Exercise 13

1

Watch the required videos on AWS codepipeline and AWS datapipeline, and read AWS's literature on the products.

In your own words:

Compare AWS data pipeline with off-the-shelf airflow in kubernetes; and AWS code pipeline with Jenkins. How does their feature set differ and how is it the same?

What are the advantages and disadvantages to using an AWS service vs the setup we have used in kubernetes?

If you were building a data pipeline, which would you choose?

2

Configure a pipeline in airflow to run "extract_mysql_fulll.py", as done last week, but make some changes:

- extract_mysql_full.py should store the csv file in s3. You can copy the code to do that from extract_mysql_incremental.py (be sure to pull the latest version from github as it was modified recently). The name of your generated s3 object should be: `week13-<username>.csv` in bucket UML.
- The dag file should be loaded into airflow from *your* github repository. Update the helm configuration as described in lab 1 and the helm documentation.

Submit

1. Airflow logs of your run
2. Output of the AWS CLI command: aws s3 ls UML
3. Changes to the "dags:" section of the helm config file
4. Your modified python program
5. A description of your solution: how you overcame problems, and how you tested.