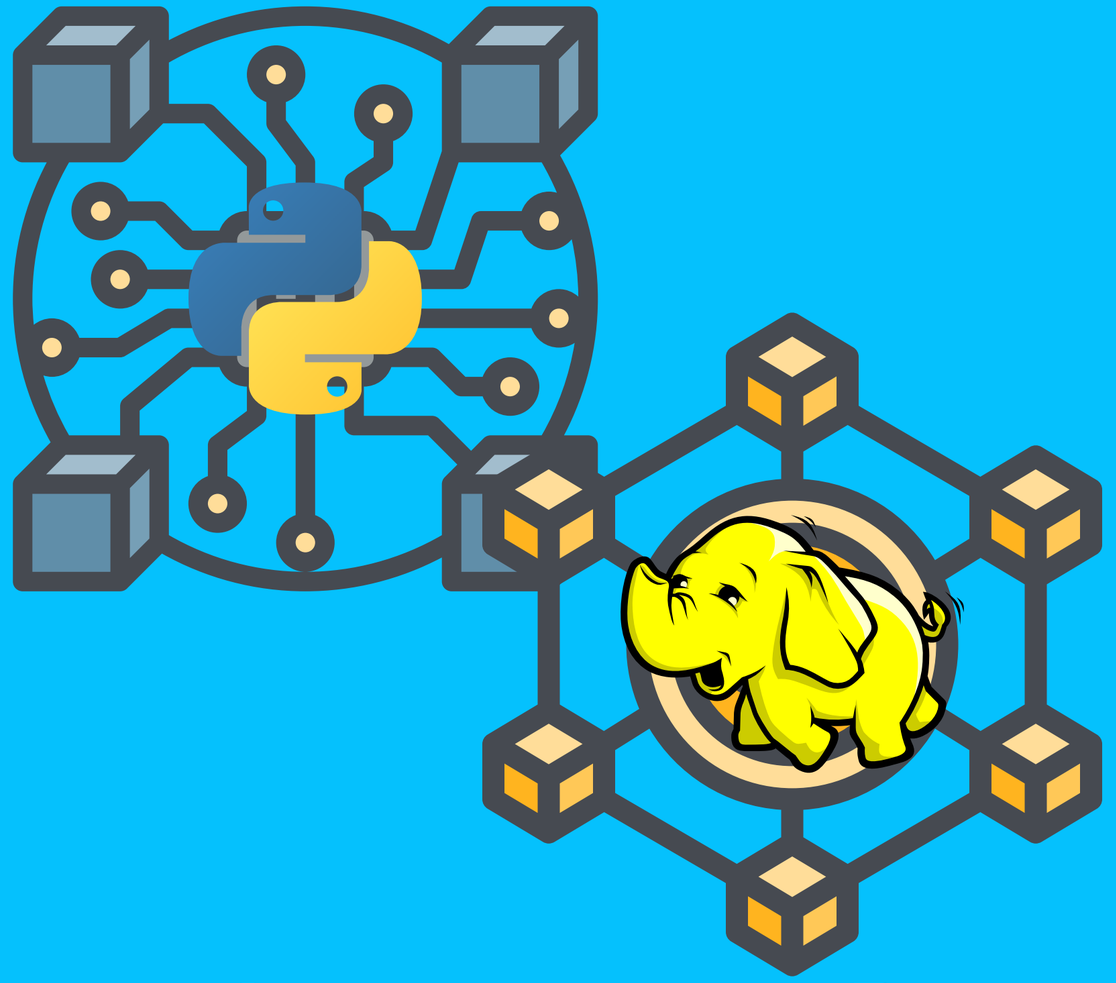
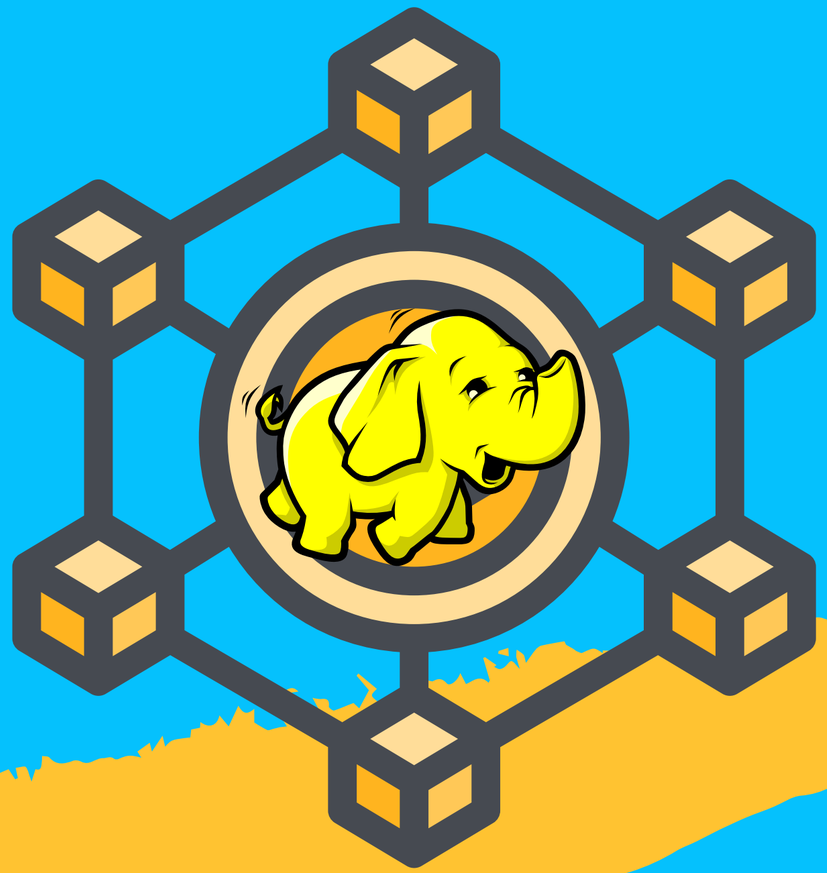
BUILDING BIG DATA APPLICATIONS WITH PYSPARK



From Data to Insight:using Pyspark

https://github.com/insightbuilder



WHAT PROBLEM WE'RE FACING

MAKING PYTHON & HADOOP TALK

Hadoop has the files distributed across many nodes, while python is capable of working on data in memory of single node(computer). Hadoop, Spark, Scala all work well in Big Data. Not Python

SERIES OF PROBLEM

- **WHEN THE FILE SIZE IS ABOVE RAM SIZE**
- **PYTHON NOT IDEAL FOR DISTRIBUTED WORK**
- **PYTHON TAKES LOT MORE MEMORY**
- **NO EFFICIENT COMPRESSION ALGO**

PYTHON ECOSYSTEM : PYSPARK

HOW TO GET IT INTO MY OS?

- Install Python
- The do "pip install Pyspark"

WHERE CAN I LEARN ABOUT STREAMLIT

- Pyspark Site

https://spark.apache.org/docs/latest/api/python/index.html

- SQL Basics

https://sqlzoo.net/

- Big Data Basics

https://spark.apache.org/docs/latest/cluster-overview.html

GUEST STARRER
AWSWRANGLER

HTTPS://GITHUB.COM/INSIGHTBUILDER

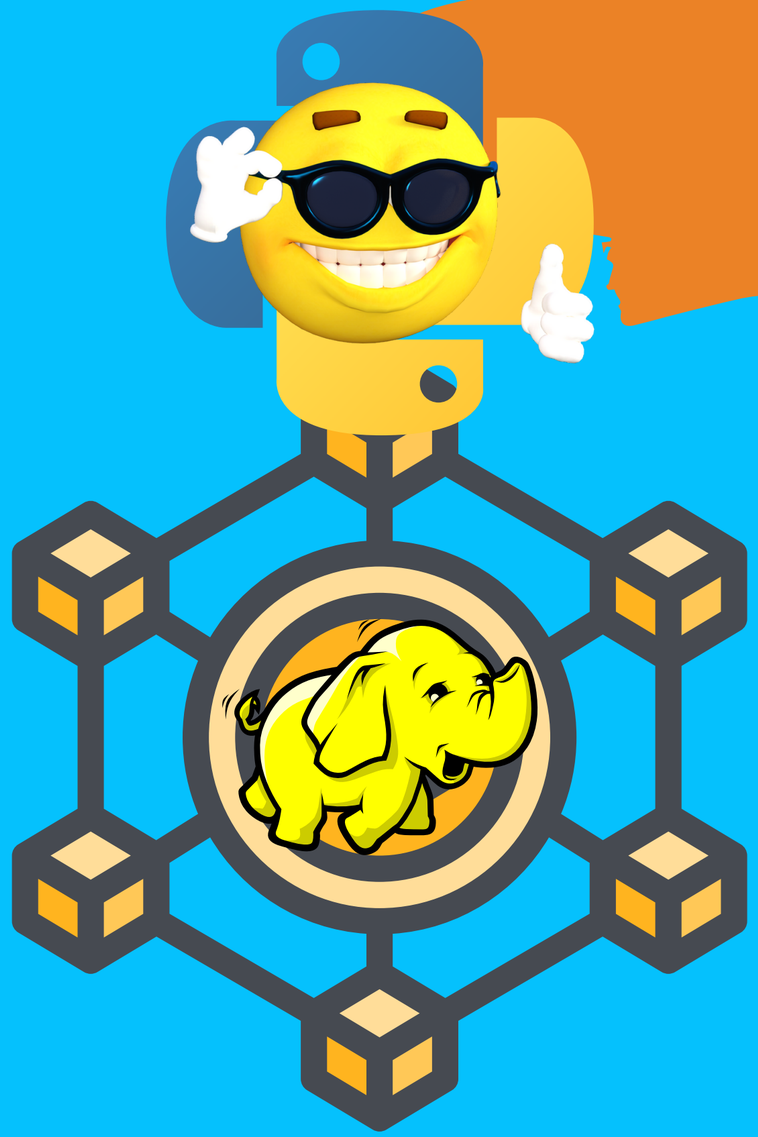
WHERE TO START

01 CREATE SPARK SESSION

- Spark session establishes a JVM inside the cluster, and pyspark connects to it.
- Data is read using the read methods. All Big data formats work

02 USING SPARK SQL

- Once data is readed, the most SQL concepts mostly work.
- Constraints don't work, since it is not implemented.



SPARK SESSION : OUR HERO

HOW DOES PYSPARK HELP

01

TWO WAYS TO QUERY DATA

- Pyspark's Select method can extensively query the data
- In-built spark-sql methods can also be used

02

CONNECTING TO DATA

- Read it through code
- Can use JDBC connectors downloaded from maven

03

EXTENSIVE TRANSFORMATIONS

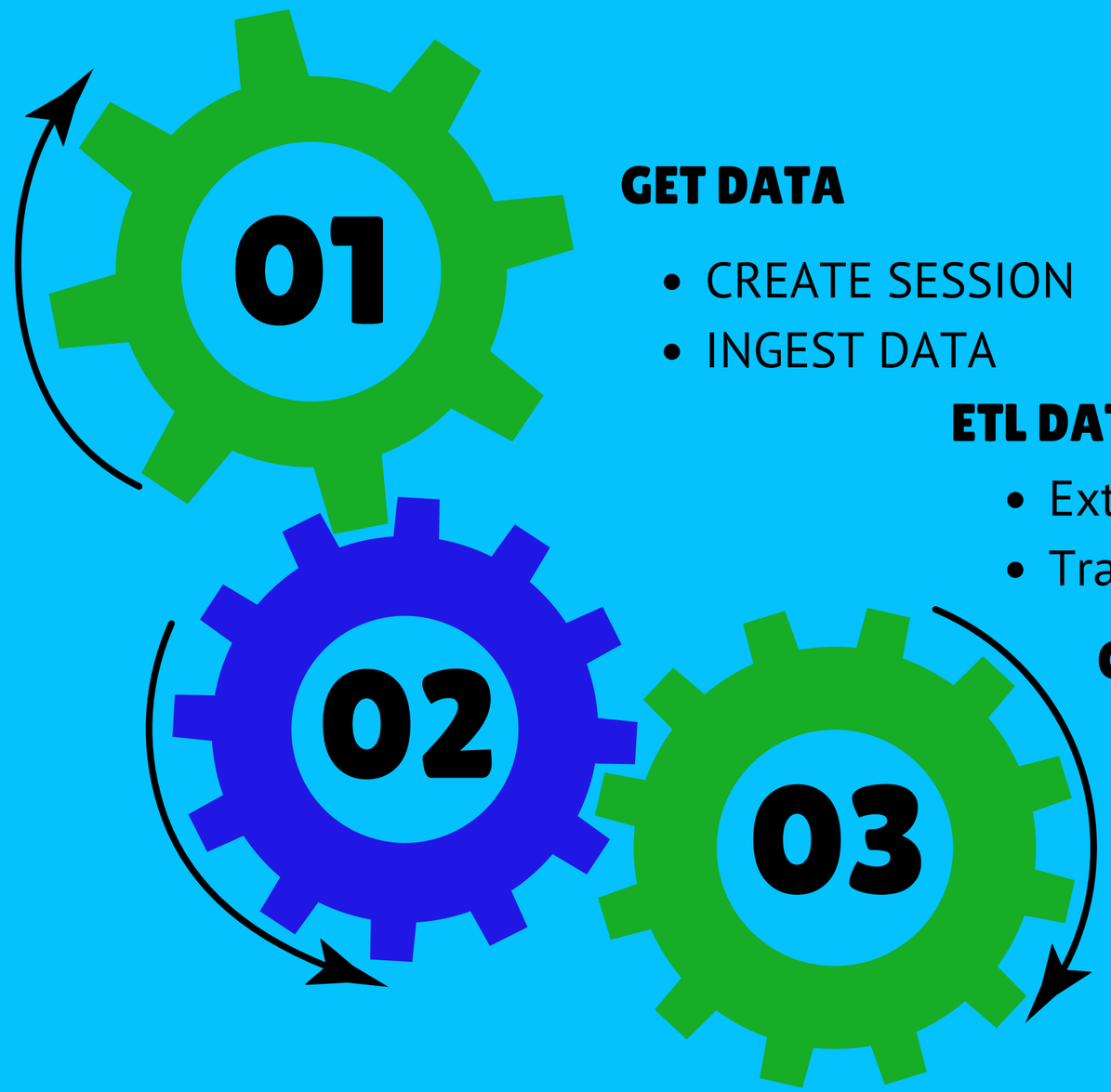
- Almost all of the sql transformations can be used
- Automate the ETL and implement pipelines

04

PARTITIONING & COMPRESSION

- Both parquet and csv files can be partitioned and written
- Writing to remote server can be configured

PYSPARK PROCESS



GET DATA

- CREATE SESSION
- INGEST DATA

ETL DATA

- Extract necessary data
- Transform the data and backup

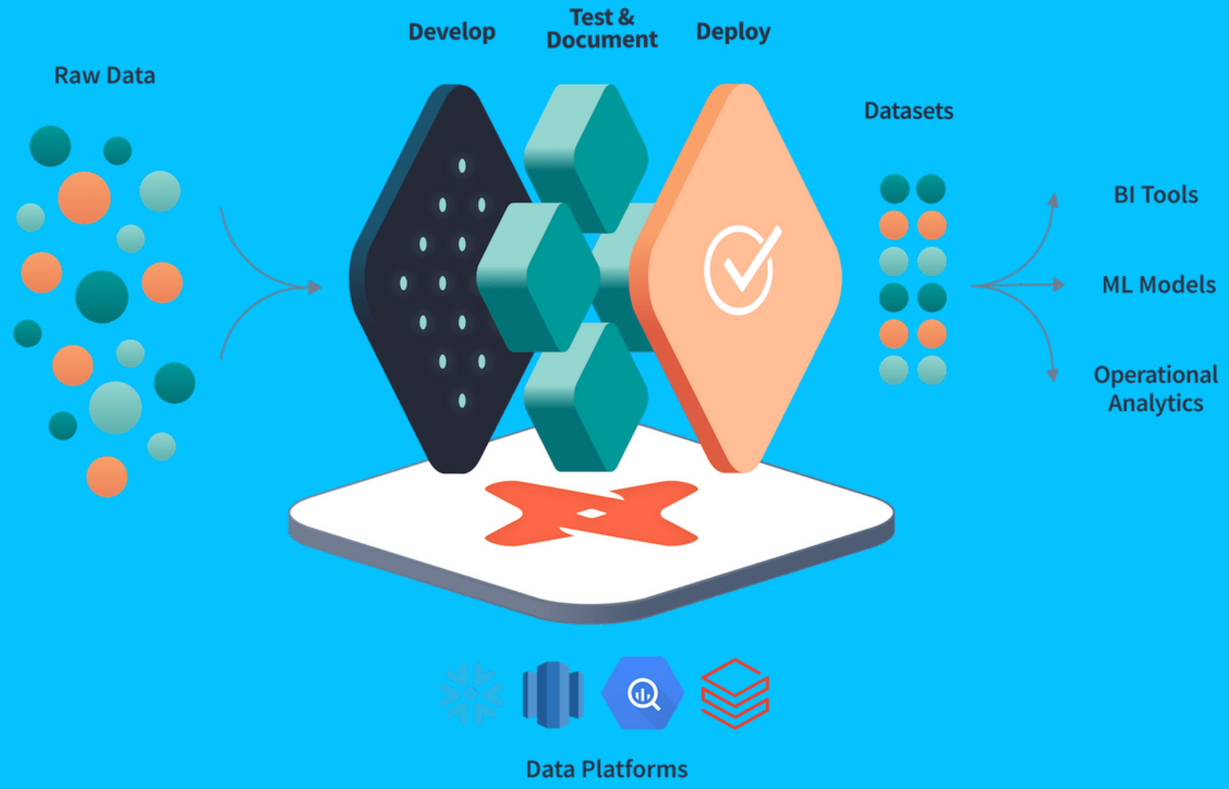
CLUSTER MANAGEMENT

- Cluster managers like yarn take care of it
- Provide access to stake holders by writing to Database or allow SSH

HEAD TO JUPYTER

QUESTIONS AND COMMENTS

WHAT NEXT ???



PYTHON IN DATA BUILD

STARRING DBT

https://github.com/Kamalabot/moreDE