DigitalSkola

Introduction to Data Engineer

By Big Three:

Ilham Ahsanudin

Ilyas Rizky Ibrahim

Harryyanto Ishaq Agasi

Table of contents

1 About Data Engineer

- What is Data Engineer?
- Data Engineer Responsibilities
- Data Engineer Workflow and Methodology

2 Big Data

- Data Engineer and Big Data
- Big Data Tools for Data Engineer

3 Python Introductory

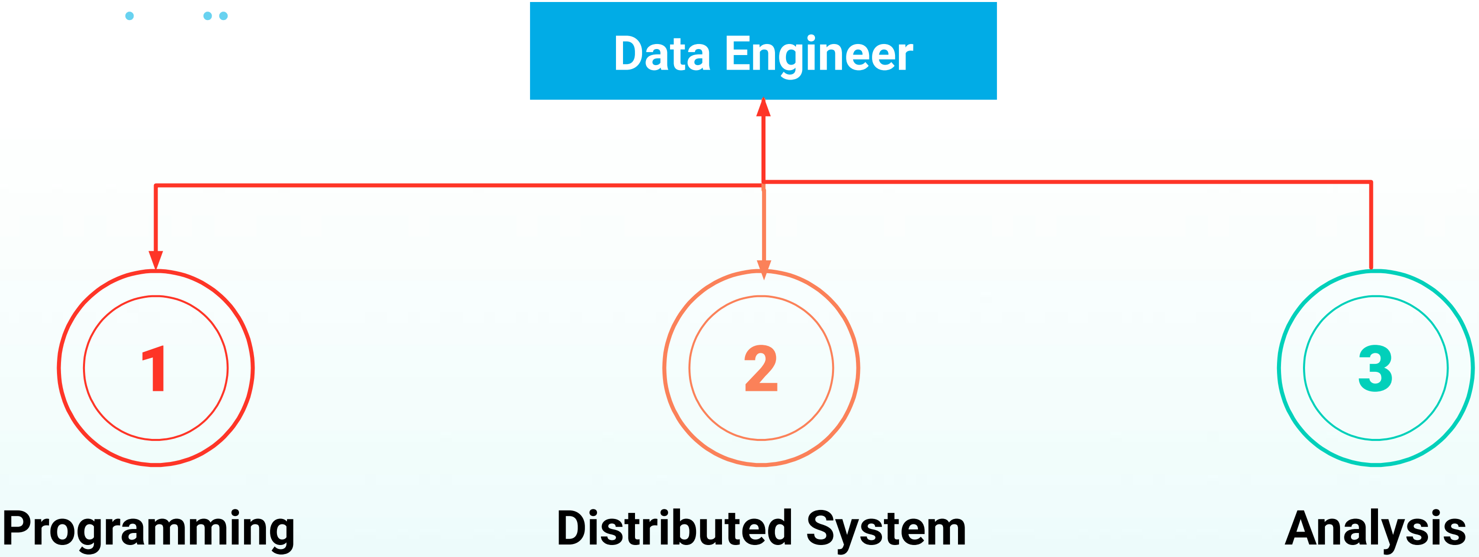
- Why choosing python?
- Variabel
- Data Types

1

What is Data Engineer?

Engineer who have responsible with data infrastructure within an organization

DigitalSkola



Data Engineer

1

Programming

2

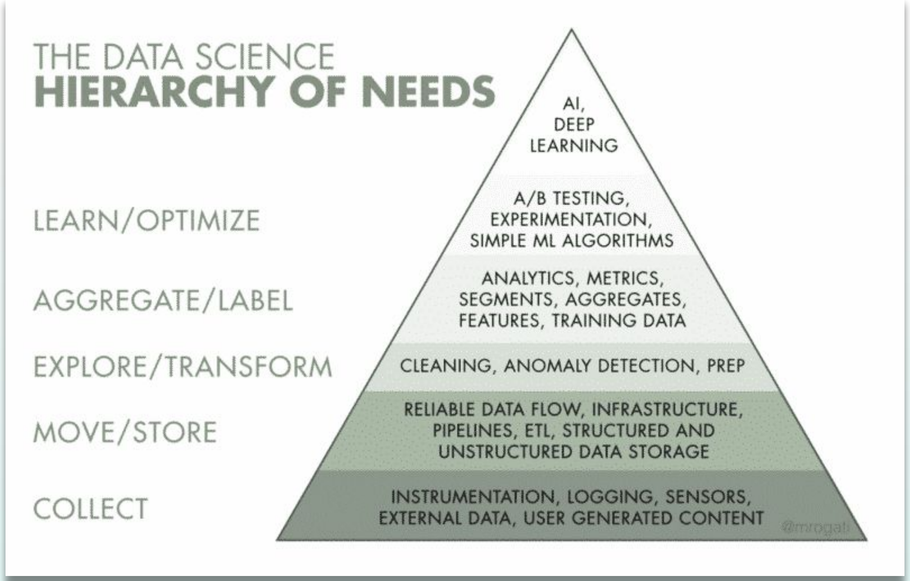
Distributed System

3

Analysis

Data Team Roles

A Data Engineer’s role is at the three bottom level; Collect, Move/Store, and Explore /Transform. A Data Engineer is like the main source in data team who provide data needs for another data role such as Data Analyst and Data Scientist.



THE DATA SCIENCE
HIERARCHY OF NEEDS

LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

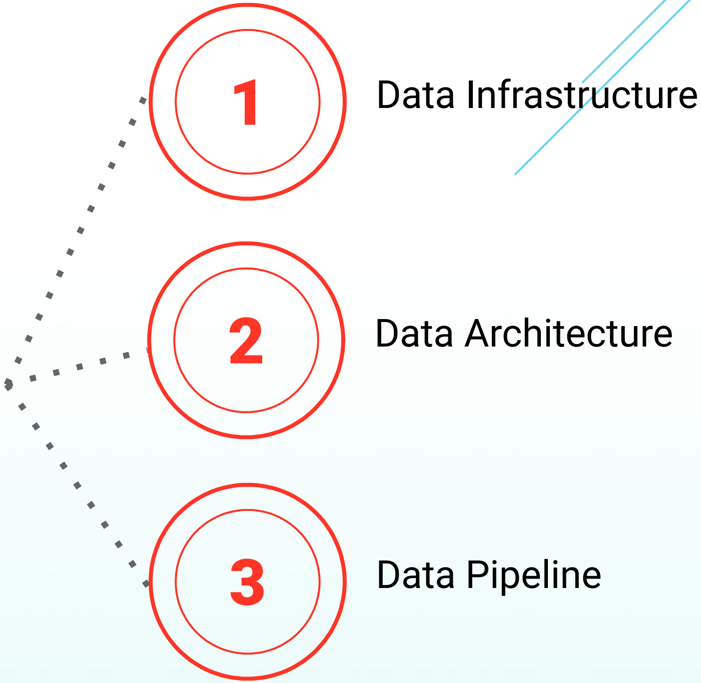
CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

@mrogati

Data Engineer Responsibility



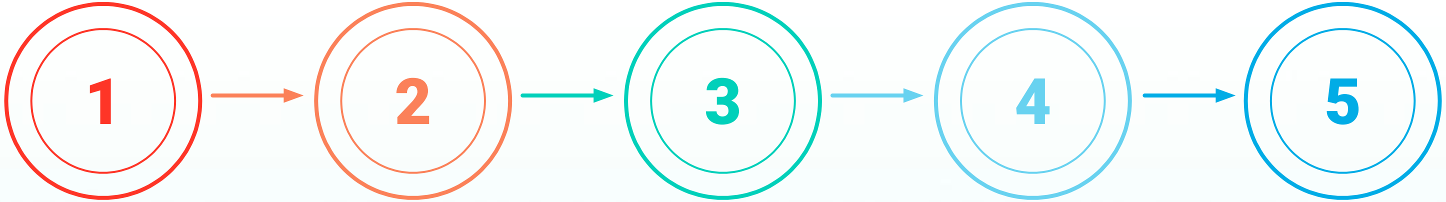
1 Data Infrastructure

2 Data Architecture

3 Data Pipeline

Besides these, a data engineer need to make sure the data flow is running smoothly.

Workflow



Data Source

It can be .csv, .json, database, etc

Extract

Collecting Data from Data Source

Transform

Transform, cleaning, detecting the data

Load

Store the data to Datawarehouse

Datawarehouse

A collection of data that can be used by Data Analyst or Data Scientist

"Automation is Our Savior".

In large organization, the needs of data is keep growing up. With automation the data can keep flowing anytime for the needs of Data Analyst or Scientist.

2

Big Data:

Data engineer & big data, and big data tools for data engineer

Data Engineer & Big Data

Big data is still data, it requires a different engineering approach and not just because of its size. Big data is tons of mixed, unstructured information that keeps piling up at high speed.

That's why traditional data transportation methods can't efficiently manage the big data flow. Big data fosters the development of new tools for transporting, storing, and analyzing vast amounts of unstructured data.

DigitalSkola

Data Engineer Big Data Tools

Apache Airflow

Apache Airflow for programmatically author, schedule, and monitor workflow.

Apache Spark

Apache Spark is a unified analytics engine for large-scale data processing.

Docker

Docker is a software platform that allows you to build, test, and, deploy applications quickly

Hadoop

Hadoop is a framework that allows for the distributed processing of large data sets

Data Engineer Big Data Tools

SQL

SQL is a standard language for storing, manipulating, and retrieving data in database

NoSQL

NoSQL is used for big data and real time web apps

MapReduce

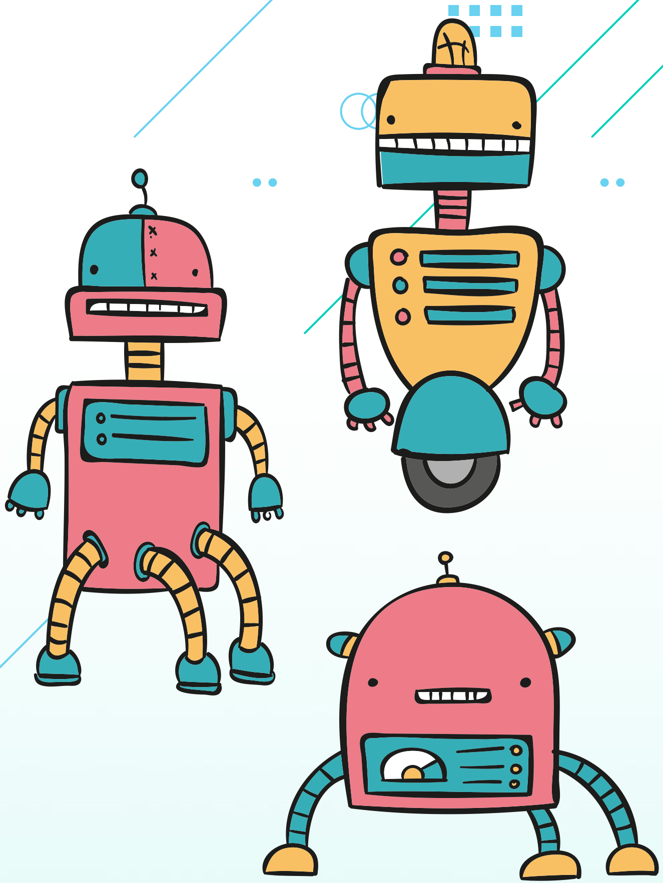
MapReduce is a programming model and an associated implementation for processing and generating big data sets

and many more...

3

Python Introductory:

About Python, Variabel, and Data Type



Why Python?

Python Usage

- Web and Internet development

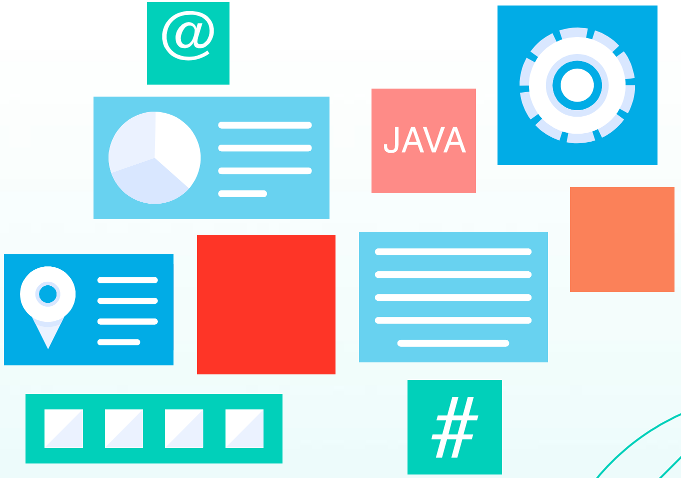
  Popular frame work are Django and Pyramid
- Scientific and Numeric

  The most popular library are numpy, pandas, SciPy
- Desktop GUI

  Some toolkits that are often used are wxWidgets, Kivy, pyqt
- Automated continuous compilation and testing

  The most popular frame work are Buildbot and Apache Gump

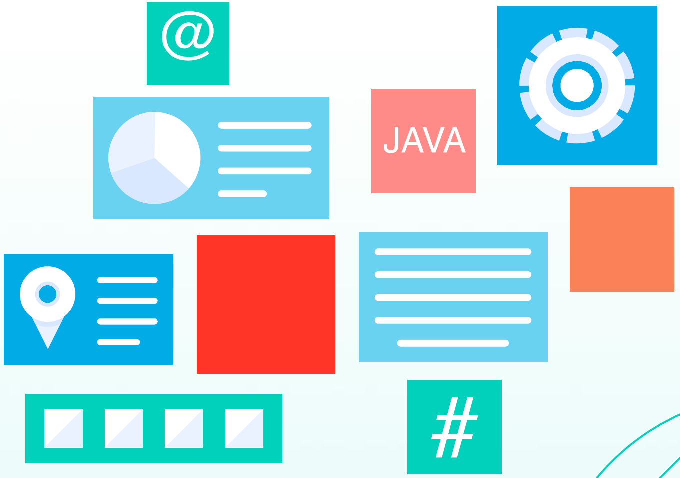
https://www.python.org/about/apps/



Why Python?

Python Feature

- Easy to code, read, also free and open source
- Interpreted
- Object Oriented and Procedure-Oriented
- High Level language
- Dynamically typed
- Support for GUI



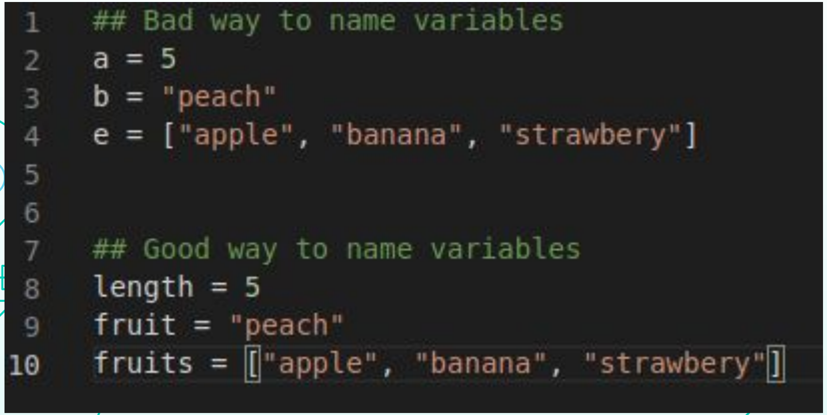
https://www.simplilearn.com/python-features-article

Variable

Variables are like a ***container*** for storing data

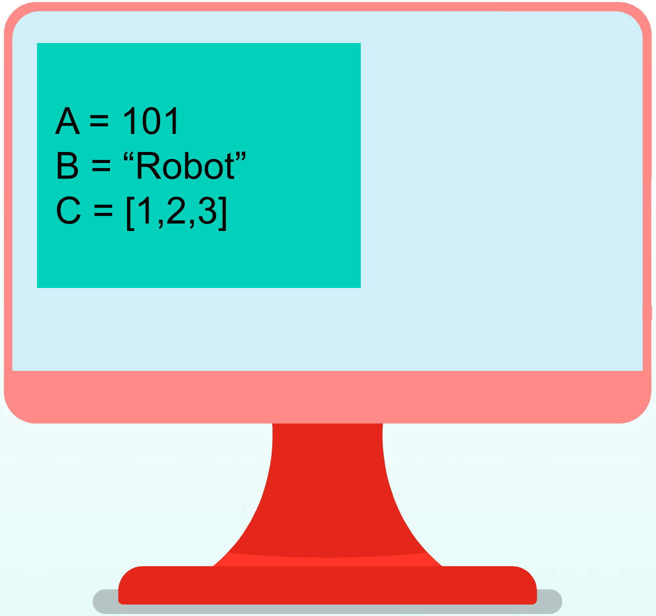
How to make a variable?

Use equals sign “=” to assign data to a variable



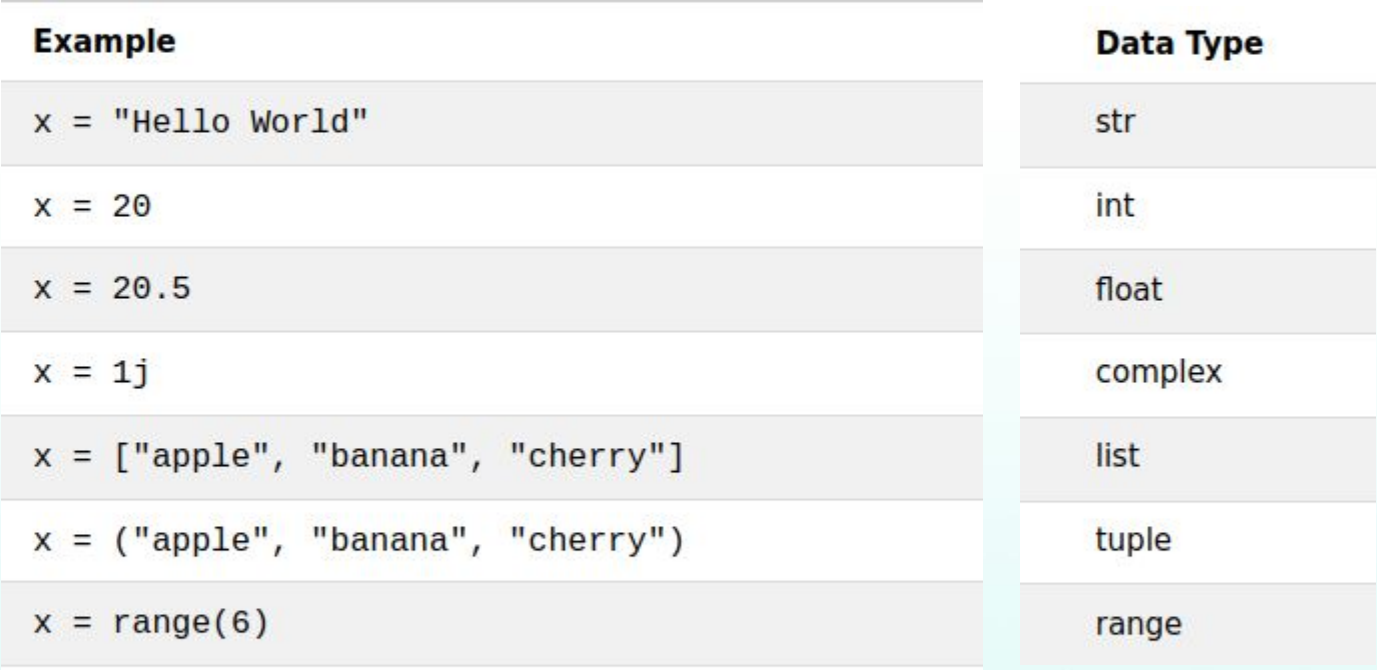
```
## Bad way to name variables
a = 5
b = "peach"
e = ["apple", "banana", "strawbery"]


## Good way to name variables
length = 5
fruit = "peach"
fruits = ["apple", "banana", "strawbery"]
```



A = 101
B = “Robot”
C = [1,2,3]

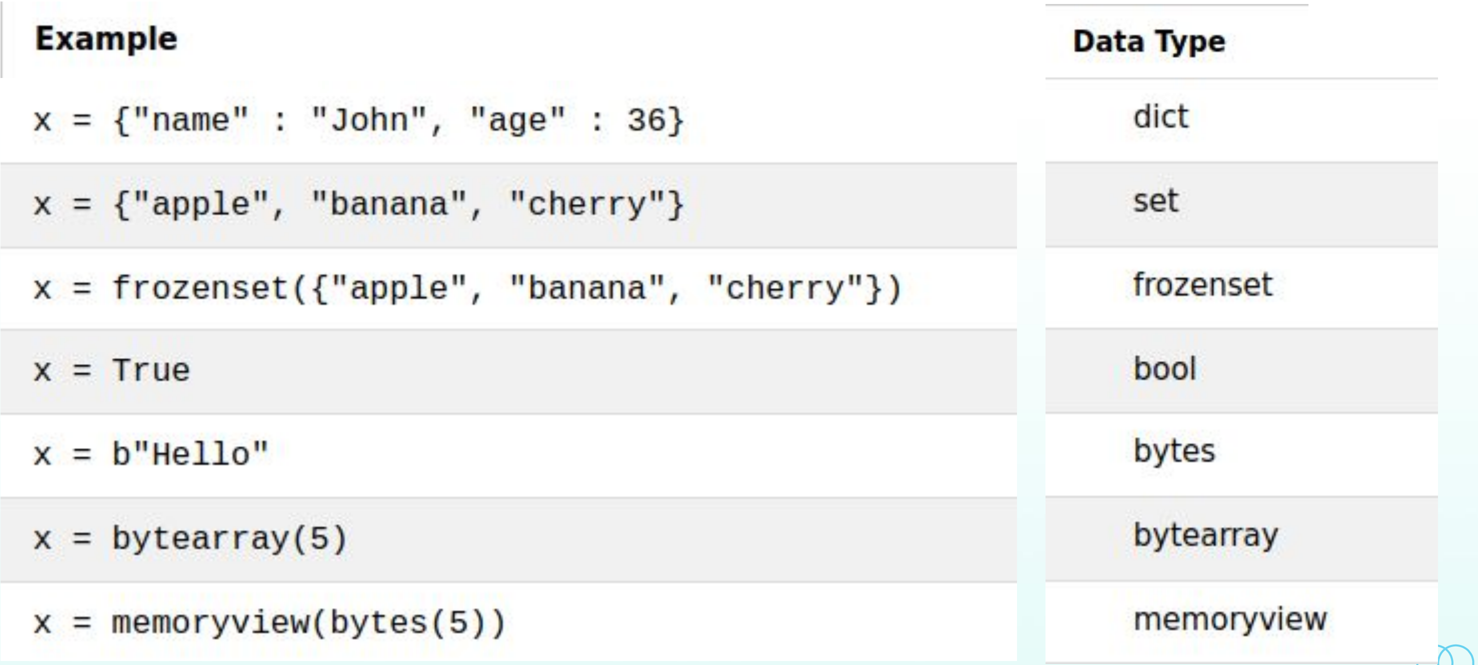
Data Type in Python



Example	Data Type
`x = "Hello World"`	str
`x = 20`	int
`x = 20.5`	float
`x = 1j`	complex
`x = ["apple", "banana", "cherry"]`	list
`x = ("apple", "banana", "cherry")`	tuple
`x = range(6)`	range

https://www.w3schools.com/python/python_datatypes.asp

Data Type in Python



Example	Data Type
x = {"name" : "John", "age" : 36}	dict
x = {"apple", "banana", "cherry"}	set
x = frozenset({"apple", "banana", "cherry"})	frozenset
x = True	bool
x = b"Hello"	bytes
x = bytearray(5)	bytearray
x = memoryview(bytes(5))	memoryview

https://www.w3schools.com/python/python_datatypes.asp

THANK YOU!