CASE 1: **Creating Data science pipelines**

Team 3 in INFO 7390

Team Members: Hongwei Hu, Guangnan Liang

GitHub: <https://github.com/Besimilar/AdvancedDataAnalysis/tree/master/DataSciencePipelines/Pipeline>

Docker Hub: besimilar/advanced-data-analysis:final

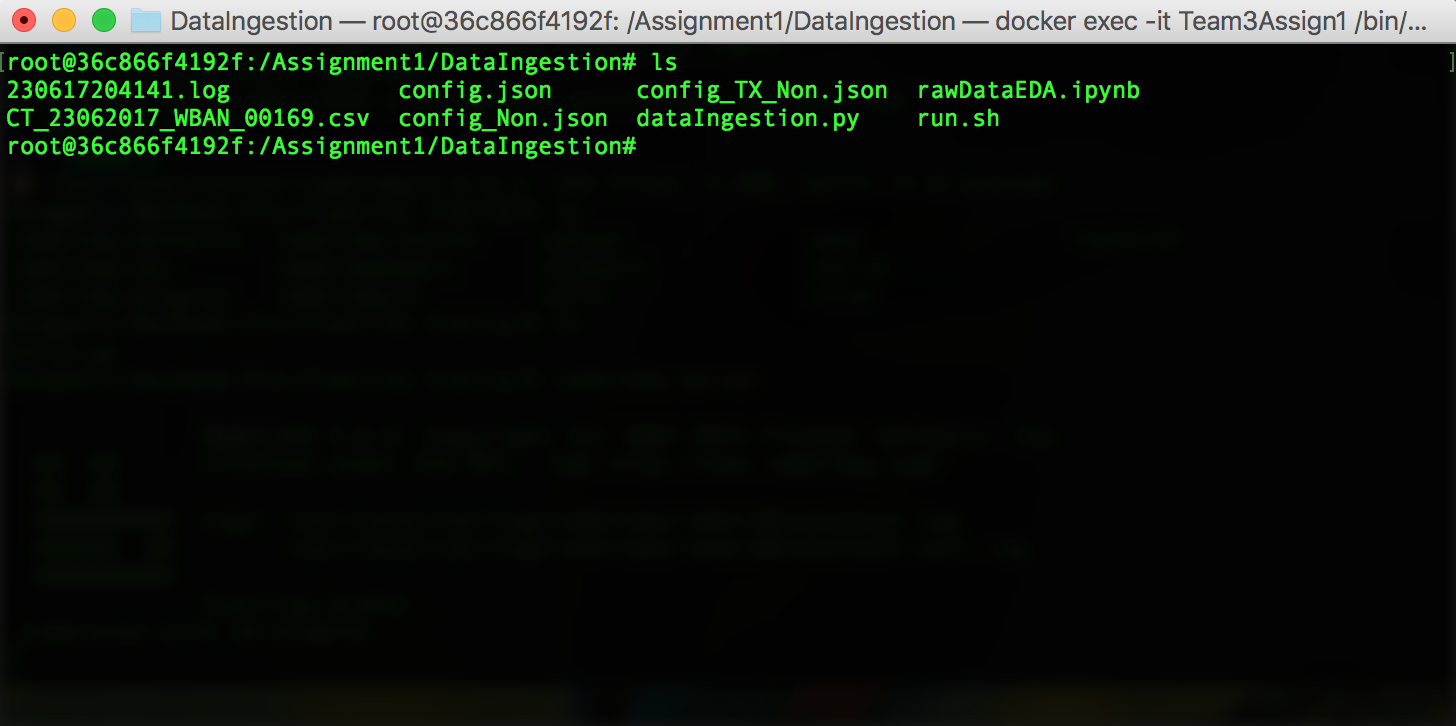
Resources:

1. Pipeline: Celery

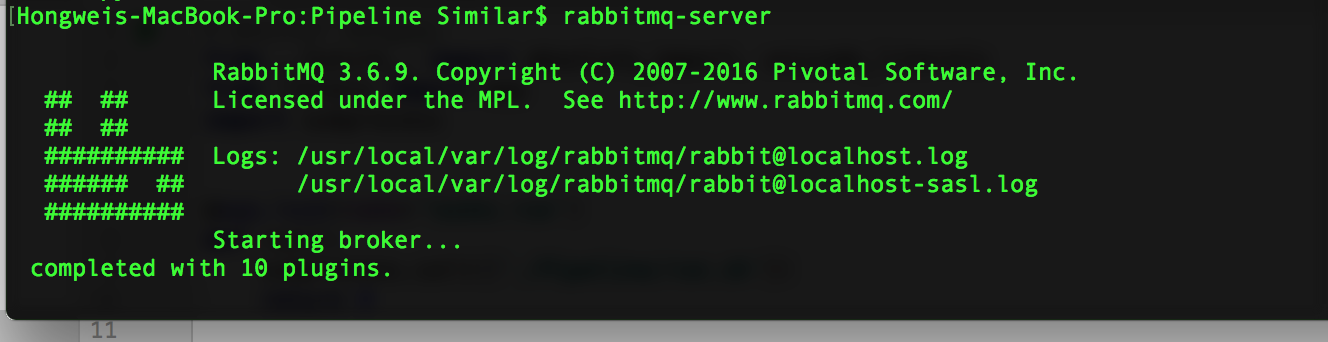
Note: Red lines means Very Important Steps.

**Create Pipeline**

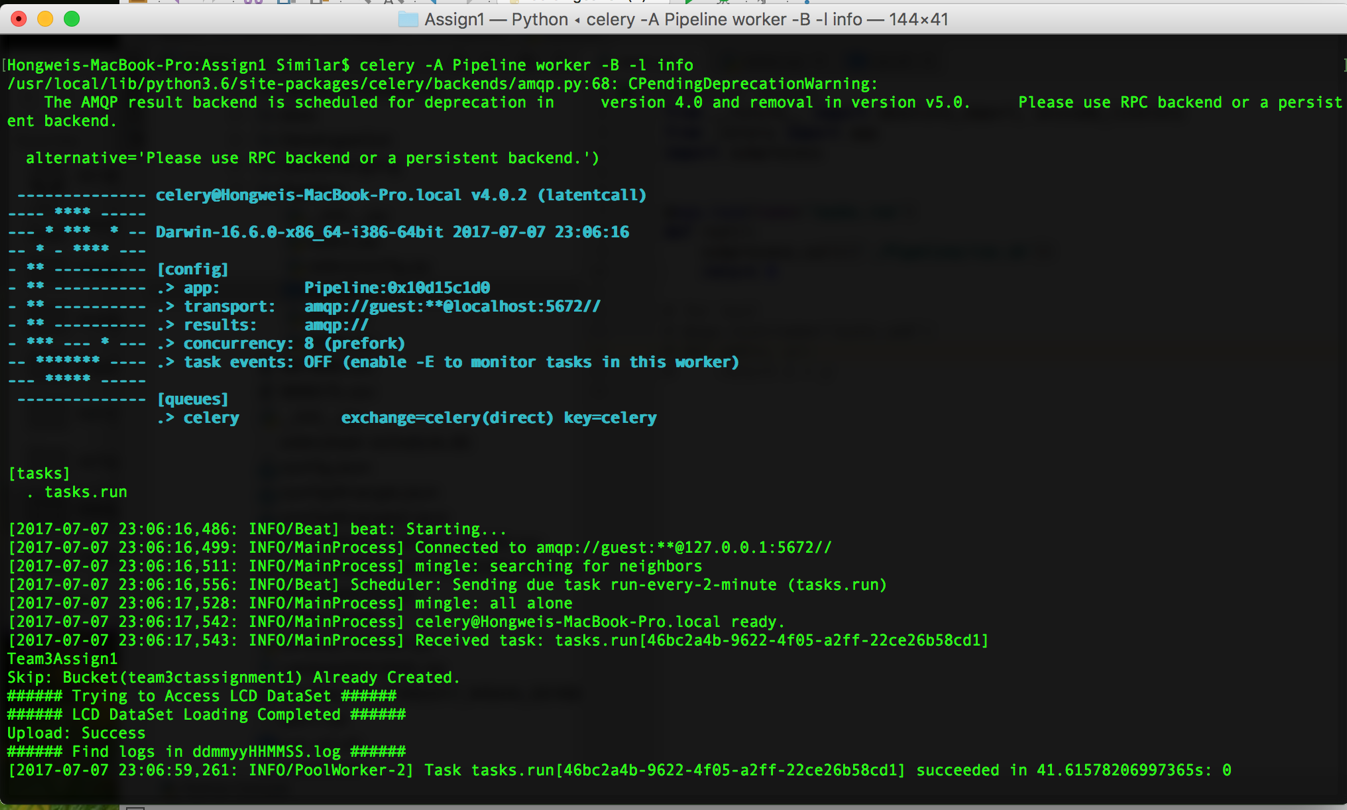
1. Configuration:
   1. Add AWS keys to config.json in container

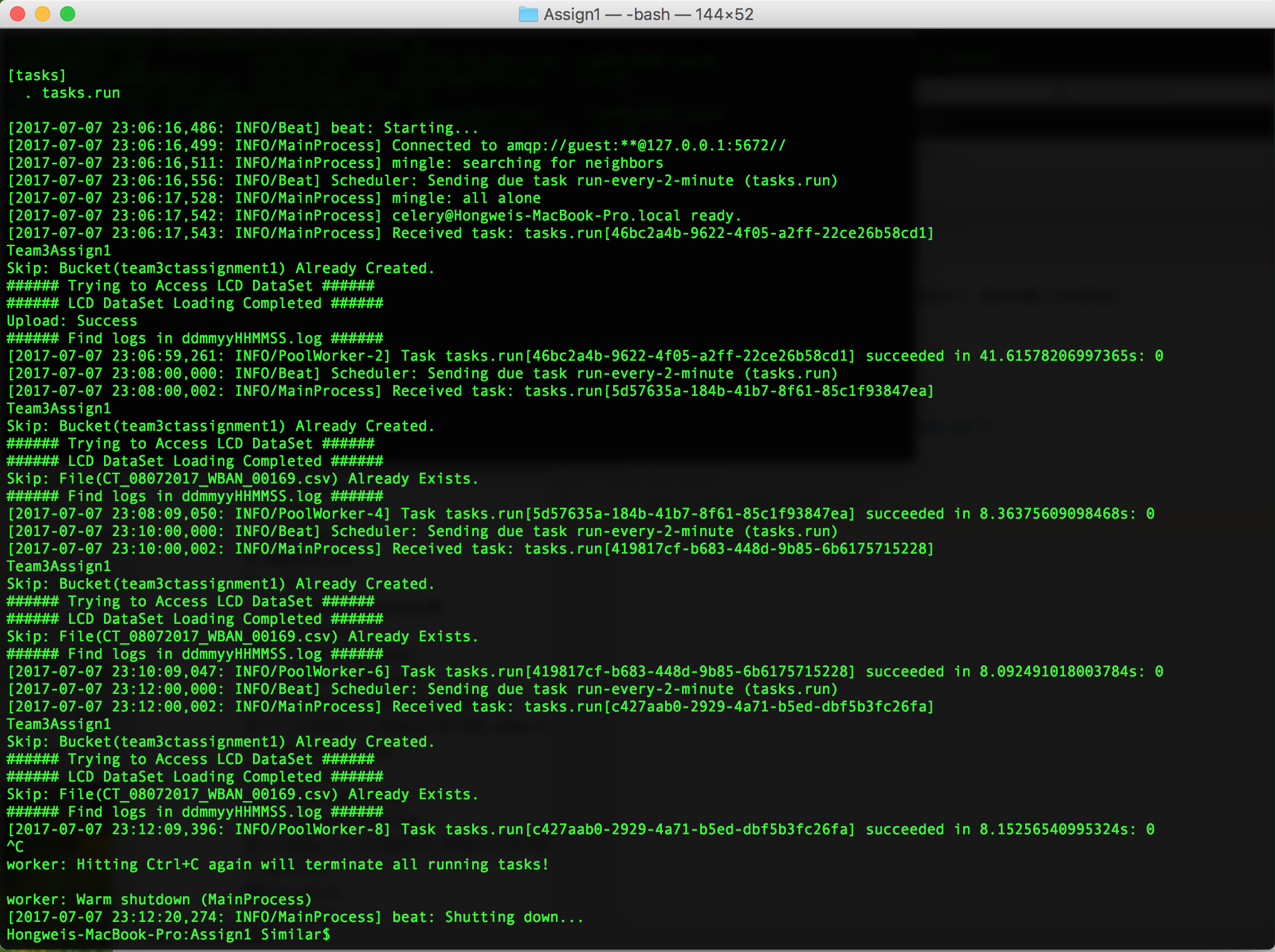


1. Start the Server:
   1. $ rabbitmq-server



1. Start Scheduling:
   1. $ celery –A <projectName> worker –B –l info
   2. Time intervals: every 2 minutes

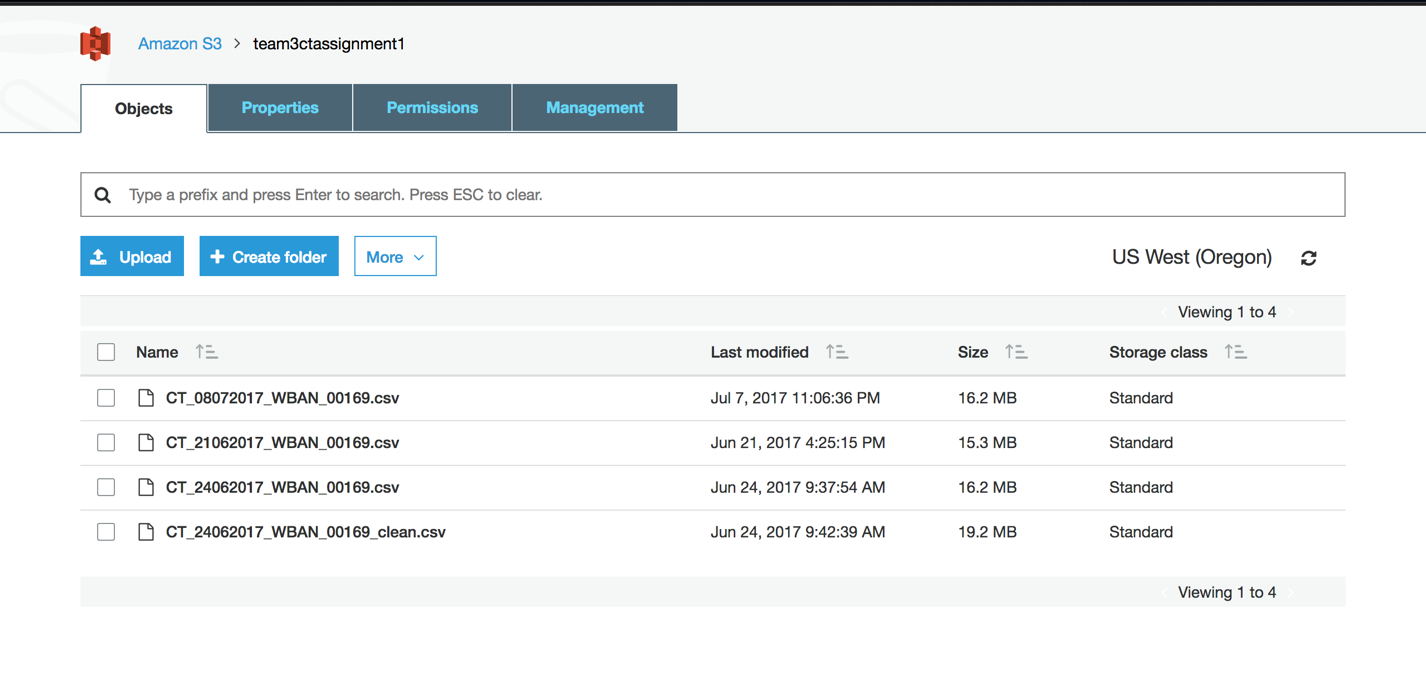




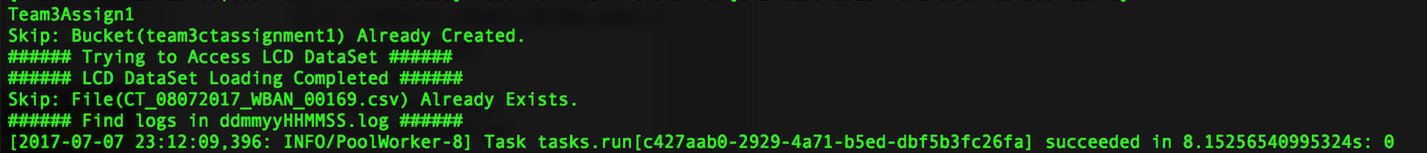
* 1. If you need to set scheduler to run every day:
     1. Simply modify the part below in ‘celery.py’:



1. Result:
   1. You can see a new dataset of today was uploaded into S3 (Jul 7).



* 1. Because each day you can only upload one version of the dataset, the latter tasks skipped uploading.



* 1. However, you will still find logs for each task:
     1. In this case, they are:
        1. 080717030618.log
        2. 080717030801.log
        3. 080717031001.log
        4. 080717031201.log
     2. Therefore, every two minutes a new task will be run through pipline.

