ADVANCED SPARK SQL TECHNIQUES: EXPLORING WINDOW CLAUSE

Unleashing the Power of Data Analytics

CHALLENGE AT HAND : SEQUENTIAL AGG

- WINDOW FUNCTION IS USEFUL WHEN THE CHANGE ACROSS TIME NEEDS TO BE ANALYSED
- RANKING THE ENTITIES RELATIVE TO TIME AND ARBITRARY DATA POINT
- UNLIKE GROUP BY NEW COLUMN IS CREATED FROM WINDOW OPERATION
- 3 DIFFERENT FUNCTIONS:

RANKING FUNCTIONS

- SYNTAX: RANK | DENSE_RANK | PERCENT_RANK | NTILE | ROW_NUMBER

ANALYTIC FUNCTIONS

- SYNTAX: CUME_DIST | LAG | LEAD | NTH_VALUE | FIRST_VALUE | LAST_VALUE

AGGREGATE FUNCTIONS

- SYNTAX: MAX | MIN | COUNT | SUM | AVG | ...

HOW WE ARE DOING IT?

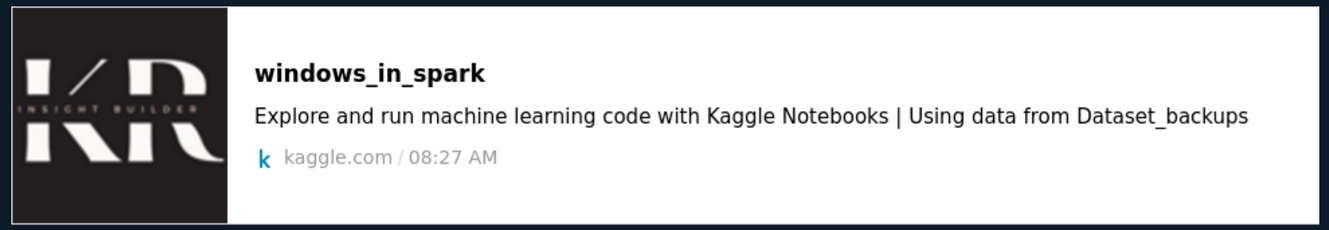
USE KAGGLE NOTEBOOK TO LOAD DATA IN PYSPARK

ABOVE COMMANDS ARE EXECUTED

DISCUSS THE RESULTS OF EXECUTION AND PROBLEM SOLVED

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