CHAPTER 2

ETL Process and OLAP -Based on CO2

By-Safa Hamdare

CSC603.2: Students should be able to use star schema for designing a data warehouse and execute appropriate OLAP queries.

Outline: Part 1

- Major steps in ETL process
- Data extraction: Techniques
- Data transformation: Basic tasks
- Major transformation types
- Data Loading: Applying Data

ETL Overview

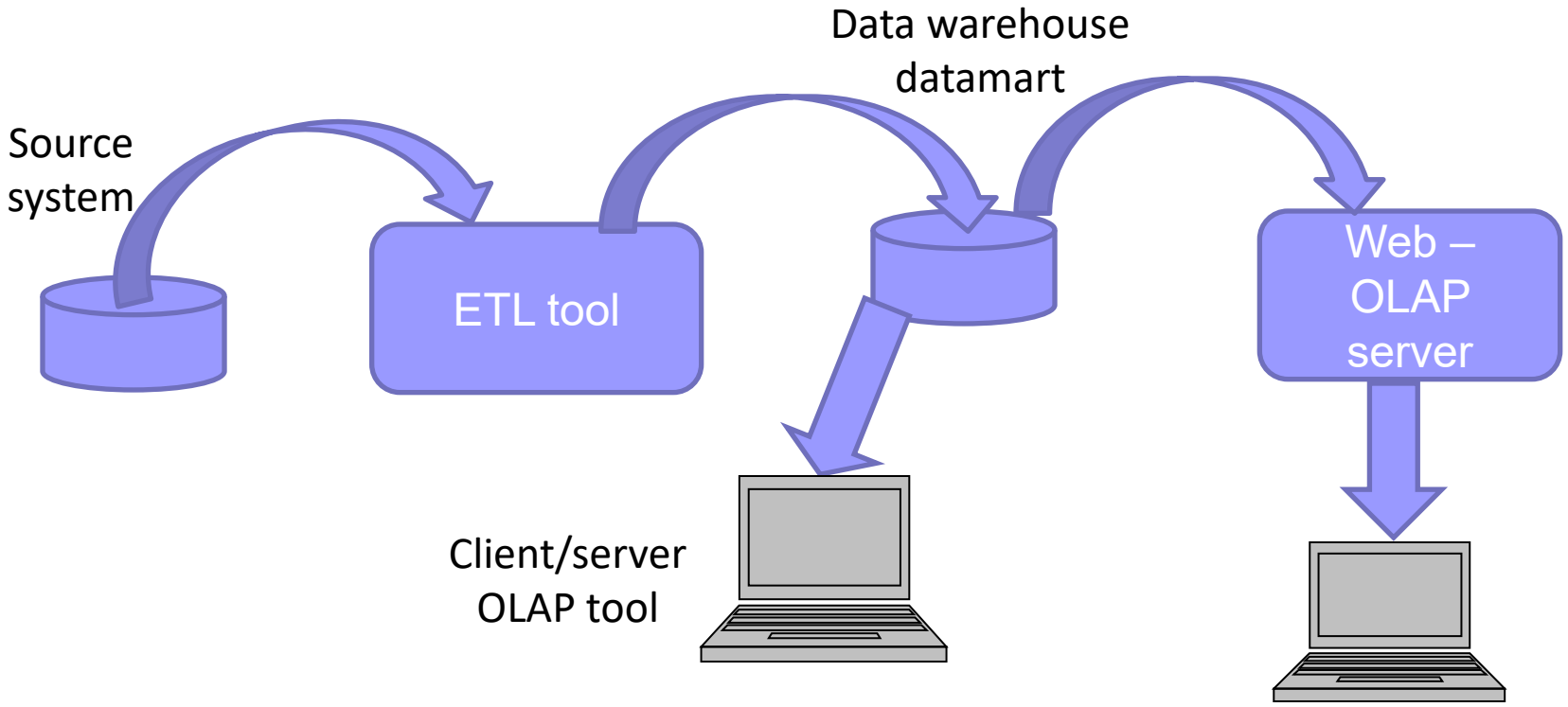
- Extraction Transformation Loading – ETL
- To get data out of the source and load it into the data warehouse – simply a process of copying data from one database to other
- Data is extracted from an OLTP database, transformed to match the data warehouse schema and loaded into the data warehouse database
- Many data warehouses also incorporate data from non-OLTP systems such as text files, legacy systems, and spreadsheets; such data also requires extraction, transformation, and loading
- When defining ETL for a data warehouse, it is important to think of ETL as a process, not a physical implementation

ETL Tool

- *An **ETL tool** is a tool that reads from one or more sources, transforms the data so that it is compatible with a destination and loads the data to the destination.* It is not a one time event; as new data is added to the Data Warehouse periodically – monthly, daily, hourly
- Because ETL is an integral, ongoing, and recurring part of a data warehouse
- Features:
  - Automated data movement across data stores
  - Extensible, Robust, Scalable Infrastructure
  - Easily changeable

ETL Tool:

- Reshapes relevant data from the source systems into the information stored in the warehouse.



Extraction Transformation Load Process

1. Explain ETL data Warehousing in detail.
2. What is meant by ETL? Explain the ETL process in detail.
3. Explain the ETL cycle for a data warehouse in detail.

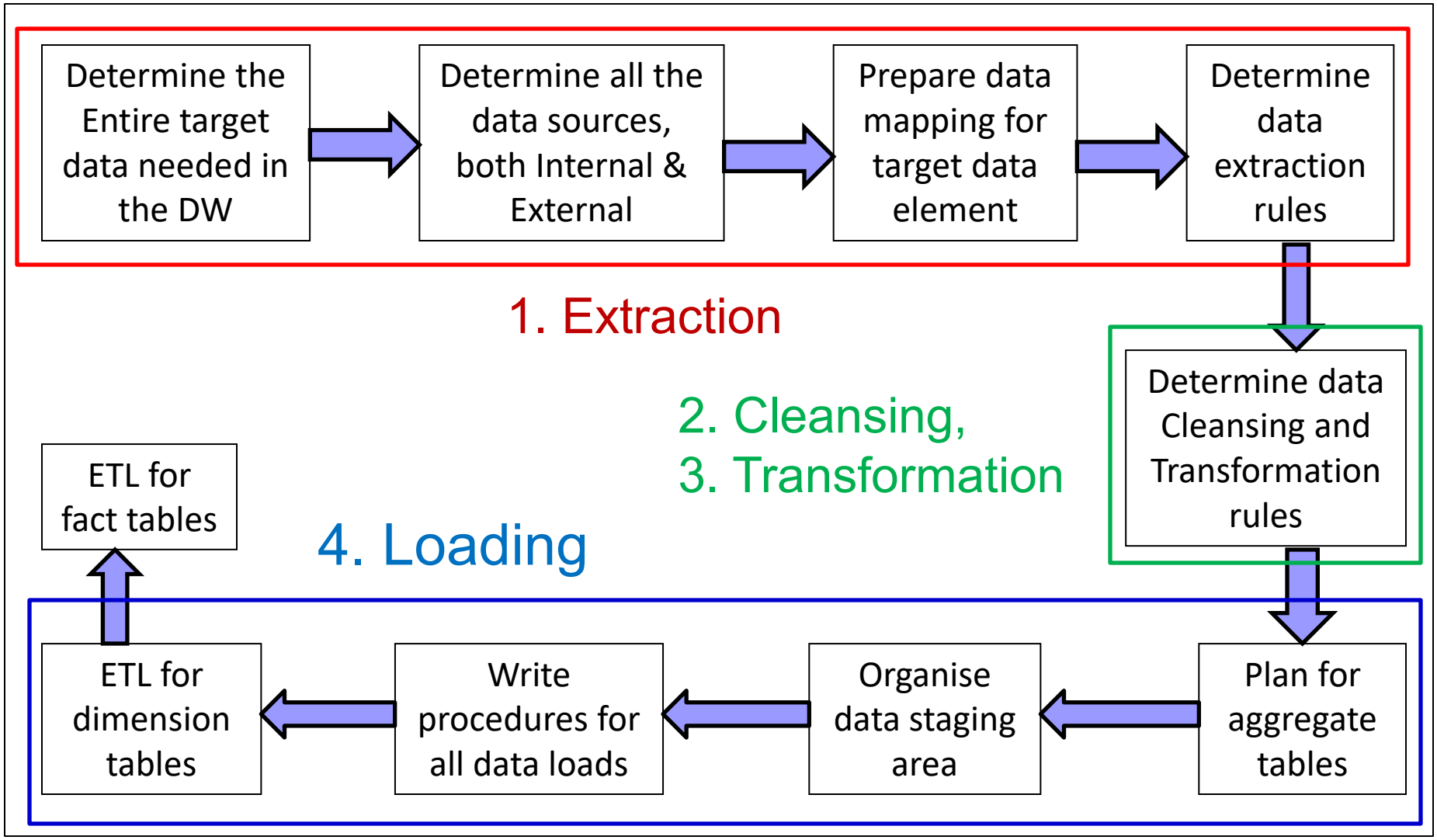
May 2010, Dec 2010, May 2011,Dec 2011, May 2012, Dec 2012, Dec 16, Dec 17, May 16, May 17, May 18

4 Major Process of the Data Warehouse: *ETL Process*

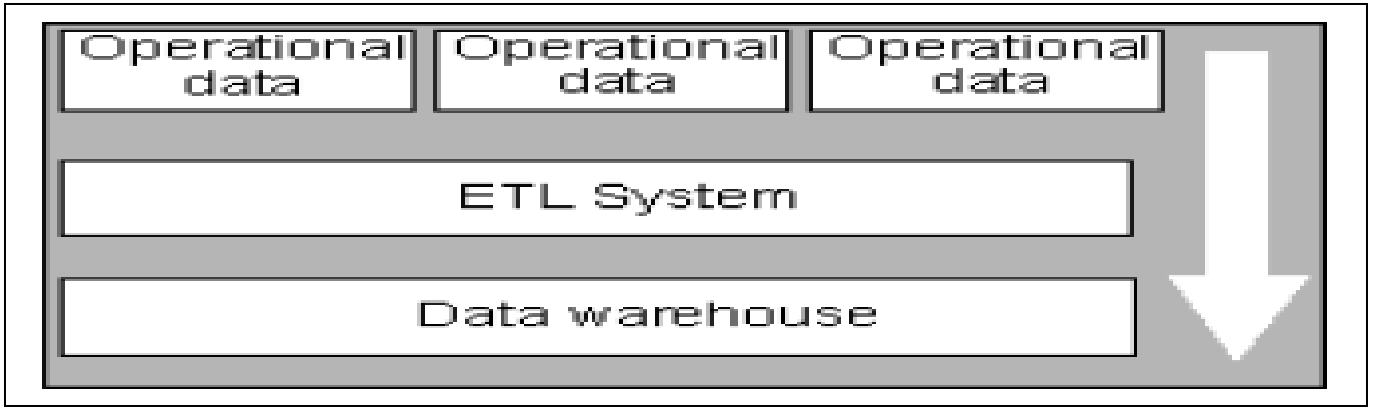
1. **Extract** – Data from the Operational Systems and bring it to data Warehouse.
2. **Cleanse** – To make sure it is of sufficient quality to be used for decision making.
3. **Transform** –Data into internal Format and structure of the data warehouse.
4. **Load** – Cleanse data is put into the data Warehouse.

4 process from Extraction through Loading often referred to as *DATA STAGING*

The ETL Process



1. Extraction



- Data is extracted from heterogeneous data sources.
- The integration of all of the disparate systems across the enterprise is the real challenge to getting the data warehouse to a state where it is usable
- For this reason, it is necessary to extract the relevant data from the operational database before bringing it to data warehouse.
- E.g.: *Extraction tool : Data Junction*

Challenges in Extracting

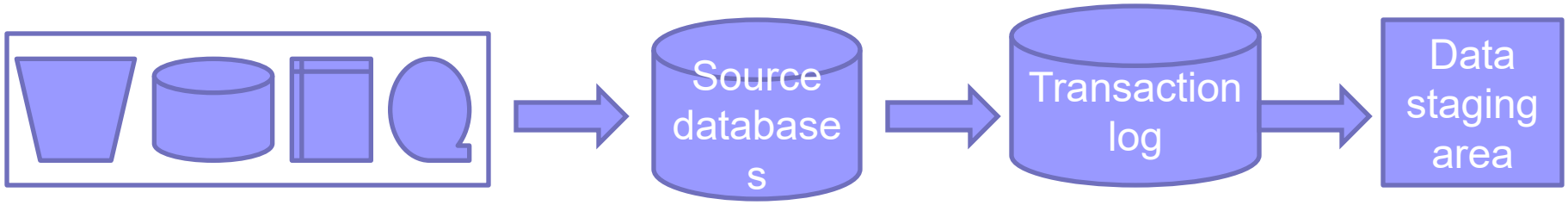
- Source systems are diverse and disparate
- Source systems run on different platforms
- Most operational systems do not preserve historical data
- Quality of data cannot be guaranteed.
- Structures of the source system keep changing over time
- Data may be ambiguous, or stored in cryptic form.

Extracting data Techniques:

1. Immediate data extraction technique
   - Real time data extraction
     a) ***Capture through transaction logs***
     b) ***Capture through triggers***
     c) ***Capture in source applications***
2. Deferred data extraction
   - Data is extracted at a later point in time
     d) ***Capture based on date and timestamp***
     e) ***Capture by comparing files***

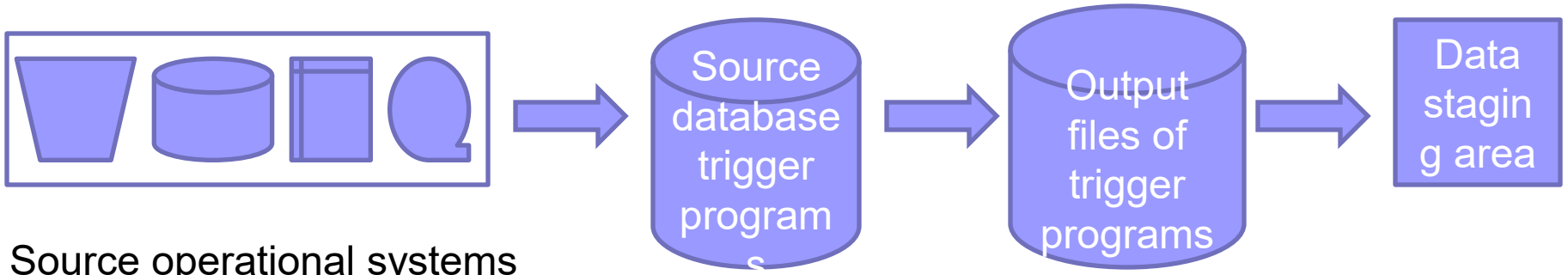
a) Capture through transaction logs

- Every *update, add or delete* is recorded in the transaction log
- These are maintained for recovery purposes
- Reads log and selects all committed transactions.



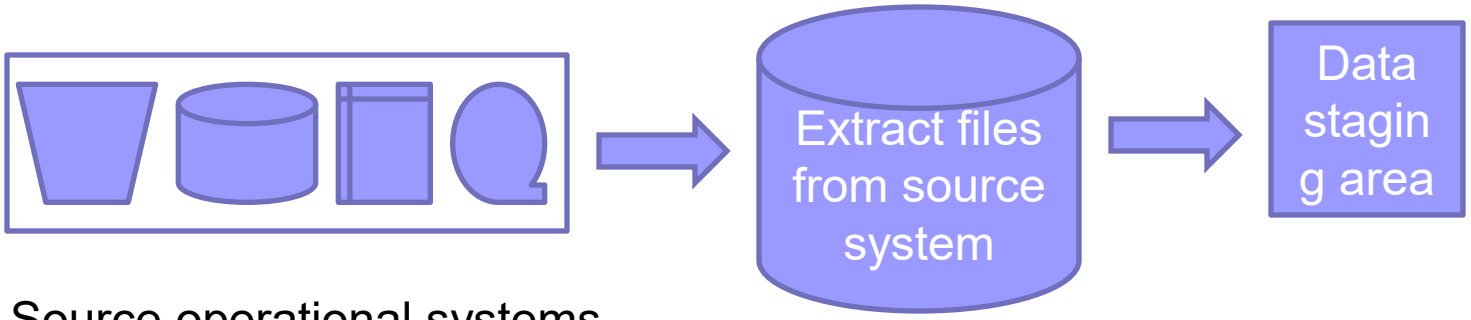
b) Capture through triggers

- Triggers can be created for *all events for which data needs to be captured.*
- Output of the trigger program is written on a separate file that will be used to extract data.
- Occurs at the source system



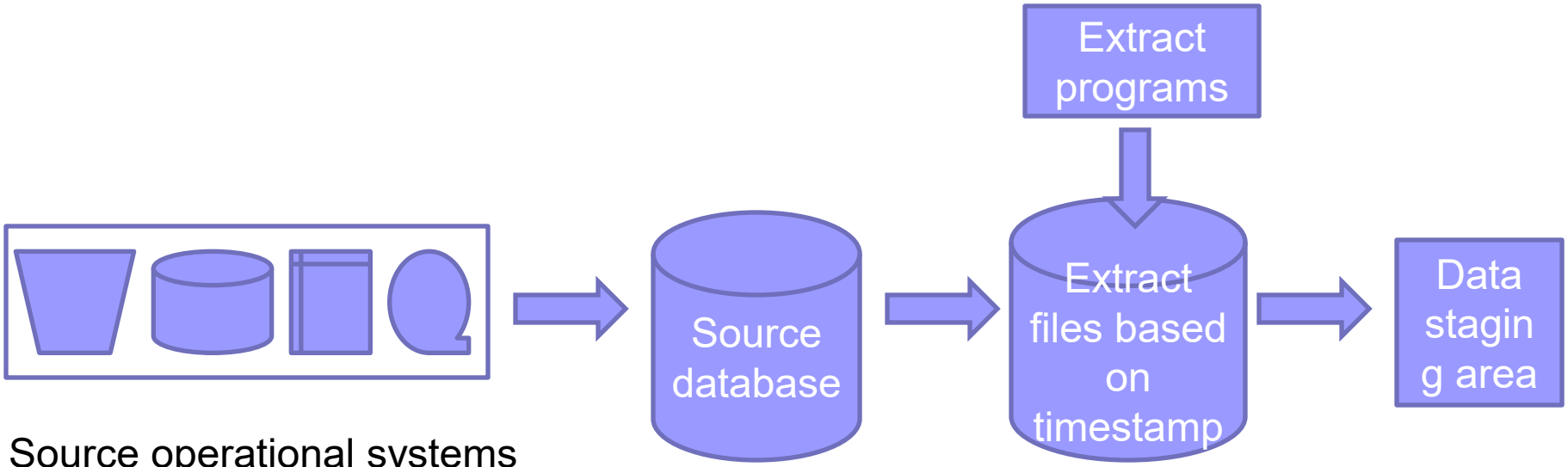
c) Capture in source applications

- A source application is used to capture data.
- All applications that write to source files, also write to the data warehouse.



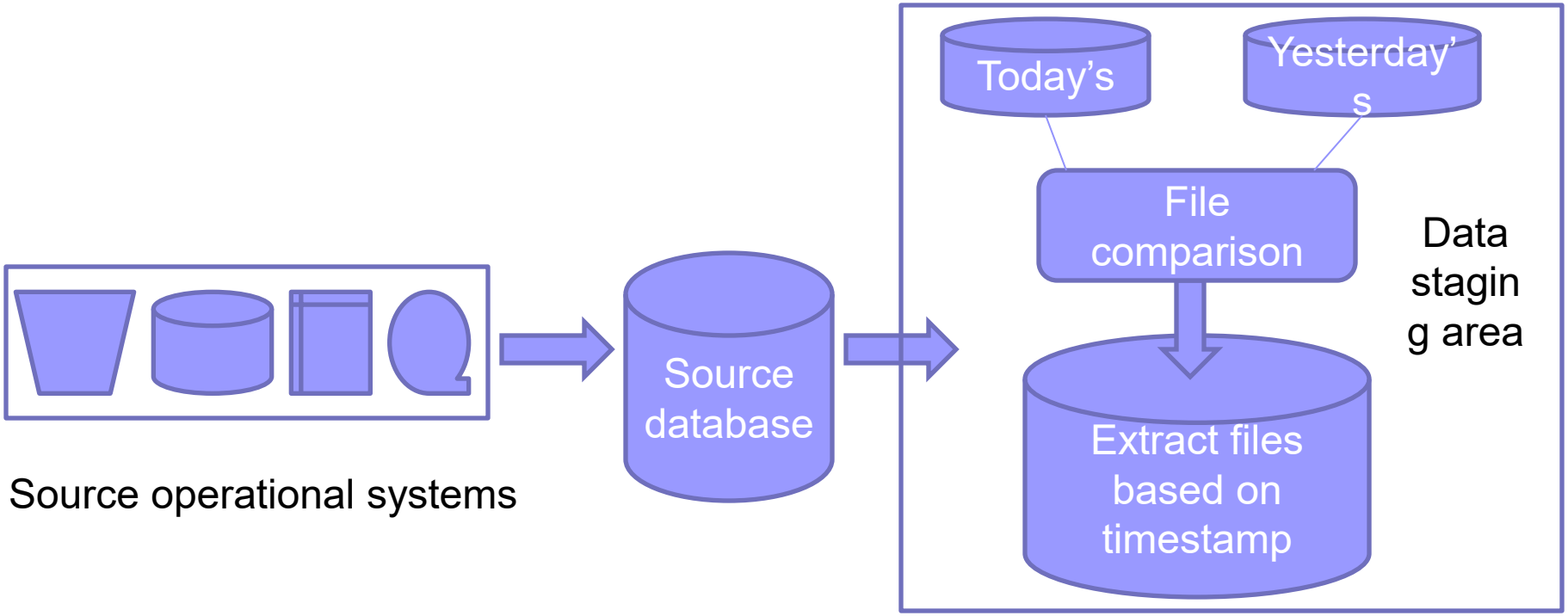
d) Capture based on date and timestamp

- Every time a record is *created, updated or deleted* in the source system, *it is marked with a timestamp*.
- Data is captured after a certain interval of time.
- Records whose timestamp is later than the previous capture, is extracted the next time.
- Records in the source can be deleted soon after the extraction.



e) Capture by comparing files

- Also called *"snapshot"* technique. It compares two snapshots of source data before capture.
- Requires source to keep prior copies of data.



Comparison of Extraction Techniques

1. Capture through transaction logs	1. Cannot capture specifications 2. Doesn't affect performance of source systems 3. Does not require any revisions to existing source applications 4. Cannot be used on file-oriented systems
2. Capture through triggers	1. Cannot capture specifications 2. Does not affect performance of source systems 3. Does not require any revisions to existing source applications 4. Cannot be used on file-oriented systems
3. Capture in source applications	1. Can capture specifications 2. Does not affect performance of source systems 3. Requires regular revisions to existing source applications 4. Can be used on file-oriented systems
4. Capture based on date & timestamp	1. Can capture specifications 2. Does not affect performance of source systems 3. Requires regular revisions to existing source applications 4. Can be used on file-oriented systems.
5. Capture by comparing files	1. Can capture specifications 2. Does not affect performance of source systems 3. Does not require any revisions to existing source applications 4. Can be used on file-oriented systems.

2. Data Cleansing

- Extracted data cannot be directly stored in the DW
  - **Data is raw** , cannot be used for analysis.
  - **Data quality is poor**, too many different sources.
- Small errors in the Operational Database system may cause huge distortions during analysis.
- Data Cleaning, also called Data Cleansing or Scrubbing, deals with ***detecting and removing errors and inconsistencies*** from data in order to improve the quality of data.
- Data Quality Problems:
  - E.g. *Misspellings, missing information or Invalid Data.*

Steps in Data Cleansing:

1. **Parsing-** Locates & identifies individual data elements in the source files and then isolates it in target files. *E.g. parsing first, middle and Last name*
2. **Correcting-** Corrects parsed individual data element using algorithms & secondary data sources. *E.g. replace an incomplete address and adding zip code*
3. **Standardizing-** Routines to transform data into its consistent format using both standard and custom business rules. *E.g. Address: Mum to Mumbai, Gender: 0/1 to M/F*
4. **Matching-** Searching & Matching records within and across the parsed, corrected and standardized data to eliminate duplications. *E.g. Identifying similar names & addresses.*
5. **Consolidating-** Analyzing & identifying relationship between matched records & consolidating them into one representation

Transformation

3. Data Transformation

- Transformation process deals with rectifying any inconsistency (if present).
- Actually changes data and provides guidance whether data can be used for its intended purposes.
- One of the most common transformation issues is ***'Attribute Naming Inconsistency'.***
  - It is common for the given data element to be referred to by different data names in different databases.
  - **Example:** Employee Name may be EMP_NAME in one database, Ename in the other.
- Thus one set of Data Names are picked and used consistently in the data warehouse.
- Once all the data elements have right names, they must be converted to common formats.

Transformation Types:

- Format revisions
  - *Changes of data type, field lengths.*
- Decoding of fields
  - *Standardize attribute values (Ename, Emp_Name)*
  - *Decode cryptic data*
- Calculated or derived values
  - *E.g., average daily balance, profit.*
- Splitting of single fields
  - *E.g., address -> city, state, pin code.*
- Merging of information
  - *E.g. product code, description, cost etc may come from different sources (merge and keep as Product_info)*

Transformation Types: Contd.

- Character set conversion
  - *Textual data -> standard data fields.*
- Conversion of units or measurements
  - *Global branches have to standardize units*
  - (i.e. Rupees to Dollar, meter to centimeter)
- Date/ time conversion
  - *Have a common format*
  - *Summarization*
  - *Choosing required granularity*
- Key restructuring

Data Mapper is one such comprehensive tool for doing these transformations automatically

Data Transformation can have the following tasks:

- ***Set of Basics Tasks are***

1. **Selection of Data:** Select desired data from source ; it can be either whole records or parts of several records.
2. **Splitting / joining:** Join operation of the selected parts during data transformation, it is data manipulation on selected records.
3. **Conversion:** Conversion of single fields to standardize the data extraction format from different source systems and understandable to the users.

Data Transformation can have the following tasks:

- ***Set of Basics Tasks are***

4. **Summarization:** Sometimes the lowest granularity for analysis or querying of data in data warehouse may not be needed.

E.g.: *For a credit card company to analyze sales patterns, it may not be necessary to store every single transaction on each credit card. Instead, you may want to summarize the daily transaction for each credit card and store summary data instead of most granular data.*

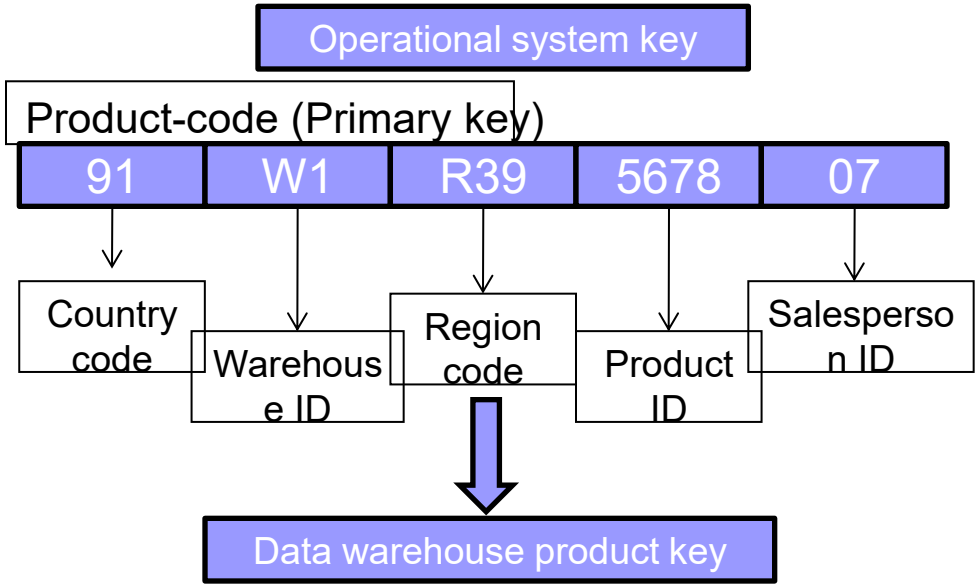
4. **Enrichment:** It is the rearranging and simplifying of individual fields to make them more useful in the ETL process.

Key Restructuring

Write Short note on Key Restructuring? May 2010

Transformation Types- *Key Restructuring*

- The Transformation of keys , which is supposed to be used as ***primary key into Generic key*** by the system is called as **Key Restructuring.**
- In this example the product-code has inherent meaning, Usage of this code as P.K can have adverse effects.
- If the Product data is moved to another D.W, the warehouse Id will have to be changed.
- This is the problem!!!
- When the keys are chosen for DW database tables avoid keys with built in meaning.
- Such keys are transformed into generic keys which are generated by the system.



Loading

Loading Dimensions

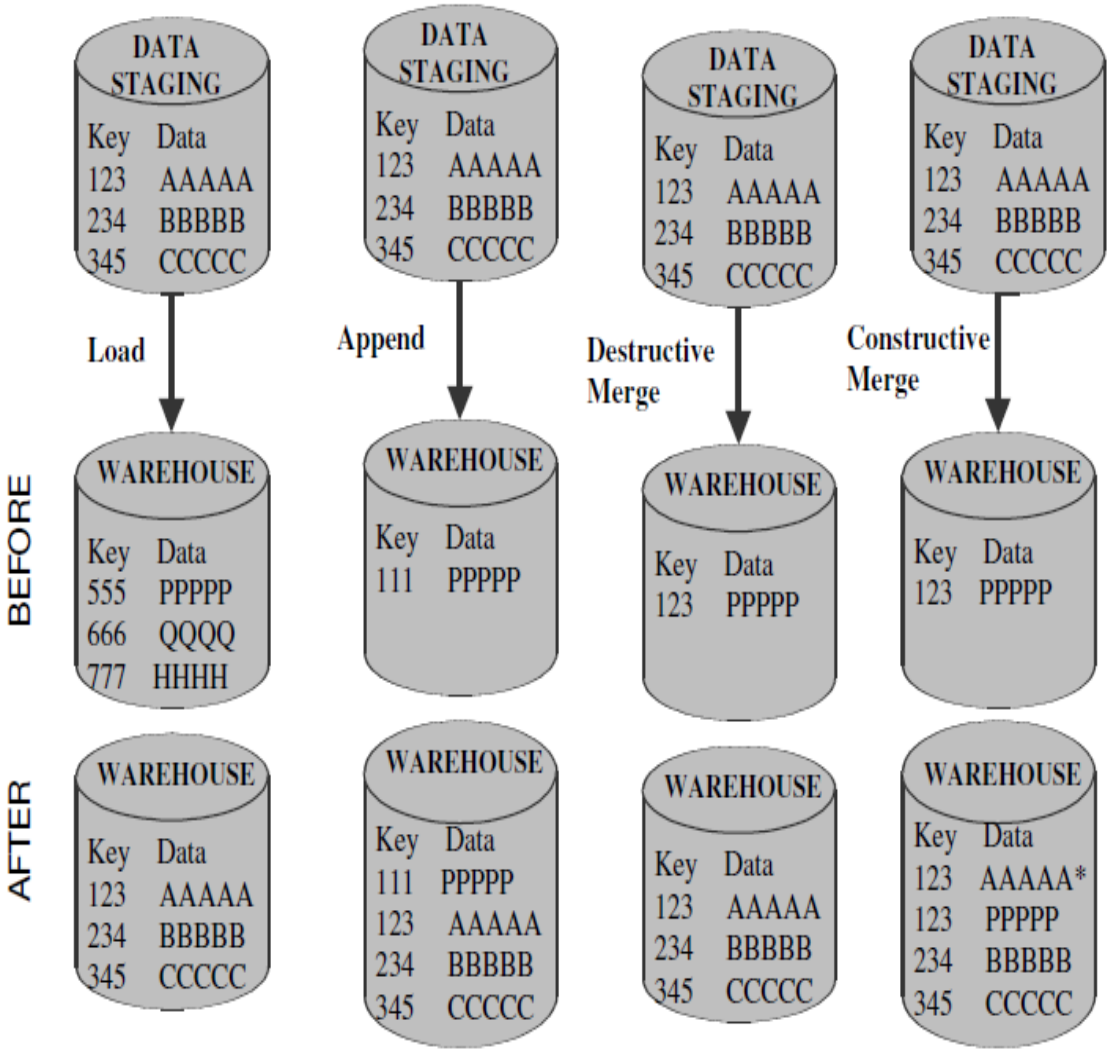
Loading Facts

Loading

- Loading process is the physical movement of the data from the ***source database(s) to the destination data warehouse***.
- **The whole process of moving data needs 3 Types of Loads:**
  1. **Initial Load:** For the very first time loading all the data warehouse tables.
     - After the initial load, the DW needs to be maintained and updated and this can be done by the following methods:
       a) **Incremental Load:** Periodically applying ongoing changes as per the requirement (inserted as new record/update the same).
       b) **Full Refresh:** Deleting the contents of a table and reloading it with fresh data.

Modes of Applying data to DW

1. Load
2. Append
3. Destructive merge
4. Constructive merge



Modes of Applying data to DW:

1. ***Load.*** If the target table to be loaded already exists and data exists in the table, the load process wipes out the existing data and applies the data from the incoming file.
   - If the table is already empty before loading, the load process simply applies the data from the incoming file.
2. ***Append.*** You may think of the append as an extension of the load. If data already exists in the table, the append process unconditionally adds the incoming data, preserving the existing data in the target table.
   - When an incoming record is a duplicate of an already existing record, you may define how to handle an incoming duplicate. The incoming record may be allowed to be added as a duplicate. In the other option, the incoming duplicate record may be rejected during the append process.

Modes of Applying data to DW:

3. ***Destructive Merge.*** In this mode, you apply the incoming data to the target data. If the primary key of an incoming record matches with the key of an existing record, update the matching target record.
   - If the incoming record is a new record without a match with any existing record, add the incoming record to the target table.
4. ***Constructive Merge.*** This mode is slightly different from the destructive merge. If the primary key of an incoming record matches with the key of an existing record, leave the existing record, add the incoming record, and mark the added record as superceding the old record.

Mapping Modes with Types of Load :

Sr. no	Types of Load	Modes of Applying Data in DW
		Further Run-Append Mode
3	Full Refresh	Load Mode followed by Append Mode

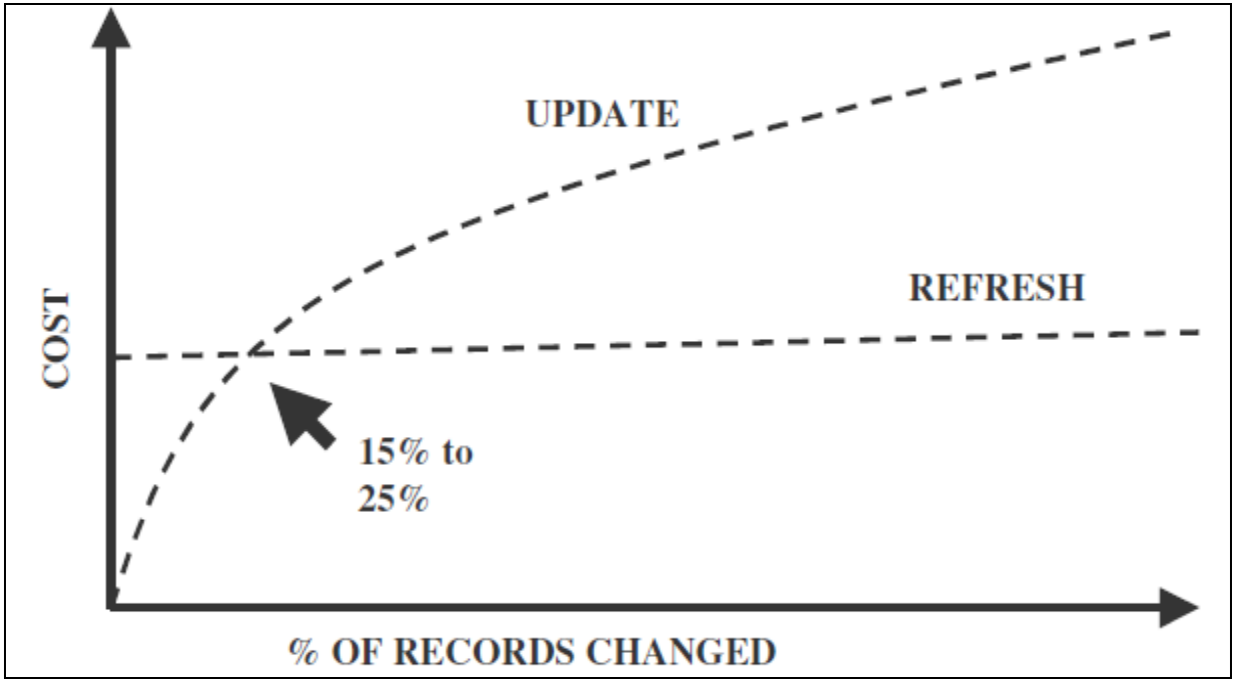
In case of **Slowly changing Dimension**: Destructive mode is used as the change to a dimension table record is meant to correct an error in the existing record. The existing record must be replaced by the corrected incoming record, so you may use the destructive merge mode.

Update/ Refresh?

- After the initial load, you may maintain the data warehouse and keep it up-to-date by using two methods:

1. **Update:** *Application of incremental changes*
2. **Refresh:** *Complete reload at specified intervals*

- Refresh is simpler to implement



Update/ Refresh?

- Technically, refresh is a much simpler option than update.
- To use the update option,
  - you have to devise the ***proper strategy to extract the changes*** from each data source.
  - Then you have to determine the ***best strategy to apply the changes*** to the data warehouse. The
- The refresh option simply involves the periodic replacement of complete data warehouse tables.
- But refresh jobs can take a long time to run.
  - If you have to run refresh jobs every day, you may have to keep the data warehouse down for unacceptably long times. The case worsens if your database has large tables.

University Questions

Dec 2016

In what way ETL style can be used in typical data warehouse, explain with suitable instance.

Dec 2017, May 2016

Explain steps in ETL Process?

May 2017, May 2018

Explain ETL of Data Warehousing in details?

THANK YOU