AWS and Data Analysis

>Data categories:

1. Structured: Perfect schema (Amazon RDS)
2. Semi-Structured: Key-Value pairs (Amazon DynamoDB)
3. Unstructured: Comes real-time from IOT devices, scrapped from web etc. (First stored in a data lake then used)

>Meta Data:

label that tells you more about an object.

ex: Details of a book like its page count, author ,book shelf no. etc to locate it.

ELT: Another Data Architecture Approach

What is ELT?

Extract, load, and transform (ELT) is an extension of extract, transform, and load (ETL) that reverses the order of operations. You can load data directly into the target system before you process it. The intermediate staging area is not required because the target data warehouse has data mapping capabilities within it. ELT

has become more popular with the adoption of cloud infrastructure, which gives target databases the processing power they need for transformations.

ETL compared to ELT

ELT works well for *high-volume* (large amounts), *high-variety* (unstructured) datasets that require *high velocity* (constant loading). It also works well for big data because you can plan for analytics after data extraction and storage. It performs most transformations in the analytics stage, and focuses on loading minimally processed raw data into the data warehouse.

Extract, transform, and load (ETL) and extract, load, and transform (ELT) are two data-processing approaches for analytics. Large organizations have several hundred (or even thousands) of data sources from all aspects of their operations —such as applications, sensors, IT infrastructure, and third-party partners. They need to filter, sort, and clean this large data volume to make it useful for analytics and business intelligence. The ETL approach uses a set of business rules to process data from several sources before they are centrally integrated. The ELT approach loads data as-is and transforms it at a later stage, depending on the use case and analytics requirements. The ETL process requires more definition at the beginning. Analytics must be involved from the start to define target data types, structures, and relationships. Data scientists mainly use ETL to load legacy databases in the data warehouse, while ELT has become the norm today.

What are the similarities between ETL and ELT?

Both extract, transform, and load (ETL) and extract, load, and transform (ELT) are sequences of processes that prepare data for further analysis. They capture, process, and load data for analysis across three steps.

Extraction

Extraction is the first step of both ETL and ELT. This step is about collecting raw data from different sources. These data sources could be databases, files, software as a service (SaaS) applications, Internet of Things (IoT) sensors, or application events. You can collect semi-structured, structured, or unstructured data at this stage.

Transformation

In the *ETL* process, transformation is the second step. However, in the *ELT* process, transformation is the third step. This step focuses on changing raw data from its original structure into a format that meets the requirements of the target system where you plan to store the data for analytics. Here are some examples of transformation:

Changing data types or formats

- Removing inconsistent or inaccurate data.
- Removing data duplication.
- You apply rules and functions to clean and prepare data for analysis in the target system.

Load

In this phase, you store data into the target database. ETL processes load data as a final step, so that reporting tools can use it directly to generate actionable reports and insights. However, in ELT, you still need to transform the extracted data after loading it.

How do the ELT and ETL processes differ from each other?

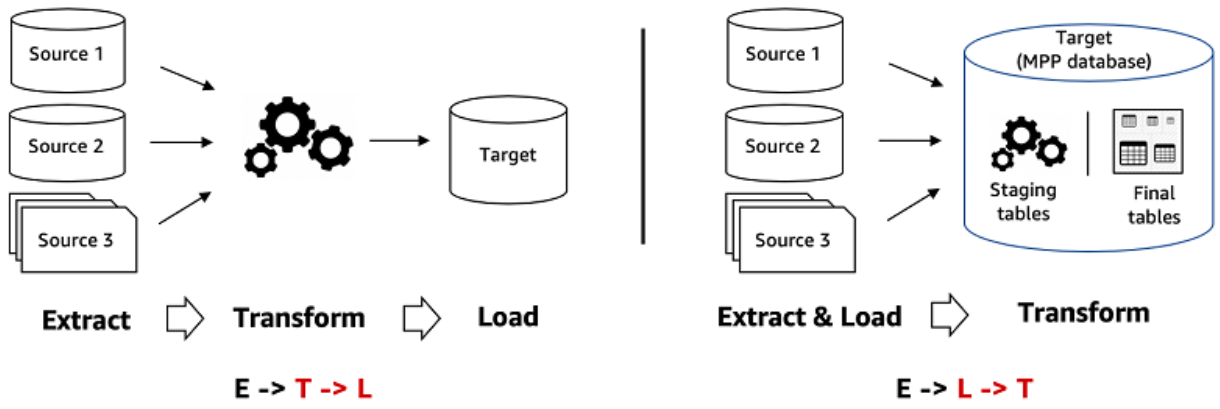
This section outlines the processes of extract, transform, and load (ETL) and extract, load, and transform (ELT).

ETL process

ETL has three steps:

1. You extract raw data from various sources.
2. You use a secondary processing server to transform that data.
3. You load that data into a target database.

The transformation stage ensures compliance with the target database's structural requirements. You only move the data after it is transformed and ready.



ELT process

ELT also has three steps:

1. You extract raw data from various sources.
2. You load it in its natural state into a data warehouse or data lake.
3. You transform it as needed while you're in the target system.

With ELT, all data cleansing, transformation, and enrichment occur within the data warehouse. You can interact with and transform the raw data as many times as you need.

When to use ETL compared to ELT

Extract, load, and transform (ELT) is the standard choice for modern analytics. However, you might consider extract, transform, and load (ETL) in the following scenarios.

Legacy databases

Legacy databases refer to older, established database systems that may use outdated technology or formats. In the context of data processing, ETL (Extract, Transform, Load) is often preferred for integrating with legacy databases due to their predetermined data structures.

Sometimes, it's more beneficial to use ETL to integrate with legacy databases or with third-party data sources that have predetermined data formats. You only

need to transform and load the data into your system one time. After the data is transformed, you can use it more efficiently for all future analytics.

Experimentation

In large organizations, data engineers conduct experiments—such as discovering hidden data sources for analytics and trying new ideas to answer business queries. ETL is useful in data experiments when you need to understand the database and its usefulness in a particular scenario.

Complex analytics

ETL and ELT can both be used together for complex analytics that use multiple data formats from varied sources. Data scientists might set up ETL pipelines from some of the sources, and use ELT with the rest. Using both processes can improve the efficiency of analytics and, in some cases, increase application performance.

IoT applications

Internet of Things (IoT) applications that use sensor data streams often benefit from using ETL instead of ELT. For example, some common use cases for ETL at the edge include the following:

- You want to receive data from different protocols and convert it into standard data formats for use in cloud workloads.
- You want to filter high-frequency data, perform averaging functions on large datasets, and then load averaged or filtered values at a reduced rate.
- You want to calculate values from disparate data sources on the local device, and send filtered values to the cloud backend.
- You want to cleanse, deduplicate, or fill missing time series data elements.

Summary of differences: ETL compared to ELT

Category	ETL	ELT
Stands for	Extract, transform, and load	Extract, load, and transform

Process	Take raw data, transform it into a predetermined format, and then load it into the target data warehouse.	Take raw data, load it into the target data warehouse, then transform it just before analysis.
Transformation and Load Locations	Transformation occurs in a secondary processing server.	Transformation takes place in the target data warehouse.
Data Compatibility	ETL works best with structured data.	ELT can handle structured, unstructured, and semi-structured data.
Speed	ETL is slower than ELT.	ELT is faster than ETL because it can use the internal resources of the data warehouse.
Costs	ETL can be time consuming and costly to set up, depending on ETL tools that are used.	ELT can be more cost efficient, depending on the ELT infrastructure that is used.
Security	ETL might require building custom applications to meet requirements for data protection.	You can use the built-in features of the target database to manage data protection.

Different techniques for data analytics

Data analytics can use many computing techniques. The following techniques are some of the most common ones.

Natural language processing (NLP)

Natural language processing is the technology that makes computers understand and respond to spoken and written human language. Data analysts use this technique to process data such as dictated notes, voice commands, and chat messages.

Text mining

Data analysts use text mining to identify trends in text data, such as email messages, Tweets, research, and blog posts. It can be used for sorting news content, customer feedback, and client email messages.

Sensor data analysis

Sensor data analysis is the examination of the data that different sensors generate. It is used for predictive machine maintenance, shipment tracking, and other business processes where machines generate data.

Outlier analysis

Outlier analysis (or anomaly detection) identifies data points and events that deviate from the rest of the data.

How data analytics is used in business

Businesses capture statistics, quantitative data, and information from multiple customer-facing and internal channels. However, finding key insights requires careful analysis of a large amount of data, which is not a small feat.

For example, data analytics and data science can add value to a business by improving customer insights. Data analytics can be conducted on datasets from various customer data sources, such as the following:

- Third-party customer surveys
- Customer purchase logs
- Social media activity
- Clickstream data
- Website cookies
- Website or application statistics

Analytics can reveal hidden information, such as customer preferences, popular pages on a website, the length of time customers spend browsing, customer feedback, and interaction with website forms. With this information, businesses can respond efficiently to customer needs and increase customer satisfaction.

Question:

A cloud consultant is working for a company that uses cloud services for different aspects of their business. The company has deployed a diverse range of resources, including virtual machines, storage buckets, and databases. However, none of these resources have tags, and the data analytics team would like to track

and optimize different components of the environment. The consultant decides to recommend the use of tag metadata to help the data analytics team perform their analysis. Which option BEST describes how the data analytics team should implement resource metadata?

Answer: **Attach specific labels to resources.**