Chapter 6 Basics of Data Integration

Content of this presentation has been taken from Book

“Fundamentals of Business Analytics”

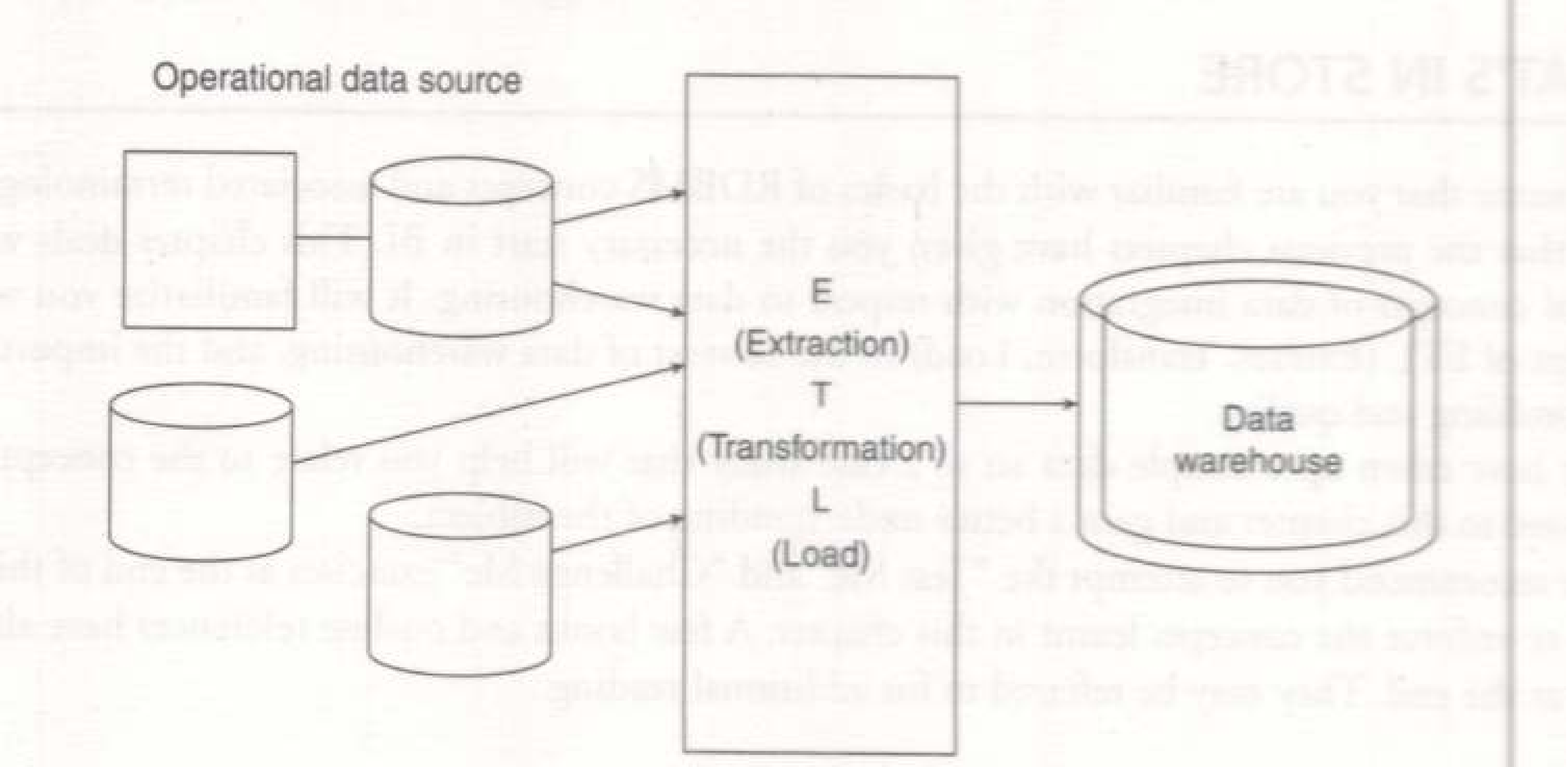
RN Prasad and Seema Acharya

Published by Wiley India Pvt. Ltd.

Need for Data Warehouse

- The Darbury Institute of Information Technology (DIIT) is an engineering institution that conducts engineering courses in Information Technology (IT), Computer Science (CS), System Engineering (SE), Information Science (IS), etc.
- Each department (IT, CS, SE, IS, etc.) has an automaetd library that meticulously handles library transactions and has good learning content in the form of DVDs, magazines, journals, several online references, etc.
- DIIT is also looking at expansion to have its branches in all major cities.
- The only downside of the library data is that it is stored differently by different departments. One department stores it in MS Excel spreadsheets, another stores it in MS Access database, and yét another department maintains a .CSV (Comma Separated Values) file. The DIIT administration is in need of report that indicates the annual spending on library purchases. The report should further drill down to the spending by each department by category (books, CDs/DVDs, magazines, journals, etc.). However, preparing such a report is not easy because of different data formats used by different departments. Prof. Frank (an expert on database technology) was called upon to suggest a possible solution to the problem at hand. He feels it would be better to start archiving the data in a data warehouse/data mart.
- The arguments put forth by him in favor of a library data warehouse are
    - Data from several heterogenous data sources (MS Excel spreadsheets, MS Access CSVfile, etc.) can be extracted and brought together in a data warehouse.
    - Even when DIIT expands into several branches in multiple cities, it still can have one ware-house to support the information needs of the institution.
    - Data anomalies can be corrected through an ETL package.
    - Missing or incomplete records can be detected and duly corrected.
    - Uniformity can be maintained over each attribute of a table.
    - Data can be conveniently retrieved for analysis and generating reports (like the report on spending requested above).
    - Fact-based decision making can be easily supported by a data warehouse.
    - Ad hoc queries can be easily supported.

Data from several heterogeneous data sources extracted and loaded in a data warehouse



When to shift to Data Warehouse solution when organization is facing following problems-

1. Lack of Information Sharing
2. Lack of information credibility
3. Reports take a longer time to be prepared
4. Little or no scope for ad hoc querying or queries that require historical data.

Data Warehouse

According to William H. Inmon, “A data warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management’s decision making process.”

Subject-oriented: A data warehouse collects data of subjects such as “customers”, “suppliers”, “partners”, “sales”, “products”, etc. spread across the enterprise or organization. A data mart on the other hand deals with the analysis of a particular subject such as “sales”.

Integrated: A typical enterprise will have a multitude of enterprise applications. It is likely that these applications are on heterogeneous technology platforms. It is also not unlikely that these applications use varied databases to house their data. Few of the applications may exist in silos. Few others may be sharing a little information between them. A data warehouse serve to bring together the data from these multiple disparate (meaning differing in the format and content of data) sources after careful cleansing and transformation into a unified format to serve the information needs of the enterprise.

Data Warehouse

According to William H. Inmon, "A data warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process."

Time-variant: A data warehouse keeps historical data while an OLTP system will usually have the most up-to-date data. From a data warehouse, one can retrieve data that is 3 months, 6 months, 12 months, or even older. For example, a system may hold the most recent address of a customer, whereas a data warehouse addresses associated with a customer recorded, say, over the last five years.

Non-volatile: We have learnt earlier that transaction processing, recovery, and concurrency control mechanisms are usually associated with OLTP systems. A data warehouse is a separate physical store of data transformed from the application data found in the operational environment.

Data Mart

The “GoodsForAll” enterprise has successfully implemented an enterprise-wide data warehouse, This data warehouse has data collected for all the customers and sales transactions from every unit/division and subsidiary in the business. The data warehouse true to its nature provides a homogenized, unified, and integrated view of information. It has proved very useful to the “GoodsForAll” enterprise.

The market research wing of the “GoodsForAll” enterprise wishes to access the data in the data warehouse. They have plans to execute predictive analytics application on the data stored in the data warehouse and look at how the analysis can help provide better business gains.

The data architect of the “GoodsForAll” enterprise has decided to create a data mart for the market research unit.

Data Mart

- A data mart is meant to provide single domain data aggregation that can then be used for analysis, reporting, and/or decision support.
- Data marts can be sourced from the enterprise-wide data warehouse or can also be sourced directly from the operational/transactional systems.
- These data marts can also perform transformations and calculations on the data housed within.
- When compared to the data warehouse, data marts are restricted in their scope and business purpose.
- Is it a good idea to go for a data mart for virtually every business process/event? The answer is “No’.
- This could result in several disparate and independent data marts. Chances are that it will become a challenge to ensure the single version of truth.

Operational Data Store (ODS)

An “operational data store” (ODS) is similar to a data warehouse in that several systems around feed operational information to it. The ODS processes this operational data to provide a unified view which can then be utilized by analysts and report-writers alike for analysis and reporting.

An ODS differs from enterprise data warehouse in that it does not store and maintain vast amounts of historical information. An ODS is meant to hold current or very recent operational data.

Why Operational Data Store is required?

Sometimes it is required to perform an instant analysis on the more recent data to allow one to respond immediately to a given situation. There are cases where some enterprises use the ODS as a staging area for the data warehouse. This would mean that the integration logic and processes are built into the ODS. On a regular basis, the data warehouse takes the current processed data from the ODS and adds it to its own historical data.

Goals of a Data Warehouse

The prime goal of a data warehouse is to enable users' appropriate access to a homogenized and comprehensive view of the organization. This in turn will support the forecasting and decision-making at the enterprise level.

Information accessibility: Data in a data warehouse must be easy to comprehend, both by the business users and developers alike. It should be properly labelled to facilitate easy access. The business users should be allowed to slice and dice the data in every possible way (slicing and dicing refers to the separation and combination of data in infinite combinations).

Information credibility: The data in the data warehouse should be credible, complete, and of desired quality.

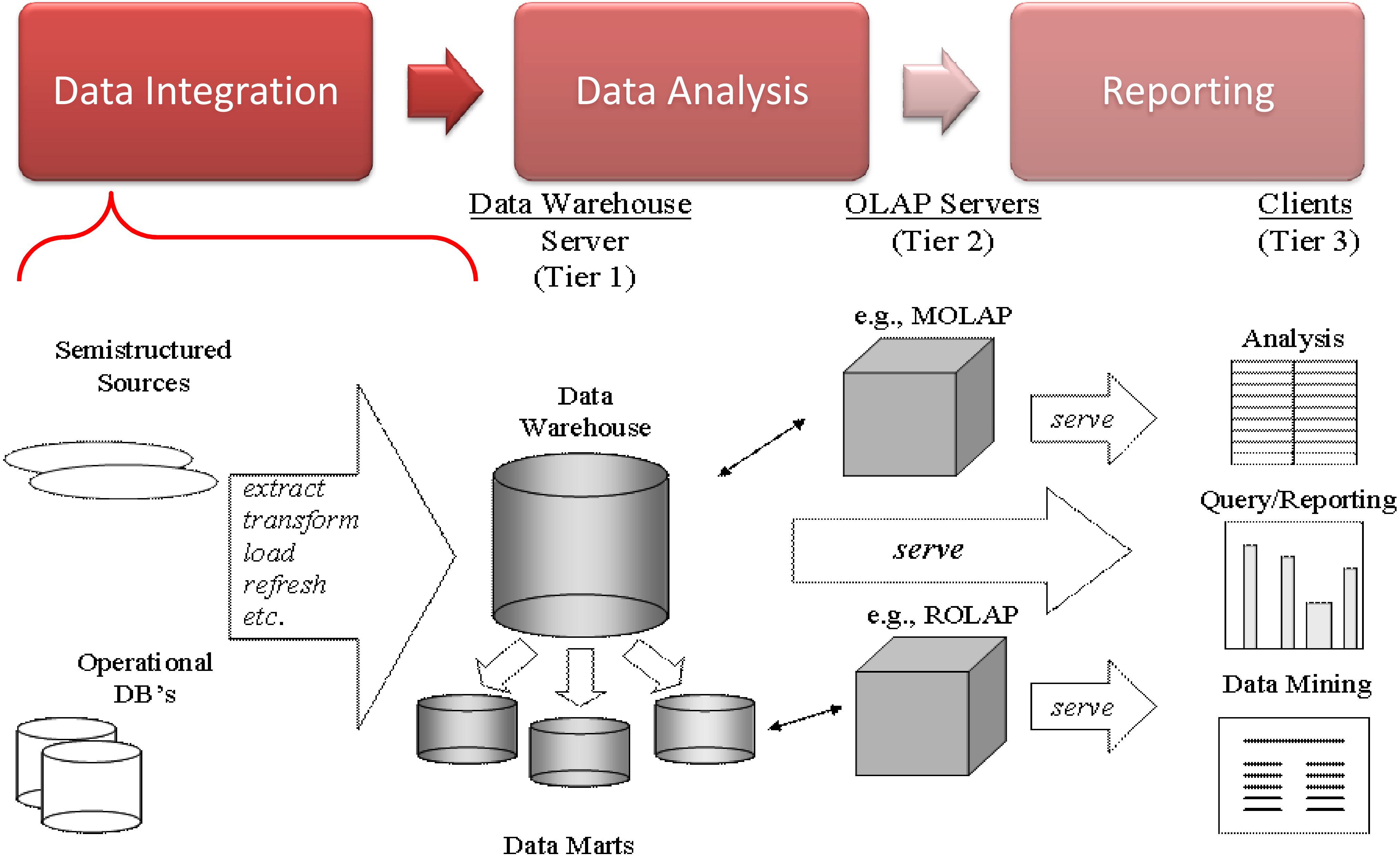
Flexible to change: Business situations change, users' requirements change, technology changes, and tools to access data may also change. The data warehouse must be adaptable to change. Addition of new data from disparate sources or new queries against the data warehouse should not invalidate the existing information in the data warehouse.

Support for more fact-based decision making: "Manage by fact" seems to be the buzzword these days. The data warehouse should have enough pertinent data to support more precise decision making. What is also required is that the business users should be able to access the data easily.

Support for the data security: The data warehouse maintains the company's confidential information. This information falling into wrong hands will do more damage than not a data warehouse at all. There should be mechanisms in place to enable the provision of information in the required format to only those who are supposed to receive it.

Information consistency: Information consistency is about a single/consistent version of truth. A data warehouse brings data from disparate data sources into a centralized repository. Users from across the organization make use of the data warehouse to view a single and consistent version of truth.

BI – The Process



What Is Data Integration?

Process of coherent merging of data from various data sources and presenting a cohesive/consolidated view to the user

- Involves combining data residing at different sources and providing users with a unified view of the data.

- Significant in a variety of situations; both

  - ➢ commercial (e.g., two similar companies trying to merge their database)

  - ➢Scientific (e.g., combining research results from different bioinformatics research repositories)

Answer a Quick Question

According to your understanding

What are the problems faced in Data Integration?

Challenges in Data Integration

- **Development challenges**
  - ➢ Translation of relational database to object-oriented applications
  - ➢ Consistent and inconsistent metadata
  - ➢ Handling redundant and missing data
  - ➢ Normalization of data from different sources

- **Technological challenges**
  - ➢ Various formats of data
  - ➢ Structured and unstructured data
  - ➢ Huge volumes of data

- **Organizational challenges**
  - ➢ Unavailability of data
  - ➢ Manual integration risk, failure

Technologies in Data Integration

.l.Integration is divided into two main approaches:

- Schema integration – reconciles schema elements

  Multiple data sources may provide data on the same entity type. The main goal is to allow applications to transparently view and query this data as one uniform data source, and this is done using various mapping rules to handle structural differences.

- Instance integration – matches tuples and attribute values

  Data integration from multiple heterogeneous data sources has become a high-priority task in many large enterprises. Hence to obtain the accurate semantic information on the data content, the information is being retrieved directly from the data. It identifies and integrates all the instance of the data items that represents the real-world entity, distinct from the schema integration.

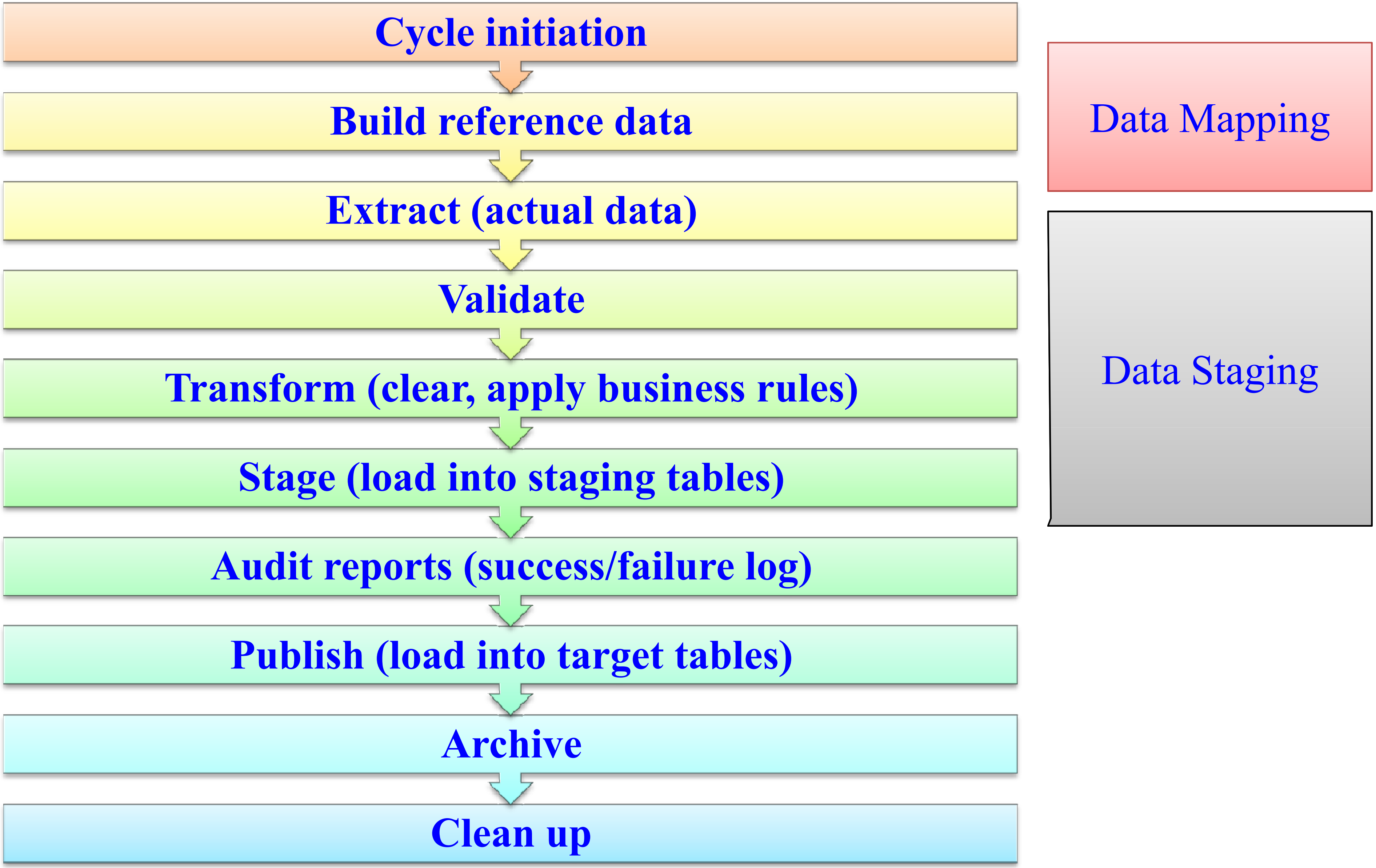
  Entity Identification (EI) and **attribute-value conflict resolution** (AVCR) comprise the instance-integration task. When common key-attributes are not available across different data sources, the rules for EI and the rules for AVCR are expressed as combinations of constraints on their attribute values.

- **Electronic Data Interchange** (EDI) :
  - It refers to the structured transmission of data between organizations by electronic means. It is used to transfer electronic documents from one computer system to another (ie) from one trading partner to another trading partner.
  - It is more than mere E-mail; for instance, organizations might replace bills of lading and even checks with appropriate EDI messages.
- **Object Brokering/Object Request Broker** (ORB)**:**
  - An ORB is a piece of middleware software that allows programmers to make programs calls from one computer to another via a network.
  - It handles the transformation of in-process data structure to and from the byte sequence.

Technologies in Data Integration

- The technologies that are used for data integration include:
  - ➢ Data interchange
  - ➢ Object Brokering
  - ➢ Modeling techniques
    - Entity-Relational Modeling
    - Dimensional Modeling

Various Stages in ETL



Cycle initiation

Build reference data

Extract (actual data)

Validate

Transform (clear, apply business rules)

Stage (load into staging tables)

Audit reports (success/failure log)

Publish (load into target tables)

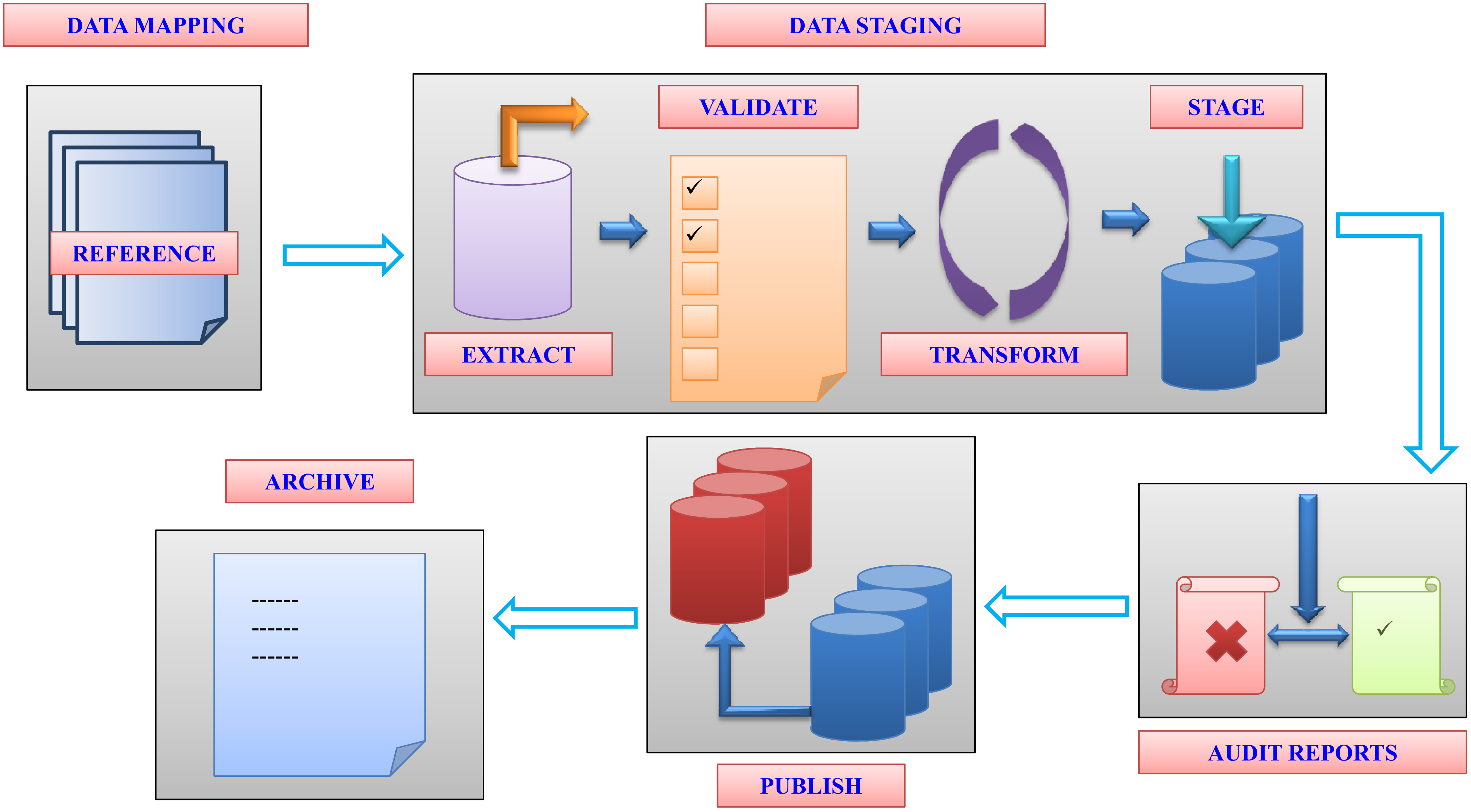
Archive

Clean up

Data Mapping

Data Staging

Various Stages in ETL



Extract, Transform and Load

- **What is ETL?**

  Extract, transform, and load (ETL) in database usage (and especially in data warehousing) involves:

  - Extracting data from different sources
  - Transforming it to fit operational needs (which can include quality levels)
  - Loading it into the end target (database or data warehouse)
- Allows to create efficient and consistent databases
- While ETL can be referred in the context of a data warehouse, the term ETL is in fact referred to as a process that loads any database.
- Usually ETL implementations store an audit trail on positive and negative process runs.

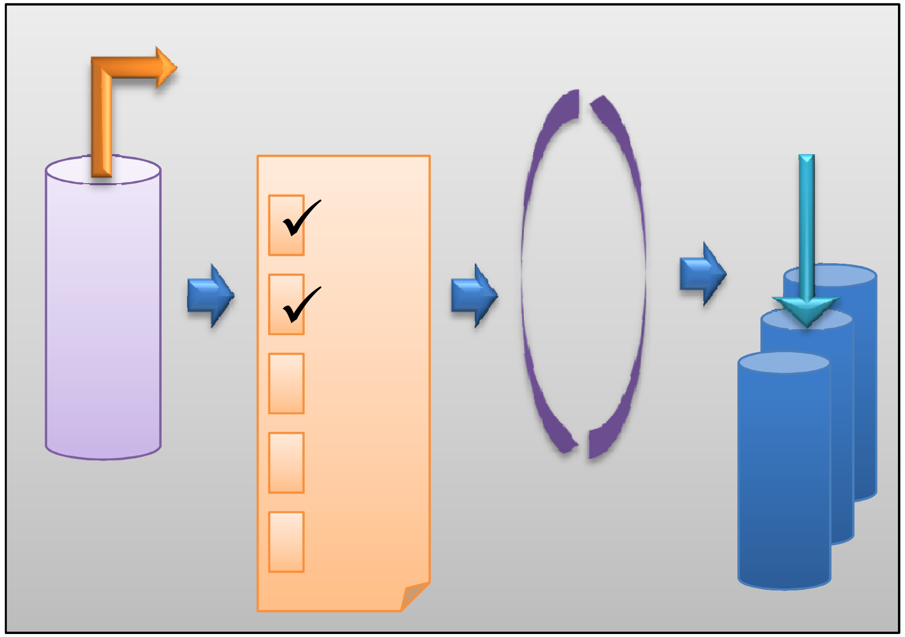
Data Mapping

- The process of creating data element mapping between two distinct data models
- It is used as the first step towards a wide variety of data integration tasks
- The various method of data mapping are
  - **Hand-coded, graphical manual**
    - Graphical tools that allow a user to “draw” lines from fields in one set of data to fields in another
  - **Data-driven mapping**
    - Evaluating actual data values in two data sources using heuristics and statistics to automatically discover complex mappings
  - **Semantic mapping**
    - A metadata registry can be consulted to look up data element synonyms
    - If the destination column does not match the source column, the mappings will be made if these data elements are listed as synonyms in the metadata registry
    - Only able to discover exact matches between columns of data and will not discover any transformation logic or exceptions between columns

Data Staging

A data staging area is an intermediate storage area between the sources of information and the Data Warehouse (DW) or Data Mart (DM)

- A staging area can be used for any of the following purposes:
  - Gather data from different sources at different times
  - Load information from the operational database
  - Find changes against current DW/DM values.
  - Data cleansing
  - Pre-calculate aggregates.



Data Extraction

- Extraction is the operation of extracting data from the source system for further use in a data warehouse environment. This the first step in the ETL process.

- Designing this process means making decisions about the following main aspects:
  - ➢ Which extraction method would I choose?
  - ➢ How do I provide the extracted data for further processing?

Data Extraction (cont…)

The data has to be extracted both logically and physically.

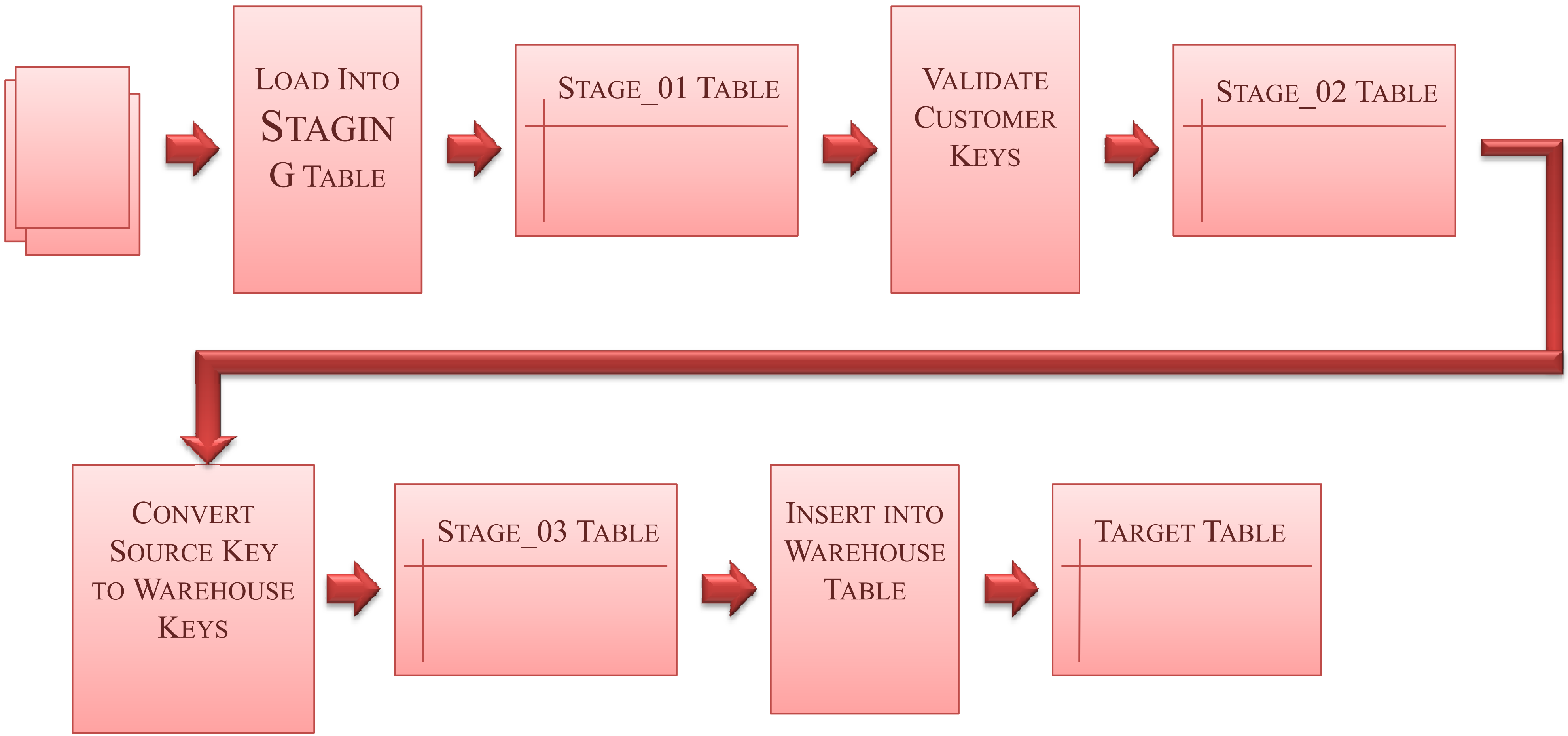
- **The logical extraction method**
  - ➢ Full extraction
  - ➢ Incremental extraction

- **The physical extraction method**
  - ➢ Online extraction
  - ➢ Offline extraction

Data Transformation

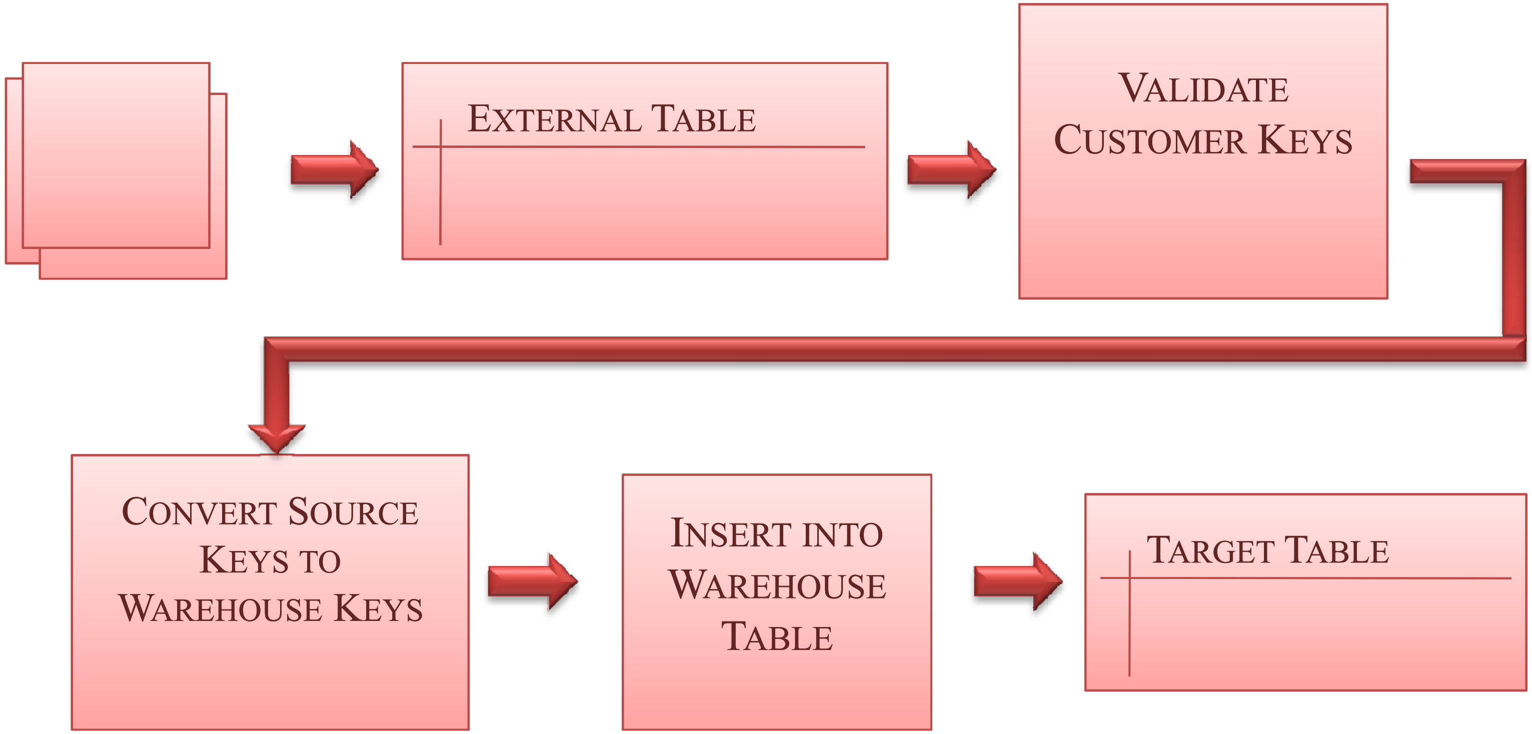
- It is the most complex and, in terms of production the most costly part of ETL process.
- They can range from simple data conversion to extreme data scrubbing techniques.
- From an architectural perspective, transformations can be performed in two ways.
  - ➢ Multistage data transformation
  - ➢ Pipelined data transformation

Data Transformation



Multistage Transformation

Data Transformation


External Table
Validate Customer Keys
Convert Source Keys to Warehouse Keys
Insert into Warehouse Table
Target Table

Pipelined Transformation

Data Loading

- The load phase loads the data into the end target, usually the data warehouse (DW). Depending on the requirements of the organization, this process varies widely.

- The timing and scope to replace or append into the DW are strategic design choices dependent on the time available and the business needs.

- More complex systems can maintain a history and audit trail of all changes to the data loaded in the DW.

Answer a Quick Question

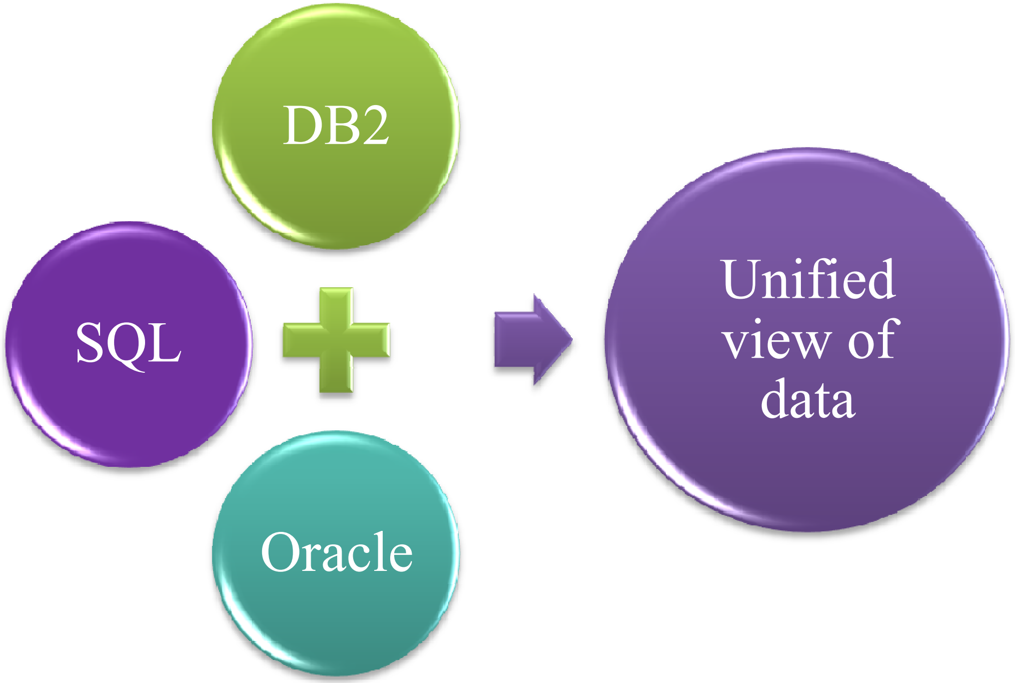
According to your understanding

What is the need for Data Integration in corporate world ?

Need for Data Integration

- It is done for providing data in a specific view as requested by users, applications, etc.
- The bigger the organization gets, the more data there is and the more data needs integration.
- Increases with the need for data sharing.

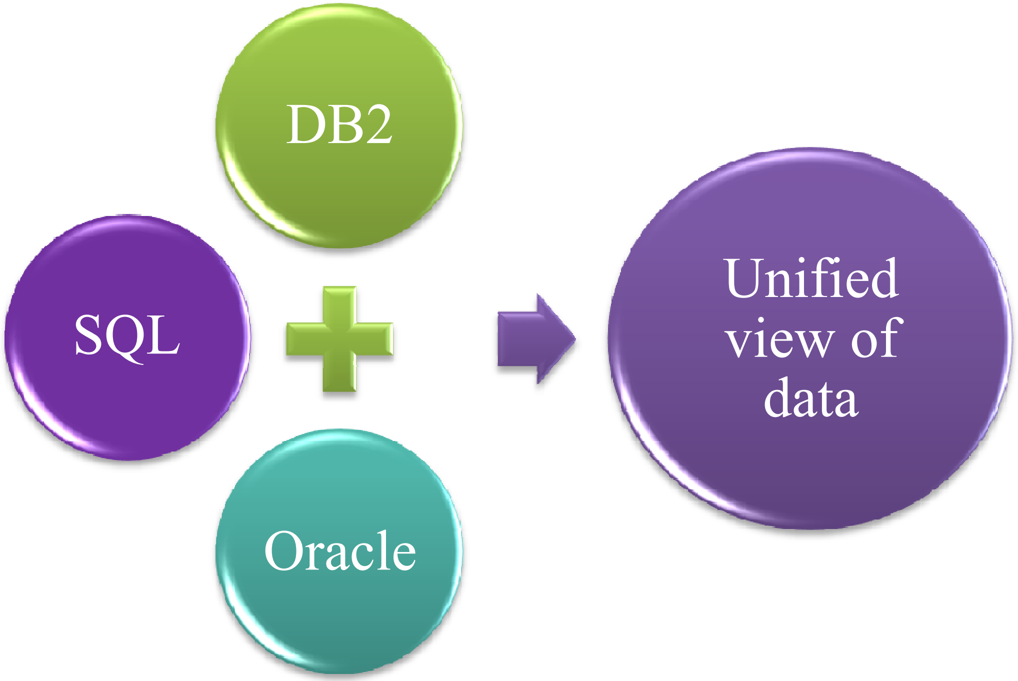
What it means?



Advantages of Using Data Integration

- Of benefit to decision-makers, who have access to important information from past studies
- Reduces cost, overlaps and redundancies; reduces exposure to risks
- Helps to monitor key variables like trends and consumer behaviour, etc.

What it means?



Common Approaches to Data Integration

Data Integration Approaches

- There are currently various methods for performing data integration.

- The most popular ones are:
  - ➢ Federated databases
  - ➢ Memory-mapped data structure
  - ➢ Data warehousing

Data Integration Approaches

- **Federated database (virtual database):**
  - Type of meta-database management system which transparently integrates multiple autonomous databases into a single federated database
  - The constituent databases are interconnected via a computer network, geographically decentralized.
  - The federated databases is the fully integrated, logical composite of all constituent databases in a federated database management system.
- **Memory-mapped data structure:**
  - Useful when needed to do in-memory data manipulation and data structure is large. It's mainly used in the dot net platform and is always performed with C# or using VB.NET
  - It's is a much faster way of accessing the data than using Memory Stream.

Data Integration Approaches

- **Data Warehousing**

  The various primary concepts used in data warehousing would be:

  - ETL (Extract Transform Load)
  - Component-based (Data Mart)
  - Dimensional Models and Schemas
  - Metadata driven

Answer a Quick Question

According to your understanding

What are the advantages and limitations of Data Warehouse?

Data Warehouse – Advantage and Limitations

ADVANTAGES

- Integration at the lowest level, eliminating need for integration queries.
- Runtime schematic cleaning is not needed – performed at the data staging environment
- Independent of original data source
- Query optimization is possible.

LIMITATIONS

- Process would take a considerable amount of time and effort
- Requires an understanding of the domain
- More scalable when accompanied with a metadata repository – increased load.
- Tightly coupled architecture

Metadata and Its Types

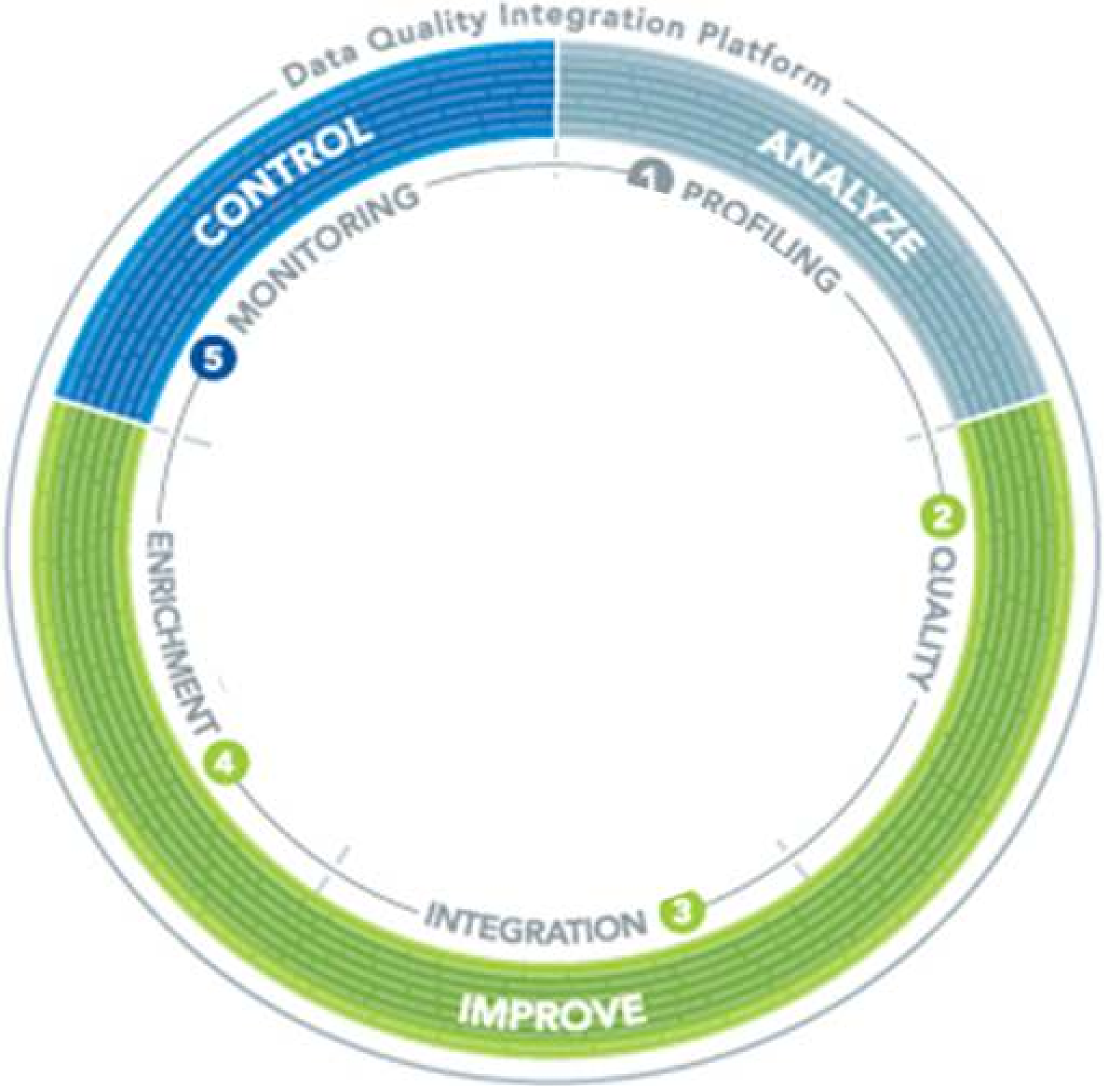
Metadata and Its Types

WHAT	**Business**	Data definitions, Metrics definitions, Subject models, Data models, Business rules, Data rules, Data owners/stewards, etc.
HOW	**Process**	Source/target maps, Transformation rules, data cleansing rules, extract audit trail, transform audit trail, load audit trail, data quality audit, etc.
TYPE	**Technical**	Data locations, Data formats, Technical names, Data sizes, Data types, indexing, data structures, etc.
WHO, WHEN	**Application**	Data access history: Who is accessing? Frequency of access? When accessed? How accessed? … , etc.

Data Quality and Data Profiling

Building Blocks of Data Quality Management

- Analyze, Improve and Control
- This methodology is used to encompass people, processes and technology.
- This is achieved through five methodological building blocks, namely:
  - ➢ Profiling
  - ➢ Quality
  - ➢ Integration
  - ➢ Enrichment
  - ➢ Monitoring



Data Profiling

- Beginning the data improvement efforts by knowing where to begin.
- **Data profiling** (sometimes called data discovery or data quality analysis) helps to gain a clear perspective on the current integrity of data. It helps:
  - Discover the quality, characteristics and potential problems
  - Reduce the time and resources in finding problematic data
  - Gain more control on the maintenance and management of data
  - Catalog and analyze metadata
- The various steps in profiling include
  - Metadata analysis
  - Outline detection
  - Data validation
  - Pattern analysis
  - Relationship discovery
  - Statistical analysis
  - Business rule validation

Data Profiling (cont…)

- Metadata profiling
  - ➢ Typical type of metadata profiling are
    - Domain: Conformation of data in column to the defined value or range
    - Type: Alphabetic or numeric
    - Pattern: The proper pattern
    - Frequency counts
    - Interdependencies:
      - Within a table:
      - Between tables:
- Data profiling analysis
  - ➢ Column profiling
  - ➢ Dependency profiling
  - ➢ Redundancy profiling

Answer a Quick Question

According to your understanding

What is data quality and why it is important?

Data Quality

- Correcting, standardizing and validating the information

- Creating business rules to correct, standardize and validate your data.

- High-quality data is essential to successful business operations.

Data Quality (cont…)

- Data quality helps you to:
  - Plan and prioritize data
  - Parse data
  - Standardize, correct and normalize data
  - Verify and validate data accuracy
  - Apply business rules
- Standardize and Transform Data
- The three components that ensure the quality and integrity of the data:
  - Data rationalization
  - Data standardization
  - Data transformation

Answer a Quick Question

What do you think are the major causes of bad data quality?

Causes of Bad Data Quality

During Process of Extraction

- Initial Conversion of Data
- Consolidation of System
- Manual Data Entry
- Batch Feeds
- Real Time Interfaces

Effect of Bad Quality

Data Decay During Loading and Archiving

- Changes Not Captured
- System Upgrades
- Use of New Data
- Loss of Expertise
- Automation Process

During Data Transformations

- Processing Data
- Data Scrubbing
- Data Purging

Data Quality in Data Integration

- Building a unified view of the database from the information.
- An effective data integration strategy can lower costs and improve productivity by ensuring the consistency, accuracy and reliability of data.
- Data integration enables to:
  - ➢ Match, link and consolidate multiple data sources
  - ➢ Gain access to the right data sources at the right time
  - ➢ Deliver high-quality information
  - ➢ Increase the quality of information

Data Quality in Data Integration

- Understand Corporate Information Anywhere in the Enterprise
- Data integration involves combining processes and technology to ensure an effective use of the data can be made.

- Data integration can include:
  - ➢ Data movement
  - ➢ Data linking and matching
  - ➢ Data house holding

Popular ETL Tools

ETL Tools

- ETL process can be create using programming language.
- Open source ETL framework tools
  - Clover.ETL
  - Enhydra Octopus
  - Pentaho Data Integration (also known as ‘Kettle’)
  - Talend Open Studio

- Popular ETL Tools
  - Ab Initio
  - Business Objects Data Integrator
  - Informatica
  - SQL Server 2005/08 Integration services