Database and Data Warehouse Creation

John Wensink

MIS450 - Data Mining

Colorado State University-Global Campus

Dr. Steven Chung

June 6, 2020

Database and Data Warehouse Creation

For this week’s assignment, I have committed to going the route of option #1 for our portfolio project. I was able to create the jigsaw operational database, as well as the jigsaw data warehouse in PostgreSQL using the scripts provided on this week’s modules page. After that, I was able to create a new connection between Hitachi’s Pentaho ETL application and PostgreSQL over the localhost network. From this point, I was able to perform 3 rudimentary ETL transformations on the manufacturers, regions, and sales tables with the target being the tables inside the jigsaw operational database, and the destination being the jigsaw data warehouse.

**PostgreSQL**

Working inside Google’s Chrom browser, the address ‘<http://127.0.0.1:49761/>‘ allows the database administrator to access PostgreSQL on the localhost network. Using the administrator’s privileges I was able to create two new databases ‘js’, as well as ‘js\_dw’, and populate them by copying and pasting the SQL scripts provided in the course module page into PostgreSQL’s built-in query tool. If this were a professional database, I would probably not recommend using the administrator’s privileges to do work inside PostgreSQL that does not require such escalated authority, users would be wise to adhere to the principle of least privilege when using or modifying production databases. This step was effortless, as I have worked with PostgreSQL before in the past, and it has come to be my favorite Relational Database Management System (RDBMS) out of the few I have tried including Microsoft’s SQL Server and MySQL. Although I have not used the latest version of PostgreSQL until now, I find working directly from the browser window to be convenient and easy to set up alongside SAS in my virtualized VMWare environment.

**Pentaho ETL**

After learning about Extract, Transform, and Load (ETL) procedures in some previous courses, as well as in last week’s interactive lecture, I was looking forward to getting some hands-on experience with Hitachi’s Pentaho ETL application. I was astonished at the size of the file, as well as the length of time required to unzip the application once it had completed downloading. My good habit of starting my coursework early in the week definitely paid off as I was not stuck waiting for these processes to complete at the last minute before my assignment was due. After the software was up and running I had virtually zero troubles making the connections required to perform ETL transformations on my jigsaw databases in the PostgreSQL RDBMS. As I am running my virtualization software on my own laptops bare metal, it didn’t make much difference to me whether to perform an ETL transformation or an Extract Load Transform (ELT) transformation, as the processing resources would be nearly identical (Ranjan, 2009). If I were making use of a cloud-based infrastructure as a service that had substantially superior processing capabilities, I might consider using the cloud’s resources to perform computationally intensive operations rather than my own PC which is decent but could never compare to the warehouses of processing power located at these cloud data centers (Homayouni, 2018). It did take me a little while to figure out that I had to select my location database and my target database from the dropdown menu, but I eventually figured it out. Although I did not experience great difficulty with this week’s assignment, I could see where students could be confused by the assignment’s instructions. Option #1 was somewhat less clearly worded than option #2 was, but the parts I would revise for the next class would include that the ‘lk\_manufavturers.ktr’ transformations along with the other transformations required for the assignment were required to be downloaded from the course content (I searched and searched), and also that the dropdown menus for the table’s location and target databases need to be adjusted for the transformation to process without errors. Otherwise, this was a great assignment for familiarization with the Pentaho ETL software, I feel confident I could put it to work for future needs, including this course’s portfolio project.

References

Homayouni, H. (2018). Testing the Extract-Transform-Load Process in Data Warehouse Systems. 2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW). Retrieved from <https://www.researchgate.net/publication/329064248_Testing_Extract-Transform-Load_Process_in_Data_Warehouse_Systems>

Ranjan, V. (2009). A Comparative Study between ETL ( Extract-Transform-Load ) and ELT ( Extract-Load-Transform ) approach for loading data into a Data Warehouse. *California State University*, 2009. Retrieved from <https://pdfs.semanticscholar.org/b6aa/6cd1aec2c36c8d7e573b109a8d1d2e87b593.pdf?_ga=2.256266955.173456915.1591587094-917221845.1583706419>