Hive Exercise (part 1): Solutions

Exercise 1

```
product_id,product_name,category,quantity,price_per_unit
1,Product A,Electronics,10,50.00
2,Product B,Clothing,15,25.00
3,Product C,Electronics,5,100.00
4,Product D,Clothing,8,30.00
```

Create a database called sales

`CREATE DATABASE sales;`

Use the created Database

`use sales;`

Create a Hive table

```
CREATE TABLE sales_data (
  product_id INT,
  product_name STRING,
  category STRING,
  quantity INT,
  price_per_unit DOUBLE
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

Load Data

`LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/hive/sales.csv' INTO TABLE sales_data;`

`SELECT * FROM sales_data;`

```
CREATE TABLE category_total_revenue (
  category STRING,
  total_revenue DOUBLE
)
STORED AS ORC;
```

Calculate total revenue for each category and insert into the new table

```
INSERT OVERWRITE TABLE category_total_revenue
SELECT category, SUM(quantity * price_per_unit) AS total_revenue
FROM sales_data
GROUP BY category;
```

This process will generate a new file in HDFS, specifically in the `/user/hive/category_total_revenue` directory, containing the aggregated total revenue data for each category. It's essential to verify that the HDFS directory path is accurate and that the Hive user possesses the necessary write permissions.

Get output locally

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/Desktop/csv_output/'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
SELECT * FROM category_total_revenue;
```

Exercise 2

orders.csv

```
order_id,customer_id,order_date,total_amount
1,101,2023-01-15,150.00
2,102,2023-01-20,200.00
3,103,2023-02-05,75.00
4,104,2023-02-10,300.00
5,105,2023-03-01,120.00
```

customers.csv

```
customer_id,customer_name,city
101,Alice,New York
102,Bob,Los Angeles
104,David,Chicago
```

Create a database called sales

`CREATE DATABASE customers;`

Use the created Database

`use customers;`

Create tables

```
CREATE TABLE orders (
  order_id INT,
  customer_id INT,
  order_date STRING,
  total_amount DOUBLE
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

CREATE TABLE customers (
  customer_id INT,
  customer_name STRING,
  city STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

Load data into tables

```
LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/hive/orders.csv' INTO TABLE orders;
LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/hive/customers.csv' INTO TABLE customers;
```

Query to join tables and handle null values

```
SELECT o.order_id, o.order_date, o.total_amount, c.customer_name, c.city
FROM orders o
LEFT OUTER JOIN customers c ON o.customer_id = c.customer_id;
```

Export the result to a local directory

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/Desktop/csv_output/'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
SELECT * FROM (
  SELECT o.order_id, o.order_date, o.total_amount, c.customer_name, c.city
  FROM orders o
  LEFT OUTER JOIN customers c ON o.customer_id = c.customer_id
) result;
```