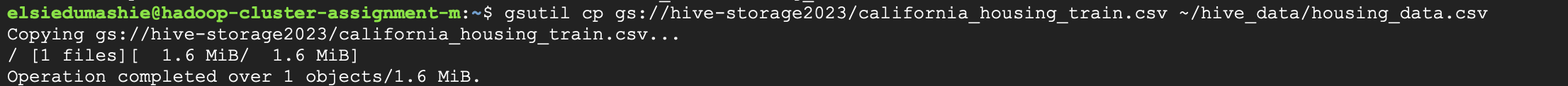
Hive assignment

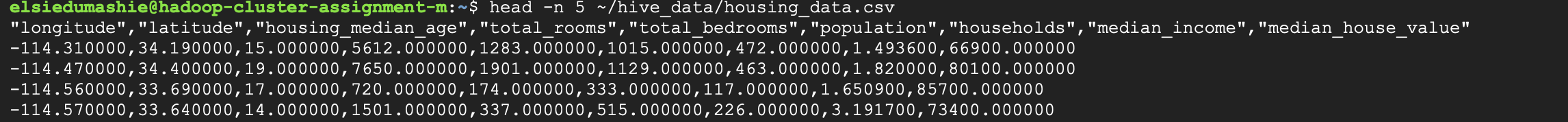
1.

* 1. **Data Preparation: ( 5 Marks)** Download the dataset "california\_housing\_train.csv."
  2. Import the dataset into your Hive environment.
  3. Define an appropriate table schema in Hive to match the structure of the dataset. Ensure that data types and column names align correctly.

A screen shot of a computer

Description automatically generated





A screenshot of a computer

Description automatically generated

**2.Query One: (5 Marks)**

Write a query to find the total number of households (**households**) and the total population (**population**) for areas where the median house value (**median\_house\_value**) exceeds $500,000.

Provide the results as two separate columns and create a hive table with the n\_a\_m\_e\_ \_o\_f\_ \_“h\_h\_\_\_p\_o\_p\_” \_in the same database

A screen shot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated



A screen shot of a computer

Description automatically generated



* 1. 3. Data Extraction: (5 Marks) After executing each query successfully, extract the results into CSV files.
  2. Name the CSV file "Query\_Results.csv,"
  3. A screenshot of a computer program

     Description automatically generated

A screenshot of a computer

Description automatically generated