**Task # 4**

- What is ETL? in detail.

- What is ELT? in detail.

- 3 Tier Architecture in DE

- ETL Tools (any 3)

**ETL**

ETL stands for Extract, transform, and load. It is the process data-driven organizations use to gather data from multiple sources and then bring it together to support discovery, reporting, analysis, and decision-making.

The data sources can be divergent in type, format, volume, and reliability, hence the data required to be processed to be helpful when delivered together. The target data stores can be databases, data warehouses, or data lakes, based on the objectives and technical execution. There are the following steps of ETL which are as follows −

**Extract** − During extraction, ETL recognizes the data and duplicate it from its sources, therefore it can transport the data to the target datastore. The data can appear from structured and unstructured sources, including files, emails, business software, databases, equipment, sensors, third parties, etc.

There are different ways to perform the extraction such as

**Partial Extraction** − The simplest way to access the information is if the source system notifies us when a record has been modified.

**Partial Extraction (with update notification)** − It is not all systems can provide an announcement in case an update has taken place; but, they can mark to those records that have been transformed and support an extract of such records.

**Full extract** − Certain systems cannot identify which data has been changed at all. In this case, a full extract is the only feasibility to extract the record out of the system. This approach needed having a copy of the final extract in the same format so it can identify the changes that have been created.

**Transform** − The second step includes transforming the raw information that has been extracted from the sources into a format that can be used by several applications. In this phase, data have cleansed, mapped, and transformed, providing to a definite schema, so it finds operational needs.

This process requires several types of transformation that provide the quality and integrity of data. Data is not generally loaded precisely into the target data source, but alternatively, it is frequent to have it uploaded into a staging database.

This step ensures a quick rollback in case something does not go as planned. During this phase, it can create audit documents for regulatory compliance, or diagnose and repair some data problems.

**Load** − ETL moves the transformed information into the target datastore. This step can require the original loading of all the source information, or it can be the loading of incremental changes in the source information. It can load the data in real-time or in scheduled batches.

**ELT**

ELT stands for Extract, Load, and Transform. It is a data integration process for transferring raw data from a source server to a data system (such as a data warehouse or data lake) on a target server and then fitting the data for downstream uses.

The extract and load procedure can be isolated from the transformation phase. Isolating the load phase from the transformation process deletes an inherent dependency between these phases. It can include the data necessary for the transformations, the extract and load process can include an element of data that can be essential in the future. The load process can take the entire source and load it into the warehouses.

**Advantages of ELT**

There are various advantages of ELT which are as follows

**Real-time, flexible data analysis.** Users have the adaptability to analyze the whole data set, such as real-time information, in various movement, without having to wait for it to extract, transform and load information.

**Lower cost and lower maintenance** ELT benefits from a powerful ecosystem of cloud-based platforms which provide much lower costs and several plan options to store and process data. The ELT procedure usually needed low maintenance given that all information is continually available and the transformation procedure is usually automated and cloud-based.

**Efficient** ELT can take advantage of the computing power of existing hardware to perform transformations.

**Flexible resulting data set** When it uses ELT, it can move the entire data set to the target. This can be useful if it doesn't want to transform the data before it can move it, or it wants flexibility in the schema for the target data.

The data is nearly simple yet massive, including log documents and sensor data. In this case, the transformations that take place in the target might be relatively simple, and the benefit comes from the ability of the target datastore to load massive volumes of data quickly.

**Simplifying management** − ELT divides the loading and transformation services, minimizing the interdependencies among these phase, lowering risk, and integrate project administration.

**Leveraging the latest technologies** − ELT solutions exploit the power of new technologies to shift development, security, and compliance across the enterprise.

**Scalability** − The scalability of cloud infrastructure and hosted services like integration platform-as-a-service (iPaaS) and software-as-a-service (SaaS) give organizations the ability to expend resources on the fly. They insert the compute time and storage space essential for even a massive data transformation service.

**Future-proofed data sets** − ELT implementations can be used precisely for data warehousing systems, but ELT is used in the data lake method in which data is collected from an area of sources. Therefore, it is mixed with the separation of the transformation process, creating it simpler to make eventual modifies to the warehouse architecture.

**3 Tier Architecture in DE**

Three-tier architecture allows any one of the three tiers to be upgraded or replaced independently.

The user interface is implemented on any platform such as a desktop PC, smartphone or tablet as a native application, web app, mobile app, voice interface, etc. It uses a standard graphical user interface with different modules running on the application server.

The relational database management system on the database server contains the computer data storage logic.

The middle tiers are usually multitiered.

Since the three are not physical but logical in nature, they may run in different servers both in on-premises based solutions, as well as in software-as-a-service (SaaS).

**ETL Tools**

**Hadoop**

[**Hadoop**](https://hadoop.apache.org/) is an open-source framework for processing and storing big data in clusters of computer servers. It is considered the foundation of big data and enables the storage and processing of large amounts of data.

The Hadoop framework consists of several modules, including the Hadoop Distributed File System (HDFS) for storing data, MapReduce for reading and transforming data, and YARN for resource management. Hive is commonly used to convert SQL queries into MapReduce operations.

Companies considering Hadoop should be aware of its costs. A significant portion of the cost of implementing Hadoop comes from the computing power required for processing and the expertise needed to maintain Hadoop ETL, rather than the tools or storage themselves.

**Oracle Data Integrator**

Oracle Data Integrator is an ETL tool that helps users build, deploy, and manage complex data warehouses. It comes with out-of-the-box connectors for many databases, including Hadoop, EREPs, CRMs, XML, JSON, LDAP, JDBC, and ODBC.

ODI includes Data Integrator Studio, which provides business users and developers with access to multiple artifacts through a graphical user interface. These artifacts offer all the elements of data integration, from data movement to synchronization, quality, and management.

### SAP BusinessObjects Data Services

[**SAP BusinessObjects Data Services**](https://www.sap.com/products/technology-platform/data-services.html) is an enterprise ETL tool that allows users to extract data from multiple systems, transform it, and load it into data warehouses.

The Data Services Designer provides a graphical user interface for defining data pipelines and specifying data transformations. Rules and metadata are stored in a repository, and a job server runs the job in batch or real time.

However, SAP data services can be expensive, as the cost of the tool, server, hardware, and engineering team can quickly add up.

SAP Data Services is a good fit for companies that use SAP as their Enterprise Resource Planning (ERP) system, as it integrates seamlessly with SAP Data Services