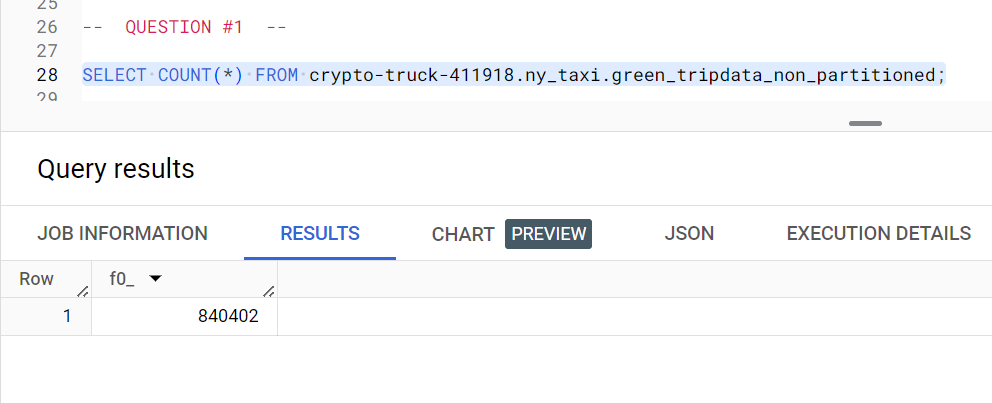
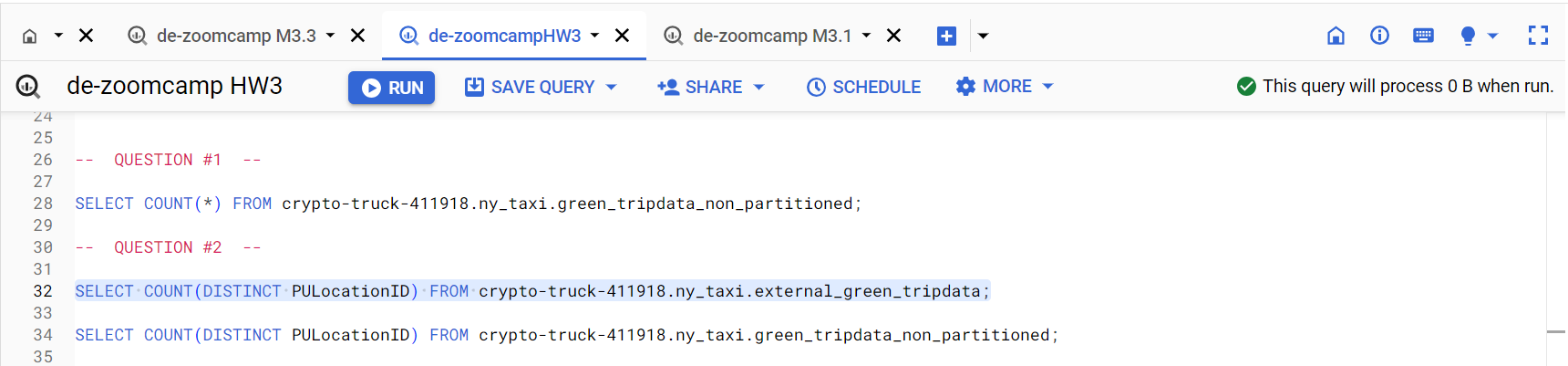
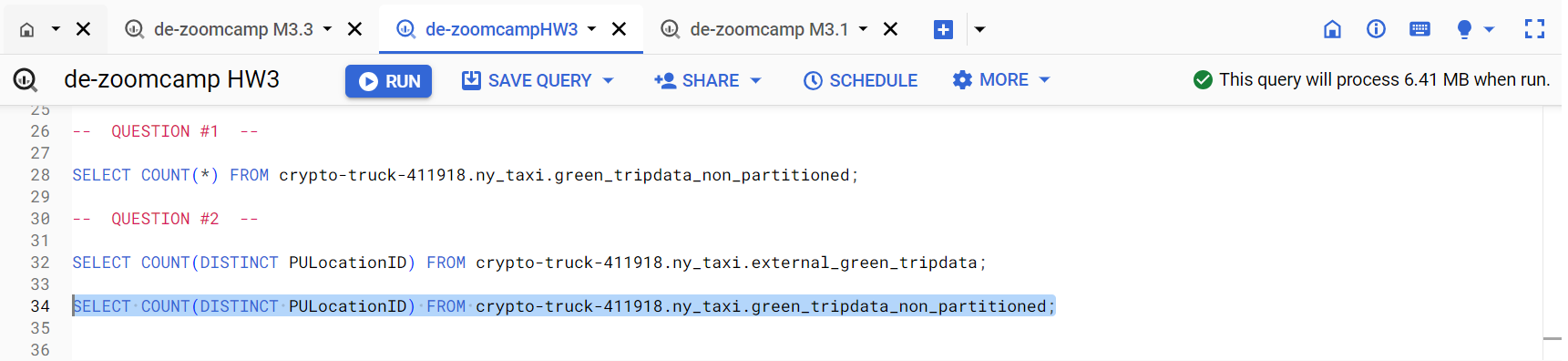
**Homework #3**

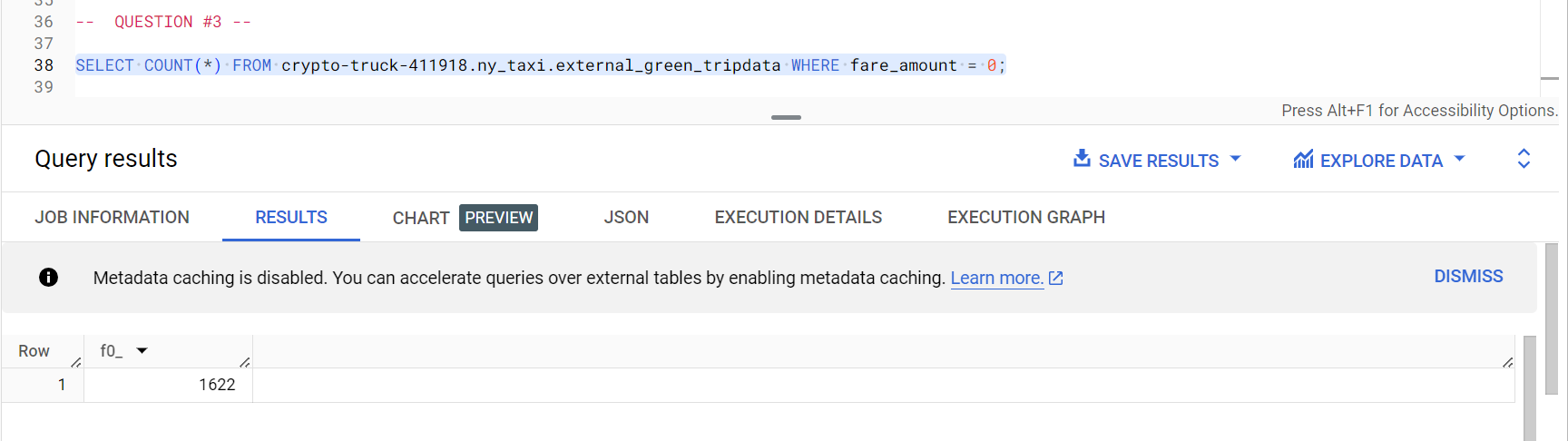
**Question #1**

* What is the count of records for the 2022 Green Taxi Data?
  + **Answer: 840,402**
  + 

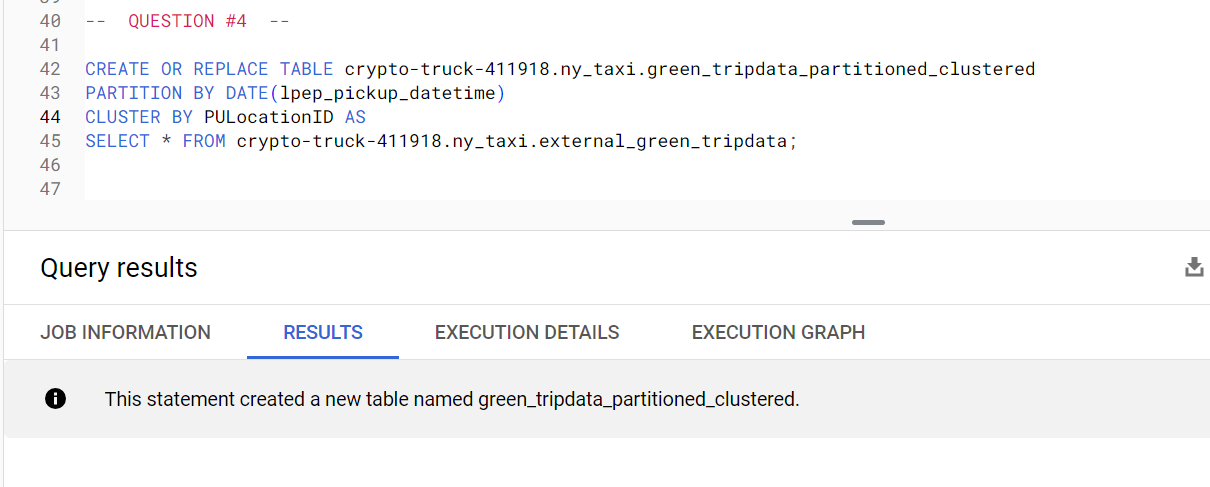
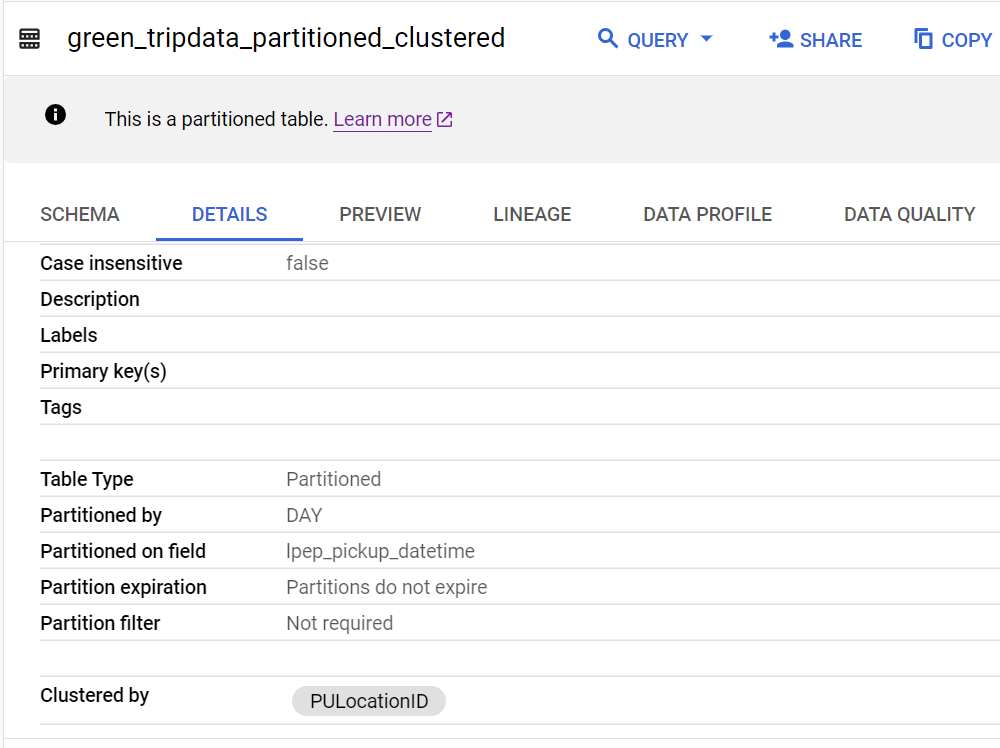
**Question #2**

* 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?
  + **Answer: 0 MB for the External Table and 6.41 MB for the Materialized Table**
  + 
  + 

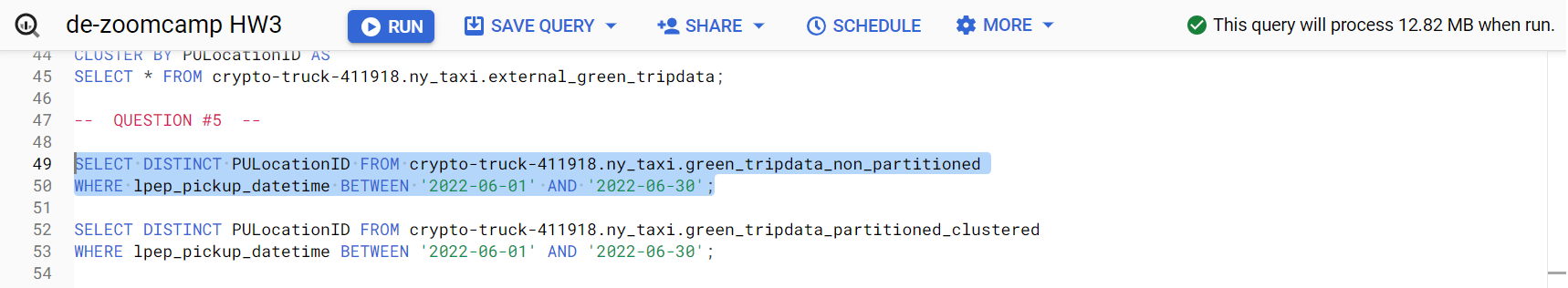
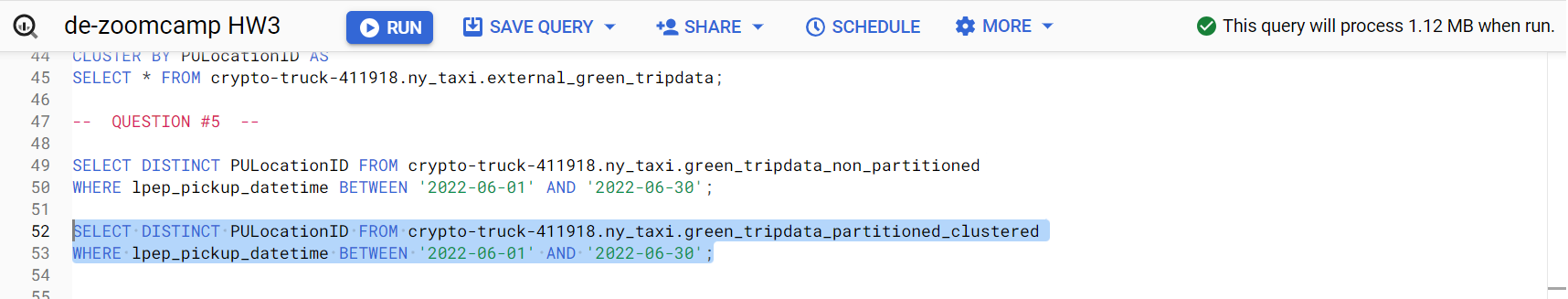
**Question #3**

* How many records have a fare\_amount of 0?
  + **Answer: 1,622**
  + 

**Question #4**

* What is the best strategy to make an optimized table in Big Query if your query will always order the results by PULocationID and filter based on lpep\_pickup\_datetime?
* Create a new table with this strategy.
  + **Answer: Partition by lpep\_pickup\_datetime Cluster on PUlocationID**
    - **Partitioning is mainly performed on an Integer Range, Ingestion Time (\_PARTITIONTIME), or a Time-Unit Column.**
  + 
  + 

**Question #5**

* Write a query to retrieve the distinct PULocationID between lpep\_pickup\_datetime 06/01/20222 and 06/30/2022 (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 4 and note the estimated bytes processed. What are these values?
  + **Answer: 12.82 MB for non-partitioned table and 1.12 MB for the partitioned table**
  + 
  + 

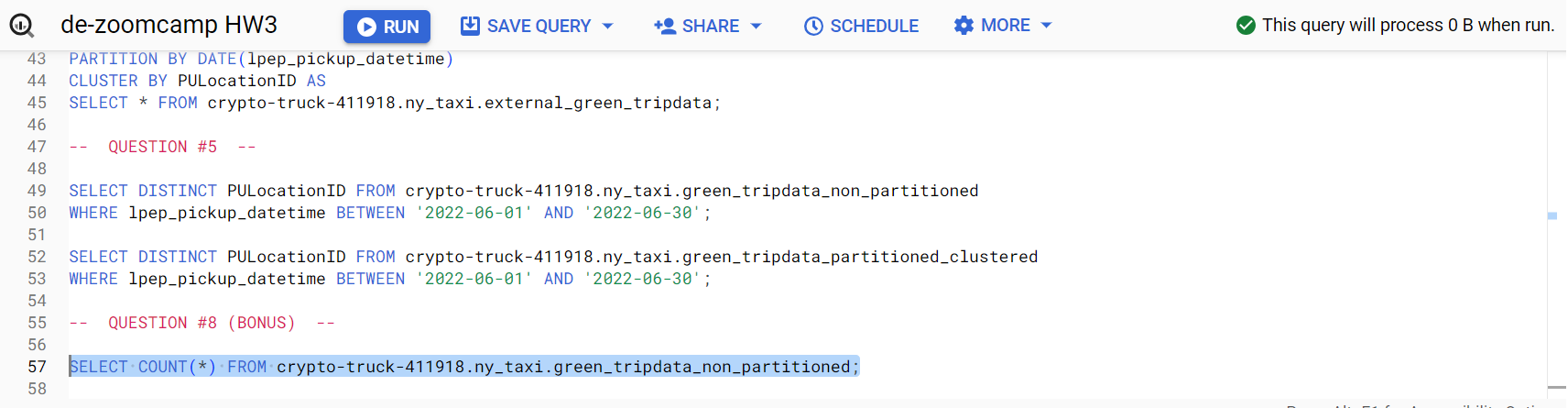
**Question #6**

* Where is the data stored in the External Table you created?
  + **Answer: GCP Bucket**
    - [**External Table Documentation**](https://cloud.google.com/bigquery/docs/external-tables)

**Question #7**

* (True or False) It is best practice in Big Query to always cluster your data.
  + **Answer: True**
    - **Clustering allows for better query optimization and cost reduction.**
    - **Clustering allows for the filtering of multiple columns or partitions.**

**Question #8 (Bonus)**

* Write a SELECT count(\*) query FROM the materialized table you created.
* How many bytes does it estimate will be read? Why?
  + **Answer: 0B**
    - **The results of the query were cached from the previous execution performed in Question #1.**
  + 
  + 