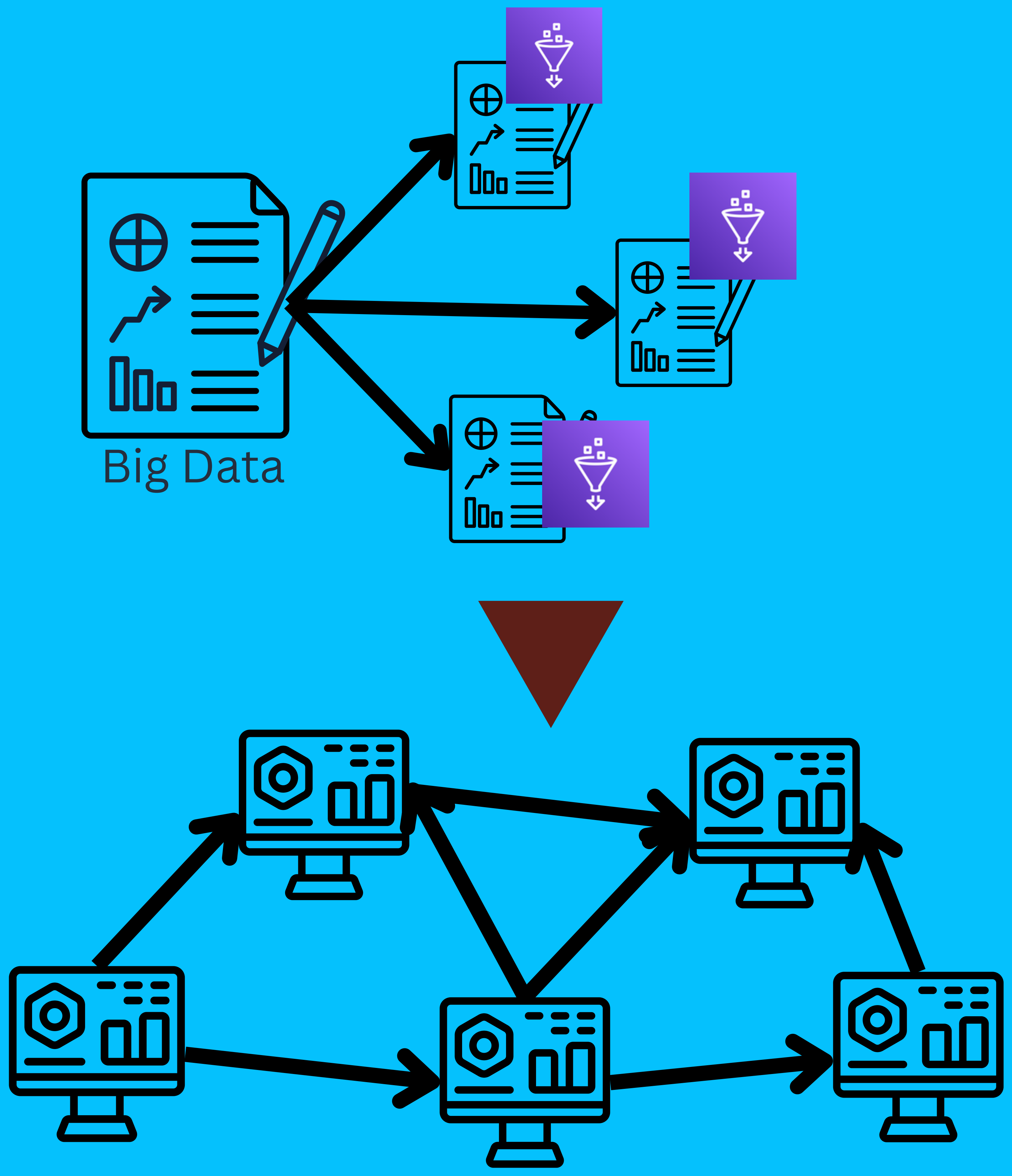
PYSPARK FOR INSIGHT BUILDERS: A PRACTICAL REVIEW

Hands On Tutorial using Jupyter Notebooks on Cloud

HTTPS://GITHUB.COM/INSIGHTBUILDER

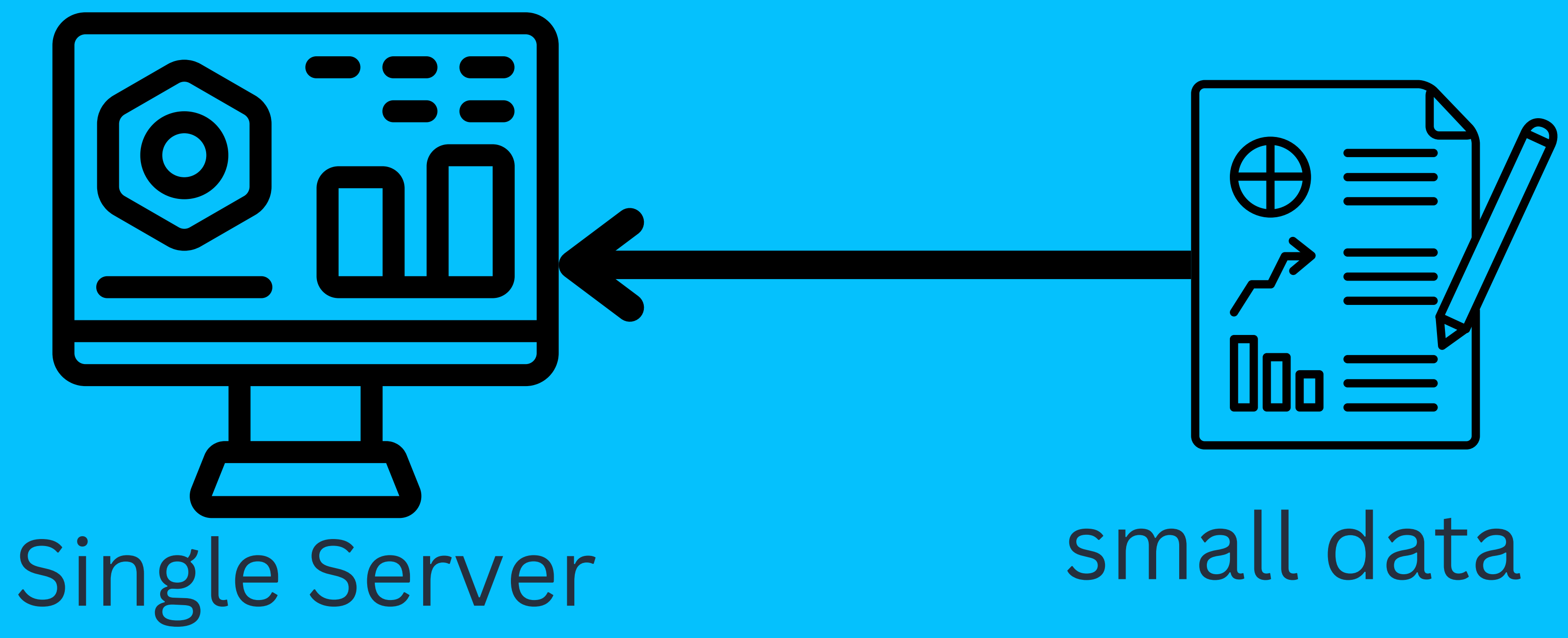
What is Big Data



A Lay man Definition

A file or Data that cannot be contained inside a RAM of one system

File that is bigger than the RAM will crash the program, when reading the data.



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. CREATE OURSELF A PLATFORM TO PRACTICE BIG DATA CONCEPT
2. UNDERSTAND THE BIG DATA CONCEPT AND THE ETL WORKFLOW
3. IMPLEMENT ETL USING PYSPARK AND SPARK SQL FROM ANY SOURCE TO DESTINATION
4. BECOME FAMILIAR WITH SPARK METASTORE AND USE IT FOR YOUR ADVANTAGE
5. BE COMFORTABLE WITH BIGDATA FILE TYPES LIKE PARQUET AND CSV

HTTPS://GITHUB.COM/INSIGHTBUILDER

WHAT WE WILL DO

1) Starting Spark Context

2) Overview of Spark read & write APIs

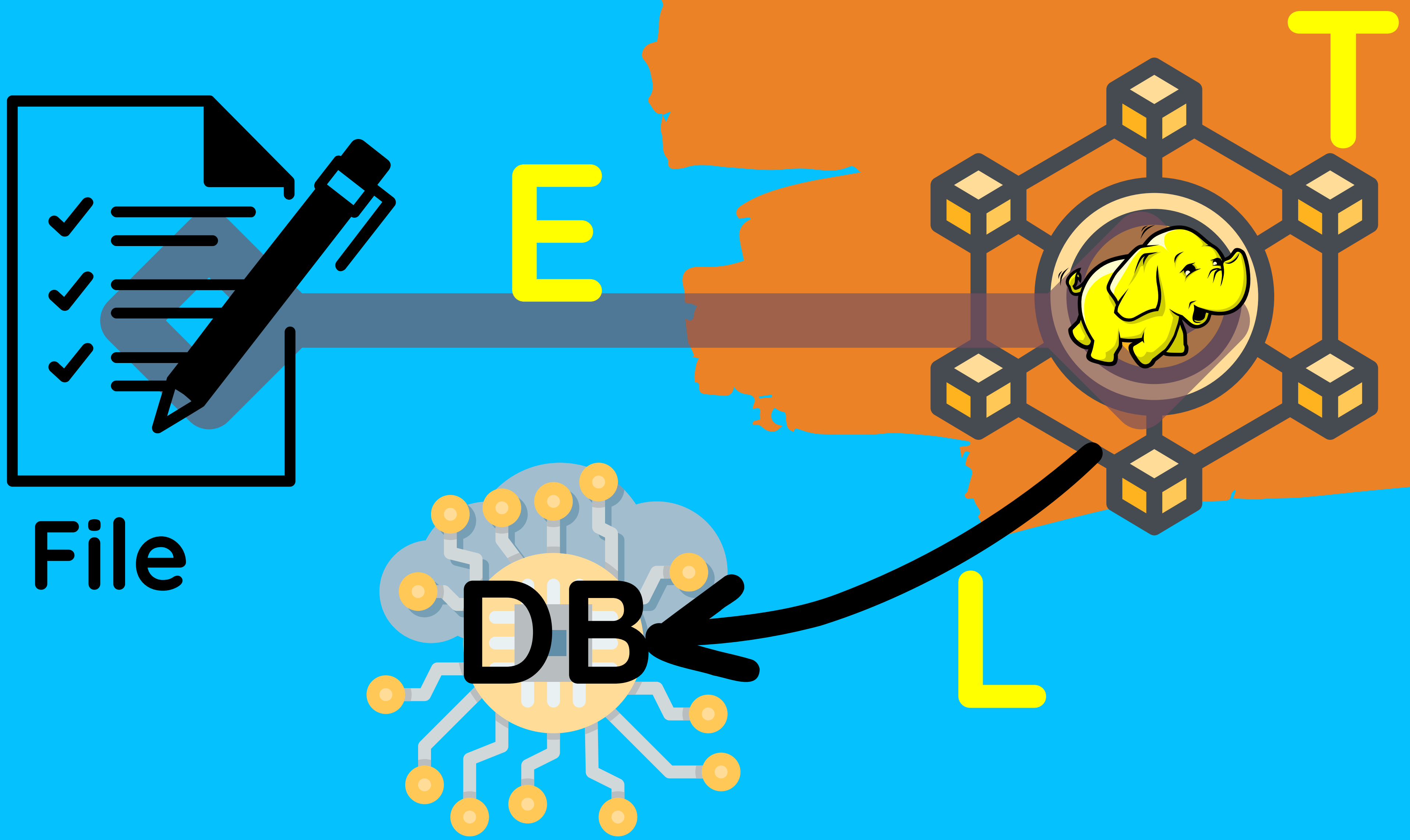
3) Inferring Schema

4) Creating Partitions

5) Overview of Data Frame API

6) Working with Spark SQL

7) Working with Spark Metastore



Spark Core - RDD and Map Reduce APIs

Spark Data Frames and Spark SQL

Spark Structured Streaming

Spark MLLib (Data Frame based)

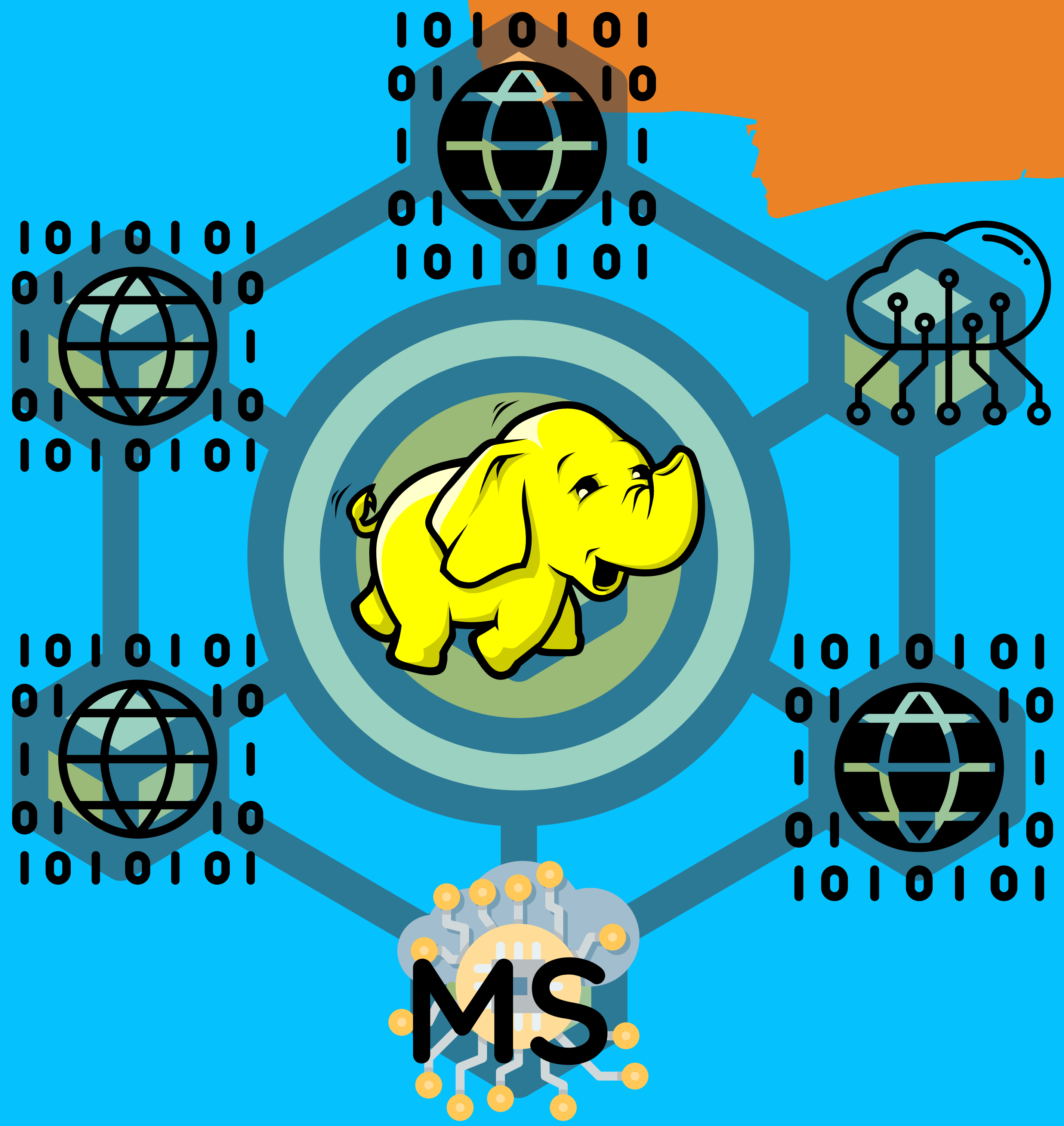
SPARK SESSION : OUR HERO

HTTPS://GITHUB.COM/INSIGHTBUILDER

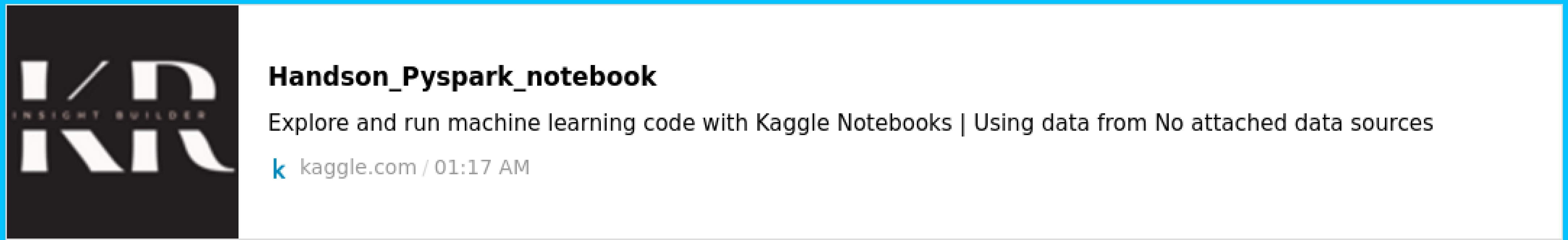
WHAT IS METASTORE

- A DATABASE that stores the location, dataset details that is stored inside the datawarehouse

- Every warehouse will have it for keeping track of file locations, so the apps like Spark/Hive/Kafka/Superset can locate it



HEAD TO kaggle / LOCAL NOTEBOOK



Handson_Pyspark_notebook

Explore and run machine learning code with Kaggle Notebooks | Using data from No attached data sources

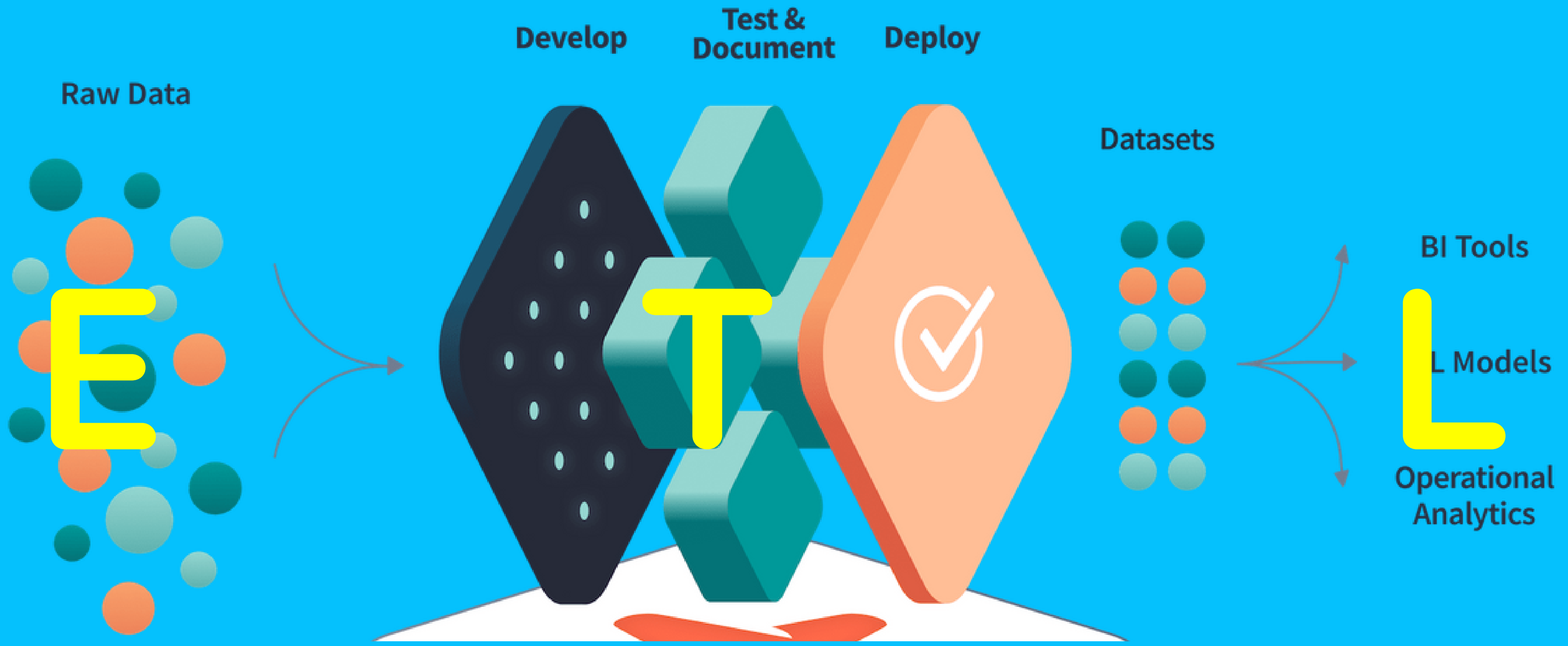
k kaggle.com / 01:17 AM

https://www.kaggle.com/code/kamaljp/handson-pyspark-notebook

QUESTIONS AND COMMENTS

HTTPS://GITHUB.COM/INSIGHTBUILDER

WHAT NEXT ???



PYSPARK ETL USING LARGER DATASETS

HTTPS://GITHUB.COM/INSIGHTBUILDER

LIKE

SHARE

SUBSCRIBE