New Era University
College of Informatics and Computing Studies
Department of Computer Science

FREE ELECTIVE 3

CCSEL3-18

Networking Academy | Course

Introduction to Data Science

ONLINE COURSE DOCUMENTATION

Marasigan, Vem Aiensi A.
BS Computer Science | 21-11295-310

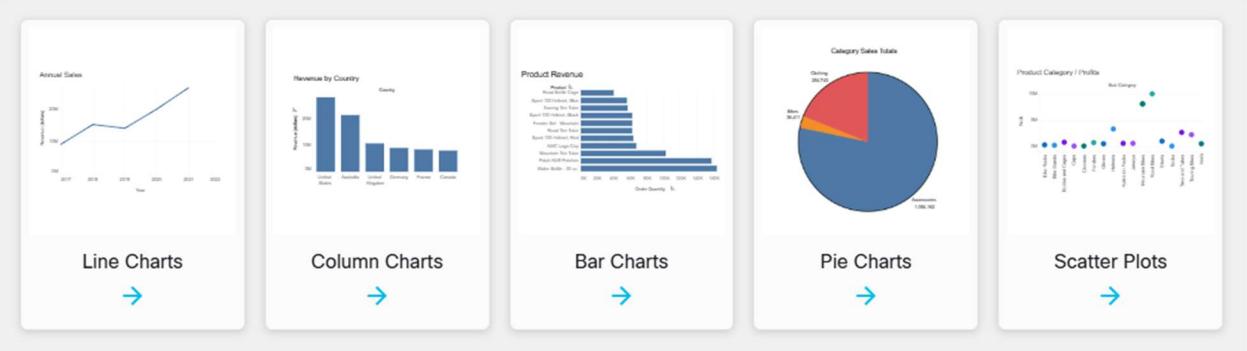
Module 1: Experience Analytics

[October 14 - 15]

This module introduces what data is and what is it for. There a two categories of data which is Qualitative and Quantitative which are used in Data analysis. It also branches to different types:

Data Type	Description
String	Data that is treated as text. It is composed of letters, numbers that are not used in computation, and symbols such as punctuation. String data also includes white space, or the spaces used to separate and format text. Examples of string data are “hello world” and “Building 153”.
Integer	Whole numbers, or numbers that do not include decimals or fractions. One use of integers is to order or rank things. Another is for counts and basic quantities. Examples of integers include 0, 1, 2, 3, and 10,546.
Floating point	Numbers with decimal places. These numbers are frequently employed in statistical analysis. Examples of floating points include 0.0003, 1.2, and -3.67.
Date and time	Stores an instant in time that is expressed as a calendar date and time of day. Date and time data is important in recording when an observation in a data set is made. Date and time formats can vary between data sources. Examples of date and time data formats include YYYY-MM-DD such as 2022-08-15, and YYYY-MM-DD hh:mm:ss such as 2022-01-01 19:24:05.
Boolean	Data that is treated as either True or False. Typically, True and False are capitalized to represent a boolean instead of a string. Boolean values can also be represented as “Yes” or “1” (for True) and “No” or “0” (for False). An example of a boolean expression is “15 is greater than 30” = False. “User John Smith has a membership account” = Yes.

These data are being collected for sake of many purposes such as making informed decisions, identifying improvements, and track and predict events. These can be done by representing data into charts:



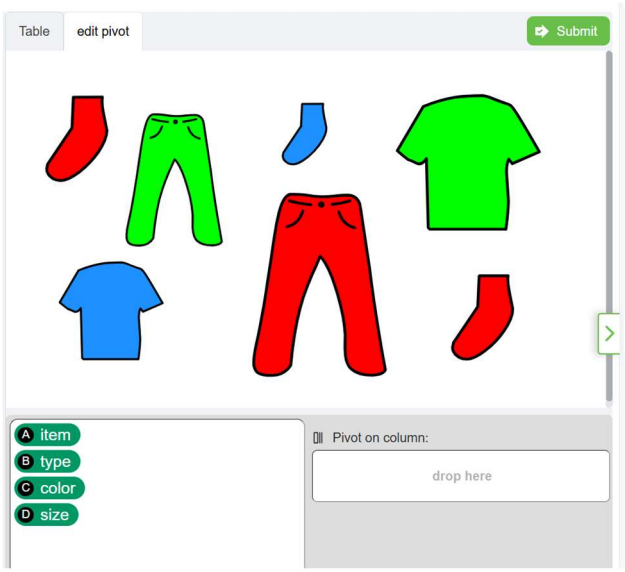
With these visualizations, data analysts can make summaries and depict key performance indicators (KPI) that are significant in decision making especially in business settings.

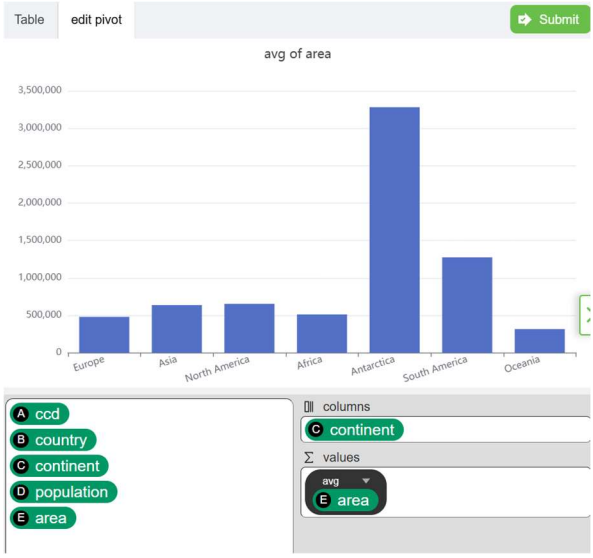
Another tool data analysts can use are pivot tables. This module also offered laboratory activities in teaching how pivot tables works and how summarizations can be of use in obtaining important information regarding a particular data set or scenario. Here are some laboratory examples for module 1

pivot_clothing

Show Pivot:

item	type	color	size
1	pants	red	L
2	shirt	green	L
3	sock	red	L
4	pants	green	S
5	sock	blue	S
6	shirt	blue	S
7	sock	red	L





Overall, this module covers the meaning of data, its types and categories, as well as how it used in data analytics, playing a role in organizations, especially in business settings.

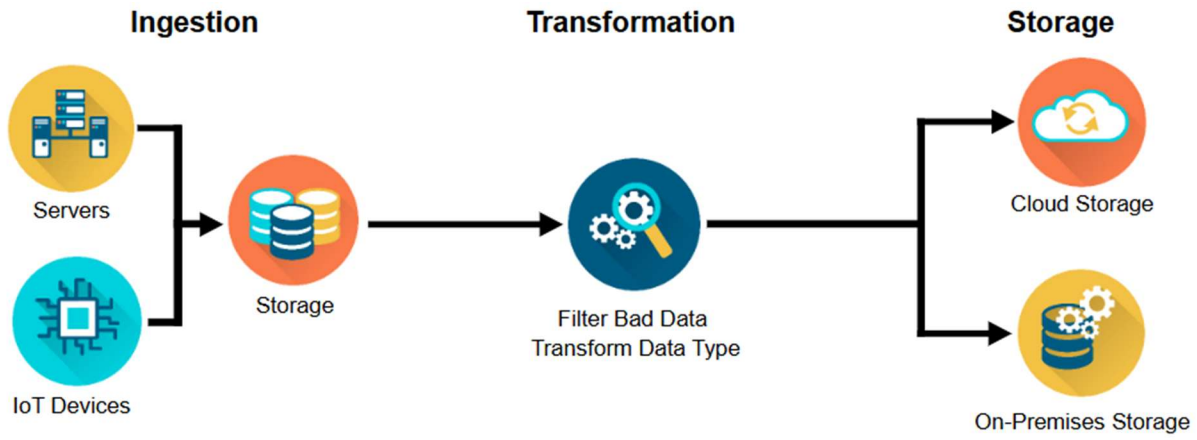
Module 2: Experience Analytics

[Completed by October 17 - 18]

Big data is a term used to describe the massive volumes of digital data generated, collected, and processed. The term big data describes data that is either moving too quickly, is simply too large, or is too complex to be stored, processed, or analyzed with traditional data storage and analytics applications. The characteristics of data, also known as the **four V's** are below:

- **Volume** describes the amount of data transported and stored.
- **Variety** describes the many forms data can take, most of which are rarely in a ready state for processing and analysis.
- **Velocity** describes the rate at which this data is generated.
- **Veracity** is the process of preventing inaccurate data from spoiling data sets.
-

DATA PIPELINE – this simply is the flow of data, from gathering to storing.

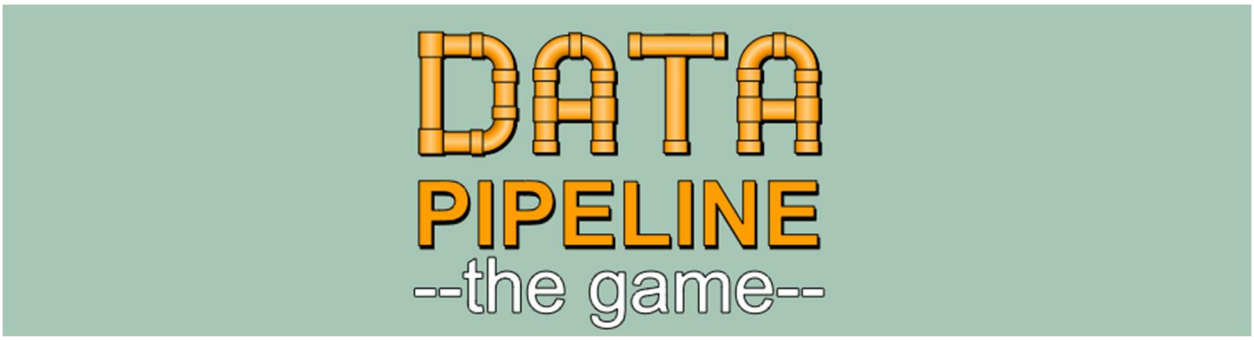


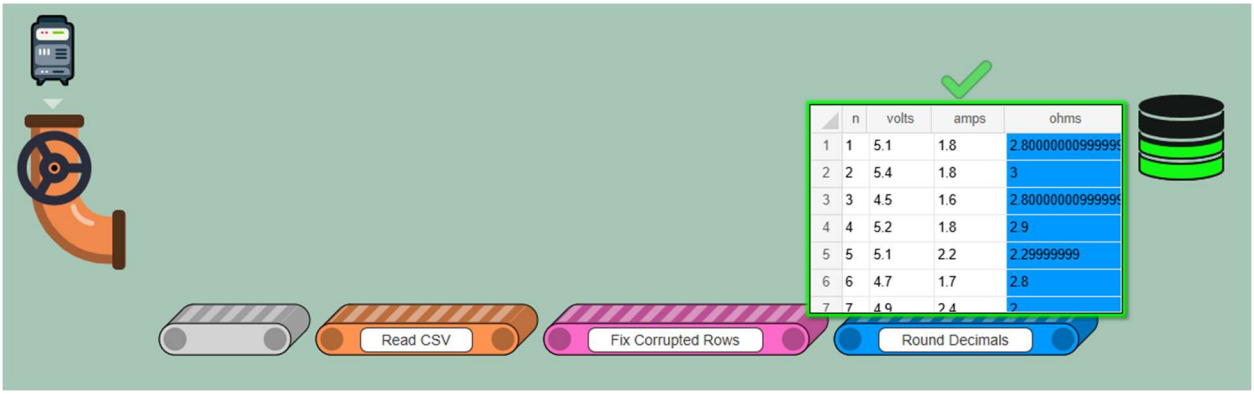
Ingestion – this involves the gathering part of data. It can also interchangeably be called as *Extraction*

Transformation – the process of data cleaning, filtering bad data that can cause anomalies to the data set.

Storage – storing clean data either on-prem or cloud storages. This is also termed as *Load* in ETL

This module of includes an interactive laboratory about data Pipeline, quite neat!





In conclusion, big data is the term used to describe massive volumes of digital data acquired from real world events through data drivers such as IoT devices, etc. Then, data engineers manage these data through a data pipeline. It has three stages: ingestion, transformation, and storage, before analysis is performed.

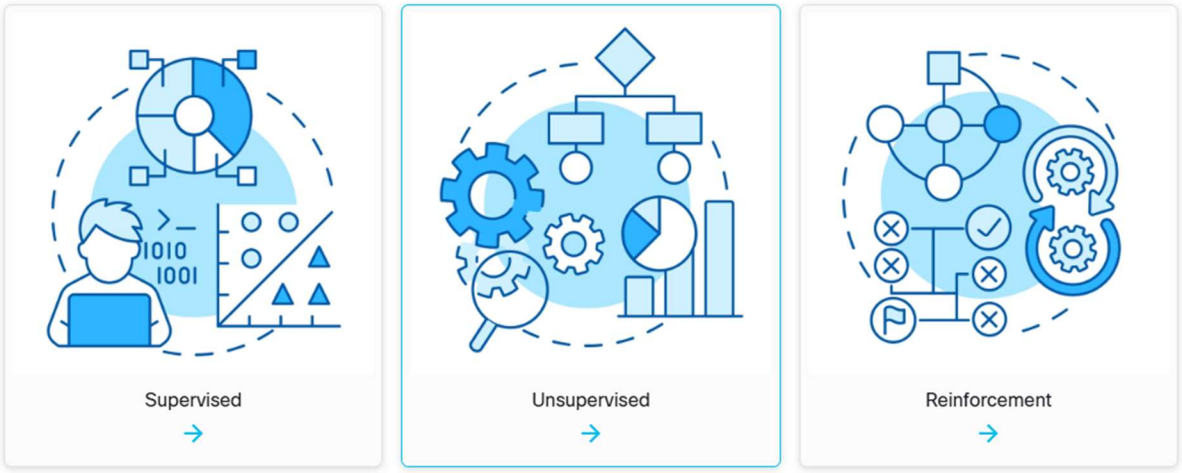
Module 3: AI and Machine Learning

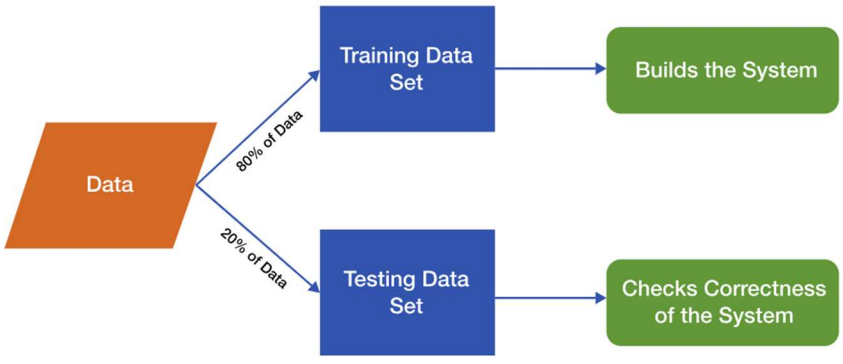
[Completed by October 18]

AI has been around us, from entertainment, agriculture, medicine, retail, and even in fitness. With AI, mundane tasks had been solved, decision making had been more accurate, and many lifestyle actions have improved thanks to AI.

However, access to this kind of technology also enables bad people to exploit Generative AIs' capabilities. In conclusion, AI has changed so much of our daily lives, opened convenience and opportunities, yet if were in the wrong hands can also be devastating.

AI models are created through machine learning. These are the types





in machine learning standard, the data set is split into two, 80% for training data, and 20% for testing data. This is done to avoid overfitting the model into the data set and be adaptive to new unseen data.

Module 4: Embarking on Data Analytics Career

[Completed by October 19]

There are three major roles in this career path which are Data Analyst, Data Engineer, and Data Scientist.

Data analysts query, process, provide reports, and summarize and visualize data. They leverage existing tools and methods to solve a problem. They help people, such as business analysts, to understand specific queries with ad-hoc reports and charts. Data analysts must understand basic statistical principles, cleaning different data types, visualization, and exploratory data analysis. In short, data analysts analyze data to help businesses and other organizations make informed decisions.

Data engineers are responsible for building and operationalizing data pipelines to collect and organize data. They ensure the accessibility and availability of quality data for data scientists and data analysts by integrating data from disparate sources and performing data cleaning and transformation. Skills needed for data engineering roles include understanding the architecture, tools, and methods of data ingestion, transformation and storage; and proficiency with multiple programming languages (including Python and Scala). In summary, data engineers build and operate the data infrastructure needed to prepare data for further analysis by data analysts and scientists.

Data scientists apply statistics, machine learning, and analytic approaches to answer critical business questions. Data scientists interpret and deliver the results of their findings by using visualization techniques, building data science apps, or narrating exciting stories about the solutions to their data (business) problems. They work with data sets of different sizes and run algorithms on large data sets. Data scientists must be current with the latest automation and machine learning technologies. The requirements to perform these roles include statistical and analytical skills, programming knowledge (Python, R, Java), and familiarity with Hadoop, a collection of open-source software utilities that facilitates working with massive amounts of data. Data scientists are data wranglers who organize and deliver value from data.

In order to proceed to a Data Analytics Career path, below are the necessary skills

- Demonstrated attention to detail
- Solid verbal and written communication skills
- Ability to work in a team
- Proficiency with spreadsheets, such as Excel
- Familiarity with SQL and databases
- Some experience with object-oriented programming (OOP) languages such as Python and Java
- Familiarity with visualization tools and presentations

COMPLETION RECORD