Work Integrated Learning Programmes

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Computer Science & Information Systems

Big Data Systems – Spark Lab Sheet 5

Spark MLLib

1. Objective:

Students should be able to

A. Get familiarity with the Spark MLLib module

B. Get hands-on experience with Machine Learning programme development and execution

This lab sheet provides a quick introduction of using Spark for Machine Learning applications. This exercise will introduce the API through MLLib package for development of classical regression model.

MLlib is Spark's machine learning (ML) library. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as:

- ML Algorithms: common learning algorithms such as classification, regression, clustering, and collaborative filtering
- Featurization: feature extraction, transformation, dimensionality reduction, and selection
- Pipelines: tools for constructing, evaluating, and tuning ML Pipelines
- Persistence: saving and load algorithms, models, and Pipelines
- Utilities: linear algebra, statistics, data handling, etc.

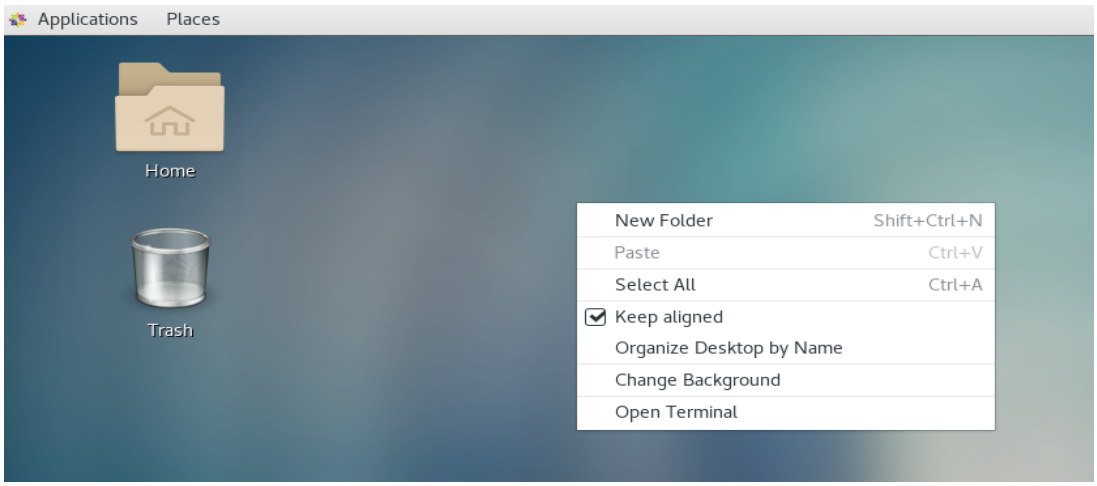
2. Steps to be performed:

Note - It's assumed that student has made a slot reservation using the slot booking interface where Apache Spark framework was selected. The details of the Apache Spark systems to be used is received through an email. If not, please contact the administrators for the same.

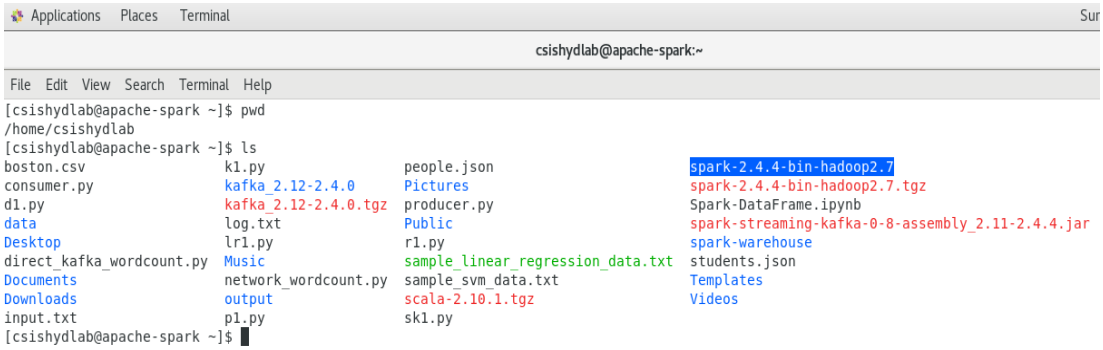
Also it's assumed that students are aware of the process of logging into these virtual machines. If not, then get access to the user manual maintained for the usage of remote lab setup.

Preparations -

a) Open the terminal by right clicking on the desktop of the virtual machine.



b) Look at the current directory and also file listings in it. It must have a spark installation directory present in it. Commands like pwd, ls can be used for it.



c) Set the SPARK_HOME and HOME variable to point to the spark installations.

[csishydlab@apache-spark bin]$ pwd

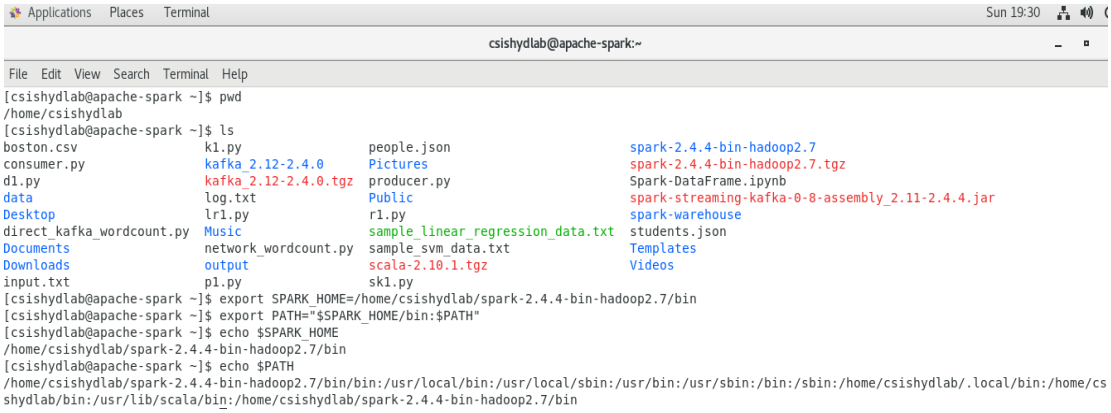
/home/csishydlab/spark-2.4.4-bin-hadoop2.7/bin

[csishydlab@apache-spark bin]$ export SPARK_HOME=/home/csishydlab/spark-2.4.4-bin-hadoop2.7/bin

[csishydlab@apache-spark bin]$ export PATH="$SPARK_HOME/bin:$PATH"

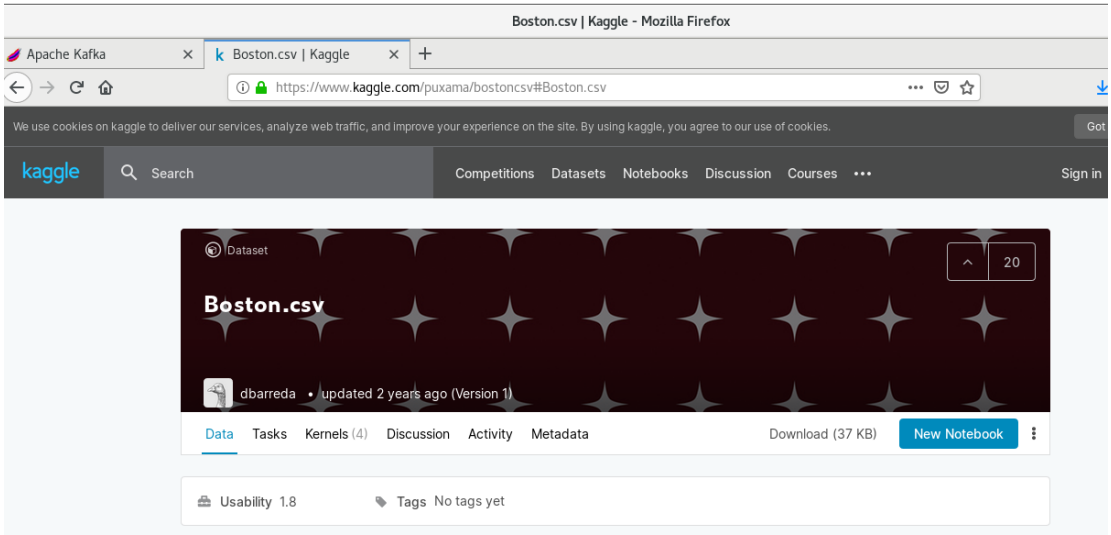
echo $SPARK_HOME

echo $PATH



d) Using the web browser , download the Boston.csv file and save it in the local file system.

https://www.kaggle.com/puxama/bostoncsv#Boston.csv



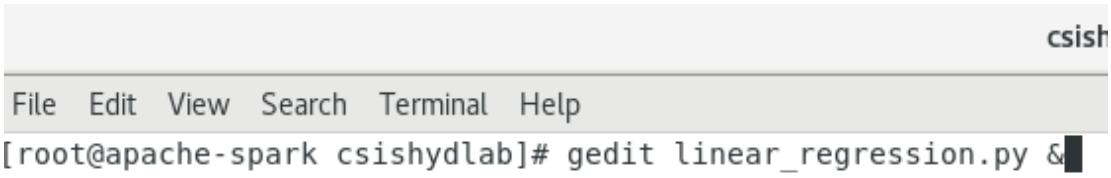
Installing pySpark

For the execution of python programmes on the Spark, a package named pyspark is required. Using the sudo previleges, install the packages with pip command.

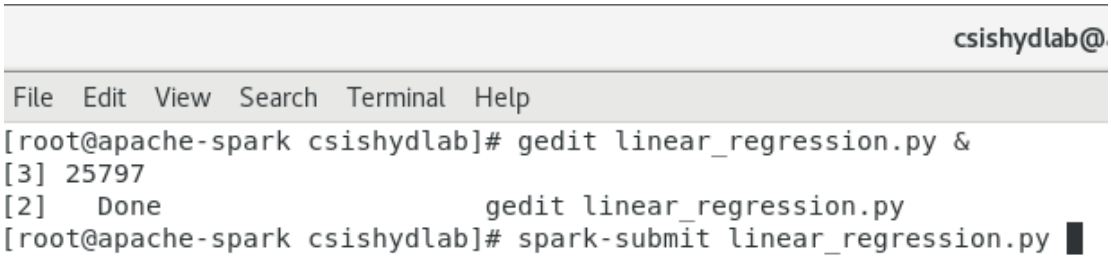
```
pip install pyspark
```

Writing Linear Regression Machine Learning programme

e) Open up the text editor and copy the code written in the attached linear_regression.py file.



f) Execute the linear_regression.py file using the spark-submit command.



g) Look at the outcome printed while the program is getting executed on the Spark cluster. It shows actual and predicted house prices.

```
+------------------+----+--------------------+
|        prediction|medv|            features|
+------------------+----+--------------------+
|27.927270908006726|22.0|[0.01096,55.0,2.2...|
|30.768510069596182|29.1|[0.01439,60.0,2.9...|
|26.497762769897413|23.1|[0.0187,85.0,4.15...|
| 40.29308710051661|50.0|[0.02009,95.0,2.6...|
|27.087495861316143|16.5|[0.02498,0.0,1.89...|
|28.147245589564633|23.9|[0.02543,55.0,3.7...|
|22.219573905760136|20.6|[0.03306,0.0,5.19...|
| 32.57075672757935|34.9|[0.03359,75.0,2.9...|
|20.383617525016447|19.5|[0.03427,0.0,5.19...|
|30.645810541006668|28.5|[0.03502,80.0,4.9...|
|27.895366964633475|22.0|[0.03537,34.0,6.0...|
|24.551339355753125|22.9|[0.03551,25.0,4.8...|
|28.756201579966294|27.9|[0.03615,80.0,4.9...|
|23.549490273145608|20.7|[0.03738,0.0,5.19...|
|  35.0073271643155|34.6|[0.03768,80.0,1.5...|
|36.411263074979686|33.3|[0.04011,80.0,1.5...|
| 26.88971216116669|22.9|[0.04203,28.0,15....|
| 24.69531323293066|20.6|[0.04294,28.0,15....|
|26.554206975416243|24.8|[0.04297,52.5,5.3...|
| 17.14499460979746|18.2|[0.04301,80.0,1.9...|
+------------------+----+--------------------+
only showing top 20 rows
```

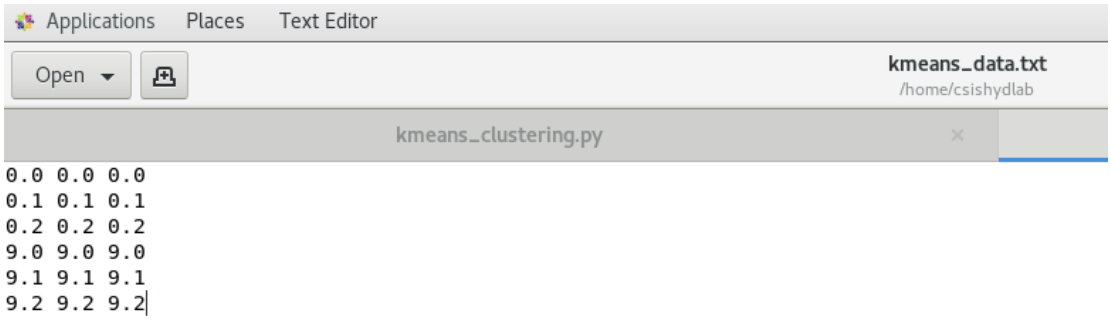
Installing numpy

h) For the execution of python programmes on the Spark, a package named numpy is required. Using the sudo previleges, install the packages with pip command.

pip install numpy

Writing k-means clustering Machine Learning programme

i) Open up the text editor and copy the data written in the attached kmeans_data.txt file.

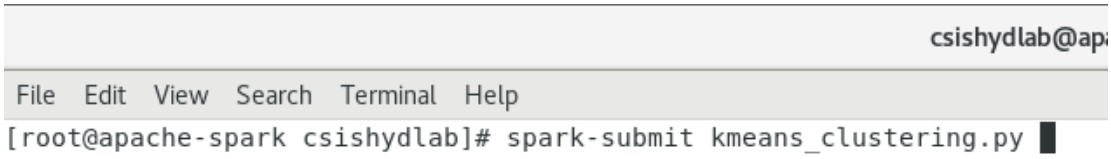


j) Open up the text editor and copy the code written in the attached kmeans_clustering.py file.

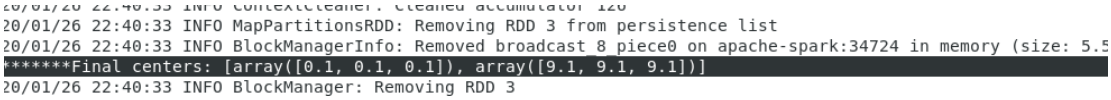
kmeans_clustering.py

```
"""
A K-means clustering program using MLlib.

This example requires NumPy (http://www.numpy.org/).
"""
from __future__ import print_function

import sys

import numpy as np
from pyspark import SparkContext
from pyspark.mllib.clustering import KMeans


def parseVector(line):
    return np.array([float(x) for x in line.split(' ')])

sc = SparkContext(appName="KMeans")
lines = sc.textFile("kmeans_data.txt")
data = lines.map(parseVector)
k = 2
model = KMeans.train(data, k)
print("*******Final centers: " + str(model.clusterCenters))
print("*******Total Cost: " + str(model.computeCost(data)))
sc.stop()
```

k) Execute the kmeans-clustering.py file using the spark-submit command.



l) Look at the outcome printed while the program is getting executed on the Spark cluster. It shows the centroids based on the number of clusters mentioned in the python code (k value) .



m) Repeat the execution by changing the value of k in the python code.

3. Outputs/Results:

Students should be able to

- Execute the linear regression programme on Spark cluster
- See the accuracy of the linear regression models in terms of error
- See the predictions done by the model for the test records
- Execute the k-means programme on Spark cluster
- Look at the centroid generated by the programme with different values of k

4. Observations:

Students carefully needs to observe

- Error statistics associated with the linear regression model
- Centroids change when the number of clusters changed

5. References:

A. Spark ML Guide

B. Spark Documentation

C. Linear Regression

D. K-means clustring