Spark Fundamentals - II

Setting up and verifying the lab environment

Contents

Setting up and verifying the lab environment

The labs for this course will be using Zeppelin, a web-based notebook that allows interactive data analysis. The notebooks have been loaded into the docker image, so when you start it up, everything is ready to go. This section will make sure you have everything that you need for the labs throughout this course. The first tutorial that you will go through will be a basic overview of the Zeppelin environment and run some Spark tasks to ensure that the Spark cluster is up and running.

After completing this hands-on lab, you should be able to:

- Use Zeppelin notebooks for the lab exercises.
- Test Spark on the cluster without any errors to verify the lab environment.

Allow 15 minutes to complete this section of lab.

1.1 Setting up the lab environment

Be sure to have downloaded and set up your docker environment prior to starting this lab. The information is on the Big Data University course page. Once you have the docker environment ready, you can use the instructions here to pull the docker to your local system.

https://registry.hub.docker.com/u/bigdatauniversity/spark2/

Note: If you are using boot2docker on windows, the command to start the Spark container is slightly different.

```
docker run -it --name bdu_spark2 -P -p 4040:4040 -p 4041:4041 -p 8080:8080 -p
8081:8081 bigdatauniversity/spark2:latest /etc/bootstrap.sh -bash
```

Note 2: You can SSH into the docker image in order to start the container using Putty. However, you must configure this prior: https://docs.docker.com/installation/windows/#login-with-putty-instead-of-using-the-cmd

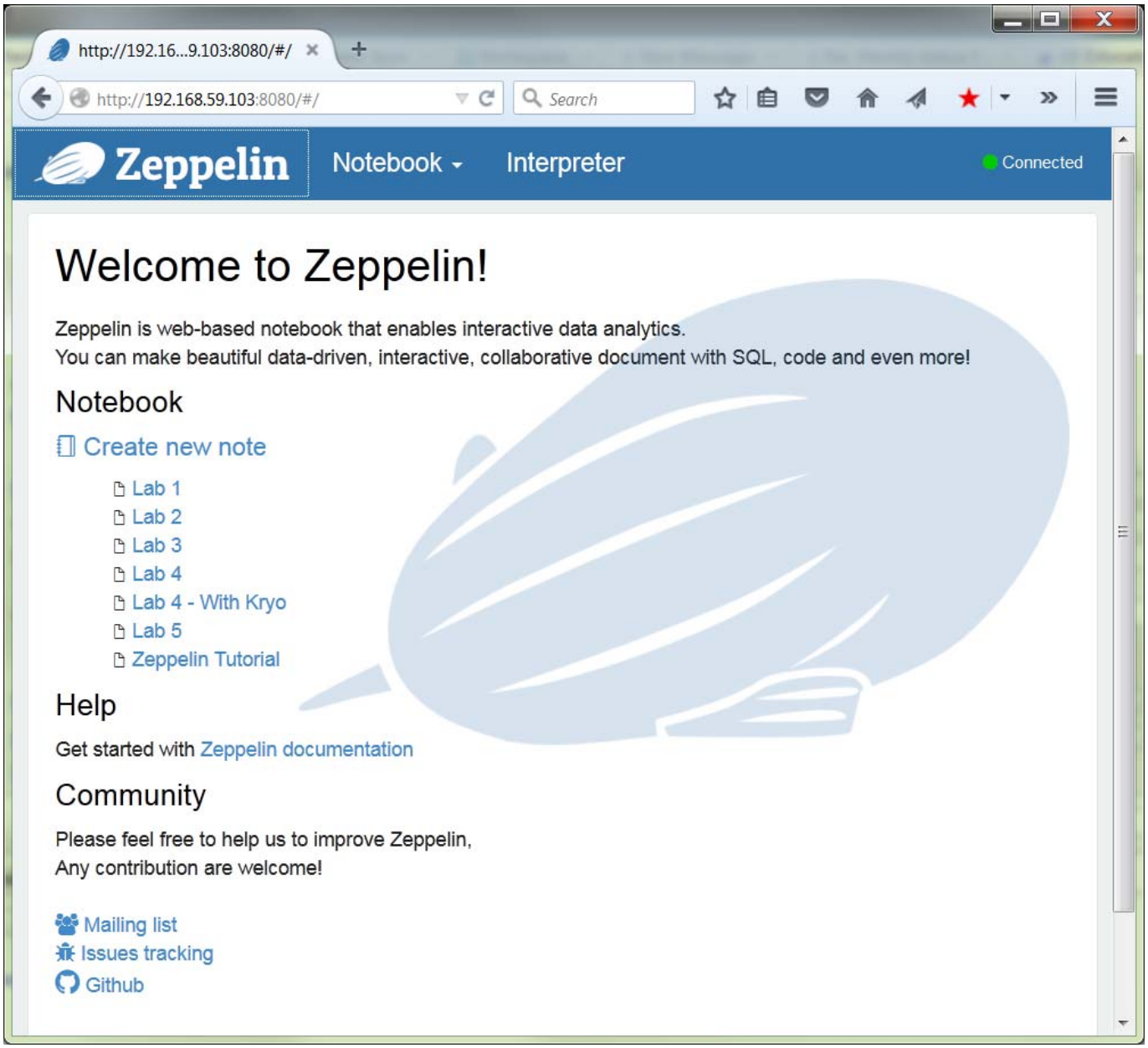
__1. Once you have started the Spark container, you should see that all the required services have started:

```
docker@boot2docker:~$ docker run -it --name bdu_spark2 -P -p 8080:8080 -p 8081:8
081 bigdatauniversity/spark2:latest
starting namenode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-namenode-b95f0ebb
4856.out
Started Hadoop namenode:                                   [  OK  ]
starting datanode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-datanode-b95f0ebb
4856.out
Started Hadoop datanode (hadoop-hdfs-datanode):            [  OK  ]
starting resourcemanager, logging to /var/log/hadoop-yarn/yarn-yarn-resourcemana
ger-b95f0ebb4856.out
Started Hadoop resourcemanager:                            [  OK  ]
starting nodemanager, logging to /var/log/hadoop-yarn/yarn-yarn-nodemanager-b95f
0ebb4856.out
Started Hadoop nodemanager:                                [  OK  ]
starting org.apache.spark.deploy.history.HistoryServer, logging to /usr/local/sp
ark-1.2.1-bin-hadoop2.4/sbin/../logs/spark-student-org.apache.spark.deploy.histo
ry.HistoryServer-1-b95f0ebb4856.out
Zeppelin start                                             [  OK  ]
```

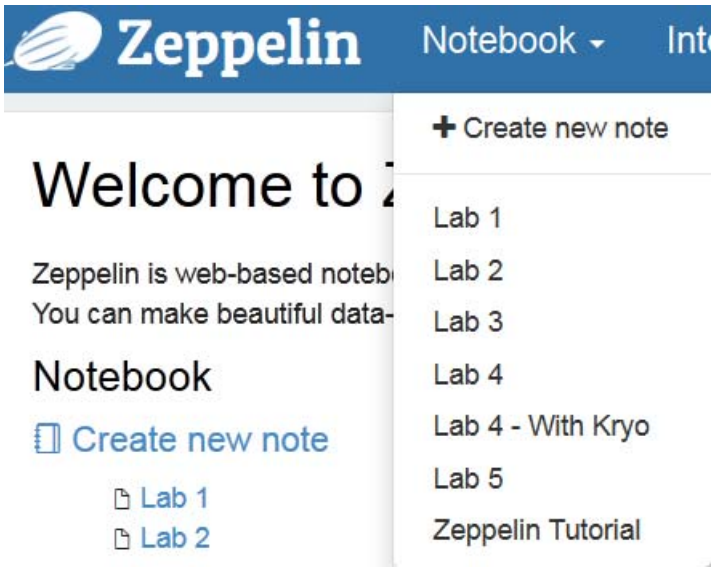
Note that Zeppelin has started. You are ready to go.

__2. Open up a web browser on your machine. Navigate to Zeppelin using this URL:

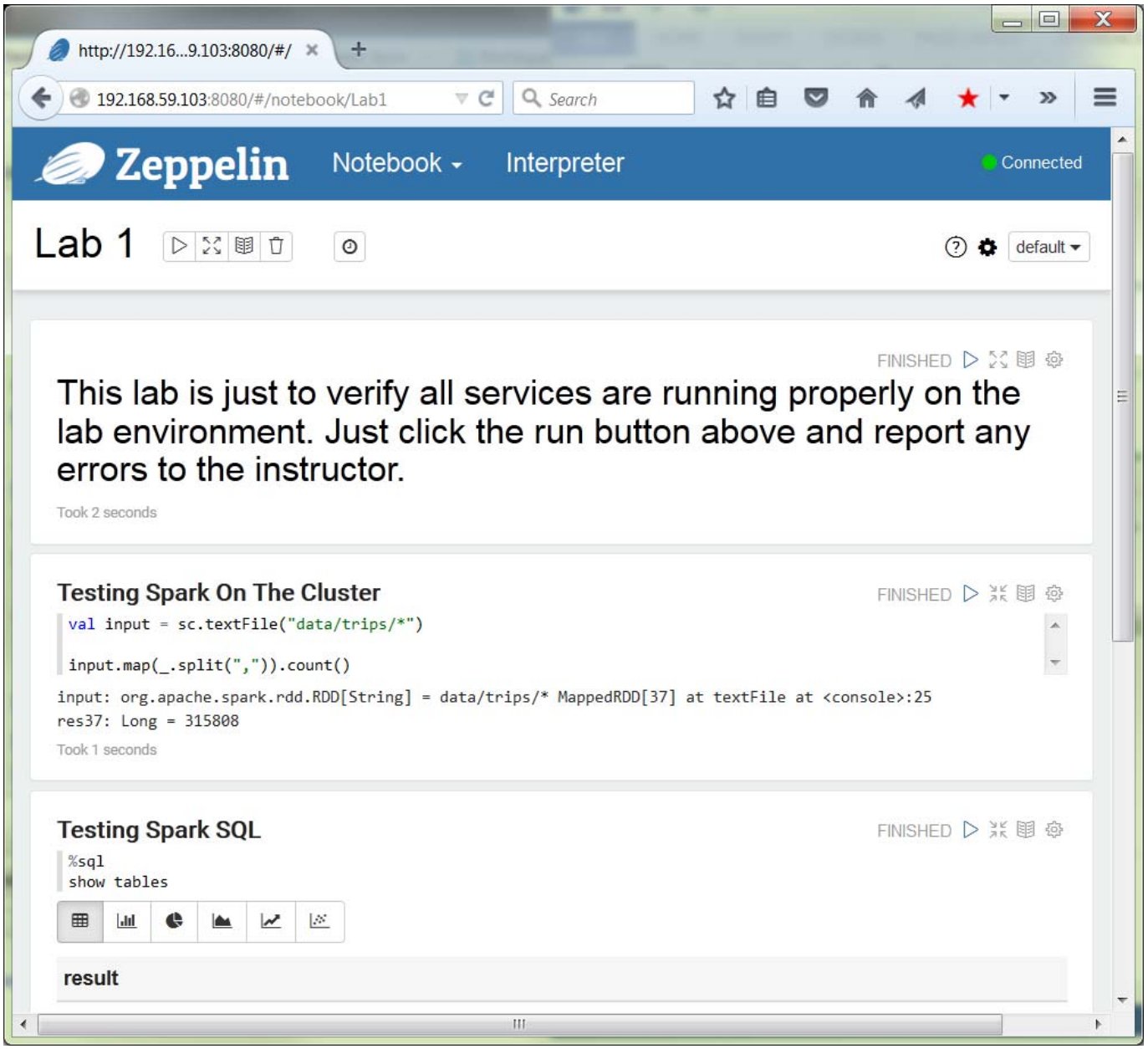
http://192.168.59.103:8080



__3. To navigate the Zeppelin notebook, you can click on the Notebook dropdown menu:



__4. All of your labs will be performed in the Zeppelin notebook, with the exception of lab 5, which is an optional labs for those that would like to use various IDEs to run Spark jobs. Click the **Lab 1** notebook to get started.



There are three tasks on this page, however for this course, you will only need to test that Spark is on the cluster. There will not be any exercises on Spark SQL and or any use of the Cassandra database, so those are not included in your docker container to keep it lightweight. Click on the **Play** button to run the **Testing Spark on the Cluster** task.

__5. You should see something similar to this when it completes:

__6. The button to the right of the play button toggles source display of the panel. Click on that button to **show/hide the Spark code**.

__7. The button that looks like a book toggles the output display of the button. Click that to **show/hide the results of the Spark tasks**.

__8. Finally, the right-most button is the panel settings. Click on it to see some of the **settings** for that particular panel.

Essentially, you can type in your own Spark commands inside the source area and click run. The entire notebook has been created for the lab exercises throughout this course. You will be asked to fill in missing source code for future labs to test your understanding of the concepts. Solutions are provided for the tasks that you are required to complete.

Summary

Having completed this exercise, you should now be able to access the Zeppelin notebook and run the task in Lab 1 without any errors. This concludes this lab.

NOTES

NOTES

The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, these materials. Nothing contained in these materials is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software. References in these materials to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. This information is based on current IBM product plans and strategy, which are subject to change by IBM without notice. Product release dates and/or capabilities referenced in these materials may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way.