Lab 6: Workflow orchestration with Prefect

1. Objective

By the end of this lab, you will:

1. Convert a monolithic data pipeline into prefect tasks and flows
2. Use prefect features to schedule, log, deploy a pipeline

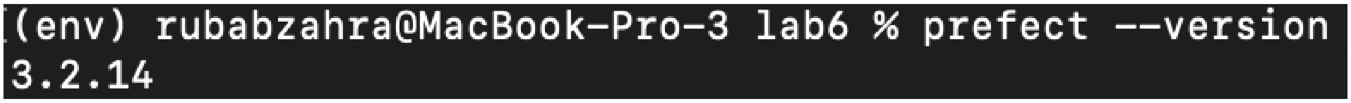
2. Prerequisites

1. **Python 3** environment and relevant packages. We'd recommend creating a new virtual environment using:
    - Navigate to the cloned folder lab5: cd <your-path>/lab6
    - Create environment: python -m venv env
    - Activate environment: source env/bin/activate
    - Install packages: pip install <package-name>
    - Packages: prefect, pandas, scikit-learn, joblib, matplotlib
2. Download the following 4 files (they are uploaded on LMS as well)
    - wget https://github.com/rubabzs/ai601-data-engineering/blob/main/labs/lab6/analytics_pipeline.py
    - wget https://github.com/rubabzs/ai601-data-engineering/blob/main/labs/lab6/ml_pipeline.py
    - wget https://github.com/rubabzs/ai601-data-engineering/blob/main/labs/lab6/Iris.csv
    - wget https://github.com/rubabzs/ai601-data-engineering/blob/main/labs/lab6/analytics_data.csv
    - wget https://github.com/rubabzs/ai601-data-engineering/blob/main/labs/lab6/prefect.yaml

3. Setup Prefect

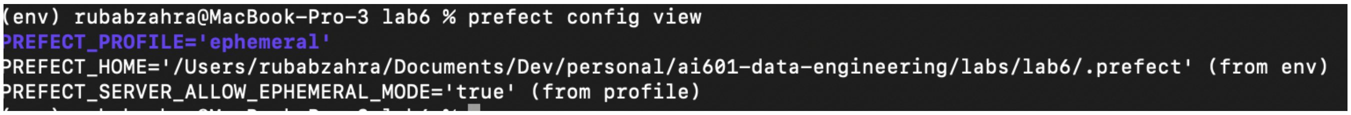
Please follow the steps below to verify prefect is configured correctly:

1. Check prefect version:



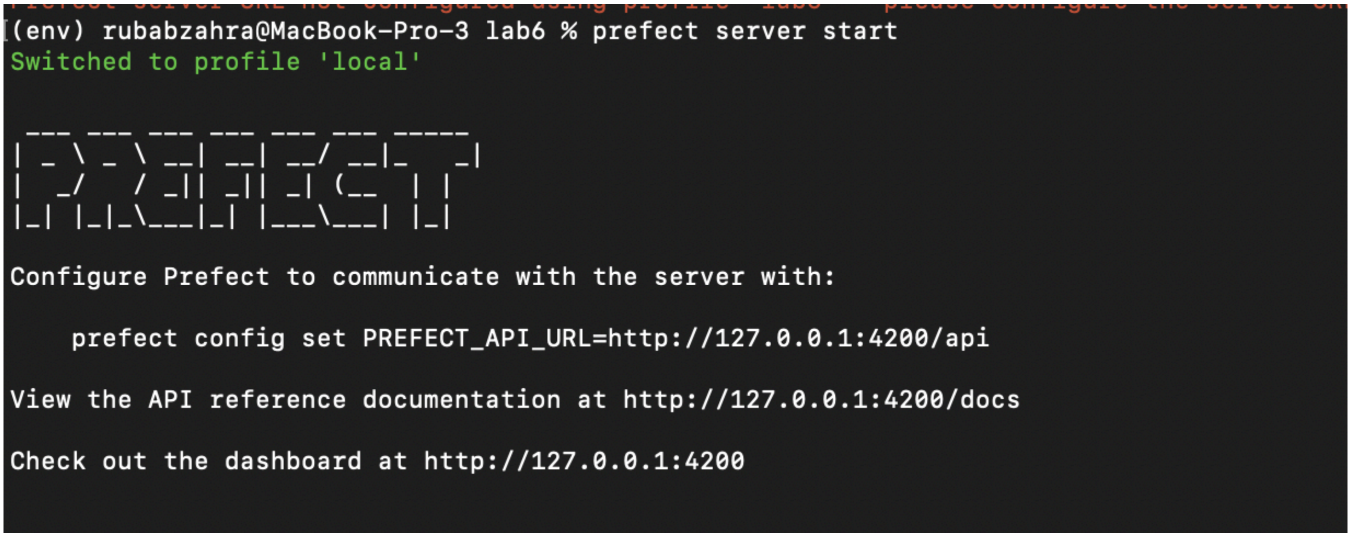
```
(env) rubabzahra@MacBook-Pro-3 lab6 % prefect --version
3.2.14
```

2. View prefect configuration:



```
(env) rubabzahra@MacBook-Pro-3 lab6 % prefect config view
PREFECT_PROFILE='ephemeral'
PREFECT_HOME='/Users/rubabzahra/Documents/Dev/personal/ai601-data-engineering/labs/lab6/.prefect' (from env)
PREFECT_SERVER_ALLOW_EPHEMERAL_MODE='true' (from profile)
(env) rubabzahra@MacBook-Pro-3 lab6 %
```

3. Start the server:

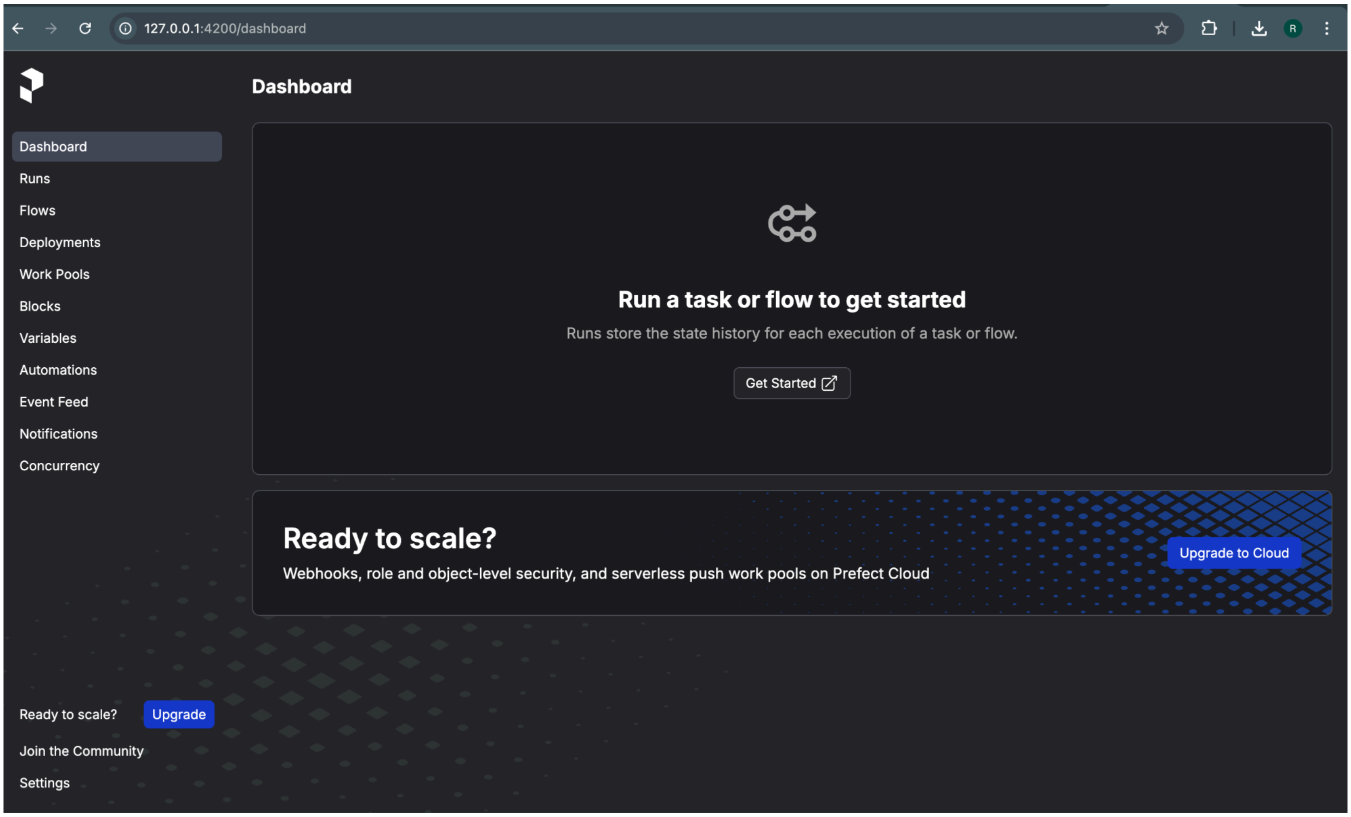


```
Prefect server URL not configured using profile 'lab6' - please configure the server URL
(env) rubabzahra@MacBook-Pro-3 lab6 % prefect server start
Switched to profile 'local'

 ___ ___ ___ ___ ___ ___ _____
| _ \ _ \ __| __| __/ __|_   _|
|  _/   / _|| _|| _| (__  | |
|_| |_|_\___|_| |___\___| |_|

Configure Prefect to communicate with the server with:

    prefect config set PREFECT_API_URL=http://127.0.0.1:4200/api

View the API reference documentation at http://127.0.0.1:4200/docs

Check out the dashboard at http://127.0.0.1:4200
```

4. Don't forget to run the command shown above:

- prefect config set PREFECT_API_URL=http://127.0.0.1:4200/api

5. Visit 12.0.0.1:4200 to access prefect web UI



Dashboard

Run a task or flow to get started

Runs store the state history for each execution of a task or flow.

Ready to scale?

Webhooks, role and object-level security, and serverless push work pools on Prefect Cloud

Congratulations! You have prefect up and running!

4. Pipeline 1

This script reads a CSV dataset, performs data validation and transformation, generates summary statistics, and produces a histogram report—all in one sequential script.

Task 1: Convert the Analytics Pipeline

- Break the monolithic `analytics_pipeline.py` into discrete tasks using Prefect's `@task` decorator.
- You will need to import some function:
  - from prefect import task, flow, get_run_logger
- Define a `@flow` to orchestrate the tasks.
- Run the flow using python analytics_pipeline.py command. You should be able to see your flow running in the UI.
- Try adding logs to a function: https://docs.prefect.io/v3/develop/logging

5. Pipeline 2

This script reads the Iris dataset, validates and transforms the data, trains a RandomForest model with a train/test split, evaluates the model's accuracy, and conditionally saves the model if the accuracy meets a threshold.

Task 2: Convert the ML Pipeline

- Refactor ml_pipeline.py into a Prefect flow.
- Create individual tasks for data fetching, validation, transformation, training (with retries), evaluation, and conditional saving.
- Use Prefect's parameterization to allow changes to parameters like dataset_path, accuracy_threshold, and test_size.
- Create a workpool:
    - prefect work-pool create "default"
- Configure a deployment using a YAML file.
    - prefect deploy
    - Follow along the prompts
- Have a look under the deployment tab your deployment should be available but in 'Not Ready' state
- You need to start a worker for it to pick up this deployment
    - prefect worker start --pool "default"
- Trigger the flow from UI

6. Submission

Zip all the changed (analytics_pipeline, ml_pipeline) files and upload the zipped folder on LMS