ETL vs. ELT - Increasing Your Data's Value in the How

For the last couple of decades ETL (extract, transform, load) has been the traditional approach for data warehousing and analytics. The ELT (extract, load, transform) approach changes the old paradigm. But, what's actually happening when the "T" and "L" are switched?

ETL and ELT solve the same need:

> *Billions of data and events need to be collected, processed and analyzed by businesses. The data needs to be clean, manageable and ready to analyze. It needs to be enriched, molded and transformed. To make it meaningful.*

But, the "*how*" is what's different and leads to new possibilities in many modern data projects. There are differences in how raw data is managed, when processing is done and how analysis is performed.

In this article, we'll demonstrate the ETL and ELT technological differences showing data engineering and analysis examples of the two approaches and summarizing 10 pros and cons.

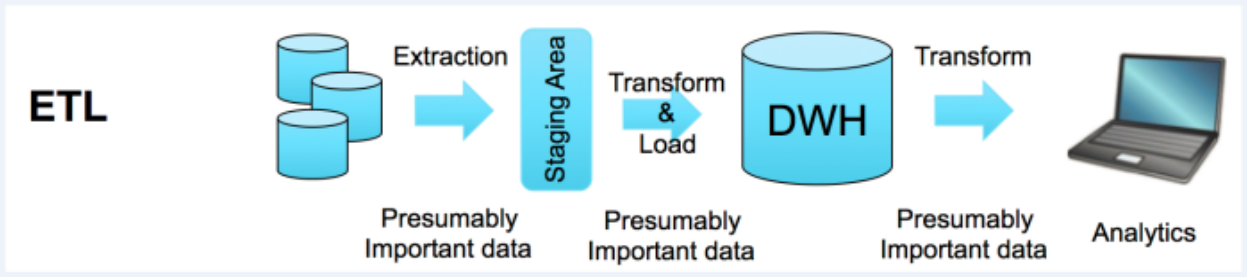
The Technological Differences

Lets first align on the 3 stages - E, T, L:

- **Extraction**: Retrieving raw data from an unstructured data pool and migrating it into a temporary, staging data repository
- **Transformation**: Structuring, enriching and converting the raw data to match the target source
- **Loading**: Loading the structured data into a data warehouse to be analyzed and used by business intelligence (BI) tools

ETL vs. ELT

ETL requires management of the raw data, including the extraction of the required information and running the right transformations to ultimately serve the business needs. Each stage - extraction, transformation and loading - requires interaction by data engineers and developers, and dealing with capacity limitations of traditional data warehouses. Using ETL, analysts and other BI users have become accustomed to waiting, since simple access to the information is not available until the whole ETL process has been completed.



ELT/ETL Image Credits: IBM

In the ELT approach, after you've extracted your data, you immediately start the loading phase - moving all the data sources into a single, centralized data repository. With today's infrastructure technologies using the cloud, systems can now support large storage and scalable compute. Therefore, a large, expanding data pool and fast processing is virtually endless for maintaining all the extracted raw data.

ELT/ETL Image Credits: IBM

In this way, the ELT approach provides a modern alternative to ETL. However, it's still evolving. Therefore, the frameworks and tools to support the ELT process are not always fully developed to facilitate load and processing of large amount of data. The upside is very promising - enabling

unlimited access to all of your data at any time and saving developers efforts and time for BI users and analysts.

A Hands-On Example

Here's an example to illustrate the technological differences between ETL and ELT, and drill down into the details.

Our demonstration will use two data tables: one for purchases and another for currencies, as below:

PURCHASES TABLE		
ip	amount	currency
2.248.0.0	100	EURO
2.248.0.0	200	GBP
72.229.28.185	300	USD

CURRENCIES TABLE	
currency	rate
EURO	1.12
GBP	1.3
USD	1

To understand the fundamentals, we'll look at how this sample is processed in ETL and ELT. For each, we'll show how to calculate a single summary table using these two tables - including the average purchase per country (based on the IP address provided).

ETL Data Transformation on Extracted Data

In the ETL process, the transform stage applies to a series of rules or functions on the extracted data to create the table that will be loaded.

Here's some code to demonstrate the preliminary data transformation process for ETL:

```
def transform(data):
    countries = {}
    for d in data:
    country = ip2country(d["ip"])
    amount = d["amount"] * currencies[p["currency"]] # where do we keep these?
    acc = countries.setdefault(country, {"sum": 0, "count": 0})
    acc["sum"] += amount
    acc["count"] += 1

    # compute the averages
    output = []
    for k, acc in countries:
    output.push({
    "country": k,
    "amount": acc["sum"] / acc["count"]
    })
    return output
```

Using this script, we are mapping the IP addresses to their related country. We are deriving a new calculated value ‘amount’ by multiplying the values of both source tables group by currency attribute. Then we are sorting data by the country column, joining the data from the purchases and currencies tables, and summing up the average values per country.

This data transformation results in a new table with the average amount per country:

AVG AMOUNT PER COUNTRY	
country	amount
USA	300
SWEDEN	372

ELT Data Transformation at Query Runtime

In contrast to ETL, with ELT all data is already loaded and can be used at any point in time.

Therefore, the transformation is done at query runtime:

```
SELECT IP2COUNTRY(ip) as country, AVG(amount * currencies.rate as amount)
NATURAL JOIN currency
```

GROUP BY 1;

In the query, we are selecting the IP address by country, multiplying amount from the purchases table and rate from the currencies table to calculate the average amount. Then, joining both tables based on the common columns of both tables and grouping by country.

This will result the same exact output table as in the ETL process above. However, in this case, since all raw data has been loaded, we can more easily continue running other queries in the same environment to test and identify the best possible data transformations that match the business requirements.

The bottom line of this hands-on example -
ELT is more efficient than ETL for development code. In addition, ELT is much more flexible than ETL. With ELT, users can run new transformations, test and enhance queries, directly on the raw data as it is required - without the time and complexity that we've become used to with ETL.

Managing Data Warehouses and Data Lakes

According to Gartner, the data management and data integration needs of businesses today require both small and big, unstructured and structured data. Here's what they suggest about what needs to change in the way of work:

> *"The traditional BI team needs to continue developing clear best practices, with well understood business objectives... there is a second mode of BI which is more fluid and … highly iterative with unforeseen data discovery and is allowed to fail fast."*

This type of conversation has created a lot of talk in the industry about data warehouses vs. data lakes. The data lake concept is a new way of thinking about big data for unstructured data made for infinite scaling - using tools like Hadoop for implementing the second mode of BI work described by Gartner. However, although enterprise still use data warehouses to support a traditional paradigm such as ETL, scalable modern data warehouses such as Redshift and Bigquery can be used to implement the ELT modern paradigm with all its inherent benefits mentioned above.

IBM talks about 5 things that modern big data projects require - showing the need for new data concepts like the data lake. It's the 5 V's:

"

1. *Volume: the volume of (raw) data*
2. *Variety: the variety (e.g. structured, unstructured, semi-structured) of data*
3. *Velocity: the speed of data processing, consummation or analytics of data*

4. *Veracity: the level of trust in the data*
5. *(Value): the value behind the data*

“

ETL continues to be a good match when dealing with legacy data warehouses - looking at smaller subsets and moving them into the data warehouse. But it’s hard to provide a solution with ETL for the 5 V’s as you go down the list - how to deal with the volumes? The unstructured data? Speed? etc.

The ELT approach opens opportunity for working in a more fluid, iterative BI environment due to its efficiency and flexibility. ELT supports both the data warehouse and extends to the data lake concept - enabling the incorporation of unstructured data into its BI solution.

Summarizing 10 Pros & Cons

To summarize the two approaches, we’ve grouped the differences into 10 criteria:

Criteria	ETL	ELT
1. Time - Load	Uses staging area and system, extra time to load data	All in one system, load only once
2. Time - Transformation	Need to wait, especially for big data sizes (as data grows, transformation time increases)	All in one system, speed is not dependant on data size
3. Time - Maintenance	High maintenance - choice of data to load and transform and must do it again if deleted or want to enhance the main data repository	Low maintenance - all data is always available
4. Implementation complexity	At early stage, requires less space and result is clean	Requires in-depth knowledge of tools and expert design of the main large repository
5. Analysis & processing style	Based on multiple scripts to create the views - deleting view means deleting data	Creating adhoc views - low cost for building and maintaining
6. Data limitation or restriction in supply	By presuming and choosing data a priori	By HW (none) and data retention policy
7. Data warehouse support	Prevalent legacy model used for on-premises and relational, structured data	Tailored to using in scalable cloud infrastructure to support structured, unstructured such big

		data sources
8. Data lake support	Not part of approach	Enables use of lake with unstructured data supported
9. Usability	Fixed tables, Fixed timeline, Used mainly by IT	Ad Hoc, Agility, Flexibility, Usable by everyone from developer to citizen integrator
10. Cost-effective	Not cost-effective for small and medium businesses	Scalable and available to all business sizes using online SaaS solutions.

Final Thoughts

ETL is outdated. It helped to cope with the limitation of the traditional rigid and data center infrastructures which with the cloud are no longer a barrier today. In organizations with large data sets of even only a few terabytes, load time can take hours, depending on the complexity of the transformation rules.

ELT is an important part of the future of data warehousing. With ELT, businesses of any size can capitalize on the current technologies. By analyzing larger pools of data with more agility and less maintenance, businesses gain key insights to create a real competitive advantages and excel in their business.