Welcome to **Introduction to ETL** course!

Since you have stumbled across this course, you're probably interested in learning how to build an ETL pipeline for your data. Before diving in, let's first learn about the theory behind ETL.

So, what is ETL? **ETL**, which stands for Extract, Transform, and Load, is an approach used to build a data engineering pipeline. It involves extracting data from various data sources, processing and transforming it into a suitable format, and then loading it into the data warehouse.

There are two other approaches: **ELT**, which loads the data into the destination after extraction and processes it later, and **reverse ETL**, where data is extracted from the data warehouse instead of from heterogeneous data sources, then transformed to a format suitable for their destinations, and fed into different applications in the company for different audiences. However, these 2 approaches will not be covered in this course.

These approaches have different distinct use cases:

1. ETL: Best for cases where data needs to be transformed before loading. Typically for legacy systems or data that requires complex transformations before it is stored.
2. ELT: Typically used with cloud data warehouses, where its faster to load raw data and then process afterwards for analytical needs.
3. Reversed ETL: Useful for operationalizing data, such as feeding processed data back into apps like CRM (customer relationship management) systems, marketing platforms, or analytics tools for decision-making.

In this course, we will be using **Mage.ai** to build an ETL pipeline for healthcare data. Afterwards, **Python** will be used to perform Exploratory Data Analysis (**EDA**) and create **data visualisations** for the processed data in the database. This will focus on identifying which infectious diseases are more prevalent in Singapore (if you're working on the batch dataset pipeline) or Malaysia (if you're working on the streaming dataset pipeline).