Data Analysis on the Cloud

Big Data and Machine Learning Fundamentals

Google Cloud Fundamentals: Big Data & Machine Learning

Google Cloud

Version #1.1

Agenda

1 Introduction

2 Fundamentals of GCP

3 Data Analysis on the Cloud

4 Scaling Data Analysis

5 Machine Learning

6 Data Processing Architecture

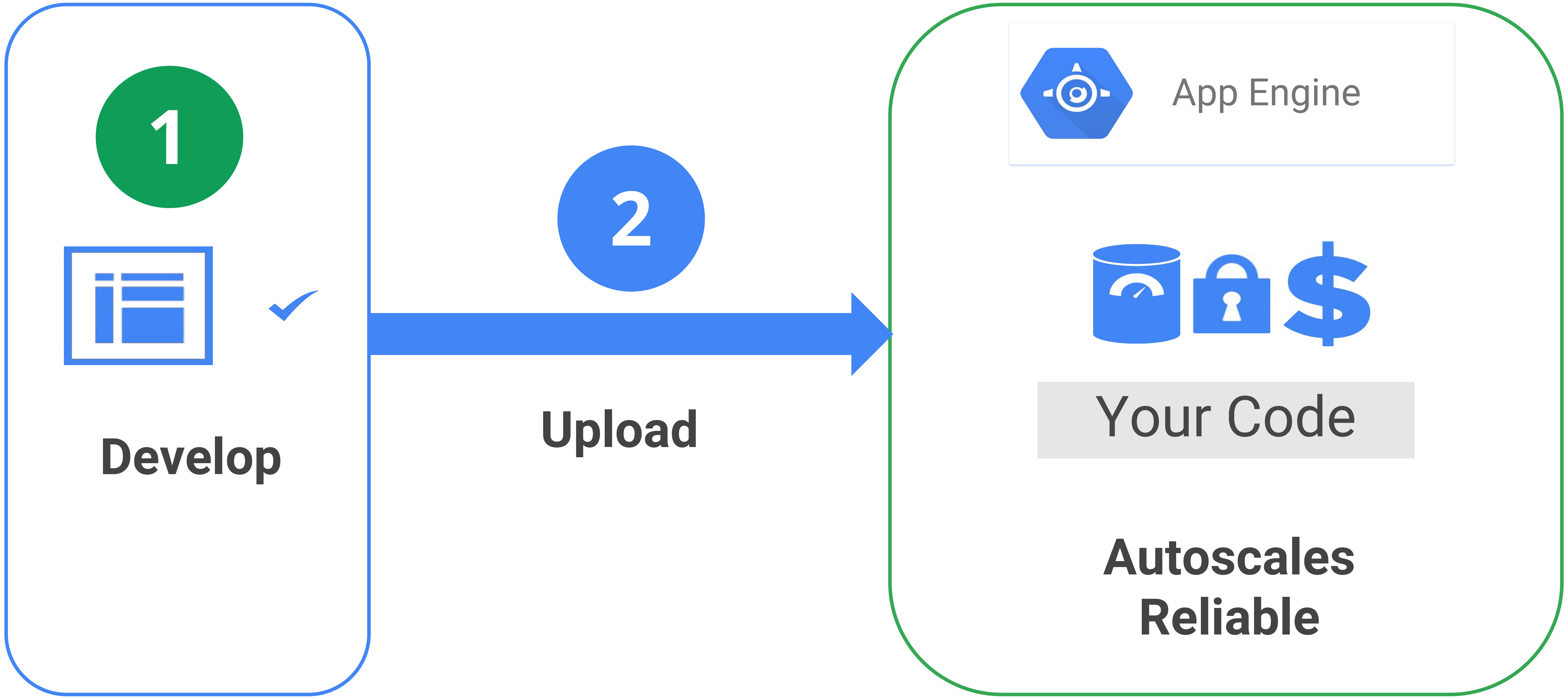
Agenda

Stepping stones to transformation

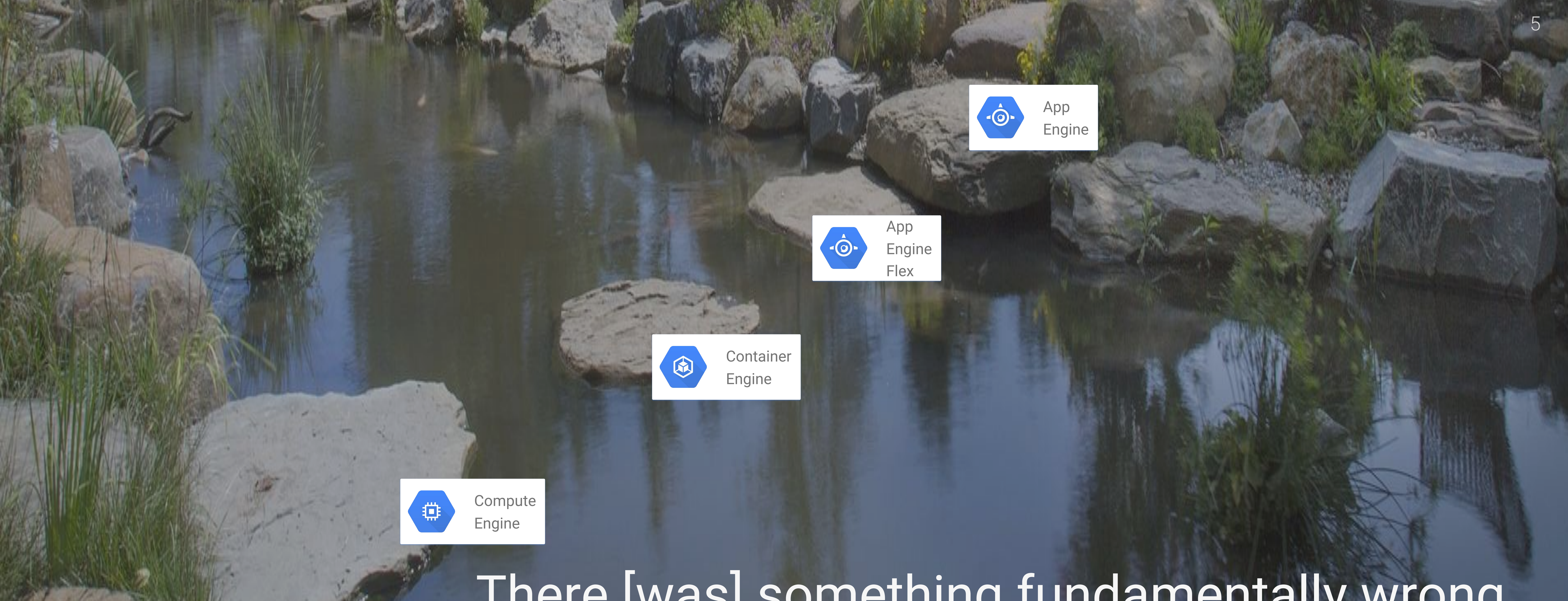
Your SQL database in the cloud + Lab

Managed Hadoop in the cloud + Lab

Google Cloud Platform began in 2008, with App Engine, a serverless way to run web applications



http://googleappengine.blogspot.com/2008/04/introducing-google-app-engine-our-new.html
http://googleappengine.blogspot.com/2013/05/the-google-app-engine-blog-is-moving.html



> There [was] something fundamentally wrong with what we were doing in 2008 … We didn't get the right stepping stones into the cloud …
>
> -- Eric Schmidt, Executive Chairman, Google

GCP now consists of a suite of products that together provide these stepping stones in a business' transformative journey

Change where you compute

Cost effective virtual machines, storage, Hadoop, and MySQL to migrate your current workloads to the public cloud.

Flexibility, scalability and reliability

Reliable, autoscaling messaging, data processing, and storage.

Change how you compute

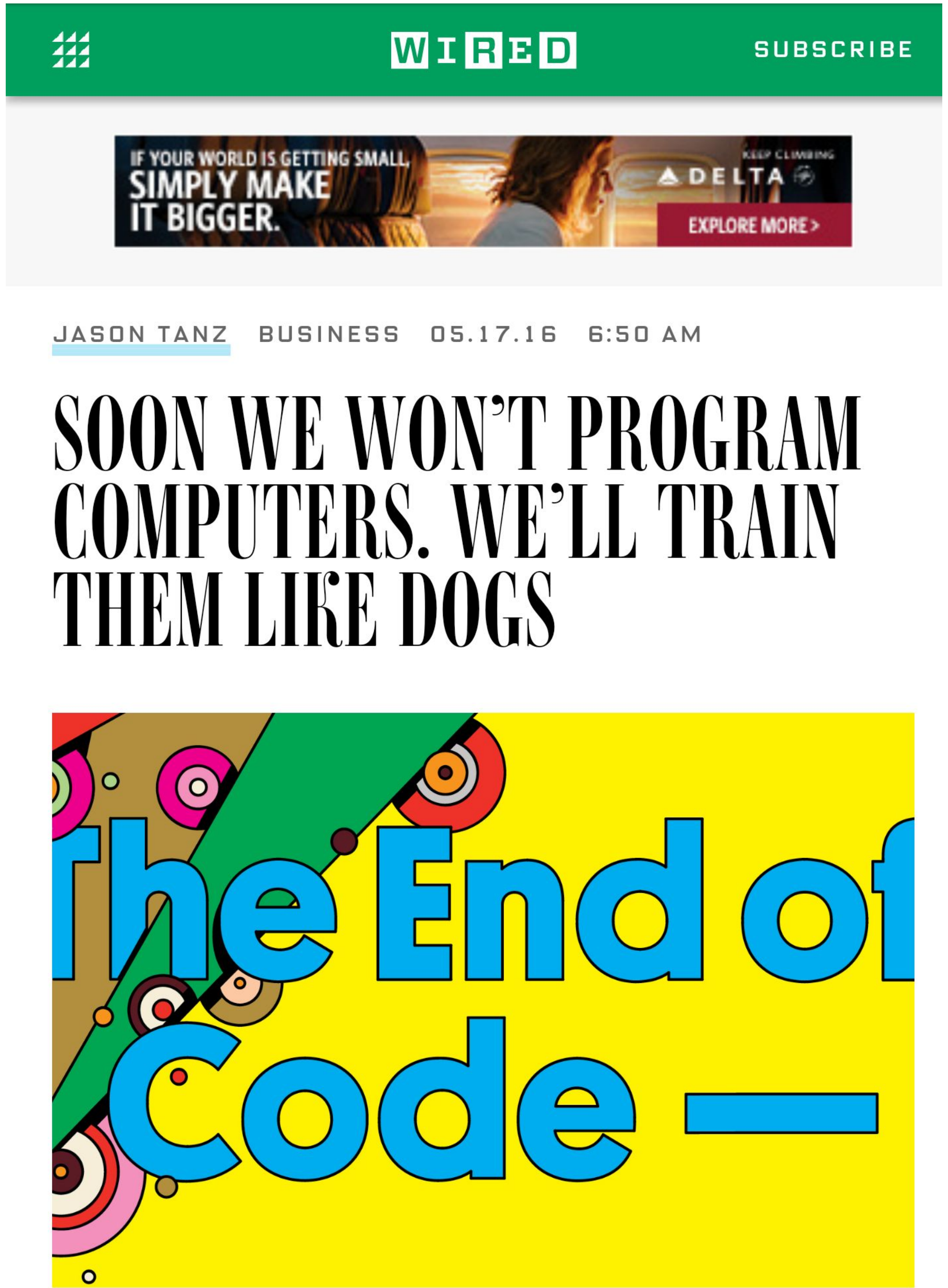
Fully managed products for data warehousing, data analysis, streaming, and machine learning.

Machine learning. This is the next transformation … the programming paradigm is changing. Instead of programming a computer, you teach a computer to learn something and it does what you want.

Eric Schmidt,
Executive Chairman,
Google

WIRED’s headline

“If you want to teach a neural network to recognize a cat, for instance, you don’t tell it to look for whiskers, ears, fur, and eyes. You simply show it thousands and thousands of photos of cats, and eventually it works things out.”



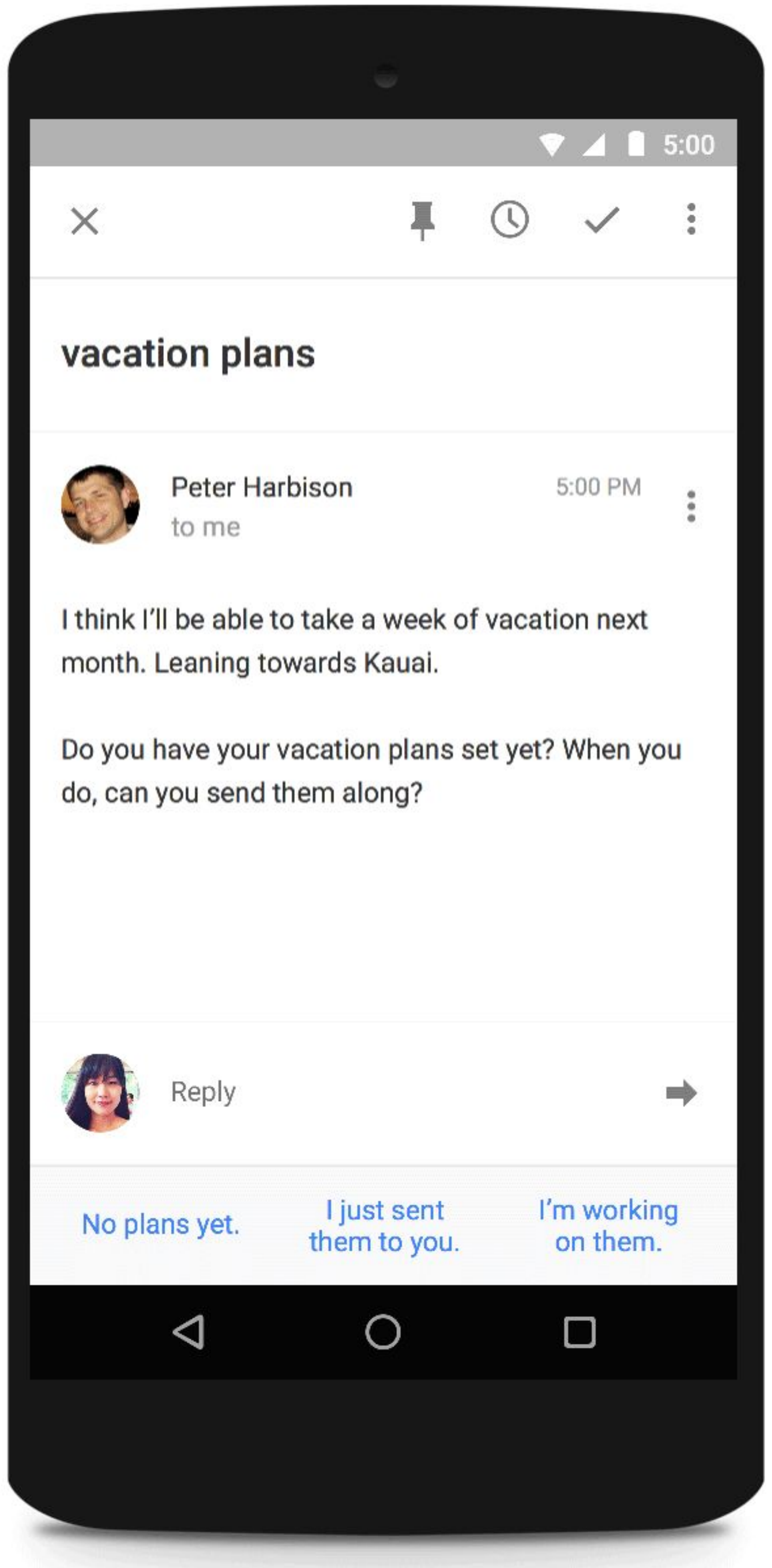
WIRED SUBSCRIBE

JASON TANZ BUSINESS 05.17.16 6:50 AM

SOON WE WON’T PROGRAM COMPUTERS. WE’LL TRAIN THEM LIKE DOGS

Google Cloud

Machine Learning is not new, but it is now mainstream



5:00

vacation plans

Peter Harbison
to me

5:00 PM

I think I'll be able to take a week of vacation next month. Leaning towards Kauai.

Do you have your vacation plans set yet? When you do, can you send them along?

Reply

No plans yet. | I just sent them to you. | I'm working on them.

Search

People who bought ...

Spam filtering

Suggest next video

Route planning

Smart Reply

? **What's common to all of these use cases of Machine Learning?**

There are three components in a recommendation system

Rating

Users rate a few houses explicitly or implicitly

Training

A machine learning model is created to predict a user's rating of a house

Recommending

For each user, the model is applied to every unrated house and the top 5 houses for that user are saved.

? How would you build a model to predict the rating of a house for a user?

The ML algorithm essentially clusters users and items

1 **Who is like this user?**

2 **Is this a good house?**

LARGE PEACEFUL PLACE

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?

In addition to the ML algorithm, you also need sophisticated data management

Data Collection	Scalable front end to collect customer actions
Data Analysis	Data that is accessible and not silo-ed
Machine Learning	(Re-)training and experimentation
Serving	Scalable, real-time system to serve recommendations

Agenda

Stepping stones to transformation

Your SQL database in the cloud + Lab

Managed Hadoop in the cloud + Lab

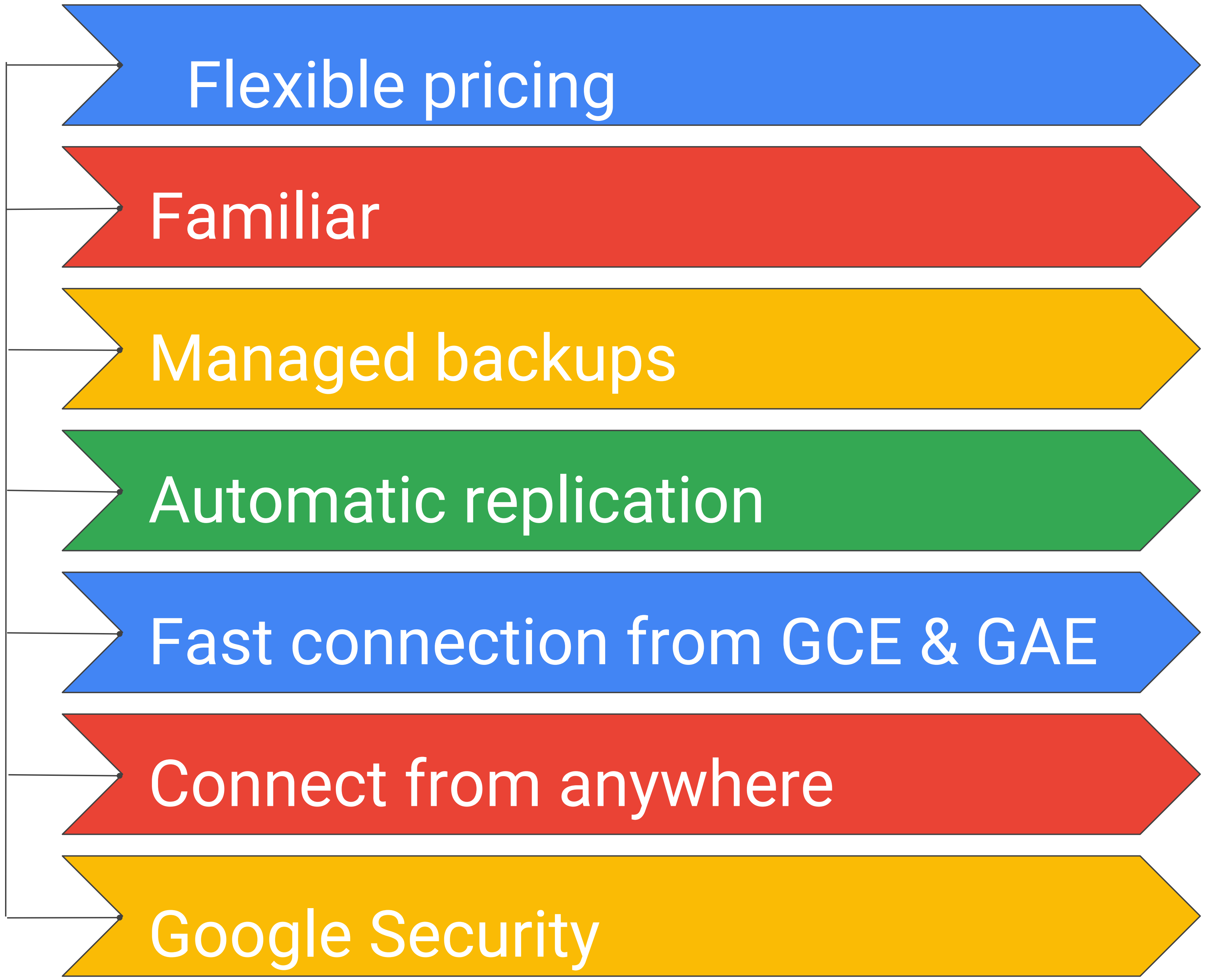
Choose your storage solution based on your access pattern

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

Cloud SQL is a fully managed database service

Cloud SQL

Google-managed MySQL or Postgres



Flexible pricing

Familiar

Managed backups

Automatic replication

Fast connection from GCE & GAE

Connect from anywhere

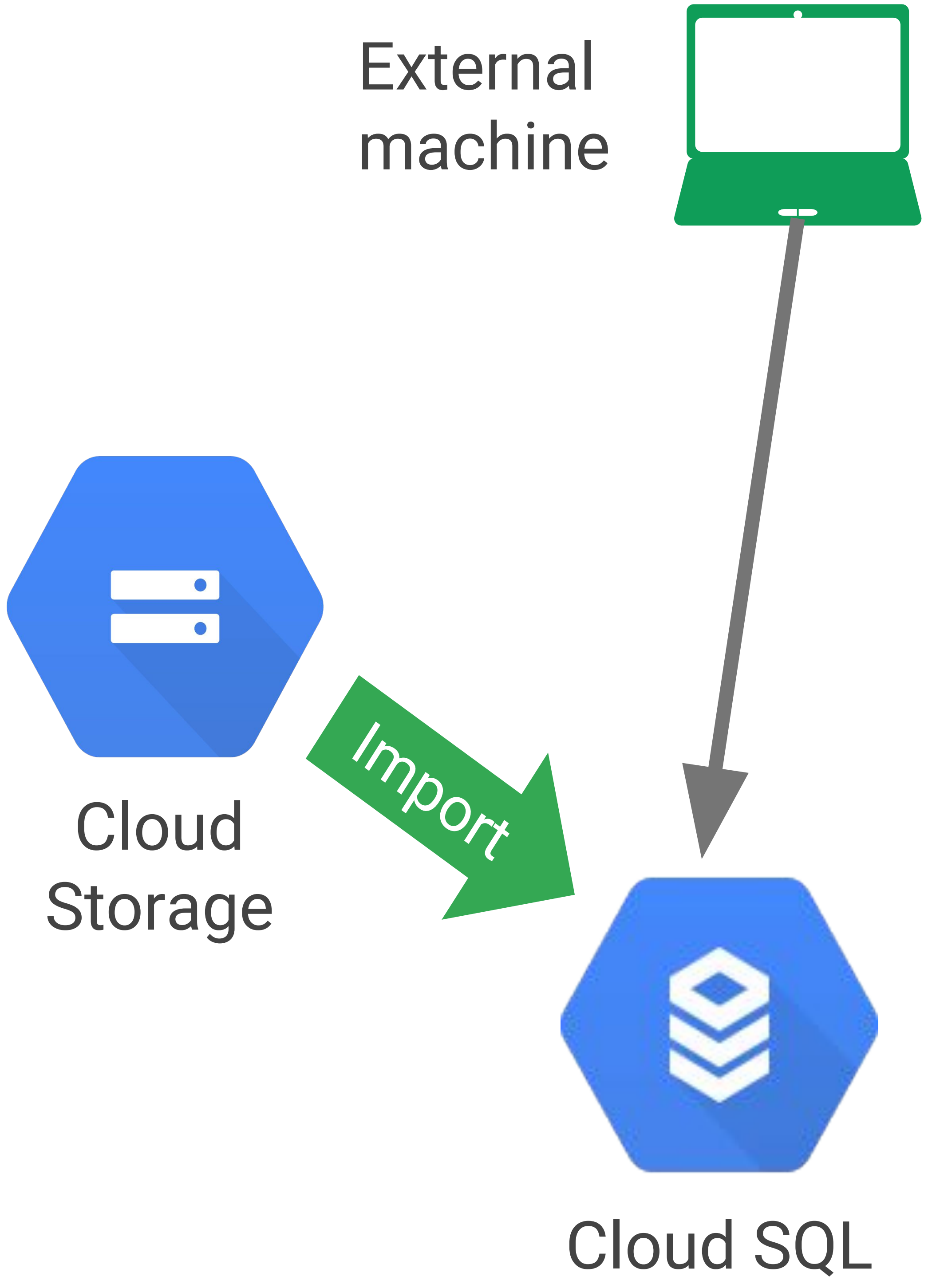
Google Security

Lab: Set up rentals data in Cloud SQL

Lab 3: Setup rentals data in Cloud SQL

In this lab, you populate rentals data in Cloud SQL for the recommendation engine to use:

1. Create Cloud SQL instance
2. Create database tables by importing .sql files from Cloud Storage
3. Populate the tables by importing .csv files from Cloud Storage
4. Allow access to Cloud SQL
5. Explore the rentals data using SQL statements from Cloud Shell



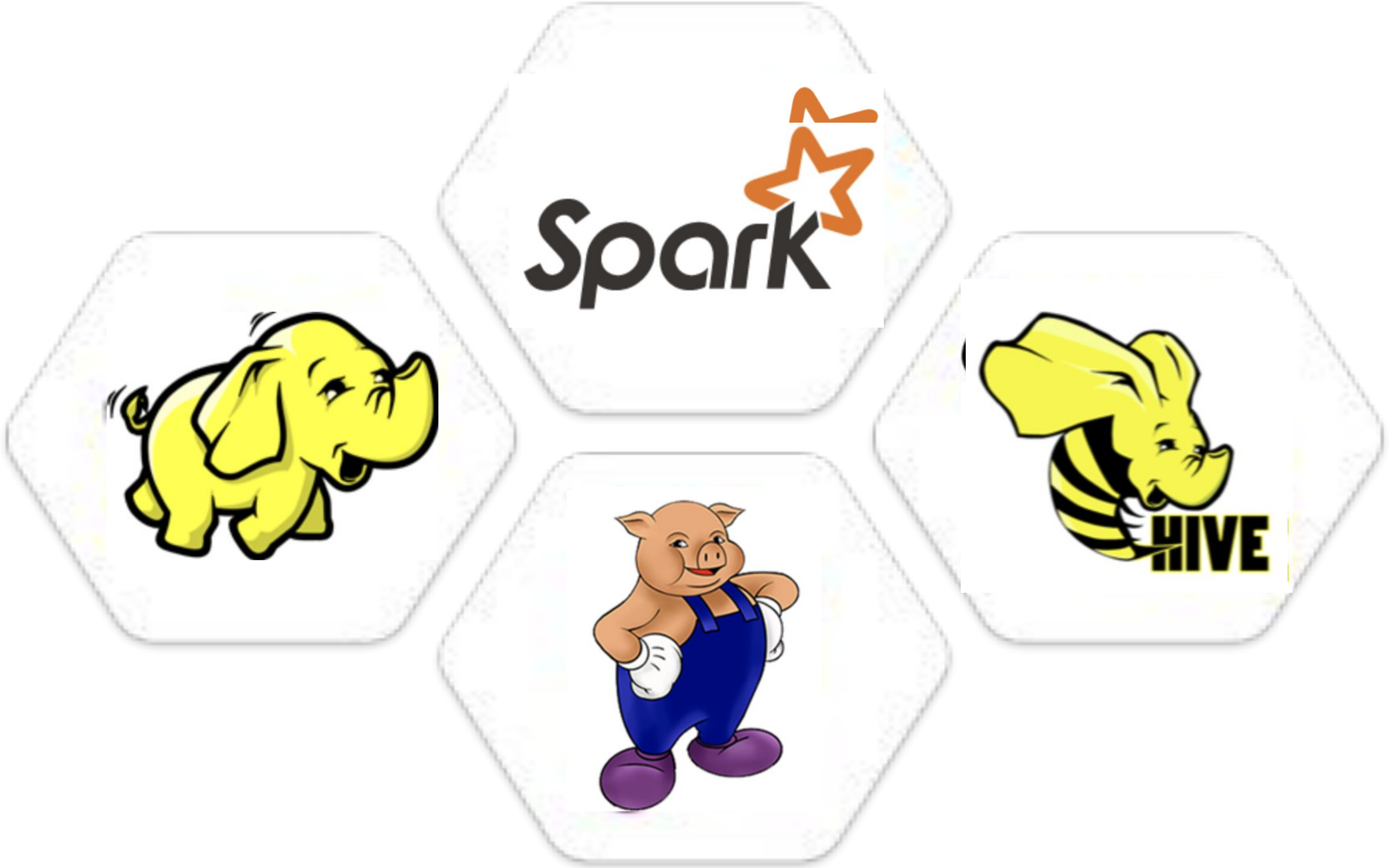
Agenda

Stepping stones to transformation

Your SQL database in the cloud + Lab

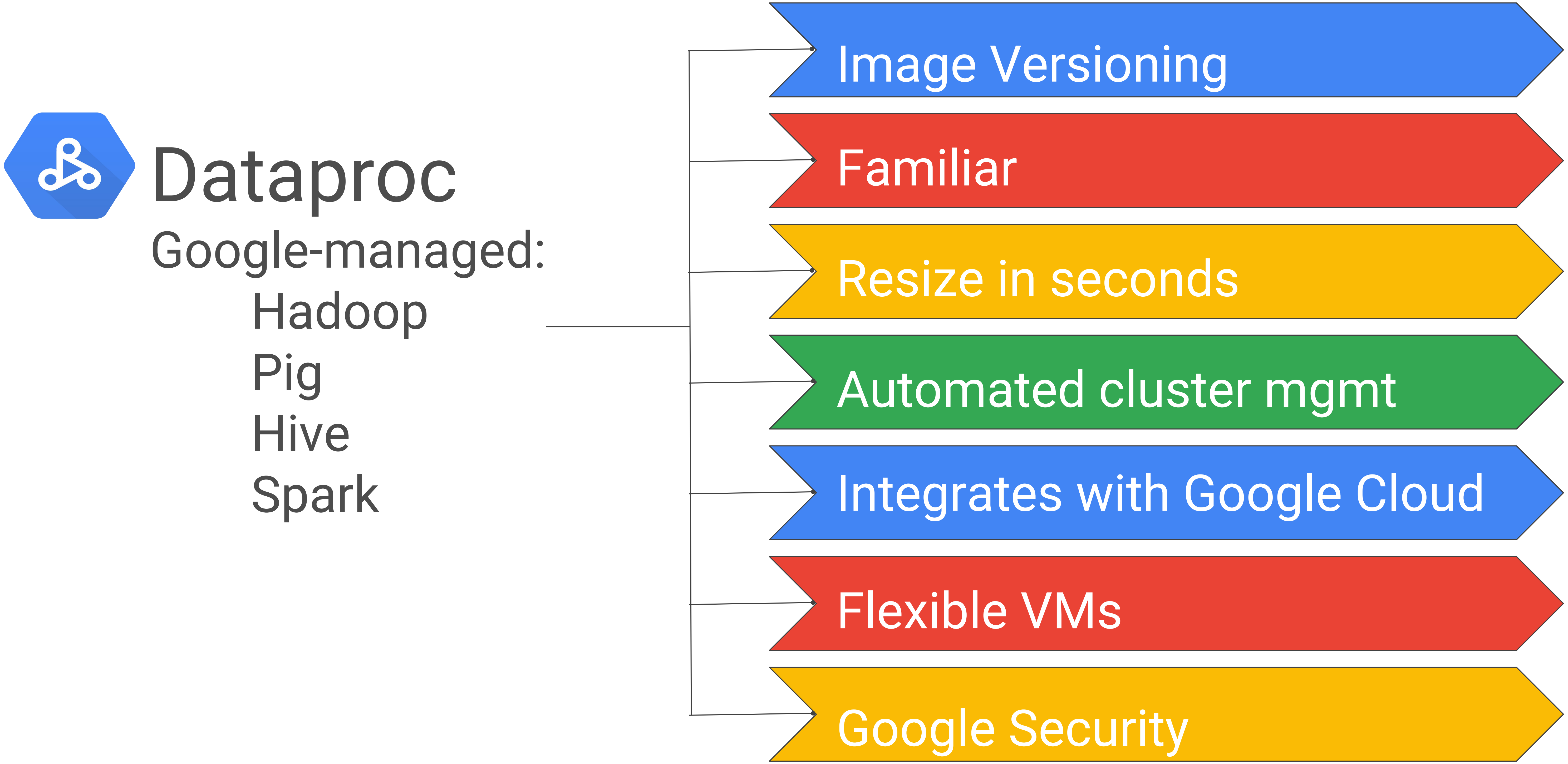
Managed Hadoop in the cloud + Lab

There is a rich open-source ecosystem for big data



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

Dataproc reduces the cost and complexity associated with Spark and Hadoop clusters



Dataproc

Google-managed:

- Hadoop
- Pig
- Hive
- Spark

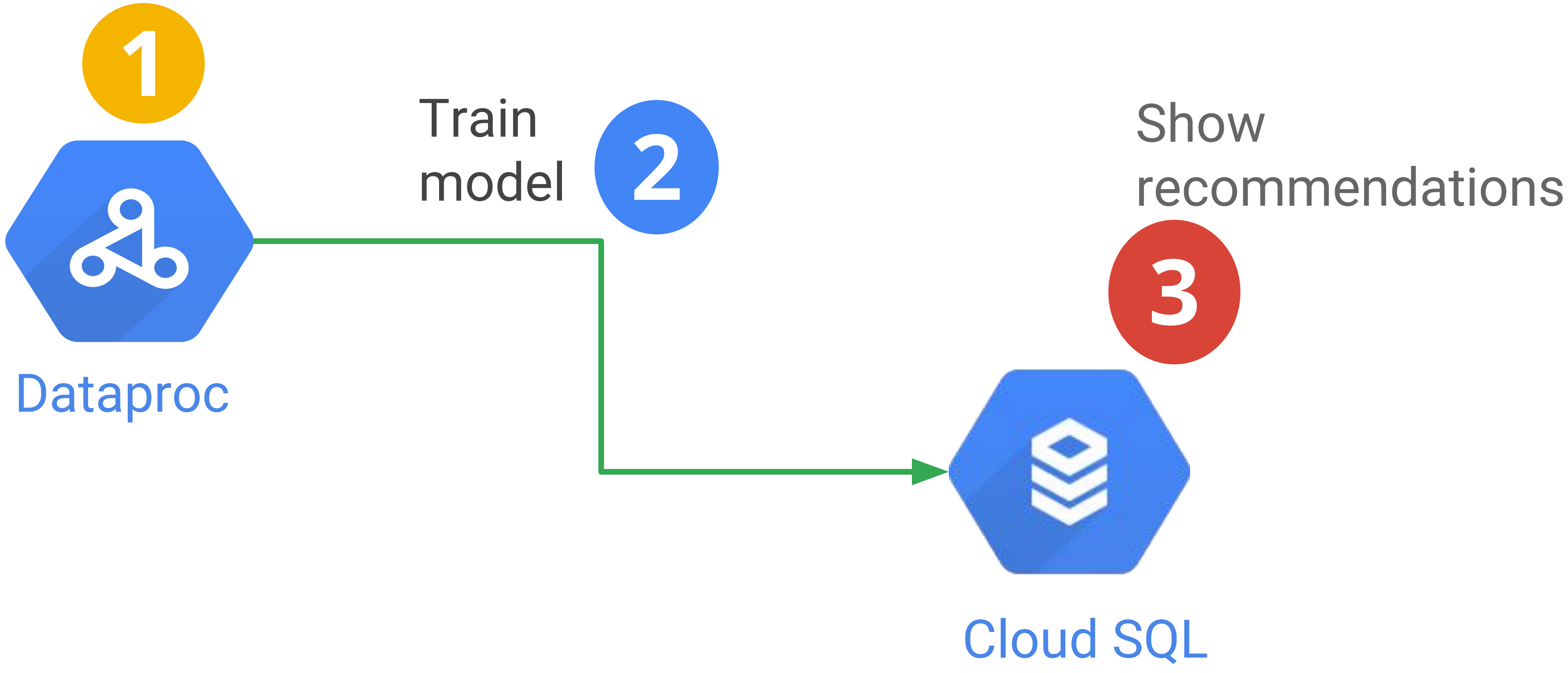
- Image Versioning
- Familiar
- Resize in seconds
- Automated cluster mgmt
- Integrates with Google Cloud
- Flexible VMs
- Google Security

Google Cloud

Lab: Recommendations ML with Dataproc

Lab 4: Recommendations ML with Cloud Dataproc

In this lab, you implement machine learning recommendations using Cloud Dataproc:



1. Launch Dataproc
2. Train and apply ML model written in PySpark to create product recommendations
3. Explore inserted rows in Cloud SQL

Google Cloud

Module Review

Module review (1 of 2)

Relational databases are a good choice when you need: (select all of the correct options)

❏ Streaming, high-throughput writes

❏ Fast queries on terabytes of data

❏ Aggregations on unstructured data

❏ Transactional updates on relatively small datasets

Module review answers (1 of 2)

Relational databases are a good choice when you need: (select all of the correct options)

❑ Streaming, high-throughput writes

❑ Fast queries on terabytes of data

❑ Aggregations on unstructured data

✓ Transactional updates on relatively small datasets

Google Cloud

Module review (2 of 2)

Cloud SQL and Cloud Dataproc offer familiar tools (MySQL and Hadoop/Pig/Hive/Spark). What is the value-add provided by Google Cloud Platform?
(select all of the correct options)

❑ It’s the same API, but Google implements it better
❑ Google-proprietary extensions and bug fixes to MySQL, Hadoop, and so on
❑ Fully-managed versions of the software offer no-ops
❑ Running it on Google infrastructure offers reliability and cost savings

Module review answers (2 of 2)

Cloud SQL and Cloud Dataproc offer familiar tools (MySQL and Hadoop/Pig/Hive/Spark). What is the value-add provided by Google Cloud Platform?
(select all of the correct options)

❏ It’s the same API, but Google implements it better
❏ Google-proprietary extensions and bug fixes to MySQL, Hadoop, and so on
✓ Fully-managed versions of the software offer no-ops
✓ Running it on Google infrastructure offers reliability and cost savings

Google Cloud

Resources

Cloud SQL	https://cloud.google.com/sql/
Cloud Dataproc	https://cloud.google.com/dataproc/
Cloud Solutions	https://cloud.google.com/solutions/

cloud.google.com

Images by Connie Zhou