**Spark-wordcount TIPS**

Table of Contents

[Install findspark on all instances! 1](#_Toc526796346)

[Modify the directory for your Anaconda libraries 1](#_Toc526796347)

[Modify the location of your Shakespeare document from previous labs. 3](#_Toc526796348)

First, you’ll want to review the spark-wordcount.ipynb file located in WorldClass. Download this juniper notebook and upload it into Junyper Notebooks. This notebook will walk you through how the work count map reduce process can be done in Spark. This will give you valuable insite for your inverted index project!

The following notes can be used as a guide to working through the spark-wordcount.ipynb notebook.

# Install findspark on all instances!

Before running the first cell, you need to install findspark on all VMs in your cluser.

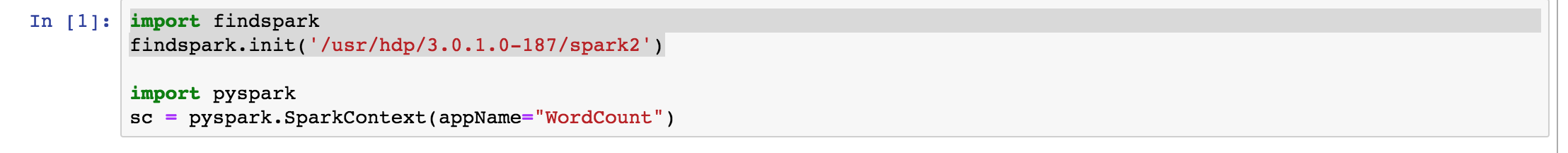
[root@c7401 ~]# pip install --upgrade pip

[root@c7401 ~]# pip install findspark

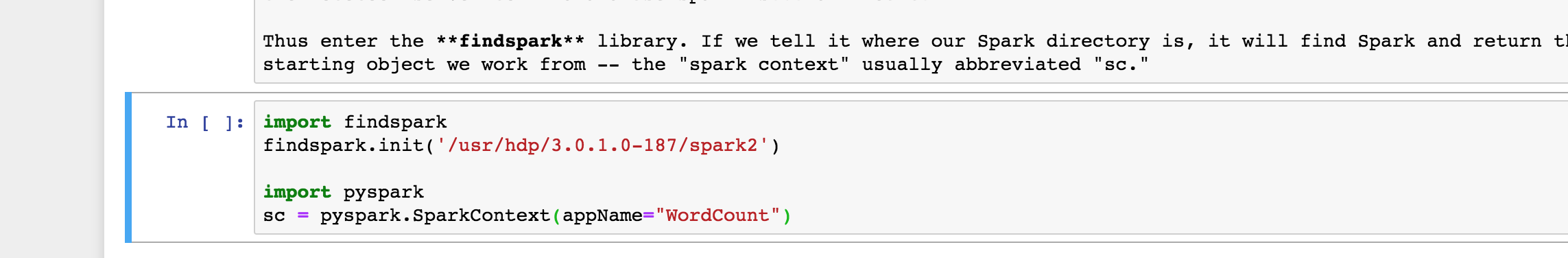
Ensure you do this step for all VMs!

# Modify the directory for your Anaconda libraries

Check your directory location on c7401 and update the following file path.



Now run the first cell. The keyboard shortcut to run the current cell is shift+enter

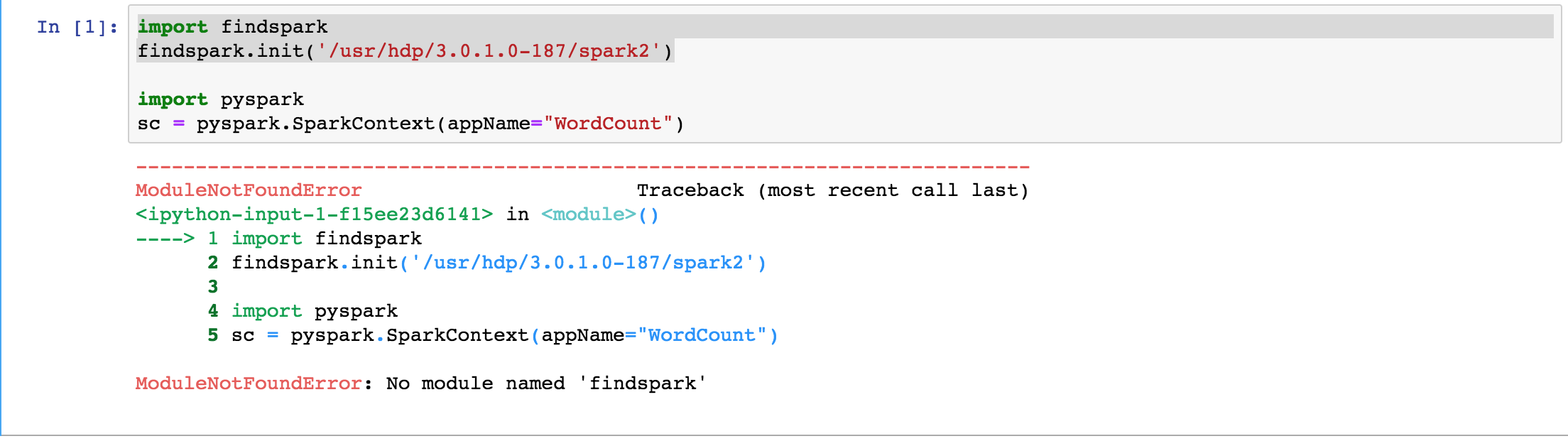


**Troubleshooting tips** if the above cell does not run successfully:

1. Ensure the following path was added to the Vagrant user after lab 5 (Anaconda install)

export PATH="/opt/anaconda3/bin:$PATH"

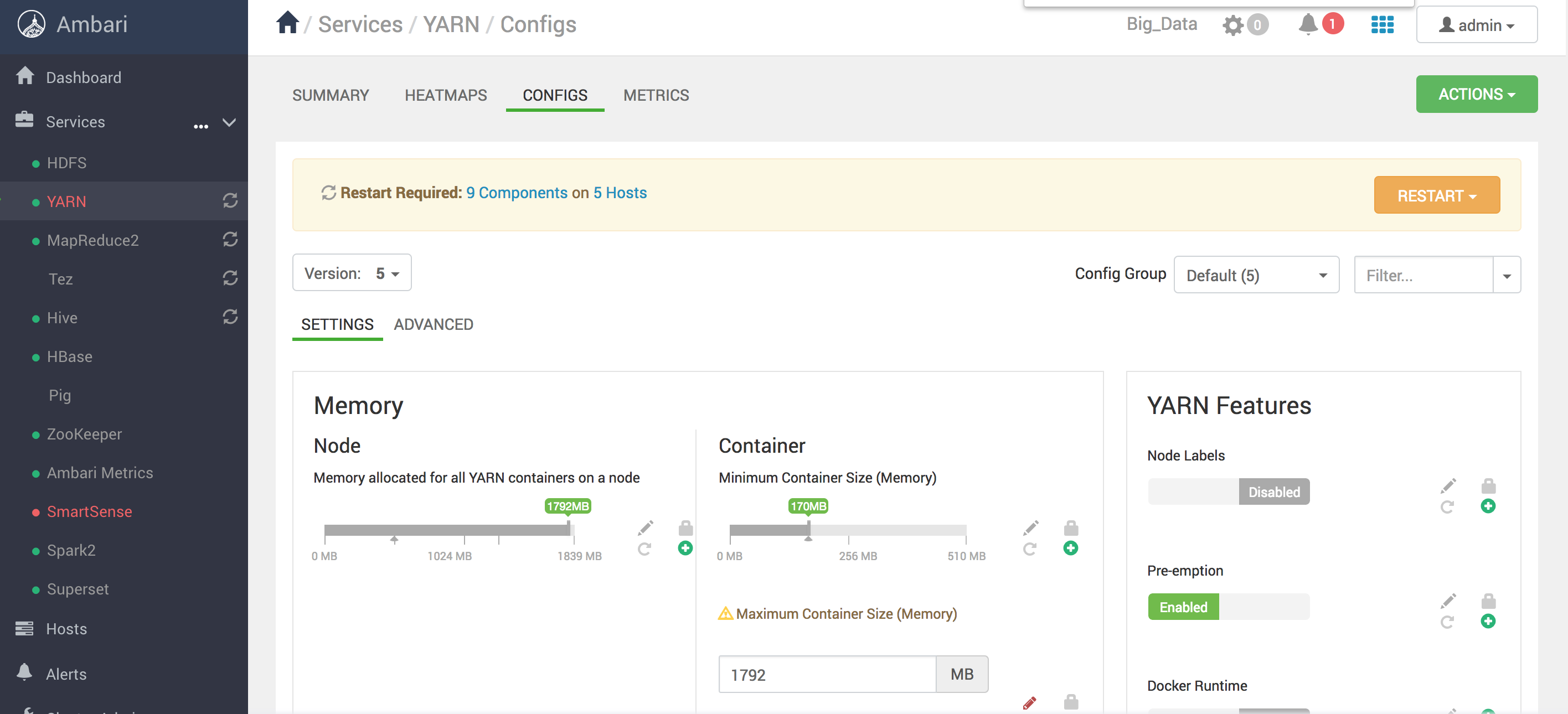
1. If you get the following error when running the first cell, you may have forgotten to install findspark. (see above)



1. If you receive the following error when creating the spark context (cell 1), and then make the following memory modifications in Ambari.



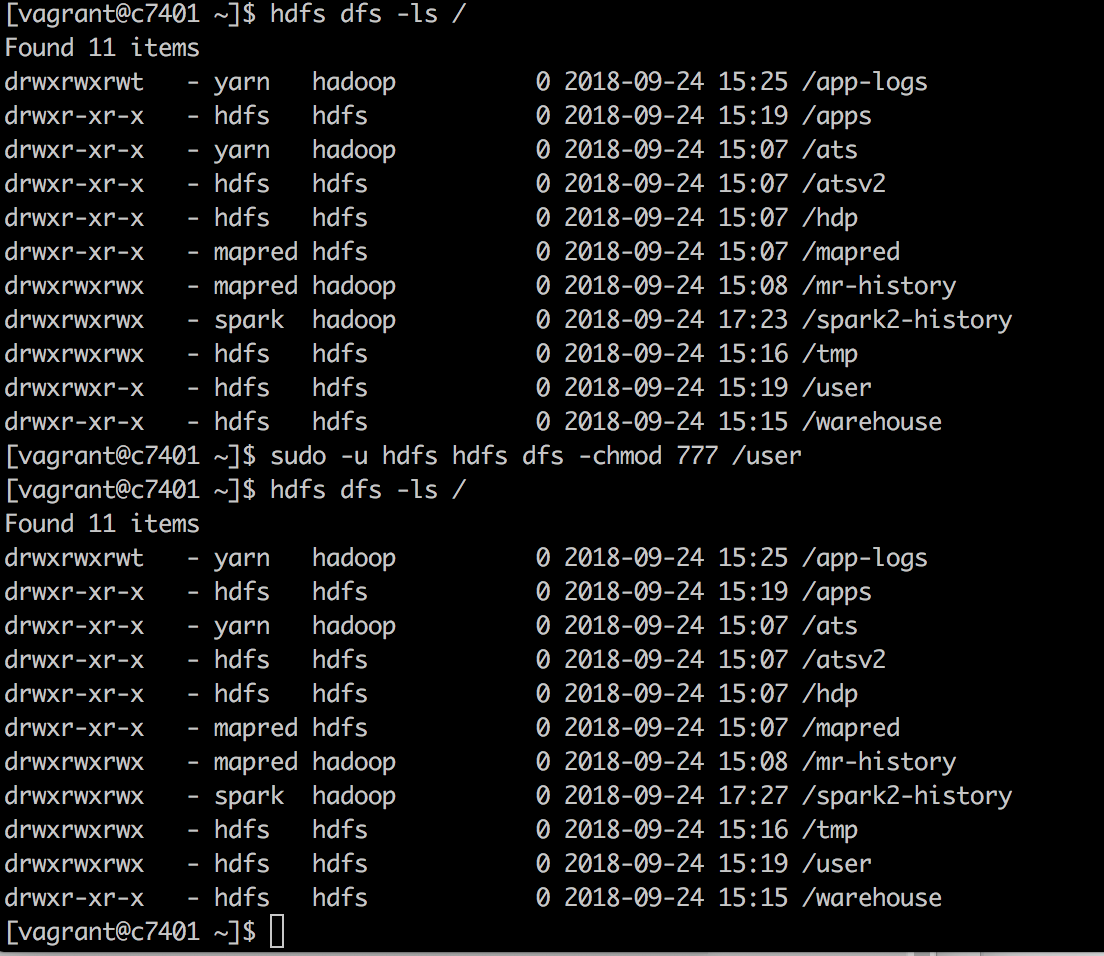
Modify the Yarn memory settings as follows.



1. If you receive the Permission denied error below, you need to modify the permissions for the /user directory within hdfs.



Change the file permissions on the /user directory in hdfs as shown below. (this should have been done in lab 4 so you may not have this issue)



# Modify the location of your Shakespeare document from previous labs.

Before running the second cell, you need to modify the location of the shakespear.txt file that you have used in previous labs.

