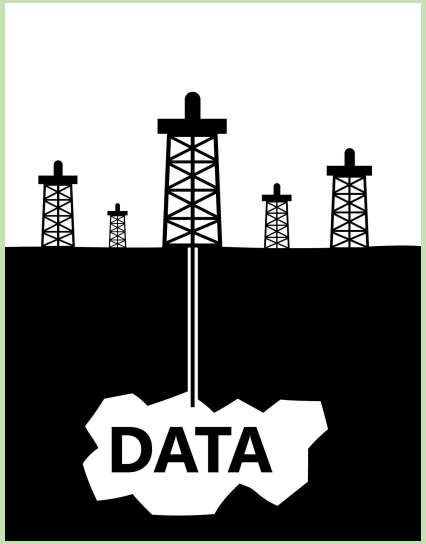
LEARNING PROGRESS REVIEW

Data Engineer Batch 9 - Week #1

Introduction to Data Engineering & Python Basics

Dwi Handoyo

DATA ENGINEERING BACKGROUND: BIG DATA



★ Data is the new oil
* Abundant data available (Big Data) can be utilized for business purpose

★ Data resources :
* Application/API
* Database
* Internet/Web

★ Data utilization:
* For supporting business process
* For getting business insight
* For predictive analysis and realtime smart system

DATA ENGINEERING BACKGROUND: BIG DATA

Big Data Characteristics (4V)

* Volume : A very large size of data
* Velocity : Enormous speed of data being generated and processed
* Variaty : Various data structures and data resources
* Veracity : Data integrity is not fully guaranted

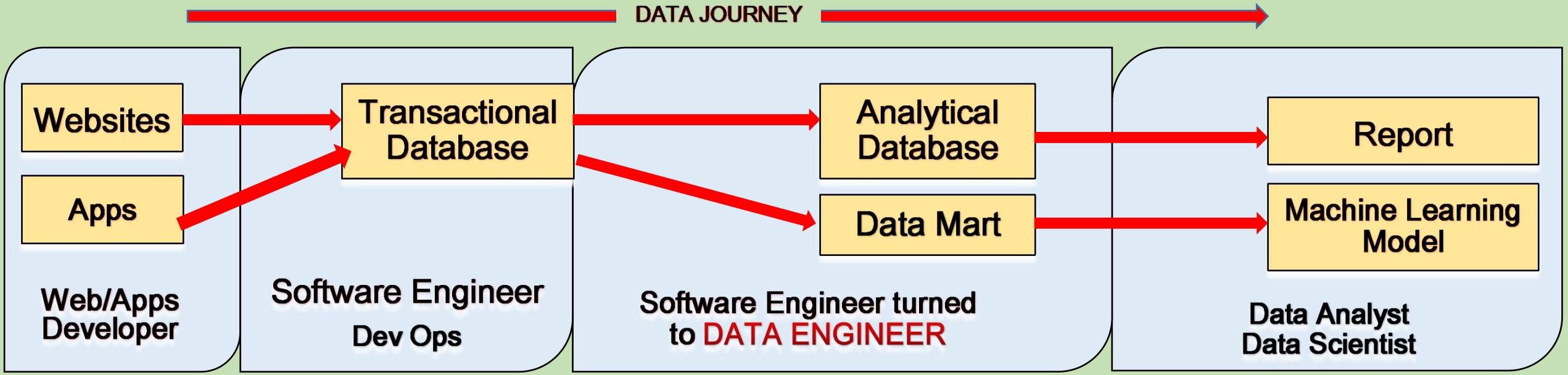
Big Data Problems

* Memory and processing time issue due to data volume
* Not all data can be processed due to data velocity

Special ways/techniques required to collect, store and process big data. A specialist is needed to do the job for analytical data infrastructure. Data Engineer is the one who fit for the job.

DATA ENGINEERING BACKGROUND: DATA JOURNEY

Data journey is data processing from data collecting up to data reporting and/or data modelling for analytical purpose.



As data became bigger and more complex (Big Data 4V), Software Engineer scope of work in developing analytical database/data mart shifted to Data Engineer.
A Data Engineer is responsible for bridging the Tech Team (Back End) with the Data Team. Data Engineer specialize in developing analytical data infrastructure and also perform deployment and monitoring of predictive model.

DATA ENGINEER AREA OF RESPONSIBILITIES

By definition, data engineer is a specialist who responsible in providing data architecture for analytics and operation purposes of a business entity. Other equal titles for data engineer are:

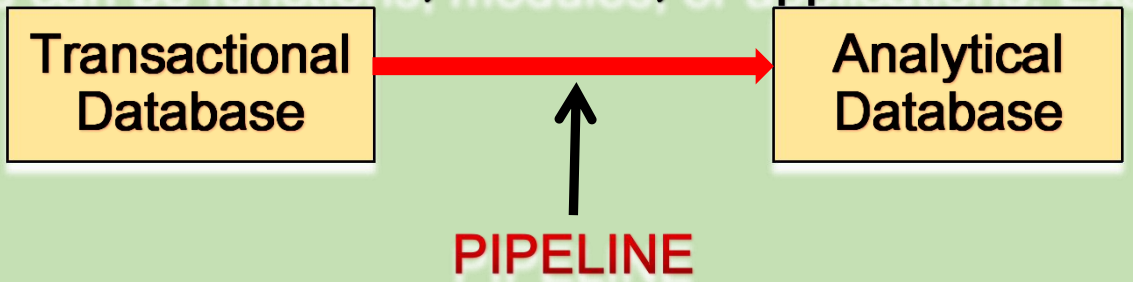
* Data Architect
* Data Platform Engineer
* Business Intelligent Engineer
* Big Data Engineer

Data engineer scope of works are including:

* Data Extraction & Processing
* Data Warehousing
* Business Intelligence
* Data Automation
* Cloud Engineering

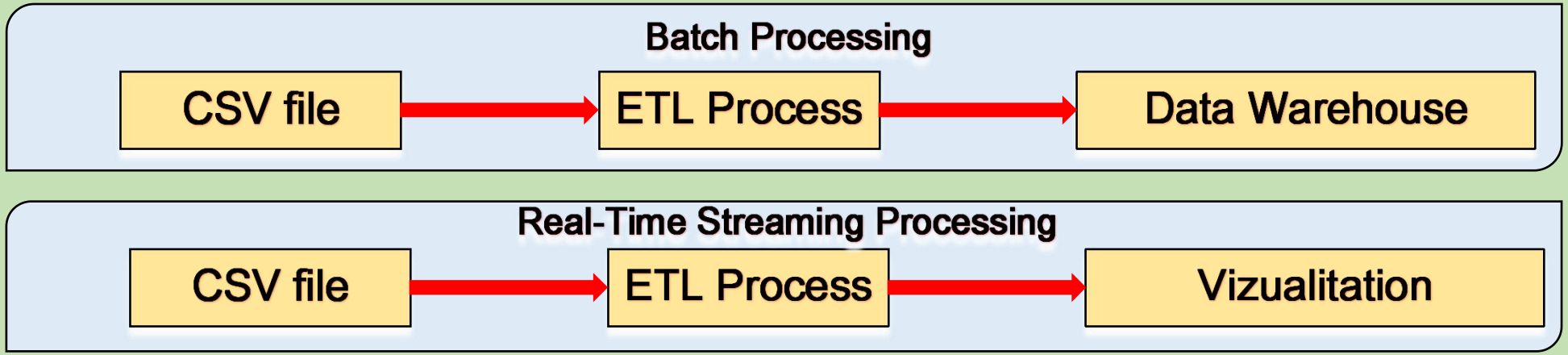
DATA EXTRACTION & PROCESSING

Data Engineer extract and process data from various resources via components called “pipeline”. Pipeline can be functions, modules, or applications. Example of pipeline:



ETL (Extract, Transform, Load) is a pipeline framework used to collect/extract and process data. The extracted source can be CSV file, database, or API.

Data processing types can be Batch Processing or Real-Time Streaming Processing



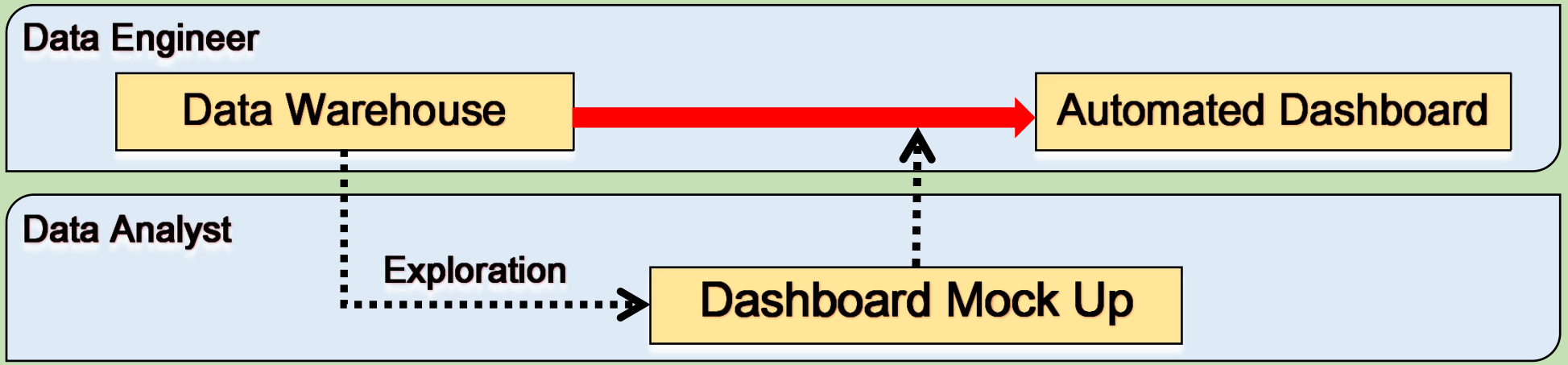
DATA WAREHOUSING

Data engineer shall design and develop data storage architecture by considering:

* Easy for getting any important information (insights) from the data
* Optimization of query time
* Optimization of storage size
* Ensure data security is in place

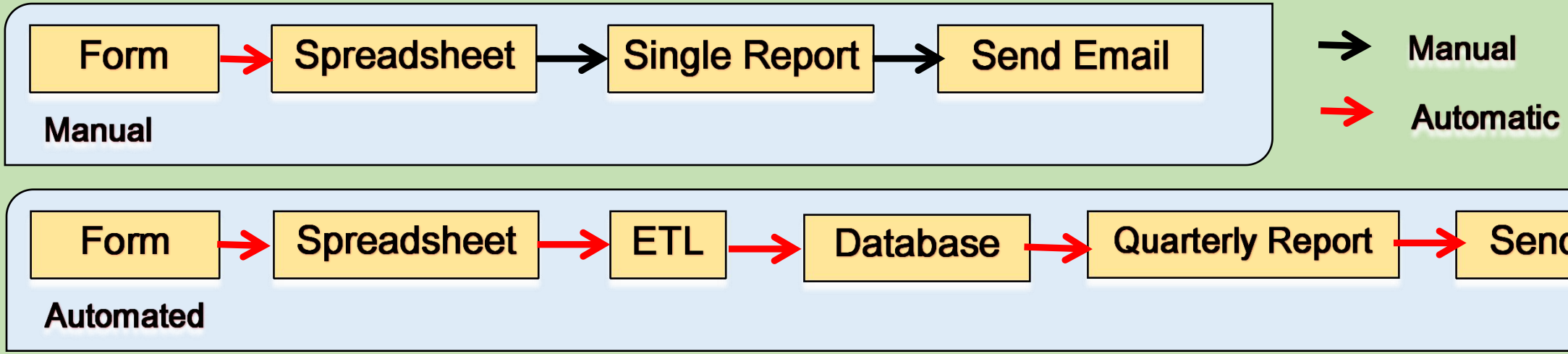
BUSINESS INTELLIGENCE

Data Dngineer shall transform data into an actionable insights, it can be a report or a dashboard. For example to create an automated dashboard based on a mock up dashboard prepared by a Data Analyst.



DATA AUTOMATION

Data engineer shall create job automation and getting insights from the data.



More comprehensive report such as Quarterly Report can be created using data automation

CLOUD ENGINEERING

If required, Data Engineer shall do data architecture migration from on-premise server to cloud servers. For example storing various data into data lake in a cloud service.

Cloud computing is computing service by using servers, storage, database, software, etc. owned by cloud service providers, such as Google Cloud Platform (GCP), Microsoft Azure, and Amazon Web Services (AWS).

DATA ENGINEERING TOOLS

DATABASE
* PostgreSQL
* BigQuery
* Amazon Redshift

ETL TOOLS & SCHEDULER
* Apache AIRFLOW
* Apache SPARK

STORAGE
* Cloud Storage
* Amazon Simple Storage Service (S3)

VIZUALISATION
* Looker
* Google Data Studio
* Metabase

TERMINOLOGY

Data Warehouse : Analytical database, cleansing already done and summarized, ready to use.
Data Mart : More specific than data warehouse, to be used by secific person or division, for example for the CEO.
Database : Table with rows and fields/columns.
Data Lake : Storage of various types of raw data.
Transactional Database : Database with traffic of data in and data out, for example back end of web application.
Analytical Database : Summaries of data for analytical purpose.
Extract : Collect data from sources (CSV file, database, API).
Transform : Data processing to comply with format and storage requirement.
Load : Data entry to storage (database).
Batch Processing : Data is processed priodically or with manual trigger (data source: files, database, API).
Real-Time Streaming Processing : Data is processed directly every time data enters the pipeline. Example implementation of streaming data is the application tracker data processing.

PYTHON BASICS

Python is widely used programming language in data engineering. Python was created by Guido Van Rossum. Its first version was released in February 20, 1991.

Variables

* snake_case
* All in lowercase
* Does not start with special characters like for example &(ampersand), $(dollars)
* Variable names can only contain the following characters A-Z, a-z, 0-9, and _
* If the name contains multiple words, it should be separated by an underscore (_) e.g. first_name
* Avoid single character variables such as a, b, c, etc.
* Variable names cannot start with a number
* Case sensitive

Comments & Docstring

* Comments are statements in code that are not executed, started with #
* Docstrings are strings written right after the definition of a function, method, class, or module, inside triple quotes

PYTHON BASICS

Primitive Data Types

* Integer (int), are all integers: 1, -2304, 0, etc.
* Float (float), are all decimal numbers: 23.5, -0.52, 3525.252, etc.
* String (str), are all letters and text: "a", "This is text. And another sentence.", etc.
* Boolean (bool), is True or False.

Non Primitive Data Types

* List - Characterics: mutable data (changeable), ordered, and allow duplicates
  list = [3,5,2,7]
* Tuple - Characterics: immutable, ordered, and allow duplicates
  tuple = ('p','e','r','m','i','t')
* Set - Characteristics: unchangeable (but items can be added/deleted), unordered/unindexed, and unique.
  set = {1, 2, 3}
* Dictionary - Characteristics: mutable, ordered, keys have to be unique
  dict = {1: 'apple', 2: 'ball', 3:'car'}

Date Data Types

Python has a module called datetime for working with dates and times

THANK YOU