

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

DATABRICKS PLATFORM – FREE TRIAL

For businesses looking for a zero-management cloud platform built around Apache Spark

- Unlimited clusters that can scale to any size
- Job scheduler to execute jobs for production pipelines
- Fully interactive notebook with collaboration, dashboards, REST API
- Advanced security, role-based access controls, and audit logs
- Single Sign On support
- Integration with BI tools such as Tableau, Qlik, and Looker
- 14-day full feature trial (excludes cloud charges)

GET STARTED

COMMUNITY EDITION

For students and educational institutions just getting started with Apache Spark

- Single cluster limited to 6GB and no worker nodes
- Basic notebook without collaboration
- Limited to 3 max users
- Public environment to share your work

GET STARTED

Login

Still Log in to Community Edition:

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

Import Lab Files

Workspace

Users



🏠 tianli@cmu.edu



tianli@cmu.edu



📄 Quickstart Notebook

Create



Clone

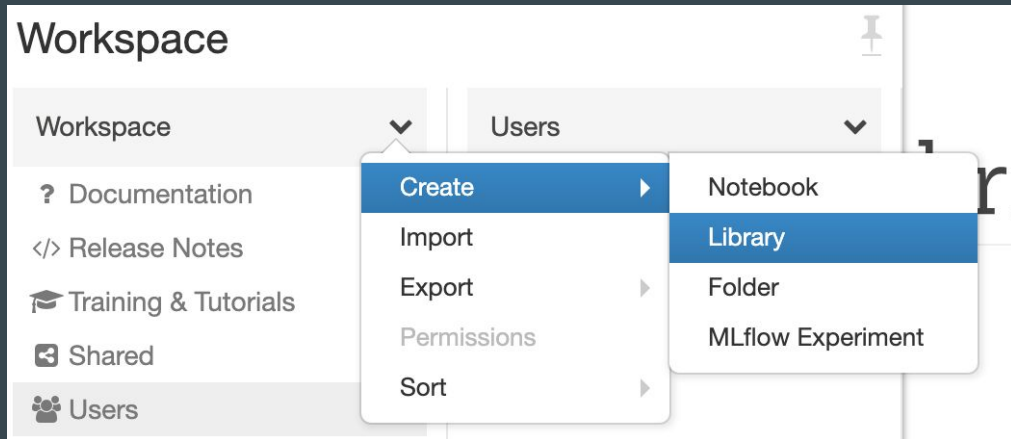
Import

Export



Permissions

Installing Third-party Packages



Create Library

Library Source

Upload DBFS/S3 **PyPI** Maven CRAN

Package "nose"

PyPI package (simplejson or simplejson==3.8.0)

Repository ?

Optional

Create Cancel

Creating a Cluster

Create Cluster

New Cluster

Cancel

Create Cluster

0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 6.0 GB Memory, 0.88 Cores, 1 DBU ?

Cluster Name

lab0

Databricks Runtime Version ?

Runtime: 6.2 (Scala 2.11, Spark 2.4.4)

New This Runtime version supports only Python 3.

Instances

Spark

Availability Zone ?

us-west-2c

Free 6GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two h
For [more configuration options](#), please [upgrade your Databricks subscription](#).

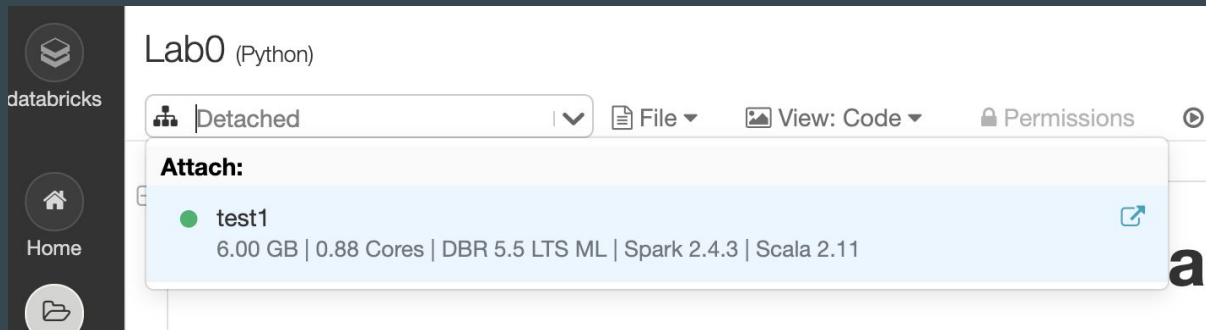
**Choose the default Spark version;
Use Python 3**

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

