Return to Table of Contents

Choose a Lesson

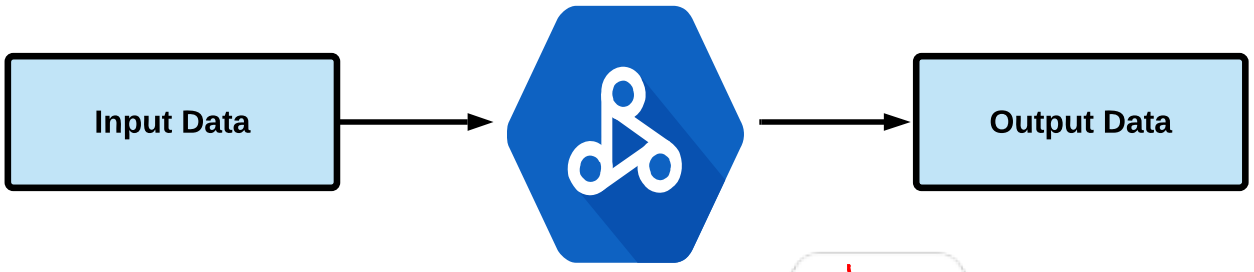
- Dataproc Overview
- Configure Dataproc Cluster and Submit Job
- Migrating and Optimizing for Google Cloud
- Best Practices for Cluster Performance

Dataproc Overview

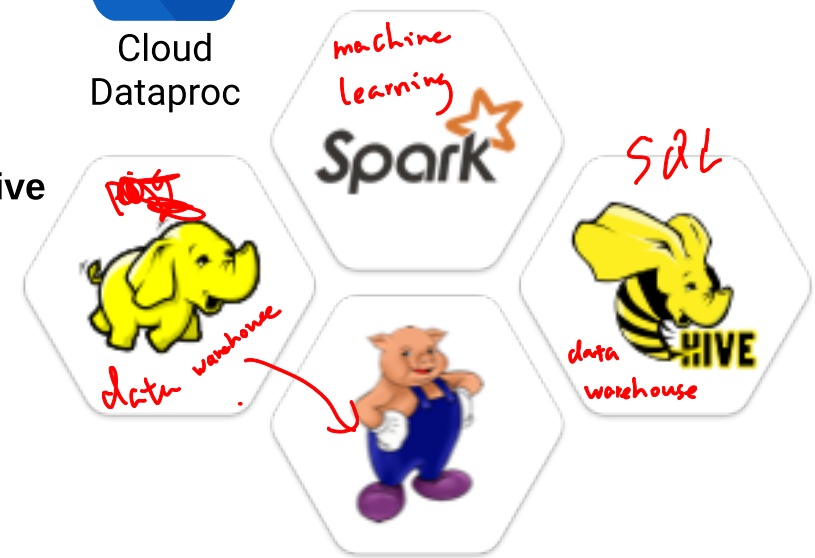
Next

What is Cloud Dataproc?

It's another transformation and data processing serv.



Cloud Dataproc



Hadoop ecosystem:

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

Managed version of hadoop and spark

Managed Hadoop/Spark Stack

Custom Code
Monitoring/Health
Dev Integration
Manual Scaling
Job Submission
Google Cloud Connectivity
Deployment
Creation

all managed by Proc!

Dataproc facts:

- On-demand, managed Hadoop and Spark clusters
- Managed, but not **no-ops**:
  - Must configure cluster, not auto-scaling (need to watch storage) } still need to set up
  - 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 such as Kafka, jupyter notebook.

Return to Table of Contents

Choose a Lesson

- Dataproc Overview
- Configure Dataproc Cluster and Submit Job
- Migrating and Optimizing for Google Cloud
- Best Practices for Cluster Performance

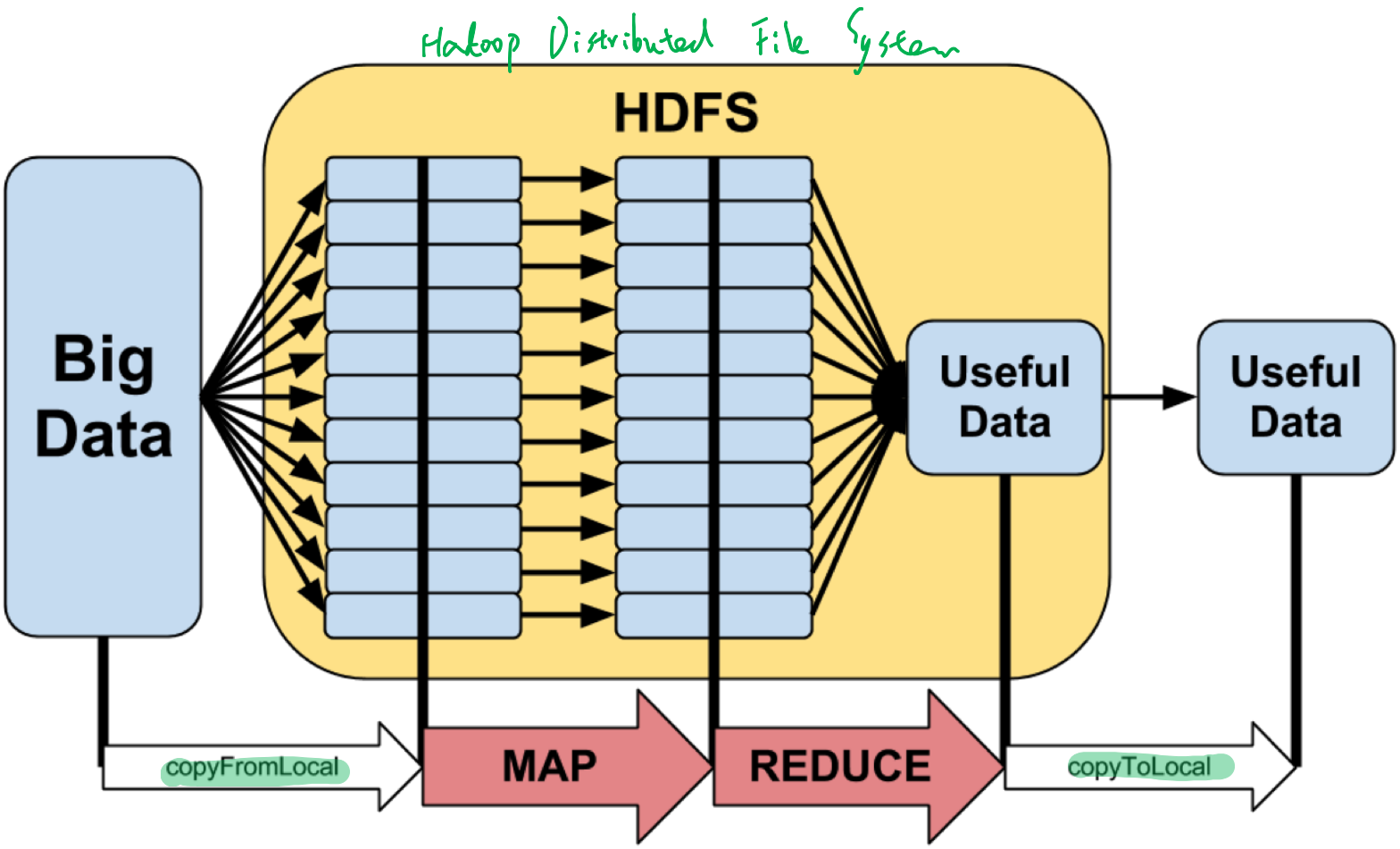
Dataproc Overview

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
- Best Practices for Cluster Performance

Dataproc Overview

Previous

Next

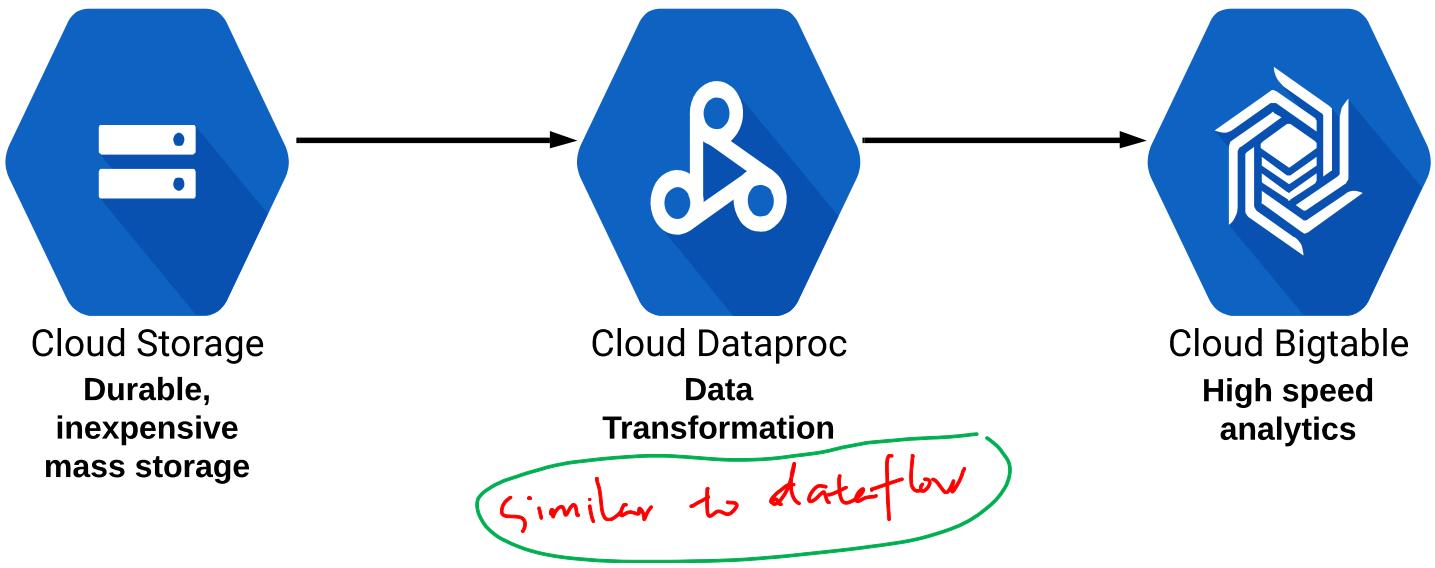
Pricing:

- **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
- Best Practices for Cluster Performance

Dataproc Overview

Previous

exam topic

Identity and Access Management (IAM):

- Project level only (to all clusters) (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
- Best Practices for Cluster Performance

Configure Dataproc Cluster

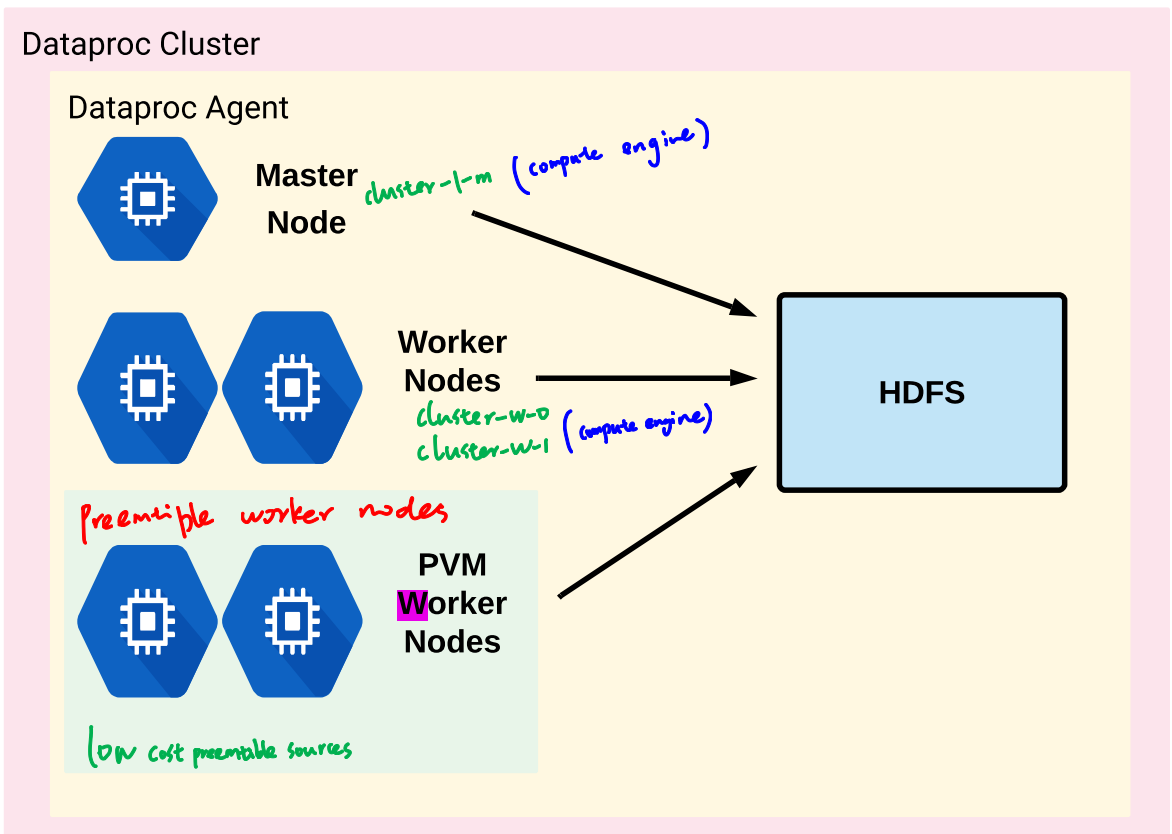
Next

Create cluster: (gcloud command)

- 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
- Best Practices for Cluster Performance

Configure Dataproc Cluster

Previous

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

port / Port

8088 : access to Hadoop cluster

9870 : Hadoop files

SOCKS proxy configuration:

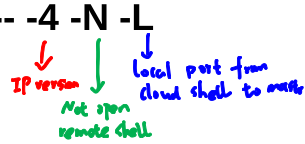
- SSH to master to enable port forwarding:
  - gcloud compute ssh *master-host-name* --project=*project-id* --zone=*master-host-zone* -- -D 1080 -N
- Open new terminal window - launch web browser with parameters (varies by OS/browser):
  - "/Applications/Google Chrome.app/Contents/MacOS/Google Chrome" --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:
  - 8088 - Hadoop
  - 9870 - HDFS

exam

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)

No exam



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

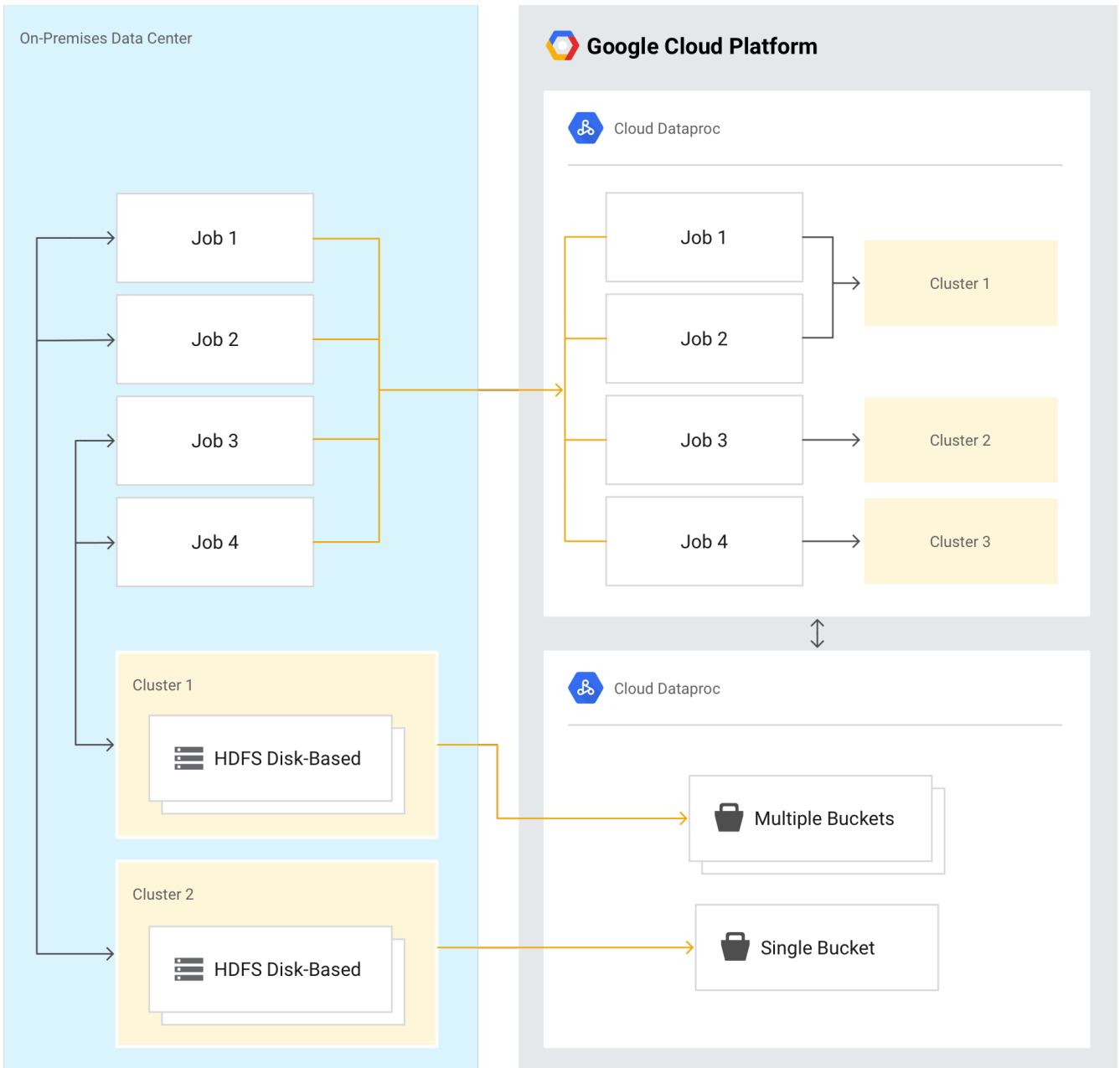
Best Practices for Cluster Performance

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

Best Practices for Cluster Performance

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

Choose a Lesson

- Dataproc Overview
- Configure Dataproc Cluster and Submit Job
- Migrating and Optimizing for Google Cloud
- Best Practices for Cluster Performance

Migrating and Optimizing for Google Cloud

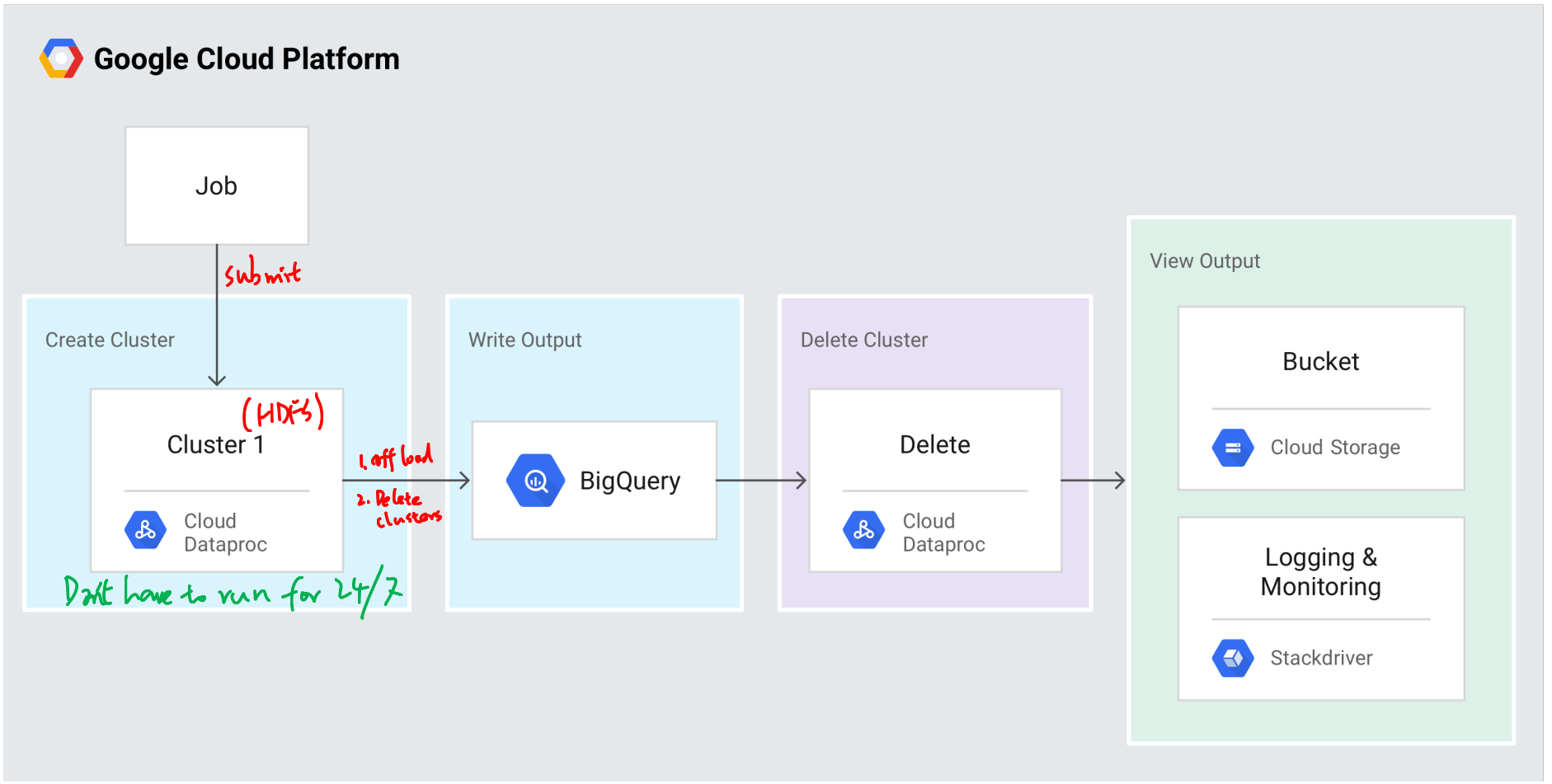
Previous

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**



Return to Table of Contents

Choose a Lesson

- Dataproc Overview
- Configure Dataproc Cluster and Submit Job
- Migrating and Optimizing for Google Cloud
- Best Practices for Cluster Performance

Migrating and Optimizing for Google Cloud

Previous

exam topic: How to ...

Converting from HDFS to Google Cloud Storage:

1. Copy data to GCS:
   - Install connector or copy manually for on-prem
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

- Dataproc Overview
- Configure Dataproc Cluster and Submit Job
- Migrating and Optimizing for Google Cloud
- Best Practices for Cluster Performance

Best Practices for Cluster Performance

Dataproc Performance Optimization

exam topic

(GCP-specific)

- Keep data close to your cluster
  - Place Dataproc cluster in the same region as storage bucket
- Larger persistent disk = better performance
  - Consider using SSD over HDD – slightly higher cost
- Allocate more VM's
  - Use preemptible VM's to save on costs

however, More VM's will come at a higher cost than larger disks if more disk throughput is needed

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 Logging and Monitoring
- BigQuery Best Practices

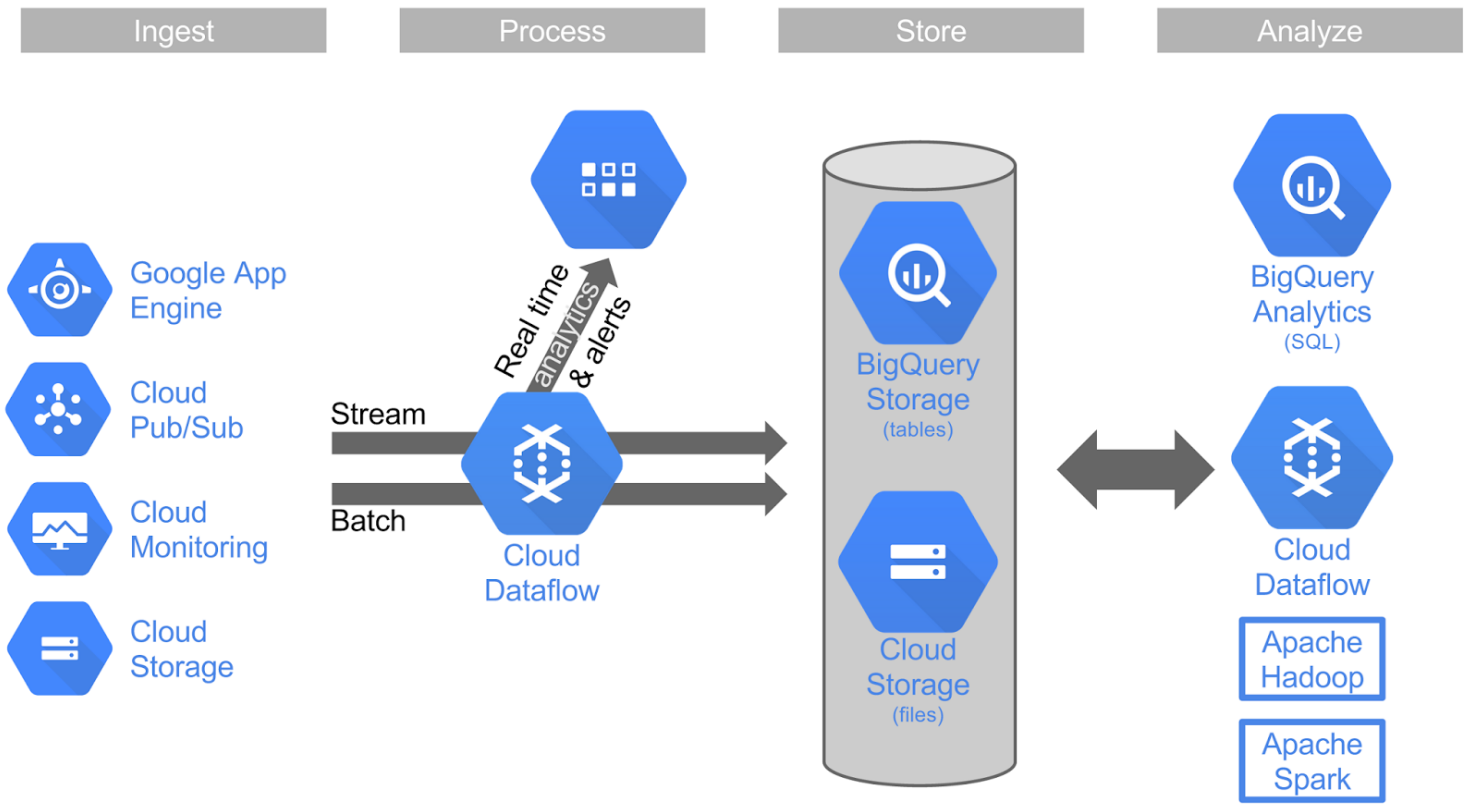
BigQuery Overview

Next

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

exam: basic 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
- BigQuery Logging and Monitoring
- BigQuery Best Practices

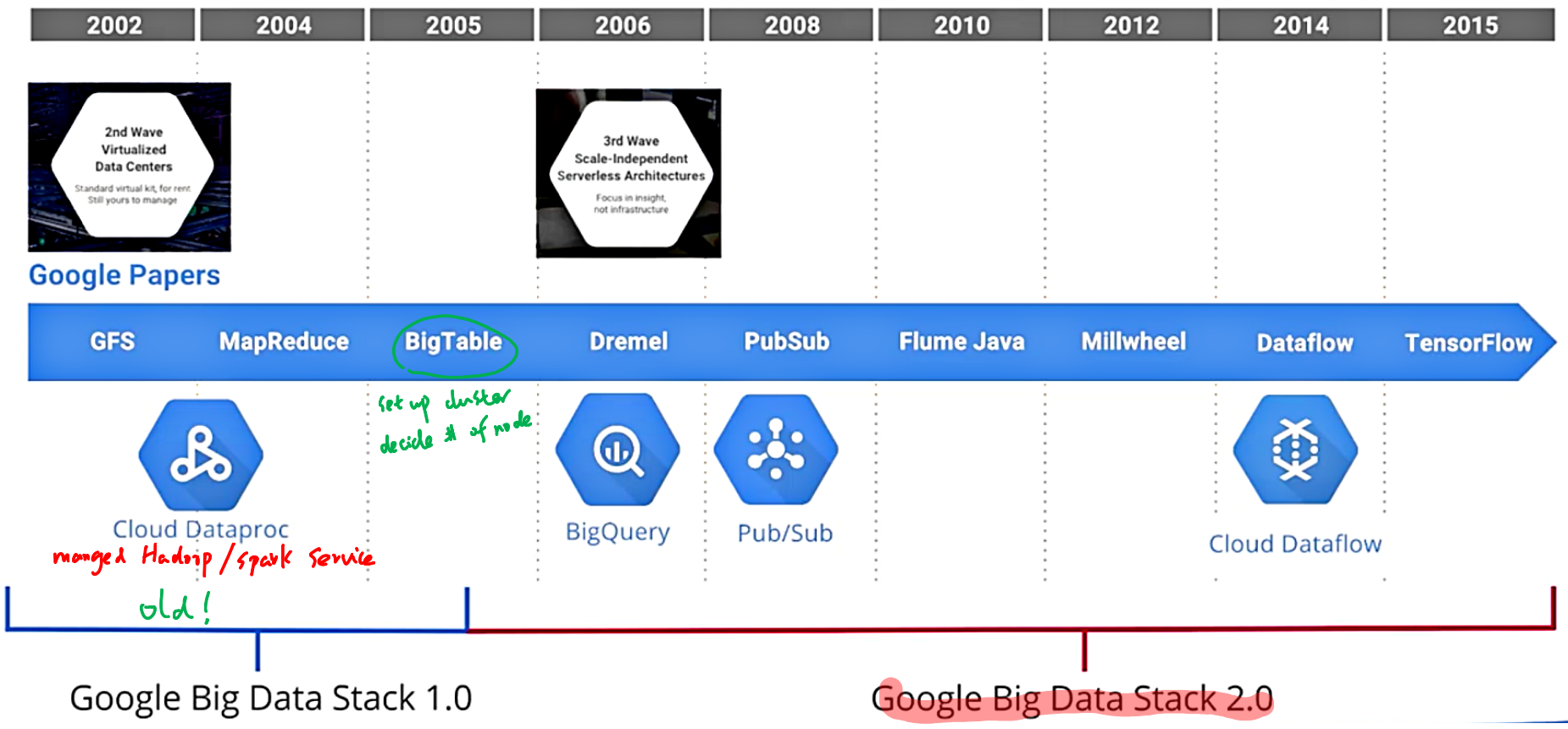
BigQuery Overview

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



Return to Table of Contents

Choose a Lesson

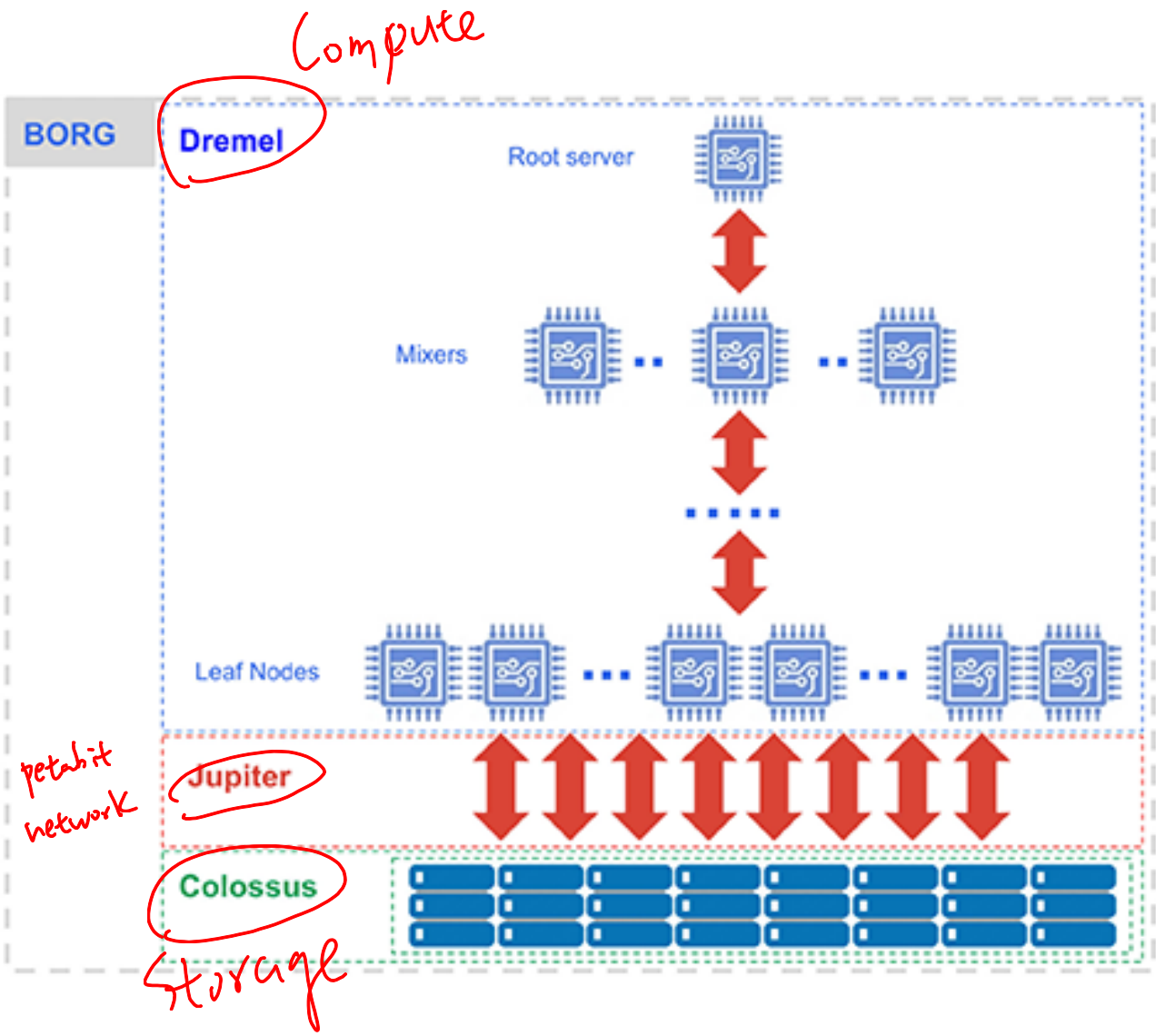
- BigQuery Overview
- Interacting with BigQuery
- Load and Export Data
- Optimize for Performance and Costs
- Streaming Insert Example
- BigQuery Logging and Monitoring
- BigQuery Best Practices

BigQuery Overview

Previous (Not in exam) 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 Logging and Monitoring
- BigQuery Best Practices

BigQuery Overview

Previous

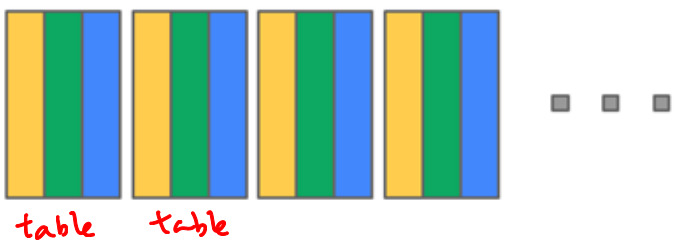
Next

How BigQuery works (cont)

- Columnar data store (different from SQL)
  - 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

just append records

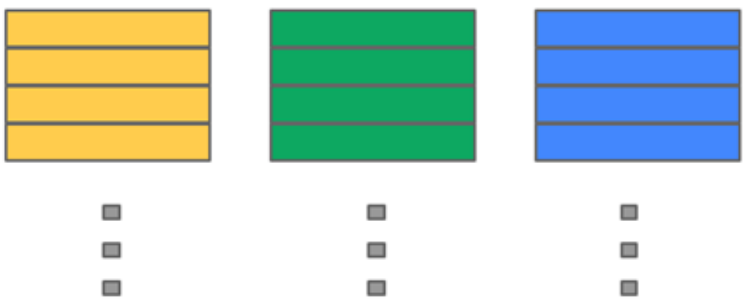
SQL



Record Oriented Storage

entire table with rows

Columnar data store



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 Logging and Monitoring
- BigQuery Best Practices

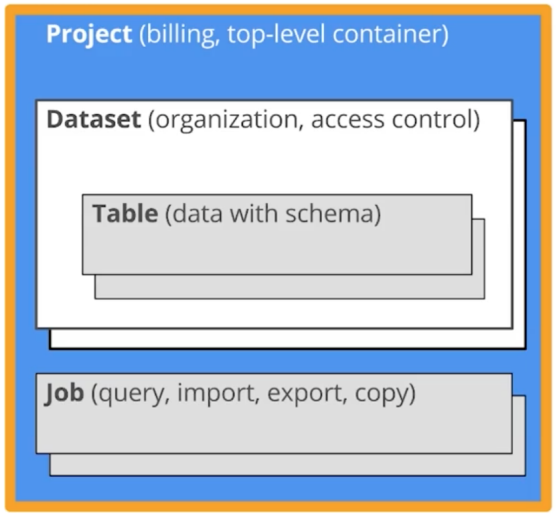
BigQuery Overview

Previous (return topic)

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 (only project level, or dataset 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 Logging and Monitoring

BigQuery Best Practices

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

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 Logging and Monitoring
- BigQuery Best Practices

Interacting with BigQuery

Next

exam topic

Interaction methods

- Web UI
- Command line (bq commands)
  - bq query --arguments 'QUERY'

create dataset:
$ bq mk --dataset <project-ID>:new_dat

- 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
- BigQuery Logging and Monitoring
- BigQuery Best Practices

Interacting with BigQuery

Previous

Views

(exam) query inside query

only contains data from previous query.

- 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

☑ use as cached results

(exam)

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

User Defined Functions (UDF)

Query Editor | UDF editor

- 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
- Command Line only with standard SQL

Return to Table of Contents

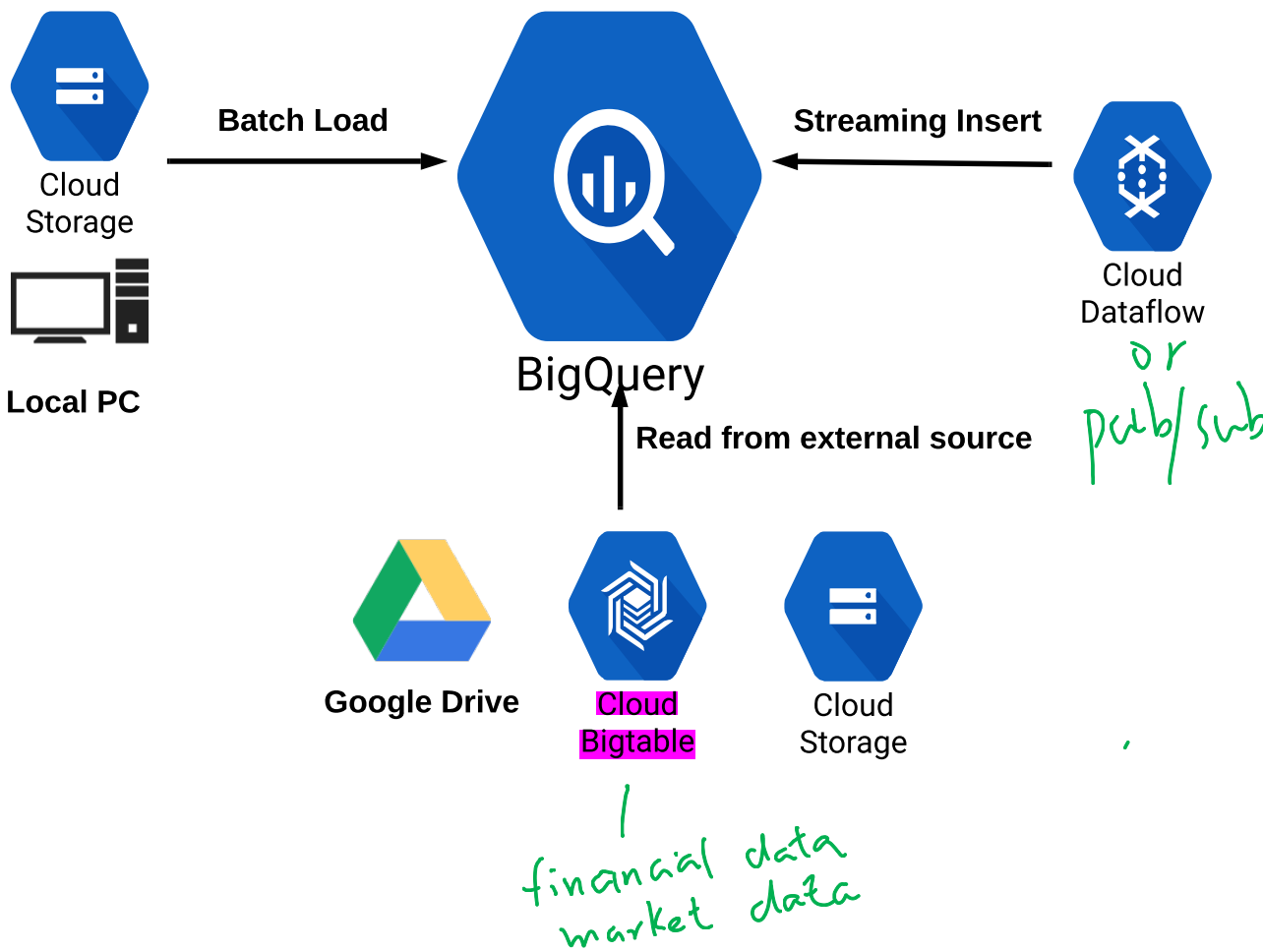
Choose a Lesson

- BigQuery Overview
- Interacting with BigQuery
- Load and Export Data
- Optimize for Performance and Costs
- Streaming Insert Example
- BigQuery Logging and Monitoring
- BigQuery Best Practices

Load and Export Data

Loading and reading sources

Next



Data formats:

Load

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

Read

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

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

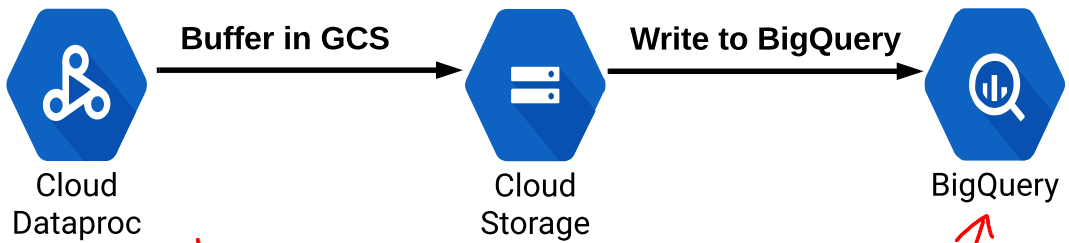
- BigQuery Overview
- Interacting with BigQuery
- Load and Export Data
- Optimize for Performance and Costs
- Streaming Insert Example
- BigQuery Logging and Monitoring
- BigQuery Best Practices

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

vs Copying 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, 1 table at a time if use web UI
- 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

> data ⇒ BigQuery

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 Logging and Monitoring

BigQuery Best Practices

Optimize for Performance and Costs

Next

Performance and costs are complementary

- 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

1. Avoid using SELECT *
2. Denormalize data when possible (counter to general relational database)
   - Grouping data into single table
   - Often with nested/repeated data (exam question)
   - Good for read performance, not for write (transactional) performance
3. Filter early and big with WHERE clause } filter out data
4. Do biggest joins first, and filter pre-JOIN
5. LIMIT does not affect cost
6. Partition data by date
   - Partition by ingest time – possible to query data by date
   - Parition by specified data columns

Return to Table of Contents

Choose a Lesson

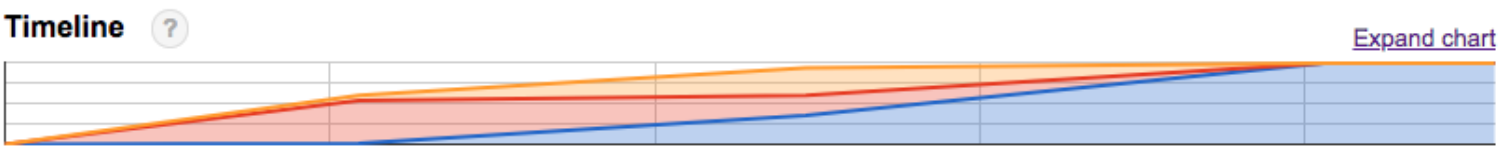
- BigQuery Overview
- Interacting with BigQuery
- Load and Export Data
- Optimize for Performance and Costs
- Streaming Insert Example
- BigQuery Logging and Monitoring
- BigQuery Best Practices

Optimize for Performance and Costs

Previous

Monitoring query performance

- Understand color codes ← for exam question
- Understand 'skew' in difference between average and max time



Execution Plan

many nodes are work in background

	Stage timing				Rows		
	Wait	Read	Compute	Write	Parallel Inputs	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)

faster performance will lower those number

⇒ less cost

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 Logging and Monitoring
- BigQuery Best Practices

Streaming Insert Example

Quick setup

cd

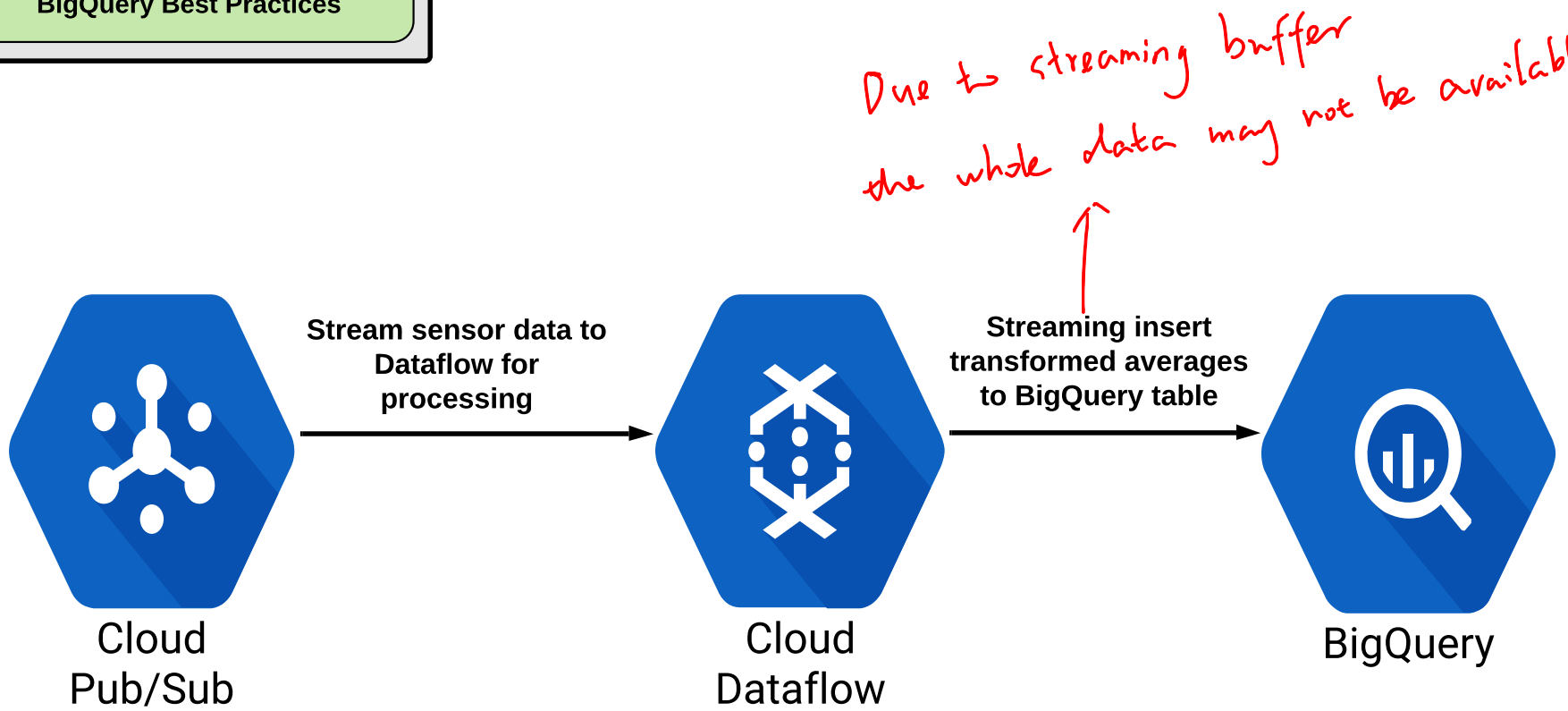
gsutil cp -r gs://gcp-course-exercise-scripts/data-engineer/* .

bash streaming-insert.sh

Clean up

bash streaming-cleanup.sh

Manually stop Dataflow job



Return to Table of Contents

Choose a Lesson

- BigQuery Overview
- Interacting with BigQuery
- Load and Export Data
- Optimize for Performance and Costs
- Streaming Insert Example
- BigQuery Logging and Monitoring
- BigQuery Best Practices

BigQuery Logging and Monitoring

Stackdriver Monitoring and Logging Differences

- Monitoring = performance/resources (usage)
- Logging = who is doing what
  - History of actions

← Two big stackdriver services!

Monitoring BigQuery Performance/Resources

- Monitoring = metrics, performance, resource capacity/usage (slots)
  - Query count, query times, slot utilization
  - Number of tables, stored and uploaded bytes over time
  - Alerts on metrics e.g., long query times
    - Example: Alerts when queries take more than one minute
- No data on **who** is doing what, or query details

find the heavy user

set an alert

(1) (2)

Stackdriver Logging: "A Paper Trail"

- Logging = who is doing what
- Record of jobs and queries associated with accounts

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 Logging and Monitoring

BigQuery Best Practices

BigQuery Best Practices

Next

Data Format for Import

- Best performance = Avro format
- Scenario: Import multi-TB databases with millions of rows

Faster
Avro - Compressed
Avro - Uncompressed
Parquet
CSV
JSON
CSV - Compressed
JSON - Compressed
Slower

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 Logging and Monitoring

BigQuery Best Practices

BigQuery Best Practices

Previous

Next

Partitioned Tables

What is a partitioned table?

- Special single table
  - Divided into segments known as “partitions” by time

Why is this important?

- Query only certain rows (partitions) instead of entire table
  - Limits amount of read data
  - Improves performance
  - Reduces costs
- Partition types
  - Ingests time — when the data/row is created
  - Includes **TIMESTAMP** or **DATE** column
- **Scenario**: A large amount of data gets generated every day, and we need to query for only certain time periods within the same table.

Why not use multiple tables (one for each day) plus wildcards?

- Limited to 1000 tables per dataset
- Substantial performance drop vs. a single table

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 Logging and Monitoring

BigQuery Best Practices

BigQuery Best Practices

Previous

Next

Clustered Tables

- Taking partitioned tables "to the next level"
- Similar to partitioning, divides table reads by a specified column field
  - Instead of dividing by date/time, divides by field
- **Scenario:** Logistics company needs to query by tracking ID
  - Cluster by tracking ID column = only reading table rows with specified tracking ID's
- Restriction: only (currently) available for partitioned tables

* the table must be partitioned before you cluster

Slots

- Computational capacity required to run a SQL query
  - Bigger/more complex queries need more slots
- Default, on-demand pricing allocates 2000 slots
  - Only an issue for extremely complex queries, or high number of simultaneous users
  - If more than 2000 slots required, switch to **flat-rate pricing**

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 Logging and Monitoring

BigQuery Best Practices

BigQuery Best Practices

Previous

Backup and Recovery

Build-in Snapshot feature

- Highly available = multi-regional dataset vs. regional
- Backup/recovery = BigQuery automatically takes continuous snapshots of tables
  - 7 day history, but 2 days if purposely deleted
- Restore to previous point in time using `@(time)`, in milliseconds
- Example: Get snapshot from one hour ago

```
#legacySQL
SELECT * FROM [PROJECT_ID:DATASET.TABLE@-3600000]
```

1 hour

- Alternatively, export table data to GCS, though not as cost effective

Return to Table of Contents

What is Machine Learning?

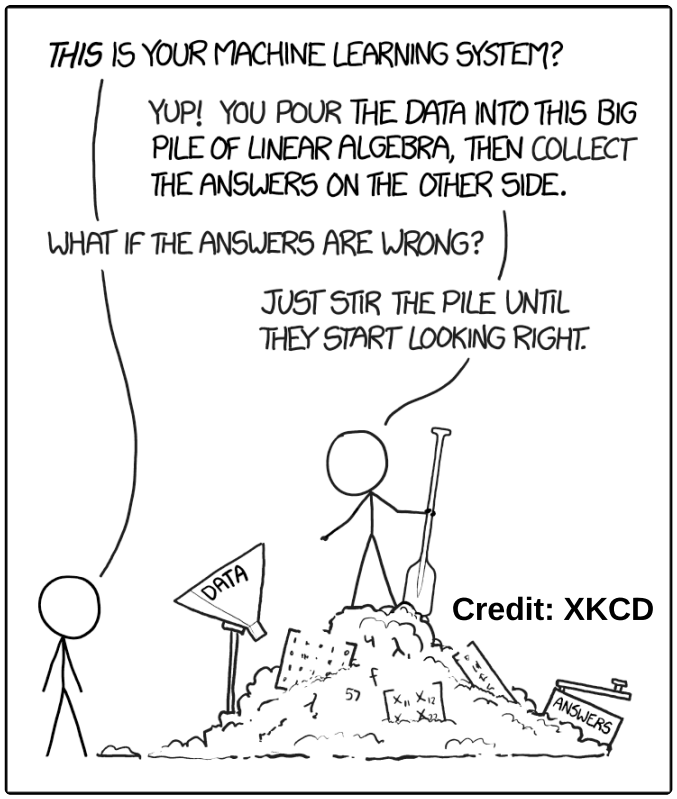
Choose a Lesson

- What is Machine Learning?
- Working with Neural Networks
- Preventing Overfitted Training Models

Popular view of machine learning...

Next

DATA →



Credit: XKCD

→ 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
- Preventing Overfitted Training Models

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

Train on ML model

TensorFlow

Match labels by adjusting weights to input features

"I think this is a cat"

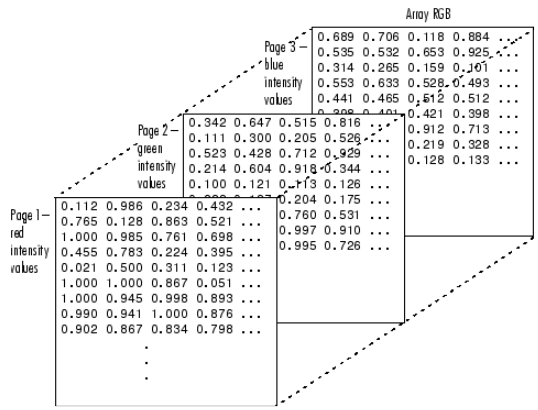
Predict with trained model

No label

Test dataset

Everything is numbers!

=



n-dimensional arrays called 'tensor', hence TensorFlow

Return to Table of Contents

What is Machine Learning?

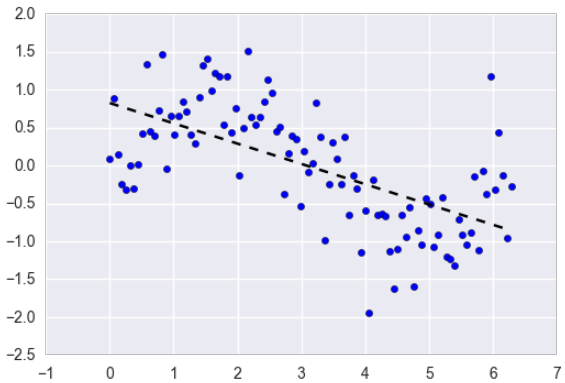
Previous

Learning types

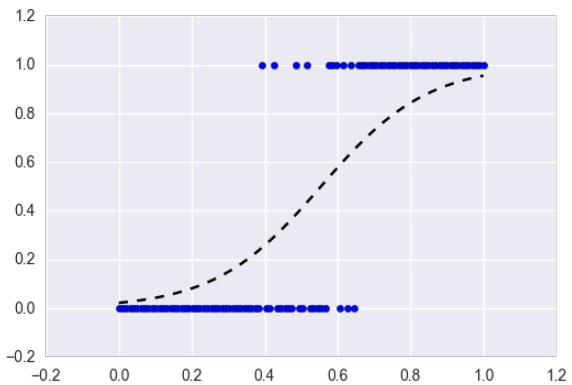
- Supervised learning
  - Apply labels to data ("cat", "spam")
  - Regression - Continuous, numeric variables:
    - Predict stock price, student test scores
  - Classification - categorical variables:
    - yes/no, decision tree
    - "is this email spam?" "is this picture a cat?"
  - Same types for 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

reward / punishment

Regression



Classification



Return to Table of Contents

Choose a Lesson

What is Machine Learning?

Working with Neural Networks

Preventing Overfitted Training Models

Working with Neural Networks

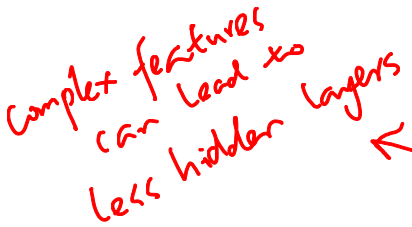
Next

Hands on learning tool

playground.tensorflow.org

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

$$w_1 x_1 + w_2 x_2 > b$$

Return to Table of Contents

Choose a Lesson

- What is Machine Learning?
- Working with Neural Networks
- Preventing Overfitted Training Models

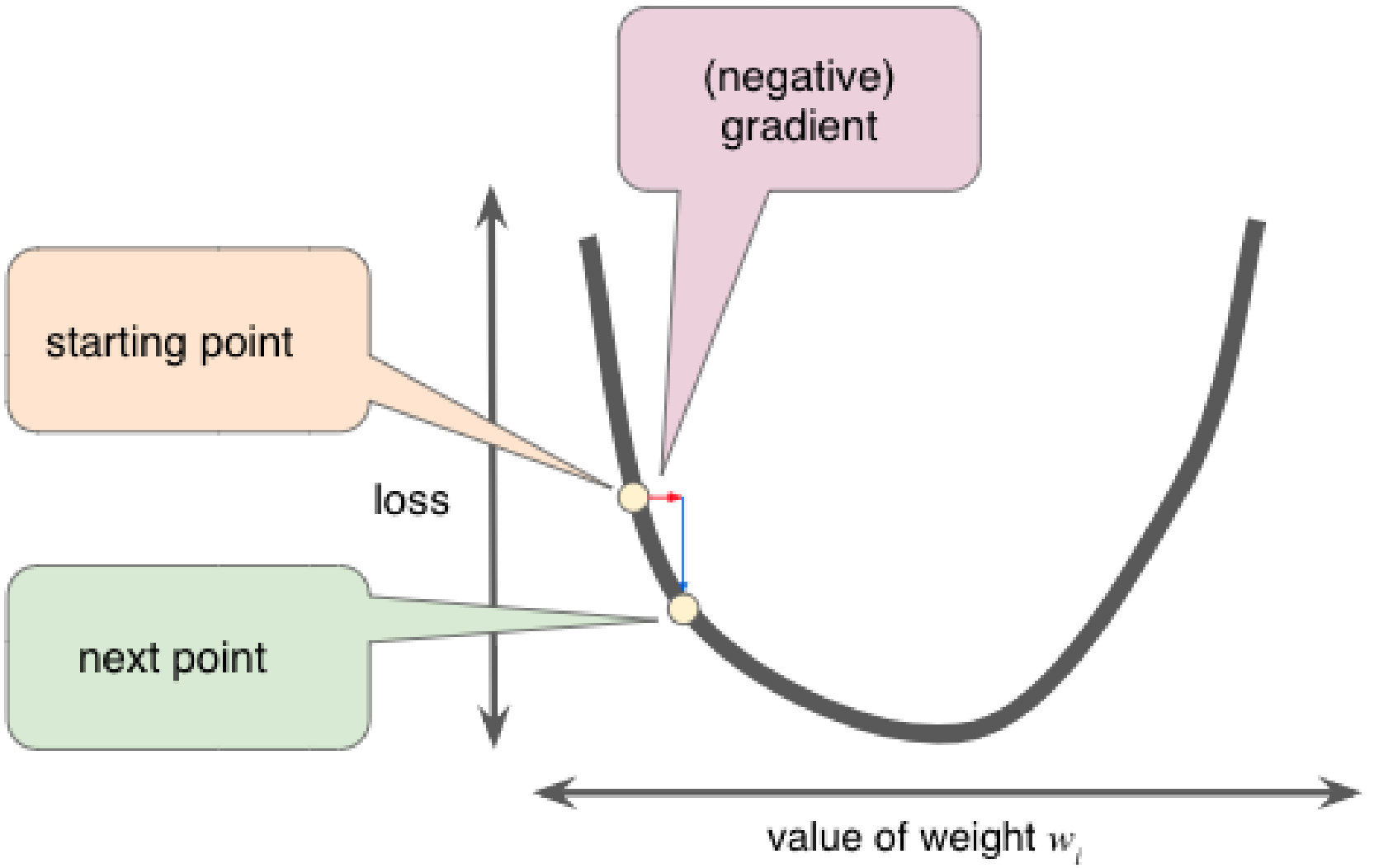
Working with Neural Networks

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

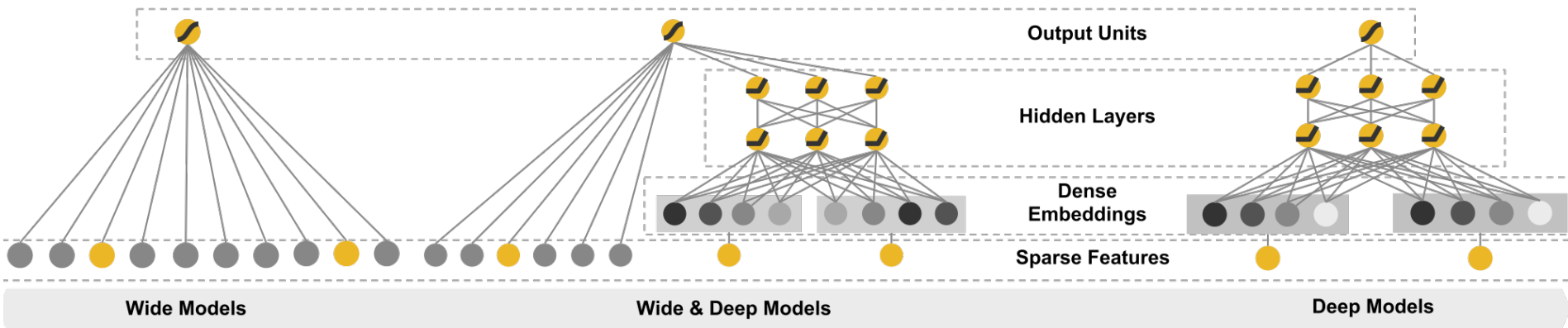
Preventing Overfitted Training Models

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

- What is Machine Learning?
- Working with Neural Networks
- Preventing Overfitted Training Models

Preventing Overfitted Training Models

Next

What is Overfitting?

- Training model *overfitted* to training data: Unable to **generalize** with new data
- Training model *fails to generalize*: Accounting for slightly different but close enough data

Causes of Overfitting:

1. Not enough training data
   - Need more variety of samples
2. Too many features
   - Too complex
3. Model fitted to unnecessary features unique to training data, a.k.a. “Noise”

Solving for Overfitted Model:

1. Use more data:
   - Add more training data.
   - More varied data allows for better generalization.
2. Make the model less complex:
   - Use less (but more relevant) features.
   - Combine multiple co-dependant/redundant features into a single representative feature:
     - This also helps reduce model training time.
3. Remove *noise*:
   - Increase **regularization** parameters

Return to Table of Contents

Choose a Lesson

What is Machine Learning?

Working with Neural Networks

Preventing Overfitted Training Models

Preventing Overfitted Training Models

Previous

Regularization?

- Adds a penalty to a model as it becomes more complex
- Penalizing **parameters** = better generalization
- Cuts out *noise* and unimportant data, to avoid overfitting

Regularization types:

- **L1 and L2 regularization** - Different approaches to tuning out noise. Each has different use case and purpose.
- **L1** - Lasso Regression: Assigns greater importance to **more influential features**
  - Shrinks less important features influence to zero
  - Good for models with many features, some more important than others
  - Example: Choosing features to predict likelihood of home selling:
    - House price more influential feature than carpet color
- **L2** - Ridge Regression: Performs better when all the input features influence the output, and with all weights being of **roughly equal size**

Return to Table of Contents

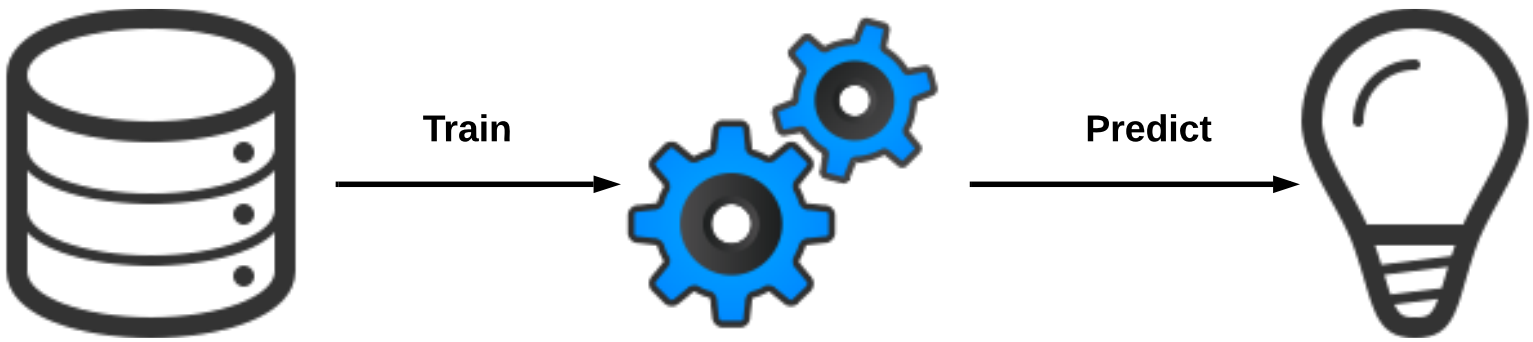
GCP Machine Learning Services

Choose a Lesson
GCP Machine Learning Services
AI Platform Overview
AI Platform Hands On

Next

Machine Learning - In a nutshell

- Algorithm that is able to learn from data



Lots of data
The more the better

ML Algorithm
Find patterns

Insight
Make intelligent decisions

Achieving this requires:
Lots of data (and data storage)
Lots of Compute
How can GCP help?

Return to Table of Contents

Choose a Lesson

- GCP Machine Learning Services
- AI Platform Overview
- AI Platform Hands On

GCP Machine Learning Services

Previous

Next

Different Roles = Different Priorities

- **Data Scientist/Machine Learning Engineer** → doing (math/parameters) details of ML Models
- **Application Developer**

Data Scientist/ML Engineer

- Works directly with ML libraries (e.g. Tensorflow)
- Creates, trains, and adjusts ML models
- "Does the math" for ML algorithms
- Gets into the ML details:
  - Parameters, biases, features, etc
- Values customization over simplicity

Application Developer

- Wants to 'plug in' ML capabilities to their app
- Avoids the mathematical details
- Values 'plug and play' solution

Return to Table of Contents

Choose a Lesson

Pre-trained ML API's

Vision API demo

Pre-trained ML API's

Next

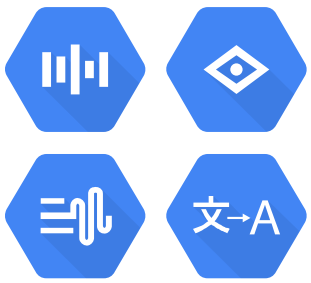
AI Platform (Formerly Cloud ML Engine)

- Train, deploy, and manage custom ML models on managed infrastructure resources.
- You create the model, then Google provides managed infrastructure for testing it.

Two main focus:

1) AI Platform

2) Pre-trained ML models



Pre-trained ML models

- Pre-trained models
- Common use cases (not customizable)
- Simply 'plug' into your application
- "Make Google do it"

Return to Table of Contents

Choose a Lesson

- GCP Machine Learning Services
- AI Platform Overview
- AI Platform Hands On

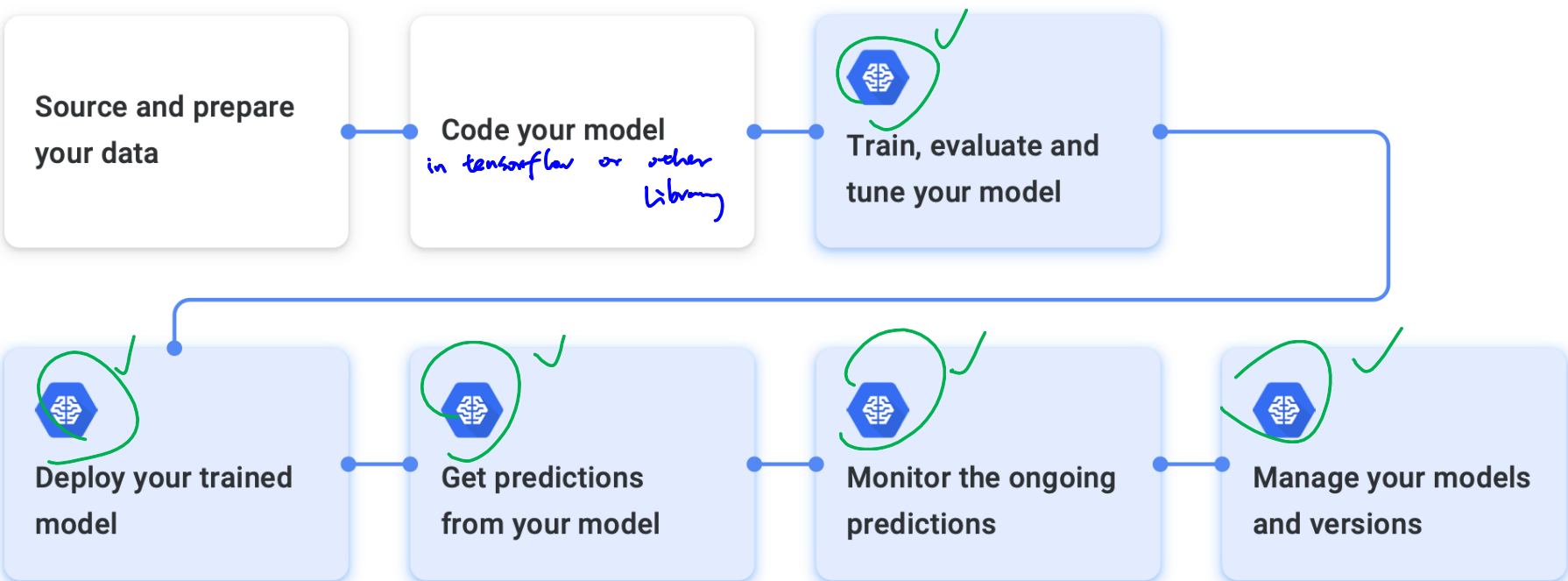
AI Platform Overview

Next

AI Platform (formerly Cloud ML Engine)

- Fully managed Tensorflow (and other ML libraries) platform
- Distributed training and prediction:
  - Breaks jobs down into pieces, distributes to multiple workers
- Scales to tens of CPUs/GPUs/TPUs
- Hyperparameter tuning with Hypertune
- Automate the "annoying bits" of machine learning
- "I want to train my own model, but automate it."

High Level Overview



Return to Table of Contents

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

AI Platform Overview

Previous

Next

Tensorflow? ML Libraries?

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

Return to Table of Contents

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

AI Platform Overview

Previous

Next

How AI Platform Works

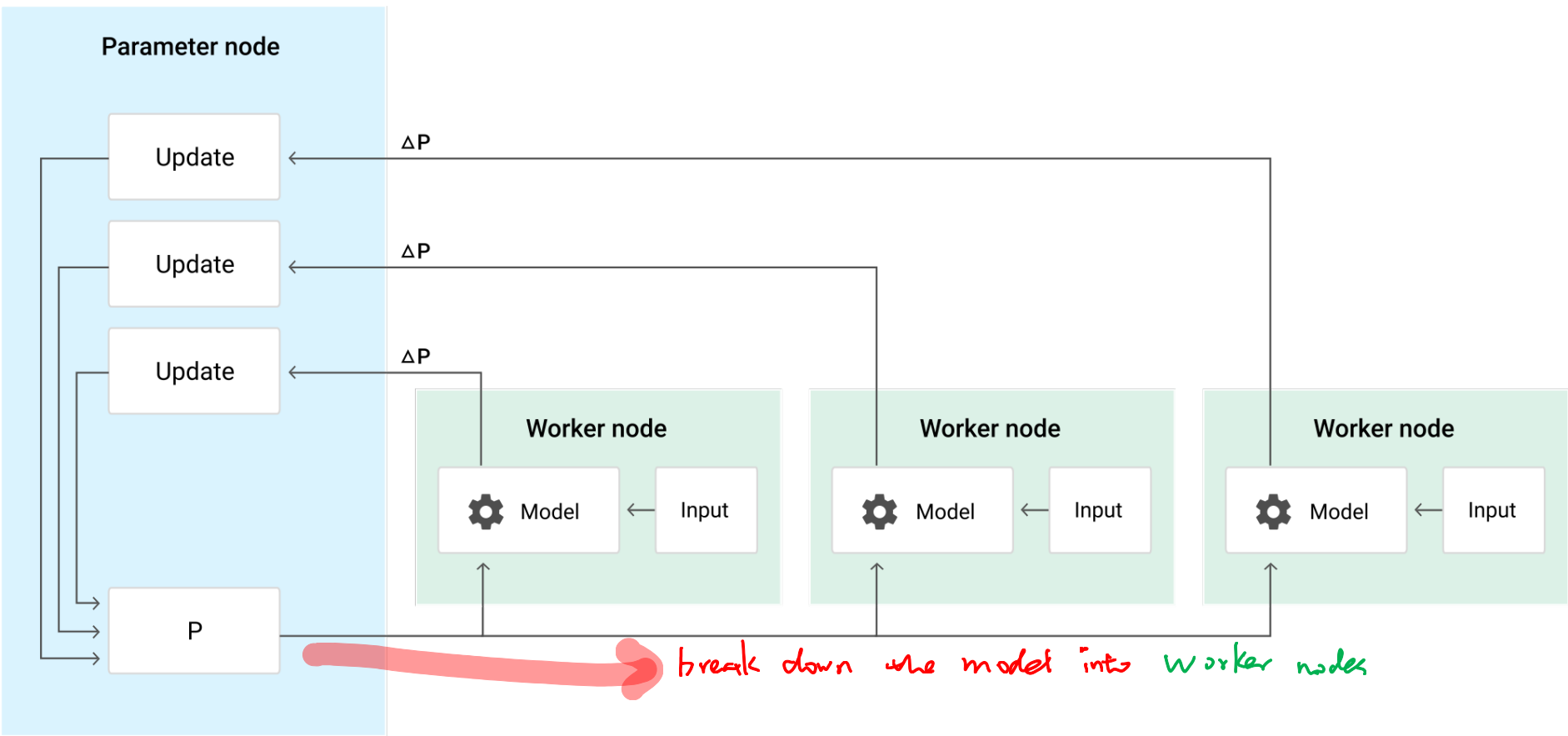
Prepare trainer and data for the cloud:

- Write training application in Tensorflow (or other ML library)
- Python is language of choice

Train your model with AI Platform:

- **Master** - Manages other nodes (1 master node for many cluster size)
- **Workers** - Works on portion of training job
- **Parameter servers** - Coordinates shared model states between workers

3 main components



Return to Table of Contents

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

AI Platform Overview

Previous

Next

Get Predictions - two types: (similar to batch and streaming)

- **Online:** (quick in + quick out)
  - 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

Key Terminology

- **Model** - Logical container of 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 AI Platform:
  - Train models: (is a job)
    - Command = 'submit job train model' on AI Platform
  - Deploy trained models: (is a job)
    - Command = 'submit job deploy trained model' on AI Platform
  - 'Failed' jobs can be monitored for troubleshooting

Return to Table of Contents

AI Platform Overview

Previous

Next

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

Typical process

Develop trainer locally

Quick iteration
No charge for cloud resources

Run trainer in AI Platform

Distributed computing across multiple nodes
Compute resources are deleted when job completes

Deploy a model from trainer output

Send prediction request

Cloud Storage

– primary storage location

Return to Table of Contents

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

AI Platform Overview

Previous

Next

Must-Know Info

- Currently supports Tensorflow, scikit-learn, and XGBoost frameworks

  Note: This list is subject to change over time
-

IAM roles:

- **Project and Models:**
  - **Admin** - Full control (assign role to people)
  - **Developer** - Create training/prediction jobs, models/versions, and send prediction requests (only can't assign role to people)
  - **Viewer** - Read-only access to above
- **Models only:**
  - **Model Owner:**
    - Full access to model and versions (assign members)
  - **Model User:**
    - Read models and use for prediction
    - Easy to share specific models

IAM

Project level

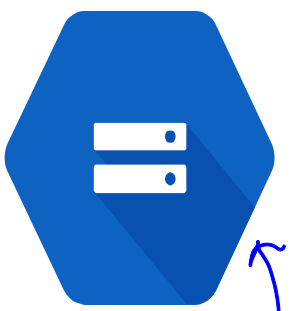
Model level

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 (not BigQuery)

direct read

BigQuery



Cloud Storage

recommend

copy to cloud storage first

AI Platform

Return to Table of Contents

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

AI Platform 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 (= Mix and match)

At most a single Master node !!!

GPU/TPU:

- Much faster processing performance

Pricing:

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

Training - Predefined scale tiers - price per hour		Training - AI Platform machine types - price per hour		Training - Compute Engine machine types - price per hour		Training - Accelerators - price per hour	
BASIC	$0.1900	standard	$0.1900	n1-standard-4	$0.1900	NVIDIA_TESLA_K80	$0.4500
STANDARD_1	$1.9880	large_model	$0.4736	n1-standard-8	$0.3800	NVIDIA_TESLA_P4 (Beta)	$0.6000
PREMIUM_1	$16.5536	complex_model_s	$0.2836	n1-standard-16	$0.7600	NVIDIA_TESLA_P100	$1.4600
BASIC_GPU	$0.8300	complex_model_m	$0.5672	n1-standard-32	$1.5200	NVIDIA_TESLA_T4 (Beta)	$0.9500
BASIC_TPU	$4.6900	complex_model_l	$1.1344	n1-standard-64	$3.0400	NVIDIA_TESLA_V100	$2.4800
CUSTOM	See the tables of machine types.	standard_gpu	$0.8300	n1-standard-96	$4.5600	Eight TPU_V2 cores*	$4.5000
		complex_model_m_gpu	$2.5600	n1-highmem-2	$0.1184	Batch prediction - price per node hour	
		complex_model_l_gpu	$3.3200	n1-highmem-4	$0.2368	$0.0791	
		standard_p100	$1.8400	n1-highmem-8	$0.4736	Online prediction - Machine types - price per node hour.	
		complex_model_m_p100	$6.6000	n1-highmem-16	$0.9472	mls1-c1-m2 (default)	$0.0401
		standard_v100	$2.8600	n1-highmem-32	$1.8944	mls1-c4-m2 (Beta)	$0.1349
		large_model_v100	$2.9536	n1-highmem-64	$3.7888		
		complex_model_m_v100	$10.6800	n1-highmem-96	$5.6832		
		complex_model_l_v100	$21.3600	n1-highcpu-16	$0.5672		
		cloud_tpu*	$4.5000	n1-highcpu-32	$1.1344		
				n1-highcpu-64	$2.2688		
				n1-highcpu-96	$3.4020		

Return to Table of Contents

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

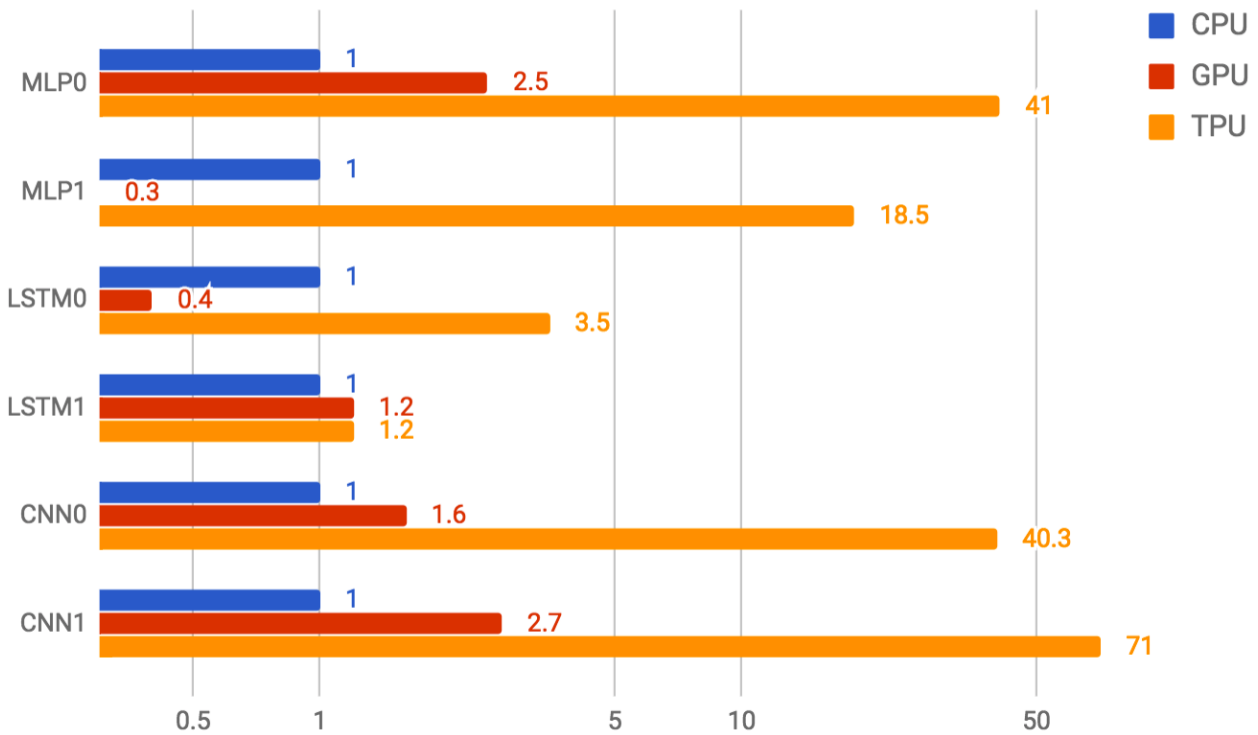
AI Platform Overview

Previous

Next

Tensor Processing Unit (TPU)

- Hardware processing specifically designed for machine learning
  - Like a GPU, but even more optimized for ML
  - Faster and more efficient

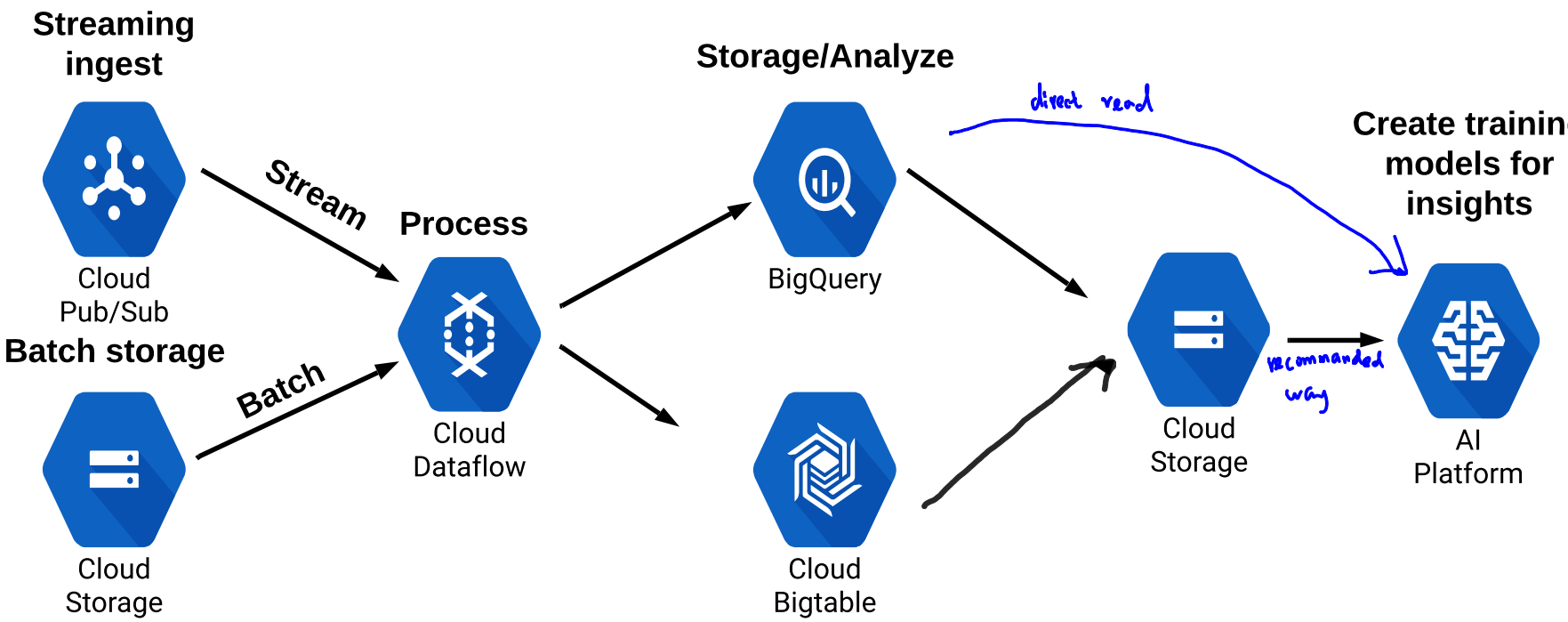


Return to Table of Contents

AI Platform Overview

Previous

Big Picture



Return to Table of Contents

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

AI Platform Hands On

What We Are Doing:

- Working with pre-packaged training model:
  - Focusing on the AI Platform aspect, not TensorFlow
- Heavy command line/`gcloud` focus, using Cloud Shell

Main steps: The big picture

- Submit training job locally using `ai-platform` commands
- Submit training job on AI Platform, both single and distributed
- Deploy trained model, and submit predictions

Instructions for Hands On

Download scripts to Cloud Shell to follow along:

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

Return to Table of Contents

Choose a Lesson

GCP Machine Learning Services

AI Platform Overview

AI Platform Hands On

ML Engine Hands On

Previous

Next

Current ML API's (page 1 of 2)

exam: familiar with different pre-train API !

How the API's are called upon.

Cloud Vision

Image recognition/analysis

Cloud Translation

Detect and translate languages

Cloud Natural Language

Text analysis
Information extraction
Understanding sentiment

Cloud Job Discovery

More relevant job searches:
Power recruitment, job boards

Cloud Speech to Text

Convert audio to text
Multi-lingual support
Understanding sentence structure

Cloud Text to Speech (Beta)

Convert text to audio
Multiple languages/voices
Natural sounding synthesis

Return to Table of Contents

ML Engine Hands On

Previous

Next

Current ML API's (page 2 of 2)

Cloud Video Intelligence

Video analysis

Labels, shot changes, explicit content

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

Current ML APIs (new ones being added)

Cloud Vision

Image recognition/analysis

Cloud Translation

Detect and translate languages

Cloud Natural Language

Text analysis
Extract information
Understand sentiment

Cloud Job Discovery

More relevant job searches:
Power recruitment, job boards

Cloud Speech to Text

Convert audio to text
Multi-lingual support
Understand sentence structure

Cloud Text to Speech

Convert text to audio
Multiple languages/voices
Natural sounding synthesis

Cloud Video Intelligence

Video analysis
Labels, shot changes, explicit content

Dialogflow for Enterprise

Conversational experiences
Virtual assistants

Return to Table of Contents

Choose a Lesson

- GCP Machine Learning Services
- AI Platform Overview
- AI Platform Hands On

GCP Machine Learning Services

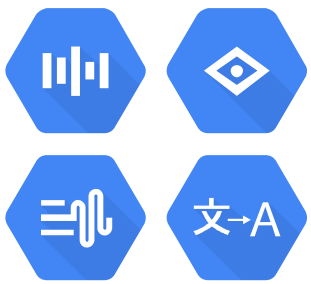
Previous

ML Options on Google Cloud Platform

- Products for ML Engineer to Developer roles, and everything in between
- Rapid expansion of solutions: Two primary ones to focus on for exam

AI Platform (Formerly Cloud ML Engine)

- Train, deploy, and manage custom ML models on managed infrastructure resources
- You create the model, Google provides managed infrastructure for testing it



Pre-trained ML models

- Pre-trained models
- Common use cases (not customizable)
- Simply 'plug' into your application
- "Make Google do it"

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 APIs (new ones being added)

Data Loss Prevention API

Detect, Manage, and Redact Sensitive data

- **Credit card numbers, SSN, birthdates, credentials**

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: A Closer Look

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

Newer ML options

1. **Auto ML**
   - Pre-trained APIs, but for custom models!
     - Example: Identify specific geographical features
   - Supply your own data to train on
   - Currently available for:
     - Vision
     - Video Intelligence
     - Natural Language
     - Translation
     - Structured Data
2. **BigQuery ML**
   - Create and train ML models inside BigQuery
   - Use SQL syntax to create models

Return to Table of Contents

Choose a Lesson

Pre-trained ML API's

Vision API demo

Pre-trained ML API's

Previous

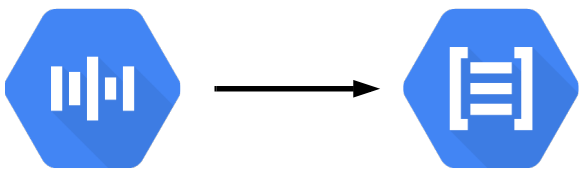
Exam Perspectives

How to covert images, video, etc. for use with API?

- Can use Cloud Storage URI for GCS stored objects
- Encodes 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

Return to Table of Contents

Choose a Lesson

Pre-trained ML API's

Vision API demo

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

Datalab Overview

Next

What is it?

Google cloud datalab

- 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
  (Google cloud Engine; Virtual private cloud network; repository)
- Cost: free - only pay for GCE resources Datalab runs on and other Google Cloud services you interact with

Java and SQL are used specifically for BigQuery

the only paid: GCE instance fee (runs datalab)

Not a feature exam topic

just how to share content to datalab

We also use bash script

Return to Table of Contents

Choose a Lesson

Datalab Overview

Datalab Overview

Previous

Next

How It Works

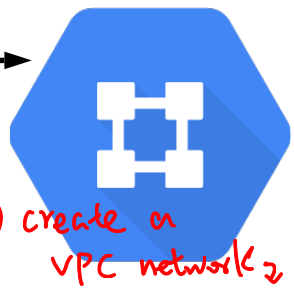
Create and connect to a Datalab instance

gcloud command:

datalab create (instance-name)

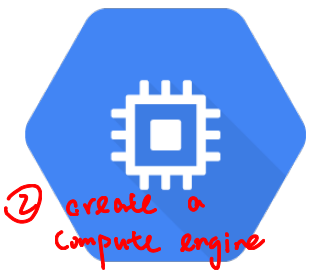
after typing command, 3 things happen:

1. Connect via SSH and open web preview
2. datalab connect (instance-name) — command
3. Open web preview - port 8081



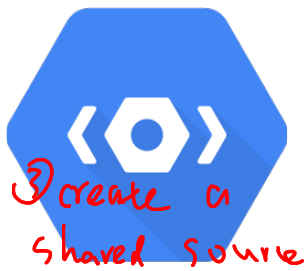
① create a VPC network

datalab-network



② create a compute engine

datalab-instance



③ create a shared source

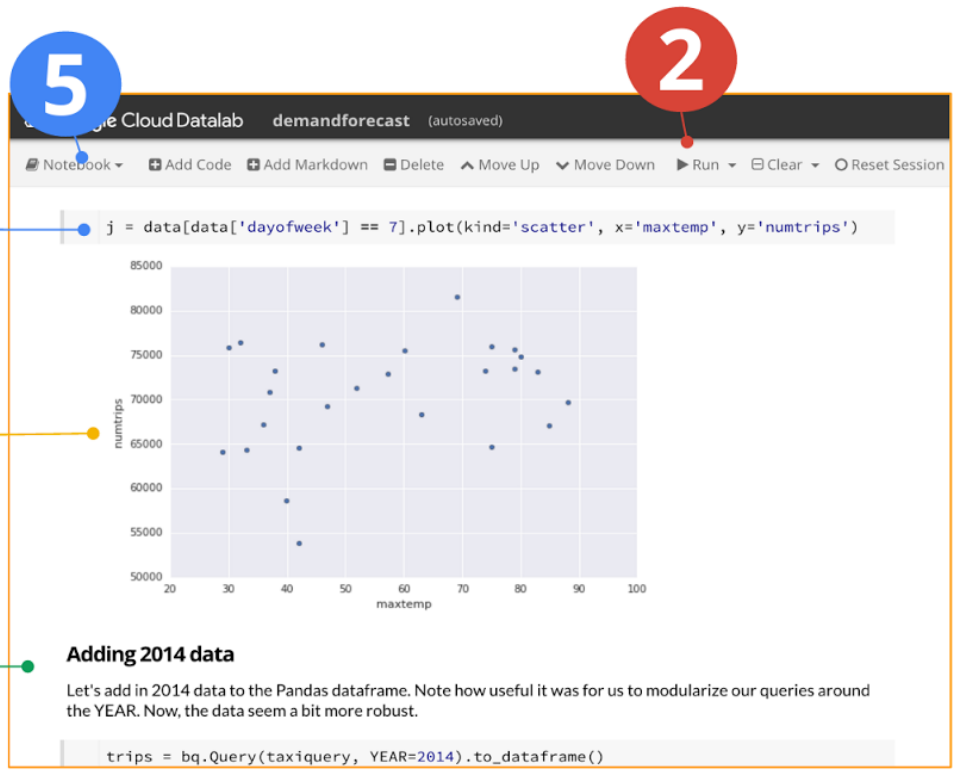
datalab-notebooks

Source repository

(share data between users)

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



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 (or Sevice account user role)
- Notebook access per user only
- Sharing data performed via shared Cloud Source Repository
- Sharing is at the project level

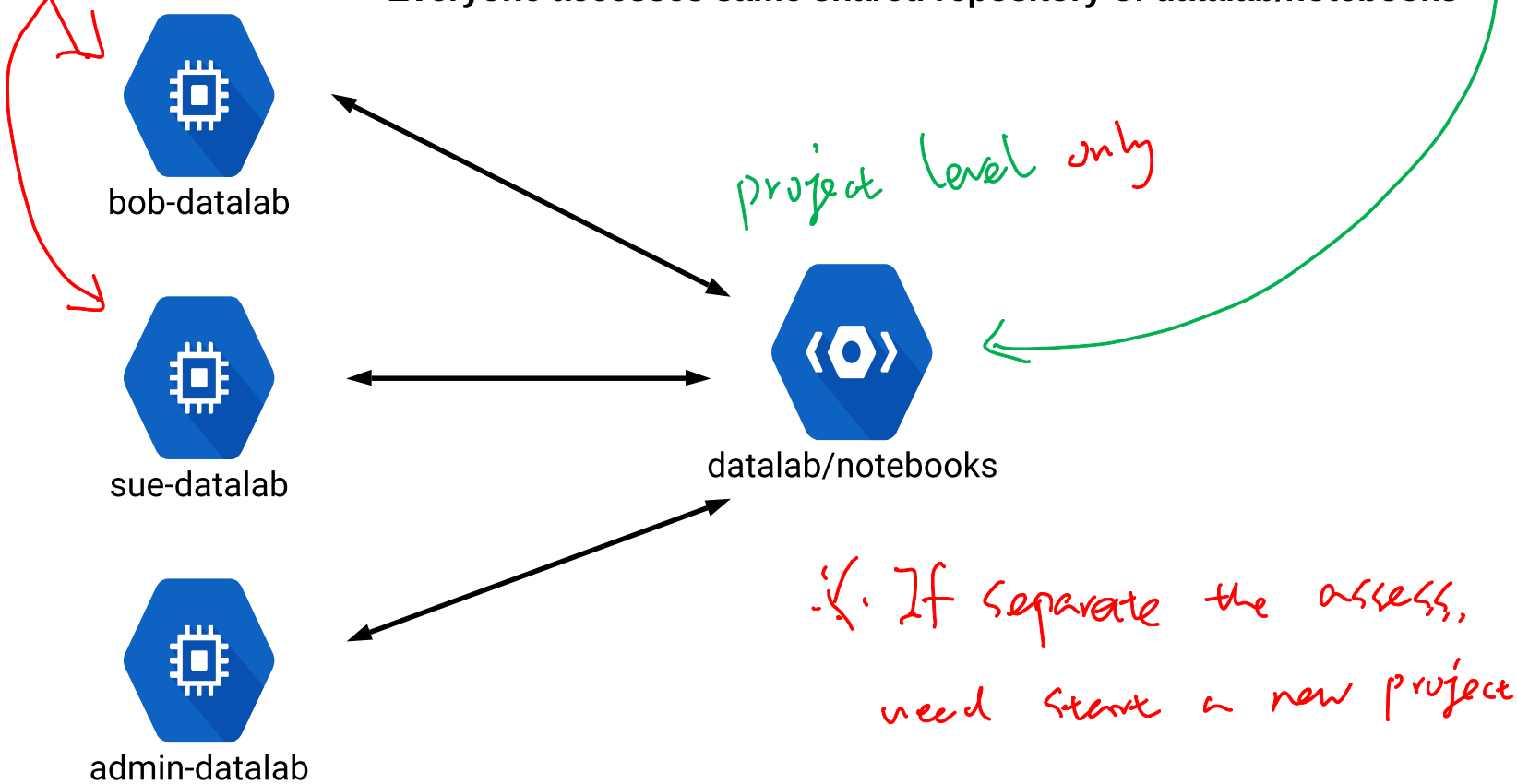
Creating team notebooks - two options:

option 1

- Team lead creates notebooks for users using --for user option:
  - datalab create [instance] --for-user bob@professionalwireless.net ← user ID

option 2

- 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?

What is Dataprep?

Next

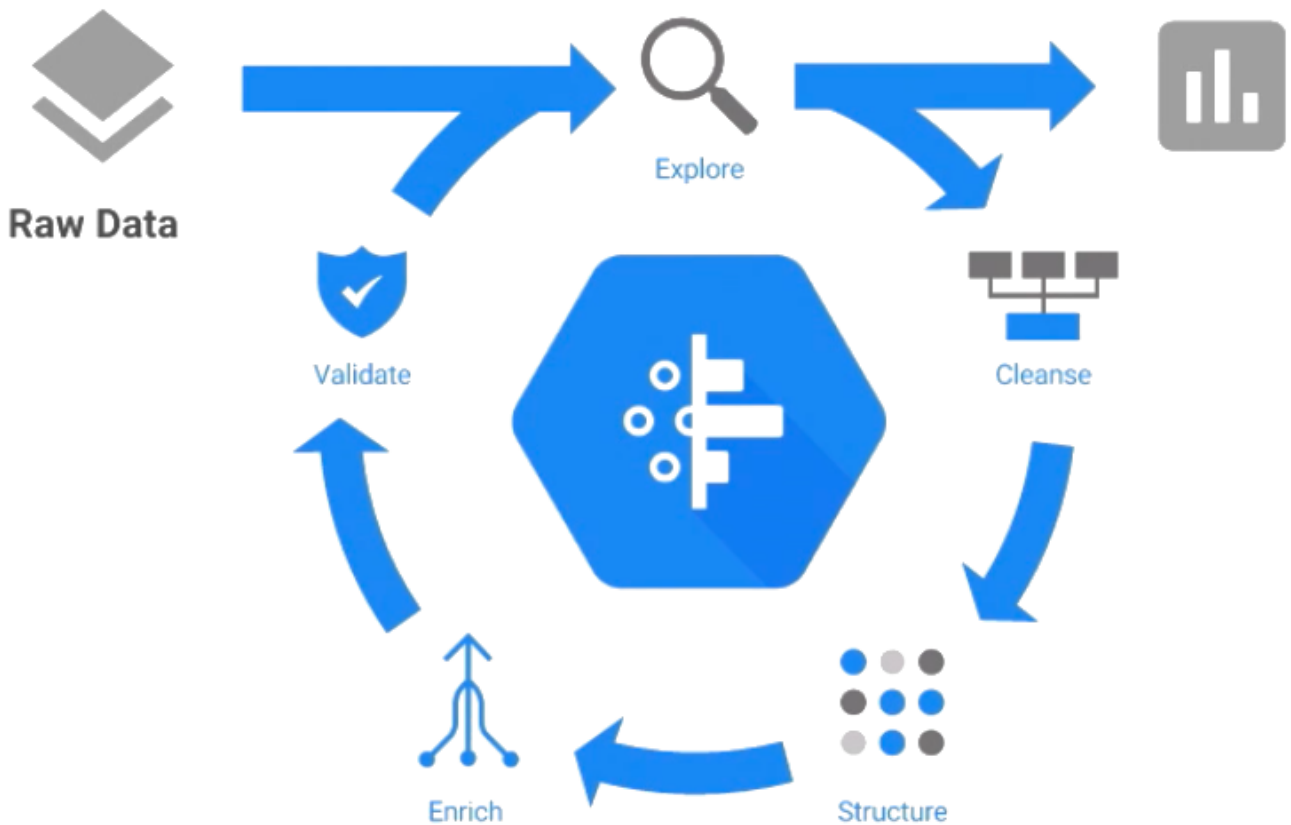
What is it?

point-and-click web interface

- Intelligent data preparation platform
- 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?

What is Dataprep?

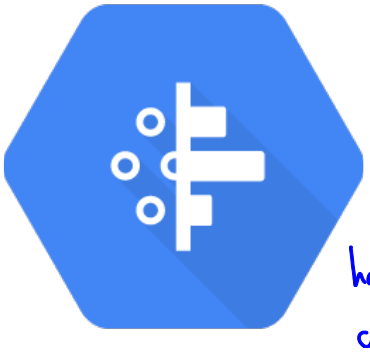
Previous

How It Works

create a dataflow job

Backed by Cloud Dataflow:

- After preparing, Dataflow processes via Apache Beam pipeline
- "User-friendly Dataflow pipeline" (Don't need to know Java)



Dataprep process:

hands-on demo

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

Intelligent suggestions: (exam topic)

suggestion + recipe

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

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

Data Studio Introduction

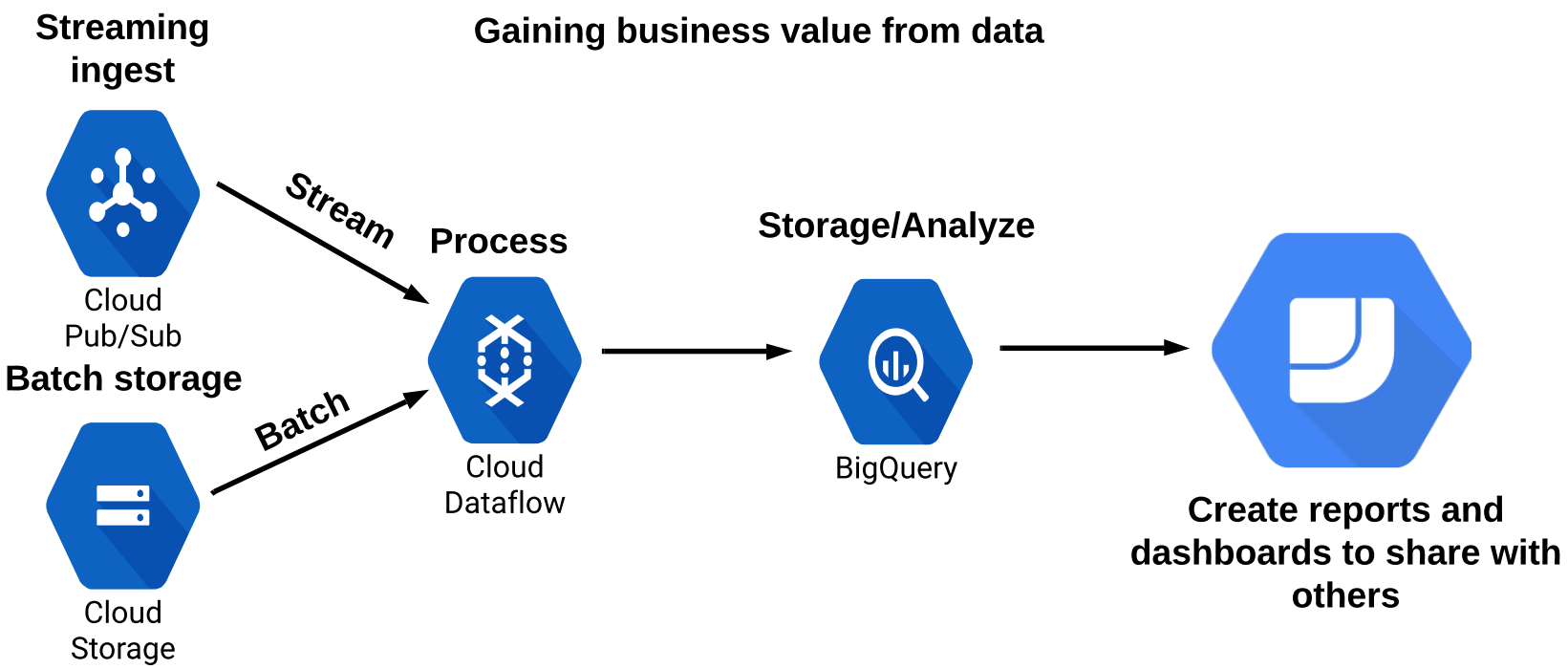
Next

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) IAM has nothing to do with Data Studio.
  - Google account permissions in GCP will determine data source access
  - Files saved in Google Drive Not Google cloud
- 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
    If pulling data from BigQuery, it's not free!

Data Lifecycle - Visualization

Gaining business value from data



Return to Table of Contents

Choose a Lesson

Data Studio Introduction

Data Studio Introduction

Previous

Not a part of Google cloud

Basic process

- Connect to data source
- Visualize data
- Share with others (G-suite permission, not GC permission)

Creating charts

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

exam topics ?

Caching - options for using cached data performance/costs (default option)

Two cache types, query cache and prefetch cache

Query cache:

- Remembers queries issues 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 ← exam point !!!
- Can be turned off

When to turn caching off:

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

Return to Table of Contents

Choose a Lesson

Cloud Composer Overview

Hands On - Cloud Composer

Cloud Composer Overview

Next

What is Cloud Composer?

- Fully managed **Apache Airflow** implementation:
  - Infrastructure/OS handled for you

open source

What is Apache Airflow?

- Programatically create, schedule, and monitor data workflows

Why is this important?

- Automation and monitoring
- Big data pipelines are often a multi-step, complex process:
  - Create resources in multiple services
  - Process and move data from one service to another
  - Remove resources when they complete a task (e.g. Dataproc cluster)
- Collaborate workflow process with other team members

How Airflow/Composer helps

- Automates the above steps, including scheduling
- Built on open source, using Python as common language
- Easy to work with, and share workflow with others
- Works with non-GCP providers (on-premises, other clouds)

Return to Table of Contents

Choose a Lesson

Cloud Composer Overview

Hands On - Cloud Composer

Cloud Composer Overview

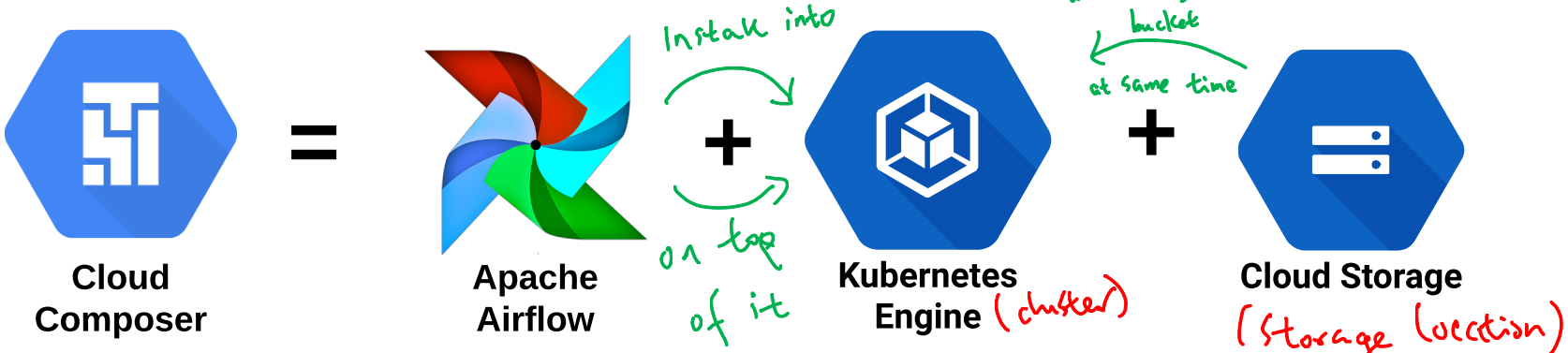
Previous

Next

How It Works

Behind the scenes:

- GKE cluster with Airflow implemented
- Cloud Storage bucket for workflow files (and other application files)



Workflows?

- Orchestrate data pipelines:
  - Like a walkthrough of tasks to run
- Format = Direct Acyclic Graph (DAG):
  - Written in Python
  - Collection of organized tasks that you want to schedule and run
- **Cloud Composer** creates **workflows** using **DAG** files

important terminology

python application

Summary!

The Process

different from Gcloud variables

- Create Composer Environment (kubernete instance cluster)
- Set Composer variables (i.e. project ID, GCS bucket, region)
- Add Workflows (DAG files), which Composer will execute

Return to Table of Contents

Choose a Lesson

Cloud Composer Overview

Hands On - Cloud Composer

Cloud Composer Overview

Previous

to cloud storage bucket (to save cost)

Examples and Exam Perspective

- Create a Dataproc cluster, submit a job, and then delete the cluster.
- Execute a Cloud Dataflow pipeline from data in GCS, and write output to BigQuery.
- Ingest third party data into Cloud Dataflow, process, then upload to GCS.
- **Exam perspective:** Know what DAGs are, and why you'd want to use workflows.

exam topics!!!

Return to Table of Contents

Choose a Lesson

Cloud Composer Overview

Hands On - Cloud Composer

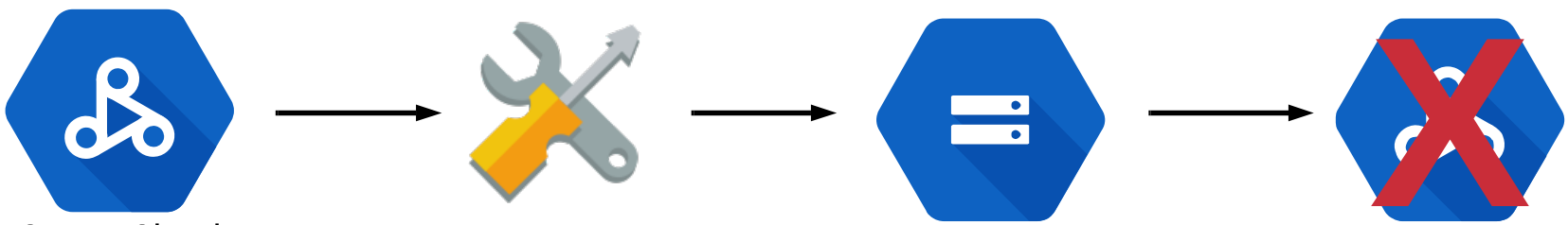
Hands On - Cloud Composer

Next

The Process:

- Create the Composer environment. (cluster)
- Then create the GCS bucket for Dataproc output.
- Assign Cloud Composer variables.
- Upload the workflow file to DAG folder.
- View the results.

Automatic processes -- Workflow



Create Composer Environment

- Enable Composer/Dataproc API
- Create environment in closest region:
  - What's happening?
  - Creating GKE cluster + GCS bucket

Create GCS bucket to output Dataproc results

- `gsutil mb -l us-central1 gs://output-$DEVSHELL_PROJECT_ID`

Return to Table of Contents

Choose a Lesson

Cloud Composer Overview

Hands On - Cloud Composer

Hands On - Cloud Composer

Previous

Configure Cloud Composer Variables

- **Format**
  - **gcloud composer environments run (ENVIRONMENT_NAME) --location (LOCATION) variables -- --set (KEY VALUE)**
- `gcloud composer environments run my-environment --location us-central1 variables -- --set gcp_project (PROJECT-ID)`
- `gcloud composer environments run my-environment --location us-central1 variables -- --set gcs_bucket gs://output-(PROJECT-ID)`
- `gcloud composer environments run my-environment --location us-central1 variables -- --set gce_zone us-central1-c`

Add workflow file (Python) to Composer DAG folder:

- **github link**

Next step? There is none! Cloud Composer will take it from here...