Data Warehouse Implementation

Project 2 Deliverable 2: ETL Processes Due Date: Thursday, May 23 (by end of day)

Overview

Implementing a data warehouse often entails for the following steps:

- Requirements analysis and capacity planning
- Choosing data warehouse infrastructure
- Data Modeling
- Identifying sources of data
- Implementing extract, transform, and load (ETL) processes.
- Populating the data warehouse
- Querying the data warehouse
- Developing user applications
- Rolling out the data warehouse and applications (i.e., deployment in a production environment).

In this project, we are focusing on only a subset of those steps, namely:

1. Data Modeling
2. Identifying sources of data
3. Implementing extract, transform, and load (ETL) processes.
4. Populating the data warehouse
5. Querying the data warehouse

I took care of identifying sources of data by choosing a given data source for you to use. Note: I encourage you to supplement this data source with other sources of related/relevant data (e.g.,).

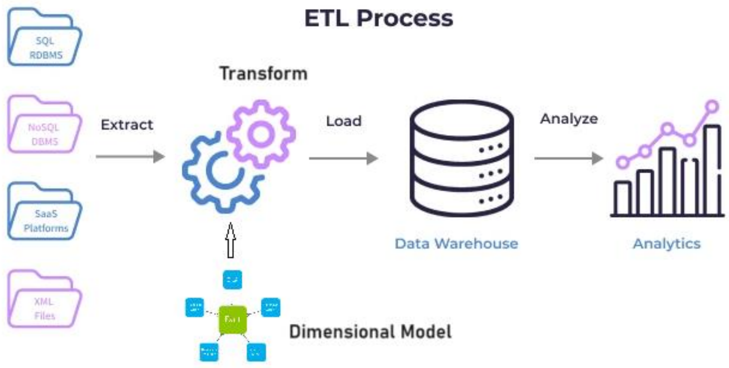
- **For deliverable 1**: data modeling.
- **For deliverable 2**: implementing ETL processes.
- **Deliverable 3**: querying the data warehouse (OLAP and reporting)

Deliverable 2: Implementing ETL Processes

The ETL Process

The ETL Process is the most underestimated step in data warehouse development and yet it is the most time-consuming step (approximately 80% of data warehouse development time is spent on ETL!).

Extract, Transform, and Load (ETL) processes extract data from several heterogeneous data sources, transform the data into a uniform format that is consistent with the dimensional model, and load it into the data warehouse (see figure on the next page).



What you should do

There are various tools to implement ETL processes, but there is no single 'best' tool. Typically, you choose what to use based on your needs. You can use any of the available ETL tools that suits your need or write your own code. While available ETL tools offer better productivity on subsequent projects, it is easy to get started with writing your own ETL code. Also, writing your own ETL code provides greater flexibility, especially with data transformation tasks.

For this project, you are going to write your own ETL code using Python libraries, most importantly the Pandas library.

Write a Python script to accomplish the following tasks (note: Not every data cleaning and transformation task may be application to your data for the project. Focus on those tasks that are applicable):

1. ***Extracting Data***
   Write code to extract the 31 country files from the emissions dataset that you are using for the project (I provided a summary of the sizes of files you need to extract in deliverable 1).
2. ***Transforming data***
   After you extract the data, most of the work you are going to do is cleaning and transforming the data. You need to clean and transform the data into a format consistent with your dimensional model.

2.1. Cleaning the data

Typical tasks you may have to do, wherever applicable include:

- Handling inconsistent data formats, such as spellings, text coding, date formats (most database systems require date to be in the format yyyy-mm-dd) etc.
- Remove unnecessary attributes, comments. Etc.
- Combining data from multiple sources with a common key.
- Normalize spelling of names, addresses, etc. (e.g., in our class example, I noticed that Turkey was misspelt as 'Turkiye').
- Removing duplicates.

- Verifying that the facts are correct and within acceptable values (remove/correct impossible or inconsistent values (e.g., a negative value where a positive value is expected).

2.2. Transforming the data

Common data transformations include:

- Data type conversions (e.g., Date/Time format conversions)
- String manipulations.
- Normalization/denormalization of data (depending on source) to the desired data warehouse format.
- Reshaping of the data (e.g., unpivoting/flattening pivoted data).
- Building dimension keys for the fact tables.
- Handling historical data correctly.
- Merging/combining data from various files/sources.

3. Loading the data to the data warehouse

After you have completed cleaning and transforming the data into a format consistent with your dimensional model, create a data warehouse in SQL Server and load the data into it.

What to submit

- Final dimensional model applied to data warehouse implementation.
- A Python Jupyter Notebook scripts for the complete ETL processes.