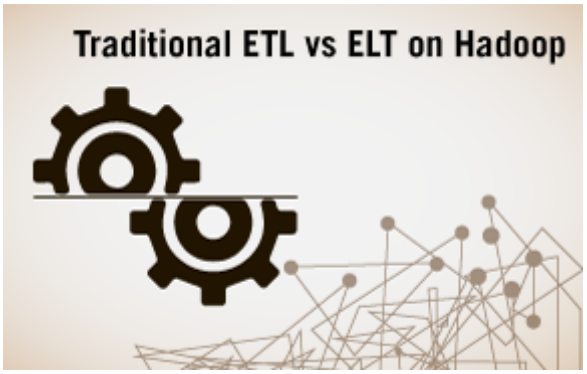
Traditional ETL vs ELT on Hadoop

JULY 04, 2017



Traditional ETL vs ELT on Hadoop

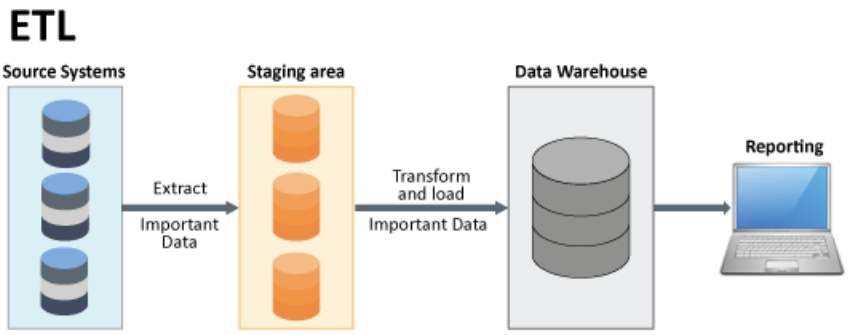
The advent of Hadoop has taken enterprises by storm. The majority of enterprises today have one or more Hadoop cluster at various stages of maturity within their organization. Enterprises are trying to cut down on infrastructure and licensing costs by offloading storage and processing to Hadoop. In almost all these cases, the warehouse area is the first candidate for Hadoop adoption primarily due to the fact that the data warehouse hosts the largest amount of data in the enterprise, and also because it is the most processor-heavy process in the enterprise.

Until fairly recently, the data warehouse area has been dominated by RDBMSes and traditional ETL tools. ETL processes form the backbone of all the data warehousing tools. This has been *the* way to process large volumes of data and prepare it for reporting and analysis. That notion, however, has been challenged of late with the rise of Hadoop. Traditional ETL tools are limited by problems related to scalability and cost overruns. These have been ably addressed by Hadoop. And while ETL processes have traditionally been solving data warehouse needs, the 3 Vs of big data (volume, variety and velocity) make a compelling use case to move to.

Let us take a comparative look at the traditional ETL process vs ELT on Hadoop at a high level.

ETL stands for **Extract**, **Transform** and **Load**. The ETL process typically extracts data from the source / transactional systems, transforms it to fit the model of data warehouse and finally loads it to the data warehouse.

The transformation process involves cleansing, enriching and applying transformations to create the desired output. Data is usually dumped to a staging area after extraction. In some cases, the transformations might be applied on the fly and loaded to the target system without the intermediate staging area. The diagram below illustrates a typical ETL process.



The development process usually starts from output, backwards, as the data model for target system (i.e. data warehouse) is predefined. Since the data model for the data warehouse is predefined, only the relevant and important data is pulled from the source system and loaded to the data warehouse.

Advantages of ETL Process

- **Ease of development**: Since the process usually involves development from the output-backwards and loading just the relevant data, it reduces complexity and time involved in development.
- **Process maturity**: This process has been *the* norm for data warehouse development and has been in practice for over two decades. The ETL process is quite mature with multiple production implementations and well defined best practices and processes.
- **Tools availability**: A prolific number of tools are available that implement ETL. This provides flexibility in choosing the most appropriate tool.
- **Availability of expertise**: The decades of existence and extensive adoption of ETL process across the board have ensured abundant availability of ETL experts.

Disadvantages of ETL Process

- **Flexibility**: The ETL process loads only the important data, as identified at design time. If there is a need to add an additional data attribute, or if a new data attribute is introduced in the system, it would involve updating and re-engineering the entire ETL routine. This adds to time and cost involved in development and maintenance of ETL process.
- **Hardware**: Most ETL tools come with their own hardware requirements. They have proprietary execution engines which do not use the existing data warehouse hardware. This leads to additional costs.
- **Cost**: The maintenance, hardware and licensing costs of the ETL tools add up to the total cost of operating and maintaining the ETL process.
- **Limited to relational data**: Traditional ETL tools are mostly limited to processing relational data. They are unable to process semi-structured and unstructured data like social media feeds, log files, etc.

ELT

ELT stands for **Extract**, **Load** and **Transform**. As opposed to loading just the transformed data in the target systems, the ELT process loads the entire data into the data lake. This results in faster load times. Optionally, the load process can also perform some basic validations and data cleansing rules. The data is then transformed for analytical reporting as per demand. Though the ELT process has been in practice for some time, it is only getting popular now with the rise of Hadoop. The diagram below illustrates a typical ELT process on Hadoop.

Advantages of ELT Process

- **Separation of concerns**: The ELT process separates the loading and transformation tasks into independent blocks and thereby minimizes the interdependencies between these processes. This makes project management easier as the project can be broken down into manageable chunks. This also minimizes the risks as a problem in one area does not affect the other.
- **Flexible and future-proof**: In ELT implementation, entire data from the source systems is already available in the data lake. This, combined with the isolation of the transformation process, guarantees that future requirements can easily be incorporated into the warehouse structure.
- **Utilizes existing hardware**: Hadoop uses the same hardware for storage as well as for processing. This helps in cutting down additional hardware cost.
- **Cost effective**: All the points mentioned above in addition to the open source Hadoop framework cuts considerable cost of operating and maintaining the ELT process.
- **Not limited to relational data**: With Hadoop, the ELT processes can process semi-structured and unstructured data.

Disadvantages of ELT Process

- **Process maturity**: Though the ELT process has been there for a while, it has not been widely adopted. However, the ELT process is gaining popularity and adoption with the rise of Hadoop. The collaboration across the industry for implementing best practices in ELT is increasing.
- **Tools availability**: As a result of limited adoption, the number of tools available to implement ELT processes on Hadoop is currently limited. One tool aimed at overcoming this limitation is Hydrograph, which was created specifically for developing ELT processes in the big data ecosystem.
- **Availability of expertise**: The limited adoption of ELT technology again has an impact on the availability of experts on ELT. The experts for ELT on Hadoop are currently scarce. However, this is changing fast. The immense popularity and adoption of Hadoop and ELT on Hadoop is increasing the number of people working on these technologies.

The Way Forward

Though the ETL process and traditional ETL tools have been serving the data warehouse needs, the changing nature of data and its rapidly growing volume have stressed the need to move to Hadoop. Apart from the obvious benefits of cost effectiveness and scalability of Hadoop, ELT on Hadoop provides flexibility in data processing environment.

Transitioning from traditional ETL tools and traditional data warehouse environments to ELT on Hadoop is a big challenge - a challenge almost all enterprises are currently facing. Apart from being a change in environment and technical skillset, it requires a change in mindset and approach. ELT is not as simple as rearranging the letters. On one hand you have developers with years of ETL tool experience and business knowledge; on the other hand you have the long term benefit of moving to ELT on Hadoop. Training the existing workforce, who is conversant with the drag-drop GUI based tools, to work on java programming is a time consuming challenge. In order to bridge this technology gap, Bitwise contributed to the development of Hydrograph, an open source ELT tool on Hadoop.