MLB Data Project Process

Set up AWS EC2. Need a t2 medium at least to have enough storage.

Need to open port 8080.

Install docker on ec2 and use docker-compose file for airflow. Need to change CeleryExecutor to LocalExecutor in docker-compose file for airflow to run well on ec2.

Create dags, logs, plugins folders for airflow to use.

Create .env file:

AIRFLOW\_UID=50000

AIRFLOW\_GID=0

<https://www.youtube.com/watch?v=J6azvFhndLg>

<https://airflow.apache.org/docs/apache-airflow/stable/docker-compose.yaml>

<https://github.com/mikestack15/orangutan-stem/wiki/Activity-1:-Open-Weather-Map-API-Data-Pipeline>

Get ssh key from ec2 instance and add it to github account so you can pull from github.

Set up docker and airflow locally as well as a development environment.

Don't need to change the docker-compose file at all.

Create S3 bucket on aws. Go to iam and create a user. Security management for aws keys.

Add keys in airflow connector extra section to upload data to S3.

Still need to improve dags and look into connecting to power bi