Data Pipelines: ETL vs ELT

Data pipeline is a generic term for moving data from one place to another. For example, it could be moving data from one server to another server.

ETL

An [ETL pipeline](https://en.wikipedia.org/wiki/Extract,_transform,_load) is a specific kind of data pipeline and very common. ETL stands for Extract, Transform, Load. Imagine that you have a database containing web [log data](https://en.wikipedia.org/wiki/Log_file). Each entry contains the IP address of a user, a timestamp, and the link that the user clicked.

What if your company wanted to run an analysis of links clicked by city and by day? You would need another data set that maps IP address to a city, and you would also need to extract the day from the timestamp. With an ETL pipeline, you could run code once per day that would extract the previous day's log data, map IP address to city, aggregate link clicks by city, and then load these results into a new database. That way, a data analyst or scientist would have access to a table of log data by city and day. That is more convenient than always having to run the same complex data transformations on the raw web log data.

Before cloud computing, businesses stored their data on large, expensive, private servers. Running queries on large data sets, like raw web log data, could be expensive both economically and in terms of time. But data analysts might need to query a database multiple times even in the same day; hence, pre-aggregating the data with an ETL pipeline makes sense.

ELT

ELT (extract, load, transform) pipelines have gained traction since the advent of cloud computing. Cloud computing has lowered the cost of storing data and running queries on large, raw data sets. Many of these cloud services, like [Amazon Redshift](https://aws.amazon.com/redshift/), [Google BigQuery](https://cloud.google.com/bigquery/), or [IBM Db2](https://www.ibm.com/cloud/db2-warehouse-on-cloud) can be queried using SQL or a SQL-like language. With these tools, the data gets extracted, then loaded directly, and finally transformed at the end of the pipeline.

However, ETL pipelines are still used even with these cloud tools. Oftentimes, it still makes sense to run ETL pipelines and store data in a more readable or intuitive format. This can help data analysts and scientists work more efficiently as well as help an organization become more data driven.