Data Pipelines Using Apache AirFlow

Scenario

Write a pipeline that analyzes the web server log file, extracts the required lines(ending with html) and fields(time stamp, size) and transforms (bytes to mb) and load (append to an existing file.)

Objectives

In this assignment you will author an Apache Airflow DAG that will:

- Extract data from a web server log file
- Transform the data
- Load the transformed data into a tar file

Tools / Software

- Apache AirFlow

Exercise 1 - Prepare the lab environment

Before you start the assignment:

- Start Apache Airflow.
- Download the dataset from the source to the destination mentioned below.

Source : https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0321EN-SkillsNetwork/ETL/accesslog.txt

Destination : /home/project/airflow/dags/capstone

Exercise 2 - Create a DAG

Task 1 - Define the DAG arguments

Create a DAG with these arguments.

- owner
- start_date
- email

You may define any suitable additional arguments.

Task 2 - Define the DAG

Create a DAG named `process_web_log` that runs daily.

Use suitable description.

Task 3 - Create a task to extract data

Create a task named `extract_data`.

This task should extract the ipaddress field from the web server log file and save it into a file named extracted_data.txt

Task 4 - Create a task to transform the data in the txt file

Create a task named `transform_data`.

This task should filter out all the occurrences of ipaddress “198.46.149.143” from extracted_data.txt and save the output to a file named `transformed_data.txt`.

Task 5 - Create a task to load the data

Create a task named `load_data`.

This task should archive the file `transformed_data.txt` into a tar file named `weblog.tar`.

Task 6 - Define the task pipeline

Define the task pipeline as per the details given below:

Task	Functionality
First task	`extract_data`
Second task	`transform_data`
Third task	`load_data`

Exercise 3 - Getting the DAG operational.

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

Task 7 - Submit the DAG

Task 8 - Unpause the DAG

Task 9 - Monitor the DAG

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-13-12	0.1	Ramesh Sannareddy	Created initial version
2022-30-01	0.2	Alison Woolford	Updated version
2022-04-14	0.2	Alison Woolford	Updated version