# Week 3 Assignment: Hands-on Hive

### Objective: Gain hands-on experience with Hive by creating tables, importing data, and executing SQL queries.

#### **1. Environment Initialization**

* Navigate to the required directory and start your Docker containers:
* cd bellevue-bigdata  
  cd hadoop-hive-spark-hbase  
  docker-compose up -d
* Access the master container:
* docker-compose exec master bash
* Load the grades.csv into HDFS:
* hadoop fs -put /data/grades.csv /

#### **2. Hive Table Creation & Data Loading**

* Start a Hive session:
* hive
* In the Hive CLI, create a table:

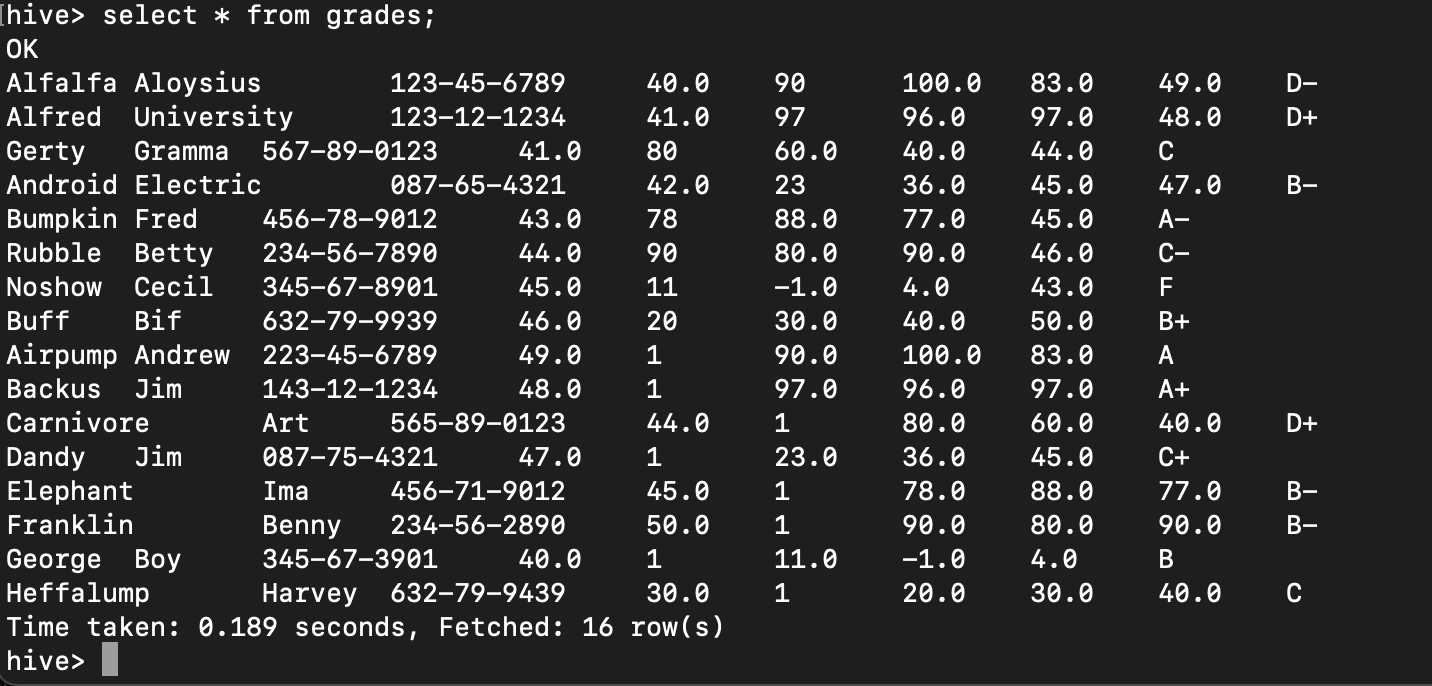
|  |
| --- |
| CREATE TABLE grades(  `Last name` STRING,  `First name` STRING,  `SSN` STRING,  `Test1` DOUBLE,  `Test2` INT,  `Test3` DOUBLE,  `Test4` DOUBLE,  `Final` DOUBLE,  `Grade` STRING)  ROW FORMAT DELIMITED  FIELDS TERMINATED BY ','  STORED AS TEXTFILE  tblproperties("skip.header.line.count"="1"); |

* Load data into the Hive table:

|  |
| --- |
| LOAD DATA INPATH '/grades.csv' INTO TABLE grades; |

* Run a query to view the data:

|  |
| --- |
| SELECT \* FROM grades; |

****

**Deliverable:** Screenshot of the query results.

#### **3. Further Exploration with SQL on Hive**

* Run 3 different SQL commands on the grades data.

First Name and Last Name of each student



Padded full table query

A screenshot of a computer

Description automatically generated

Top 5 students with highest final grade

A screenshot of a computer

Description automatically generated

Average test grade:

A screenshot of a computer

Description automatically generated

**Deliverable:** Screenshots of each command and its output.

#### **4. Working with a New Dataset**

* Choose a CSV dataset of your choice from an open data set. Some data sources to consider:
  1. [Kaggle Datasets](https://www.kaggle.com/datasets)
  2. [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/index.php)
  3. [Government Datasets](https://www.data.gov/)
  4. [AWS Public Datasets](https://registry.opendata.aws/)
* After downloading your dataset, copy it to your Docker container:
* docker compose cp /path/to/your/csv/file master:/data/
* Load the data into HDFS:
* hadoop fs -put /data/your\_csv\_file.csv /
* Return to the Hive CLI and create a table for your dataset, then load the data. Remember, you can choose to define each column as a STRING for simplicity, or use the appropriate data types for a greater challenge. Visit the [Hive Language Manual](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types) to learn about different data types available in Hive.
* Run 3 SQL commands on your new data.

**Deliverable:** Screenshots of each command and its output. Provide a brief explanation of why you chose the particular dataset and what insights you hoped to gain from your queries.

## Shutting Down

Ensure all Docker containers are turned off with docker-compose down for each directory. If you’re using google cloud, please shut down your virtual machine to preserve cloud costs.