Assignment Overview: Data Pipelines using Apache Airflow

Welcome to the Data Engineering Capstone Project.
This video will give you an overview of lesson two of the fifth module – Data Pipelines Using Apache Airflow.

In this assignment, you will perform a couple of exercises. But before proceeding with the assignment, you will prepare the lab environment by starting the Apache Airflow and then downloading the data set from the source (using the link provided) to the appropriate destination.

In the first exercise, you will perform a series of tasks to create a directed acyclic graph (DAG) that runs daily. You will create a task that extracts the IP address and the date fields from the web server log file and then saves them to a CSV file.
The next task requires you to transform the date field into YYYYMMMDD format and save the output into a CSV file.
In the final task of the first exercise, you will load the data by archiving the transformed CSV file into a TAR file. Before moving on to the next exercise, you will define the task pipeline as per the given details.

In the second exercise, you will get the DAG operational by saving the defined DAG into a .PY file. Further, you will submit, unpause and then monitor that the DAG runs for the Airflow console. After performing each task, take a screenshot of the command you used and its output, and name the screenshot.
Good luck and let’s get started!