Satellite ETL

Brian Parish and Kate Engard

University of Arizona/Trillogy

Introduction to ETL

Advances in modern web and app-based technology have significantly reduced the barriers to data collection. Today, data is sourced from seemingly every part of our daily lives. For data to be valuable, it needs to be strategically warehoused in a way that allows decision makers to look at data in meaningful ways. As a result, data warehousing has become a common means of collecting data from multiple sources over long periods of time and transforming data into not only compatible formats, but also organizing data in a way that is relevant and efficient in providing decision makers with meaningful insights.

Data warehousing typically involves three primary steps: 1) Extracting data from multiple sources to create a comprehensive dataset; 2) Transforming this data into formats that are both logical and efficient in providing relevant insights when queried; And 3) loading this data into the intended database or warehouse where it can be queried for insights into relational data. The concept of extracting, transforming and loading data (ETL) seems relatively straight forward, however within the ETL framework, there are many different ways to execute each of the steps in this process. The most effective ETLs are those that provide decision-makers with rapid insights that result in improved data-informed decisions.

For this project, our ETL goal was to extract satellite data and match it with weather data to provide insights into weather conditions using geocoordinates from the satellite launch site.

Extract

Satellite launch information and weather data from Open Weather API was extracted from Heavens-Above.com and OpenWeatherMap.org respectively to determine the current weather at the launch site for each satellite listed on Heavens-Above.com .

Satellite data was extracted using BeautifulSoup to perform web scraping of a series of satellites from Heavens-Above.com. Web scraping collected satellite ID, Title, launch date and launch site. This web scrape proved to be difficult for extracting some of the data, as the HTML markup was not thorough. Using a table index to reference the nested table containing the essential data it was possible to use BeautifulSoup to find the launch site location data. After web scraping for satellite data, we then manually identified the latitude and longitude for each of the launch sites using google maps and created a separate data frame to hold launch site coordinates.

Using launch date and launch site geocoordinates, we attempted to pull historic weather data from Open Weather Map’s API. The original goal was to produce the historic records of weather conditions at the time of the satellite launch, however due to subscription requirements for historical data, we were not able to obtain the historic weather API information for the day of the launch. Instead, we ran an API for the current weather at each of the launch sites listed for each satellite. Weather data was pulled from the Open Weather Map API for each of the launch site locations.

Transform

After web scraping for satellite data and calling APIs for current weather conditions at each of the launch sites, data from both sources was merged into a single data frame containing relevant information for API and search queries: Satellite Name, Satellite ID, Launch Date, Launch Site, Latitude, Longitude.

Duplicates and NaN line items were removed from the data set and the data set was exported to a .csv file. There was little data manipulation needed because we tried to execute the web scrape as efficiently as possible. After cleaning/transforming data, we used pandas to create a SQL engine connection to export data into a SQL database.

Load

To create a searchable database, we created a table in SQL using PostgreSQL and specified table columns, data types, and primary keys. SQL databases are typically used for managing relational or structured data. Satellite launches and tracking systems involve many relational data points over the time-course of each satellite's usage – from launch data and weather, to speed and trajectory. Establishing a data warehouse in SQL provides the best means of searching for relational satellite data. SQL differs from other databases in that tables are lists of rows rather than sets of tuples. In this way, the same row of data can be assessed across multiple queries. This can be particularly advantageous for satellite data where you may have satellite-specific data (technical stats) that, at any given time, may be associated with variables such as trajectory, speed, payload, weather etc. Furthermore, the ability to store common search schemas is a highly efficient means of providing rapid insights to meaningful data.

Conclusions

ETL is a highly effective means to provide meaningful insights to decision makers. By pulling relevant data from multiple sources, robust data sets can be generated and warehoused it in a way that makes large data sets able to provide highly specific insights to relational data that might otherwise be missed.

References

Mukherjee, R., and Kar, P. (2017). A comparative review of data warehousing ETL tools with new trends and industry insight. *IEEE 7th International Advance Computing Conference (IACC)*, Hyderabad. p. 943-948, doi: 10.1109/IACC.2017.0192.

Kakish, K., and Kraft, T. (2012). ETL evolution for real-time data warehousing. *2012 Proceedings of the Conference on Information Systems Applied Research (*v5 n2214), New Orleans Louisiana, USA.