Name: Moiz Zulfiqar

Task 4

ETL

ETL, or extract, transform, and load, is a data integration process that combines data from various sources into a single, consistent data store that is loaded into a data warehouse or other target system. It is the primary method to process data for data warehousing projects and provides the foundation for data analytics and machine learning workstreams. ETL cleanses and organizes data through a series of business rules in a way that addresses specific business intelligence needs.

In the extract step, raw data is copied or exported from source locations to a staging area. Data management teams can extract data from various sources, including structured or unstructured data sources such as SQL or NoSQL servers, CRM and ERP systems, flat files, email, and web pages.

In the transform step, the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. Tasks in this phase may include filtering, cleansing, de-duplicating, validating, and authenticating the data, performing calculations, translations, or summarizations based on the raw data, conducting audits to ensure data quality and compliance, and formatting the data into tables or joined tables to match the schema of the target data warehouse.

In the load step, the transformed data is moved from the staging area into a target data warehouse. This typically involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse. For most organizations that use ETL, the process is automated, well-defined, continuous, and batch-driven. ETL usually takes place during off-hours when traffic on the source systems and the data warehouse is at its lowest.

ETL solutions improve quality by performing data cleansing prior to loading the data to a different repository. ETL is recommended more often for creating smaller target data repositories that require less frequent updating, while other data integration methods—including ELT (extract, load, transform), change data capture (CDC), and data virtualization—are used to integrate increasingly larger volumes of data that changes or real-time data streams.

There are various other data integration methods used to facilitate data integration workflows, including change data capture (CDC), data replication, data virtualization, and stream data integration (SDI). CDC identifies and captures only the source data that has changed and moves that data to the target system. Data replication copies changes in data sources in real-time or in batches to a central database. Data virtualization uses a software abstraction layer to create a unified, integrated, fully usable view of data—without physically copying, transforming, or loading the source data to a target system. SDI continuously consumes data streams in real-time, transforms them, and loads them to a target system for analysis.

There are now many open source and commercial ETL tools and cloud services available with capabilities such as comprehensive automation and ease of use, a visual drag-and-drop interface, support for complex data management, and security and compliance features.

ELT

ELT is a data integration process that stands for “Extract, Load, Transform,” similar to ETL (Extract, Transform, Load), but with a fundamentally different approach to data pre-processing. The ELT process moves raw data from a source system to a destination resource such as a data warehouse. In the Extract stage, data is copied or exported from source locations to a staging area. In the Load stage, the transformed data is moved from the staging area into a data storage area such as a data warehouse or data lake. Finally, in the Transform stage, a schema-on-write approach is employed, which applies the schema for the data using SQL, or transforms the data, prior to analysis.

ELT is preferred over ETL when large amounts of streaming data are generated as ELT allows that data to be loaded immediately and transformed after it reaches its destination. This prevents any slowdown that can often occur if the transformation occurs before the Load function, such as in ETL. ELT is also preferred when the data needs to be transformed after it reaches its destination as the recipient of the data can control data manipulation. ELT allows for on-demand flexibility and scalability and requires a less-powerful server for data transformation, resulting in cost savings and resource efficiencies.

ELT is ideal for environments that require fast access to data, particularly in cloud environments that will often include applications that are accessed on-demand continuously. ELT is used in high-volume or real-time data use environments such as stock exchanges or large-scale wholesale distributors of stocks, industrial components, and other materials, which require real-time access to current data for immediate access to business intelligence. ELT is also used by meteorological systems such as weather services that collect, collate, and use large amounts of data.

In conclusion, ELT is a data integration process that has gained adoption with the transition to cloud environments. ELT provides several advantages over ETL such as faster availability, decoupling of the transformation and load stages, scalability, cost savings, resource efficiencies, and flexibility. ELT is well-suited for data utilized within cloud environments that require fast access to data, particularly in high-volume or real-time data use environments.

3 Tier Architecture in DE

The three-tier architecture is a widely adopted approach in data engineering that separates different layers of data processing into three distinct tiers. Each tier is responsible for a different aspect of data processing, and they work together to create an efficient and scalable data processing system.

The three tiers of data processing are as follows:

1. Data Storage Tier: The data storage tier, also known as the data layer or the persistence layer, is responsible for storing and retrieving data. It is usually a database

management system (DBMS) that provides the interface to store, access, and manage data. The data storage tier is where the raw data is initially stored before it is processed, transformed, and analyzed. This tier includes all types of databases like relational databases, NoSQL databases, data lakes, and data warehouses. Data in this tier is typically stored in a format optimized for efficient storage and retrieval, with minimal processing.

2. Data Processing Tier: The data processing tier, also known as the business layer or the middle tier, is responsible for processing, transforming, and analyzing the data stored in the data storage tier. It is where the data is cleaned, validated, and transformed before being analyzed. This tier includes various data processing tools and frameworks like Apache Spark, Apache Flink, Hadoop, and other big data processing engines. The data processing tier is also responsible for performing complex data operations like machine learning and data mining, to extract meaningful insights from the data.
3. Presentation Tier: The presentation tier, also known as the application layer or the client tier, is responsible for presenting the data processed in the data processing tier to the end-user. This tier includes various data visualization tools, dashboards, and reporting tools that provide a user-friendly interface for the end-user to access and interact with the data. The presentation tier provides data in a format that is easily understandable to the user, and it can be accessed through various devices like laptops, desktops, tablets, and smartphones.

The three-tier architecture provides several benefits to data engineering, including scalability, modularity, and flexibility. It enables data engineers to design a system that can be easily scaled up or down based on the changing needs of the business. The modularity of the architecture enables easy maintenance and updates to each tier without affecting the other tiers. The flexibility of the architecture allows data engineers to use different tools and technologies to build each tier, based on the specific requirements of the business.

In summary, the three-tier architecture provides a framework for building efficient and scalable data processing systems that can handle large volumes of data. By separating the data processing into distinct layers, data engineers can optimize each tier for its specific function, resulting in a more efficient and flexible data processing system.

ETL Tools

ETL tools are essential components in data engineering, and they play a critical role in the development and maintenance of data pipelines. ETL tools are used to extract data from multiple sources, transform the data to suit specific requirements, and load the data into target applications.

1. Informatica PowerCenter is one of the leading ETL tools available in the market today. It offers a wide range of connectors for cloud data warehouses and lakes, including AWS, Azure, Google Cloud, and SalesForce. PowerCenter includes several services that help users design, deploy, and monitor data pipelines, such as the Repository Manager, the Designer, and the Workflow Manager. These tools are designed to save time and simplify workflows, making PowerCenter an efficient and powerful ETL tool.
2. Apache Airflow is an open-source ETL tool that allows users to programmatically author, schedule, and monitor workflows. Airflow uses directed acyclic graphs

(DAGs) to define workflows, which allows for clear visualization and management of tasks and dependencies. Airflow also integrates with other tools commonly used in data engineering and data science, such as Apache Spark and Pandas. Companies using Airflow can benefit from its ability to scale and manage complex workflows, as well as its active open-source community and extensive documentation.

3. IBM Infosphere Datastage is an enterprise ETL tool that offers a graphical framework for designing data pipelines. With Infosphere Datastage, users can extract data from multiple sources, perform complex transformations, and deliver the data to target applications. Infosphere Datastage is known for its speed, thanks to features like load balancing and parallelization. It also supports metadata, automated failure detection, and a wide range of data services, from data warehousing to AI applications. Infosphere Datastage offers a range of connectors for integrating different data sources and integrates seamlessly with other components of the IBM Infosphere Information Server.

References

- ETL: https://www.ibm.com/topics/etl#:~:text=the%20next%20step-,What%20is%20ETL%3F,warehouse%20or%20other%20target%20system
- ELT: https://www.ibm.com/topics/elt#:~:text=ELT%2C%20which%20stands%20for%20%E2%80%9CExtract,such%20as%20a%20data%20warehouse
- ETL Tools: https://www.datacamp.com/blog/a-list-of-the-16-best-etl-tools-and-why-to-choose-them