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

1) Download the attached note in STAGING\_AREA

2) Create a folder $HOME/test-jupyter/P2/M1/SM1 in Ubuntu machine

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

4) Open it in Jupyter Notebook running on Ubuntu Machine

Dataset - Spark/P2/M1/\*.ipynb

**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 $HOME/test-jupyter/P2/M1/SM2

4) Import the notes in Jupyter Notebooks

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

Dataset - Spark/P2/M1/SM2/2\_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) Extrat the RAR

3) Put the .ipynb files in $HOME/test-jupyter/P2/M1/SM3

Import in Jupyter Notebook and implement

Dataset - Spark/P2/M1/SM3/3\_ReviewOfFunctionalProgramming.rar

Time = 20 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 $HOME/test-jupyter/P2/M2/sm2

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P2/M2/SM1/1\_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 $HOME/test-jupyter/P2/M2/SM2

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P2/M2/SM2/2\_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 $HOME/test-jupyter/P2/M2/SM3

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P2/M2/SM3/3\_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 $HOME/test-jupyter/P2/M2/sm4

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P2/M2/SM4/4\_AdvancedRddActions.rar

TimeLine = 30 Minutes

**8] WordCount Application**

Apply your newly acquired knowledge of RDD and PairRDD

1) Implement the wordcount application in PySpark,Which you implemented using Java MapReduce previously

2) Refer the same dataset which you use it forJava MR application.

3) Implement the Lab in a Jupyter Notebook

TimeLine = 30 Minutes

**9] PairRdd Application**

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Refer the pdf file which is a Lab Manual for instructions

4) Implement the Lab in a Jupyter Notebook

Dataset - Spark/RDD/Pair-RDD.rar

TimeLine = 30 Minutes

**10] PySpark SQL & DataFrames**

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in $HOME/test-jupyter/P2/M3/SM1

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P2/M3/SM1/1\_AbstractingDatawithDataFrames.rar

TimeLine = 15 Minutes

**11] PySpark Sql and DataFrame**

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

Dataset - Spark/P2/M3/SM3/3\_InteractingwithDataFramesusingPySparkSQL.rar

TimeLine = 20 Minutes

**12] Study and Run the application (Self Study and Implementation)**

1. Refer the Spark-Hbase.ipynb file for instructions

Dataset - Spark/Spark-HBase/spark-hbase.rar

**13] Intro to data cleaning with Apache Spark**

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in $HOME/test-jupyter/P3/M1/SM1

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P3/M1/SM1/1\_IntrotodatacleaningwithApacheSpark.rar

TimeLine = 20 Mins

**14] Immutability and lazy processing**

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

Dataset - Spark/P3/M1/SM2/2\_Immutabilityandlazyprocessing.rar

TimeLine = 20 Mins

**15] Understanding Parquet**

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in $HOME/test-jupyter/P3/M1/SM3

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P3/M1/SM3/3\_UnderstandingParquet.rar

TimeLine = 20 Mins

**16] PySpark MlLib**

1) Pull the extracted notes in $HOME/test-jupyter/P2/M4/SM1

2) Import the notes in Jupyter notebook.

3) Follow the instructions based in the notes

Dataset - Spark/P2/M4/SM1/1\_Overviewofpysparkmllib.rar

TimeLine = 20 Mins

**17] PySpark MlLib**

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

Dataset - Spark/P2/M4/SM2/2\_CollaborativeFiltering.rar

TimeLine = 30 Mins

**18] DataFrame column operations**

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in $HOME/test-jupyter/P3/M2/SM1

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P3/M2/SM1/1\_DataFramecolumnoperations.rar

TimeLine = 20 Mins

**19] Conditional 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/sm2

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P3/M2/SM2/2\_ConditionalDataFramecolumnoperations.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

Dataset - Spark/P3/M2/SM3/3\_UserDefinedFunctions.rar

TimeLine = 20 Mins

**21] Spark Integration with Hive**

Download the attached file in STAGING\_AREA

Pull the file in LABS\_AREA/test-jupyter/spark-hive

Dataset - Spark/Saprk-Hive/\*

**22] PySpark MlLib**

1) Pull the extracted notes in UBUNTU\_HOME/test-jupyter/P2/M4/SM3

2) Import the notes in the Jupyter notebook.

3) Follow the instructions based in the notes

Dataset- Spark/P2/M4/SM2/3\_Classification.rar

TimeLine = 30 Mins

**23] Partitioning and lazy processing**

1) Download RAR on STAGING\_AREA

2) Extrat the RAR

3) Put the \*.ipynb files in $HOME/test-jupyter/P3/M2/SM4

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P3/M2/SM4/4\_Partitioningandlazyprocessing\_Solution.rar

TimeLine = 20 Mins

**24] Caching**

1) Download RAR on STAGING\_AREA

2) Extrat the RAR

3) Put the \*.ipynb files in $HOME/test-jupyter/P3/M3/SM1

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P3/M3/SM1/Caching.rar

TimeLine = 15 Mins

**25) Improve import performance**

1) Download RAR on STAGING\_AREA

2) Extrat the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M3/sm2

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

Dataset - Spark/P3/M3/SM2/2\_Improveimportperformance.rar

TimeLine = 15 Mins

**26] Spark Submit**

1)Download and pull to Sandbox home directory

Dataset - Spark/Spark-Submit/

27] Spark-MySql

1. Download,extract and pull the rar to Sandbox home directory
2. Open the ipynb file and follow the instructions.

Dataset- Spark/Spark-MySql

**28] Spark Streaming**

1)Download and pull to Sandbox home directory

Dataset - Spark/Spark-Streaing/PySpark-Kafka.rar

**29]** **Spark Streaming**

1)Download and pull to Sandbox home directory

Dataset - Spark/Spark-Streaing/Spark-Structured-Streaming.rar

30] Airflow

1) Download Airflow/dags folder and pull to Sandbox Desktop directory

Follow the following steps to setup the environment for running Airflow

steps to set airflow

~$ cd Desktop/airflow-tutorial/

~Desktop/airflow-tutorial$ source ~/unset\_jupyter.sh

~Desktop/airflow-tutorial$ echo $PYSPARK\_DRIVER\_PYTHON

~Desktop/airflow-tutorial$ echo $PYSPARK\_DRIVER\_PYTHON\_OPTS

~Desktop/airflow-tutorial$ unset PYTHONPATH

~Desktop/airflow-tutorial$ conda activate airflow-tutorial

(airflow-tutorial)~Desktop/airflow-tutorial$ echo $AIRFLOW\_HOME

(airflow-tutorial)~Desktop/airflow-tutorial$ code .

(airflow-tutorial)~Desktop/airflow-tutorial$ airflow resetdb / airflow initdb

(airflow-tutorial)~Desktop/airflow-tutorial$ airflow webserver

(airflow-tutorial)~Desktop/airflow-tutorial$ airflow scheduler

Dataset - Airflow/dags/