1] BigData Fundamentals With Pyspark (M1 -> Introduction to Big Data analysis with Spark)

(M1->SM1)

1) Download the attached note in STAGING\_AREA and extract

2) Create a folder UBUNTU\_HOME/test-jupyter/P2/M1/SM1 in Ubuntu machine

3) Pull the extracted folder in above location from your STAGING\_AREA

4) Open it in Jupyter Notebook running on Ubuntu Machine

* Code/Dataset

data/1\_Whatisbigdata.zip

**TimeLine -** 5 Mins

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

PySpark: Spark with Python (M1->SM2)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the Jupyter notes in UBUNTU\_HOME/test-jupyter/P2/M1/SM2

4) Import the notes in Jupyter Notebooks

5) Follow the instructions in the notes and give the solutions

* Code/Dataset -

data/2\_Pyspark-SparkWithPython.rar

Time = 20 Mins.

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 Note book and implement

* Code/Dataset

data/3\_ReviewOfFunctionalProgramming.rar

Time = 15 Mins.

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

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

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

Advanced RDD Actions (M2 -> SM4)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P2/M2/sm4

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

data/4\_AdvancedRddActions.rar

TimeLine = 30 Minutes

8] Pair Rdd

* Code/Dataset

data/Pair-RDD.zip

9] PySpark SQL & DataFrames

(M3->sm1)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P2/M3/sm1

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

data/1\_AbstractingDatawithDataFrames.rar

TimeLine = 15 Minutes

10] PySpark Sql and DataFrame

(M3->sm2)

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P2/M3/sm2

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

data/2\_OperatingonDataFramesinPySpark.rar

TimeLine = 25 Minutes

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

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

TimeLine = 20 Mins

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

TimeLine = 15 Mins

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

15] 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

data/1\_DataFramecolumnoperations.rar

TimeLine = 20 Mins

16] Conditional DataFrame column operations

P3->M2->sm21) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M2/sm2

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

data/2\_ConditionalDataFramecolumnoperations.rar

TimeLine = 25 Mins

17] Spark HBase (Demo)

1. Extract the zip file in STAGING\_AREA
2. Pull the extracted folder in ~/test-jupyter
3. Wait for the instructions from the trainer

* Code/Dataset

data/spark-hbase.zip

18] PySpark MlLib - Overview

1) Pull the extracted notes in UBUNTU\_HOME/test-jupyter/P2/M4/SM1

2) Import the notes in the Jupyter notebook.

3) Follow the instructions based in the notes

* Code/Dataset

data/1\_Overviewofpysparkmllib.rar

TimeLine = 5 Mins

19] PySpark MlLib - Collaborative Filtering

1) Pull the extracted notes in UBUNTU\_HOME/test-jupyter/P2/M4/SM2

2) Import the notes in the Jupyter notebook.

3) Follow the instructions based in the notes

* Code/Dataset

data/2\_CollaborativeFiltering.rar

data/2\_CollaborativeFiltering\_Solution.rar

TimeLine = 25 Mins

20] User defined functions

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M2/sm3

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

data/3\_UserDefinedFunctions.rar

TimeLine = 30 Mins

21] Partitioning and lazy processing

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M2/sm4

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

data/4\_Partitioningandlazyprocessing.rar

TimeLine = 15 Mins

22] Spark Integration with Hive

1. Download the attached file in STAGING\_AREA
2. Pull the file in LABS\_AREA/test-jupyter/spark-hive

* Code/Dataset

data/Spark-Hive.zip

23] Spark JDBC

1. Download the attached file in STAGING\_AREA
2. Pull the file in LABS\_AREA/test-jupyter/spark-mysql

* Code/Dataset

data/Spark-MySql.zip

24] Spark Structured Streaming

1. Download the attached file in STAGING\_AREA
2. Pull the file in LABS\_AREA/test-jupyter/pyspark-kafka

* Code/Dataset

data/PySpark-Kafka.zip

25] Surprise Lab - Wordcount using Spark DF

* Code/Dataset

data/Surprise\_Lab\_WordCount.zip

26] Spark Submit

Download and pull to Sandbox home directory

* Code/Dataset

data/spark-submit.zip