MASTERING DATA CLUSTERING IN SPARK: 5 WAYS TO CLUSTER DATA IN SPARK SQL

Unleash

The

Power Of

Big Data

CHALLENGE AT HAND

- THERE IS A RIVER OF FLOWING DATA, WHICH IS GENERATED REAL TIME, RANDOM AND FULL OF INSIGHTS.
- TO GATHER THOSE INSIGHTS THE FIRST STEP TO IS TO CATEGORIZE THE DATA IN SOME FASHION. IN SPARK THERE ARE 5 DIFFERENT WAYS TO DO IT.

CLUSTER BY

DISTRIBUTE BY

GROUP BY

PARTITION BY

BUCKET BY

NOTE BOOK CONTAINS THE NECESSARY DETAILS AND DEFINITIONS, ALONG WITH THE DATA

HOW WE ARE DOING IT?

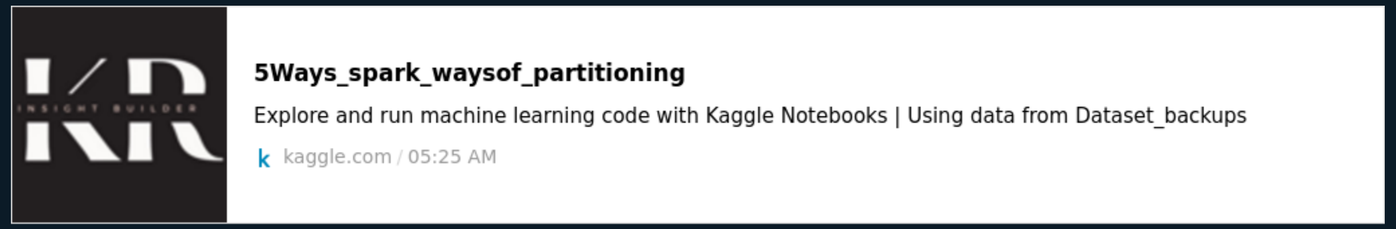
USE KAGGLE NOTEBOOK TO LOAD DATA IN PYSPARK

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