10-405/10-605
Machine Learning with Large Datasets
Homework Setups

...

Tian Li
tianli@cmu.edu
01/17

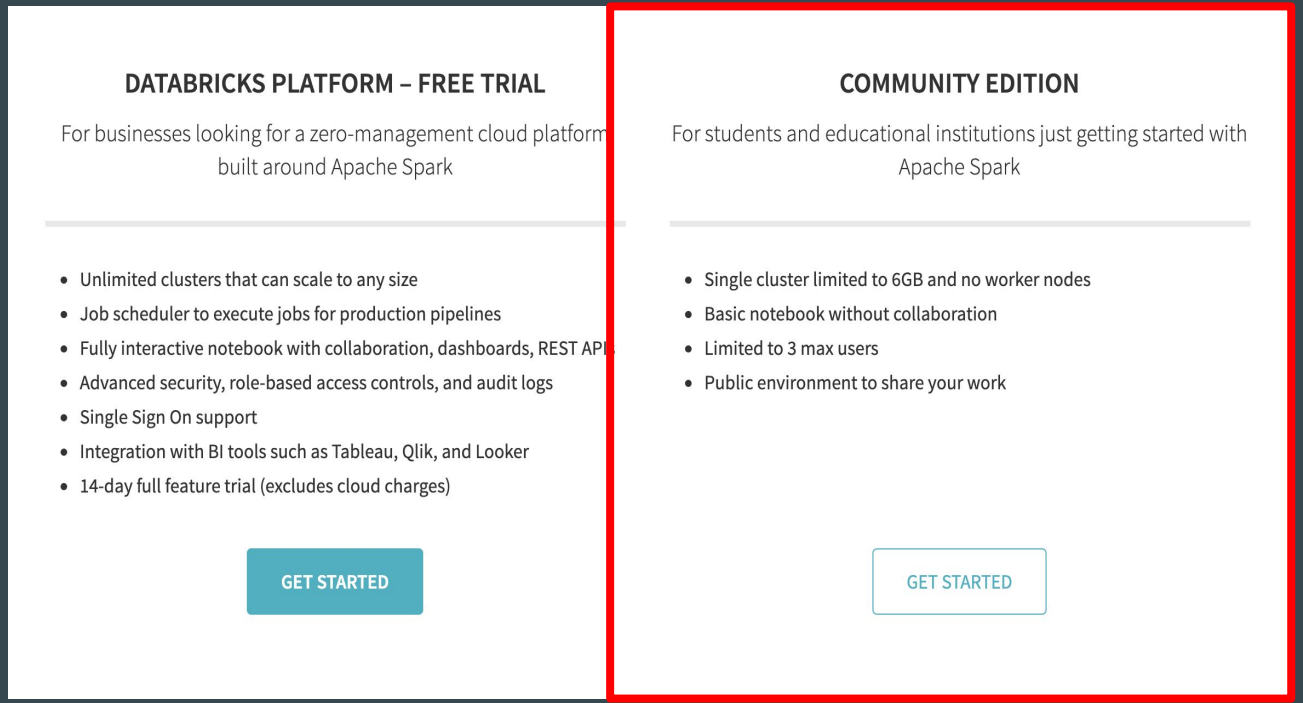
Homework Overview

- **Programming Assignments (3 on Spark; 2 on Tensorflow)**
  - Spark homeworks: Databricks
- **To-dos:**
  - Register for a free community version of Databricks
  - Import the IPython Notebook file we provide
  - Configure the environment according to instructions in the writeup (creating a cluster, installing a third-party package, and starting running)
  - Hand in the solution to Gradescope (see the writeup)
- **Tensorflow homeworks: provide information later in the course**

Registration

https://databricks.com/try-databricks

Make sure to choose the community edition

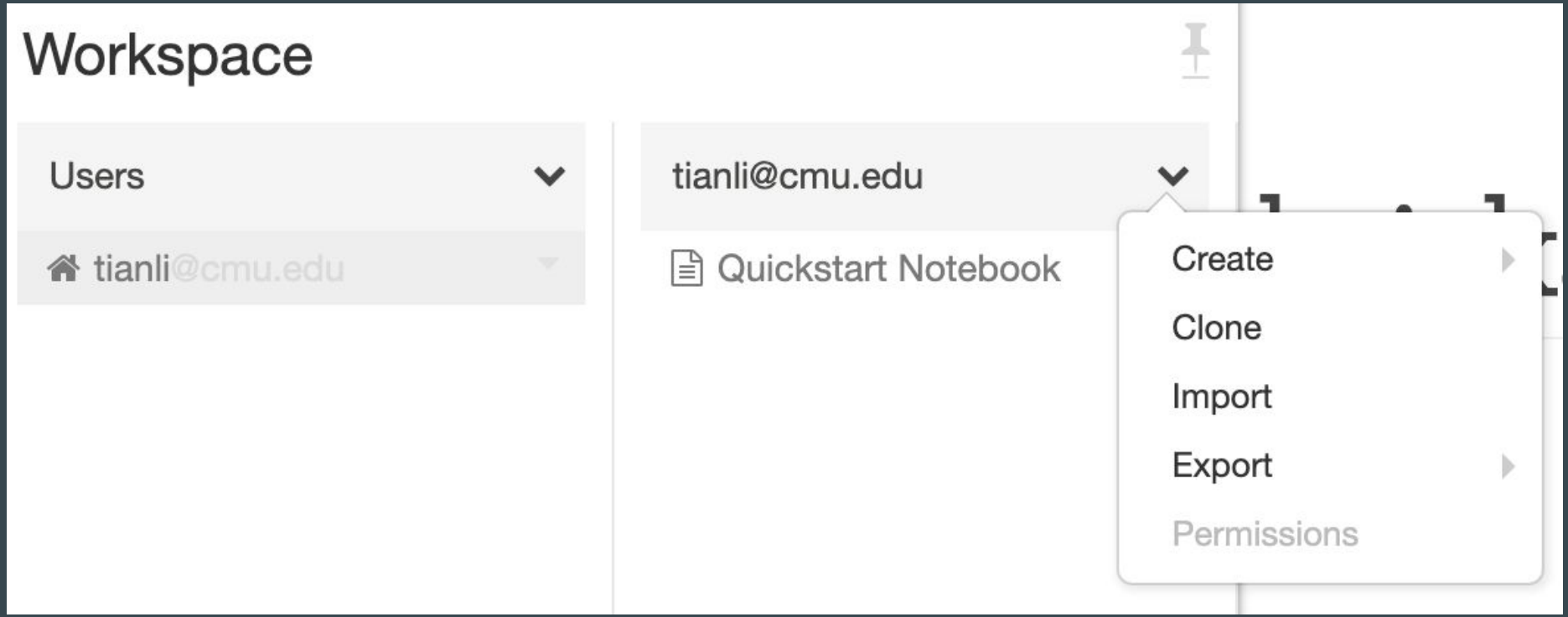


Login

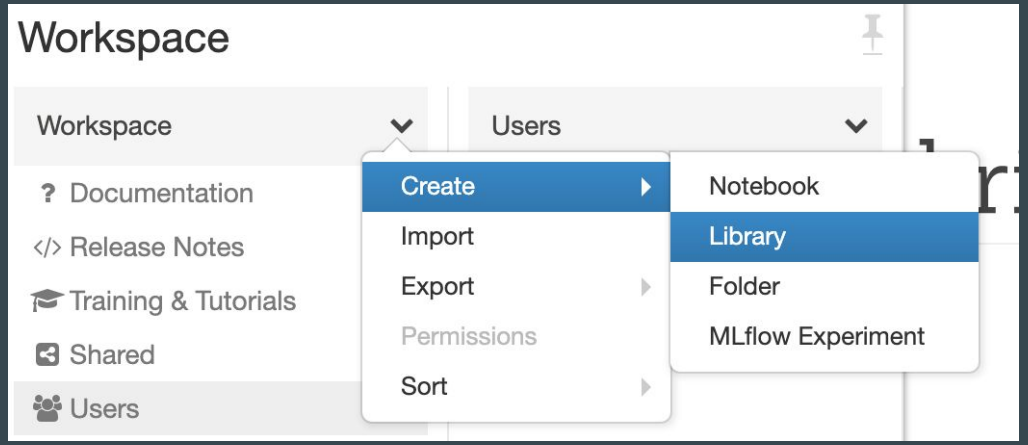
Still Log in to Community Edition:

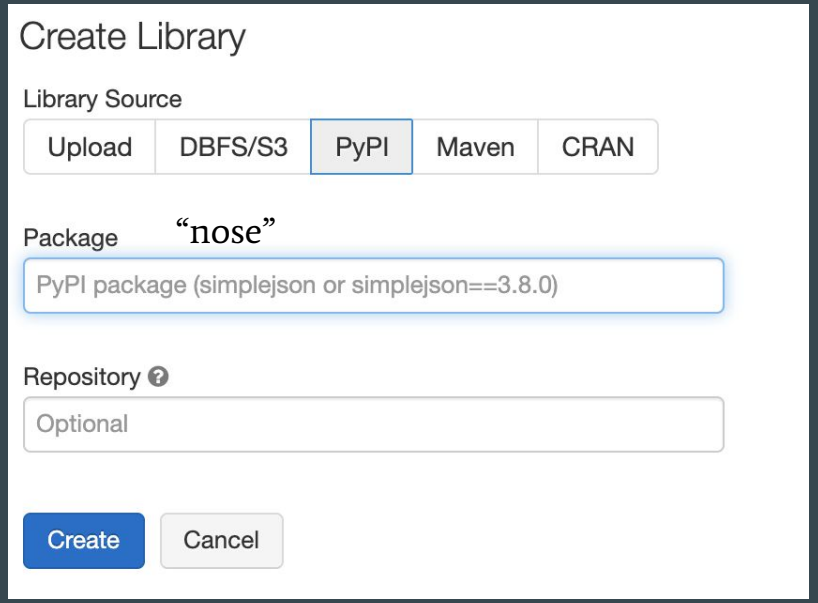
https://community.cloud.databricks.com/login.html

Import Lab Files

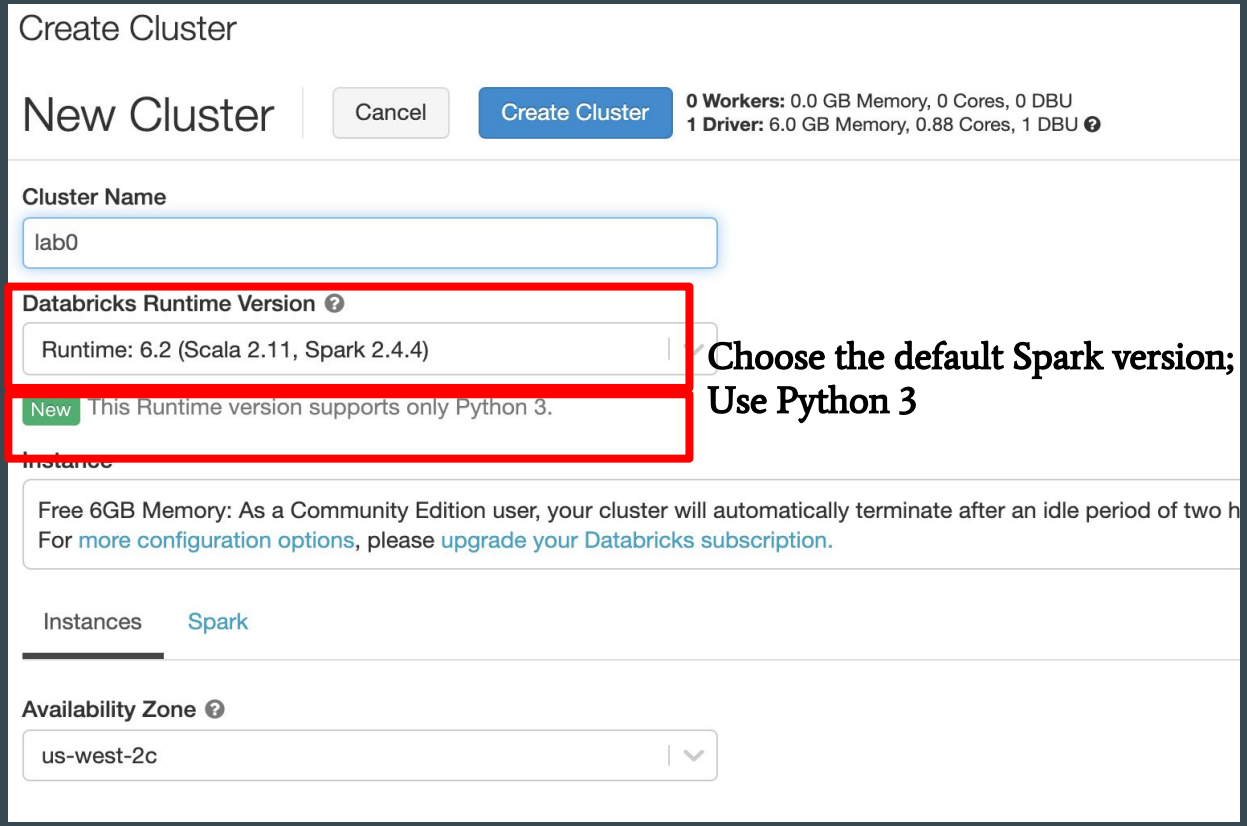


Installing Third-party Packages





Creating a Cluster



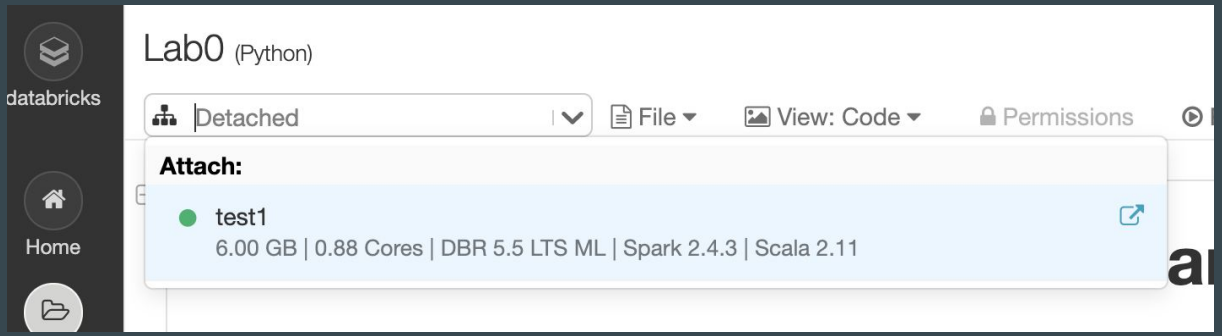
Choose the default Spark version;
Use Python 3

Notes about Clusters

- Spark version: default on Databricks (2.4.4); Python 3
- It may take a while to launch the cluster (e.g., 20 seconds)
- The cluster status should be ‘active’ for it to be functional
- The community edition only allows for one cluster, which is essentially a single machine
    - When you start a second notebook, either delete the current cluster and create a new one; or attach to and activate the existing (terminated) cluster
- Max memory: 6GB (enough for our homeworks)

Interact with Notebooks

- Attach to the cluster



- Similar with interacting with Jupyter Notebook
- Export the homework as an IPython file, and submit it to Gradescope