Product Recommendations using Cloud SQL and Spark

Agenda

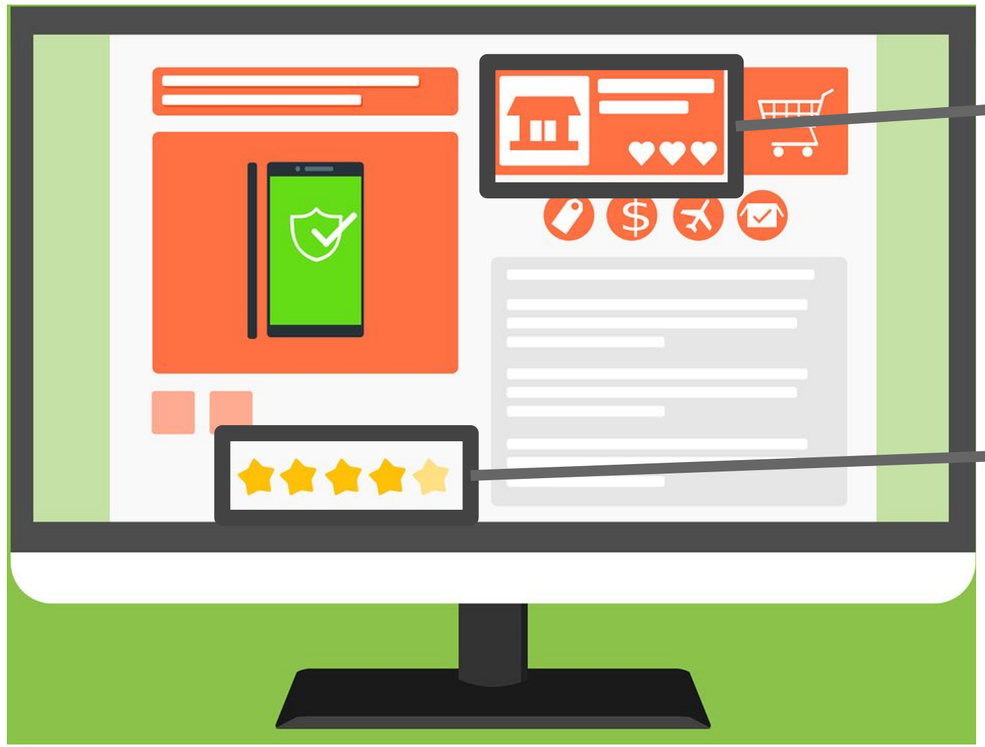
Recommendation systems

- Business applications
- Scenario: ML for housing rentals

Choosing the right solution approach

- On-premise to Google Cloud Platform
- Challenge: Utilizing and tuning on-premise clusters
- Off-cluster storage with Google Cloud Storage
- Storing recommendations

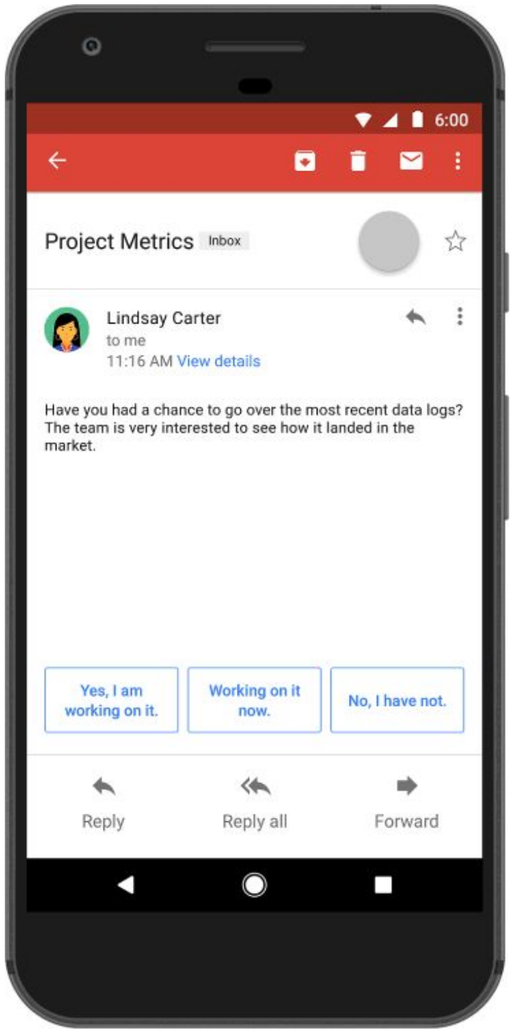
Where have you seen recommendation systems before?



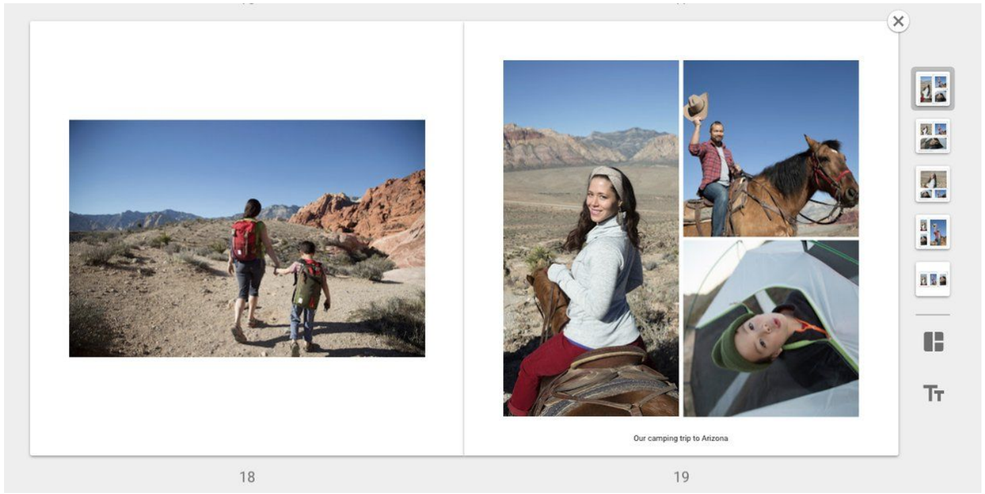
Model recommends products you may like based on your preferences

You can give feedback to better inform the model

Recommendation systems are used in many applications

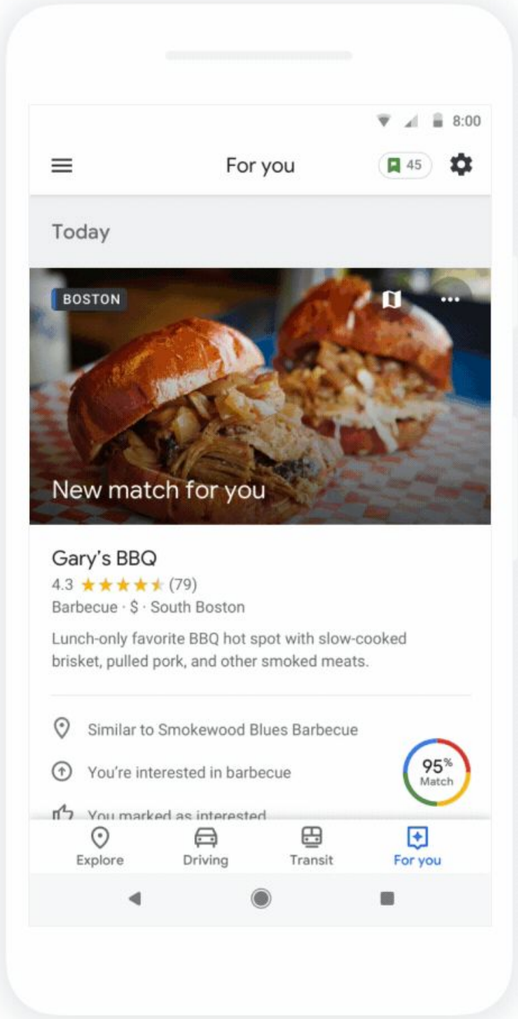


Smart Reply in Gmail recommends 3 possible answers to your emails

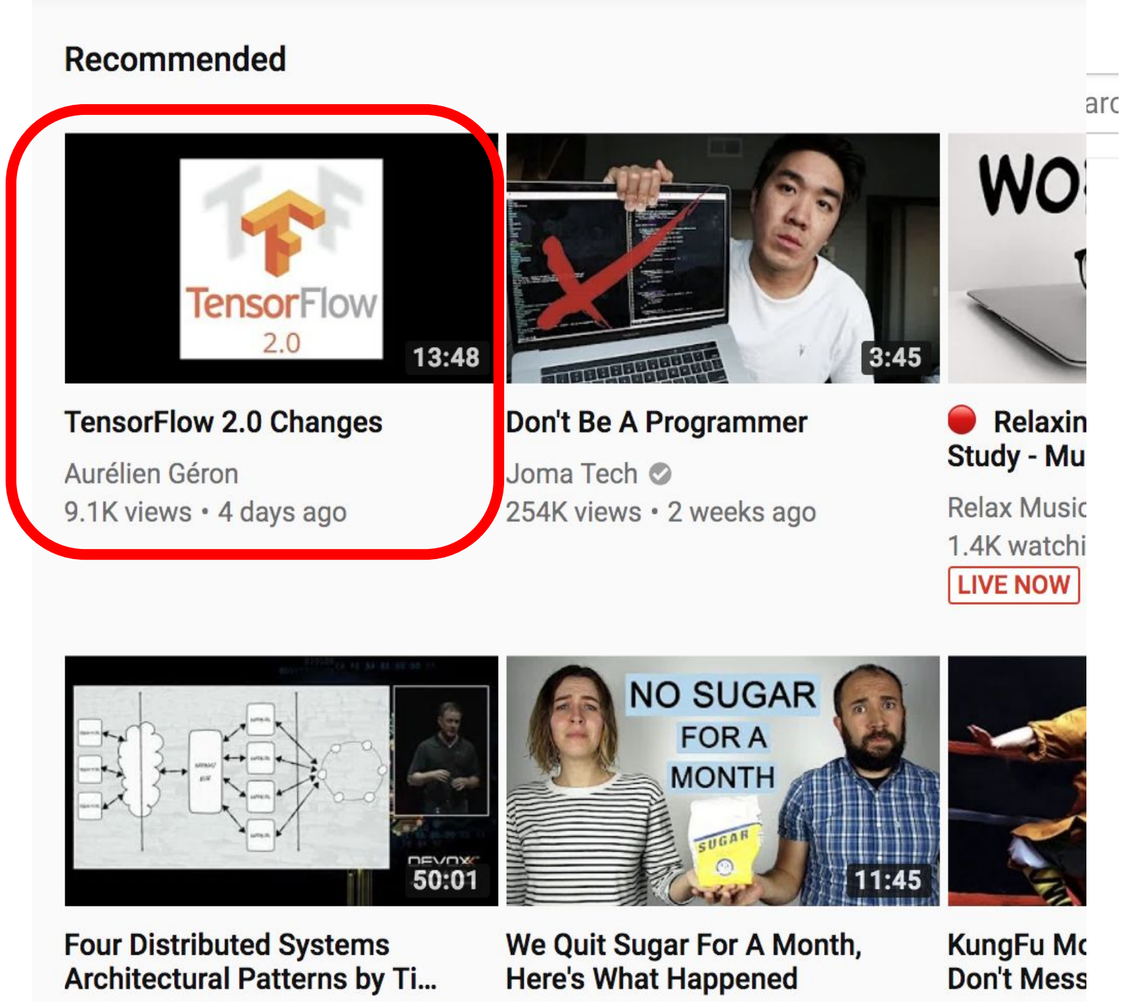


Google Photos recommends similar pictures to include in an album

Google Maps recommends restaurants based on what you like



Recommendation systems must scale to meet demand



Agenda

Recommendation systems

- Business applications
- Scenario: ML for housing rentals

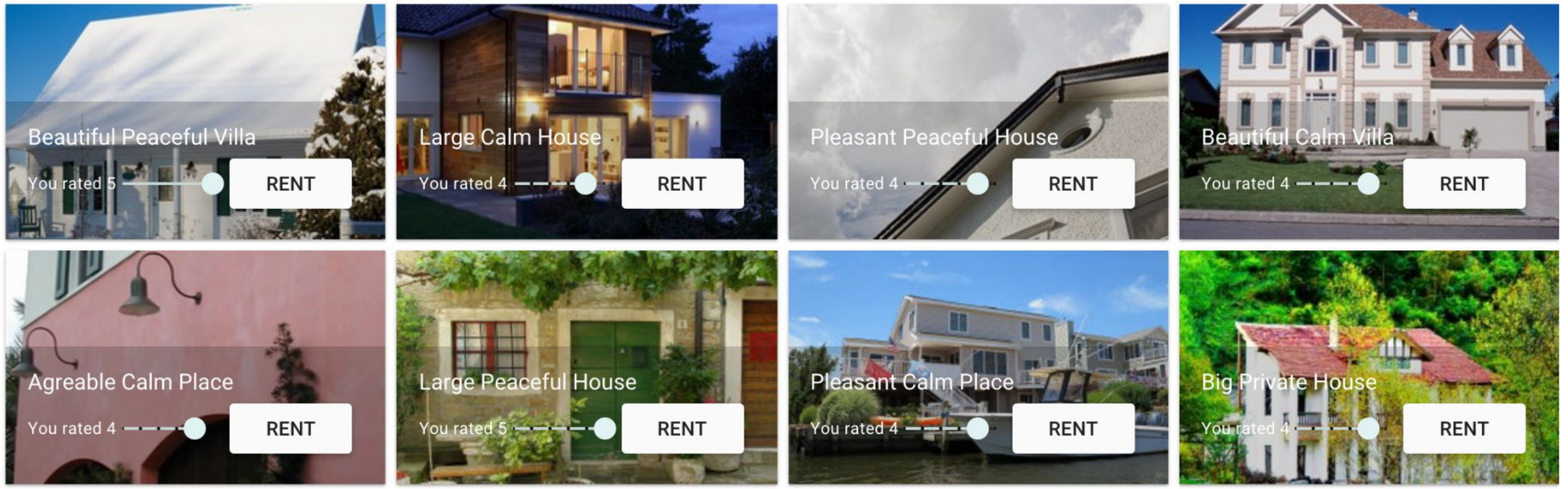
Choosing the right solution approach

- On-premise to Google Cloud Platform
- Challenge: Utilizing and tuning on-premise clusters
- Off-cluster storage with Google Cloud Storage
- Storing recommendations

Recommendation systems require **data**, a **model**,
and training/serving **infrastructure**

Use case: Recommending housing rental options

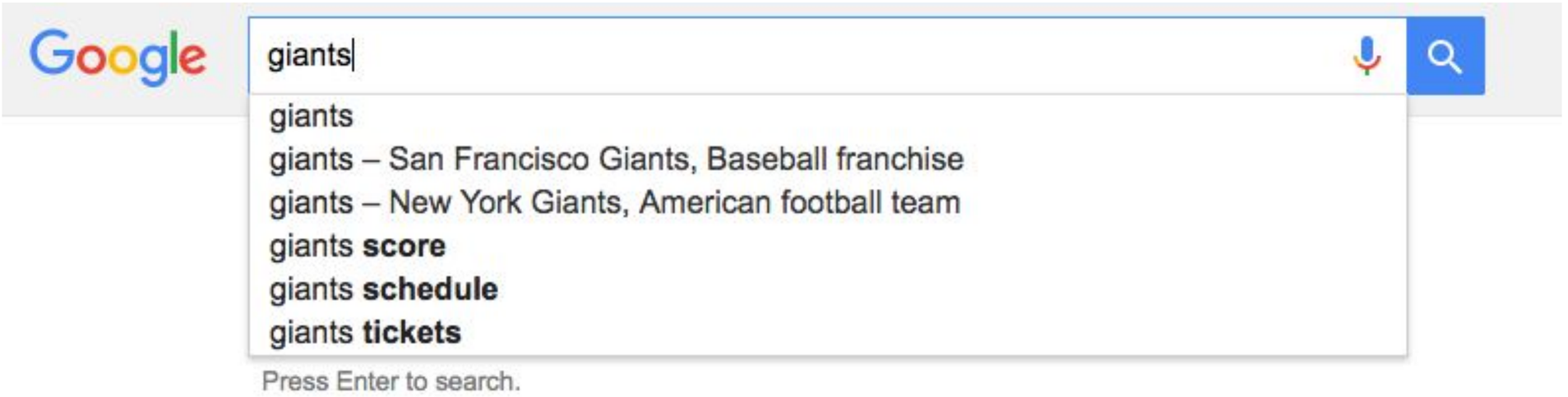
Here are some housing rentals you may be interested in:



Train machine learning models on data not rules

X

```
IF house = 'beach_house'
AND season = 'summer'
AND user_pref = 'cozy'
THEN recommend = house#22
```


Google
giants
giants
giants – San Francisco Giants, Baseball franchise
giants – New York Giants, American football team
giants score
giants schedule
giants tickets
Press Enter to search.

Google

giants

giants
giants – San Francisco Giants, Baseball franchise
giants – New York Giants, American football team
giants **score**
giants **schedule**
giants **tickets**

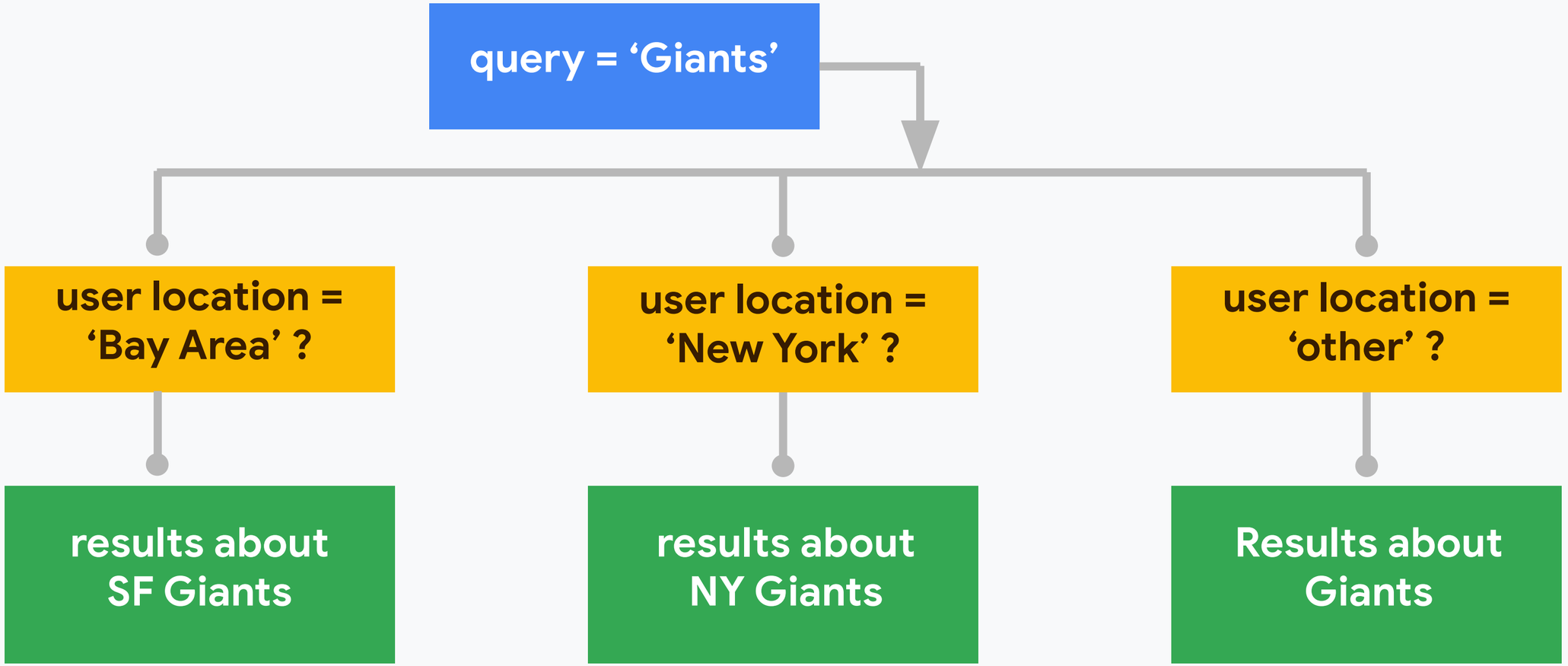
Press Enter to search.

Google

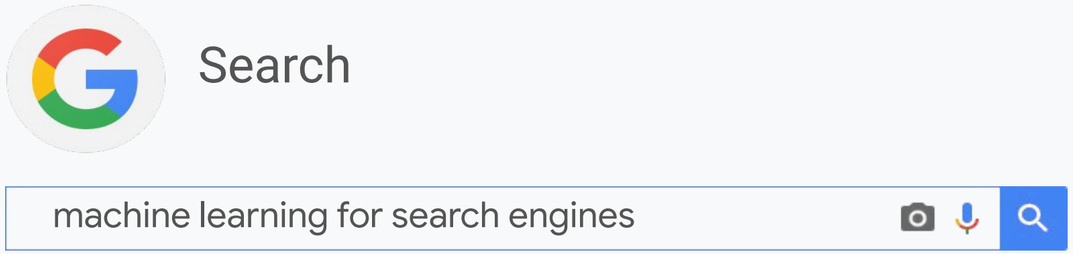
giants

giants
giants – San Francisco Giants, Baseball franchise
giants – New York Giants, American football team
giants **score**
giants **schedule**
giants **tickets**

Press Enter to search.


query = ‘Giants’
user location = ‘Bay Area’ ?
user location = ‘New York’ ?
user location = ‘other’ ?
results about SF Giants
results about NY Giants
Results about Giants

RankBrain (ML for search ranking) improved performance significantly



#3

signal
for Search ranking, out of hundreds

#1

improvement
to ranking quality
in 2+ years

Machine Learning =

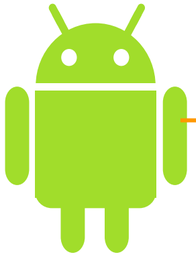
Examples, not rules

How would housing recommendations work?

Here are some rentals that you might interested in

Cluster users and items to combat rating sparsity

1 **Who is this user like?**



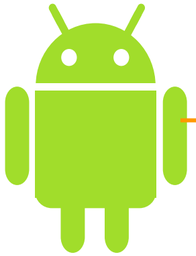
Learn from history of all houses liked by that user

2 **Is this a good house?**

Choose a new house from inventory to recommend

Cluster users and items to combat rating sparsity

1 **Who is this user like?**



2 **Is this a good house?**

3 **Predict rating**

Is this house similar to houses that people similar to this user like?

Predicted rating = user-preference * item-quality

How often do you need to **compute** the predicted ratings?

Where would you **save** them?

Agenda

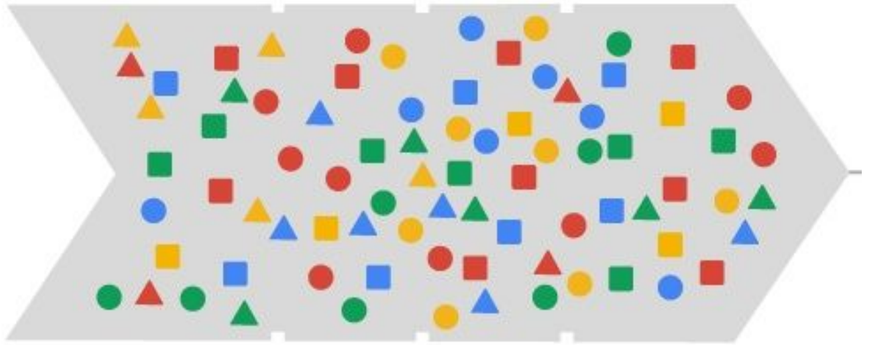
Recommendation systems

- Business applications
- Scenario: ML for housing rentals

Choosing the right solution approach

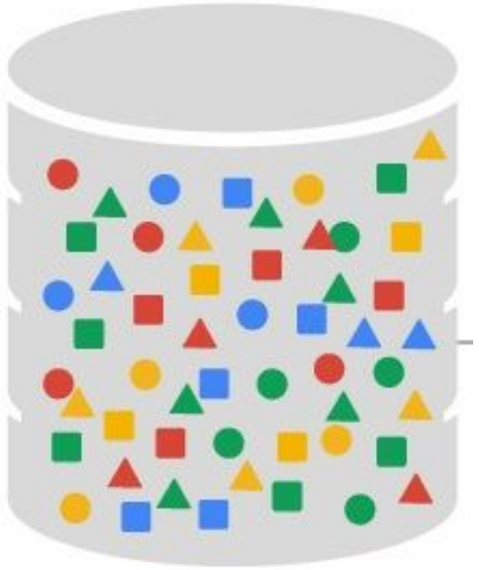
- On-premise to Google Cloud Platform
- Challenge: Utilizing and tuning on-premise clusters
- Off-cluster storage with Google Cloud Storage
- Storing recommendations

How often and where will you compute the predicted ratings?



STREAM

or



BATCH

Let’s compute ratings once per day

Compute ratings quickly using a Hadoop cluster

white and brown house during daytime

Garden Roof Building

landscape photo of 2-storey house

Building Modern Architecture

white house under maple trees

Tree Window Home

white concrete house near green tree during daytime

Home Architecture Building

red and white house surround green grass field

Red Cabin Iceland

white and grey concrete building near swimming pool under clear sky during daytime

Villa Home Pool

wooden house near pine trees and pond coated with snow during daytime

Cabin Wooden Building

aerial photography of rural

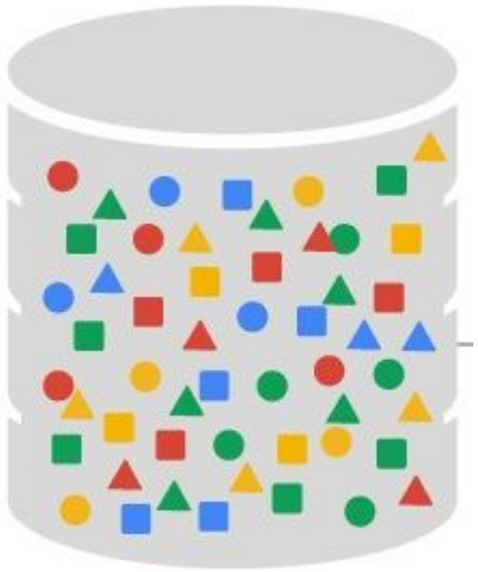
Home Roof Rooftop

brown wooden sectional bench on top of white concrete house balcony

Rooftop Seat Sofa

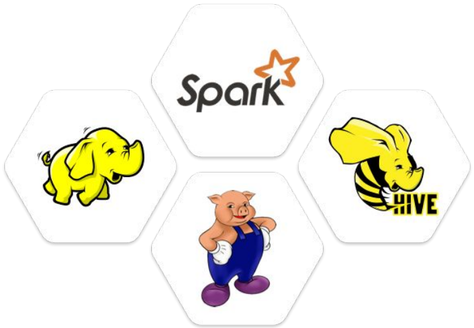
black SUV park beside white and brown house during daytime

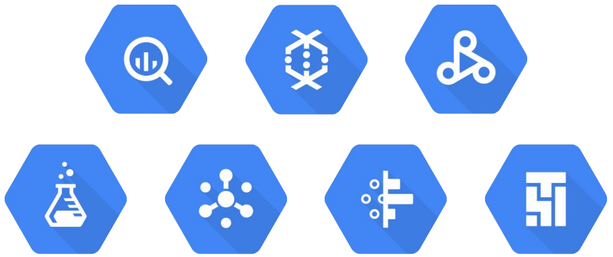
Architecture Structure Tree



BATCH

Hadoop is a popular big data software solution





Database

Bring the data to the processor

Hadoop

Distribute the processing

Store the data with the processors

Cloud services

Petabit bisectional bandwidth network

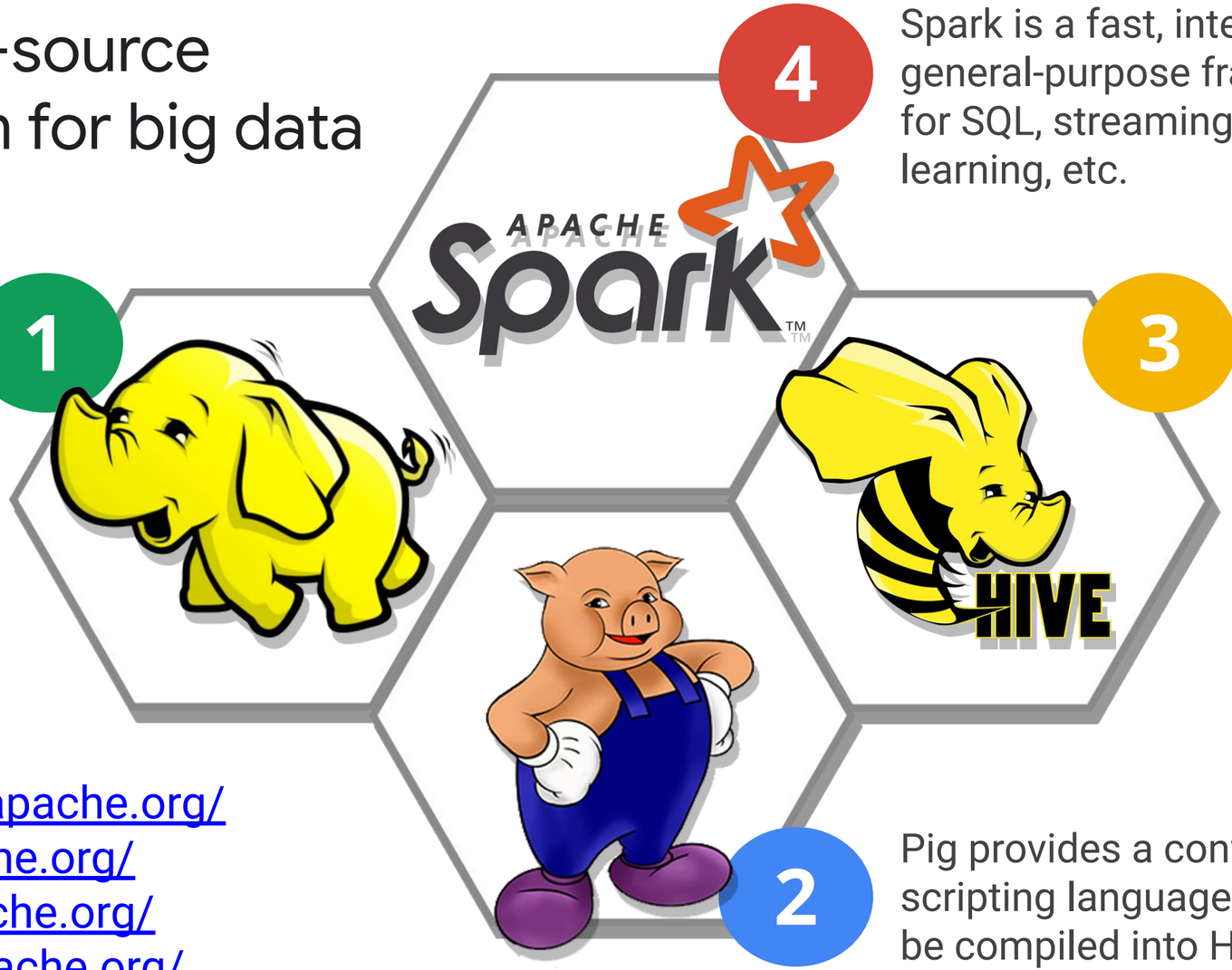
Separate, specialize, and connect

Data storage services
Data processing services
Messaging services

2006 — 2010

Rich open-source ecosystem for big data

1 Hadoop is the canonical open-source MapReduce framework.



4 Spark is a fast, interactive, general-purpose framework for SQL, streaming, machine learning, etc.

3 Hive is a data warehousing system and query language.

2 Pig provides a convenient scripting language that can be compiled into Hadoop MapReduce jobs.

http://hadoop.apache.org/
http://pig.apache.org/
http://hive.apache.org/
http://spark.apache.org/

Demo

Creating a Hadoop Cluster in 10 Minutes or Less

Agenda

Recommendation systems

- Business applications
- Scenario: ML for housing rentals

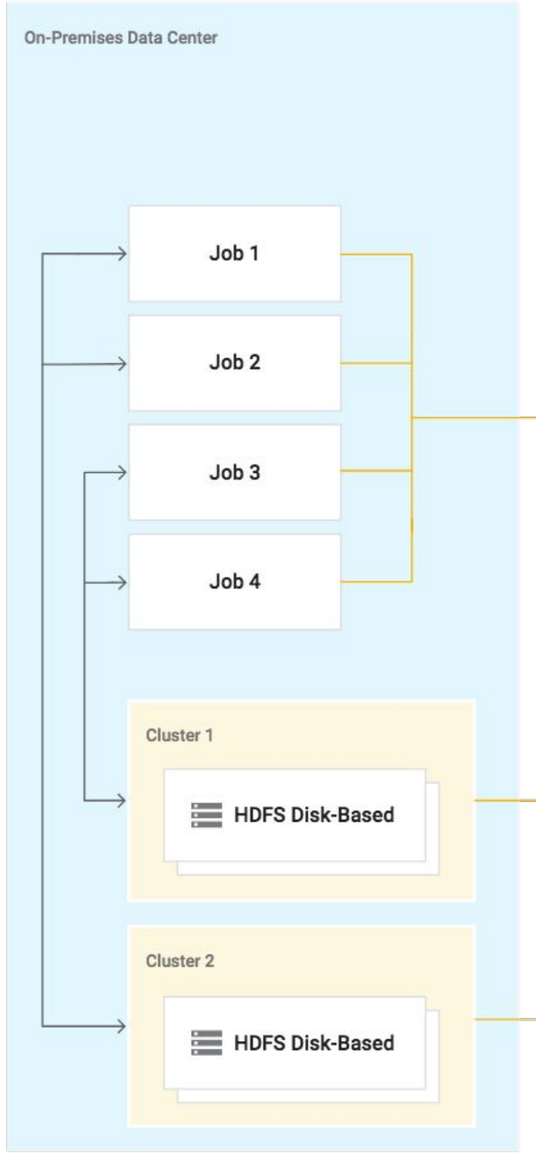
Choosing the right solution approach

- On-premise to Google Cloud Platform
- Challenge: Utilizing and tuning on-premise clusters
- Off-cluster storage with Google Cloud Storage
- Storing recommendations

On-Premise Data Center

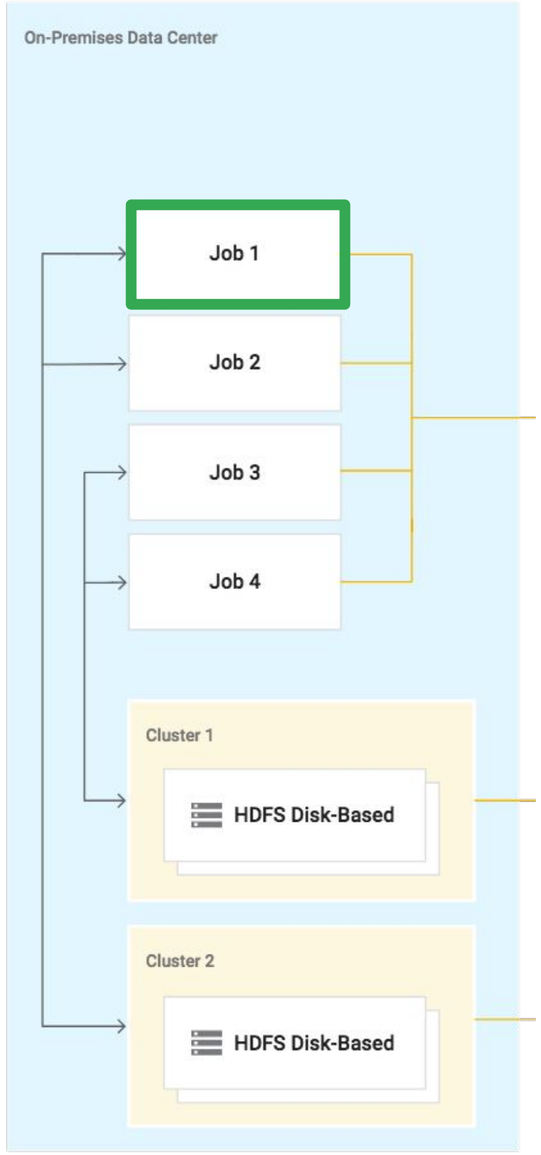
What if I already have an on-premise cluster?

You manage your on-premises cluster and have **four** jobs



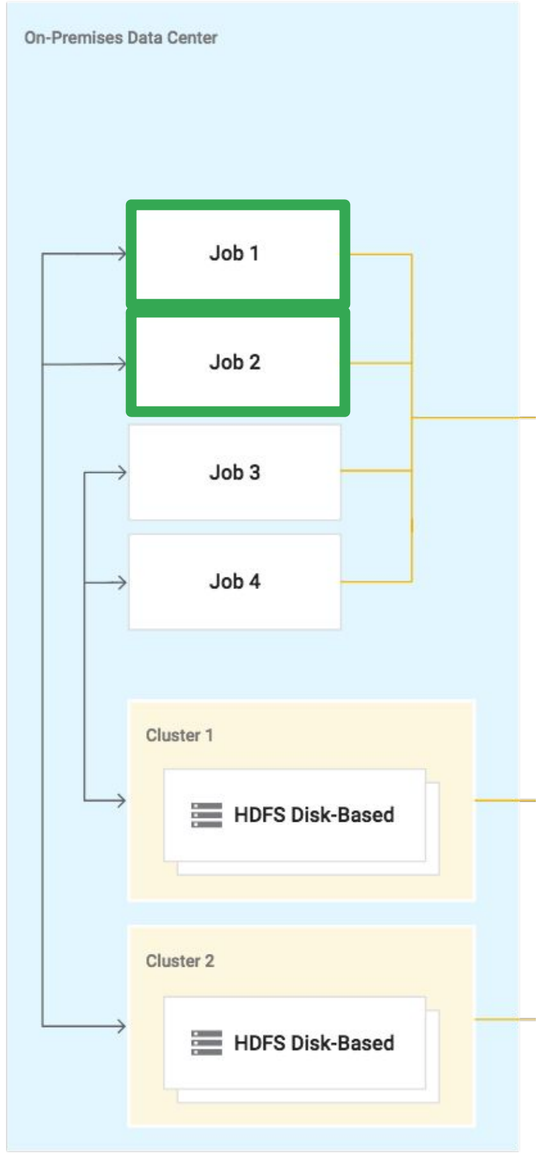
On-Premise Data Center

Scenario 1:
Job #1 starts and consumes 50% of cluster resources



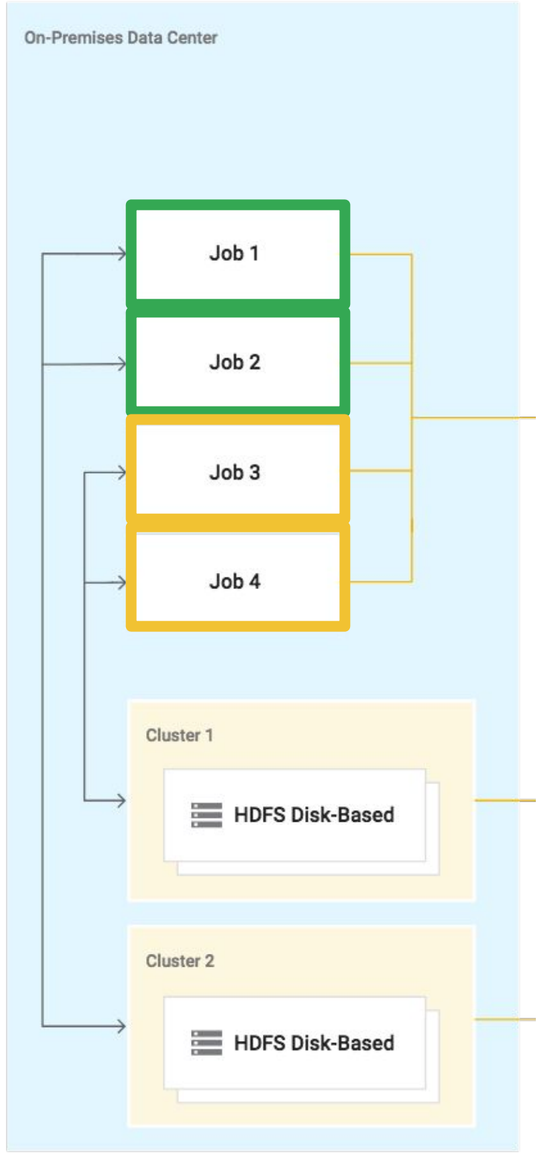
On-Premise Data Center

Scenario 1:
Job #2 starts and consumes 50% of cluster resources



On-Premise Data Center

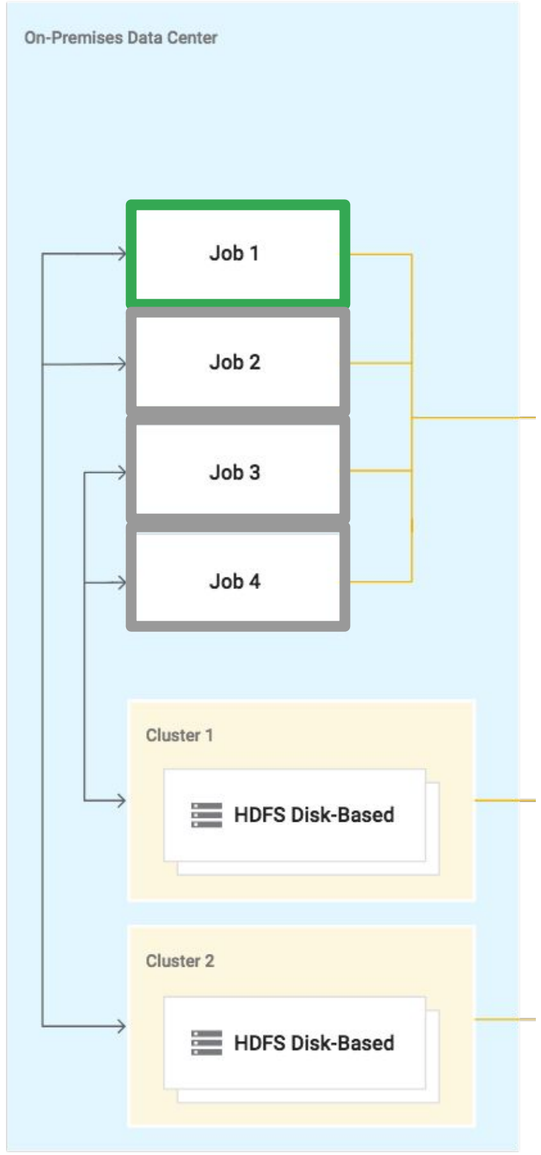
Scenario 1:
Jobs #3 and #4 attempt to start but are starved



Underprovisioned

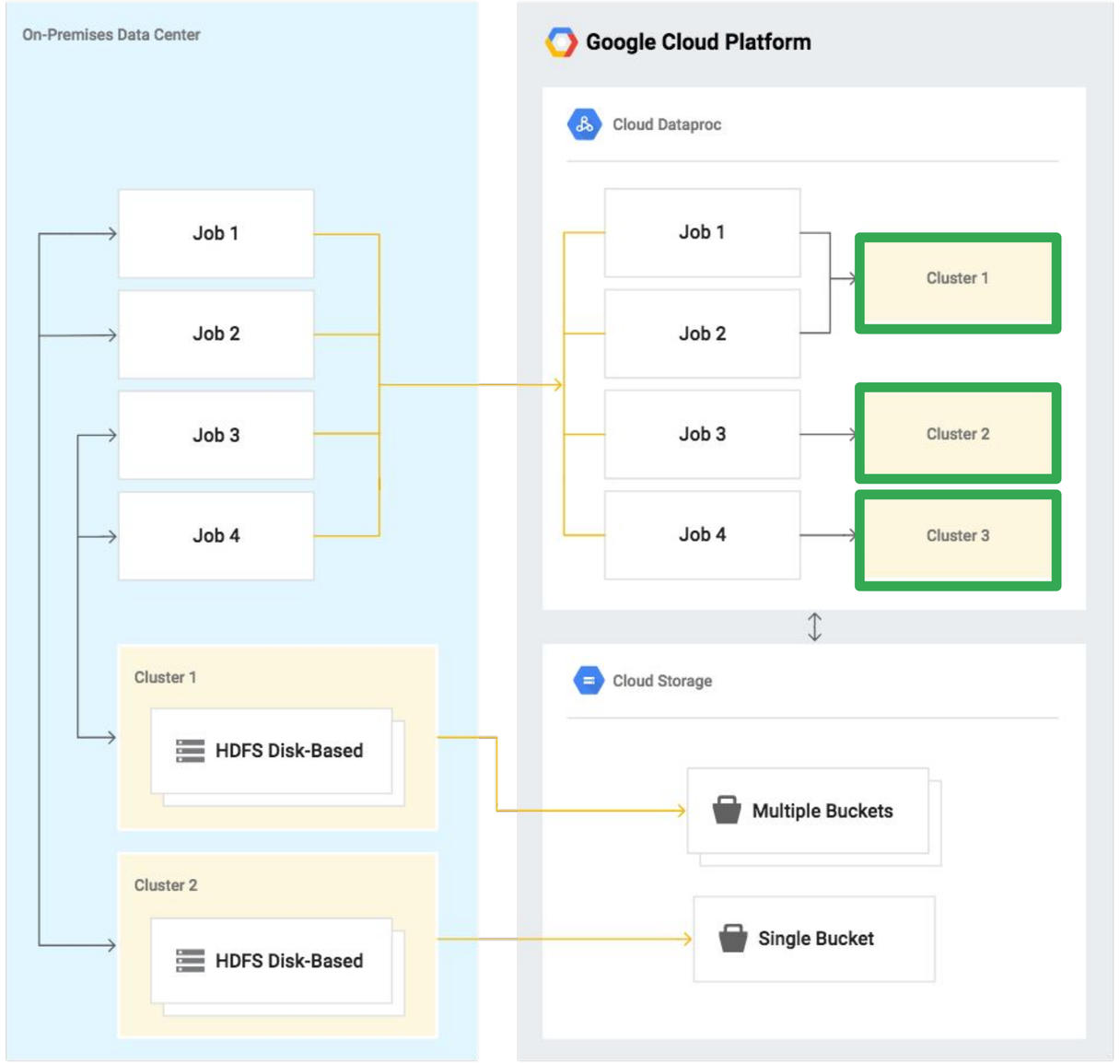
On-Premise Data Center

Scenario 2:
Job #1 is consuming 50% of cluster resources and jobs 2, 3, 4 are not needed



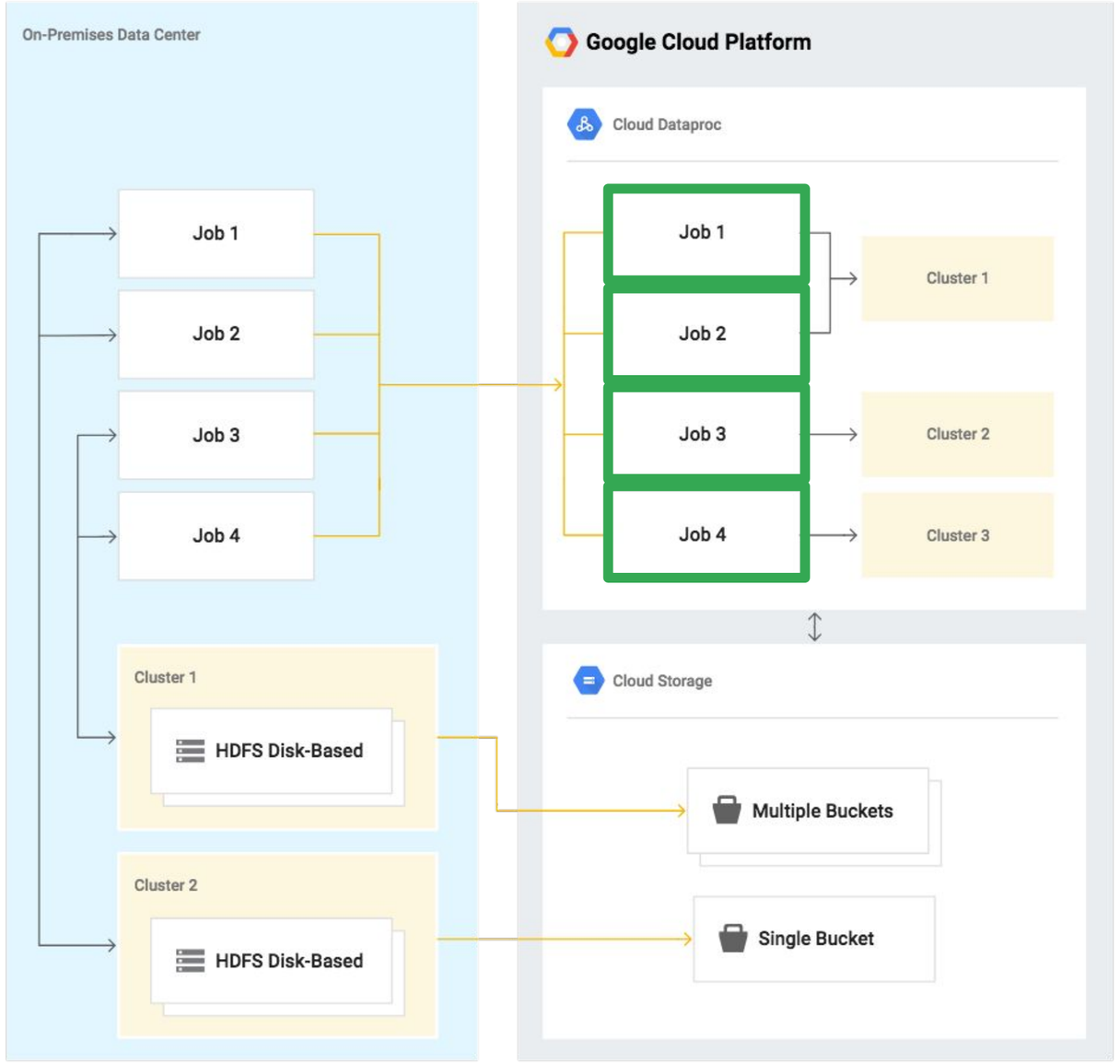
Overprovisioned

With Cloud Dataproc, Hadoop clusters are now flexible resources

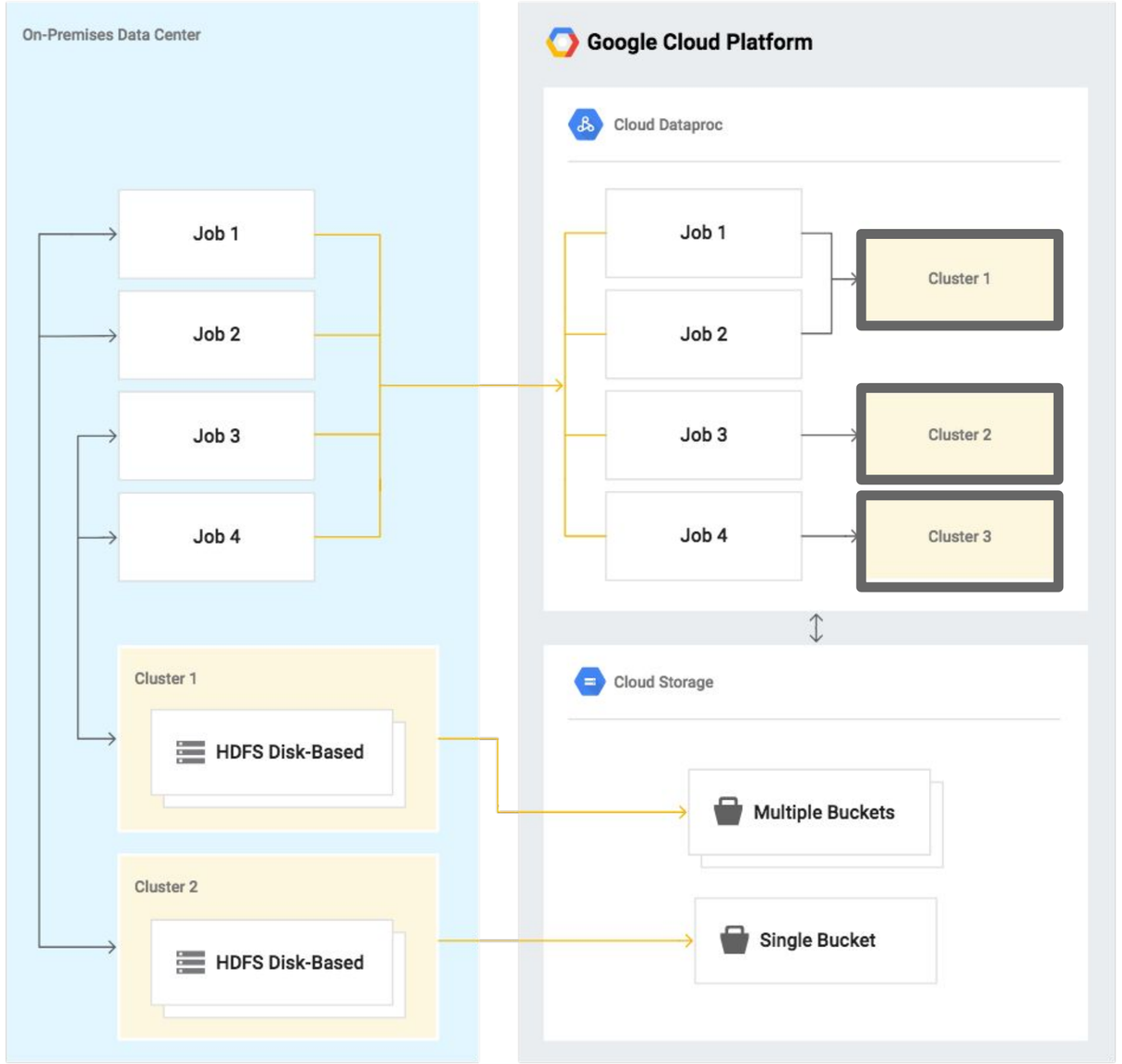


Google Cloud

Jobs can get the resources they need

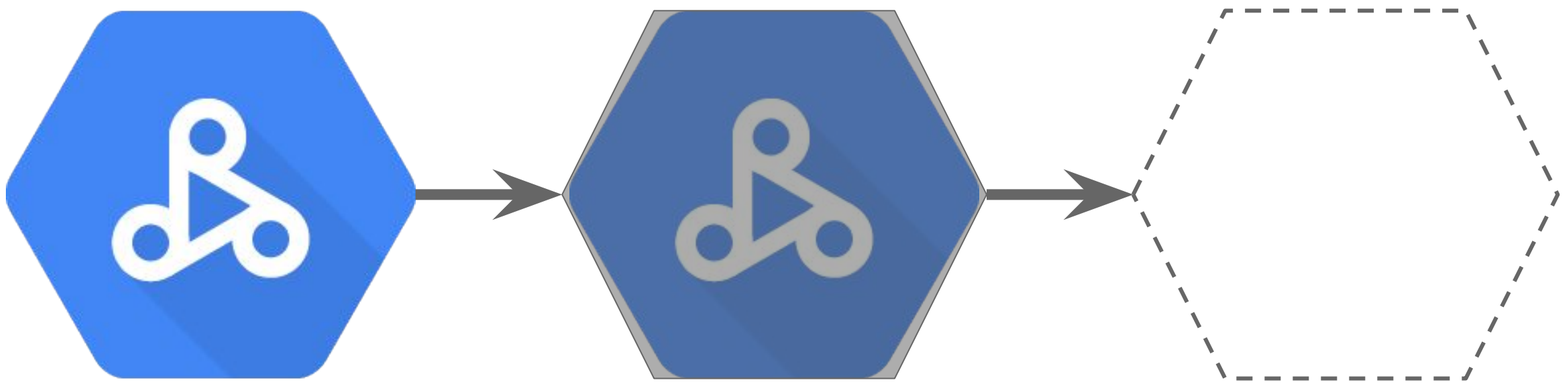


Clusters can be turned down when no jobs are running



Google Cloud

Turn down clusters automatically with Scheduled Deletion



Idle

Timestamp

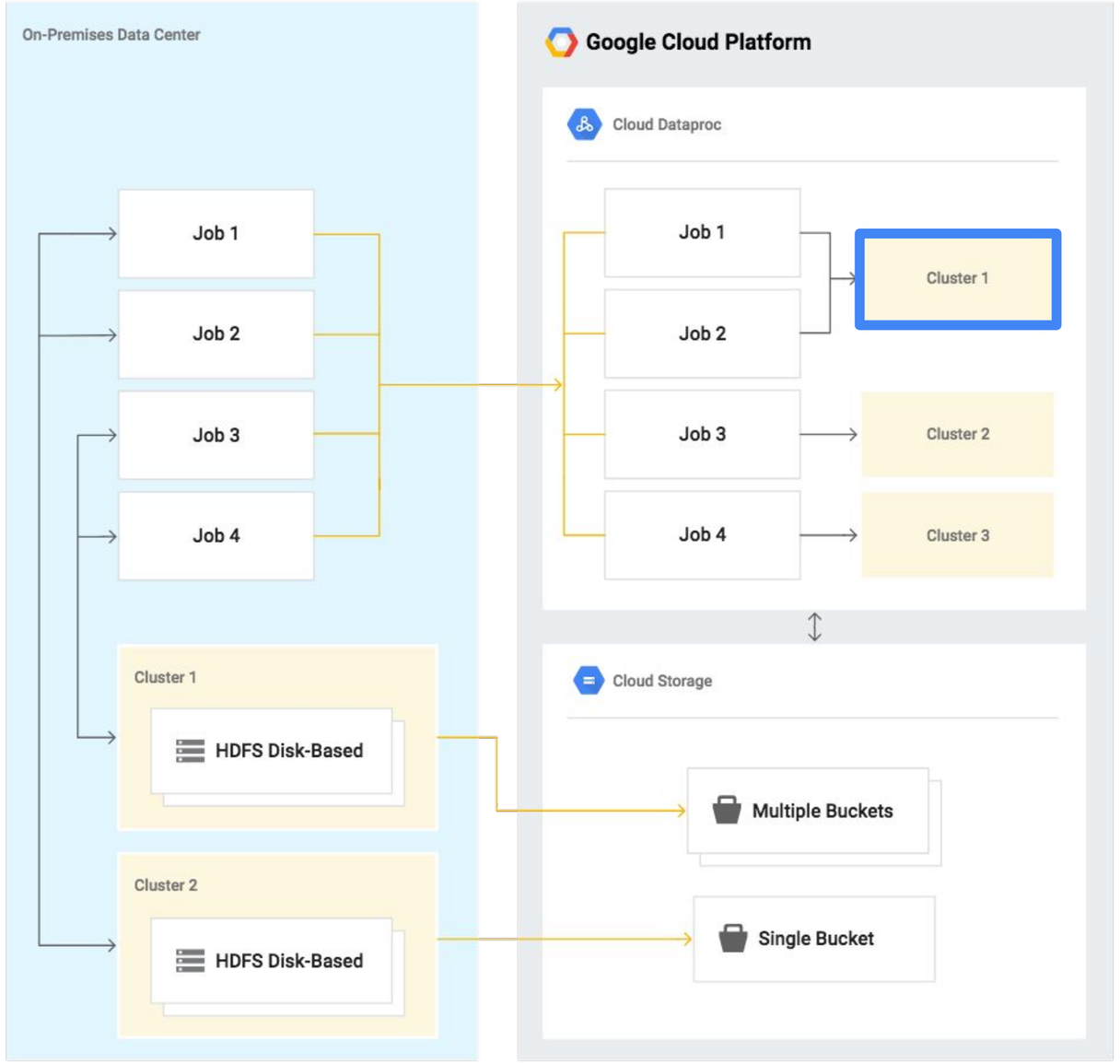
Duration

Minimum 10 minutes

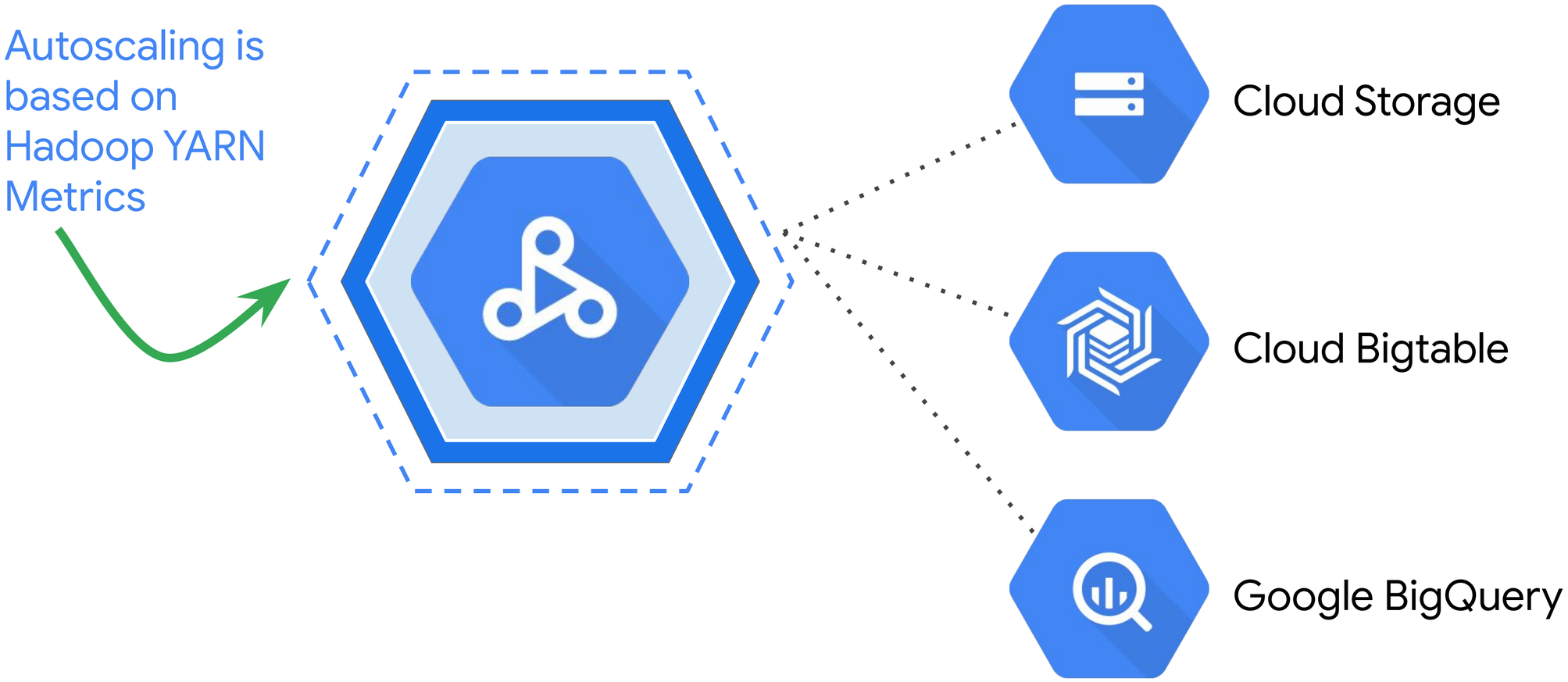
Maximum 14 days

Granularity 1 second

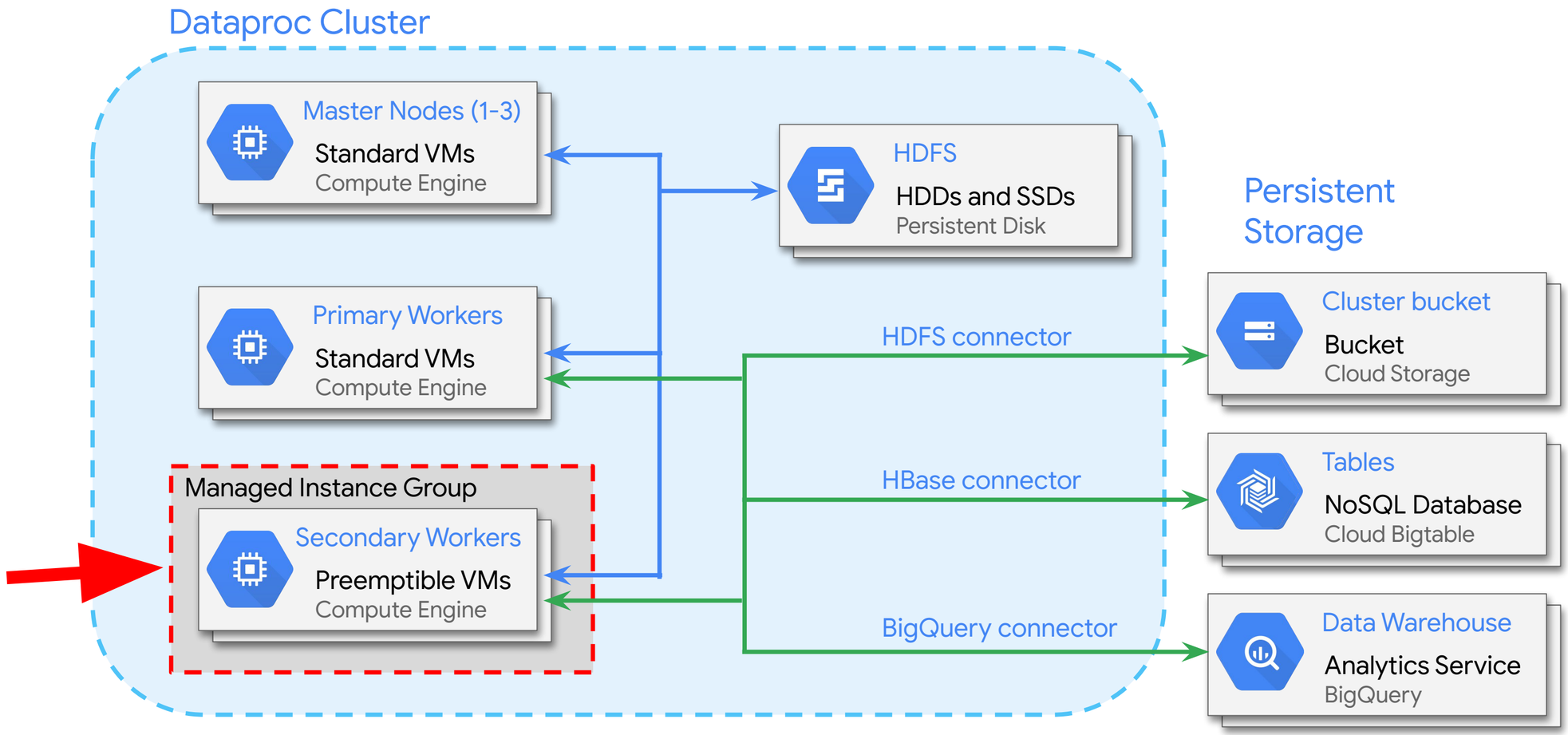
Clusters can be resized if job needs change



Cloud Dataproc Autoscaling (alpha) provides flexible capability



Utilize PVMs to significantly reduce costs for fault-tolerant workloads



Recap: Google Cloud Platform enables on-demand scalability

On-premises Hadoop Cluster

Processing Time (120s, 1 Machine)

Underprovisioned (demand > capacity)

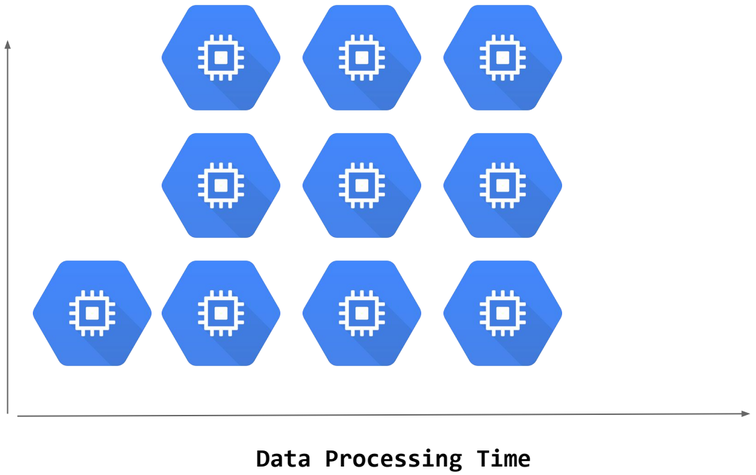
Processing Time (10s, 120 Machines)

Overprovisioned (demand < capacity)

Cloud Dataproc

Serverless and easy to resize

= 1 VM



Data Processing Time

Agenda

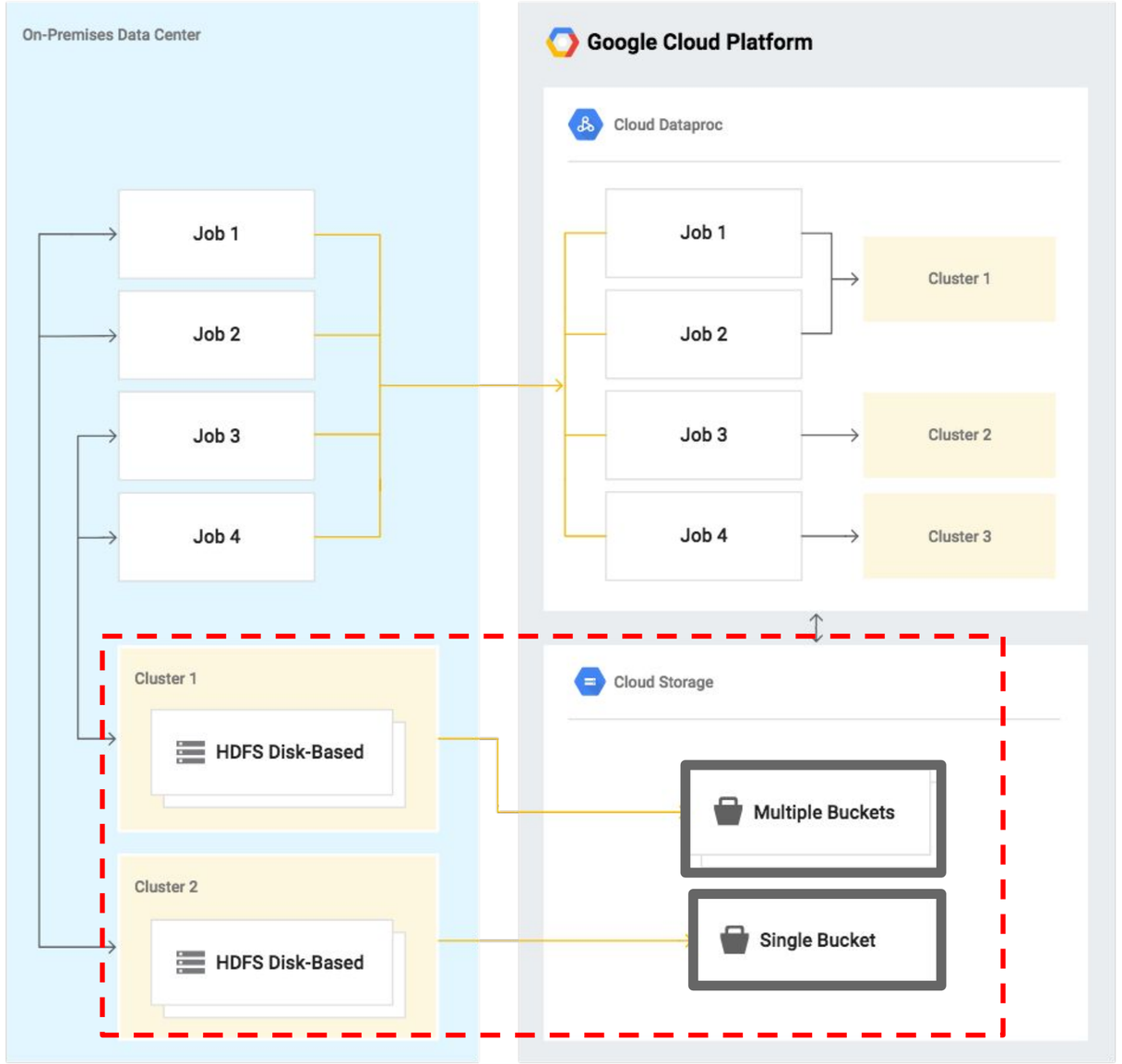
Recommendation systems

- Business applications
- Scenario: ML for housing rentals

Choosing the right solution approach

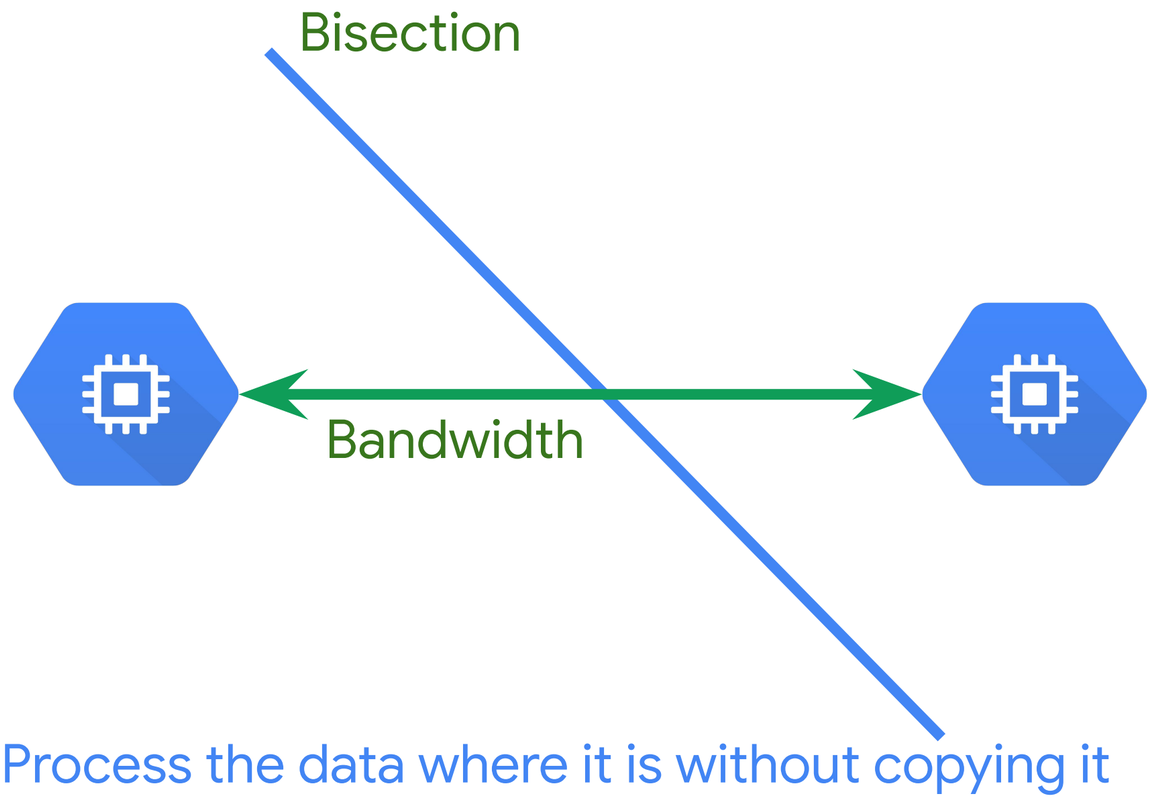
- On-premise to Google Cloud Platform
- Challenge: Utilizing and tuning on-premise clusters
- Off-cluster storage with Google Cloud Storage
- Storing recommendations

Store data persistently in Google Cloud Storage



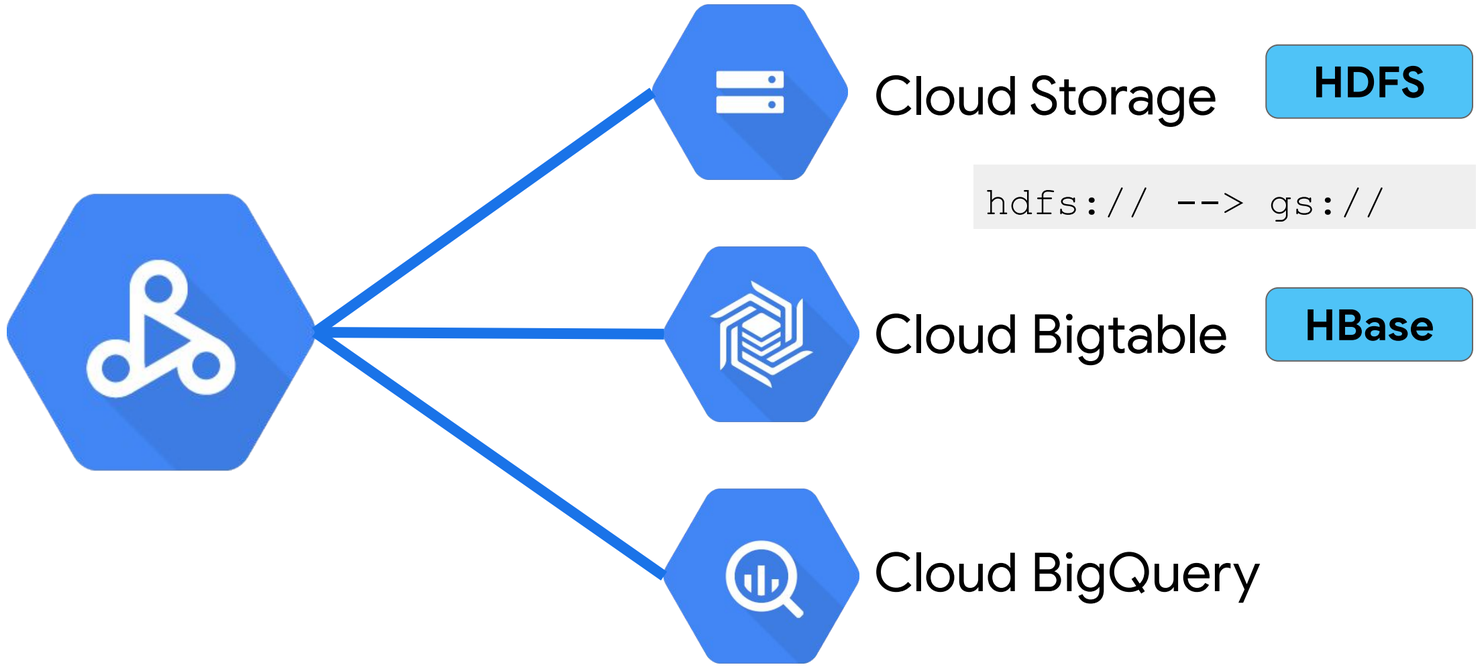
Google Cloud

Google's data center network speed enables the separation of compute and storage



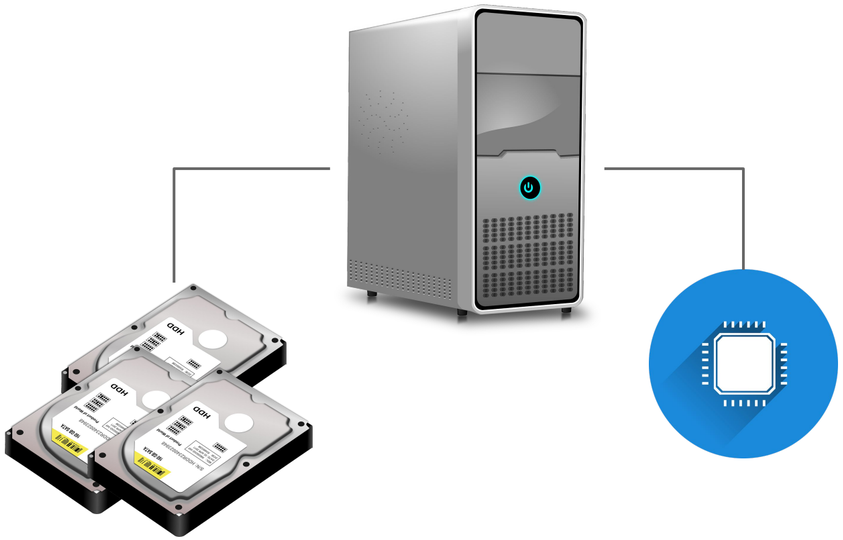
Off-cluster storage is the gateway to efficiency

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

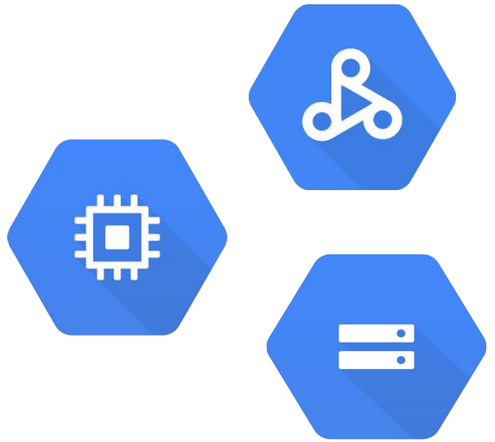


Recap: Separation of storage and computing power enables efficient resource allocation

On-premises



Google Cloud Platform



Paying for static amount of compute power even when no pipelines or queries are running

Pay for only the resources you are using

Google Cloud

Dataproc

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

Agenda

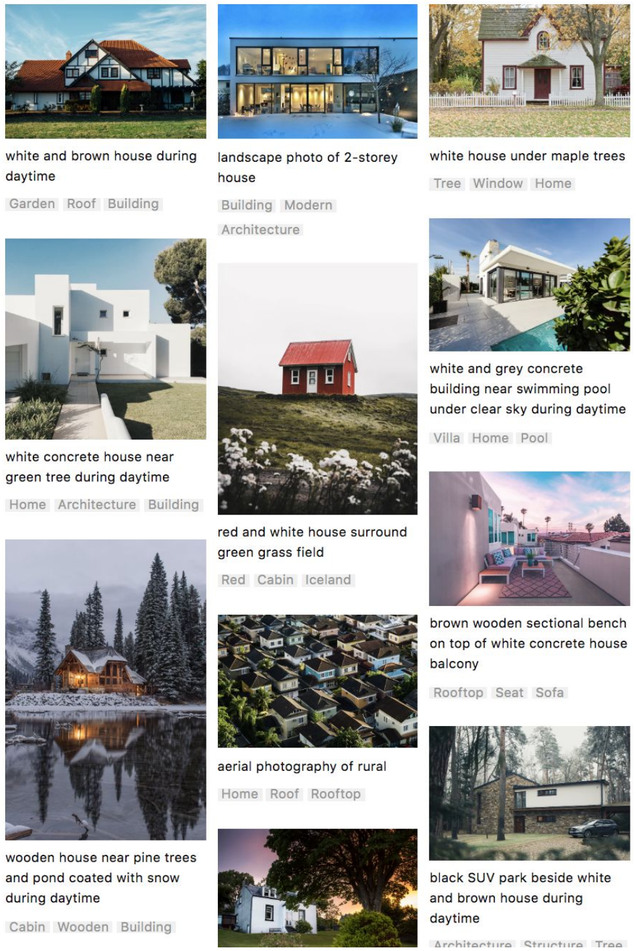
Recommendation systems

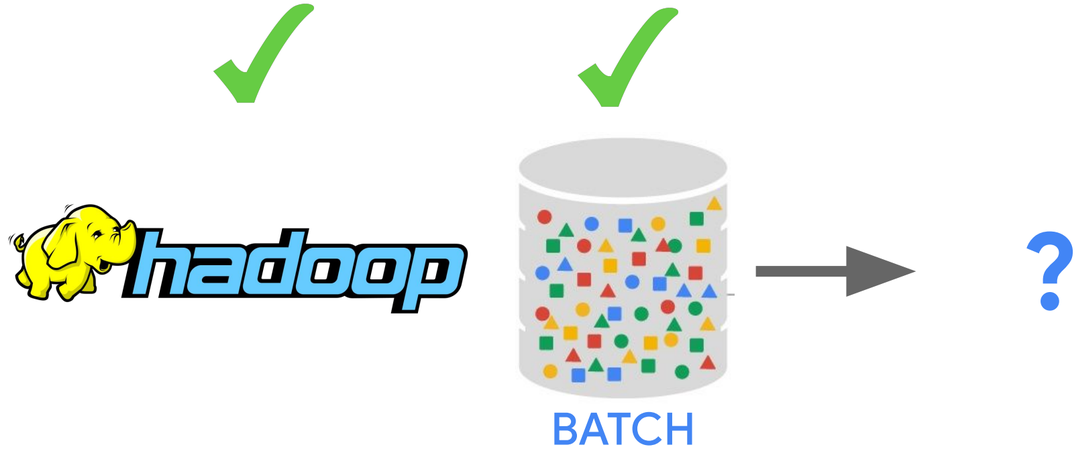
- Business applications
- Scenario: ML for housing rentals

Choosing the right solution approach

- On-premise to Google Cloud Platform
- Challenge: Utilizing and tuning on-premise clusters
- Off-cluster storage with Google Cloud Storage
- Storing recommendations

Once the recommendations are calculated, where will we **store** them?





Store the ratings, users, and inventory in a RDBMS

white and brown house during daytime

Garden Roof Building

landscape photo of 2-storey house

Building Modern Architecture

white house under maple trees

Tree Window Home

white concrete house near green tree during daytime

Home Architecture Building

red and white house surround green grass field

Red Cabin Iceland

white and grey concrete building near swimming pool under clear sky during daytime

Villa Home Pool

brown wooden sectional bench on top of white concrete house balcony

Rooftop Seat Sofa

wooden house near pine trees and pond coated with snow during daytime

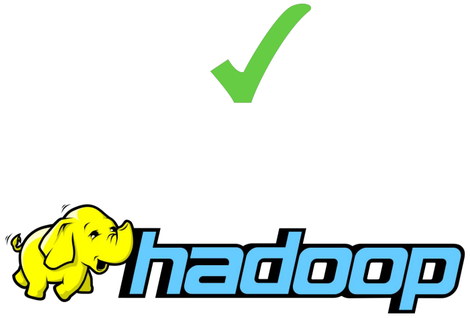
Cabin Wooden Building

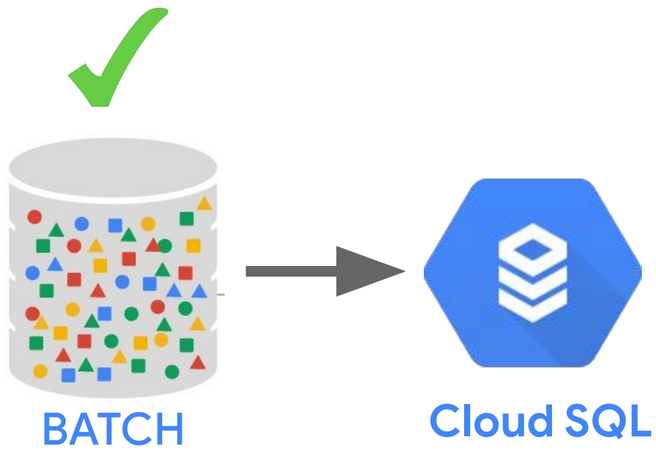
aerial photography of rural

Home Roof Rooftop

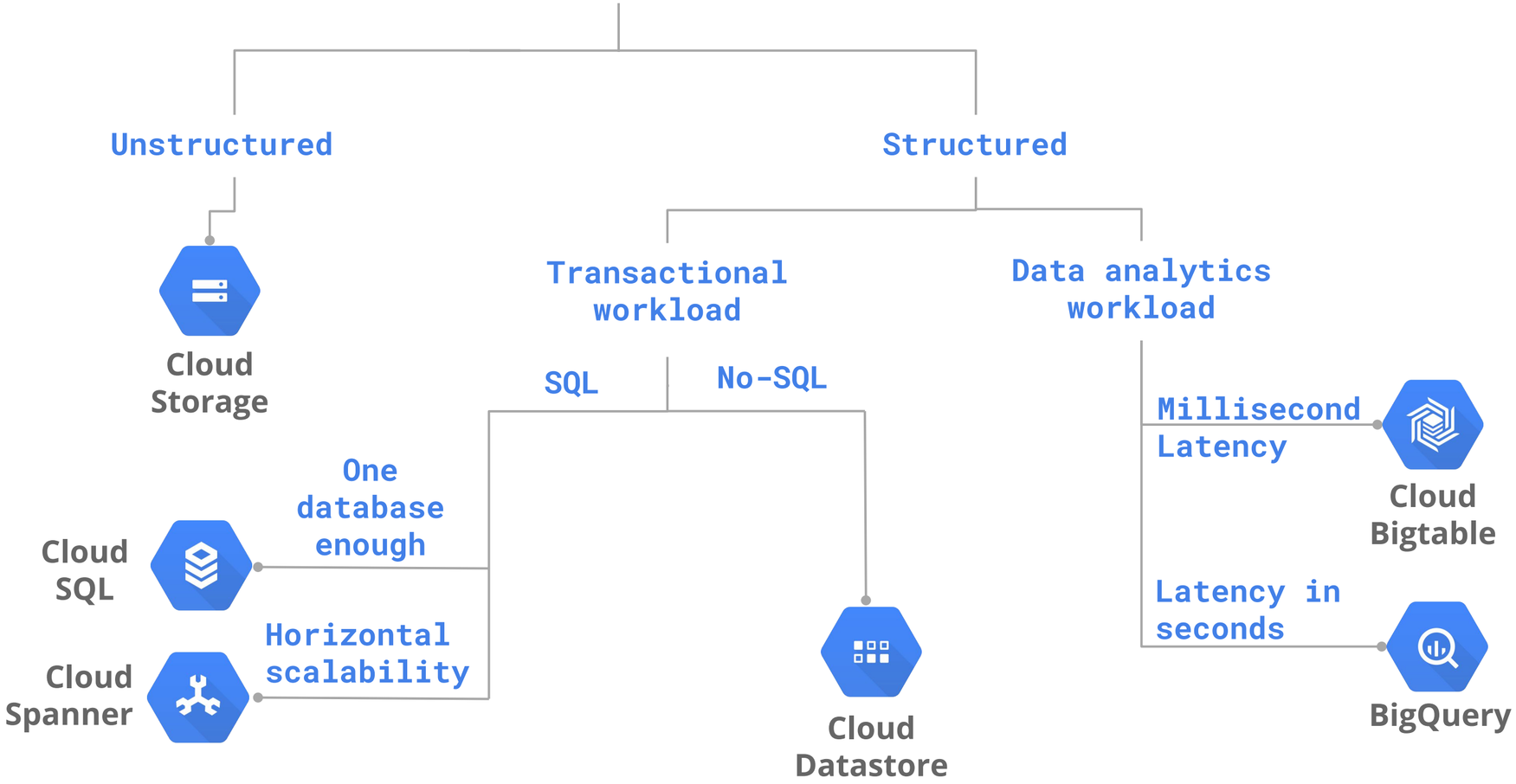
black SUV park beside white and brown house during daytime

Architecture Structure Tree





If your data is....



Cloud SQL is fully managed RDBMS

Cloud SQL
Google-managed MySQL

- Familiar
- Flexible pricing
- Managed backups
- Connect from anywhere
- Automatic replication
- Fast connection from GCE & GAE
- Google security

Lab: Product Recommendations using Cloud SQL and Spark

What housing rentals should I recommend to my customers based on their history?

Cloud SQL

Cloud Dataproc

Lab

Product Recommendations using Cloud SQL and Spark

- Create Cloud SQL instance and populate tables
- Explore the rentals data using SQL statements from Cloud Shell
- Launch Dataproc
- Train and apply machine learning model written in PySpark to create product recommendations
- Explore inserted rows in Cloud SQL