Final Assignment (Part 1) - Creating ETL Data Pipelines using Bash with Apache Airflow

Estimated time needed: **90** minutes.

About This SN Labs Cloud IDE

This Skills Network Labs Cloud IDE provides a hands-on environment for course and project related labs. It utilizes Theia, an open-source IDE (Integrated Development Environment) platform, 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 Apache Airflow and MySQL database running in a Docker container. You will also need an instance of DB2 running in IBM Cloud.

Important Notice about this lab environment

Please be aware that sessions for this lab environment are not persistent. A new environment is created for you every time you connect to this lab. Any data you may have saved in an earlier session will get lost. To avoid losing your data, please plan to complete these labs in a single session.

Scenario

You are a data engineer at a data analytics consulting company. You have been assigned to a project that aims to de-congest the national highways by analyzing the road traffic data from different toll plazas. Each highway is operated by a different toll operator with a different IT setup that uses different file formats. Your job is to collect data available in different formats and consolidate it into a single file.

Objectives

In this assignment you will create a shell script using bash commands to:

- Extract data from a csv file
- Extract data from a tsv file
- Extract data from a fixed width file
- Transform the data
- Load the transformed data into a new csv file

You will then create a `DAG` to call the shell script.

Note - Screenshots

Throughout this lab you will be prompted to take screenshots and save them on your own device. These screenshots will need to be uploaded for peer review in the next section of the course. You can use various free screengrabbing tools or your operating system's shortcut keys (Alt + PrintScreen in Windows, for example) to capture the required screenshots. The screenshots can be saved with either the `.jpg` or `.png` extension.

Exercise 1 - Prepare the lab environment

1. Start Apache Airflow.

Open Apache Airflow in IDE

2. Download the dataset from the source to the destination `/home/project/airflow/dags` using `wget` command.

Source: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Final%20Assignment/tolldata.tgz

Note: While downloading the file in the terminal use the **sudo** command before the command used to download the file.

Exercise 2 - Create a DAG

Task 1.1 - Define DAG arguments

Define the DAG arguments as per the following details:

Parameter	Value
owner	< You may use any dummy name>
start_date	today
email	< You may use any dummy email>
email_on_failure	True
email_on_retry	True
retries	1
retry_delay	5 minutes

Take a screenshot of the task code.

Name the screenshot `dag_args.jpg`.

Task 1.2 - Define the DAG

Create a DAG as per the following details.

Parameter	Value
DAG id	`ETL_toll_data`

Parameter	Value
Schedule	Daily once
default_args	as you have defined in the previous step
description	Apache Airflow Final Assignment

Take a screenshot of the command you used and the output.

Name the screenshot `dag_definition.jpg`.

Task 1.3 - Create a shell script `Extract_Transform_data.sh` and add the following commands to your tasks:

Write a command to `unzip the data`.
Use the downloaded data from the url given in the first part of this assignment in exercise 1 and uncompress it into the destination directory.

```
1
2
3
4
5
```

```
1. Take a screenshot of the task code.
2.
3. Name the screenshot `unzip_data.jpg`.
4.
5. Read through the file `fileformats.txt` to understand the column details.
```

Copied!

Task 1.4 - Update the shell script to add a command to `extract data from csv file`

```
1
2
3
4
5
```

```
1. You should extract the fields `Rowid`, `Timestamp`, `Anonymized Vehicle number`, and `Vehicle type` from the `vehicle-data.csv` file and save them into a file named `csv_data.csv`.
2.
3. Take a screenshot of the task code.
4.
5. Name the screenshot `extract_data_from_csv.jpg`.
```

Copied!

Task 1.5 - Update the shell script to add a command to `extract data from tsv file`

```
1
2
3
4
5
```

```
1. You should extract the fields `Number of axles`, `Tollplaza id`, and `Tollplaza code` from the `tollplaza-data.tsv` file and save it into a file named `tsv_data.csv`.
2.
3. Take a screenshot of the task code.
4.
5. Name the screenshot `extract_data_from_tsv.jpg`.
```

Copied!

Task 1.6 - Update the shell script to add a command to `extract data from fixed width file`

```
1
2
3
4
5
```

```
1. You should extract the fields `Type of Payment code`, and `Vehicle Code` from the fixed width file `payment-data.txt` and save it into a file named `fixed_width_data.csv`.
2.
3. Take a screenshot of the task code.
4.
5. Name the screenshot `extract_data_from_fixed_width.jpg`.
```

Copied!

Task 1.7 -Update the shell script to add a command to consolidate data `extracted from previous tasks`

```
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
```

```
1. You should create a single csv file named `extracted_data.csv` by combining data from the following files:
2.
3. - `csv_data.csv`
4. - `tsv_data.csv`
5. - `fixed_width_data.csv`
6.
7. The final csv file should use the fields in the order given below:
8.
9. `Rowid`, `Timestamp`, `Anonymized Vehicle number`, `Vehicle type`, `Number of axles`, `Tollplaza id`, `Tollplaza code`, `Type of Payment code`, and `Vehicle Code`
10.
11. Hint: Use the bash `paste` command.
```

```
12.
13. `paste` command merges lines of files.
14.
15. Example : `paste file1 file2 > newfile`
16.
17. The above command merges the columns of the files file1 and file2 and sends the output to newfile.
18.
19. You can use the command `man paste` to explore more.
20.
21. Take a screenshot of the command you used and the output.
22.
23. Name the screenshot `consolidate_data.jpg`.
```

Copied!

Task 1.8 -. Update the shell script to add a command to `Transform and load the data`

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

```
1. You should transform the vehicle_type field in `extracted_data.csv` into capital letters and save it into a file named `transformed_data.csv` in the staging directory.
2.
3. Take a screenshot of the command you used and the output.
4.
5. Name the screenshot `transform.jpg`.
```

Copied!

Copy the shell script to `/home/project/airflow/dags` folder

Task 1.9 - Create a task `extract_transform_load` in the `ETL_toll_data.py` to call the shell script.

Define the task pipeline as per the details given below:

Task	Functionality
First task	`extract_transform_load`

Take a screenshot of the task pipeline section of the DAG.

Name the screenshot `task_pipeline.jpg`.

Exercise 3 - Getting the DAG operational.

Save the DAG you defined into a file named `ETL_toll_data.py`.

Task 1.10 - Submit the DAG

Take a screenshot of the command you used and the output.

Name the screenshot `submit_dag.jpg`.

Task 1.11 - Unpause the DAG

Take a screenshot of the command you used and the output.

Name the screenshot `unpause_dag.jpg`.

Task 1.12 - Monitor the DAG

Take a screenshot of the DAG runs for the Airflow console.

Name the screenshot `dag_runs.jpg`.

This concludes the assignment.

Authors

Ramesh Sannareddy

Niveditha Pandith

Other Contributors

Rav Ahuja

Lakshmi Holla

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-08-06	0.1	Ramesh Sannareddy	Created initial version
2022-08-26	0.2	Lakshmi Holla	Updated sudo commands
2023-02-25	0.3	Niveditha Pandith	Converted initial version

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