

Getting Started with Hive



Estimated time needed: 20 minutes

In this lab you will explore Apache Hive, a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale. You will be creating a table and running SQL commands on it.

Learning Objectives

At the end of this lab, you will be able to:

- Create a table in Hive
- Add data to the table using file
- Add data to the table using insert
- Query the data in the table using SQL commands

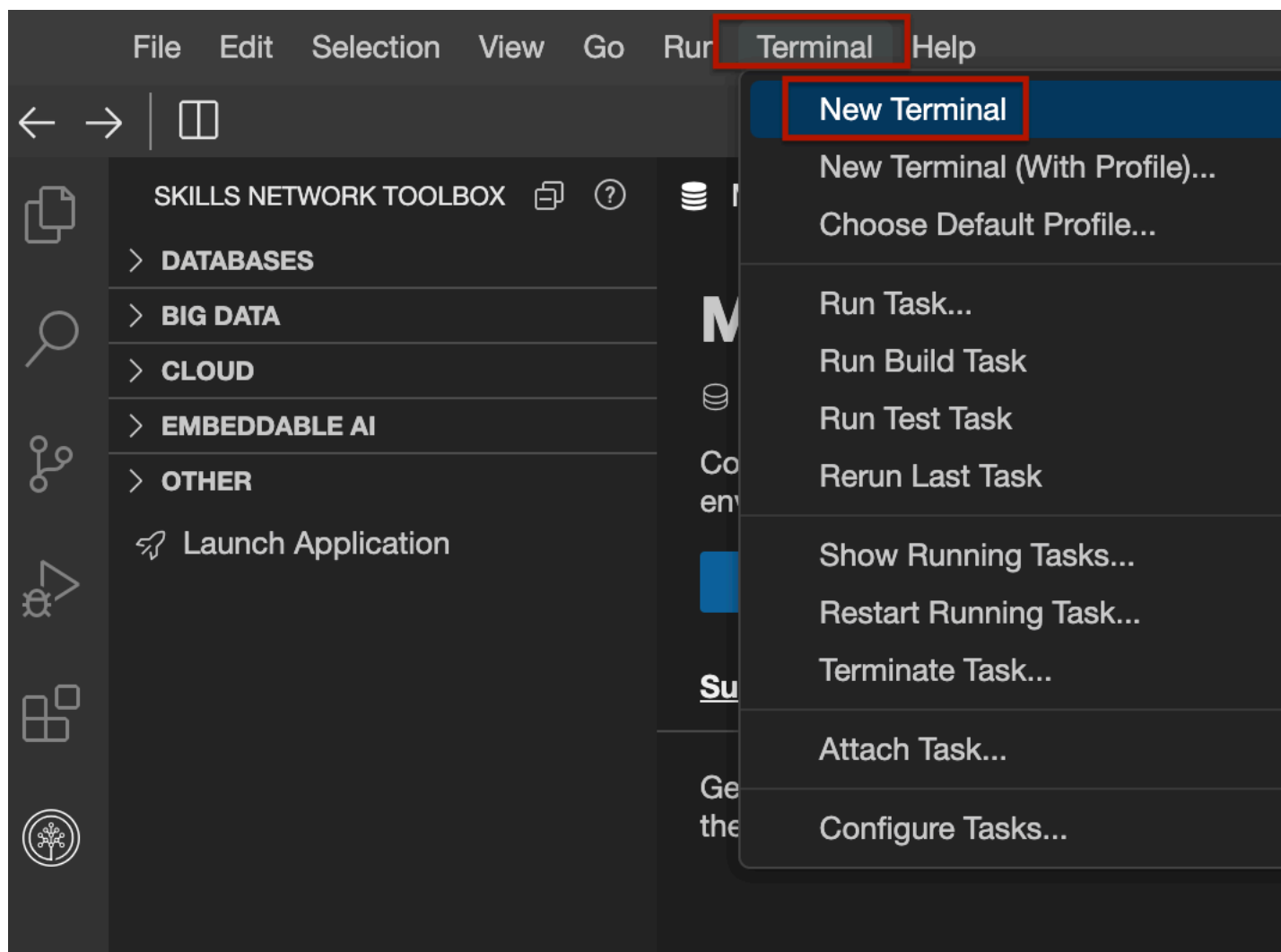
Prerequisites

- You should comfortable working with the Linux terminal
- Prior knowledge of SQL will be helpful

While all the terminal commands can be copy pasted and run, it is highly recommended for you to type the commands for better learning.

Step 1: Get a copy of the CSV file

1. You will run the commands in the terminal. If you don't have a terminal open, open a new terminal, by clicking on Terminal and choosing New Terminal from the submenu.



2. Create a directory named data under /home/project by running the following command.

```
mkdir /home/project/data
```

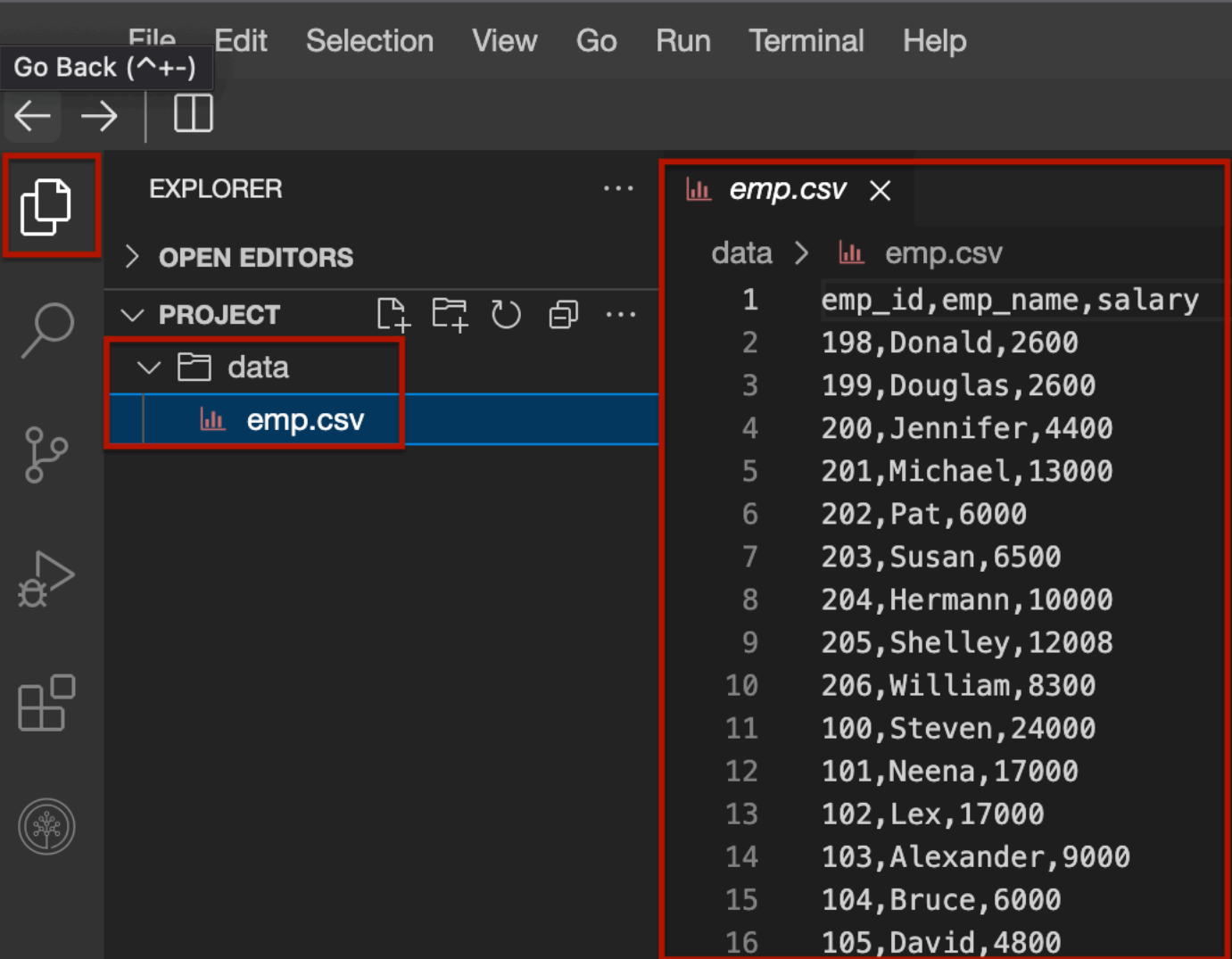
3. Change to the /home/project/data directory.

```
cd /home/project/data
```

4. Run the following command to get the emp.csv, a data file with Employee data, in a comma-separated file which you will use later to infuse data into the table you create.

```
wget https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-BD0225EN-SkillsNetwork/data/emp.csv
```

5. Open the file in editor and view the file.



The screenshot shows the Visual Studio Code (VS Code) interface. The Explorer sidebar on the left is open, showing the 'data' folder expanded and the 'emp.csv' file selected. The main editor area displays the contents of 'emp.csv', which is a comma-separated file containing employee data. The data is as follows:

emp_id	emp_name	salary
198	Donald	2600
199	Douglas	2600
200	Jennifer	4400
201	Michael	13000
202	Pat	6000
203	Susan	6500
204	Hermann	10000
205	Shelley	12008
206	William	8300
100	Steven	24000
101	Neena	17000
102	Lex	17000
103	Alexander	9000
104	Bruce	6000
105	David	4800

Step 2: Setup Hive and Bee

1. You will use the hive from the docker hub for this lab. Pull the hive image into your system by running the following command.

```
docker pull apache/hive:4.0.0-alpha-1
```

This will take a few seconds, depending on the speed of your internet connection.

- Now, you will run the hive server on port 10002. You will name the server instance myhiveserver. We will mount the local data folder in the hive server as hive_custom_data. This would mean that the whole data folder that you created locally, along with anything you add in the data folder, is copied into the container under the directory hive_custom_data.

```
docker run -d -p 10000:10000 -p 10002:10002 --env SERVICE_NAME=hiveserver2 -v /home/project/data:/hive_custom_data --name myhiveserv
```

- You can open and take a look at the Hive server with the GUI. Click the button to open the HiveServer2 GUI.

HiveServer2 GUI

- Now run the following command, which allows you to access beeline. This is a SQL cli where you can create, modify, delete table, and access data in the table.

```
docker exec -it myhiveserver beeline -u 'jdbc:hive2://localhost:10000/'
```

Step 3: Create table, add and view data

- To create a new table Employee with three columns as in the csv you downloaded - emp_id, emp_name and salary, run the following command.

```
create table Employee(emp_id string, emp_name string, salary int) row format delimited fields terminated by ',' ;
```

You may notice that there is an explicit mention for the fields delimited by , just as in the csv file.

- Run the following command to check if the table is created.

```
show tables;
```

This should list the Employee table that you just created.

- Now load the data into the table from the csv file by running the following command.

```
LOAD DATA INPATH '/hive_custom_data/emp.csv' INTO TABLE Employee;
```

3. Run the following command to list all the rows from the table to check if the data has been loaded from the CSV.

```
SELECT * FROM employee;
```

4. You can view the details of the commands and the outcome in the HiveServer2 GUI.

HiveServer2 GUI

5. To quit from the beehive prompt in the terminal, press `ctrl+D`.

Hive internally uses MapReduce to process and analyze data. When you execute a Hive query, it generates MapReduce jobs that run on the Hadoop cluster.

Conclusion

In this lab you created a table in hive, added data to the table from csv and listed the data contained in the table.

Next Steps

You can explore more SQL commands with table and see how it works.

Author(s)

Lavanya T S

© IBM Corporation. All rights reserved.