**Task 4**

Name: Ayesha Bibi

Date: 16/3/2023

**ETL:**

ETL is a data integration process that stands for extract, transform, and load.

* Extracting data from varied data sources
* Perform transformation to optimize it
* Loading data into the desired database

ETL is a traditional method used by data warehouses for integration and loading data for computation and analysis. ETL is recommended more often for creating smaller target data repositories that require less frequent updates.

**How Does ETL Process Work?**

1. **Extract:**

In this stage of ETL, data is extracted from disparate systems which can be structured and unstructured and present in its raw form. These data sources may include SQL or [NoSQL](https://www.ibm.com/topics/nosql-databases) servers, CRM and ERP systems, Flat files, Email, and Web pages. etc.

1. **Transform:**

In this stage, raw data is processed to make it useful and insightful. Transformation includes Filtering, cleaning, de-duplicating, splitting, merging, sorting, Formatting, summarizing, validating, and authenticating the data.

1. **Load**

In the last step of ETL, the transformed data is moved into a target data warehouse. Depending on your business needs, data can be loaded in batches or all at once. The exact nature of the loading will depend upon the data source, ETL tools, and various other factors.

**Challenges:**

Overall, the ETL process is an essential process in data warehousing that helps to ensure that the data in the data warehouse is accurate, complete, and up-to-date. However, it also comes with its own set of challenges which are as follows:

1. Management of different Data sources
2. Scalability of huge amounts of day-to-day data of organizations
3. Ensure the Reliability of Data
4. Accuracy in Transformation

**ELT**

ETL abbreviated as extract, Load, and Transform. ELT is often used when large volumes of data are involved and generally costs less.

* Extract: Extract raw data from multiple sources
* Load – Unlike the ETL method, after data extraction, it is simply loaded into a data store without being cleaned or standardized.
* Transform – In ELT, transformation is done at the last, meaning data is cleaned and standardized in the data store rather than a separate staging area.

ELT is particularly useful for high-volume, unstructured datasets as loading can occur directly from the source. ELT can be more ideal for big data management since it doesn’t need much upfront planning for data extraction and storage.

**The Biggest Advantages of ELT**

Here are some other benefits of ELT:

1. **Flexibility:**

The primary advantage of ELT is flexibility and ease of storing any type of information either structured or unstructured data. It is very useful at the time when you don’t have the time or ability to transform and structure it first.

1. **Fast Speed**

ELT is the faster approach of accessing the exact data that users need to transform and analyze after all the data go into the system.

1. **Low Maintenance**

In ELT, users generally won't need "high-touch" maintenance since it is cloud-based, and utilizes automated solutions.

1. **Quicker Loading**

Data is quickly loaded into the data store due to transformation occurring at the end.

**ETL Tools:**

1. **Talend Open Studio**
2. Oracle Data Integrator (ODI).
3. Skyvia
4. Hevo
5. **Hadoop**

**3-Tier Architecture:**

Three-tier architecture consists of three tiers which are as follows:

1. **Presentation tier**

Presentation tier is the top tier of 3-tier architecture. It is the user interface, where the end-user interacts with the application. It is basically designed for taking the query from the user and displaying its result to it. The user can see its query result on a web browser, desktop application, or mobile GUI. HTML, CSS, and JavaScript are used for developing Web presentation tiers. While the desktop application can be written in a variety of high-level languages depending on the platform. It communicates with the application tier to give the query and retrieve the result from it to give back to the end-user on UI.

1. **Application tier**

The application tier, also known as the logic tier or middle tier, is the heart of the application. In this tier, real processing of data occurs using business logic. The application tier communicates with the data tier using [API](https://www.ibm.com/topics/api) calls. The application tier act as a bridge between the presentation tier and the data tier cannot through which all the communication occurs.

1. **Data tier**

The data tier, also known as the database tier, is the back end, where the information processed by the application is stored and managed. This can be a [relational database management system](https://www.ibm.com/topics/relational-databases) such as [PostgreSQL](https://www.ibm.com/topics/postgresql), MySQL, MariaDB, Oracle, DB2, Informix, or Microsoft SQL Server, or in a [NoSQL](https://www.ibm.com/topics/nosql-databases) Database server such as Cassandra, [CouchDB](https://www.ibm.com/topics/couchdb) or [MongoDB](https://www.ibm.com/topics/mongodb).

**Tier vs. layer:**

A 'layer' refers to a functional division of the software, but a 'tier' refers to a functional division of the software that runs on infrastructure separate from the other divisions. for example, The Contacts app on your phone, is a three-layer application, but a single-tier application, because all three layers run on your phone.

**Full load:**

In full load, the bulk of data is dumped, or loaded into the desired system without any increments. If in future any updates occur then completely replaced the old data with the new one. This is traditional approach inefficient, although relatively easy to set up and maintain.

**For Example:**

take a hotel that uploads all of its sales through the ETL process in data warehouse at the end of each day. Let’s say 50 sales were made on a Tuesday, so that on Tuesday night a table of 50 records would be uploaded. Then, on Wednesday, another 30 sales were made which need to be added. So, on Wednesday night, assuming a full load, Tuesday’s 50 records as well as Wednesday’s 30 records are uploaded.

**Incremental Load:**

Unlike the full load, In incremental load instead of loading all the data, only the updated data is loaded to the target store in the form of batches. This is the faster approach that saves time and resources but increases complexity and maintenance issues due to the nature of having to manage them as individual batches rather than one big group.

**For Example:**

Following the previous example, the hotel that made 30 sales on Wednesday will load only the additional 30 records to the sales table, instead of reloading all records.

**Historical Data:**

Historical Data is the data collected and stored in data stores about past events and circumstances. Storing data takes space and resources so the organization decides which data to store for short-term history and which for long-term history.

Mostly, It includes the data generated manually or automatically. These might be the press releases, [log files](https://www.techtarget.com/whatis/definition/log-log-file), [financial reports](https://www.techtarget.com/searcherp/definition/financial-reporting), project and product documentation and email and other communications.

**How is historical data used and why is it important?**

Historical data is used to make crucial strategic decisions about the present and future of business.

Managers use historical data to:

* Track organizational performance over time
* Identify areas of improvement
* Make predictions about future trends.

**How is historical data stored?**

Historical data can be stored in a variety of ways, including [databases](https://www.techtarget.com/searchdatamanagement/definition/database), [spreadsheets](https://www.techtarget.com/whatis/definition/spreadsheet), and text files. Since storing, data takes space and resources. So, It is vital to choose a [storage method](https://www.techtarget.com/searchstorage/definition/storage) that is efficient and scalable so historical data can be easily accessed and analyzed.