

Hands-on Lab: ETL using shell scripts



Estimated time needed: **30** minutes

Objectives

After completing this lab you will be able to:

- Extract data from a delimited file.
- Transform text data.
- Load data into a database using shell commands.

About Skills Network Cloud IDE

Skills Network Cloud IDE (based on Theia and Docker) provides an environment for hands on labs for course and project related labs. Theia is an open-source IDE (Integrated Development Environment), that can be run on desktop or on the cloud. to complete this lab, we will be using the Cloud IDE based on Theia and Postgres running in a Docker container.

Important Notice about this lab environment

Please be aware that sessions for this lab environment are not persisted. Every time you connect to this lab, a new environment is created for you. Any data you may have saved in the earlier session would get lost. Plan to complete these labs in a single session, to avoid losing your data.

Getting the environment ready

Open a new terminal, by clicking on the menu bar and selecting **Terminal->New Terminal**, as in the image below. This will open a new terminal at the bottom of the screen.

The screenshot shows the Theia IDE interface. At the top is a menu bar with 'File', 'Edit', 'Selection', 'View', 'Go', 'Run', 'Terminal', and 'Help'. The 'Terminal' menu is open, displaying options: 'New Terminal' (highlighted in blue), 'Split Terminal', 'Run Task...', 'Run Build Task', 'Run Test Task', 'Rerun Last Task', 'Show Running Task', 'Restart Running Task', 'Terminate Task...', 'Attach Task...', and 'Configure Tasks...'. On the left is a sidebar with icons for file explorer, search, source control, and a list of databases: 'SKILLS NETWORK...', 'DATABASES', 'MySQL INACTIVE', 'PostgreSQL IDLE' (highlighted), 'Cassandra INACTIVE', and 'MongoDB INACTIVE'. Below these are sections for 'BIG DATA', 'CLOUD', 'EMBEDDABLE AI', and 'OTHER', followed by a 'Launch Application' button. The main area on the right shows a terminal window with the prompt 'theia@theiadosker-lavanyas: /home/p' and a green prompt 'theia@theiadosker-lavanyas: /h'.

Run all the commands on the newly opened terminal. (You can copy the code by clicking on the little copy button on the bottom right of the codeblock below and then paste it, wherever you wish.)

Exercise 1 - Extracting data using 'cut' command

The filter command `cut` helps us extract selected characters or fields from a line of text.

1. Extracting characters.

The command below shows how to extract the first four characters.

```
1. 1
1. echo "database" | cut -c1-4
```

Copied! Executed!

You should get the string 'data' as output.

The command below shows how to extract 5th to 8th characters.

```
1. 1
1. echo "database" | cut -c5-8
```

Copied! Executed!

You should get the string 'base' as output.

Non-contiguous characters can be extracted using the comma.

The command below shows how to extract the 1st and 5th characters.

```
1. 1
1. echo "database" | cut -c1,5
```

Copied! Executed!

You get the output : 'db'

2. Extracting fields/columns

We can extract a specific column/field from a delimited text file, by mentioning

- the delimiter using the -d option, or
- the field number using the -f option.

The /etc/passwd is a “:” delimited file.

The command below extracts usernames (the first field) from /etc/passwd.

```
1. 1
1. cut -d":" -f1 /etc/passwd
```

Copied! Executed!

The command below extracts multiple fields 1st, 3rd, and 6th (username, userid, and home directory) from /etc/passwd.

```
1. 1
1. cut -d":" -f1,3,6 /etc/passwd
```

Copied! Executed!

The command below extracts a range of fields 3rd to 6th (userid, groupid, user description and home directory) from /etc/passwd.

```
1. 1
1. cut -d":" -f3-6 /etc/passwd
```

Copied! Executed!

Exercise 2 - Transforming data using 'tr'

tr is a filter command used to translate, squeeze, and/or delete characters.

1. Translate from one character set to another

The command below translates all lower case alphabets to upper case.

```
1. 1
1. echo "Shell Scripting" | tr "[a-z]" "[A-Z]"
```

Copied! Executed!

You could also use the pre-defined character sets also for this purpose:

```
1. 1
1. echo "Shell Scripting" | tr "[:lower:]" "[:upper:]"
```

Copied! Executed!

The command below translates all upper case alphabets to lower case.

- 1
1. `echo "Shell Scripting" | tr "[A-Z]" "[a-z]"`

Copied! Executed!

2. Squeeze repeating occurrences of characters

The `-s` option replaces a sequence of a repeated characters with a single occurrence of that character.

The command below replaces repeat occurrences of 'space' in the output of `ps` command with one 'space'.

- 1
1. `ps | tr -s " "`

Copied! Executed!

In the above example, the space character within quotes can be replaced with the following : `"[\:space\:]"`.

3. Delete characters

We can delete specified characters using the `-d` option.

The command below deletes all digits.

- 1
1. `echo "My login pin is 5634" | tr -d "[:digit:]"`

Copied! Executed!

The output will be : 'My login pin is'

Exercise 3 - Start the PostgreSQL database.

1. From the SkillsNetwork tools, under Databases choose PostgreSQL Database server and click Start to start the server. This will take a few mins.

The screenshot shows the SkillsNetwork Labs interface. On the left, a sidebar lists various tools under the 'Databases' category. 'PostgreSQL INACTIVE' is highlighted with a red box. Below it, other databases like MySQL, Cassandra, and MongoDB are listed as 'INACTIVE'. Further down, there are sections for 'BIG DATA', 'CLOUD', 'EMBEDDABLE AI', and 'OTHER'. At the bottom of the sidebar, there is a 'Launch Application' button. The main panel displays the 'PostgreSQL' status as 'INACTIVE'. It shows versions v13.2, v5.0, and v14.5. A description states: 'Connect to PostgreSQL and pgAdmin directly in your Skills Network Labs environment.' A prominent blue 'Start' button is highlighted with a red box. Below this, there are tabs for 'Summary', 'Connection Information', and 'Details'. The 'Summary' tab is active, showing instructions: 'Get started with PostgreSQL in a faster, easier way. To launch your database, hit the Start button.'

2. Click PostgreSQL CLI on the screen to start interacting with the PostgreSQL server.

The screenshot shows the Skills Network Tools interface. On the left, a sidebar lists various tools under categories: DATABASES, BIG DATA, CLOUD, EMBEDDABLE AI, and OTHER. PostgreSQL is listed under DATABASES and is marked as 'ACTIVE'. Other tools like MySQL, Cassandra, MongoDB, Apache Airflow, Apache Hive, and Apache Spark are marked as 'INACTIVE' or 'COMING SOON'. A red box highlights the PostgreSQL icon in the sidebar. The main panel displays the PostgreSQL configuration page. It shows the version as v13.2, v5.0, and v14.5. A green 'ACTIVE' button is present. Below this, there is a 'Stop' button. The page has three tabs: 'Summary', 'Connection Information', and 'Details'. The 'Summary' tab is selected, showing a message that the database and pgAdmin server are ready to use. It provides login credentials: Username: captainfedo1 and Password: Mjk2MDktY2FwdGFp. The password field is highlighted with a red box. Below the credentials, there are buttons for 'pgAdmin' and 'PostgreSQL CLI'. The 'PostgreSQL CLI' button is also highlighted with a red box. There is also a 'New Terminal' button.

This will start the interactive psql client which connects to the PostgreSQL server with postgres=# prompt as shown below.

```
theia@theiadocker-lavanyas:/home/project$ psql --username=postgres --host=localhost
psql (15.2 (Ubuntu 15.2-1.pgdg18.04+1), server 13.2)
Type "help" for help.

postgres=#
```

Exercise 4 - Create a table

In this exercise we will create a table called users in the PostgreSQL database using PostgreSQL CLI. This table will hold the user account information.

The table users will have the following columns:

1. uname
2. uid

3. home

4. You will connect to `template1` database which is already available by default. To connect to this database, run the following command at the `'postgres=#'` prompt.

```
1. 1
1. \c template1
```

Copied!

You will get the following message.

You are now connected to database "template1" as user "postgres".

Also, your prompt will change to `'template1=#'`.

2. Run the following statement at the `'template1=#'` prompt to create the table.

```
1. 1
1. create table users(username varchar(50),userid int,homedirectory varchar(100));
```

Copied!

If the table is created successfully, you will get the message below.

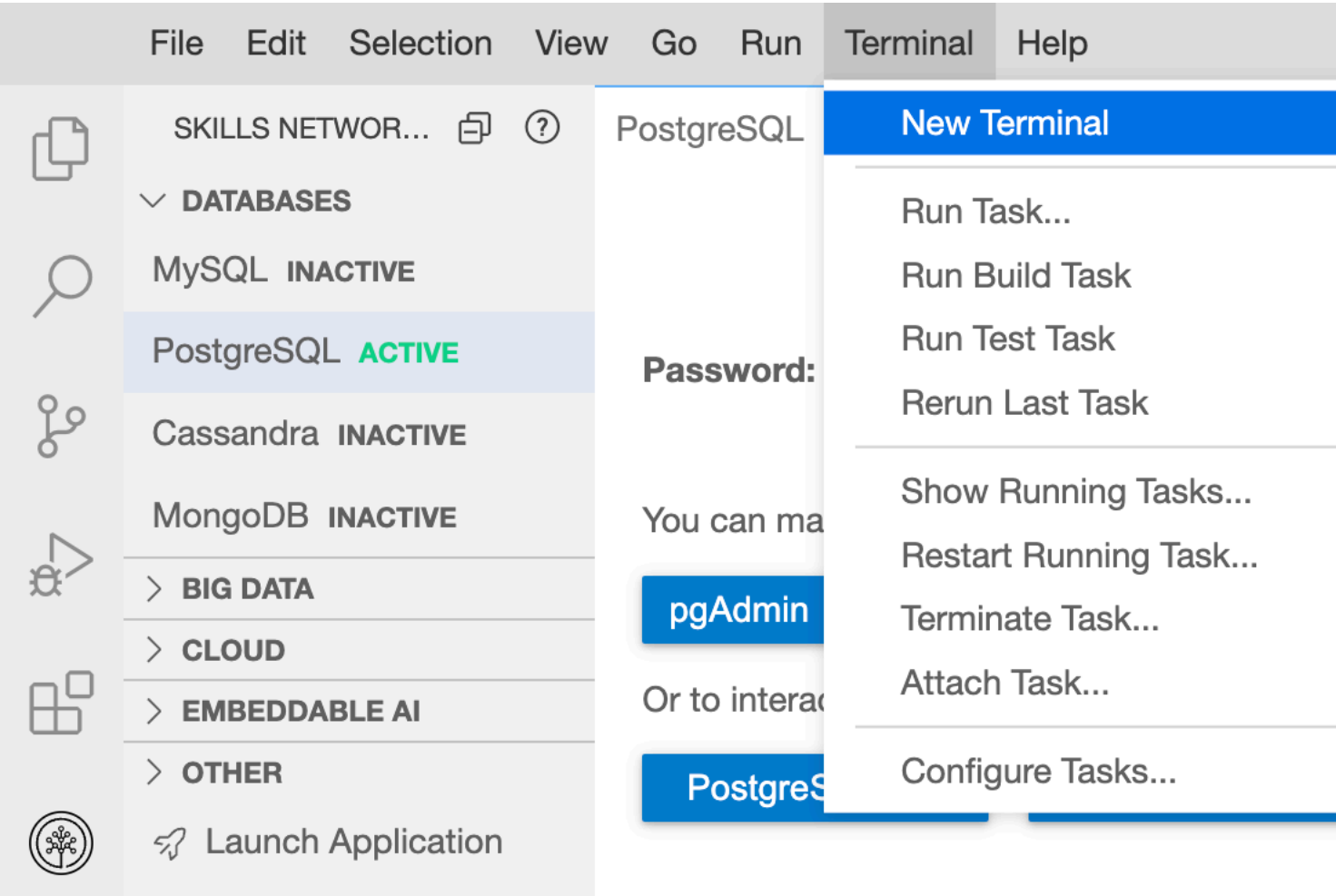
CREATE TABLE

Exercise 5 - Loading data into a PostgreSQL table.

In this exercise, you will create a shell script which does the following.

- Extract the user name, user id, and home directory path of each user account defined in the `/etc/passwd` file.
- Save the data into a comma separated (CSV) format.
- Load the data in the csv file into a table in PostgreSQL database.

1. Open a new Terminal.



2. In the terminal, run the following command to create a new shell script named `csv2db.sh`.

```
1. 1
```

```
1. touch csv2db.sh
```

Copied! Executed!

3. Open the file in the editor. Copy and paste the following lines into the newly created file.

Open **csv2db.sh** in IDE

```
1. 1
2. 2
3. 3
4. 4
5. 5
6. 6
7. 7
8. 8

1. # This script
2. # Extracts data from /etc/passwd file into a CSV file.
3.
4. # The csv data file contains the user name, user id and
5. # home directory of each user account defined in /etc/passwd
6.
7. # Transforms the text delimiter from ":" to ",".
8. # Loads the data from the CSV file into a table in PostgreSQL database.
```

Copied!

4. Save the file by pressing Ctrl+s or by using the **File->Save** menu option.

5. You need to add lines of code to the script that will extract user name (field 1), user id (field 3), and home directory path (field 6) from /etc/passwd file using the cut command.

Copy the following lines and paste them to the end of the script and save the file.

```
1. 1
2. 2
3. 3
4. 4
5. 5
6. 6
7. 7
8. 8

1. # Extract phase
2.
3. echo "Extracting data"
4.
5. # Extract the columns 1 (user name), 2 (user id) and
6. # 6 (home directory path) from /etc/passwd
7.
8. cut -d":" -f1,3,6 /etc/passwd
```

Copied!

6. Run the script.

```
1. 1

1. bash csv2db.sh
```

Copied! Executed!

7. Verify that the output contains the three fields, that you extracted.

8. Change the script to redirect the extracted data into a file named extracted-data.txt

Replace the cut command at end of the script with the following command.

```
1. 1

1. cut -d":" -f1,3,6 /etc/passwd > extracted-data.txt
```

Copied!

9. Run the script.

```
1. 1

1. bash csv2db.sh
```

Copied! Executed!

10. Run the command below to verify that the file extracted-data.txt is created, and has the content.

```
1. 1

1. cat extracted-data.txt
```

Copied! Executed!

- The extracted columns are separated by the original ":" delimiter. You need to convert this into a "," delimited file. Add the below lines at the end of the script and save the file.

```
1. 1
2. 2
3. 3
4. 4
5. 5

1. # Transform phase
2. echo "Transforming data"
3. # read the extracted data and replace the colons with commas.
4.
5. tr ":" " ," < extracted-data.txt > transformed-data.csv
```

Copied!

- Run the script.

```
1. 1

1. bash csv2db.sh
```

Copied! Executed!

- Run the command below to verify that the file transformed-data.csv is created, and has the content.

```
1. 1

1. cat transformed-data.csv
```

Copied!

- To load data from a shell script, you will use the psql client utility in a non-interactive manner. This is done by sending the database commands through a command pipeline to psql with the help of echo command.

PostgreSQL command to copy data from a CSV file to a table is COPY.

The basic structure of the command which we will use in our script is,

```
COPY table_name FROM 'filename' DELIMITERS 'delimiter_character' FORMAT;
```

Now, add the lines below to the end of the script 'csv2db.sh' and save the file.

```
1. 1
2. 2
3. 3
4. 4
5. 5
6. 6

1. # Load phase
2. echo "Loading data"
3. # Send the instructions to connect to 'template1' and
4. # copy the file to the table 'users' through command pipeline.
5.
6. echo "\c template1;\COPY users FROM '/home/project/transformed-data.csv' DELIMITERS ',' CSV;" | psql --username=postgres --host=localhost
```

Copied!

Exercise 6 - Execute the final script

- Run the script.

```
1. 1

1. bash csv2db.sh
```

Copied!

- Go to Postgres SQL CLI that you used in the Run the command below to verify that the table users is populated with the data.

```
1. 1

1. echo "\c template1; \SELECT * from users;" | psql --username=postgres --host=localhost
```

Copied!

Congratulations! You have created an ETL script using shell scripting.

Practice exercises

- Copy the data in the file 'web-server-access-log.txt.gz' to the table 'access_log' in the PostgreSQL database 'template1'.

The file is available at the location : <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Bash%20Scripting/ETL%20using%20shell%20scripting/web-server-access-log.txt.gz>

The following are the columns and their data types in the file:

- a. timestamp - TIMESTAMP
- b. latitude - float
- c. longitude - float
- d. visitorid - char(37)
- e. accessed_from_mobile - boolean
- f. browser_code - int

The columns which we need to copy to the table are the first four columns : timestamp, latitude, longitude and visitorid.

NOTE: The file comes with a header. So use the 'HEADER' option in the 'COPY' command.

The problem may be solved by completing the following tasks:

1. Go to the SkillsNetwork Tools menu and start the Postgres SQL server if it is not already running.
2. Create a table named access_log to store the timestamp, latitude, longitude and visitorid.

- [Click here for Hint](#)
- [Click here for Solution](#)

Task 3. Create a shell script named cp-access-log.sh and add commands to complete the remaining tasks to extract and copy the data to the database.

Create a shell script to add commands to complete the rest of the tasks.

- [Click here for Hint](#)

Task 4. Download the access log file.

Add the wget command to the script to download the file.

1. 1
1. wget "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Bash%20Scripting/ETL%20using%20shell%20scrip

Copied!

Task 5. Unzip the gzip file.

Add the code, to run the gunzip command to unzip the .gz file and extract the .txt file, to the script.

1. 1
2. 2
1. # Unzip the file to extract the .txt file.
2. gunzip -f web-server-access-log.txt.gz

Copied!

The -f option of gunzip is to overwrite the file if it already exists.

Task 6. Extract required fields from the file.

Extract timestamp, latitude, longitude and visitorid which are the first four fields from the file using the cut command.

The columns in the web-server-access-log.txt file is delimited by '#'.

- [Click here for Hint](#)
- [Click here for Solution](#)

Task 7. Redirect the extracted output into a file.

Redirect the extracted data into a file named extracted-data.txt

- [Click here for Hint](#)
- [Click here for Solution](#)

Task 8. Transform the data into CSV format.

The extracted columns are separated by the original “#” delimiter.

We need to convert this into a “,” delimited file.

- [Click here for Hint](#)
- [Click here for Solution](#)

Task 9. Load the data into the table access_log in PostgreSQL

PostgreSQL command to copy data from a CSV file to a table is COPY.

The basic structure of the command is,

1. 1
1. COPY table_name FROM 'filename' DELIMITERS 'delimiter_character' FORMAT;

Copied!

The file comes with a header. So use the ‘HEADER’ option in the ‘COPY’ command.

Invoke this command from the shellscript, by sending it as input to ‘psql’ filter command.

- ▶ Click here for Hint
- ▶ Click here for Solution

Task 10. Execute the final script.

Run the final script.

- ▶ Click here for Solution

Task 11. Verify by querying the database.

- ▶ Click here for Hint
- ▶ Click here for Solution

Authors

Ramesh Sannareddy

Other Contributors

Rav Ahuja

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2022-07-27	0.3	Lakshmi Holla	Updated Linux command
2021-09-06	0.2	Ramesh Sannareddy	Incorporated the beta feedback.
2021-06-07	0.1	Ramesh Sannareddy	Created initial version of the lab

Copyright (c) 2021 IBM Corporation. All rights reserved.