

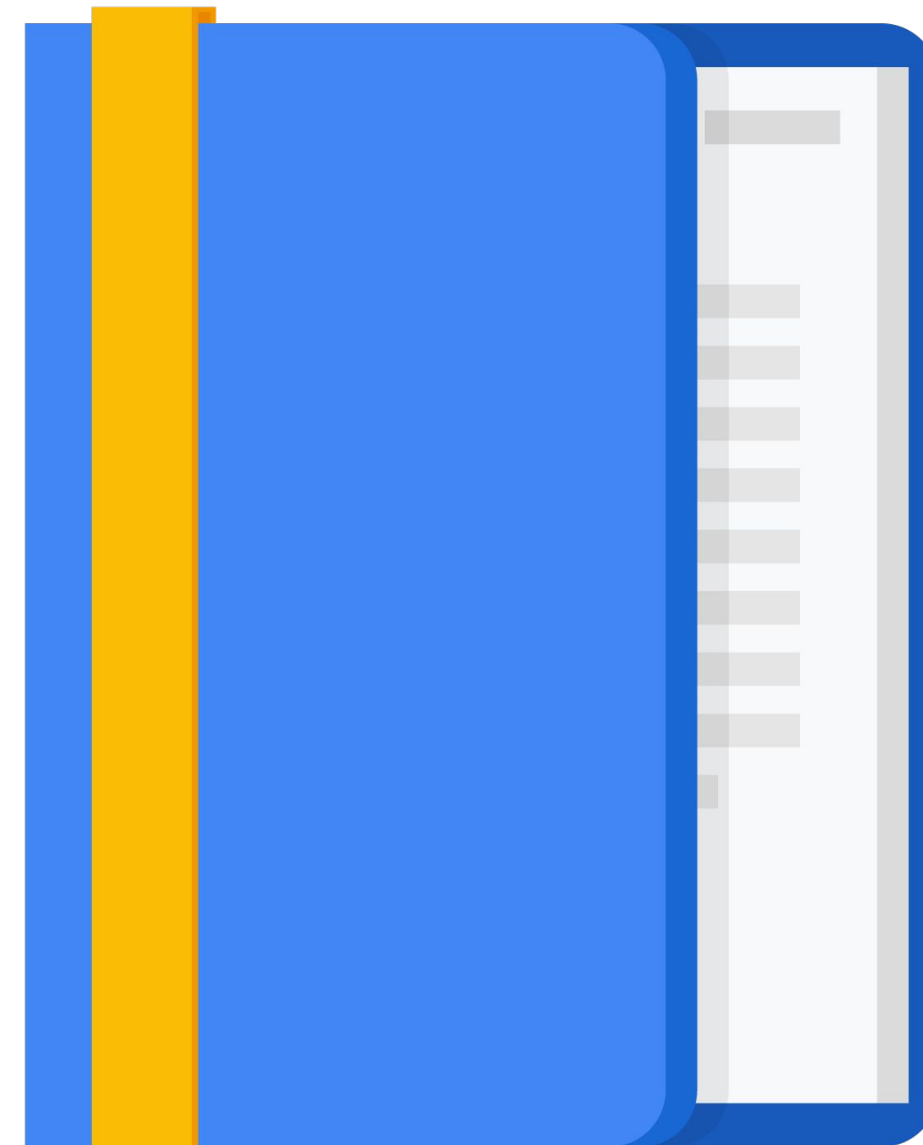


Big Data Analytics with Cloud AI Platform Notebooks

Agenda

What's a Notebook

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

The screenshot shows a Jupyter Notebook interface with a sidebar on the left and a main content area on the right. The sidebar contains a file browser, a share icon, and a vertical toolbar with icons for file operations, code execution, and output. The main content area displays a notebook titled 'Untitled.ipynb' with a menu bar (File, Edit, View, Run, Kernel, Git, Tabs, Settings, Help) and a toolbar (New, Open, Save, Copy, Paste, Run, Undo, Redo, Close, Git). The notebook content includes a code cell with the following code:

```
[2]: %matplotlib inline
import matplotlib.pyplot as plt
import numpy as np
```

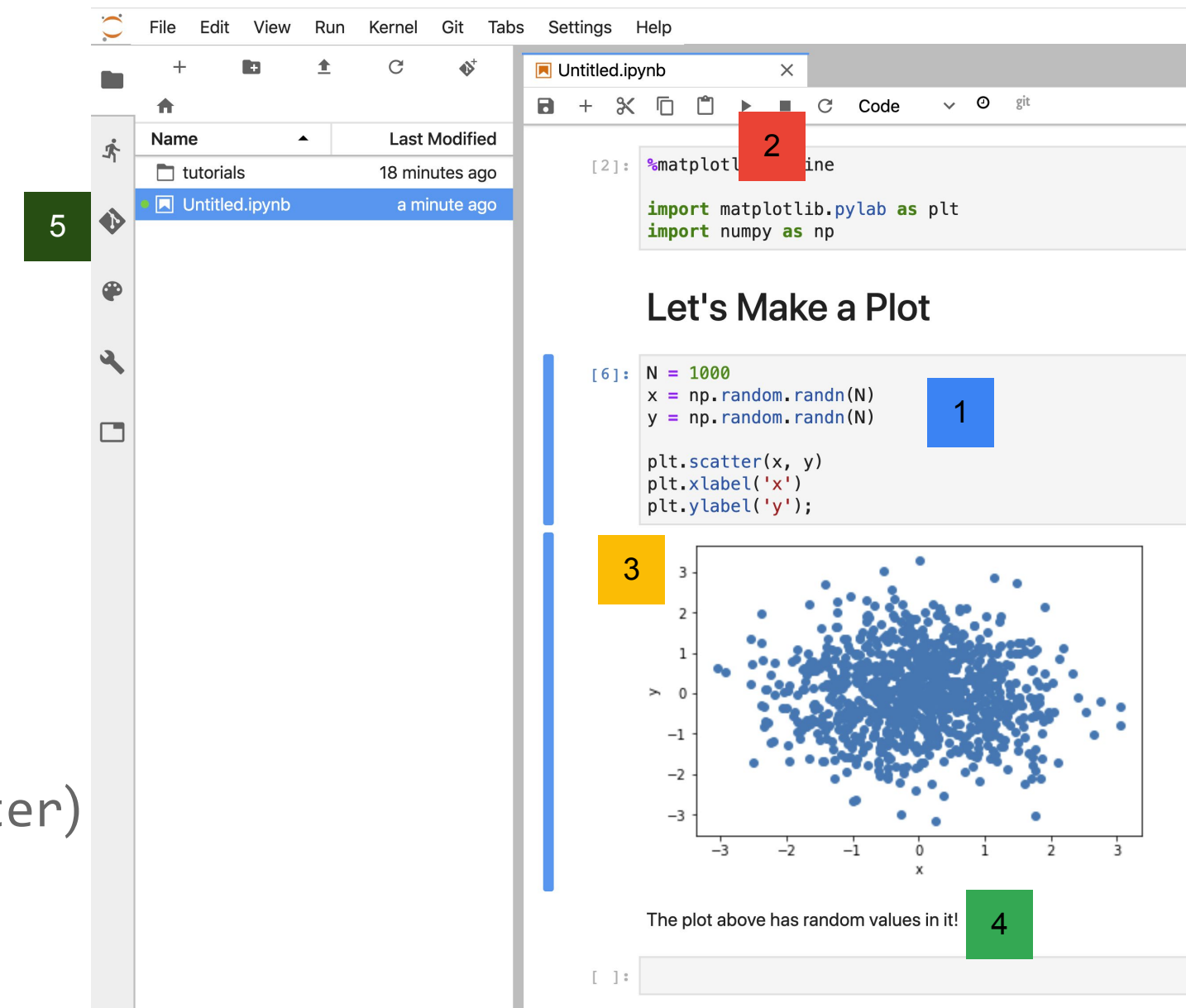
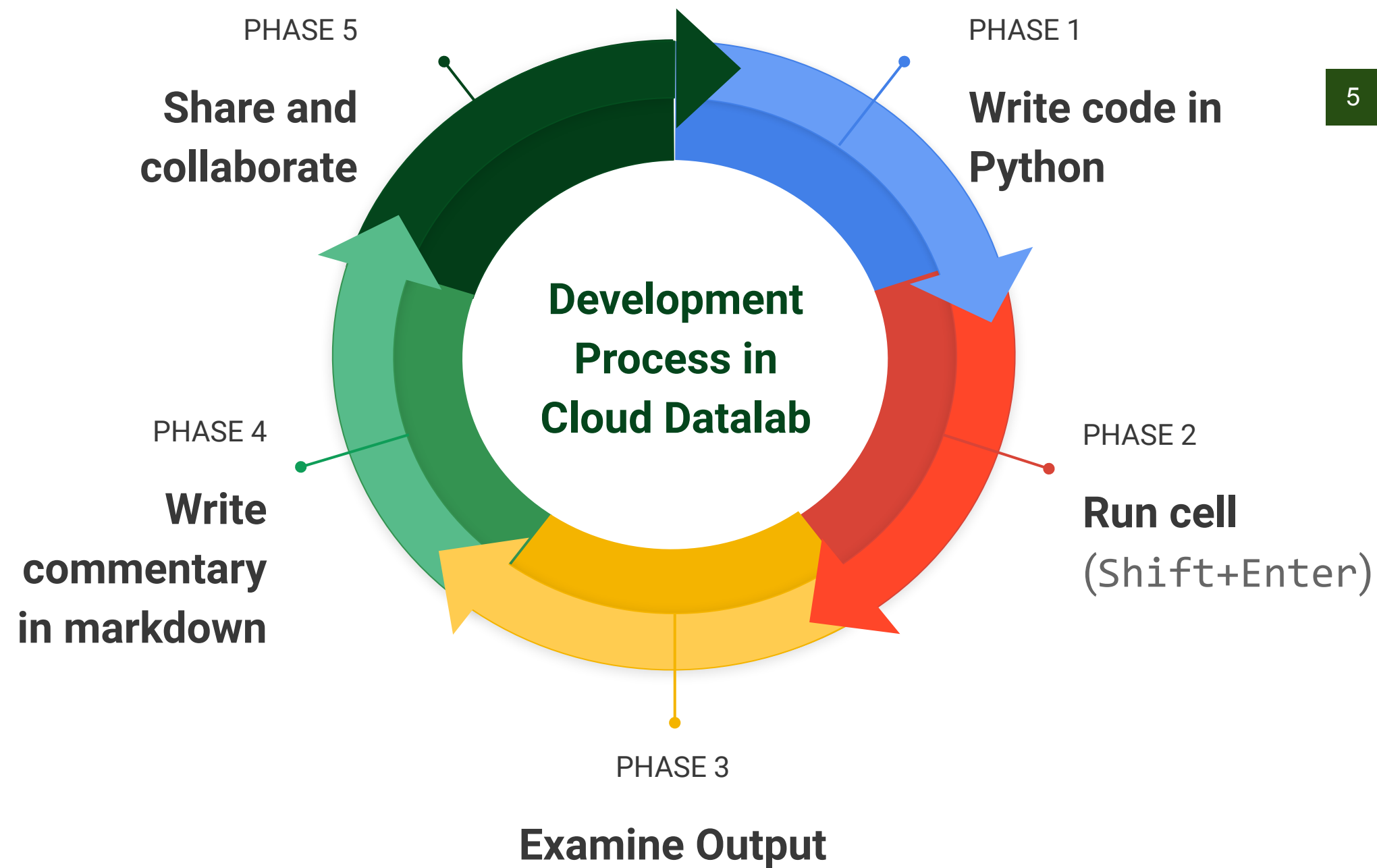
Below the code cell is a text cell with the heading 'Let's Make a Plot'. This is followed by another code cell:

```
[6]: N = 1000
x = np.random.randn(N)
y = np.random.randn(N)

plt.scatter(x, y)
plt.xlabel('x')
plt.ylabel('y');
```

The output of the code cell is a scatter plot showing a dense cluster of blue points centered around the origin (0,0). The x-axis is labeled 'x' and ranges from -3 to 3. The y-axis is labeled 'y' and ranges from -3 to 3. Below the plot is a text cell containing the explanation: 'The plot above has random values in it!'. The sidebar on the left has four green arrows pointing to specific elements: 'Share' points to the share icon, 'Code' points to the code execution icon, 'Output' points to the output icon, and 'Markdown' points to the text cell below the plot.

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.

Google Cloud Platform

Project

ML Engine

Notebook instances

Jobs

Models

Notebook instances are Compute Engine VM instances pre-configured with machine learning frameworks. It comes with access to the VM instance.

Filter

Instance name

Region

ML framework

Machine type

GPUs

Labels

No notebook runtimes to display

+ NEW INSTANCE

EDIT

START

STOP

RESET

DELETE

SHOW INFO PANEL

TensorFlow

PyTorch

More options

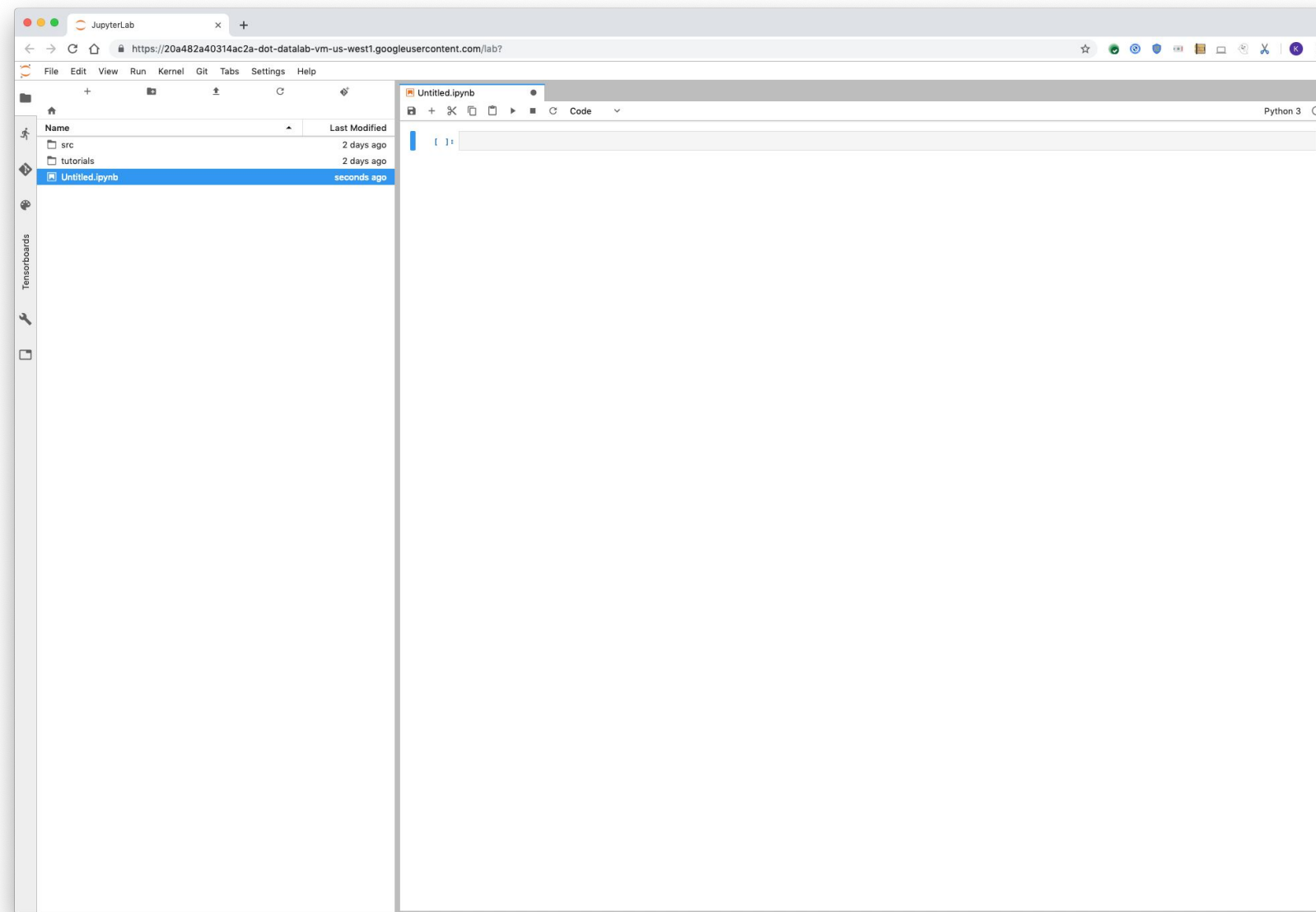
Standard

us-east1-b, 4 vCPUs, 15 GB Memory, 100 GB disk

With GPU

us-east1-b, 4 vCPUs, 15 GB Memory, 1 NVIDIA Tesla K80, 100 GB disk

AI Platform Notebooks uses the latest open-source version of the industry-standard JupyterLab



You can easily change hardware including adding and removing GPUs

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	OPEN JUPYTERLAB	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>	

Use any GCE instance type. You can pick the hardware that makes sense, and scale up or down as needed

Google Cloud Platform

Project

Edit notebook instance

Instance name: [ml-notebook-runtime-1](#)

Region: us-east1 (South Carolina)

Zone: us-east1-b

ML framework: TensorFlow 1.12

Machine type *
4 vCPUs, 15 GB Memory

\$28.27 per month estimated
Effective hourly rate \$0.039 (730 hours per month)

[Details](#)

GPUs

The number of GPU dies is linked to the number of CPU cores and memory selected for this instance. For this machine type, you can select no fewer than 1 GPU die. [Learn more](#)

Number of GPUs: None

GPU type: NVIDIA Tesla K80

Machines with GPUs can't migrate on host maintenance

Boot disk

Boot disk type: Standard Persistent Disk

Boot disk size in GB: 100

SAVE CANCEL

You can even add and remove GPUs

caip-dexter-bugbash

Notebook instances BETA

+ NEW INSTANCE

REFRESH

START

STOP

RESET

DELETE

SHOW INFO PANEL

Filter table

<input type="checkbox"/>	<input type="radio"/>	Instance name		Region	ML framework	Machine type	GPUs	Labels
<input type="checkbox"/>	<input type="radio"/>	lquera-dexter-tf	OPEN JUPYTERLAB	us-west1-a	TensorFlow	4 vCPUs, 15 GB RAM	None	No labels
<input type="checkbox"/>	<input type="radio"/>	rpasricha-dexter	OPEN JUPYTERLAB	us-west1-a	TensorFlow	4 vCPUs, 15 GB RAM	None	
<input type="checkbox"/>	<input type="radio"/>	mgorner	OPEN JUPYTERLAB	us-west1-b	TensorFlow	4 vCPUs, 15 GB RAM	None	
<input type="checkbox"/>	<input checked="" type="radio"/>	sivaibhav-dexter	OPEN JUPYTERLAB	us-west1-b	TensorFlow	4 vCPUs, 15 GB RAM	None	
<input type="checkbox"/>	<input checked="" type="radio"/>	ramachandrank-dexter	OPEN JUPYTERLAB	us-east1-b	TensorFlow	4 vCPUs, 15 GB RAM	None	
<input type="checkbox"/>	<input checked="" type="radio"/>	sunparty-dexter-test	OPEN JUPYTERLAB	us-east1-c	TensorFlow	4 vCPUs, 15 GB RAM	None	

None

NVIDIA Tesla P100

NVIDIA Tesla P100 Virtual Workstation

NVIDIA Tesla T4

NVIDIA Tesla T4 Virtual Workstation

NVIDIA Tesla V100

Notebook instances are standard GCE instances that live in your projects

Google Cloud Platform

ramachandr...

Compute Engine

VM instances

Instance groups

Instance templates

Sole tenant nodes

Disks

Snapshots

Images

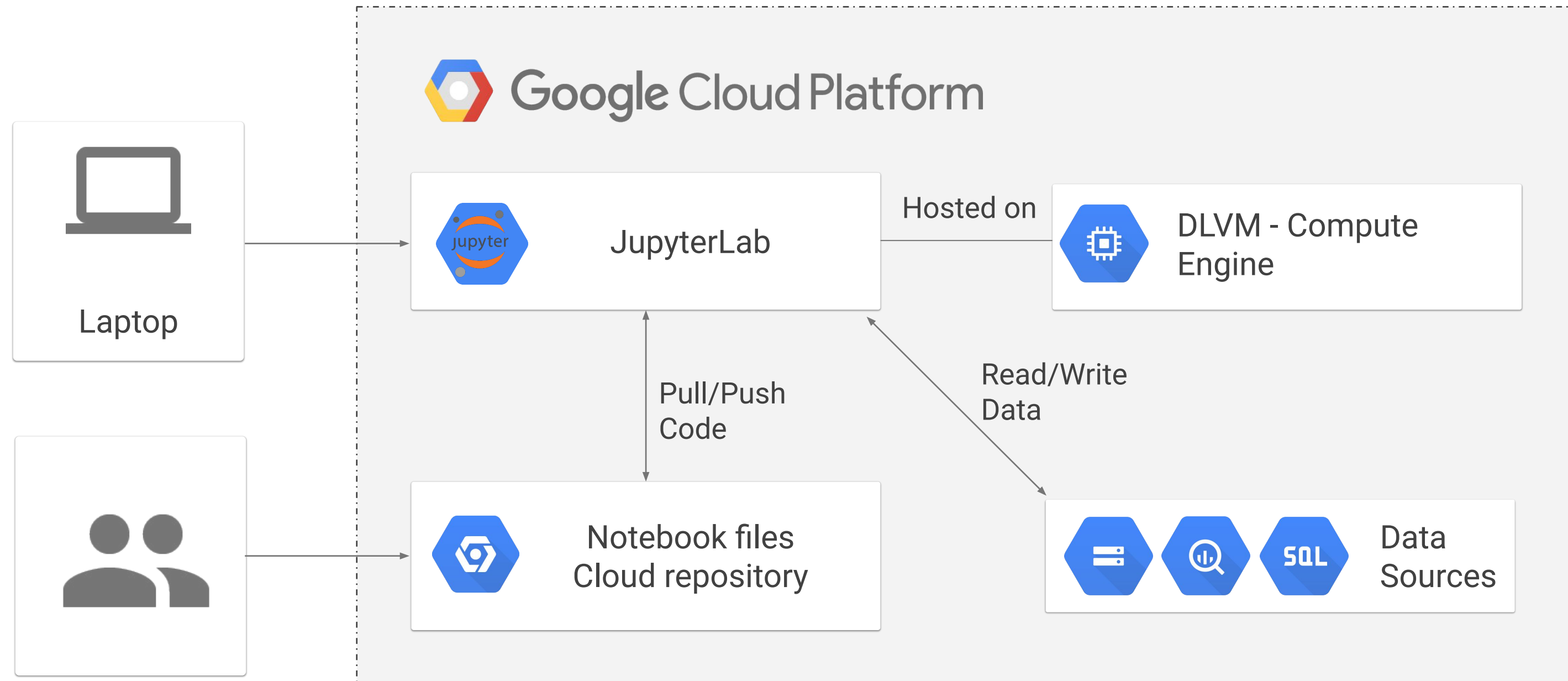
VM instances

CREATE INSTANCEIMPORT VM

Filter VM instances

Name	Zone	Recommendation	Internal IP	External IP	Connect
pytorch-2-vm	us-west1-b		10.138.0.2 (nic0)	35.185.251.12	SSH
tensorflow-1549311103062	us-central1-c	Save \$74 / mo	10.128.0.5 (nic0)	35.226.219.116	SSH

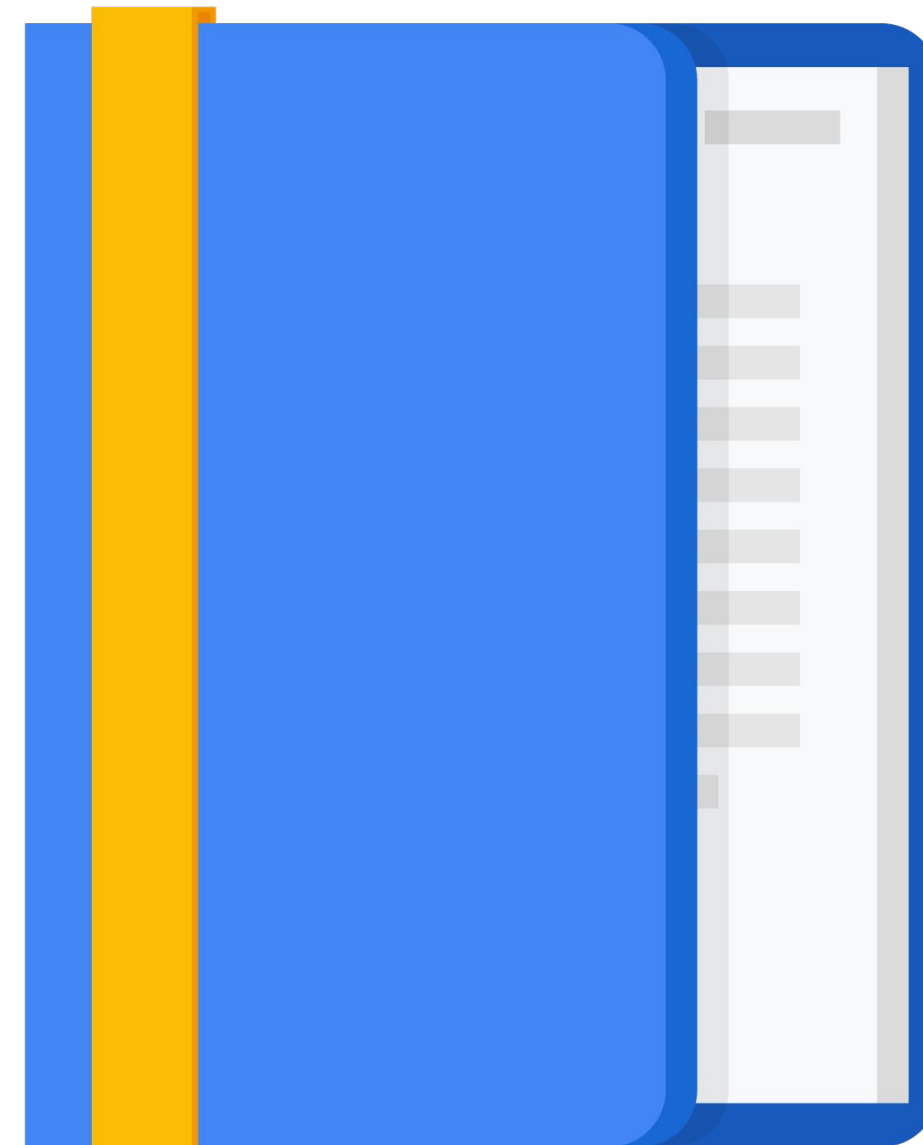
How does it work?



Agenda

What's a Notebook

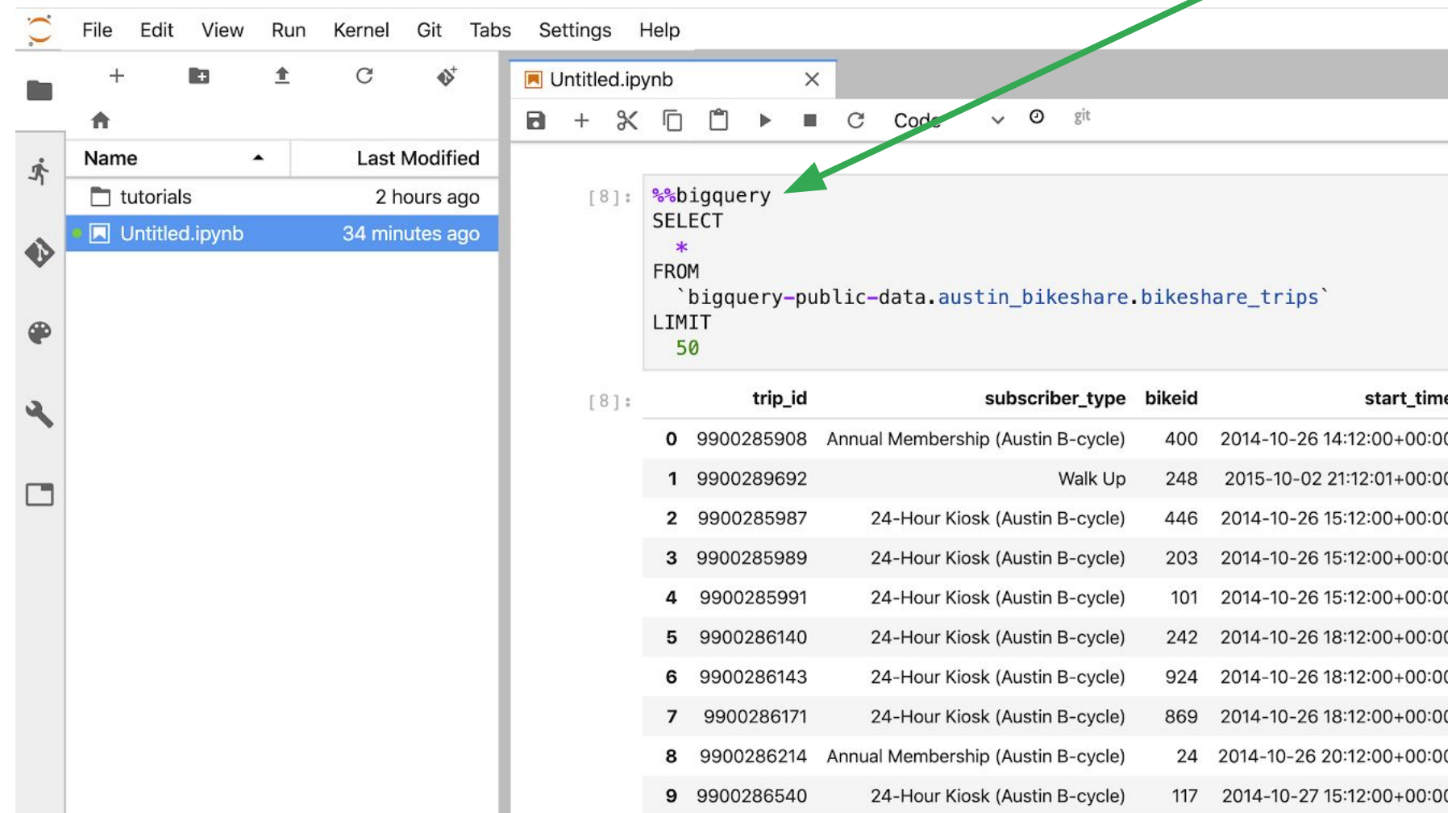
BigQuery Magic and Ties to
Pandas



You can execute BigQuery commands from AI Platform Notebooks

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

Jupyter “magic” function



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

	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]:
```

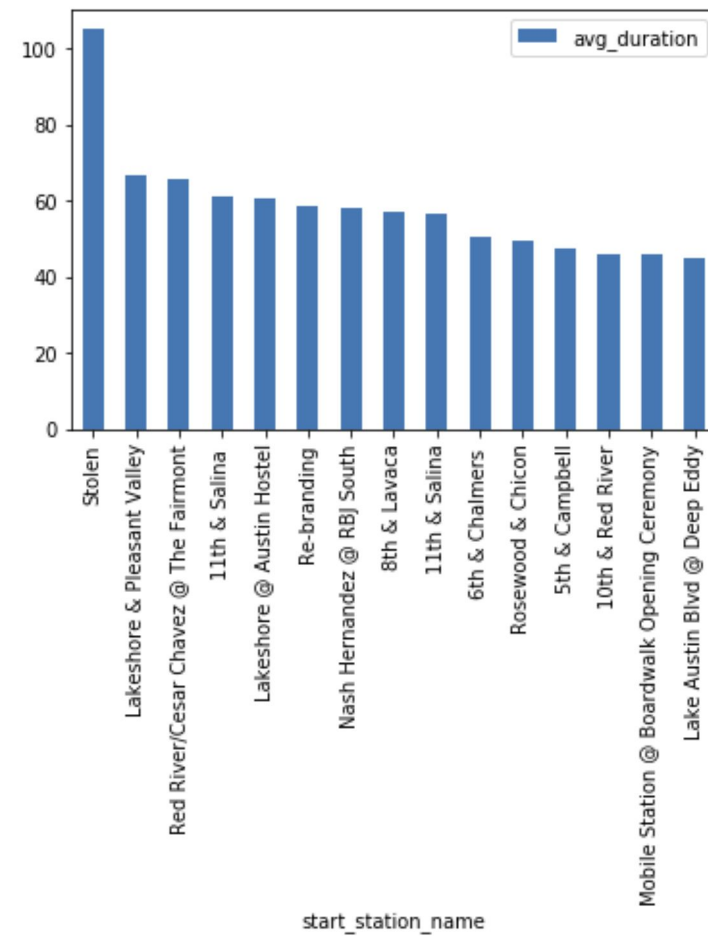
	trip_id	subscriber_type	bikeid	start_time	start_station_id	start_station_name	end_station_id	end_station_name	duration_minutes
0	9900259257	Walk Up	93	2015-09-18 08:12:05+00:00	2712	Toomey Rd @ South Lamar	None	Stolen	2863
1	16898448	Walk Up	1857	2018-03-18 22:51:20+00:00	2501	5th & Bowie	None	Stolen	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');
```





BigQuery in Jupyter Labs on AI Platform

Objectives

- Instantiate a Jupyter notebook on AI Platform
- Execute a BigQuery query from within a Jupyter notebook and process the output using Pandas

Module Summary

- AI Platform Notebooks are ideal for prototyping machine learning pipelines and models
- Notebooks integrate nicely with BigQuery and other GCP services