Module 4
Data Integration Concepts, Processes, and Techniques

Lesson 1: Concepts of Data Integration Processes

Lesson Objectives

- Explain diagrams for refresh processing and typical tasks

- Discuss difficulties of initial population of a data warehouse

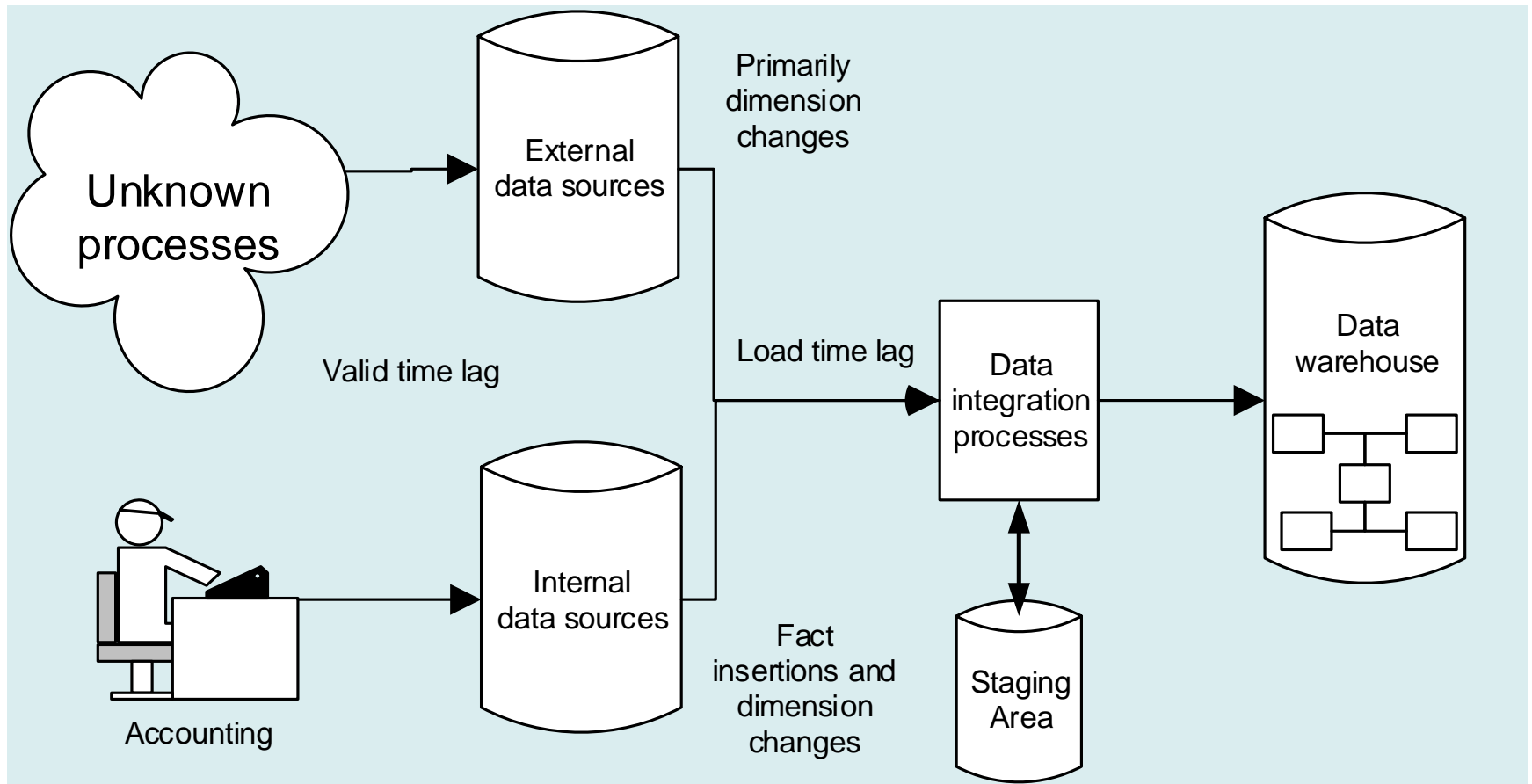- Understand tradeoffs and constraints in managing refresh processing
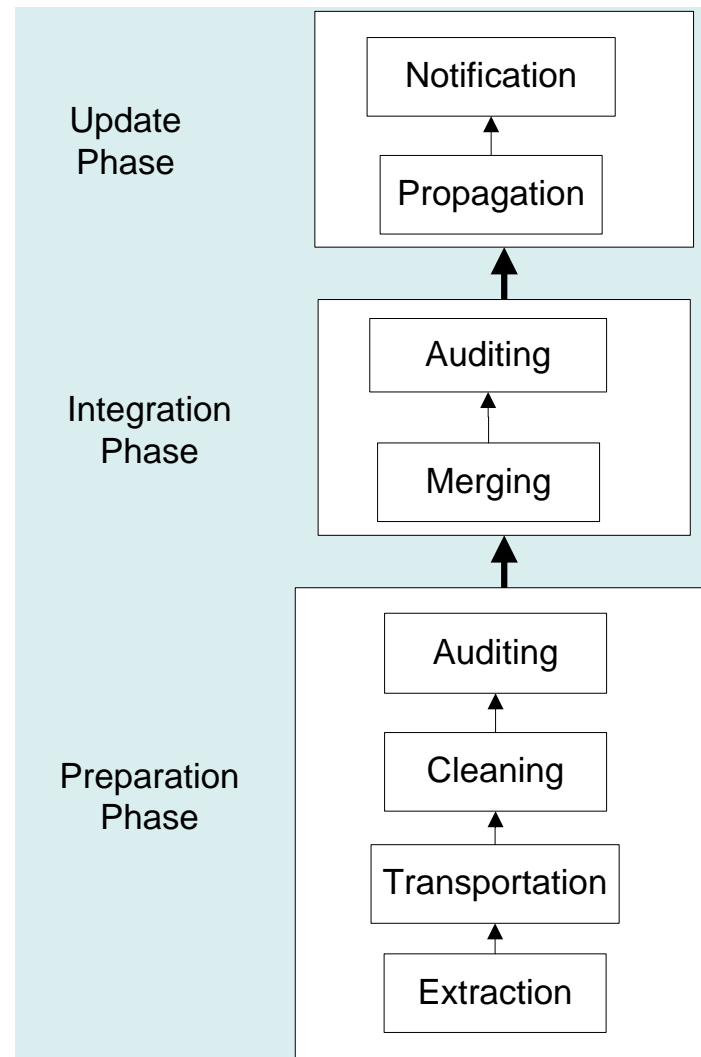
Motivation for Data Integration

- 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

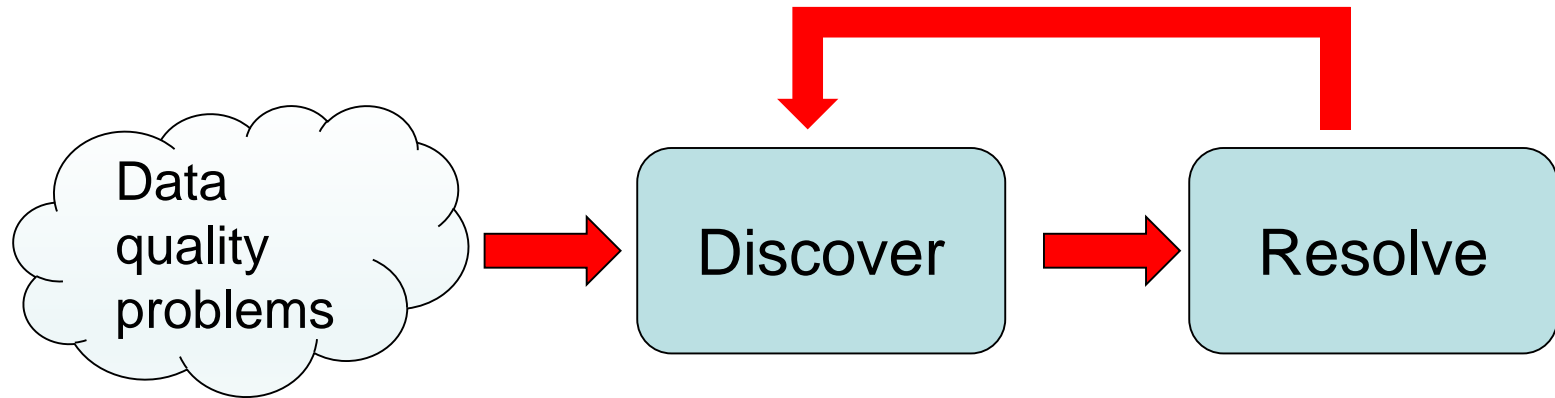# Refresh Processing

Unknown processes

External data sources

Primarily dimension changes

Valid time lag

Accounting

Internal data sources

Fact insertions and dimension changes

Load time lag

Data integration processes

Staging Area

Data warehouse

Refresh Workflow

Update Phase

```
Notification
    ↑
Propagation
```

Integration Phase

```
Auditing
    ↑
Merging
```

Preparation Phase

```
Auditing
    ↑
Cleaning
    ↑
Transportation
    ↑
Extraction
```

Initial Data Warehouse Load

```
Data quality problems  →  Discover  →  Resolve
                              ↑_____|
```
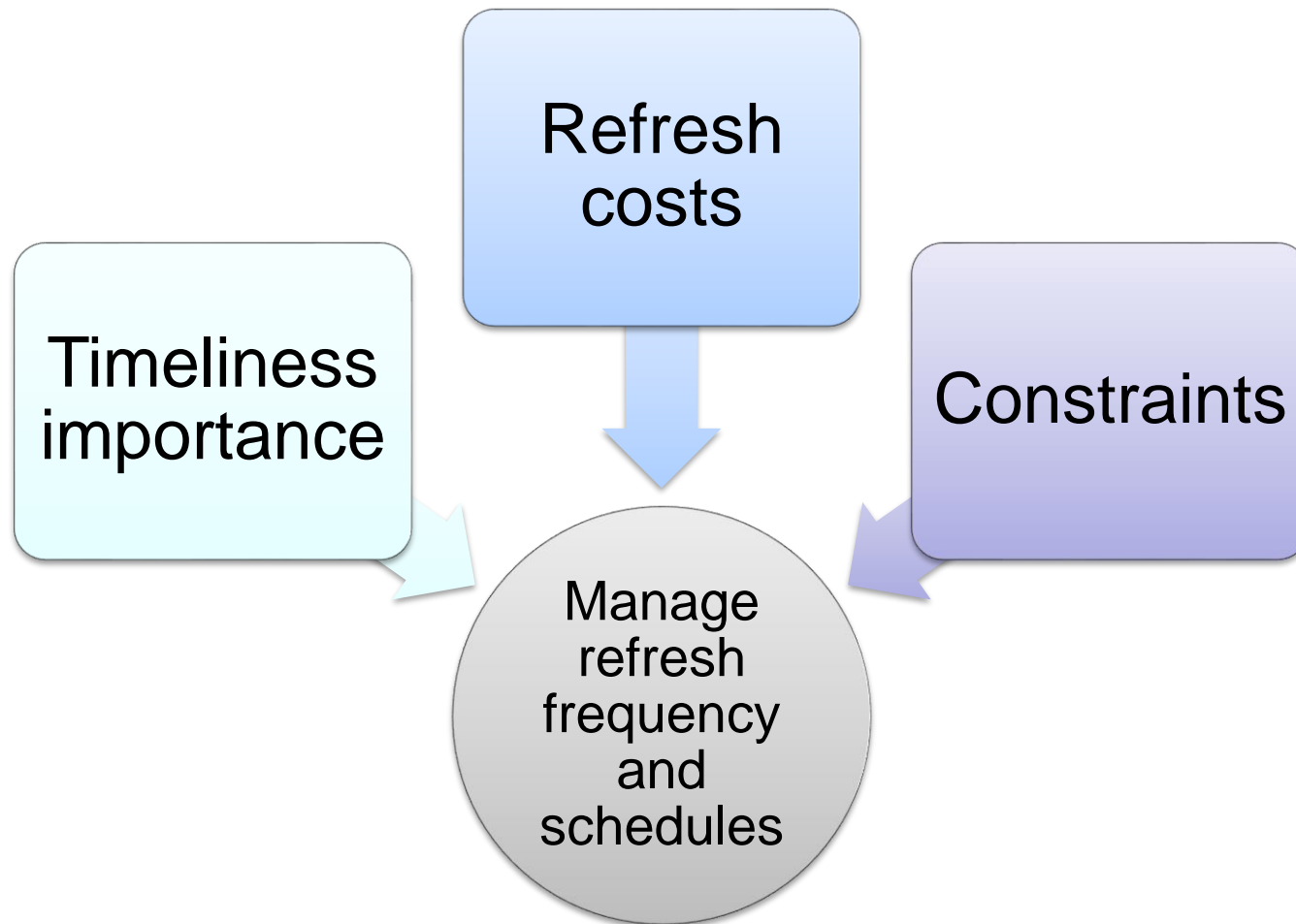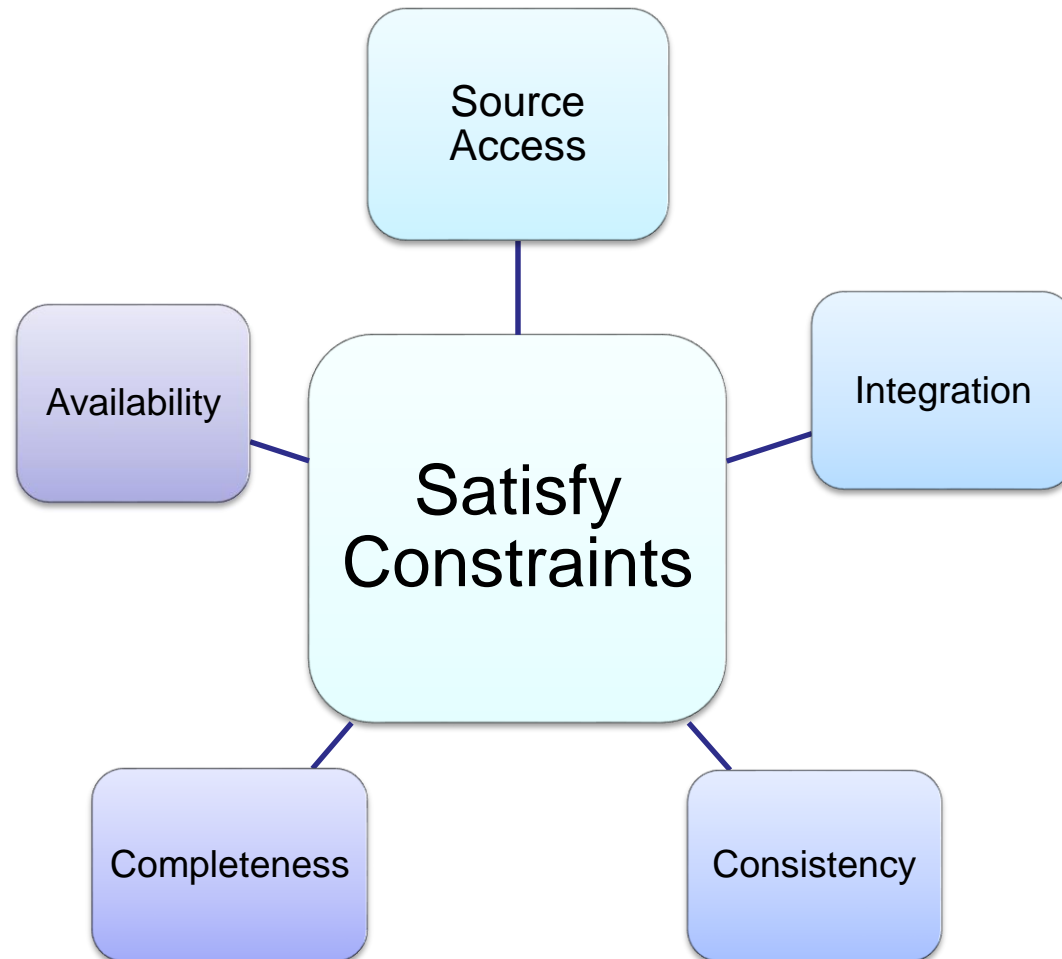
- Major development activity
- More open ended than refresh with difficult to estimate time requirements
- Use profiling tools to discover data quality problems
- Perform for major data warehouse extensions

Refresh Processing Decision Making

Refresh costs

Timeliness importance

Constraints

Manage refresh frequency and schedules

Refresh Constraints

Business School
UNIVERSITY OF COLORADO **DENVER**

Information Systems Program

Summary

- Add value by transformations
- Tedious but important processes for population and refreshment
- Substantial investments in technology and effort
- Data integration tools to improve productivity