MUST HAVE SPARK DML COMMANDS FOR DATA ENGINEERS & ANALYSTS

Solving Common Data Analysis Challenges

CHALLENGE SOLUTION APPROACH

1) INGESTING THE DATA INTO THE SPARK TABLE : READ APIS

2) MULTIPLE WAYS OF SELECTING COLUMNS INSIDE TABLES

FUNCTIONS:

3) HANDING NULL VALUES IN COLUMNS : COALESCE

4) WORKING WITH DATE COLUMNS: DATE_FORMAT / DATE_TRUNC

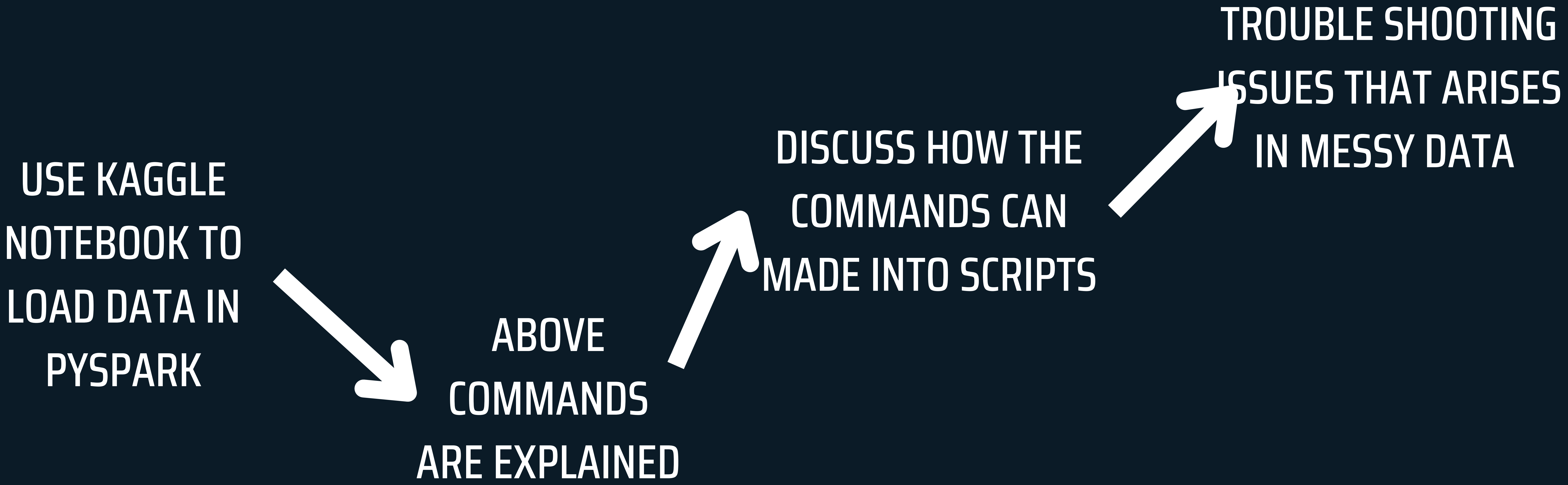
5) CREATING DATE_SERIES: SEQUENCE, EXPR AND WITHCOLUMN

6) STATISTICAL AGGREGATION: MAX, MIN, GROUPBY AND STATISTICAL FUNCTION

7) FILTERING THE ROWS : WHERE AND HAVING

7) CHECKING MULTIPLE CONDITIONS AND PROVIDING OUTPUT: CASE WHEN THEN

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

spark_DML_commands_part1

Explore and run machine learning code with Kaggle Notebooks | Using data from multiple data sources

kaggle.com / 01:14 AM

REAL CLUSTER IS NOT NECESSARY FOR LEARNING THE DML

THANKS FOR WATCHING

PRACTICE

PRACTICE

LIKE

SHARE

SUBSCRIBE

PRACTICE

PRACTICE

HTTPS://GITHUB.COM/INSIGHTBUILDER