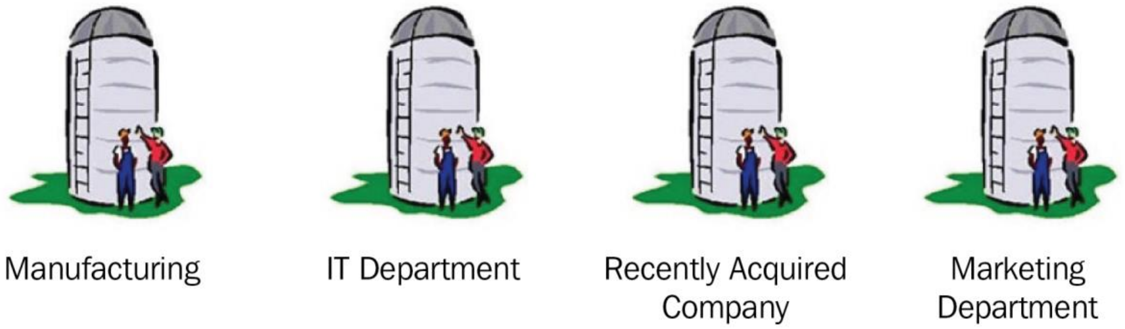
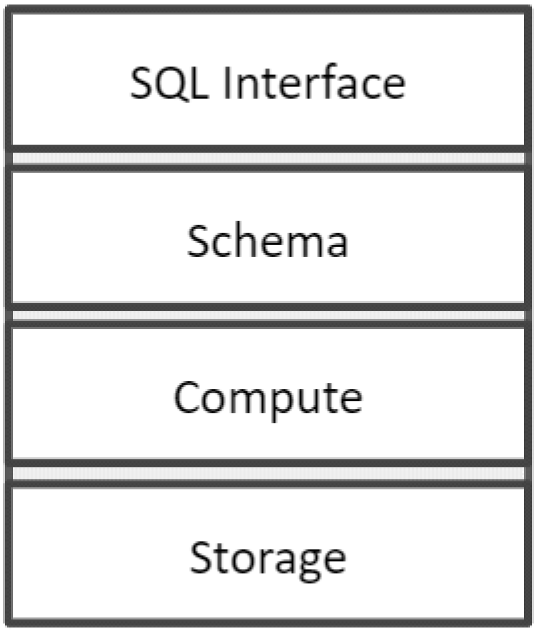
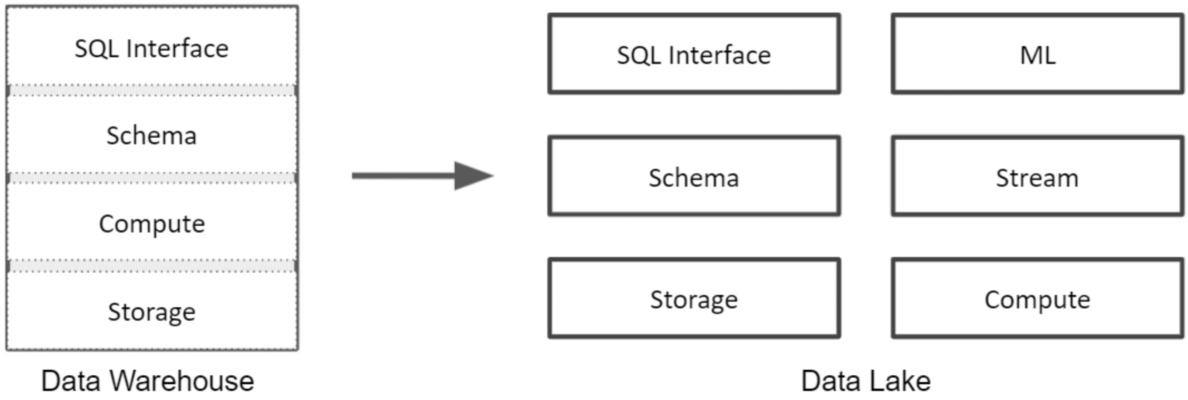
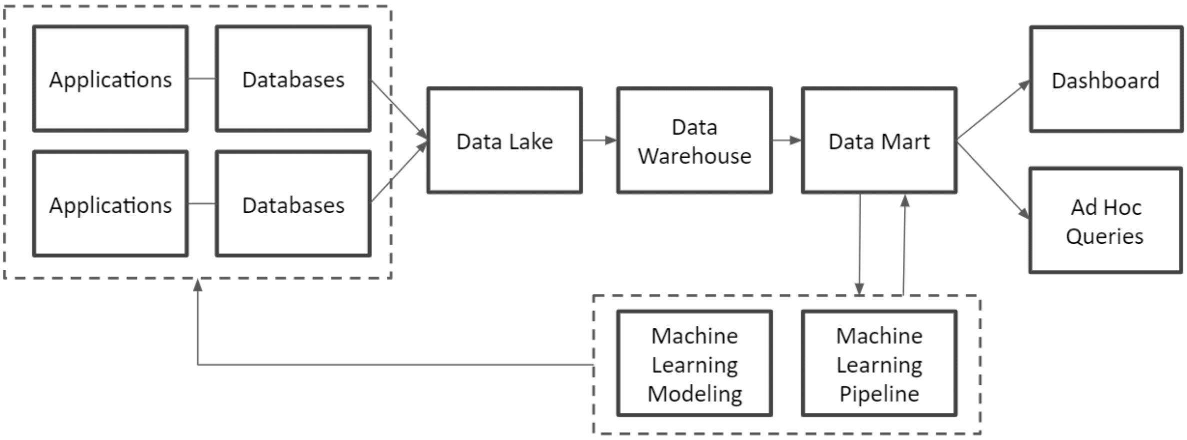
Chapter 1: Fundamentals of Data Engineering

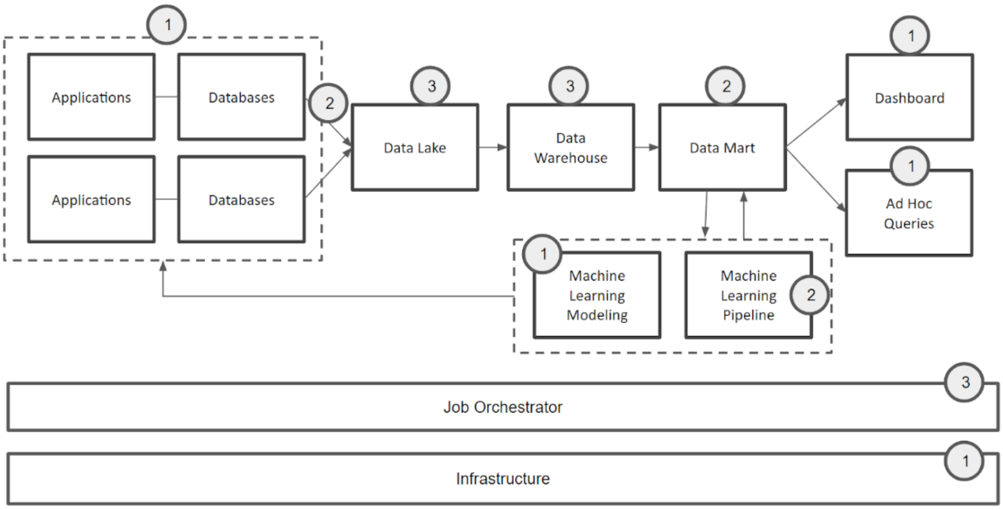


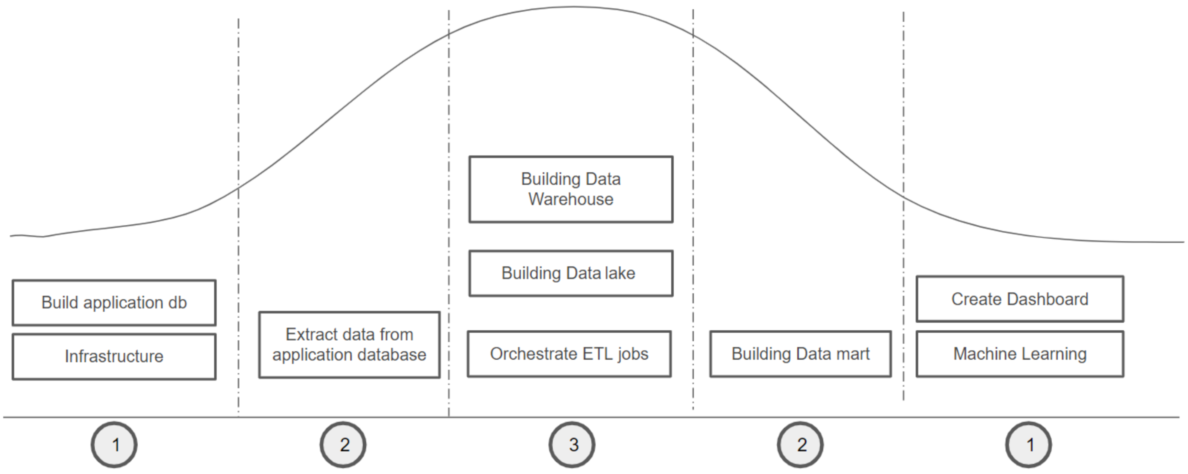


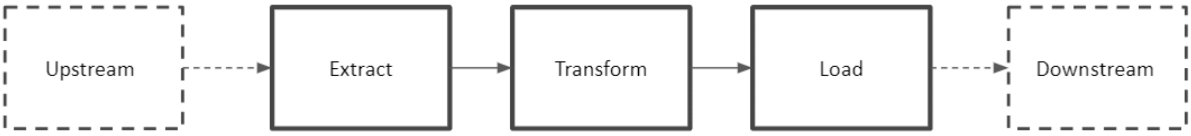


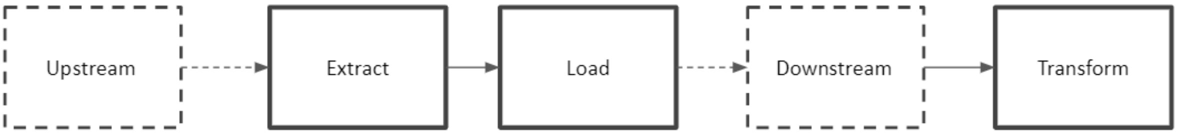
Data Lake	Data Warehouse
Schema is not mandatory	Schema is mandatory
Possibility to compute using different technologies for the same underlying storage	With all access using SQL, the developer doesn’t have control over how to compute and store the data
First focus is to store as much data as possible. Business relevancy and data model are defined later	First focus is business relevancy and data models. Only store data based on the business needs

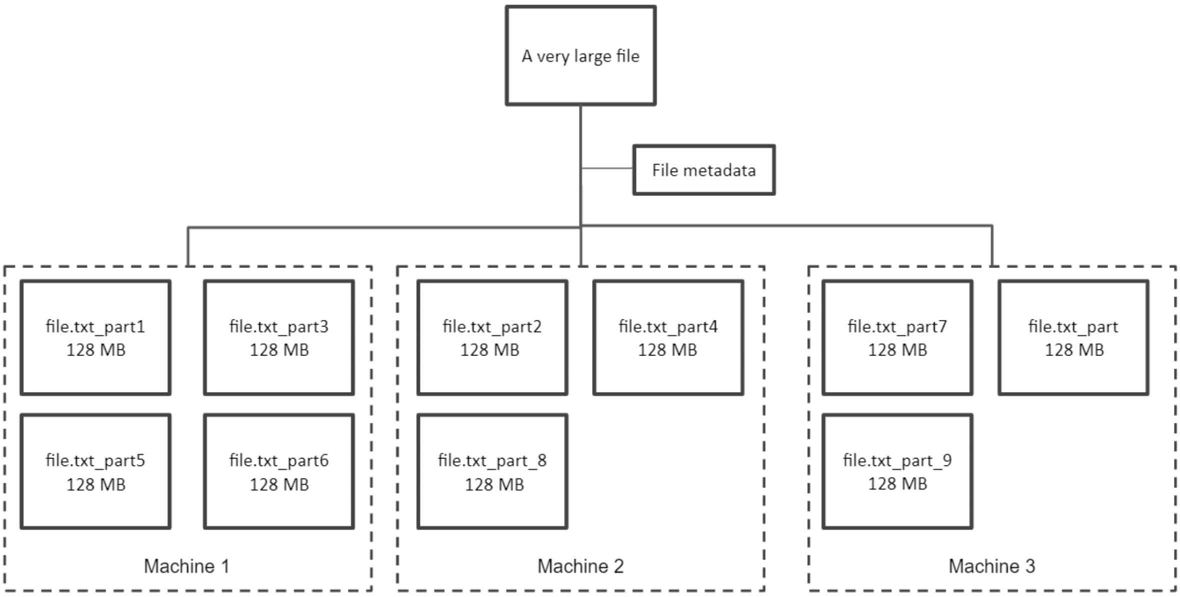


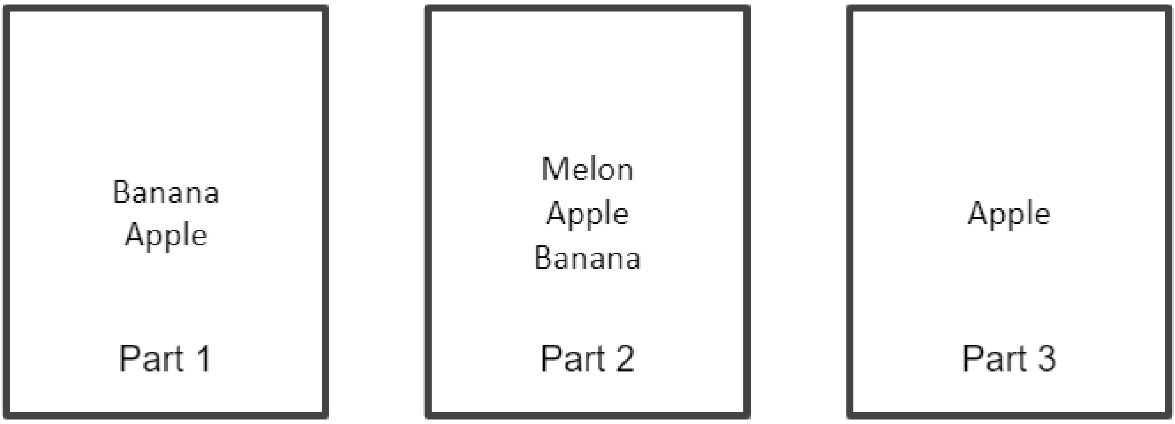


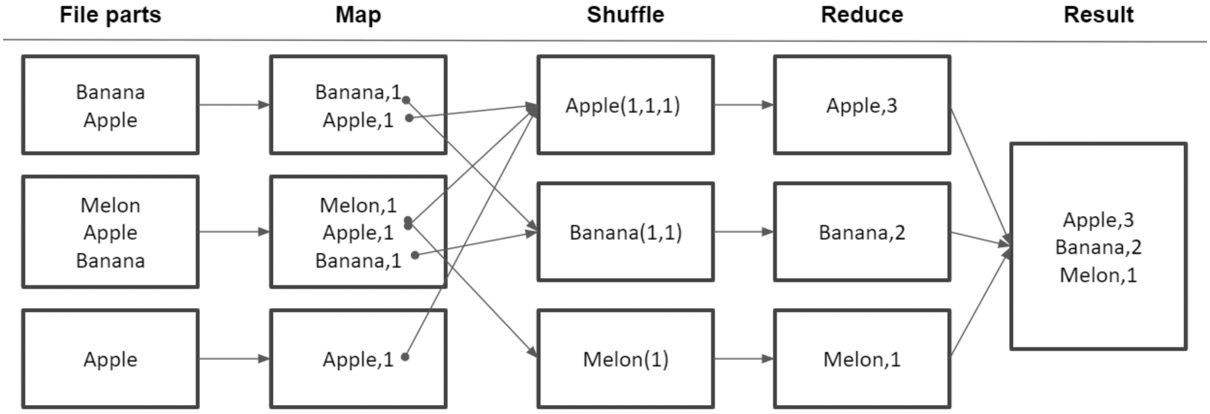

Building Data Warehouse
Building Data lake
Build application db
Extract data from application database
Orchestrate ETL jobs
Building Data mart
Create Dashboard
Infrastructure
Machine Learning
1
2
3
2
1


Upstream
Extract
Transform
Load
Downstream

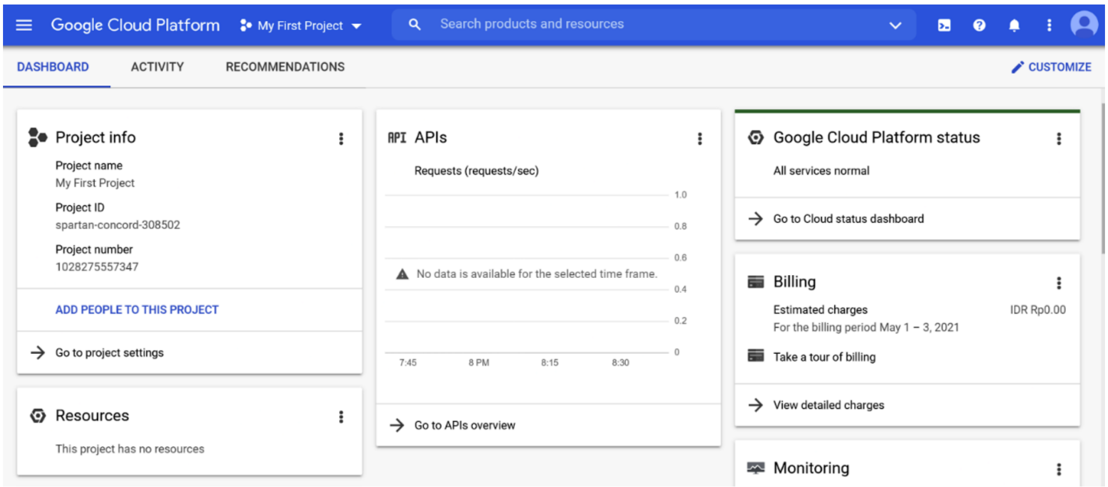

Upstream
Extract
Load
Downstream
Transform


A very large file
File metadata
file.txt_part1 128 MB
file.txt_part3 128 MB
file.txt_part5 128 MB
file.txt_part6 128 MB
Machine 1
file.txt_part2 128 MB
file.txt_part4 128 MB
file.txt_part_8 128 MB
Machine 2
file.txt_part7 128 MB
file.txt_part 128 MB
file.txt_part_9 128 MB
Machine 3


Banana
Apple
Part 1
Melon
Apple
Banana
Part 2
Apple
Part 3


File parts
Map
Shuffle
Reduce
Result
Banana
Apple
Melon
Apple
Banana
Apple
Banana,1
Apple,1
Melon,1
Apple,1
Banana,1
Apple,1
Apple(1,1,1)
Banana(1,1)
Melon(1)
Apple,3
Banana,2
Melon,1
Apple,3
Banana,2
Melon,1

Chapter 2: Big Data Capabilities on GCP



Project name
My First Project

Project ID
spartan-concord-308502

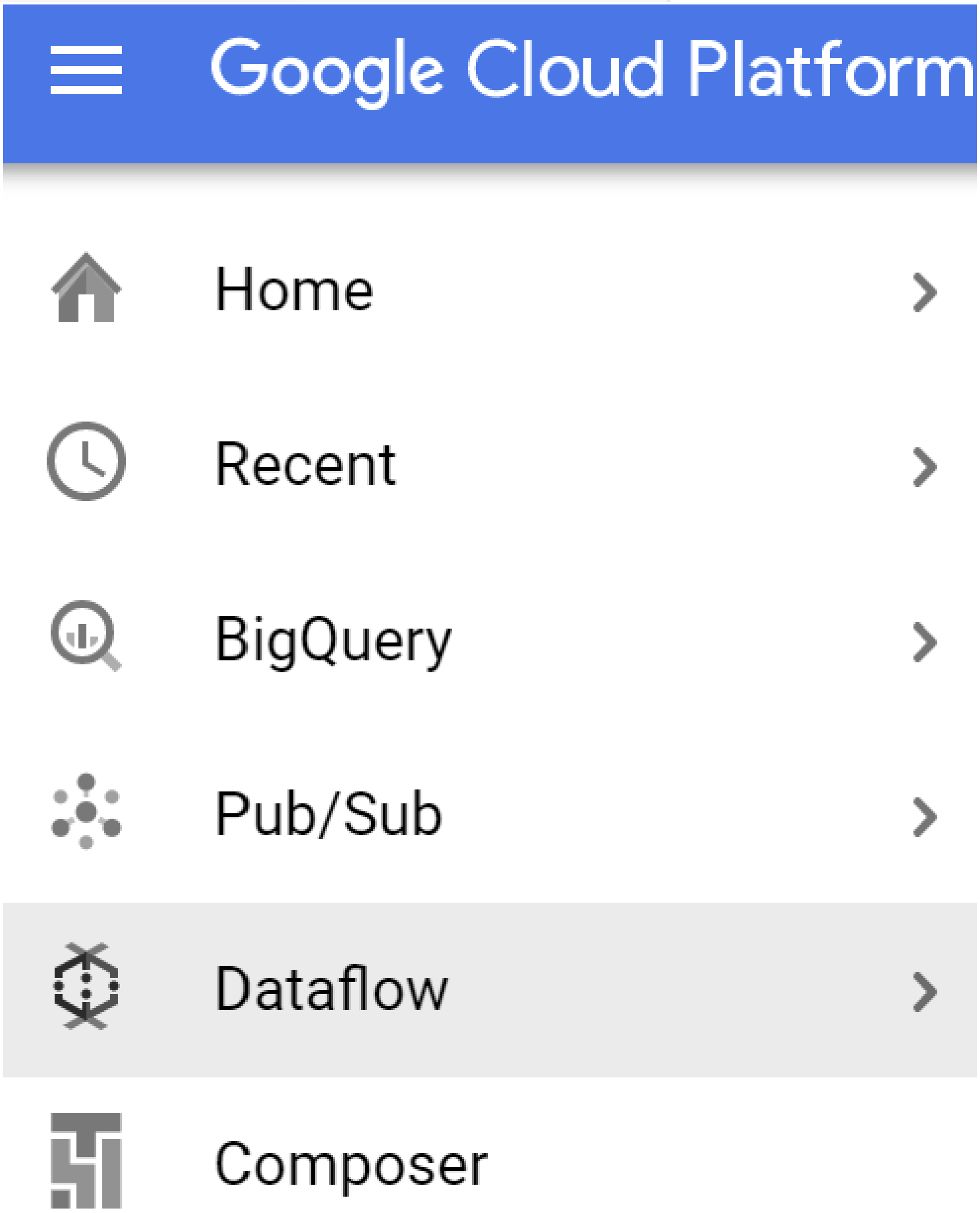
Project number
1028275557347

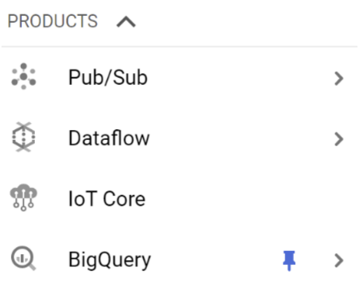
This project has no resources

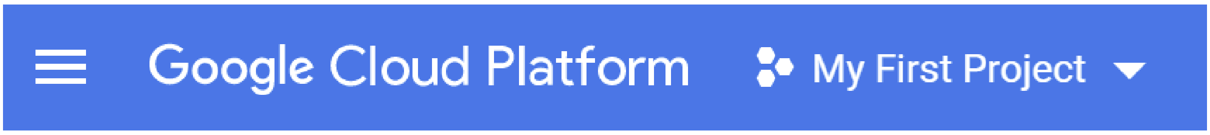
No data is available for the selected time frame.

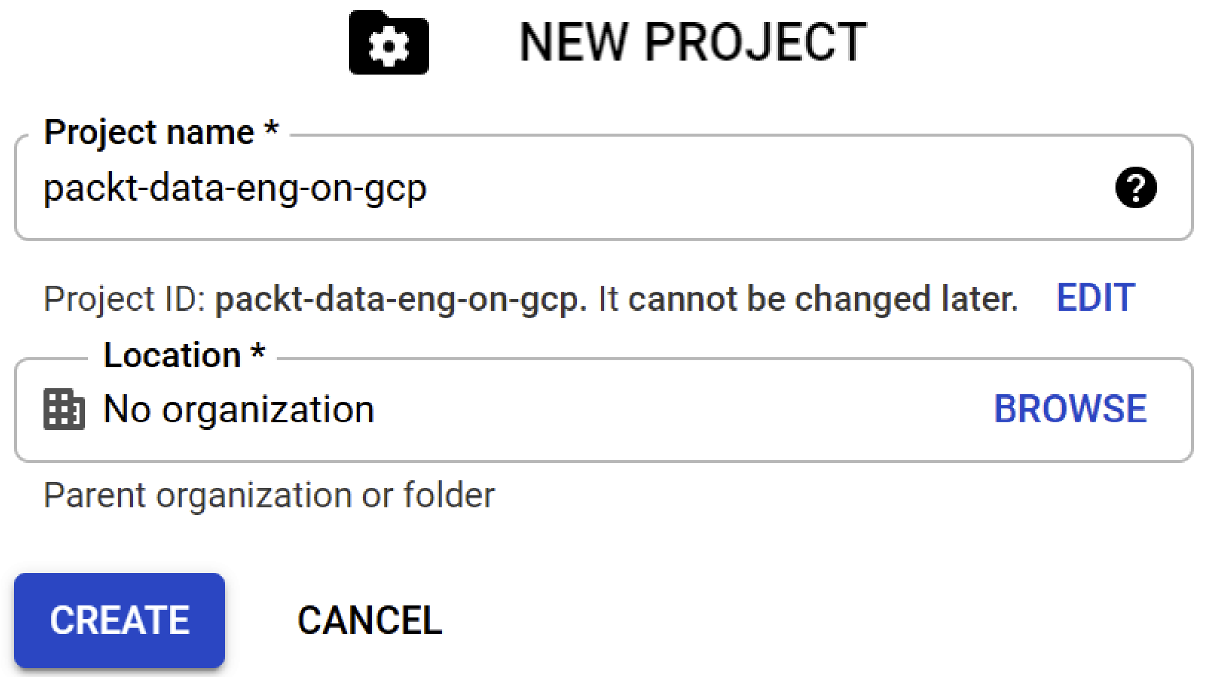
All services normal

Estimated charges IDR Rp0.00
For the billing period May 1 – 3, 2021







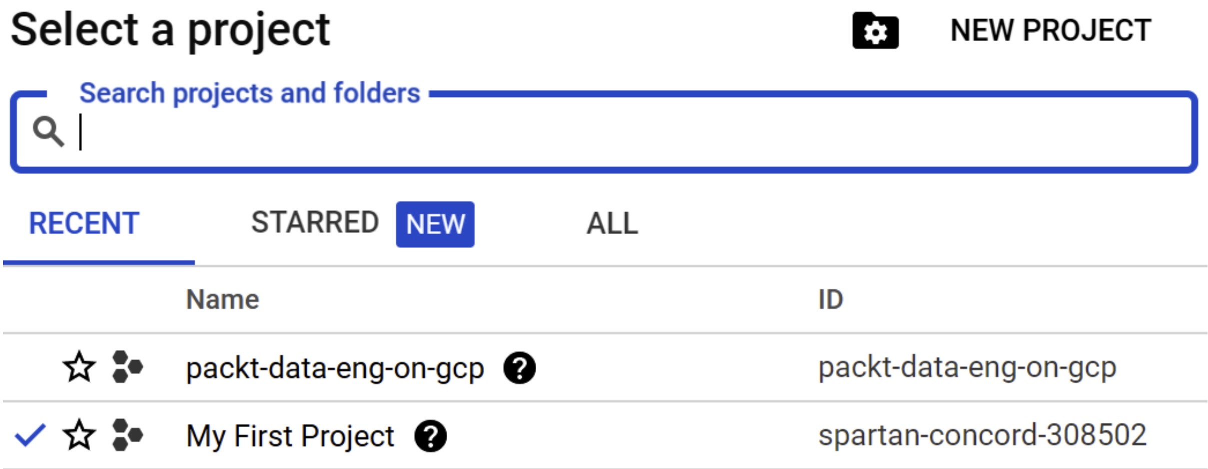


Project name *
packt-data-eng-on-gcp

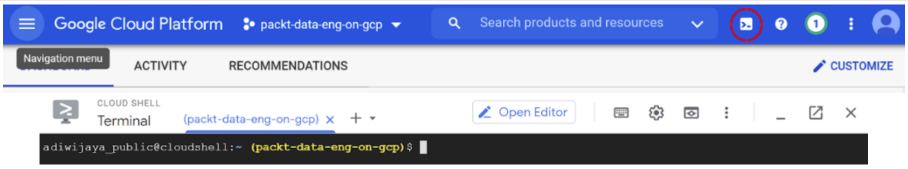
Project ID: packt-data-eng-on-gcp. It cannot be changed later.

Location *
No organization

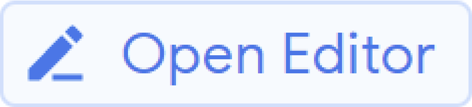
Parent organization or folder

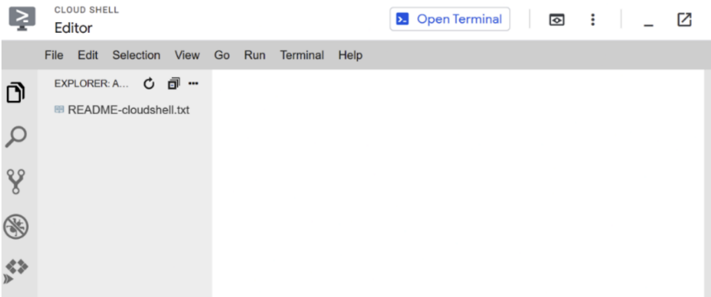


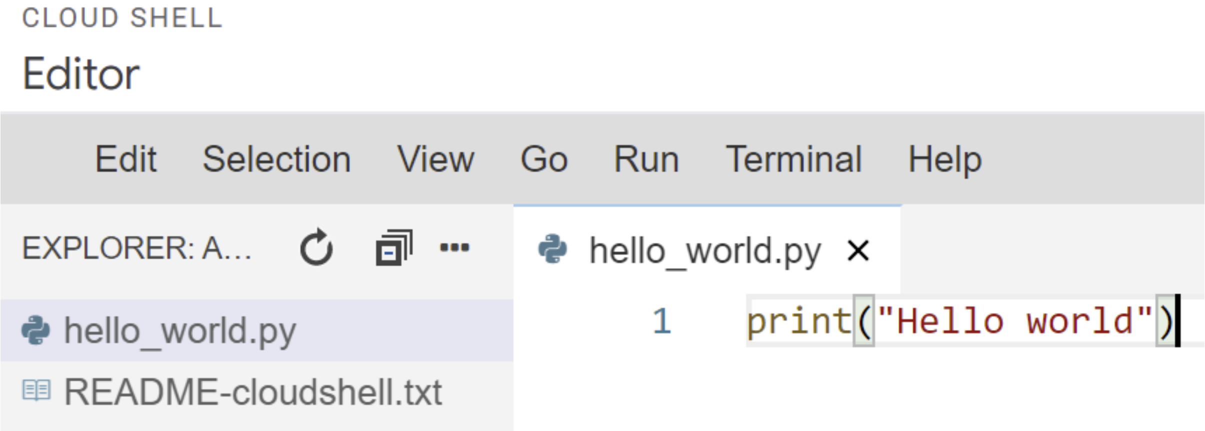
Name	ID
packt-data-eng-on-gcp	packt-data-eng-on-gcp
✓ My First Project	spartan-concord-308502



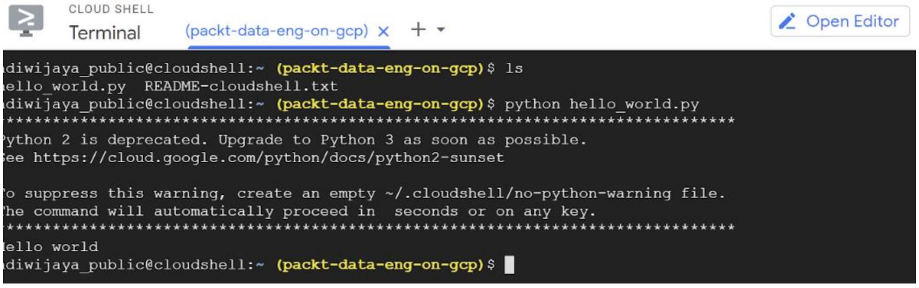
```
adiwijaya_public@cloudshell:~ (packt-data-eng-on-gcp)$
```





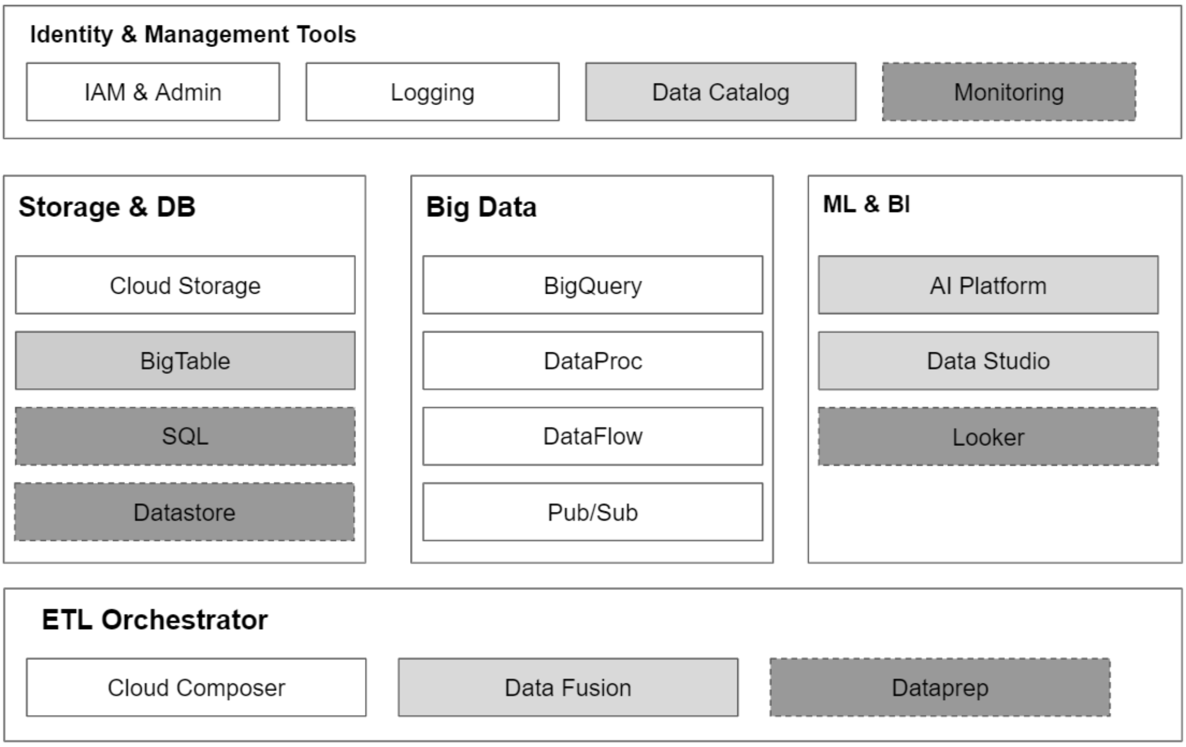


```python
print("Hello world")
```

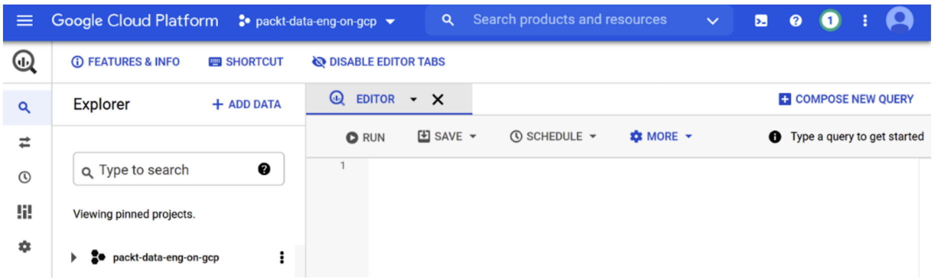


```
diwijaya_public@cloudshell:~ (packt-data-eng-on-gcp)$ ls
ello_world.py  README-cloudshell.txt
diwijaya_public@cloudshell:~ (packt-data-eng-on-gcp)$ python hello_world.py
*******************************************************************************
ython 2 is deprecated. Upgrade to Python 3 as soon as possible.
ee https://cloud.google.com/python/docs/python2-sunset

o suppress this warning, create an empty ~/.cloudshell/no-python-warning file.
he command will automatically proceed in  seconds or on any key.
*******************************************************************************
ello world
diwijaya_public@cloudshell:~ (packt-data-eng-on-gcp)$ 
```

	Manage physical infrastructure	Manage virtual machines	Manage application service	Develop solution on top of the service
On-premises (non-cloud)	O	O	O	O
VM-based	X	O	O	O
Managed service	X	X	O	O
Fully managed service	X	X	X	O

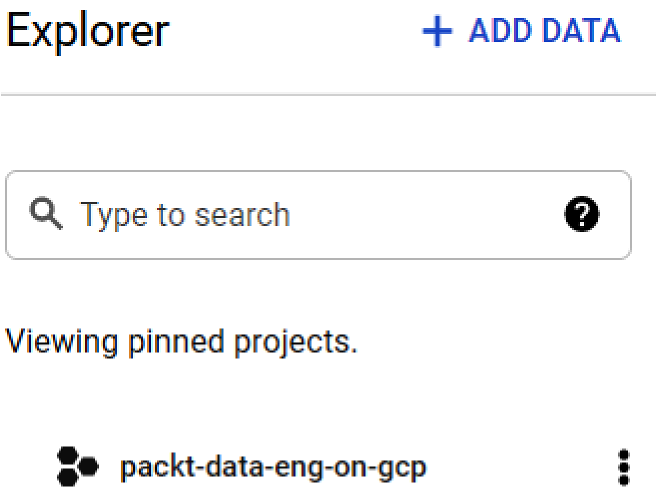


Chapter 3: Building a Data Warehouse in BigQuery

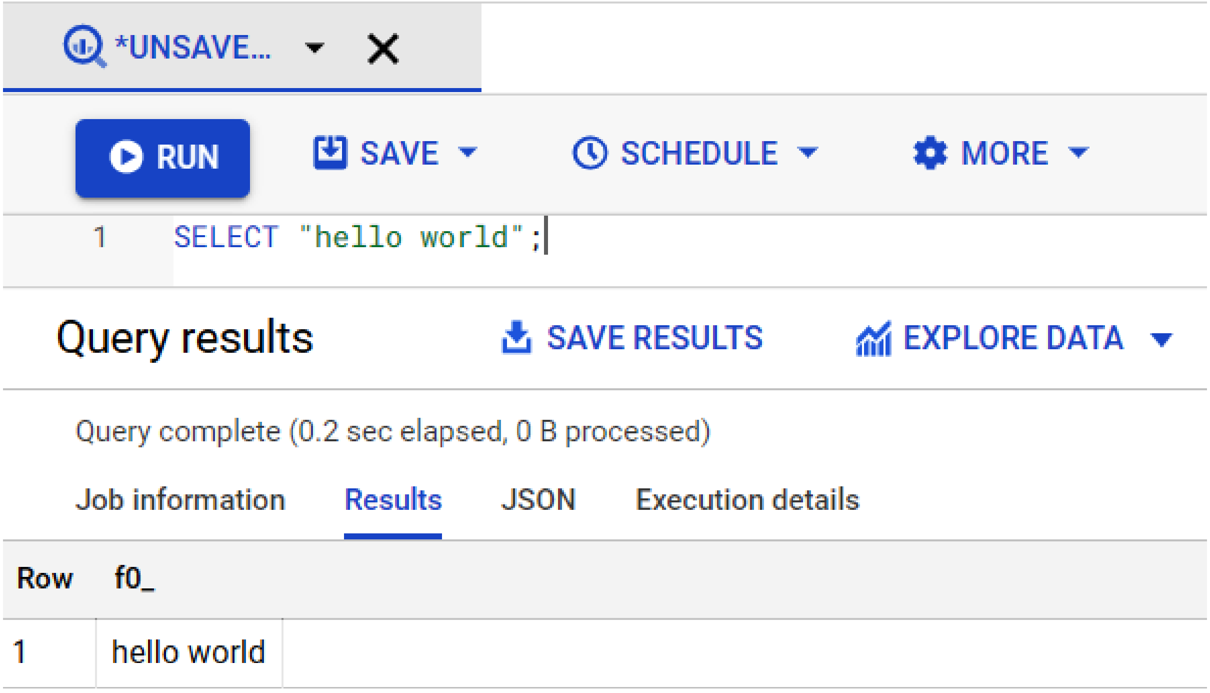


Type a query to get started

Viewing pinned projects.



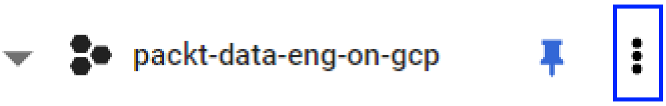
Viewing pinned projects.

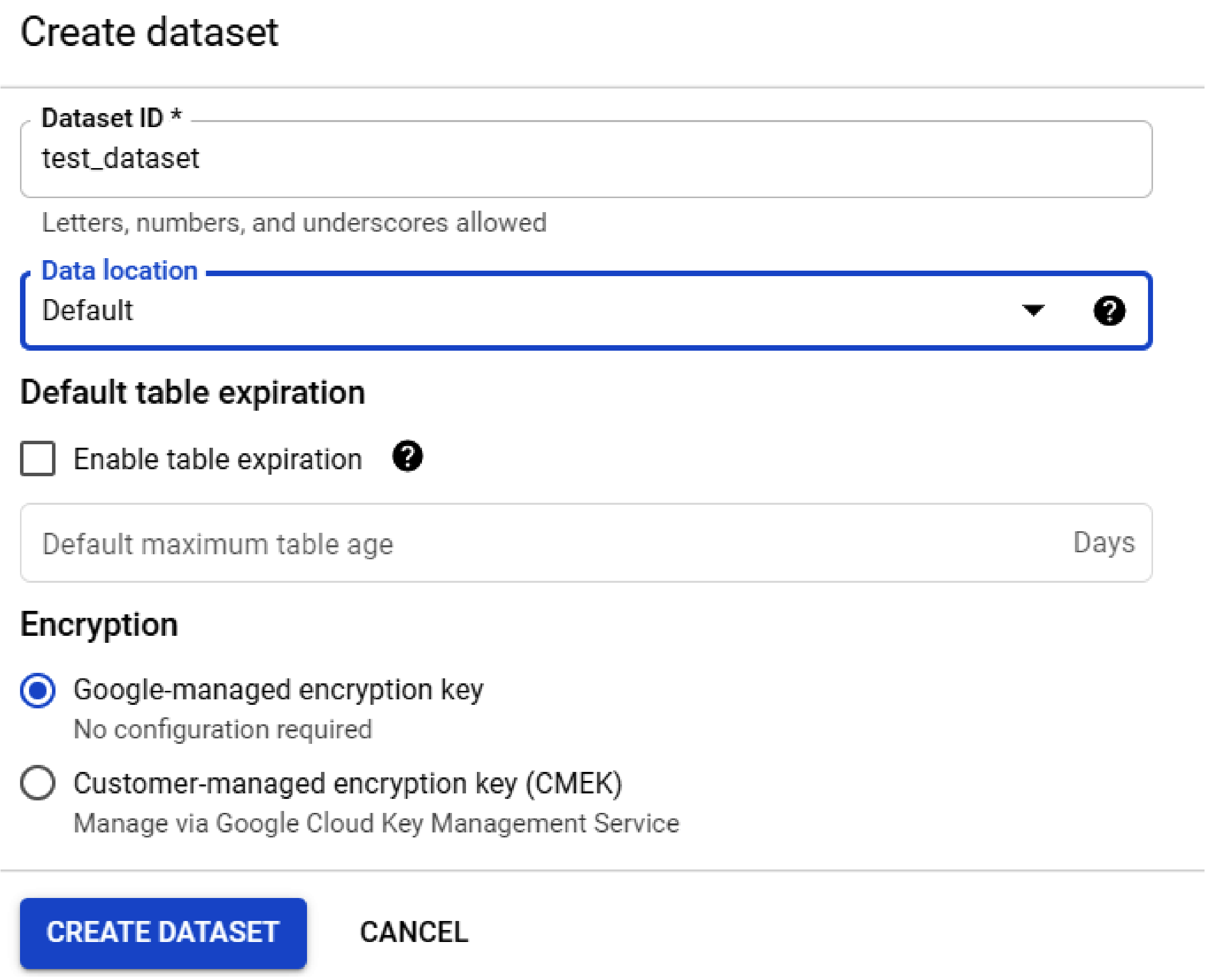


```
SELECT "hello world";
```

Query complete (0.2 sec elapsed, 0 B processed)

Row	f0_
1	hello world





Create dataset

Dataset ID * test_dataset

Letters, numbers, and underscores allowed

Data location: Default

Default table expiration

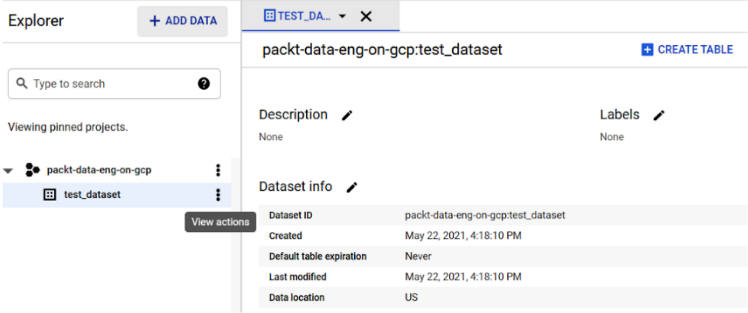
□ Enable table expiration

Default maximum table age — Days

Encryption

● Google-managed encryption key
No configuration required

○ Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service



Explorer

Type to search

Viewing pinned projects.

packt-data-eng-on-gcp
test_dataset

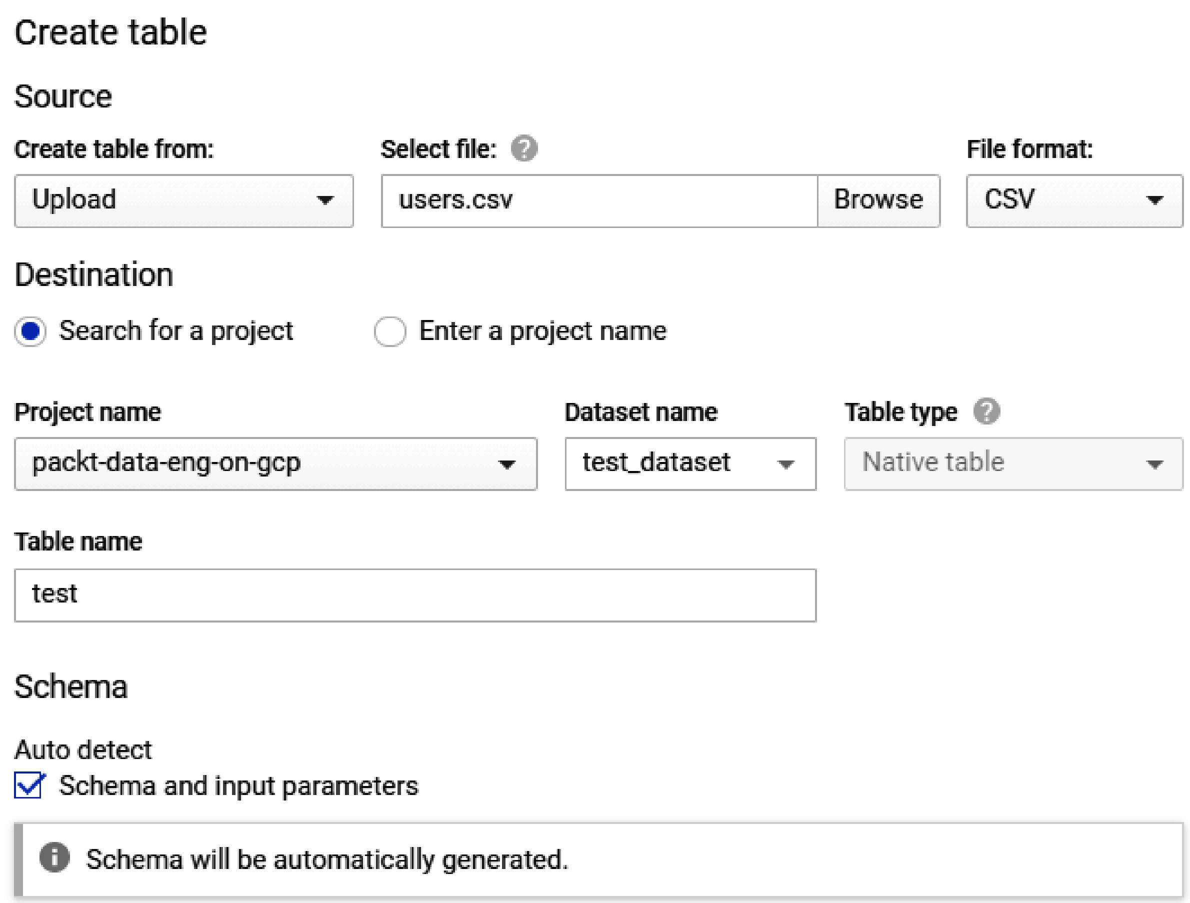
packt-data-eng-on-gcp:test_dataset

Description: None

Labels: None

Dataset info

Dataset ID	packt-data-eng-on-gcp:test_dataset
Created	May 22, 2021, 4:18:10 PM
Default table expiration	Never
Last modified	May 22, 2021, 4:18:10 PM
Data location	US



Create table

Source

Create table from: Upload

Select file: users.csv

File format: CSV

Destination

● Search for a project ○ Enter a project name

Project name: packt-data-eng-on-gcp

Dataset name: test_dataset

Table type: Native table

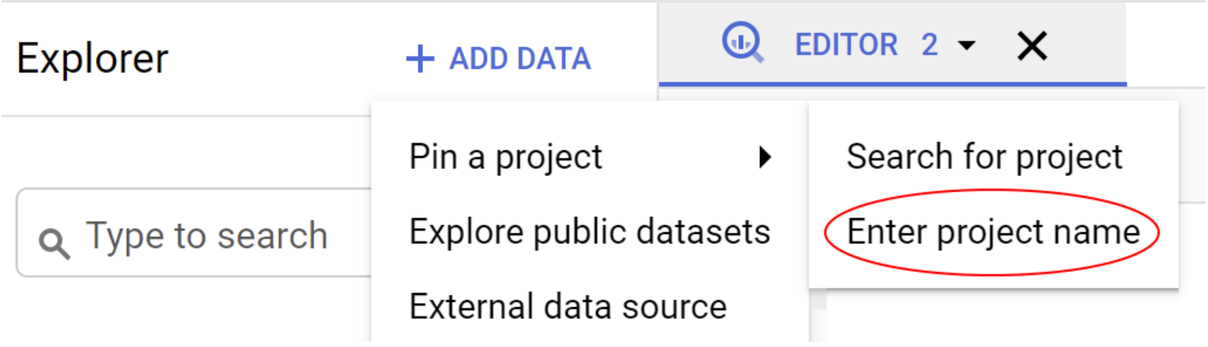
Table name: test

Schema

Auto detect

☑ Schema and input parameters

Schema will be automatically generated.



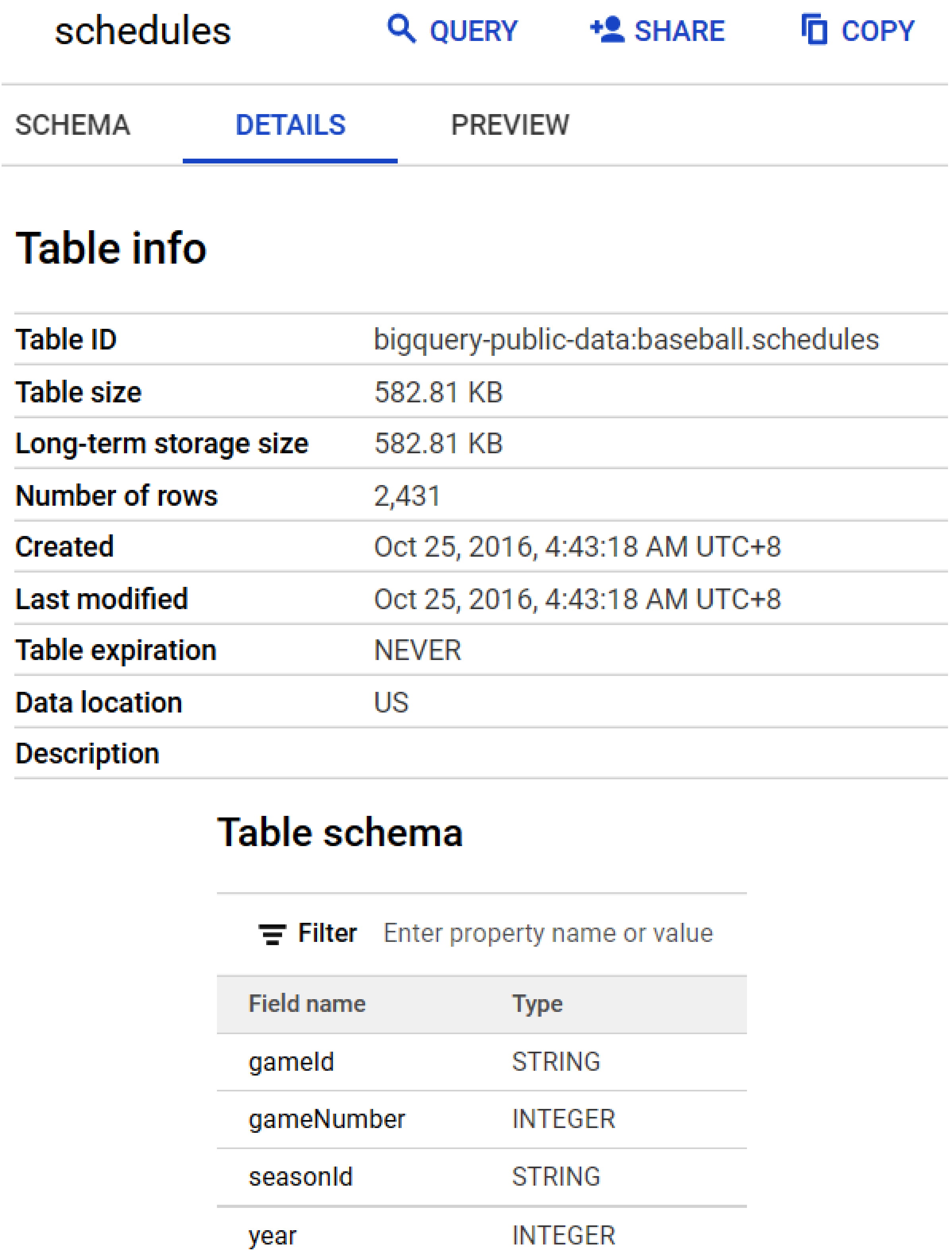
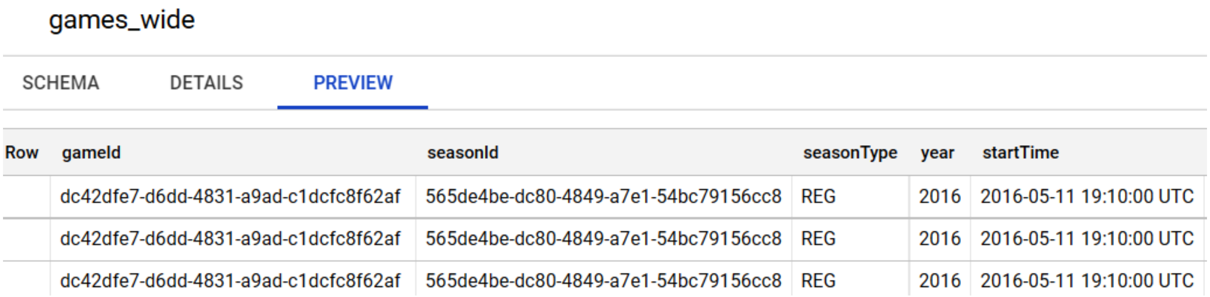


Table info

Table ID	bigquery-public-data:baseball.schedules
Table size	582.81 KB
Long-term storage size	582.81 KB
Number of rows	2,431
Created	Oct 25, 2016, 4:43:18 AM UTC+8
Last modified	Oct 25, 2016, 4:43:18 AM UTC+8
Table expiration	NEVER
Data location	US
Description	

Table schema

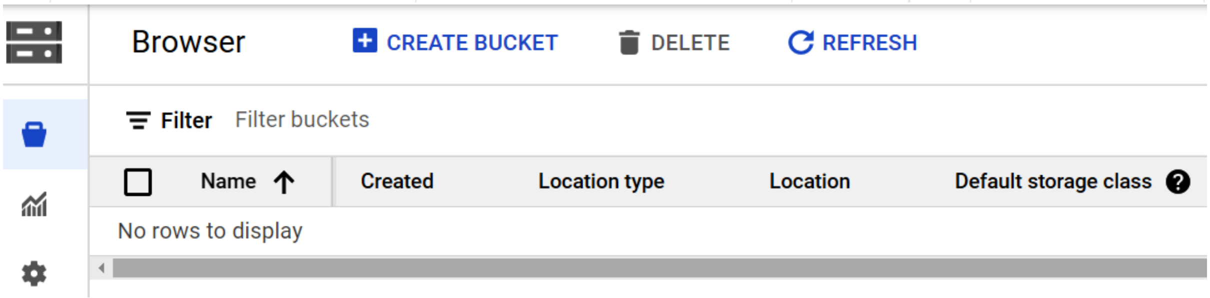
Field name	Type
gameId	STRING
gameNumber	INTEGER
seasonId	STRING
year	INTEGER



games_wide

SCHEMA DETAILS PREVIEW

Row	gameId	seasonId	seasonType	year	startTime
	dc42dfe7-d6dd-4831-a9ad-c1dcfc8f62af	565de4be-dc80-4849-a7e1-54bc79156cc8	REG	2016	2016-05-11 19:10:00 UTC
	dc42dfe7-d6dd-4831-a9ad-c1dcfc8f62af	565de4be-dc80-4849-a7e1-54bc79156cc8	REG	2016	2016-05-11 19:10:00 UTC
	dc42dfe7-d6dd-4831-a9ad-c1dcfc8f62af	565de4be-dc80-4849-a7e1-54bc79156cc8	REG	2016	2016-05-11 19:10:00 UTC

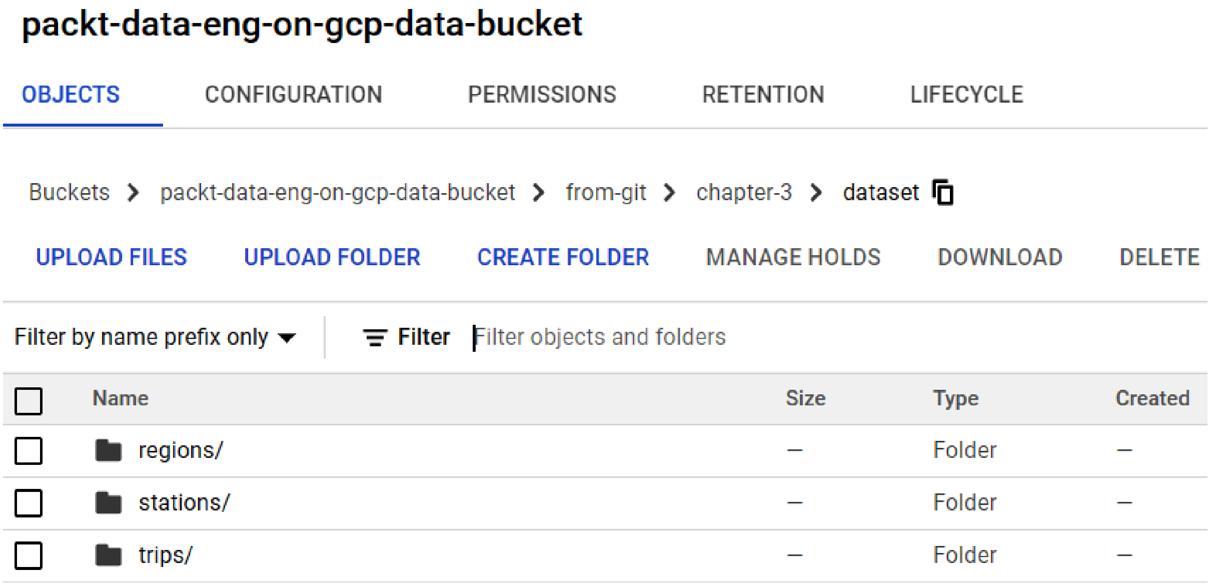


Browser CREATE BUCKET DELETE REFRESH

Filter Filter buckets

□	Name ↑	Created	Location type	Location	Default storage class

No rows to display



packt-data-eng-on-gcp-data-bucket

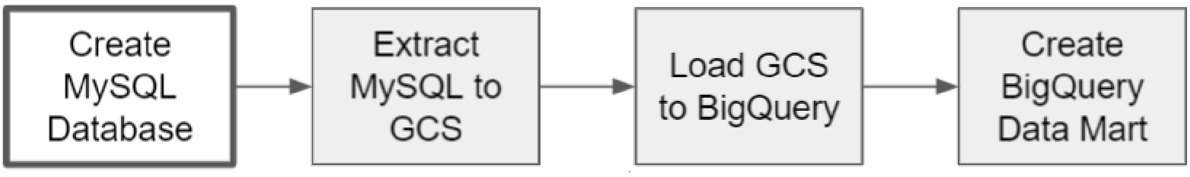
OBJECTS CONFIGURATION PERMISSIONS RETENTION LIFECYCLE

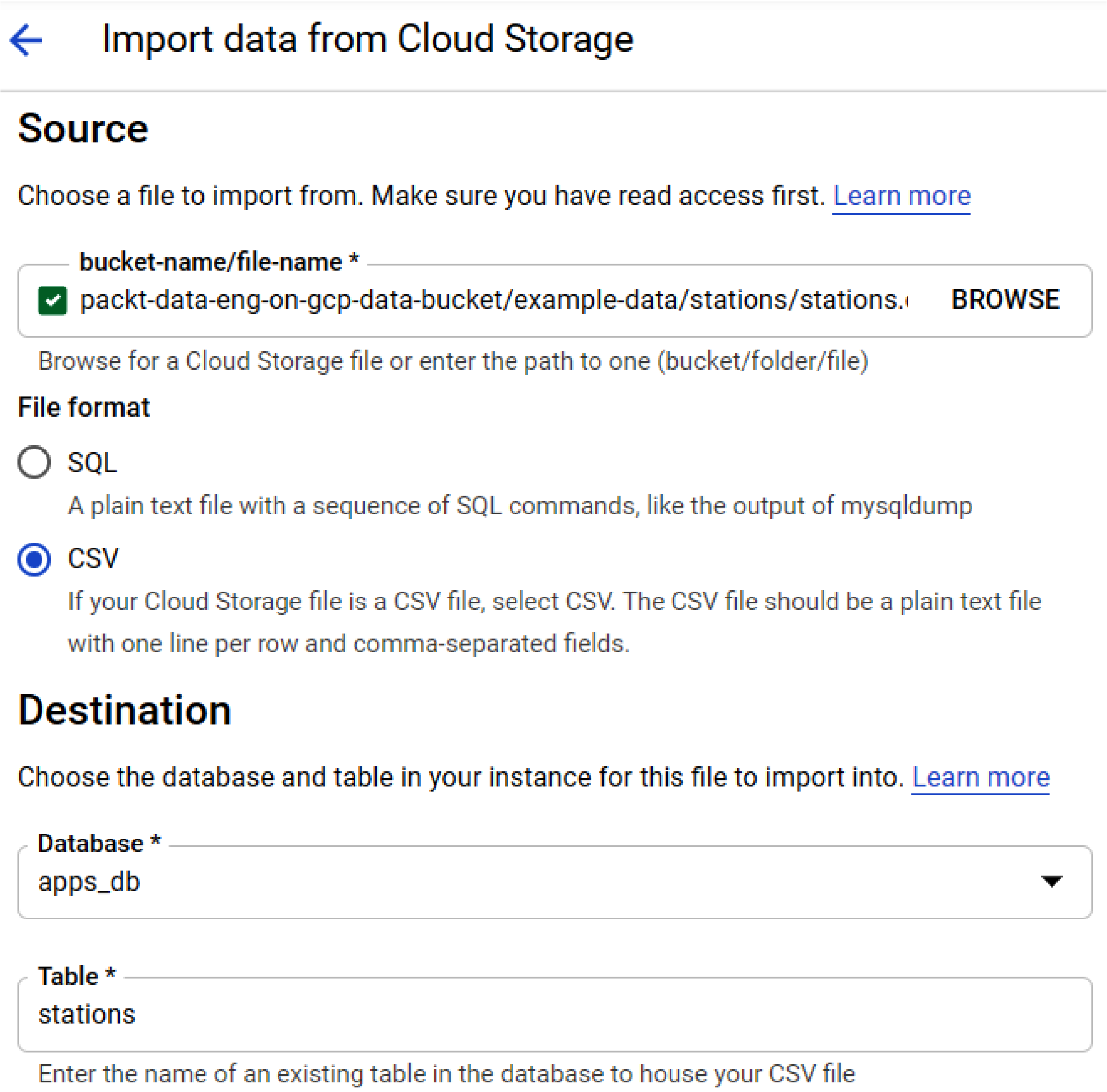
Buckets > packt-data-eng-on-gcp-data-bucket > from-git > chapter-3 > dataset

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE

Filter by name prefix only ▾ Filter Filter objects and folders

□	Name	Size	Type	Created
□	regions/	—	Folder	—
□	stations/	—	Folder	—
□	trips/	—	Folder	—





Import data from Cloud Storage

Source

Choose a file to import from. Make sure you have read access first. Learn more

bucket-name/file-name *
packt-data-eng-on-gcp-data-bucket/example-data/stations/stations.(BROWSE
Browse for a Cloud Storage file or enter the path to one (bucket/folder/file)

File format

○ SQL
A plain text file with a sequence of SQL commands, like the output of mysqldump

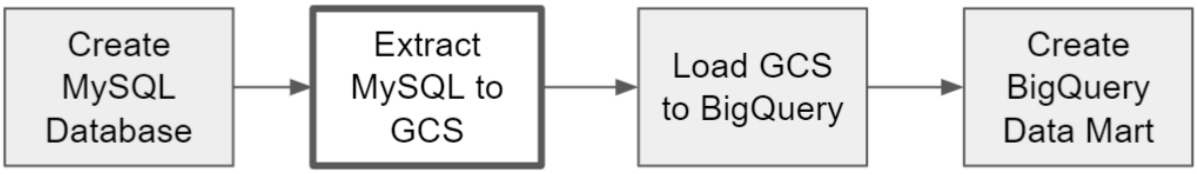
● CSV
If your Cloud Storage file is a CSV file, select CSV. The CSV file should be a plain text file with one line per row and comma-separated fields.

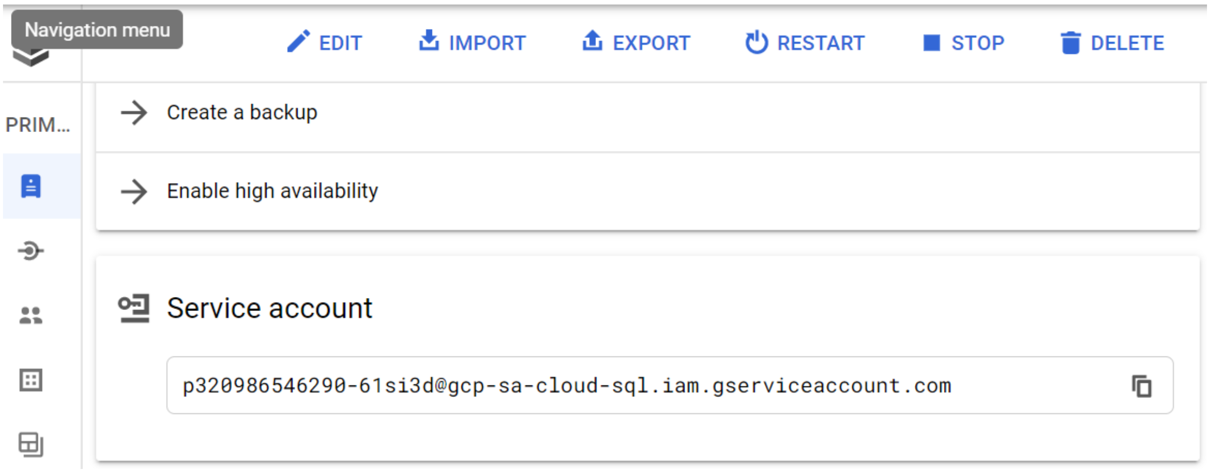
Destination

Choose the database and table in your instance for this file to import into. Learn more

Database *
apps_db

Table *
stations
Enter the name of an existing table in the database to house your CSV file



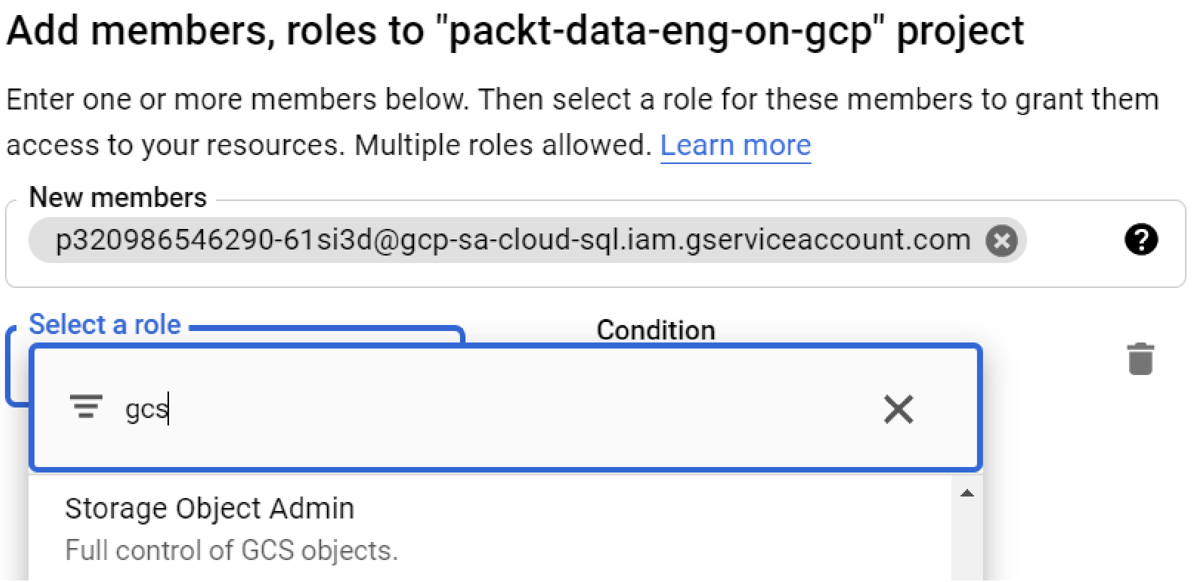


Create a backup

Enable high availability

Service account

p320986546290-61si3d@gcp-sa-cloud-sql.iam.gserviceaccount.com



Add members, roles to "packt-data-eng-on-gcp" project

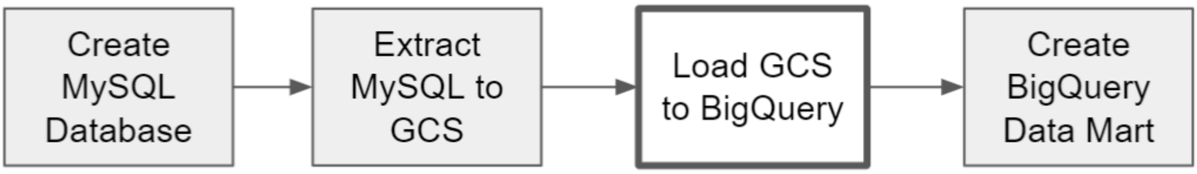
Enter one or more members below. Then select a role for these members to grant them access to your resources. Multiple roles allowed. Learn more

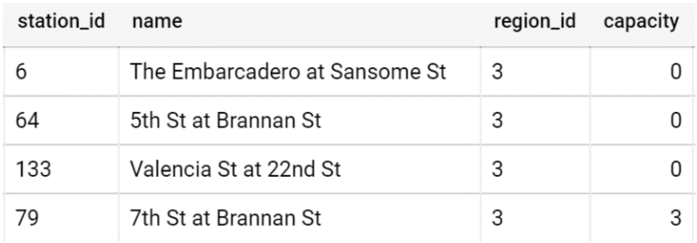
New members: p320986546290-61si3d@gcp-sa-cloud-sql.iam.gserviceaccount.com

Select a role: gcs

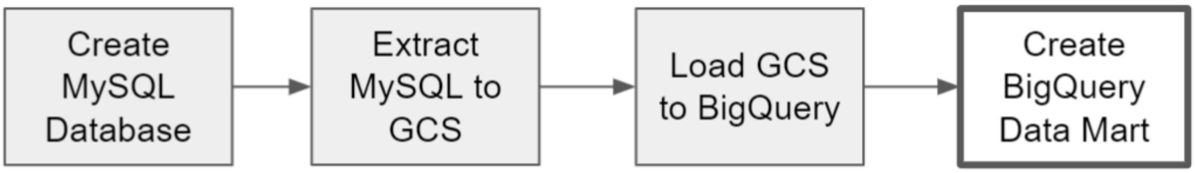
Condition

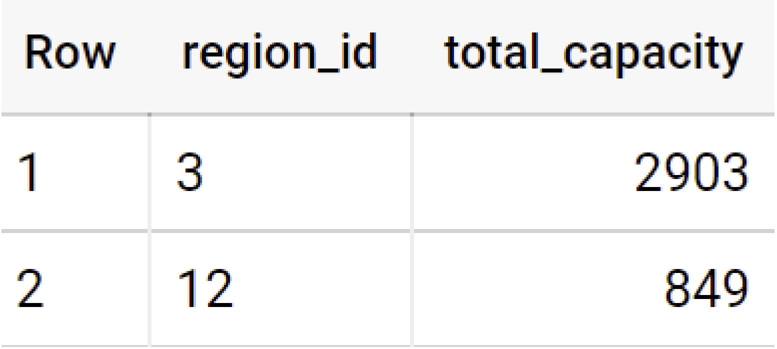
Storage Object Admin
Full control of GCS objects.



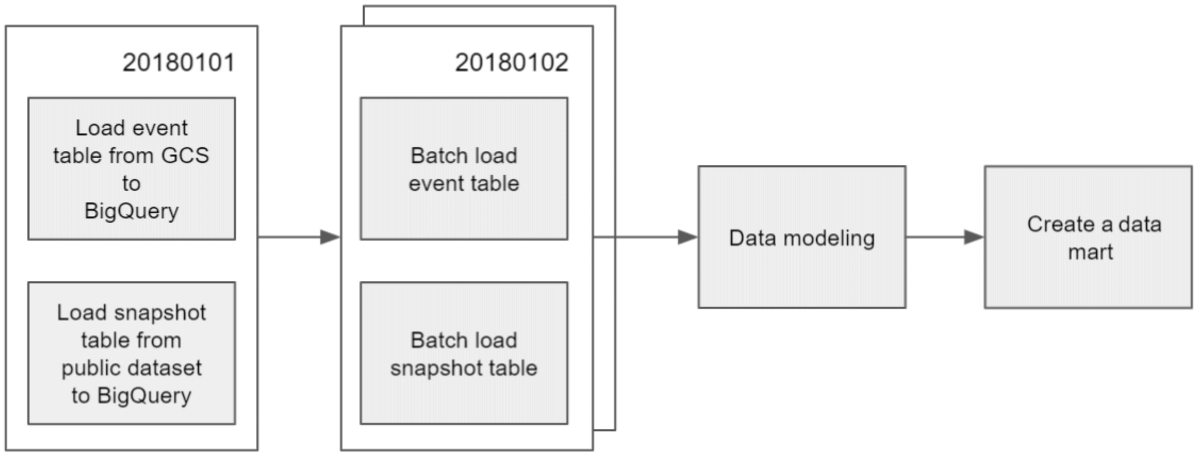


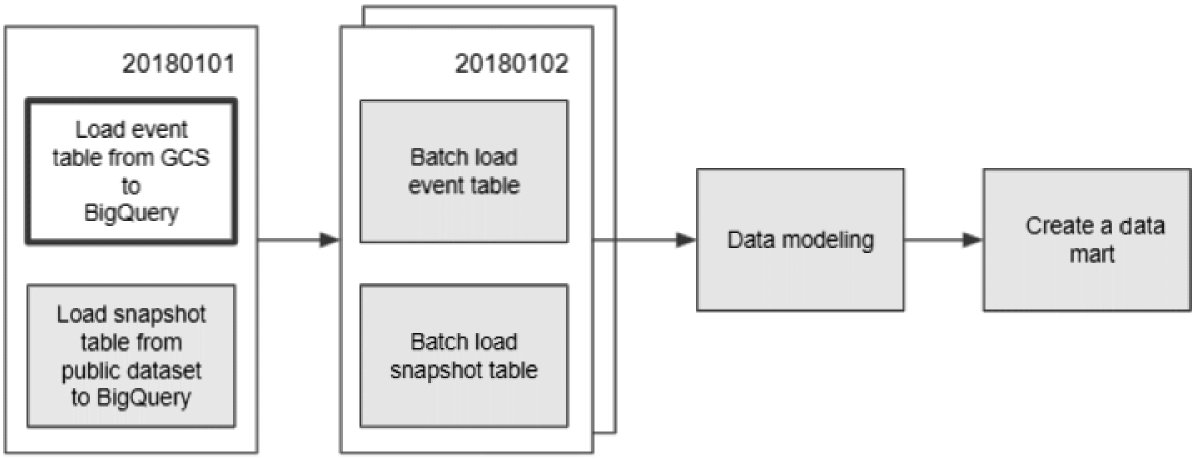
station_id	name	region_id	capacity
6	The Embarcadero at Sansome St	3	0
64	5th St at Brannan St	3	0
133	Valencia St at 22nd St	3	0
79	7th St at Brannan St	3	3

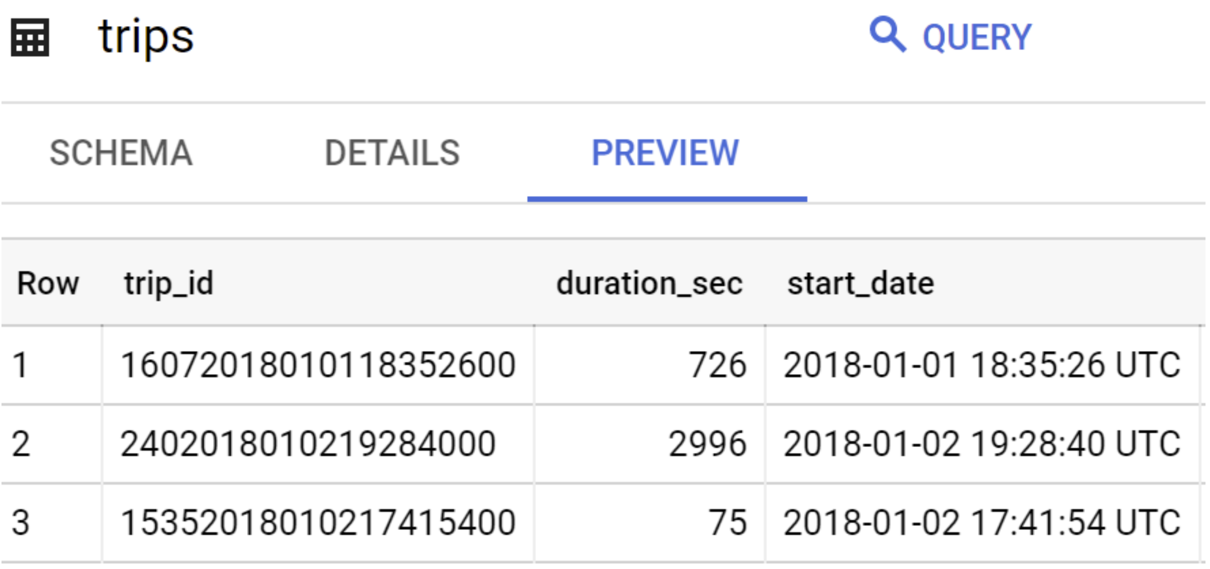




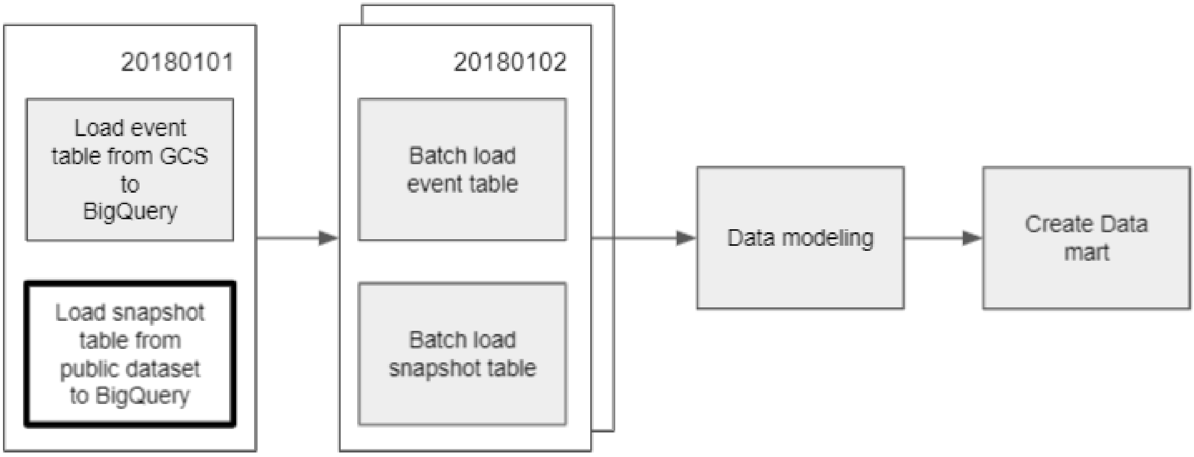
Row	region_id	total_capacity
1	3	2903
2	12	849

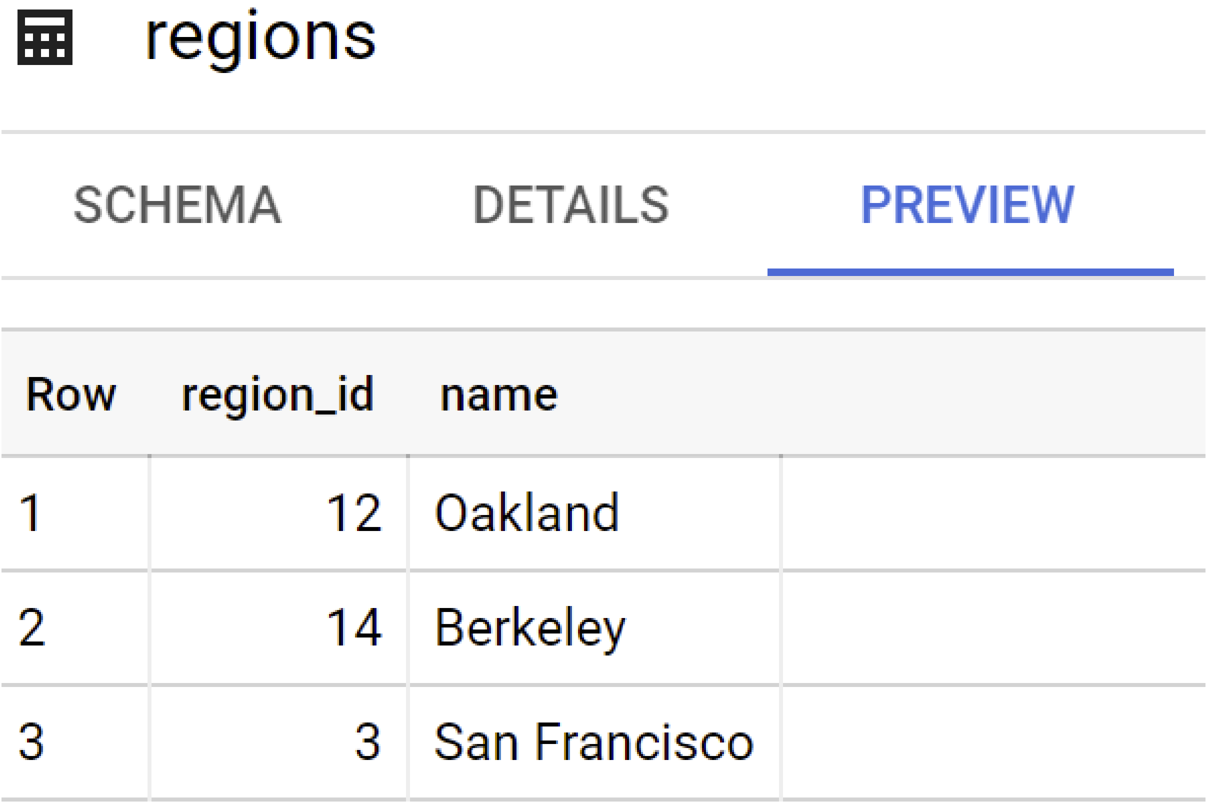

20180101
Load event table from GCS to BigQuery
Load snapshot table from public dataset to BigQuery
20180102
Batch load event table
Batch load snapshot table
Data modeling
Create a data mart


20180101
Load event table from GCS to BigQuery
Load snapshot table from public dataset to BigQuery
20180102
Batch load event table
Batch load snapshot table
Data modeling
Create a data mart

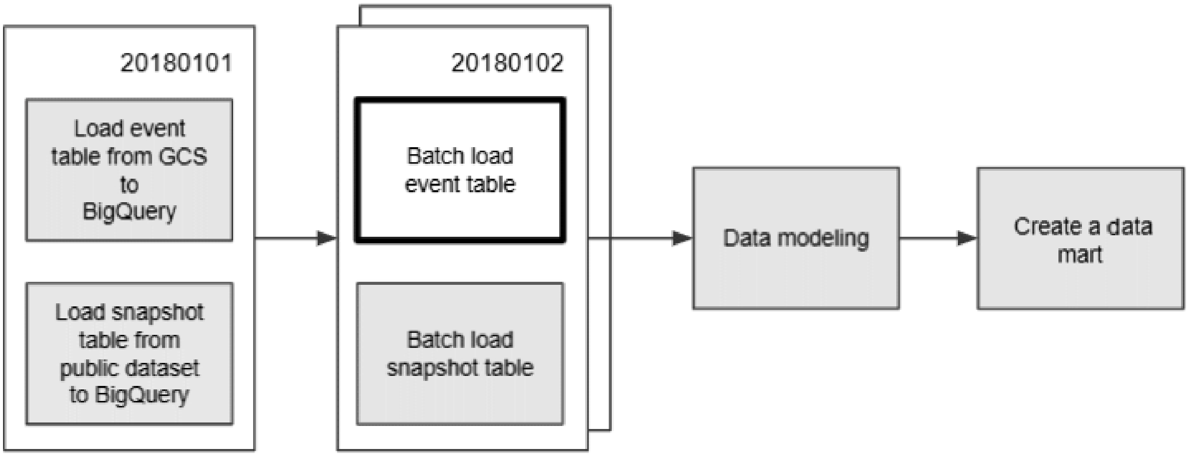


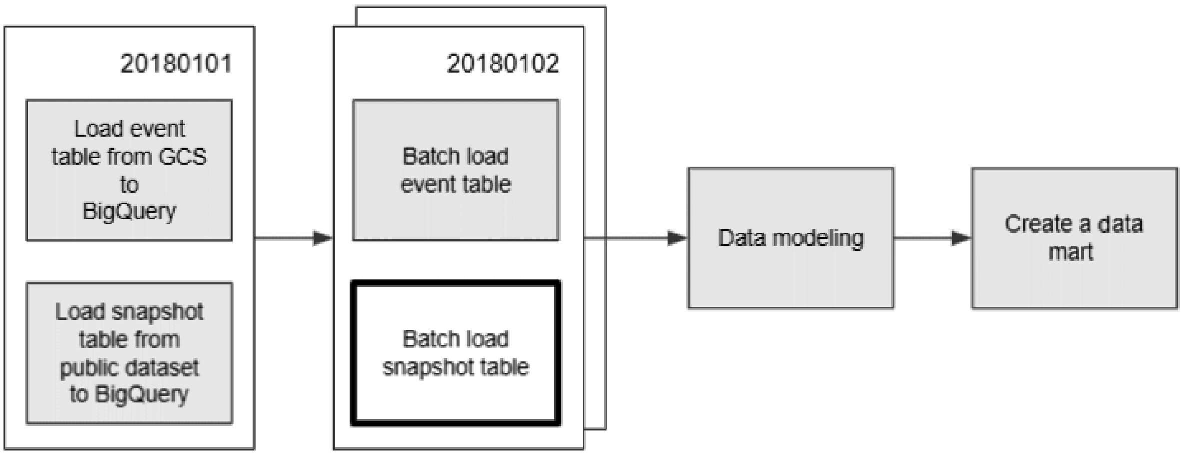
Row	trip_id	duration_sec	start_date
1	16072018010118352600	726	2018-01-01 18:35:26 UTC
2	2402018010219284000	2996	2018-01-02 19:28:40 UTC
3	15352018010217415400	75	2018-01-02 17:41:54 UTC

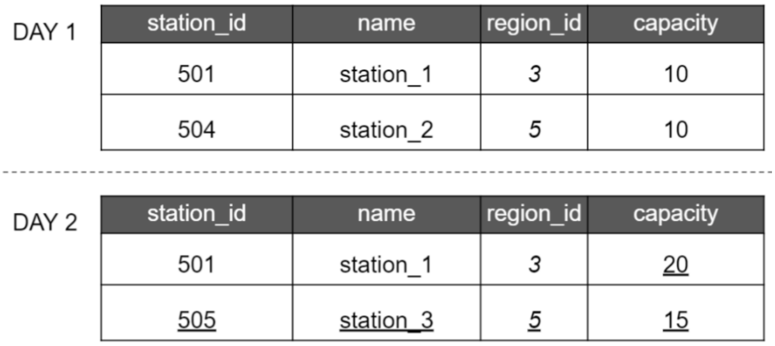

20180101
Load event table from GCS to BigQuery
Load snapshot table from public dataset to BigQuery
20180102
Batch load event table
Batch load snapshot table
Data modeling
Create Data mart



Row	region_id	name
1	12	Oakland
2	14	Berkeley
3	3	San Francisco


20180101
Load event table from GCS to BigQuery
Load snapshot table from public dataset to BigQuery
20180102
Batch load event table
Batch load snapshot table
Data modeling
Create a data mart


20180101
Load event table from GCS to BigQuery
Load snapshot table from public dataset to BigQuery
20180102
Batch load event table
Batch load snapshot table
Data modeling
Create a data mart

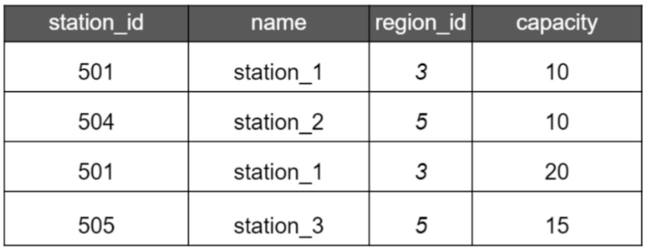


DAY 1

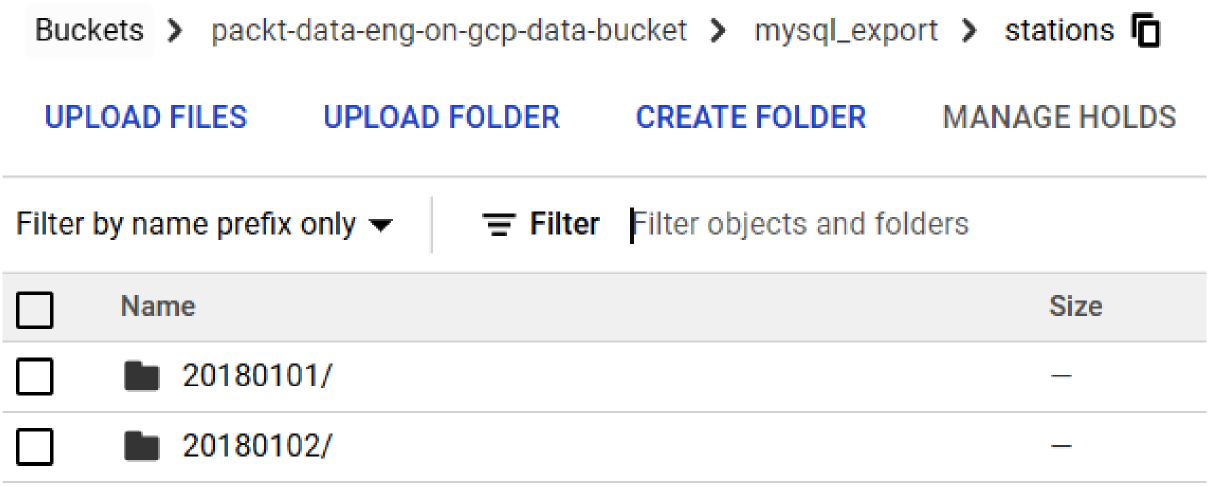
station_id	name	region_id	capacity
501	station_1	3	10
504	station_2	5	10

DAY 2

station_id	name	region_id	capacity
501	station_1	3	20
505	station_3	5	15



station_id	name	region_id	capacity
501	station_1	3	10
504	station_2	5	10
501	station_1	3	20
505	station_3	5	15

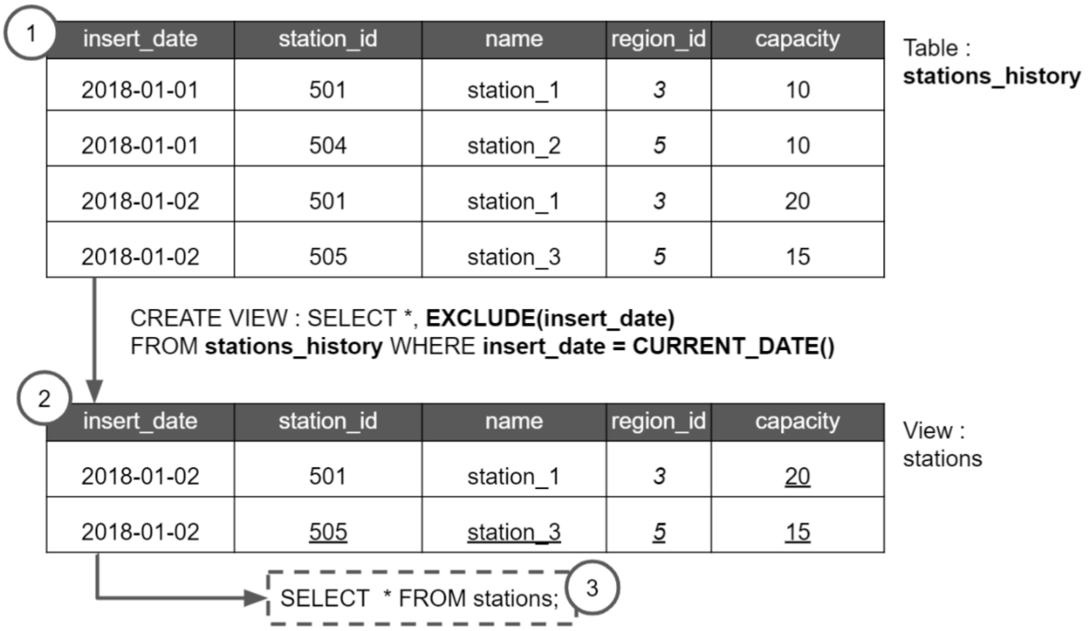


Buckets > packt-data-eng-on-gcp-data-bucket > mysql_export > stations

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS

Filter by name prefix only ▾ Filter Filter objects and folders

□	Name	Size
□	20180101/	—
□	20180102/	—



insert_date	station_id	name	region_id	capacity
2018-01-01	501	station_1	*3*	10
2018-01-01	504	station_2	*5*	10
2018-01-02	501	station_1	*3*	20
2018-01-02	505	station_3	*5*	15

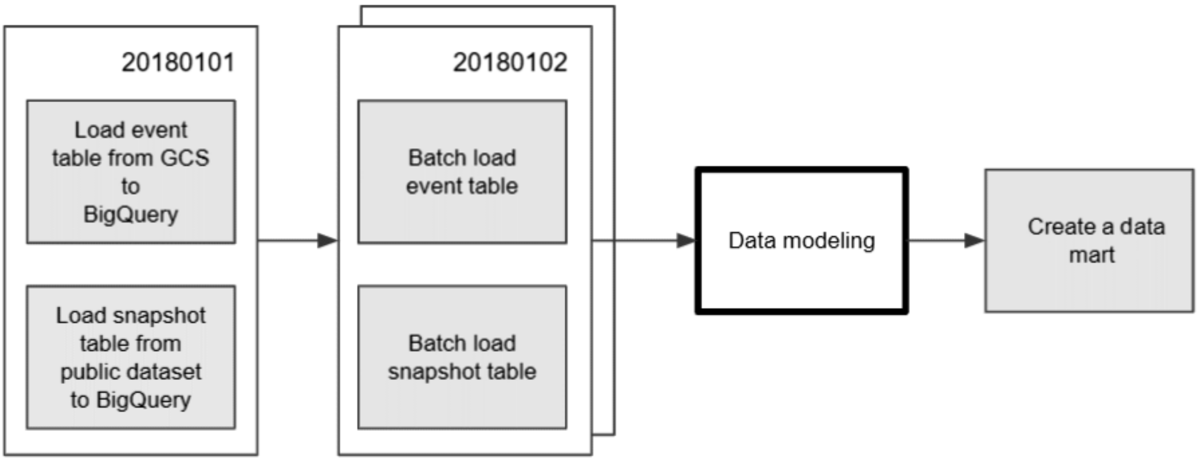
Table : **stations_history**

CREATE VIEW : SELECT *, **EXCLUDE(insert_date)**
FROM **stations_history** WHERE **insert_date = CURRENT_DATE()**

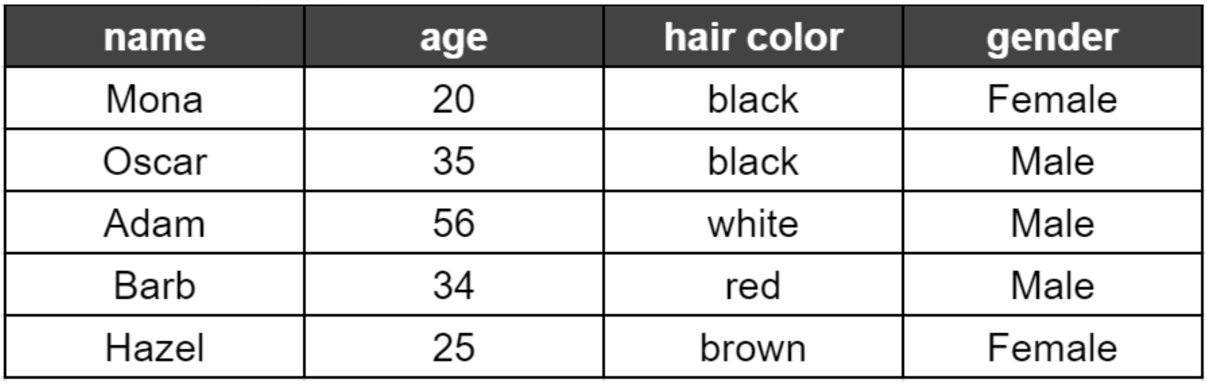
insert_date	station_id	name	region_id	capacity
2018-01-02	501	station_1	*3*	20
2018-01-02	505	station_3	*5*	15

View : stations

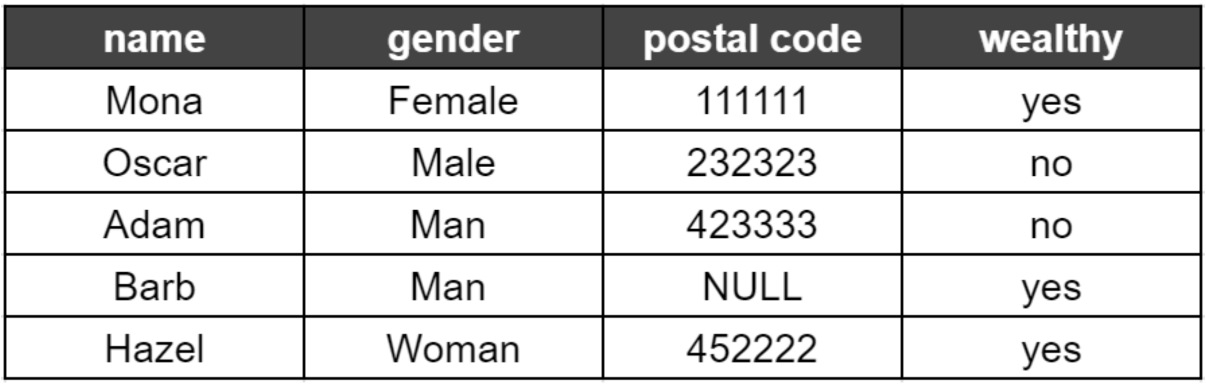
SELECT * FROM stations;



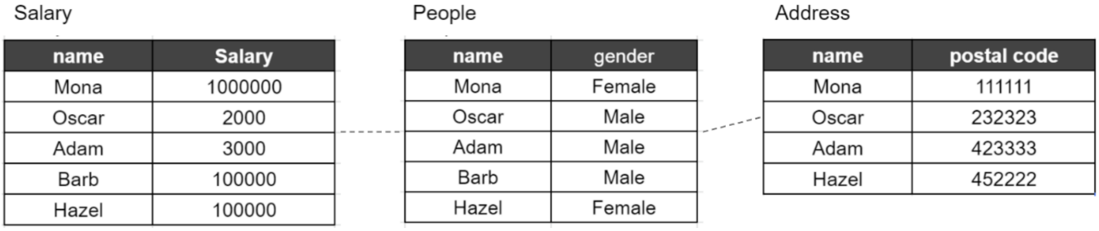
20180101
Load event table from GCS to BigQuery
Load snapshot table from public dataset to BigQuery
20180102
Batch load event table
Batch load snapshot table
Data modeling
Create a data mart



name	age	hair color	gender
Mona	20	black	Female
Oscar	35	black	Male
Adam	56	white	Male
Barb	34	red	Male
Hazel	25	brown	Female



name	gender	postal code	wealthy
Mona	Female	111111	yes
Oscar	Male	232323	no
Adam	Man	423333	no
Barb	Man	NULL	yes
Hazel	Woman	452222	yes



Salary

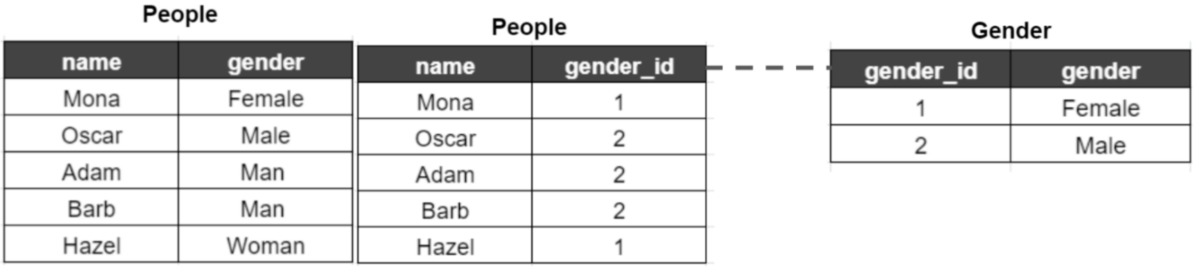
name	Salary
Mona	1000000
Oscar	2000
Adam	3000
Barb	100000
Hazel	100000

People

name	gender
Mona	Female
Oscar	Male
Adam	Male
Barb	Male
Hazel	Female

Address

name	postal code
Mona	111111
Oscar	232323
Adam	423333
Hazel	452222



People

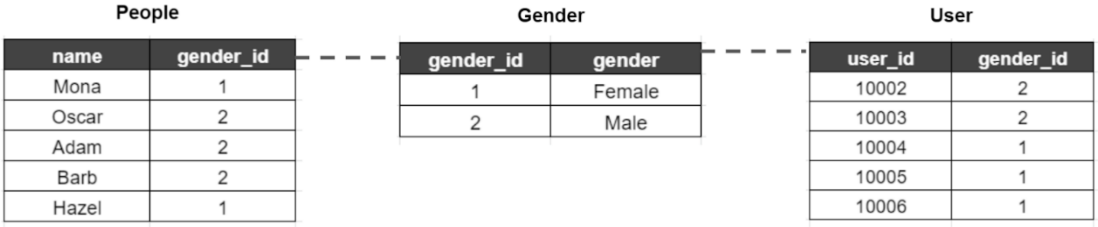
name	gender
Mona	Female
Oscar	Male
Adam	Man
Barb	Man
Hazel	Woman

People

name	gender_id
Mona	1
Oscar	2
Adam	2
Barb	2
Hazel	1

Gender

gender_id	gender
1	Female
2	Male



People

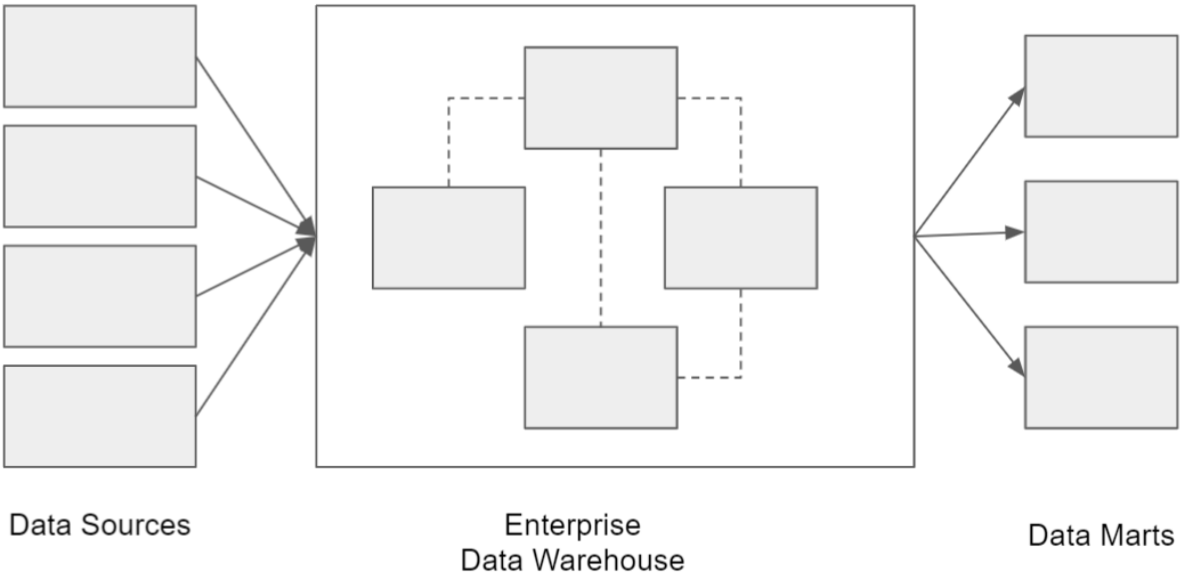
name	gender_id
Mona	1
Oscar	2
Adam	2
Barb	2
Hazel	1

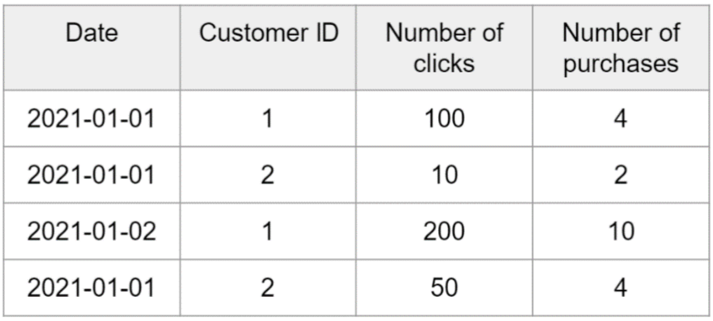
Gender

gender_id	gender
1	Female
2	Male

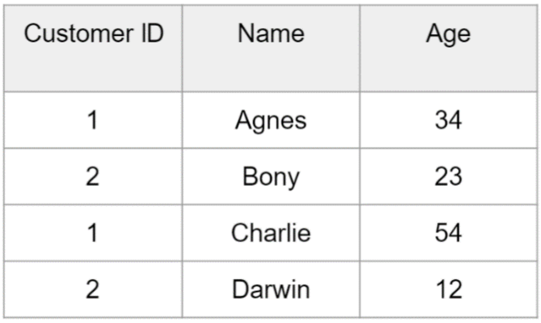
User

user_id	gender_id
10002	2
10003	2
10004	1
10005	1
10006	1

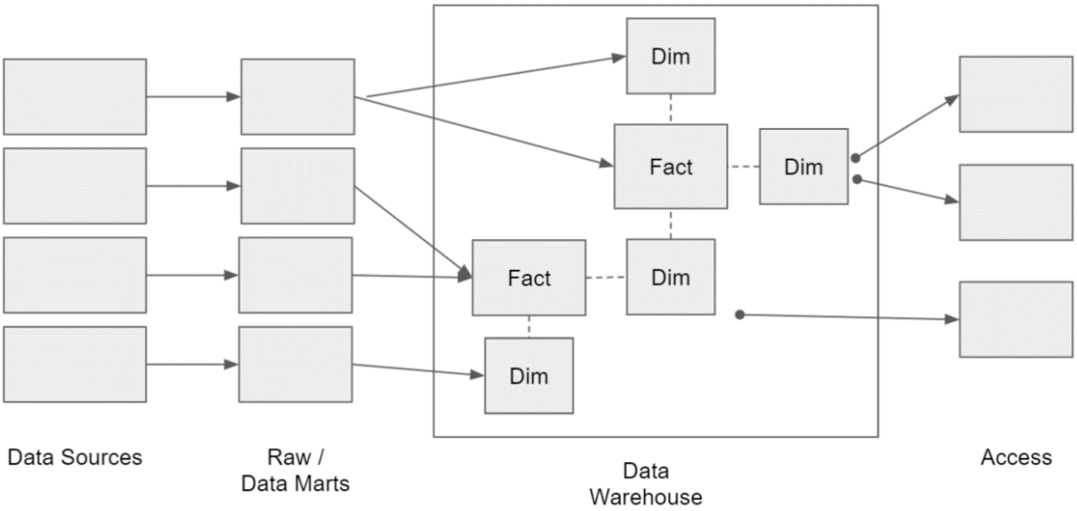


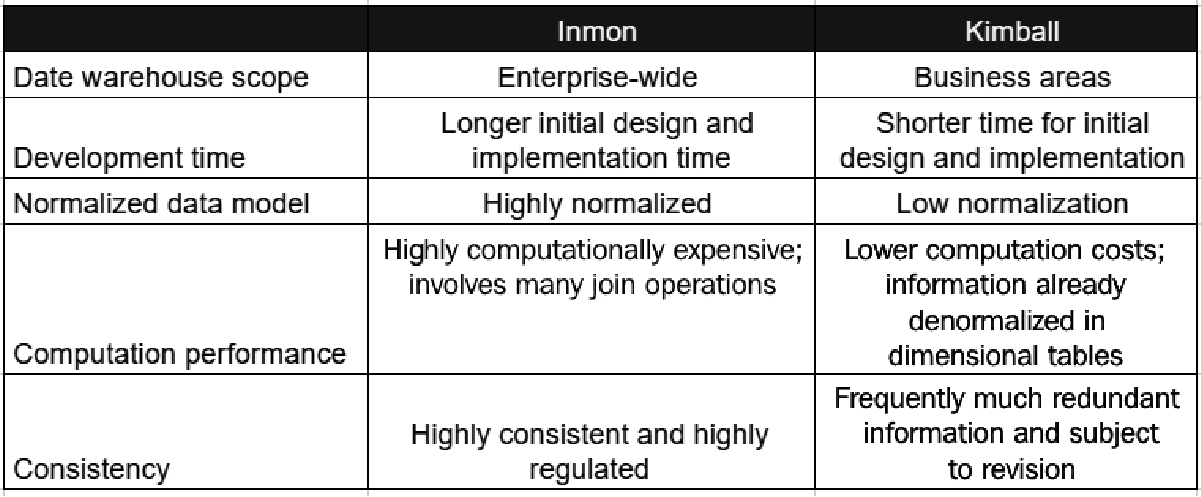


Date	Customer ID	Number of clicks	Number of purchases
2021-01-01	1	100	4
2021-01-01	2	10	2
2021-01-02	1	200	10
2021-01-01	2	50	4

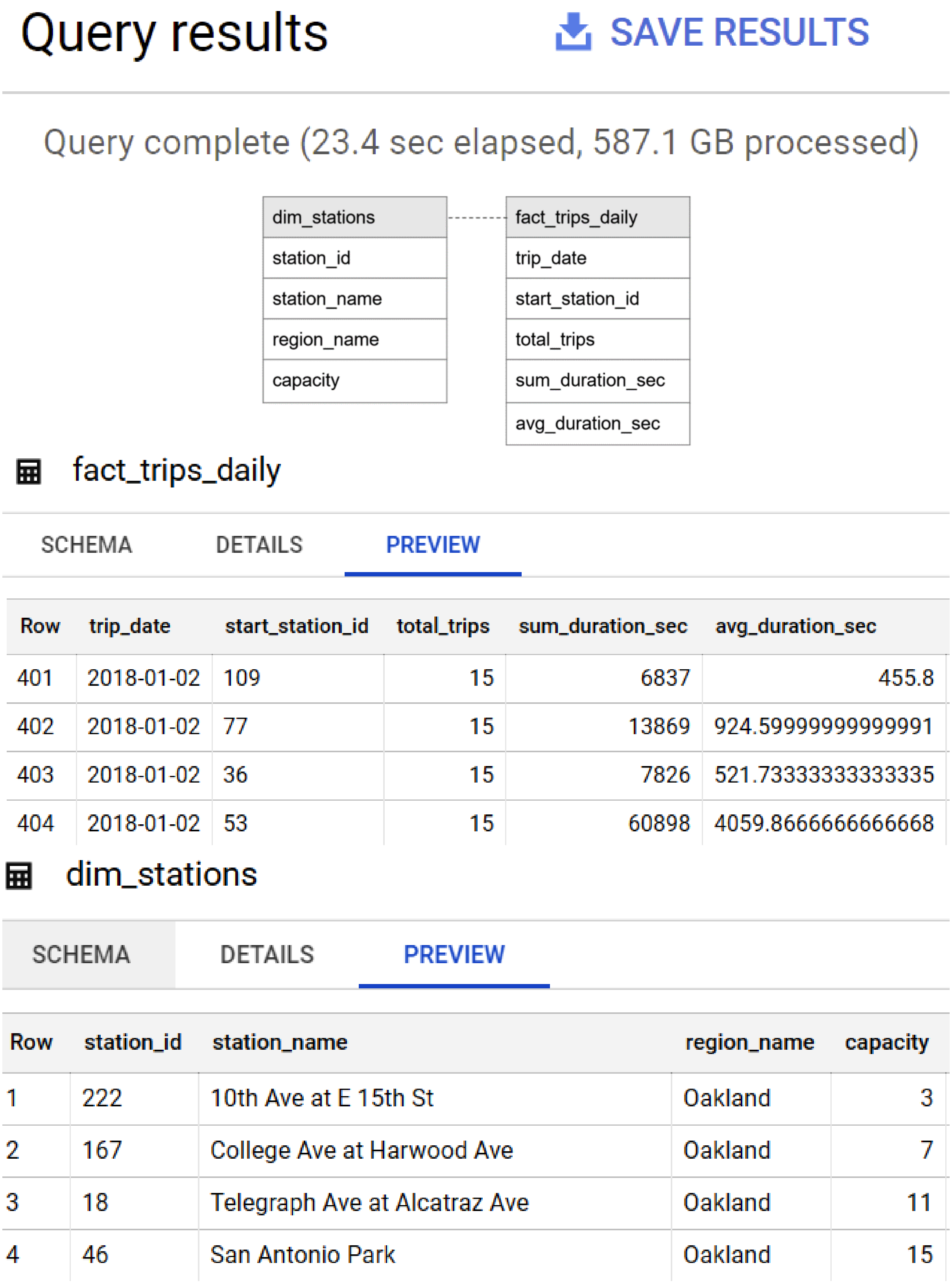


Customer ID	Name	Age
1	Agnes	34
2	Bony	23
1	Charlie	54
2	Darwin	12





	Inmon	Kimball
Date warehouse scope	Enterprise-wide	Business areas
Development time	Longer initial design and implementation time	Shorter time for initial design and implementation
Normalized data model	Highly normalized	Low normalization
Computation performance	Highly computationally expensive; involves many join operations	Lower computation costs; information already denormalized in dimensional tables
Consistency	Highly consistent and highly regulated	Frequently much redundant information and subject to revision



Query results

Query complete (23.4 sec elapsed, 587.1 GB processed)

dim_stations
station_id
station_name
region_name
capacity

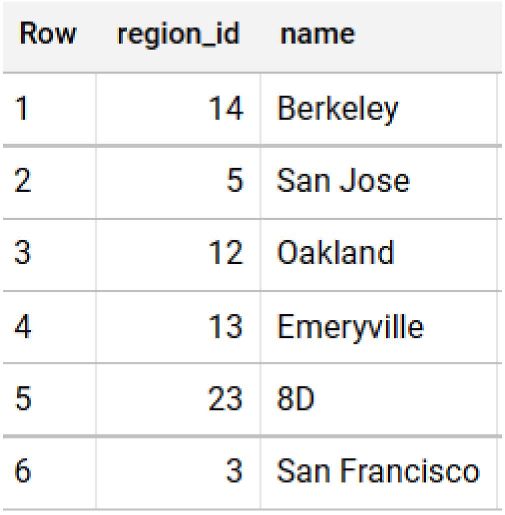
fact_trips_daily
trip_date
start_station_id
total_trips
sum_duration_sec
avg_duration_sec

fact_trips_daily

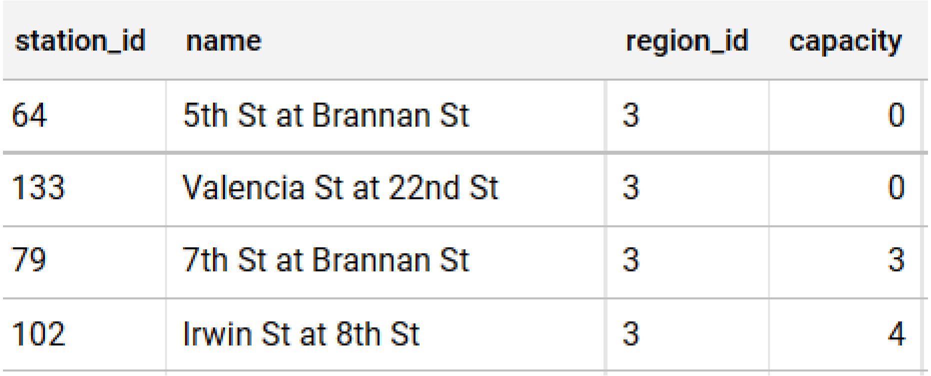
Row	trip_date	start_station_id	total_trips	sum_duration_sec	avg_duration_sec
401	2018-01-02	109	15	6837	455.8
402	2018-01-02	77	15	13869	924.59999999999991
403	2018-01-02	36	15	7826	521.73333333333335
404	2018-01-02	53	15	60898	4059.8666666666668

dim_stations

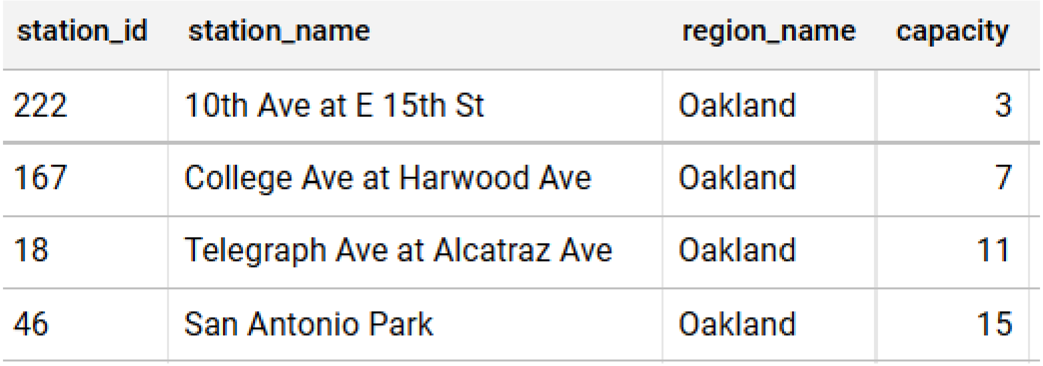
Row	station_id	station_name	region_name	capacity
1	222	10th Ave at E 15th St	Oakland	3
2	167	College Ave at Harwood Ave	Oakland	7
3	18	Telegraph Ave at Alcatraz Ave	Oakland	11
4	46	San Antonio Park	Oakland	15



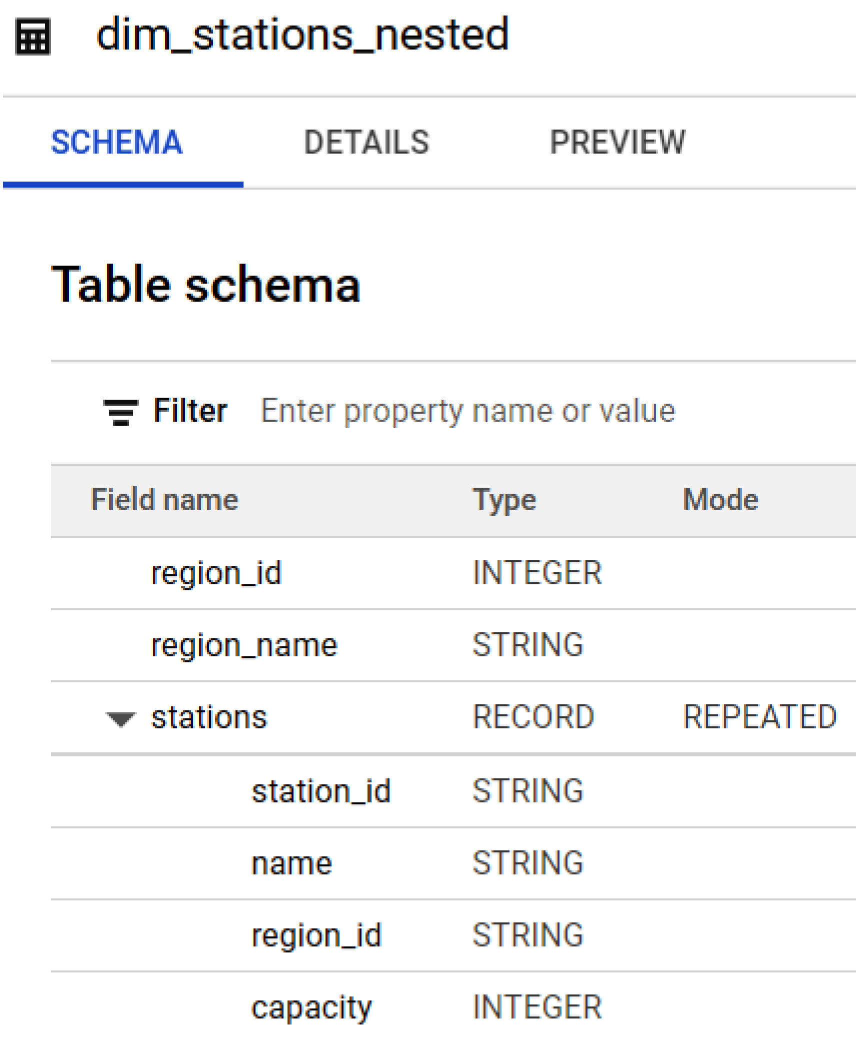
Row	region_id	name
1	14	Berkeley
2	5	San Jose
3	12	Oakland
4	13	Emeryville
5	23	8D
6	3	San Francisco



station_id	name	region_id	capacity
64	5th St at Brannan St	3	0
133	Valencia St at 22nd St	3	0
79	7th St at Brannan St	3	3
102	Irwin St at 8th St	3	4



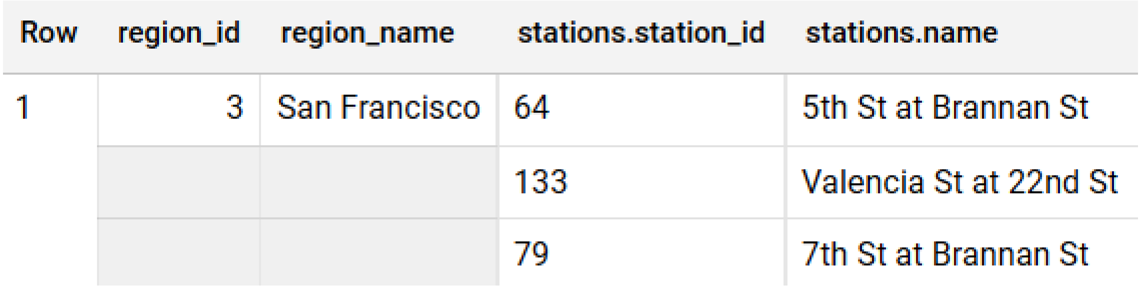
station_id	station_name	region_name	capacity
222	10th Ave at E 15th St	Oakland	3
167	College Ave at Harwood Ave	Oakland	7
18	Telegraph Ave at Alcatraz Ave	Oakland	11
46	San Antonio Park	Oakland	15



dim_stations_nested

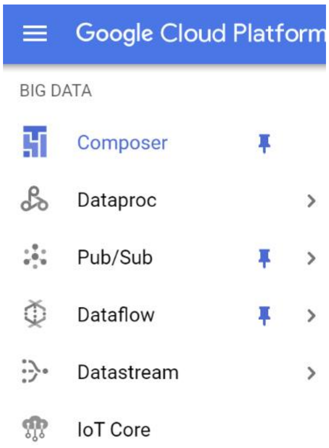
Table schema

Field name	Type	Mode
region_id	INTEGER	
region_name	STRING	
stations	RECORD	REPEATED
station_id	STRING	
name	STRING	
region_id	STRING	
capacity	INTEGER	



Row	region_id	region_name	stations.station_id	stations.name
1	3	San Francisco	64	5th St at Brannan St
			133	Valencia St at 22nd St
			79	7th St at Brannan St

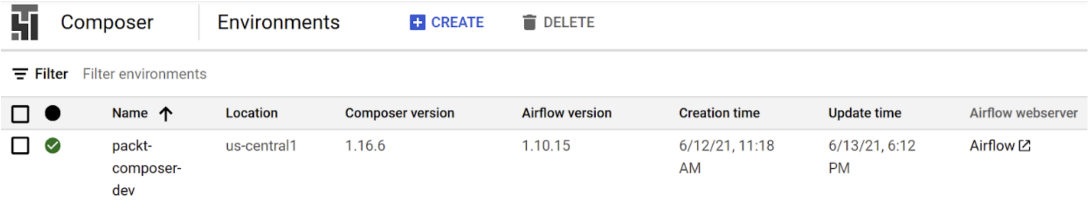
Chapter 4: Building Orchestration for Batch Data Loading Using Cloud Composer



Google Cloud Platform

BIG DATA

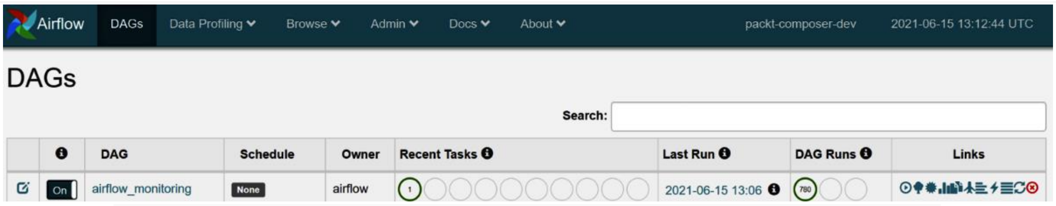
- Composer
- Dataproc
- Pub/Sub
- Dataflow
- Datastream
- IoT Core



Composer | Environments | CREATE | DELETE

Filter Filter environments

□	●	Name ↑	Location	Composer version	Airflow version	Creation time	Update time	Airflow webserver
□	✓	packt-composer-dev	us-central1	1.16.6	1.10.15	6/12/21, 11:18 AM	6/13/21, 6:12 PM	Airflow

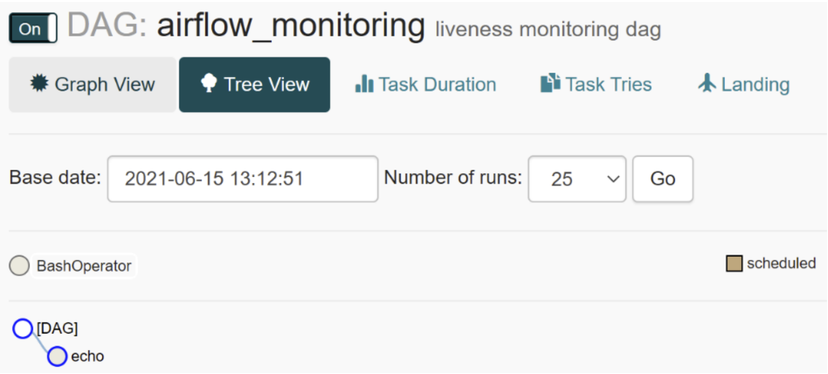


Airflow | DAGs | Data Profiling | Browse | Admin | Docs | About | packt-composer-dev | 2021-06-15 13:12:44 UTC

DAGs

Search:

		DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
	On	airflow_monitoring	None	airflow	1	2021-06-15 13:06	780	



On DAG: airflow_monitoring liveness monitoring dag

Graph View | Tree View | Task Duration | Task Tries | Landing

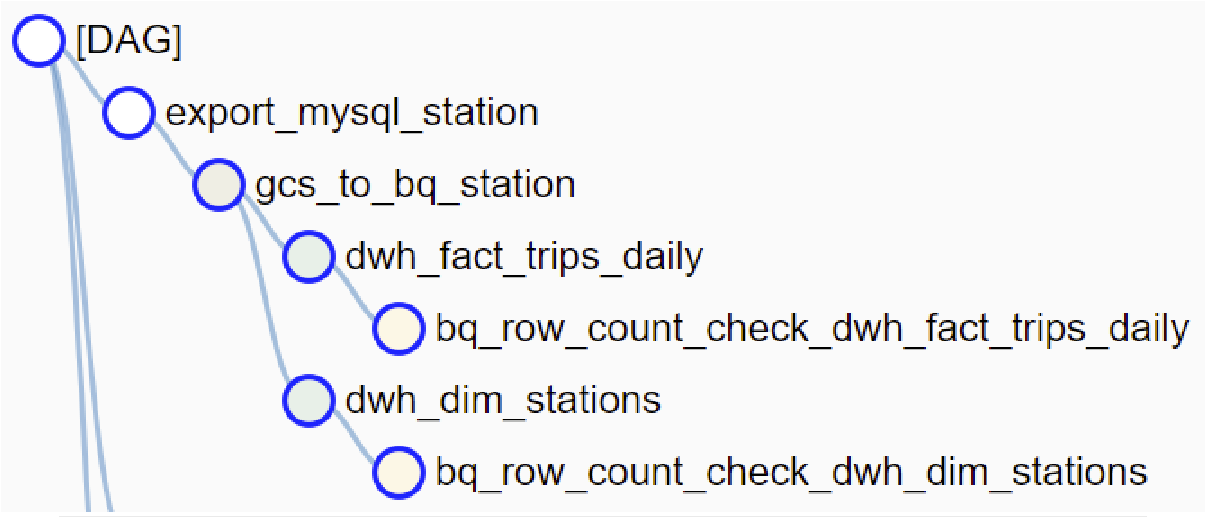
Base date: 2021-06-15 13:12:51 Number of runs: 25 Go

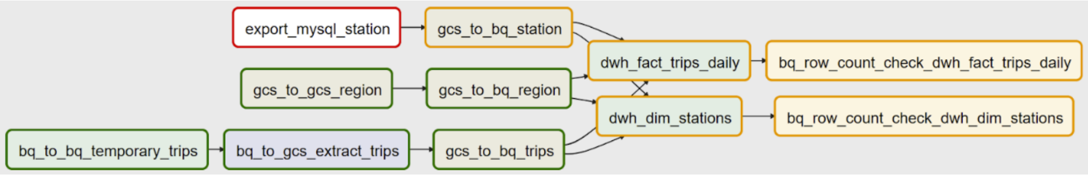
○ BashOperator

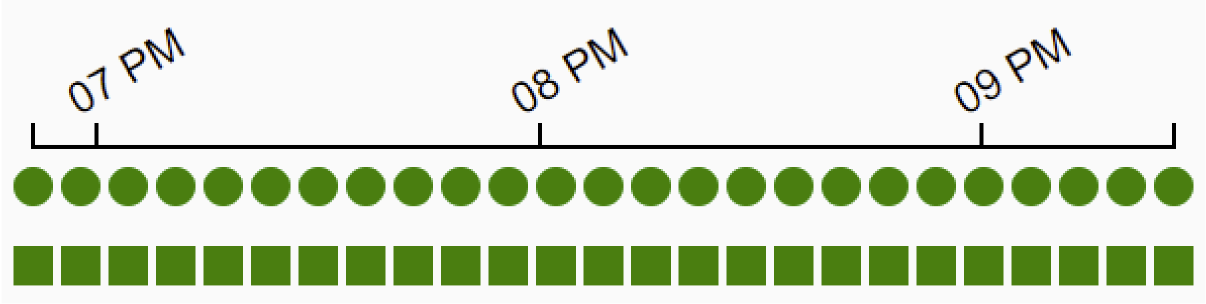
scheduled

[DAG]

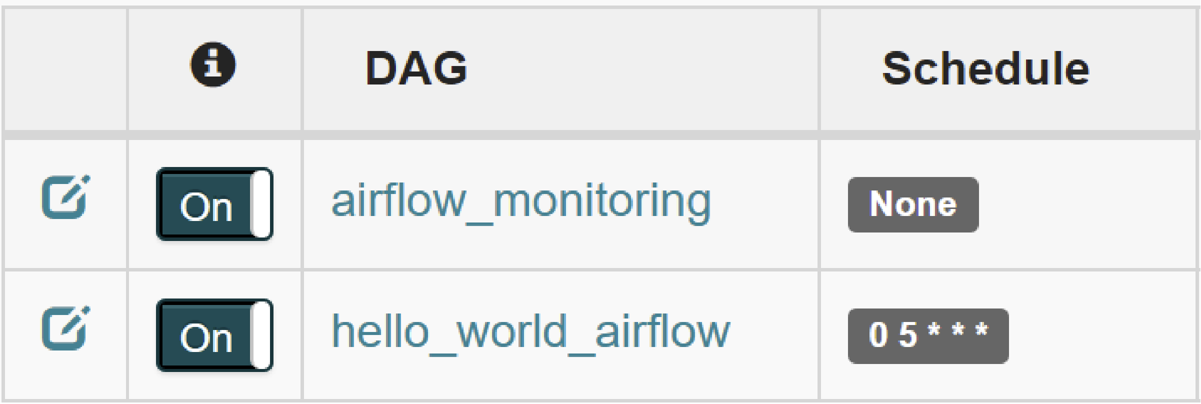
echo



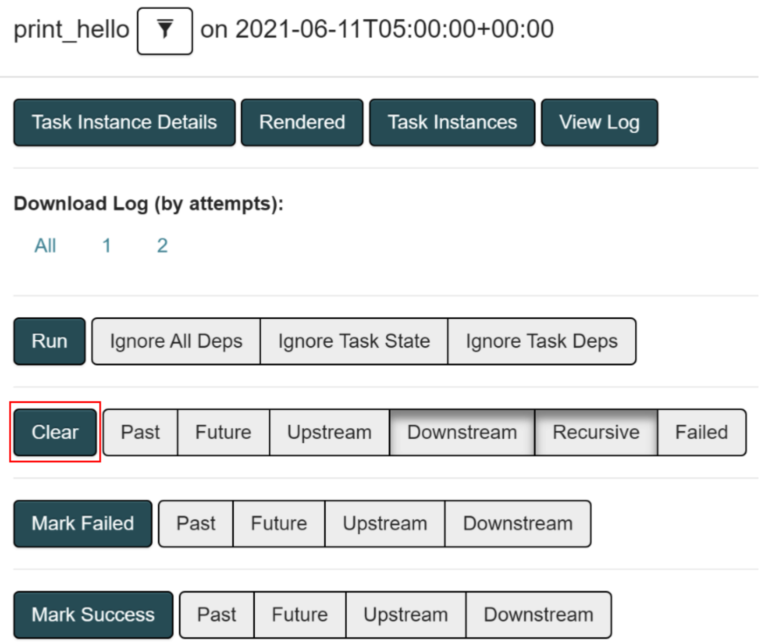


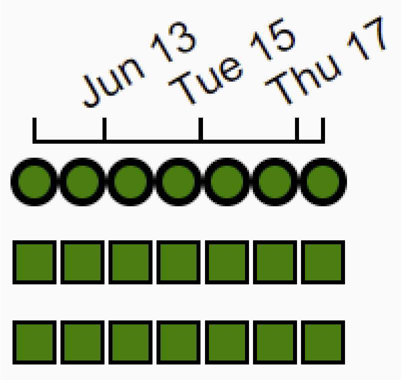


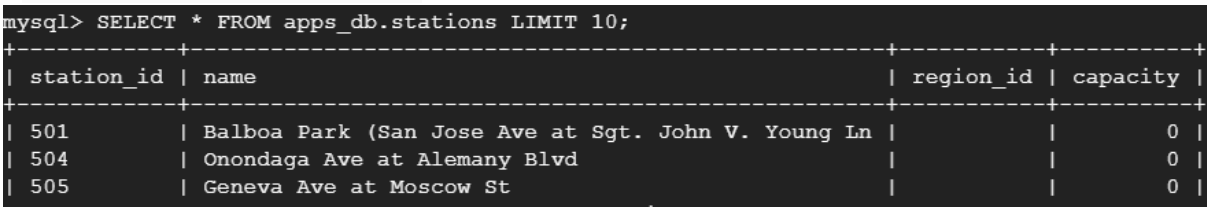
GCS directories	Mapped Local Directory	Usage
gs://{composer-bucket}/dags	/home/airflow/gcs/dags	DAGs
gs://{composer-bucket}/plugins	/home/airflow/gcs/plugins	Airflow plugins
gs://{composer-bucket}/data	/home/airflow/gcs/data	Workflow-related data
gs://{composer-bucket}/logs	/home/airflow/gcs/logs	Airflow task logs



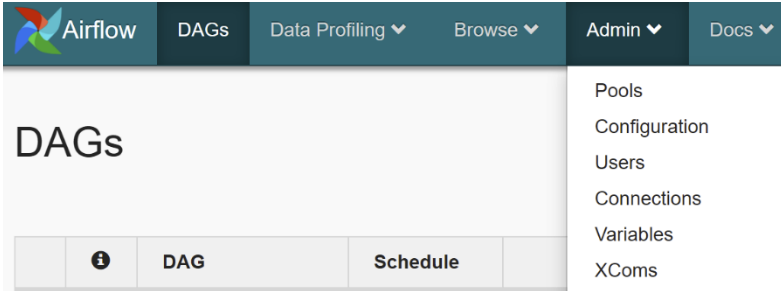
	ℹ	DAG	Schedule
✎	On	airflow_monitoring	None
✎	On	hello_world_airflow	0 5 * * *

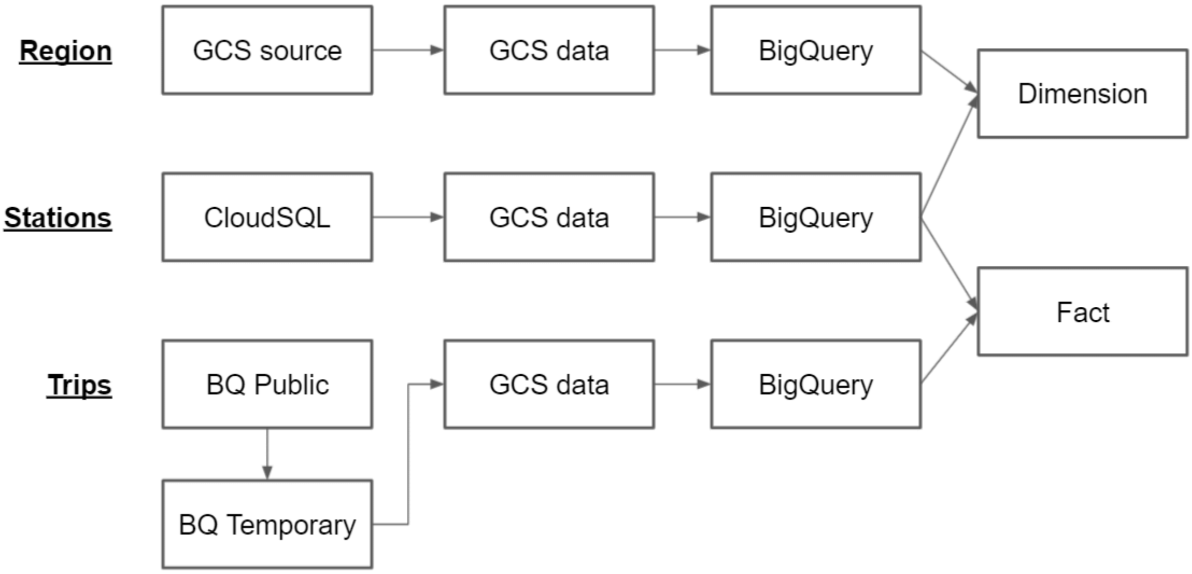


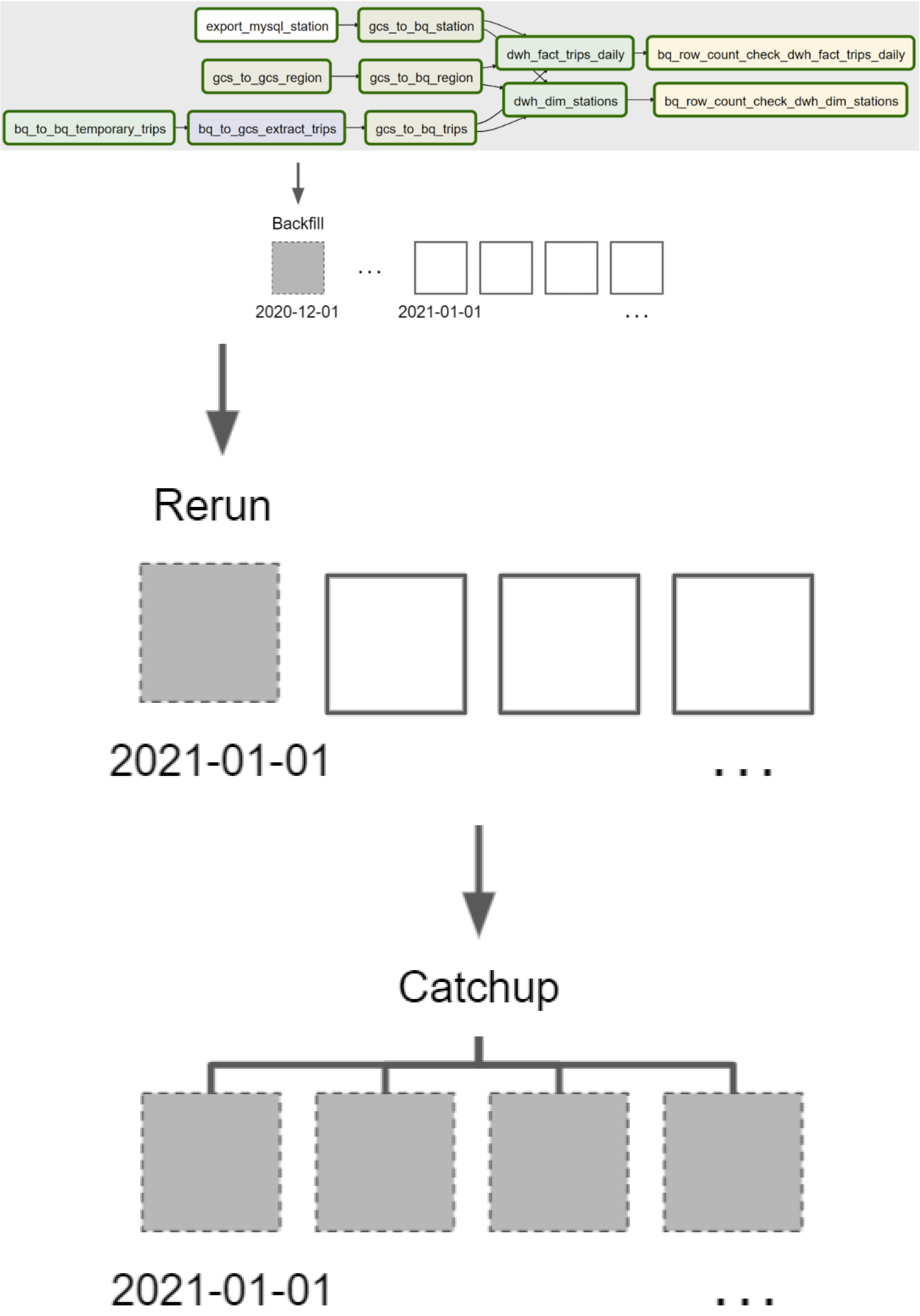


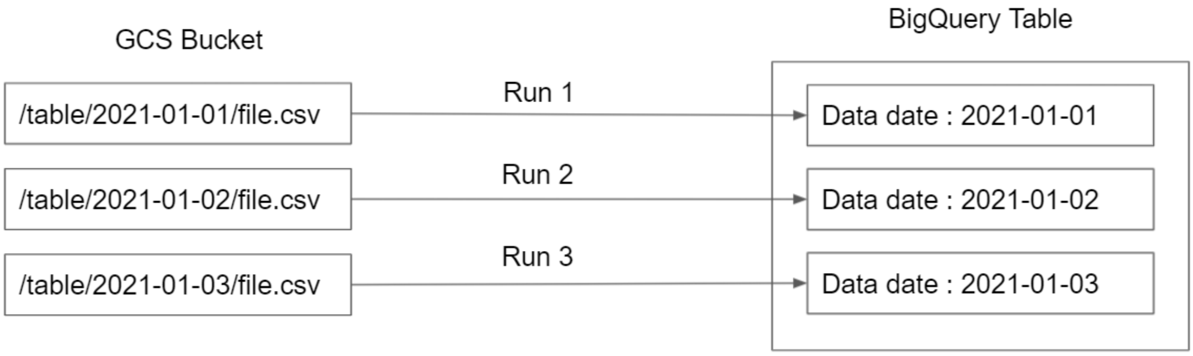


```
mysql> SELECT * FROM apps_db.stations LIMIT 10;
+------------+----------------------------------------------------+-----------+----------+
| station_id | name                                               | region_id | capacity |
+------------+----------------------------------------------------+-----------+----------+
| 501        | Balboa Park (San Jose Ave at Sgt. John V. Young Ln |           |        0 |
| 504        | Onondaga Ave at Alemany Blvd                       |           |        0 |
| 505        | Geneva Ave at Moscow St                            |           |        0 |
```

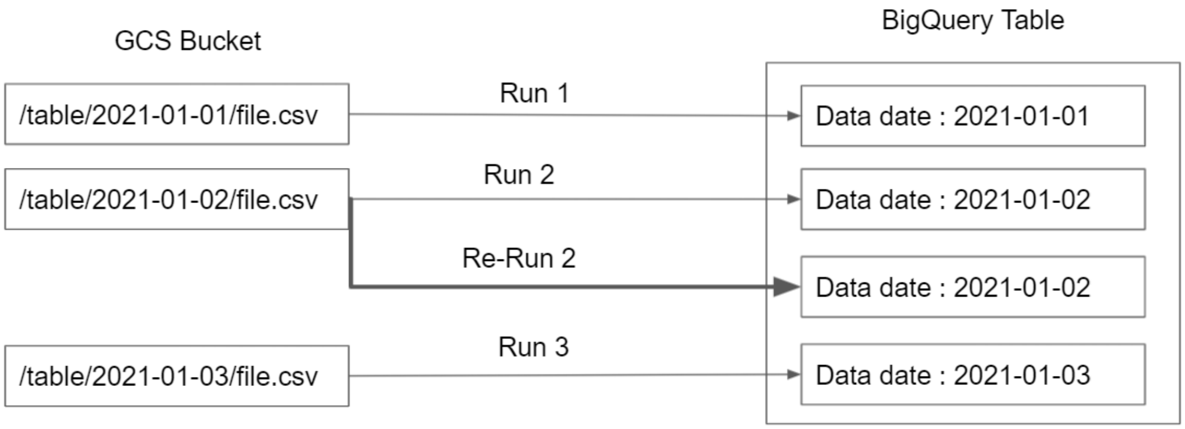




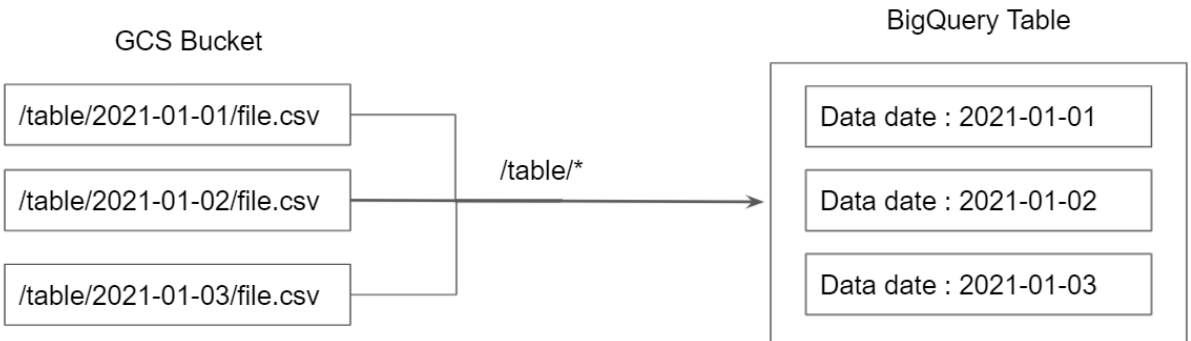

export_mysql_station
gcs_to_bq_station
dwh_fact_trips_daily
bq_row_count_check_dwh_fact_trips_daily
gcs_to_gcs_region
gcs_to_bq_region
dwh_dim_stations
bq_row_count_check_dwh_dim_stations
bq_to_bq_temporary_trips
bq_to_gcs_extract_trips
gcs_to_bq_trips
Backfill
2020-12-01
...
2021-01-01
...
Rerun
2021-01-01
...
Catchup
2021-01-01
...



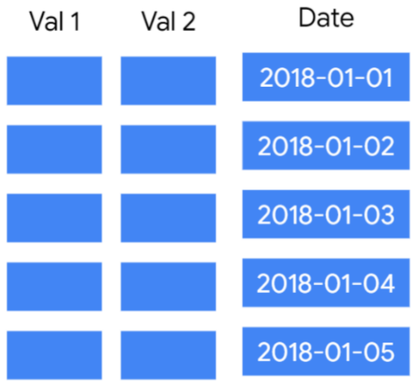
WRITE APPEND



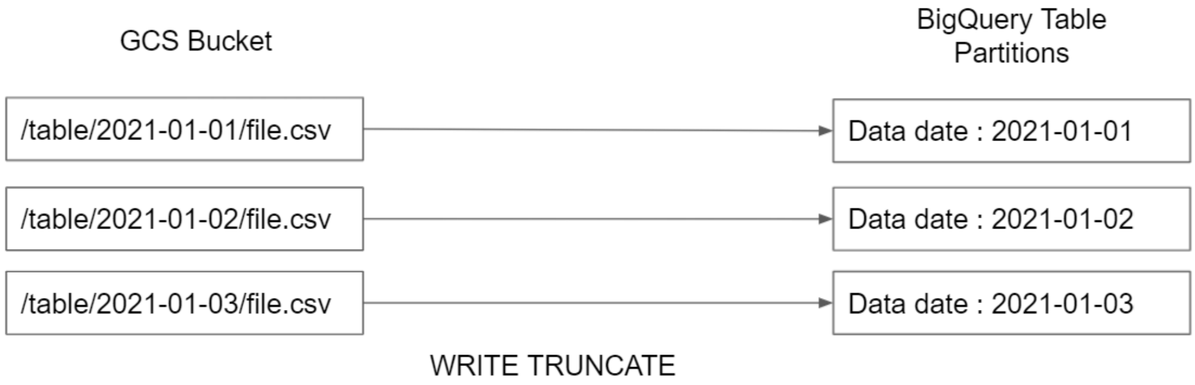
WRITE APPEND



WRITE TRUNCATE



Val 1	Val 2	Date
		2018-01-01
		2018-01-02
		2018-01-03
		2018-01-04
		2018-01-05



GCS Bucket

BigQuery Table Partitions

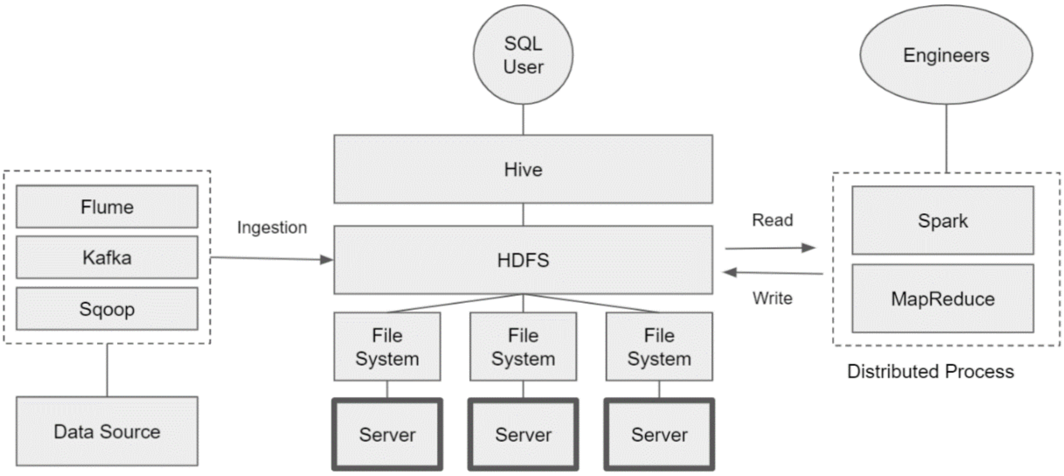
/table/2021-01-01/file.csv → Data date : 2021-01-01

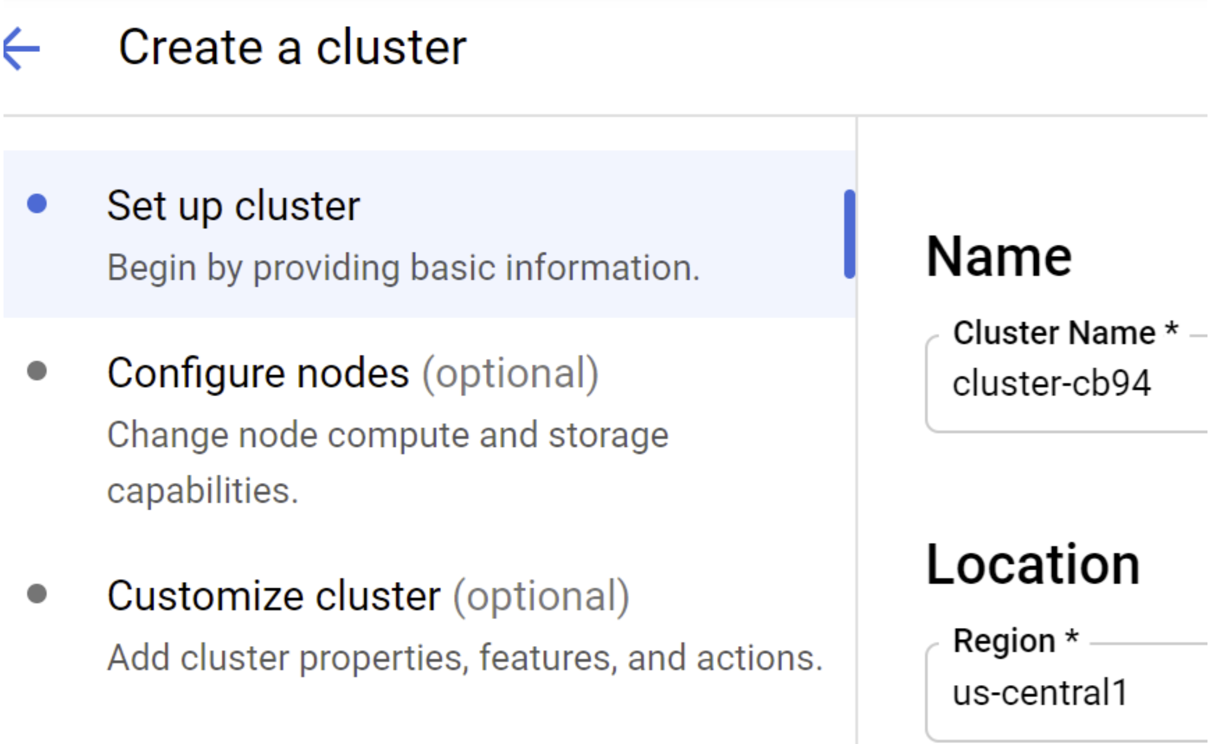
/table/2021-01-02/file.csv → Data date : 2021-01-02

/table/2021-01-03/file.csv → Data date : 2021-01-03

WRITE TRUNCATE

Chapter 5: Building a Data Lake Using Dataproc





Create a cluster

● Set up cluster
Begin by providing basic information.

● Configure nodes (optional)
Change node compute and storage capabilities.

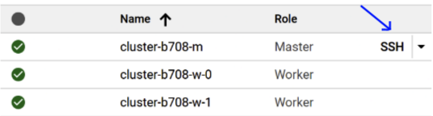
● Customize cluster (optional)
Add cluster properties, features, and actions.

Name

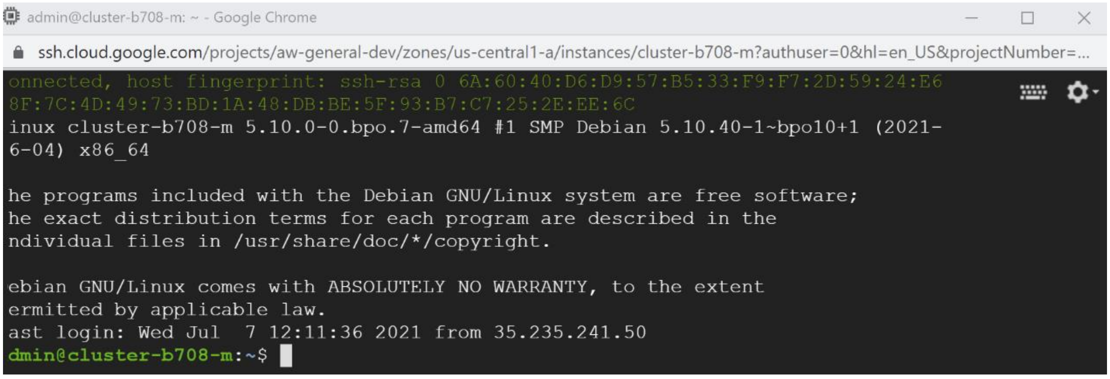
Cluster Name *
cluster-cb94

Location

Region *
us-central1

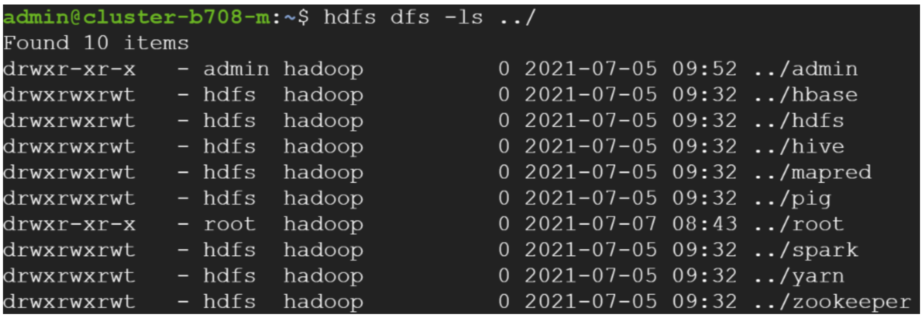


●	Name ↑	Role
✓	cluster-b708-m	Master
✓	cluster-b708-w-0	Worker
✓	cluster-b708-w-1	Worker

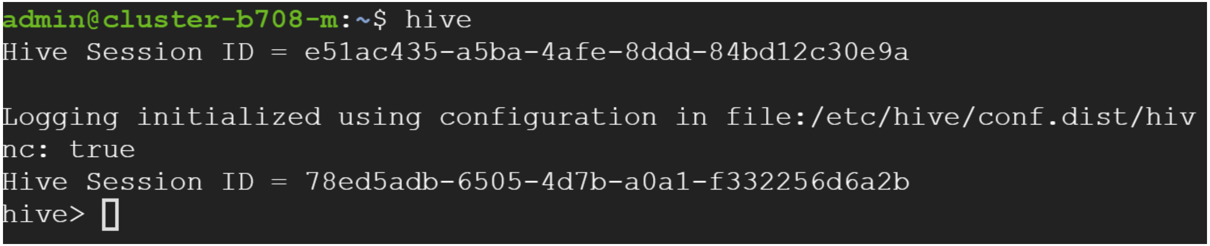


ssh.cloud.google.com/projects/aw-general-dev/zones/us-central1-a/instances/cluster-b708-m?authuser=0&hl=en_US&projectNumber=...

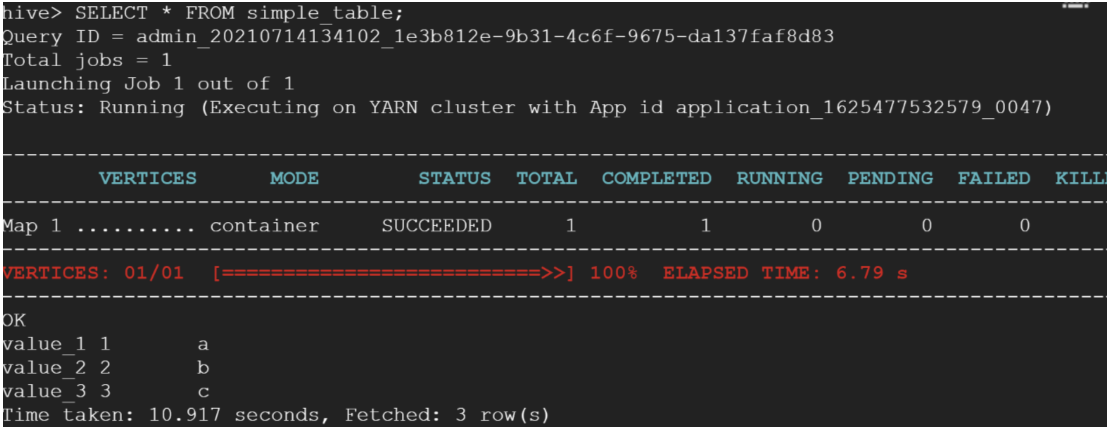
```
onnected, host fingerprint: ssh-rsa 0 6A:60:40:D6:D9:57:B5:33:F9:F7:2D:59:24:E6
8F:7C:4D:49:73:BD:1A:48:DB:BE:5F:93:B7:C7:25:2E:EE:6C
inux cluster-b708-m 5.10.0-0.bpo.7-amd64 #1 SMP Debian 5.10.40-1~bpo10+1 (2021-
6-04) x86_64

he programs included with the Debian GNU/Linux system are free software;
he exact distribution terms for each program are described in the
ndividual files in /usr/share/doc/*/copyright.

ebian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
ermitted by applicable law.
ast login: Wed Jul  7 12:11:36 2021 from 35.235.241.50
dmin@cluster-b708-m:~$ █
```



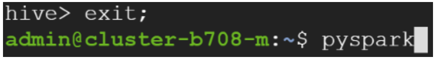
```
admin@cluster-b708-m:~$ hdfs dfs -ls ../
Found 10 items
drwxr-xr-x   - admin hadoop          0 2021-07-05 09:52 ../admin
drwxrwxrwt   - hdfs  hadoop          0 2021-07-05 09:32 ../hbase
drwxrwxrwt   - hdfs  hadoop          0 2021-07-05 09:32 ../hdfs
drwxrwxrwt   - hdfs  hadoop          0 2021-07-05 09:32 ../hive
drwxrwxrwt   - hdfs  hadoop          0 2021-07-05 09:32 ../mapred
drwxrwxrwt   - hdfs  hadoop          0 2021-07-05 09:32 ../pig
drwxr-xr-x   - root  hadoop          0 2021-07-07 08:43 ../root
drwxrwxrwt   - hdfs  hadoop          0 2021-07-05 09:32 ../spark
drwxrwxrwt   - hdfs  hadoop          0 2021-07-05 09:32 ../yarn
drwxrwxrwt   - hdfs  hadoop          0 2021-07-05 09:32 ../zookeeper
```



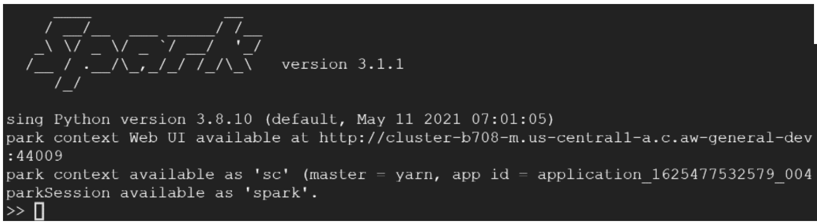
```
admin@cluster-b708-m:~$ hive
Hive Session ID = e51ac435-a5ba-4afe-8ddd-84bd12c30e9a

Logging initialized using configuration in file:/etc/hive/conf.dist/hiv
nc: true
Hive Session ID = 78ed5adb-6505-4d7b-a0a1-f332256d6a2b
hive> ▯
```



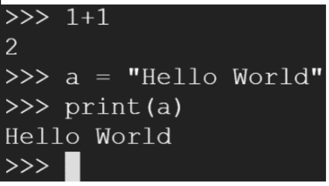
```
hive> SELECT * FROM simple_table;
Query ID = admin_20210714134102_1e3b812e-9b31-4c6f-9675-da137faf8d83
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1625477532579_0047)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILL
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.79 s
----------------------------------------------------------------------------------------------
OK
value_1 1       a
value_2 2       b
value_3 3       c
Time taken: 10.917 seconds, Fetched: 3 row(s)
```



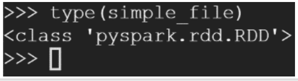
```
hive> exit;
admin@cluster-b708-m:~$ pyspark
```



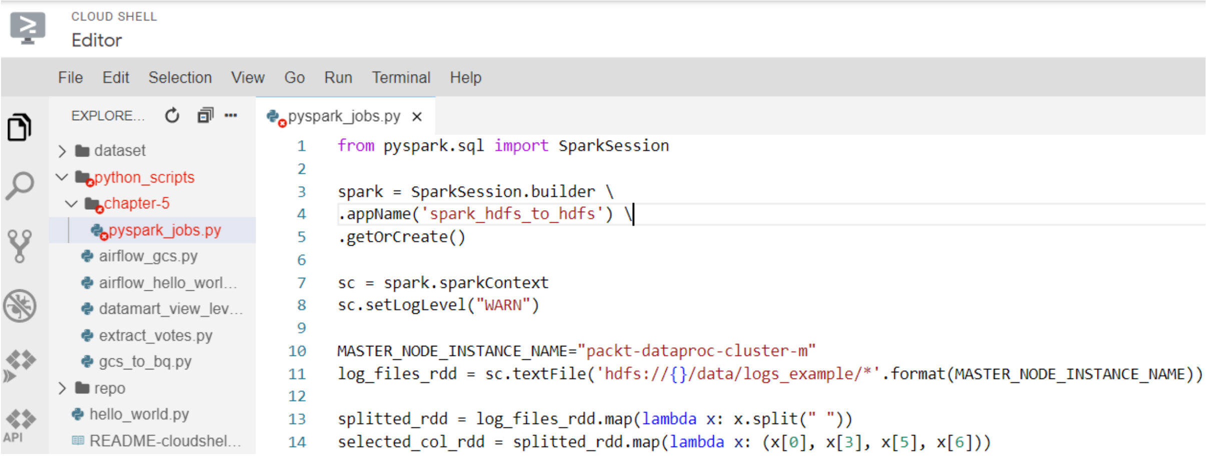
```
    ____              __
   / __/__  ___ _____/ /__
  _\ \/ _ \/ _ `/ __/  '_/
 /__ / .__/\_,_/_/ /_/\_\   version 3.1.1
    /_/

sing Python version 3.8.10 (default, May 11 2021 07:01:05)
park context Web UI available at http://cluster-b708-m.us-central1-a.c.aw-general-dev
:44009
park context available as 'sc' (master = yarn, app id = application_1625477532579_004
parkSession available as 'spark'.
>> ▯
```



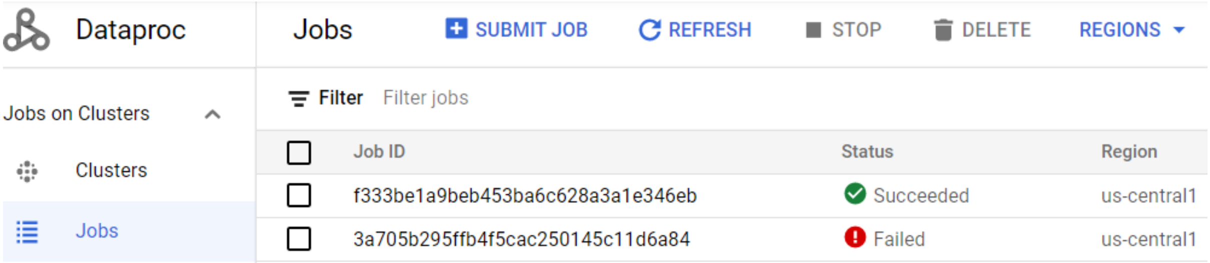
```
>>> 1+1
2
>>> a = "Hello World"
>>> print(a)
Hello World
>>> 
```



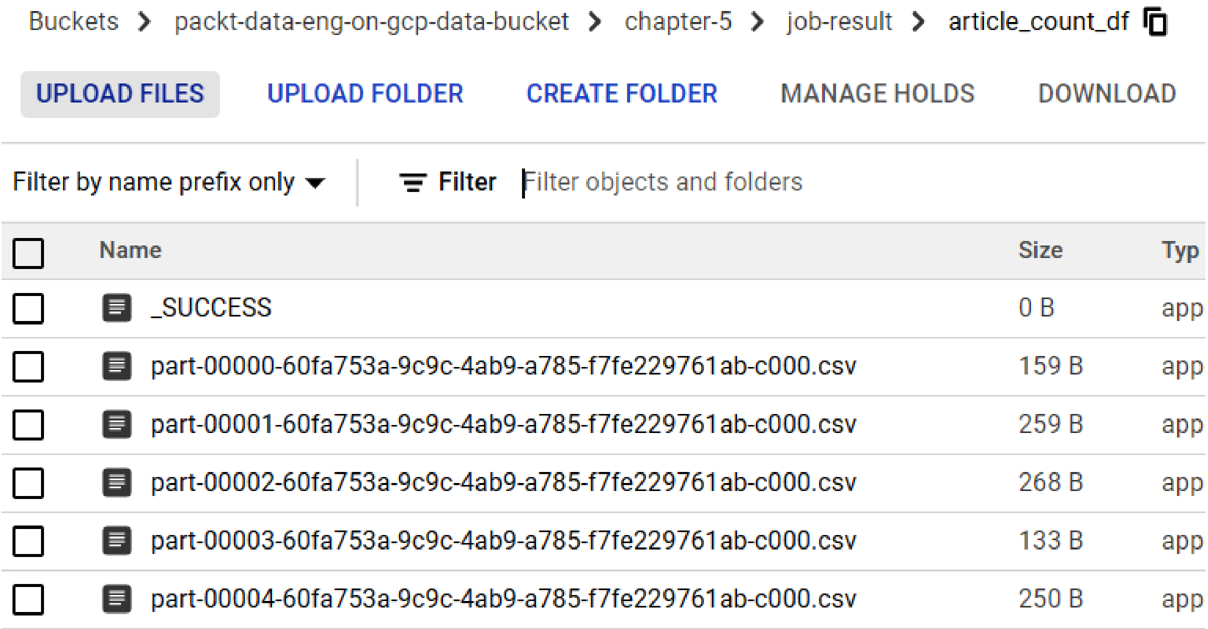
```
>>> type(simple_file)
<class 'pyspark.rdd.RDD'>
>>> ▯
```



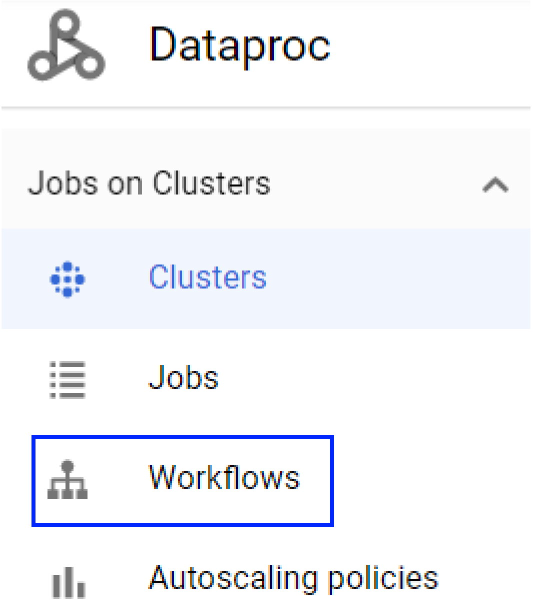
```python
from pyspark.sql import SparkSession

spark = SparkSession.builder \
.appName('spark_hdfs_to_hdfs') \
.getOrCreate()

sc = spark.sparkContext
sc.setLogLevel("WARN")

MASTER_NODE_INSTANCE_NAME="packt-dataproc-cluster-m"
log_files_rdd = sc.textFile('hdfs://{}/data/logs_example/*'.format(MASTER_NODE_INSTANCE_NAME))

splitted_rdd = log_files_rdd.map(lambda x: x.split(" "))
selected_col_rdd = splitted_rdd.map(lambda x: (x[0], x[3], x[5], x[6]))
```

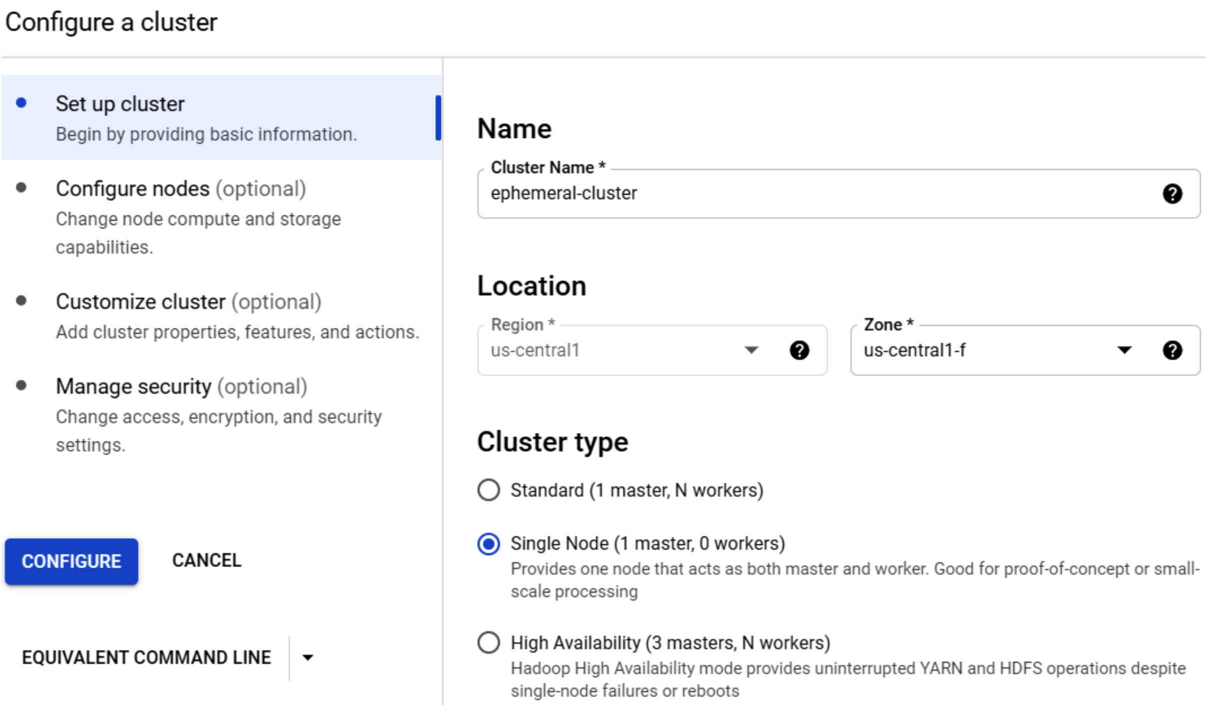


□	Job ID	Status	Region
□	f333be1a9beb453ba6c628a3a1e346eb	✓ Succeeded	us-central1
□	3a705b295ffb4f5cac250145c11d6a84	! Failed	us-central1


Buckets > packt-data-eng-on-gcp-data-bucket > chapter-5 > job-result > article_count_df
UPLOAD FILES
UPLOAD FOLDER
CREATE FOLDER
MANAGE HOLDS
DOWNLOAD
Filter by name prefix only
Filter
Filter objects and folders

□	Name	Size	Typ
□	_SUCCESS	0 B	app
□	part-00000-60fa753a-9c9c-4ab9-a785-f7fe229761ab-c000.csv	159 B	app
□	part-00001-60fa753a-9c9c-4ab9-a785-f7fe229761ab-c000.csv	259 B	app
□	part-00002-60fa753a-9c9c-4ab9-a785-f7fe229761ab-c000.csv	268 B	app
□	part-00003-60fa753a-9c9c-4ab9-a785-f7fe229761ab-c000.csv	133 B	app
□	part-00004-60fa753a-9c9c-4ab9-a785-f7fe229761ab-c000.csv	250 B	app





Configure a cluster

- Set up cluster
  Begin by providing basic information.
- Configure nodes (optional)
  Change node compute and storage capabilities.
- Customize cluster (optional)
  Add cluster properties, features, and actions.
- Manage security (optional)
  Change access, encryption, and security settings.

CONFIGURE   CANCEL

EQUIVALENT COMMAND LINE

Name

Cluster Name *: ephemeral-cluster

Location

Region *: us-central1

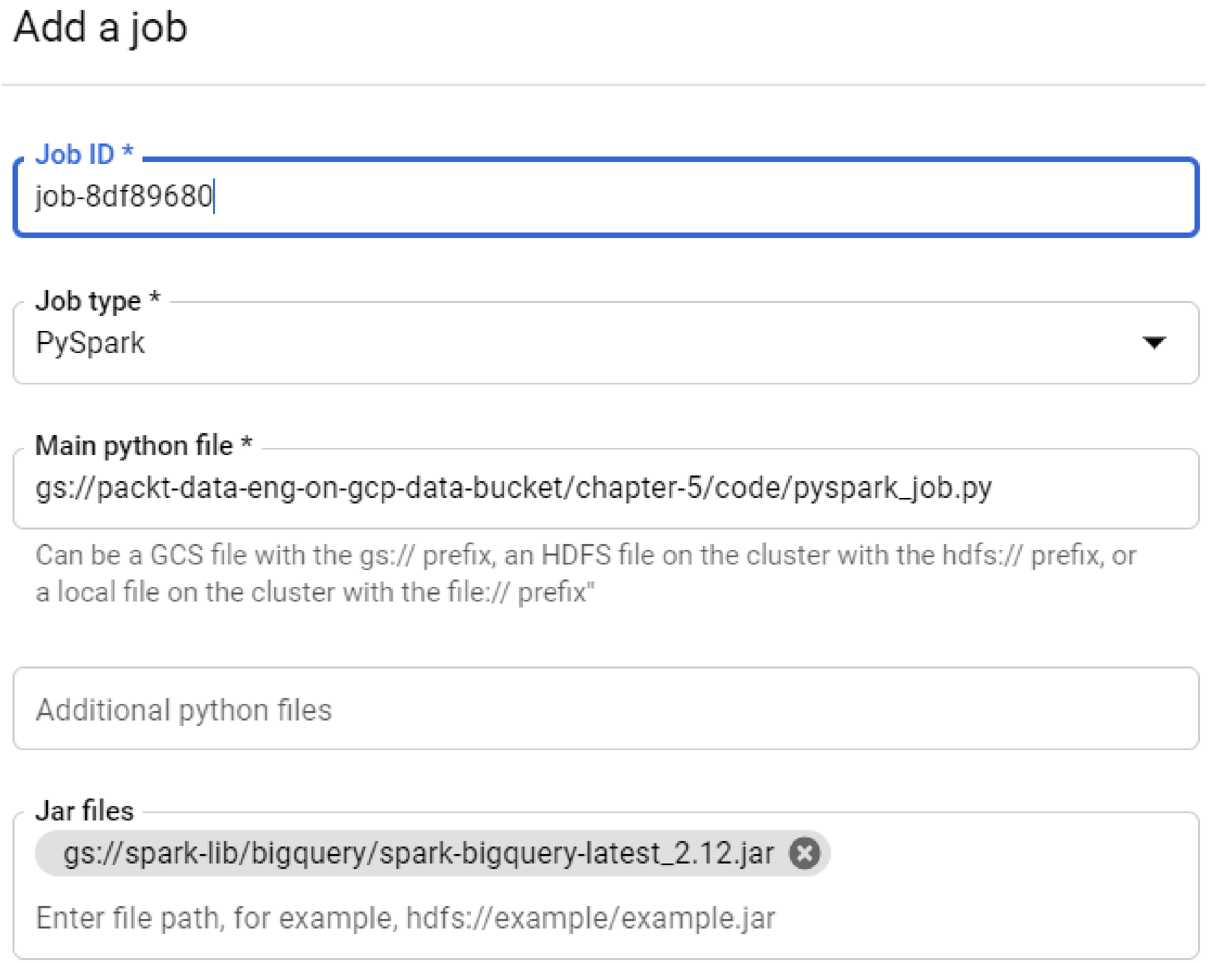
Zone *: us-central1-f

Cluster type

○ Standard (1 master, N workers)

● Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

○ High Availability (3 masters, N workers)
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots



Add a job

Job ID *: job-8df89680

Job type *: PySpark

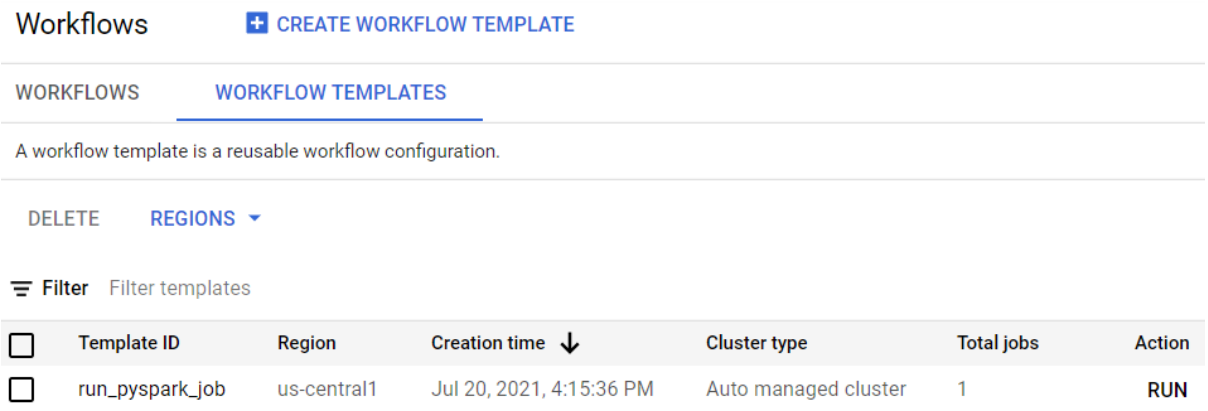
Main python file *: gs://packt-data-eng-on-gcp-data-bucket/chapter-5/code/pyspark_job.py

Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix"

Additional python files

Jar files: gs://spark-lib/bigquery/spark-bigquery-latest_2.12.jar

Enter file path, for example, hdfs://example/example.jar



Workflows CREATE WORKFLOW TEMPLATE

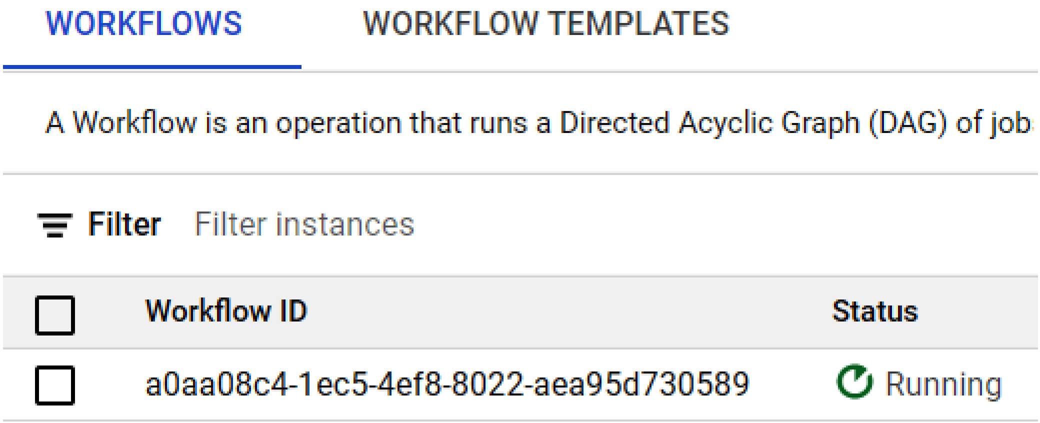
WORKFLOWS WORKFLOW TEMPLATES

A workflow template is a reusable workflow configuration.

DELETE REGIONS ▾

Filter Filter templates

□	Template ID	Region	Creation time ↓	Cluster type	Total jobs	Action
□	run_pyspark_job	us-central1	Jul 20, 2021, 4:15:36 PM	Auto managed cluster	1	RUN

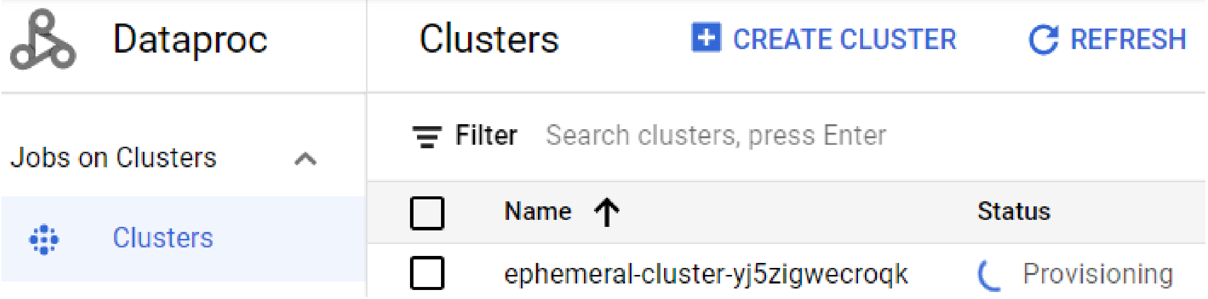


WORKFLOWS WORKFLOW TEMPLATES

A Workflow is an operation that runs a Directed Acyclic Graph (DAG) of job

Filter Filter instances

□	Workflow ID	Status
□	a0aa08c4-1ec5-4ef8-8022-aea95d730589	Running



Dataproc Clusters CREATE CLUSTER REFRESH

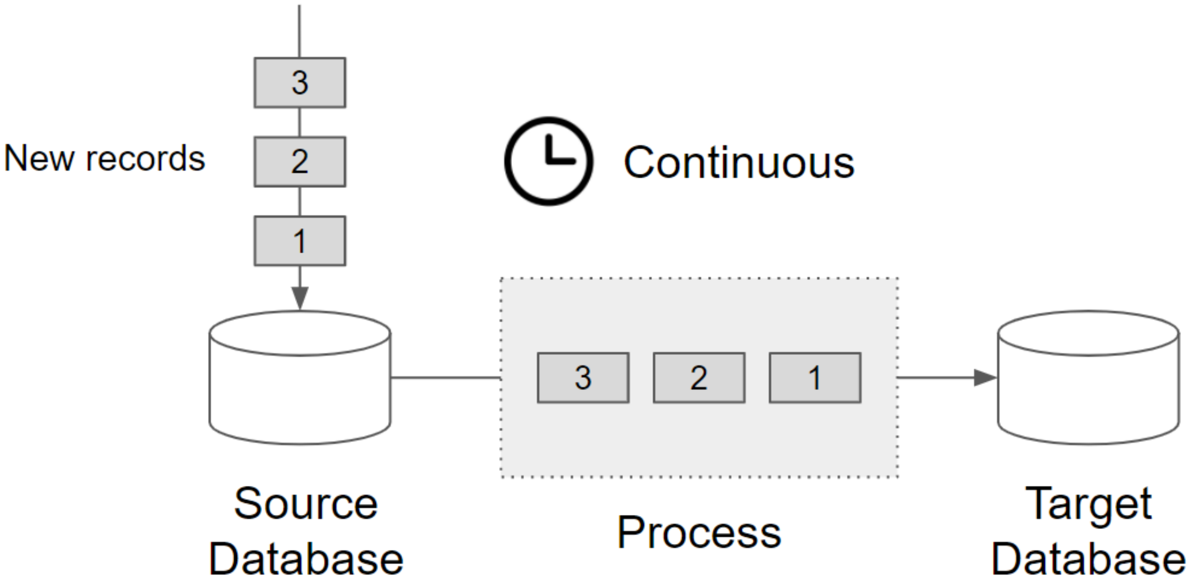
Jobs on Clusters

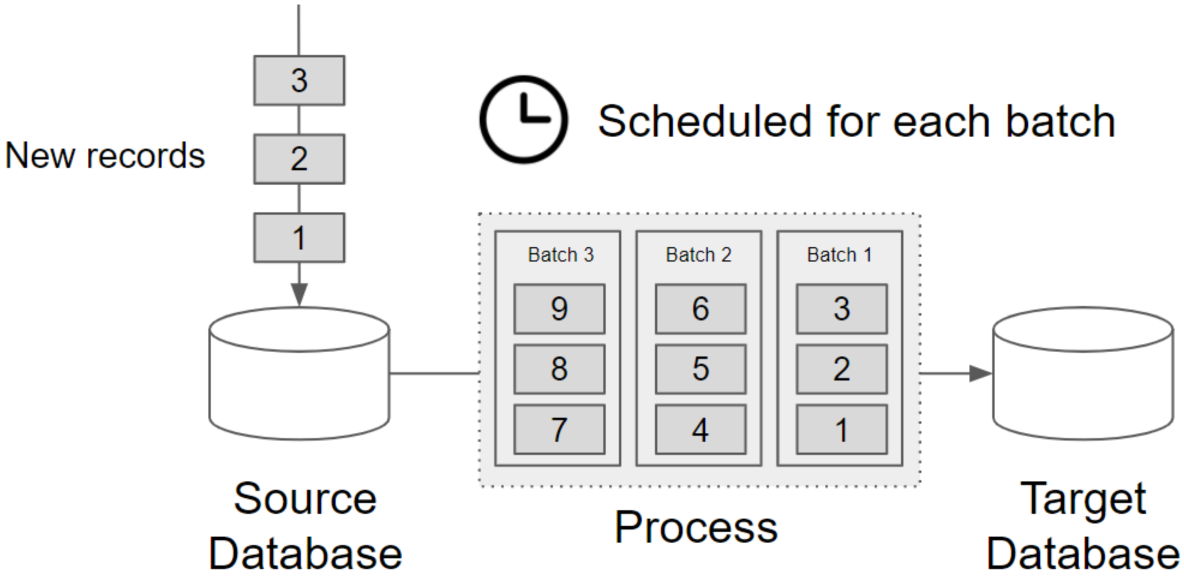
Clusters

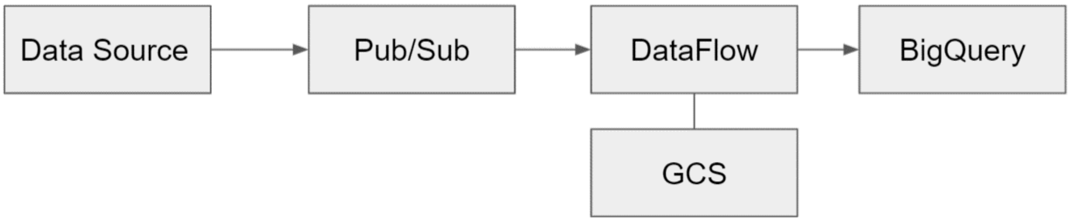
Filter Search clusters, press Enter

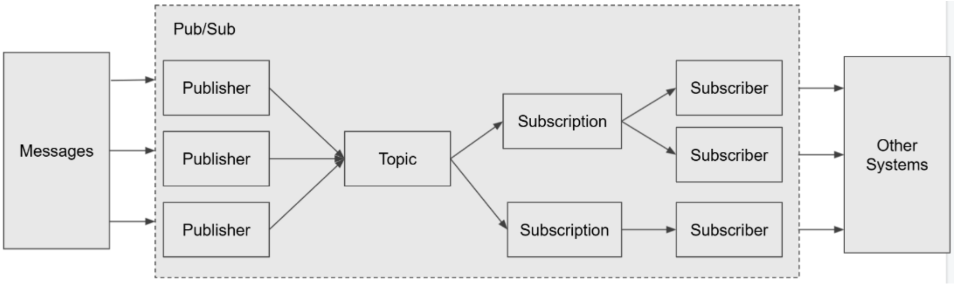
□	Name ↑	Status
□	ephemeral-cluster-yj5zigwecroqk	Provisioning

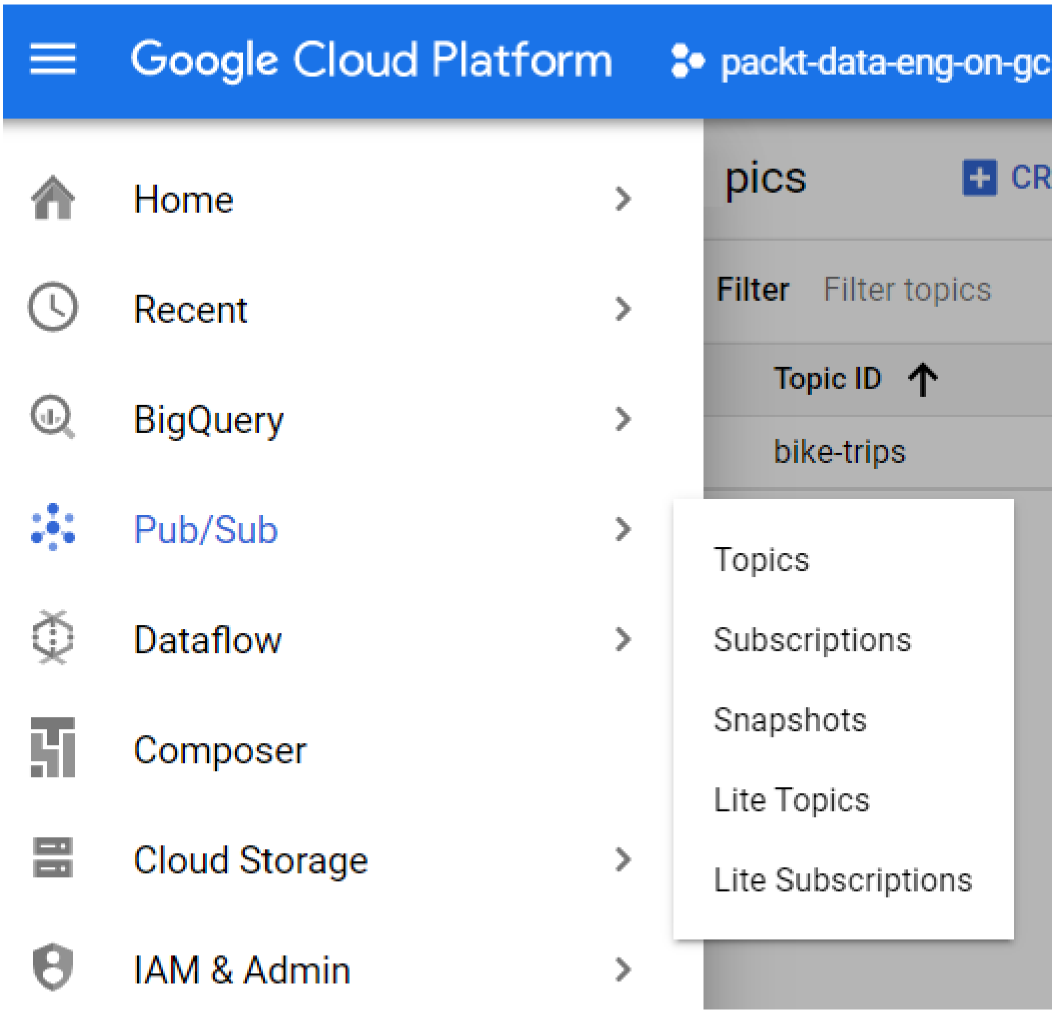
Chapter 6: Processing Streaming Data with Pub/Sub and Dataflow

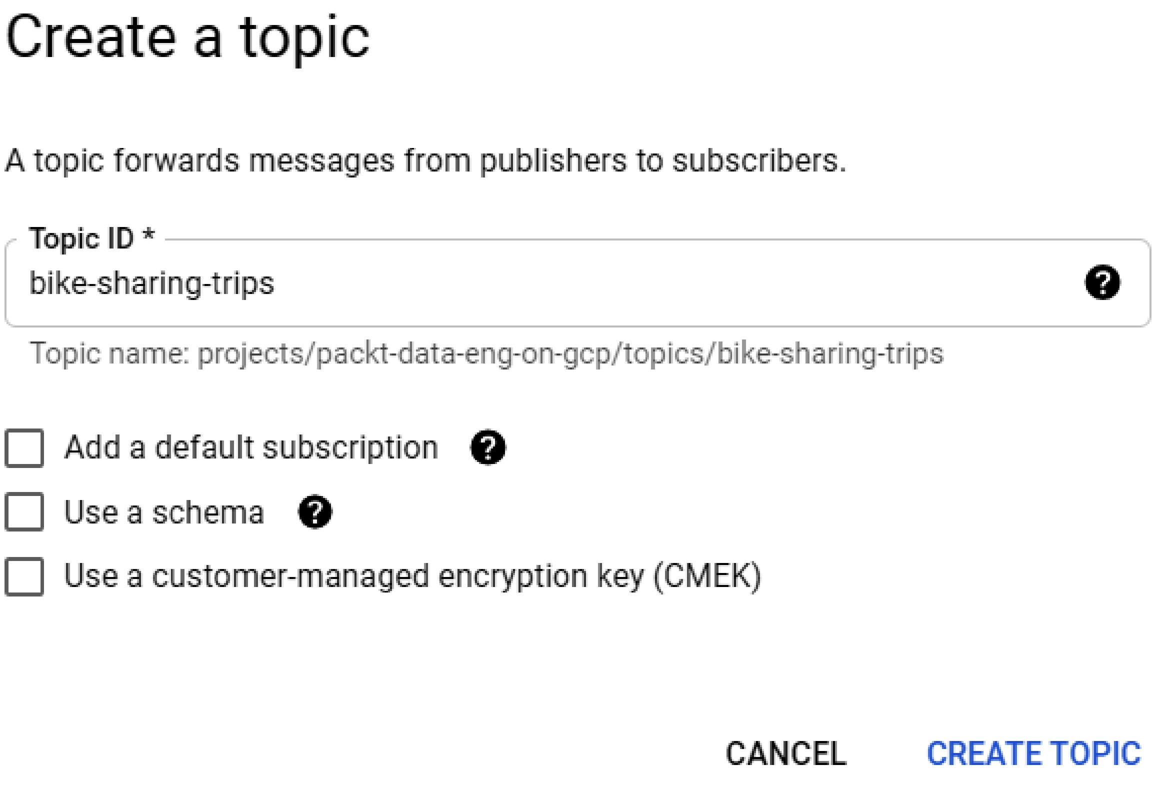











Create a topic

A topic forwards messages from publishers to subscribers.

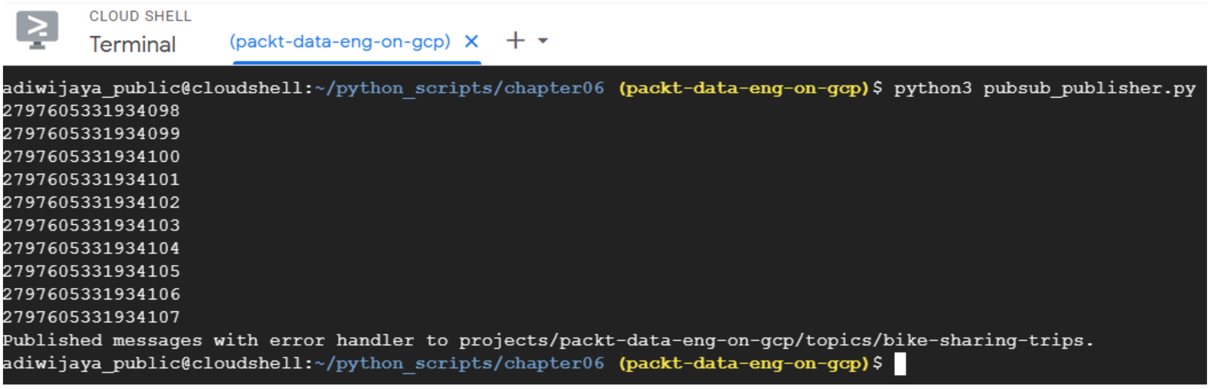
Topic ID *: bike-sharing-trips

Topic name: projects/packt-data-eng-on-gcp/topics/bike-sharing-trips

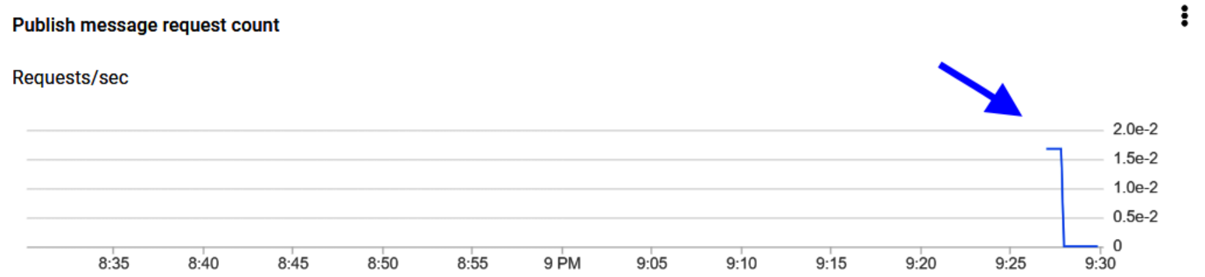
□ Add a default subscription

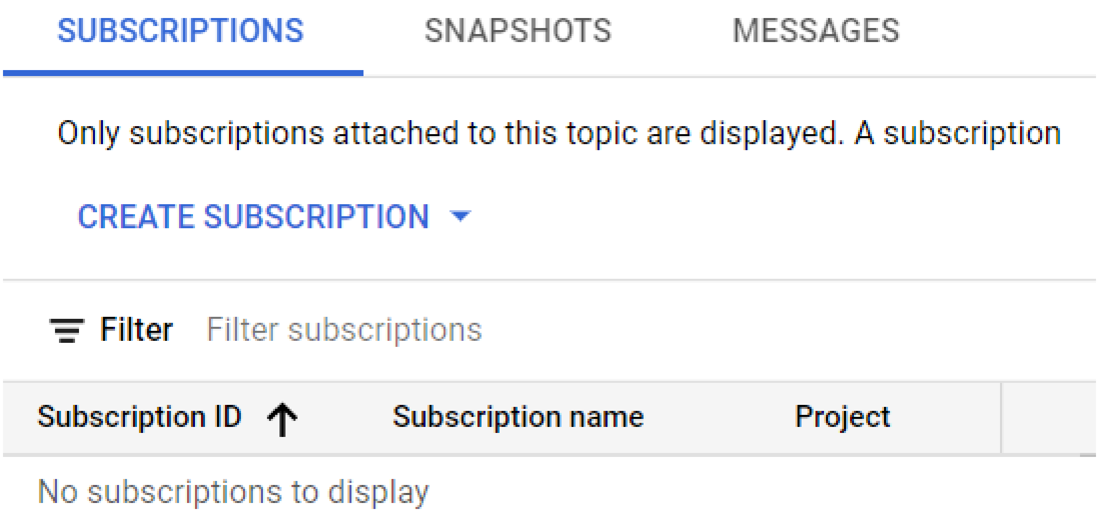
□ Use a schema

□ Use a customer-managed encryption key (CMEK)



```
adiwijaya_public@cloudshell:~/python_scripts/chapter06 (packt-data-eng-on-gcp)$ python3 pubsub_publisher.py
2797605331934098
2797605331934099
2797605331934100
2797605331934101
2797605331934102
2797605331934103
2797605331934104
2797605331934105
2797605331934106
2797605331934107
Published messages with error handler to projects/packt-data-eng-on-gcp/topics/bike-sharing-trips.
adiwijaya_public@cloudshell:~/python_scripts/chapter06 (packt-data-eng-on-gcp)$
```



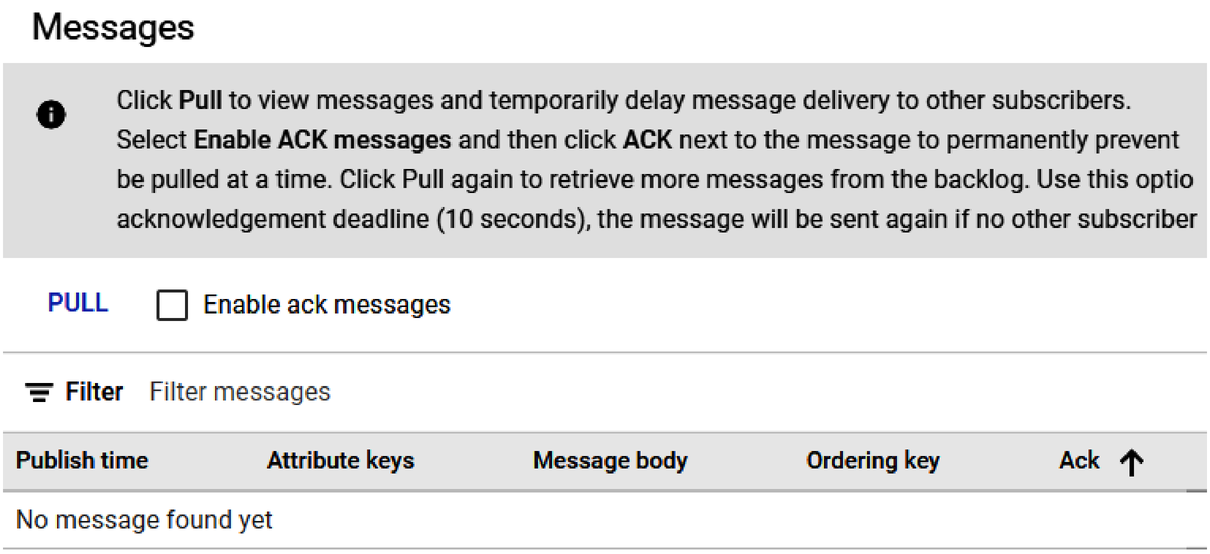


Only subscriptions attached to this topic are displayed. A subscription

Filter subscriptions

Subscription ID ↑	Subscription name	Project

No subscriptions to display


Messages

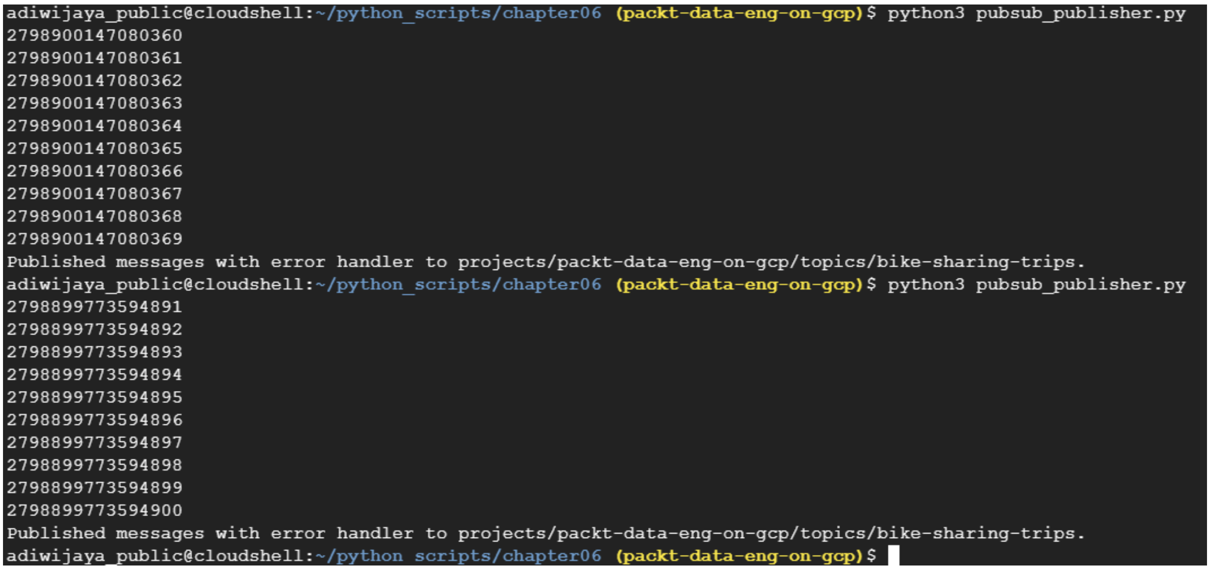
Click **Pull** to view messages and temporarily delay message delivery to other subscribers. Select **Enable ACK messages** and then click **ACK** next to the message to permanently prevent be pulled at a time. Click Pull again to retrieve more messages from the backlog. Use this optio acknowledgement deadline (10 seconds), the message will be sent again if no other subscriber

□ Enable ack messages

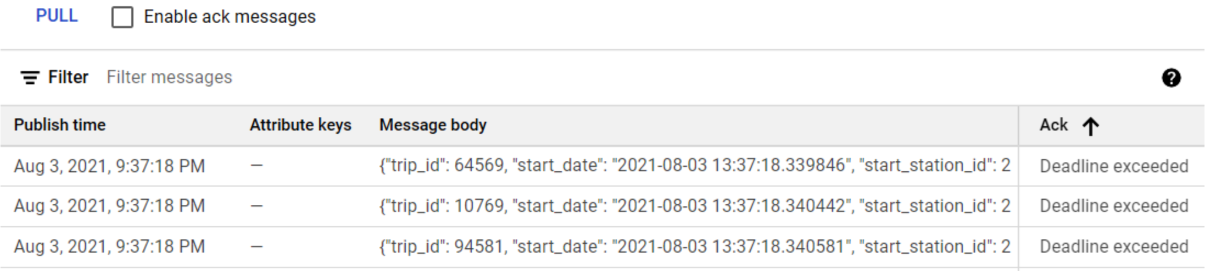
Filter messages

Publish time	Attribute keys	Message body	Ordering key	Ack ↑

No message found yet



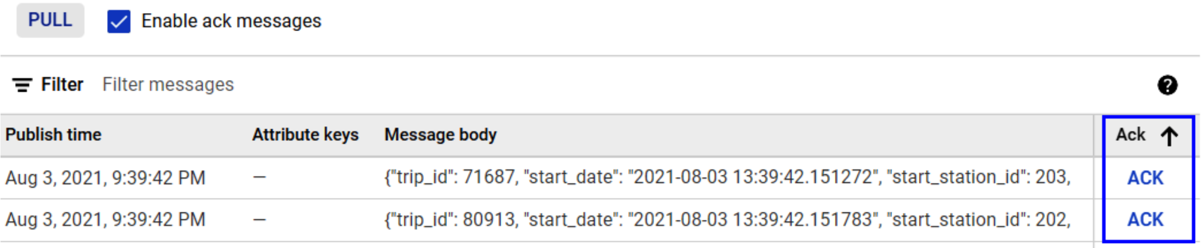
```
adiwijaya_public@cloudshell:~/python_scripts/chapter06 (packt-data-eng-on-gcp)$ python3 pubsub_publisher.py
2798900147080360
2798900147080361
2798900147080362
2798900147080363
2798900147080364
2798900147080365
2798900147080366
2798900147080367
2798900147080368
2798900147080369
Published messages with error handler to projects/packt-data-eng-on-gcp/topics/bike-sharing-trips.
adiwijaya_public@cloudshell:~/python_scripts/chapter06 (packt-data-eng-on-gcp)$ python3 pubsub_publisher.py
2798899773594891
2798899773594892
2798899773594893
2798899773594894
2798899773594895
2798899773594896
2798899773594897
2798899773594898
2798899773594899
2798899773594900
Published messages with error handler to projects/packt-data-eng-on-gcp/topics/bike-sharing-trips.
adiwijaya_public@cloudshell:~/python_scripts/chapter06 (packt-data-eng-on-gcp)$
```



☐ Enable ack messages

Filter Filter messages

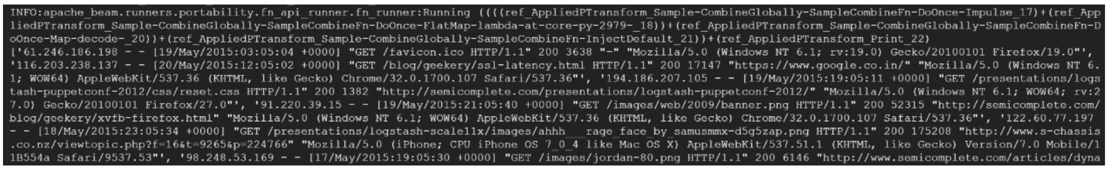
Publish time	Attribute keys	Message body	Ack ↑
Aug 3, 2021, 9:37:18 PM	—	{"trip_id": 64569, "start_date": "2021-08-03 13:37:18.339846", "start_station_id": 2	Deadline exceeded
Aug 3, 2021, 9:37:18 PM	—	{"trip_id": 10769, "start_date": "2021-08-03 13:37:18.340442", "start_station_id": 2	Deadline exceeded
Aug 3, 2021, 9:37:18 PM	—	{"trip_id": 94581, "start_date": "2021-08-03 13:37:18.340581", "start_station_id": 2	Deadline exceeded



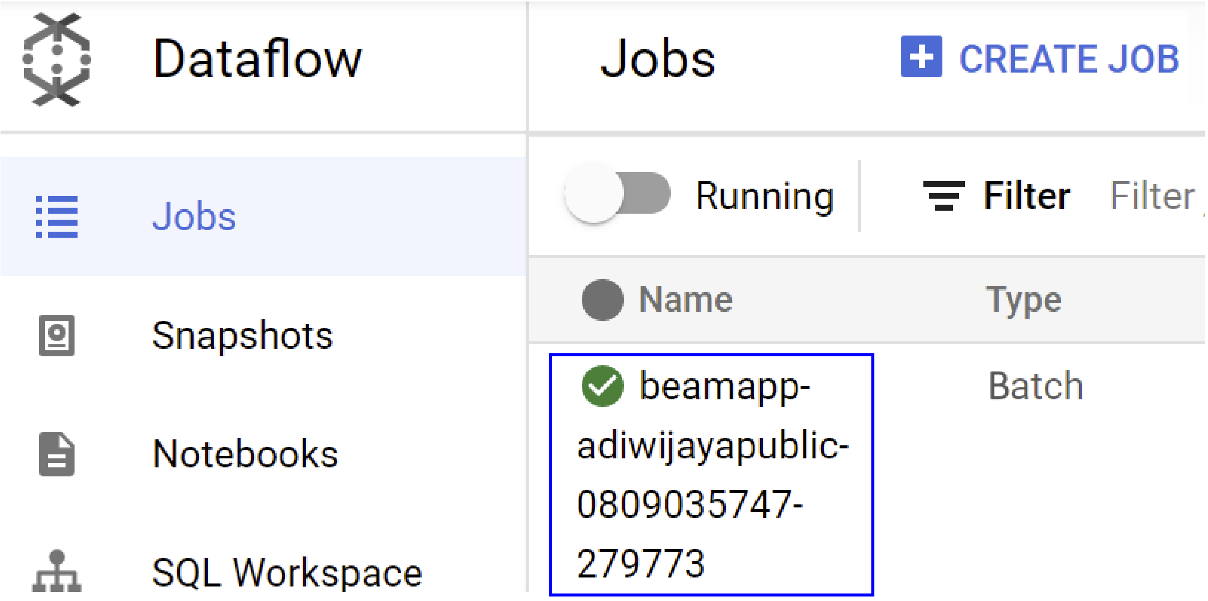
☑ Enable ack messages

Filter Filter messages

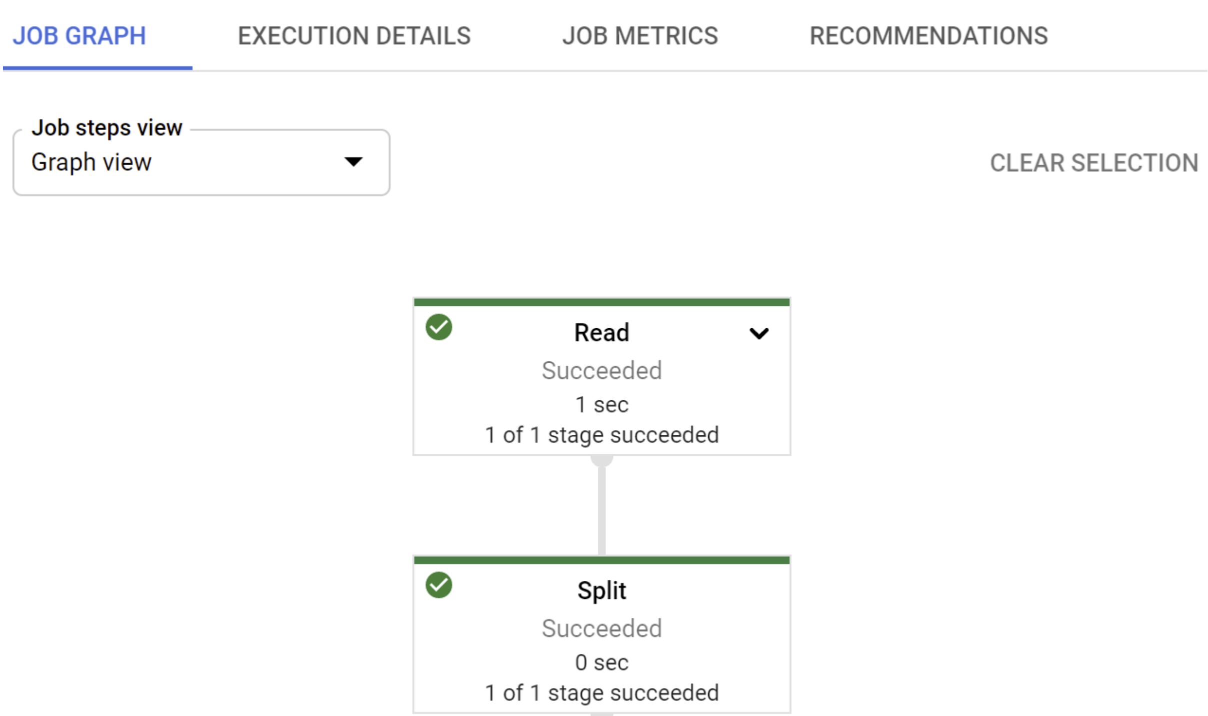
Publish time	Attribute keys	Message body	Ack ↑
Aug 3, 2021, 9:39:42 PM	—	{"trip_id": 71687, "start_date": "2021-08-03 13:39:42.151272", "start_station_id": 203,	ACK
Aug 3, 2021, 9:39:42 PM	—	{"trip_id": 80913, "start_date": "2021-08-03 13:39:42.151783", "start_station_id": 202,	ACK



```
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running ((((ref_AppliedPTransform_Sample-CombineGlobally-SampleCombineFn-DoOnce-Impulse_17)+(ref_App
liedPTransform_Sample-CombineGlobally-SampleCombineFn-DoOnce-FlatMap-lambda-at-core-py-2979-_18))+(ref_AppliedPTransform_Sample-CombineGlobally-SampleCombineFn-D
oOnce-Map-decode-_20))+(ref_AppliedPTransform_Sample-CombineGlobally-SampleCombineFn-InjectDefault_21))+(ref_AppliedPTransform_Print_22)
['61.246.186.198 - - [19/May/2015:03:05:04 +0000] "GET /favicon.ico HTTP/1.1" 200 3638 "-" "Mozilla/5.0 (Windows NT 6.1; rv:19.0) Gecko/20100101 Firefox/19.0"',
'116.203.238.137 - - [20/May/2015:12:05:02 +0000] "GET /blog/geekery/ssl-latency.html HTTP/1.1" 200 17147 "https://www.google.co.in/" "Mozilla/5.0 (Windows NT 6.
1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/537.36"', '194.186.207.105 - - [19/May/2015:19:05:11 +0000] "GET /presentations/logs
tash-puppetconf-2012/css/reset.css HTTP/1.1" 200 1382 "http://semicomplete.com/presentations/logstash-puppetconf-2012/" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:2
7.0) Gecko/20100101 Firefox/27.0"', '91.220.39.15 - - [19/May/2015:21:05:40 +0000] "GET /images/web/2009/banner.png HTTP/1.1" 200 52315 "http://semicomplete.com/
blog/geekery/xvfb-firefox.html" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/537.36"', '122.60.77.197
- - [18/May/2015:23:05:34 +0000] "GET /presentations/logstash-scale11x/images/ahhh___rage_face_by_samusmmx-d5g5zap.png HTTP/1.1" 200 175208 "http://www.s-chassis
.co.nz/viewtopic.php?f=16&t=9265&p=224766" "Mozilla/5.0 (iPhone; CPU iPhone OS 7_0_4 like Mac OS X) AppleWebKit/537.51.1 (KHTML, like Gecko) Version/7.0 Mobile/1
1B554a Safari/9537.53"', '98.248.53.169 - - [17/May/2015:19:05:30 +0000] "GET /images/jordan-80.png HTTP/1.1" 200 6146 "http://www.semicomplete.com/articles/dyna
```



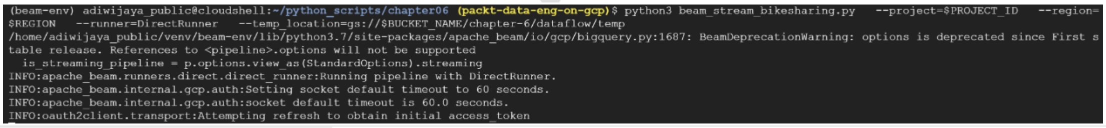
Name	Type
beamapp-adiwijayapublic-0809035747-279773	Batch



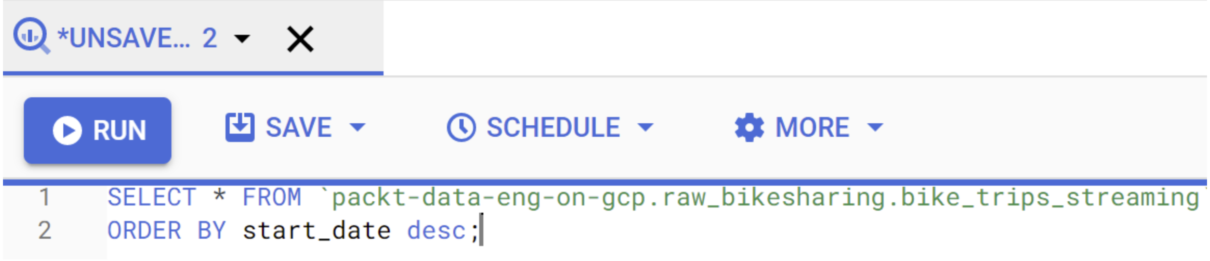
Job steps view: Graph view

Read
Succeeded
1 sec
1 of 1 stage succeeded

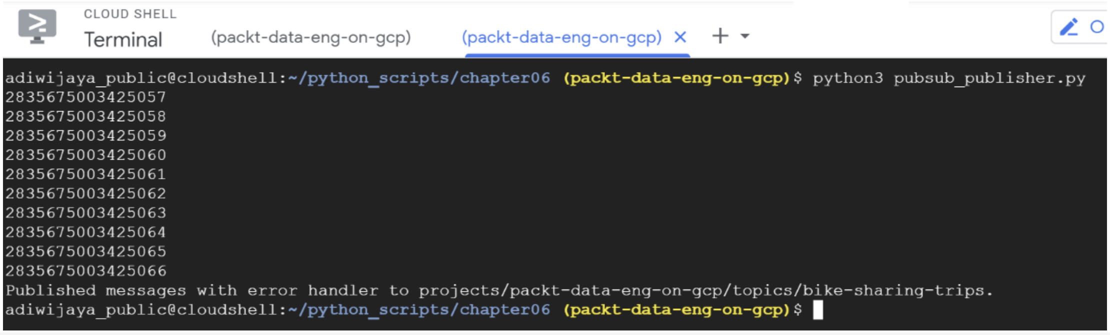
Split
Succeeded
0 sec
1 of 1 stage succeeded



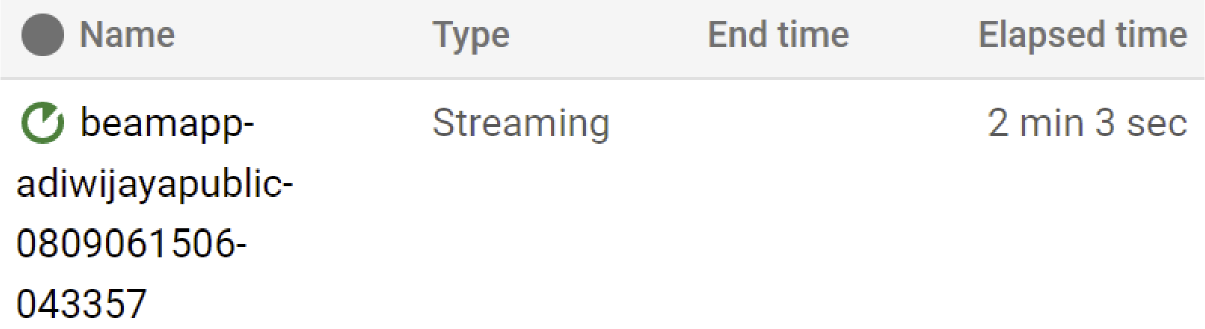
```
(beam-env) adiwijaya_public@cloudshell:~/python_scripts/chapter06 (packt-data-eng-on-gcp)$ python3 beam_stream_bikesharing.py   --project=$PROJECT_ID   --region=
$REGION   --runner=DirectRunner   --temp_location=gs://$BUCKET_NAME/chapter-6/dataflow/temp
/home/adiwijaya_public/venv/beam-env/lib/python3.7/site-packages/apache_beam/io/gcp/bigquery.py:1687: BeamDeprecationWarning: options is deprecated since First s
table release. References to <pipeline>.options will not be supported
  is_streaming_pipeline = p.options.view_as(StandardOptions).streaming
INFO:apache_beam.runners.direct.direct_runner:Running pipeline with DirectRunner.
INFO:apache_beam.internal.gcp.auth:Setting socket default timeout to 60 seconds.
INFO:apache_beam.internal.gcp.auth:socket default timeout is 60.0 seconds.
INFO:oauth2client.transport:Attempting refresh to obtain initial access_token
```



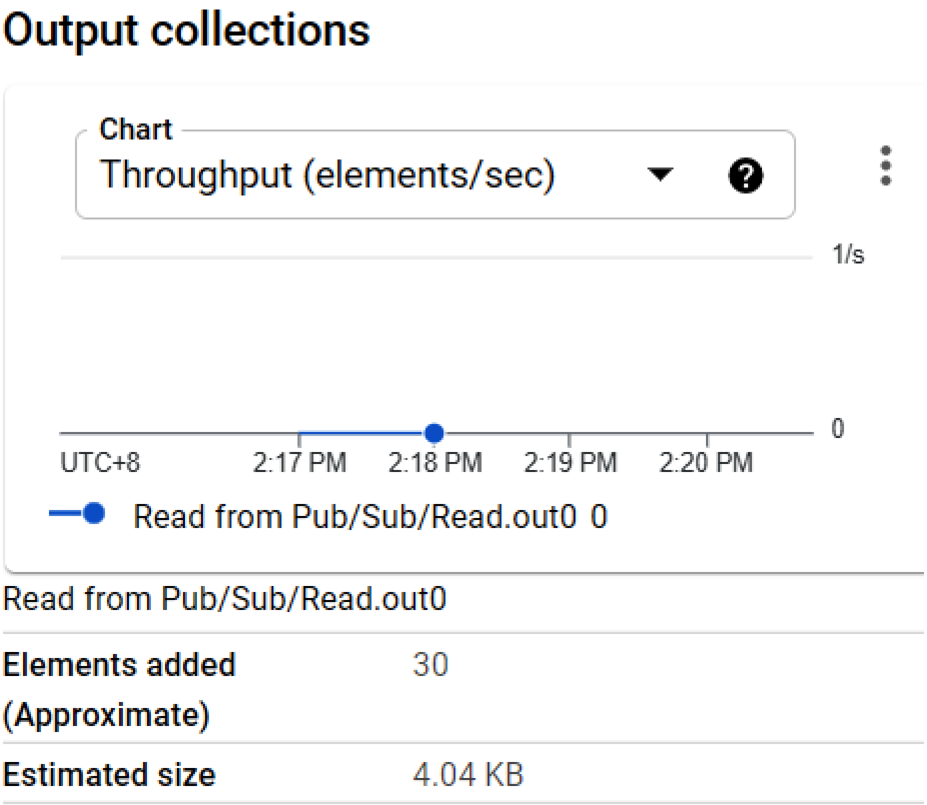
```sql
SELECT * FROM `packt-data-eng-on-gcp.raw_bikesharing.bike_trips_streaming
ORDER BY start_date desc;
```



```
adiwijaya_public@cloudshell:~/python_scripts/chapter06 (packt-data-eng-on-gcp)$ python3 pubsub_publisher.py
2835675003425057
2835675003425058
2835675003425059
2835675003425060
2835675003425061
2835675003425062
2835675003425063
2835675003425064
2835675003425065
2835675003425066
Published messages with error handler to projects/packt-data-eng-on-gcp/topics/bike-sharing-trips.
adiwijaya_public@cloudshell:~/python_scripts/chapter06 (packt-data-eng-on-gcp)$
```



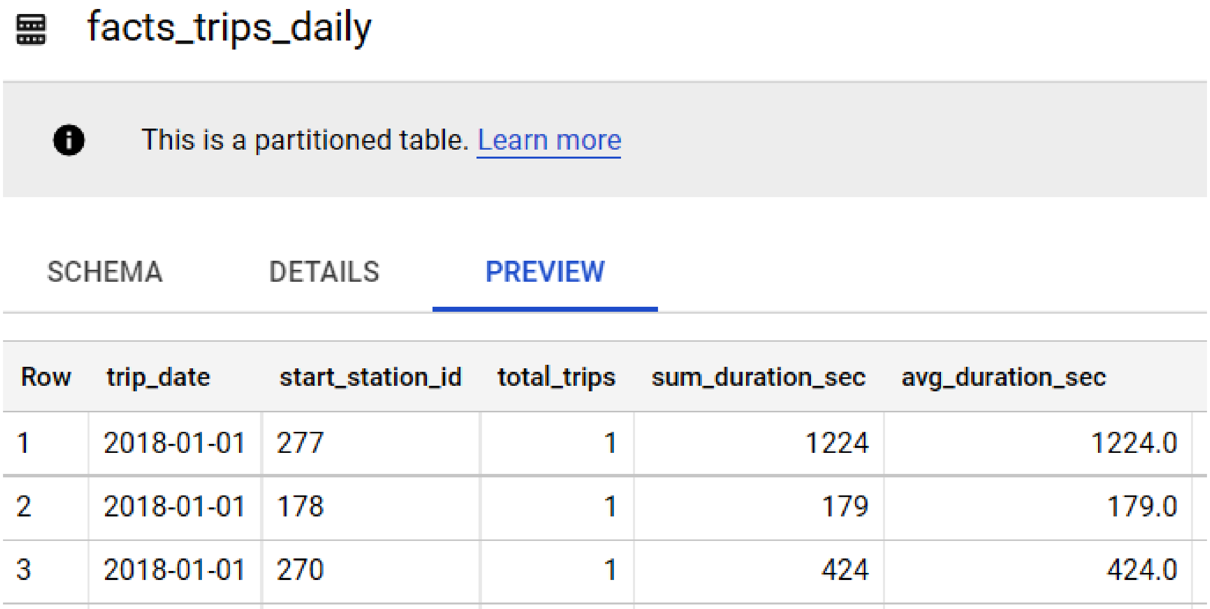
Name	Type	End time	Elapsed time
beamapp-adiwijayapublic-0809061506-043357	Streaming		2 min 3 sec



Output collections

Read from Pub/Sub/Read.out0

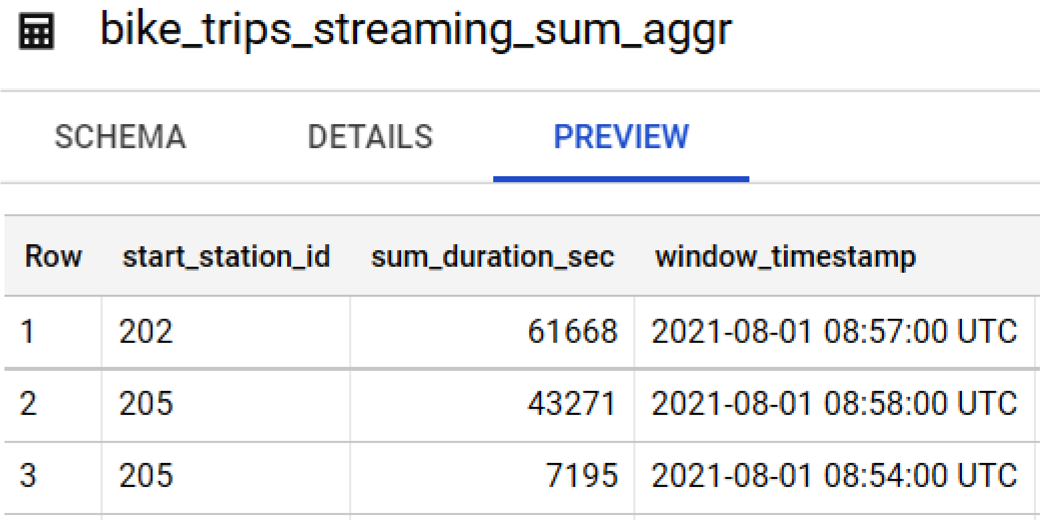
Elements added (Approximate)	30
Estimated size	4.04 KB



facts_trips_daily

This is a partitioned table. Learn more

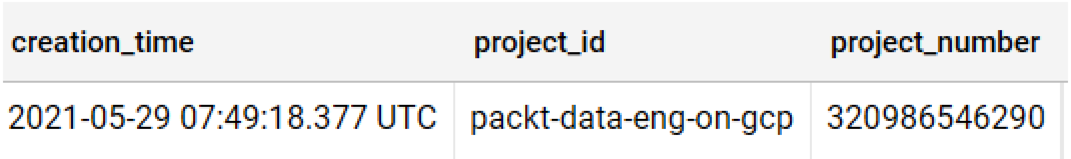
Row	trip_date	start_station_id	total_trips	sum_duration_sec	avg_duration_sec
1	2018-01-01	277	1	1224	1224.0
2	2018-01-01	178	1	179	179.0
3	2018-01-01	270	1	424	424.0



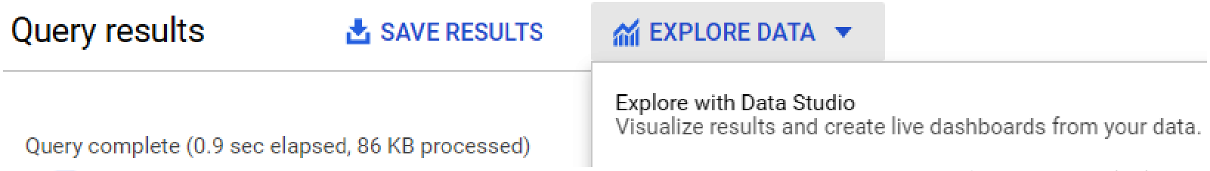
bike_trips_streaming_sum_aggr

Row	start_station_id	sum_duration_sec	window_timestamp
1	202	61668	2021-08-01 08:57:00 UTC
2	205	43271	2021-08-01 08:58:00 UTC
3	205	7195	2021-08-01 08:54:00 UTC

Chapter 7: Visualizing Data for Making Data-Driven Decisions with Data Studio

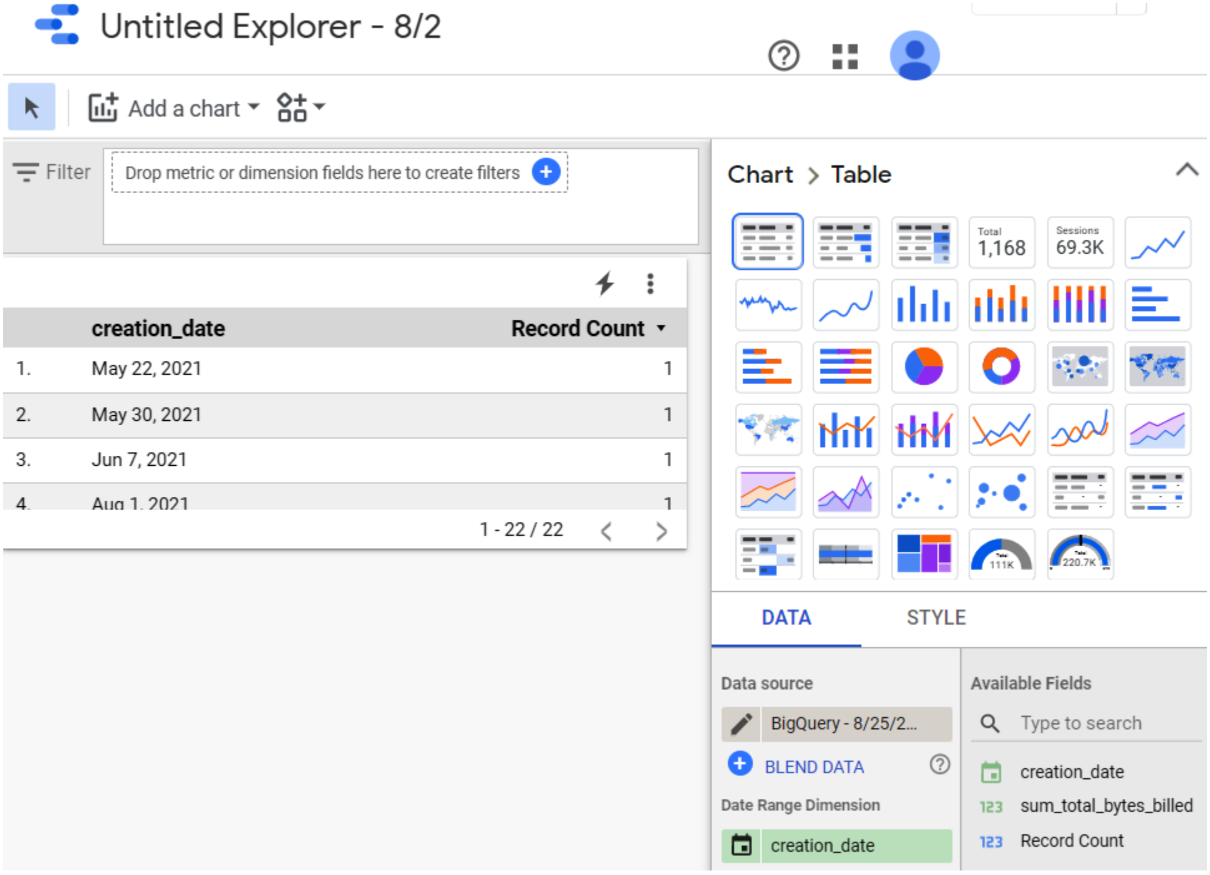


creation_time	project_id	project_number
2021-05-29 07:49:18.377 UTC	packt-data-eng-on-gcp	320986546290



Explore with Data Studio
Visualize results and create live dashboards from your data.

Query complete (0.9 sec elapsed, 86 KB processed)



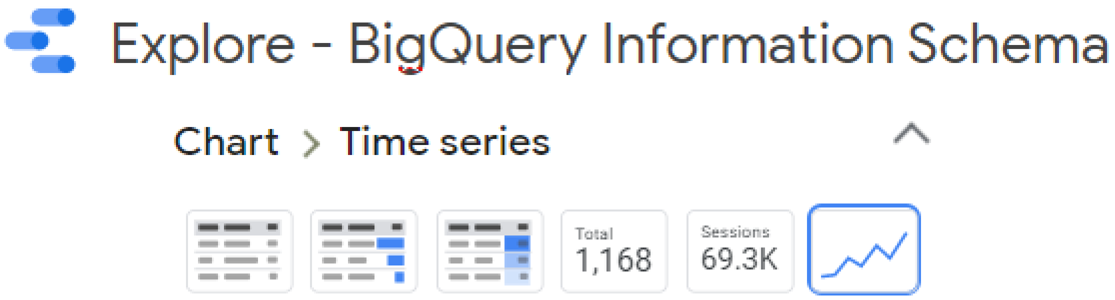
	creation_date	Record Count
1.	May 22, 2021	1
2.	May 30, 2021	1
3.	Jun 7, 2021	1
4.	Aug 1, 2021	1

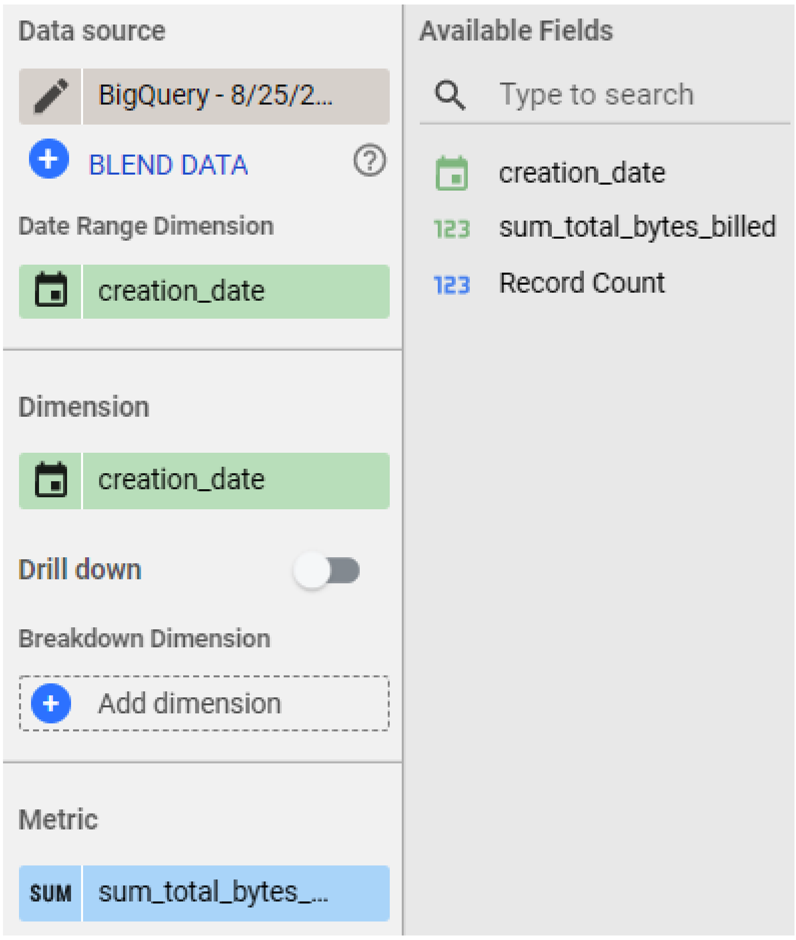
1 - 22 / 22

Data source: BigQuery - 8/25/2...

Date Range Dimension: creation_date

Available Fields: creation_date, sum_total_bytes_billed, Record Count





Data source

BigQuery - 8/25/2...

Date Range Dimension

creation_date

Dimension

creation_date

Drill down

Breakdown Dimension

Metric

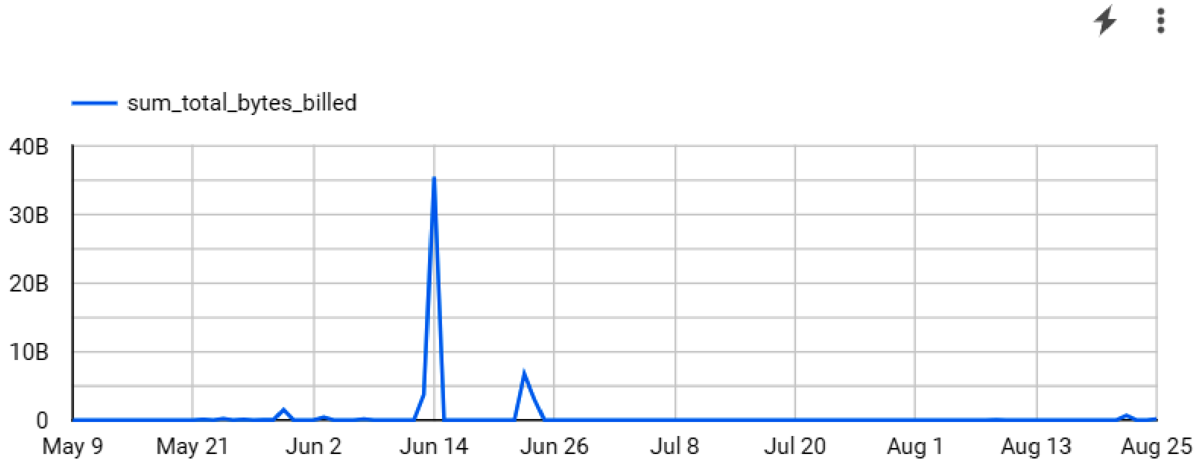
SUM sum_total_bytes_...

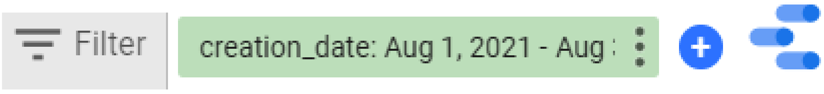
Available Fields

creation_date

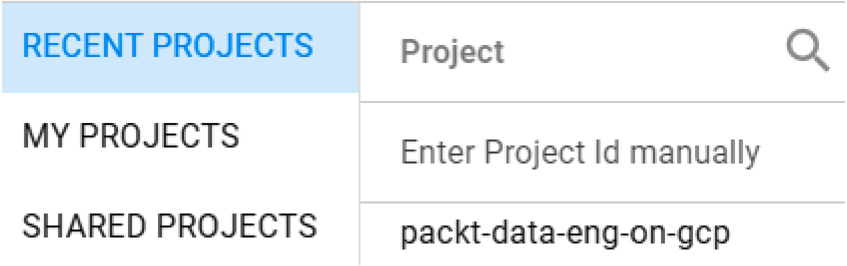
123 sum_total_bytes_billed

123 Record Count

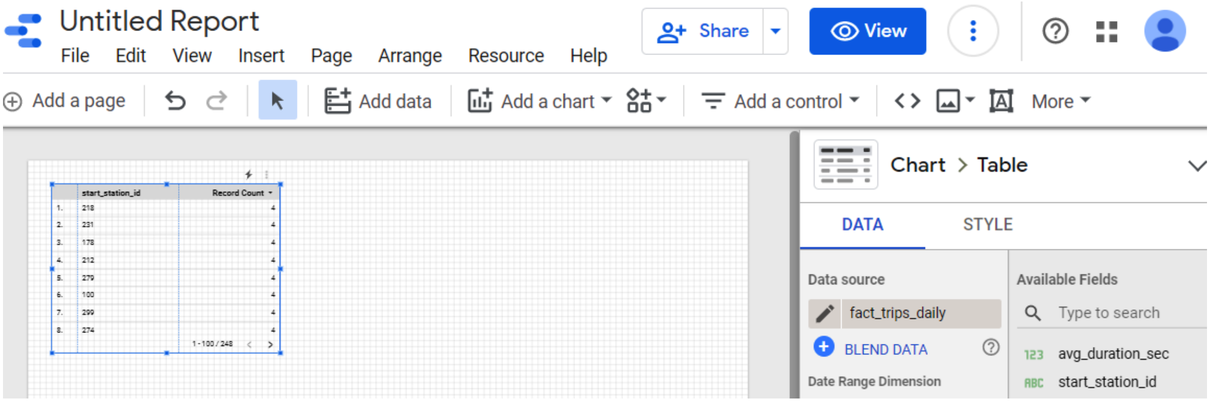




creation_date: Aug 1, 2021 - Aug :

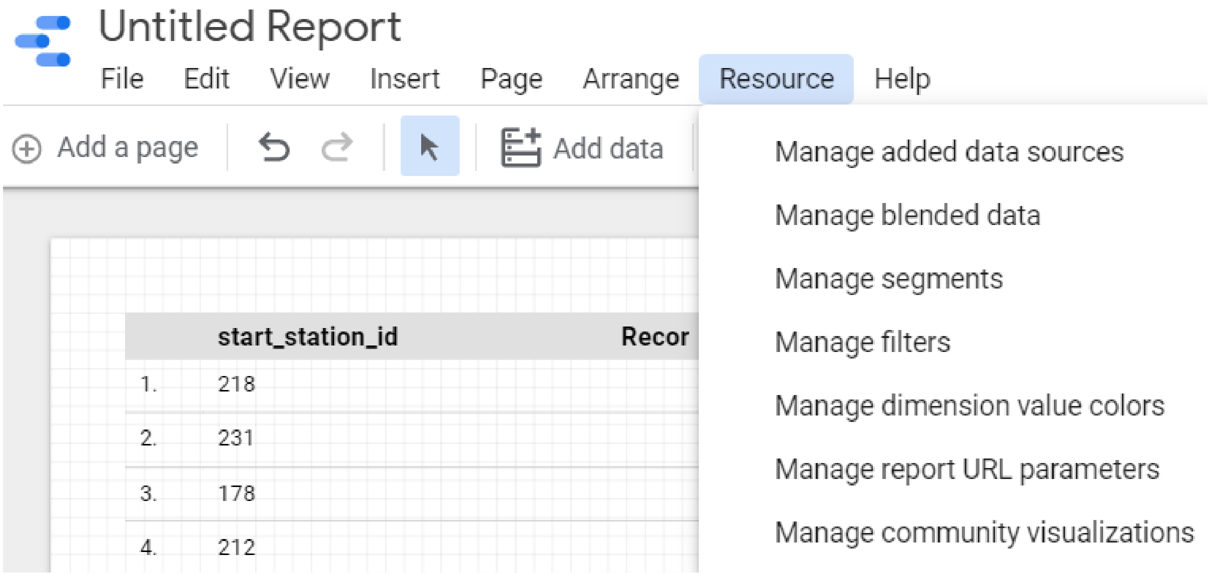


packt-data-eng-on-gcp

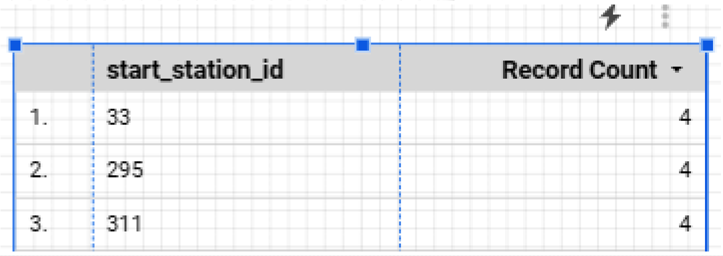


	start_station_id	Record Count
1.	218	4
2.	231	4
3.	178	4
4.	212	4
5.	279	4
6.	100	4
7.	299	4
8.	274	4

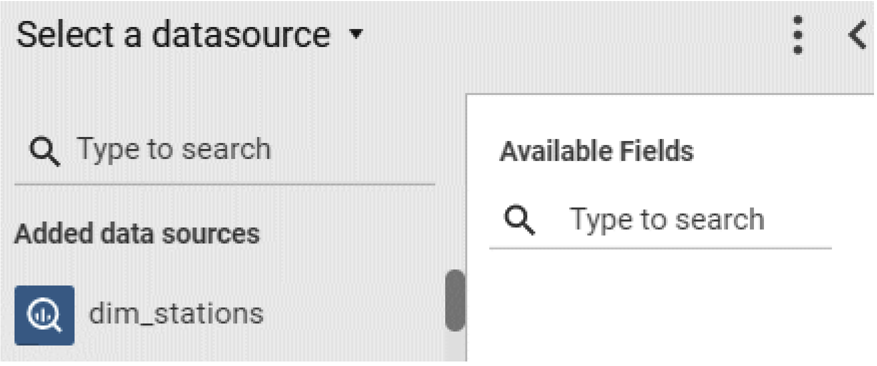
1 - 100 / 248

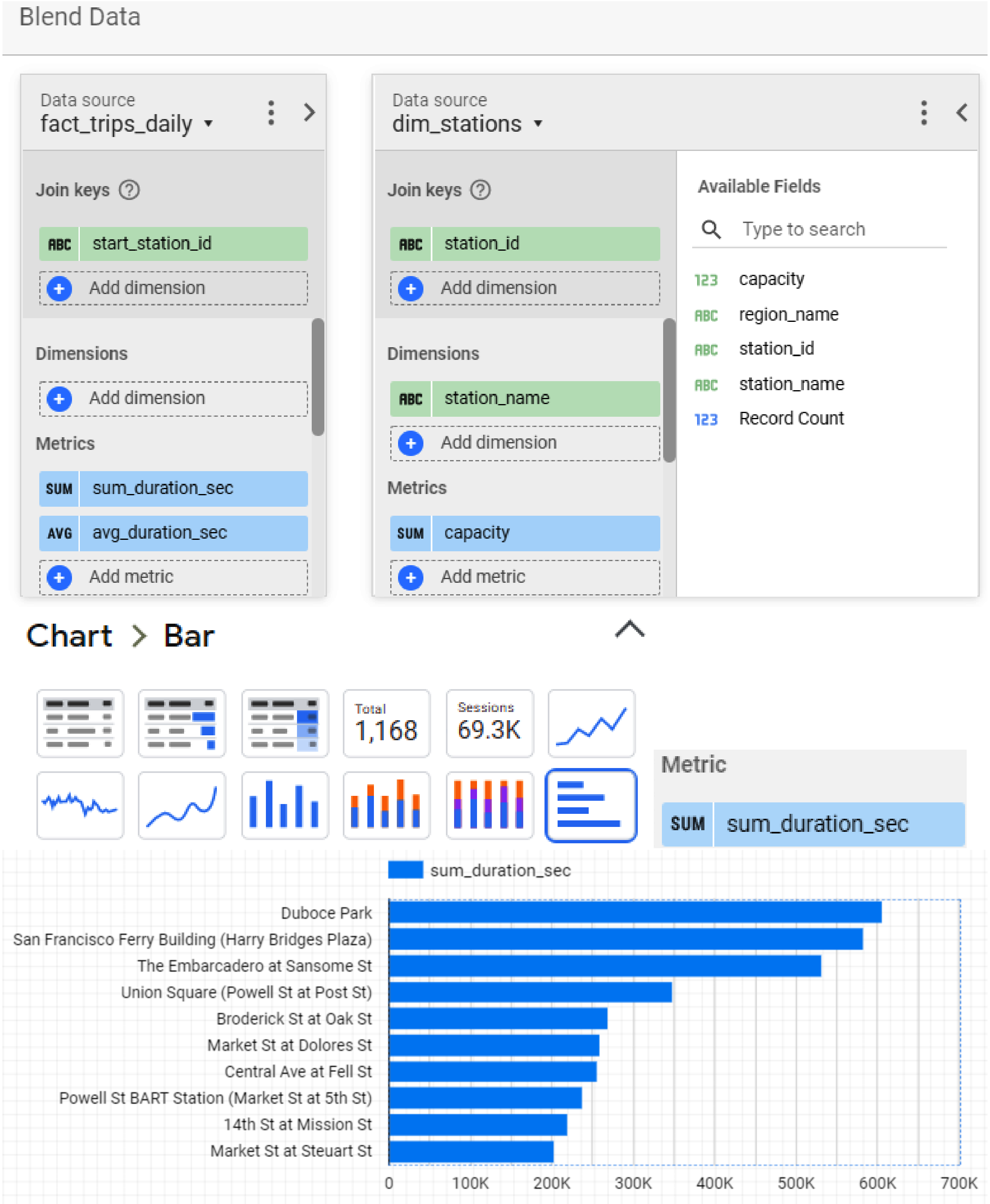


	start_station_id	Recor
1.	218	
2.	231	
3.	178	
4.	212	



	start_station_id	Record Count
1.	33	4
2.	295	4
3.	311	4





Blend Data

Data source: fact_trips_daily

Join keys: start_station_id

Dimensions: Add dimension

Metrics: SUM sum_duration_sec, AVG avg_duration_sec

Data source: dim_stations

Join keys: station_id

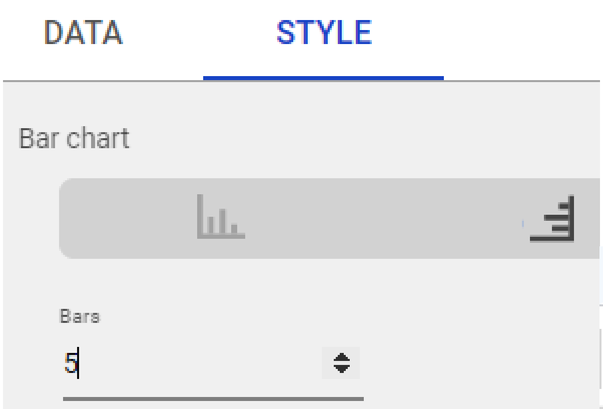
Dimensions: station_name

Metrics: SUM capacity

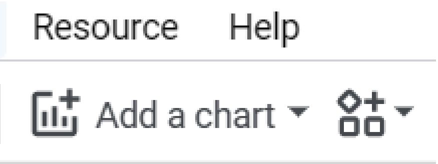
Available Fields: capacity, region_name, station_id, station_name, Record Count

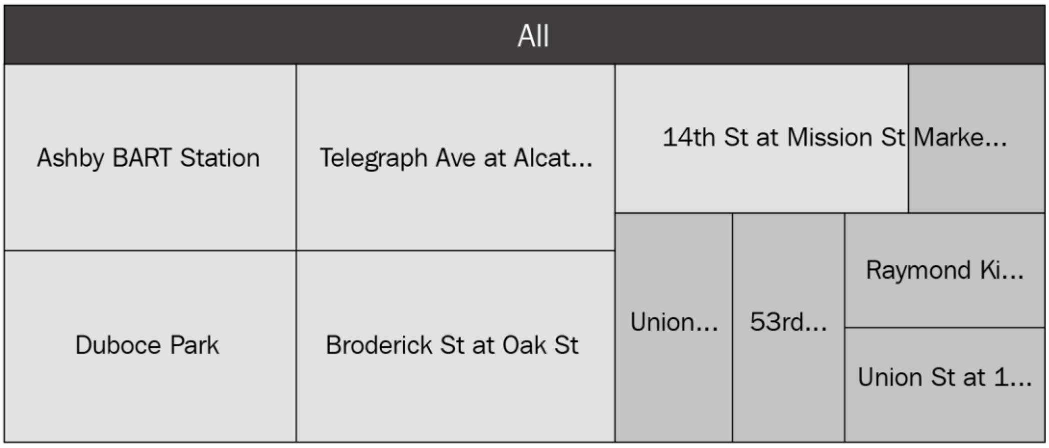
Chart > Bar

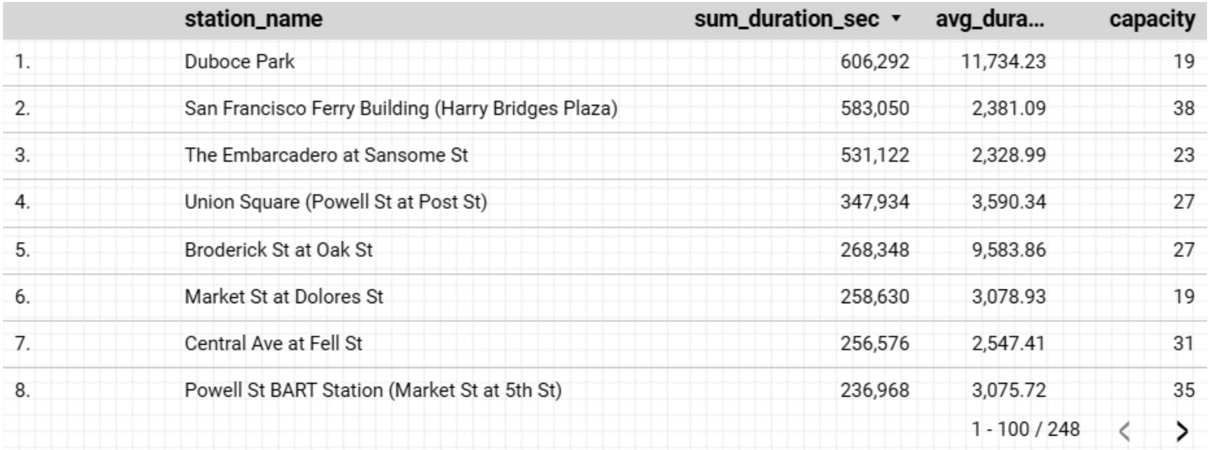
Metric: SUM sum_duration_sec



Bars
5

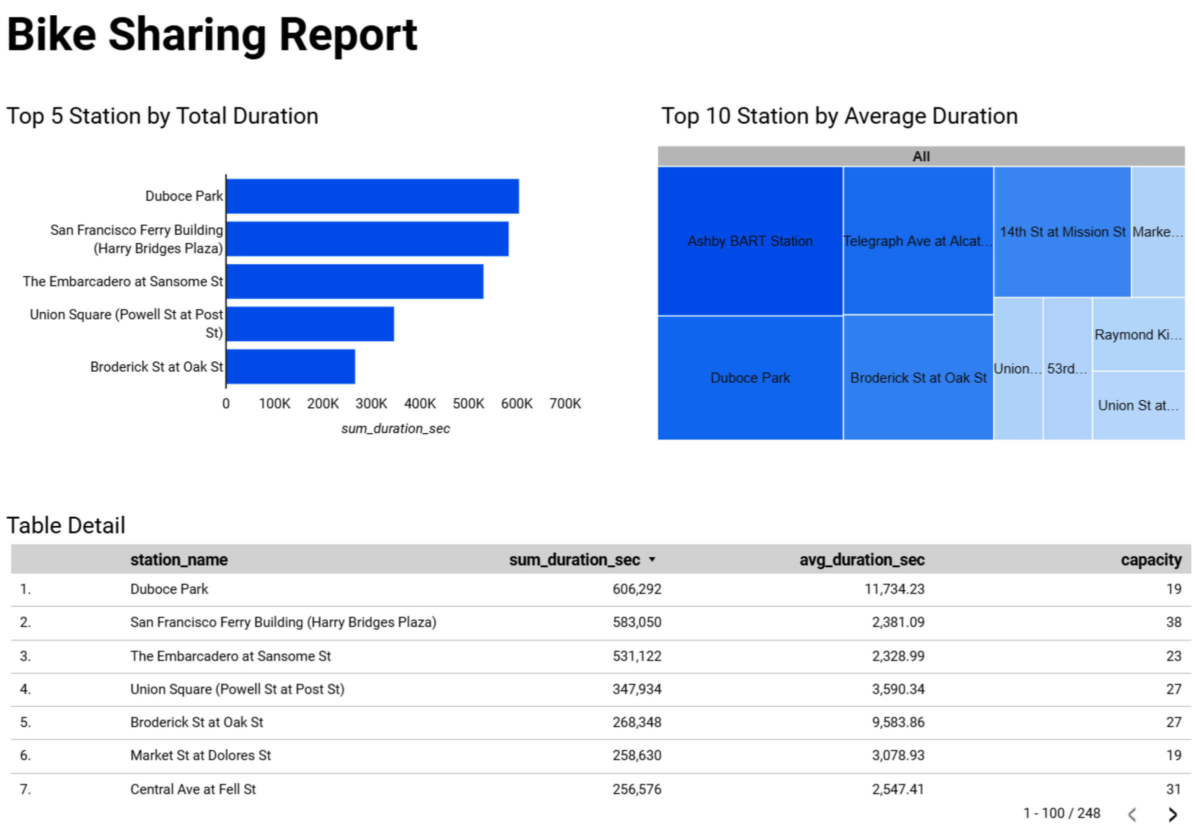






	station_name	sum_duration_sec ▾	avg_dura...	capacity
1.	Duboce Park	606,292	11,734.23	19
2.	San Francisco Ferry Building (Harry Bridges Plaza)	583,050	2,381.09	38
3.	The Embarcadero at Sansome St	531,122	2,328.99	23
4.	Union Square (Powell St at Post St)	347,934	3,590.34	27
5.	Broderick St at Oak St	268,348	9,583.86	27
6.	Market St at Dolores St	258,630	3,078.93	19
7.	Central Ave at Fell St	256,576	2,547.41	31
8.	Powell St BART Station (Market St at 5th St)	236,968	3,075.72	35

1 - 100 / 248



Bike Sharing Report

Top 5 Station by Total Duration

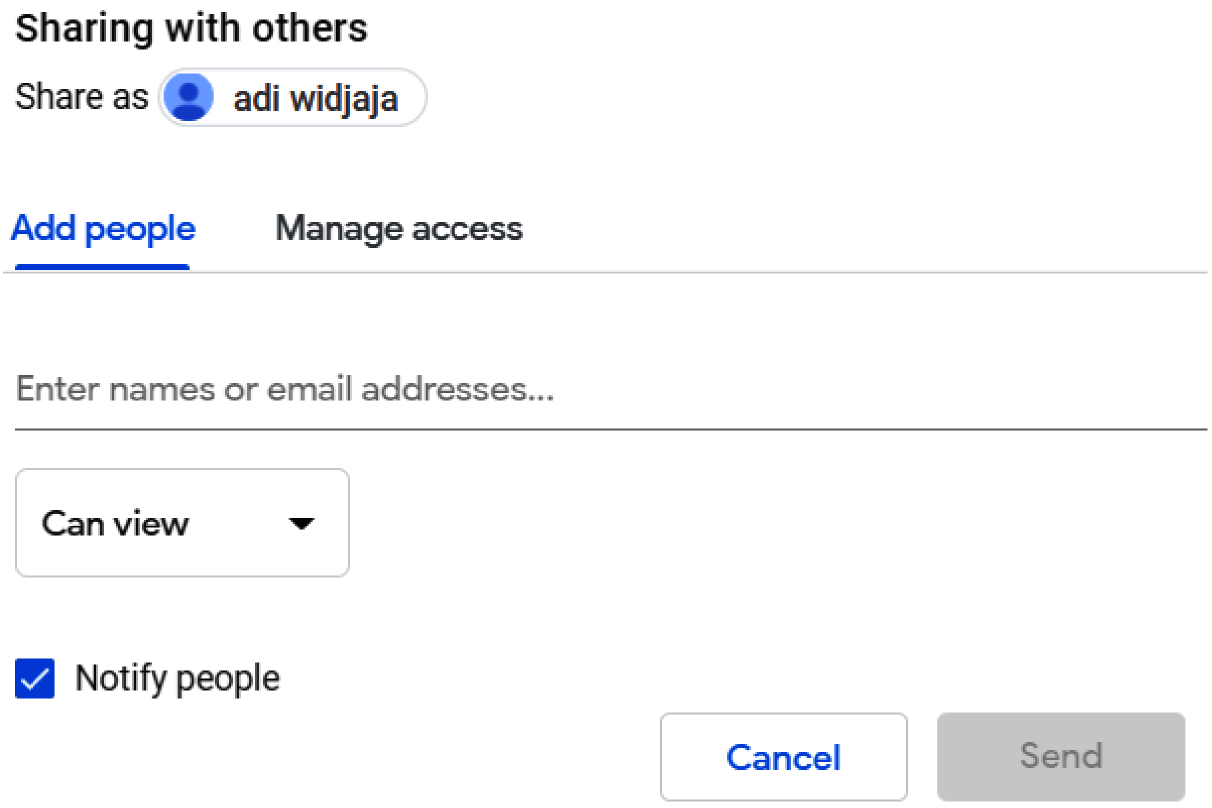
Duboce Park
San Francisco Ferry Building (Harry Bridges Plaza)
The Embarcadero at Sansome St
Union Square (Powell St at Post St)
Broderick St at Oak St
0
100K
200K
300K
400K
500K
600K
700K
sum_duration_sec

Top 10 Station by Average Duration

Table Detail

	station_name	sum_duration_sec ▾	avg_duration_sec	capacity
1.	Duboce Park	606,292	11,734.23	19
2.	San Francisco Ferry Building (Harry Bridges Plaza)	583,050	2,381.09	38
3.	The Embarcadero at Sansome St	531,122	2,328.99	23
4.	Union Square (Powell St at Post St)	347,934	3,590.34	27
5.	Broderick St at Oak St	268,348	9,583.86	27
6.	Market St at Dolores St	258,630	3,078.93	19
7.	Central Ave at Fell St	256,576	2,547.41	31

1 - 100 / 248 ‹ ›



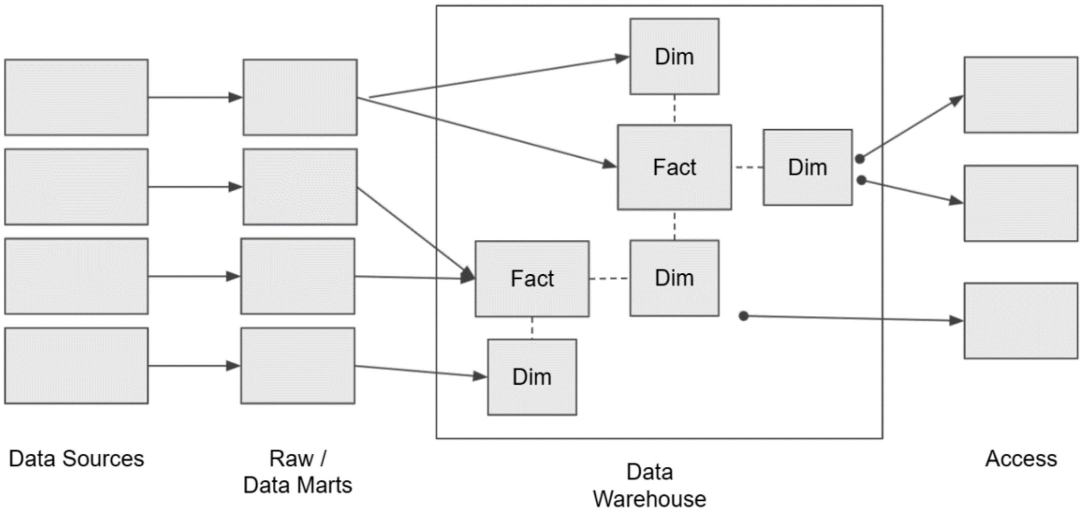
Sharing with others

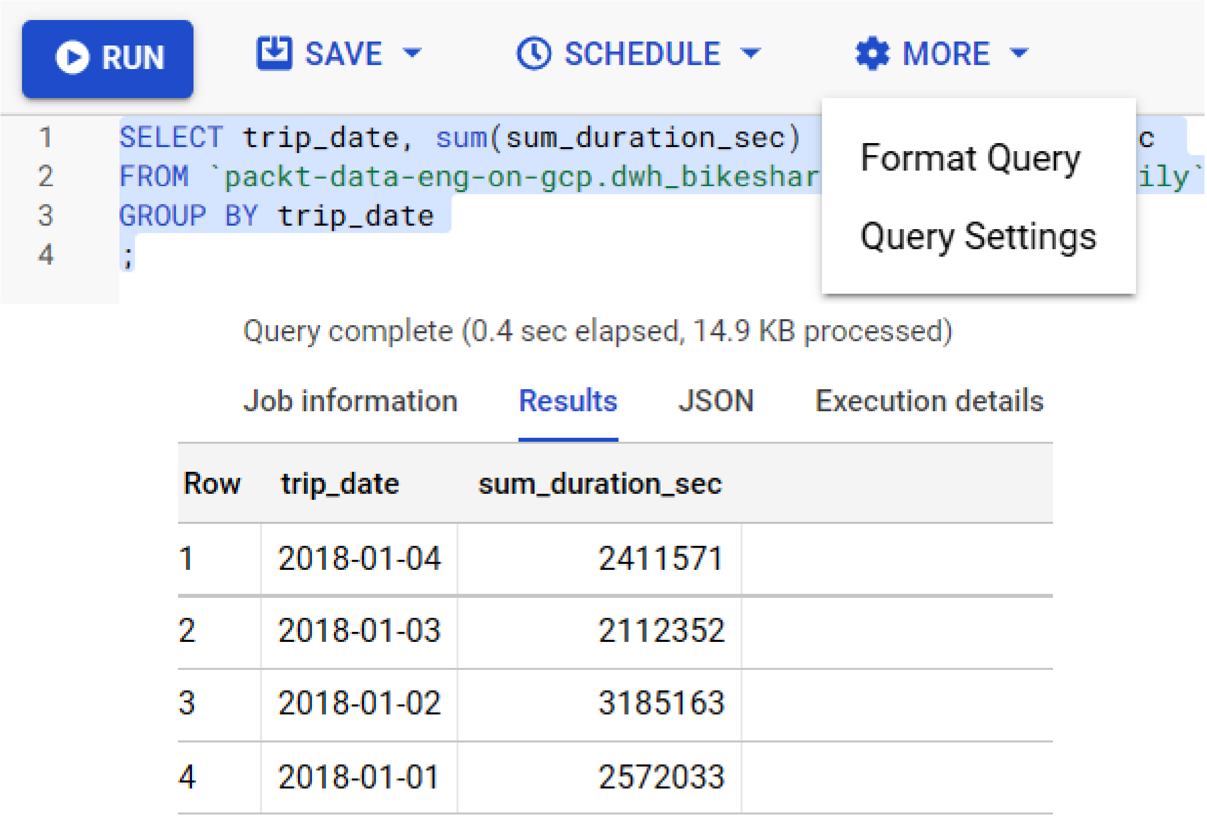
Share as adi widjaja

Enter names or email addresses...

Can view ▾

☑ Notify people



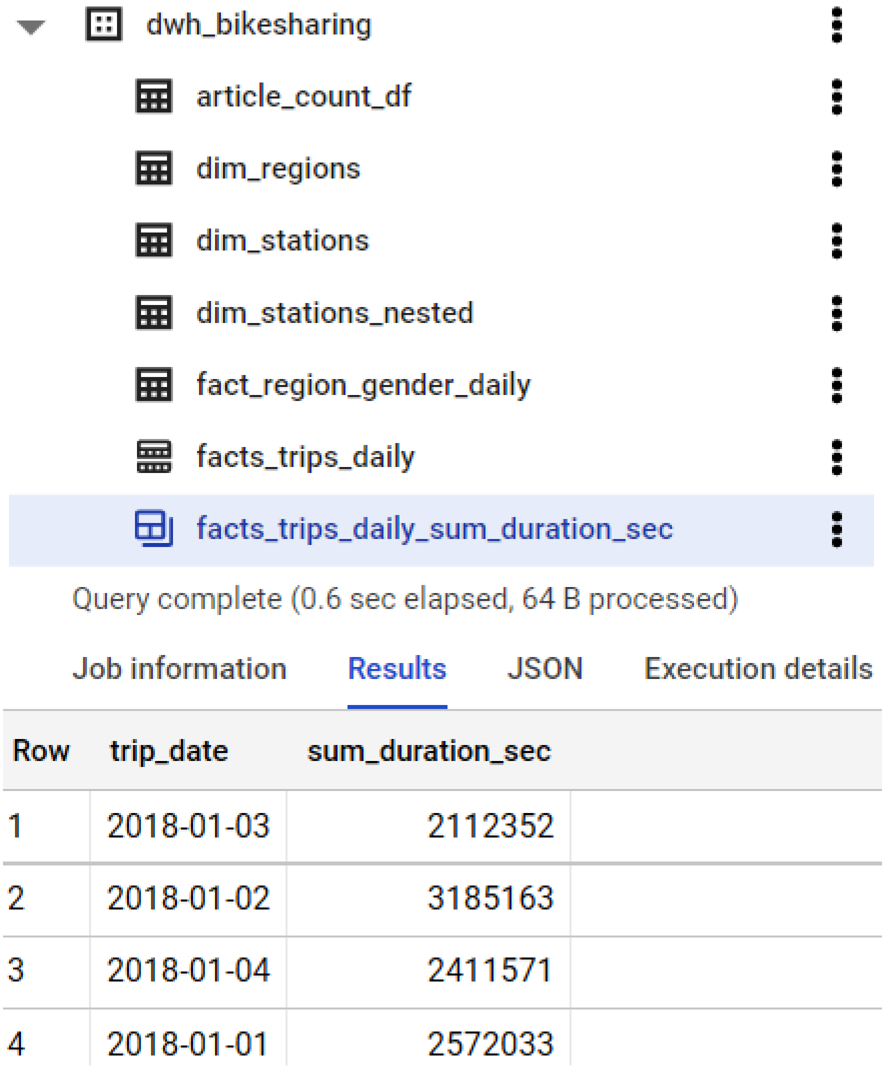


```
SELECT trip_date, sum(sum_duration_sec)                         c
FROM `packt-data-eng-on-gcp.dwh_bikeshar                         ily`
GROUP BY trip_date
;
```

Query complete (0.4 sec elapsed, 14.9 KB processed)

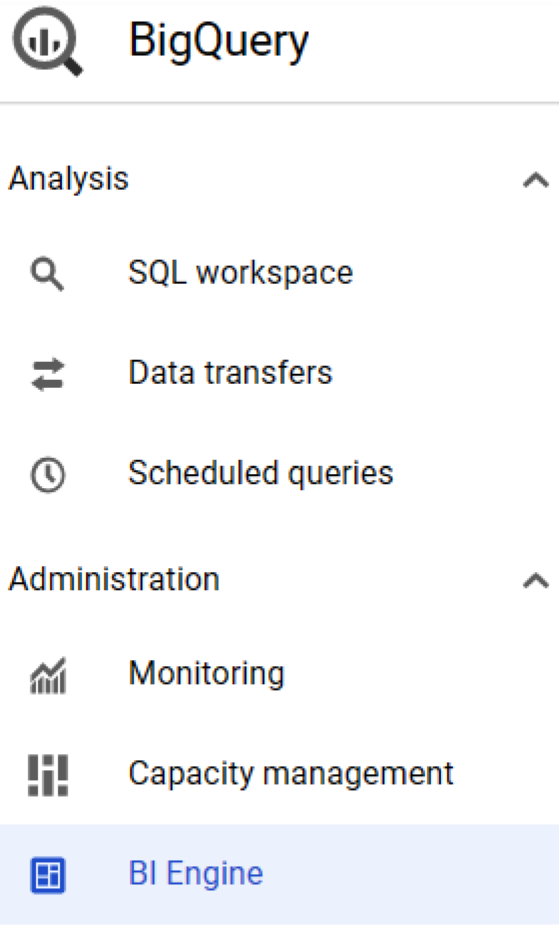
Job information   Results   JSON   Execution details

Row	trip_date	sum_duration_sec
1	2018-01-04	2411571
2	2018-01-03	2112352
3	2018-01-02	3185163
4	2018-01-01	2572033

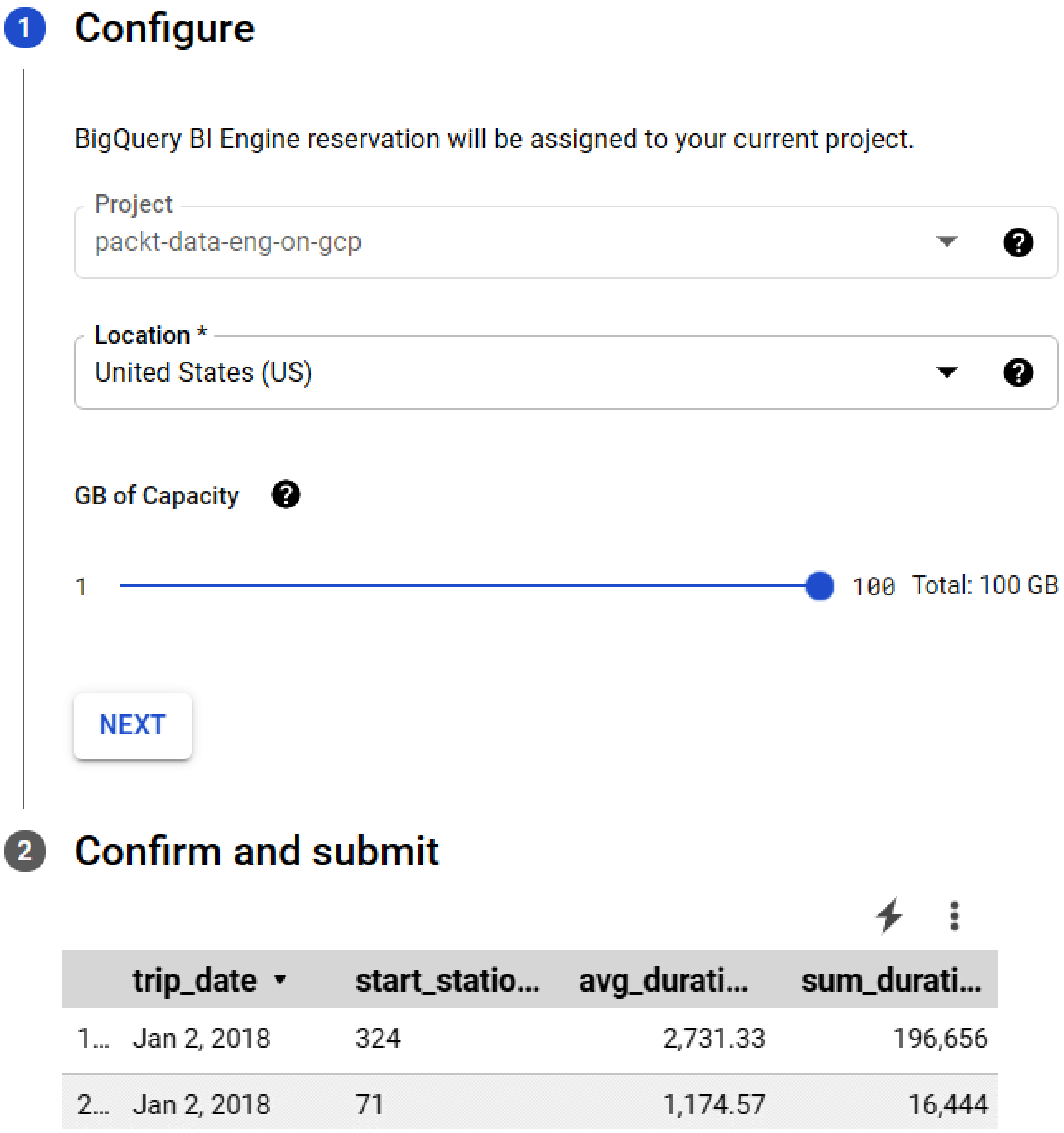


Query complete (0.6 sec elapsed, 64 B processed)

Row	trip_date	sum_duration_sec
1	2018-01-03	2112352
2	2018-01-02	3185163
3	2018-01-04	2411571
4	2018-01-01	2572033



BigQuery
Analysis
SQL workspace
Data transfers
Scheduled queries
Administration
Monitoring
Capacity management
BI Engine



1 Configure

BigQuery BI Engine reservation will be assigned to your current project.

Project: packt-data-eng-on-gcp

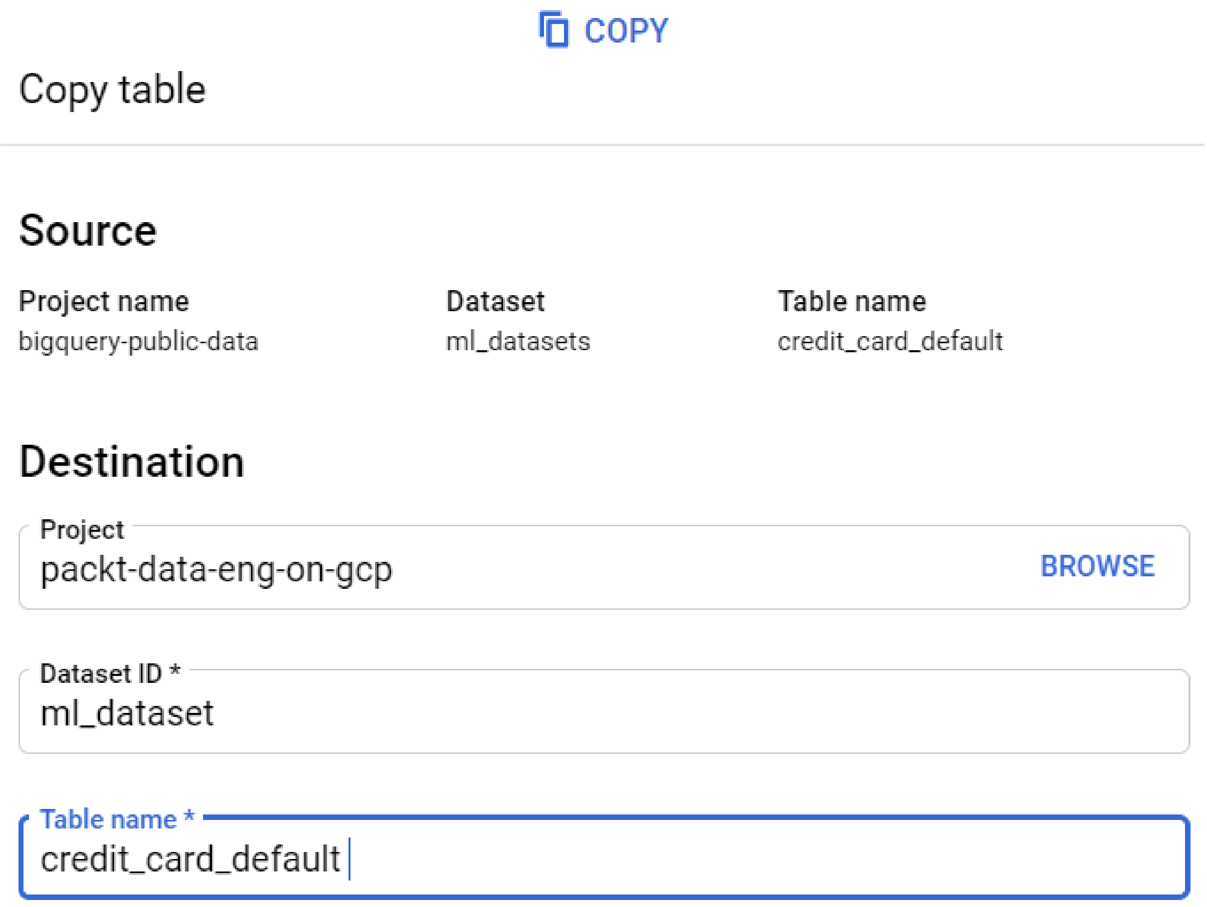
Location *: United States (US)

GB of Capacity

1 — 100 Total: 100 GB

2 Confirm and submit

	trip_date ▾	start_statio...	avg_durati...	sum_durati...
1...	Jan 2, 2018	324	2,731.33	196,656
2...	Jan 2, 2018	71	1,174.57	16,444



Copy table

Source

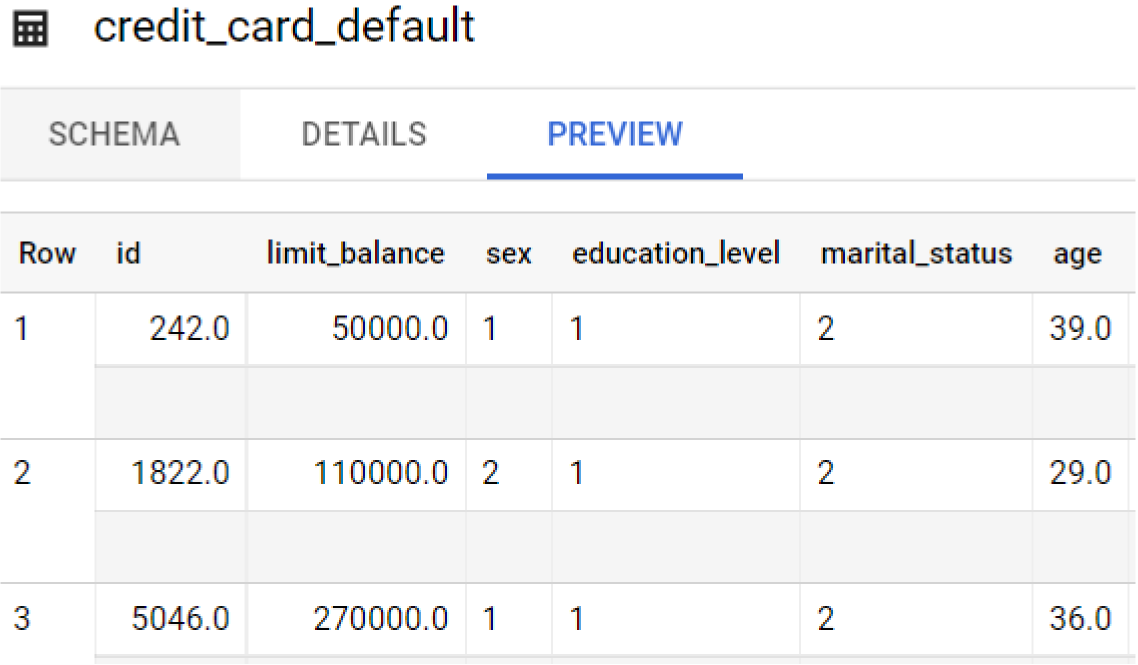
Project name	Dataset	Table name
bigquery-public-data	ml_datasets	credit_card_default

Destination

Project: packt-data-eng-on-gcp

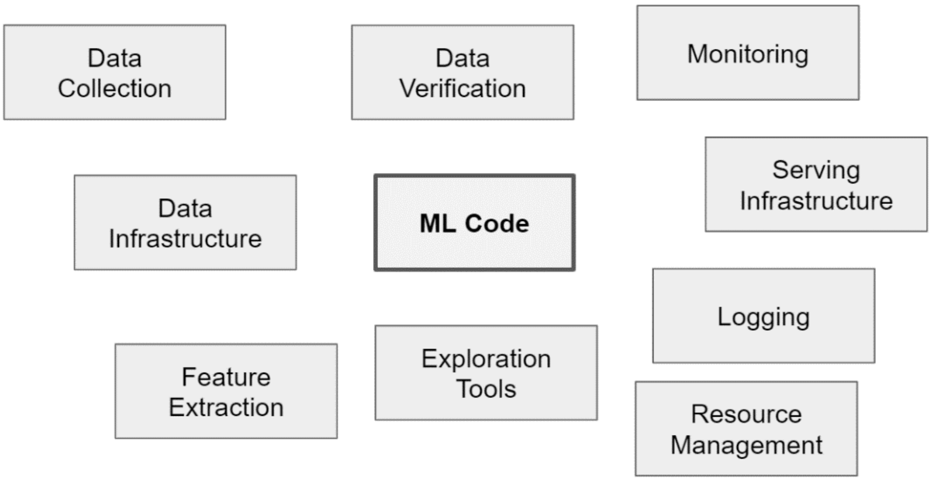
Dataset ID *: ml_dataset

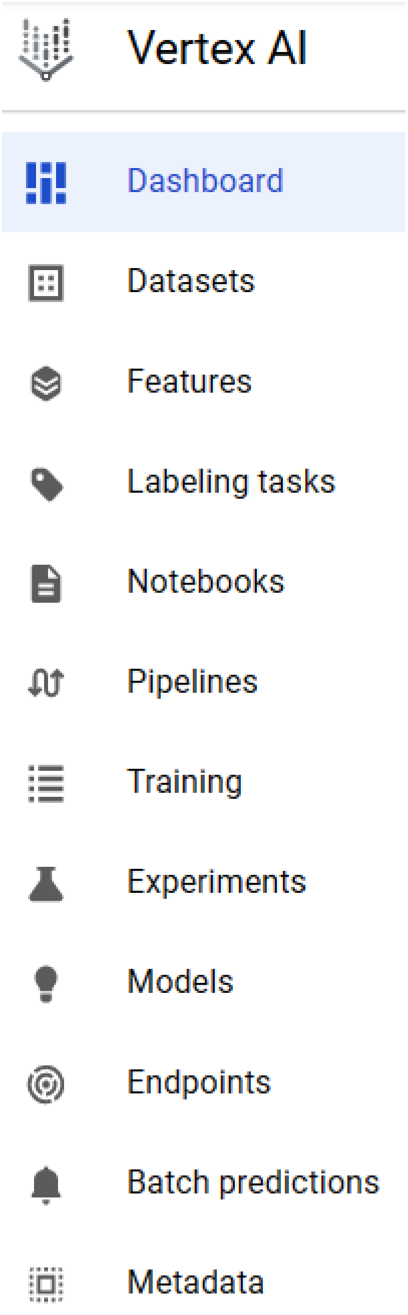
Table name *: credit_card_default

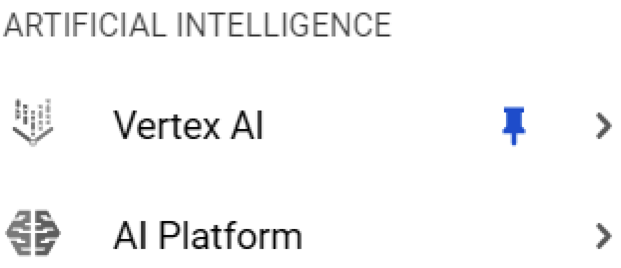


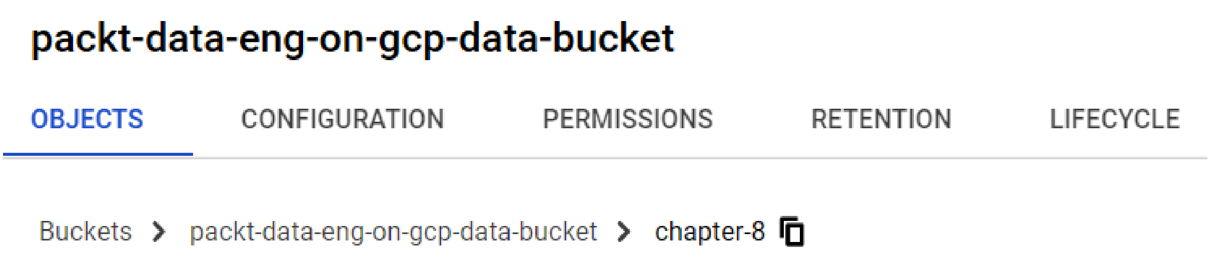
credit_card_default

Row	id	limit_balance	sex	education_level	marital_status	age
1	242.0	50000.0	1	1	2	39.0
2	1822.0	110000.0	2	1	2	29.0
3	5046.0	270000.0	1	1	2	36.0


Data Collection
Data Verification
Monitoring
Serving Infrastructure
Data Infrastructure
ML Code
Logging
Feature Extraction
Exploration Tools
Resource Management

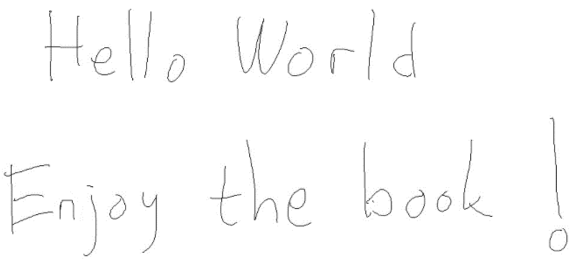

Vertex AI
Dashboard
Datasets
Features
Labeling tasks
Notebooks
Pipelines
Training
Experiments
Models
Endpoints
Batch predictions
Metadata


ARTIFICIAL INTELLIGENCE
Vertex AI
AI Platform



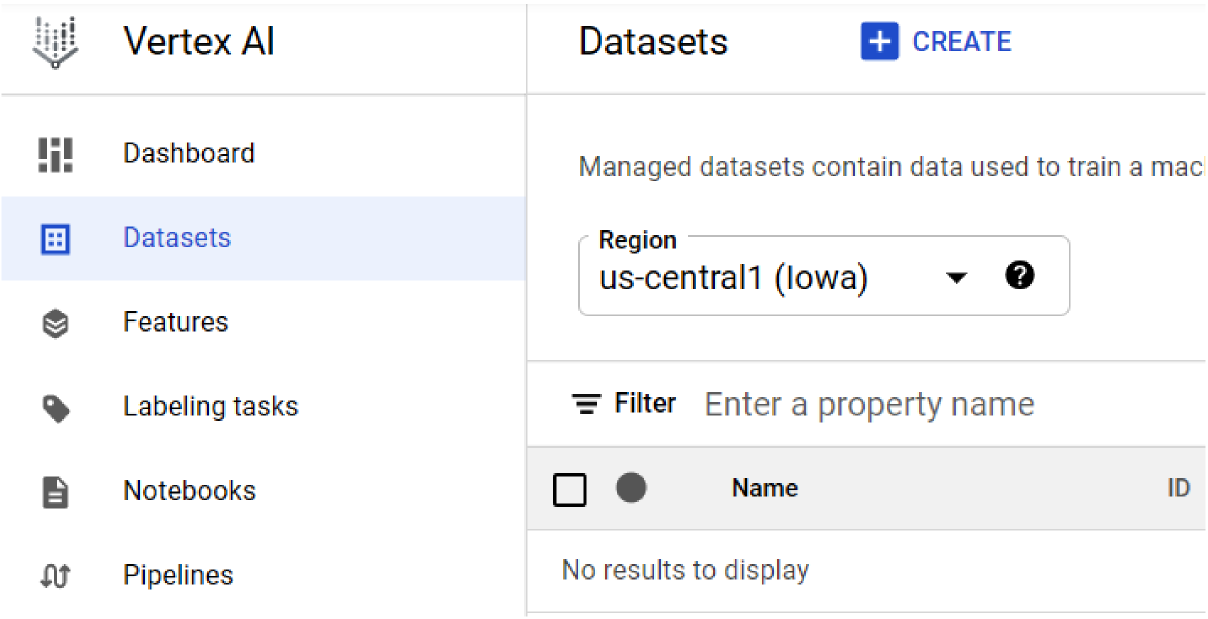
packt-data-eng-on-gcp-data-bucket

Buckets > packt-data-eng-on-gcp-data-bucket > chapter-8



Hello World

Enjoy the book!

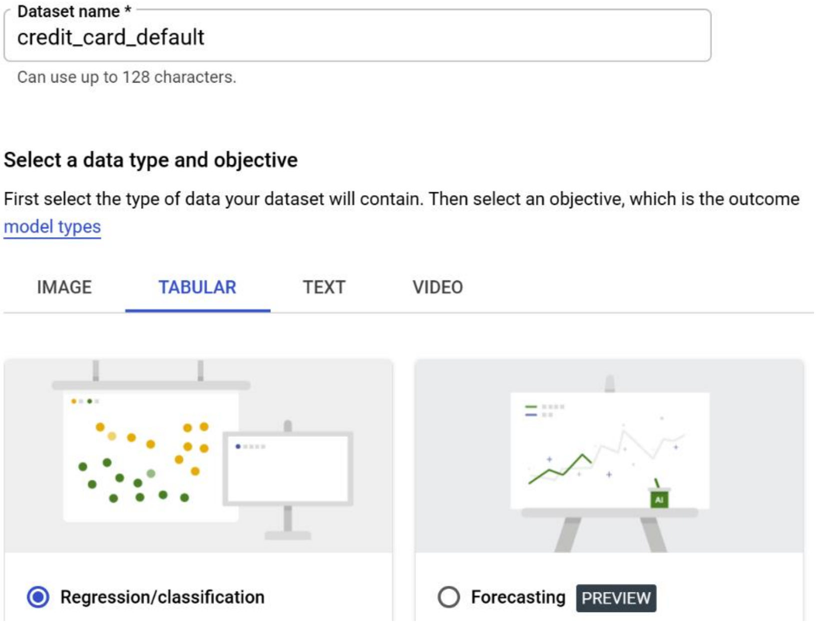


Managed datasets contain data used to train a mac

Region: us-central1 (Iowa)

Enter a property name

No results to display



Dataset name *: credit_card_default

Can use up to 128 characters.

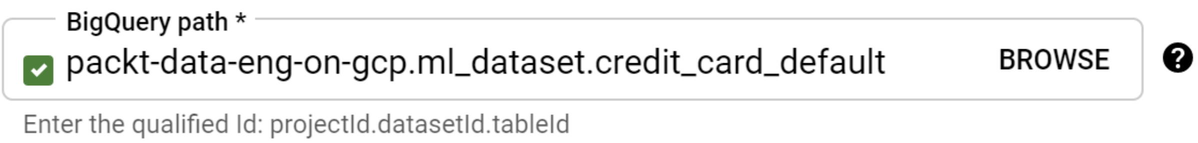
Select a data type and objective

First select the type of data your dataset will contain. Then select an objective, which is the outcome

model types

● Regression/classification

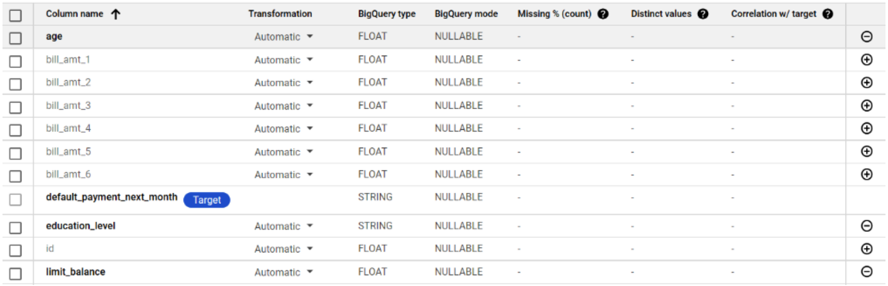
○ Forecasting PREVIEW



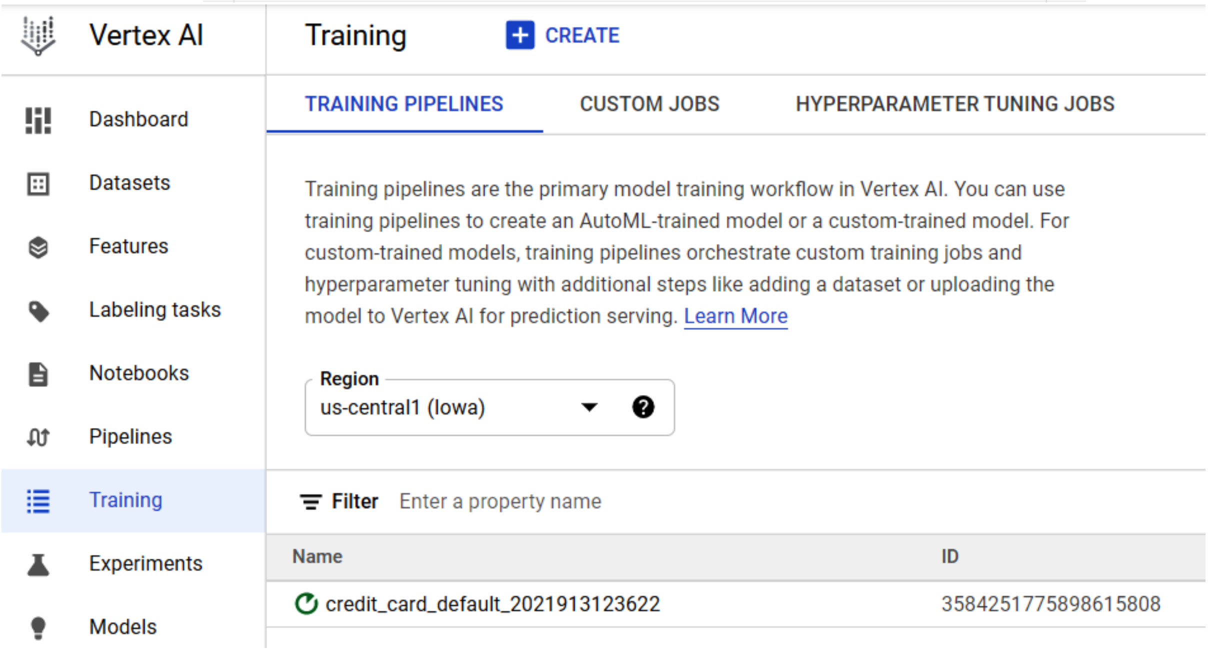
BigQuery path *

☑ packt-data-eng-on-gcp.ml_dataset.credit_card_default

Enter the qualified Id: projectId.datasetId.tableId



□	Column name ↑	Transformation	BigQuery type	BigQuery mode	Missing % (count)	Distinct values	Correlation w/ target	
□	age	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊖
□	bill_amt_1	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊕
□	bill_amt_2	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊕
□	bill_amt_3	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊕
□	bill_amt_4	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊕
□	bill_amt_5	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊕
□	bill_amt_6	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊕
□	default_payment_next_month Target		STRING	NULLABLE	-	-	-	
□	education_level	Automatic ▾	STRING	NULLABLE	-	-	-	⊖
□	id	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊕
□	limit_balance	Automatic ▾	FLOAT	NULLABLE	-	-	-	⊖

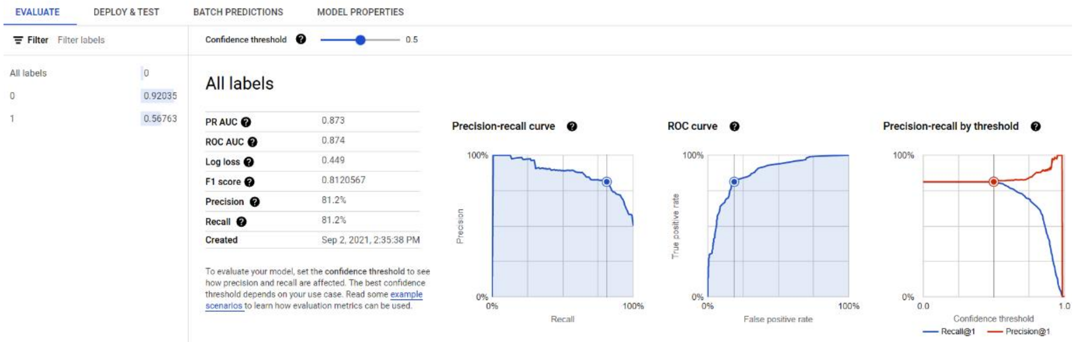


Training pipelines are the primary model training workflow in Vertex AI. You can use training pipelines to create an AutoML-trained model or a custom-trained model. For custom-trained models, training pipelines orchestrate custom training jobs and hyperparameter tuning with additional steps like adding a dataset or uploading the model to Vertex AI for prediction serving. Learn More

Region: us-central1 (Iowa)

Filter: Enter a property name

Name	ID
credit_card_default_2021913123622	3584251775898615808

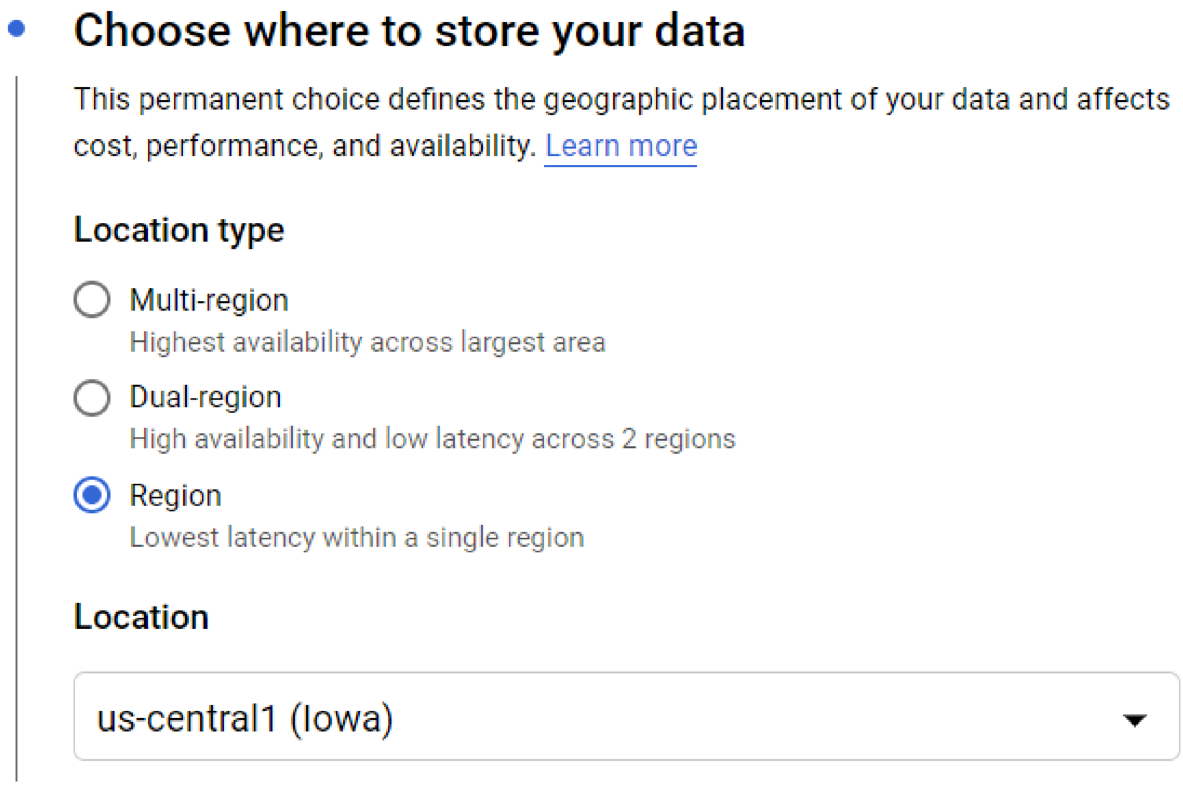


All labels	0
0	0.92035
1	0.56763

All labels

PR AUC	0.873
ROC AUC	0.874
Log loss	0.449
F1 score	0.8120567
Precision	81.2%
Recall	81.2%
Created	Sep 2, 2021, 2:35:38 PM

To evaluate your model, set the **confidence threshold** to see how precision and recall are affected. The best confidence threshold depends on your use case. Read some example scenarios to learn how evaluation metrics can be used.



Choose where to store your data

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. Learn more

Location type

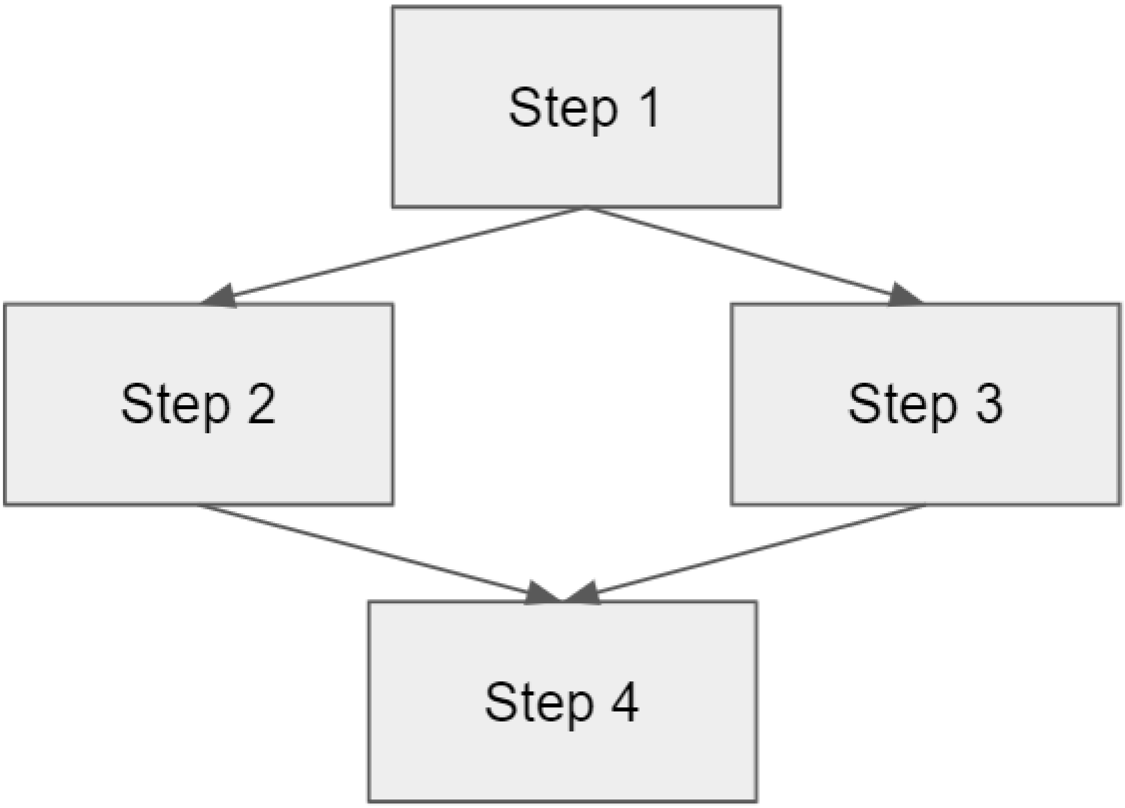
○ Multi-region
Highest availability across largest area

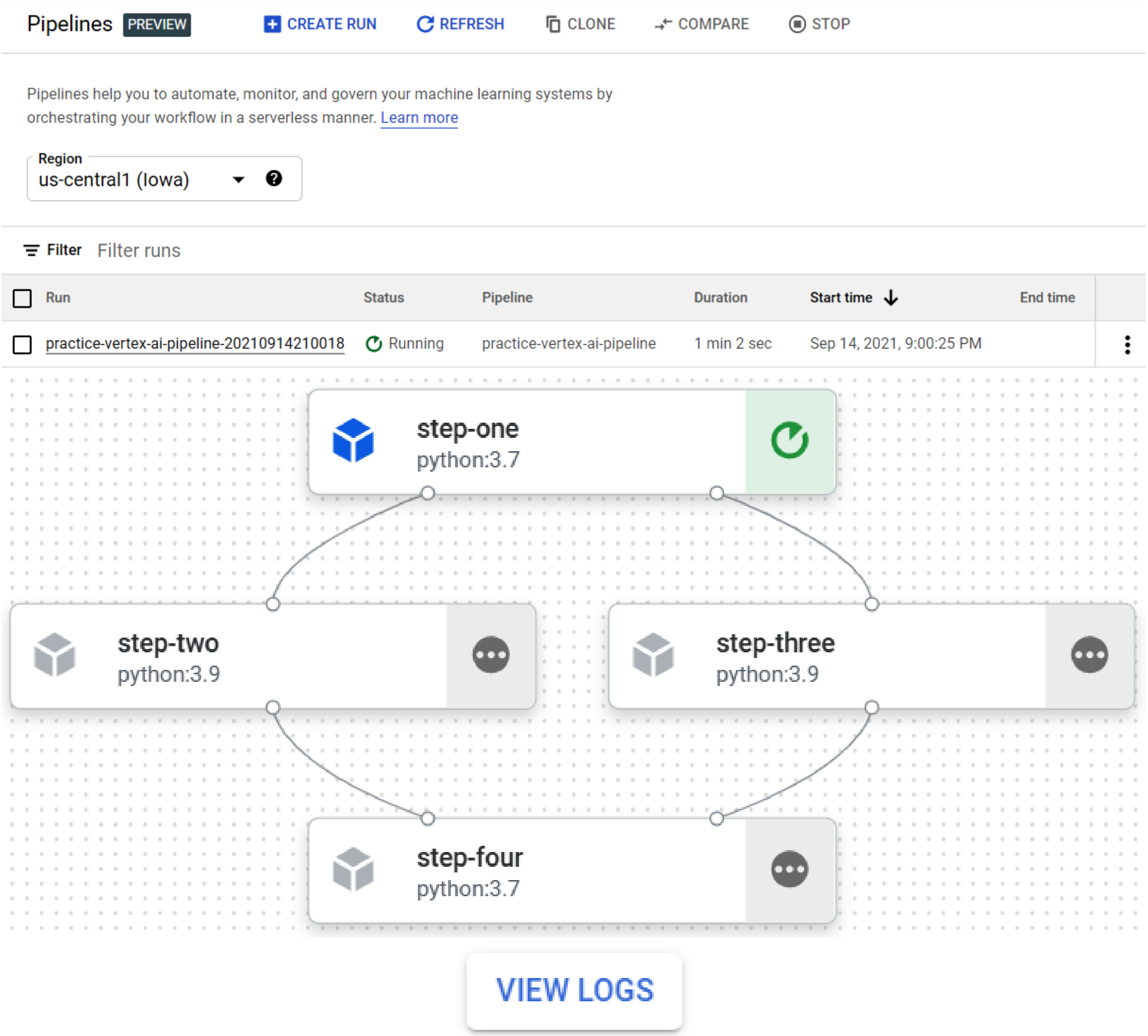
○ Dual-region
High availability and low latency across 2 regions

● Region
Lowest latency within a single region

Location

us-central1 (Iowa)





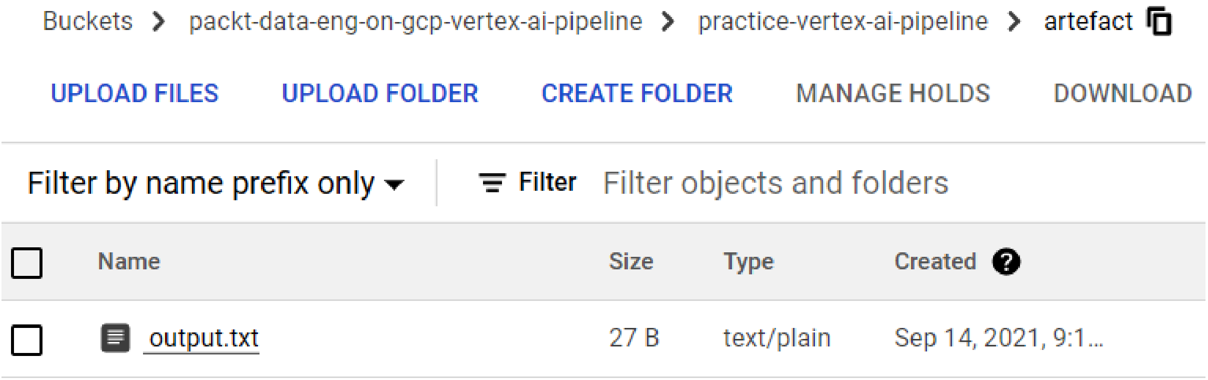
Pipelines PREVIEW CREATE RUN REFRESH CLONE COMPARE STOP

Pipelines help you to automate, monitor, and govern your machine learning systems by orchestrating your workflow in a serverless manner. Learn more

Region: us-central1 (Iowa)

Filter Filter runs

□ Run	Status	Pipeline	Duration	Start time ↓	End time
□ practice-vertex-ai-pipeline-20210914210018	Running	practice-vertex-ai-pipeline	1 min 2 sec	Sep 14, 2021, 9:00:25 PM	

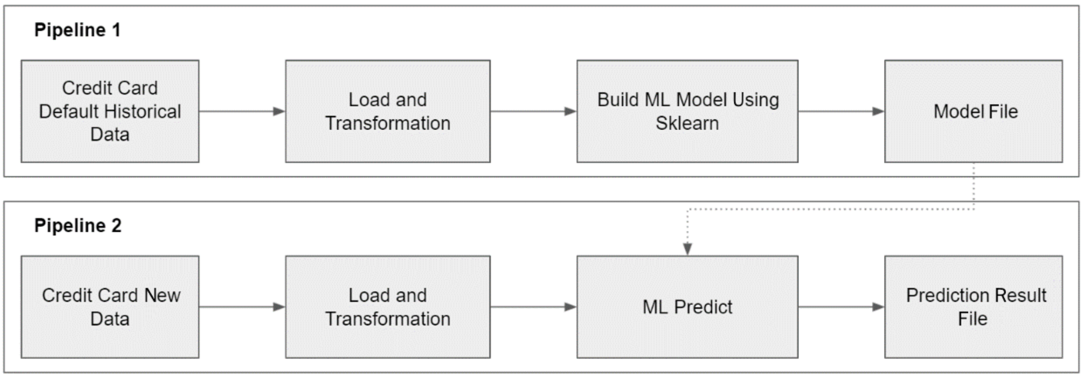


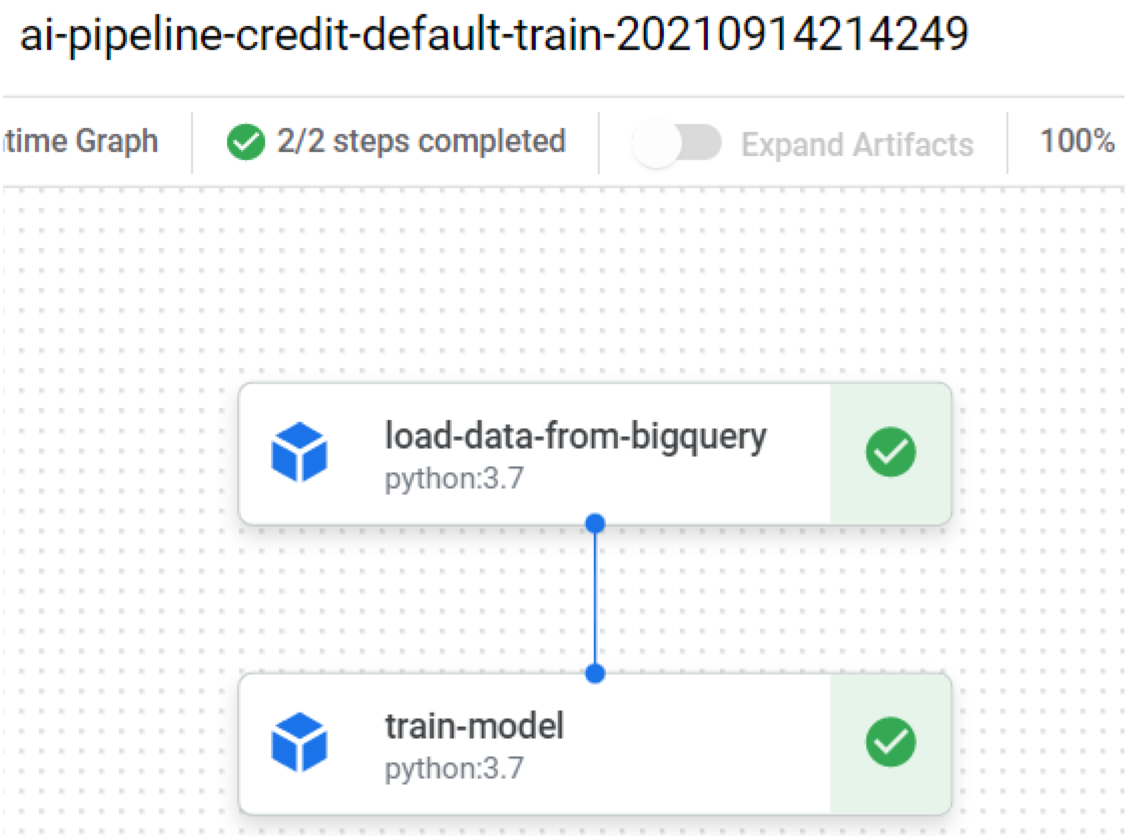
Buckets > packt-data-eng-on-gcp-vertex-ai-pipeline > practice-vertex-ai-pipeline > artefact

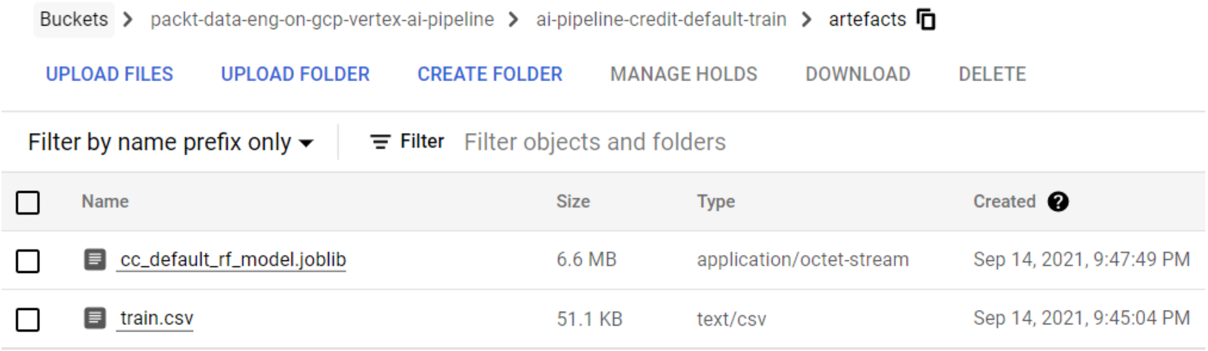
UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD

Filter by name prefix only ▾ Filter Filter objects and folders

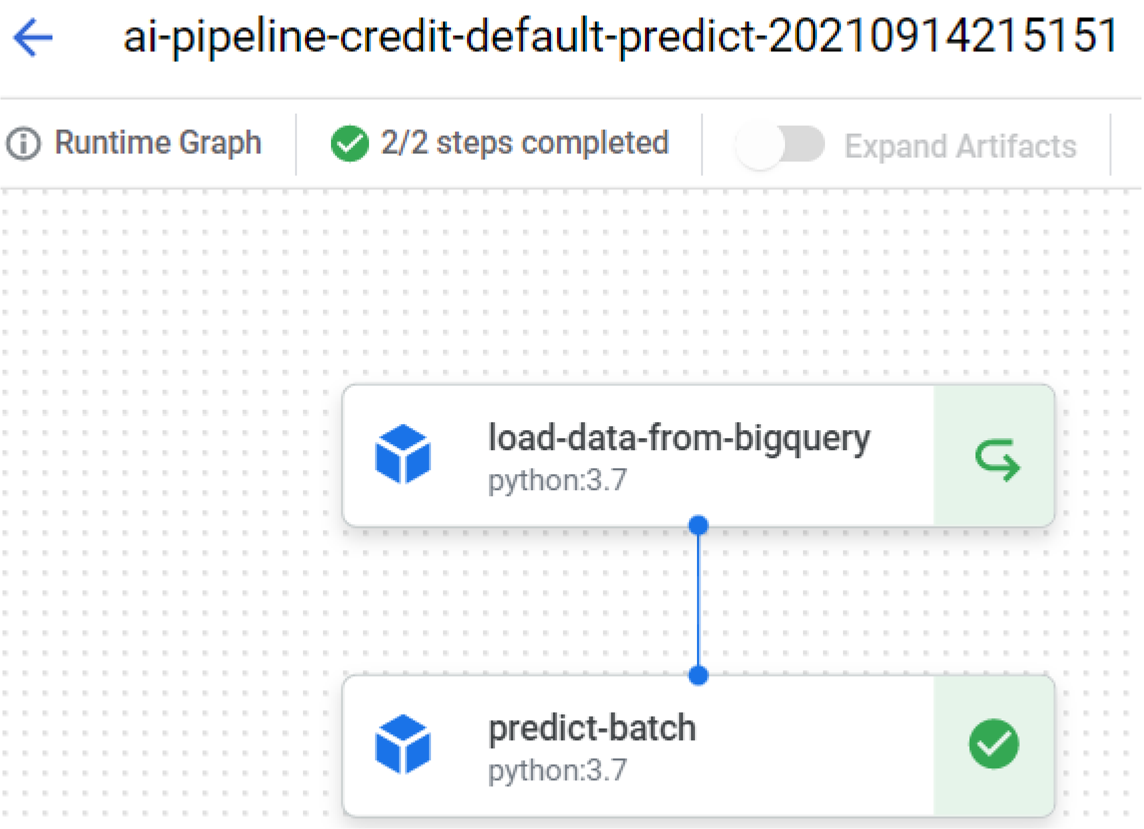
□	Name	Size	Type	Created
□	output.txt	27 B	text/plain	Sep 14, 2021, 9:1…





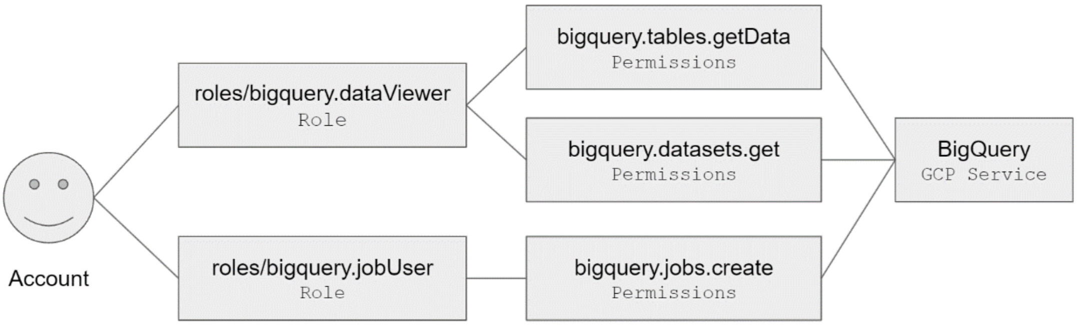

Buttons
packt-data-eng-on-gcp-vertex-ai-pipeline
ai-pipeline-credit-default-train
artefacts
UPLOAD FILES
UPLOAD FOLDER
CREATE FOLDER
MANAGE HOLDS
DOWNLOAD
DELETE
Filter by name prefix only
Filter
Filter objects and folders

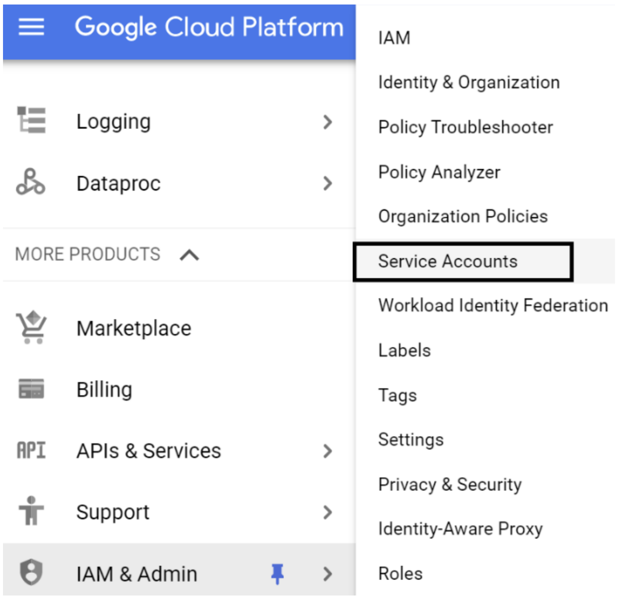
□	Name	Size	Type	Created
□	cc_default_rf_model.joblib	6.6 MB	application/octet-stream	Sep 14, 2021, 9:47:49 PM
□	train.csv	51.1 KB	text/csv	Sep 14, 2021, 9:45:04 PM

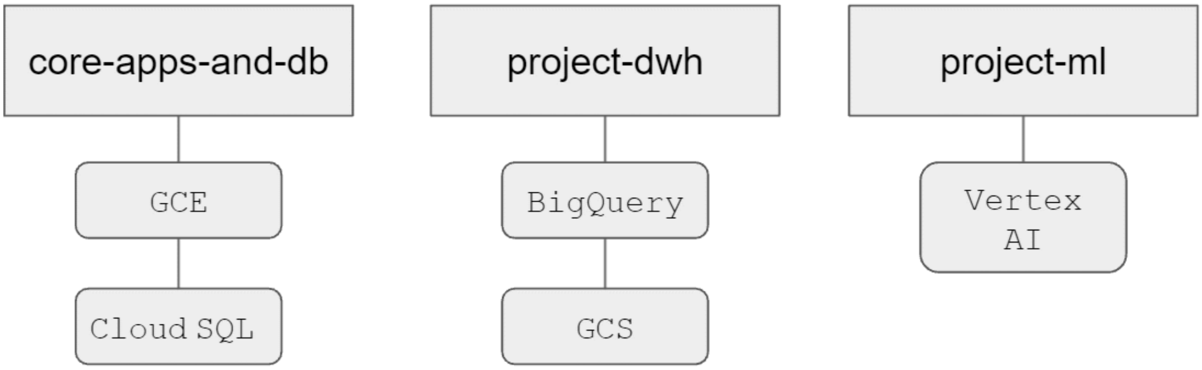


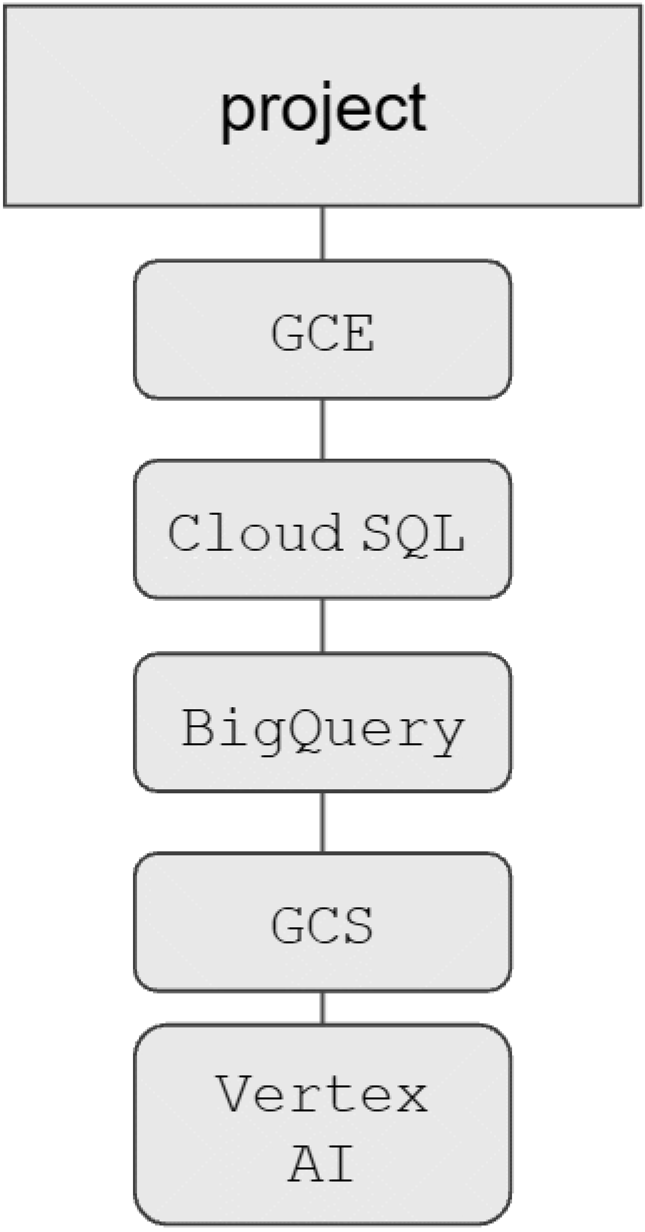
ai-pipeline-credit-default-predict-20210914215151

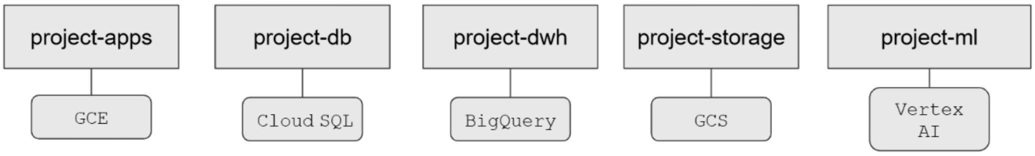
Chapter 9: G User and Project Management in GCP

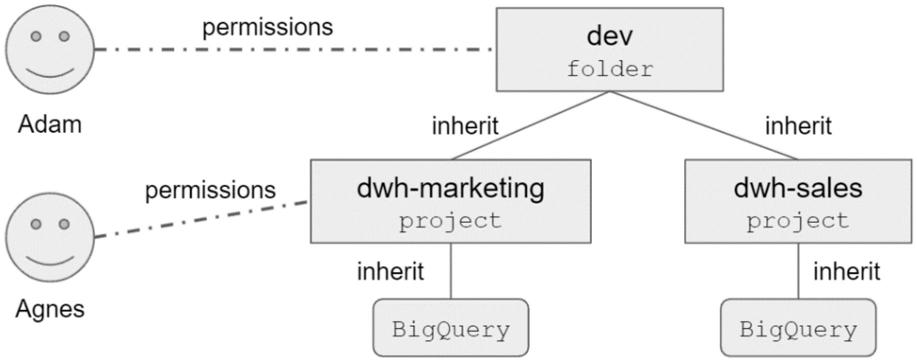


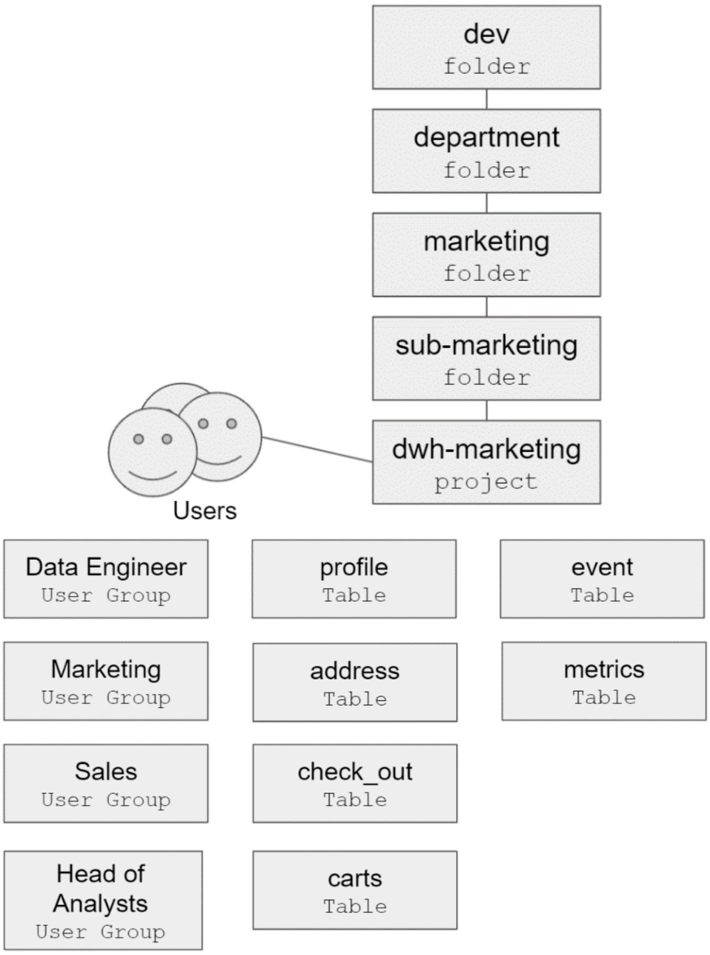


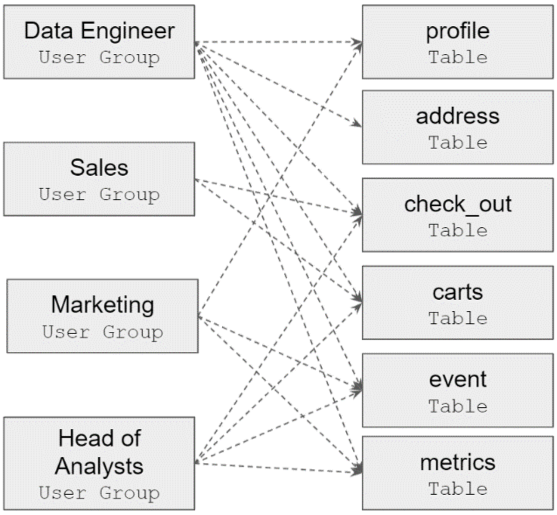



project
GCE
Cloud SQL
BigQuery
GCS
Vertex AI

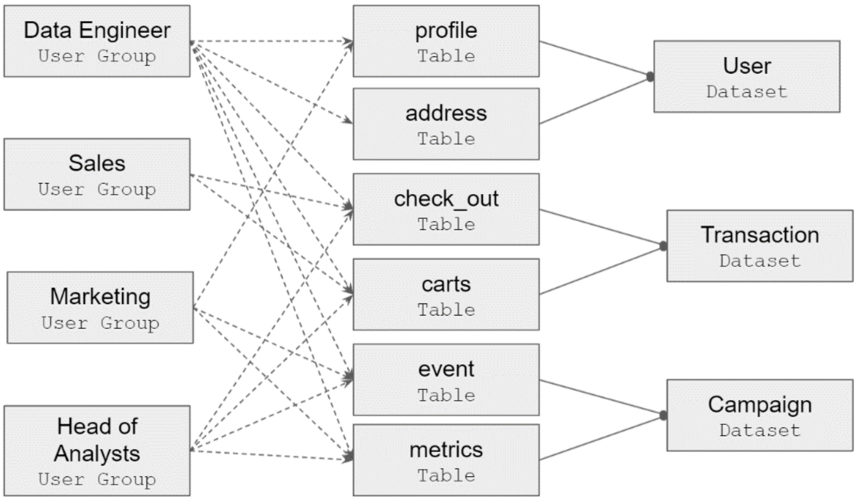

project-apps
GCE
project-db
Cloud SQL
project-dwh
BigQuery
project-storage
GCS
project-ml
Vertex AI


Adam
permissions
dev
folder
inherit
inherit
Agnes
permissions
dwh-marketing
project
dwh-sales
project
inherit
inherit
BigQuery
BigQuery


dev
folder
department
folder
marketing
folder
sub-marketing
folder
dwh-marketing
project
Users
Data Engineer
User Group
profile
Table
event
Table
Marketing
User Group
address
Table
metrics
Table
Sales
User Group
check_out
Table
Head of Analysts
User Group
carts
Table

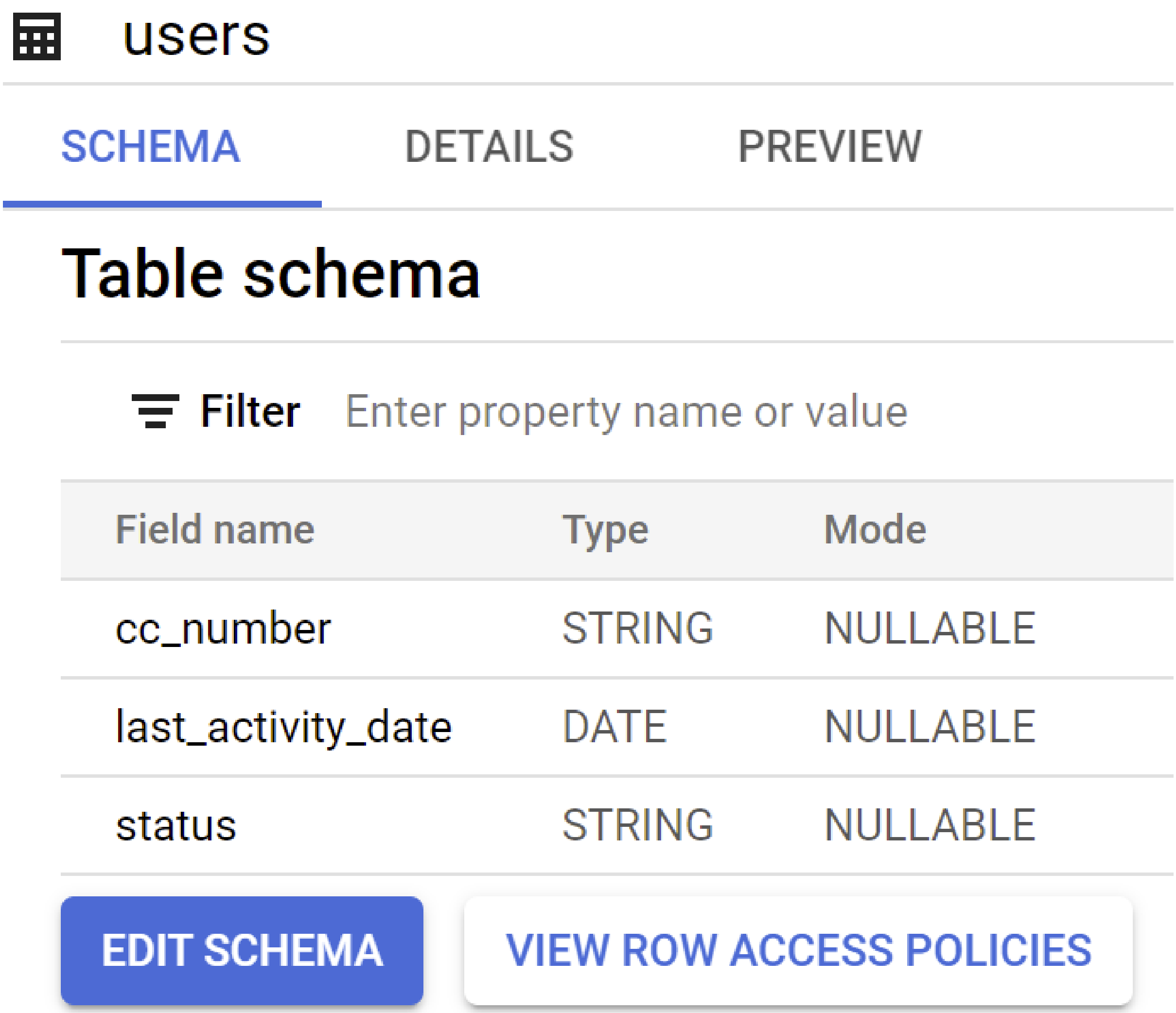

Data Engineer
User Group
Sales
User Group
Marketing
User Group
Head of Analysts
User Group
profile
Table
address
Table
check_out
Table
carts
Table
event
Table
metrics
Table

Access



Access Dataset

User Group	Role	Resource
Data Engineer	BigQuery Data Editor `roles/bigquery.dataEditor`	Project level
Head of Analytics	BigQuery Data Viewer `roles/bigquery.dataViewer`	Transaction dataset Campaign dataset
Sales	BigQuery Data Viewer `roles/bigquery.dataViewer`	Transaction dataset
Marketing	BigQuery Data Viewer `roles/bigquery.dataViewer`	Campaign dataset user_profile table



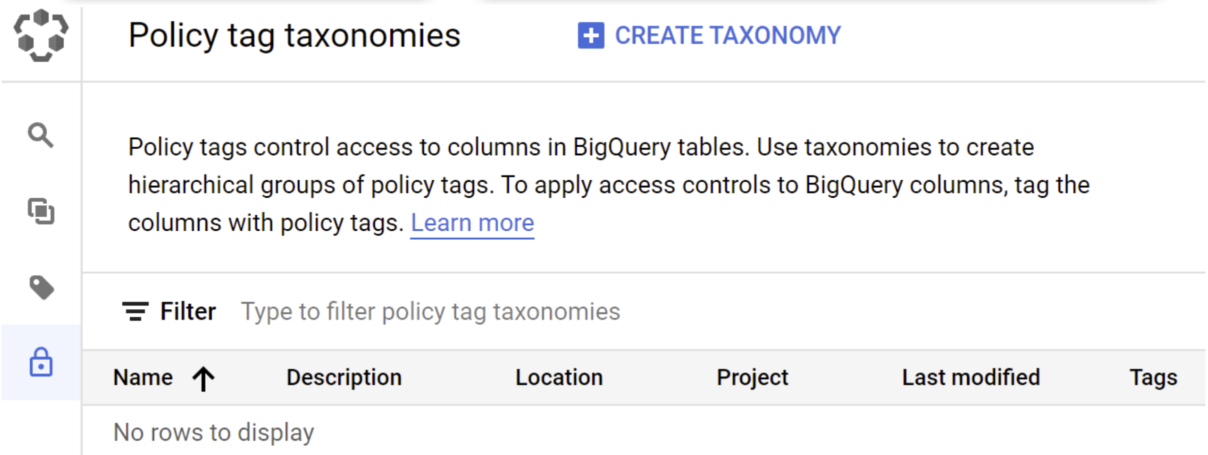
users

SCHEMA DETAILS PREVIEW

Table schema

Filter Enter property name or value

Field name	Type	Mode
cc_number	STRING	NULLABLE
last_activity_date	DATE	NULLABLE
status	STRING	NULLABLE



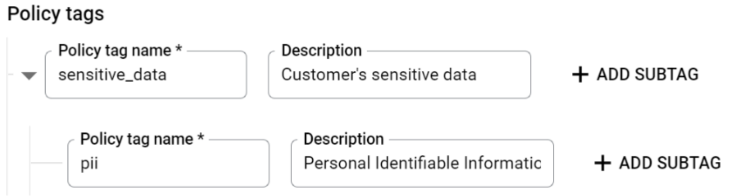
Policy tag taxonomies

Policy tags control access to columns in BigQuery tables. Use taxonomies to create hierarchical groups of policy tags. To apply access controls to BigQuery columns, tag the columns with policy tags. Learn more

Filter Type to filter policy tag taxonomies

Name ↑	Description	Location	Project	Last modified	Tags

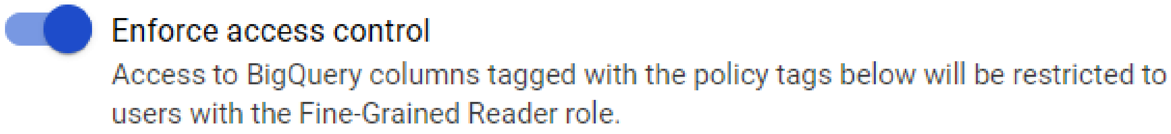
No rows to display



Policy tags

Policy tag name * sensitive_data
Description Customer's sensitive data

Policy tag name * pii
Description Personal Identifiable Informatic



Enforce access control

Access to BigQuery columns tagged with the policy tags below will be restricted to users with the Fine-Grained Reader role.

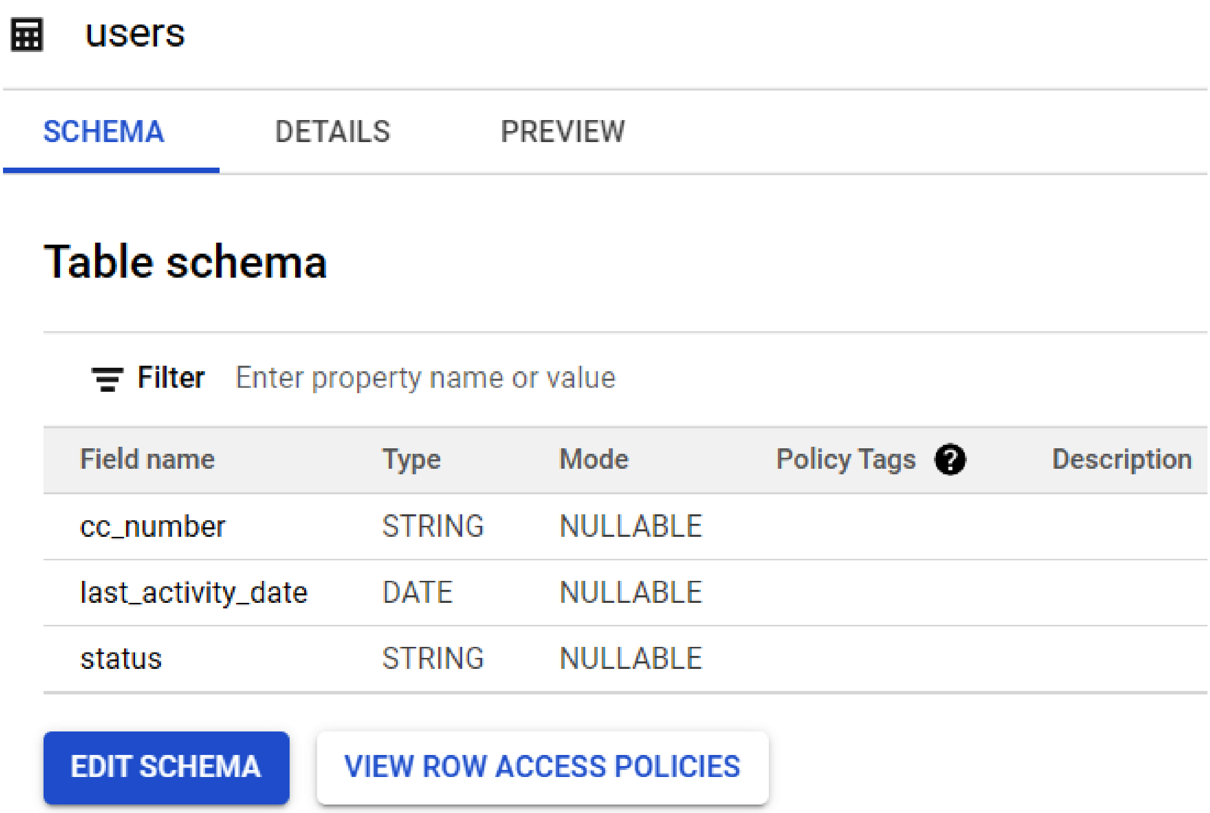
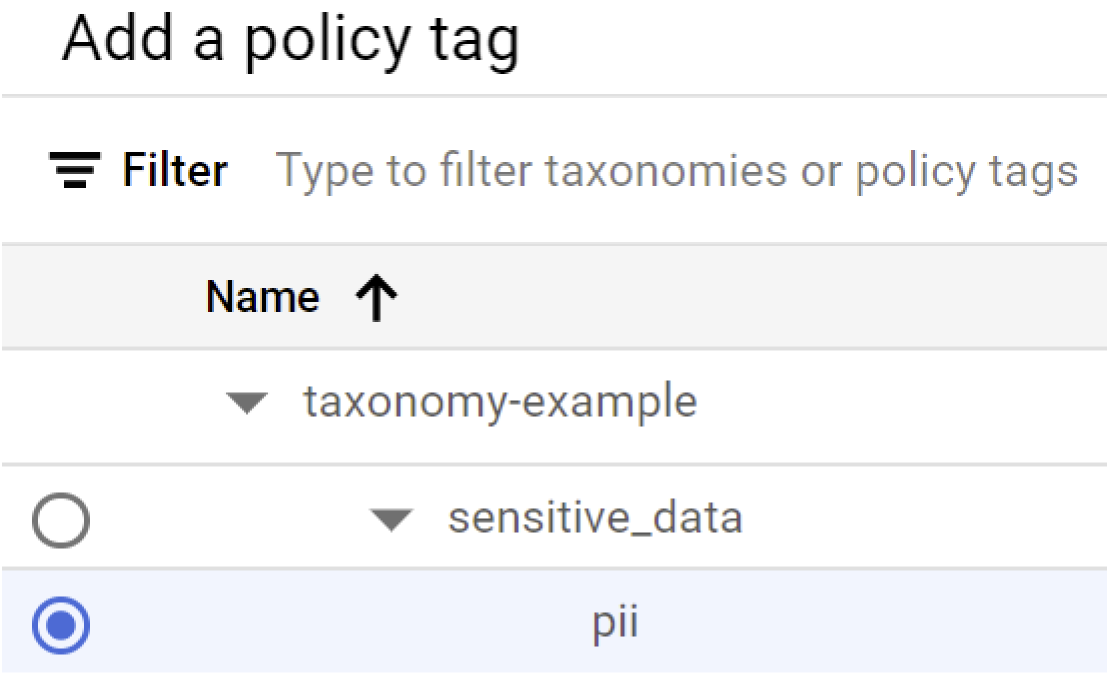
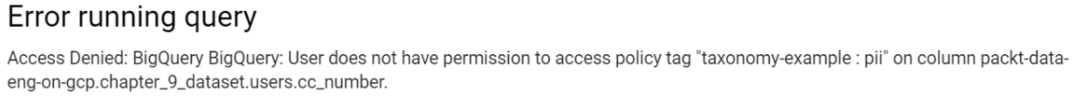


Table schema

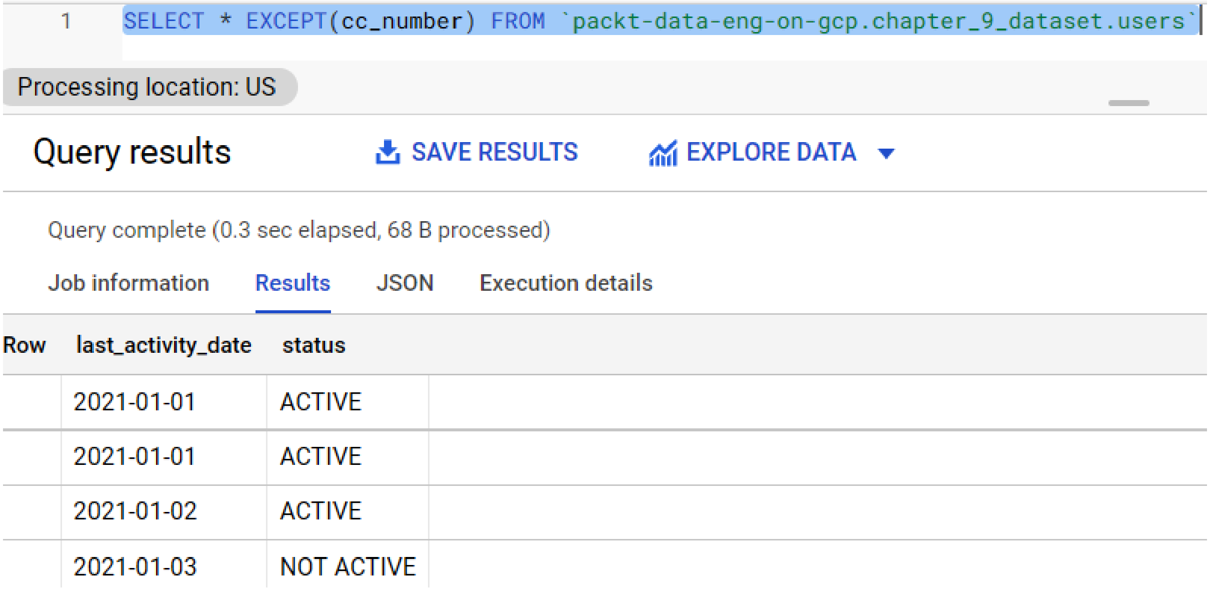
Field name	Type	Mode	Policy Tags	Description
cc_number	STRING	NULLABLE		
last_activity_date	DATE	NULLABLE		
status	STRING	NULLABLE		


Add a policy tag

- taxonomy-example
- ○ sensitive_data
- ● pii


Error running query

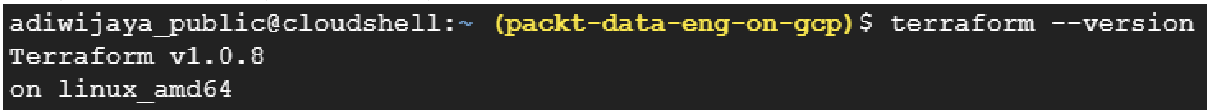
Access Denied: BigQuery BigQuery: User does not have permission to access policy tag "taxonomy-example : pii" on column packt-data-eng-on-gcp.chapter_9_dataset.users.cc_number.



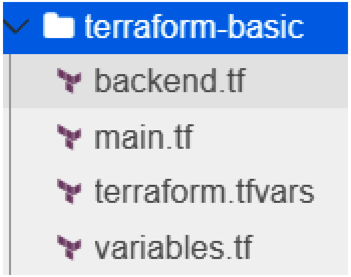
```
SELECT * EXCEPT(cc_number) FROM `packt-data-eng-on-gcp.chapter_9_dataset.users`
```

Query complete (0.3 sec elapsed, 68 B processed)

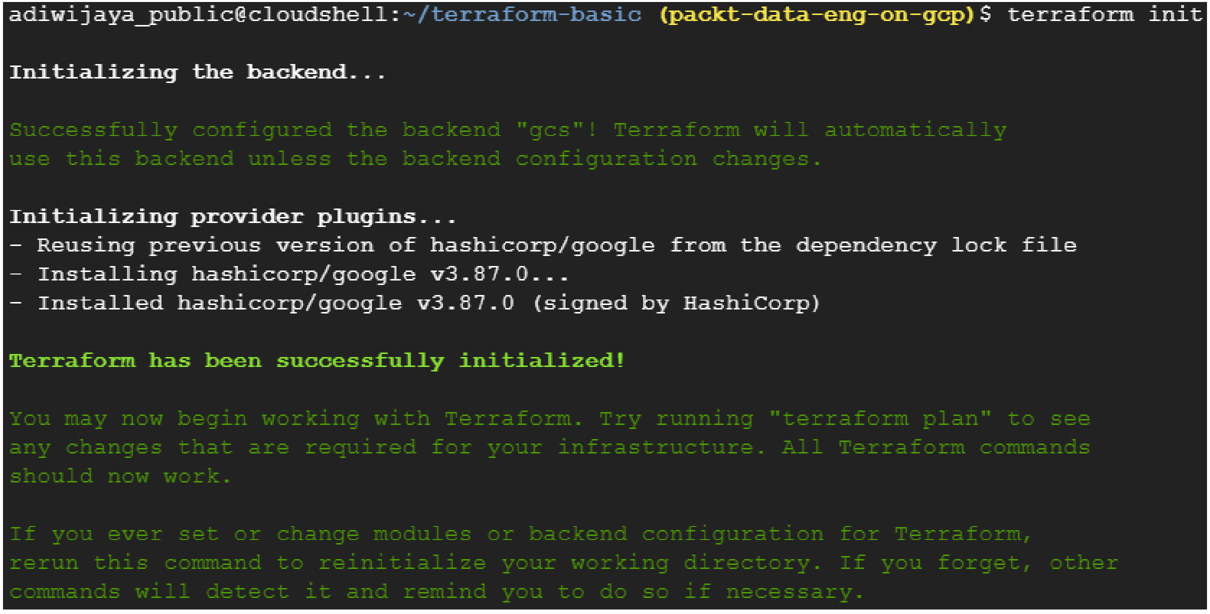
Row	last_activity_date	status
	2021-01-01	ACTIVE
	2021-01-01	ACTIVE
	2021-01-02	ACTIVE
	2021-01-03	NOT ACTIVE



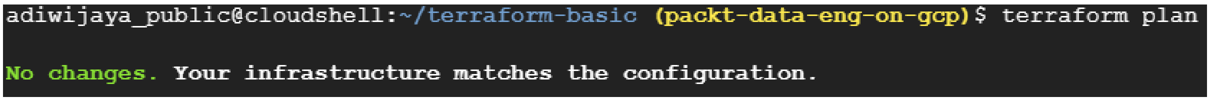
```
adiwijaya_public@cloudshell:~ (packt-data-eng-on-gcp)$ terraform --version
Terraform v1.0.8
on linux_amd64
```



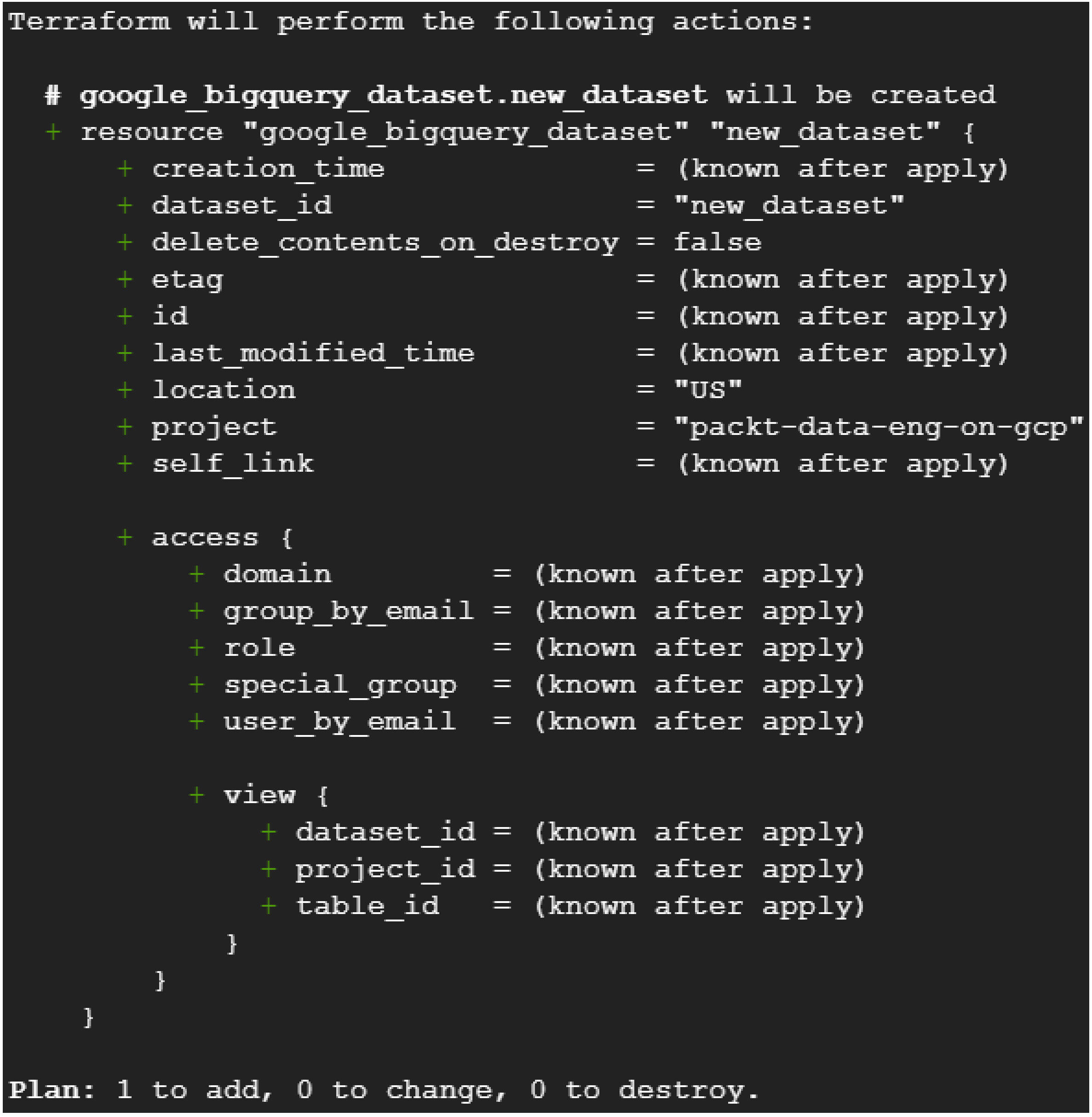
```
terraform-basic
    backend.tf
    main.tf
    terraform.tfvars
    variables.tf
```



```
adiwijaya_public@cloudshell:~/terraform-basic (packt-data-eng-on-gcp)$ terraform init

Initializing the backend...

Successfully configured the backend "gcs"! Terraform will automatically
use this backend unless the backend configuration changes.

Initializing provider plugins...
- Reusing previous version of hashicorp/google from the dependency lock file
- Installing hashicorp/google v3.87.0...
- Installed hashicorp/google v3.87.0 (signed by HashiCorp)

Terraform has been successfully initialized!

You may now begin working with Terraform. Try running "terraform plan" to see
any changes that are required for your infrastructure. All Terraform commands
should now work.

If you ever set or change modules or backend configuration for Terraform,
rerun this command to reinitialize your working directory. If you forget, other
commands will detect it and remind you to do so if necessary.
```

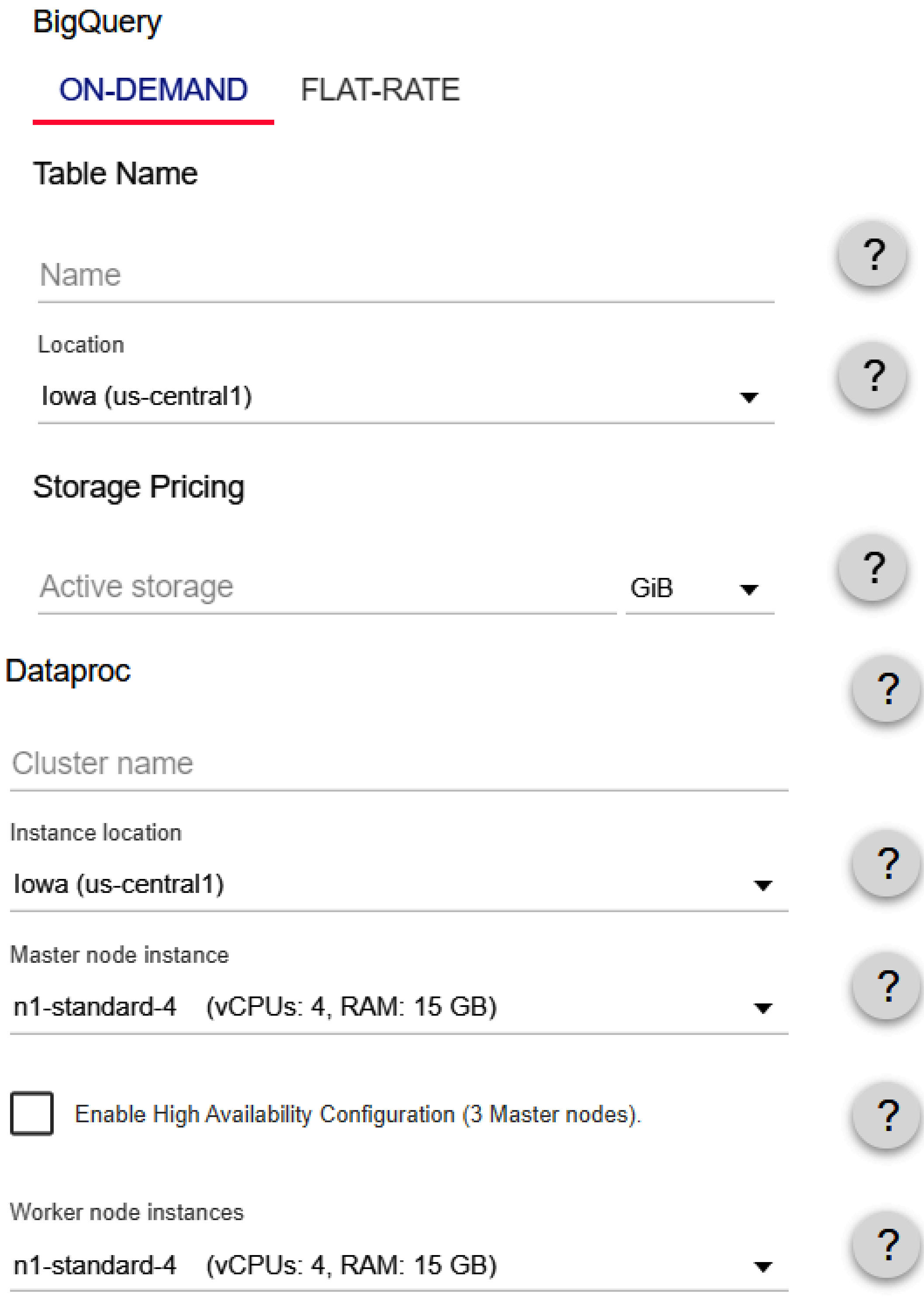


```
adiwijaya_public@cloudshell:~/terraform-basic (packt-data-eng-on-gcp)$ terraform plan

No changes. Your infrastructure matches the configuration.
```



```
Terraform will perform the following actions:

  # google_bigquery_dataset.new_dataset will be created
  + resource "google_bigquery_dataset" "new_dataset" {
      + creation_time              = (known after apply)
      + dataset_id                 = "new_dataset"
      + delete_contents_on_destroy = false
      + etag                       = (known after apply)
      + id                         = (known after apply)
      + last_modified_time         = (known after apply)
      + location                   = "US"
      + project                    = "packt-data-eng-on-gcp"
      + self_link                  = (known after apply)

      + access {
          + domain         = (known after apply)
          + group_by_email = (known after apply)
          + role           = (known after apply)
          + special_group  = (known after apply)
          + user_by_email  = (known after apply)

          + view {
              + dataset_id = (known after apply)
              + project_id = (known after apply)
              + table_id   = (known after apply)
            }
        }
    }

Plan: 1 to add, 0 to change, 0 to destroy.
```

Chapter 10: Cost Strategy in GCP



BigQuery

ON-DEMAND | FLAT-RATE

Table Name

Name

Location: Iowa (us-central1)

Storage Pricing

Active storage: GiB

Dataproc

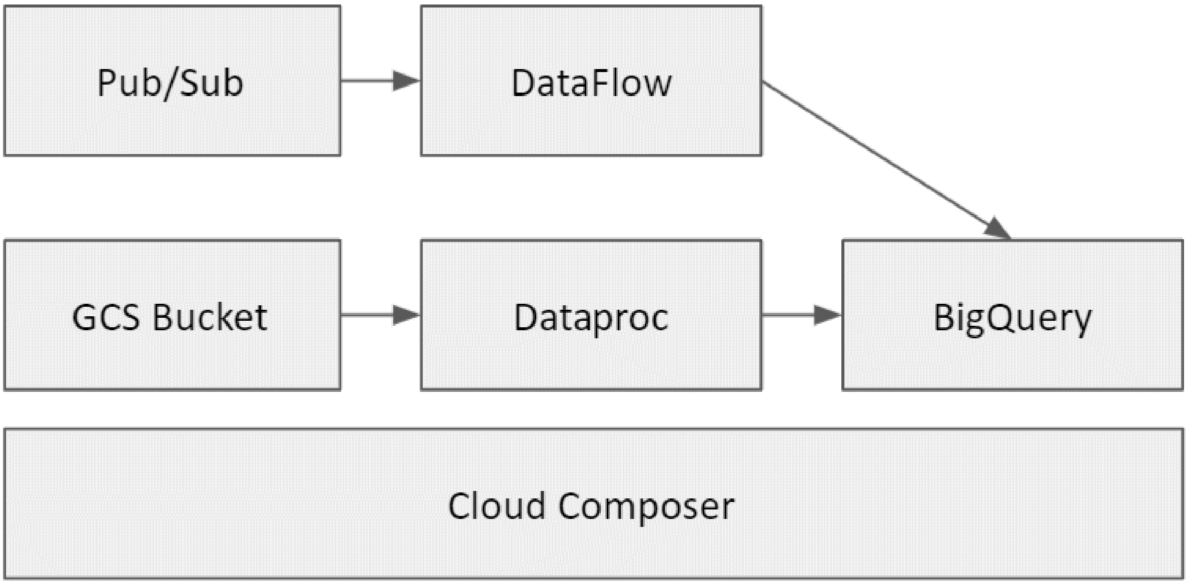
Cluster name

Instance location: Iowa (us-central1)

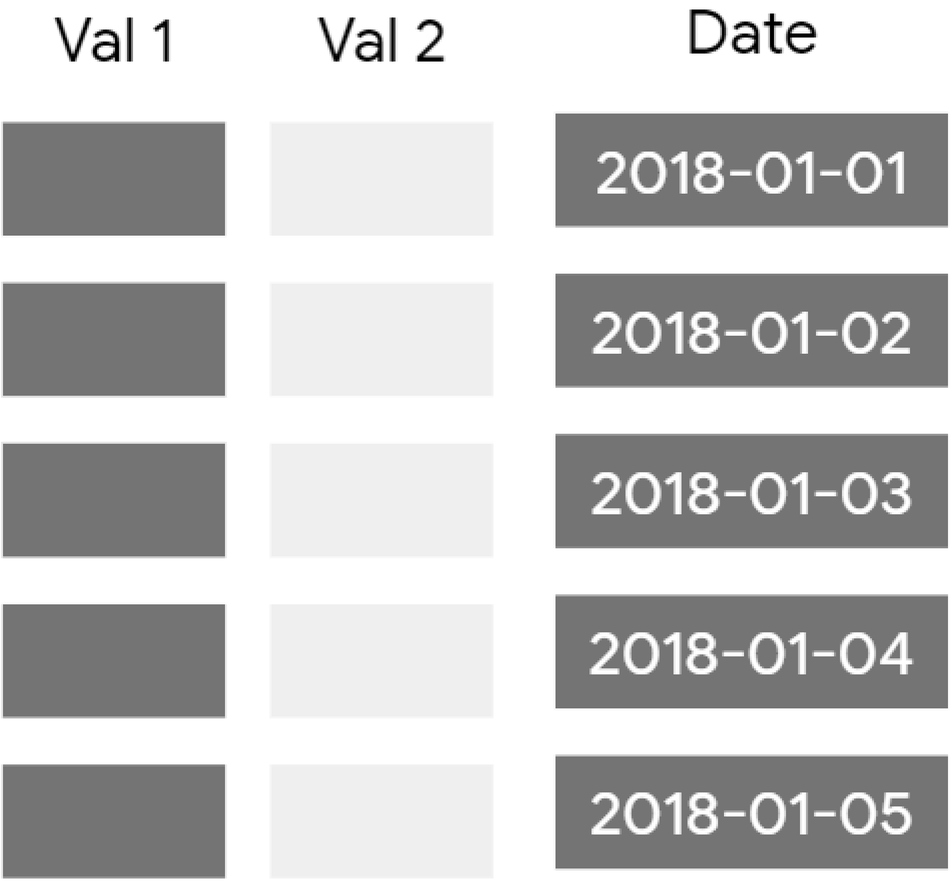
Master node instance: n1-standard-4 (vCPUs: 4, RAM: 15 GB)

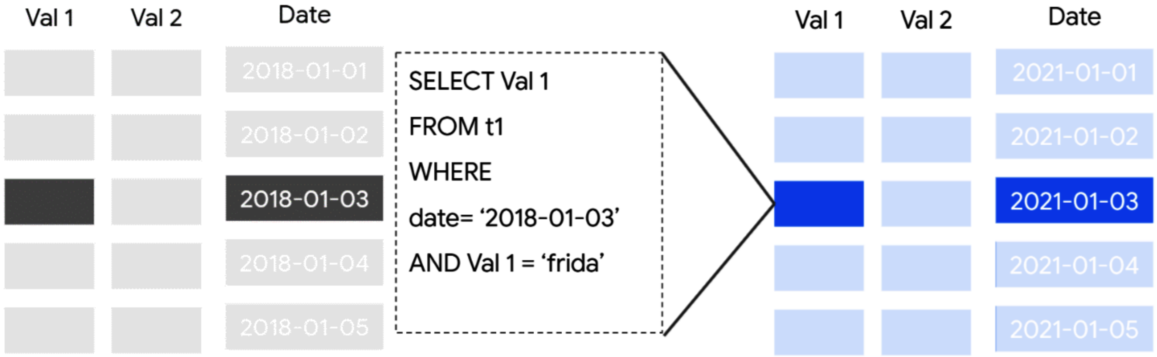
□ Enable High Availability Configuration (3 Master nodes).

Worker node instances: n1-standard-4 (vCPUs: 4, RAM: 15 GB)

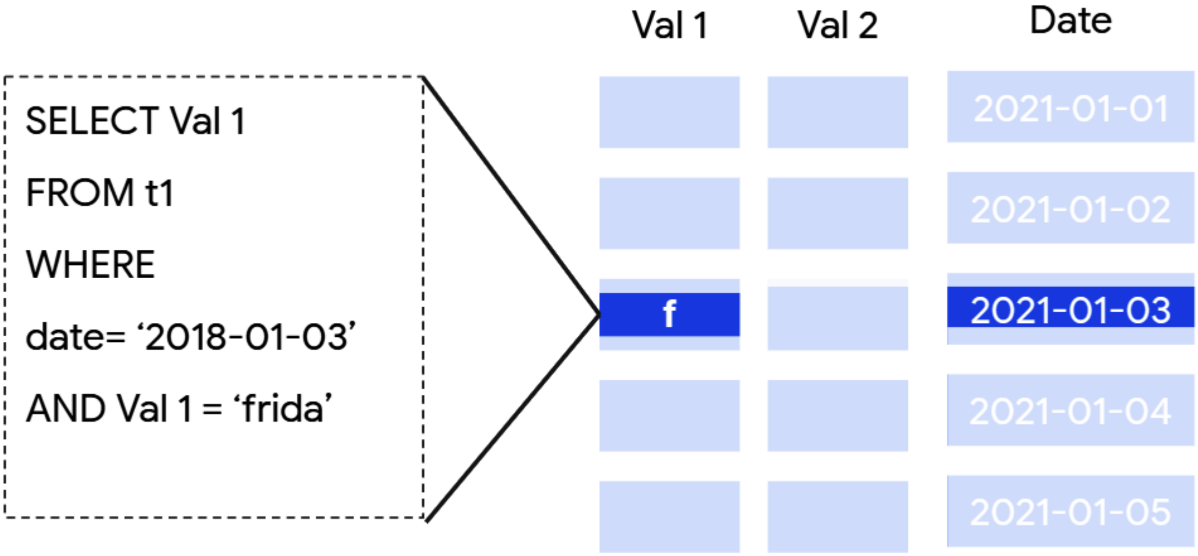


Val 1	Val 2	Date
		2018-01-01
		2018-01-02
		2018-01-03
		2018-01-04
		2018-01-05

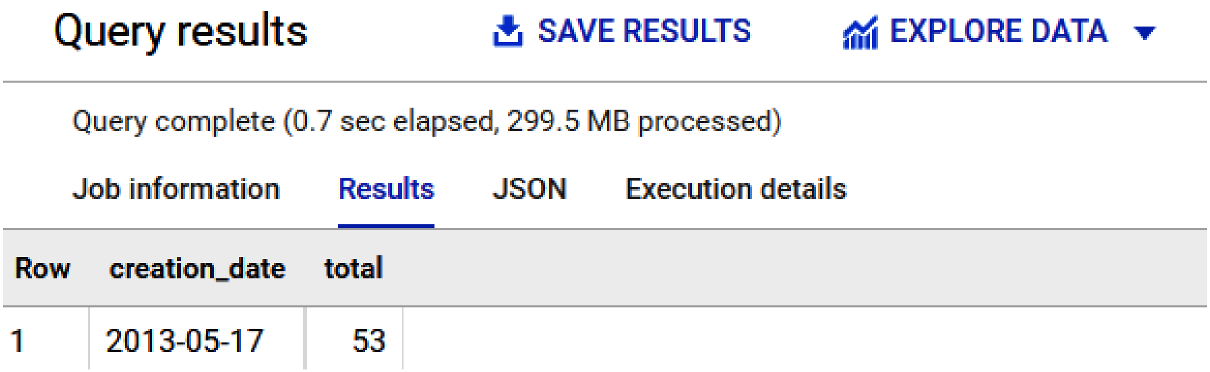




```
SELECT Val 1
FROM t1
WHERE
date= ‘2018-01-03’
AND Val 1 = ‘frida’
```



```
SELECT Val 1
FROM t1
WHERE
date= ‘2018-01-03’
AND Val 1 = ‘frida’
```



Query results

Query complete (0.7 sec elapsed, 299.5 MB processed)

Row	creation_date	total
1	2013-05-17	53

Table	Billed Bytes
Standard table	299.5 MB
Partitioned table	60.1 MB
Partitioned + Clustered table	57.2 MB

GCP Service	Cost Component	Requirements	Cost
Pub/Sub	The volume of bytes published daily	48 GB (2 GB x 24 hours)	$112.11
	Number of subscriptions	1	

GCP Service	Cost Component	Requirements	Cost
Dataflow	Job type	Streaming	$189.55
	Data processed	2 GB	
	Hours the job runs per month	720 hours (24 hours x 30 days)	
	Number of worker nodes used by the job	3	
	Worker node instance type	n1-standard-1	

GCP Service	Cost Component	Requirements	Cost
Cloud Storage	Total amount of storage	3,000 GB (100 GB x 30 days)	$60

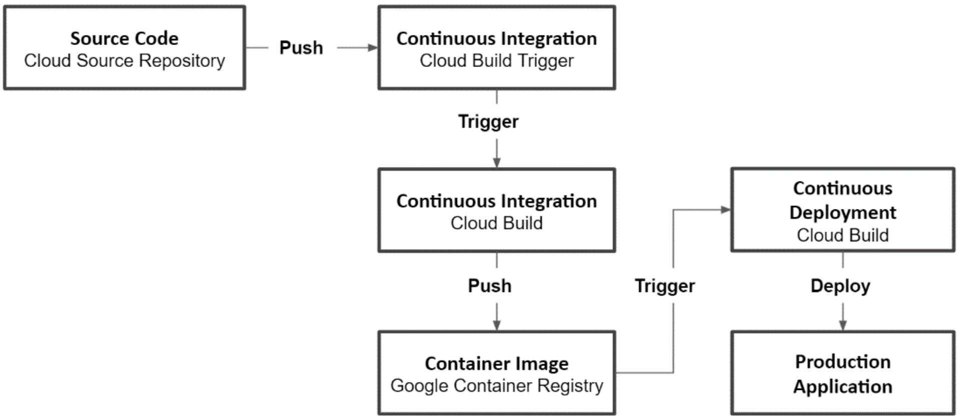
GCP Service	Cost Component	Requirements	Cost
Dataproc	Master node instance	n1-standard-4	$2,062.63
	Enable High Availability Configuration (three master nodes)	Yes	
	Worker node instance	n1-standard-4	
	Number of normal worker nodes	10	
	Hours the cluster runs per month	720 hours (24 hours x 30 days)	
	Storage (per node)	PD SSD – 200 GiB	

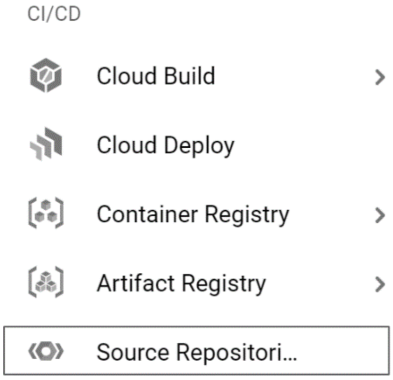
GCP Service	Cost Component	Requirements	Cost
Cloud Composer	Number of workers	3	$298.73
	Average hours per day each server is running	24	
	Average days per week each server is running	7	

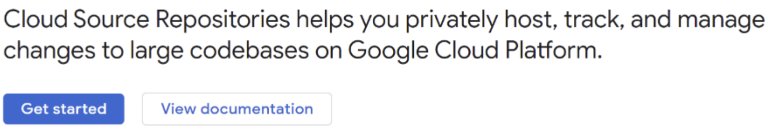
GCP Service	Cost Component	Requirements	Cost
BigQuery	Queries	20 end users x 5 days x 4 weeks x 100 GB (40,000 GB)	$397.08
	Active storage	300 GB x 30 days (9,000 GB)	

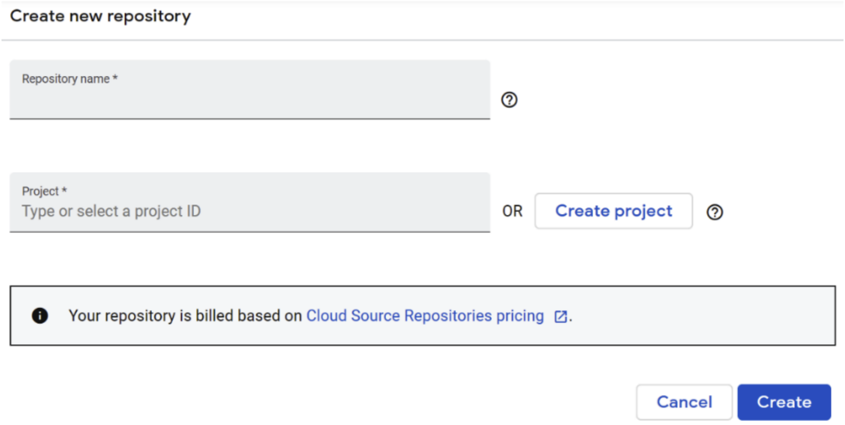
Service	Cost Monthly
Pub/Sub	$112.11
Dataflow	$189.55
Cloud Storage	$60
Dataproc	$2,062.63
Cloud Composer	$298.73
BigQuery	$397.08
Total	$3,120.1

Chapter 11: CI/CD on Google Cloud Platform for Data Engineers






Cloud Source Repositories helps you privately host, track, and manage changes to large codebases on Google Cloud Platform.

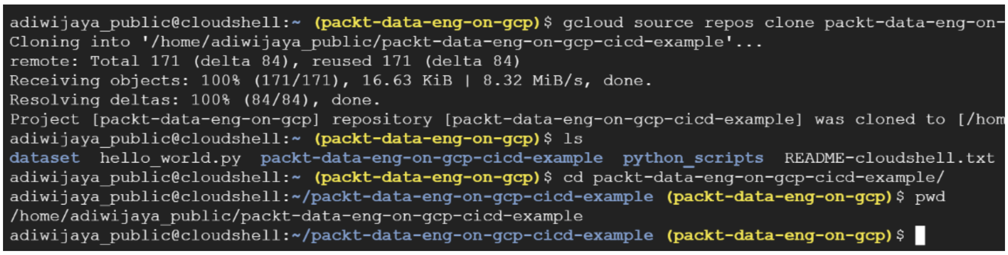

Create new repository

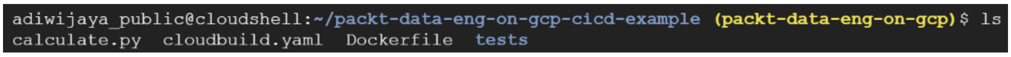
Repository name *

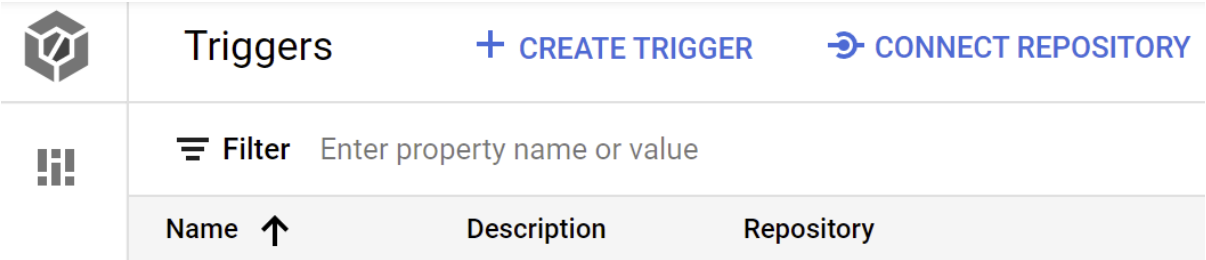
Project *
Type or select a project ID

OR

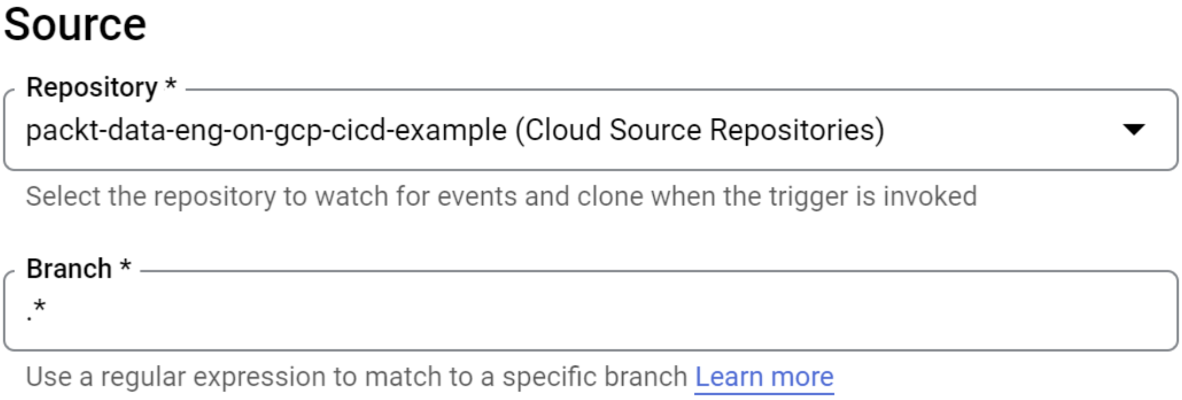
Your repository is billed based on Cloud Source Repositories pricing.


```
adiwijaya_public@cloudshell:~ (packt-data-eng-on-gcp)$ gcloud source repos clone packt-data-eng-on-
Cloning into '/home/adiwijaya_public/packt-data-eng-on-gcp-cicd-example'...
remote: Total 171 (delta 84), reused 171 (delta 84)
Receiving objects: 100% (171/171), 16.63 KiB | 8.32 MiB/s, done.
Resolving deltas: 100% (84/84), done.
Project [packt-data-eng-on-gcp] repository [packt-data-eng-on-gcp-cicd-example] was cloned to [/hom
adiwijaya_public@cloudshell:~ (packt-data-eng-on-gcp)$ ls
dataset  hello_world.py  packt-data-eng-on-gcp-cicd-example  python_scripts  README-cloudshell.txt
adiwijaya_public@cloudshell:~ (packt-data-eng-on-gcp)$ cd packt-data-eng-on-gcp-cicd-example/
adiwijaya_public@cloudshell:~/packt-data-eng-on-gcp-cicd-example (packt-data-eng-on-gcp)$ pwd
/home/adiwijaya_public/packt-data-eng-on-gcp-cicd-example
adiwijaya_public@cloudshell:~/packt-data-eng-on-gcp-cicd-example (packt-data-eng-on-gcp)$
```


```
adiwijaya_public@cloudshell:~/packt-data-eng-on-gcp-cicd-example (packt-data-eng-on-gcp)$ ls
calculate.py  cloudbuild.yaml  Dockerfile  tests
```

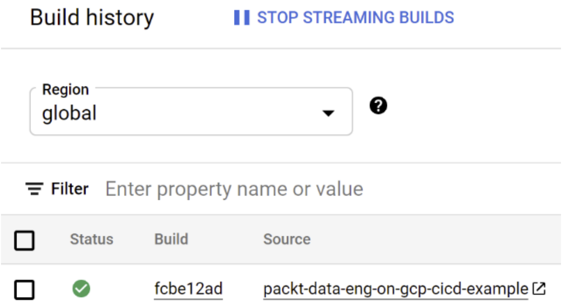


Name ↑	Description	Repository



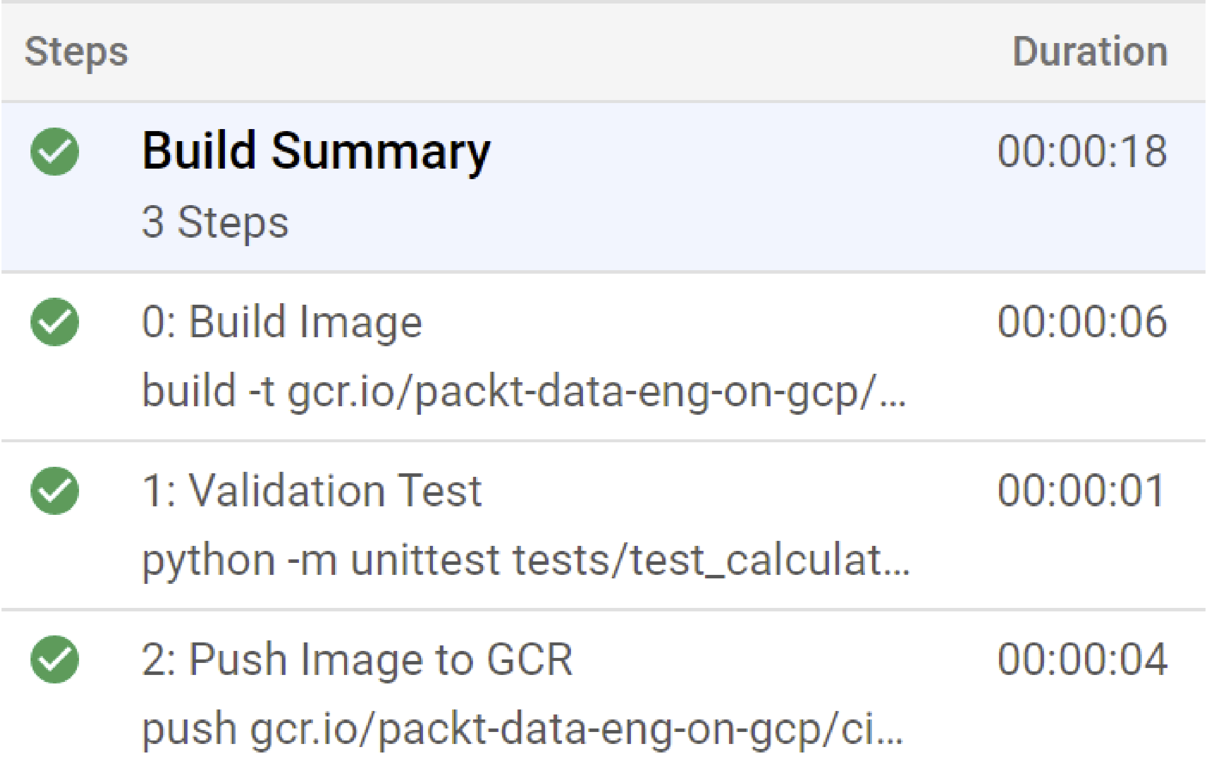
Repository *
packt-data-eng-on-gcp-cicd-example (Cloud Source Repositories)
Select the repository to watch for events and clone when the trigger is invoked

Branch *
.*
Use a regular expression to match to a specific branch Learn more

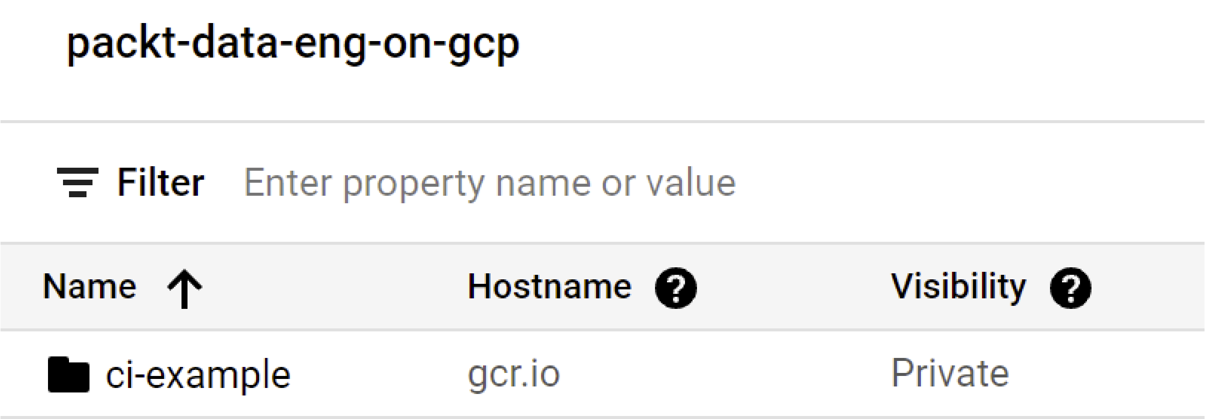


Region
global

□	Status	Build	Source
□	✓	fcbe12ad	packt-data-eng-on-gcp-cicd-example

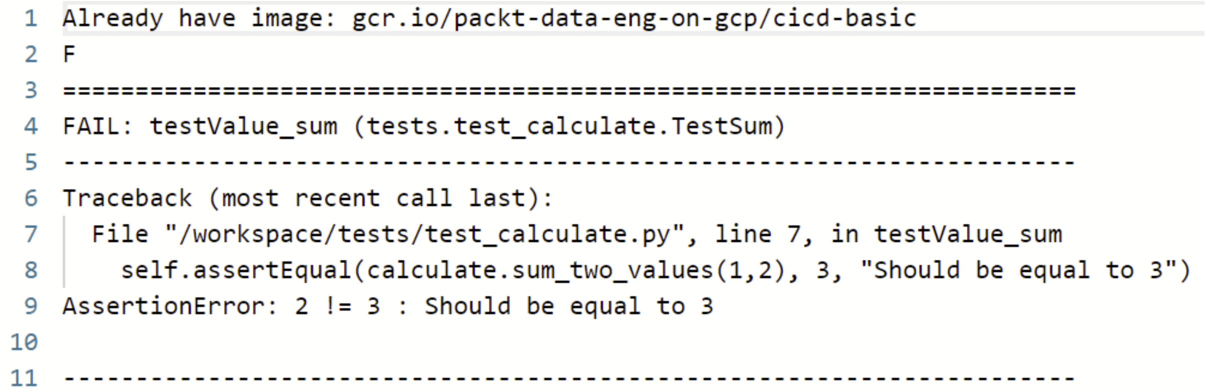


Steps	Duration
✓ **Build Summary** 3 Steps	00:00:18
✓ 0: Build Image build -t gcr.io/packt-data-eng-on-gcp/...	00:00:06
✓ 1: Validation Test python -m unittest tests/test_calculat...	00:00:01
✓ 2: Push Image to GCR push gcr.io/packt-data-eng-on-gcp/ci...	00:00:04

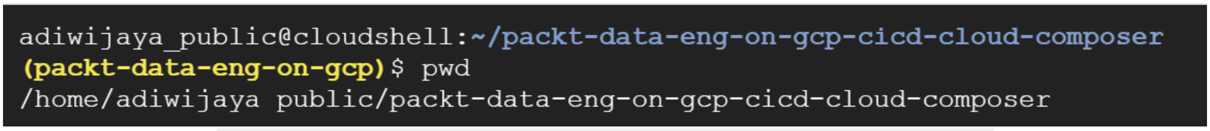


packt-data-eng-on-gcp

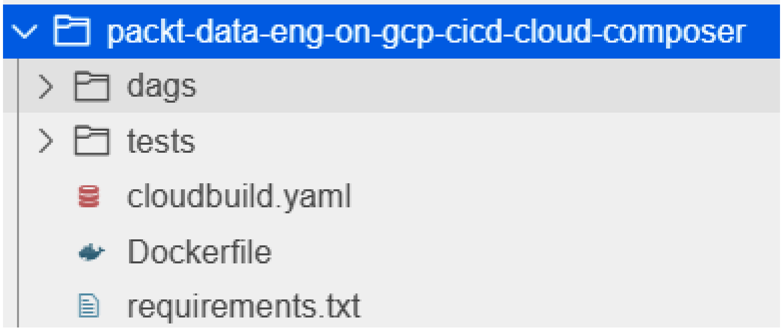
Name	Hostname	Visibility
ci-example	gcr.io	Private



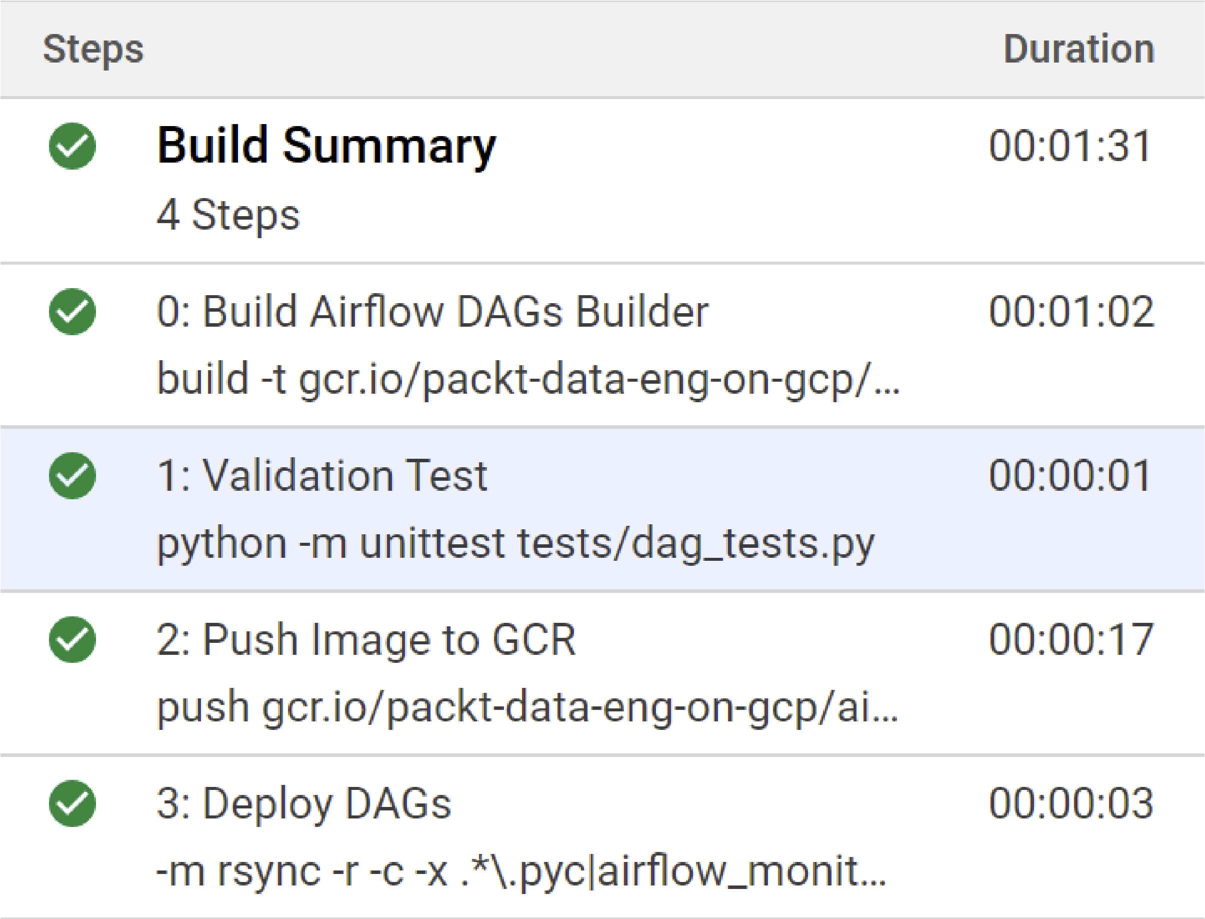
```
Already have image: gcr.io/packt-data-eng-on-gcp/cicd-basic
F
======================================================================
FAIL: testValue_sum (tests.test_calculate.TestSum)
----------------------------------------------------------------------
Traceback (most recent call last):
  File "/workspace/tests/test_calculate.py", line 7, in testValue_sum
    self.assertEqual(calculate.sum_two_values(1,2), 3, "Should be equal to 3")
AssertionError: 2 != 3 : Should be equal to 3

----------------------------------------------------------------------
```



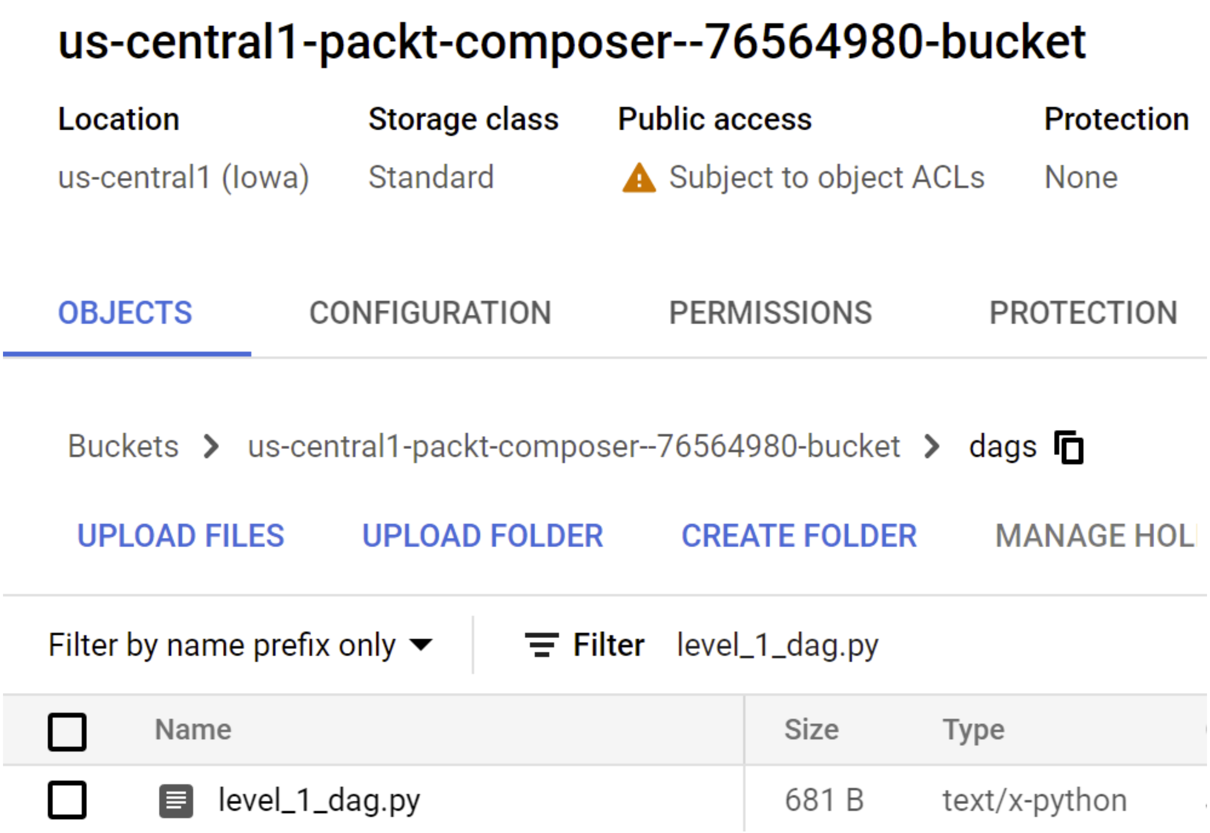
```
adiwijaya_public@cloudshell:~/packt-data-eng-on-gcp-cicd-cloud-composer
(packt-data-eng-on-gcp)$ pwd
/home/adiwijaya public/packt-data-eng-on-gcp-cicd-cloud-composer
```



```
packt-data-eng-on-gcp-cicd-cloud-composer
  dags
  tests
  cloudbuild.yaml
  Dockerfile
  requirements.txt
```



Steps	Duration
✓ **Build Summary** 4 Steps	00:01:31
✓ 0: Build Airflow DAGs Builder build -t gcr.io/packt-data-eng-on-gcp/...	00:01:02
✓ 1: Validation Test python -m unittest tests/dag_tests.py	00:00:01
✓ 2: Push Image to GCR push gcr.io/packt-data-eng-on-gcp/ai...	00:00:17
✓ 3: Deploy DAGs -m rsync -r -c -x .*\.pyc|airflow_monit...	00:00:03



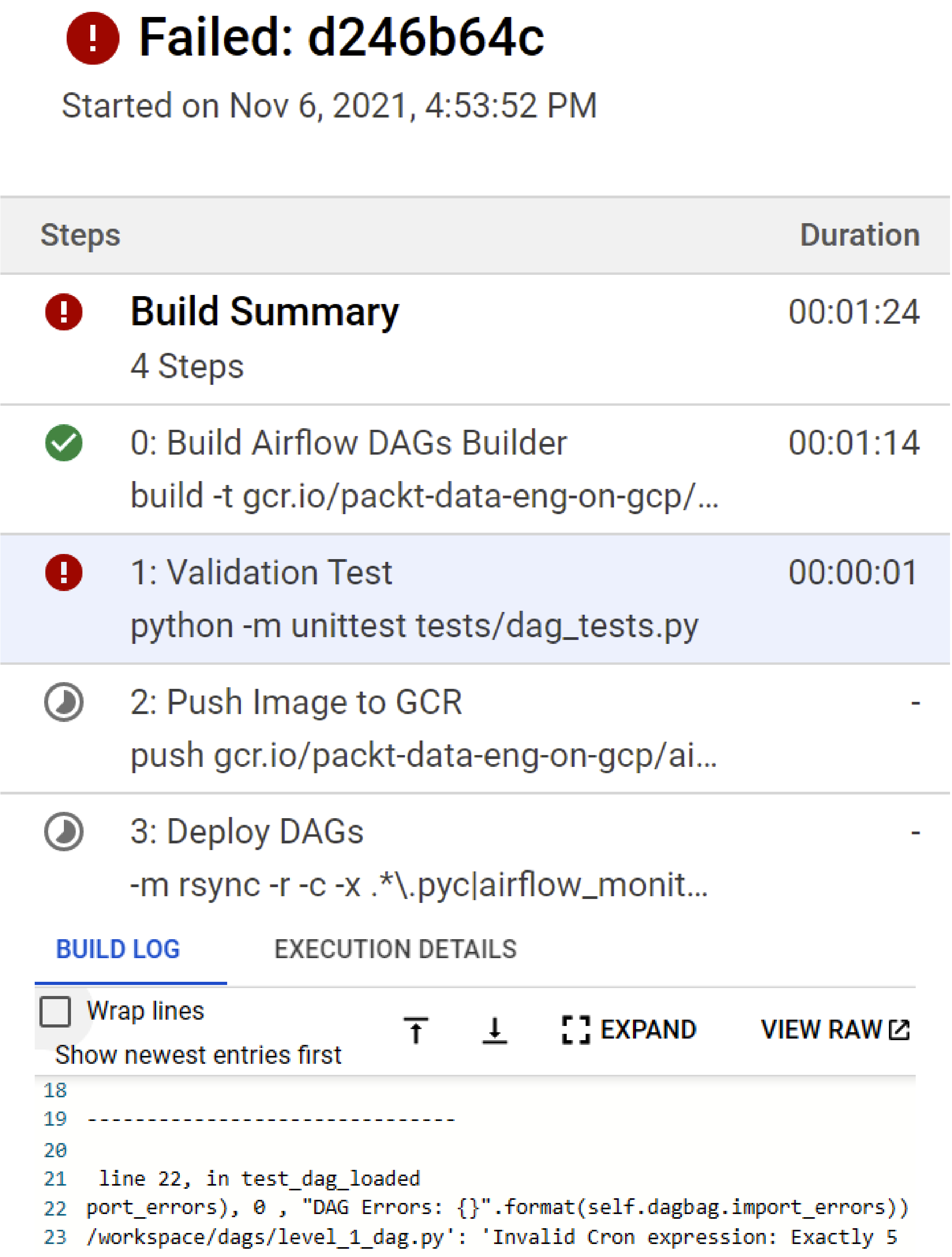
us-central1-packt-composer--76564980-bucket

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	⚠ Subject to object ACLs	None

Buckets > us-central1-packt-composer--76564980-bucket > dags

Filter by name prefix only ▾ Filter level_1_dag.py

□	Name	Size	Type
□	level_1_dag.py	681 B	text/x-python



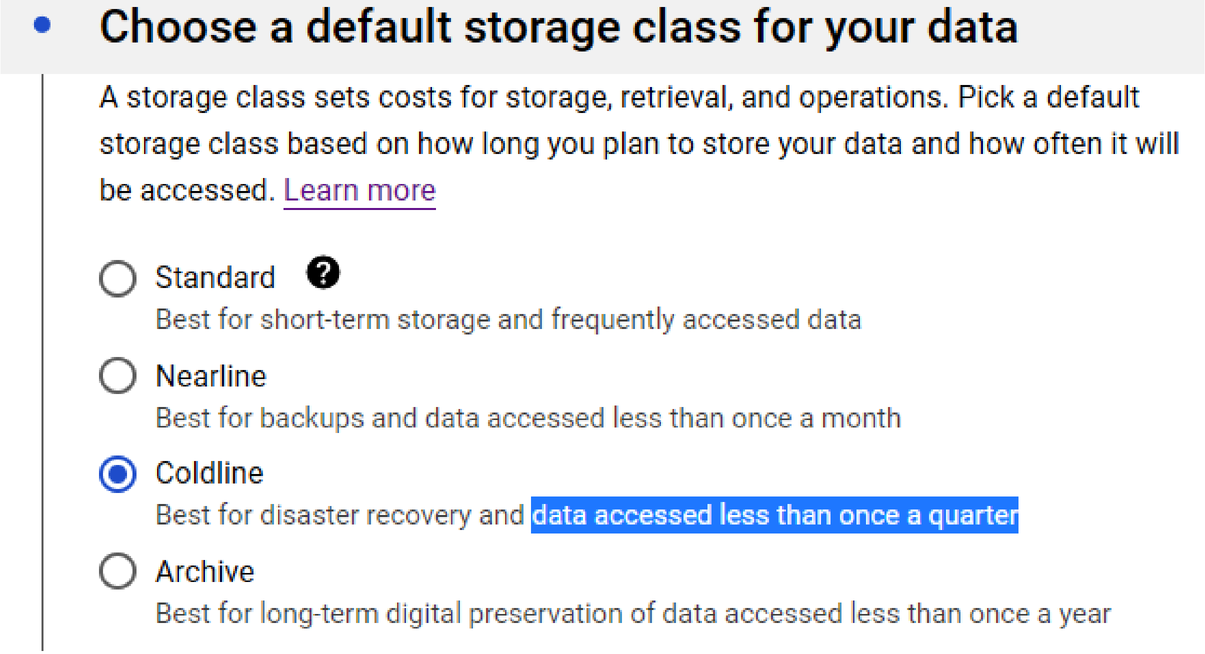
Failed: d246b64c

Started on Nov 6, 2021, 4:53:52 PM

Steps	Duration
Build Summary 4 Steps	00:01:24
0: Build Airflow DAGs Builder build -t gcr.io/packt-data-eng-on-gcp/...	00:01:14
1: Validation Test python -m unittest tests/dag_tests.py	00:00:01
2: Push Image to GCR push gcr.io/packt-data-eng-on-gcp/ai...	-
3: Deploy DAGs -m rsync -r -c -x .*\.pyc|airflow_monit...	-

```

------------------------------

 line 22, in test_dag_loaded
port_errors), 0 , "DAG Errors: {}".format(self.dagbag.import_errors))
/workspace/dags/level_1_dag.py': 'Invalid Cron expression: Exactly 5
```

Chapter 12: Boosting Your Confidence as a Data Engineer



Choose a default storage class for your data

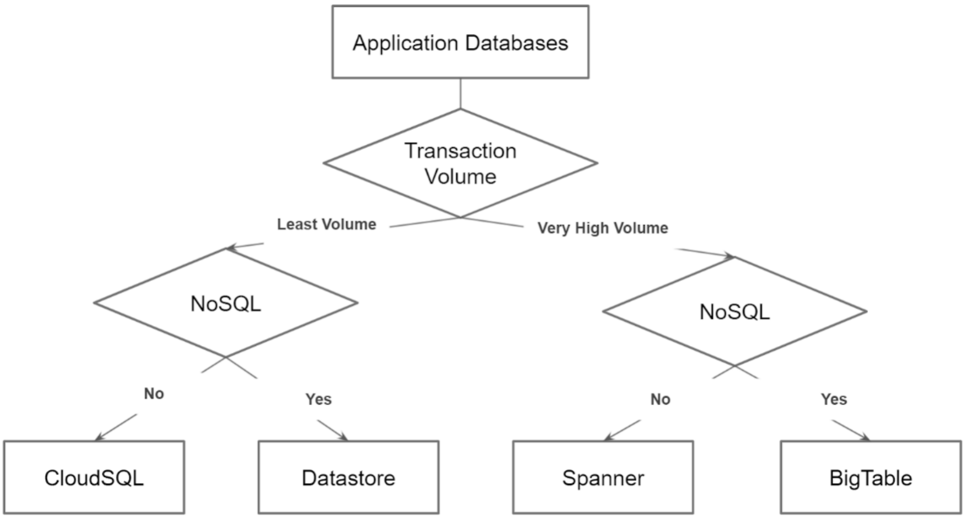
A storage class sets costs for storage, retrieval, and operations. Pick a default storage class based on how long you plan to store your data and how often it will be accessed. Learn more

○ Standard
Best for short-term storage and frequently accessed data

○ Nearline
Best for backups and data accessed less than once a month

● Coldline
Best for disaster recovery and data accessed less than once a quarter

○ Archive
Best for long-term digital preservation of data accessed less than once a year


Application Databases
Transaction Volume
Least Volume
Very High Volume
NoSQL
NoSQL
No
Yes
No
Yes
CloudSQL
Datastore
Spanner
BigTable