Apache Spark and Google Cloud

Brad Miro - March 2022

Agenda

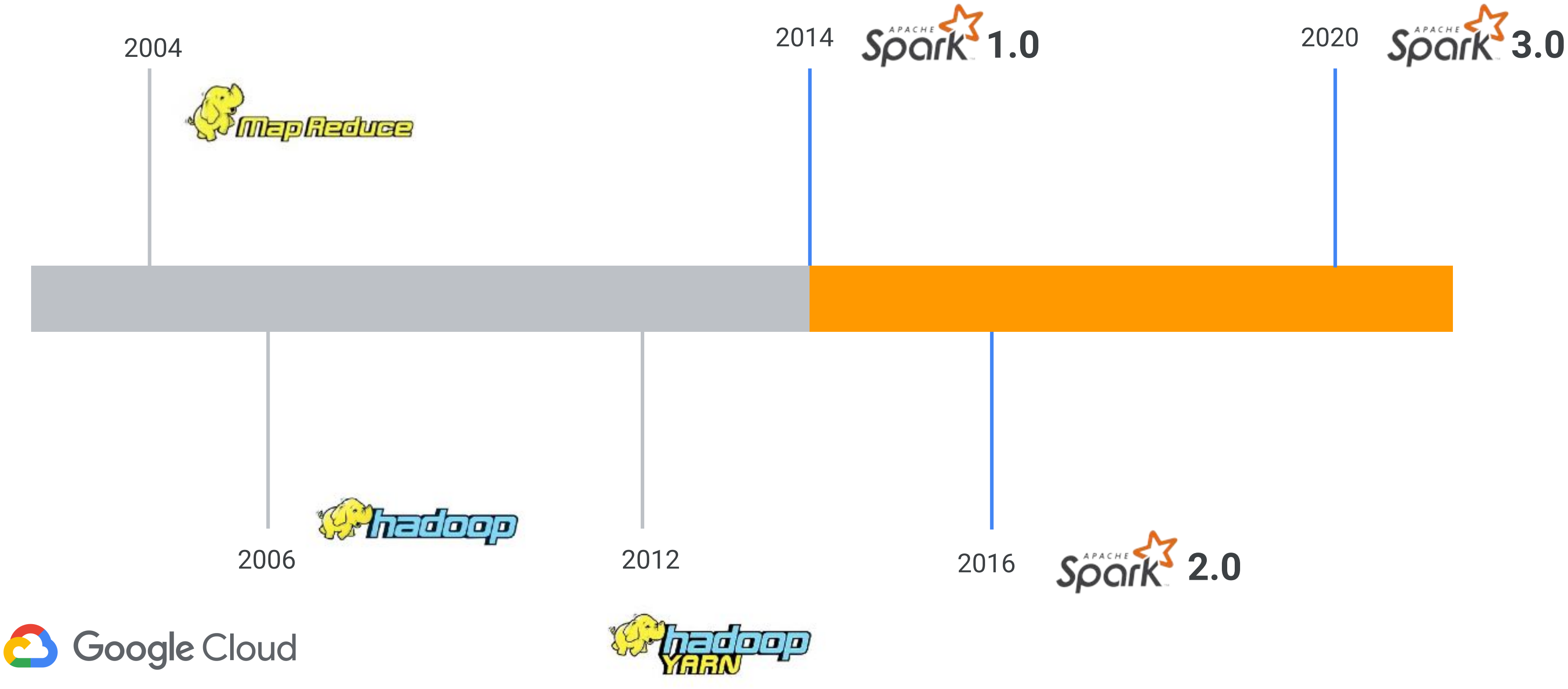
Part 1: Intro to Apache Spark

Part 2: Spark on Google Cloud

Part 3: Getting Started

Part 1: Apache Spark

Data processing history



2004

2006

2012

2014 1.0

2016 2.0

2020 3.0

Apache Spark

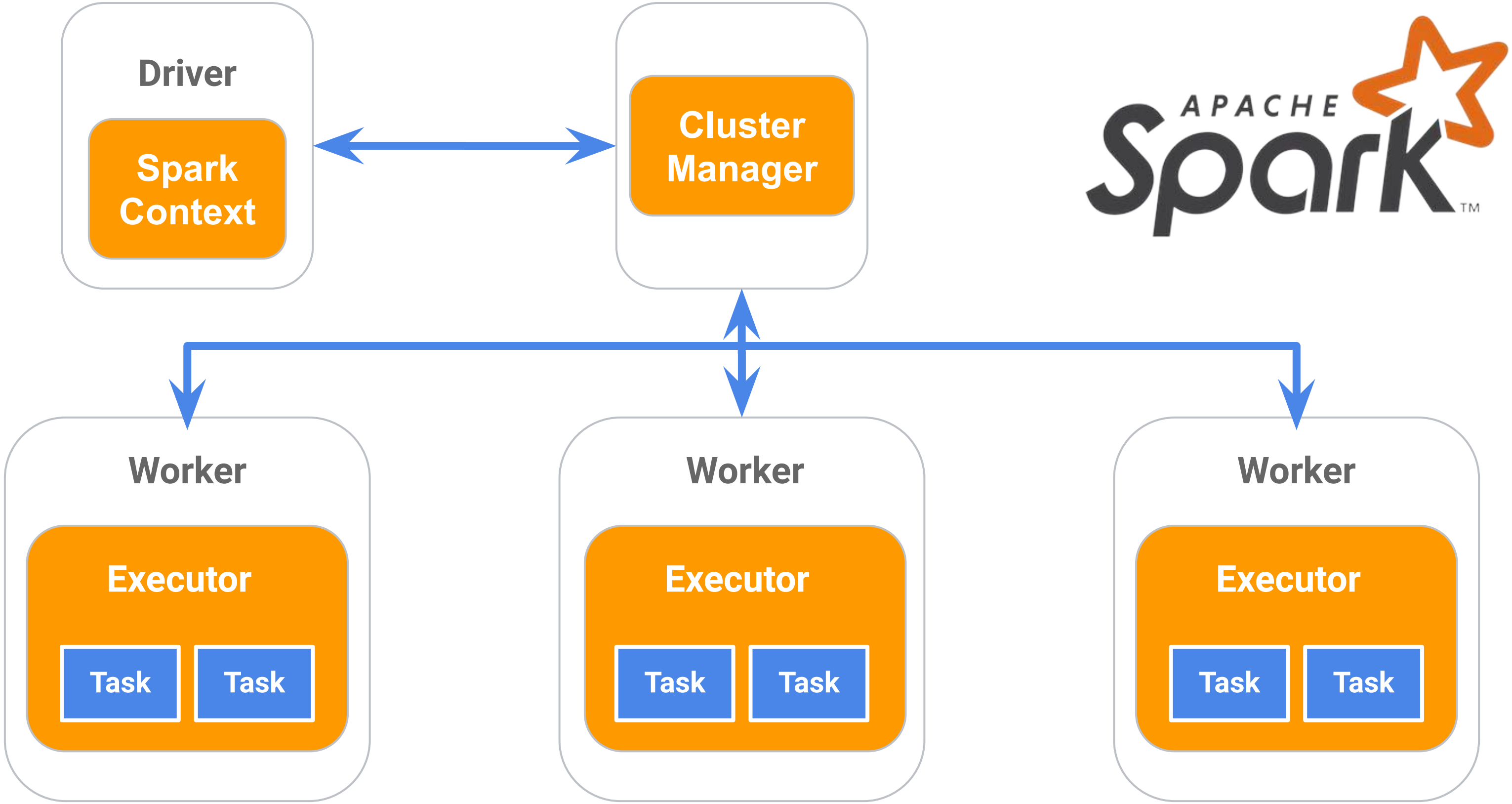
OSS “Unified analytics engine for large-scale data processing”

In-memory distributed data processing

Rich ecosystem

Python, Java, Scala, and R

Abstracted parallelization



Driver

Spark Context

Cluster Manager

APACHE Spark™

Worker

Executor

Task Task

Worker

Executor

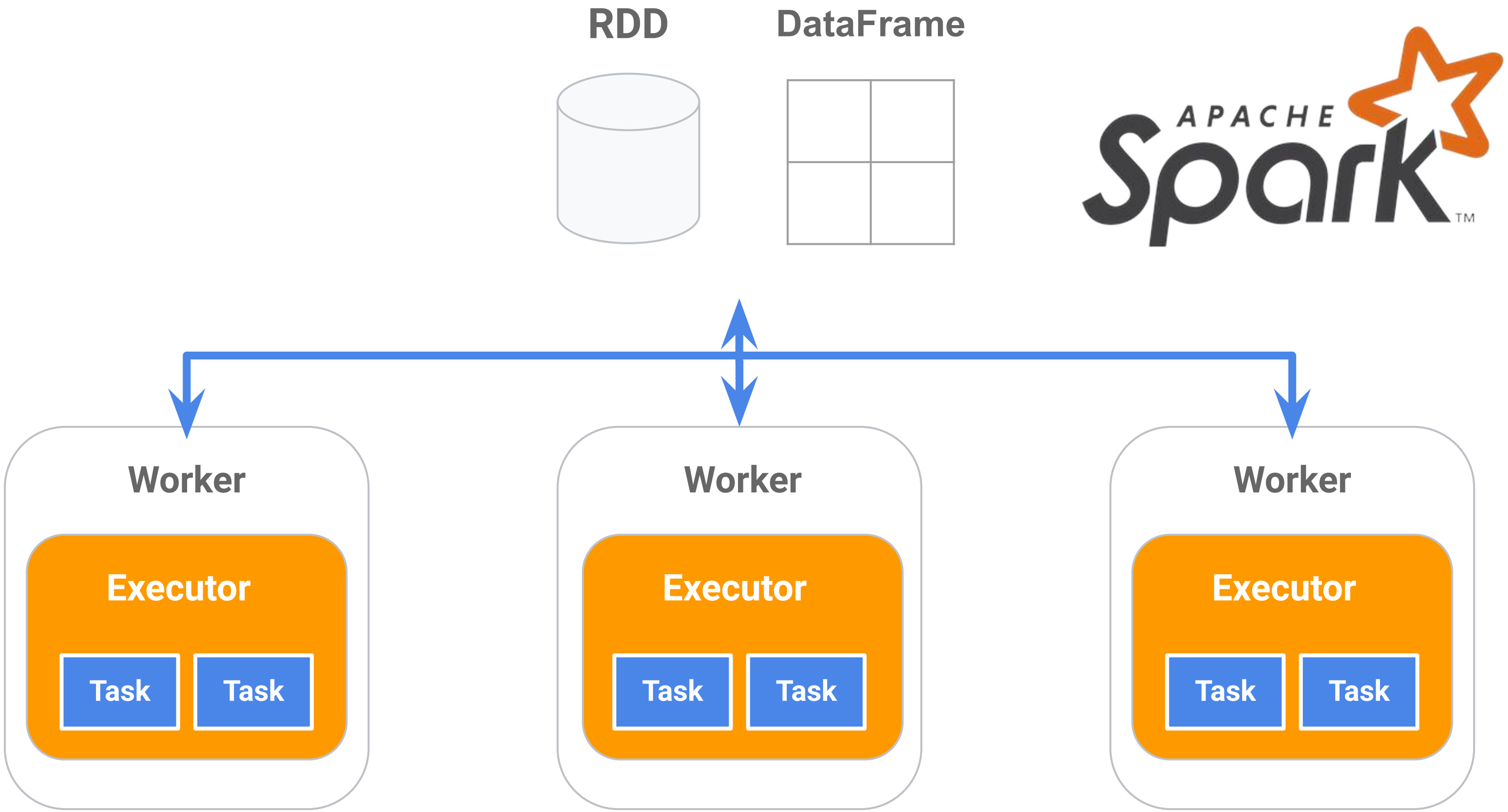
Task Task

Worker

Executor

Task Task

Google Cloud



RDD

DataFrame

APACHE Spark™

Worker

Executor

Task Task

Worker

Executor

Task Task

Worker

Executor

Task Task

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName(“Hello World!”) \
    .getOrCreate()

df = spark.read.option(inferSchema=True).csv(“data.csv”)

df.where(“age > 21”).select(“name.first”).show()
```

Datasources

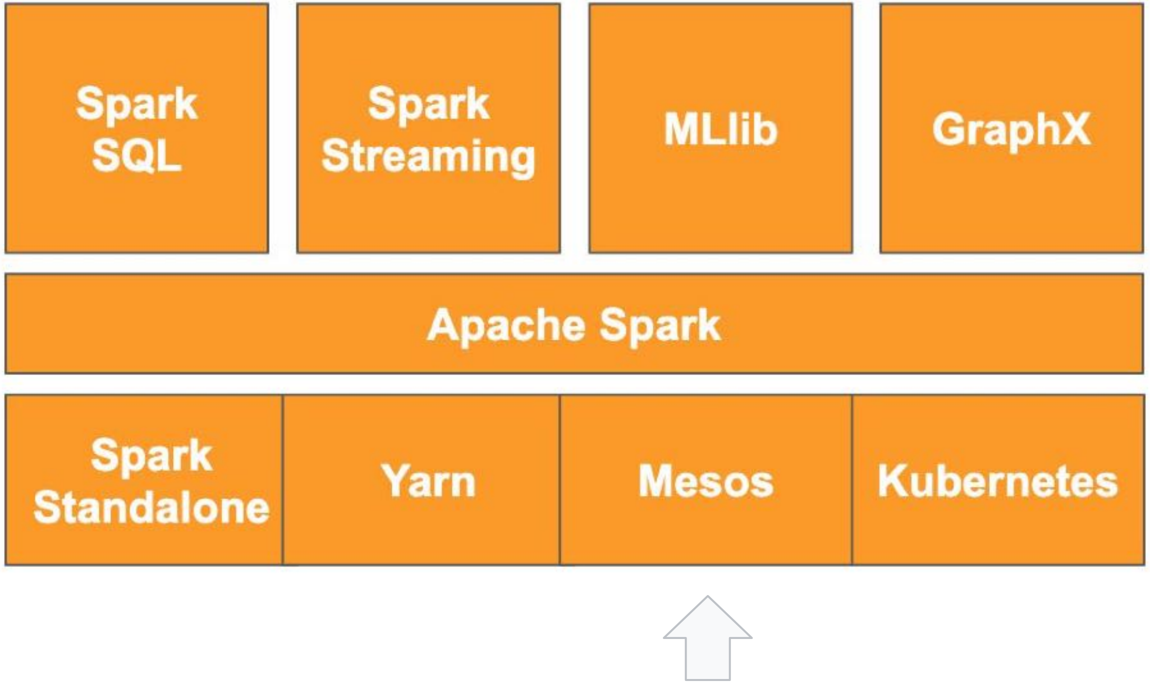
csv / json / parquet / avro

Blobstore (GCS, S3, etc)

HDFS

Iceberg, Delta Lake, Hudi

Data warehouses (BigQuery, Snowflake, etc.)



Ecosystem

Spark SQL

Spark Streaming

MLlib

GraphX

Apache Spark

Runtimes

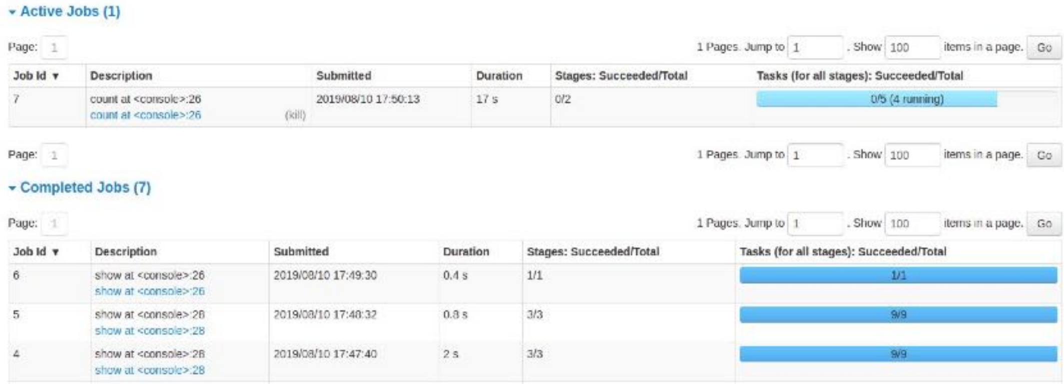
Spark Standalone

Yarn

Mesos

Kubernetes

Deprecated!



▾ **Active Jobs (1)**

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7	count at <console>:26 count at <console>:26 (kill)	2019/08/10 17:50:13	17 s	0/2	0/5 (4 running)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

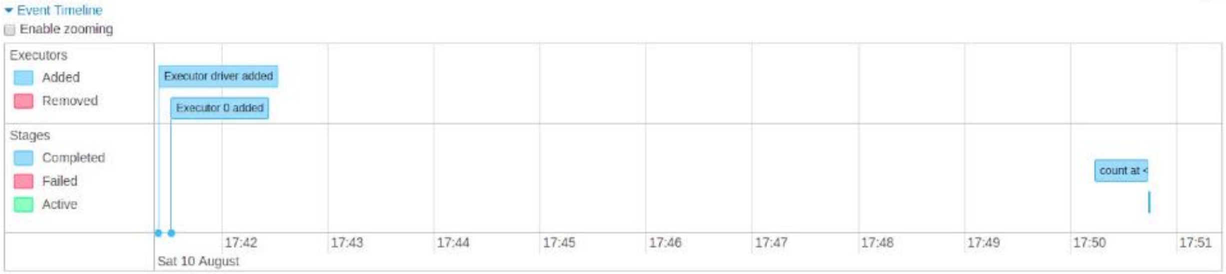
▾ **Completed Jobs (7)**

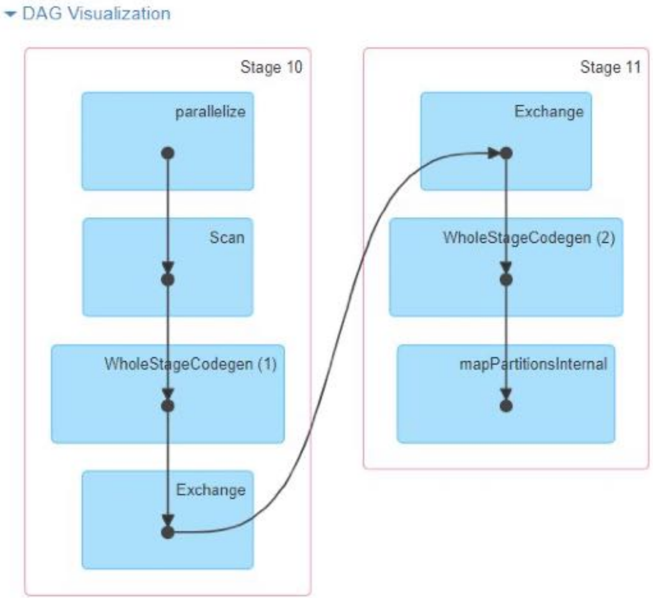
Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
6	show at <console>:26 show at <console>:26	2019/08/10 17:49:30	0.4 s	1/1	1/1
5	show at <console>:28 show at <console>:28	2019/08/10 17:48:32	0.8 s	3/3	9/9
4	show at <console>:28 show at <console>:28	2019/08/10 17:47:40	2 s	3/3	9/9

Details for Job 7

Status: SUCCEEDED
Associated SQL Query: 8
Completed Stages: 2





Google Cloud

Other features

Native GPU support (with significant NVIDIA investment)

Clusters can be single or multi-tenant

Transactional writes to prevent data loss during processing

Apache Spark vs Apache Beam

Beam: Maintained by Google, hides much of the internal happenings of Spark

Both fundamentally do the same thing

Spark is more popular, better for batch

Beam stronger for streaming

Why use Spark

Scale data processing off local machines to a larger cluster

Parallelize your data processing

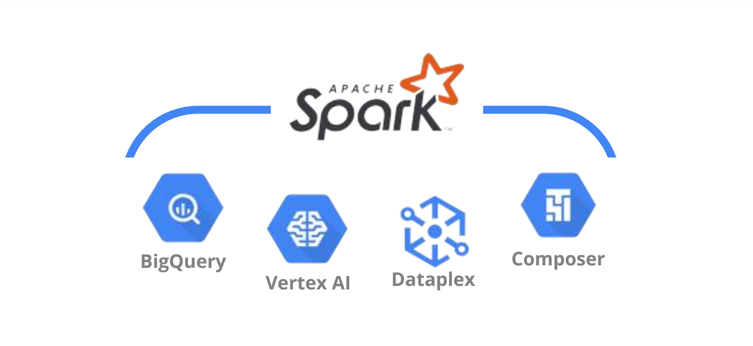
Spark is well-established: many open source add-ons

Cloud providers make deployment easy

Part 2:
Spark on
Google Cloud

Spark on Google Cloud

Industry's first autoscaling Serverless Spark, integrated with the best of Google Cloud.
Run and write spark where you need it across all use-cases: ETL, data science and exploration.



Industry's First Serverless Spark for All Workloads

Auto-scale, without any manual infrastructure provisioning or tuning for Spark. Empowers customers to shift from managing clusters to workloads.

Pervasive Spark Experience

Connect, analyze and execute Spark jobs from BigQuery, Vertex AI or Dataplex in 2 clicks, without any custom integrations, using the best of Google-native and Open Source tools.

Flexibility of Consumption

One size does not fit all. Choose between Serverless, Google Kubernetes Engine (GKE), and compute clusters for your Spark applications.

Serverless Spark

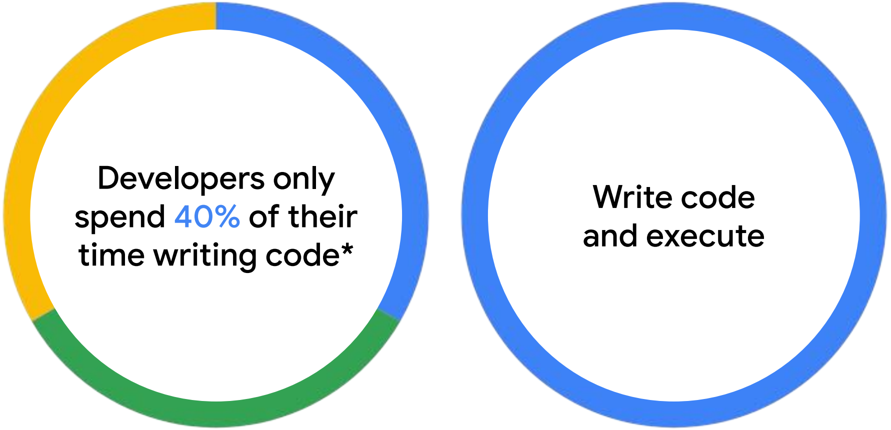
Focus on Spark, not infrastructure

Today

Manage clusters 🟦

Write code 🟨

Decide infrastructure 🟩
Pay while it is running



Developers only spend 40% of their time writing code*

Write code and execute

Spark with Serverless

- Job auto-scales
- No infrastructure to tune
- No clusters to manage
- Only pay for the job duration

*: PaaSView and the Developer 2021, May 2021, IDC

Spark through BigQuery

Google’s Cloud Data Warehouse

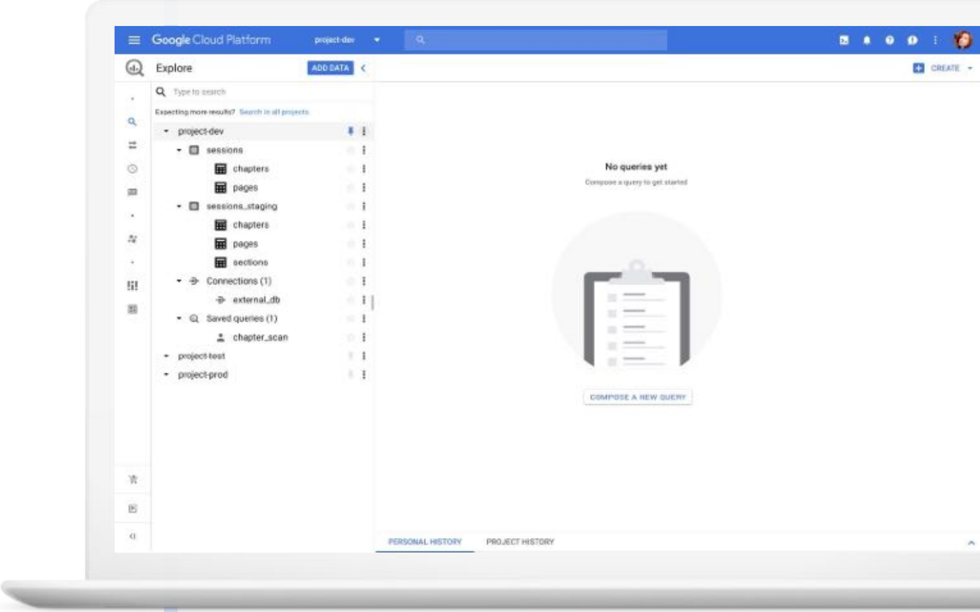
Serverless, highly scalable, multicloud data warehouse designed for business agility.

Unified SQL and Spark experience

Enable data warehousing users to easily write and execute Spark on BigQuery data without exporting it

Serverless Spark and SQL analytics

No infrastructure management required for either Spark or SQL analytics. Both autoscale.



No queries yet

Compose a query to get started

Spark through Dataplex

Intelligent Data Fabric

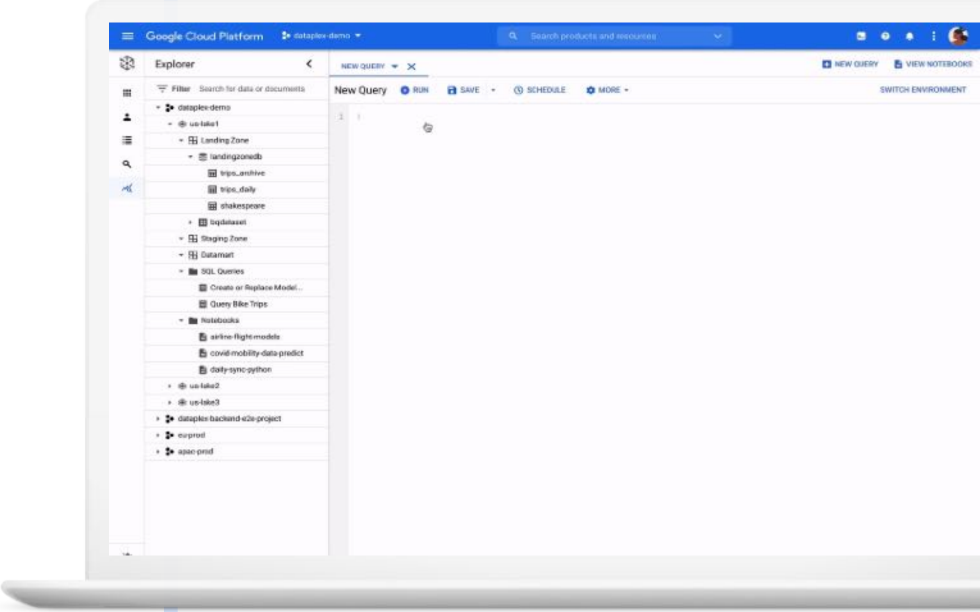
Centrally manage, monitor, and govern data across multiple data lakes and warehouses

Collaborative analytics environments

1-click access to SparkSQL, Notebooks, or PySpark. Easy collaboration with ability to save, share, search notebooks and scripts alongside data

Built-in governance across data lakes

Leverage the governance policies defined on your data lakes automatically



Spark through Vertex AI

Suite of tools for data science and ML

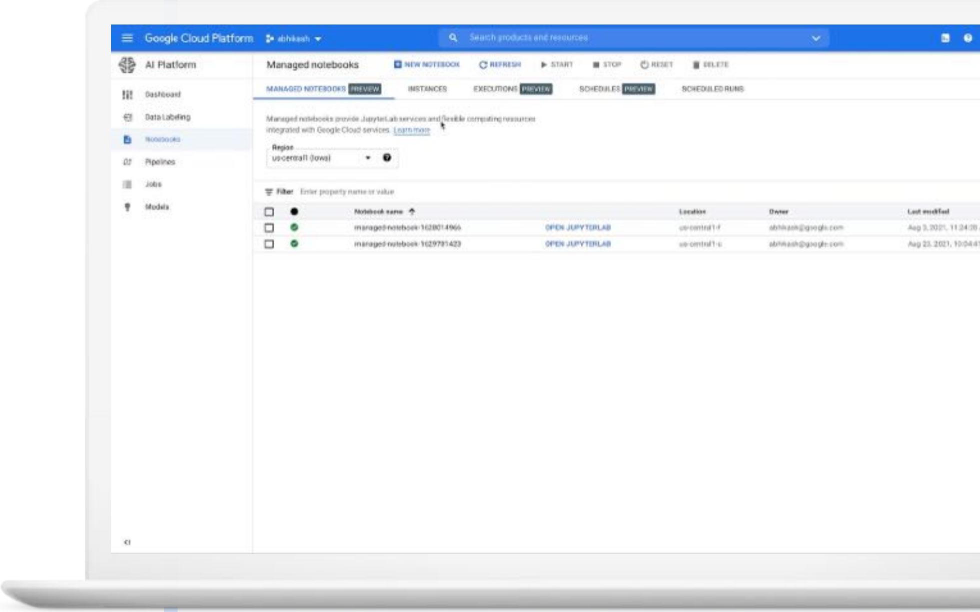
Build, deploy, and scale ML models faster, with pre-trained and custom tooling.

Built-in security and authentication

GCP security and user access are automatically applied from Vertex AI to Spark

Integrate Spark with MLOps

Execute Spark code through notebook executor, integrate with Vertex AI pipelines



Managed notebooks provide JupyterLab services and flexible computing resources integrated with Google Cloud services. Learn more

Region: us-central1 (Iowa)

□	●	Notebook name ↑		Location	Owner	Last modified
□	●	managed-notebook-1628014966	OPEN JUPYTERLAB	us-central1-f	abhikash@google.com	Aug 3, 2021, 11:24:20
□	●	managed-notebook-1629781423	OPEN JUPYTERLAB	us-central1-c	abhikash@google.com	Aug 23, 2021, 10:04:4

Flexibility of consumption with Dataproc

01

Dataproc on GCE

- **YARN** runtime
- Create managed clusters on GCP
- Fine grained cost and performance control

02

Dataproc Serverless

- **Standalone** runtime
- Developers can easily use Spark
- No clusters, no infra tuning

03

Dataproc on GKE

- **Kubernetes** runtime
- Simplify infrastructure management across the enterprise

Part 3:
Getting Started

Continued Learning

- spark.apache.org
- cloud.google.com/solutions/spark
- cloud.google.com/data-science
- *Spark: Cluster Computing with Working Sets*

  - Zaharia et al. (2010)

Youtube Channels:

- Sundog Education
- Simplilearn
- Databricks

Thank you!

Brad Miro
twitter.com/bradmiro
linkedin.com/in/bradmiro

Google Cloud