Database Administration
Lab 05: Extraction, Transformation, and Load.

Andrés Calderón, Ph.D.

March 3, 2025

1 Introduction

Data integration is a fundamental process in data engineering, enabling organizations to consolidate information from multiple sources into a unified format for analysis and decision-making. In this lab, we will explore the process of extracting, transforming, and loading (ETL) data from heterogeneous sources using **Pentaho Data Integration (PDI)**.

The lab is structured around a practical scenario in which we collect and process data from *two different sources*:

1. **GPX files** containing location records of moving entities, including timestamps and object identifiers.
2. **TSV files** containing sensor data with similar attributes.

The objective of this lab is to demonstrate how to *clean, transform, merge, and load* these datasets into a structured database. We will work with a sample database called '`sensors`' and create tables ('`sensor`' and '`sensor2`') to store the integrated data. The lab will guide you through a step-by-step approach, including:

- Extracting relevant data from the files.
- Cleaning and transforming the data to match the database schema.
- Merging both data sources into a unified dataset.
- Loading the final dataset into the database.
- Correcting data types, such as converting timestamps stored as strings into proper datetime formats.

By the end of this lab, you will have hands-on experience with PDI and an understanding of the ETL workflow. Additionally, you will explore alternative ETL tools and create a tutorial report based on a different data integration solution.

2 Pentaho Data Integration Installation

We will explore two methods for installing PDI: one using the *Nube Privada Javeriana*, which you can replicate on your own machine, and the other using a Docker image from hiromuhota/webspoon. You can follow this video to see the step-by-step procedure.

3 Data Extraction

We will use two data sources and assume that we want to collect data on moving entities. One source is GPX tracks from a GPS device (download the file from here). The other data source is sensor data provided in a TSV file (download the file from here).

The GPX file contains 86 locations, including the ID of the moving object and the timestamp when each sample was taken. The TSV file contains 284 records with the same fields. In total, we will load 300 tuples into our database. We will use a dummy sample database named sensors and create two tables (sensor and sensor2) to simulate the loading of records.

We need to clean and extract some data from the files, so please follow these two videos to see the process for the GPX file and the TSV file, respectively.

4 Data Transformation

Before loading the data into the database, we need to merge both data sources. To do this, follow this video, where we explain how to append two different data streams.

5 Data Load

Finally, once we have extracted the data from the files and transformed it according to the schema of our tables, we will proceed with loading the integrated data. Watch this video for details on the process.

6 Additional Details

You may have noticed that we store the event timestamps as a String data type, which is not the correct approach. In the next video, we show you how to transform and integrate this particular field attribute. You will need to complete this step on your own.

7 Individual Work

Guess what? In addition to Pentaho Data Integration, there are plenty of alternatives, both open-source and commercial. Read the following document and choose one of them. Then, write a well-structured tutorial report similar to the one you just completed, but using your chosen alternative.

We expect you to submit your report by **March 17, 2025**.

Happy Hacking ☺!