Tasks:

Data Ingestion Tasks:

Task 1. Create an RDS instance in your AWS account and upload the data to the RDS instance (Note: Instructions on how to work with RDS can be found here.)

Since the dataset is huge, you need to upload the data from only two files (*i.e.* `yellow_tripdata_2017-01.csv` & `yellow_tripdata_2017-02.csv`) from the dataset.

Note: You will need to create an appropriate schema for the data sets to upload them to RDS (you can find the data dictionary in the previous segments. The steps to work with RDS in given in the Additional Resource).

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.

Note: For the above task 3, you just need to import data from the subsequent 2 csv files (*i.e.* `yellow_tripdata_2017-03.csv` & `yellow_tripdata_2017-04.csv`) on your EMR cluster.

MapReduce Tasks:

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

- Which vendors have the most trips, and what is the total revenue generated by that vendor?
- Which pickup location generates the most revenue?

- What are the different payment types used by customers and their count? The final results should be in a sorted format.
- What is the average trip time for different pickup locations?
- Calculate the average tips to revenue ratio of the drivers for different pickup locations in sorted format.
- 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).

NOTE: It's recommended to use MRJob for completing the MapReduce taks above.

Optional Task:

Task 5. Use Sqoop export command to export the results of each MapReduce tasks above to your RDS instance. Use the RDS connection string connection to visualise the dataset using a dashboarding tool (Google Data Studio, Tableau or PowerBI) ***(Optional)***

NOTE: Please note that Task 5 is optional and purely to demonstrate how RDS and Sqoop.

NOTE: The Data Ingestion tasks and the MapReduce tasks are separate. The MapReduce tasks must be run with the local data downloaded to the cluster.