15] 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

day5/2\_ConditionalDataFramecolumnoperations.zip

TimeLine = 25 Mins

16] 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

day5/3\_UserDefinedFunctions.zip

TimeLine = 30 Mins

Kafka

—------

1. Topic

In third CLI give kafka-topics.sh command

>>>>$kafka-topics.sh

$kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic first\_topic --create

$kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic first\_topic --create --partitions 3

$kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic first\_topic --create --partitions 3 --replication-factor 2

$kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic first\_topic --create --partitions 3 --replication-factor 1

$kafka-topics.sh --zookeeper 127.0.0.1:2181 --list

$kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic first\_topic --describe

1. Producer

$kafka-console-producer.sh

Produce messages

So we are going to launch our first kafka-console-producer.

>>>>$kafka-console-producer.sh --broker-list 127.0.0.1:9092 --topic first\_topic

$kafka-console-producer.sh --broker-list 127.0.0.1:9092 --topic first\_topic --producer-property acks=all

1. Consumer

$kafka-console-consumer.sh

$kafka-console-consumer.sh --bootstrap-server 127.0.0.1:9092 --topic first\_topic

$kafka-console-consumer.sh --bootstrap-server 127.0.0.1:9092 --topic first\_topic --from-beginning

1. Consumer group

$kafka-console-consumer.sh --bootstrap-server 127.0.0.1:9092 --topic first\_topic --group my-first-application

<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html>

17] Caching

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M3/sm1

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

day5/1\_Caching.zip

TimeLine = 20 Mins

18] Improving the import performance

1) Download RAR on STAGING\_AREA

2) Extract 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

* Code/Dataset

day5/2\_Improveimportperformance.zip

TimeLine = 20 Mins

19] Improving the import performance

1) Download RAR on STAGING\_AREA

2) Extract the RAR

3) Put the \*.ipynb files in UBUNTU\_HOME/test-jupyter/P3/M3/sm3

4) Import the notes in Jupyter Notebook

5) Follow the instructions and give the solutions

* Code/Dataset

day5/3\_clusterconfigurations.zip

TimeLine = 20 Mins

20] Improving the import performance

1) Download RAR on STAGING\_AREA

2) Extract 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

* Code/Dataset

day5/4\_Performanceimprovements.zip

TimeLine = 20 Mins

21] Spark Submit

Download and pull extracted folder to Sandbox home directory

* Code/Dataset

data/spark-submit.zip

22] 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

23] 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

day5/2\_CollaborativeFiltering.rar

day5/2\_CollaborativeFiltering\_Solution.rar

TimeLine = 30 Mins

24] PySpark MlLib - Classification

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

* Code/Dataset

data/3\_Classification.rar

data/3\_Classification\_Solution.rar

TimeLine = 30 Mins

25] PySpark MlLib - Clustering

1) Pull the extracted notes in UBUNTU\_HOME/test-jupyter/P2/M4/SM4

2) Import the notes in the Jupyter notebook.

3) Follow the instructions based in the notes

* Code/Dataset

data/4\_Clustering.rar

data/4\_Clustering\_Solution.rar

TimeLine = 40 Mins

ML References -

—------------------------

1) PySpark Machine Learning Tutorial for Beginners (1)

https://www.projectpro.io/hadoop-tutorial/pyspark-machine-learning-tutorial

2) Your First Apache Spark ML Model

https://towardsdatascience.com/your-first-apache-spark-ml-model-d2bb82b599dd (2)

3) Machine Learning Library (MLlib) Guide

https://spark.apache.org/docs/latest/ml-guide.html (3)