AD Module 4 (ETL Process)

* Introduction to ETL Process.

→ ETL Stands for Extract, Transform, Load, which refers to the process of moving data from one place to another, transforming it into a desired format and loading it into the target system.

→ Steps in ETL.

① Extract (E) :- This is the first step in the process where data is extracted from one or more sources. This could be a database, a file or even a web service.

→ Data is typically extracted to a staging area where it can be cleaned and prepared for the next step.

② Transform (T) :- once the data is extracted, it needs to be transformed into a format which is suitable for the target system. This involves cleaning, filtering joining and aggregating the data, as well as applying business rules and data quality checks on the same.

③ Load (L) :- Finally, the transformed data is loaded onto the system, such as a data warehouse, or a business intelligence application. This could involve writing the data to a database or a file and streaming it real-time to a dashboard.

* Advantages of ETL

① Improved data quality :- By transforming and cleaning the data during ETL, organizations ensure the data is accurate and consistent across all systems.

② Simplified reporting :- ETL can help to simplify reporting by consolidating data from multiple sources into a single data warehouse or application.

* Disadvantages of ETL

① Complexity :- ETL processes can be complex, data may be lost due to incorrect mappings, transformations, and other errors.

② Cost - ETL processes are expensive to build and maintain.

* Data Extraction. (E)

→ Data Extraction is the first step in the ETL process. and it involves extracting data from one or more sources into a staging area where it can be cleaned and prepared for future steps

① Identification of Data sources and Types

→ The first step in data extraction is to identify the sources from which data is extracted and the type of data which is being extracted.

→ There are mainly 3 types of data.

a) Structured Data :- Data which is organized in the form of relational tables along with rows and columns and primary keys etc.

b) Semi-Structured Data :- Data which is not stored in relational model but has some organisational properties. eg. JSON Data.

c) Unstructured Data :- Data which cannot be analyzed as it has no organizational properties. Eg. Word doc, PDF etc.

→ Some common sources from which data is extracted is are :-

a) Databases
b) Flat Files
c) Web Services.
d) Message Queues.

* Types of Data Extraction.

① Immediate Data Extraction :- This type of data extraction involves extracting data as soon as it becomes available (when data is real-time)

→ This extraction also uses ~~data~~ real-time extraction techniques like change-data-capture (CDC) or data streaming

→ CDC (change Data capture) is a method used to capture changes to data in real-time from database, enabling applications to access the most recent data available.

→ Data Streaming is a method to continuously deliver data in real-time from source system to the target system, it invdves sending data in small, frequent and incremental updates.

→ Some common sources from which data is extracted is are :-

a) Databases
b) Flat Files
c) Web services.
d) Message Queues.

* Types of Data Extraction.

① Immediate Data Extraction :- This type of data extraction involves extracting data as soon as it becomes available (when data is real-time)

→ This extraction also uses ~~data~~ real-time extraction techniques like change-data-capture (CDC) or data streaming

→ CDC (change Data capture) is a method used to capture changes to data in real-time from database, enabling applications to access the most recent data available.

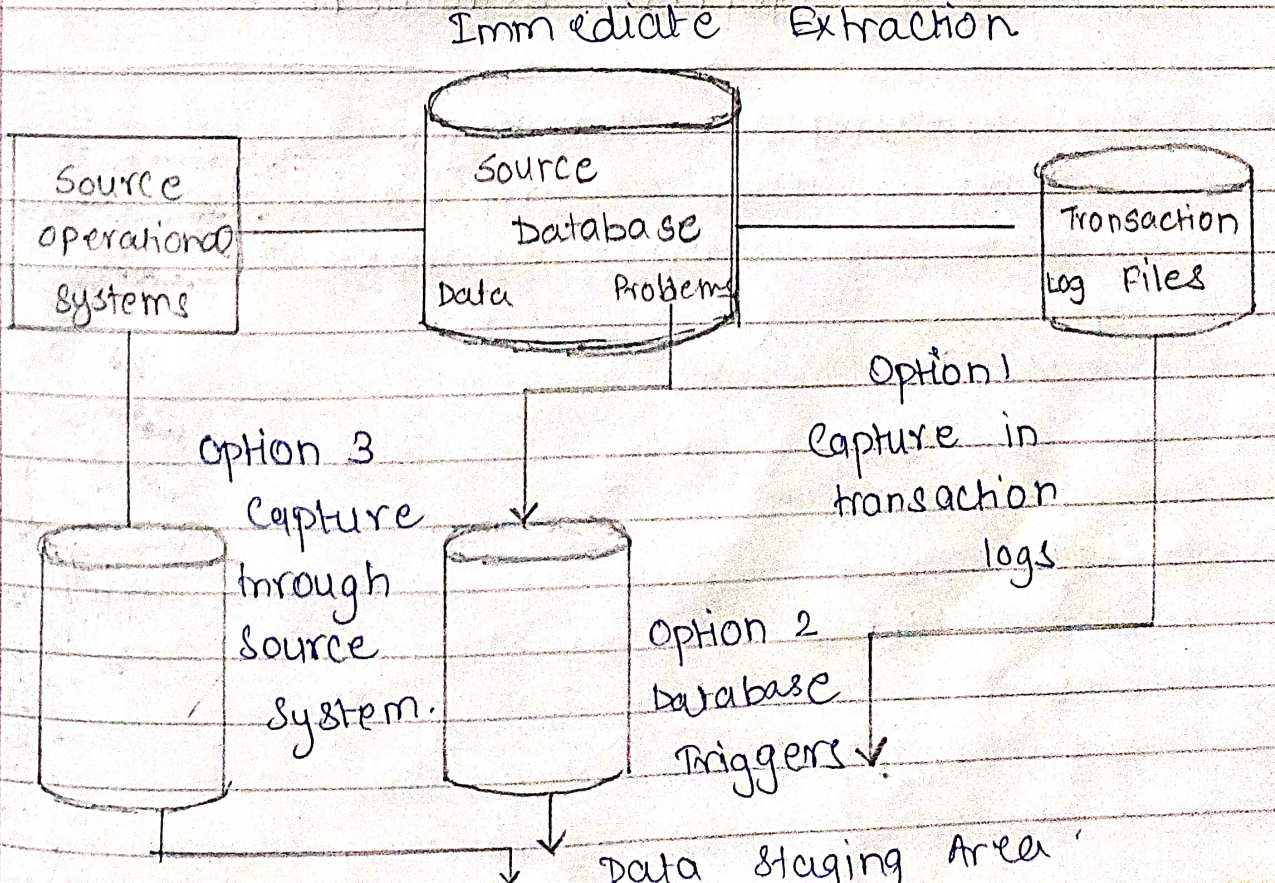
→ Data Streaming is a method to continuously deliver data in real-time from source system to the target system, it invdves sending data in small, frequent and incremental updates.

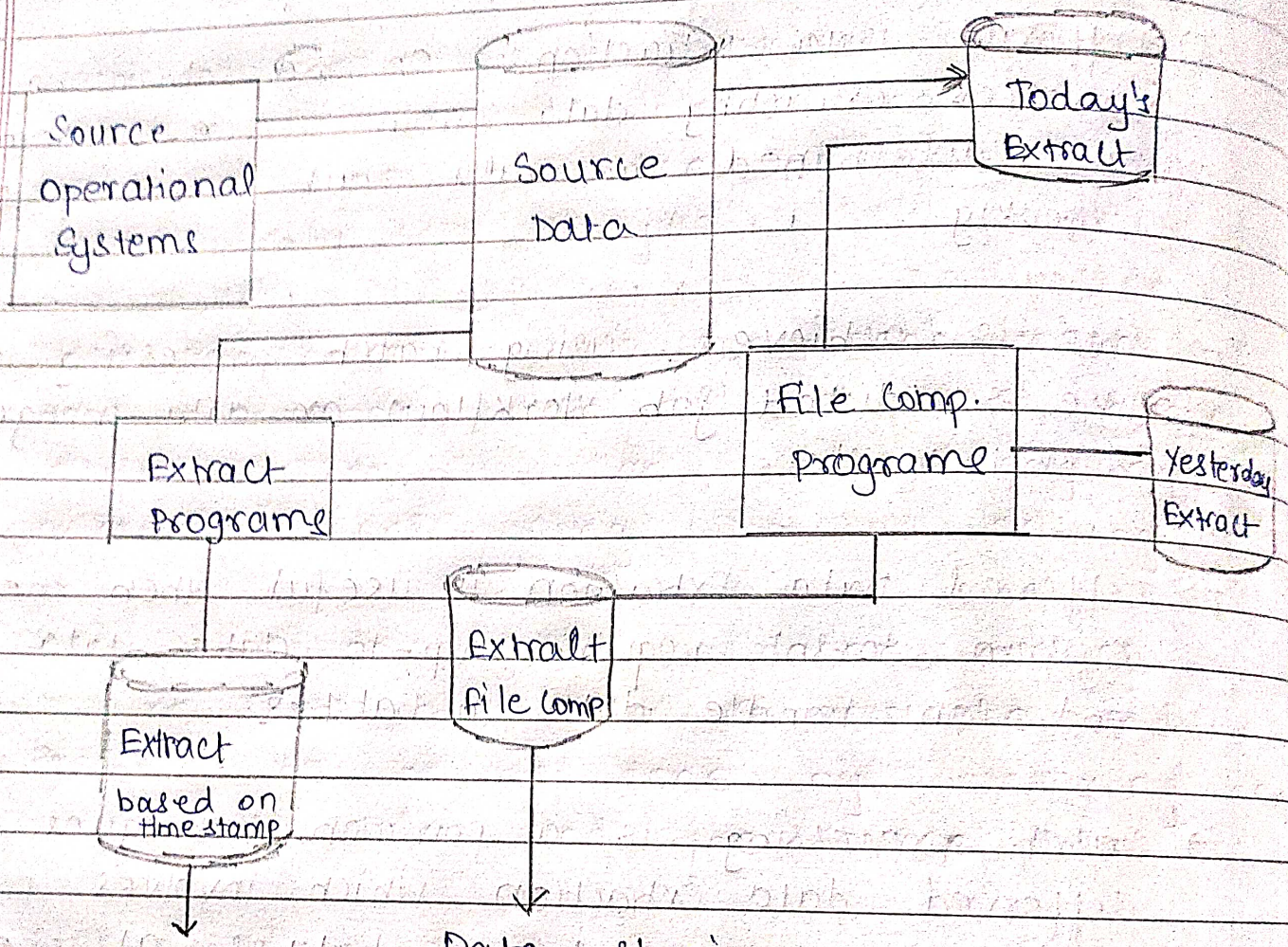
② Deffered Data Extraction :- on the other hand, involves extracting data from source system at pre-delermined intervals, such as daily or weekly.

→ This is achieved using batch processing techniques like scheduled job workflow or data integration workflow etc.

→ Deffered Data Extraction is useful when the target system doesnt require up-to-date data and can handle data in batches.

→ Batch processing is a common techniques used in deferred data extraction which involves processing large volumes of data in batches at pre-determined intervals, such as daily weekly or monthly.




Source operational systems
Source Data
Today's Extract
Extract Programe
File Comp. programe
Yesterday Extract
Extract file comp
Extract based on Time stamp
Data staging Area.
Deferred Extraction

* Data Transformation. (ETL)

→ Data transformation is the second step, following data extraction, where the raw data collected from the source system is transformed into the format which can be accessed by the target system.

→ After the data is transformed, it is loaded onto a data mart or a data warehouse.

* Tasks in Data Transformation.

① Data Mapping :- The first task in data transformation is to map the source data elements to the target data elements. This involves identify the source data fields and matching with the corresponding target fields.

② Data Cleansing :- Data cleansing is the process of identifying and correcting or removing errors, inconsistencies and inaccuracies in the data. Also involves tasks like removing duplicates correcting misspellings and typos and filling missing values.

③ Data filtering :- Data filtering involves selecting a subset of data from the source system that meets a specific criteria. The criterias can be date ranges, product types, or geographic regions etc.

④ Data aggregation :- This involves combining multiple data records into a single record. This is often done to reduce the size of dataset and simplify data analysis.

⑤ Data enrichment :- Data enrichment means adding new information which did not exist in the source system. This can be done by adding external sources along with the existing ones.

⑥ Data Validation :- Data validation involves ensuring that the data is accurate and complete. This can be done by comparing the transformed data and the source data.

---

* Data Loading (L)

→ Data Loading is the process of inserting or updating data from the transformed data set into the target database or data warehouse.

→ Steps in Data Loading

① Connection with target system :- The first step is to establish a connection with the target system. It involves setting up a connection string, specifying the database or data warehouse to be loaded and providing validation.

② Creation of the target schema :- Before loading data, a target schema has to be created so that data can be mapped into it.

This involves creating tables, views, indexes and other database objects.

③ Loading the data :- Once the target schema is set up, data is loaded onto the target system.

→ This can be done by various methods; bulk loading, incremental loading or real-time data streaming.

④ Validation of data :- After data is loaded, it must be ensured that it meets the quality and integrity of the target system.

→ This involves running data quality checks and performing data profiling.

⑤ Error Handling :- Errors which occur due to data inconsistencies, network failures or other issues.

→ This involves logging errors and ensuring timely resolutions.

* Techniques of Data Loading

① Bulk Loading :- It involves loading a large amount of data into the target system in a single operation. It is often used when the volume of data is too large to be processed.

→ Done by using tools as SQL Loader, BCP or bulk insert statements.

② Incremental Loading :- Incremental Loading is a technique which involves loading only the changed or updated data into the target system.

→ Often used when the data is changing frequently. Tools used are CDC, Timestamp-based data filtering.

③ Real-Time Data Streaming :- Real-time data streaming allows loading data as soon as it is generated in the source system.

→ Often used when target system requires real time updates. Tools used are Apache Nifi, AWS Kinesis etc.

* Fact Loading
* Fact Tables and Dimension Tables.

① Fact Table :- A fact table is the central table in a data warehouse or data mart that contains the quantitative or numerical measures of buisness process.

→ A fact table typically contains one or more measures or metrics, such as sales, revenue sold quantity, expenses etc.

② Dimensions Tables :- A Dimension Table is a table containing descriptive attributes that define dimensions of a business process. These attributes provide the context for the quantitative or numerical measurement stored in the Fact Table.

→ Typically contains attributes such as time, products, geography, customers, etc.

* Loading data ~~onto~~ of Fact and Dimension Tables

→ once the data is transformed and cleansed it is loaded onto the dimension tables.

→ This involves mapping dimensional attributes to columns in the target database into the appropriate dimension table.

→ After the Dimension Table is loaded, the data is loaded into the Facts table.

→ This involves loading the numeric values of the corresponding Dimensions into the target database.

→ Lastly, tables can be indexed to improve the performance of queries and to optimize the storage of data.

---

* Data Quality

→ Data Quality in ETL refers to the accuracy, completeness, consistency, relevance and timeliness of data being extracted.

* Issues in Data Cleansing

① Incomplete or missing data

② Inaccurate data

③ Duplicate data

④ outliers

⑤ Integration Issues.