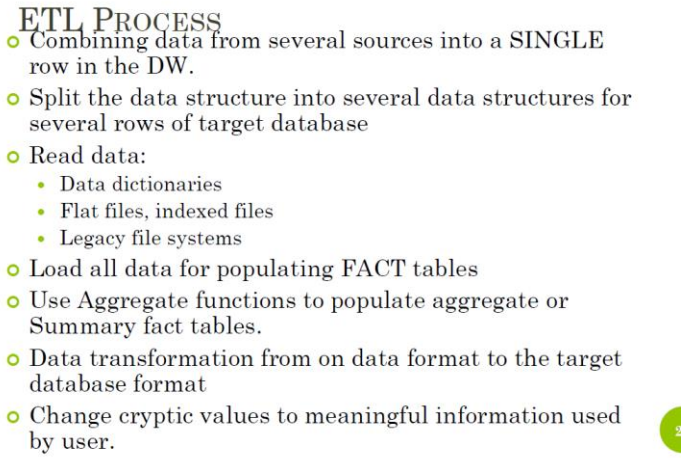
Mod-4

What is ETL?

The mechanism of extracting information from source systems and bringing it into the data warehouse is commonly called **ETL**, which stands for **Extraction, Transformation and Loading**.



Extraction

- Extraction is the operation of extracting information from a source system for further use in a data warehouse environment. This is the first stage of the ETL process.
- Extraction process is often one of the most time-consuming tasks in the ETL.

DATA CLEANSING

Data cleansing (also known as *data scrubbing*) is the name of a process of correcting and - if necessary - eliminating inaccurate records from a particular database. The purpose of data cleansing is to detect so called dirty data (incorrect, irrelevant or incomplete parts of the data) to either modify or delete it to ensure that a given set of data is accurate and consistent with other sets in the system.

Transformation

Transformation is the core of the reconciliation phase. It converts records from its operational source format into a particular data warehouse format. If we implement a three-layer architecture, this phase outputs our reconciled data layer.

Loading

The **Load** is the process of writing the data into the target database. During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible.

1)Extraction:

When performing data extraction for a data warehouse or data integration process, it is important to identify the relevant data sources that contain the data you need. Here are some steps to help identify data sources for extraction:

1. Understand the Business Requirements: Start by understanding the specific business requirements and objectives of the data extraction process. Identify the data elements and entities that are necessary for analysis or integration.
2. Identify Internal Systems: Determine which internal systems or applications within your organization hold the required data. This could include databases, enterprise resource planning (ERP) systems, customer relationship management (CRM) systems, human resources management systems (HRMS), and other operational systems.
3. Explore Data Documentation: Review any available documentation, such as system documentation, data dictionaries, or data catalogs, that describe the data sources within your organization. These documents may provide insights into the data structure, relationships, and sources.
4. Engage with Data Stewards and Subject Matter Experts: Collaborate with data stewards, subject matter experts (SMEs), or business users who have knowledge about the data sources. They can provide valuable information about the systems, databases, and applications that hold the required data.
5. Analyze Data Flow and Integration Points: Analyze the flow of data within your organization's systems and identify integration points where data is shared or transferred between different systems. This can help identify the systems involved in the data flow.

Immediate data extraction:

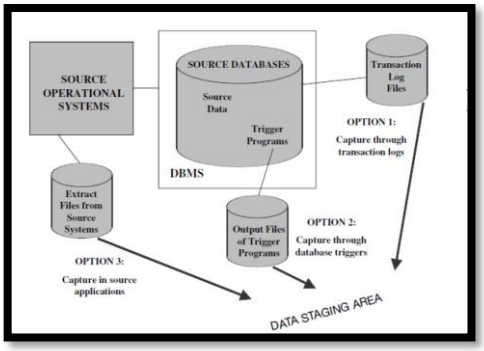
Immediate Data Extraction:

Immediate data extraction, also known as real-time or near real-time extraction, involves capturing and transferring data from source systems as soon as the transaction or event occurs. The data is extracted in real time or with minimal latency, providing up-to-date information for analysis, reporting, or other purposes. Immediate data extraction is suitable when real-time or near real-time data is required for immediate decision-making, monitoring, or operational needs. It often utilizes techniques like change data capture (CDC), event-driven architectures, or streaming technologies to capture and transfer data in real time.

Advantages of Immediate Data Extraction:

1. Timeliness: Provides the most up-to-date data for immediate analysis or action.
2. Real-Time Insights: Enables real-time analytics, monitoring, and operational reporting.
3. Rapid Response: Supports immediate decision-making based on current data.
4. Operational Efficiency: Facilitates near real-time integration and synchronization of data across systems.

- Types:
  - Capture through Transaction Logs
  - Capture through Database Triggers
  - Capture in Source Applications



Deffered data extraction:

Deferred Data Extraction:

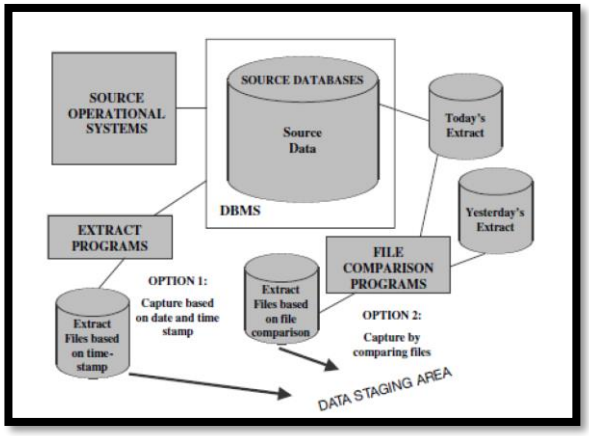
Deferred data extraction, also known as batch processing or scheduled extraction, involves extracting data from source systems at predefined intervals or schedules. Data is extracted in batches, usually at less frequent intervals such as daily, hourly, or weekly. Deferred data extraction is suitable when the immediacy of data is not critical, and there is a need to process and analyze data in bulk or at regular intervals. It is commonly used for data warehousing, data integration, and large-scale data processing.

Advantages of Deferred Data Extraction:

1. Scalability: Allows processing large volumes of data in batch mode, reducing the strain on source systems.
2. Resource Efficiency: Enables the optimization of resources by batching data extraction and processing tasks.

TYPES:

- Capture Based on Date and Time Stamp
- Capture by Comparing Files



2)Transformation:

Data transformation is a crucial step in data integration and data warehousing processes. It involves converting and manipulating data from its source format into a format that is suitable for analysis, reporting, or storage. Here are some common tasks involved in data transformation:

1. Data Cleansing: Data cleansing involves identifying and correcting or removing errors, inconsistencies, or anomalies in the data. This can include handling missing values, standardizing data formats, removing duplicate records, and resolving data quality issues.
2. Data Integration: Data integration involves combining data from multiple sources into a unified format. This task includes mapping and matching data elements, resolving conflicts or inconsistencies between different data sources, and merging data based on common identifiers or keys.
3. Data Formatting and Parsing: Data formatting involves converting data into a standardized format, such as date formats, numerical formats, or text formats. Parsing involves extracting specific elements or fields from a data record, such as splitting a full name into first name and last name.
4. Data Aggregation: Data aggregation involves summarizing or consolidating data at a higher level of granularity. This can include calculating totals, averages, counts, or other statistical measures based on specific dimensions or criteria.
6. Data Derivation and Calculation: Data derivation involves creating new data elements or attributes based on existing data. This can include calculating derived metrics, creating calculated fields, or applying business rules to transform the data.
7. Data Filtering and Selection: Data filtering involves selecting and retaining only the relevant subset of data based on specific criteria or conditions. This can include applying filters to exclude unwanted records or selecting data that meets certain criteria.

3)Data Loading:

Full load
In full load, the entire data from the source is transformed and moved to the data warehouse. The full load usually takes place the first time you load data from a source system into the data warehouse.

Incremental load
In incremental load, the ETL tool loads the delta (or difference) between target and source systems at regular intervals. It stores the last extract date so that only records added after this date are loaded. There are two ways to implement incremental load.

Streaming incremental load
If you have small data volumes, you can stream continual changes over data pipelines to the target data warehouse. When the speed of data increases to millions of events per second, you can use event stream processing to monitor and process the data streams to make more-timely decisions.

Batch incremental load
If you have large data volumes, you can collect load data changes into batches periodically. During this set period of time, no actions can happen to either the source or target system as data is synchronized.

Bulk Loading: Bulk loading involves loading data in large volumes or batches. This technique is suitable when dealing with significant amounts of data. It typically offers faster loading speeds compared to other methods. Bulk loading is commonly used when initially populating a data warehouse or loading large periodic data updates.

3. Loading Dimension Tables: Dimension tables contain descriptive attributes that provide context to the data in the fact tables. Each dimension table represents a specific aspect of the business, such as time, location, product, or customer. The dimension tables are loaded first to establish the reference points for the fact tables.
   - Create or Truncate: If the dimension table already exists, it may need to be truncated or cleared before loading new data. If the table does not exist, it needs to be created with the appropriate structure.
   - Load Data: The dimension table is populated with the relevant data. This can involve inserting new records, updating existing records, or a combination of both, depending on the type of changes being made.
4. Loading Fact Tables: Fact tables contain the numerical measures or metrics that represent the business transactions or events. The fact tables are loaded after the dimension tables have been loaded and the necessary reference keys are established.
   - Create or Truncate: Similar to the dimension tables, the fact table may need to be created or truncated before loading data.
   - Load Data: The fact table is loaded with the measures or metrics, along with the corresponding foreign keys that link to the dimension tables. This can involve inserting new records or updating existing records, depending on the type of data being loaded.

Data quality measures how well a dataset meets criteria for accuracy, completeness, validity, consistency, uniqueness, timeliness, and fitness for purpose.

Issues in data cleansing:

Data cleansing is a critical step in data management that involves identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data. However, there can be various challenges and issues that arise during the data cleansing process. Some common issues in data cleansing include:

1. Missing Values: Missing values occur when data is not available or not recorded for certain attributes. Handling missing values requires decision-making on how to impute or handle those missing values based on the data and domain knowledge.
2. Inconsistent Data Formats: Data may be inconsistent in terms of formats, such as different date formats, inconsistent units of measurement, or variations in naming conventions. Resolving these inconsistencies requires standardizing the data to ensure consistency and compatibility.
3. Duplicates: Duplicate records or entries can occur due to data entry errors, system glitches, or merging of data from multiple sources. Identifying and removing duplicates is important to maintain data accuracy and prevent distortions in analysis or reporting.
4. Data Integrity Issues: Data integrity issues can arise due to constraints violations, referential integrity problems, or logical inconsistencies in the data. It is crucial to identify and resolve these issues to ensure data reliability and validity.
5. Outliers: Outliers are data points that significantly deviate from the normal or expected patterns. Outliers can distort analysis and statistical calculations. Identifying and handling outliers appropriately is essential to avoid misleading results.

9. Scalability and Performance: Data cleansing processes can be computationally intensive and time-consuming, especially when dealing with large volumes of data. Ensuring scalability and performance of data cleansing techniques is crucial for efficient data processing.