-- loading data from cloud storage bucket to bigquery external table

CREATE OR REPLACE EXTERNAL TABLE `ny-rides-mukesh.trips\_data\_all.external\_yellow\_tripdata`

OPTIONS (

  format = 'CSV',

  uris = ['gs://dtc\_data\_lake\_ny-rides-mukesh/fhv\_tripdata\_2019-\*.csv.gz']

);

-- Question 1: count for fhv vehicle records for year 2019

SELECT count(\*) FROM

    ny-rides-mukesh.trips\_data\_all.external\_yellow\_tripdata;

-- creating BQ table from external table

CREATE OR REPLACE TABLE ny-rides-mukesh.trips\_data\_all.yellow\_tripdata\_non\_partitoned AS

SELECT \* FROM ny-rides-mukesh.trips\_data\_all.external\_yellow\_tripdata;

-- Question 2: count distinct number of affiliated\_base\_number from external and BQ table

SELECT COUNT(DISTINCT Affiliated\_base\_number) FROM

    ny-rides-mukesh.trips\_data\_all.external\_yellow\_tripdata;

SELECT COUNT(DISTINCT Affiliated\_base\_number) FROM

    ny-rides-mukesh.trips\_data\_all.yellow\_tripdata\_non\_partitoned;

--Question 3: records having both null PULocationID and DOLocationID

SELECT

  COUNT(\*) as total\_rows

  FROM ny-rides-mukesh.trips\_data\_all.external\_yellow\_tripdata

WHERE PUlocationID IS NULL AND DOlocationID IS NULL;

--Question 4: Create a partined table on pickup\_datetime and cluster on affiliated\_base\_number\_column

CREATE OR REPLACE TABLE `ny-rides-mukesh.trips\_data\_all.fhv\_partitioned\_tripdata`

PARTITION BY DATE(pickup\_datetime)

CLUSTER BY Affiliated\_base\_number AS (

  SELECT \* FROM `ny-rides-mukesh.trips\_data\_all.yellow\_tripdata\_non\_partitoned`

);

--Question 5: Query to retrieve distinct affiliated base\_number from partitioned and non\_partitioned table for given range

SELECT COUNT(DISTINCT Affiliated\_base\_number) FROM  `ny-rides-mukesh.trips\_data\_all.fhv\_partitioned\_tripdata`

WHERE pickup\_datetime BETWEEN '2019-03-01' AND '2019-03-31';

SELECT COUNT(DISTINCT Affiliated\_base\_number) FROM  `ny-rides-mukesh.trips\_data\_all.yellow\_tripdata\_non\_partitoned`

WHERE pickup\_datetime BETWEEN '2019-03-01' AND '2019-03-31';