

03



# Big Data Analytics with Notebooks

# Big Data Analytics with Notebooks

01

What's a Notebook?

02

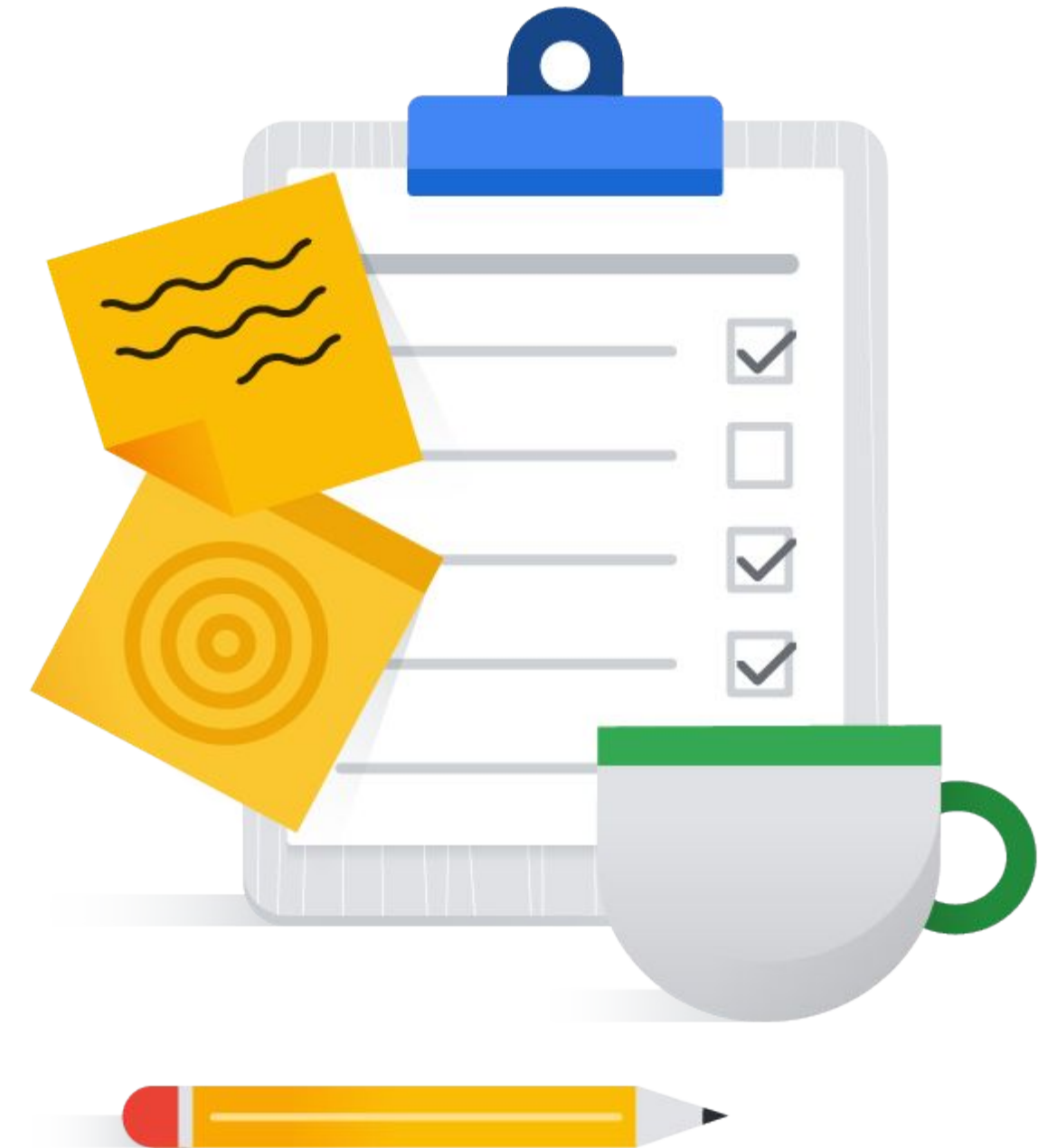
BigQuery magic and ties to Pandas



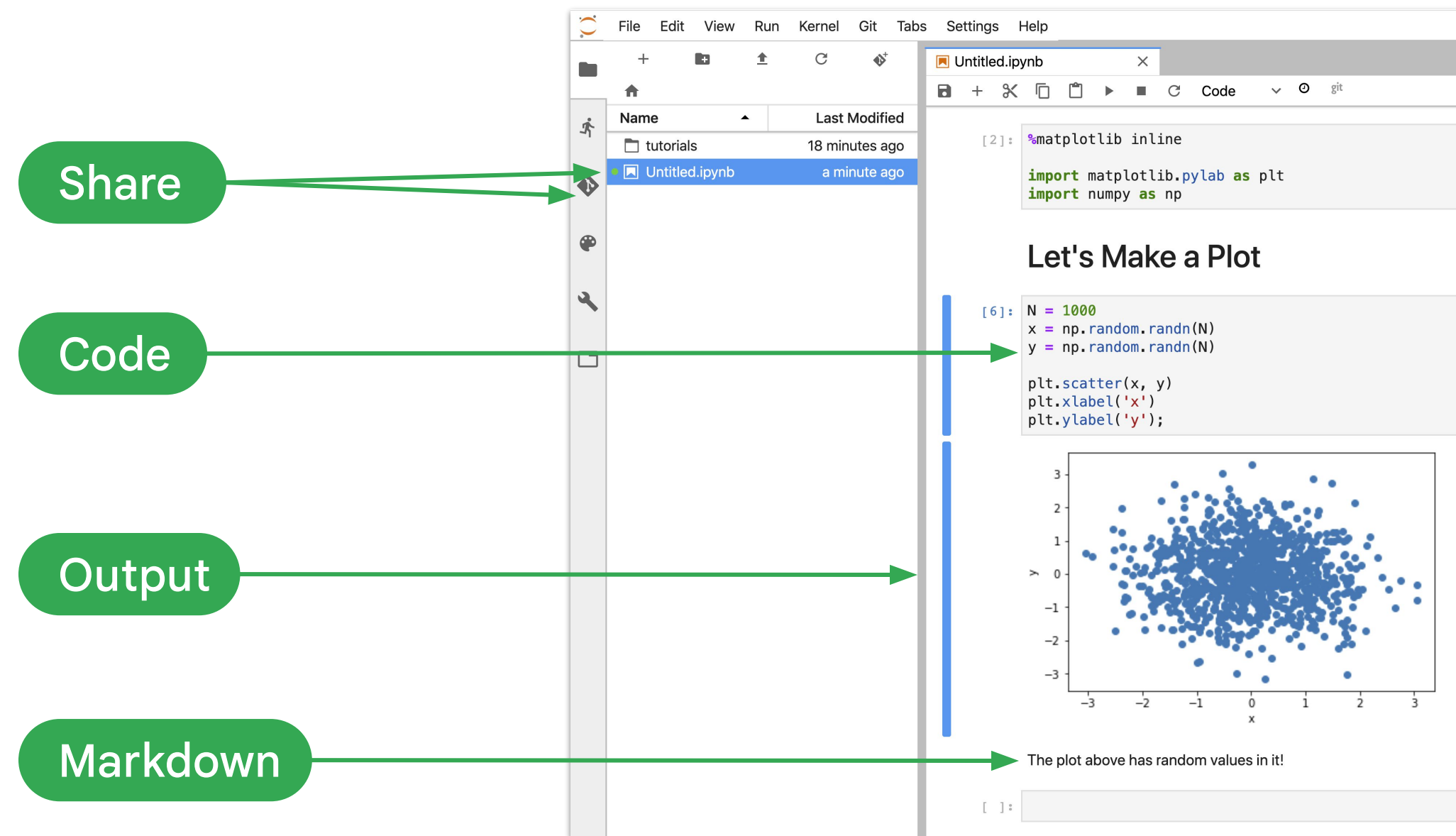
# Big Data Analytics with Notebooks

01 What's a Notebook?

02 BigQuery magic and ties to Pandas

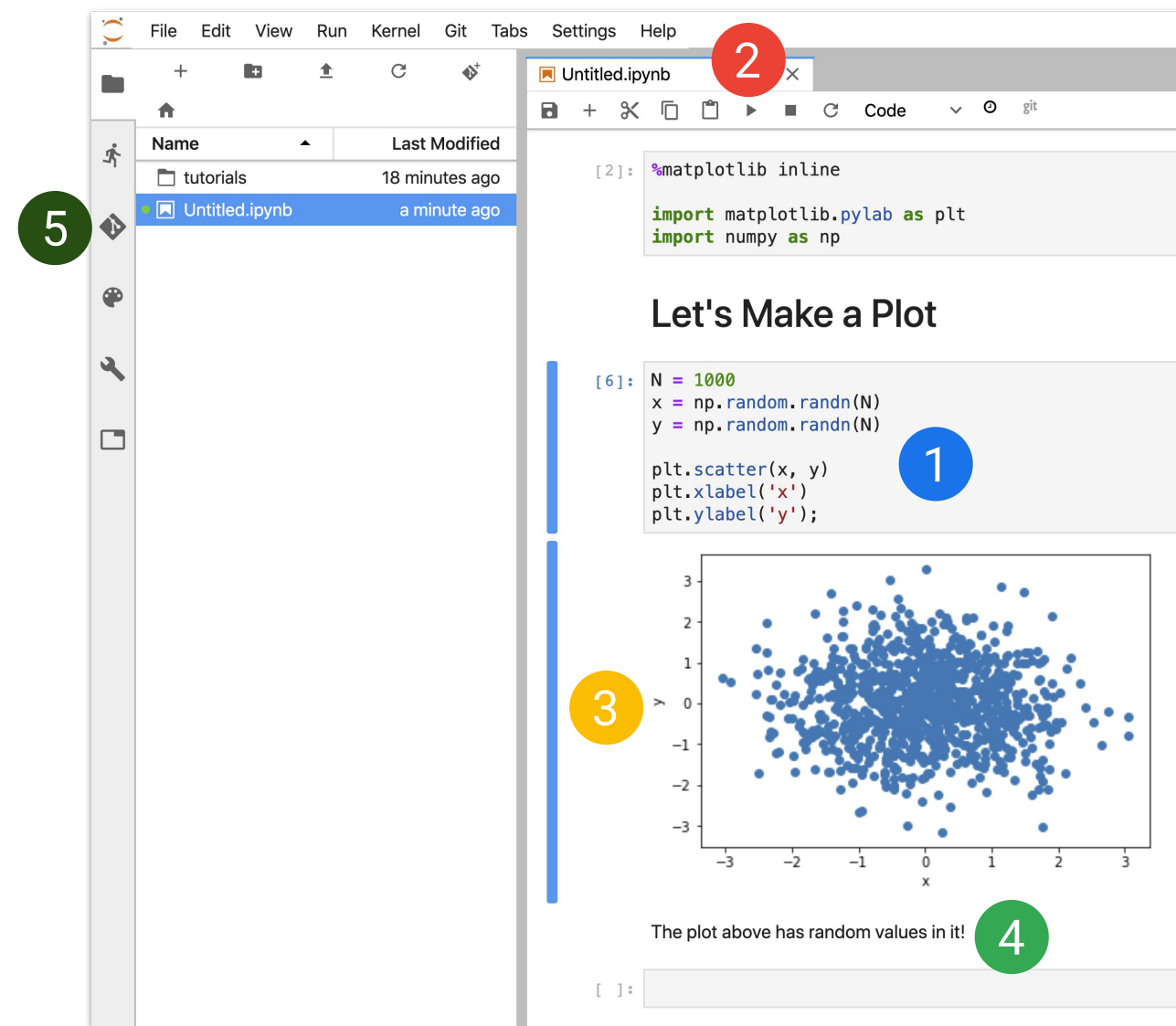
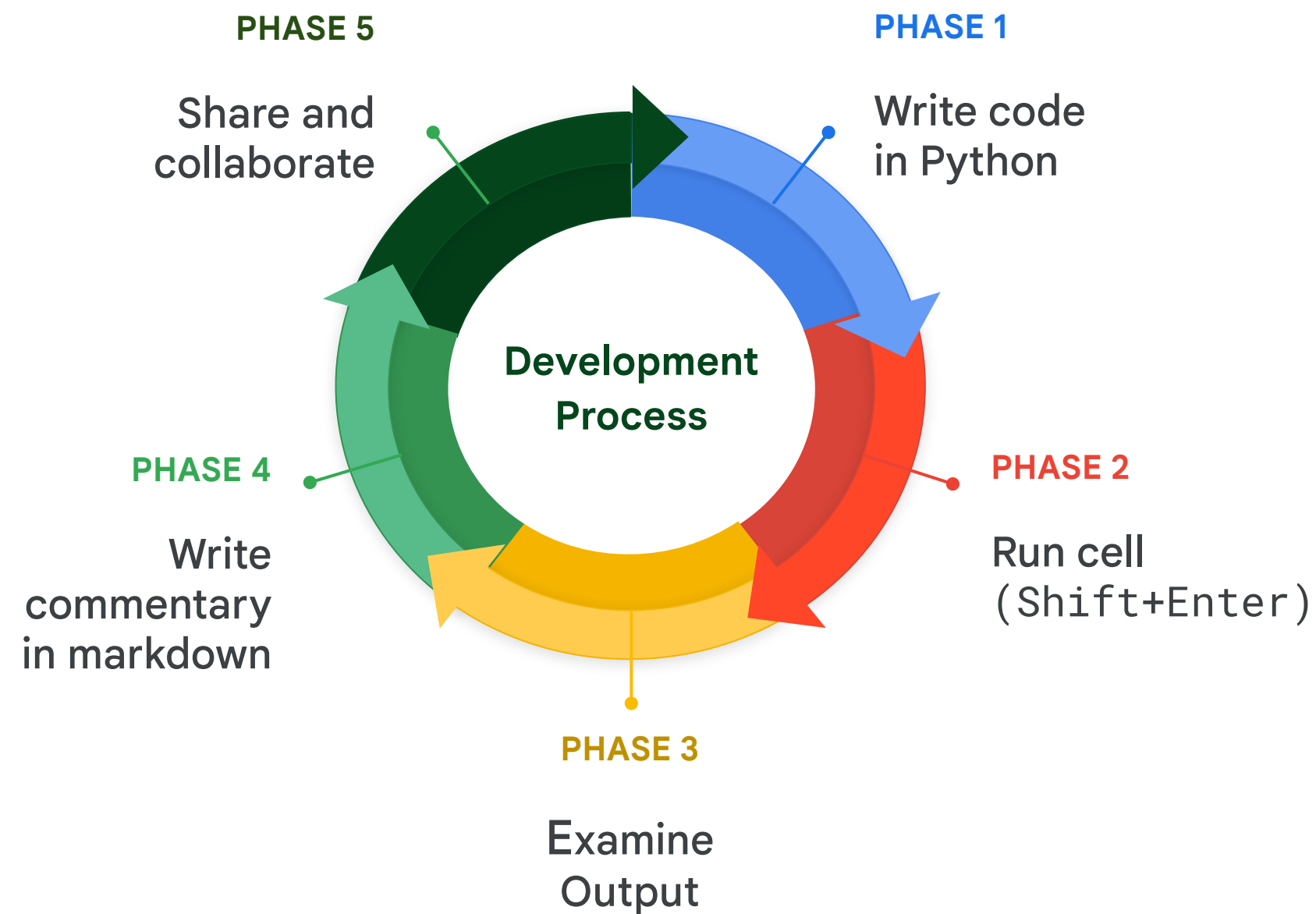


# Increasingly, data analysis and machine learning are carried out in self-descriptive, shareable, executable notebooks



A typical notebook contains code, charts, and explanations

# Notebooks are developed in an iterative, collaborative process



# Spin up a JupyterLab instance, pre-configured with the latest machine learning and data science frameworks in one click

The screenshot displays the Google Cloud Platform (GCP) AI Platform Notebooks interface. The top navigation bar includes the Google Cloud Platform logo, the project ID 'qwiklabs-gcp-04-bb1acaaefe54', a search bar, and user profile icons. The left sidebar shows the 'AI Platform' menu with options like Dashboard, Data Labeling, Notebooks (selected), Pipelines, Jobs, and Models. The main content area is titled 'Notebooks' and features a '+ NEW INSTANCE' button, a 'REFRESH' button, and status buttons (START, STOP, RESET, DELETE). Below these, a 'Customize instance' dialog is open, listing various pre-configured frameworks: Python 3, Python 3 (CUDA Toolkit 11.0), TensorFlow Enterprise, PyTorch 1.9, R 4.0, RAPIDS 0.18 [EXPERIMENTAL], Kaggle Python [BETA], Theia IDE [EXPERIMENTAL], and Smart Analytics Frameworks. A 'CREATE INSTANCE' button is at the bottom of the dialog. On the right, an 'Info panel' is visible with tabs for 'DOCUMENTATION' and 'LABELS', and a link to 'Notebook instances'.

Google Cloud Platform

qwiklabs-gcp-04-bb1acaaefe54

Search products and resources

AI Platform

Notebooks

+ NEW INSTANCE REFRESH START STOP RESET DELETE

HIDE INFO PANEL

Dashboard

Data Labeling

Notebooks

Pipelines

Jobs

Models

Create and use JupyterLab pre-installed frameworks. [Learn more](#)

Filter Enter property

Customize instance

Python 3  
Includes scikit-learn, pandas and more

Python 3 (CUDA Toolkit 11.0)  
Optimized for NVIDIA GPUs

TensorFlow Enterprise  
Includes Keras, scikit-learn, pandas, NLTK and more

PyTorch 1.9  
Includes scikit-learn, pandas, NLTK and more

R 4.0  
Includes basic R packages, scikit-learn, pandas, NLTK and more

RAPIDS 0.18 [EXPERIMENTAL]  
Optimized for NVIDIA GPUs

Kaggle Python [BETA]  
Python image for Kaggle Notebooks, supporting hundreds of machine learning libraries popular on Kaggle

Theia IDE [EXPERIMENTAL]  
IDE with notebook support including scikit-learn, pandas, and more

Smart Analytics Frameworks  
BigQuery, Apache Beam, Apache Spark, Apache Hive, and more

CREATE INSTANCE

Info panel

DOCUMENTATION LABELS

[Notebook instances](#)

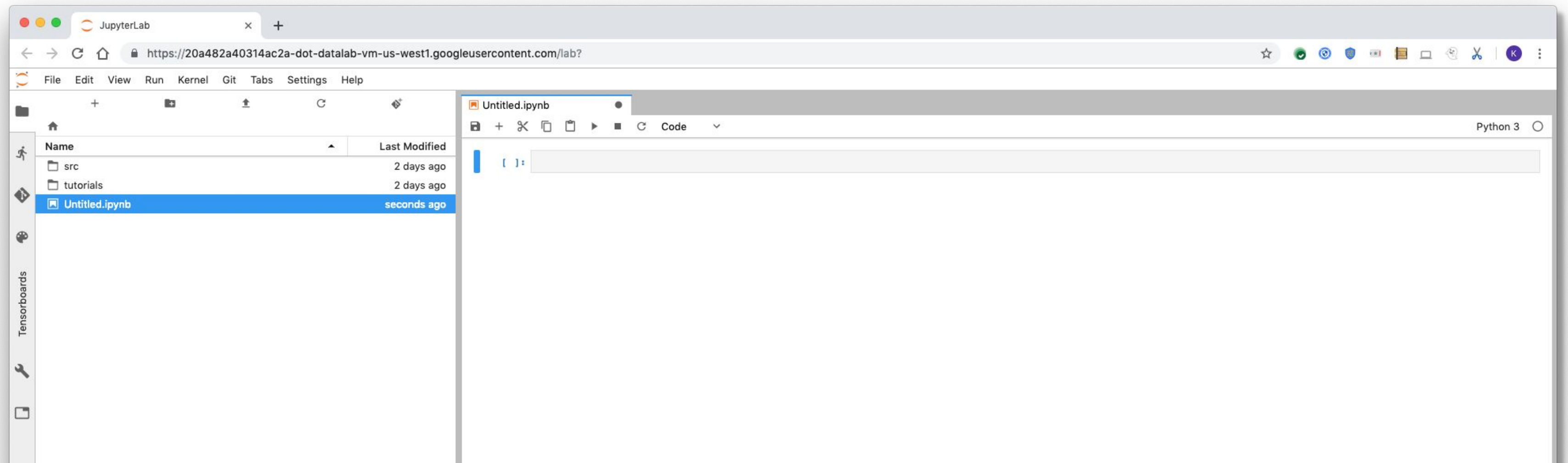
TensorFlow Enterprise 1.15 (with LTS) Without GPUs

TensorFlow Enterprise 2.1 (with LTS) With 1 NVIDIA Tesla T4

TensorFlow Enterprise 2.3 (with LTS)

TensorFlow Enterprise 2.5

# Notebooks use the latest open-source version of the industry-standard JupyterLab





# Use any Compute Engine instance type

←

Editing Notebook instance details

OPEN JUPYTERLAB

CANCEL

VIEW VM DETAILS

▶ START

■ STOP

↺ RESET

⬆ UPGRADE

🗑 DELETE

● tensorflow-1-15-20210701-152344

BASIC INFO

Region	us-west1 (Oregon)
Zone	us-west1-b
Environment ?	TensorFlow Enterprise 1.15 (with LTS and Intel® MKL-DNN/MKL)
Environment version	M74
Boot disk	100 GB disk
Data disk	100 GB disk
Backup	Not specified
Subnetwork	default
Service account	673971842877-compute@developer.gserviceaccount.com
Permission mode	Service account
Sudo access	Enabled
File downloads	Enabled
nbconvert	Enabled
Shielded VM	Secure Boot not enabled vTPM enabled Integrity Monitoring enabled

Machine configuration

Machine type \*  
n1-standard-4 (4 vCPUs, 15 GB RAM)

GPUs

GPU type  
None

Based on the zone, environment, and machine type selected above, the available GPU types and the minimum number of GPUs that can be selected may vary. [Learn more](#)

System

Environment auto-upgrade

If a new environment version is available, an upgrade will be conducted for your running instance at the time specified. It will be skipped if instances are stopped. **All the data in your data disk will be kept. Backward compatibility is not guaranteed.** [Learn more](#)

☐ Enable environment auto-upgrade

Instance upgrade history

Includes all the environment upgrades conducted for this instance. If needed, you could rollback the latest upgrade.

Date	Description
No upgrades to display	

SUBMIT

CANCEL



# You can easily change hardware

Project

Notebook instances

+ NEW INSTANCE

▶ START

■ STOP

⏻ RESET

🗑 DELETE

SHOW INFO PANEL

Filter

?

↺

⏏

<input type="checkbox"/>	Instance name		Region	ML framework	Machine type	GPUs	Labels
<input type="checkbox"/>	ml-notebook-runtime-1	<a href="#">OPEN JUPYTERLAB</a>	us-east1-b	TensorFlow 1.12	<div><div>1</div><div>2</div><div>4</div><div>8</div></div>	<div><div>✓ None</div><div>NVIDIA Tesla K80 12 GB GDDR5, \$1,065.80 per month</div><div>NVIDIA Tesla P100 12 GB HBM2, \$1,065.80 per month</div><div>NVIDIA Tesla P4 8 GB GDDR5, \$1,065.80 per month</div></div>	

# You can even add and remove GPUs

Notebooks

NEW INSTANCE

REFRESH

START

STOP

RESET

DELETE

SHOW INFO PANEL

Create and use Jupyter Notebooks with a notebook instance. Notebook instances have JupyterLab pre-installed and are configured with GPU-enabled machine learning frameworks. [Learn more](#)

Filter

Enter property name or value

?

⋮

		Instance name ↑		Zone	Environment Version	Auto-upgrade	Environment	Machine type	GPUs	Permission
<input type="checkbox"/>	✓	python-20210701-151535	<a href="#">OPEN JUPYTERLAB</a>	us-west1-b	M73	—	NumPy/SciPy/scikit-learn	4 vCPUs, 15 GB RAM ▾	None ▾	Service account
<input checked="" type="checkbox"/>	⏸	<a href="#">tensorflow-1-15-20210701-152344</a>	OPEN JUPYTERLAB	us-west1-b	M74	—	TensorFlow:1.15	4 vCPUs, 15 GB RAM ▾	None ▾	Service account

✓ None

NVIDIA Tesla K80

\$328.50 per month

▶

NVIDIA Tesla P100

\$1,065.80 per month

▶

NVIDIA Tesla T4

\$255.50 per month

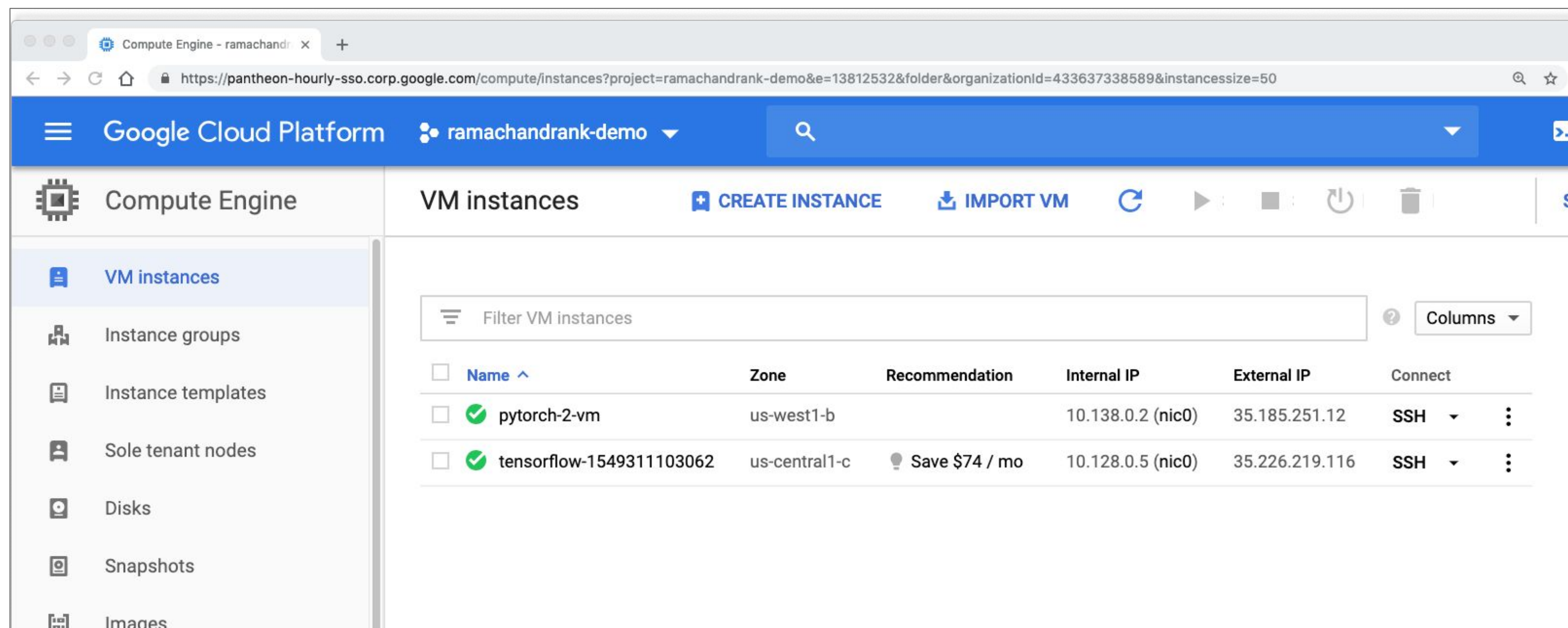
▶

NVIDIA Tesla V100

\$1,810.40 per month

▶

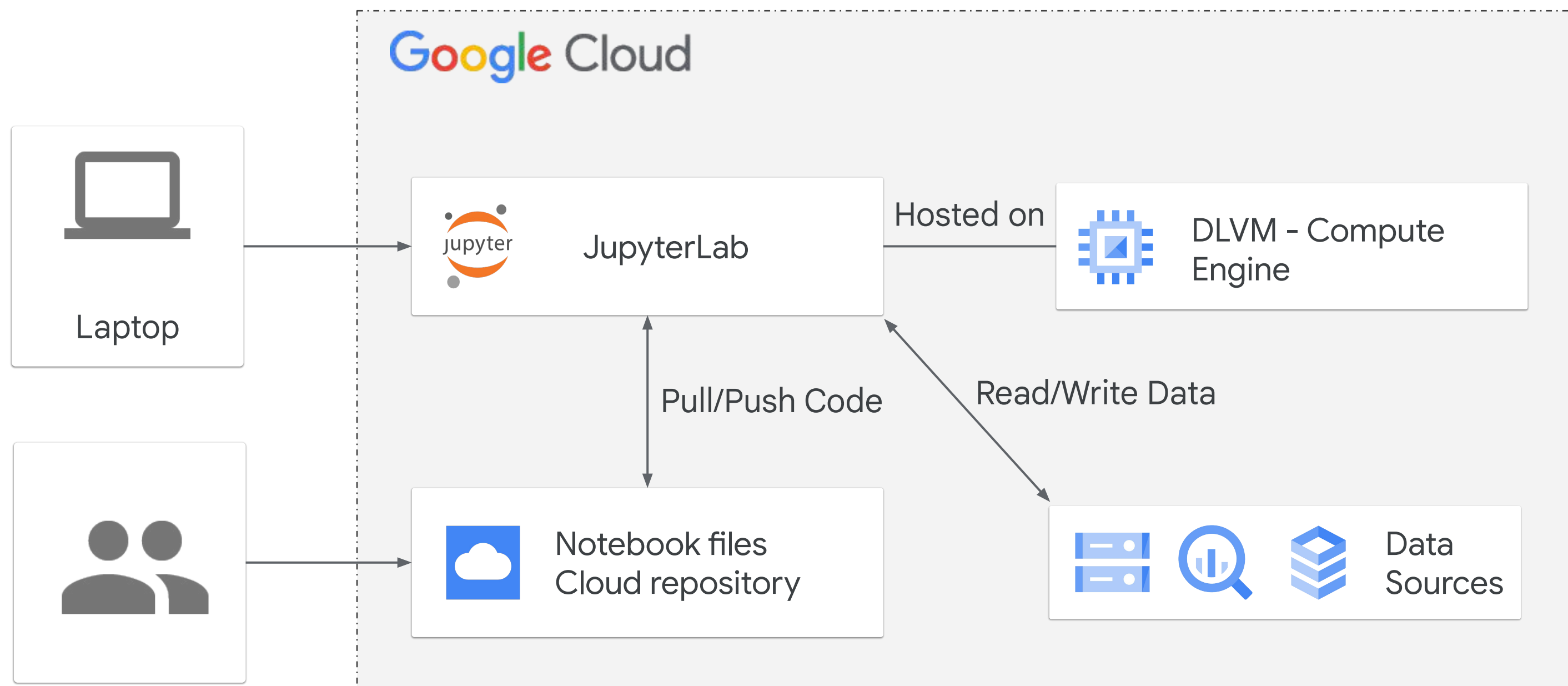
# Notebook instances are standard Compute Engine instances that live in your projects



The screenshot shows the Google Cloud Platform interface for VM instances. The left sidebar contains a navigation menu with options: VM instances (selected), Instance groups, Instance templates, Sole tenant nodes, Disks, Snapshots, and Images. The main content area is titled 'VM instances' and includes buttons for 'CREATE INSTANCE', 'IMPORT VM', and other actions. A table lists two VM instances:

<input type="checkbox"/>	Name ^	Zone	Recommendation	Internal IP	External IP	Connect
<input type="checkbox"/>	✓ pytorch-2-vm	us-west1-b		10.138.0.2 (nic0)	35.185.251.12	SSH ▾ ⋮
<input type="checkbox"/>	✓ tensorflow-1549311103062	us-central1-c	💡 Save \$74 / mo	10.128.0.5 (nic0)	35.226.219.116	SSH ▾ ⋮

# How does it work?



# Big Data Analytics with Notebooks

01

What's a Notebook?

02

BigQuery magic and ties to Pandas

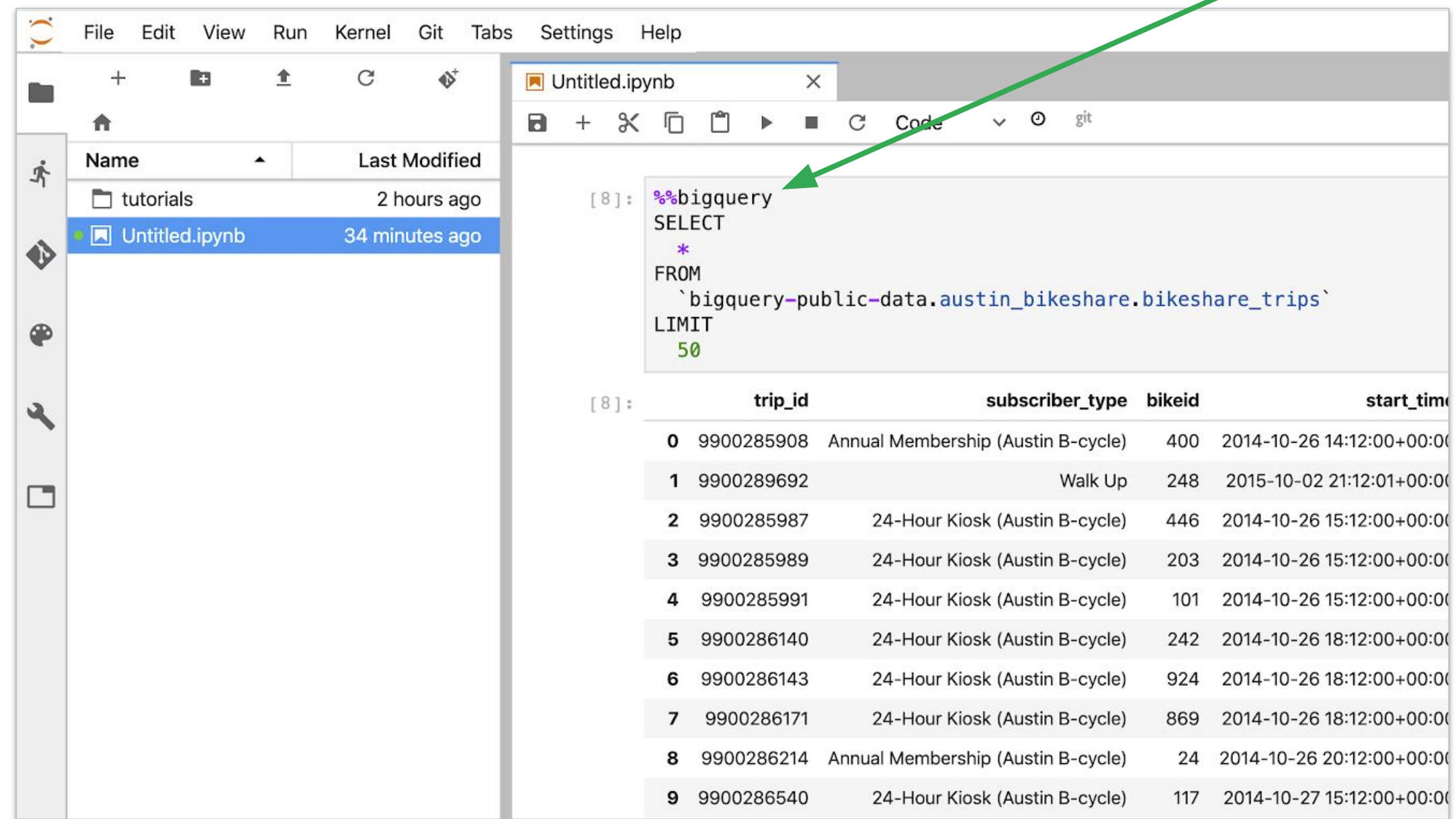




# You can execute BigQuery commands from Notebooks

- Useful for checking query validity
- Viewing query output
- But... can't use query output for anything

Jupyter “magic” function



The screenshot shows a Jupyter Notebook window with a file explorer on the left and a code editor on the right. The file explorer shows a folder named 'tutorials' and a file named 'Untitled.ipynb'. The code editor shows a Jupyter Notebook cell with the following code:

```
[8]: %%bigquery
      SELECT
      *
      FROM
      `bigquery-public-data.austin_bikeshare.bikeshare_trips`
      LIMIT
      50
```

A green arrow points from the text 'Jupyter “magic” function' to the `%%bigquery` line in the code. Below the code, the output of the query is displayed as a table with 5 columns: `trip_id`, `subscriber_type`, `bikeid`, and `start_time`. The table contains 10 rows of data.

	trip_id	subscriber_type	bikeid	start_time
0	9900285908	Annual Membership (Austin B-cycle)	400	2014-10-26 14:12:00+00:00
1	9900289692	Walk Up	248	2015-10-02 21:12:01+00:00
2	9900285987	24-Hour Kiosk (Austin B-cycle)	446	2014-10-26 15:12:00+00:00
3	9900285989	24-Hour Kiosk (Austin B-cycle)	203	2014-10-26 15:12:00+00:00
4	9900285991	24-Hour Kiosk (Austin B-cycle)	101	2014-10-26 15:12:00+00:00
5	9900286140	24-Hour Kiosk (Austin B-cycle)	242	2014-10-26 18:12:00+00:00
6	9900286143	24-Hour Kiosk (Austin B-cycle)	924	2014-10-26 18:12:00+00:00
7	9900286171	24-Hour Kiosk (Austin B-cycle)	869	2014-10-26 18:12:00+00:00
8	9900286214	Annual Membership (Austin B-cycle)	24	2014-10-26 20:12:00+00:00
9	9900286540	24-Hour Kiosk (Austin B-cycle)	117	2014-10-27 15:12:00+00:00



# Can use the BigQuery API in Notebooks to return query results as a Pandas DataFrame

```
[44]: %%bigquery df
      SELECT
      *
      FROM
      `bigquery-public-data.austin_bikeshare.bikeshare_trips`
      WHERE
      end_station_name = 'Stolen'
```

```
[46]: print(type(df))
      df.head()
```

<class 'pandas.core.frame.DataFrame'>

```
[46]:
```

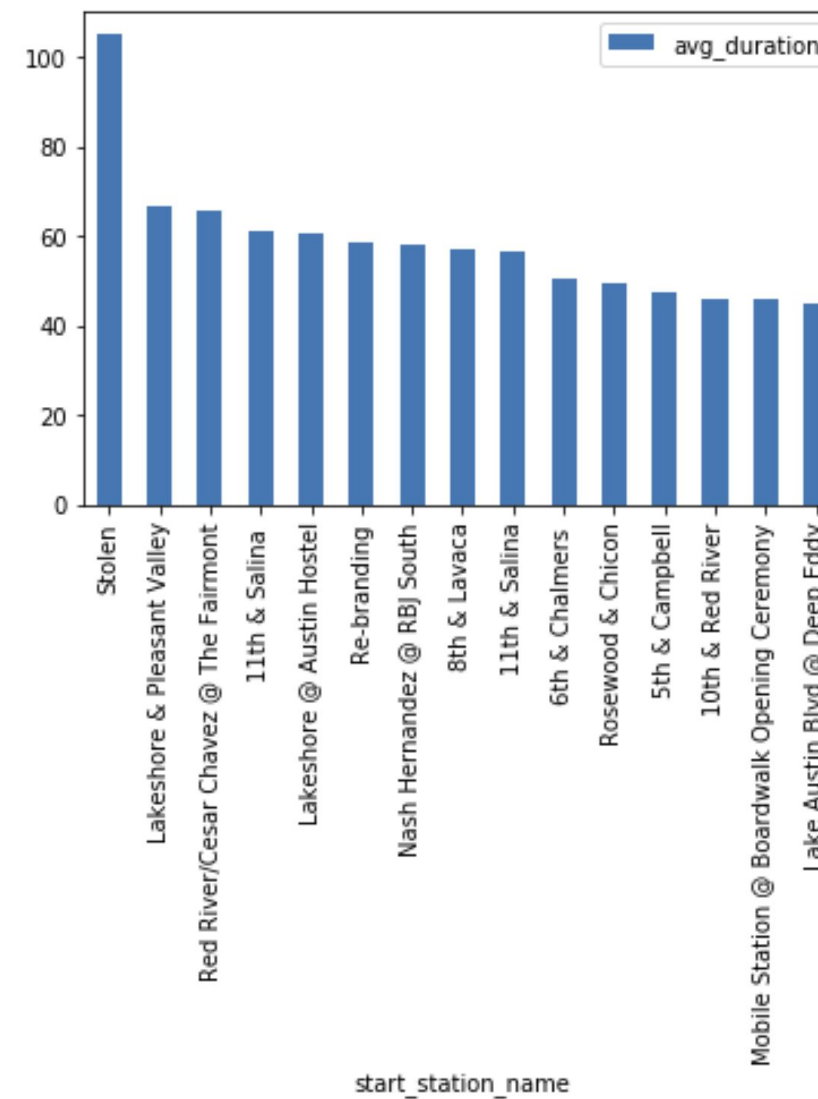
	trip_id	subscriber_type	bikeid	start_time	start_station_id	start_station_name	end_station_id	end_station_name	duration_minutes
0	9900259257	Walk Up	93	2015-09-18 08:12:05+00:00	2712	Toomey Rd @ South Lamar	None	Stolen	2863
1	16898448	Walk Up	1857	2018-03-18 22:51:20+00:00	2501	5th & Bowie	None	Stolen	3806
2	9900298869	Walk Up	127	2015-10-10 19:12:38+00:00	2574	Zilker Park	None	Stolen	3632
3	9900290440	Local365	277	2015-10-02 22:12:06+00:00	2494	2nd & Congress	None	Stolen	8
4	9900322570	Walk Up	439	2015-11-01 02:12:28+00:00	2496	8th & Congress	None	Stolen	6609

Pandas DataFrame

# Pandas + BigQuery in Notebook rocks!

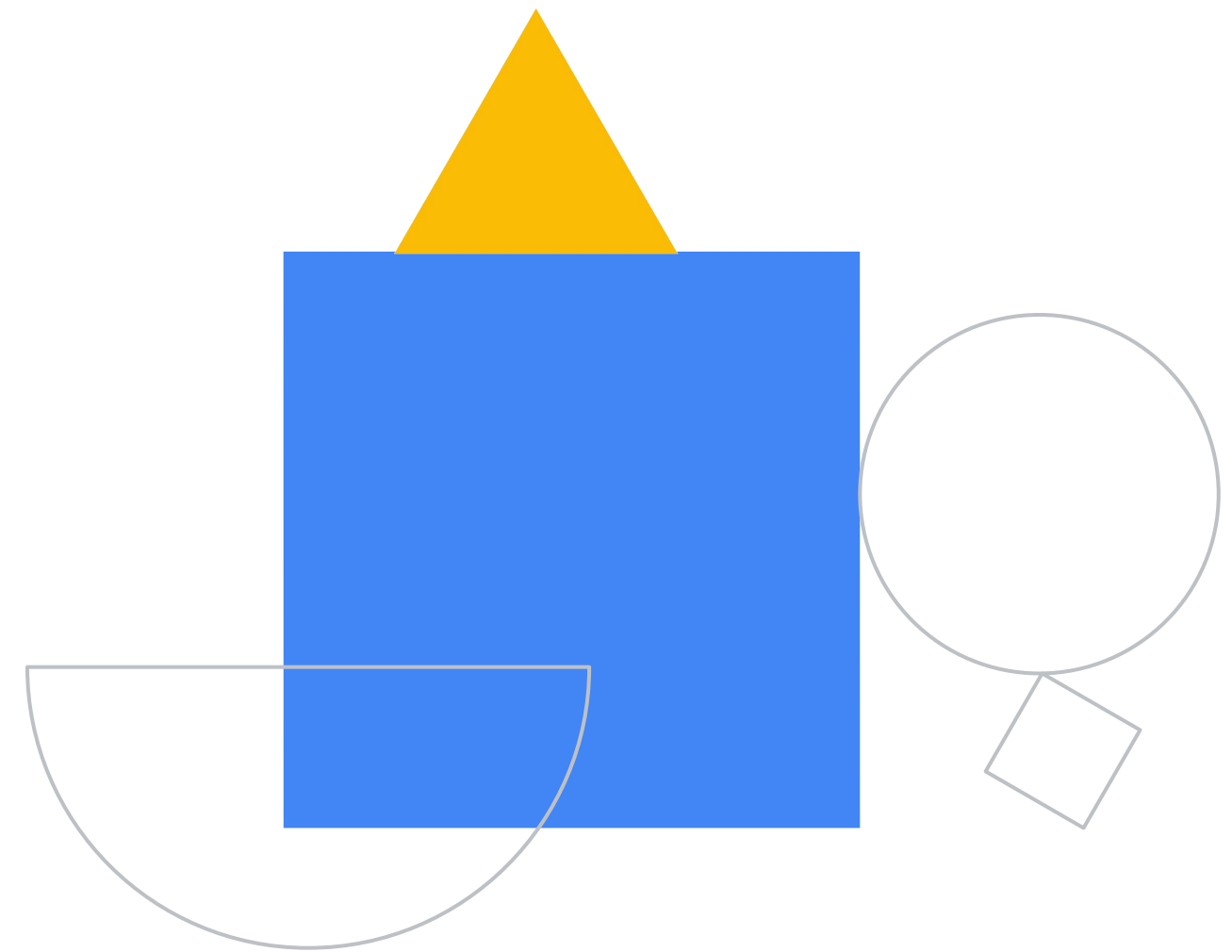
```
[47]: %%bigquery avg_dur_by_station
SELECT
  start_station_name,
  AVG(duration_minutes) as avg_duration
FROM
  `bigquery-public-data.austin_bikeshare.bikeshare_trips`
GROUP BY
  start_station_name
ORDER BY
  avg_duration
DESC
LIMIT 15
```

```
[48]: avg_dur_by_station.plot(x='start_station_name', y='avg_duration', kind='bar');
```



# Lab Intro

BigQuery in JupyterLab  
on Vertex AI



# Lab objectives

01

Instantiate a Jupyter notebook on AI Platform

02

Execute a BigQuery query from within a Jupyter notebook and process the output using Pandas



