MASTER ART OF BIG DATA ETL WITH PYSPARK

ETL 175Million Records using Kaggle Environment

HTTPS://GITHUB.COM/INSIGHTBUILDER

WHAT PROBLEM WE'RE FACING

WE TALKING WITH BIG DATA ENVIRONMENT

Spark program can function using our NTFS / LFS without the Distributed feature of the file storage and processing.

Only by getting hands on with a tech, we can master it

SERIES OF PROBLEM

- NOT HAVING BIG ENOUGH FILE
- NOT HAVING THE COMPUTE POWER
- PAID WAREHOUSES FOR LIMITED TIME
- OVERVIEW OF HOW EVERY PART OF ECOSYSTEM WORKS

PYSPARK & THIS VIDEO WILL SOLVE

HTTPS://GITHUB.COM/INSIGHTBUILDER

OUR OBJECTIVE

1. PRACTICE ON MASSIVE DATASETS ON KAGGLE. CHOOSING UBER NY DATASET
2. LOAD THE DATASET INTO SPARK SESSION AND DO TRANSFORMATION
3. EXPLORE GROUPBY, WITHCOLUMN AND SELECT CLAUSES IN SPARK DATAFRAME
4. CREATE TABLE ON SPARK CATALOG IN TWO DIFFERENT METHODS
5. WORK WITH PARTITIONED DATA AND RECOVER THE DATA

HTTPS://GITHUB.COM/INSIGHTBUILDER

HEAD TO kaggle / LOCAL NOTEBOOK

Pyspark_UberNY_ETL_to_SparkCTG

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

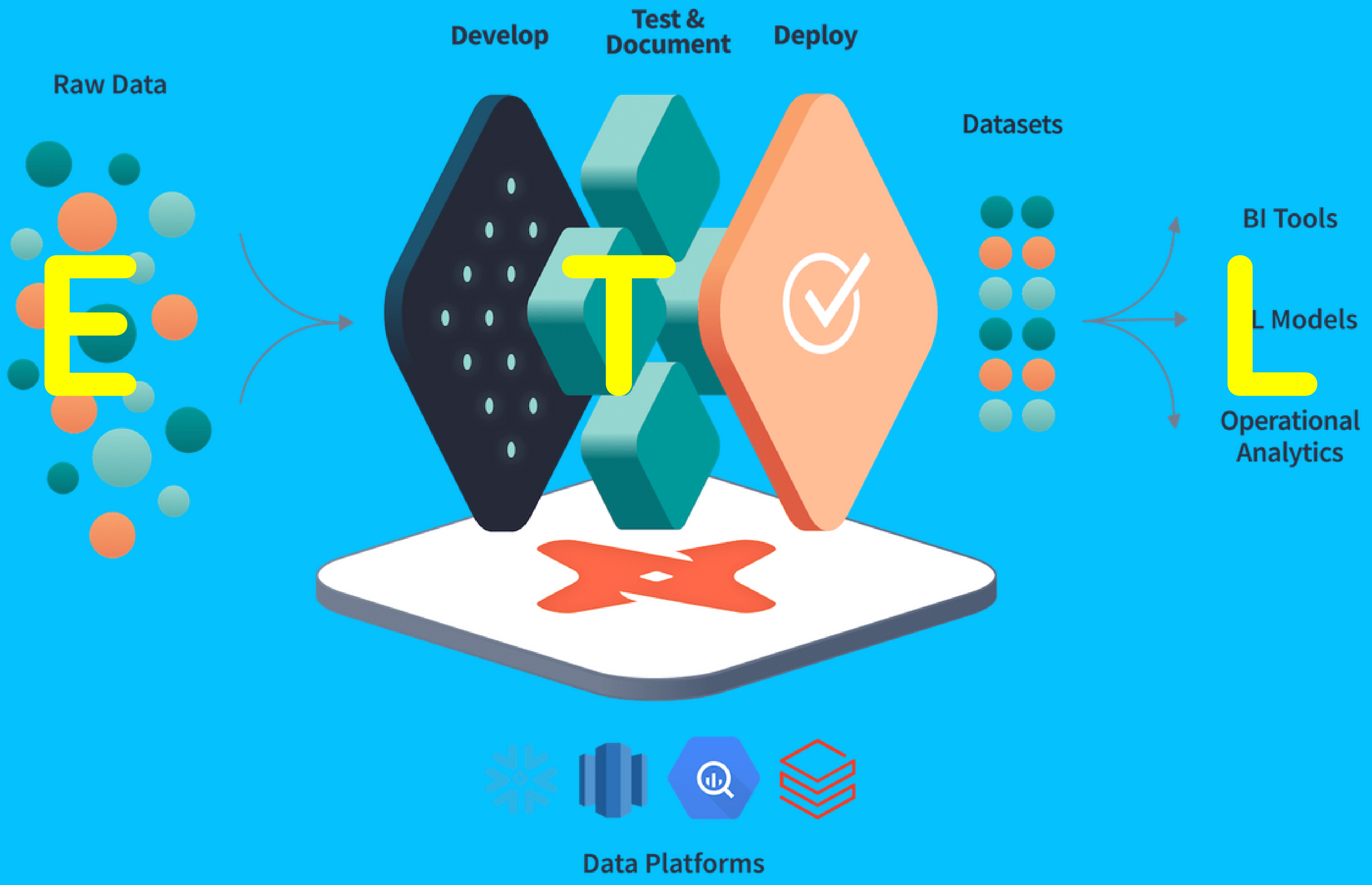
kaggle.com / 08:26 AM

https://www.kaggle.com/kamaljp/pyspark-uberny-etl-to-sparkctg

QUESTIONS AND COMMENTS

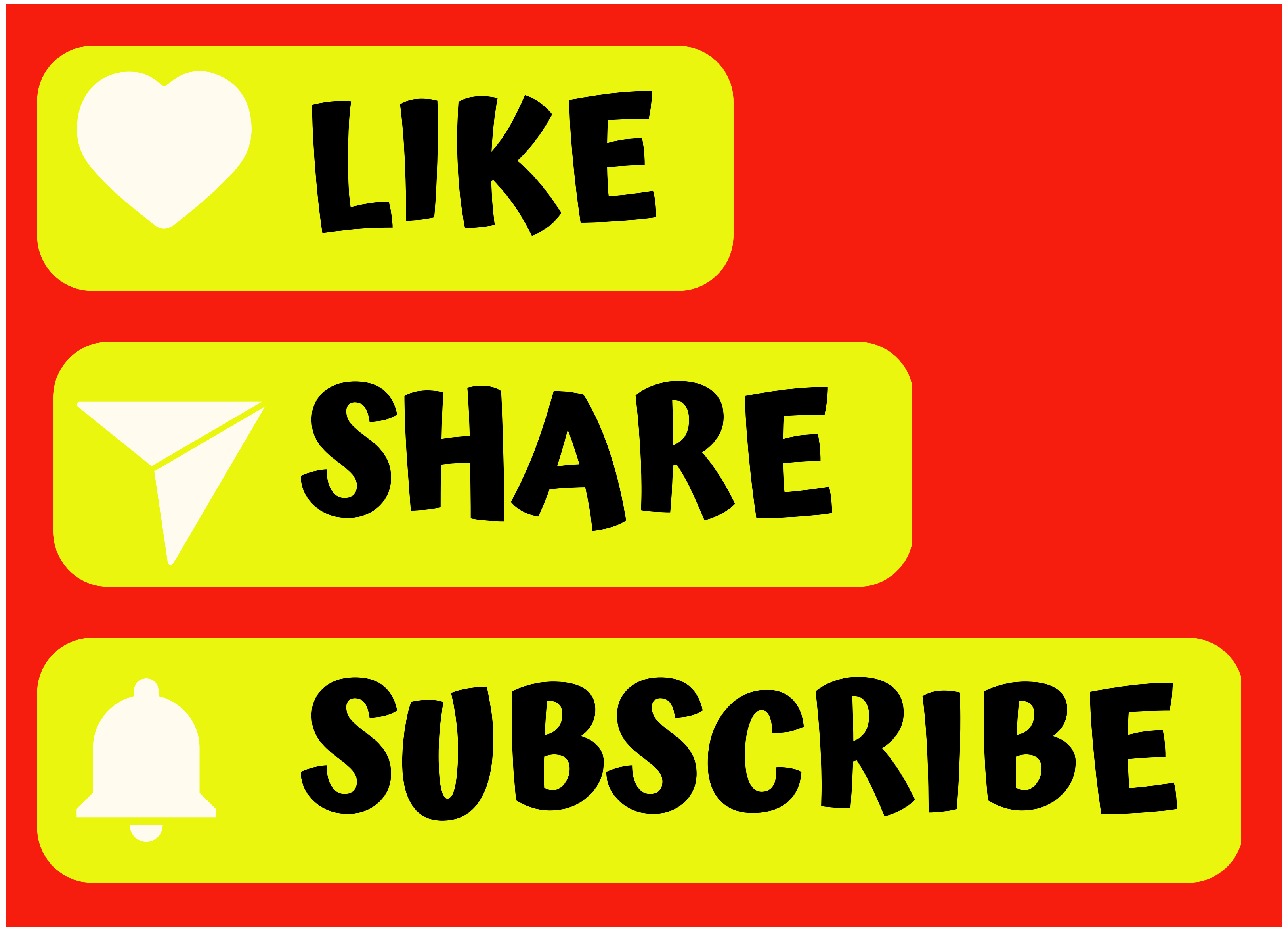
HTTPS://GITHUB.COM/INSIGHTBUILDER

WHAT NEXT ???



MANAGING THE ETL USING DBT (DATA BUILD TOOL)

HTTPS://GITHUB.COM/INSIGHTBUILDER


LIKE
SHARE
SUBSCRIBE