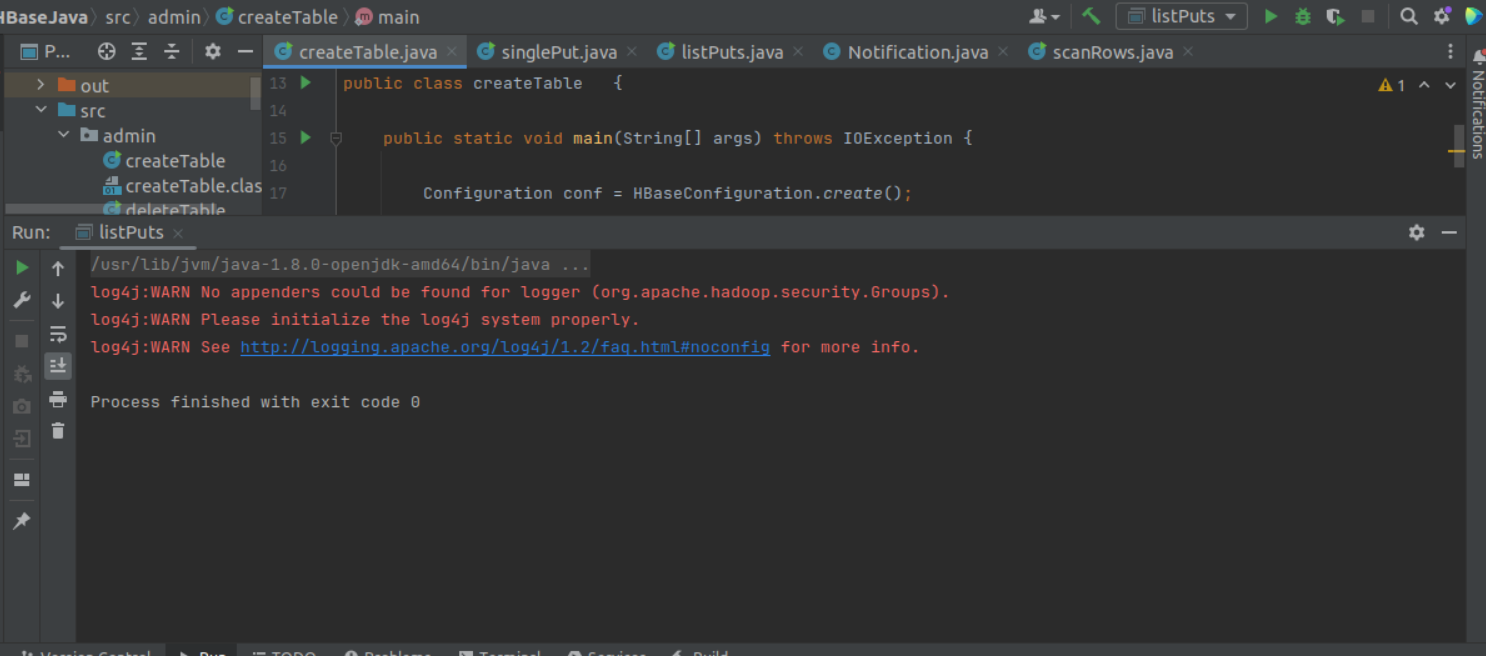
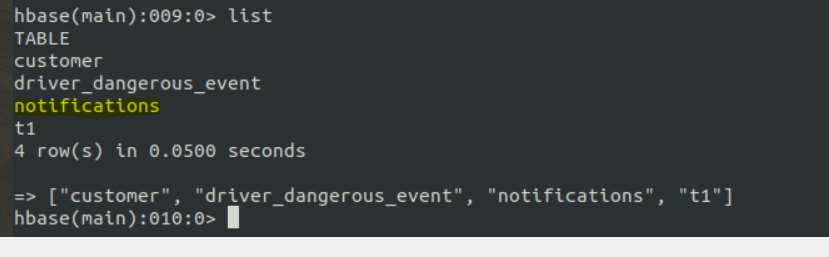


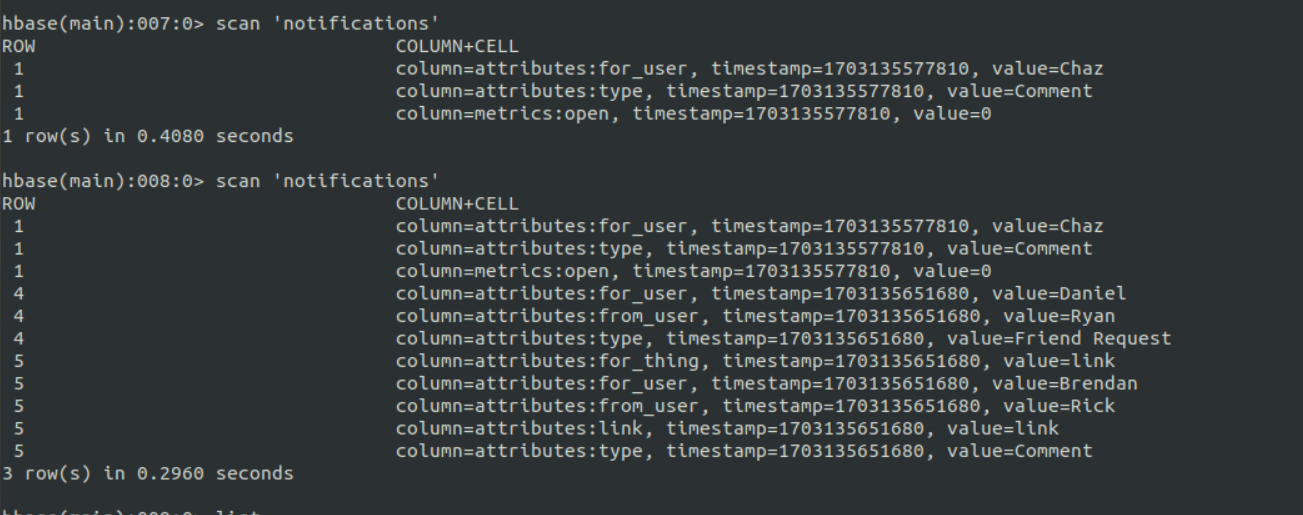
Run Create table file



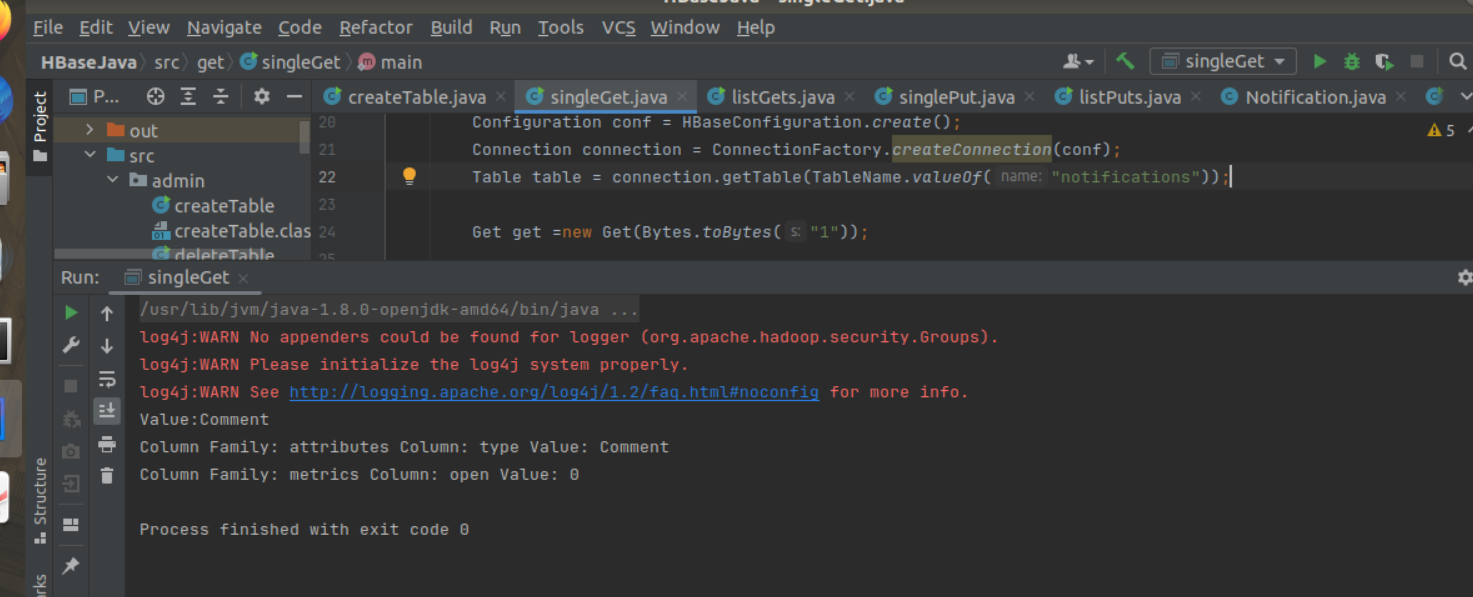


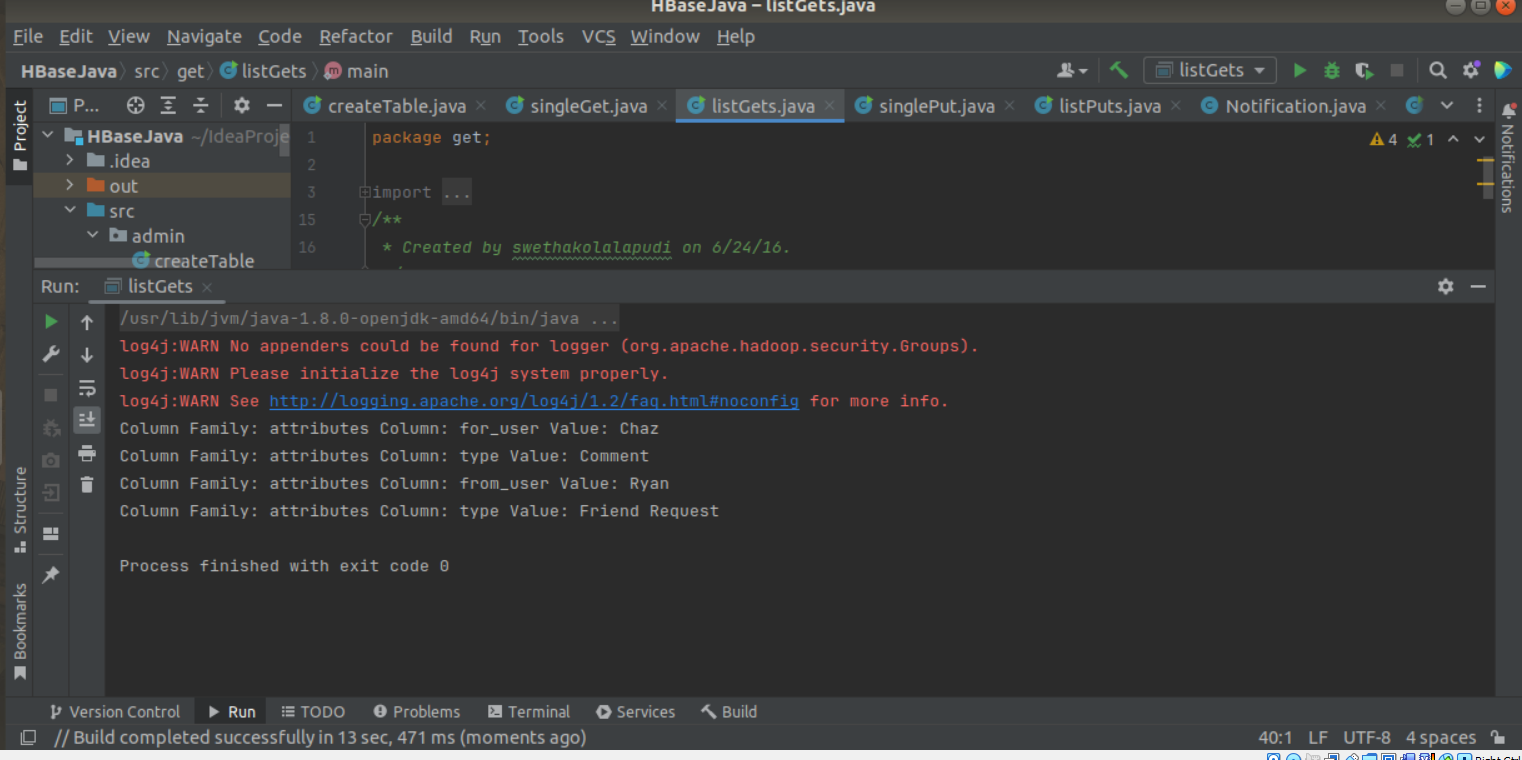
Run Single Put

And listput



Change the row id in the code- given-2 change to 1





Run remaining files

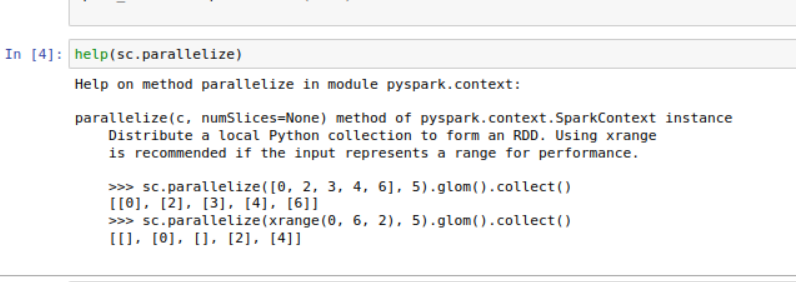
------------------------------------------------------------------------------------------------------------------------------------------

In file Ex2\_Interactive Use of Pyspark

Run paragraphs

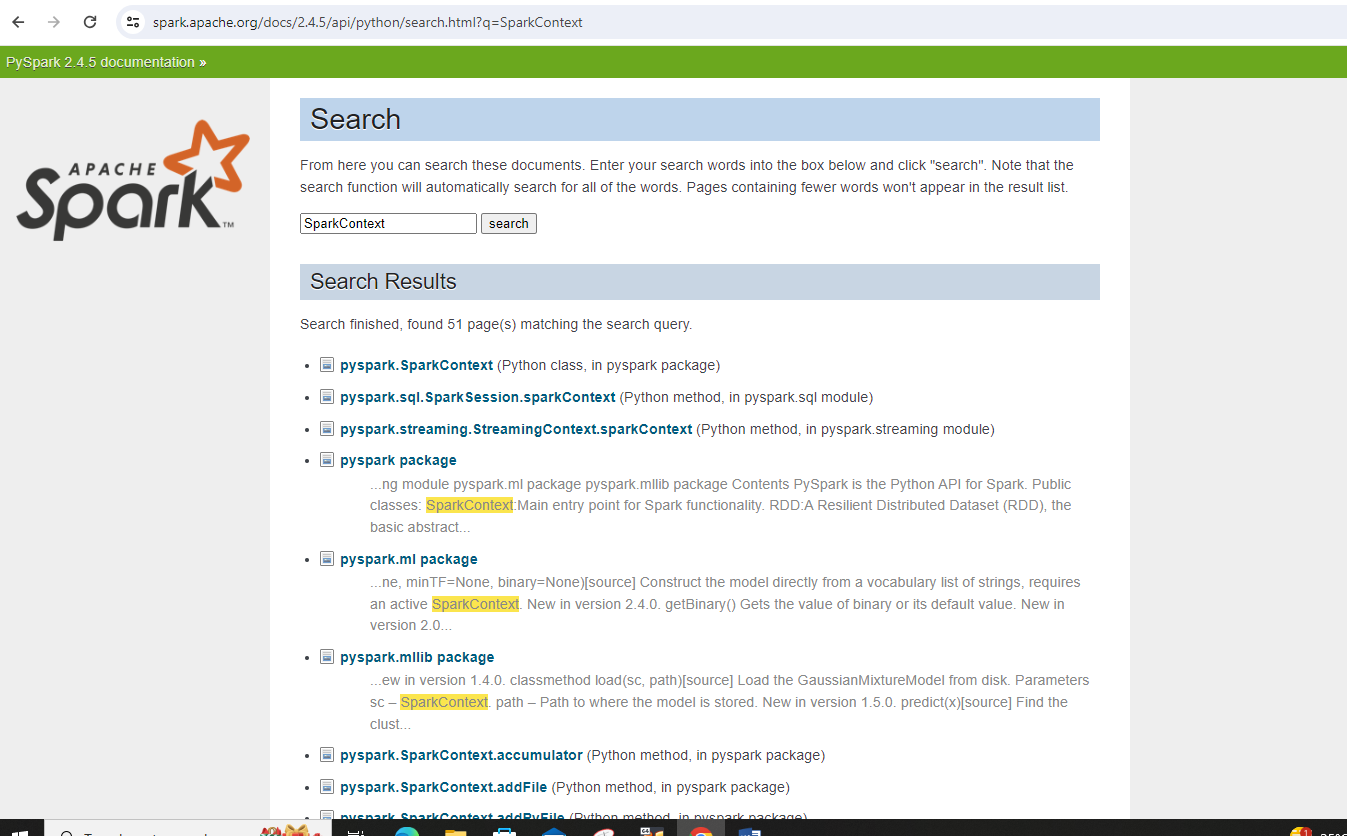
And then

Help(sc.parallelize) to view documentation

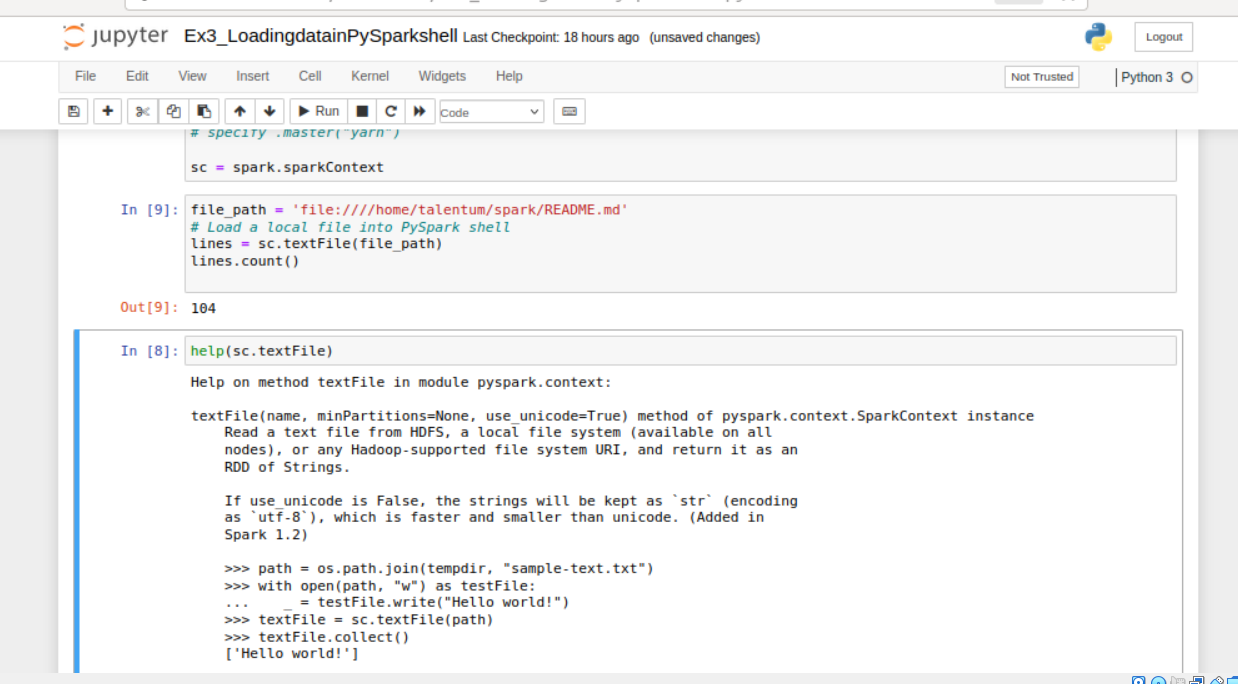


Another way

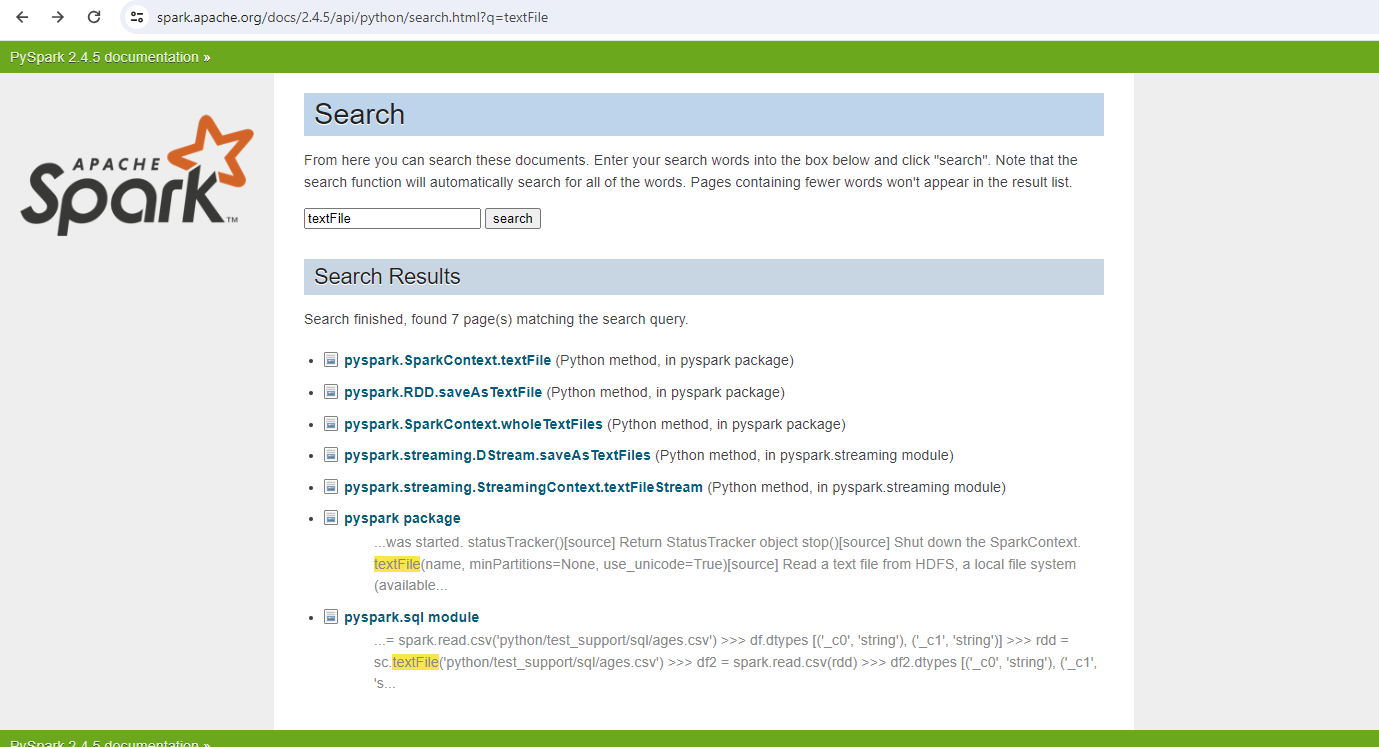
See doucumentation

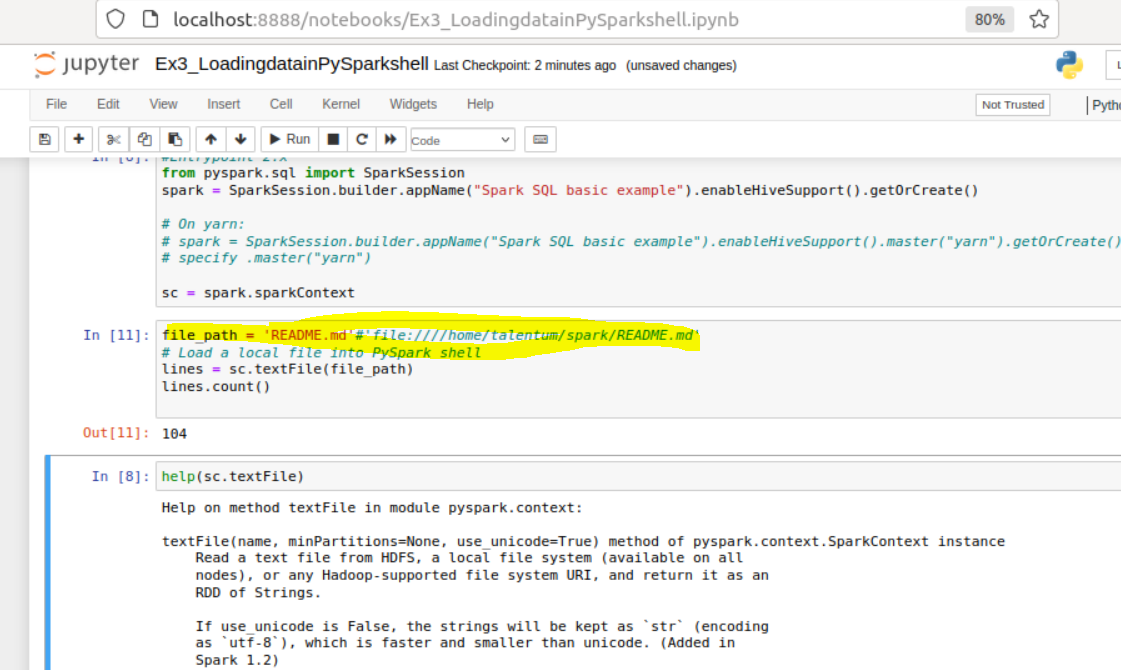


In Ex3\_LoadingdatainPySparkshell



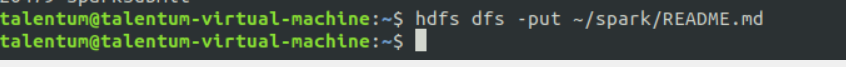
See documentation





Will Give error

Save this file in Hadoop



Make changes and run again

It will run now

3] BigData Fundamentals With Pyspark (M1 Introduction to Big Data analysis with Spark

Review of functional programming in Python (M1->SM3)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

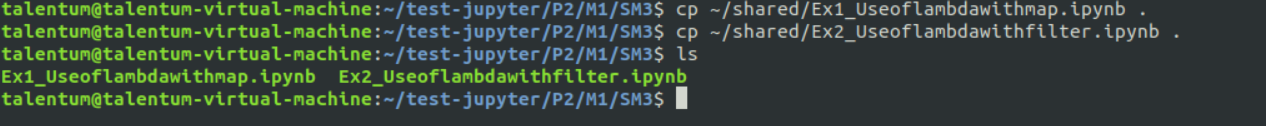
3) Put the .ipynb files in UBUNTU\_HOME/test-jupyter/P2/M1/SM3

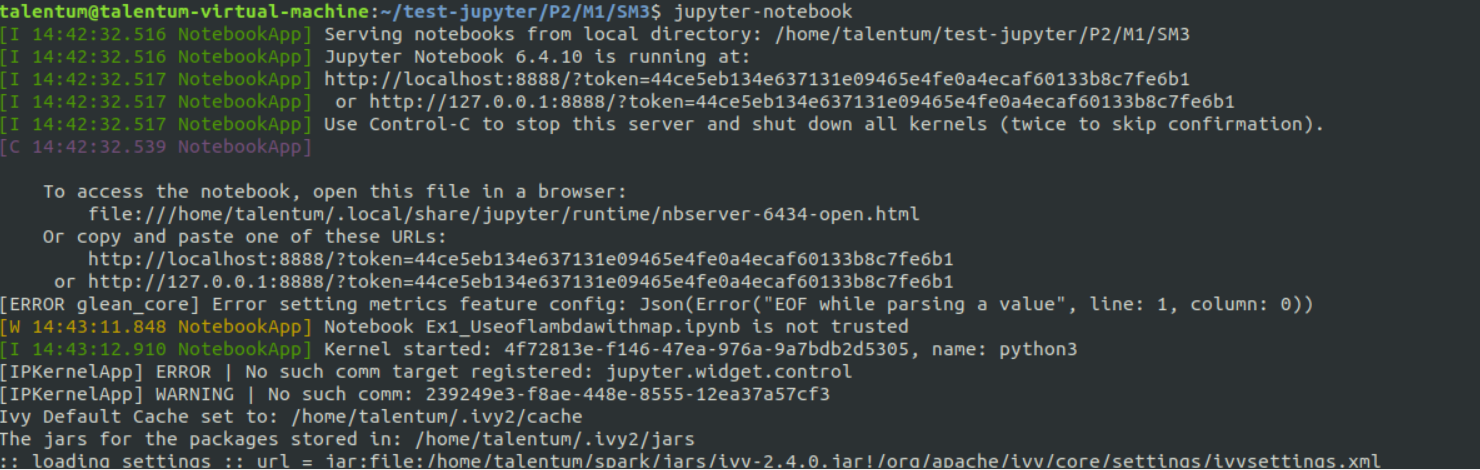
Import in Jupyter Notebook and implement

* Code/Dataset

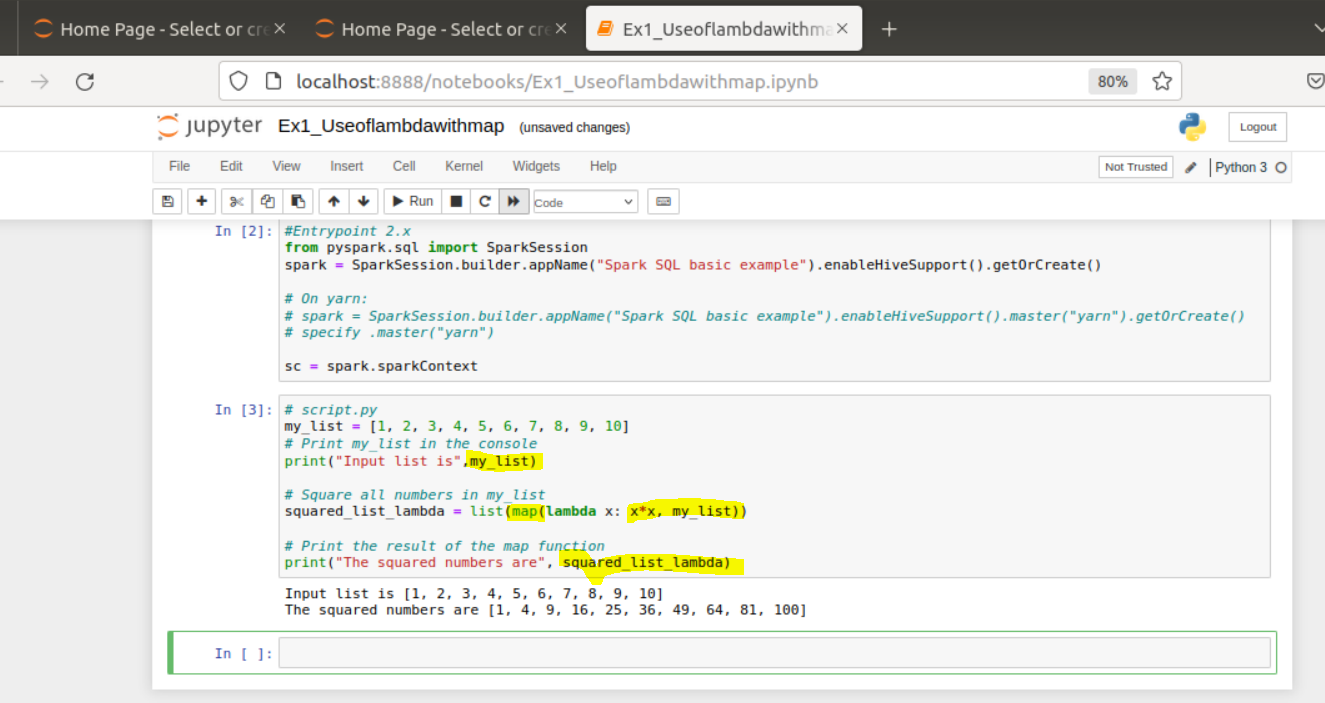
data/3\_ReviewOfFunctionalProgramming.rar

Time = 15 Mins.

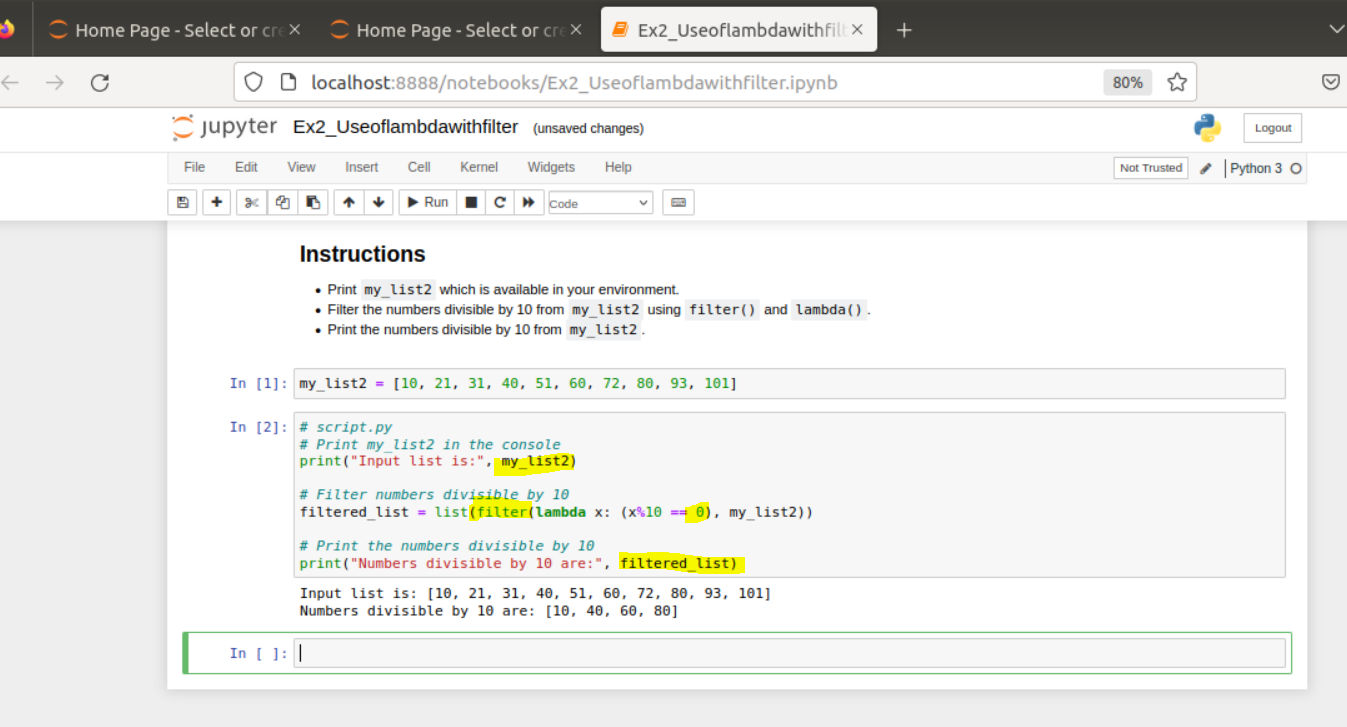




Ex1\_U



Ex2\_Use of Lambda with filter



4] Big Data Fundamentals with PySpark (M2 Programming in PySpark RDD’s

Abstracting Data with RDDs (M2->SM1)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P2/M2/SM1

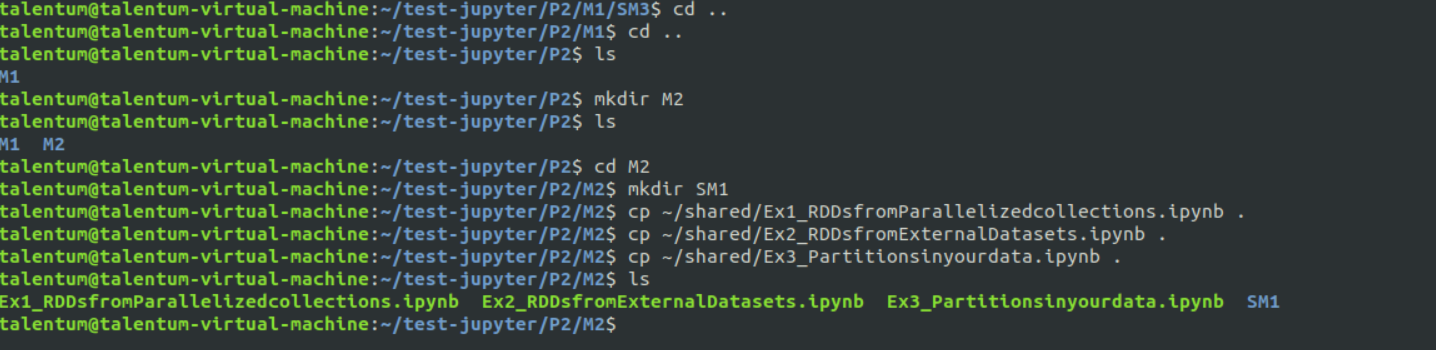
4) Import the notes in Jupyter Note book

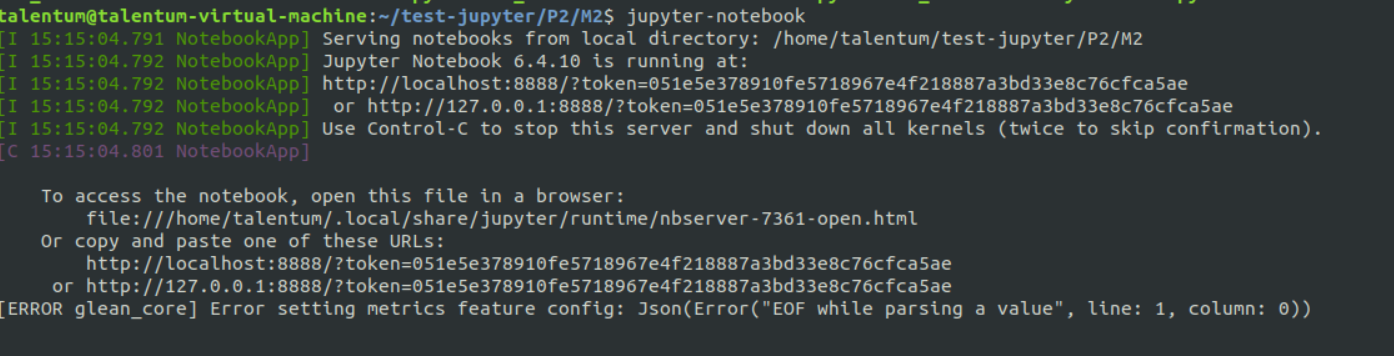
5) Follow the instructions and give the solutions

* Code/Dataset

data/1\_AbstractingDatawithRDDs.rar

Time == 20 Mins

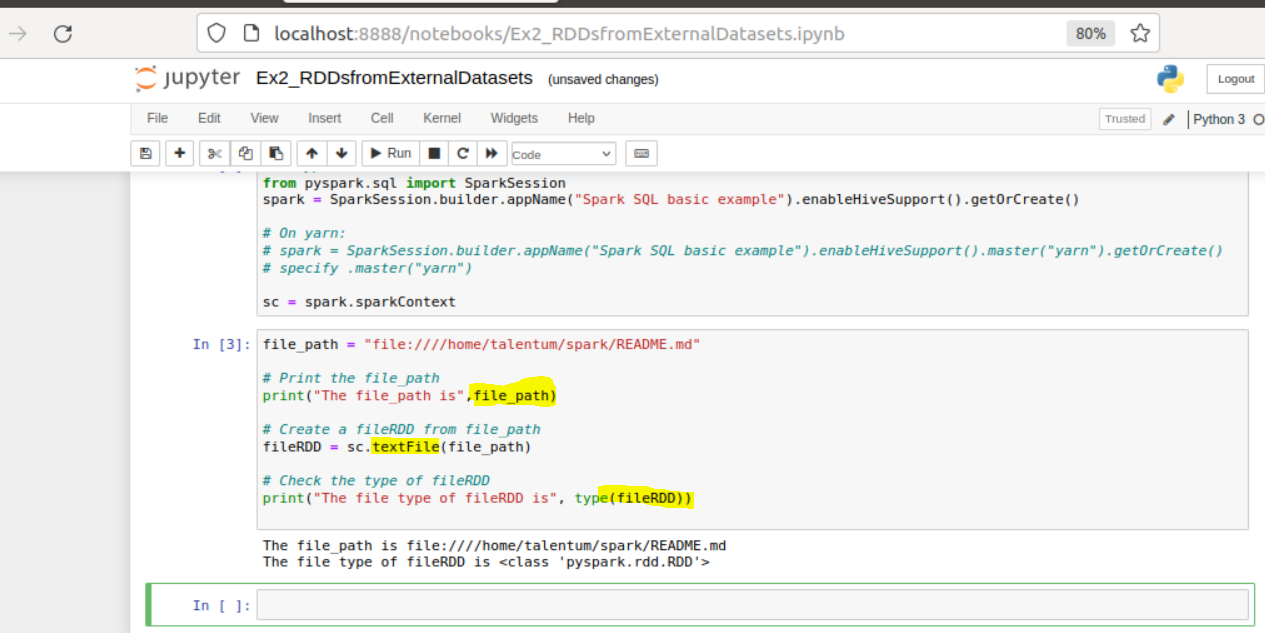




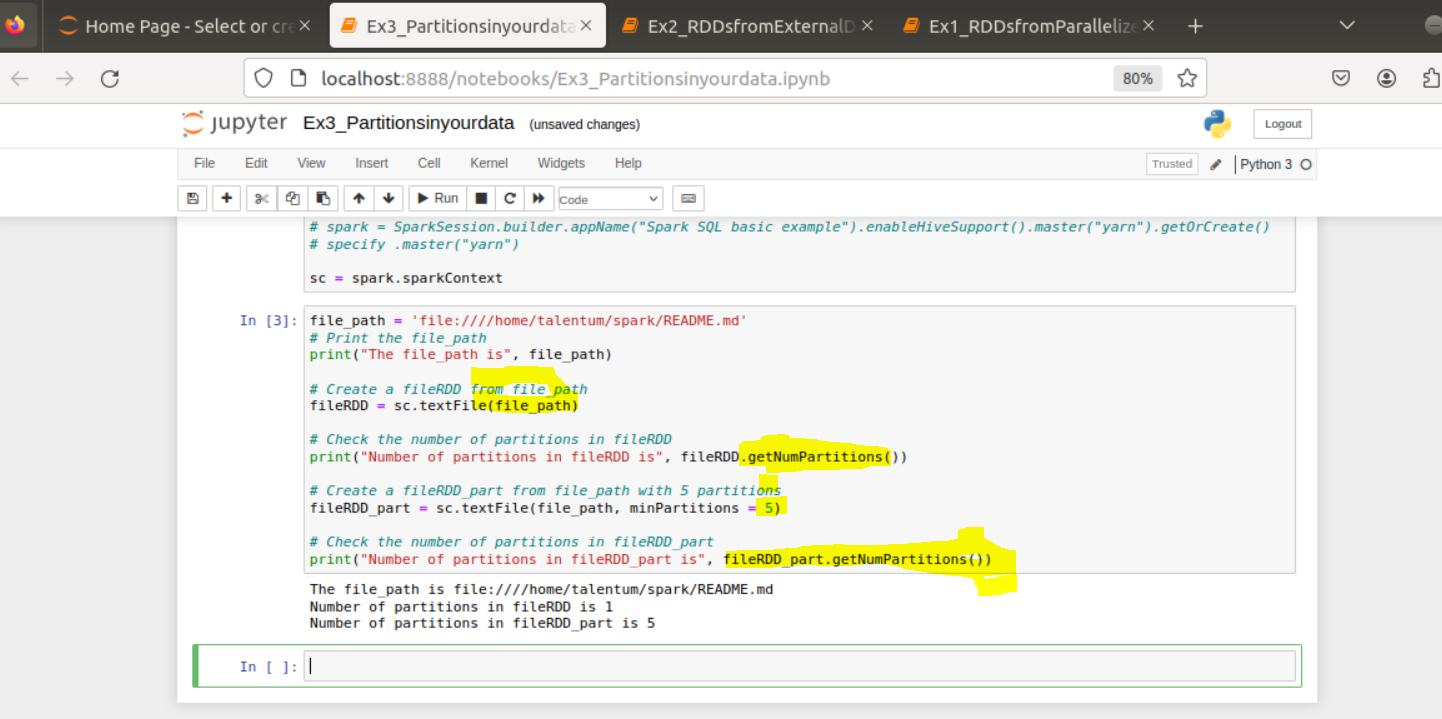
EX1



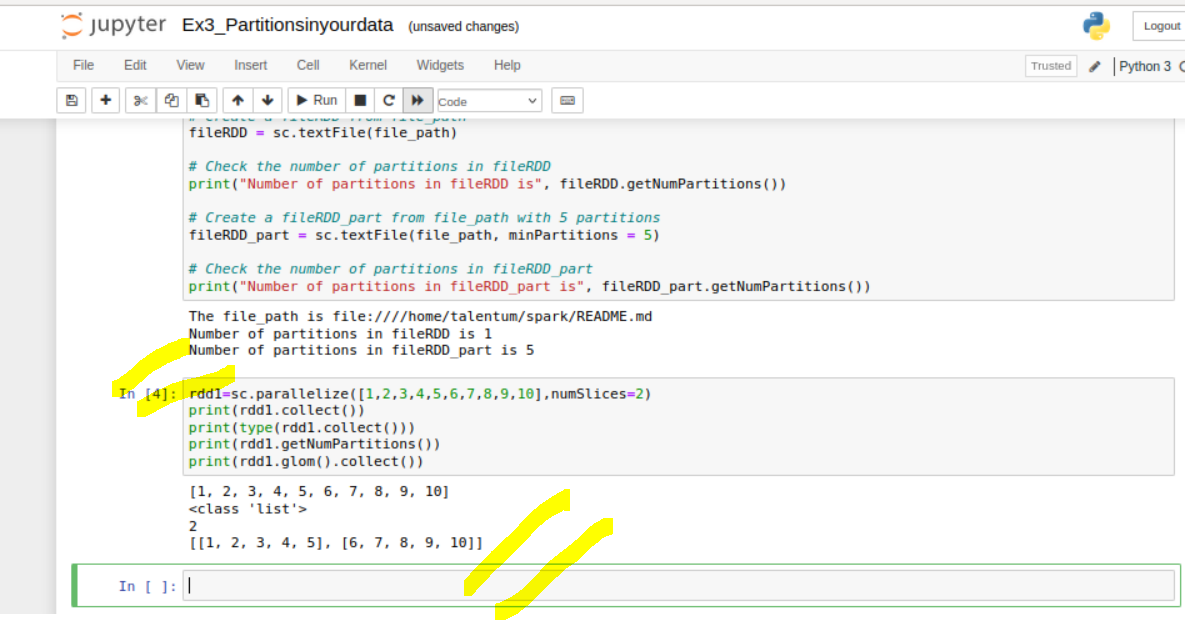
EX2



EX3

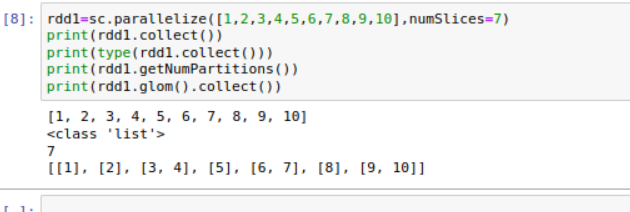


Try this



What is glom in pyspark?

glom ()[source] Return an RDD created by coalescing all elements within each partition into a list.



5] Big Data Fundamentals with PySpark (M2 Programming in PySpark RDD’s )

Basic RDD Transformations and Actions (M2 -> SM2)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P2/M2/sm2

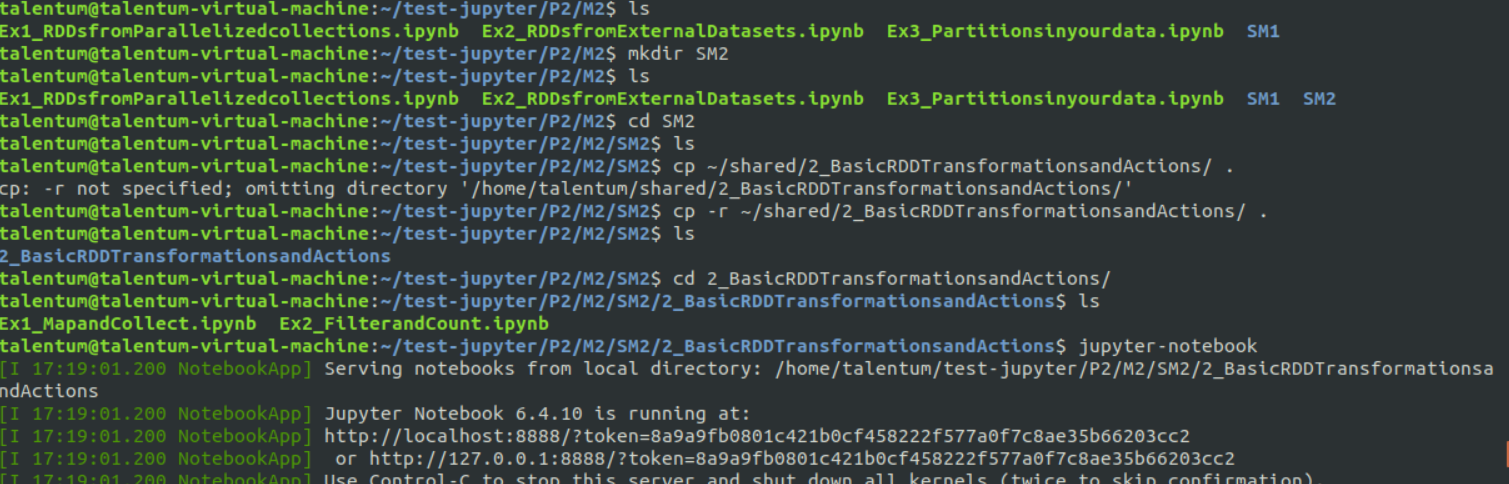
4) Import the notes in Jupyter Notebook

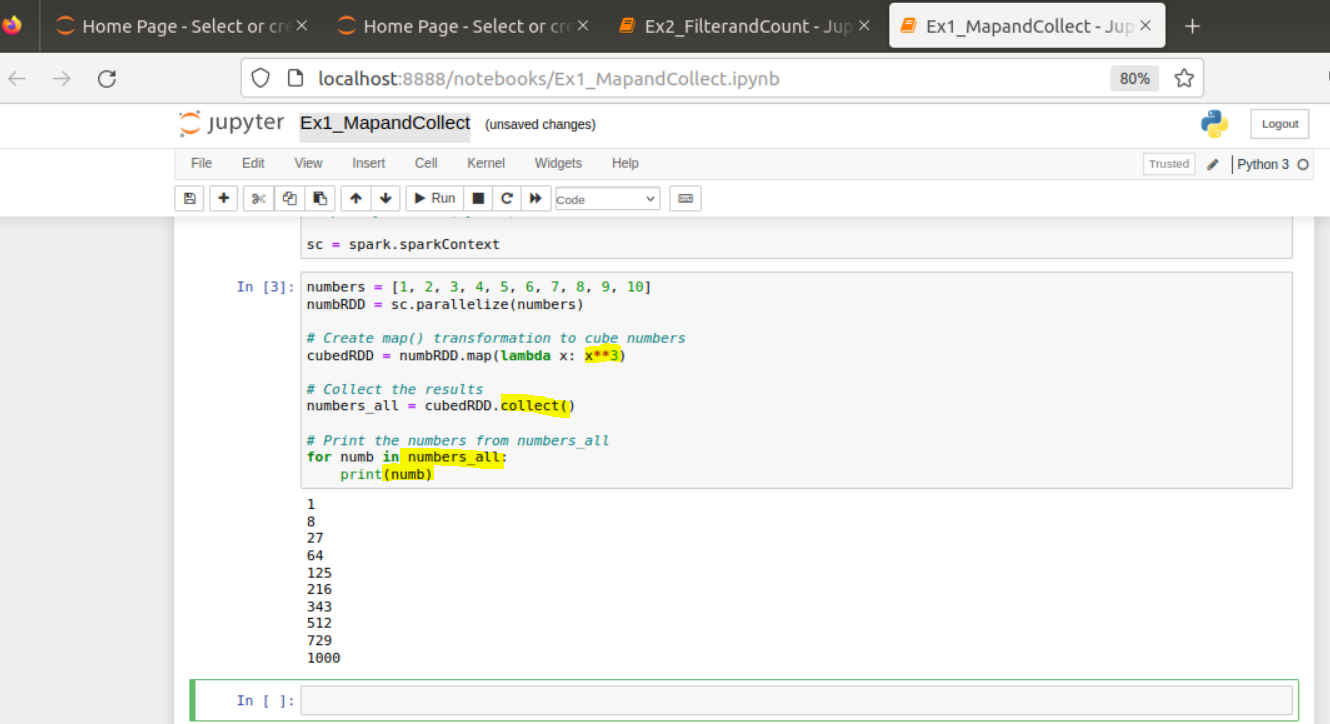
5) Follow the instructions and give the solutions

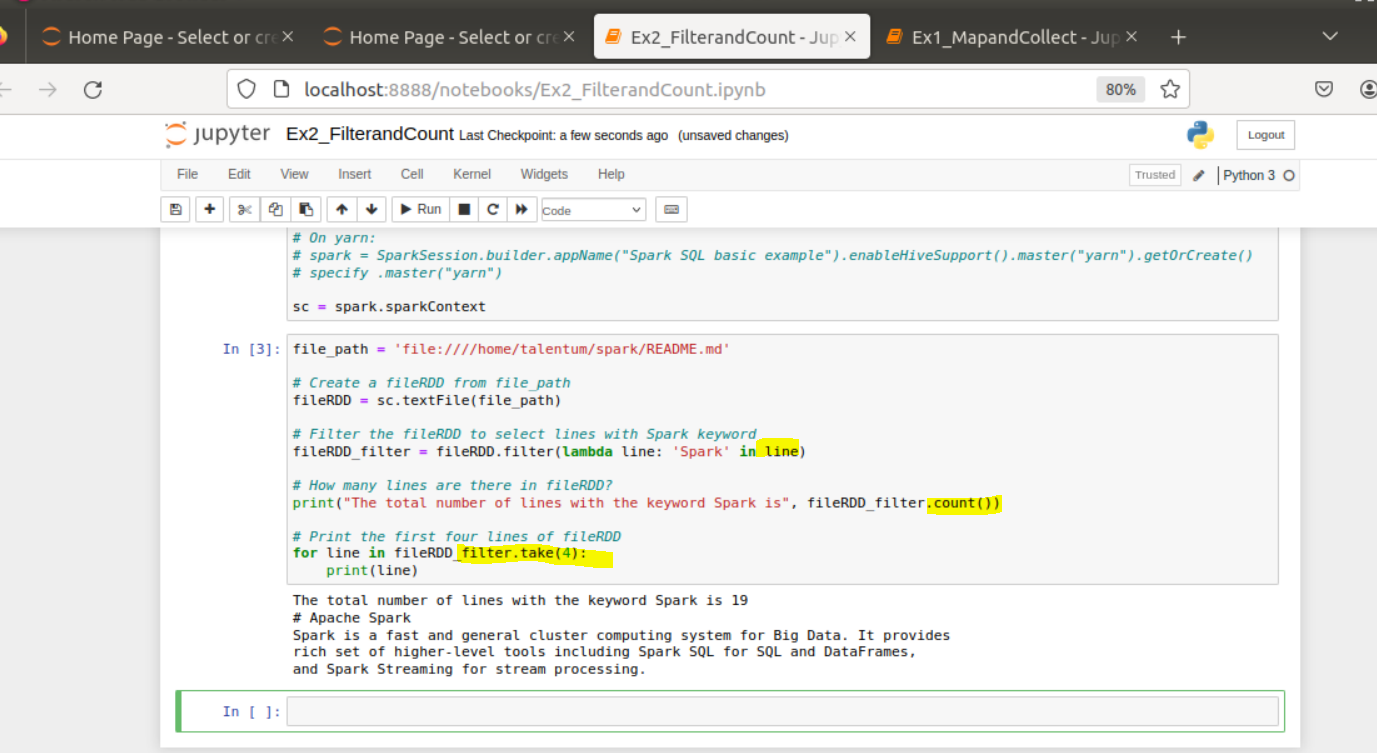
* Code/Dataset

data/2\_BasicRDDTransformationsandActions.rar

**TimeLine** = 20 Minutes







6] Big Data Fundamentals with PySpark (M2 Programming in PySpark RDD’s )

Pair RDDs in PySpark (M2 -> SM3)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P2/M2/sm3

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

data/3\_PairRDDsinPySpark.rar

TimeLine = 20 Minutes

