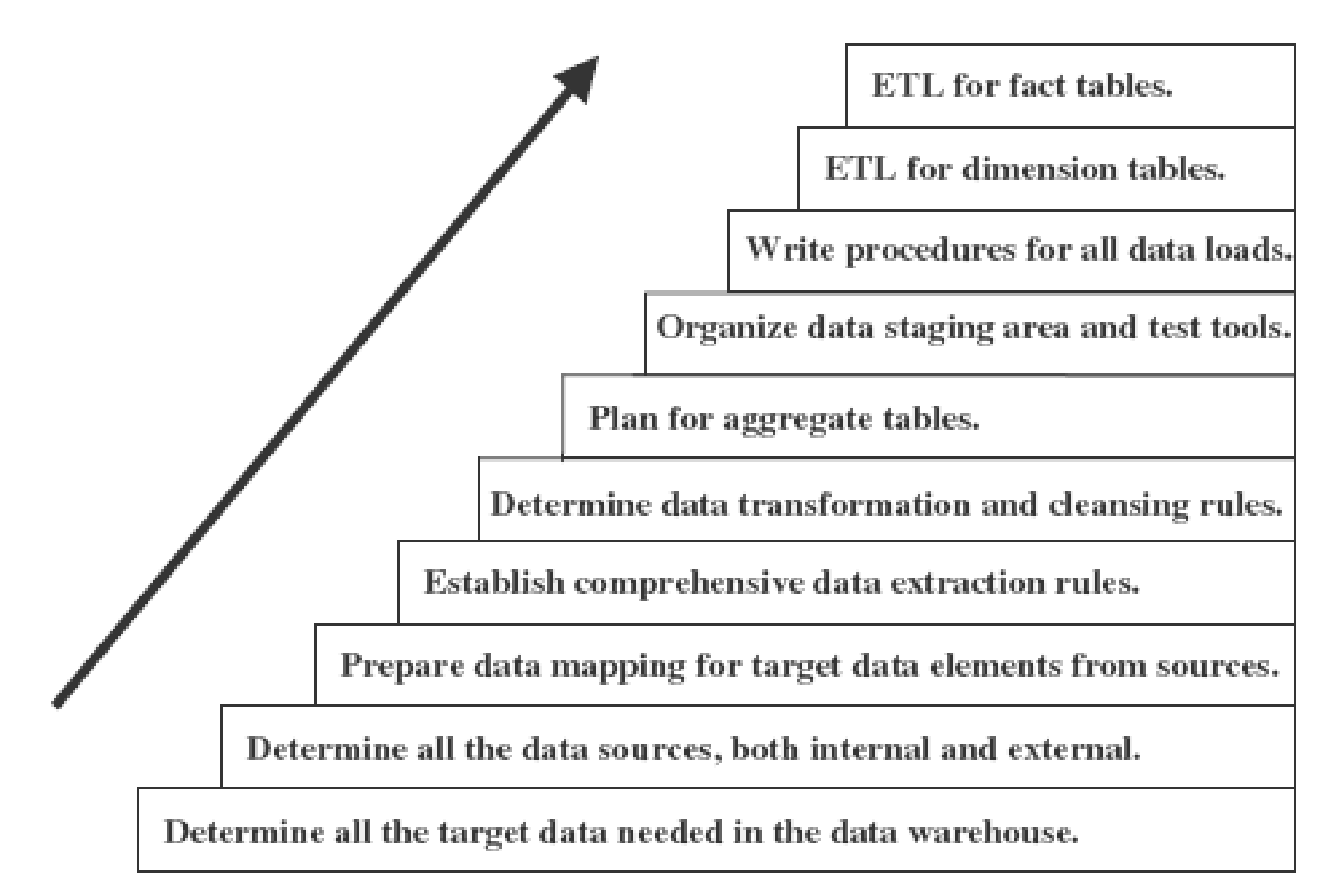
DATA WAREHOUSING
LECTURE 9

ENGR. MADEHA MUSHTAQ
DEPARTMENT OF COMPUTER SCIENCE
IQRA NATIONAL UNIVERSITY

EXTRACT TRANSFORM LOAD (ETL)

- Data extraction, transformation, and loading encompass the areas of data acquisition and data storage.
- When the project team designs the ETL functions, tests the various processes, and deploys them, we will find that these consume a very high percentage of the total project effort.
- It is not uncommon for a project team to spend as much as 50–70% of the project effort on ETL functions.

ETL REQUIREMENTS AND STEPS



ETL for fact tables.

ETL for dimension tables.

Write procedures for all data loads.

Organize data staging area and test tools.

Plan for aggregate tables.

Determine data transformation and cleansing rules.

Establish comprehensive data extraction rules.

Prepare data mapping for target data elements from sources.

Determine all the data sources, both internal and external.

Determine all the target data needed in the data warehouse.

DATA EXTRACTION

- Two major factors differentiate the data extraction for a new operational system from the data extraction for a data warehouse.
  - First, for a data warehouse, you have to extract data from many disparate sources.
  - Next, for a data warehouse, you have to extract data on the changes for ongoing incremental loads as well as for a one-time initial full load.

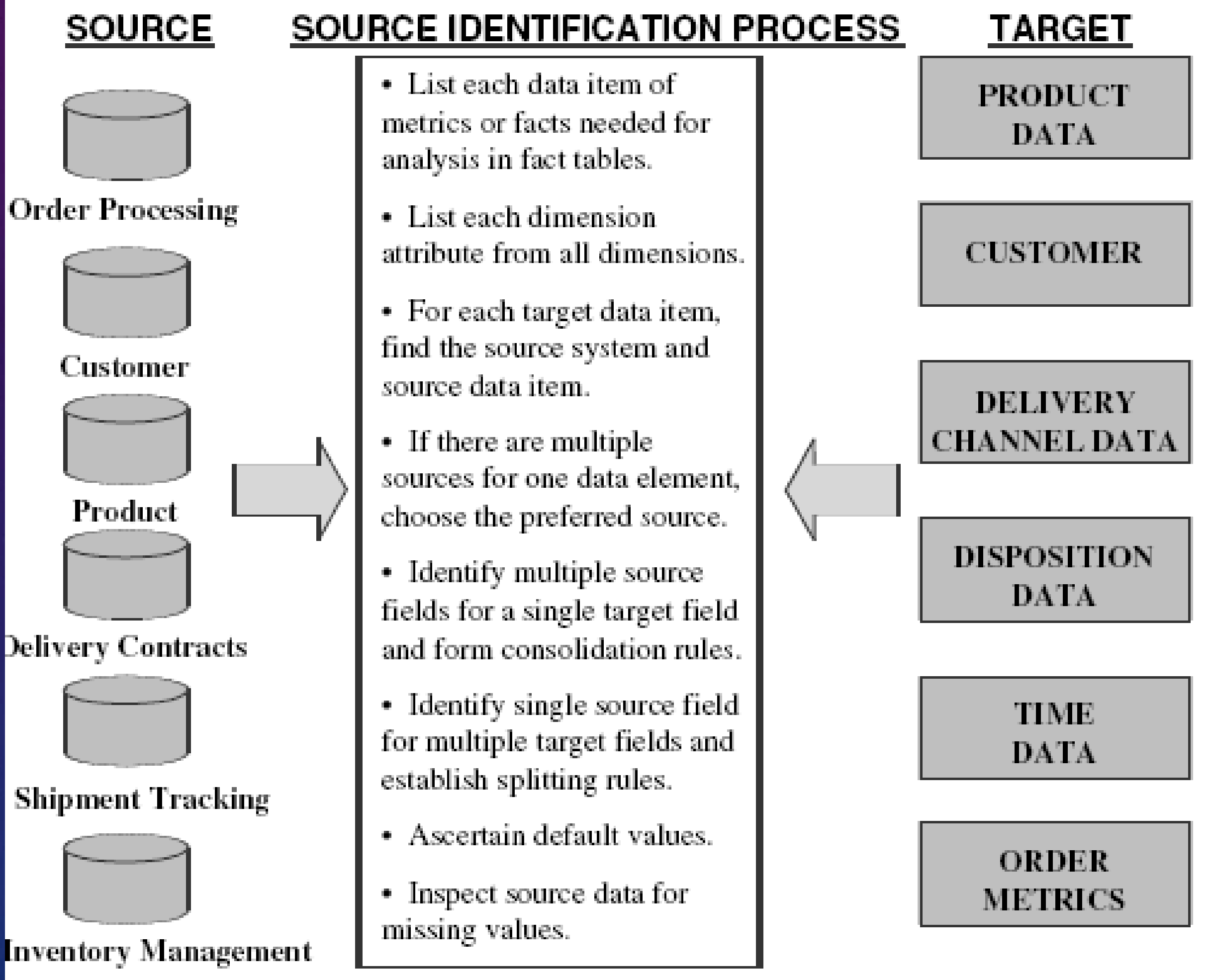
DATA EXTRACTION ISSUES

- Source Identification:
  - Identify source applications and source structures.
- Method of extraction:
  - For each data source, define whether the extraction process is manual or tool-based.
- Extraction frequency:
  - For each data source, establish how frequently the data extraction must by done—daily, weekly, quarterly, and so on.

DATA EXTRACTION ISSUES

- Time window
  - For each data source, denote the time window for the extraction process.
- Job sequencing:
  - Determine whether the beginning of one job in an extraction job stream has to wait until the previous job has finished successfully.
- Exception handling:
  - Determine how to handle input records that cannot be extracted.

SOURCE IDENTIFICATION PROCESS



SOURCE

- Order Processing
- Customer
- Product
- Delivery Contracts
- Shipment Tracking
- Inventory Management

SOURCE IDENTIFICATION PROCESS

- List each data item of metrics or facts needed for analysis in fact tables.
- List each dimension attribute from all dimensions.
- For each target data item, find the source system and source data item.
- If there are multiple sources for one data element, choose the preferred source.
- Identify multiple source fields for a single target field and form consolidation rules.
- Identify single source field for multiple target fields and establish splitting rules.
- Ascertain default values.
- Inspect source data for missing values.

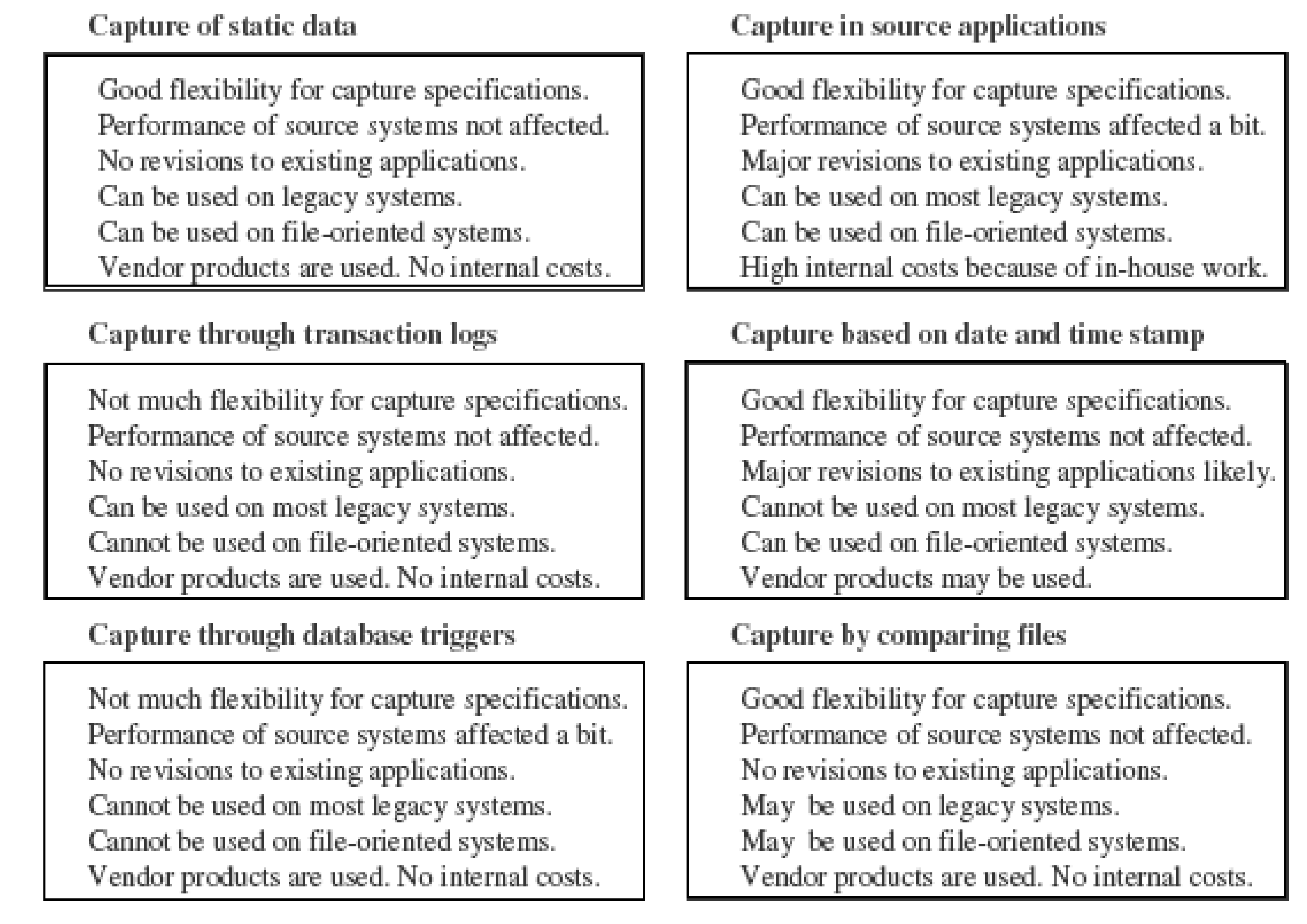
TARGET

- PRODUCT DATA
- CUSTOMER
- DELIVERY CHANNEL DATA
- DISPOSITION DATA
- TIME DATA
- ORDER METRICS

DATA EXTRACTION TECHNIQUES

- The following options are available for data extraction:
  - Capture of static data
  - Capture through transaction logs
  - Capture through database triggers
  - Capture in source applications
  - Capture based on date and time stamp
  - Capture by comparing files

DATA EXTRACTION TECHNIQUES



Capture of static data

Good flexibility for capture specifications.
Performance of source systems not affected.
No revisions to existing applications.
Can be used on legacy systems.
Can be used on file-oriented systems.
Vendor products are used. No internal costs.

Capture in source applications

Good flexibility for capture specifications.
Performance of source systems affected a bit.
Major revisions to existing applications.
Can be used on most legacy systems.
Can be used on file-oriented systems.
High internal costs because of in-house work.

Capture through transaction logs

Not much flexibility for capture specifications.
Performance of source systems not affected.
No revisions to existing applications.
Can be used on most legacy systems.
Cannot be used on file-oriented systems.
Vendor products are used. No internal costs.

Capture based on date and time stamp

Good flexibility for capture specifications.
Performance of source systems not affected.
Major revisions to existing applications likely.
Cannot be used on most legacy systems.
Can be used on file-oriented systems.
Vendor products may be used.

Capture through database triggers

Not much flexibility for capture specifications.
Performance of source systems affected a bit.
No revisions to existing applications.
Cannot be used on most legacy systems.
Cannot be used on file-oriented systems.
Vendor products are used. No internal costs.

Capture by comparing files

Good flexibility for capture specifications.
Performance of source systems not affected.
No revisions to existing applications.
May be used on legacy systems.
May be used on file-oriented systems.
Vendor products are used. No internal costs.

DATA TRANSFORMATION

- Here is the set of basic tasks for data transformation:
- Selection:
  - This takes place at the beginning of the whole process of data transformation.
  - You select either whole records or parts of several records from the source systems.

DATA TRANSFORMATION

- Splitting/joining:
  - This task includes the types of data manipulation you need to perform on the selected parts of source records.
  - Sometimes (uncommonly), you will be splitting the selected parts even further during data transformation.
- Conversion:
  - This is an all-inclusive task.
  - It includes a large variety of rudimentary conversions of single fields for two primary reasons—one to standardize among the data extractions from disparate source systems, and the other to make the fields usable and understandable to the users.

DATA TRANSFORMATION

- Summarization:
  - Sometimes you may find that it is not feasible to keep data at the lowest level of detail in your data warehouse.
  - So the data transformation function includes summarization of daily sales by product and by store.
- Enrichment:
  - This task is the rearrangement and simplification of individual fields to make them more useful for the data warehouse environment.

USING MANUAL TECHNIQUES

- This was the predominant method until recently when transformation tools began to appear in the market.
- Manual techniques are still in use for smaller data warehouses.
- Here manually coded programs and scripts perform every data transformation.
- A major disadvantage relates to metadata.
- Automated tools record their own metadata, but in-house programs have to be designed differently if you need to store and use metadata.

USING TRANSFORMATION TOOLS

- Use of automated tools improves efficiency and accuracy.
- A major advantage from using a transformation tool is the recording of metadata by the tool.
- When we specify the transformation parameters and rules, these are stored as metadata by the tool.
- This metadata then becomes part of the overall metadata component of the data warehouse.

DATA LOADING

- The whole process of moving data into the data warehouse repository is referred to as data loading.
  - Initial Load: populating all the data warehouse tables for the very first time
  - Incremental Load: applying ongoing changes as necessary in a periodic manner
  - Full Refresh: completely erasing the contents of one or more tables and reloading with fresh data (initial load is a refresh of all the tables).

DATA LOADING TECHNIQUES AND PROCESSES

- There are four modes of loading data:
- 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 empty before loading, the load process simply applies the data from the incoming file.
- Append:
  - If data already exists in the table, the append process unconditionally adds the incoming data, preserving the existing data in the target table.

DATA LOADING TECHNIQUES AND PROCESSES

- 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.
- Constructive 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 superseding the old record.

TYPES OF LOADS

- There are three types of loads:
- Initial Load:
  - Loading the whole data warehouse in a single run.
  - As a variation of this single run, let us say you are able to split the load into separate subloads and run each of these subloads as single loads.
  - In other words, every load run creates the database tables from scratch.
  - In these cases, you will be using the load mode discussed above.

TYPES OF LOADS

- Incremental Loads:
  - These are the applications of ongoing changes from the source systems.
  - We need a method to preserve the periodic nature of the changes in the data warehouse.
  - The constructive merge mode is an appropriate method for incremental loads.

TYPES OF LOADS

- Full Refresh:
  - This type of application of data involves periodically rewriting the entire data warehouse.
  - In the case of full refreshes, data exists in the target tables before incoming data is applied.
  - The existing data must be erased before applying the incoming data.
  - Just as in the case of the initial load, the load and append modes are applicable to full refresh.

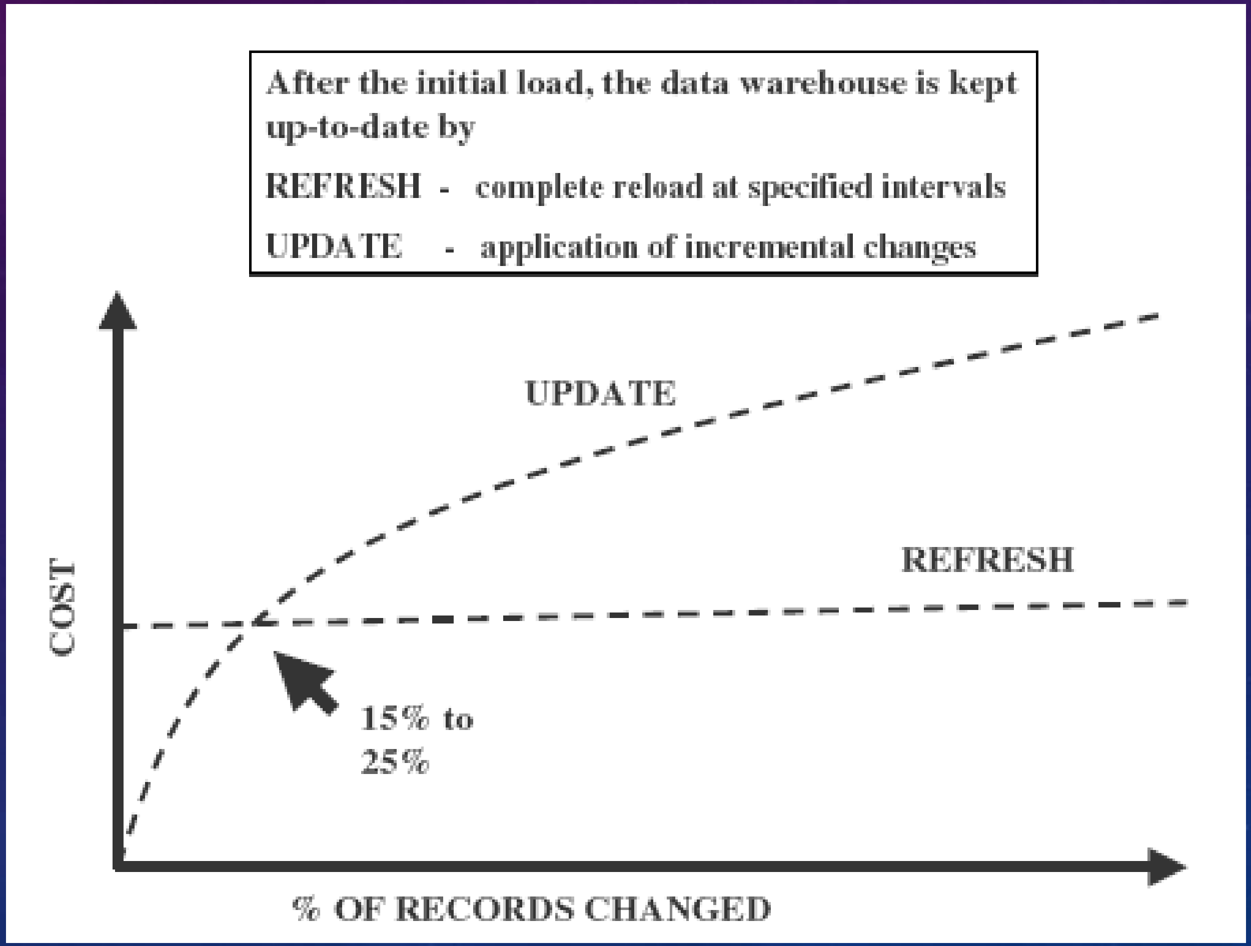
DATA REFRESH VERSUS UPDATE

- After the initial load, we may maintain the data warehouse and keep it up-to-date by using two methods:
- Update:
  - Application of incremental changes in the data sources.
  - To use the update option, we have to devise the proper strategy to extract the changes from each data source.
  - Then we have to determine the best strategy to apply the changes to the data warehouse.

DATA REFRESH VERSUS UPDATE

- Refresh:
  - Complete reload at specified intervals
  - Technically, refresh is a much simpler option than update
  - 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.

DATA REFRESH VERSUS UPDATE



After the initial load, the data warehouse is kept up-to-date by

REFRESH - complete reload at specified intervals

UPDATE - application of incremental changes

PROCEDURE FOR DIMENSION TABLES

- The procedure for maintaining the dimension tables includes two functions:
- First, the initial loading of the tables
- Thereafter, applying the changes on an ongoing basis.
- The issue with dimension tables is about the keys of the records in the source systems and the keys of the records in the data warehouse.
- We do not use the production system keys for the records in the data warehouse.
- In the data warehouse, you use system generated keys.

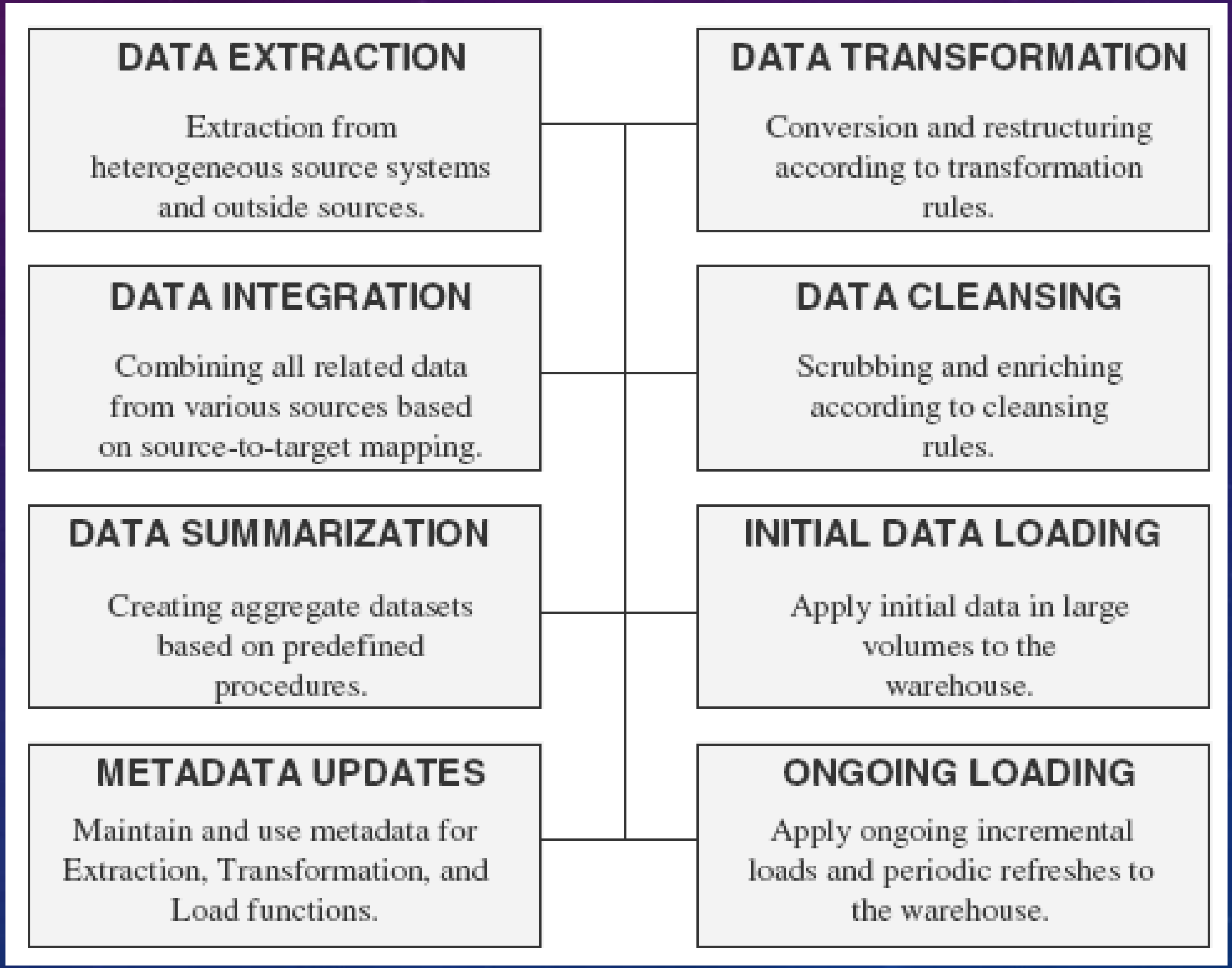
PROCEDURE FOR DIMENSION TABLES

- Before source data can be applied to the dimension tables, whether for the initial load or for ongoing changes, the production keys must be converted to the system-generated keys in the data warehouse.
- Key conversion may be done as part of the transformation functions or
- We may do it separately before the actual load functions.

PROCEDURE FOR FACT TABLES

- The key of the fact table is the concatenation of the keys of the dimension tables.
- Therefore, dimension records are loaded first.
- Then, before loading each fact table record, you have to create the concatenated key for the fact table record from the keys of the corresponding dimension records.

ETL SUMMARY



DATA EXTRACTION

Extraction from heterogeneous source systems and outside sources.

DATA TRANSFORMATION

Conversion and restructuring according to transformation rules.

DATA INTEGRATION

Combining all related data from various sources based on source-to-target mapping.

DATA CLEANSING

Scrubbing and enriching according to cleansing rules.

DATA SUMMARIZATION

Creating aggregate datasets based on predefined procedures.

INITIAL DATA LOADING

Apply initial data in large volumes to the warehouse.

METADATA UPDATES

Maintain and use metadata for Extraction, Transformation, and Load functions.

ONGOING LOADING

Apply ongoing incremental loads and periodic refreshes to the warehouse.

END OF SLIDES