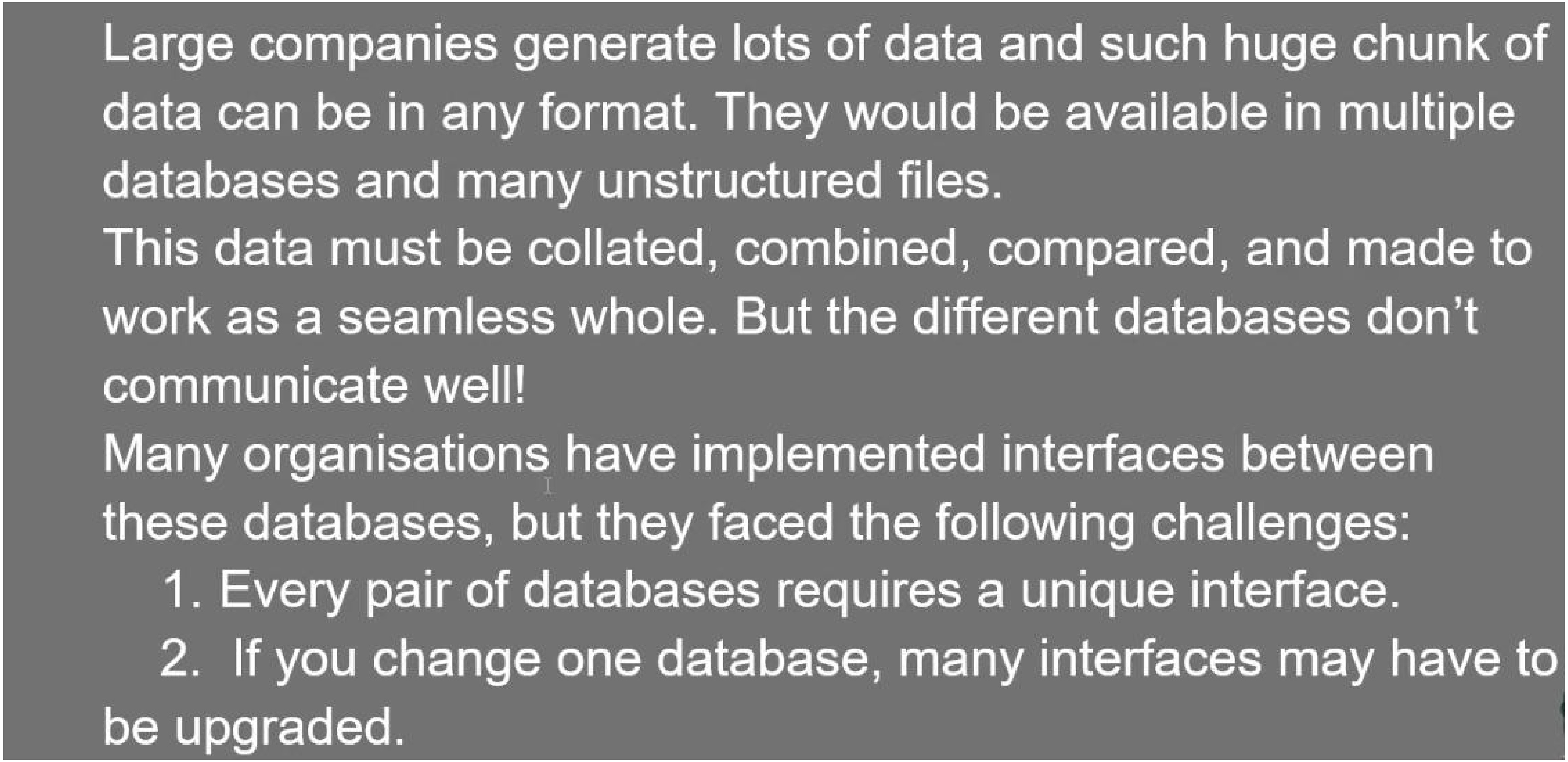
Why Do We Need ETL?

Every company these days have to process large sets of data from varied sources. This data needs to be processed to give insightful information for making business decisions. But, quite often such data have following challenges:

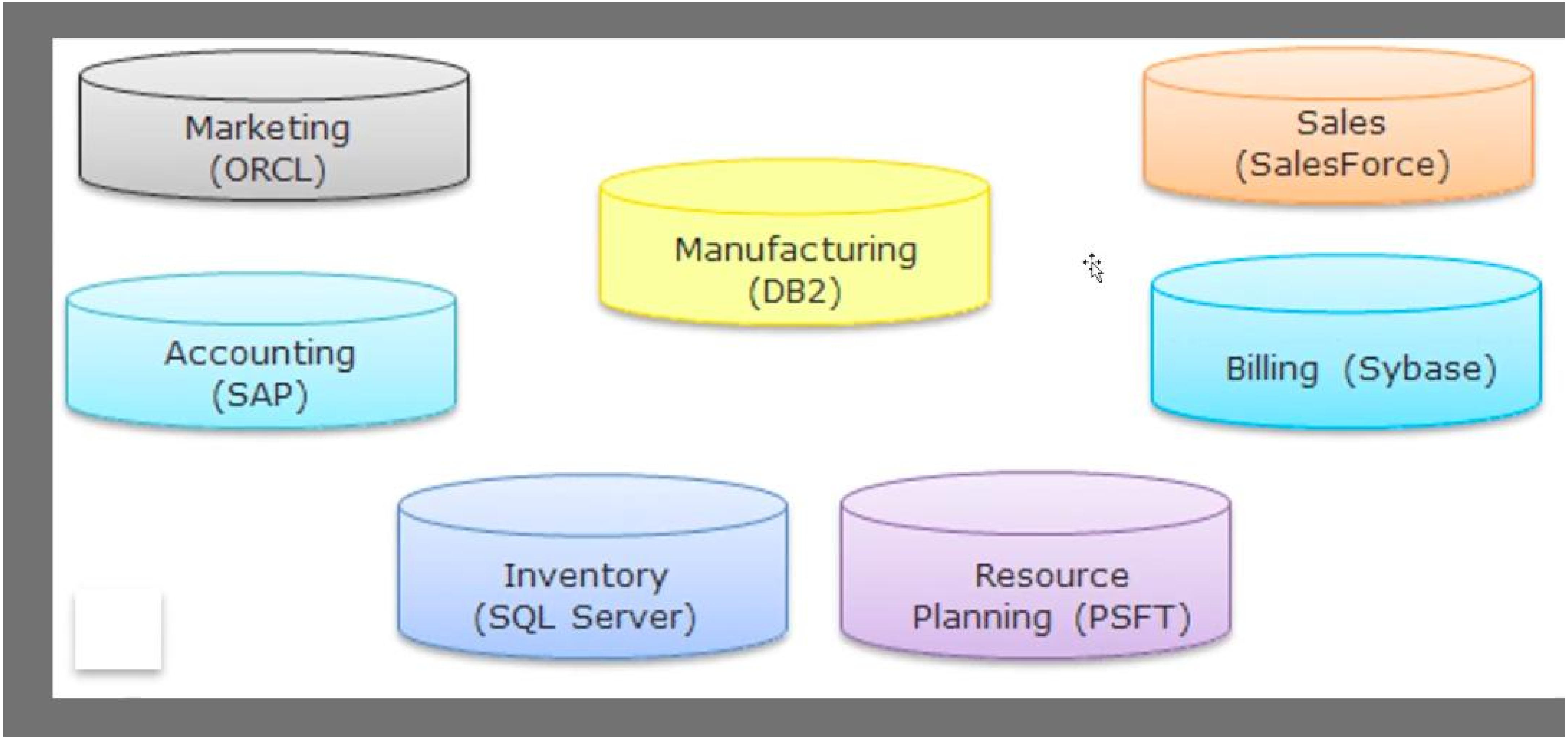


Large companies generate lots of data and such huge chunk of data can be in any format. They would be available in multiple databases and many unstructured files.

This data must be collated, combined, compared, and made to work as a seamless whole. But the different databases don’t communicate well!

Many organisations have implemented interfaces between these databases, but they faced the following challenges:

1. Every pair of databases requires a unique interface.
2. If you change one database, many interfaces may have to be upgraded.



Marketing
(ORCL)

Manufacturing
(DB2)

Sales
(SalesForce)

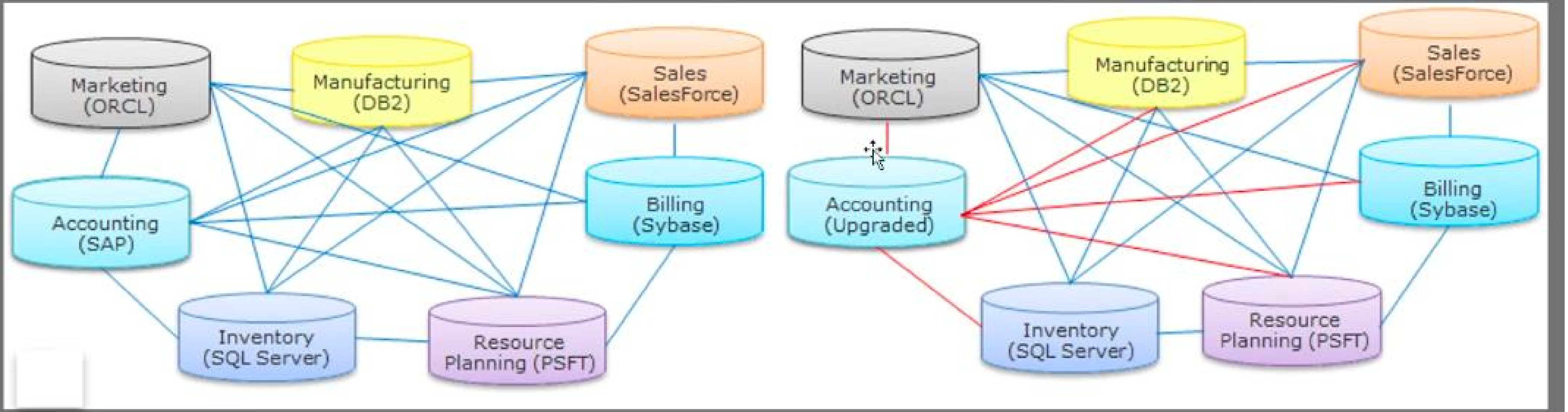
Accounting
(SAP)

Billing (Sybase)

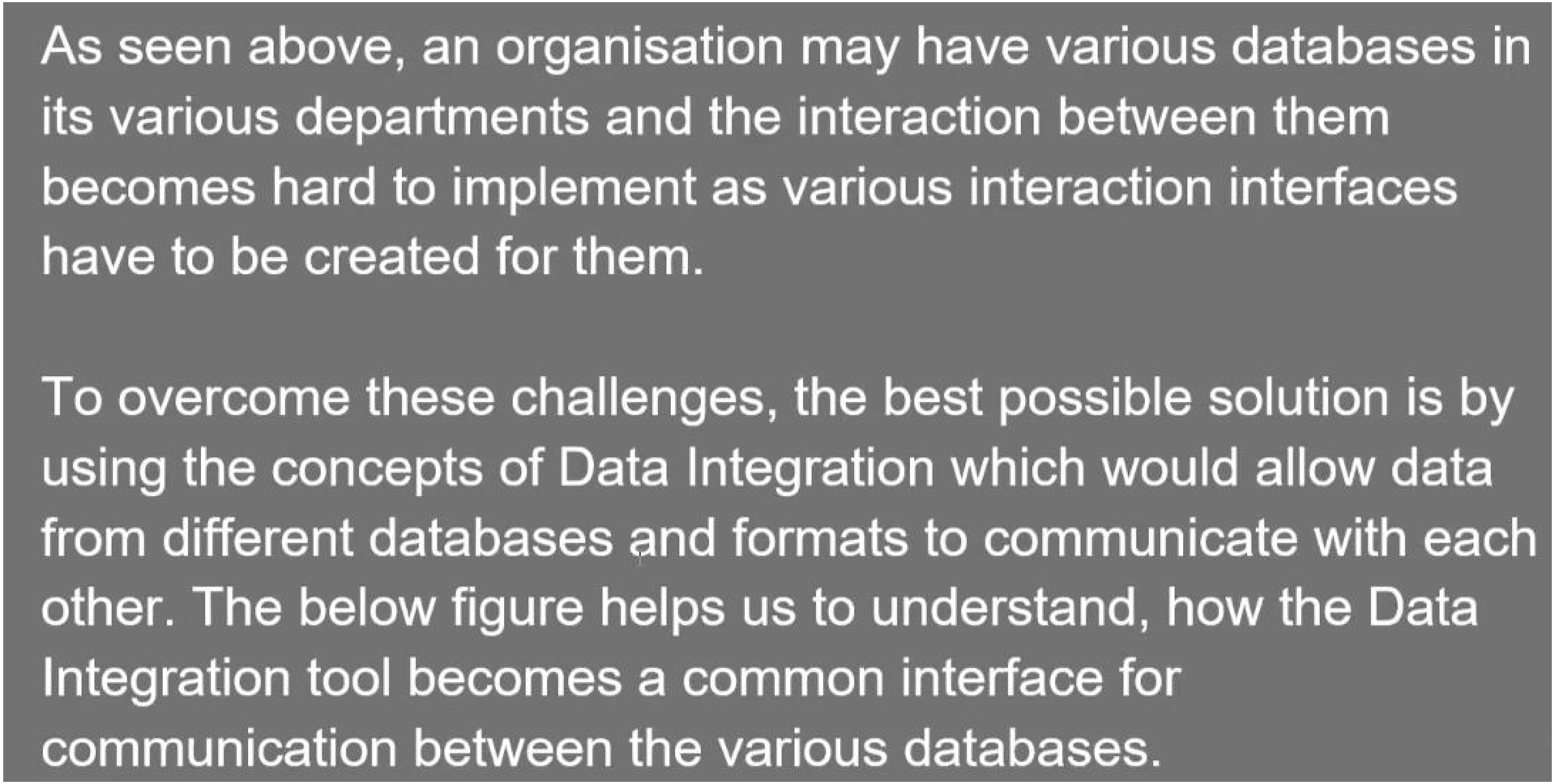
Inventory
(SQL Server)

Resource
Planning (PSFT)

Various Databases used by different departments of an organization

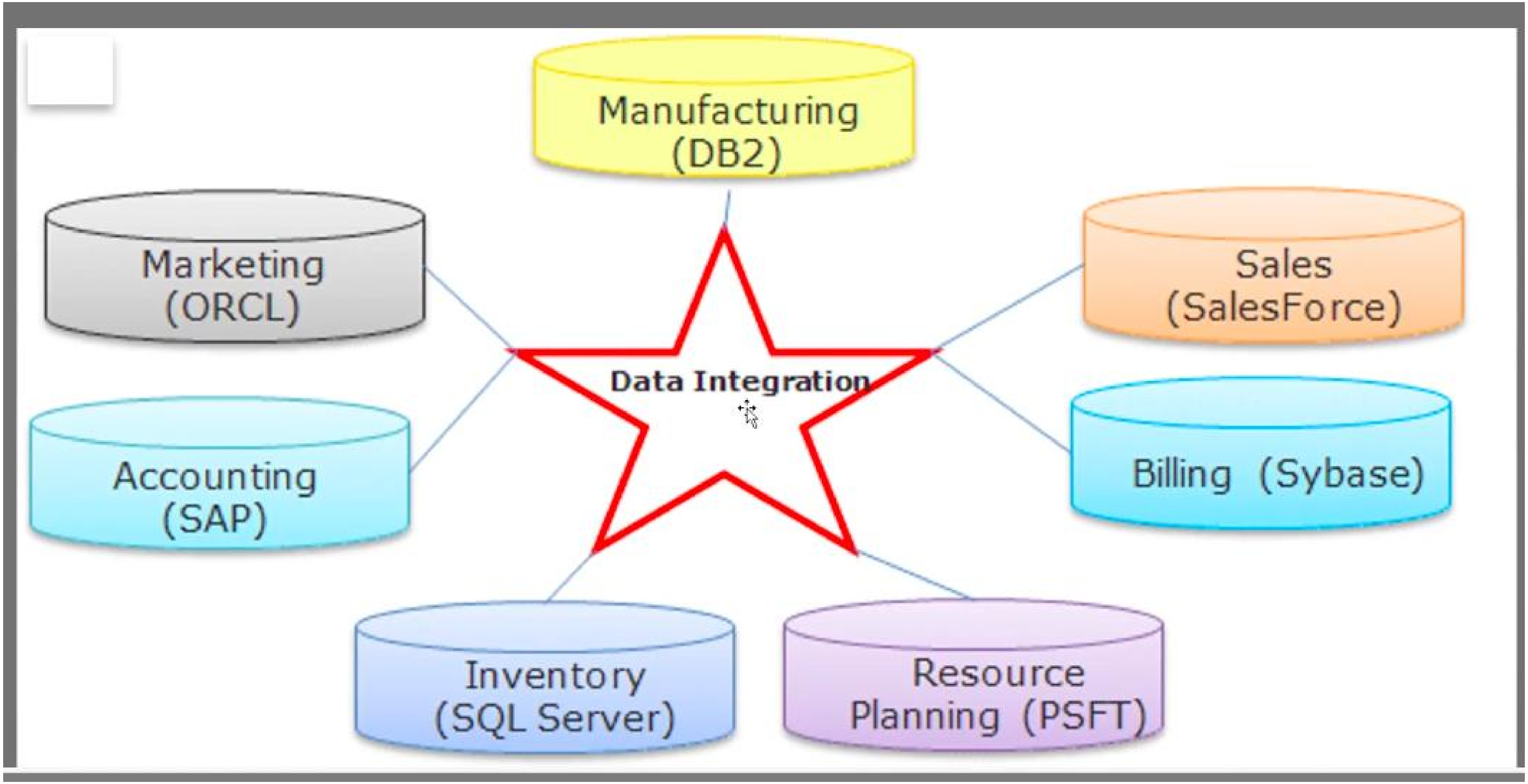


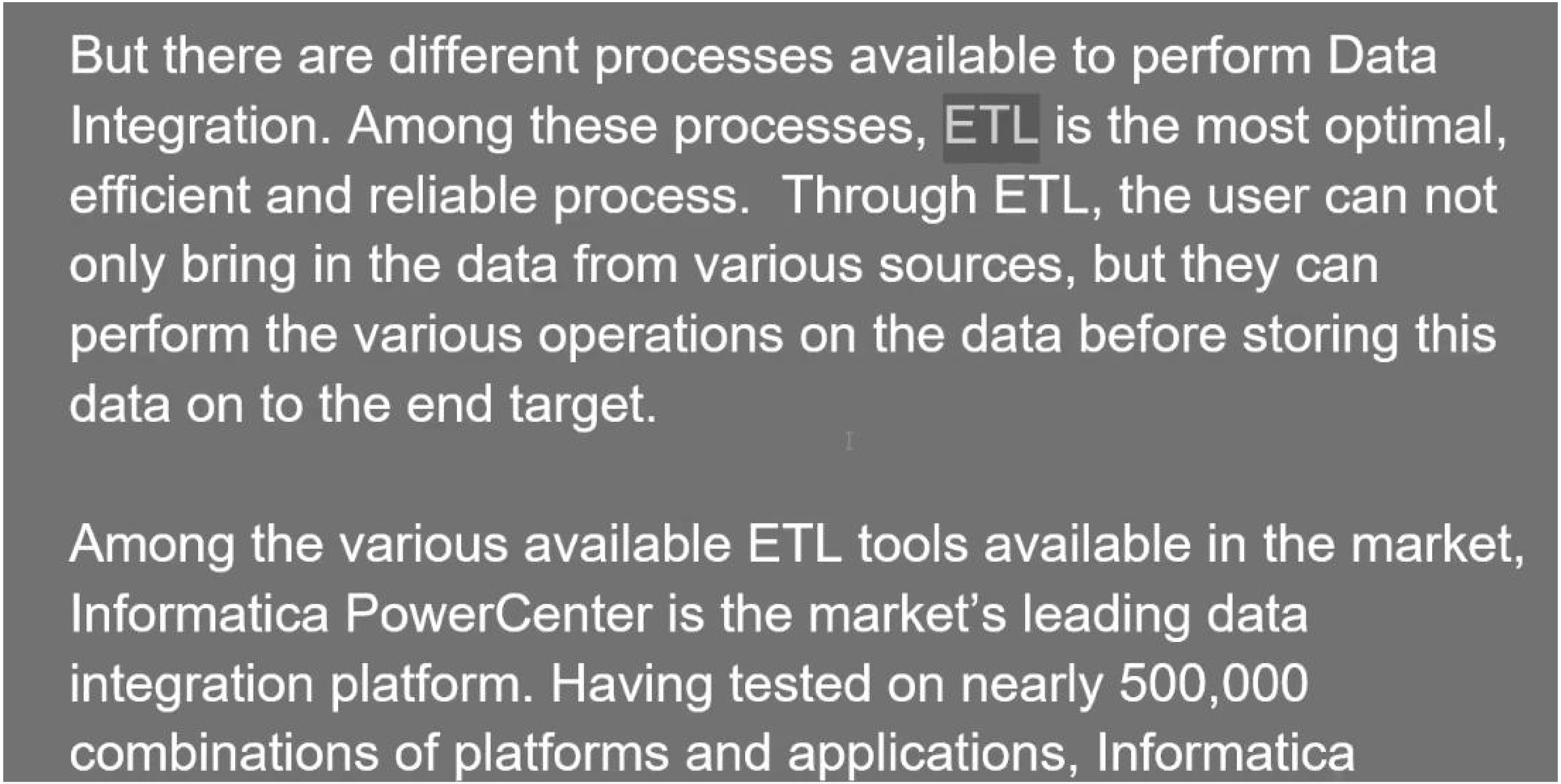
As seen above, an organisation may have various databases in its various departments and the interaction between them



As seen above, an organisation may have various databases in its various departments and the interaction between them becomes hard to implement as various interaction interfaces have to be created for them.

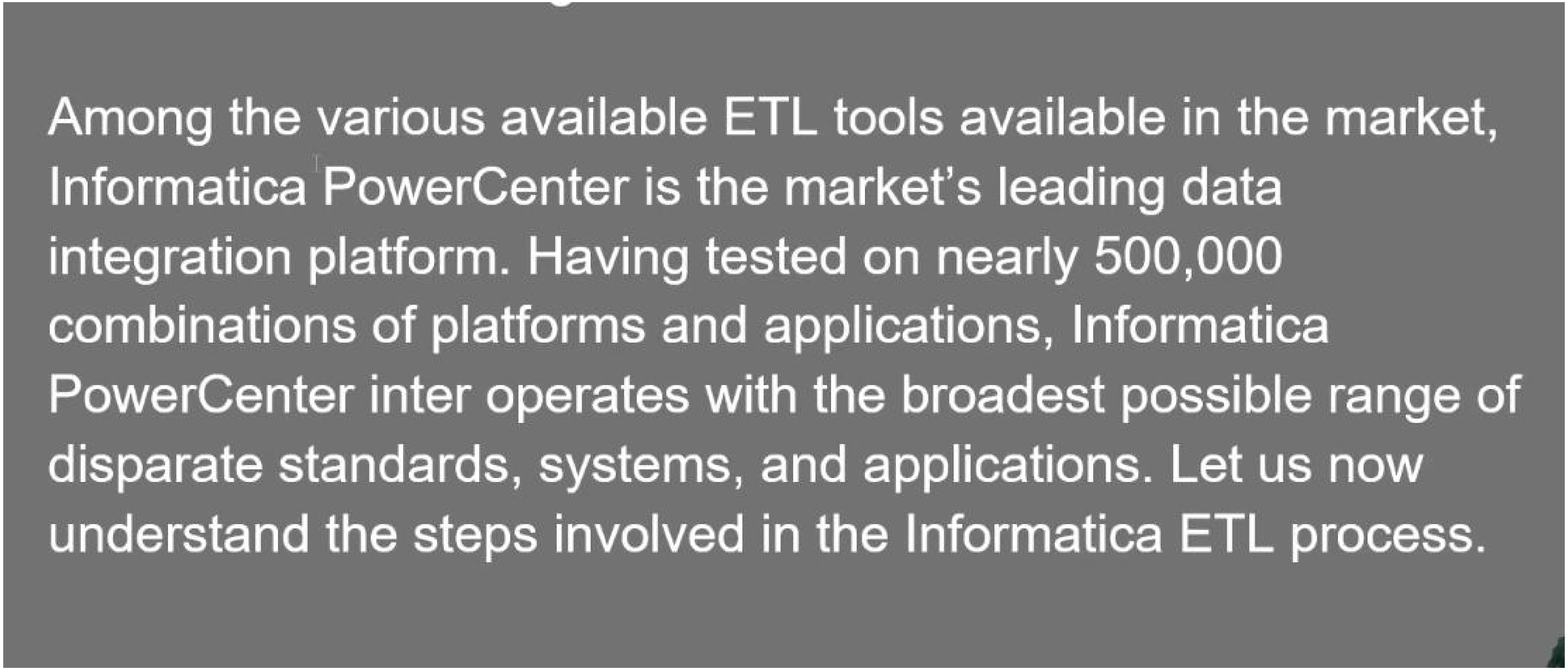
To overcome these challenges, the best possible solution is by using the concepts of Data Integration which would allow data from different databases and formats to communicate with each other. The below figure helps us to understand, how the Data Integration tool becomes a common interface for communication between the various databases.


Manufacturing (DB2)
Marketing (ORCL)
Sales (SalesForce)
Data Integration
Accounting (SAP)
Billing (Sybase)
Inventory (SQL Server)
Resource Planning (PSFT)

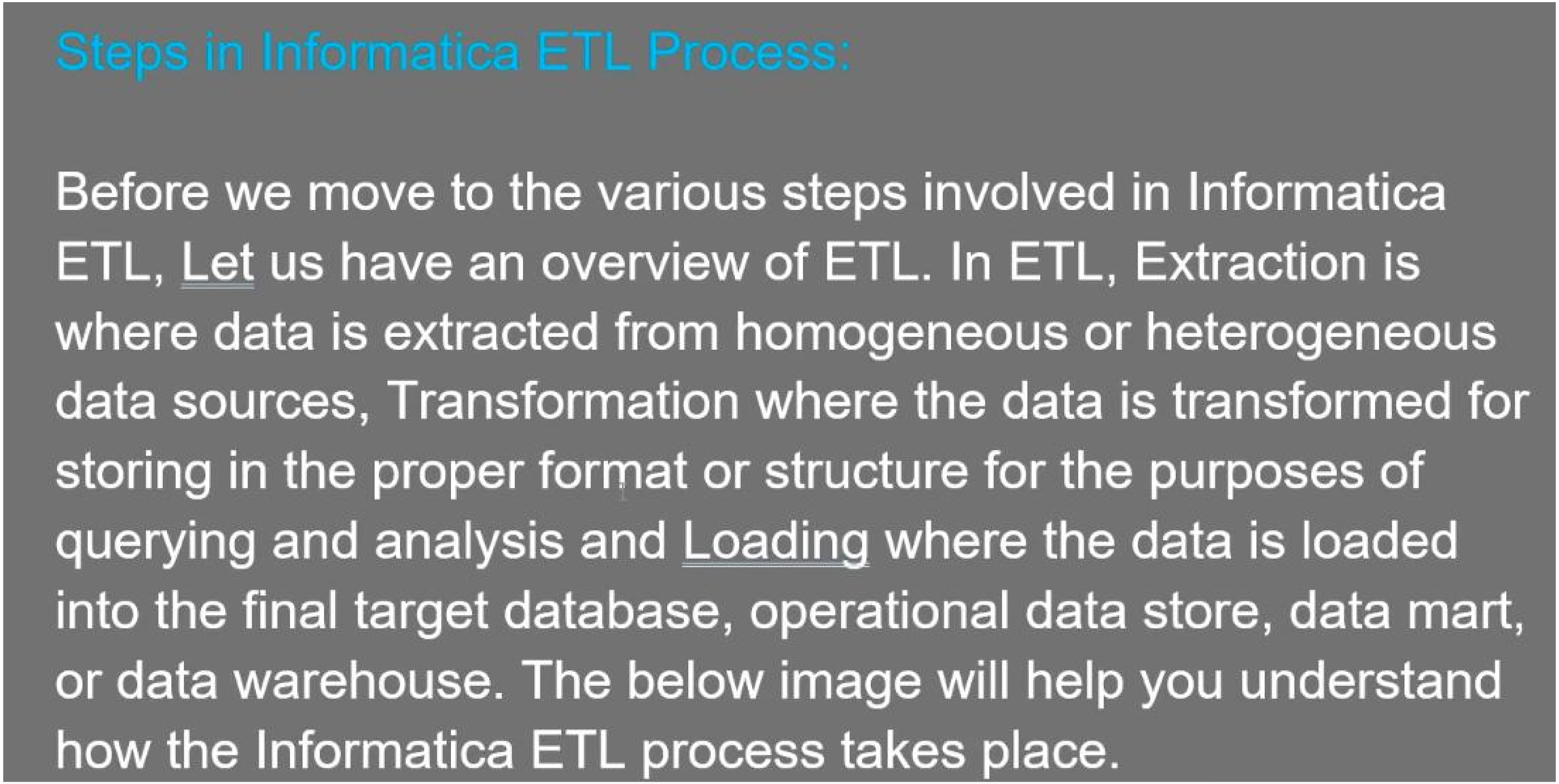


But there are different processes available to perform Data Integration. Among these processes, ETL is the most optimal, efficient and reliable process. Through ETL, the user can not only bring in the data from various sources, but they can perform the various operations on the data before storing this data on to the end target.

Among the various available ETL tools available in the market, Informatica PowerCenter is the market’s leading data integration platform. Having tested on nearly 500,000 combinations of platforms and applications, Informatica

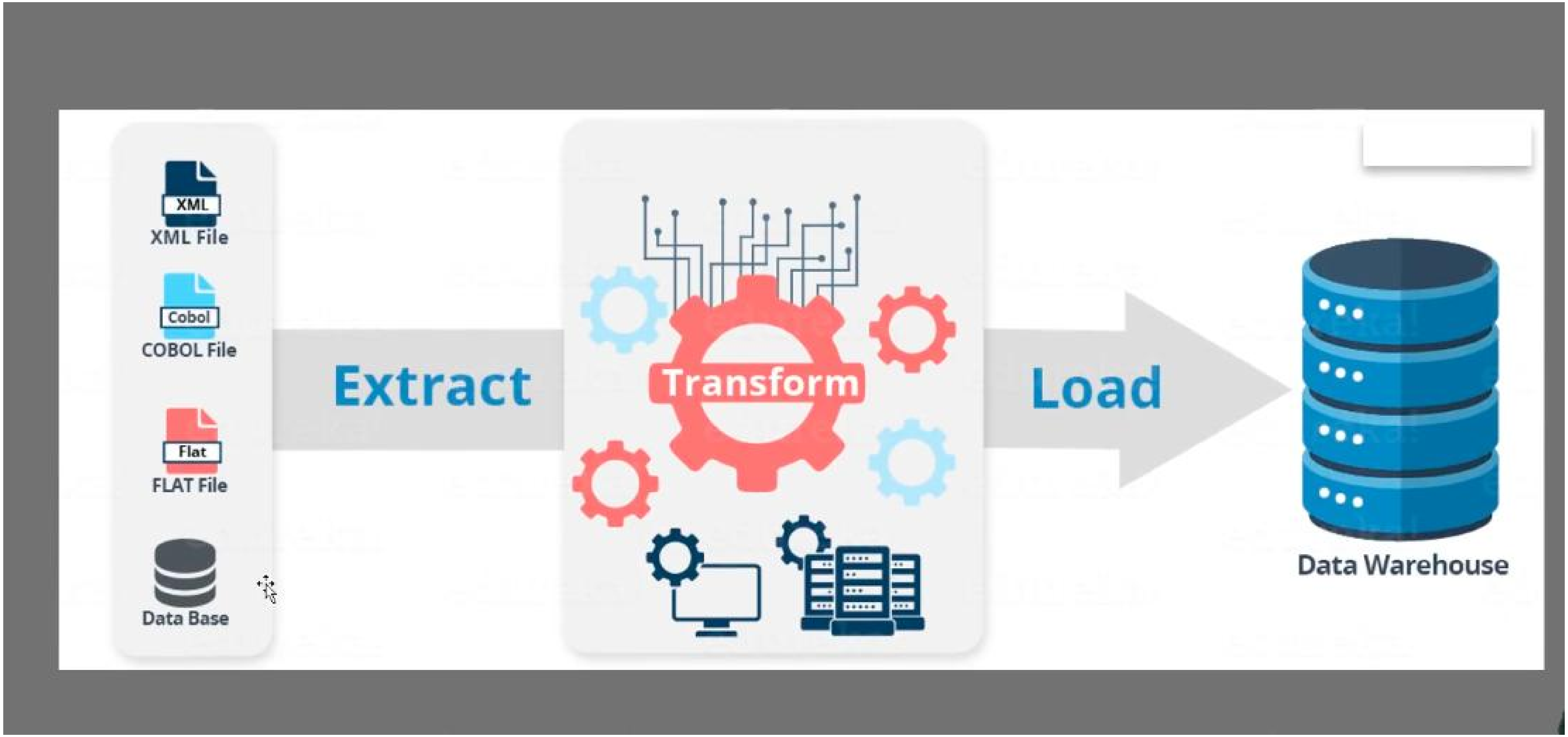


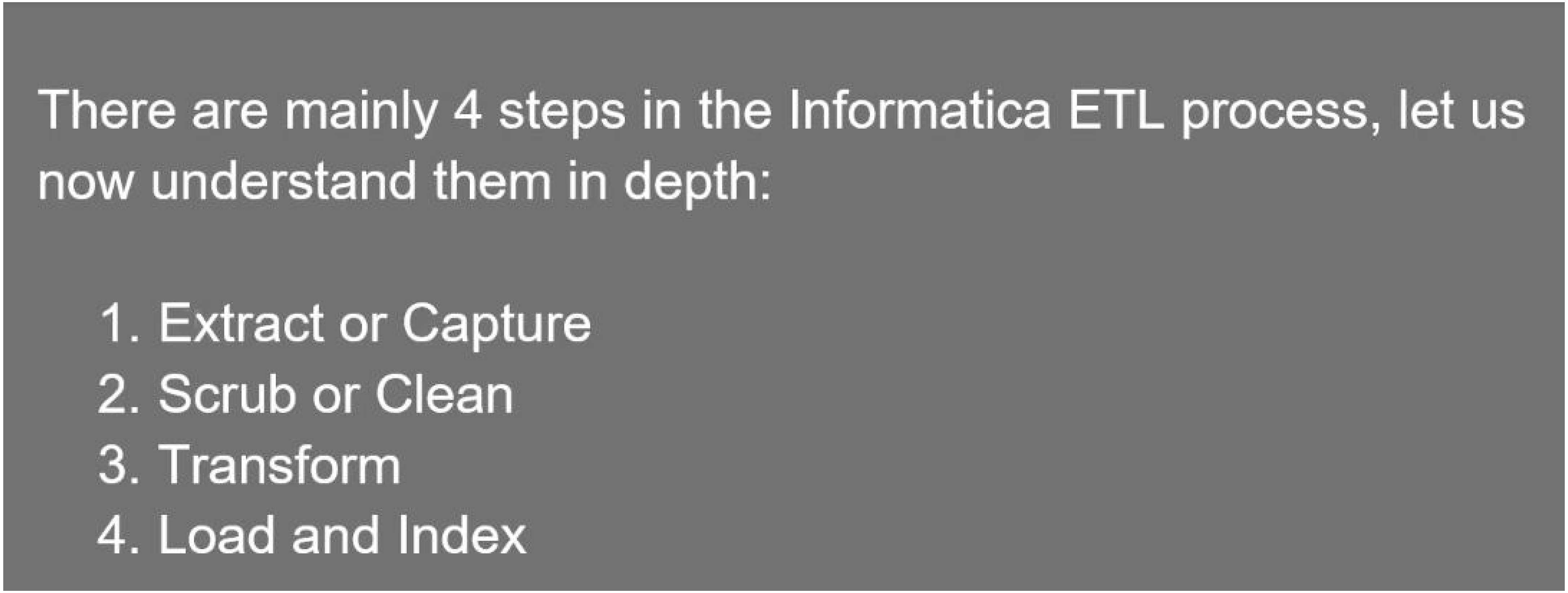
Among the various available ETL tools available in the market, Informatica PowerCenter is the market’s leading data integration platform. Having tested on nearly 500,000 combinations of platforms and applications, Informatica PowerCenter inter operates with the broadest possible range of disparate standards, systems, and applications. Let us now understand the steps involved in the Informatica ETL process.



Steps in Informatica ETL Process:

Before we move to the various steps involved in Informatica ETL, Let us have an overview of ETL. In ETL, Extraction is where data is extracted from homogeneous or heterogeneous data sources, Transformation where the data is transformed for storing in the proper format or structure for the purposes of querying and analysis and Loading where the data is loaded into the final target database, operational data store, data mart, or data warehouse. The below image will help you understand how the Informatica ETL process takes place.

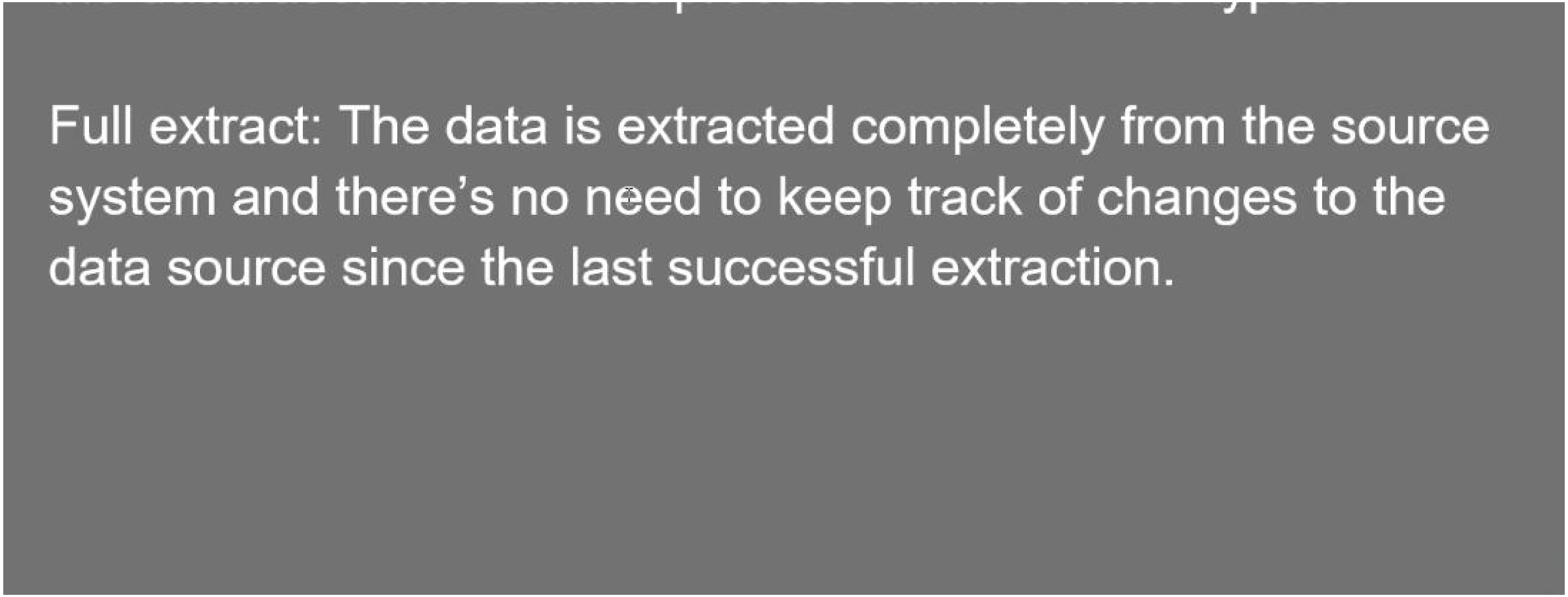

XML
XML File
Cobol
COBOL File
Flat
FLAT File
Data Base
Extract
Transform
Load
Data Warehouse



There are mainly 4 steps in the Informatica ETL process, let us now understand them in depth:

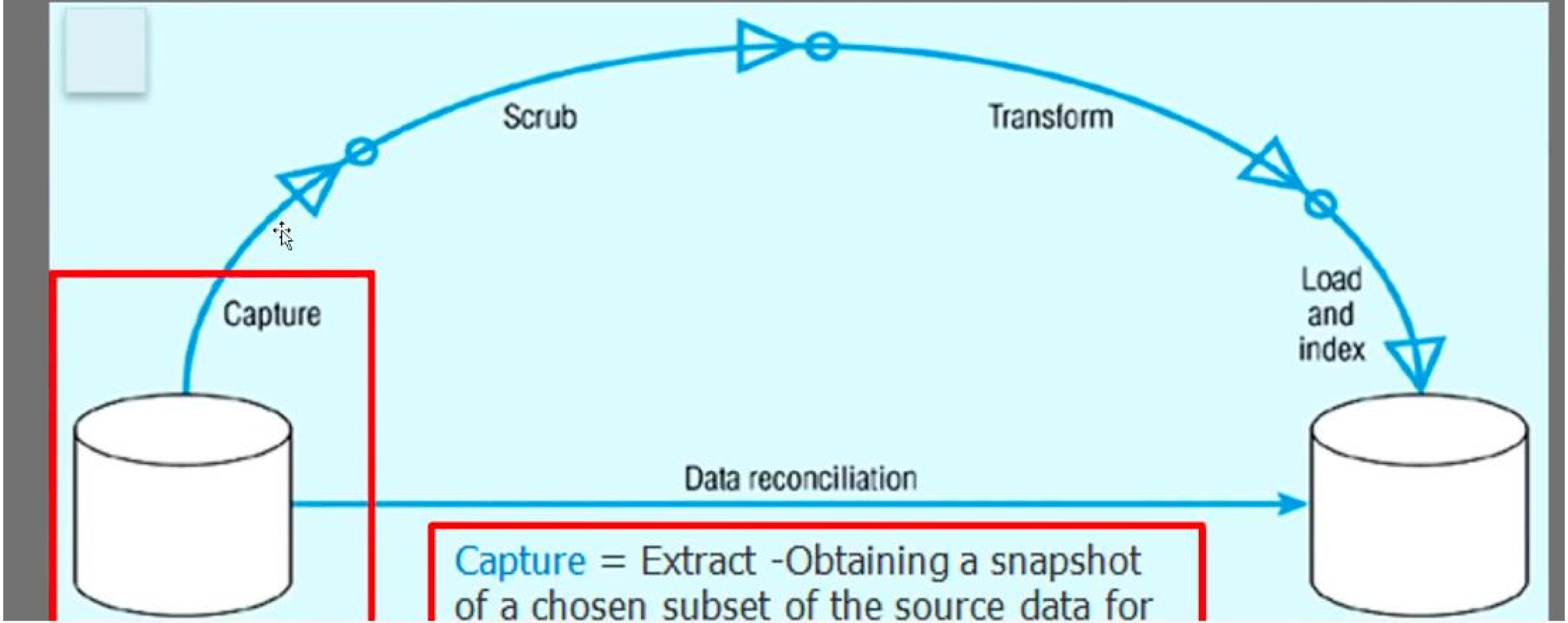
1. Extract or Capture
2. Scrub or Clean
3. Transform
4. Load and Index

1. Extract or Capture: As seen in the image below, the Capture or Extract is the first step of Informatica ETL process. It is the process of obtaining a snapshot of the chosen subset of data from the source, which has to be loaded into the data warehouse. A snapshot is a read-only static view of the data in the database. The Extract process can be of two types:

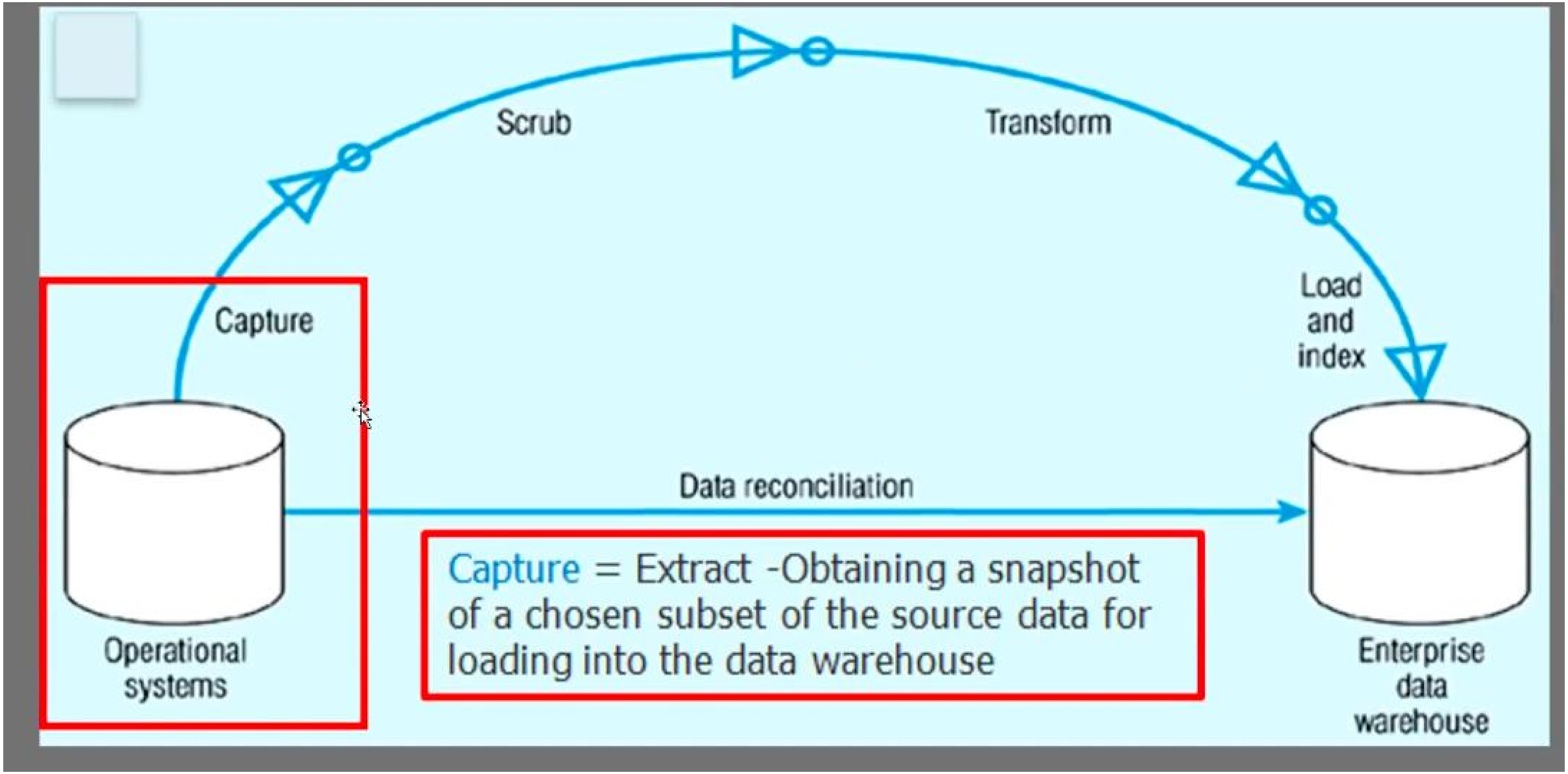


Full extract: The data is extracted completely from the source system and there's no need to keep track of changes to the data source since the last successful extraction.

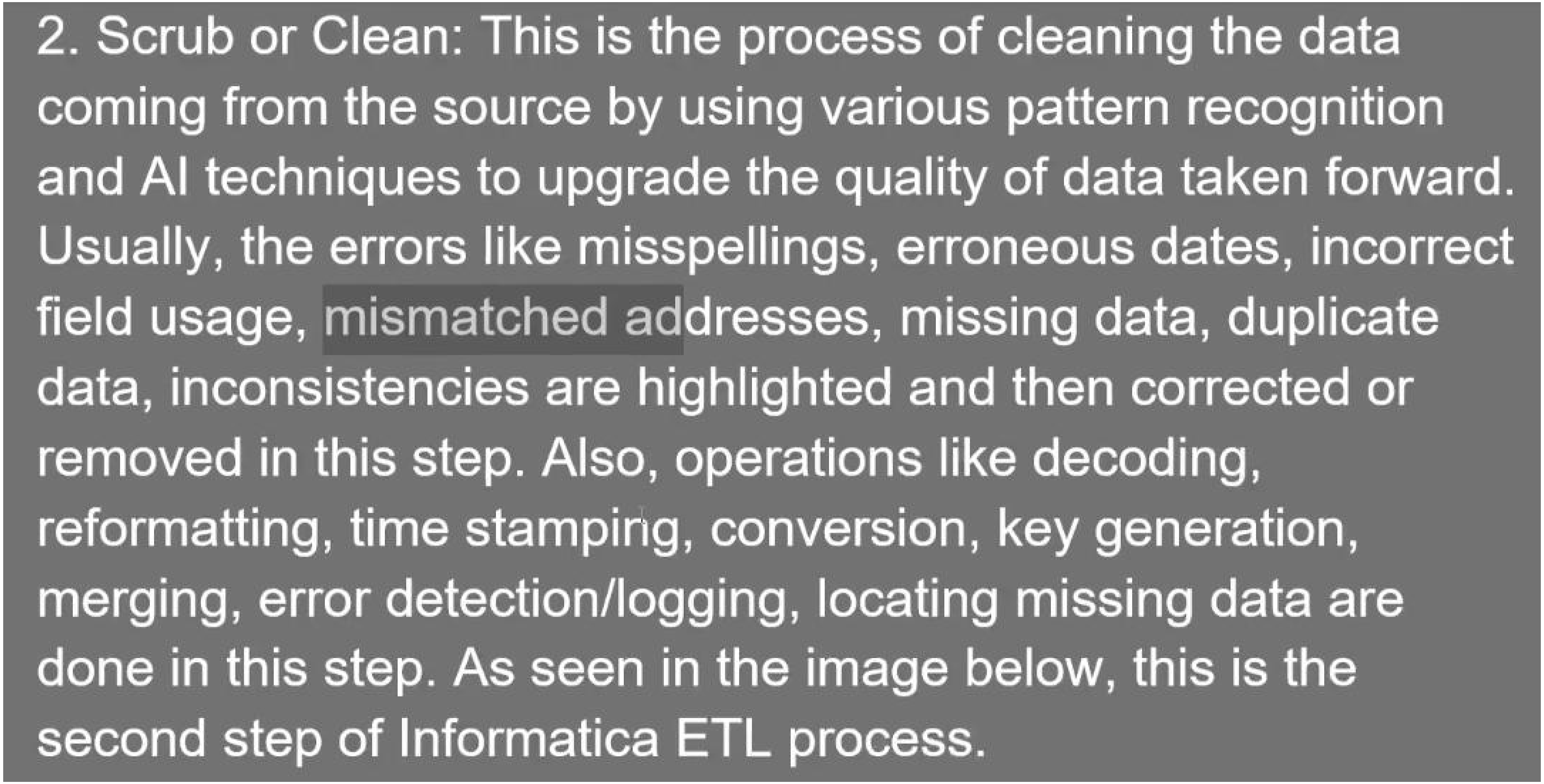
Incremental extract: This will only capture changes that have occurred since the last full extract.



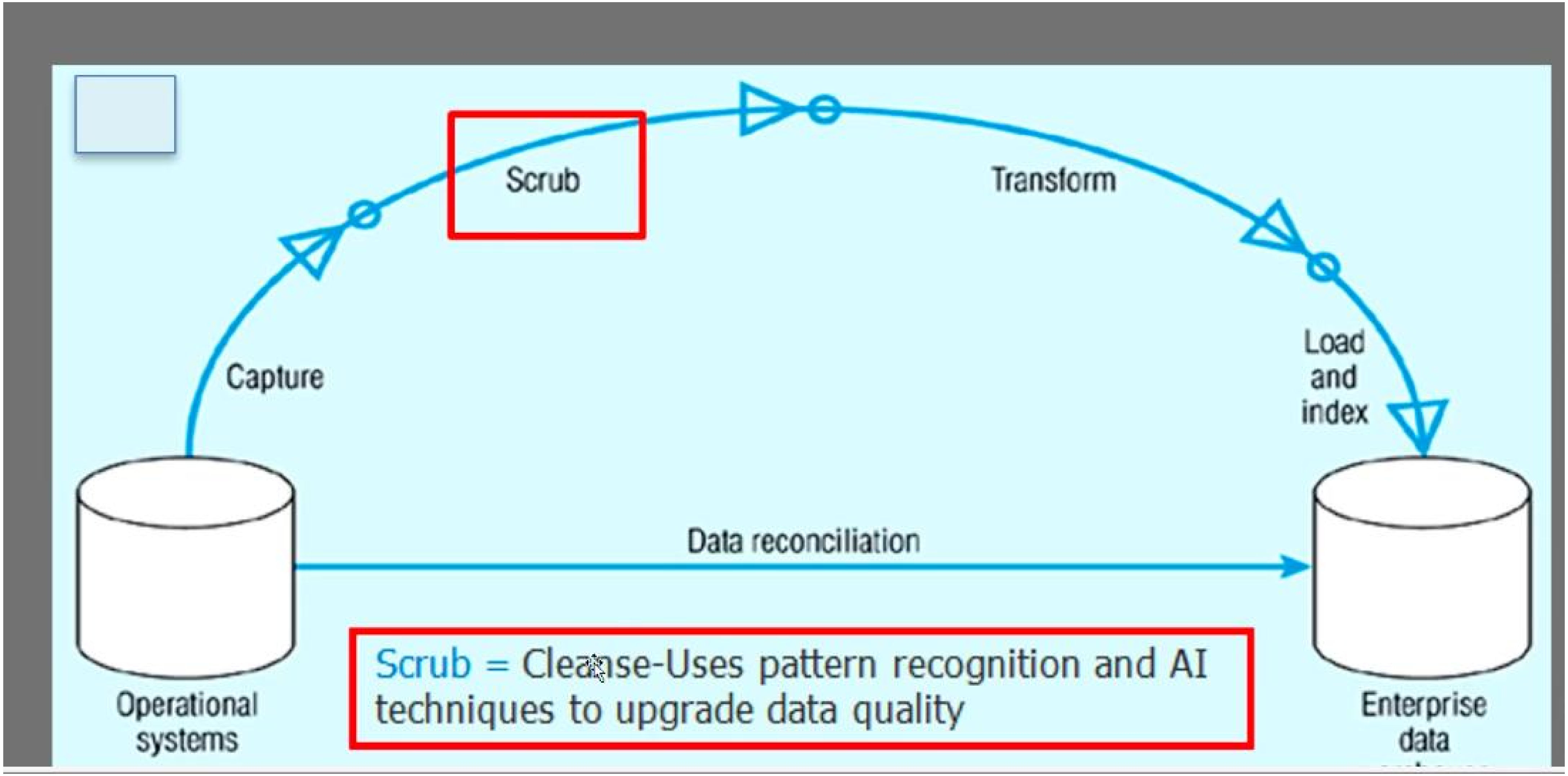
Capture = Extract -Obtaining a snapshot of a chosen subset of the source data for



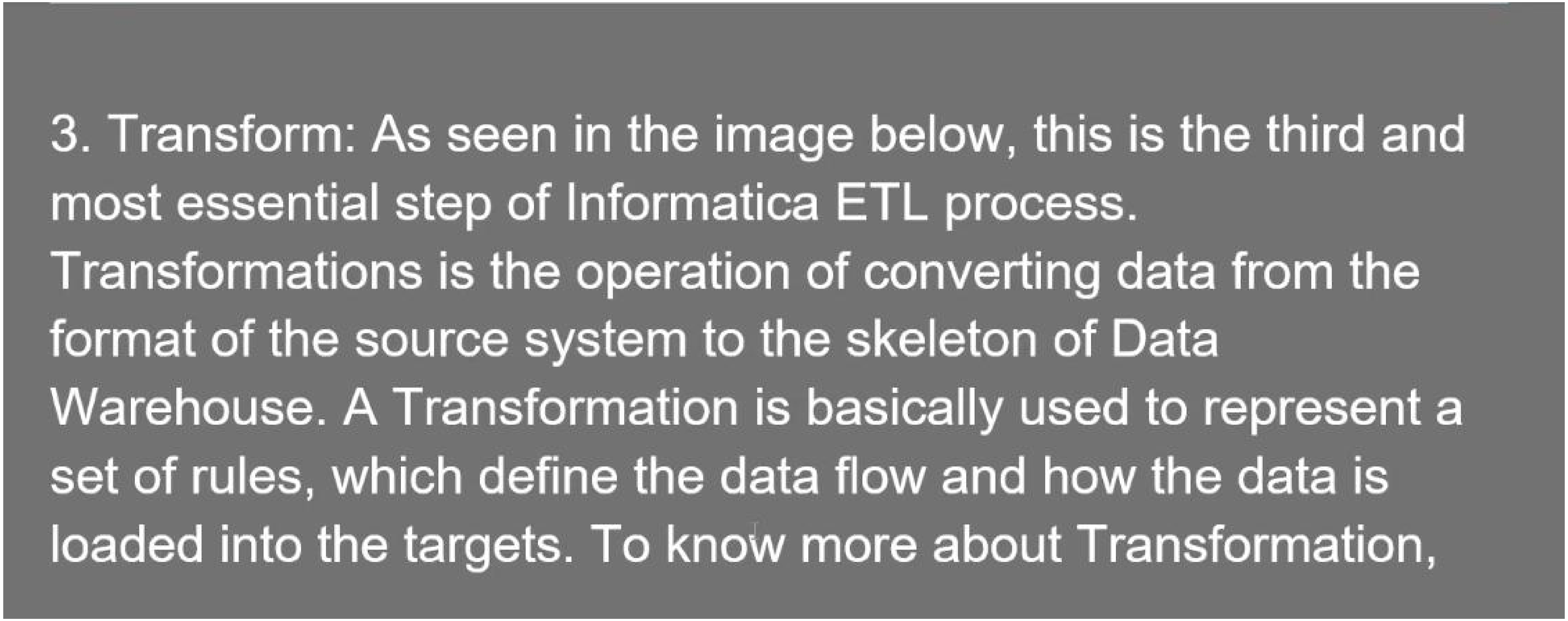
Capture = Extract -Obtaining a snapshot of a chosen subset of the source data for loading into the data warehouse



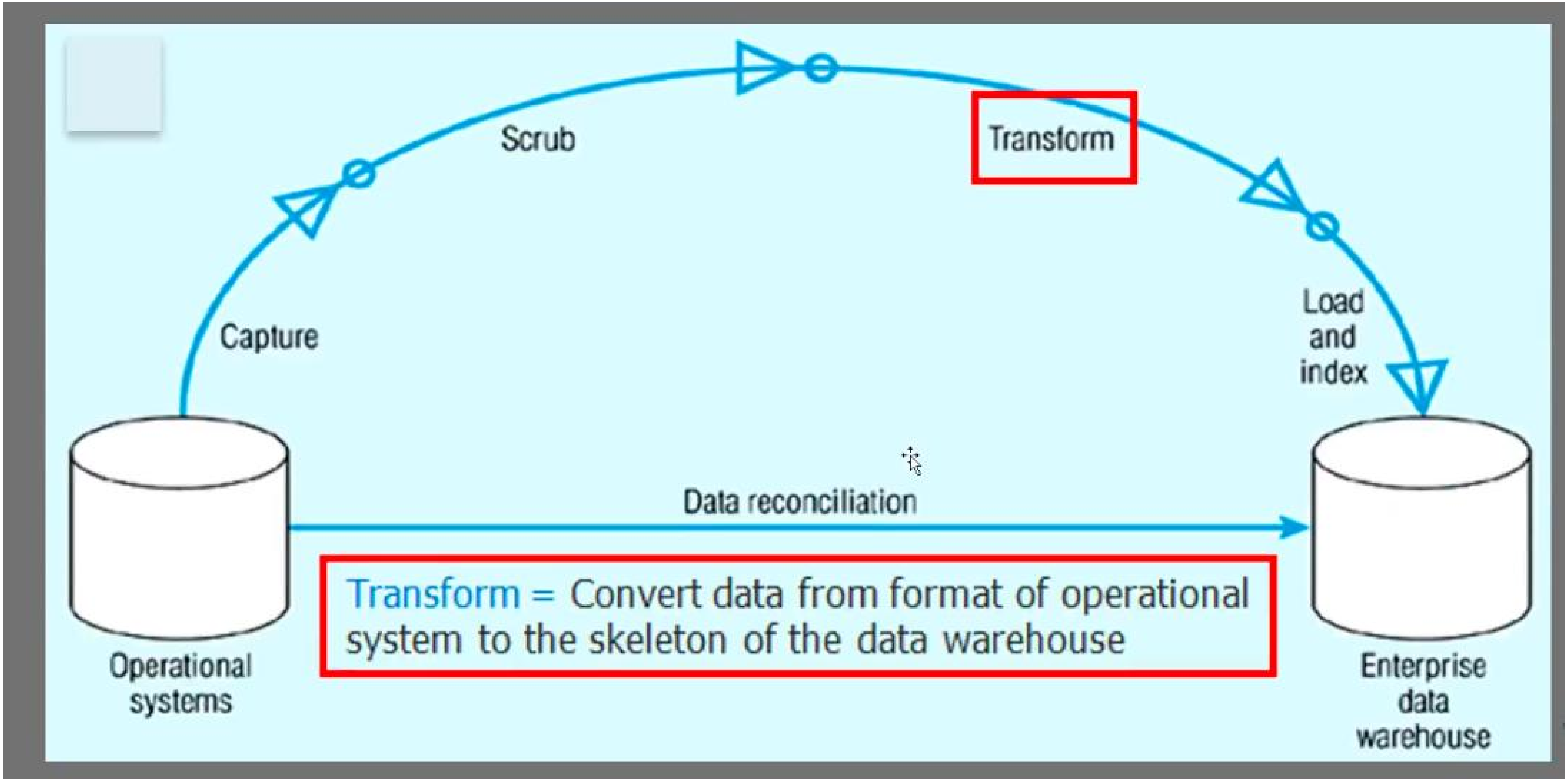
2. Scrub or Clean: This is the process of cleaning the data coming from the source by using various pattern recognition and AI techniques to upgrade the quality of data taken forward. Usually, the errors like misspellings, erroneous dates, incorrect field usage, mismatched addresses, missing data, duplicate data, inconsistencies are highlighted and then corrected or removed in this step. Also, operations like decoding, reformatting, time stamping, conversion, key generation, merging, error detection/logging, locating missing data are done in this step. As seen in the image below, this is the second step of Informatica ETL process.



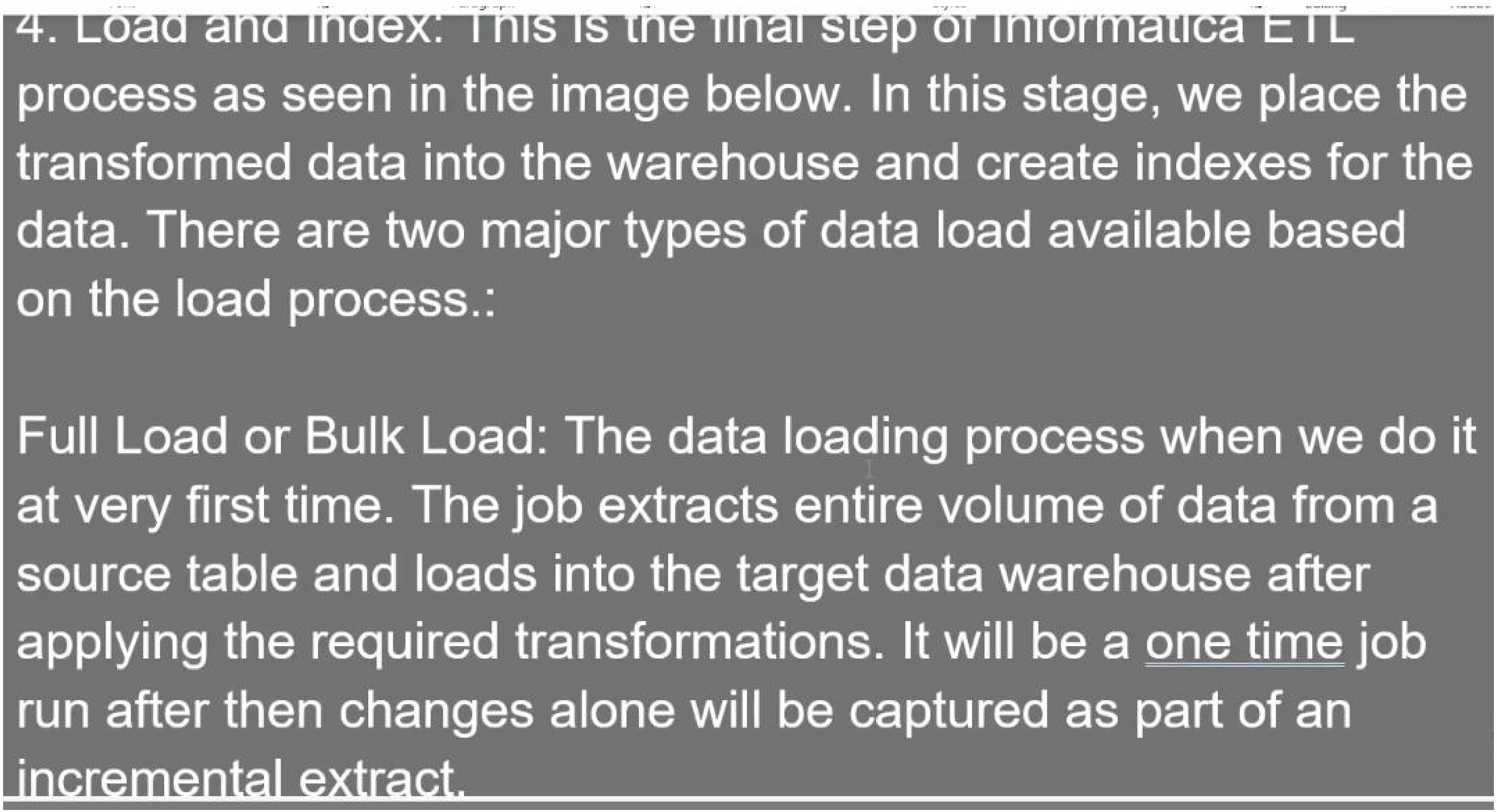
Scrub = Cleanse-Uses pattern recognition and AI techniques to upgrade data quality



3. Transform: As seen in the image below, this is the third and most essential step of Informatica ETL process. Transformations is the operation of converting data from the format of the source system to the skeleton of Data Warehouse. A Transformation is basically used to represent a set of rules, which define the data flow and how the data is loaded into the targets. To know more about Transformation,

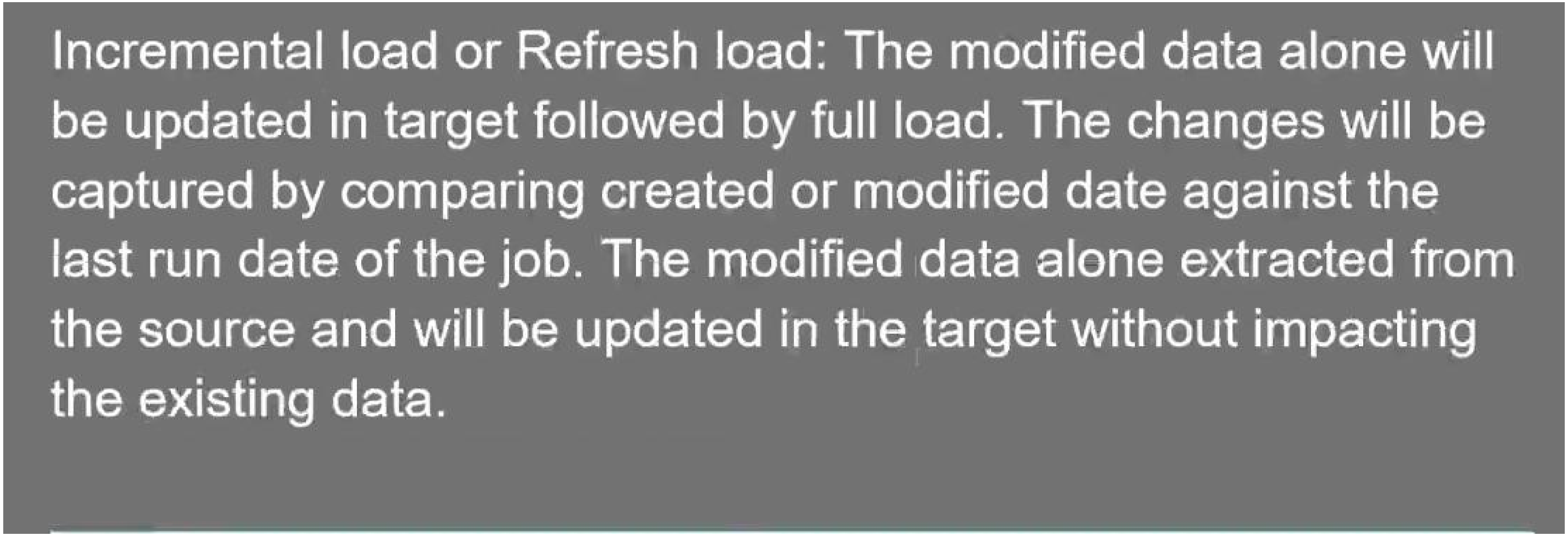


Transform = Convert data from format of operational system to the skeleton of the data warehouse

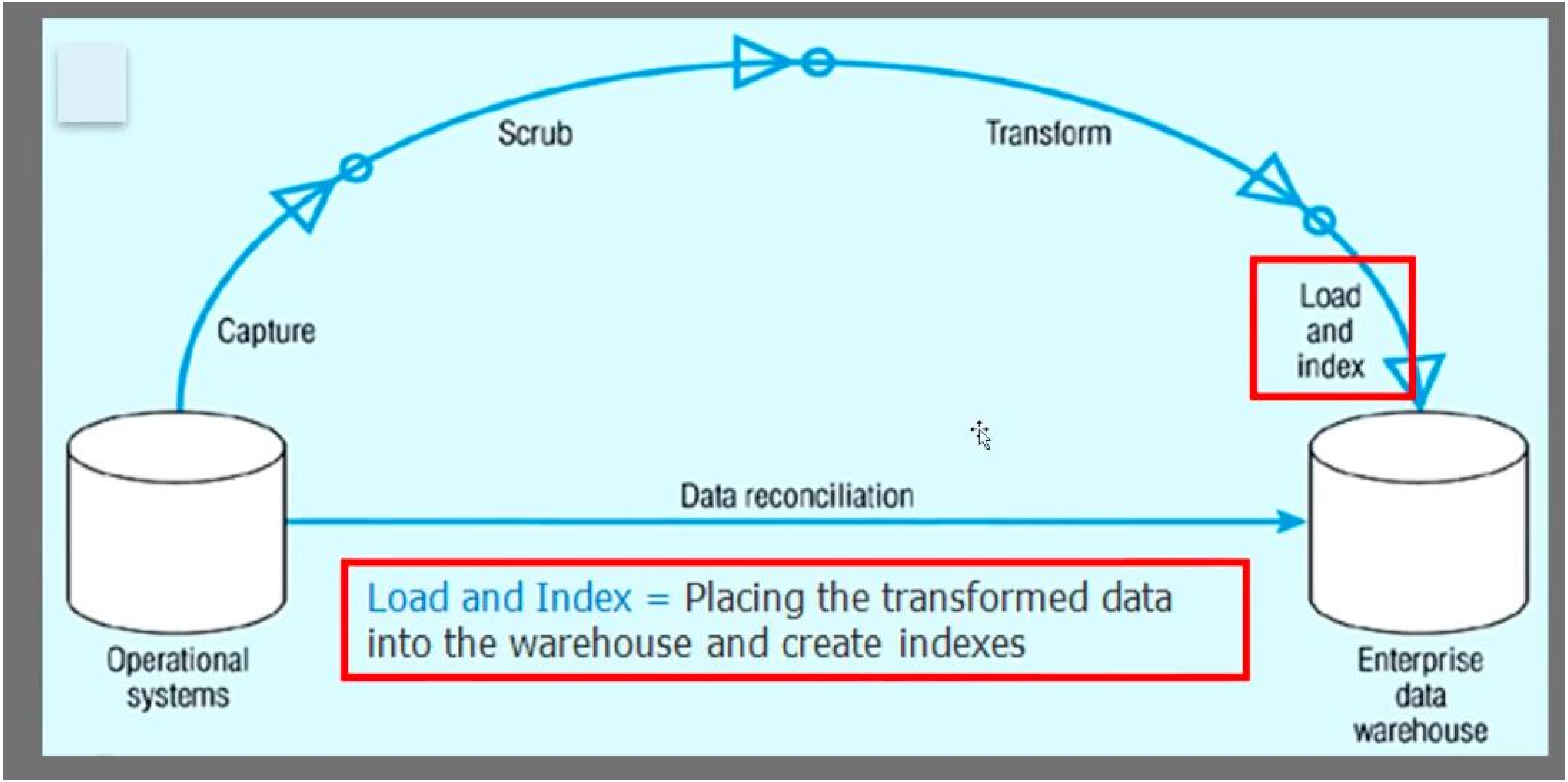


4. Load and index: This is the final step of informatica ETL process as seen in the image below. In this stage, we place the transformed data into the warehouse and create indexes for the data. There are two major types of data load available based on the load process.:

Full Load or Bulk Load: The data loading process when we do it at very first time. The job extracts entire volume of data from a source table and loads into the target data warehouse after applying the required transformations. It will be a one time job run after then changes alone will be captured as part of an incremental extract.



Incremental load or Refresh load: The modified data alone will be updated in target followed by full load. The changes will be captured by comparing created or modified date against the last run date of the job. The modified data alone extracted from the source and will be updated in the target without impacting the existing data.



Load and Index = Placing the transformed data into the warehouse and create indexes