ETL Project

Presented by Cintia, Berenice and Carlos.

### Project Introduction

For this project, we use 2 sources of data, both were found on Kaggle:

<https://www.kaggle.com/ihelon/lego-minifigures-classification?select=index.csv>

### E: Extract sources

Description: The data source “metadata.csv” is a dataset pulled for lego detail, including: id, name and minifigure name from every product and “index.csv” is a dataset for collection path images for every lego figure.

Files format: csv

Columns used in metadata.csv: class\_id, lego\_id, lego\_name, minifigure\_name

Columns used in index.csv: image\_id, image\_path, class\_id

### T: Transform

Data was edited and cleaned

Drop index column that was a sequential numbers for every row in the tables

Lambda function (apply) was implemented to clean special characters like brackets, double quotes in columns: lego\_ids and lego\_names.

Pandas library was used to format, remove and edit columns.

Each csv file was written into mongo collections.

### L: Load

Using a jupyter notebook to create a database connection in a mongoDB to transfer data in a non-relational database.

Collections were created in mongo DB