1.

Step1 Download All datafiles into a folder

Step2 Upload them to S3

Step3 Enable Bucket Versioning

Step4 Create a database on Athena

Step5 Create a crawler to generate a table and load all the data into it while creating an AMI role.

Step6 Once the crawler is created then run the crawler.

Step7 Once the table (all\_data) is created

Run the below query

create table traffic\_data as

select date("tpep\_pickup\_datetime") as "pickup\_date",

date("tpep\_dropoff\_datetime") as "dropoff\_date",

date\_format("tpep\_pickup\_datetime", '%H:%i:%s') as "dropoff\_time",

date\_format("tpep\_dropoff\_datetime", '%H:%i:%s') as "pickup\_time",

"pulocationid" as "pickup\_location",

"dolocationid" as "dropoff\_location",

"fare\_amount",

"ratecodeid" as "rate\_type",

"payment\_type",

"passenger\_count"

from " all\_data ";

2.

CREATE TABLE partitioned\_data

WITH (

format = 'PARQUET',

partitioned\_by = ARRAY['pickup\_date'],

external\_location = 's3://jackson-hk-bucket-1/results/'

) AS

SELECT "dropoff\_date","pickup\_time" , "dropoff\_time", "pickup\_location","dropoff\_location","fare\_amount", "rate\_type", "payment\_type","passenger\_count", "pickup\_date"

FROM " traffic\_data";

I assumed that the pickup\_date column is of the DATE data type.

3.

SELECT \*,

AVG("daily\_trip\_count") OVER() AS "daily\_avg\_trip\_count",

RANK() OVER (ORDER BY "total\_revenue" DESC) AS "revenue\_rank"

FROM (SELECT

DATE("tpep\_pickup\_datetime") AS "pickup\_date",

COUNT(\*) AS "daily\_trip\_count",

SUM("total\_amount") - SUM("fare\_amount") AS "total\_revenue"

from "traffic\_data"

GROUP BY DATE("tpep\_pickup\_datetime")

)

ORDER BY "total\_revenue" DESC;