Chapter 6
Basics of Data Integration

Learning Objectives and Learning Outcomes

Learning Objectives	Learning Outcomes
1. Concepts of data integration	(a) To realize the importance of metadata
2. Needs and advantages of using data integration	(b) To understand data quality
3. Introduction to common data integration approaches	(c) To be able to perform scrubbing/cleaning of data
4. Metadata – types and sources	(d) To be able to apply de-duplication
5. Introduction to data quality	(e) To be able to enhance the quality of data
6. Data profiling concepts and applications	

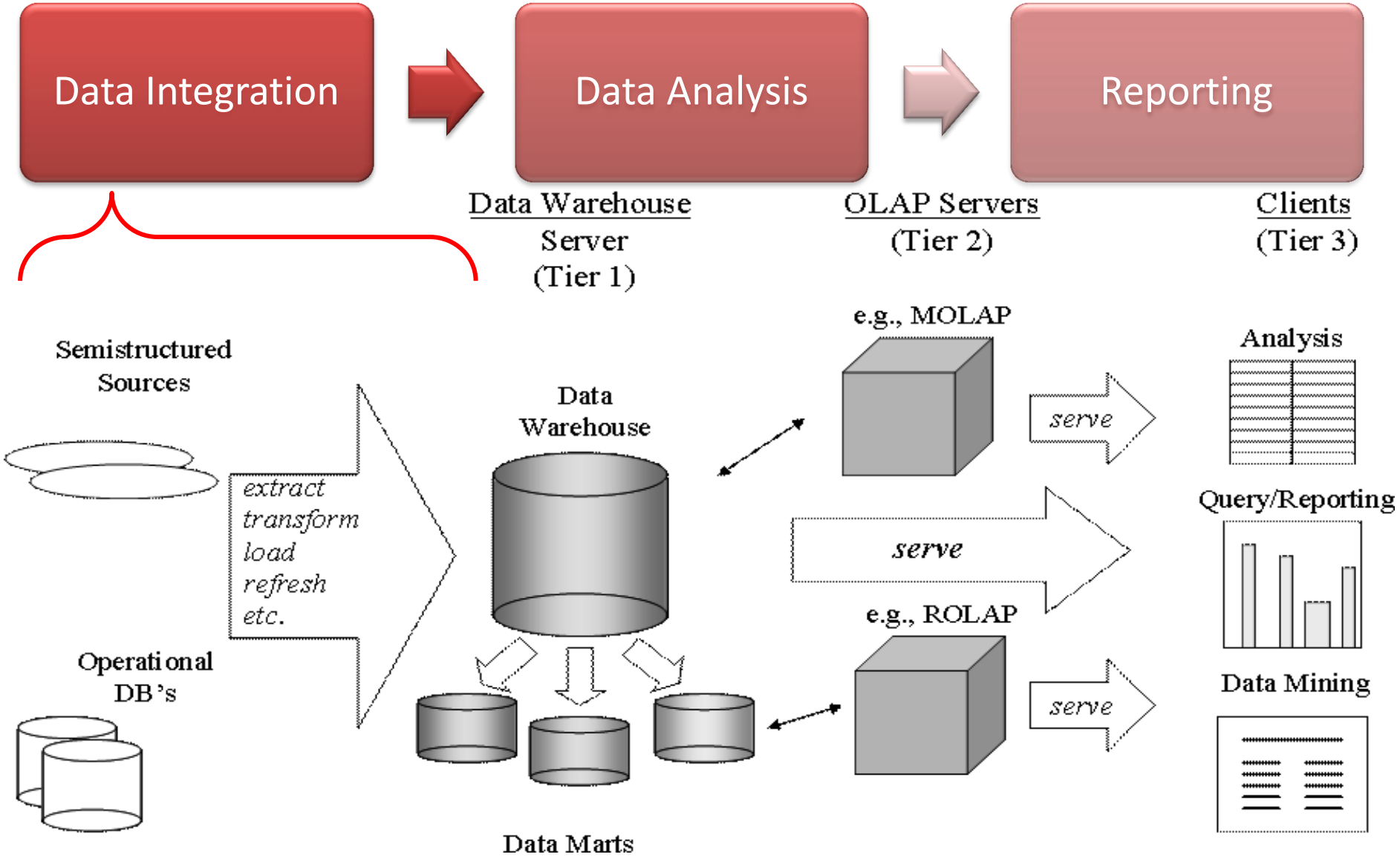
Session Plan

Lecture time	:	90 minutes
Q/A	:	15 minutes

Agenda

- BI – the process
- What is Data Integration?
  - Challenges in Data Integration
  - Technologies in Data Integration
- ETL: Extract, Transform, Load
  - Various stages in ETL
- Need for Data Integration
- Advantages of using Data Integration
- Common approaches to Data Integration
- Metadata and its types
- Data Quality and Data Profiling concepts

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

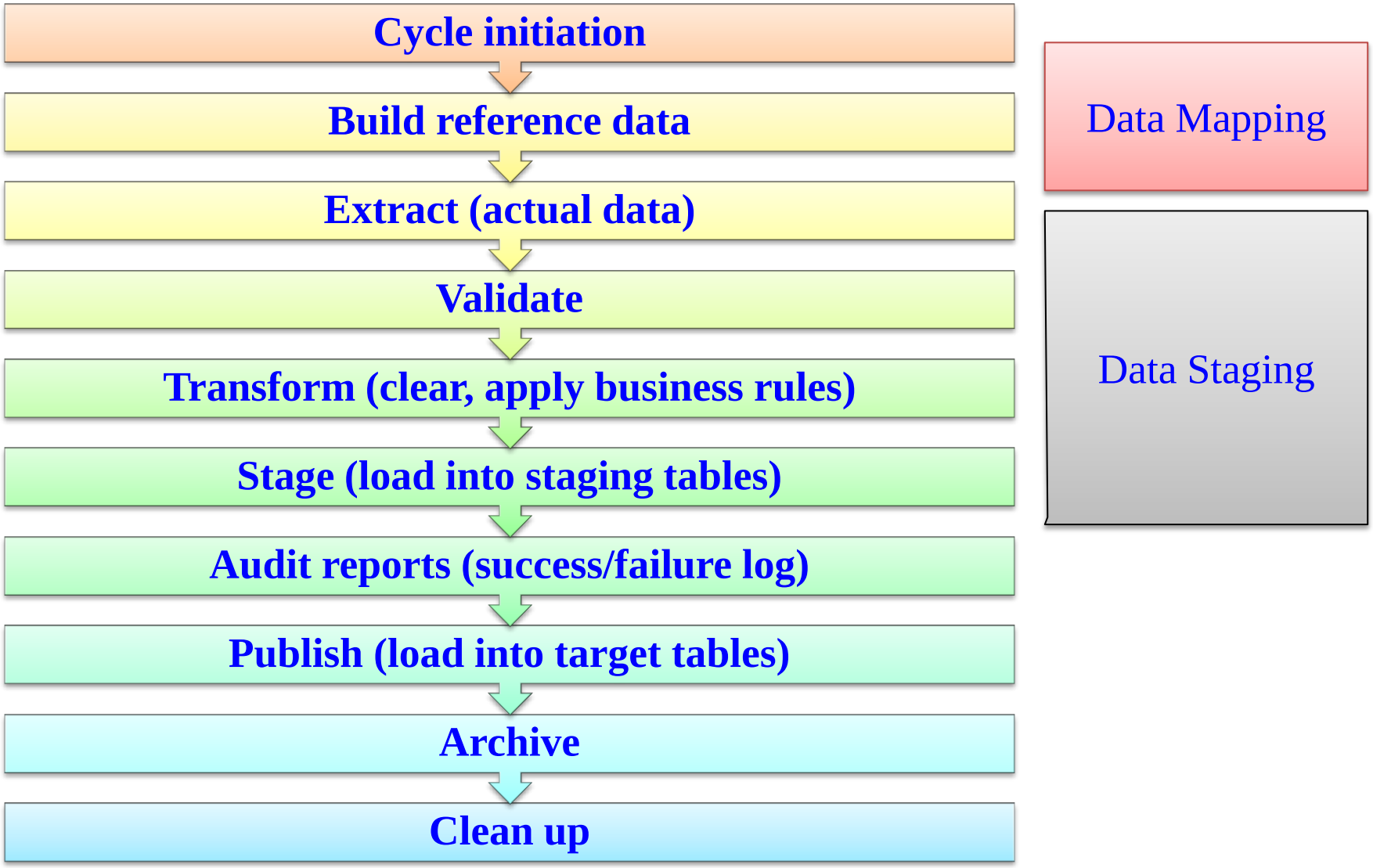
- **Organizational challenges**
  - Unavailability of data
  - Manual integration risk, failure

Technologies in Data Integration

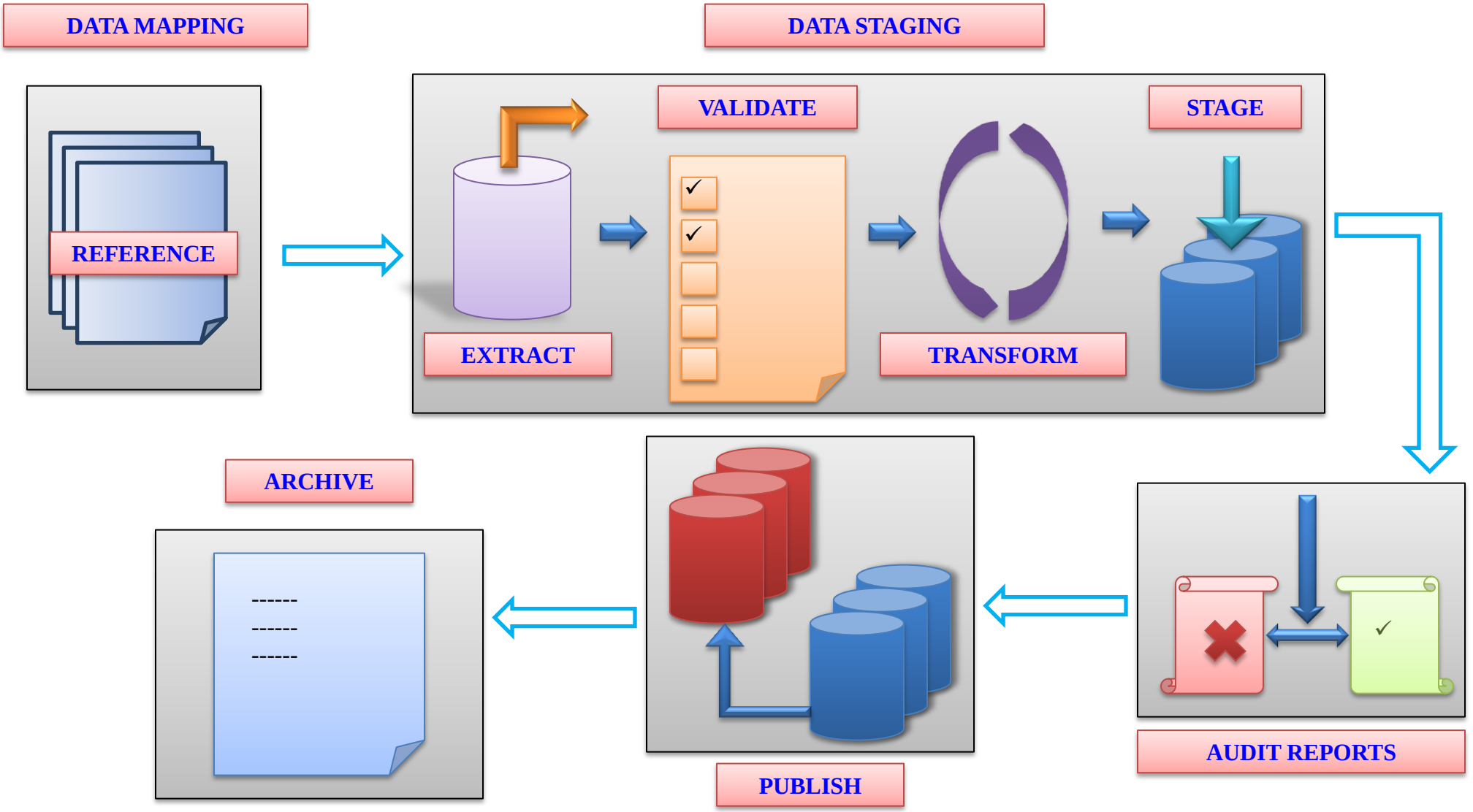
- Integration is divided into two main approaches:
  - Schema integration – reconciles schema elements
  - Instance integration – matches tuples and attribute values

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

Various Stages in ETL



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.

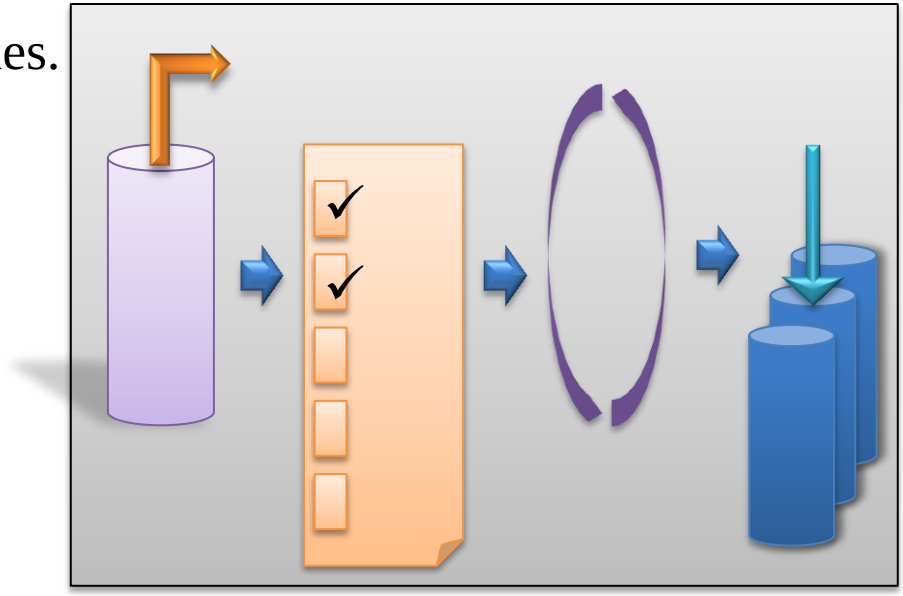
Data Mapping

- 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.

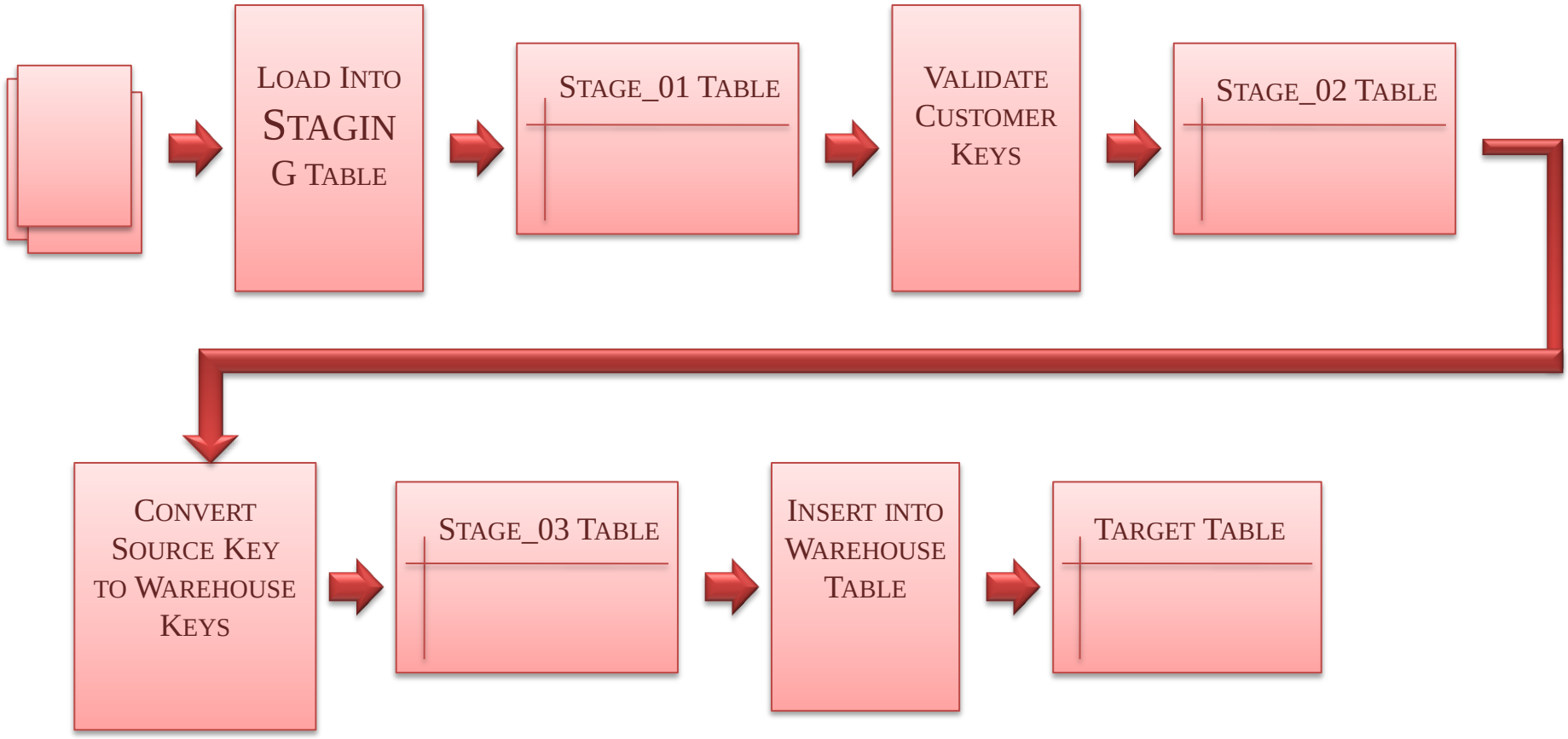
- **The logical extraction method**
  - Full extraction
  - Incremental extraction

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

Data Transformation

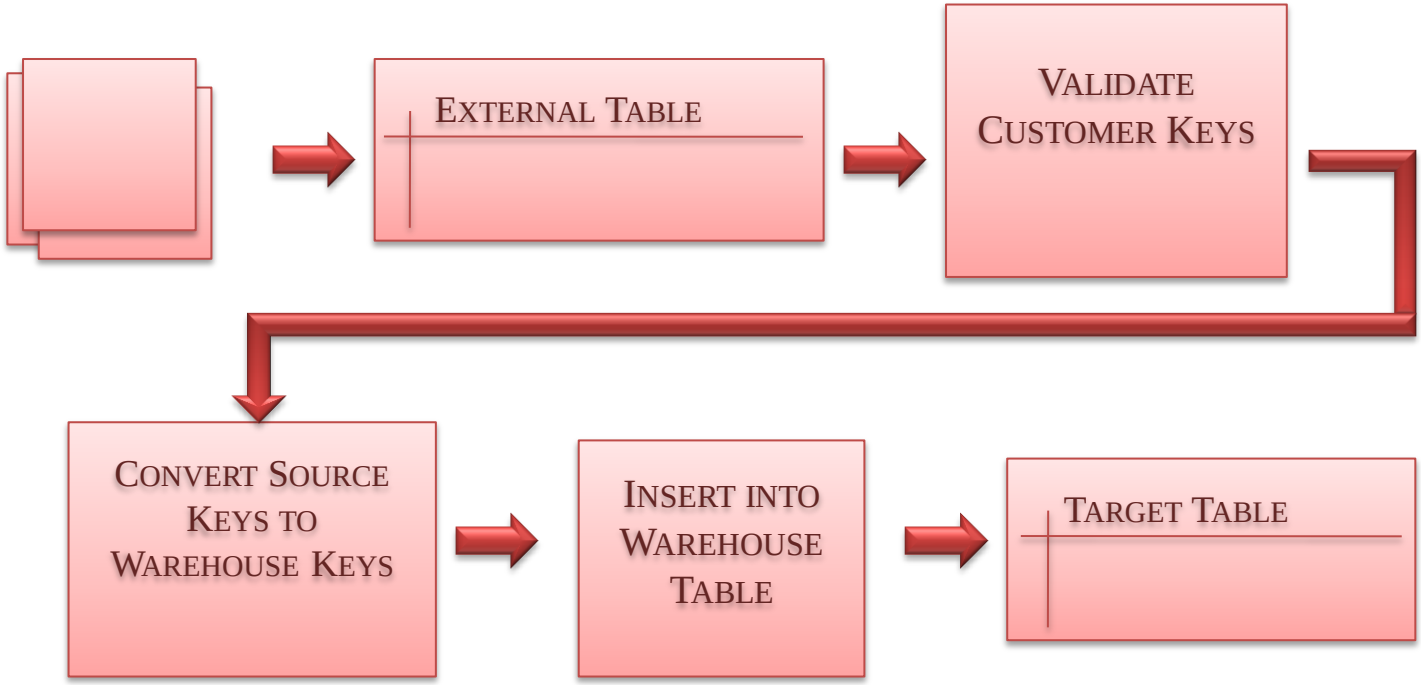
- It is the most complex and, in terms of production the most costly part of ETL process.
- They can range from simple data conversion to extreme data scrubbing techniques.
- From an architectural perspective, transformations can be performed in two ways.
  - Multistage data transformation
  - Pipelined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

Data Transformation



MULTISTAGE TRANSFORMATION

Data Transformation



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

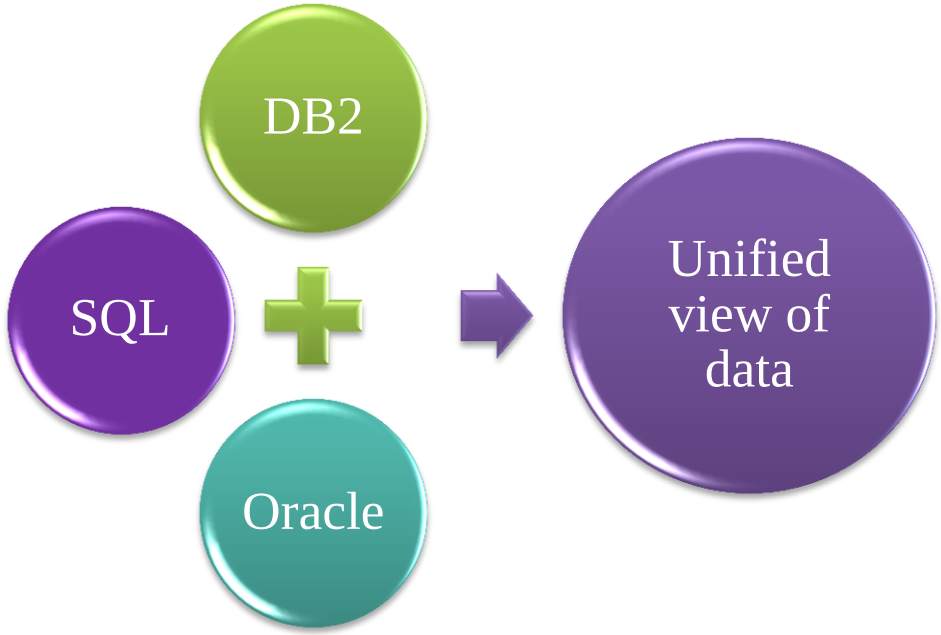
According to your understanding

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.

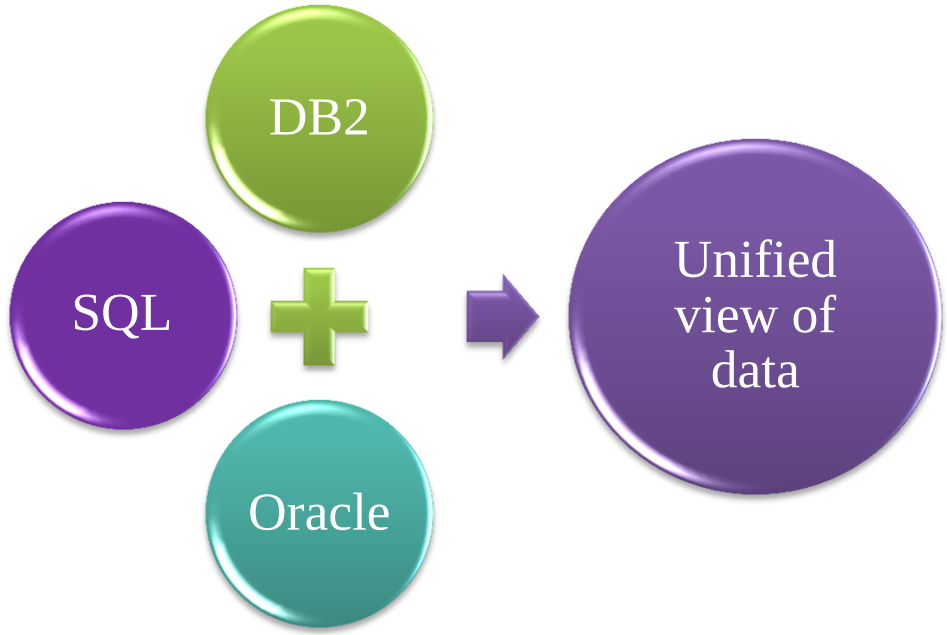
What it means?



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

"Fundamentals of Business Analytics"
RN Prasad and Seema Acharya

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

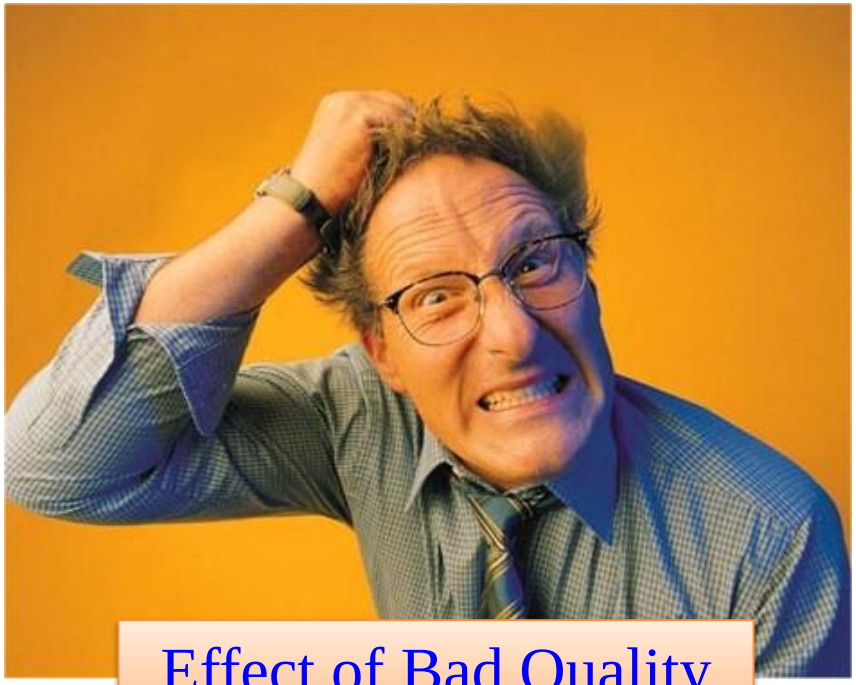
Answer a Quick Question

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

Summary please…

Ask a few participants of the learning program to summarize the lecture.