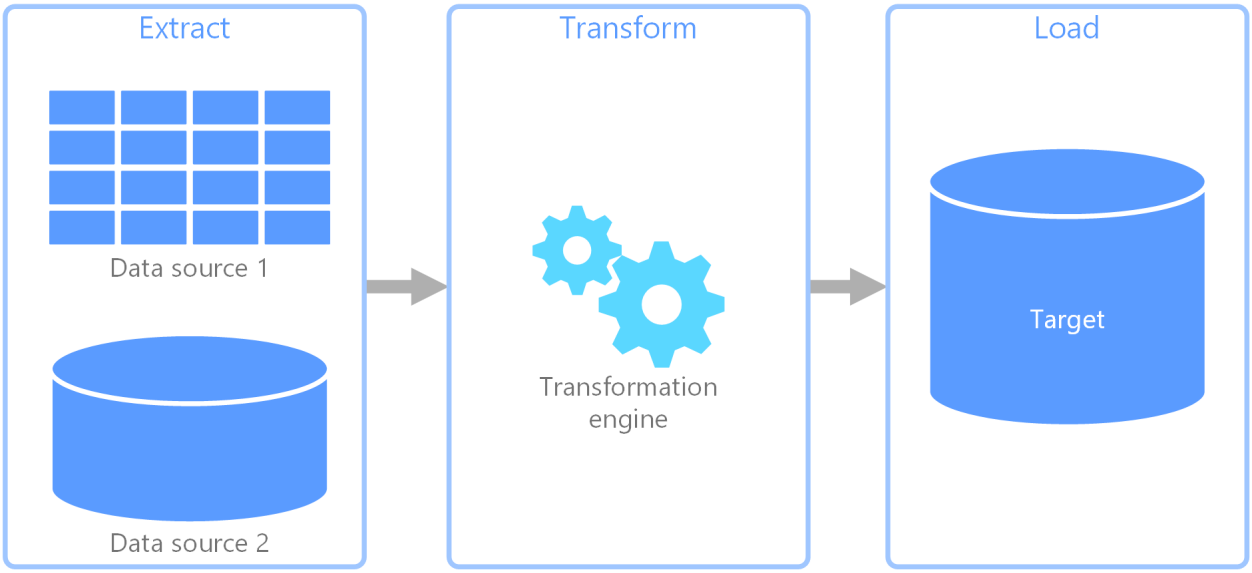
ITCS 3190 Final Project
Fall 2021

I. Introduction: ETL

As a culmination of your semester-long study into Big Data and how industry solutions handle analyzing large data sets, your final project will involve utilizing the ETL process. ETL, which stands for extract, transform, and load, is the process data engineers use to extract data from different sources, transform the data into a usable and trusted resource, and load that data into the systems end-users can access and use downstream to solve business problems. After finishing the ETL process, you will be required to showcase your data transformations in a visualization.



From a high-level perspective, you will be creating a Python application in the IDE of your choice. You have been shown examples of how to use AWS Sagemaker throughout this semester, but you are free to select any IDE. Your Python application will take a data source, perform transformations on the data, and load it into a dashboard or report to showcase your learned ability to operate on data sets.

Your submission will include your project files (or shared link) and a video of 2-5 minutes, demonstrating your resource chosen, the various transformations you performed, and the loaded data (either in a dashboard or extract report).

II. EXTRACT

In this project, you will **choose a data source** of interest to you from Pragmatic AI, Chapters 6 & 10. You may choose a data source from one of the following two links:
https://github.com/noahgift/socialpowernba/tree/master/data
https://www.zillow.com/research/data/

These data sources have a large number of dimensions which allow you to get hands-on experience with querying large datasets, a focus of this class.

To begin, you will need to select an in-memory schema for your data to be held. For example, you used DataFrames to hold data throughout this semester. Several options to choose from for holding your data are listed below. Note the parallel processing capabilities of Dask when you are considering how to extract your data.

- Pandas DataFrames (https://pandas.pydata.org/)
- Dask DataFrames (https://docs.dask.org/en/stable/dataframe.html)

III. TRANSFORM

For the transform phase, we are giving you a lot of flexibility to choose how you want to transform your data. Data transformation is the process of changing the format, structure, or values of data. For larger data analytics projects, data may be transformed at multiple stages of the data pipeline. Processes such as data integration, data migration, data warehousing, and data wrangling all may involve data transformation. In this project, you will select at least **FIVE** transformations to perform on your data.

These transformations should be done with consideration for your end result (which will be a dashboard or Excel/CSV extract)

> *Use Case Example of a Transformation:*
> As a business analyst for Zillow, I want to know how many homes were purchased in the last five years, so that I can make an informed estimate of what home purchased amounts may be in 2022.
>
> Your transformation would then involve operating on the dataframe to determine the number of homes purchased each year in a specified range of years.

Be sure to look at the DataFrames documentation and take advantage of the capabilities offered with dataframes to transform data held in a DataFrame. This significantly speeds up query times and reduces the amount of code you have to write.

IV. LOAD

In the last phase of the ETL process, you will be “loading” your transformed data into a view. Data loading is the process of loading data or data sets from a source to its destination. In production scenarios, companies often utilize data warehouses. For this project, you will simply be delivering your transformed data to a Dashboard or Report extract.

Your options are to show your transformed data in a Dashboard or a Report extract. Both options are possible with Pandas/Dask if you have chosen to utilize DataFrames for this project.

The requirements for both options are that you **visualize** in some way (either through charts or graphs) your data transformations.

Option 1: Dashboard

One open-source example to consider is Plotly's Dash (https://dash.plotly.com/) which allows you to deliver real time responses to the user in a Dashboard interface. Dash is an excellent microcosm for utilizing a Python data pipeline for ingesting table data, transforming it and delivering the transformed data back to the user in a simple user interface. Feel free to use the sample code provided in order to generate your own Dashboard and pass in your transformed data. Note, there are many solutions for a simple Dashboard other than Plotly's Dash, and you are not in any way restricted to just Dash. Another great visualization tool is Amazon Redshift which you have interacted with in AWS Academy. You may want to do some research on other dashboards available or create your own!

Option 2: Report Extract

If you choose to do a report extract to return your transformed data, you will want to ensure you have sufficiently shown charts/graphs that demonstrate your data transformations. You may want to utilize the Jupyter notebooks .ipynb files that you used in previous activities with Amazon SageMaker to show each visualization of transformations you've done. You can also use an Excel extract so long as it contains the charts/graphs of your five transformations.

V. Rubric

Category	Criteria	Points
Extract	Data source chosen is suitable for the project and has 1000+ records	10
Transform	Five transformations performed on data that provide insight into your data source	25
Load	Data is visualized in a Dashboard or report extract	25
Technical	Project Files submitted	20
Technical	Video/Screen capture of 2-5 minutes in length submitted that accurately shows all stages of the ETL process	20