11] PySpark Sql and DataFrame

(M3->sm3)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P2/M3/sm3

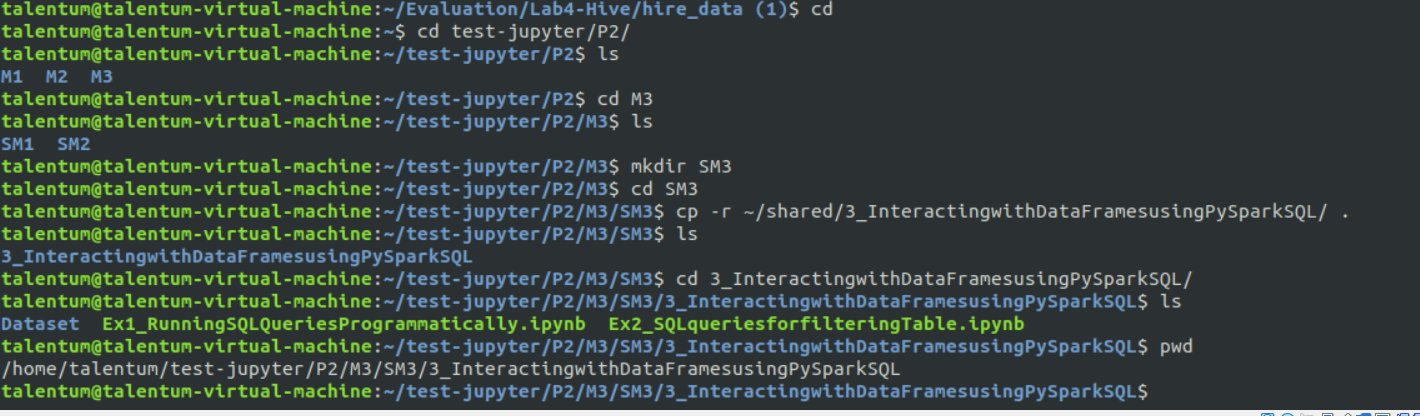
4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

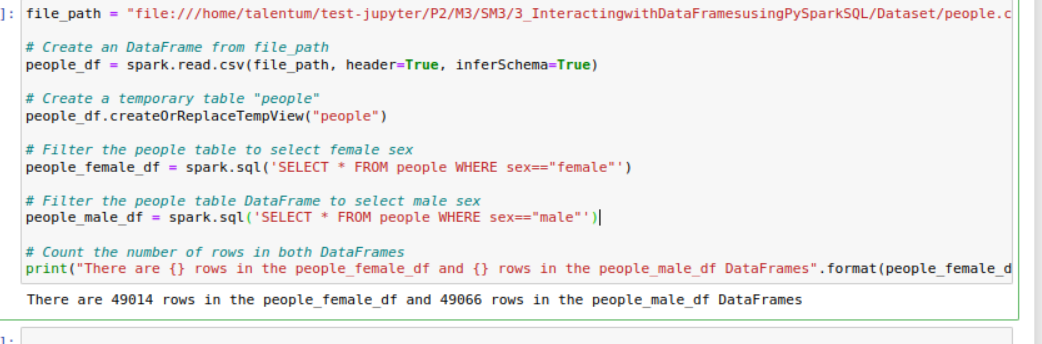
* Code/Dataset

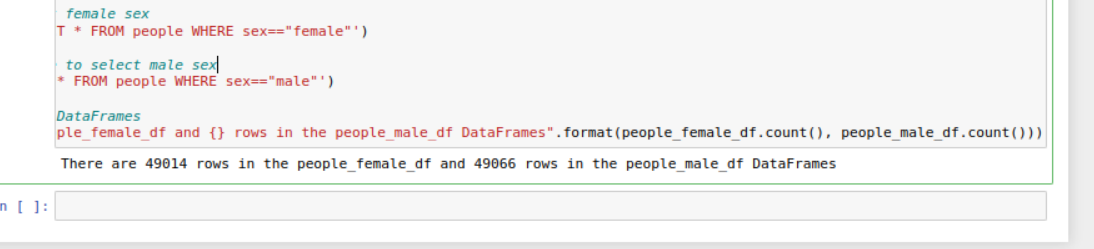
data/3\_InteractingwithDataFramesusingPySparkSQL.rar

Time = 20 Mins









Hdfs processes- NameNode, DataNode, SecondaryNameNode

Yarn – Resourcemanager

Hive – RunJar

Hbase- Hmaster, HquoremPeer can also have HRegionService

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

12] Intro to data cleaning with Apache Spark

(P3->M1)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

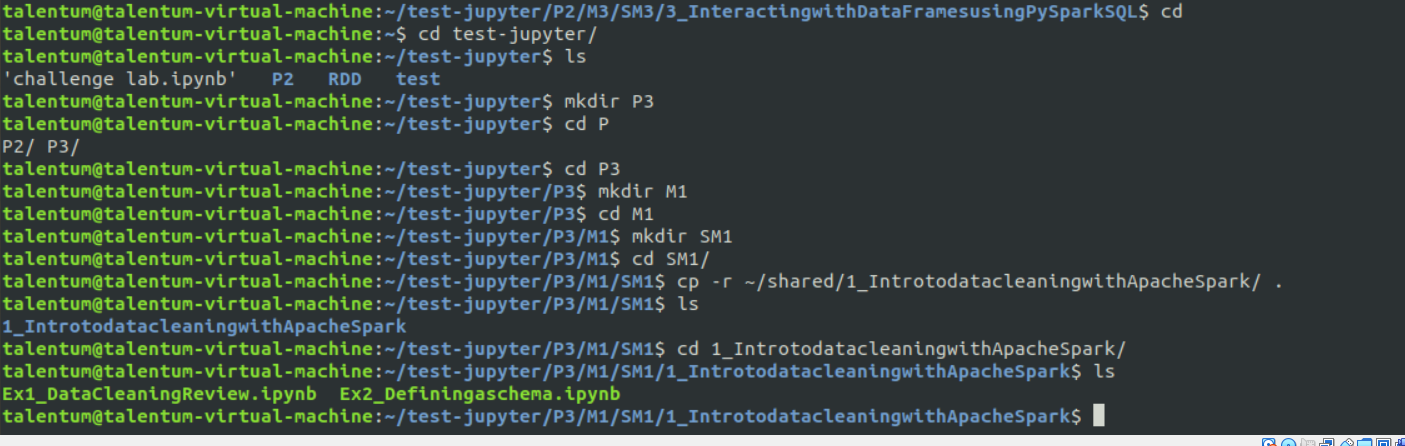
3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M1/sm1

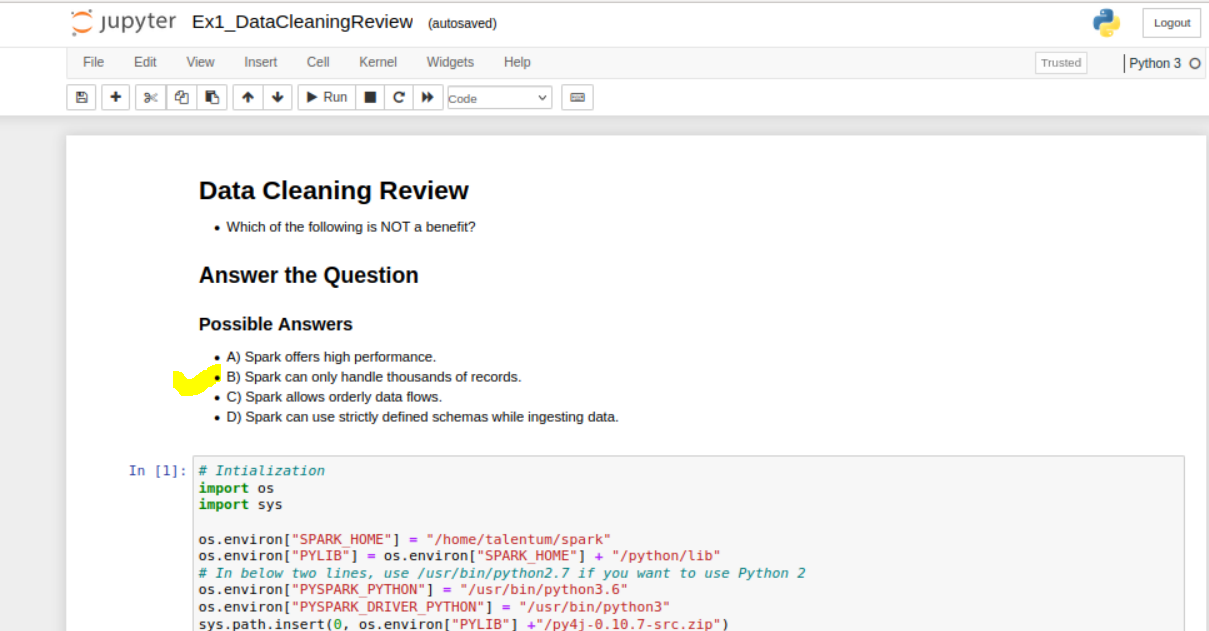
4) Import the notes in Jupyter Notebook

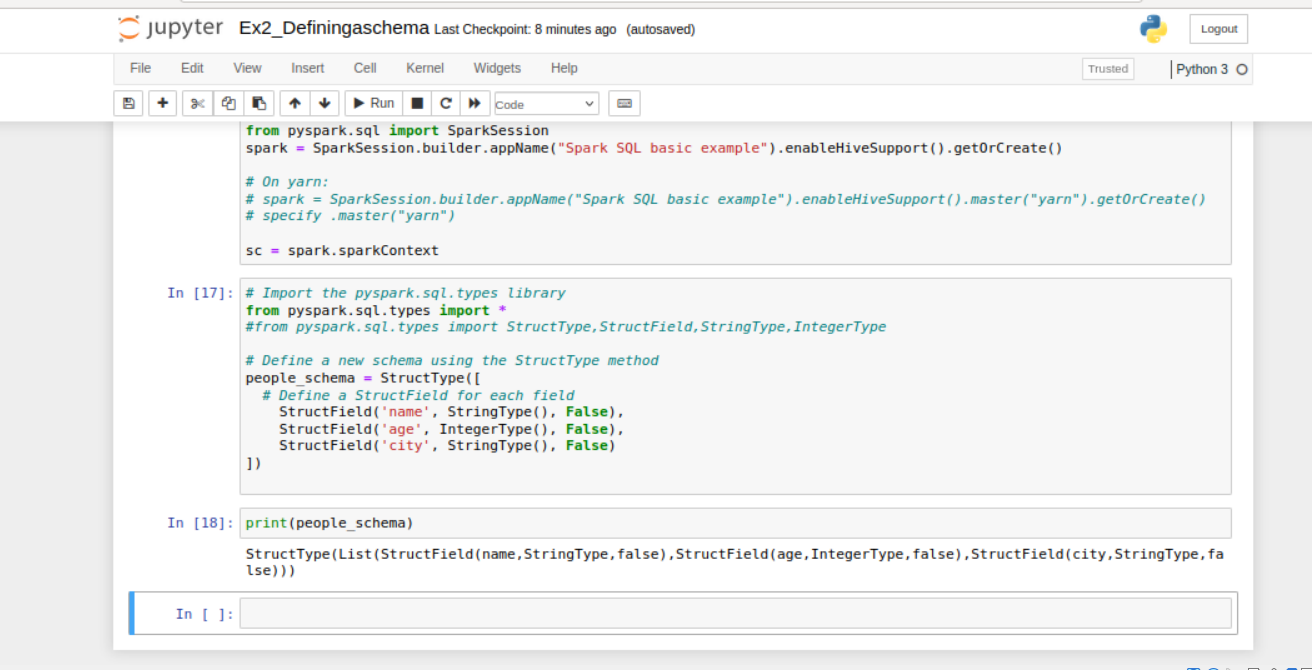
5) Follow the instructions and give the solutions

* Code/Dataset

data/1\_IntrotodatacleaningwithApacheSpark.rar







13] Immutability and lazy processing

(P3 -> M1 -> sm2)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

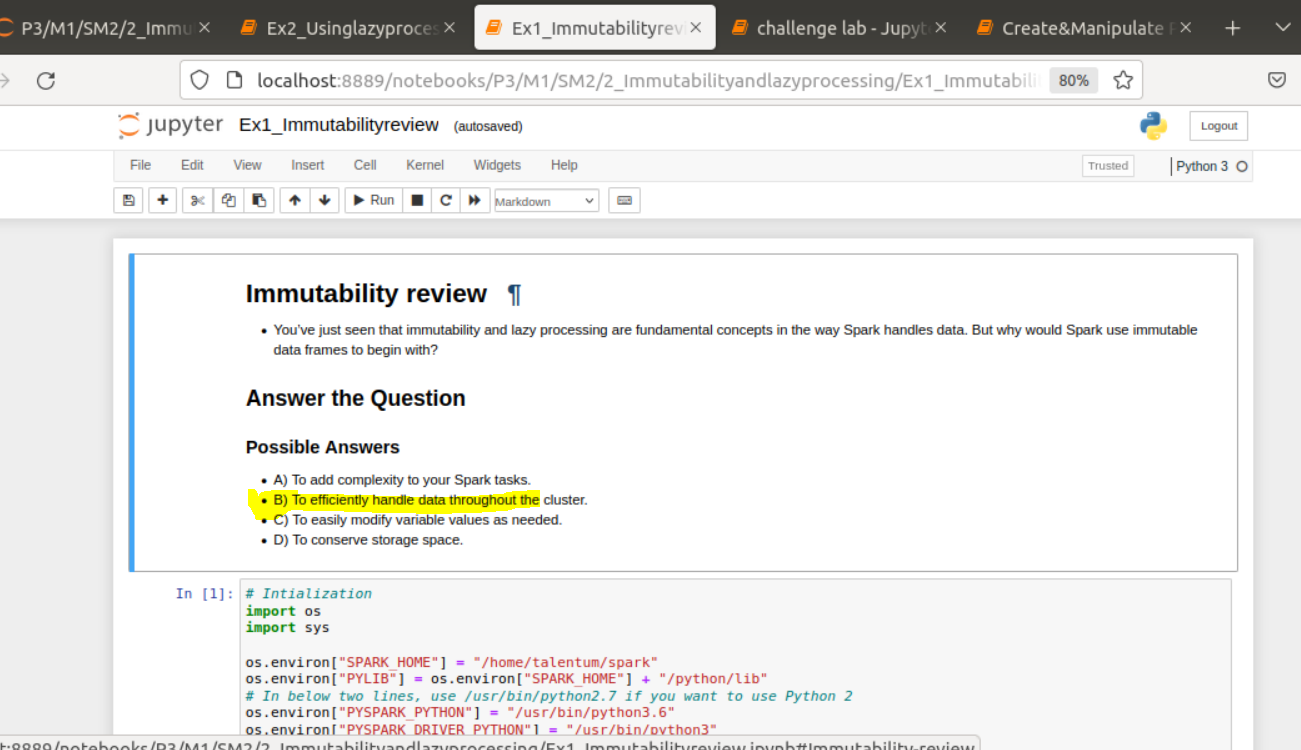
3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M1/sm2

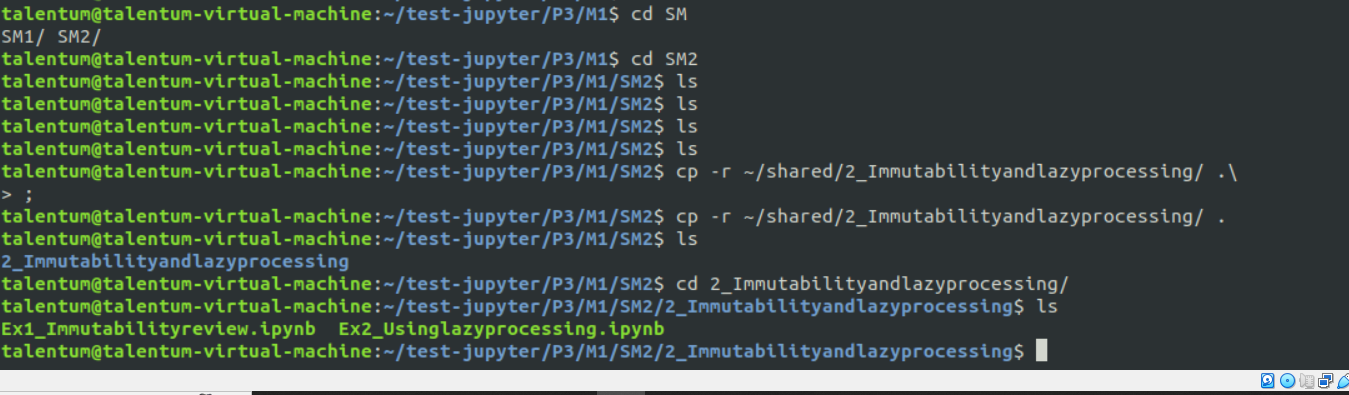
4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

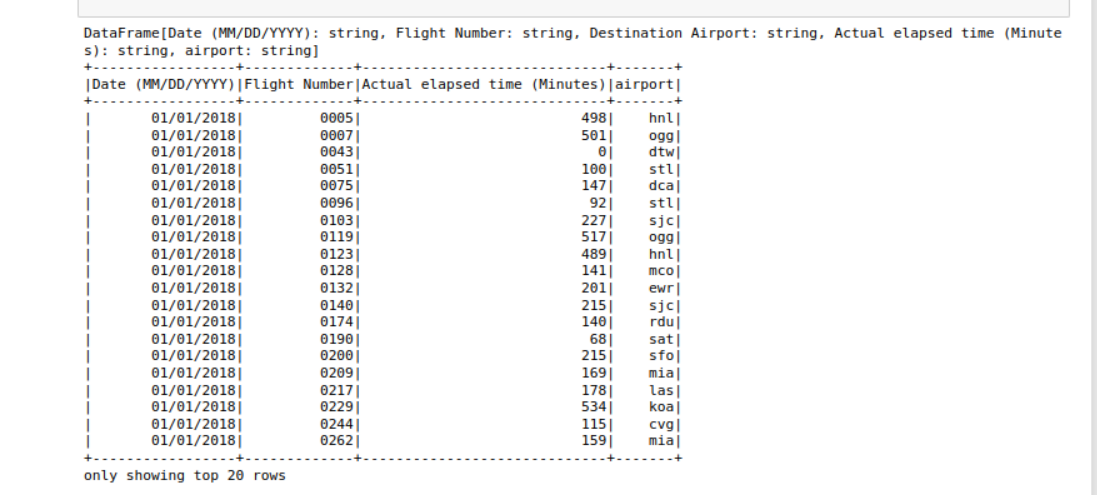
* Code/Dataset

data/2\_Immutabilityandlazyprocessing.rar









14] Understanding Parquet

1) Download RAR on STAGING\_AREA

2) Extrat the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M1/sm3

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

data/3\_UnderstandingParquet.rar

TimeLine = 20 Mins

