BIG DATA CLUSTER DEMYSTIFIED

Advanced INTRO

Exploration of Configuration & Hadoop File System

What is Big Data

Single Server

small data

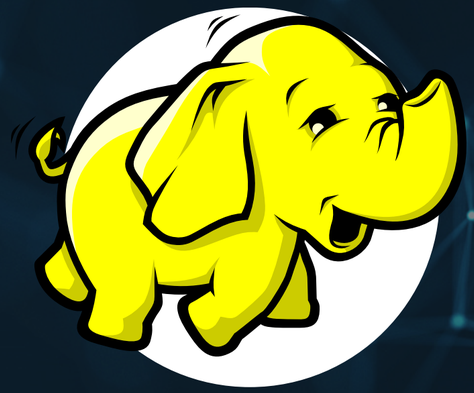
A Lay man Definition

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.

The Ecosystem

Its all written by humans for humans...



File System

HADOOP

Query Engine

Spark

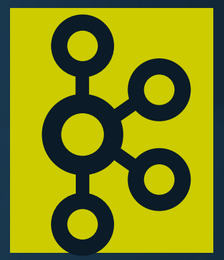
Programming

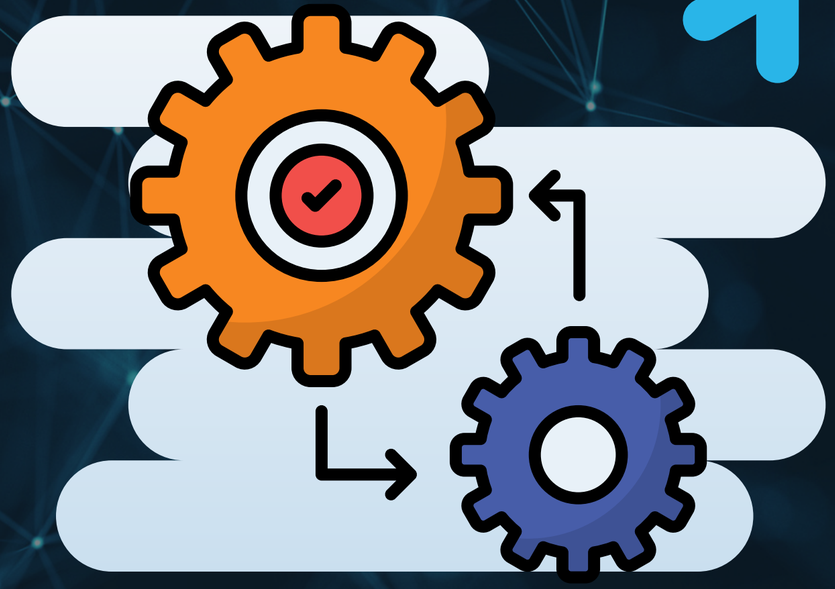
Python

HTTPS://GITHUB.COM/INSIGHTBUILDER

Multi Purpose + Multi Project

Different Warehouses : Same Architecture





ETL + LEARNING

INTEGRATION

HTTPS://GITHUB.COM/INSIGHTBUILDER

WHY EXPLORE CLUSTER

- WE HAVE TO KNOW HOW THE CLUSTER IS ARCHITECTED SO THAT WE CAN PLACE OUR DATA INSIDE IT
- MIGRATE YOUR DATA AND ETL PIPELINES WITH TO EMR.
- LEARN HOW THE GLUE CATALOG AND EMR INSTANCE INTERACT
- WHEN YOUR PYSPARK SCRIPT FAILS, YOU C DEBUG LOGICALLY
- TO DESCRIBE AND HANDLE THE TABLE DETAILS WITH PRECISION

I DON'T OWN A CLUSTER

YOU CAN BUILD ONE USING
THE BELOW PLAYLIST

YOU CAN FOLLOW ALONG THIS
VIDEO USING AWS EMR CLUSTER

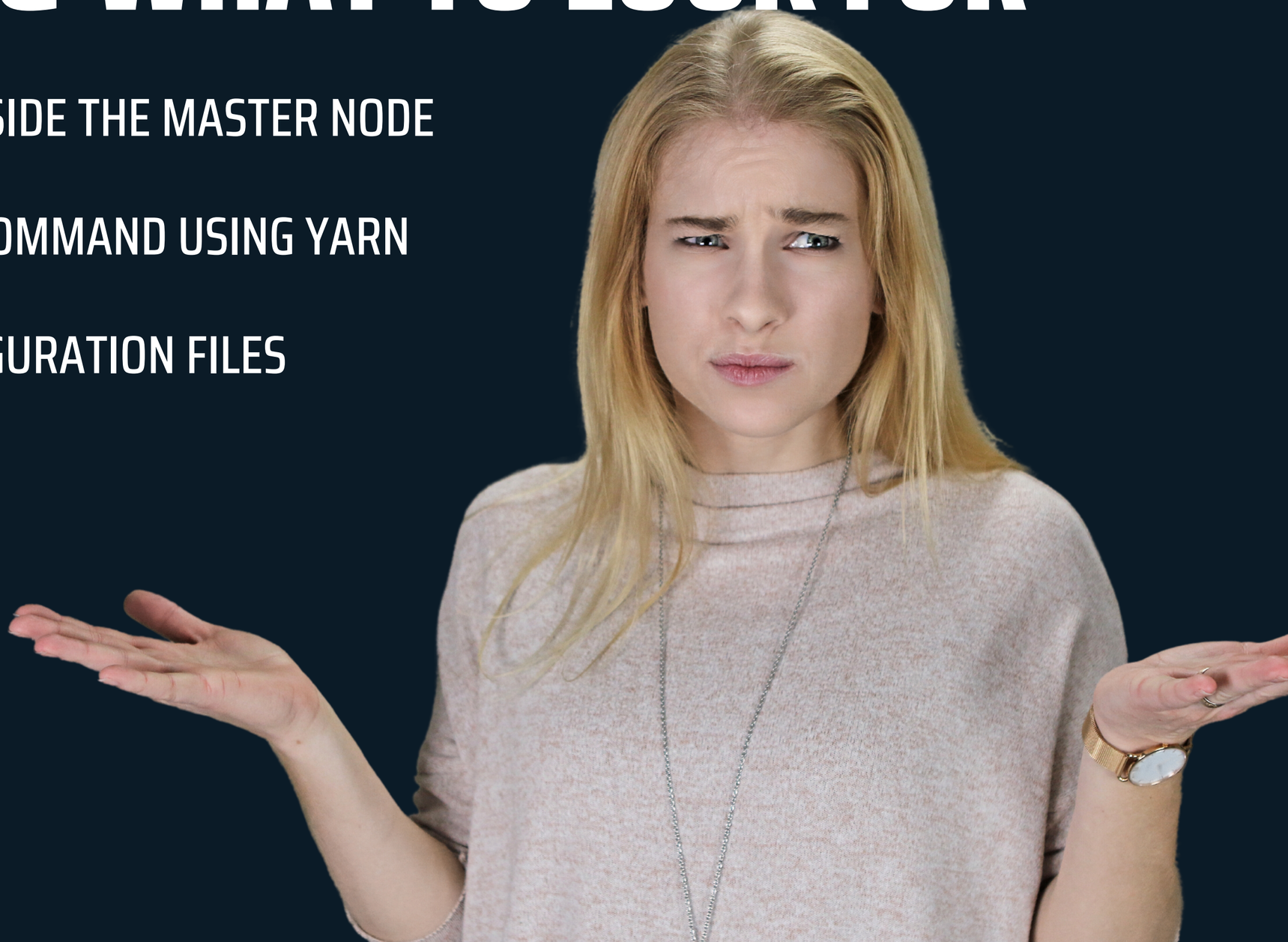
WHERE & WHAT TO LOOK FOR

START BY THE .PROFILE FILE INSIDE THE MASTER NODE

THEN EXPLORE THE PYSPARK COMMAND USING YARN

LOCATE THE FOLLOWING CONFIGURATION FILES

- HDFS-SITE.CONF
- HADOOP-ENV.SH
- CORE-SITE.CONF
- HIVE-SITE.CONF
- YARN-SITE.CONF
- SPARK-DEFAULTS.CONF
- SPARK-ENV.SH



WHAT YOU WILL LEARN

WHAT IS THE IP OF YOUR WORKERNODE, SO YOU CAN CONNECT WITH THEM FROM THE MASTER

WHICH DATABASE IS USED FOR GLUE CATALOG IN YOUR CLUSTER

DESCRIBE THE TABLES IN YOUR GLUE CATALOG AND UNDERSTAND THEM

HOW TO ADD NEW TABLES TO GLUE CATALOG USING YOUR CODE (ONLY TO EXPERIMENT)

EXPERIMENTING WITH MULTIPLE SPARK CONFIGURATIONS AND EXPERIMENT

MIGRATION AND RECOVERY OF YOUR CLUSTER BECOMES EASIER

READ AND UNDERSTAND THE HELP PROVIDED BY AWS EMR

USE AWS WRANGLER PYTHON LIBRARY AND INTERACT WITH EMR

LETS GET OURSELF A CLUSTER AND DIG IN

THANKS FOR WATCHING

PRACTICE

PRACTICE

PRACTICE

PRACTICE

HTTPS://GITHUB.COM/INSIGHTBUILDER