Produced by:
Dr. Brenda Mullally bmullally@wit.ie
Ruth Barry rbarry@wit.ie

Department Computing Maths and Physics
Waterford Institute of Technology
www.wit.ie
moodle.wit.ie

MSc Enterprise Software Systems

Business Intelligence

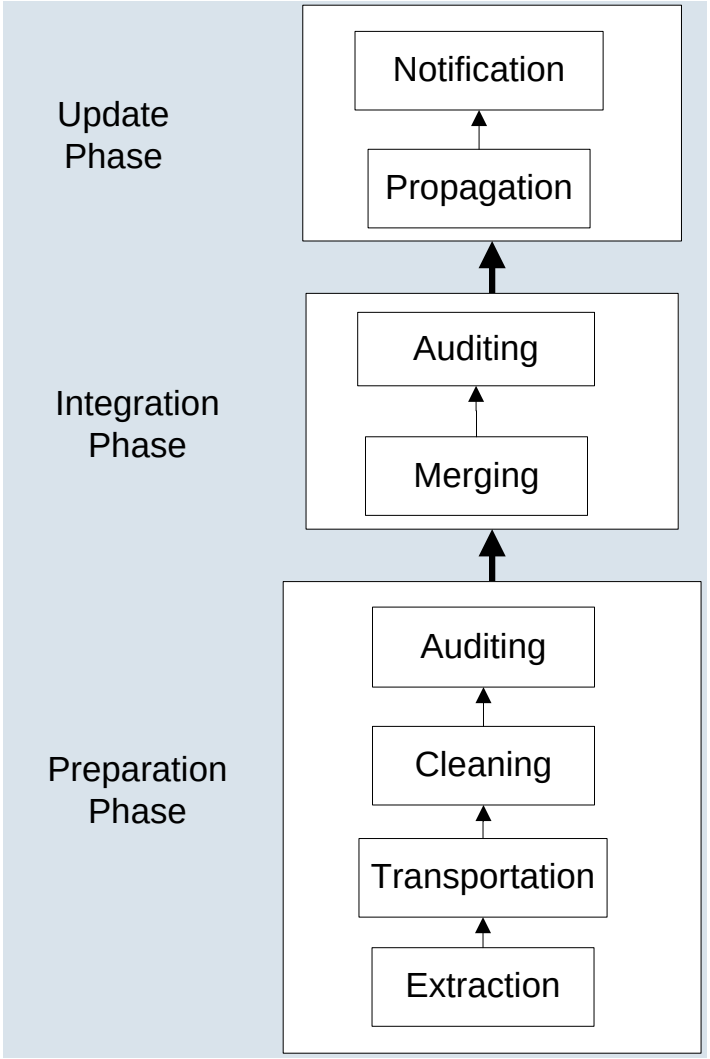
Objectives

- Discuss difficulties of initial population of a data warehouse
- Understand the tradeoffs and constraints in managing refresh processing
- Understand the types of data sources
- Explain the differences between ETL and ELT architectures
- Common features of data integration tools
- Vendors

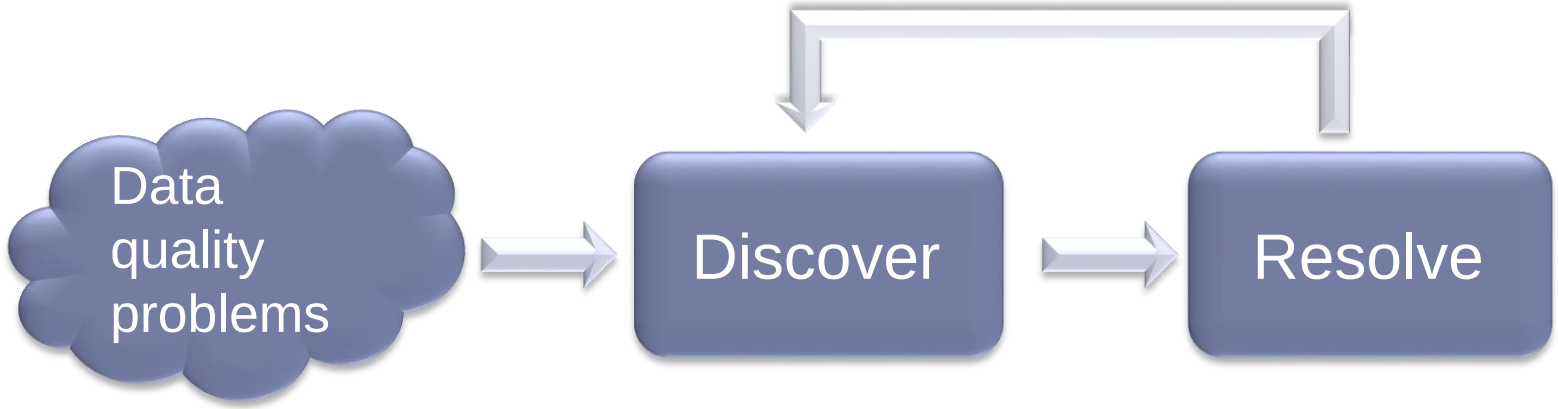
Motivation for Data Integration

- Add value by data transformations
- Find single source of truth
- Overcome challenges
- Critical success factor for data warehouse projects
- Significant investments in effort, hardware, and software

Refresh Workflow

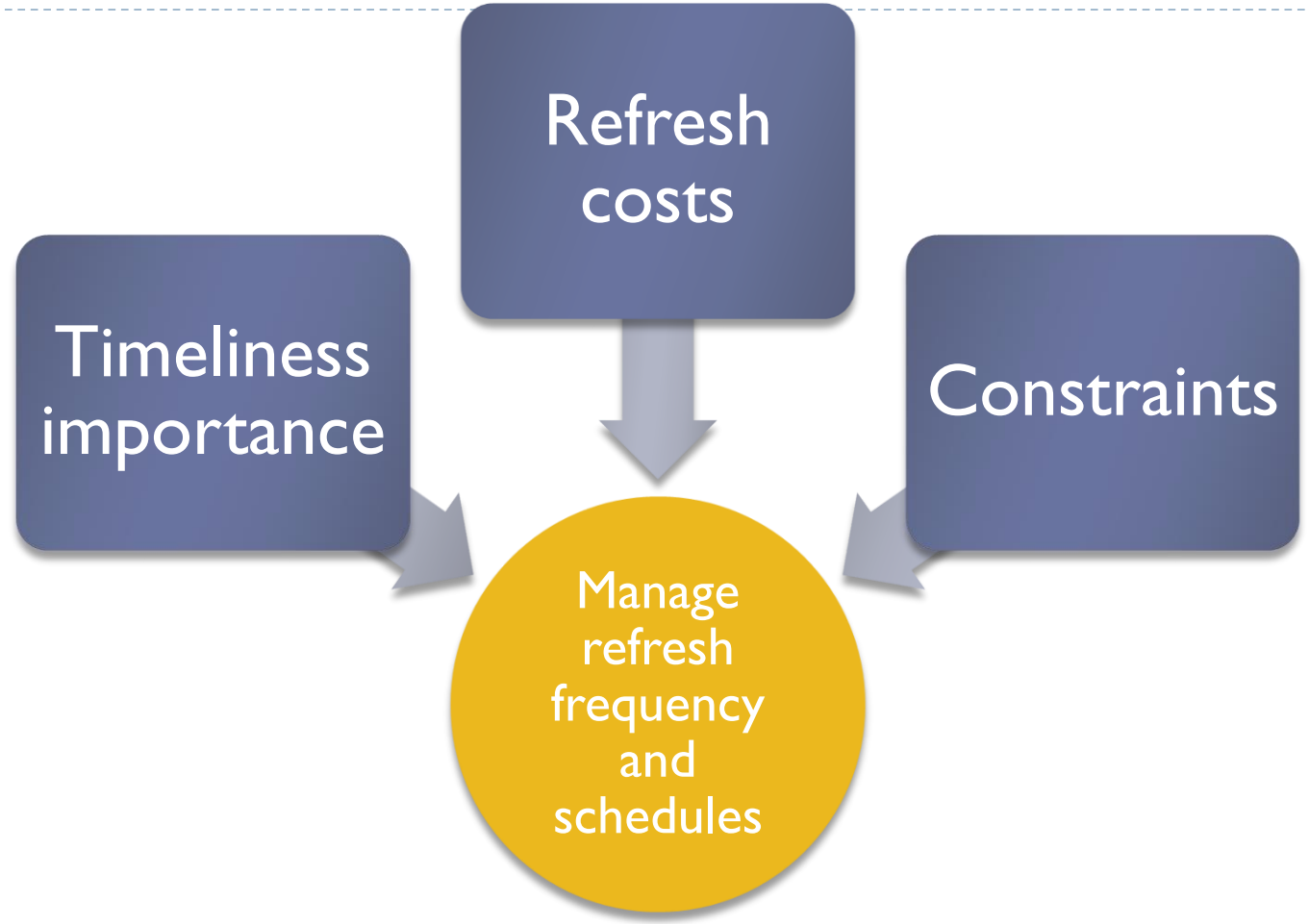


Initial Data Warehouse Load

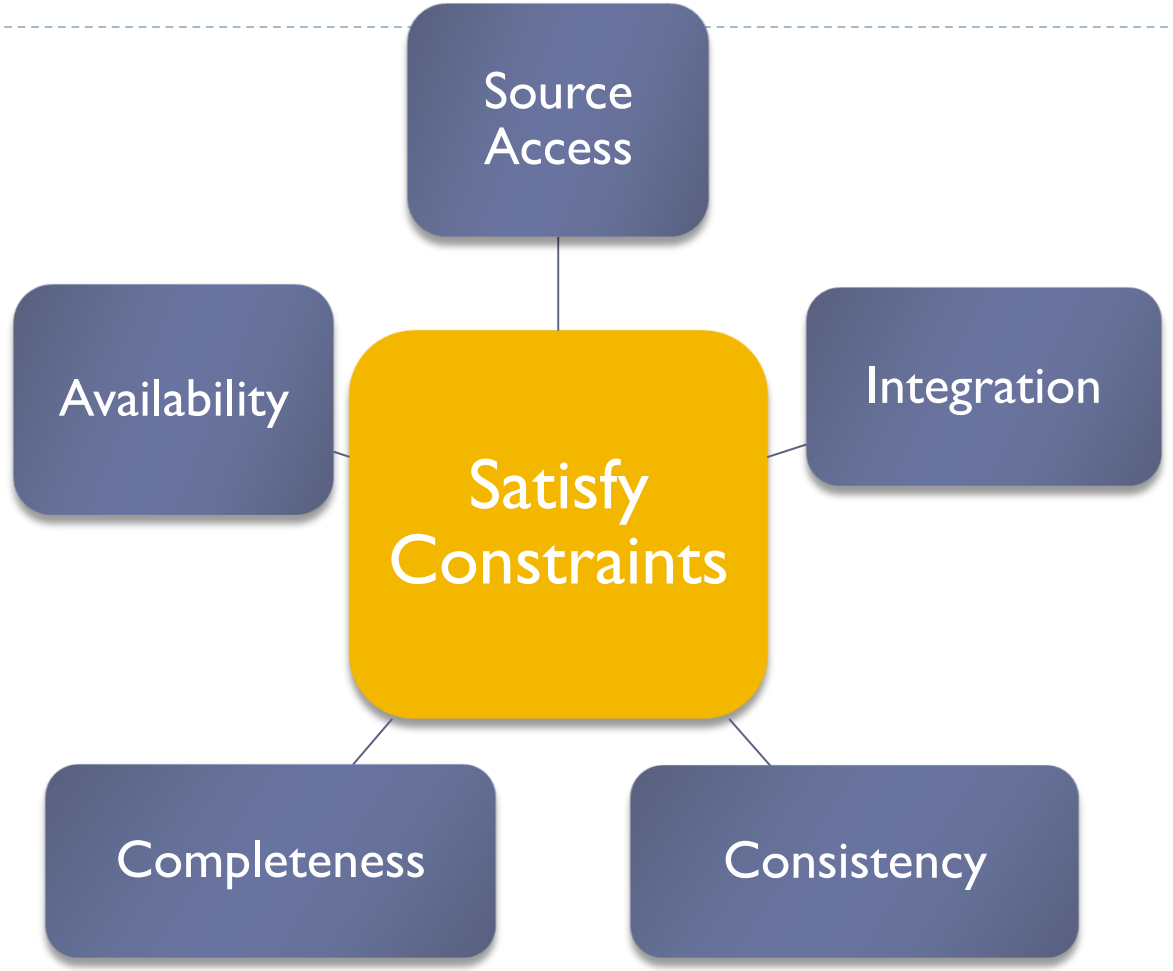


- Major development activity
- More open ended than refresh with difficult to estimate time requirements
- Use profiling tools to discover data quality problems
- Perform for major data warehouse extensions

Refresh Processing Decision Making



Refresh Constraints



Basics of Change Data

- Derived from internal and external data sources
- Used to populate and refresh a data warehouse
  - Insert rows in fact and dimension tables
  - Update rows in dimension tables
- Challenges
  - Difficult to change source systems especially external systems
  - Lack of SQL access and descriptive (meta) data especially for legacy data

Data Quality Problems

- Multiple identifiers
- Different units
- Missing values
- Text data with different components and formats
- Conflicting data
- Different update times

Data cleaning & transforming

- Parsing
- Missing Values
- Conflicting Values
- Standardisation
- Entity matching and consolidation

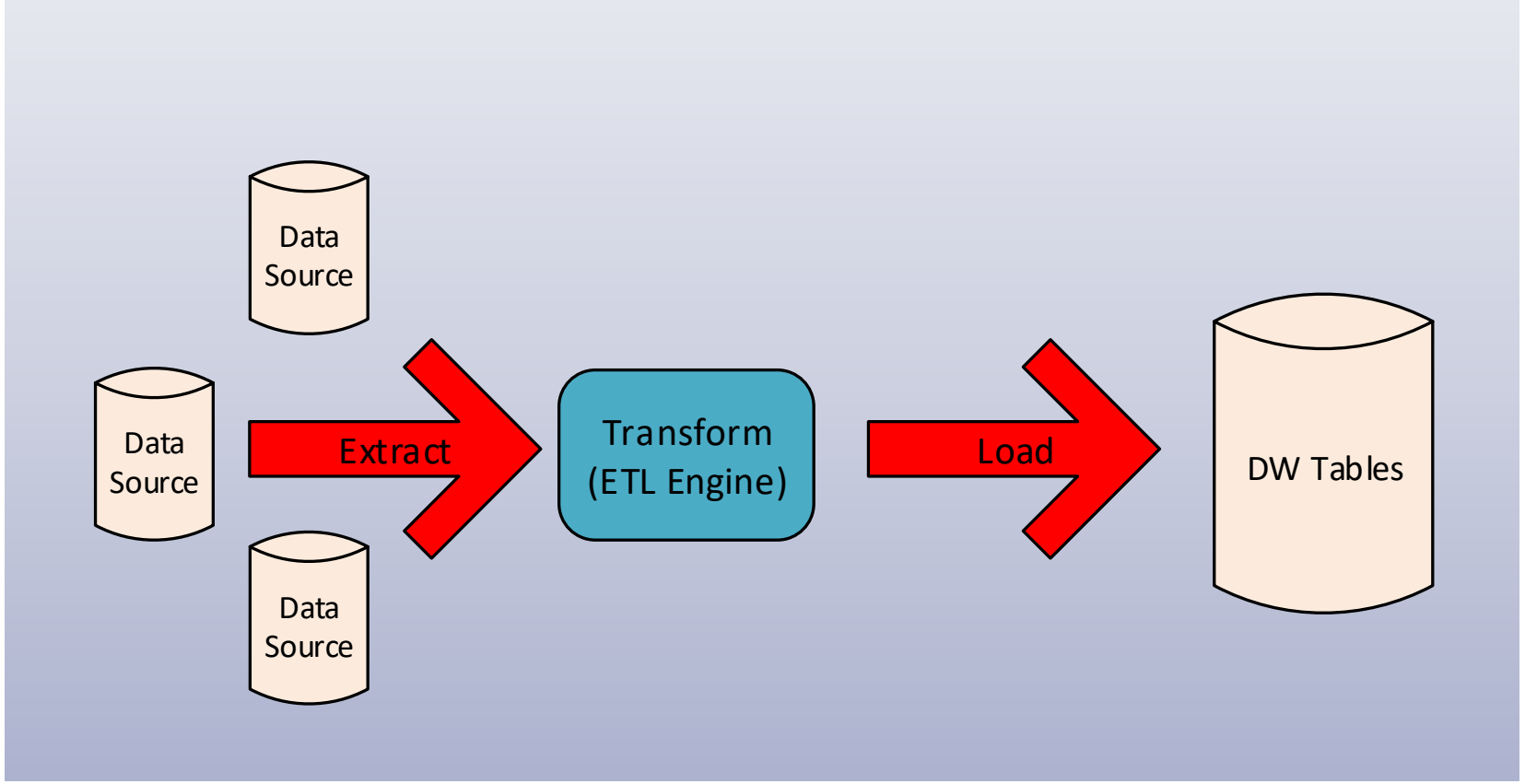
Motivation for Data Integration Tools

- Support initial population and refresh processes
- Project failures partly due to lack of tools and poor performance
- Improve software development productivity
  - Complete development environment
  - Full range of data integration tasks
  - Minimize custom coding
- Achieve high performance

ETL

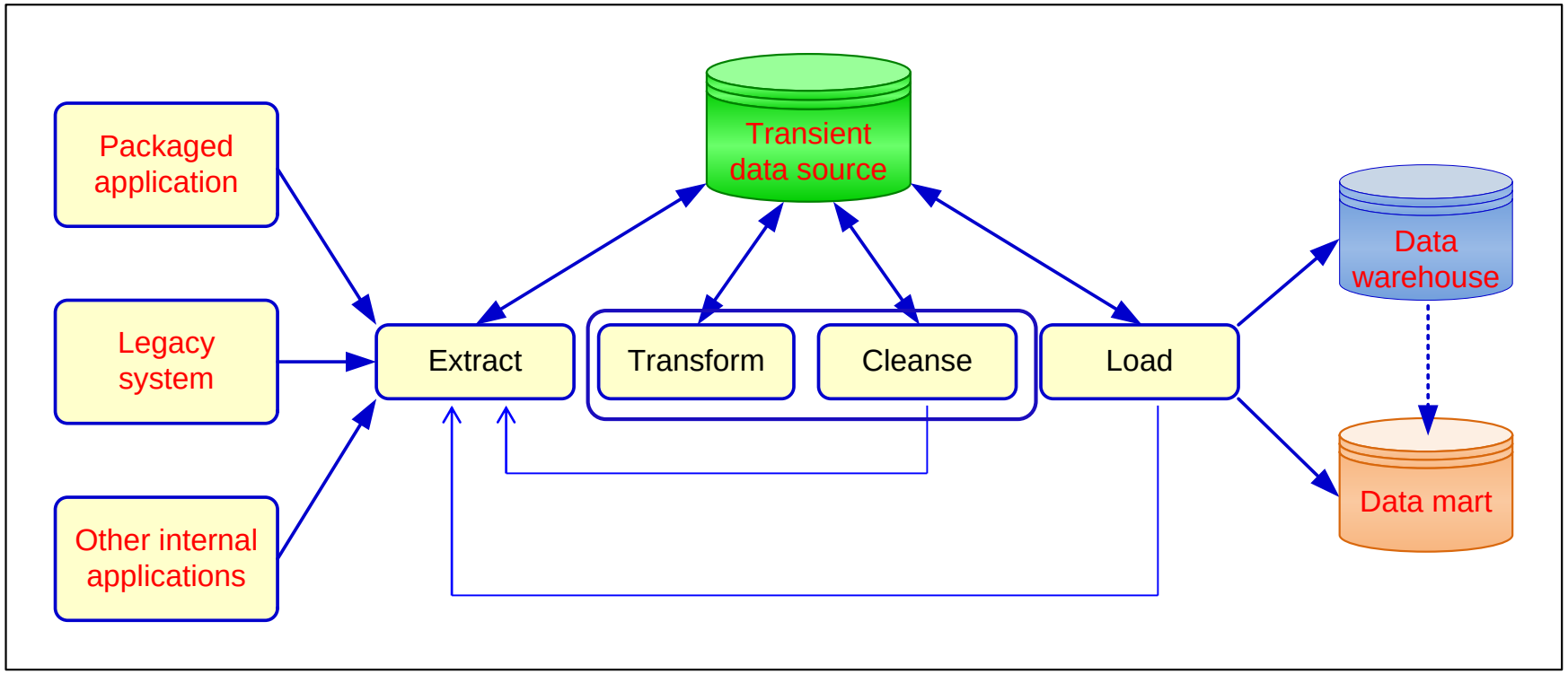
- Extraction
  - sources
- Transform & Clean
  - rules
- Load

ETL Architecture

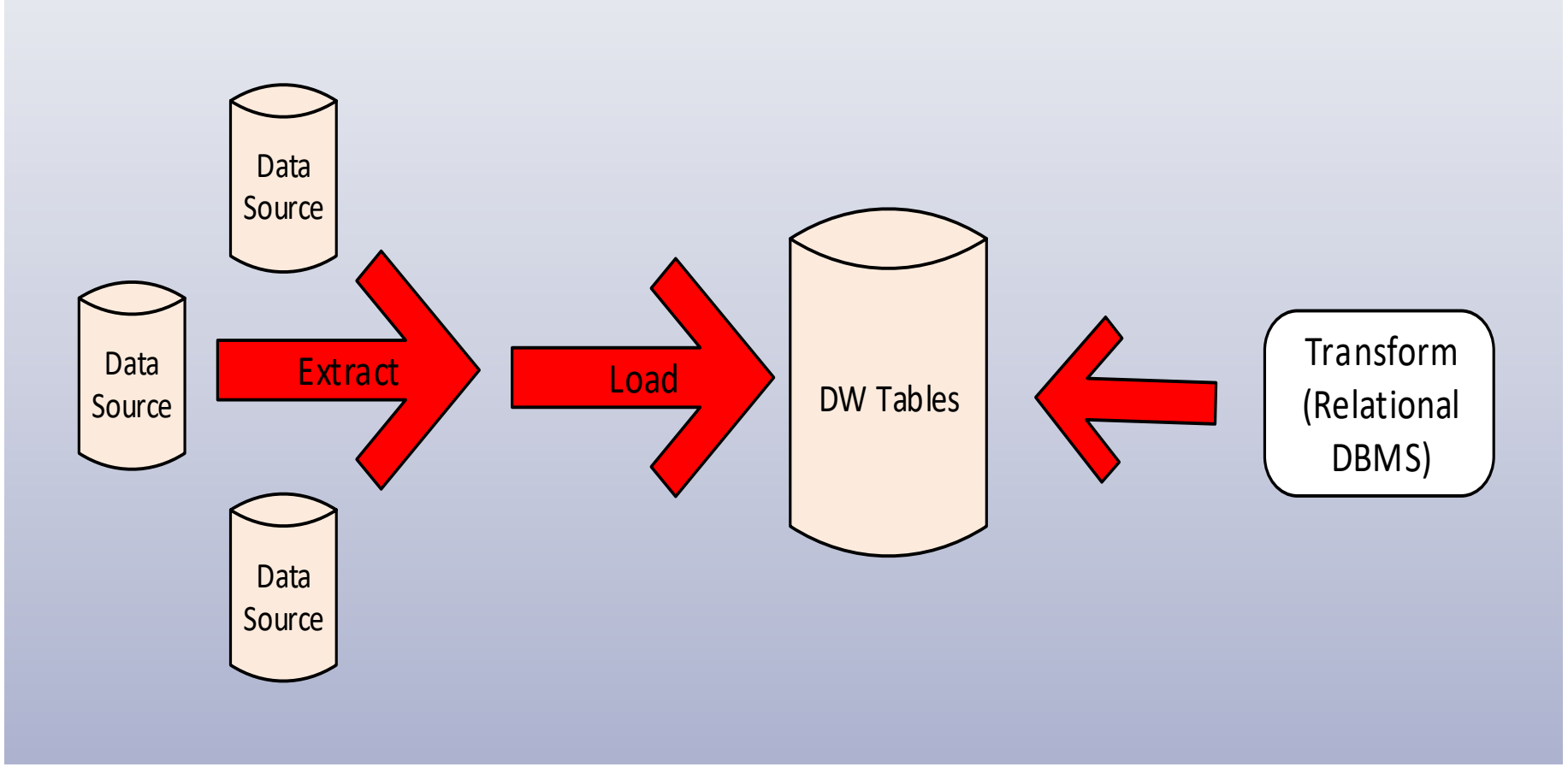


Data Integration and the Extraction, Transformation, and Load (ETL) Process

Extraction, transformation, and load (ETL)



ELT Architecture



Architecture Evaluation

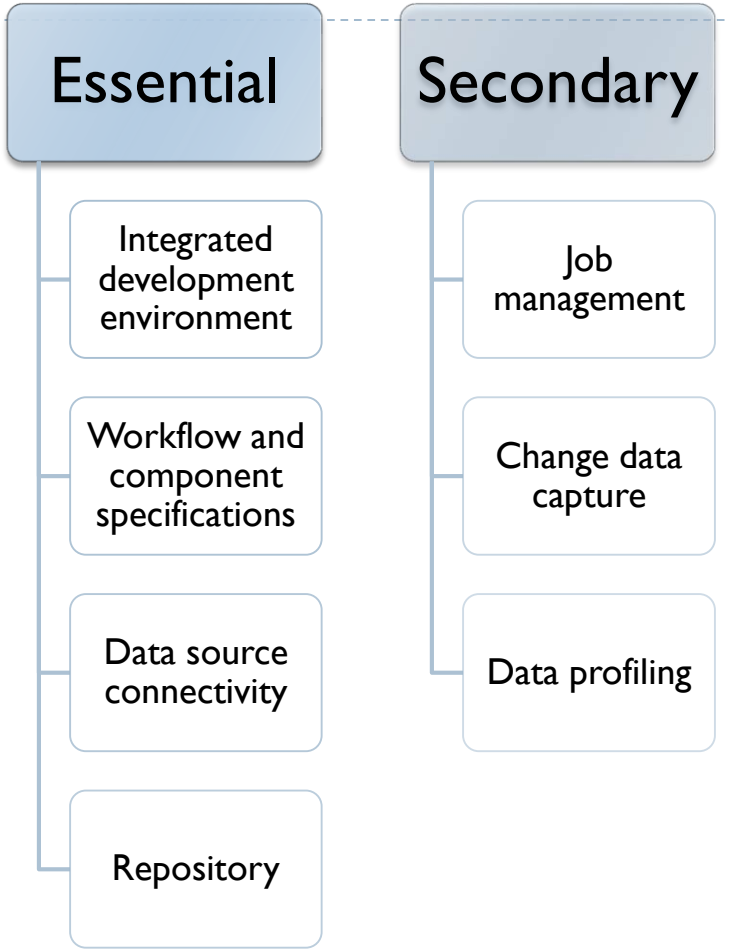
- ETL
  - DBMS independence for ETL
  - More complex operations for ETL in transformations ELT
- ELT
  - Superior optimization technology in relational DBMS engines
  - Less network bandwidth for ELT

- Combination of architectures possible

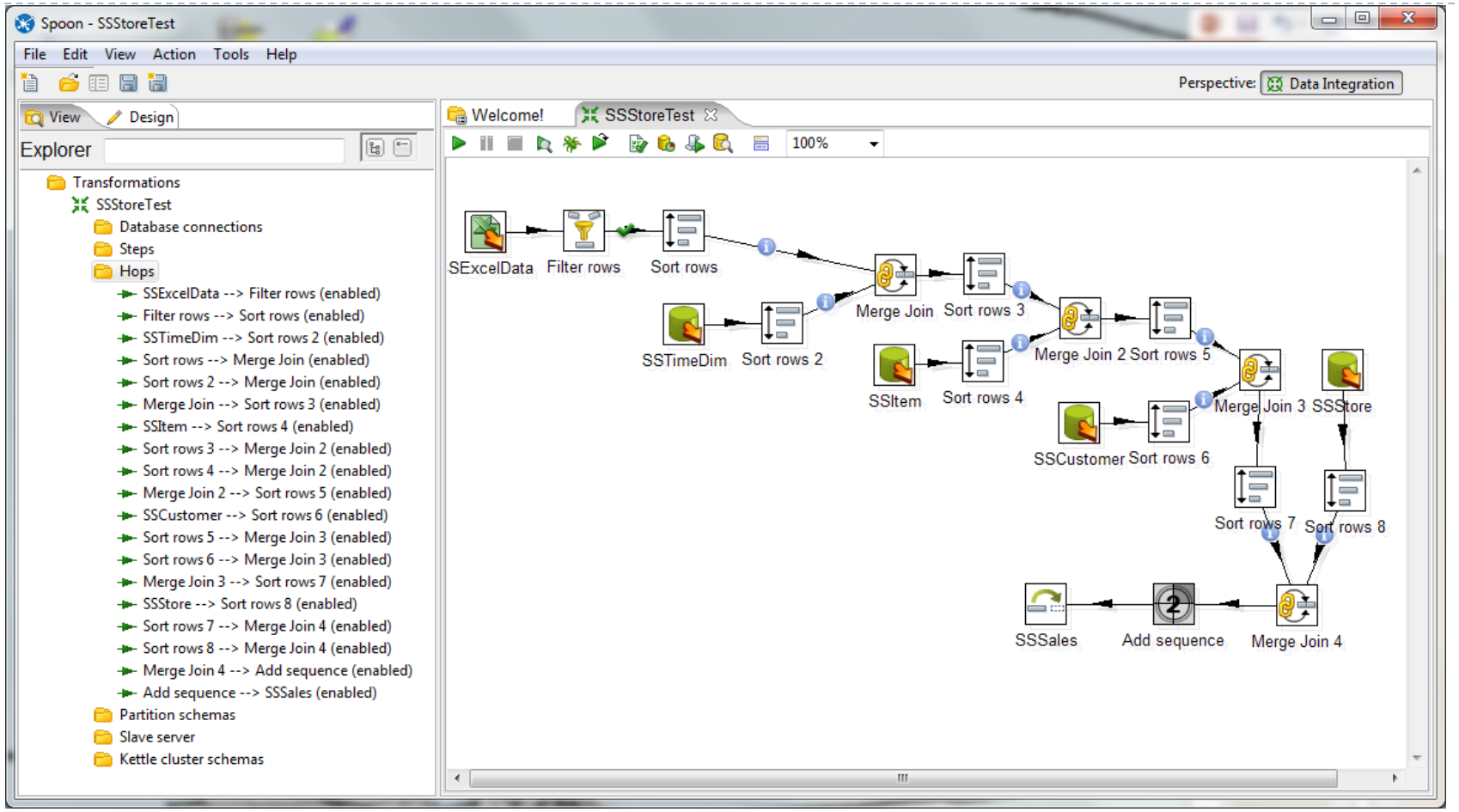
Data Integration Tool Vendors

- Traditional vendor products
  - Database vendors: Oracle, IBM, Microsoft
  - Other vendors: SAP, Informatica, SAS, Information Builders
- Open source with subscription services
  - Pentaho Data Integration
  - Talend Open Studio for Data Integration
  - CloverETL

Feature Overview

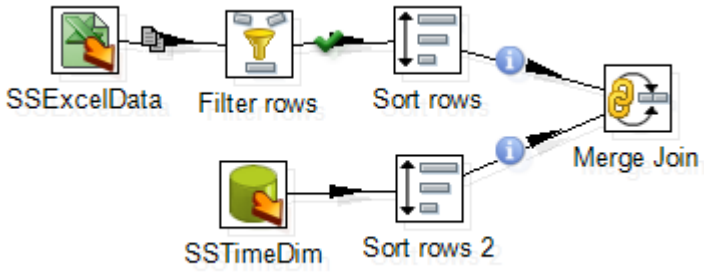


IDE Overview

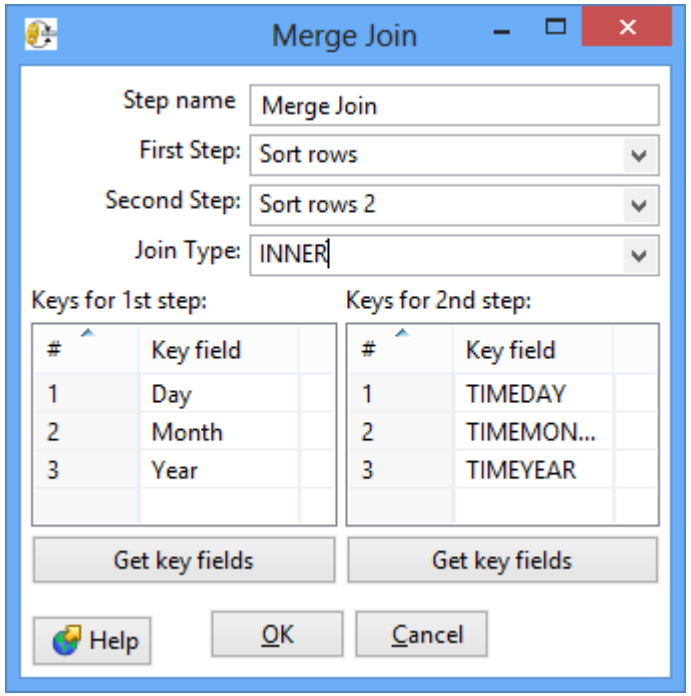


Workflow and Component Specification

Workflow



Specification Window



Workflow Components

Processing

Orchestration

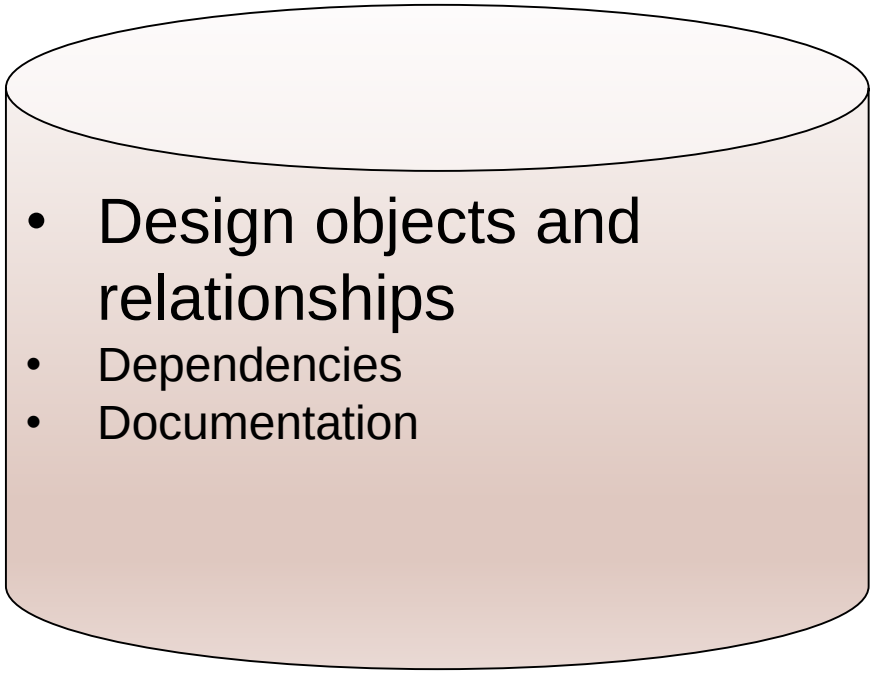
Business intelligence

Database

Data quality

File processing

Repository

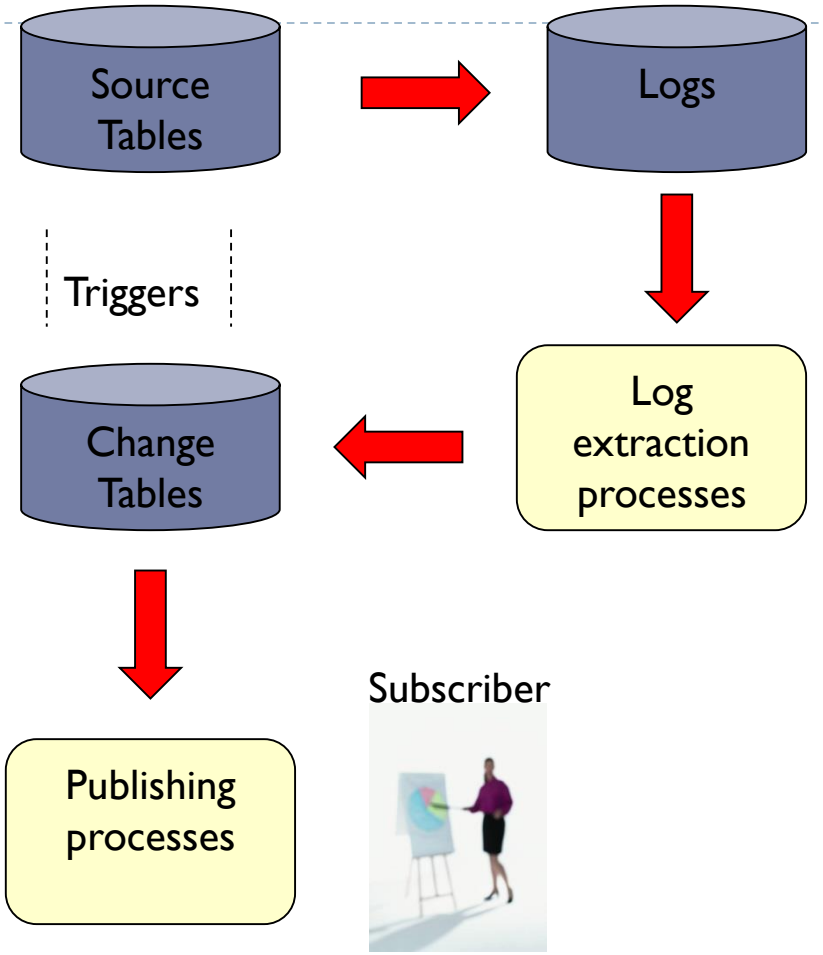

• Design objects and relationships
• Dependencies
• Documentation

Data Profiling

- Descriptive statistics and distribution
- Null values
- Uniqueness
- Pattern matching coverage
- Field relationships

Change Data Capture

Publisher



The Future of DW

- Sourcing…
  - Web, social media, big data
  - Open source software
  - SaaS (software as a service)
  - Cloud computing
  - DW appliances
- Infrastructure…
  - Real-time DW
  - Data management practices/technologies
  - In-memory storage
  - In-memory processing
  - Sandboxes
  - Advanced analytics

Real-time/active/right-time DW/BI

- Enabling real-time data updates for real-time analysis and real-time decision making is growing rapidly
  - Push vs. Pull (of data)
- Concerns about real-time BI
  - Not all data should be updated continuously
  - Mismatch of reports generated minutes apart
  - May be cost prohibitive
  - May also be unachievable

Teradata active DW

Real time/active/right time

For many people, the "real-time" term is synonymous with "instantaneous."

This interpretation, however, is incorrect when applied to data warehousing. While some warehouse data may be captured and entered into the warehouse in seconds or minutes, much of it is not.

For example, some source systems, such as a legacy COBOL program that is updated once a month, can never be more real-time than when last updated. Some data may be prohibitively expensive or difficult to make real-time. Most importantly, there may not be a business need for real-time data. Data only needs to be as fresh as the business requirements. For these reasons, some people prefer the "right time" term.