10] 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

day4/3\_InteractingwithDataFramesusingPySparkSQL.zip

Time = 20 Mins

11] 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

day4/1\_IntrotodatacleaningwithApacheSpark.zip

TimeLine = 20 Mins

12] 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

day4/2\_Immutabilityandlazyprocessing.zip

TimeLine = 15 Mins

13] Understanding Parquet

1) Download RAR on STAGING\_AREA

2) Extract 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

day4/3\_UnderstandingParquet.zip

TimeLine = 20 Mins

14] DataFrame column operations

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M2/sm1

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

day4/1\_DataFramecolumnoperations.zip

TimeLine = 25 Mins