# **Big Data : Homework1**

# **Sindhuja Pulyala**

1. Assignments directory creation:

A screen shot of a computer screen

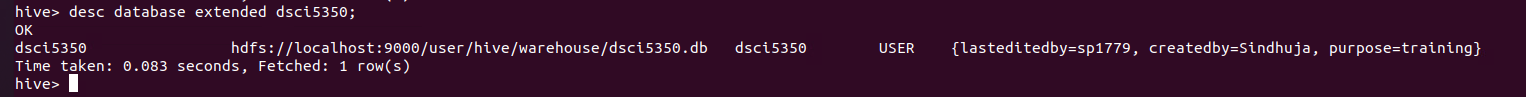
Description automatically generated

After putting documents in assignments directory

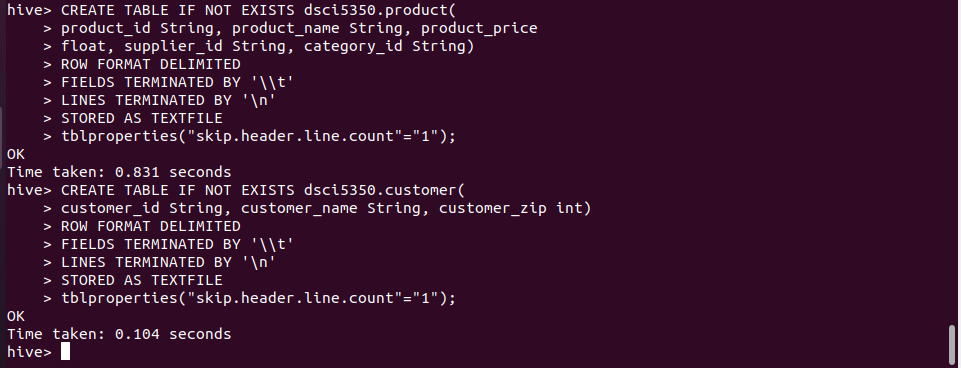
A screenshot of a computer screen

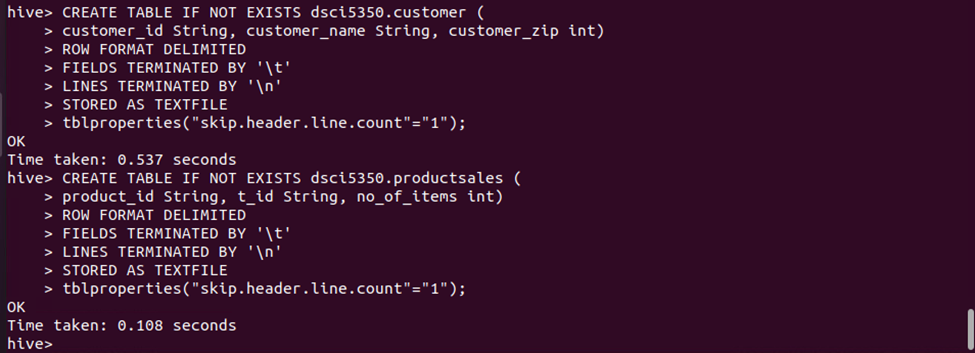
Description automatically generated

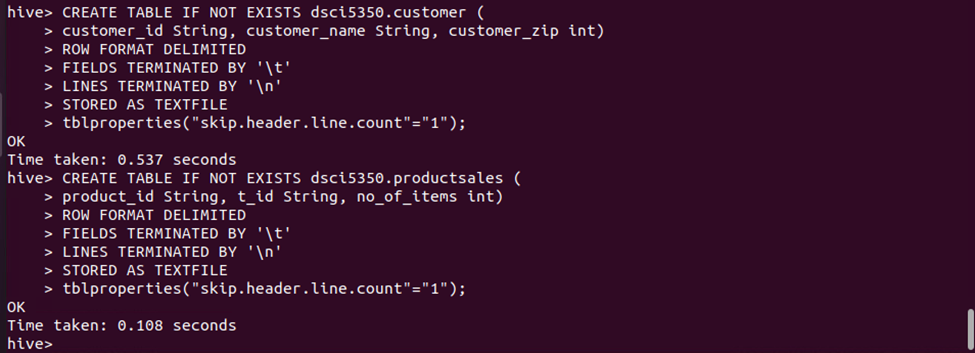
1. Describe Database:

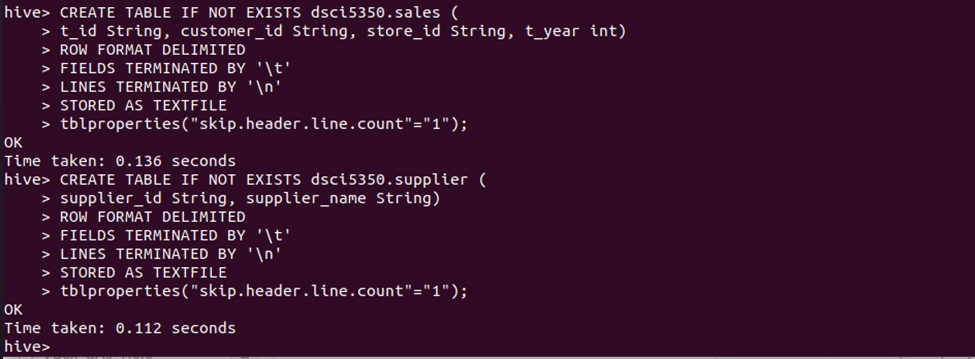


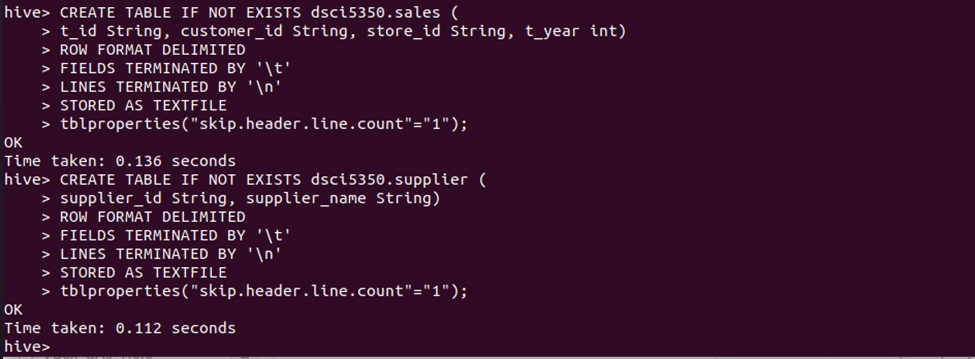
Create tables:







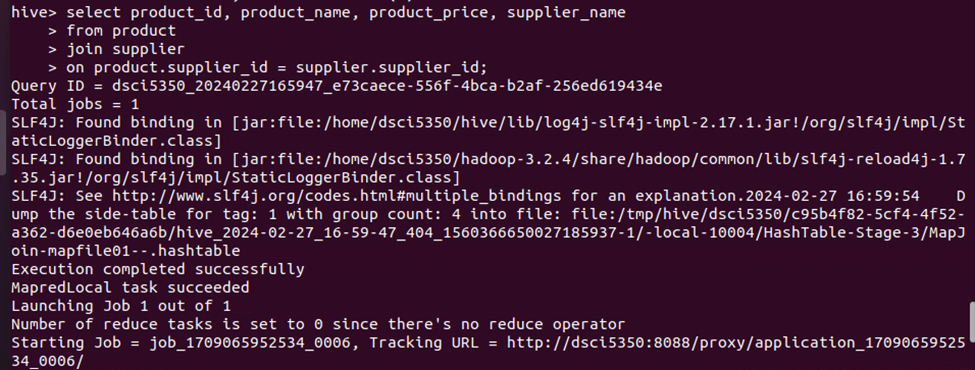




A computer screen shot of a computer

Description automatically generated

1. Makes Yourself familiar with the product and supplier lists by making a table with product id, product name, product price, and supplier name for all products (Take a screenshot of your query and the results



A screenshot of a computer

Description automatically generated

1. Write a query that can find and rank customers sales showing the most number of sales on the top. Make sure that you display the Customer name and ID (Take a screenshot of your results)

A computer screen shot of white text

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Write a query showing the total number of sales and the payments received from sales (based on product\_price and number of sales) for each product. Be sure to display the Product ID, product name, supplier, as well as total number of sales and payments received ranked by total payments received.

A computer screen with text

Description automatically generated

A screenshot of a computer

Description automatically generated

**Q1. Were any MapReduce jobs involved in HW1.**

Yes, there are map reduce jobs involved in HW1 with number of mappers-1 and number of reducers -1.

**Q2. Though Hadoop/Hive were not necessarily needed for this task, assume you are an analyst on the job and write a justification.**

The volume of data sets being collected and analyzed in the companies for business intelligence is growing quickly, making traditional warehousing solutions prohibitively expensive and as the volume of the data increases, traditional databases may struggle to handle the processing demands properly. The Hadoop ecosystem, particularly tools like Hive [1], can be highly useful for analyzing structured data in different kinds of scenarios. One primary reason is scalability. Hadoop's distributed file system model allows for seamless scaling across clusters of commodities, qualify organizations to analyze huge amounts of structured data effectively.

Another main advantage of the Hadoop ecosystem, including Hive, lies in its Affordability By utilizing commodity hardware and open-source software [2], companies can decrease the infrastructure costs associated with data storage and analysis. Hadoop's ability to store and process data across nodes also increases fault tolerance and reliability, ensuring that analysis tasks can continue uninterrupted even in the event of hardware failures or network issues. This cost-effectiveness makes Hadoop a great solution for businesses looking to derive insights from their structured data without facing excessive expenses.

# References

|  |  |
| --- | --- |
| [1] | T. White, Hadoop: The Definitive Guide, O'Reilly Media, Inc., 2012. |
| [2] | Thusoo, Ashish and Sarma, J. Sen and N. Jain, "Hive: a warehousing solution over a map-reduce framework," *Proceedings of the VLDB Endowment,* vol. 2, no. VLDB Endowment, 2009. |