The Role of ETL in Data Mining and Statistical Analysis

John Wensink

MIS450 - Data Mining

Colorado State University-Global Campus

Dr. Steve Chung

May 31, 2020

The Role of SQL in Data Mining and Statistical Analysis

Pentaho is a type of business intelligence software distributed by Hitachi Data Systems in both an enterprise and community format. Much like the SAS software we have been working with in the weeks previous, Pentaho is a grouping of many software products each with their own specialization. Some examples of the offerings include the Business Analysis (BA) Platform, Pentaho Analysis Services aka Montrain, Pentaho Dashboard Designer, Pentaho Analyzer, Pentaho Interactive Reporting, Pentaho Data Access Wizard, as well as the newly offered Pentaho Mobile. It seems that for this class we will be dealing mostly with the BA suite of tools which makes up the core of the Pentaho software.

**The Role of ETL Tools in the Data Mining Process**

ETL stands for Extract, Transform, Load. These three steps, if done correctly, are able to provide clean data that has been transformed with explicit purpose to our statistical analysis software such as SAS. ETL became popular in the ‘70s when businesses started to store their data in different purpose-built repositories for different types of data. Integrating data that had been housed in different locations and in different formats proved to be a formidable challenge at the time. The role of ETL was then as it is now to extract data from multiple sources, transform that data into a format compatible with the target’s infrastructure, and load the transformed data onto the target. In the modern era, tremendous volumes of data are stored in data warehouses small and large in all corners of the world. Different regions of the globe have since developed alternative data extraction tools, and ETL is no longer the only game in town. Yet ETL remains relevant as businesses continue to rely on the process to gather business intelligence and perform predictive analytics on incredible data sets that are as diverse as they are large. (SAS, n.d.). The reason for ETL’s continued popularity is that it offers an efficient, proven solution for companies to consolidate and make use of their warehoused data at rest.

In this class, we will be using ETL alongside a Relational Database Management System (RDMBS), namely PostgreSQL to extract data from the Northwind database, transform the SQL data into a data type compatible with our statistical analysis suite SAS such as XML, CSV, or flat/text file types. By using Pentaho’s ETL software, we will ensure we are importing high-quality data from our data warehouse that is complete, accurate, valid, and we are even able to prove this to a legal standard using Pentaho’s data quality tools such as data mapping and data lineage (Pentaho, 2017). Although for the purposes of this class we will be using SAS to perform our data analytics, perhaps unsurprisingly, ETL is also compatible with other big-data services such as Spark, Hadoop, BigQuery, and Tensor Flow. In my situation, it makes no difference whether I use ETL or ELT, as I am running both systems on the same computer, although SAS will technically be running in a virtualized environment, the processor resources are the same. Say for example I was using a cloud provider’s platform as a service to do my analytics processing. It might make more sense to use the Extract Load Transform (ELT) process to take advantage of the cloud provider’s superior processing capabilities.

In summary, ETL will almost certainly be a powerful and efficient way to extract data from the Northwind data warehouse, transform it into a CSV file, and load that data into SAS where analytics can be performed such as regressions and clustering. Although Pentaho took almost an hour to download and extract, it appears to be operating nominally and I am excited to get started working with this tool.

References

Pentaho. (2017, May 21). Pentaho Data Integration. Retrieved May 31, 2020, from <https://help.pentaho.com/Documentation/7.1/0D0/Pentaho_Data_Integration>

SAS. (n.d.). What Is ETL? Retrieved May 31, 2020, from <https://www.sas.com/en_us/insights/data-management/what-is-etl.html#etl-importance>