LOADING DATA DEMISTIFIED: 4 WAYS OF READING DATA IN TO SPARK SQL

Let the FILES do

All the work in

Cluster

CHALLENGE AT HAND

- DATA PROVIDED CAN BE IN MULTIPLE FORMAT, AND IT COULD HAVE COME DIRECTLY FROM ANOTHER CLUSTER
- WE NEED TO UNDERSTAND HOW TO CREATE DATA FOR LOADING INTO THE SPARK CLUSTER

1. SPARK.READ.PARQUET
2. SPARK.READ.CSV
3. SPARK.CATALOG.CREATETABLE

MSCK REPAIR TABLE TBL_NAME

HOW WE ARE DOING IT?

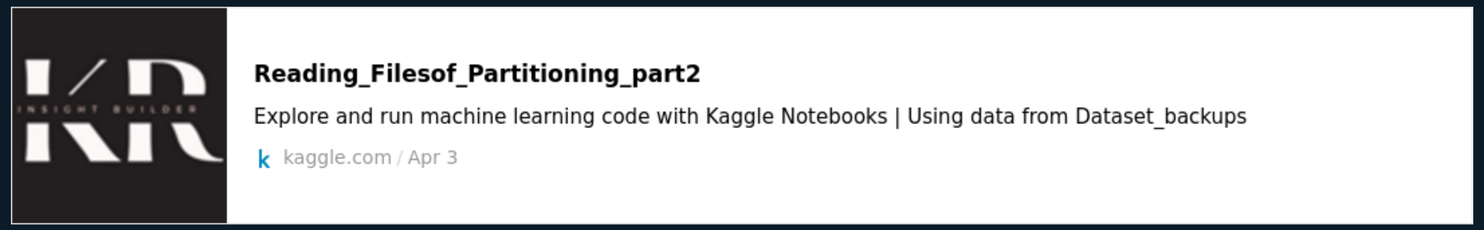
USE KAGGLE NOTEBOOK TO LOAD DATA IN PYSPARK

ABOVE COMMANDS ARE EXPLAINED

DISCUSS HOW THE COMMANDS CAN MADE INTO SCRIPTS

TROUBLE SHOOTING ISSUES THAT ARISES IN MESSY DATA

LETS GET OURSELF A PYSPARK NOTEBOOK AND DIG IN



REAL CLUSTER IS NOT NECESSARY FOR LEARNING THE DML

THANKS FOR WATCHING

PRACTICE

PRACTICE

LIKE

SHARE

SUBSCRIBE

PRACTICE

PRACTICE

HTTPS://GITHUB.COM/INSIGHTBUILDER