Return to Table of Contents

Choose a Lesson

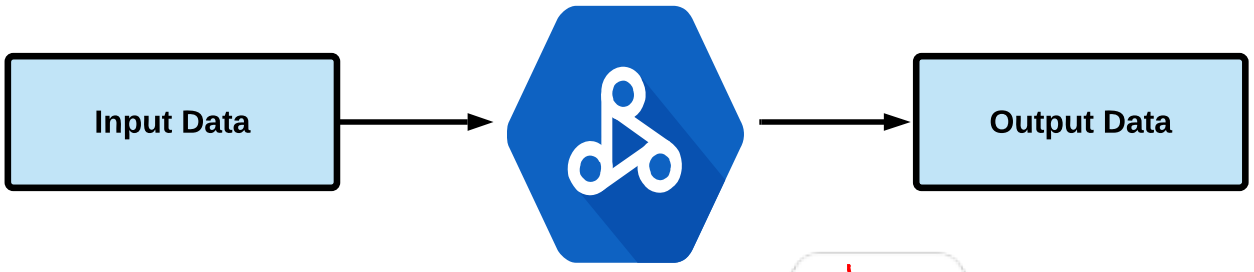
- Dataproc Overview
- Configure Dataproc Cluster and Submit Job
- Migrating and Optimizing for Google Cloud
- Best Practices for Cluster Performance

Dataproc Overview

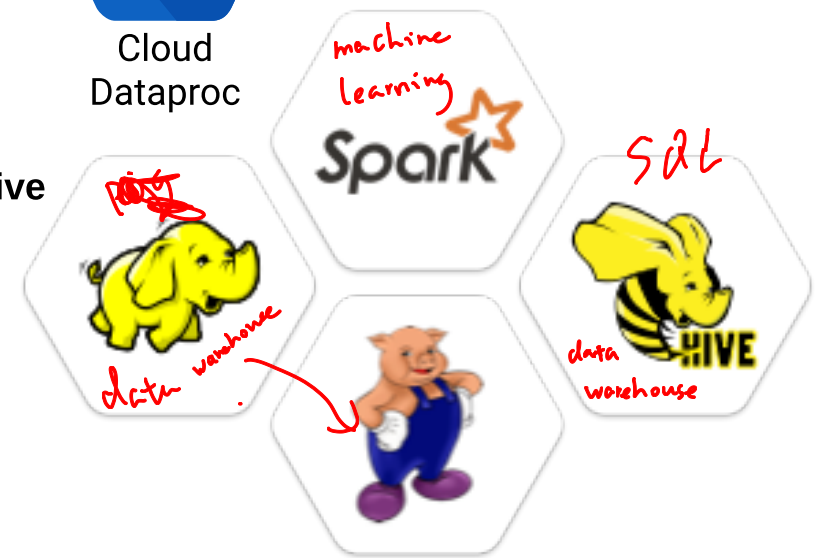
Next

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)
  - Greatly reduces administrative overhead
  - (still need to set up)
- 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

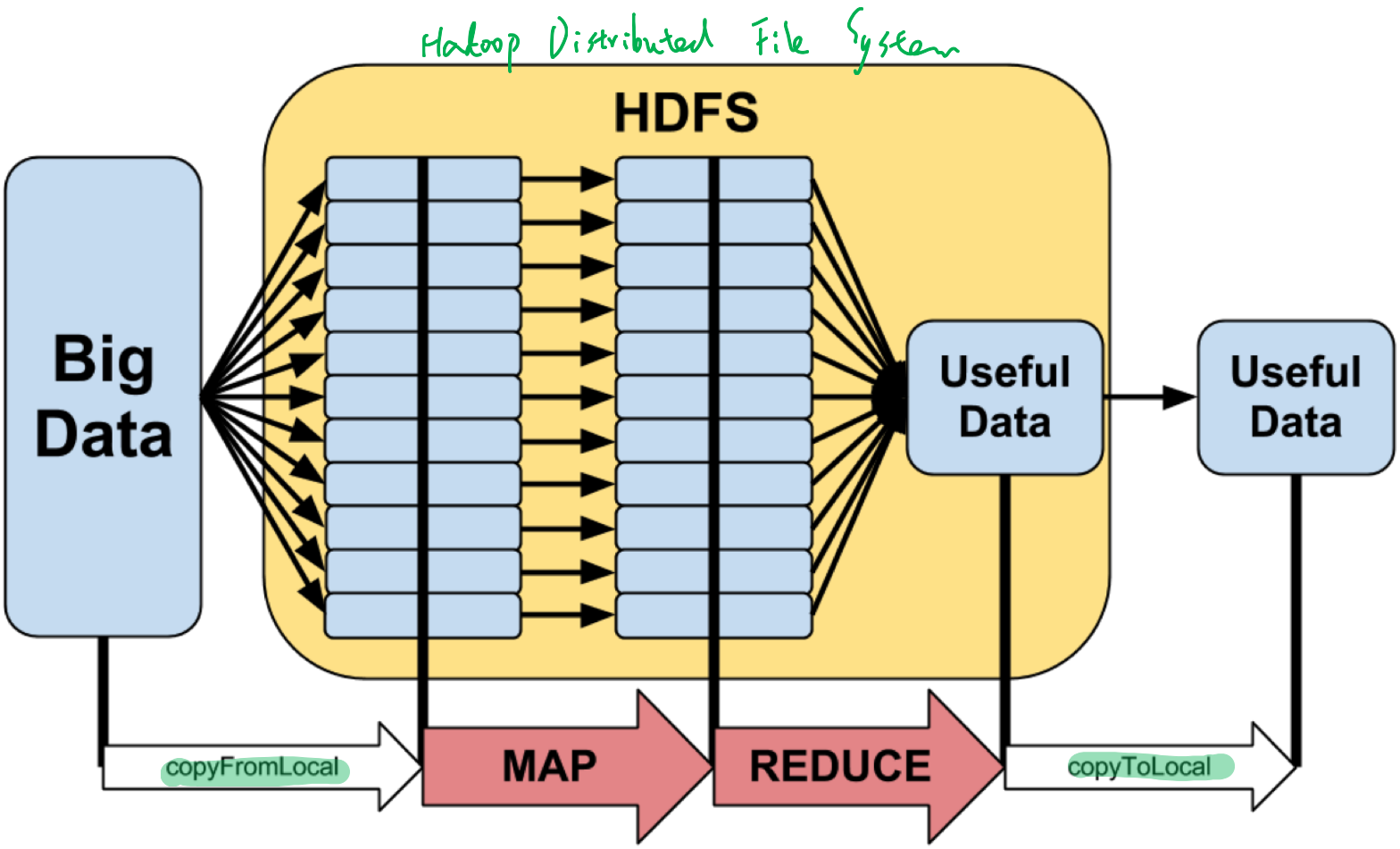
Dataproc Overview

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

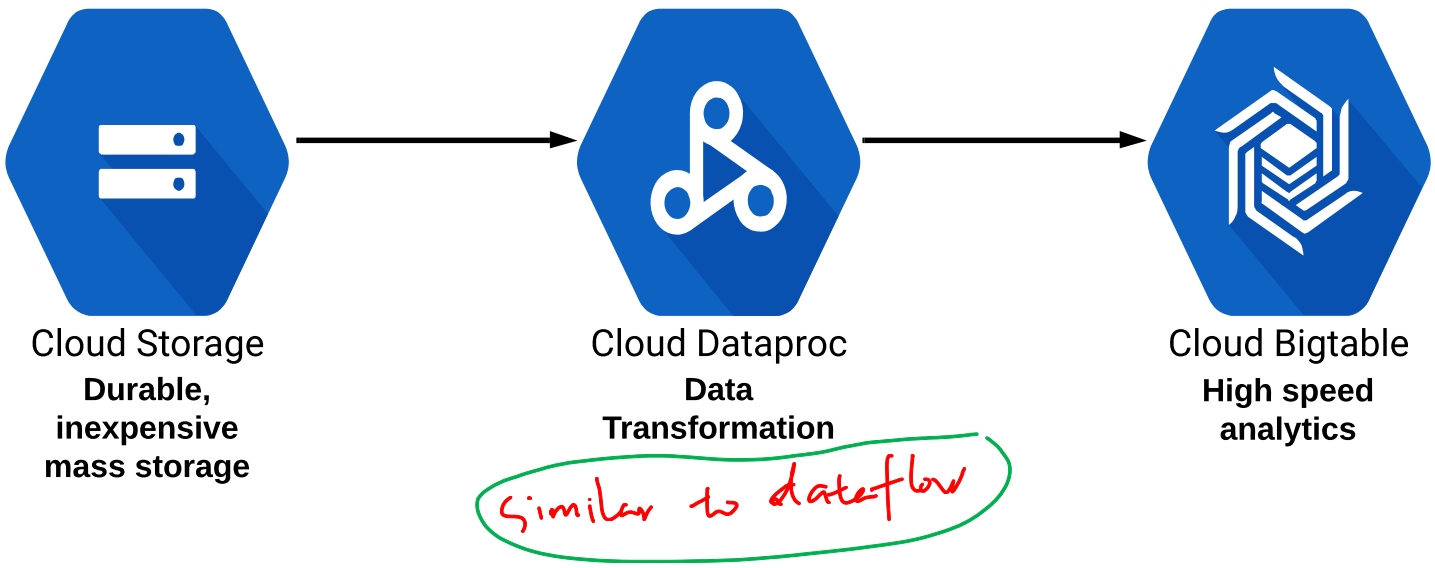
Pricing:

- **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

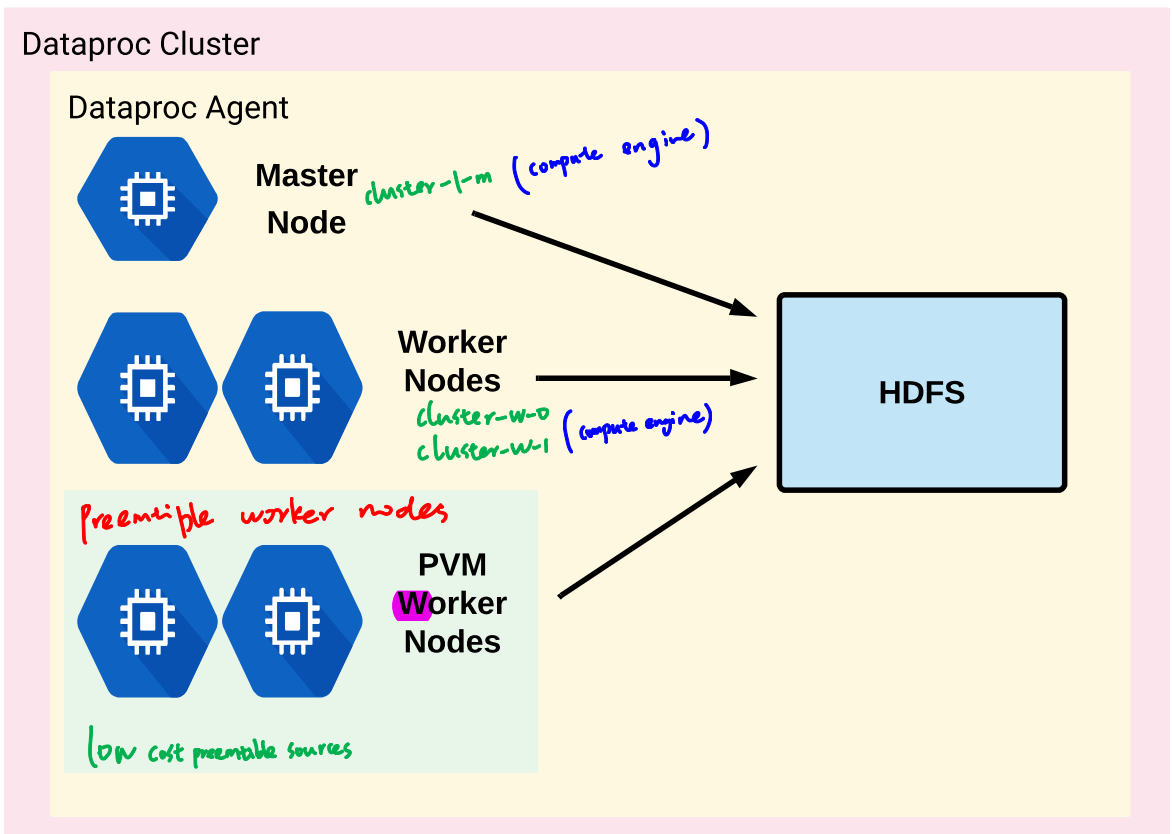
Next

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:

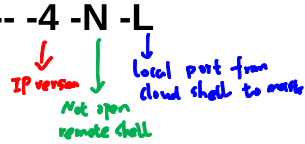
- SSH to master to enable port forwarding:
  - gcloud compute ssh *master-host-name* --project=*project-id* --zone=*master-host-zone* -- -D 1080 -N
- Open new terminal window - launch web browser with parameters (varies by OS/browser):
  - "/Applications/Google Chrome.app/Contents/MacOS/Google Chrome"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

Choose a Lesson

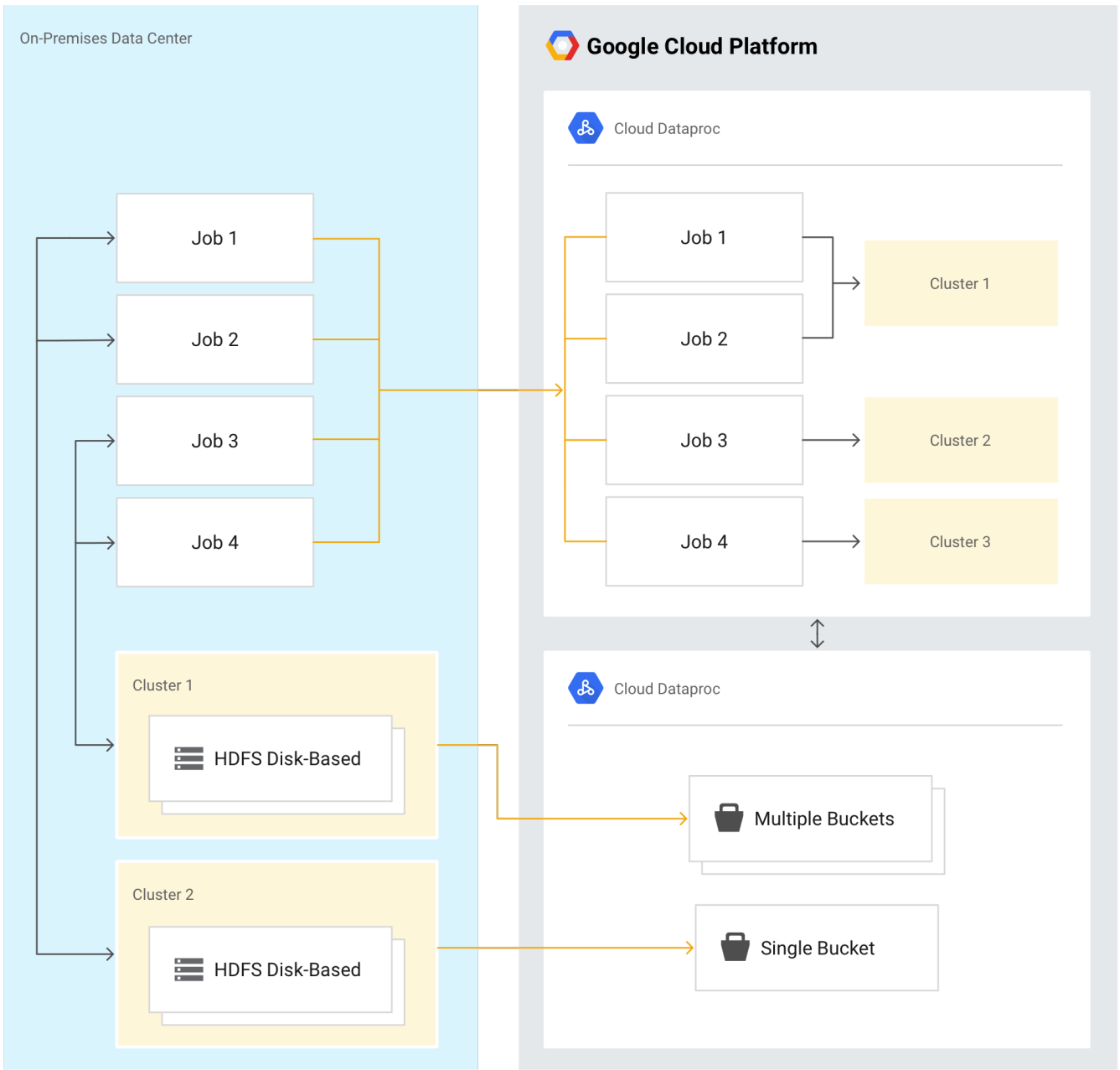
Migrating and Optimizing for Google Cloud

Next

Migrating to Cloud Dataproc

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

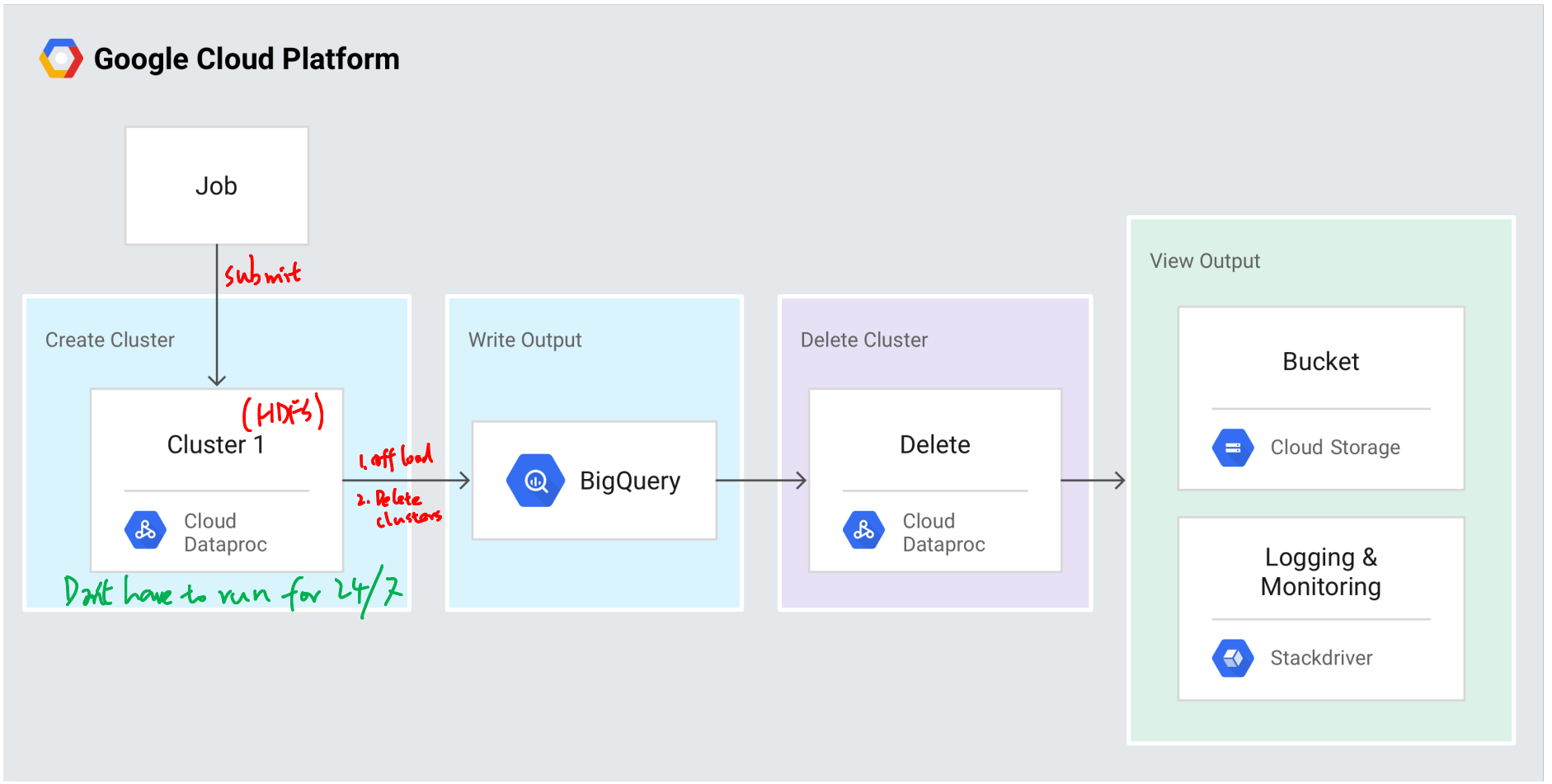
Previous

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

multi-region ~~single region~~

- 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