ETL / ELT / 3 TIER + DATA LOADING

SUBMITTED BY: MUHAMMAD FAHAD

TASKS DAY 3 [16/03/2023]

Task # 4:

What is ETL? in detail.

What is ELT? in detail

3 Tier Architecture in DE

ETL Tools (any 3)

Task # 5:

What is Historical Load

What is Full Load

What is Incremental Load

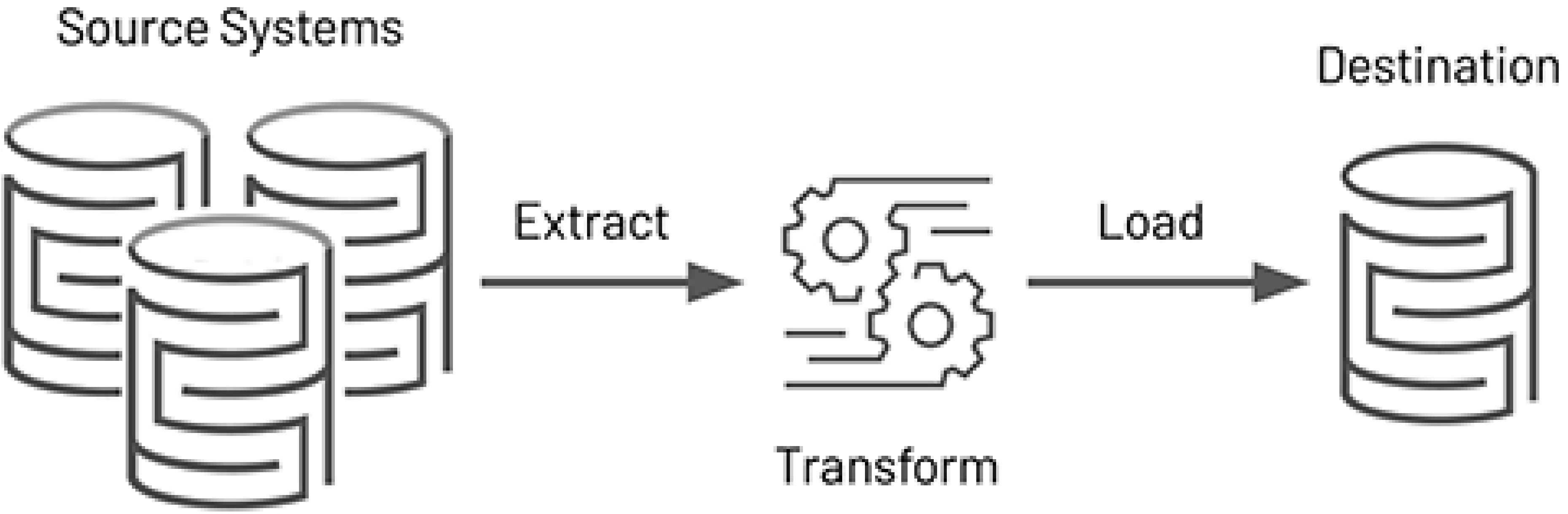
TASK # 4
WHAT IS ETL?

ETL stands for EXTRACT, TRANSFORM and LOAD.

- Extract data from it's original source
- Transform data by deduplicating it, combining it and ensuring quality
- Load data into the target database

ETL is a traditional approach that involves extracting data from various sources, transforming it into a consistent format and then loading it into the target data warehouse/data mart. ETL is best suited for batch processing and large volumes of data and is often used in situations where data needs to be cleaned, normalized and aggregated before it's loaded into target data warehouse or data mart.

ETL Process



STEPS OF ETL
STEP 1: EXTRACTION

This is the first step of ETL, in this process the data is extracted from the target sources that are usually heterogenous such as business systems, APIs. Sensor data, marketing tools and transaction database or other sources. As you can see, some of the data types are likely to be structured outputs of widely used systems while other are semi structured JSON server logs. Moving further there are different methods to perform the extraction.

1. Partial Extraction: This method involves obtaining data when the source system notifies you of any record changes.

2. Partial Extraction with Update Notification: This method involves obtaining data when the source system points to the changed records and provides an extract of those records.

3. Full Extract: This method involves obtaining all data from the source system when there is no way to identify which data has been changed. To use this method, a copy of the previous extract in the same format is required to identify the changes that have been made.

STEPS OF ETL
STEP 2: TRANSFORM

The data extracted from the source server is typically raw and not useful in its original form. Therefore, it must be processed through a series of steps to cleanse, map, and transform it to make it useful for generating BI reports. This transformation step is a crucial part of the ETL process that enhances the value of the data. Some data requires no transformation and can be directly moved to the target system. However, for most data, a set of functions must be applied during transformation to convert it into a usable form.

These functions may include customized calculations or the use of formulas like SUM. For instance, if you need to calculate the total sales revenue, you can use the SUM formula during transformation to obtain the desired result.

Similarly, if the first and last names of customers are stored in separate columns, you can concatenate them before loading to create a single name field.

STEPS OF ETL
TRANSFORMATION TYPES

Basic Transformations:

- Cleaning: Mapping NULL to 0 or "Male" to "M" and "Female" to "F," date format consistency, etc.
- Deduplication: Identifying and removing duplicate records
- Format revision: Character set conversion, unit of measurement conversion, date/time conversion, etc.
- Key restructuring: Establishing key relationships across tables.

Advanced Transformations:

- Derivation: Applying business rules to your data that derive new calculated values from existing data – for example, deriving age from date of birth.
- Filtering: Selecting only certain rows and/or columns
- Joining: Linking data from multiple sources
- Splitting: Splitting a single column into multiple columns
- Aggregation: Data elements are aggregated from multiple data sources and databases. (for example, total sales for each store, and for each region.)
- Integration: Give each unique data element one standard name with one standard definition. Data integration reconciles different data names and values for the same data element.

STEPS OF ETL
STEP 3: LOAD

Loading data into the target Datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period. Hence, load process should be optimized for performance. In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.

Types of Loading:

- Initial Load: Populating all the Data Warehouse tables
- Incremental Load: Applying ongoing changes as when needed periodically.
- Full Refresh: Erasing the contents of one or more tables and reloading with fresh data.

WHAT IS ELT?

ELT is an acronym for Extract, Load, and Transform. ELT is a modern variation on the older process of extract, transform, and load (ETL), in which transformations take place before the data is loaded. It's a process that extracts raw data from a source system to a target system, and the information is then transformed into the source or destination system for downstream applications. Unlike ETL, where data transformation processes occur on a staging area before being loaded into the target system, in ELT, data is loaded directly into the target system and converted there. In this way, ELT is most useful for handling enormous datasets and using them for business intelligence and data analytics.

ELT PROCESS

1. Extract: This step works similarly in both ETL and ELT data management approaches. Raw streams of data from virtual infrastructure, software, and applications are ingested either in their entirety or according to predefined rules.

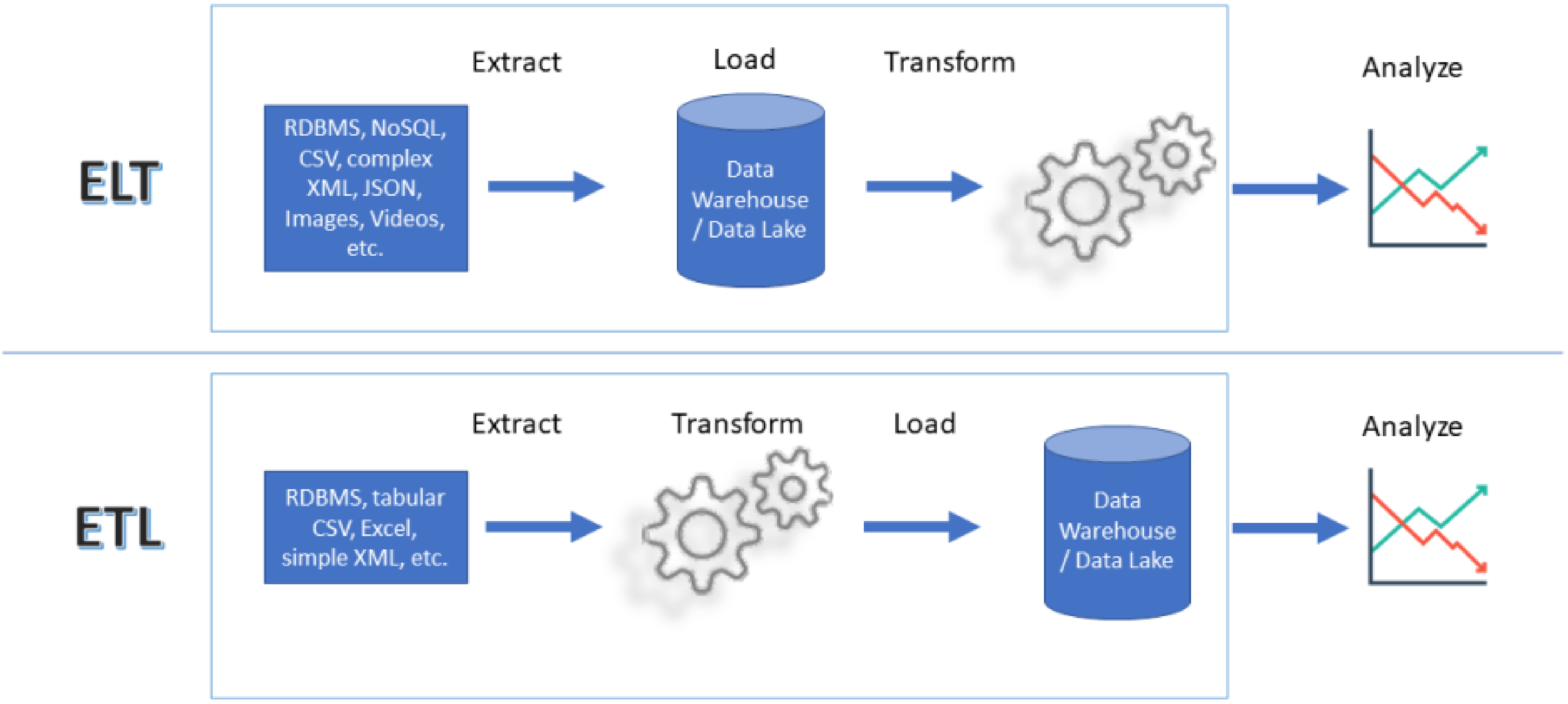
2. Load: Here is where ELT branches off from its ETL cousin. Rather than deliver this mass of raw data and load it to an interim processing server for transformation, ELT delivers it directly to the target storage location. This shortens the cycle between extraction and delivery.

3. Transform: The database or data warehouse sorts and normalizes the data, keeping part or all of it on hand and accessible for customized reporting. The overhead for storing this much data is higher, but it offers more opportunities to mine it for relevant business intelligence in near real-time.

COMPARISON BETWEEN ETL AND ELT

ETL	ELT
Extract, Transform and Load	Extract, Load and Transform
Data is extracted from source systems, transformed to the desired format, and then loaded into the target system.	Data is extracted from source systems and loaded into the target system. Transformation is applied on the data after it has been loaded into the target system.
Suitable for large data volumes	Suitable for small to medium data volumes
Typically used in traditional data warehousing	Typically used in modern data warehousing
Data is transformed before loading, making analysis faster and more efficient	Data is transformed after loading, so analysis may take longer
Requires more storage as data is transformed before loading	Requires less storage as data is transformed after loading

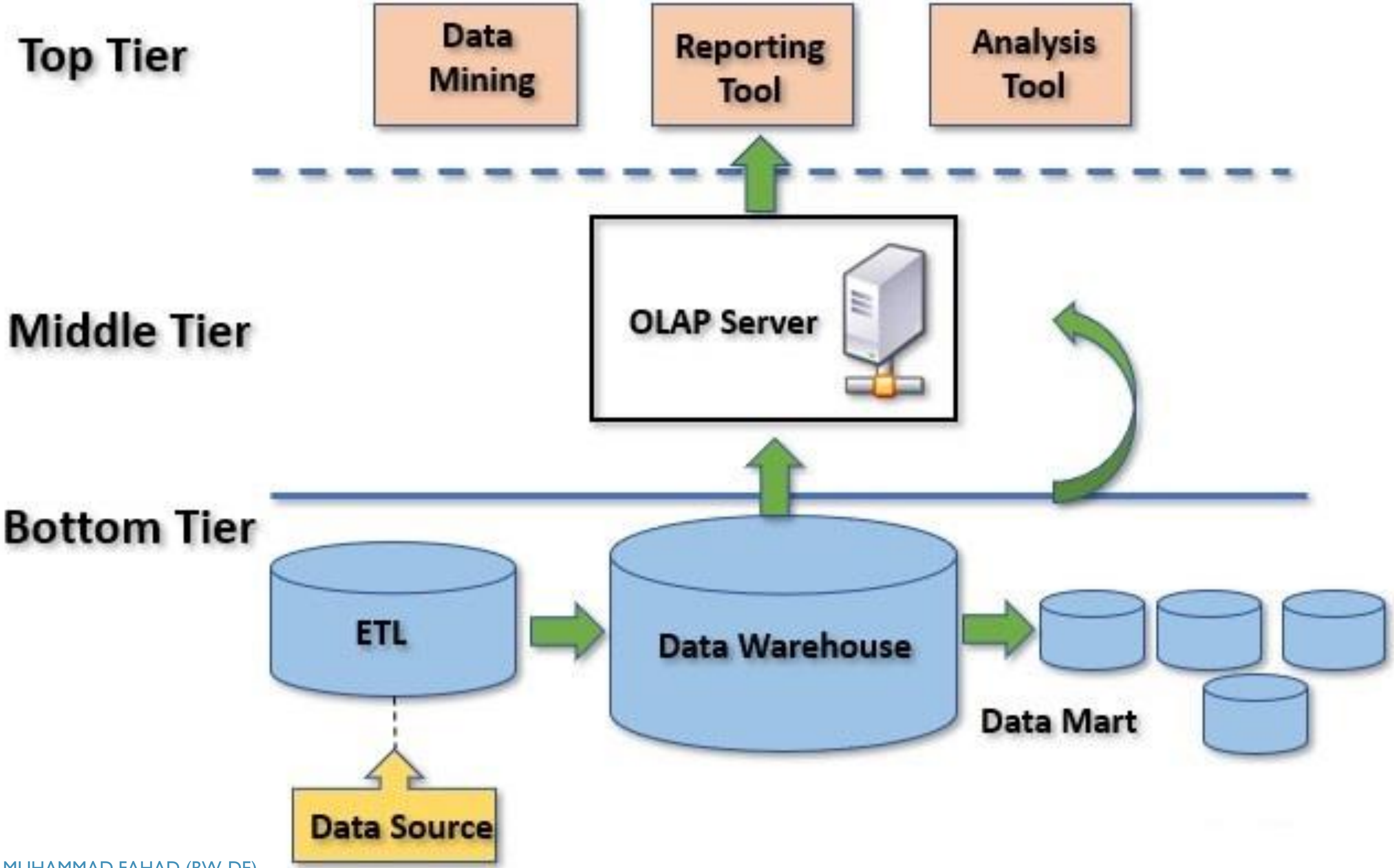
ELT vs ETL


ELT
Extract
Load
Transform
Analyze
RDBMS, NoSQL, CSV, complex XML, JSON, Images, Videos, etc.
Data Warehouse / Data Lake
ETL
Extract
Transform
Load
Analyze
RDBMS, tabular CSV, Excel, simple XML, etc.
Data Warehouse / Data Lake

3 TIER ARCHITECTURE IN DATA ENGINEERING

Three-tier architecture: This is the most widely used architecture. It consists of the Top, Middle and Bottom Tier.

- Presentation Tier (Top): The presentation layer is the top layer of the architecture, also known as the user interface layer. This layer is responsible for presenting the data to end-users, typically through a web or mobile application. The presentation layer interacts with the middle tier to retrieve the data needed to display the user interface.
- Application Tier (Middle): The middle tier in Data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database
- Data Tier (Bottom): The data layer is the bottom layer and is responsible for storing and managing the data. This layer contains the databases, data warehouses, or other data storage solutions used to store the data. The data layer is responsible for retrieving the data requested by the application layer and returning it for processing.


Top Tier
Data Mining
Reporting Tool
Analysis Tool
Middle Tier
OLAP Server
Bottom Tier
ETL
Data Warehouse
Data Mart
Data Source

ETL TOOLS

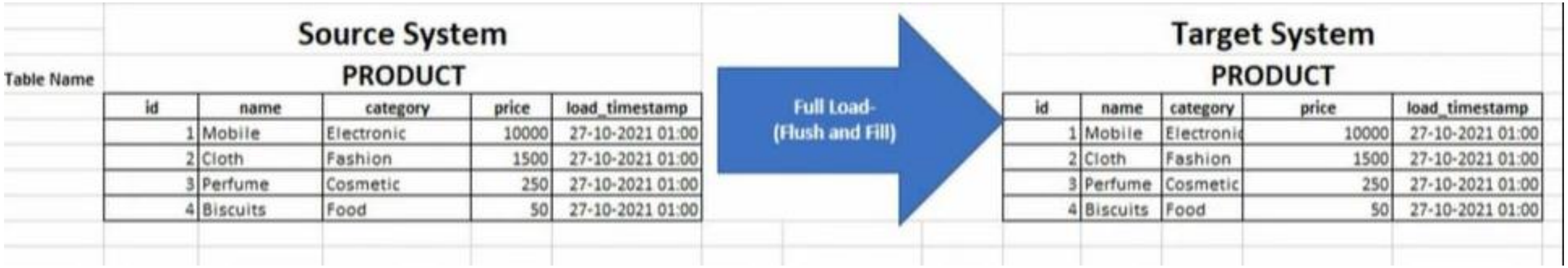
ETL (Extract, Transform, Load) tools are software applications that automate the process of extracting data from various sources, transforming the data to fit specific requirements, and loading it into a target system such as a data warehouse or database.

Some popular ETL tools are:

- Informatica PowerCenter
- Microsoft SQL Server Integration Services
- IBM Infosphere DataStage

TASK # 5
WHAT IS FULL LOAD?

This involves loading all the data from the source system into the target system, regardless of whether it has changed or not. This is typically used when the target system needs to be completely refreshed, and all the data gets updated. This can be time-consuming and resource-intensive, especially if the data volume is large.



Source System

Table Name PRODUCT

id	name	category	price	load_timestamp
1	Mobile	Electronic	10000	27-10-2021 01:00
2	Cloth	Fashion	1500	27-10-2021 01:00
3	Perfume	Cosmetic	250	27-10-2021 01:00
4	Biscuits	Food	50	27-10-2021 01:00

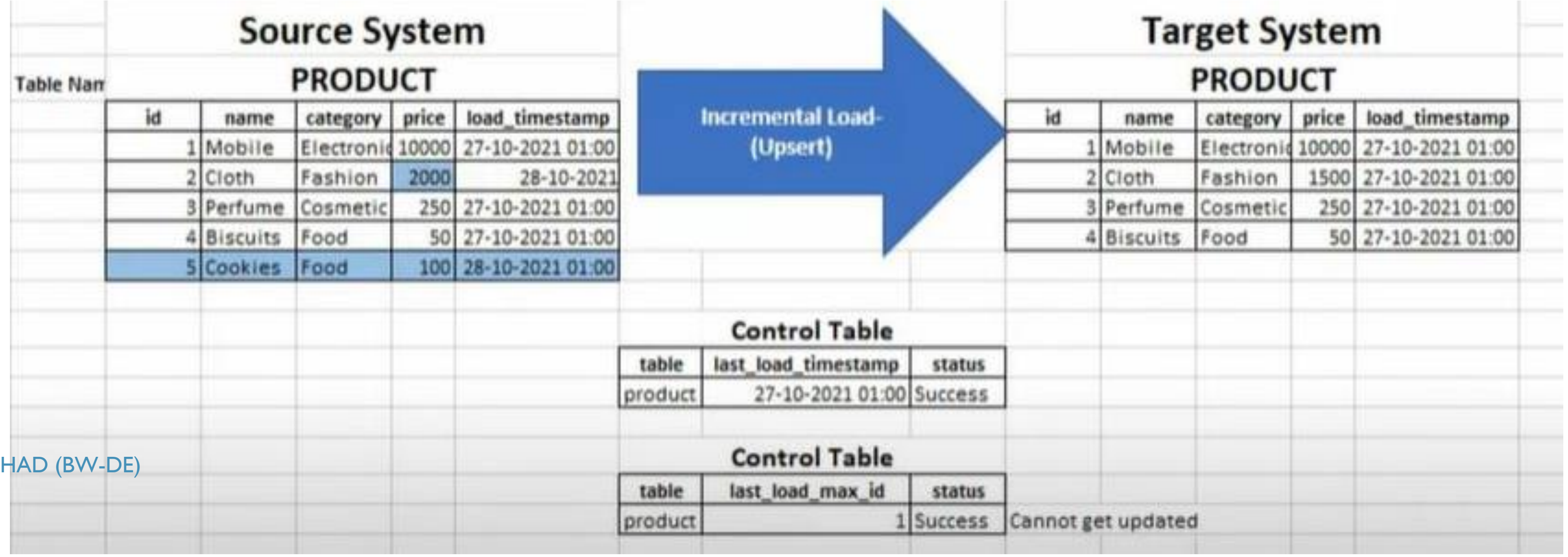
Target System

PRODUCT

id	name	category	price	load_timestamp
1	Mobile	Electronic	10000	27-10-2021 01:00
2	Cloth	Fashion	1500	27-10-2021 01:00
3	Perfume	Cosmetic	250	27-10-2021 01:00
4	Biscuits	Food	50	27-10-2021 01:00

WHAT IS INCREMENTAL LOAD?

This involves loading only the data that has changed since the last load into the target system. This is typically used when the target system needs to be kept up-to-date and only the changes in the source system need to be reflected in the target system. Incremental loads are more efficient than full loads, as they reduce the amount of data that needs to be processed and loaded.



Source System

Table Nan

PRODUCT

id	name	category	price	load_timestamp
1	Mobile	Electronic	10000	27-10-2021 01:00
2	Cloth	Fashion	2000	28-10-2021
3	Perfume	Cosmetic	250	27-10-2021 01:00
4	Biscuits	Food	50	27-10-2021 01:00
5	Cookies	Food	100	28-10-2021 01:00

Target System

PRODUCT

id	name	category	price	load_timestamp
1	Mobile	Electronic	10000	27-10-2021 01:00
2	Cloth	Fashion	1500	27-10-2021 01:00
3	Perfume	Cosmetic	250	27-10-2021 01:00
4	Biscuits	Food	50	27-10-2021 01:00

Control Table

table	last_load_timestamp	status
product	27-10-2021 01:00	Success

Control Table

table	last_load_max_id	status
product	1	Success

Cannot get updated

WHAT IS HISTORICAL LOAD?

Historic load involves loading all the historical data available from the source system into the target system. This is typically used when building a data warehouse for the first time or when starting a new project. The historical load can be a large amount of data and it may take longer to load and transform the data.