ITECH2302 Big Data Management Laboratory – Data Mining 1

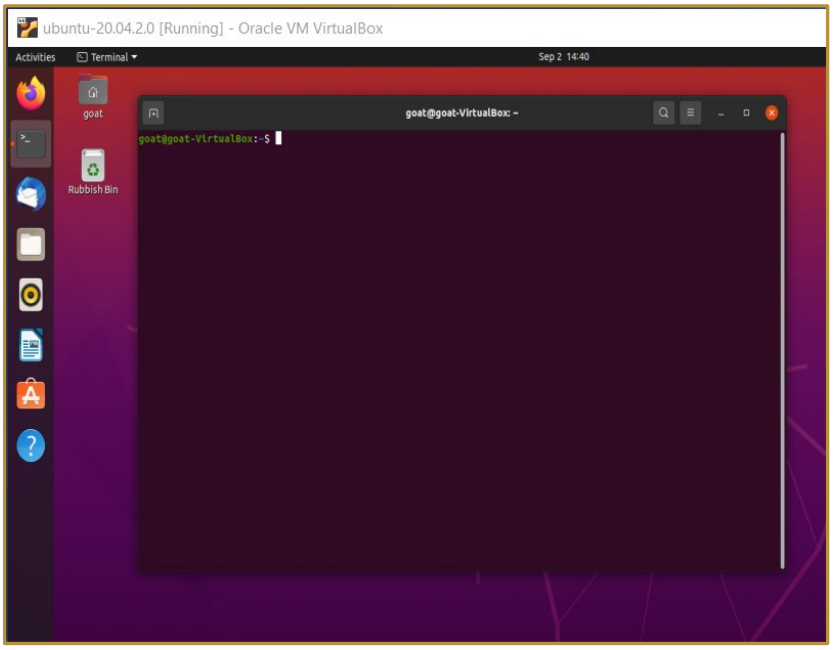
Objectives:

- Introduction to Spark MLlib using pyspark
- Introduction to scikit-learn and python

Activity 1

Apache Spark

1. Start Apache Hadoop



Write the following commands to open up python notebooks:
$ pyspark

Firefox should open at the following URL. This is where you can upload your python notebooks (with extension .pynb)
It is also possible to start the dashboard on any system via the command prompt (or terminal on Unix systems) by entering the command jupyter notebook; in this case, the current working directory will be the start-up directory. With Jupyter Notebook open in your browser, you may have noticed that the URL for the dashboard is something like https://localhost:8888/tree. Localhost is not a website, but indicates that the content is being served from your *local* machine: your own computer.
Jupyter's Notebooks and dashboard are web apps, and Jupyter starts up a local Python server to serve these apps to your web browser, making it essentially platform-independent and opening the door to easier sharing on the web.

Run and examine the following ipynb files in the spark directory:

The following notebooks come from https://github.com/jadianes/spark-py-notebooks and can be found in this folder: /home/goat/hadoop_spark/hadoop/lab_data/ spark-py-notebooks-master

Work your way through the notebooks:

- nb4-rdd-set
- nb5-rdd-aggregations
- nb6-rdd-key-value
- nb7-mllib-statistics

Activity 2

Spark ML/MLlib:

Review the material about Spark ML libraries:

- https://spark.apache.org/docs/latest/ml-guide.html

Activity 3

scikit-learn:

Review the material about scikit-learn algorithms:

- https://scikit-learn.org/stable/modules/svm.html
- https://scikit-learn.org/stable/modules/clustering.html
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
- https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html