Big Data Workshop FAQ

Workshop Questions

- Recordings - will a recording of this session be available?
  The XSEDE Training youtube channel has recent recordings of the Big Data workshop, it may not be this current workshop but it is the same content. A playlist of these recordings can be found here: https://www.youtube.com/c/XSEDETraining/playlists

- Slides - are the slides from today's talks available?
  Slides can be found linked on the Big Data Workshop agenda page. A link to the current workshop can be found on this page:
  https://www.psc.edu/resources/training/hpc-workshop-series/xsede-hpc-workshop-big-data-december-7-8-2021/

- Setup - I came in late/got logged out, how do I get set up again?
  Here are the commands you need to get started:

```
ssh -l MYUSERNAME bridges2.psc.edu # (substituting MYUSERNAME.)
# (You may be connecting via another method like putty)
~training/Setup # (copies files into ~/BigData for you)
interact # (connects you to a job on a compute node)
module load spark
cd BigData/Shakespeare # (or the appropriate directory)
pyspark
```

Spark Technical Questions

- py4j errors
  If you are getting py4j errors usually this means either you have a typo in your file name or you're not in the right directory. You don't get an error when you run the sc.textFile() because it doesn't attempt to read the file until you perform an action later in your code (usually rdd.count()). The solution is to make sure you're in the right directory (usually ~/BigData/Shakespeare) and that you have spelled the file name correctly.

- sc not defined
  These errors usually occur when you attempt to run spark on the login node instead of a compute node. Exit pyspark and run

```
interact # wait for session to start
module load spark
pyspark
```

- standalone script
  How do I run spark programs without the pyspark shell?
  You need to set up the spark context yourself in your code, in the shell this is done for you automatically. That means adding in the following lines to your python script:

  ```
  from pyspark import SparkContext
  sc=SparkContext()
  ```

- batch mode/multi-node spark
  Please see the Bridges2 User Guide for information about running Spark jobs in non-interactive mode. https://www.psc.edu/resources/software/spark

Account Issues

- My password doesn't work
  You can reset your password at https://apr.psc.edu/

- I don't have an account on Bridges-2
  Check the materials you received via email before the workshop for information on how to log in to Bridges-2. If you still have questions contact tmaiden@psc.edu

- How can I compute after the workshop?
  The allocation for the workshop will be available until it is exhausted, which could be a matter of days or weeks depending on how judiciously those resources are used. Instructions for an easy application for a startup grant are here: https://www.psc.edu/resources/allocations/apply-for-an-xsede-startup-grant/