Tasks

Data Ingestion Tasks:

Task 1. Create an RDS instance in your AWS account and upload the data to the RDS instance.

Task 2. Use Sqoop command to ingest the data from RDS into the HBase Table.

Task 3. Bulk import data from next two files in the dataset on your EMR cluster to your HBase Table using the relevant codes.

MapReduce Tasks:

Task 4. Write MapReduce codes to perform the tasks using the files you've downloaded on your EMR Instance:

1. Which vendors have the most trips, and what is the total revenue generated by that vendor?

2. Which pickup location generates the most revenue?

3. What are the different payment types used by customers and their count? The final results should be in a sorted format.

4. What is the average trip time for different pickup locations?

5. Calculate the average tips to revenue ratio of the drivers for different pickup locations in sorted format.

6. How does revenue vary over time? Calculate the average trip revenue per month - analysing it by hour of the day (day vs night) and the day of the week (weekday vs weekend).