High Performance Research Computing

A Resource for Research and Discovery

TEXAS A&M UNIVERSITY

Spark for Big Data

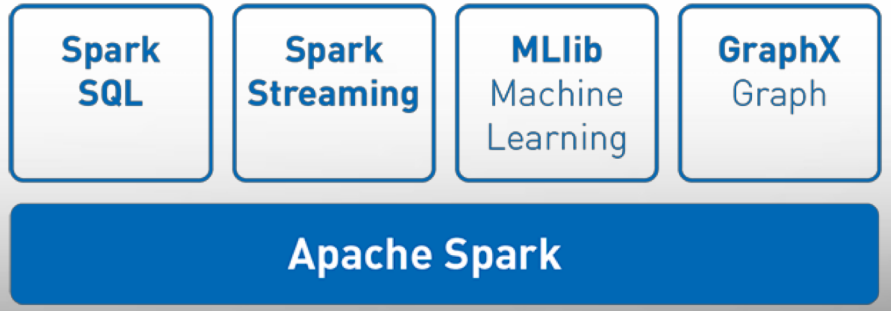
Rick McMullen, Associate Director HPRC
mcmullen@tamu.edu

help@hprc.tamu.edu

DIVISION OF RESEARCH
TEXAS A&M UNIVERSITY

What is Spark?

- General-purpose distributed data processing engine
- Data management primitives plus
  - SQL, machine learning, graph computation
- Machine learning
- Interactive analytics
- Data integration
- Stream processing



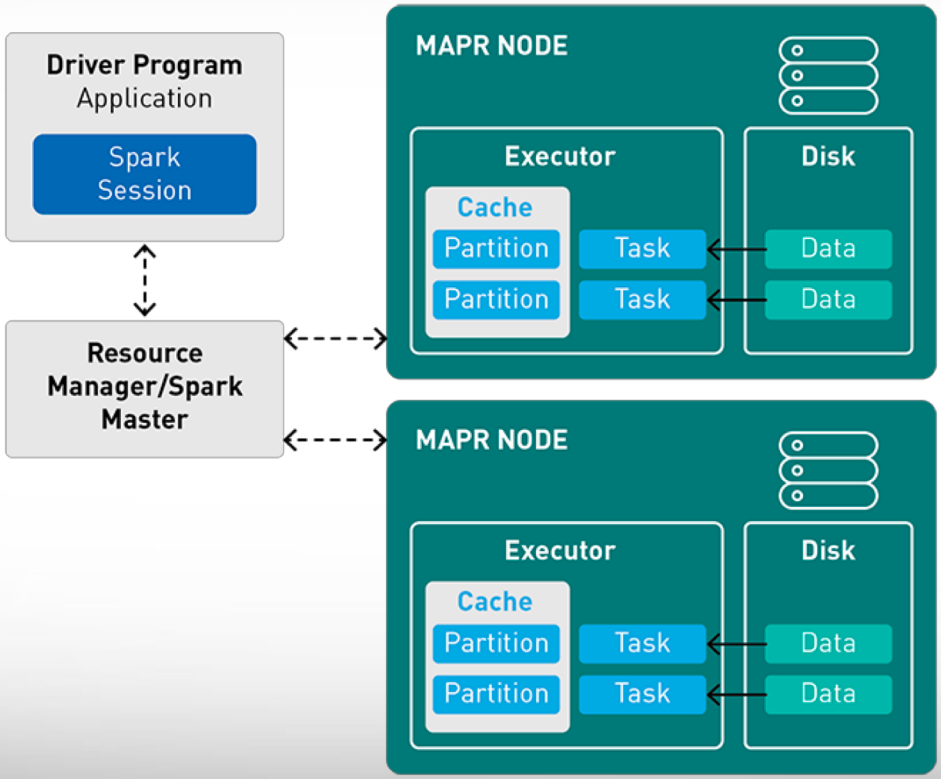
Lots of useful bindings

- Python 2,3
- Scala
- Java
- C#
- Clojure DSL
- Haskell
- Back-end for R

(SQL) connectors

- CSV
- XML
- Mongodb
- Cassandra
- Inline SQL queries
- JDBC & ODBC

How Spark works on a cluster



https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/

Basic Spark

- Create or load an RDD
  - Load a data set from some file or URL
- Transform the RDD
  - Filter and select data from RDD
- Wait until data is needed before doing anything
  - Stack up more transformations
- Perform Actions that return data
  - Return something specific about the data

Example… in python

```
from pyspark import SparkConf, SparkContext
conf = SparkConf().setMaster("local").setAppName("Test_App")
sc = SparkContext(conf = conf)

lines_rdd = sc.textFile("nasa_19950801.tsv")
stanfordLines_rdd = lines_rdd.filter(lambda line: "stanford" in line)

stanfordLines_rdd.count()
47
stanfordLines_rdd.first()
u'glim.stanford.edu\t-\t807258357\tGET\t/shuttle/missions/61-c/61-c-patch-small.gif\t‘
```

Setup

Create rdd

Transform data

Perform *Actions* on data

Lambdas - “small” filter/selector operators

```
from pyspark import SparkConf, SparkContext
conf = SparkConf().setMaster("local").setAppName("Test_App")
sc = SparkContext(conf = conf)

lines_rdd = sc.textFile("nasa_19950801.tsv")
stanfordLines_rdd = lines_rdd.filter(lambda line: "stanford" in line)

stanfordLines_rdd.count()
47
stanfordLines_rdd.first()
u'glim.stanford.edu\t-\t807258357\tGET\t/shuttle/missions/61-c/61-c-patch-small.gif\t‘
```

“Select only lines with the string ‘stanford’ in them”

Common *Transformations*

Transformation	Result
map(func)	Return a new RDD by passing each element through func
filter(func)	Return a new RDD by selecting the elements for which func returns true
flatMap(func)	func can return multiple items, and generate a sequence, in order to “flatten” nested entries in JSON into a Python list then return as RDD
distinct()	Return an RDD with only distinct entries
sample(...)	Sample to create a subset of the RDD
union(RDD)	Return a union of the RDDs
intersection(RDD)	Return an intersection of the RDDs
subtract(RDD)	Remove contents of RDD from other
cartesian(RDD)	Cartesian product of the RDDs
parallelize(list)	Create an RDD from Python list using the current spark context

Common *Actions*

Transformation	Result
collect()	Return all the elements from the RDD
count()	Number of elements in RDD
countByValue()	Count of instances of each value in the RDD
reduce(func)	Aggregate elements of the RDD using a function which combines any two into one (sum, min, max, etc.)
first(), take(n)	Return the first, or first n elements
top(n)	Return the n highest valued elements of the RDDs
takeSample(...)	Various options to return a subset of the RDD
saveAsTextFile(path)	Write the elements as a text file
foreach(func)	Run the *func* on each element. Used for side-effects (updating accumulator variables) or interacting with external systems

Key/value – Pair RDDs

- Key/value pairs are simple and efficient
- Spark supports this with special operations on key/value “pair” RDDs
- Implemented in a language as 2-tuples

Pair RDD Transformations

Transformation	Result
reduceByKey(func)	Reduce values using *func*, but on a key by key basis, i.e. combine values with the same key
groupByKey()	Combine values with same key. Each key ends up with a list.
sortByKey()	Return an RDD sorted by key
mapValues(func)	Use *func* to change values, but not key
keys()	Return an RDD of only keys
values()	Return an RDD of only values

Two Pair RDD Transformations

Transformation	Result
subtractByKey(otherRDD)	Remove elements with a key present in other RDD
join(otherRDD)	Inner join: Return an RDD containing all pairs of elements with matching keys in self and other. Each pair of elements will be returned as a (k, (v1, v2)) tuple, where (k, v1) is in self and (k, v2) is in other
leftOuterJoin(otherRDD)	For each element (k, v) in self, the resulting RDD will either contain all pairs (k, (v, w)) for w in other, or the pair (k, (v, None)) if no elements in other have key k
rightOuterJoin(otherRDD)	For each element (k, w) in other, the resulting RDD will either contain all pairs (k, (v, w)) for v in this, or the pair (k, (None, w)) if no elements in self have key k
cogroup(otherRDD)	Group data from both RDDs by key

Pair RDD Actions

Action	Result
countByKey()	Count the number of elements for each key
lookup(key)	Return all the values for this key

Spark input and output options

- Text
- CSV
- JSON
- Database-like things
  - Sequence files (key/value)
  - Old and new Hadoop API
  - Hbase
  - Compression (gzip...)
  - SQL
  - MongoDB
- Protocol Buffers (serialization of structured data for Google apps)

Some fun with Spark and Jupyter

- We will be using the HPRC portal
- **https://portal.hprc.tamu.edu**
- You must have an Ada cluster account

- Log in using CAS and Duo **with your NetID and password**

Dashboard

https://portal.hprc.tamu.edu/pun/sys/dashboard

TAMU HPRC OnDemand (Ada) | Files | Jobs | Clusters | Interactive Apps | My Interactive Sessions | Help | Logged in as mcmullen | Log Out

OnDemand provides an integrated, single access point for all of your HPC resources.

Message of the Day

IMPORTANT POLICY INFORMATION

- **Unauthorized use of HPRC resources is prohibited and subject to criminal prosecution.**
- **Use of HPRC resources in violation of United States export control laws and regulations is prohibited. Current HPRC staff members are US citizens and legal residents.**
- **Sharing HPRC account and password information is in violation of State Law. Any shared accounts will be DISABLED.**
- **Authorized users must also adhere to ALL policies at: https://hprc.tamu.edu/policies**

!! WARNING: There are NO active backups of user data. !!

powered by OPEN OnDemand

Dashboard

https://portal.hprc.tamu.edu/pun/sys/dashboard

TAMU HPRC OnDemand (Ada) | Files | Jobs | Clusters | Interactive Apps | My Interactive Sessions | Help | Logged in as mcmullen | Log Out

OnDemand provides an integrated, s…our HPC resources.

ADA
- ANSYS Workbench
- Abaqus/CAE
- IGV
- LS-PREPOST
- MATLAB
- ParaView
- VNC

Galaxy
- Galaxy (maroon)
- Galaxy (reveille)

Servers
- Jupyter Notebook (Ada)
- RStudio Server (Singularity)
- Spark-Jupyter Notebook (Ada)

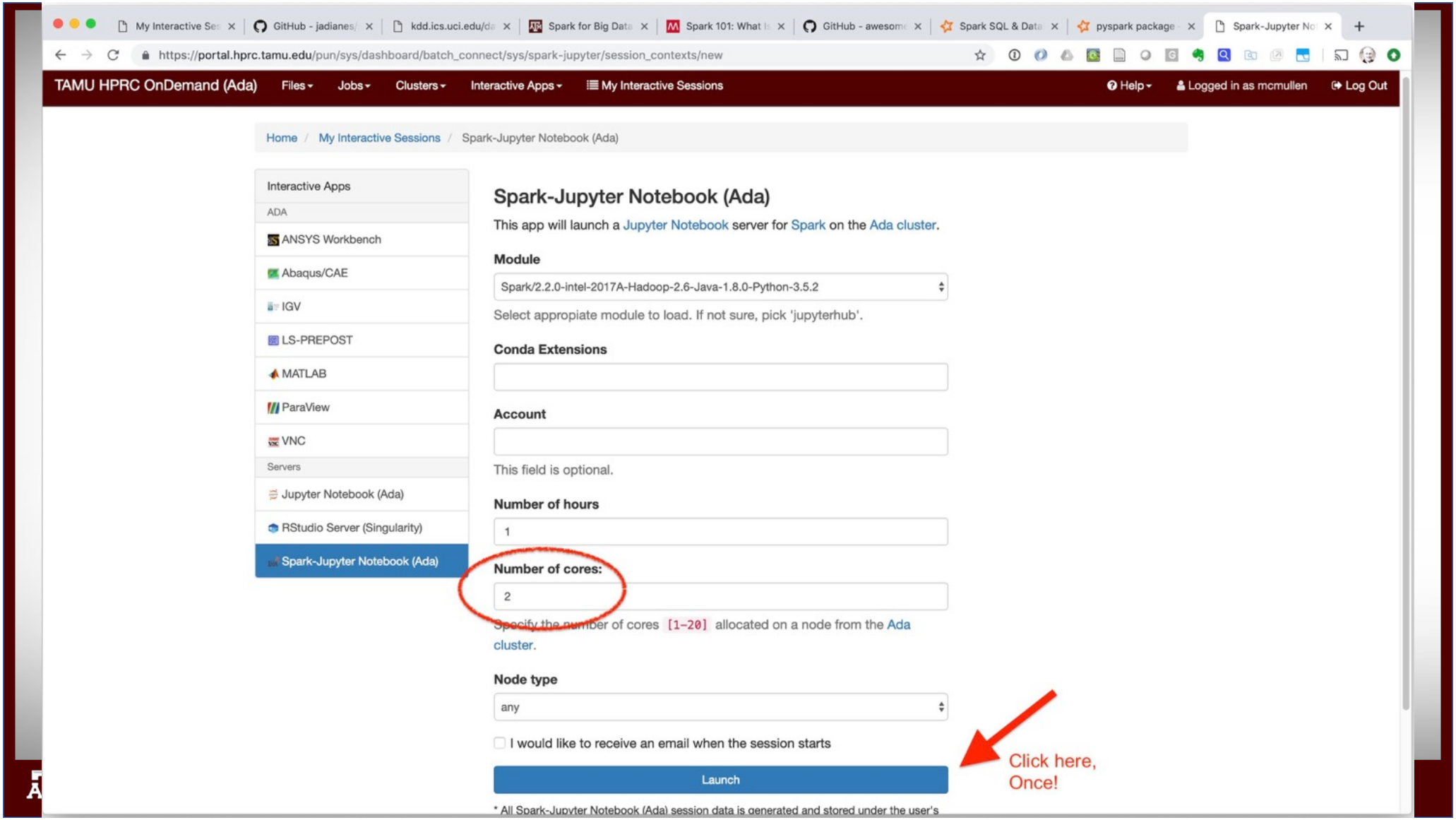
Message of the Day

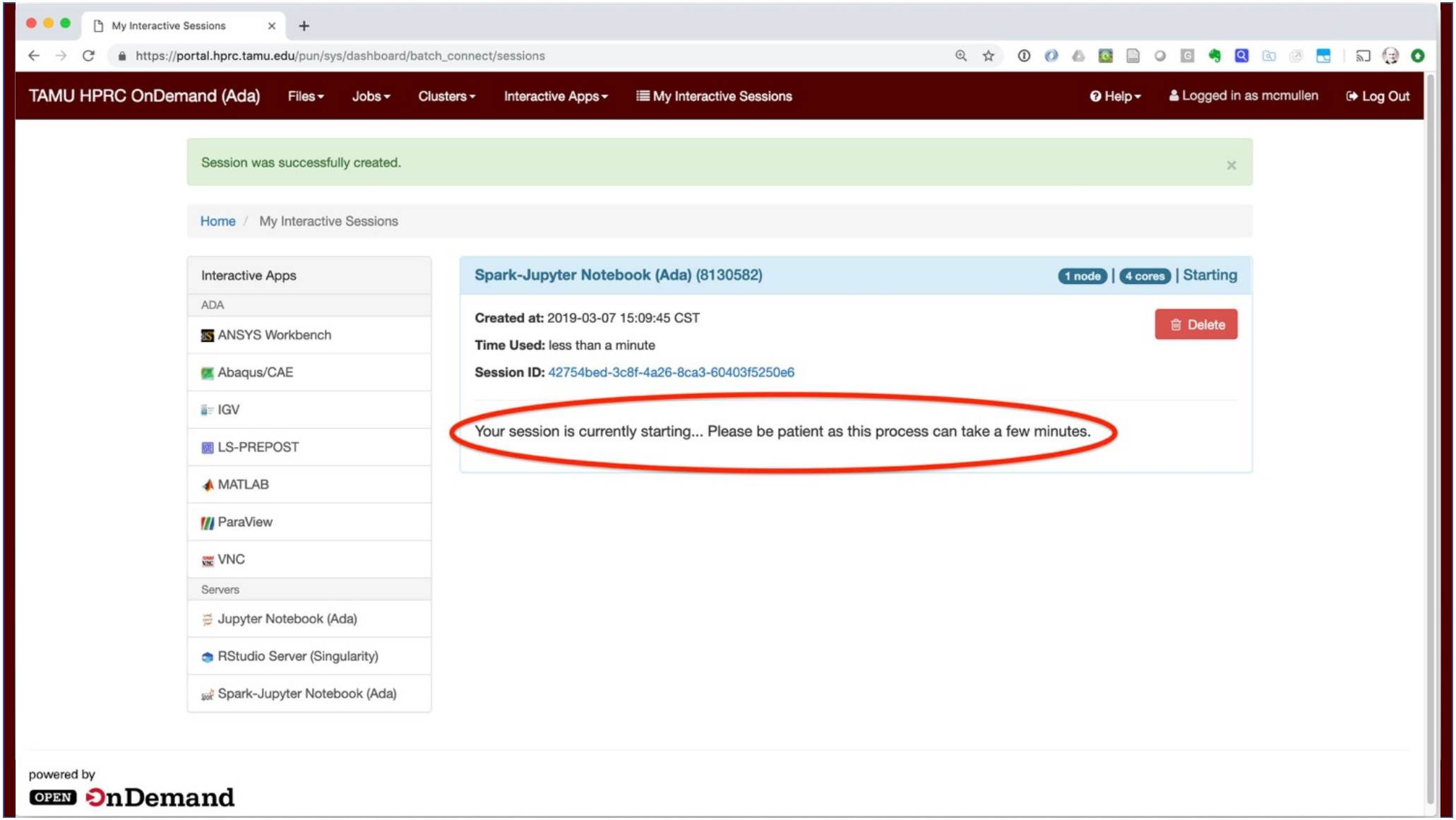
IMPORTANT POLICY INFORMATION

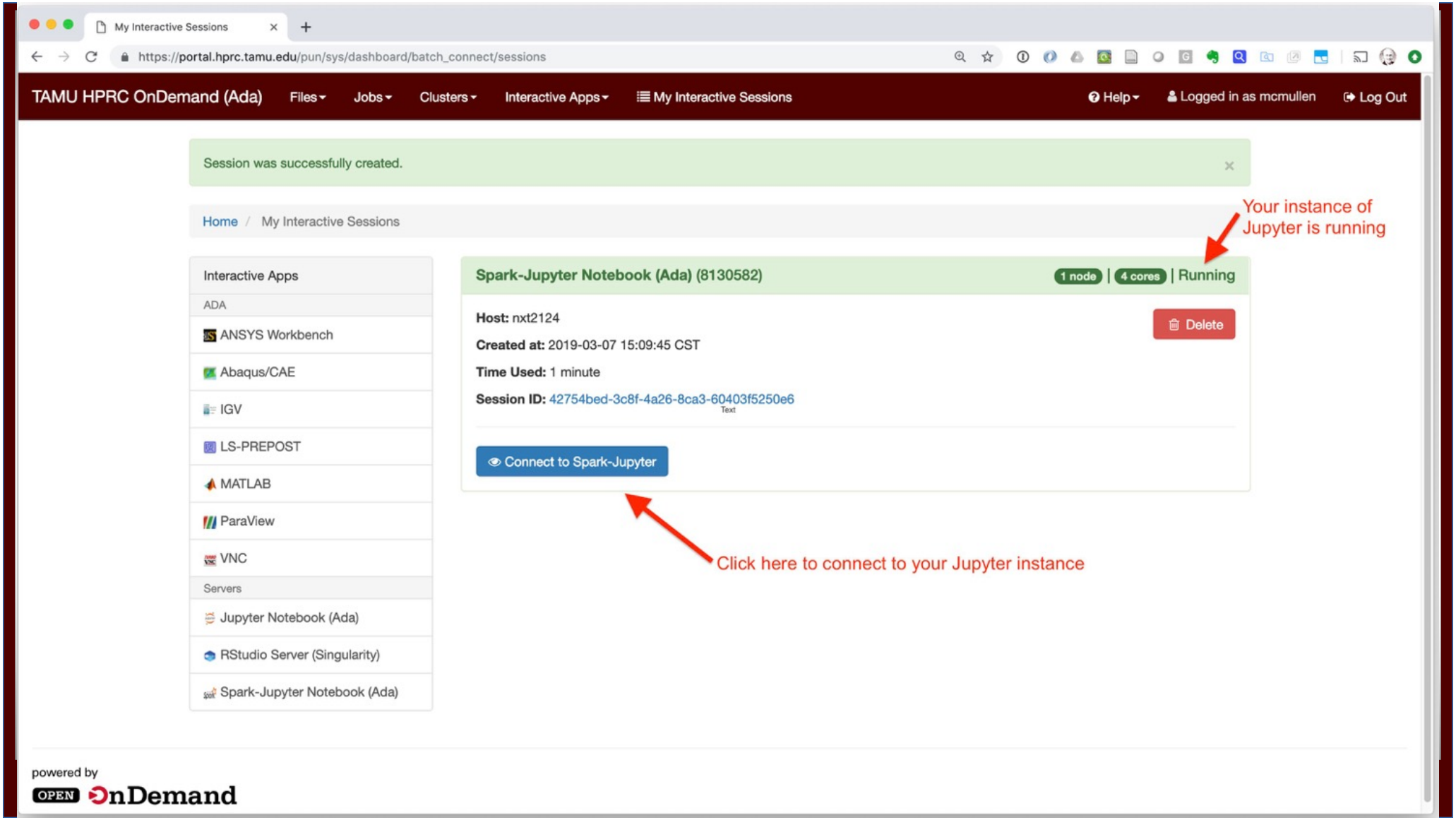
- **Unauthorized use of HPRC resources is pr…ecution.**
- **Use of HPRC resources in violation of Uni…gulations is prohibited. Current HPRC staff members are US citizens and legal residents.**
- **Sharing HPRC account and password info…ny shared accounts will be DISABLED.**
- **Authorized users must also adhere to ALL policies at: https://hprc.tamu.edu/policies**

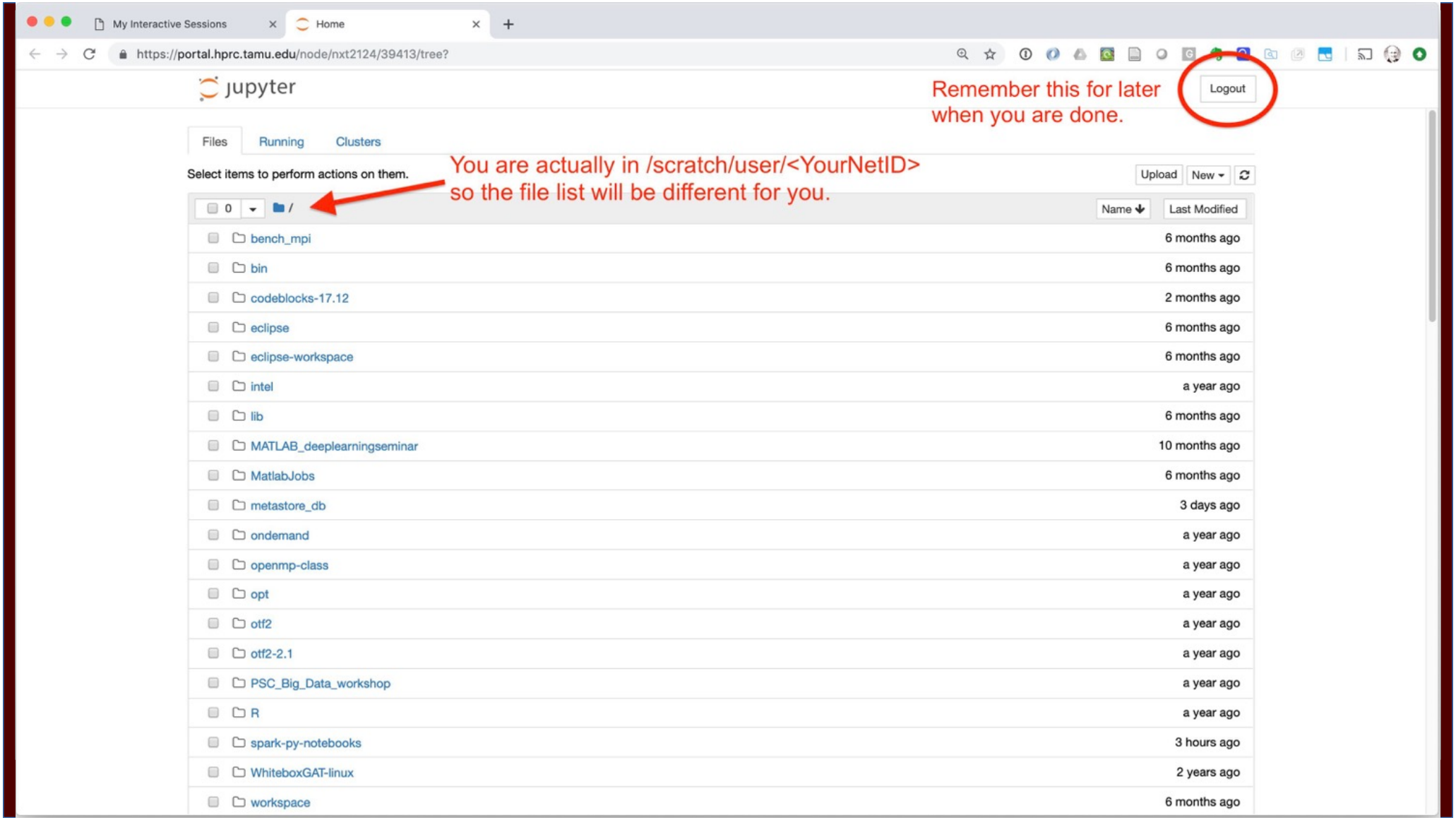
!! WARNING: There are NO active backups of user data. !!

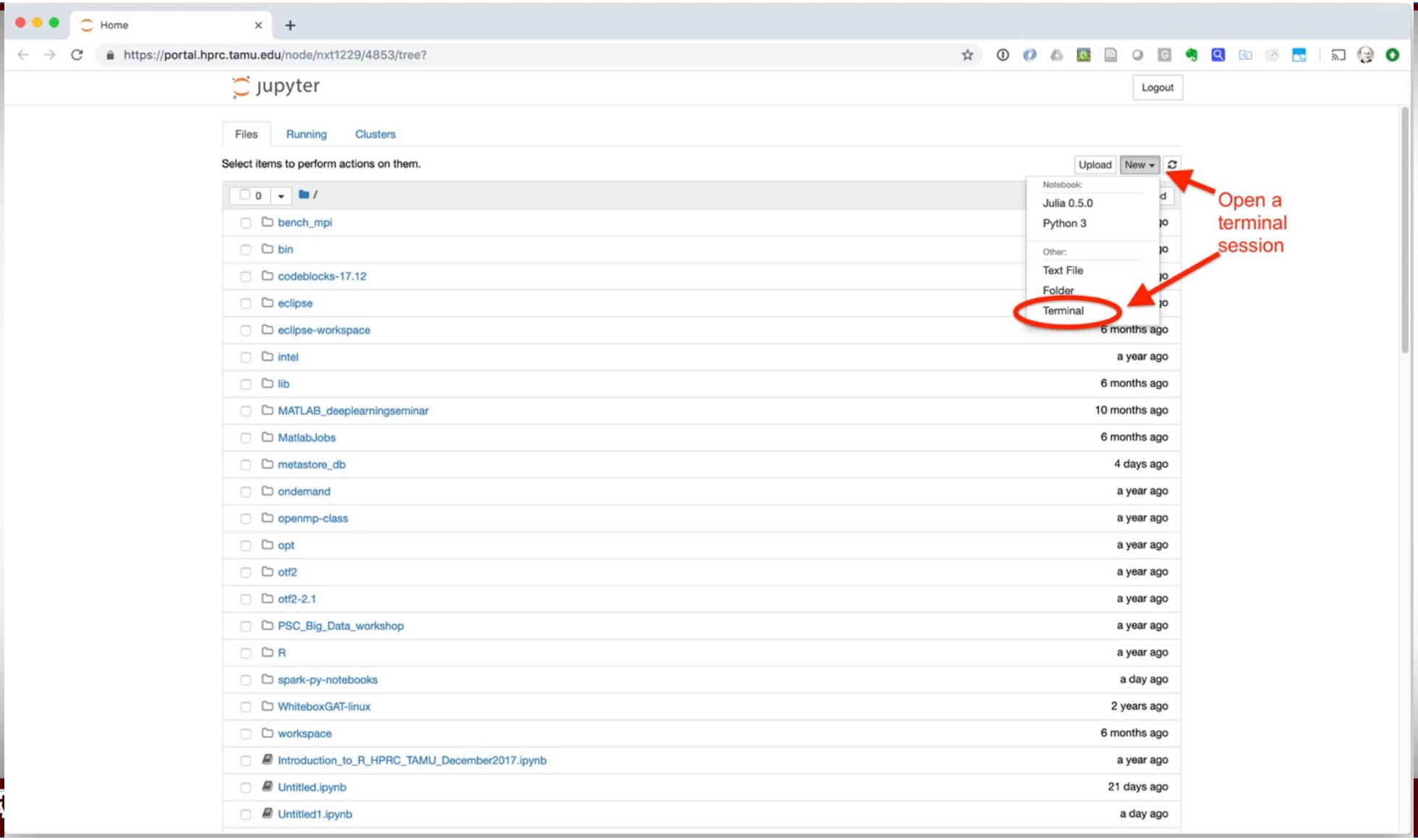
powered by OPEN OnDemand


My Interactive Ses
GitHub - jadianes/
kdd.ics.uci.edu/da
Spark for Big Data
Spark 101: What Is
GitHub - awesome
Spark SQL & Data
pyspark package
Spark-Jupyter No
https://portal.hprc.tamu.edu/pun/sys/dashboard/batch_connect/sys/spark-jupyter/session_contexts/new
TAMU HPRC OnDemand (Ada)
Files
Jobs
Clusters
Interactive Apps
My Interactive Sessions
Help
Logged in as mcmullen
Log Out
Home / My Interactive Sessions / Spark-Jupyter Notebook (Ada)
Interactive Apps
ADA
ANSYS Workbench
Abaqus/CAE
IGV
LS-PREPOST
MATLAB
ParaView
VNC
Servers
Jupyter Notebook (Ada)
RStudio Server (Singularity)
Spark-Jupyter Notebook (Ada)
Spark-Jupyter Notebook (Ada)
This app will launch a Jupyter Notebook server for Spark on the Ada cluster.
Module
Spark/2.2.0-intel-2017A-Hadoop-2.6-Java-1.8.0-Python-3.5.2
Select appropiate module to load. If not sure, pick 'jupyterhub'.
Conda Extensions
Account
This field is optional.
Number of hours
1
Number of cores:
2
Specify the number of cores [1-20] allocated on a node from the Ada cluster.
Node type
any
I would like to receive an email when the session starts
Launch
Click here, Once!
* All Spark-Jupyter Notebook (Ada) session data is generated and stored under the user's


My Interactive Sessions
https://portal.hprc.tamu.edu/pun/sys/dashboard/batch_connect/sessions
TAMU HPRC OnDemand (Ada)
Files
Jobs
Clusters
Interactive Apps
My Interactive Sessions
Help
Logged in as mcmullen
Log Out
Session was successfully created.
Home / My Interactive Sessions
Interactive Apps
ADA
ANSYS Workbench
Abaqus/CAE
IGV
LS-PREPOST
MATLAB
ParaView
VNC
Servers
Jupyter Notebook (Ada)
RStudio Server (Singularity)
Spark-Jupyter Notebook (Ada)
Spark-Jupyter Notebook (Ada) (8130582)
1 node | 4 cores | Starting
Created at: 2019-03-07 15:09:45 CST
Time Used: less than a minute
Session ID: 42754bed-3c8f-4a26-8ca3-60403f5250e6
Delete
Your session is currently starting... Please be patient as this process can take a few minutes.
powered by
OPEN OnDemand


My Interactive Sessions
https://portal.hprc.tamu.edu/pun/sys/dashboard/batch_connect/sessions
TAMU HPRC OnDemand (Ada)
Files
Jobs
Clusters
Interactive Apps
My Interactive Sessions
Help
Logged in as mcmullen
Log Out
Session was successfully created.
Home / My Interactive Sessions
Your instance of Jupyter is running
Interactive Apps
ADA
ANSYS Workbench
Abaqus/CAE
IGV
LS-PREPOST
MATLAB
ParaView
VNC
Servers
Jupyter Notebook (Ada)
RStudio Server (Singularity)
Spark-Jupyter Notebook (Ada)
Spark-Jupyter Notebook (Ada) (8130582)
1 node | 4 cores | Running
Host: nxt2124
Created at: 2019-03-07 15:09:45 CST
Time Used: 1 minute
Session ID: 42754bed-3c8f-4a26-8ca3-60403f5250e6
Text
Delete
Connect to Spark-Jupyter
Click here to connect to your Jupyter instance
powered by
OPEN OnDemand


My Interactive Sessions
Home
https://portal.hprc.tamu.edu/node/nxt2124/39413/tree?
jupyter
Logout
Remember this for later when you are done.
Files
Running
Clusters
Select items to perform actions on them.
You are actually in /scratch/user/<YourNetID> so the file list will be different for you.
Upload
New
0
/
Name
Last Modified
bench_mpi 6 months ago
bin 6 months ago
codeblocks-17.12 2 months ago
eclipse 6 months ago
eclipse-workspace 6 months ago
intel a year ago
lib 6 months ago
MATLAB_deeplearningseminar 10 months ago
MatlabJobs 6 months ago
metastore_db 3 days ago
ondemand a year ago
openmp-class a year ago
opt a year ago
otf2 a year ago
otf2-2.1 a year ago
PSC_Big_Data_workshop a year ago
R a year ago
spark-py-notebooks 3 hours ago
WhiteboxGAT-linux 2 years ago
workspace 6 months ago


Home
https://portal.hprc.tamu.edu/node/nxt1229/4853/tree?
jupyter
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
Notebook:
Julia 0.5.0
Python 3
Other:
Text File
Folder
Terminal
Open a terminal session
bench_mpi
bin
codeblocks-17.12
eclipse
eclipse-workspace 6 months ago
intel a year ago
lib 6 months ago
MATLAB_deeplearningseminar 10 months ago
MatlabJobs 6 months ago
metastore_db 4 days ago
ondemand a year ago
openmp-class a year ago
opt a year ago
otf2 a year ago
otf2-2.1 a year ago
PSC_Big_Data_workshop a year ago
R a year ago
spark-py-notebooks a day ago
WhiteboxGAT-linux 2 years ago
workspace 6 months ago
Introduction_to_R_HPRC_TAMU_December2017.ipynb a year ago
Untitled.ipynb 21 days ago
Untitled1.ipynb a day ago

In the terminal window type:

```
(.venv) cp /scratch/user/mcmullen/sparkfiles.tgz .
```

There is a period '.' after "tgz". Then type

```
(.venv) tar xzvf sparkfiles.tgz
(.venv) ls -l sparkfiles
(.venv) exit
```

This unpacks the archive in your directory. "`ls`" should show a new directory "sparkfiles".

Time to start doing something with Spark

- In the Jupyter file navigator window, go into the sparkfiles directory
- Click on “spark_basic.ipynb”

Next fun thing: Shakespeare!

- While in the sparkfiles directory, go into "Shakespeare" directory
- Click on "shakespeare.ipynb"

More fun, if there is time:

- While in the sparkfiles directory, go into “spark-py-notebooks/nb1-rdd-creation” directory
- These are a collection of exercises demonstrating features of Spark by Jose A. Dianes, a data scientist in the UK
- https://www.linkedin.com/in/jadianes/?originalSubdomain=uk
- Click on “nb1-rdd-creation.ipynb”
- There are other exercises in nb2 to nb10. Some are not correct for python3 and spark 2.2.0

Contact the HPRC Helpdesk

Website:

hprc.tamu.edu

Email:

help@hprc.tamu.edu

Help us, help you -- we need more info

- Which Cluster
- UserID/NetID
- Job id(s) if any
- Location of your jobfile, input/output files
- Application used if any
- Module(s) loaded if any
- Error messages
- Steps you have taken, so we can reproduce the problem

https://hprc.tamu.edu
help@hprc.tamu.edu

Thanks!

Questions?