Subject Name: DATA WAREHOUSING AND MINING

Ms.Puja S Vakhare

(Assistant Professor)

INDEX

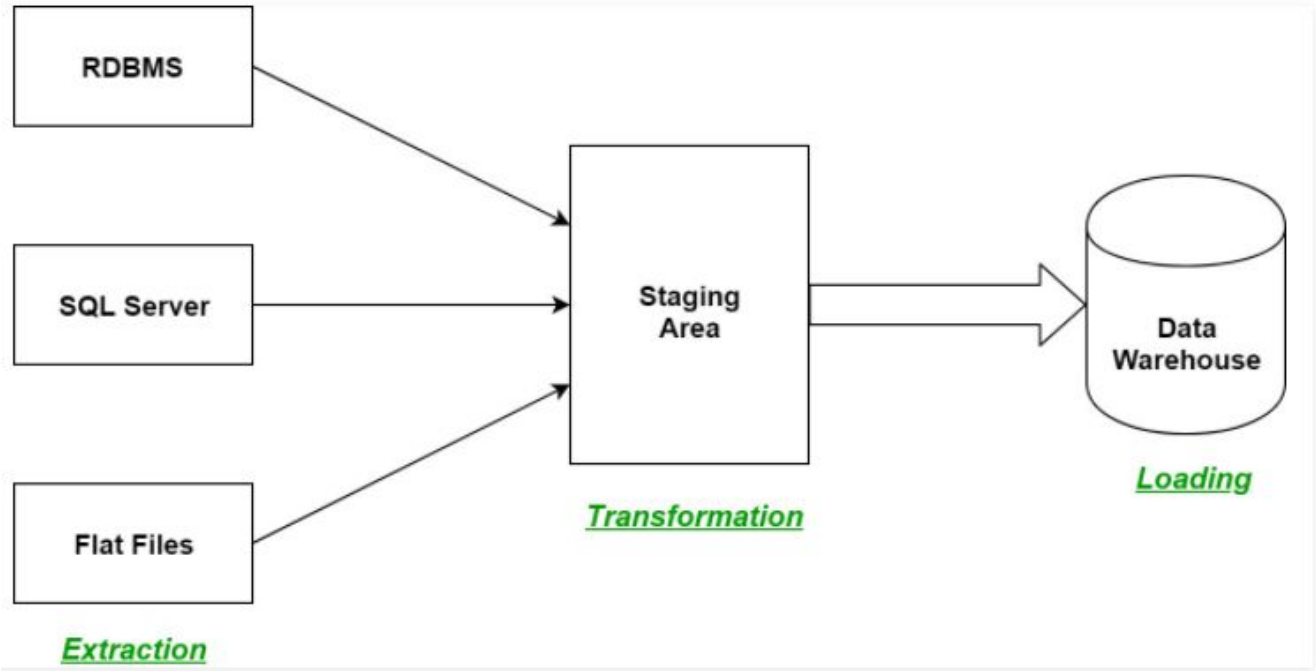
Lecture 10: ETL : Steps in ETL Process

Lecture 11: Data Extraction ;Task involved in Data Transformation.

Lecture 12 : Techniques of Data Loading

ETL (Extract, Transform, and Load) Process

It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area and then finally, loads it into the Data Warehouse system.



ETL - Extraction

- The first step of the ETL process is extraction.
- Data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML and flat files into the staging area.
- It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also.
- Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult.
- Therefore, this is one of the most important steps of ETL process.

ETL - Transformation

- The second step of the ETL process is transformation.

- In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format.

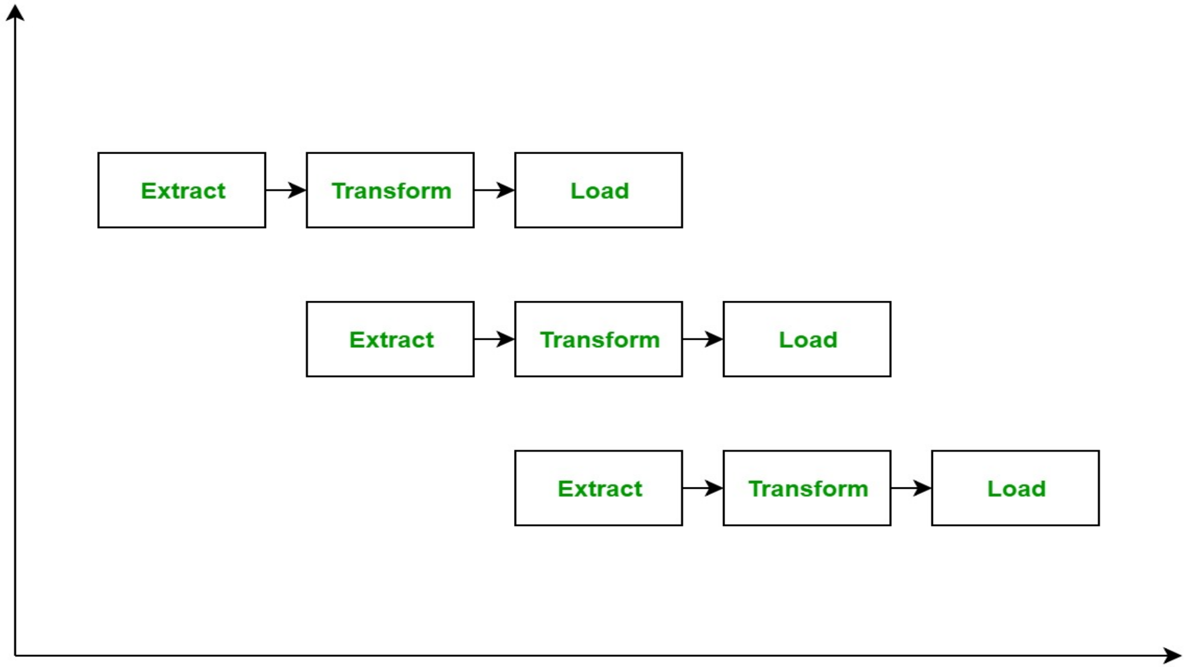
- It may involve following processes/tasks:
  - Filtering – loading only certain attributes into the data warehouse.
  - Cleaning – filling up the NULL values with some default values
  - Joining – joining multiple attributes into one.
  - Splitting – splitting a single attribute into multiple attributes.
  - Sorting – sorting tuples on the basis of some attribute (generally key-attribute).

ETL - Loading

- The third and final step of the ETL process is loading.
- In this step, the transformed data is finally loaded into the data warehouse.
- Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals.
- The rate and period of loading solely depends on the requirements and varies from system to system

ETL - Pipelining

- ETL process can also use the pipelining concept

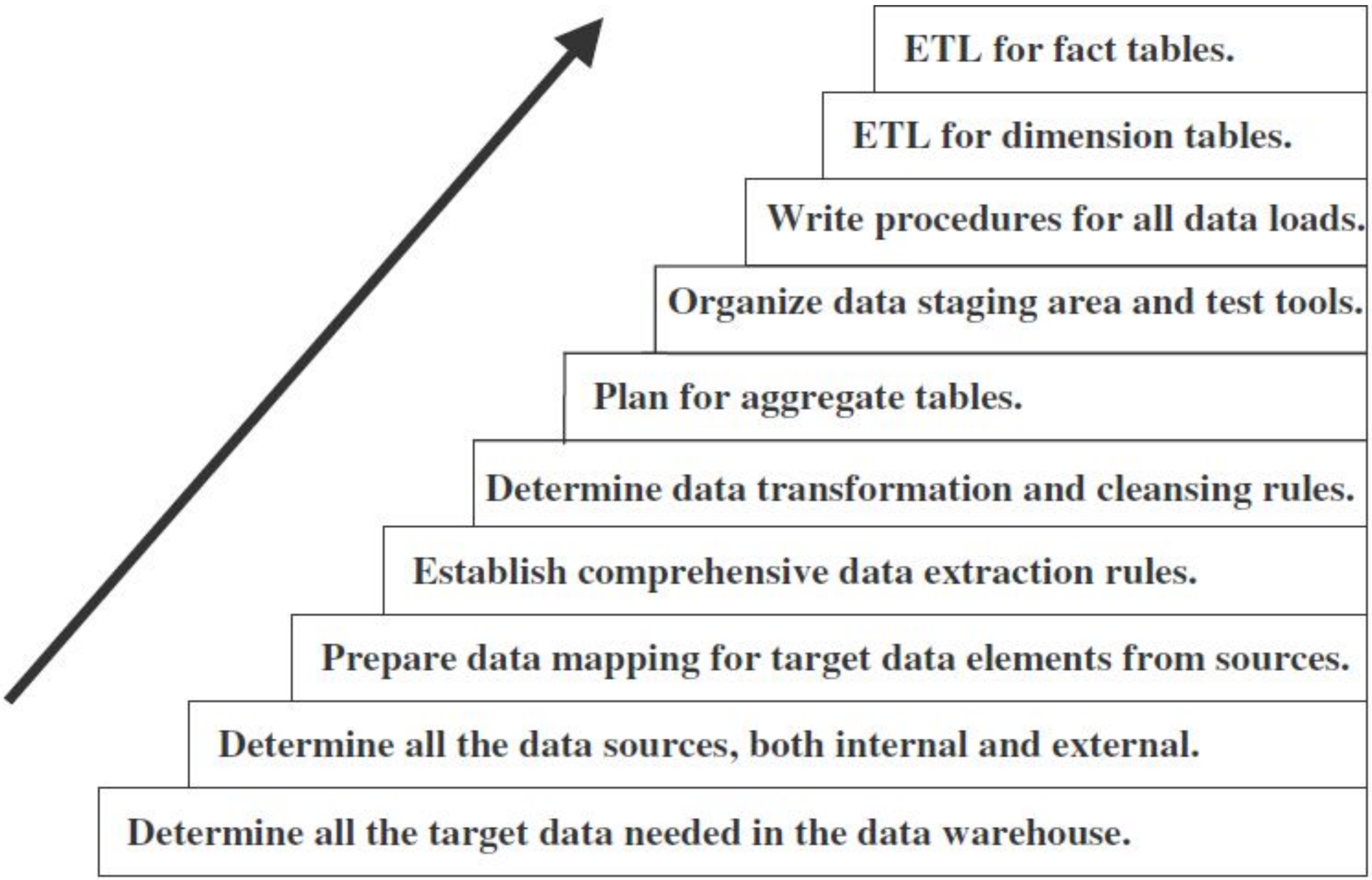


- **ETL Tools:** Most commonly used ETL tools are Sybase, Oracle Warehouse builder, CloverETL and MarkLogic.

Challenges in ETL Functions

- Source systems are very diverse and disparate
- A need to deal with sources systems on multiple platforms and different operating systems
- Source systems are running on old obsolete technologies
- Historical information is important, however source operation system do not maintain them
- Quality of data is uncertain in many old systems
- Source system structures keep changing over time because of new business conditions – ETL functions had to be modified accordingly
- Lack of consistency can be seen among source systems
- Most source systems do not represent data in format meaningful to the users

Major Steps in ETL Process



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 factors increase** the **complexity** of **data extraction**
  - extract data from **many disparate sources**
  - extract data on the **changes for ongoing incremental loads** as well as for a one-time initial full load

- Effective data extraction is a key to the success of your data warehouse

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.
- **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.

Data Extraction Methods

Logical Extraction Methods

There are two kinds of logical extraction:

- Full Extraction
- Incremental Extraction

Physical Extraction Methods

- Online Extraction
- Offline Extraction

Data Extraction Methods – Full Extraction

- The data is extracted completely from the source system.
- It reflects all the data currently available on the source system, there’s no need to keep track of changes to the data source since the last successful extraction.
- The source data will be provided as-is and no additional logical information (for example, timestamps) is necessary on the source site.
- An example for a full extraction may be an export file of a distinct table or a remote SQL statement scanning the complete source table.

INCREMENTAL EXTRACTION

- At a specific point in time, only the data that has changed since a well-defined event back in history will be extracted.
- This event may be the last time of extraction or a more complex business event like the last booking day of a fiscal period.

PHYSICAL EXTRACTION METHODS

Depending on the chosen logical extraction method and the capabilities and restrictions on the source side, the extracted data can be physically extracted by two mechanisms.

Online Extraction

- The data is extracted directly from the source system itself.
- The extraction process can connect directly to the source system to access the source tables themselves or to an intermediate system that stores the data in a preconfigured manner (for example, snapshot logs or change tables)

PHYSICAL EXTRACTION METHODS

Offline Extraction

- The data is not extracted directly from the source system but is staged explicitly outside the original source system.

- The data already has an existing structure (for example, redo logs, archive logs or transportable tablespaces) or was created by an extraction routine.

Data Transformation

- **Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed**.

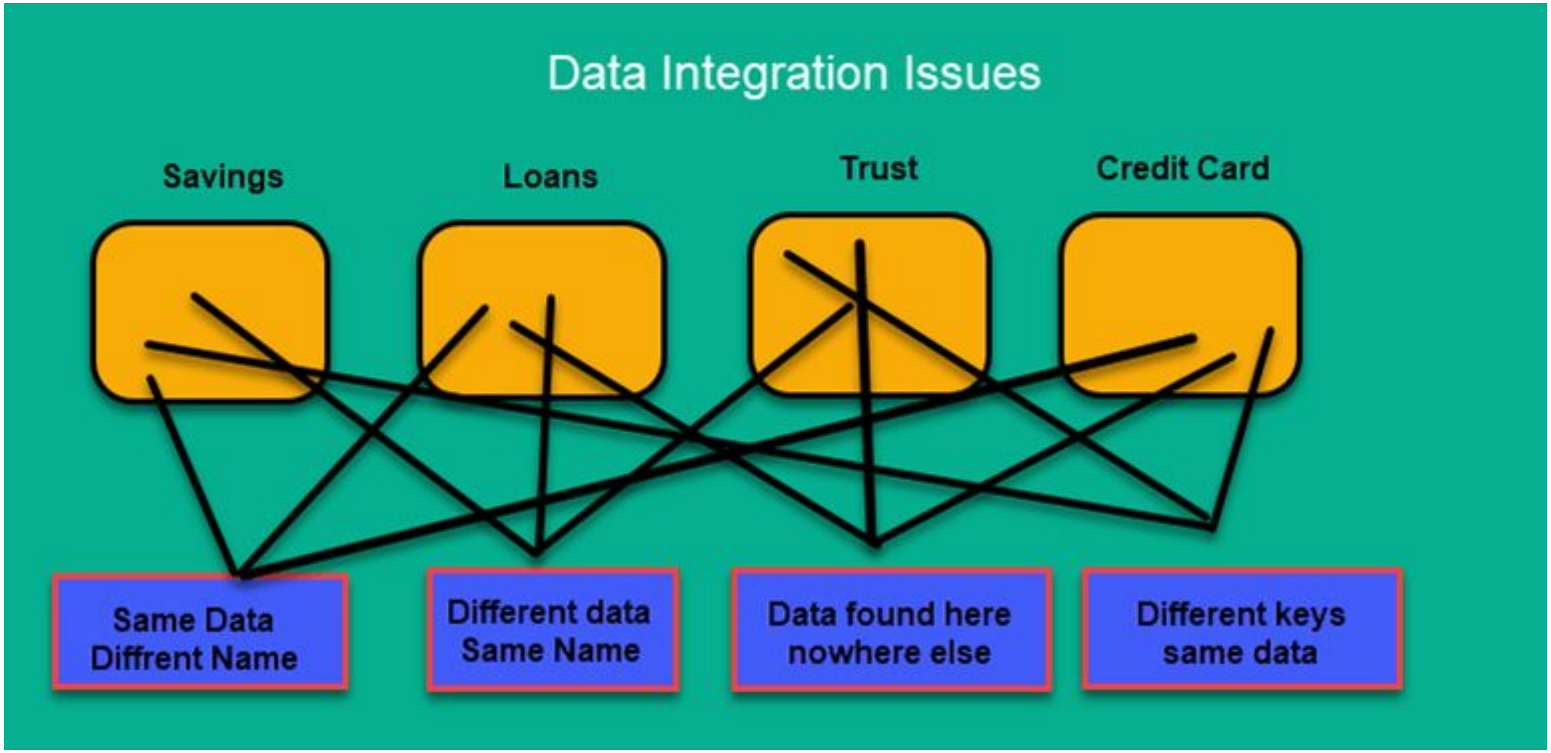
- It is one of the important ETL concepts where you apply a set of functions on extracted data.

- Analyzing information requires structured and accessible data for best results. Data transformation enables organizations to alter the structure and format of raw data as needed.

- Data transformation is the process of changing the format, structure, or values of data. For data analytics projects, data may be transformed at two stages of the data pipeline.

DATA TRANSFORMATION – DATA INTEGRATION ISSUE

In transformation step, we can perform customized operations on data. For example, if the user wants sum-of-sales revenue which is not in the database or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading.



DATA TRANSFORMATION – DATA INTEGRATION ISSUE

Following are Data Integrity Problems:

1. Different spelling of the same person like Jon, John, etc.
2. There are multiple ways to denote company name like Google, Google Inc.
3. Use of different names like Cleaveland, Cleveland.
4. There may be a case that different account numbers are generated by various

   applications for the same customer.

1. In some data required files remains blank
2. Invalid product collected at POS as manual entry can lead to mistakes.

DATA TRANSFORMATION

Validations are done during this stage-

- Filtering – Select only certain columns to load
- Using rules and lookup tables for Data standardization
- Character Set Conversion and encoding handling
- Conversion of Units of Measurements like Date Time Conversion, currency conversions, numerical conversions, etc.
- Data threshold validation check. **For example, age cannot be more than two digits**.
- Data flow validation from the staging area to the intermediate tables.
- Required fields should not be left blank.
- **Cleaning (for example, mapping NULL to 0 or Gender Male to "M" and Female to "F" etc.)**
- Split a column into multiples and merging multiple columns into a single column.
- Transposing rows and columns,
- Use lookups to merge data
- Using any complex data validation (e.g., if the first two columns in a row are empty then it automatically reject the row from processing)

HOW TO TRANSFORM DATA

- Data transformation can increase the efficiency of analytic and business processes and enable better data-driven decision-making.
- The first phase of data transformations should include things like data type conversion and flattening of hierarchical data.
- These operations shape data to increase compatibility with analytics systems.
- Data analysts and data scientists can implement further transformations additively as necessary as individual layers of processing.
- Each layer of processing should be designed to perform a specific set of tasks that meet a known business or technical requirement.

WHAT IS LOADING?

- The extracted data and the transformed data are loaded into the target database.
- All three steps in the ETL process can be run parallel. Data extraction takes time and therefore the second phase of the transformation process is executed simultaneously.
- This prepared the data for the third stage of loading. As soon as some data is ready, it is loaded without waiting for the previous steps to be completed.

TYPES OF LOADING

1. **Initial Load**: For the very first time loading all the data warehouse tables.
2. **Incremental Load**: Periodically applying ongoing changes as per the requirement. After the data is loaded into the data warehouse database, verify the referential integrity between the dimensions and the fact tables to ensure that all records belong to the appropriate records in the other tables. The DBA must verify that each record in the fact table is related to one record in each dimension table that will be used in combination with that fact table.
3. **Full Refresh**: Deleting the contents of a table and reloading it with fresh data.

Refresh Vs Update

Update– application of incremental changes in the data sources.
Refresh– complete reloads at specified intervals.

METHODS FOR DATA LOADING

- **Cloud-based:** ETL solutions in the cloud are frequently able to process data in real-time and are designed for speed and scalability.
- **Batch processing:** Data is moved every day or every week via ETL systems that use batch processing. Large data sets and organizations that don't necessarily require real-time access to their data are the greatest candidates for it.
- **Open-source:** Since their codebases are shared, editable, and publicly available, many open-source ETL systems are extremely affordable. Despite being a decent substitute for commercial solutions, many tools may still need some hand-coding or customization.

ETL TOOLS

- Skyvia
- IRI Voracity
- Xtract.io
- **Sprinkle**
- DBConvert Studio By SLOTIX s.r.o.
- **Informatica – PowerCenter**
- **IBM – Infosphere Information Server**
- **Oracle Data Integrator**
- **Microsoft – SQL Server Integrated Services (SSIS)**
- Ab Initio

WHY DATA MINING

world is data rich but information poor.

- Data mining is the process of searching and analyzing a large batch of raw data in order to **identify patterns and extract useful information**. Companies use data mining software to learn more about their customers. It can help them to **develop more effective marketing strategies, increase sales, and decrease costs**.
- Data mining is used to explore large data volumes to find patterns and insights that can be used for specific purposes. These purposes might include improving sales and marketing, optimizing manufacturing, detecting fraud, and enhancing security.

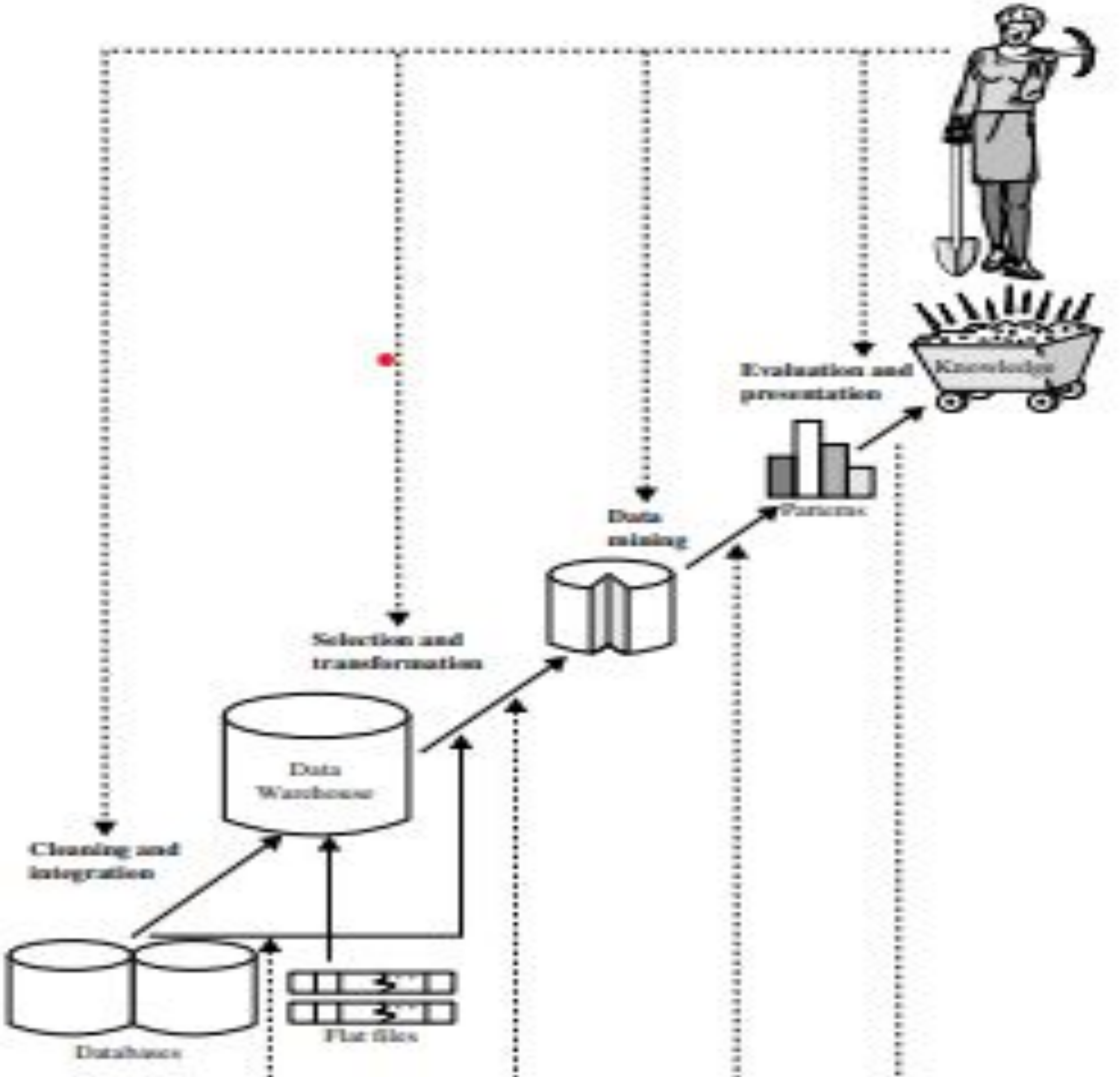
WHAT IS DATA MINING

Data mining—searching for knowledge (interesting patterns) in data.

- Data mining is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information.
- Companies use data mining software to learn more about their customers.
- It can help them to develop more effective marketing strategies, increase sales, and decrease costs.

Data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD.

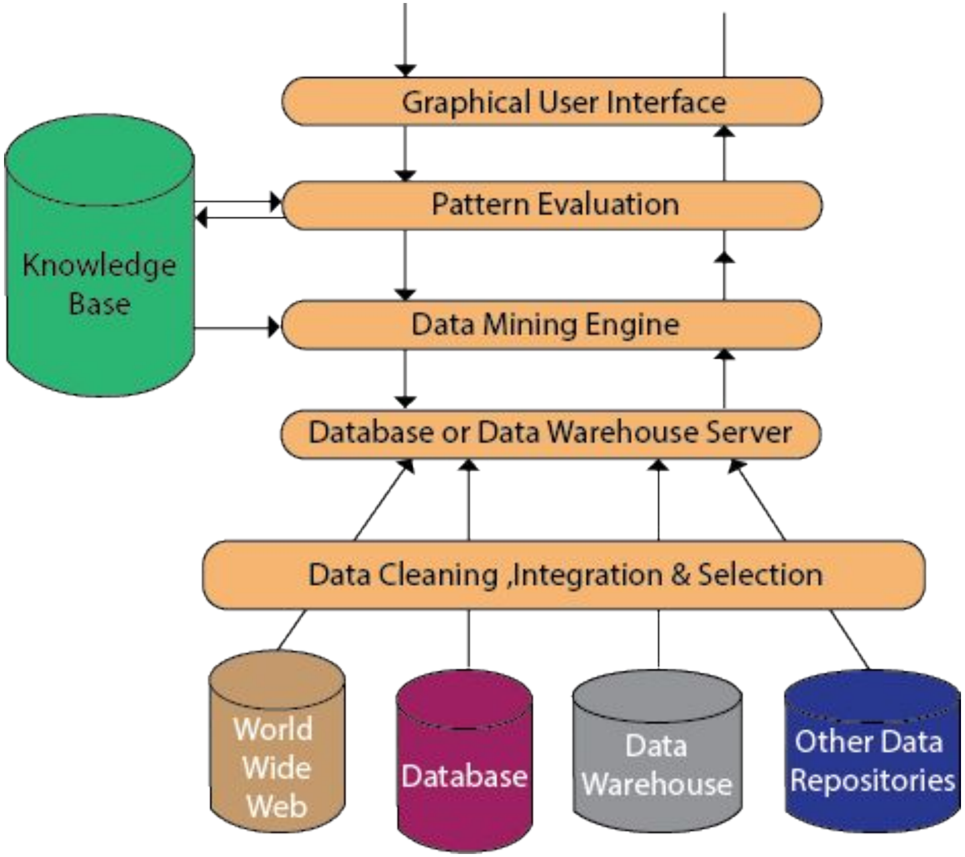
KDD PROCESS : ITERATIVE SEQUENCE OF THE FOLLOWING STEPS:



STEPS OF KDD

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

DATA MINING ARCHITECTURE



Thank You