Technical Test: Deployment Design Proposal

Created by:

Fatemeh Haeri

Prepared for:

Career Foundry

The solution for deploying the project to the production environment

Introduction:

In order to deploy the project in the production environment, data collection operations must be performed and inserted in the database automatically and continuously, and then display the final output of the updated data. For this purpose, via Apache Airflow and creating workflows with scheduler at the desired time intervals, the mentioned process can be automated for ETL operations.

ETL operations include three steps as following:

- **First step (Extract data):** In this step, by connecting to the source, which here is the link of a web page with web scraping method, the data can be collected in JSON format with different time intervals.
- **Second step (Transform data):** In this step, by standard deviation function (std) from NumPy library, we calculate price volatility.
- **Third step (Load data):** In this step, we insert the data in the PostgreSQL database as a data repository.

For displaying the results in the form of charts, so we can use the PowerBI tool or Matplotlib plotting Python library.

Tools:

- **Extract from web site:** Scrapy library for crawling the web site
- **ETL operation:** Apache Airflow
- **Programming language:** Python
- **Database:** PostgreSQL
- **Visualizations:** PowerBI – Python Matplotlib library
- **API for displaying charts with Matplotlib output:** Flask

Conclusion:

As a result of the above progress, the project deployment operation can be performed effortlessly with high performance.