

**dataengineer\_yellow\_taxi\_002**

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Regional	Subject to object ACLs	Soft Delete

< **OBJECTS** CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE OBSERVABILITY INVENTORY REPORTS >

Folder browser **dataengineer\_yellow\_taxi\_002**

Buckets > dataengineer\_yellow\_taxi\_002

CREATE FOLDER UPLOAD TRANSFER DATA OTHER SERVICES

Filter by name prefix only Filter objects and folders Show **Live objects only**

Uploads and dataengineer operations

- Uploading 3 files
  - yellow\_tripdata\_2024-04.parquet
  - yellow\_tripdata\_2024-05.parquet
  - yellow\_tripdata\_2024-06.parquet

1 file successfully uploaded

```
CREATE SCHEMA `dataengineer.us_central1_dataset`
OPTIONS(location="us-central1");
```

```
CREATE OR REPLACE EXTERNAL TABLE
`dataengineer.us_central1_dataset.external_yellow_taxi_002`
OPTIONS(
  format = 'PARQUET',
  uris = ['gs://dataengineer_yellow_taxi_002/*.parquet']
);
```

```
SELECT COUNT(*)
From dataengineer.us_central1_dataset.external_yellow_taxi_002;
```

**Question 1: What is count of records for the 2024 Yellow Taxi Data?**

- ☐ 65,623
- ☐ 840,402
- ☒ 20,332,093
- ☐ 85,431,289

Untitled query RUN SAVE DOWNLOAD

```
1 CREATE SCHEMA `dataengineer.us_central1_dataset`
2 OPTIONS(location="us-central1");
3
4 CREATE OR REPLACE EXTERNAL TABLE `dataengineer.us_central1_dataset.external_yellow_taxi_002`
5 OPTIONS(
6   format='PARQUET',
7   uris=['gs://dataengineer_yellow_taxi_002/*.parquet']
8 );
9
10 select count(*)
11 from dataengineer.us_central1_dataset.external_yellow_taxi_002;
12
```

Press Alt+F1 for Accessibility Option

Query results SAVE RESULTS OPEN IN

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Row	f0_					
1	20332093					

**Question 2:**Write a query to count the distinct number of PULocationIDs for the entire dataset on both the tables.

**What is the estimated amount of data that will be read when this query is executed on the External Table and the Table?**

- ☐ 18.82 MB for the External Table and 47.60 MB for the Materialized Table
- ☒ 0 MB for the External Table and 155.12 MB for the Materialized Table
- ☐ 2.14 GB for the External Table and 0MB for the Materialized Table
- ☐ 0 MB for the External Table and 0MB for the Materialized Table

Untitled query RUN SAVE DOWNLOAD SHARE SCHEDULE OPEN IN MO

```
1 SELECT COUNT(DISTINCT PULocationID) FROM `dataenigneer.us_central1_dataset.external_yellow_taxi_002`;
2
```

No cached results × Press Alt+F1 for help

Query results SAVE RESULTS 📊

JOB INFORMATION	RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Job ID	dataenigneer:us-central1.bqjob_5b7c9c0d_194f576f8d1				
User	<span style="background-color: red; color: black;">[REDACTED]</span>				
Location	us-central1				
Creation time	Feb 11, 2025, 9:43:22 AM UTC-5				
Start time	Feb 11, 2025, 9:43:22 AM UTC-5				
End time	Feb 11, 2025, 9:43:23 AM UTC-5				
Duration	1 sec				
Bytes processed	155.12 MB				
Bytes billed	156 MB				
Slot milliseconds	5109				

**Question 3: Write a query to retrieve the PULocationID from the table (not the external table) in BigQuery. Now write a query to retrieve the PULocationID and DOLocationID on the same table. Why are the estimated number of Bytes different?**

- ☒ BigQuery is a columnar database, and it only scans the specific columns requested in the query. Querying two columns (PULocationID, DOLocationID) requires reading more data than querying one column (PULocationID), leading to a higher estimated number of bytes processed.
- ☐ BigQuery duplicates data across multiple storage partitions, so selecting two columns instead of one requires scanning the table twice, doubling the estimated bytes processed.
- ☐ BigQuery automatically caches the first queried column, so adding a second column increases processing time but does not affect the estimated bytes scanned.
- ☐ When selecting multiple columns, BigQuery performs an implicit join operation between them, increasing the estimated bytes processed.

**Question 4: How many records have a fare\_amount of 0?**

- ☐ 128,210
- ☐ 546,578
- ☐ 20,188,016
- ☒ 8,333

```

8
9 SELECT COUNT(fare_amount) FROM `dataengineer.us_central1_dataset.external_yellow_taxi_002`
10 WHERE fare_amount = 0;

```

Query results [SAVE RESULTS](#)

JOB INFORMATION **RESULTS** CHART JSON EXECUTION DETAILS EXECUTION GRAPH

**i** Metadata caching is disabled. You can accelerate queries over external tables by enabling metadata caching. [Learn more.](#)

Row	f0_
1	8333

**Question 5: What is the best strategy to make an optimized table in Big Query if your query will always filter based on tpep\_dropoff\_datetime and order the results by VendorID (Create a new table with this strategy)**

- ☒ Partition by tpep\_dropoff\_datetime and Cluster on VendorID
- ☐ Cluster on by tpep\_dropoff\_datetime and Cluster on VendorID
- ☐ Cluster on tpep\_dropoff\_datetime Partition by VendorID
- ☐ Partition by tpep\_dropoff\_datetime and Partition by VendorID

**Question 6: Write a query to retrieve the distinct VendorIDs between tpep\_dropoff\_datetime 2024-03-01 and 2024-03-15 (inclusive)**

**Use the materialized table you created earlier in your from clause and note the estimated bytes. Now change the table in the from clause to the partitioned table you created for question 5 and note the estimated bytes processed. What are these values?**

**Choose the answer which most closely matches.**

- ☐ 12.47 MB for non-partitioned table and 326.42 MB for the partitioned table
- ☒ 310.24 MB for non-partitioned table and 26.84 MB for the partitioned table
- ☐ 5.87 MB for non-partitioned table and 0 MB for the partitioned table
- ☐ 310.31 MB for non-partitioned table and 285.64 MB for the partitioned table

external\_...002 X \*Untitled query X +

Untitled query RUN SAVE DOWNLOAD SHARE SCHEDULE OPEN IN MORE

```

1 SELECT DISTINCT VendorID
2 FROM `dataengineer.us_central1_dataset.external_yellow_taxi_002`
3 WHERE tpep_dropoff_datetime BETWEEN '2024-03-01' AND '2024-03-15';
4

```

No cached results Press Alt+F1 for /

## Query results

SAVE RESULTS OPE

JOB INFORMATION	RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Job ID	dataengineer:us-central1.bqjob_75e00b7a_194f57c329a				
User					
Location	us-central1				
Creation time	Feb 11, 2025, 9:49:05 AM UTC-5				
Start time	Feb 11, 2025, 9:49:05 AM UTC-5				
End time	Feb 11, 2025, 9:49:07 AM UTC-5				
Duration	1 sec				
Bytes processed	310.24 MB				
Bytes billed	311 MB				
Slot milliseconds	12267				

**Question 7: Where is the data stored in the External Table you created?**

- ☐ Big Query
- ☐ Container Registry
- ☒ GCP Bucket
- ☐ Big Table

**Question 8 :It is best practice in Big Query to always cluster your data:**

- ☐ True
- ☒ False