DATA ENGINEERING

Dinda Adilfi Wirahmi

What is Data Engineer?

Engineer who have goal to solve problem
of data from engineering side

3 MAIN PART OF DATA ENGINEER

DISTRIBUTED SYSTEM

Handling big data, collection of task that located in different machines that share message each other to achieve problem goals

PROGRAMMING

ANALYSIS

Is there a problem with the pipeline? Why the records of data consist null value? Why data in strange format?

DATA ENGINEER METHODOLOGY

Data engineer handle big data, so it makes us to understand how to **automate** tasks. Data Engineer create data pipelines that connect data from one system to another and also responsible for data transformation. It can help data analyst/scientist to pull data from different systems for analysis. For example, if we want to running pipeline at certain time, we can make schedule for that instead of running it manually.

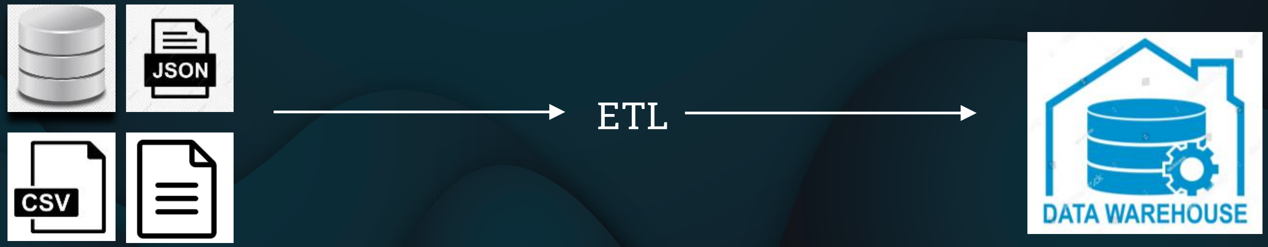
WORKFLOW

Extract – Transfrom – Load (ETL)

Extract : collect data from any source (database, csv, json file, etc)

Transform : cleaning data, standarize data, join several tables or any other tansformation techniques

Load : store data that has been extract and/or transform



So, what is data warehouse?

Data warehouse is large collection of business data used to help an organization make decisions

Example of Data Engineer – Big Data Tools

AIRFLOW

Scheduling and running data pipelines

Spark

Processing big data

DOCKER

Containers to make it easier to create, deploy, and run applications.

HADOOP

To efficiently store and process large datasets

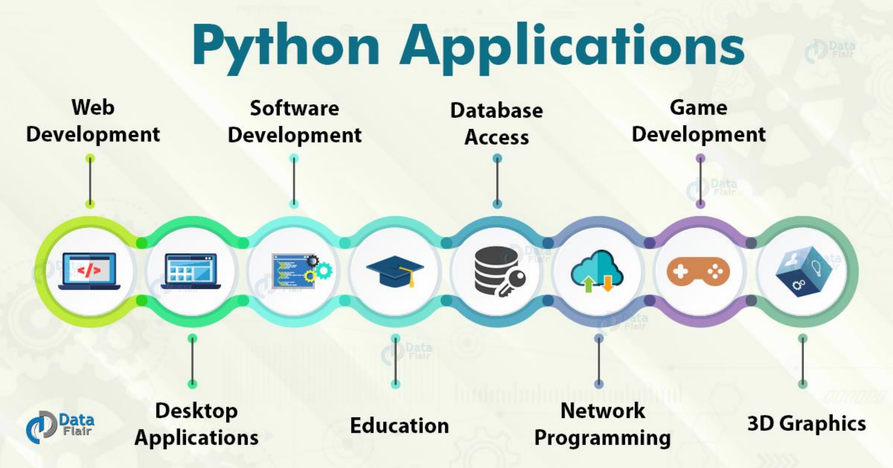
SQL AND NO SQL

Database

MANY MORE

PYTHON

One of main part for data engineering is programming. **Python** is the most common language.



https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2018/02/python-applications.jp

THANKS!