

# BUILD BIG DATA CLUSTER IN CLOUD



# What is Big Data



Single Server

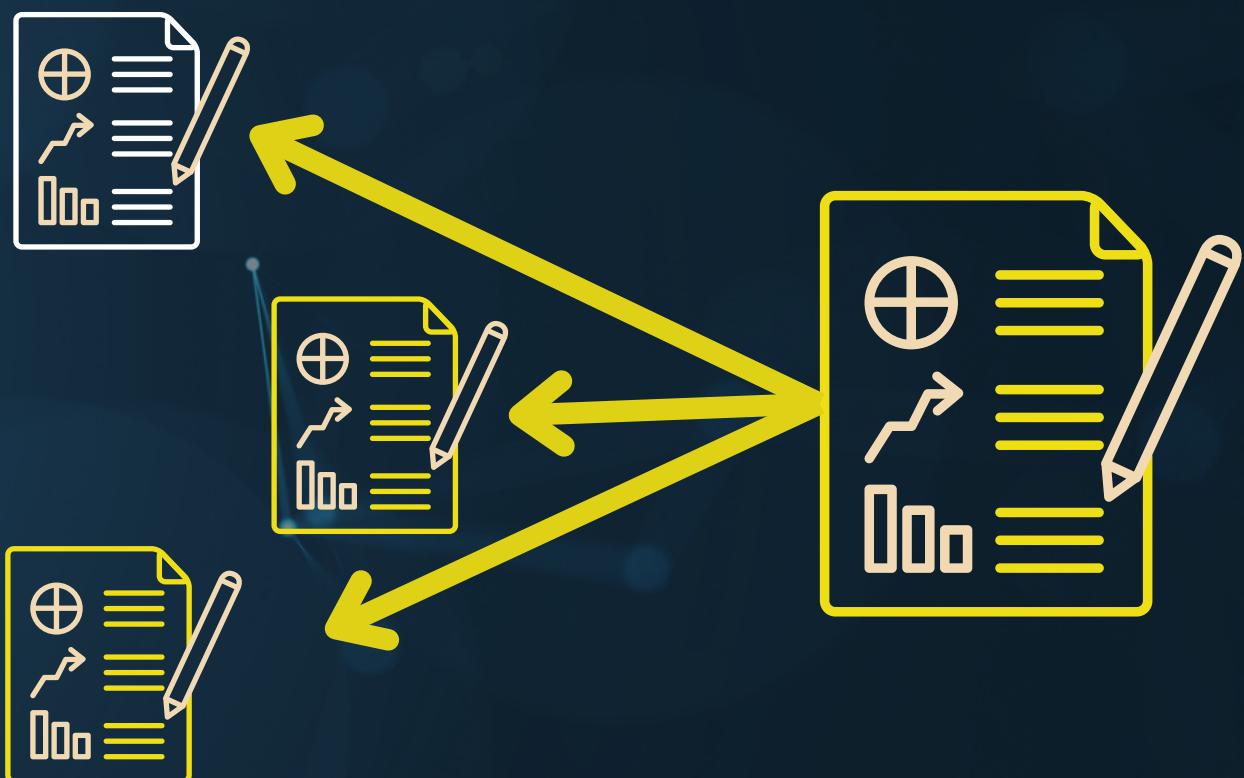


small data

## 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.





# Why BYOD ???

- Clarity on the Ecosystem
- Skillset to build Local Environment
- How each application connect + talk
- Faster Recovery and Troubleshooting
- Can touch and move the raw data
- See through the Technology Abstraction



# Steps To Build

## Get your Cloud

Sign up for any of the Cloud Services

## Understand Compute

2 VCPUs with 8GB of RAM  
22GB of RAM

## Get ubuntu on Compute

Setup passwordless login inside Linux

01

02

03

04

05

06

## Setting up Python

Setting up Jupyterlab and Pyspark and ensure connectivity

## Download installers

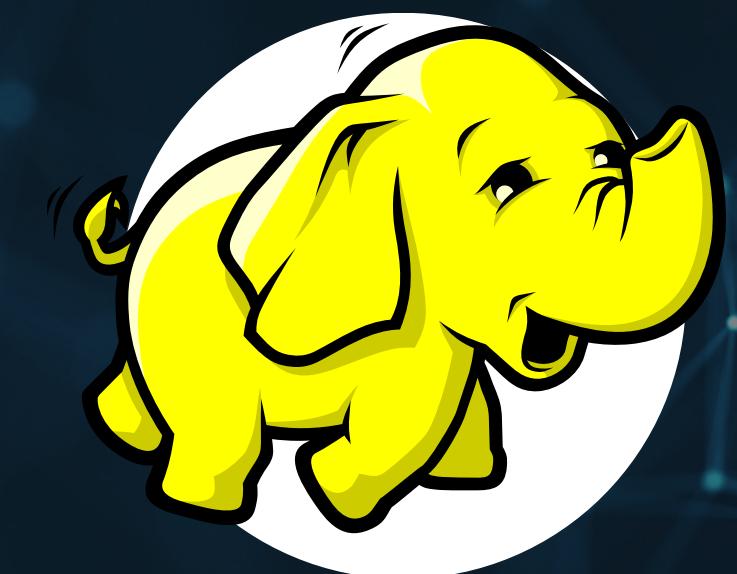
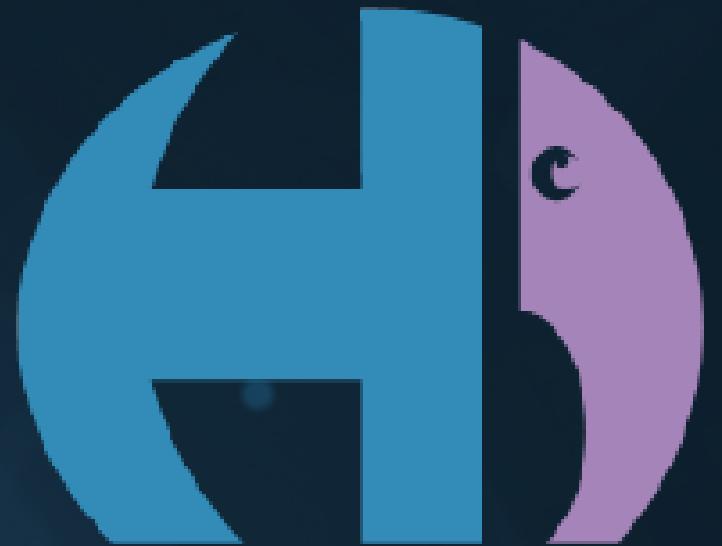
Hadoop, Hive, Spark3 and Postgres Docker

## Initiate the installation

Requires good grasp of Linux commands

# The Ecosystem

Its all written by humans for humans...



File System

HADOOP



Query Engine

Spark



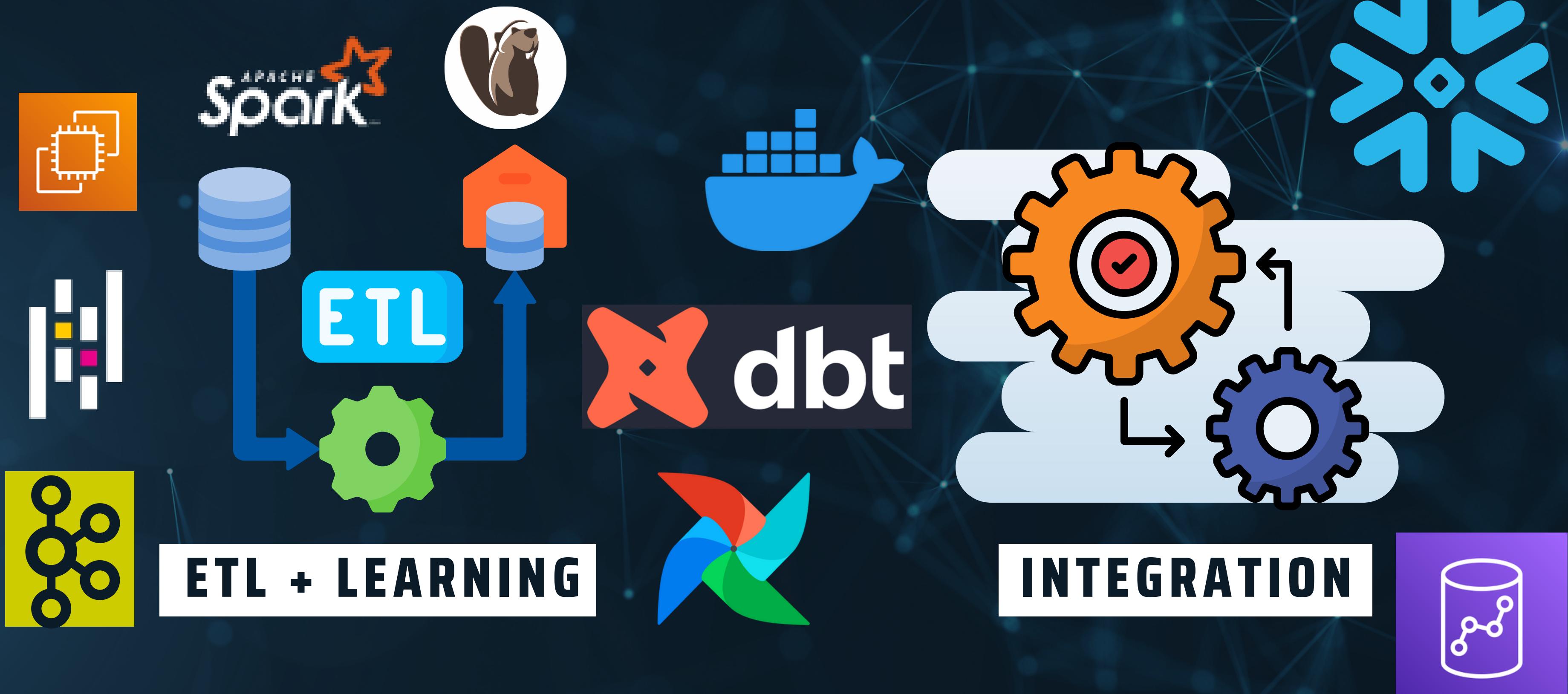
Programming

Python



# Multi Purpose + Multi Project

# Different Warehouses : Same Architecture



# THANKS FOR WATCHING

PRACTICE

PRACTICE

 LIKE

 SHARE

 SUBSCRIBE

PRACTICE

PRACTICE