud Shell to follow along

```
gsutil -m cp gs://gcp-course-exercise-scripts/data-engineer/cloud-ml_engine/* .
```

[Return to Table of Contents](#)

## **Choose a Lesson**

[Pre-trained ML API's](#)[Vision API demo](#)

[Return to Table of Contents](#)

## *Pre-trained ML API's*

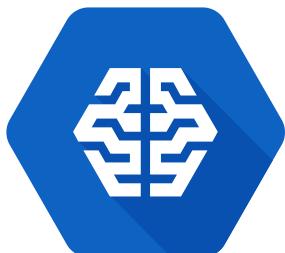
### Choose a Lesson

[Pre-trained ML API's](#)[Vision API demo](#)[Next](#)

### TensorFlow

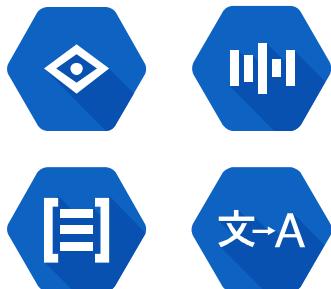
ML Researcher

Design custom ML model  
Design algorithm/neural network



Train and deploy custom ML model with cloud resources

Data Engineer/Scientist



"Make Google do it"  
Integrate Google's pre-trained models into your app  
Plug and play machine learning solutions

App Developer

[Return to Table of Contents](#)

## Choose a Lesson

[Pre-trained ML API's](#)[Vision API demo](#)

## *Pre-trained ML API's*

[Previous](#)[Next](#)

### Current ML API's (new ones being added)



Image recognition/analysis

Cloud Vision



Detect and translate languages

Cloud Translation



Text analysis  
Extract information  
Understand sentiment

Cloud Natural Language



More relevant job searches:  
Power recruitment, job boards

Cloud Job Discovery



Convert audio to text  
Multi-lingual support  
Understand sentence structure

Cloud Speech to Text



Convert text to audio  
Multiple languages/voices  
Natural sounding synthesis

Cloud Text to Speech (Beta)



Video analysis  
Labels, shot changes, explicit content

Cloud Video Intelligence



Dialogflow

Dialogflow for Enterprise

Conversational experiences  
Virtual assistants

[Return to Table of Contents](#)

## Choose a Lesson

[Pre-trained ML API's](#)[Vision API demo](#)

## *Pre-trained ML API's*

[Previous](#)[Next](#)

### Cloud Vision - closer look



<b>Label Detection</b>	Extract info in image across categories: Plane, sports, cat, night, recreation
<b>Text Detection (OCR)</b>	Detect and extract text from images
<b>Safe Search</b>	Recognize explicit content: Adult, spoof, medical, violent
<b>Landmark Detection</b>	Identify landmarks
<b>Logo Detection</b>	Recognize logos
<b>Image Properties</b>	Dominant colors, pixel count
<b>Crop Hints</b>	Crop coordinates of dominant object/face
<b>Web Detection</b>	Find matching web entries

[Return to Table of Contents](#)

## Choose a Lesson

[Pre-trained ML API's](#)[Vision API demo](#)

## *Pre-trained ML API's*

[Previous](#)[Next](#)

**When to use pre-trained API's?**

**Does your use case fit in pre-packaged model?**

**Do you need custom insights outside of pre-packaged models?**

I want to....	ML API	ML Engine
Detect B2B company products in photos		X
Detect objects or product logos in living room photos	X	
Recommend products based on purchase history		X
Interpret company sentiment on social media	X	
Capture customer sentiment in customer support calls	X	
Optimize inventory levels in multiple locations based on multiple factors (region, weather, demand)		X
Extract text data from receipt images	X	
Determine receipt type		X

[Return to Table of Contents](#)

## Choose a Lesson

[Pre-trained ML API's](#)[Vision API demo](#)

## *Pre-trained ML API's*

[Previous](#)

### Exam perspectives:

#### When to use pre-trained API vs. Cloud ML Engine?

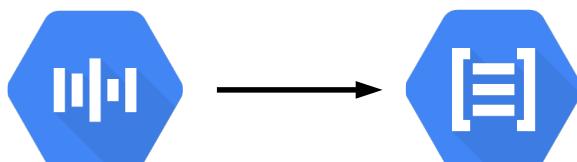
- I need a quick solution
- I don't know how to train an ML model
- I don't have time to train an ML model
- The pre-trained APIs fit my use case

### How to convert images, video, etc for use with API?

- Can use Cloud Storage URI for GCS stored objects
- Encode in base64 format

### How to combine API's for scenarios?

- Search customer service calls and analyze for sentiment



Convert call audio to text  
Make searchable

Analyze text  
for sentiment

### Pricing:

- Pay per API request per feature

Feature	Price per 1000 units		
	First 1000 units/month	Units 1001 - 5,000,000 / month	Units 5,000,001 - 20,000,000 / month
Label Detection	Free	\$1.50	\$1.00
Text Detection	Free	\$1.50	\$0.60
Document Text Detection	Free	\$1.50	\$0.60
Safe Search (explicit content) Detection	Free	Free with Label Detection, or \$1.50	Free with Label Detection, or \$0.60
Facial Detection	Free	\$1.50	\$0.60
Landmark Detection	Free	\$1.50	\$0.60
Logo Detection	Free	\$1.50	\$0.60
Image Properties	Free	\$1.50	\$0.60
Crop Hints	Free	Free with Image Properties, or \$1.50	Free with Image Properties, or \$0.60
Web Detection	Free	\$3.50	Contact Google for more information

[Return to Table of Contents](#)

### Choose a Lesson

[Pre-trained ML API's](#)[Vision API demo](#)

# The Data Dossier

## *Vision API Demo*

### Basic steps for most APIs:

- Enable the API
- Create API key
- Authenticate with API key
- Encode in base64 (optional)
- Make an API request
- Requests and outputs via JSON

Commands will be in lesson description

[Return to Table of Contents](#)

## Choose a Lesson

[Datalab Overview](#)

[Return to Table of Contents](#)

## Choose a Lesson

[Datalab Overview](#)

## Datalab Overview

[Next](#)

### What is it?

- Interactive tool for exploring and visualizing data:
  - Notebook format
  - Great for data engineering, machine learning
- Built on Jupyter (formerly iPython):
  - Open source - Jupyter ecosystem
  - Create documents with live code and visualizations
- Visual analysis of data in BigQuery, ML Engine, Compute Engine, Cloud Storage, and Stackdriver
- Supports Python, SQL, and JavaScript
- Runs on GCE instance, dedicated VPC and Cloud Source Repository
- Cost: free - only pay for GCE resources Datalab runs on and other Google Cloud services you interact with



[Return to Table of Contents](#)

## Datalab Overview

### Choose a Lesson

[Datalab Overview](#)[Previous](#)[Next](#)

### How It Works

#### Create and connect to a Datalab instance

`datalab create (instance-name)` →

- Connect via SSH and open web preview
- `datalab connect (instance-name)`
- Open web preview - port 8081



datalab-network



datalab-instance



datalab-notebooks

Source repository

## Working with Datalab

1 Write code in Python

2 Run cell (Shift+Enter)

3 Examine output

4 Write commentary in markdown

5 Share and collaborate



The screenshot shows a Jupyter Notebook interface with the following elements:

- Header:** "Single Cloud Datalab demandforecast (autosaved)"
- Toolbar:** Notebook, Add Code, Add Markdown, Delete, Move Up, Move Down, Run, Clear, Reset Session.
- Cell 1 (Python code):** `j = data[data['dayofweek'] == 7].plot(kind='scatter', x='maxtemp', y='numtrips')`
- Output 1:** A scatter plot showing the relationship between maximum temperature (x-axis, 20-100) and number of trips (y-axis, 50,000-85,000).
- Cell 2 (Python code):** `Adding 2014 data`
- Output 2:** Text: "Let's add in 2014 data to the Pandas dataframe. Note how useful it was for us to modularize our queries around the YEAR. Now, the data seem a bit more robust."
- Cell 3 (Python code):** `trips = bq.Query(taxiquery, YEAR=2014).to_dataframe()`

[Return to Table of Contents](#)

## Choose a Lesson

[Datalab Overview](#)

## Datalab Overview

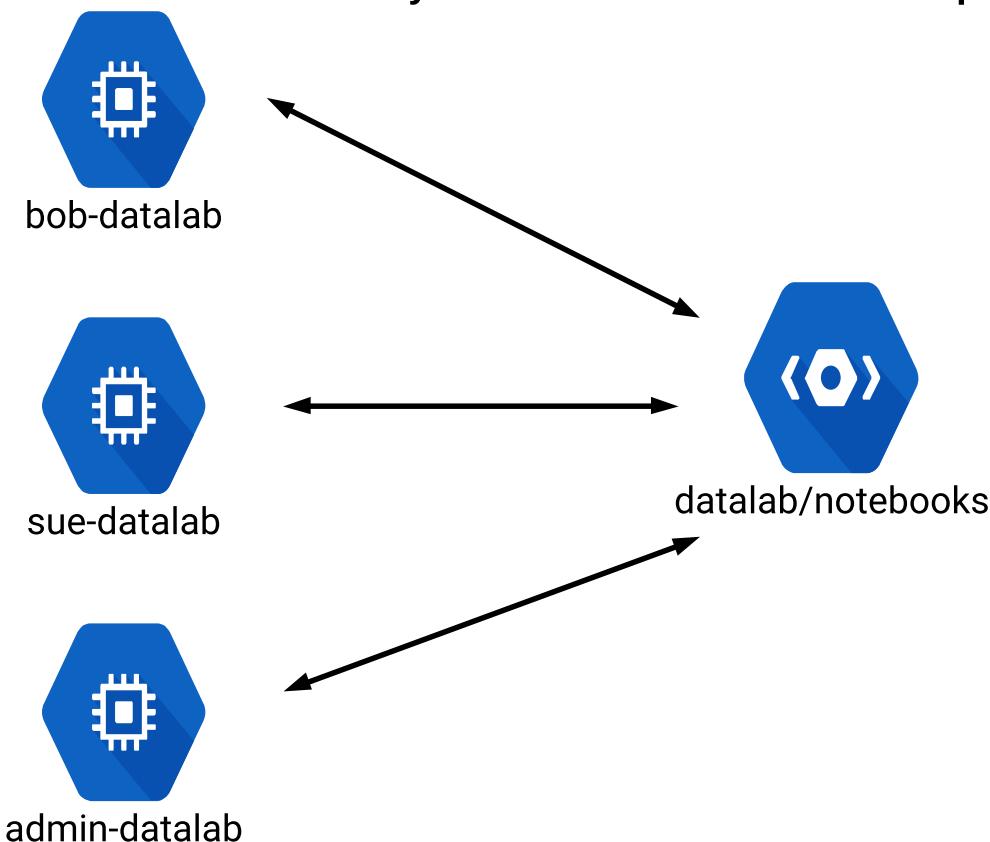
[Previous](#)

### Sharing notebook data:

- GCE access based on GCE IAM roles:
  - Must have Compute Instance Admin and Service Account Actor roles
- Notebook access per user only
- Sharing data performed via shared Cloud Source Repository
- Sharing is at the project level

### Creating team notebooks - two options:

- Team lead creates notebooks for users using --for user option:
  - `datalab create [instance] --for-user bob@professionalwireless.net`
- Each user creates their own datalab instance/notebook
- Everyone accesses same shared repository of datalab/notebooks



[Return to Table of Contents](#)

## Choose a Lesson

[What is Dataprep?](#)

[Return to Table of Contents](#)

## Choose a Lesson

[What is Dataprep?](#)

## What is Dataprep?

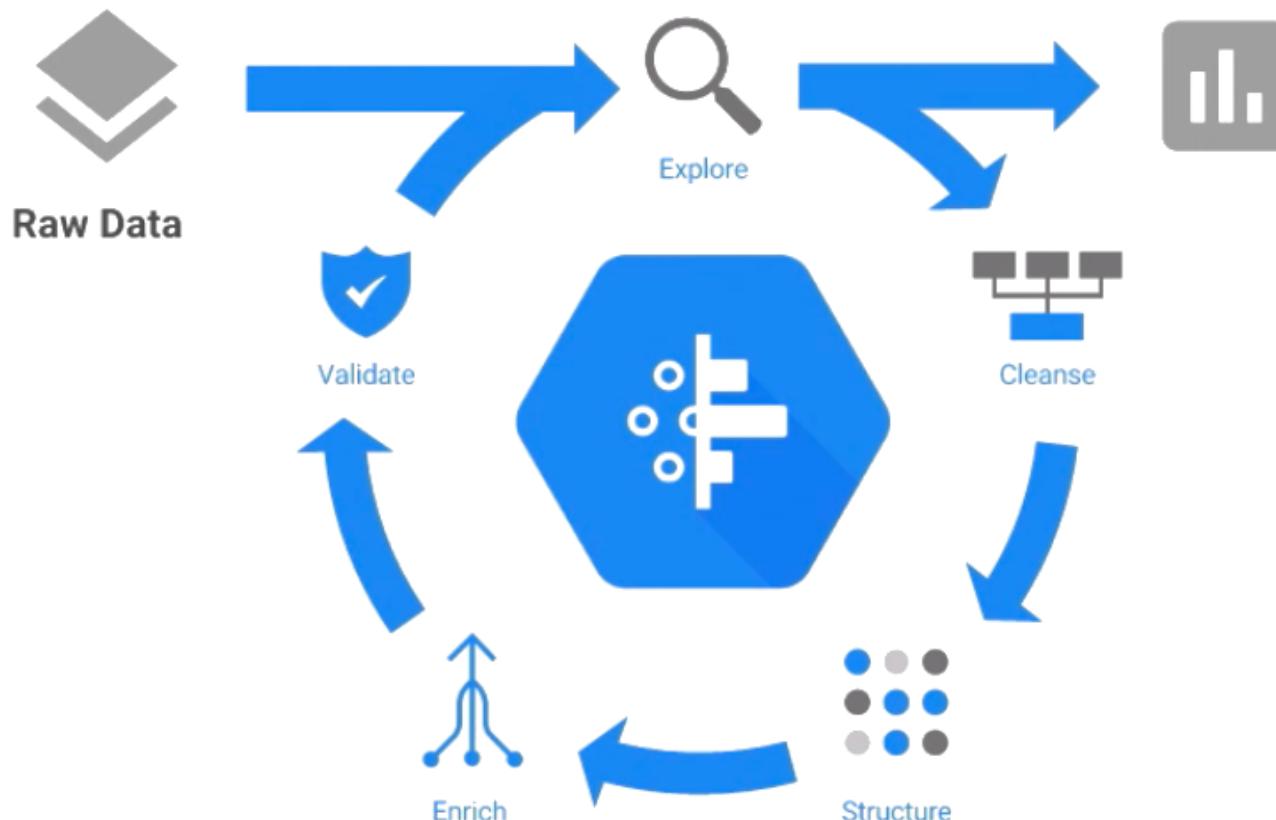
[Next](#)

### What is it?

- Intelligent data preparation
- Partnered with Trifacta for data cleaning/processing service
- Fully managed, serverless, and web-based
- User-friendly interface:
  - Clean data by clicking on it
- Supported file types:
  - Input - CSV, JSON (including nested), Plain text, Excel, LOG, TSV, and Avro
  - Output - CSV, JSON, Avro, BigQuery table:
    - CSV/JSON can be compressed or uncompressed

### Why is this important?

- Data Engineering requires high quality, cleaned, and prepared data
- 80% - time spent in data preparation
- 76% - view data preparation as the least enjoyable part of work
- Dataprep democratizes the data preparation process



[Return to Table of Contents](#)

## Choose a Lesson

[What is Dataprep?](#)[Previous](#)

## *What is Dataprep?*



### How It Works

Backed by Cloud Dataflow:

- After preparing, Dataflow processes via Apache Beam pipeline
- "User-friendly Dataflow pipeline"

Dataprep process:

- Import data
- Transform sampled data with recipes
- Run Dataflow job on transformed dataset
- Export results (GCS, BigQuery)

### Intelligent suggestions:

- Selecting data will often automatically give the best suggestion
- Can manually create recipes, however simple tasks (remove outliers, de-duplicate) should use auto-suggestions

### IAM:

- Dataprep User - Run Dataprep in a project
- Dataprep Service Agent - Gives Trifecta necessary access to project resources:
  - Access GCS buckets, Dataflow Developer, BigQuery user/data editor
  - Necessary for cross-project access + GCE service account

### Pricing:

- **1.16 \* cost of Dataflow job**

[Return to Table of Contents](#)

## Choose a Lesson

[Data Studio Introduction](#)

[Return to Table of Contents](#)

## Choose a Lesson

[Data Studio Introduction](#)

## Data Studio Introduction

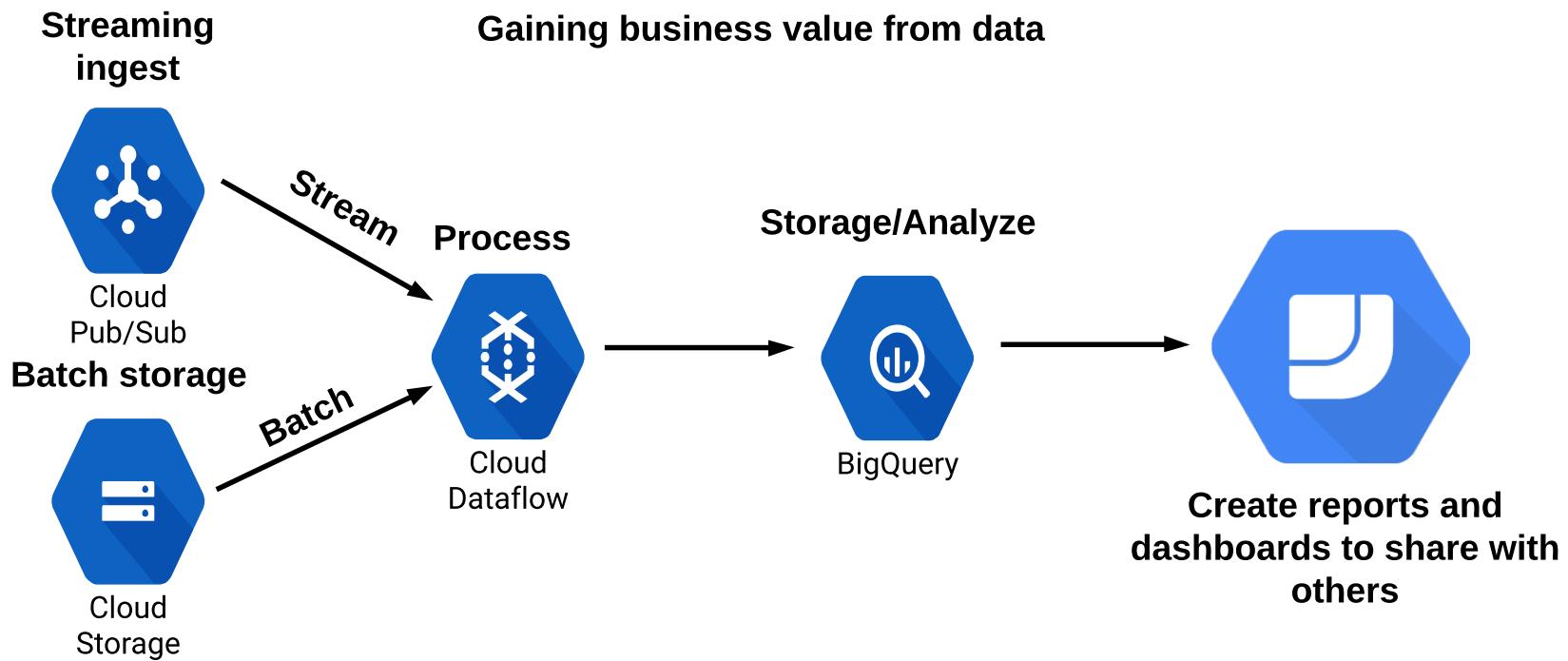
[Next](#)

### What is Data Studio?

- Easy to use data visualization and dashboards:
  - Drag and drop report builder
- Part of G Suite, not Google Cloud:
  - Uses G Suite access/sharing permissions, not Google Cloud (no IAM)
  - Google account permissions in GCP will determine data source access
  - Files saved in Google Drive
- Connect to many Google, Google Cloud, and other services:
  - BigQuery, Cloud SQL, GCS, Spanner
  - YouTube Analytics, Sheets, AdWords, local upload
  - Many third party integrations
- Price - Free:
  - BigQuery access run normal query costs

### Data Lifecycle - Visualization

Gaining business value from data



[Return to Table of Contents](#)

## Choose a Lesson

[Data Studio Introduction](#)

## *Data Studio Introduction*

[Previous](#)

### Basic process

- Connect to data source
- Visualize data
- Share with others

### Creating charts

- Use combinations of dimensions and metrics
- Create custom fields if needed
- Add date range filters with ease



## Caching - options for using cached data performance/costs

Two cache types, query cache and prefetch cache

Query cache:

- Remembers queries issued by reports components (i.e. charts)
- When performing same query, pulls from cache
- If query cache cannot help, goes to prefetch cache
- Cannot be turned off

Prefetch cache:

- 'Smart cache' - predicts what 'might' be requested
- If prefetch cache cannot serve data, pulls from live data set
- Only active for data sources that use owner's credentials for data access
- Can be turned off

When to turn caching off:

- Need to view 'fresh data' from rapidly changing data set

# The Data Dossier

[Return to Table of Contents](#)

[Return to Table of Contents](#)

## ***Additional Study Resources***

### **SQL deep dive**

- Course - SQL Primer
- <https://linuxacademy.com/cp/modules/view/id/52>

### **Machine Learning**

- Google Machine Learning Crash Course (free)
- <https://developers.google.com/machine-learning/crash-course/>

### **Hadoop**

- Hadoop Quick Start
- <https://linuxacademy.com/cp/modules/view/id/294>

### **Apache Beam (Dataflow)**

- Google's guide to designing your pipeline with Apache Beam (using Java)
- <https://cloud.google.com/dataflow/docs/guides/beam-creating-a-pipeline>