DATA WAREHOUSING AND MINING

T.E. CSE-DS, Sem V
Academic Year: 2022-23

Data Warehousing Fundamentals: ETL

The ETL process

1. **Extraction**
2. **Transformation**
3. **Loading**

ETL takes place in the data staging area of DW

(Data staging area: is the place where all the extracted data from the source systems is temporarily stored and prepared/reshape the relevant data into useful information for loading into the datawarehouse)

- ETL exceeds 70% of total effort to build a DW
- Query processing is the backbone of the DW
- Query processing will not efficiently takes place if the source data is not extracted correctly, cleansed properly and integrated in proper format

Challenges in ETL process

A lot of disparities in the source systems make the ETL functions a challenging task

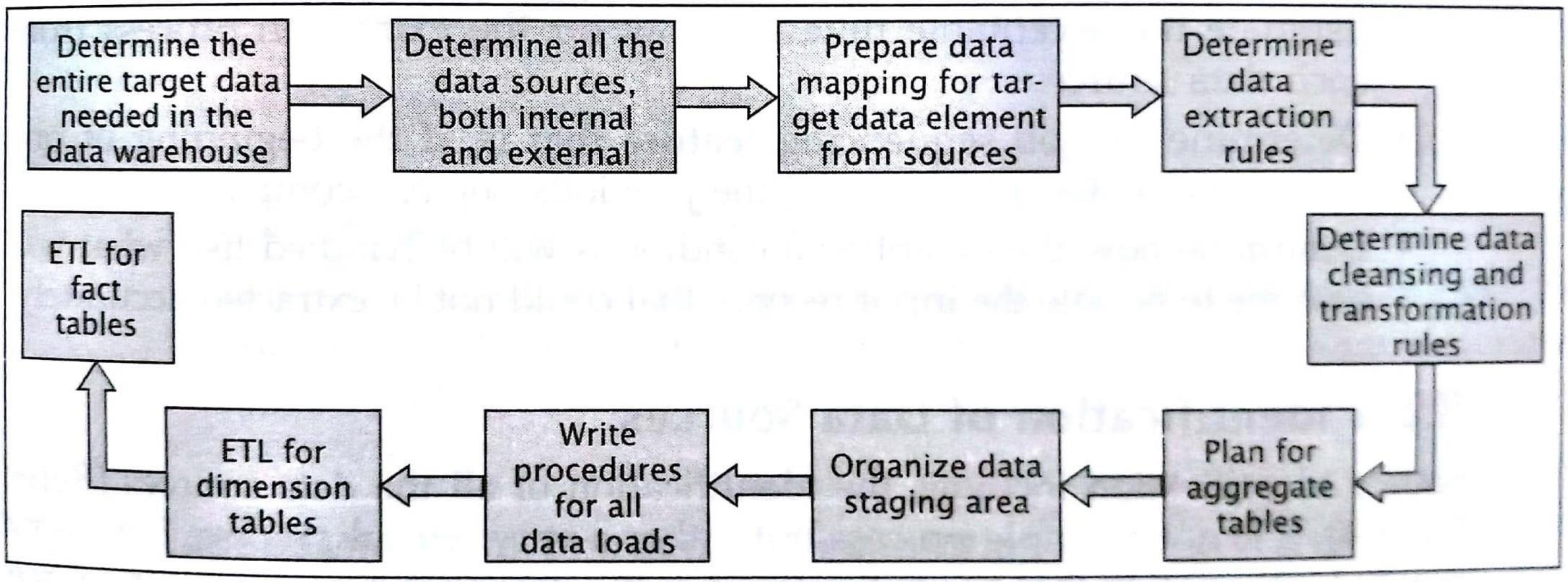
Following are the challenges in the ETL functions

1. Source systems are diverse and disparate
2. Source systems runs on different platforms and have different operating systems installed
3. Most of the operational systems do not preserve historical data
4. Quality of the data can not guaranteed in the older operational source systems.(legacy systems)
5. Structure of the source system keep changing over time with new technology

Challenges in ETL process

6. Same data elements represented differently in different source systems
7. Lack of consistency and lack of tools for resolving these consistencies
8. Data in the source system may be ambiguous or stored in cryptic(difficult to understand) form which does not provide any help to the user
9. The data type, format and naming convention may be different in different source systems

Major Steps in ETL process



Determine the entire target data needed in the data warehouse

Determine all the data sources, both internal and external

Prepare data mapping for target data element from sources

Determine data extraction rules

Determine data cleansing and transformation rules

Plan for aggregate tables

Organize data staging area

Write procedures for all data loads

ETL for dimension tables

ETL for fact tables

Major steps in the ETL process

Data Extraction

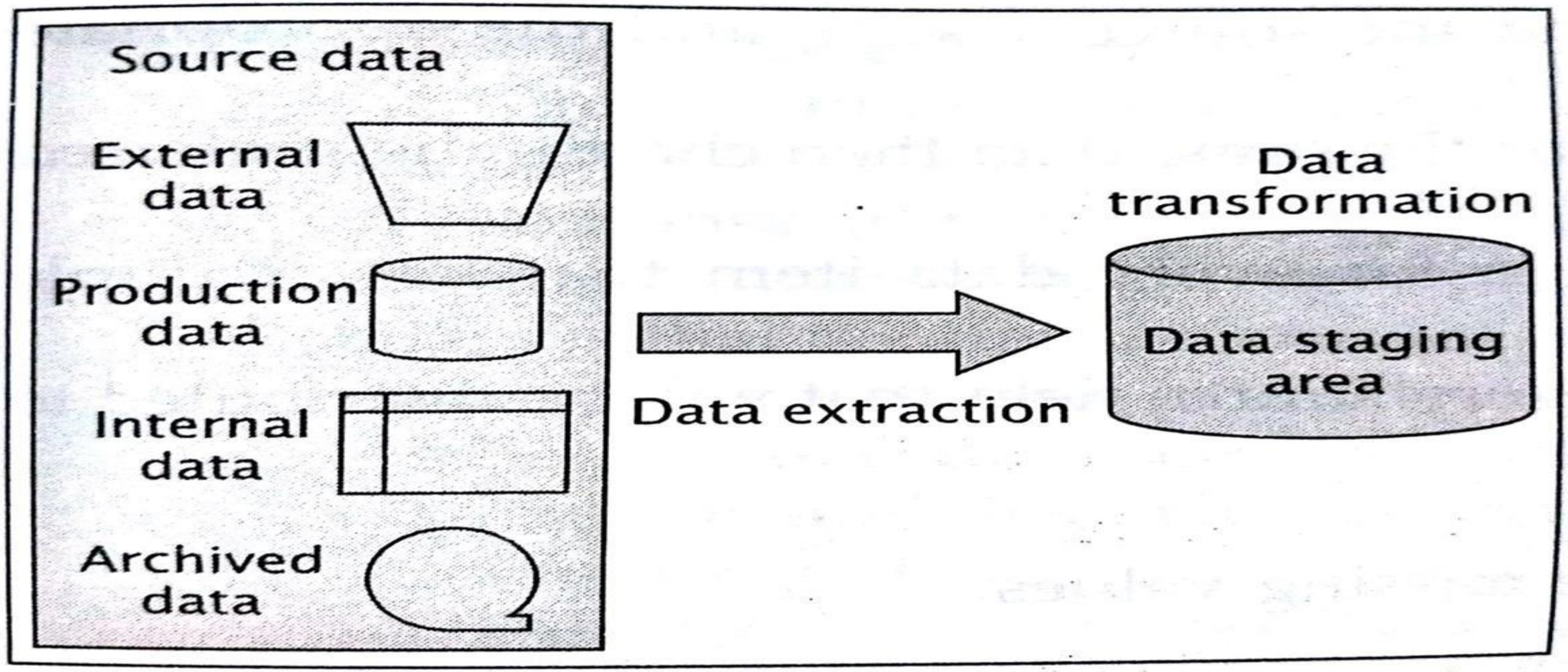
- In this data extraction stage the data flows from the different data source and pause at the staging area
- The majority of primary data sources consists of the enterprise’s operational systems
- Operational systems (legacy systems, client-server architecture, ERP) also called as data production systems are the resources of row data and it is not easy to extract all these data
- Data from outside/external sources requires temporary files to hold the data
- Effective data extraction is a key to the success of DW

Data Extraction

One has to consider following issues to formulate data extraction strategies

1. Identify the applications and systems from which the data will be extracted
2. For each identified data source, determine the method for data extraction(manually or by using tools)
3. For each data source, determine the extraction frequency (data extracted daily, weekly, monthly)
4. Estimate the acceptable time span(window) for the extraction process
5. Determine job sequencing feature (wait until previous completed)
6. Determine how to handle exceptional conditions (like what will be done to handle incorrectly extracted records)

Architecture of Data Extraction



Technical architecture for data extraction

Identification of data sources

For every piece of information that has to be stored in the data warehouse, first its source has to be identified

Sequence of steps to identify data sources

1. List every fact needed for analysis in fact tables
2. For every dimension table, list each and every attribute
3. For each target data item, find the source system and the appropriate source data item
4. If there are multiple sources for the same data then choose the preferred sources e.g. age, DOB
5. Formulate consolidation rules for every data item that has multiple sources (age=current date – date of birth)

Identification of data sources

6.Formulate splitting rules for every source field that will be distributed to multiple fields (Name→F_name, M_name, L_name)

7. Determine the default values (height, weight on the basis of age)

8. Search the source data for the missing values

Data Extraction Techniques:

Data extraction techniques are classified into two categories

1. Immediate data extraction (IDE)
2. Deferred data extraction (DDE)

Immediate Data Extraction

Immediate data extraction is a real-time extraction. It occurs as the transaction happens at the source databases and files

There are three options for immediate data extraction

1. **Capture through transaction logs**

(In the field of databases, a transaction log is a history of actions executed by a database management system used to guarantee ACID properties over crashes or hardware failures)

2. Capture through database triggers

(A database trigger is procedural code that is automatically executed in response to certain events on a particular table or view in a database. The trigger is mostly used for maintaining the integrity of the information on the database)

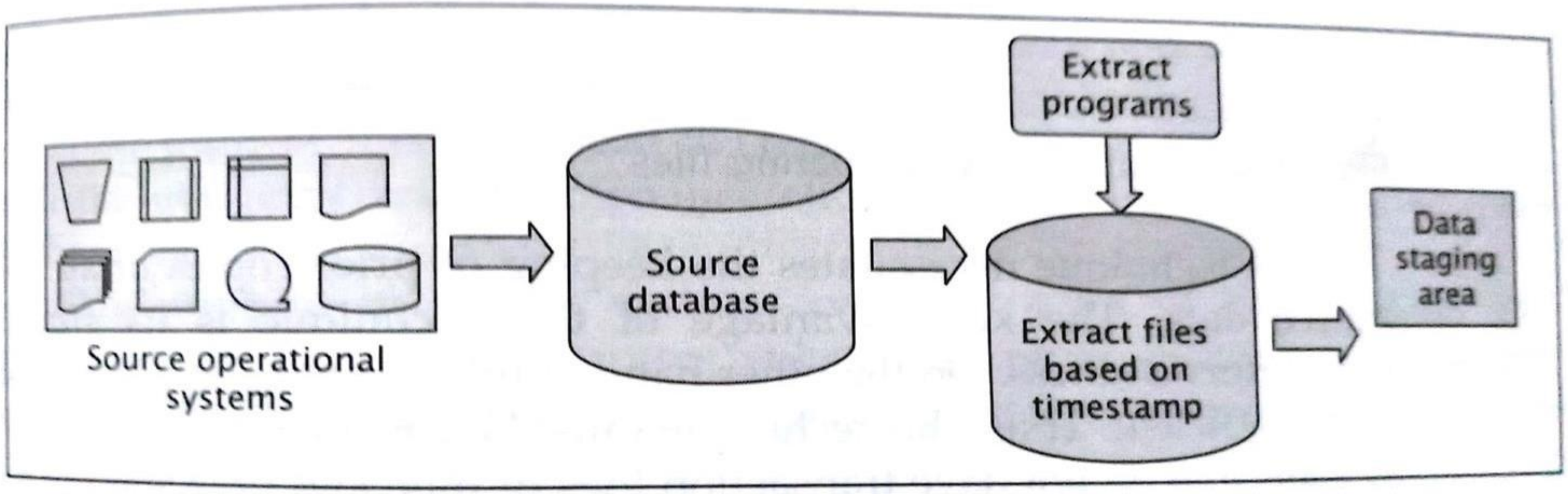
Deferred Data Extraction

The data captured does not take place in real time. The capture is done at a later point of time

Capture based on data and timestamp

- Every transaction in source system is marked with a timestamp and that will be used for selecting the records for data extraction
- The timestamp shows the date and time at which the source record was created or updated
- The data is usually extracted during the midnight
- This technique will work with every type of source system
- It captures the latest state of the source data

Deferred Data Extraction



Capture based on data and timestamps

Transformation

Extracted data from source server is raw and not usable in its original form.

Therefore the data should be mapped, cleansed, and transformed.

Transformation is an important step where the ETL process adds values and change the data, such as the BI reports, can be generated.

Transformation

- **Filtering:** Only specific attributes are loading into the data warehouse.
- **Cleaning:** Filling up the null values with specific default values.
- **Joining:** Join the multiple attributes into the one.
- **Splitting:** Splitting the single attribute into multiple attributes.
- **Sorting:** Sort the tuples based on the attributes.

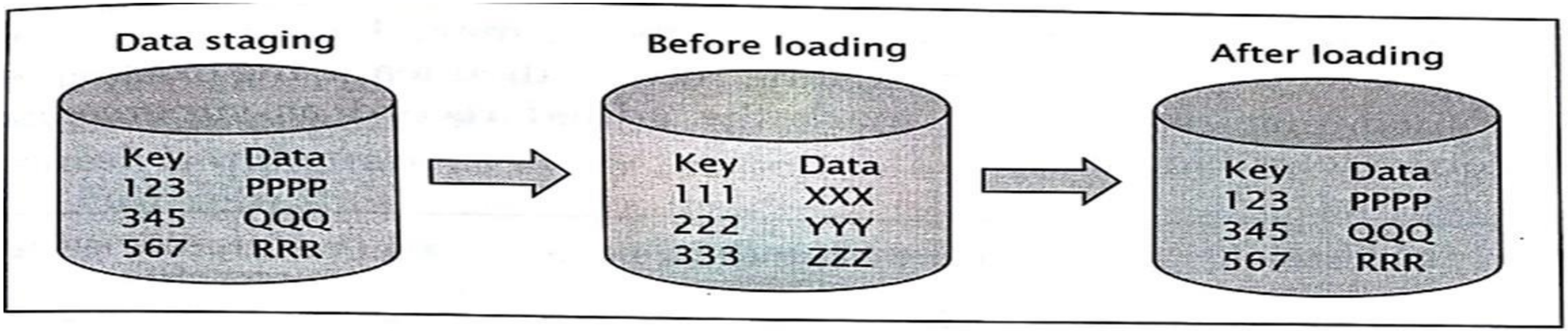
Data loading:

- Once the extraction and transformation of data has been done the next major set of functions consists of taking the prepared data, applying it to the DW and storing it in the DW repository.
- During the data loads the DW has to be offline for the duration of the loading process
- If you want to keep some parts of the DW online you can divide the load process into smaller chunks and populate only few tables at a time

Data Loading

Different modes of data loading

1. Load : if the target table to be loaded is already exists and contains some data records, then the load process will wipe out the existing data and store the data from incoming file



Data staging

Key	Data
123	PPPP
345	QQQ
567	RRR

Before loading

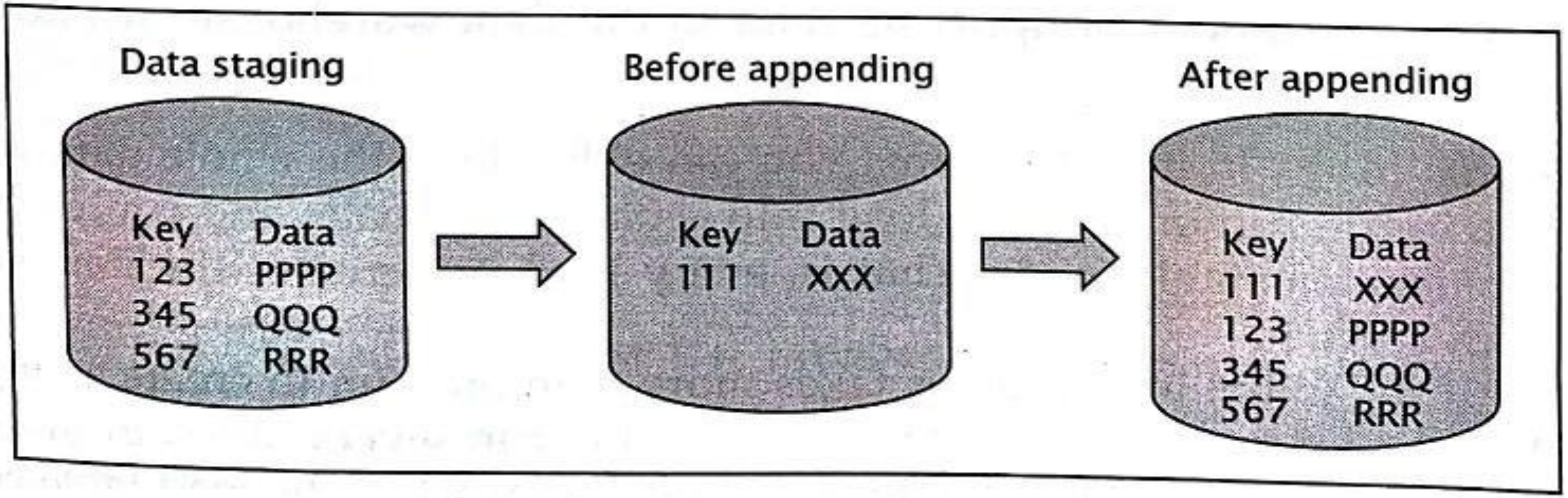
Key	Data
111	XXX
222	YYY
333	ZZZ

After loading

Key	Data
123	PPPP
345	QQQ
567	RRR

Load mode of data loading

APPEND



Data staging

Key	Data
123	PPPP
345	QQQ
567	RRR

Before appending

Key	Data
111	XXX

After appending

Key	Data
111	XXX
123	PPPP
345	QQQ
567	RRR

Append mode of data loading

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.
- Loading can be carried in two ways:

1. **Refresh:** Data Warehouse data is completely rewritten. This means that older file is replaced. Refresh is usually used in combination with static extraction to populate a data warehouse initially.
2. **Update:** Only those changes applied to source information are added to the Data Warehouse. An update is typically carried out without deleting or modifying preexisting data. This method is used in combination with incremental extraction to update data warehouses regularly.