ETL vs. Data Wrangling

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

Every data science project or a project that deals with data starts with the gathering of pertinent data for the project. Also, ever since the analysis of data has been deemed essential in the operation of businesses and organizations alike, methodologies have been conceived to contribute to the efficacy of the analysis. ETL/ELT has been a traditional approach over the years. However, with the advent of data science and the complexity of data, the traditional approaches do not suffice and have been joined by a new approach called data preparation or data wrangling. This more modernized approach is to meet the demand of the present and the future applications of data science.

The purpose of this paper is to describe each approach and draw a comparison between the two.

Keywords:

ETL, ELT, Data Wrangling

ETL vs. Data Wrangling

Introduction

Extract, Transform, and Load (or Extract Load, and Transform) approach to make data available have been in use since the 1970's. As the name implies, data from various sources with various formats are extracted, then transformed into a format that has been formalized, and then finally loading it into a designated environment where the end users can access them. It has traditionally been the job of a team of IT professionals to perform the ETL operations. These professionals work with business intelligence experts in the field to provide them with the data they require. They use ETL tools to gather and deliver the data in the required format. The ETL process produces high quality data because of its rigid structure and the end user can readily use business intelligence tools such as Excel or professional reporting tool to analyze the data. However, because the ETL process is predefined in its data provision, it falls short when new use cases arise that require new data or new transformations in a timely fashion.

In today's world, customers and therefore data scientists require quick answers to "what if's", "how long before?", "is x related to x?" questions. To answer these questions, the IT barrier in the ETL model would inhibit the analyst to extract the information directly from the data, but rather must work with IT to implement it.

Data Wrangling allows the data scientist to directly work with the data that is produced by ETL process or obtained elsewhere and combined with the ETL data. Oftentimes, the ETL data is made as raw as possible. And it is left to the data wrangling process to make it useful.

The idea of data wrangling is rather simple. It is basically preparing the data in such a way to make the modeling easier and more efficient and accurate. It is an accepted fact in the

industry that Data Wrangling takes up anywhere from 80% to even 100% of the project. It is very critical to the success of the projects on which money and sometimes lives depend on.

Data Wrangling is typically done by people who have domain knowledge in the data they analyze, so they can ask the right questions. For this, there may be need to recategorize some elements in the data, or create new element using other elements. For example, turn country to coded value where instead of having one column for country, have a column for each country and the value to be one or zero accordingly. Another example could be to take log value of another column to see if it would fit the model better or its correlation with another column would change. These are the operations we can perform in data wrangling process. We can do them quickly, change use cases and perform other operations until we've gotten the results we were looking for. It is also not inconceivable that new questions might arise that will require more wrangling.

Conclusion

The difference between ETL and data wrangling is essentially the end user, the use cases, and the data themselves. In ETL, the end user is the IT professional who has developed the tools necessary to extract, transform, and load the data per the use cases provided by the business intelligence person. The data is typically well structured and is stored enterprise-wide in the organization. With data wrangling the end user is the business intelligence person, the data could be the ETL data or otherwise, and the use cases are more dynamic than the ones in the ETL model.

References

1. ETL vs. Data Preparation: What Are the Differences? - https://blog.syncsort.com/2019/02/big-data/etl-data-preparation-differences/
2. Data Wrangling Versus ETL: What's the Difference? - https://tdwi.org/articles/2017/02/10/data-wrangling-and-etl-differences.aspx
3. THE VALUE IS IN THE DATA (WRANGLING) - https://www.darkhorseanalytics.com/blog/the-value-is-in-the-data-wrangling
4. What exactly is Data wrangling? - https://www.quora.com/What-exactly-is-Data-wrangling
5. What is ETL? - https://www.youtube.com/watch?v=a5C-Bw8y9gM