

Data and Information Integration

CIS 8020 Systems Integration

Jack G. Zheng
September 2009

Data Integration Problems

- The need to bring together different data/information (sources)
 - Autonomous
 - Distributed
 - Different
- Level of differences (heterogeneity)
 - System
 - Syntax
 - Structural
 - Semantic

System Heterogeneity

- Difference of data storage and management hardware and software (physical difference)
 - Hardware
 - Operating system
 - File system
 - Database management system
- Example
 - Data in SQL Server on a Windows server, and data in Oracle on a Linux server.

Syntax Heterogeneity

- Difference of syntactic format of data information (logical difference)
 - Relational
 - Hierarchical
 - XML (tag or namespace difference)
 - Flat file
 - CSV
 - Objects
 - Binary
- Example
 - Data in SQL Server, and Data in MS Excel
 - RSS 1.0 vs. RSS 2.0

Structure Heterogeneity

- Difference of data/information representation (logical or conceptual difference)
 - Rooted from the human understanding.
 - How a news item should be represented (described)?
 - What data are needed for employee performance?
- Example
 - RSS and Atom

Semantic Heterogeneity

- Difference of human understanding: the meaning, interpretation or intended use (conceptual difference)
 - Naming difference
 - Meaning difference
 - Value domain difference
- Examples
 - What do you mean by “GPA”?
 - How “GPA” is calculated?
 - Are “GPA”s directly comparable?

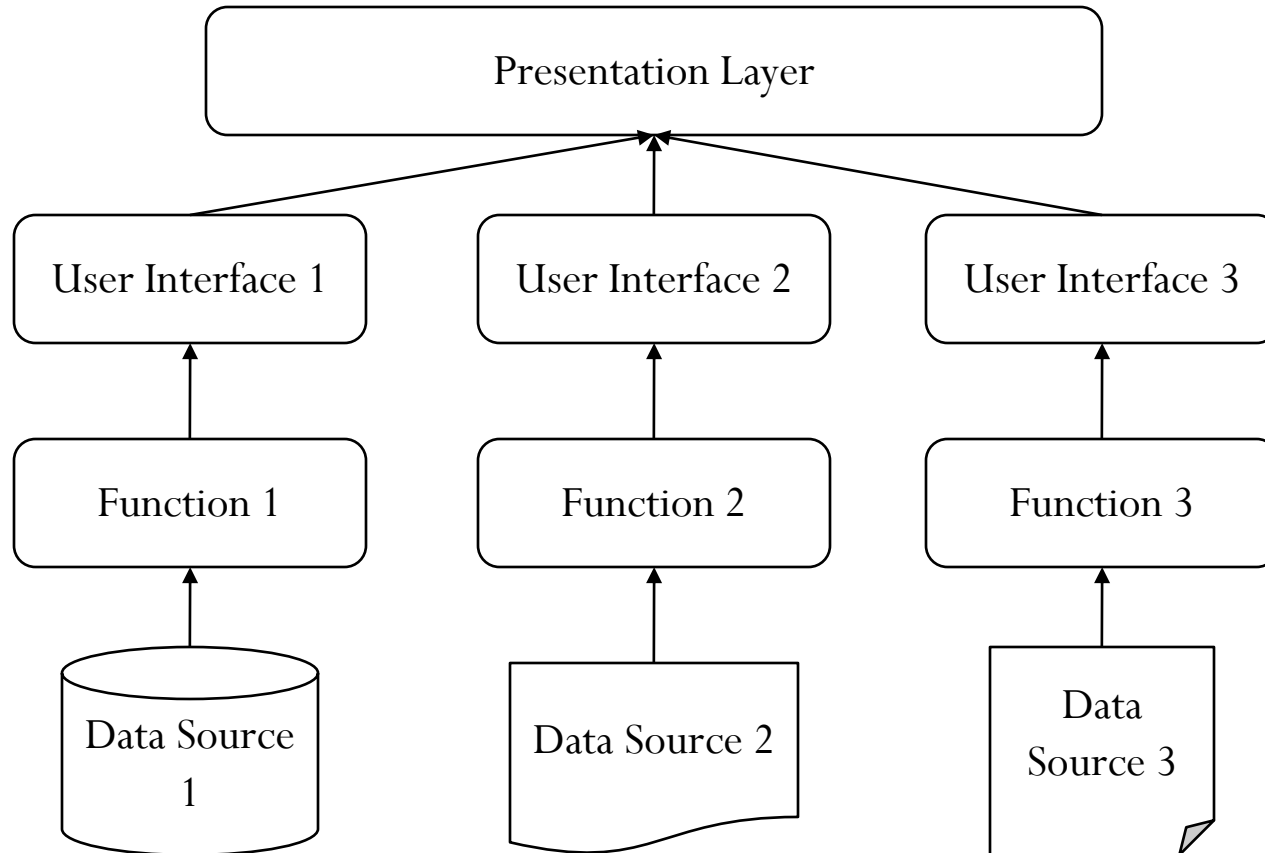
Integration Strategies

- Integration at different layers
- Integration approaches
- Dealing with heterogeneities

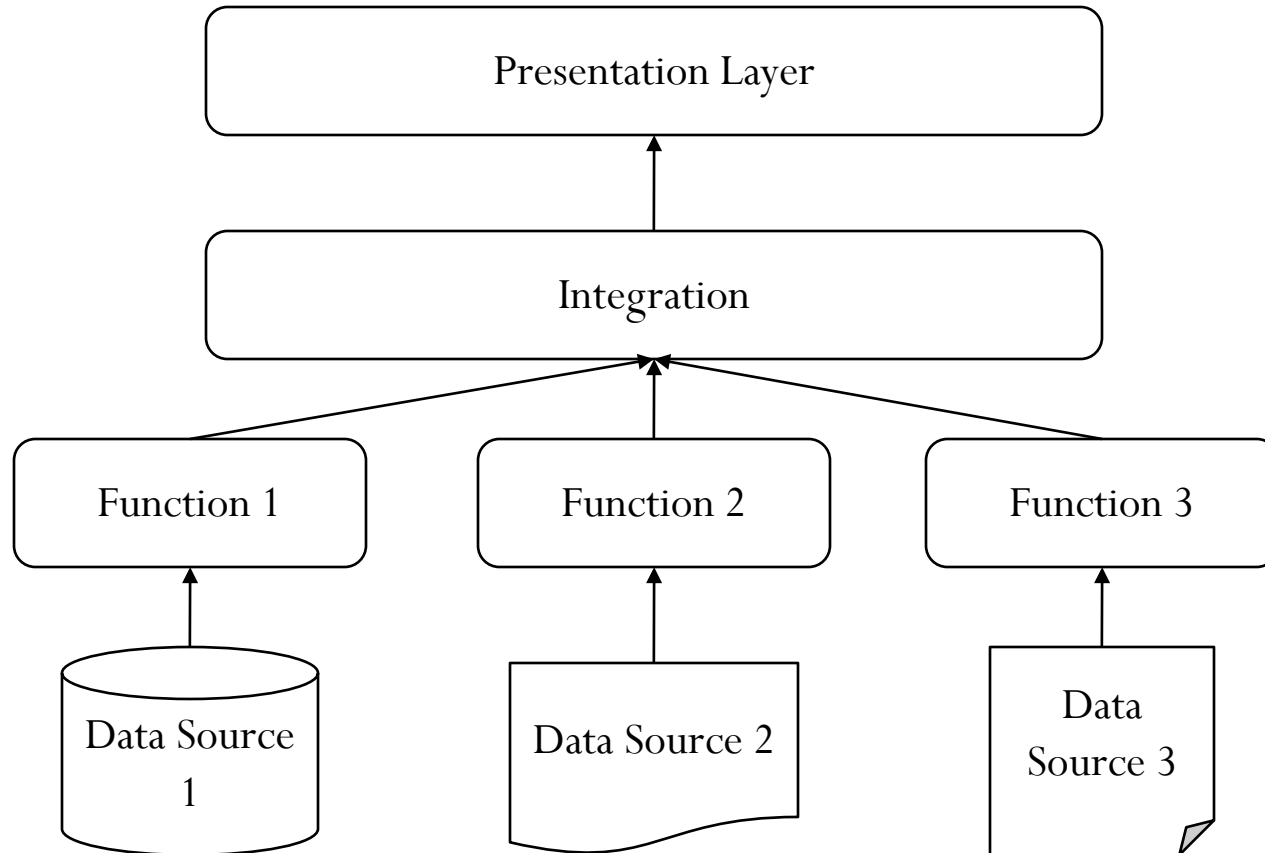
Integration at Different Layers

- Presentation layer
- Data layer
- Function/logic layer

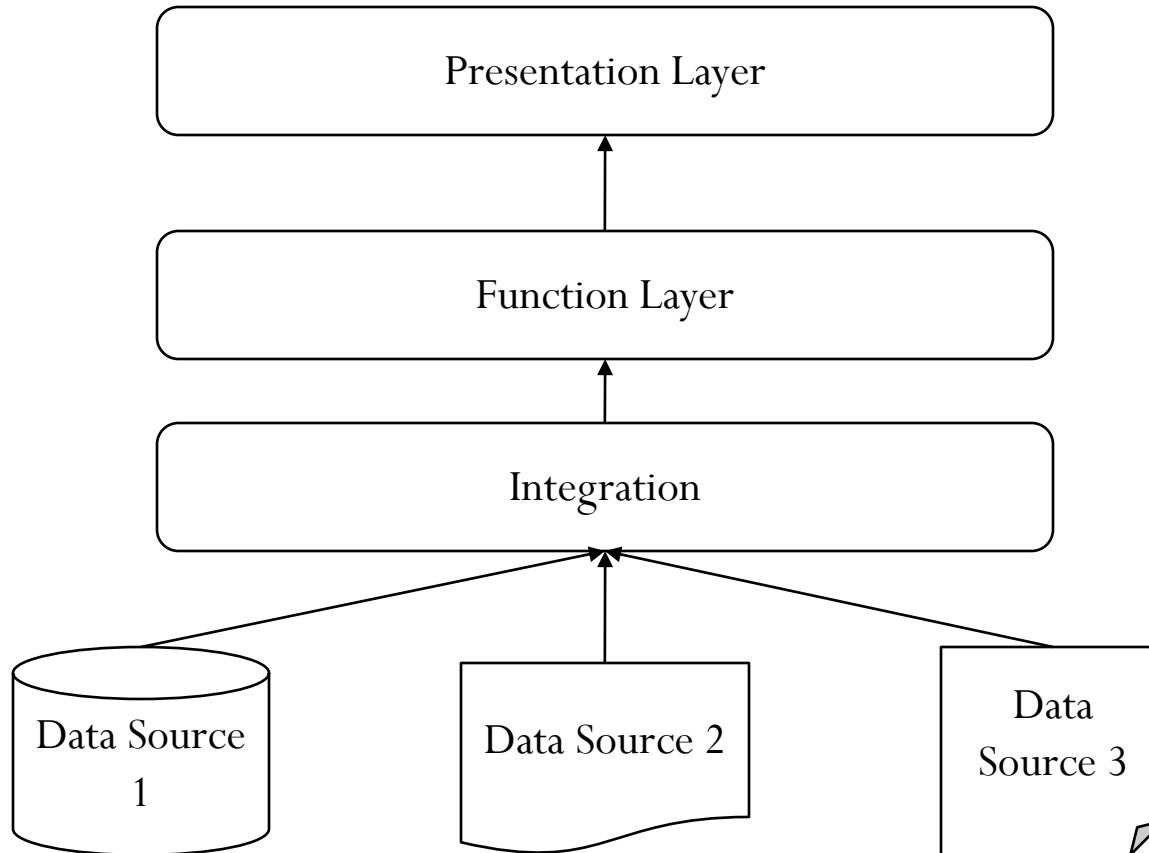
Integration at the Presentation Layer



Integration at the Function Layer



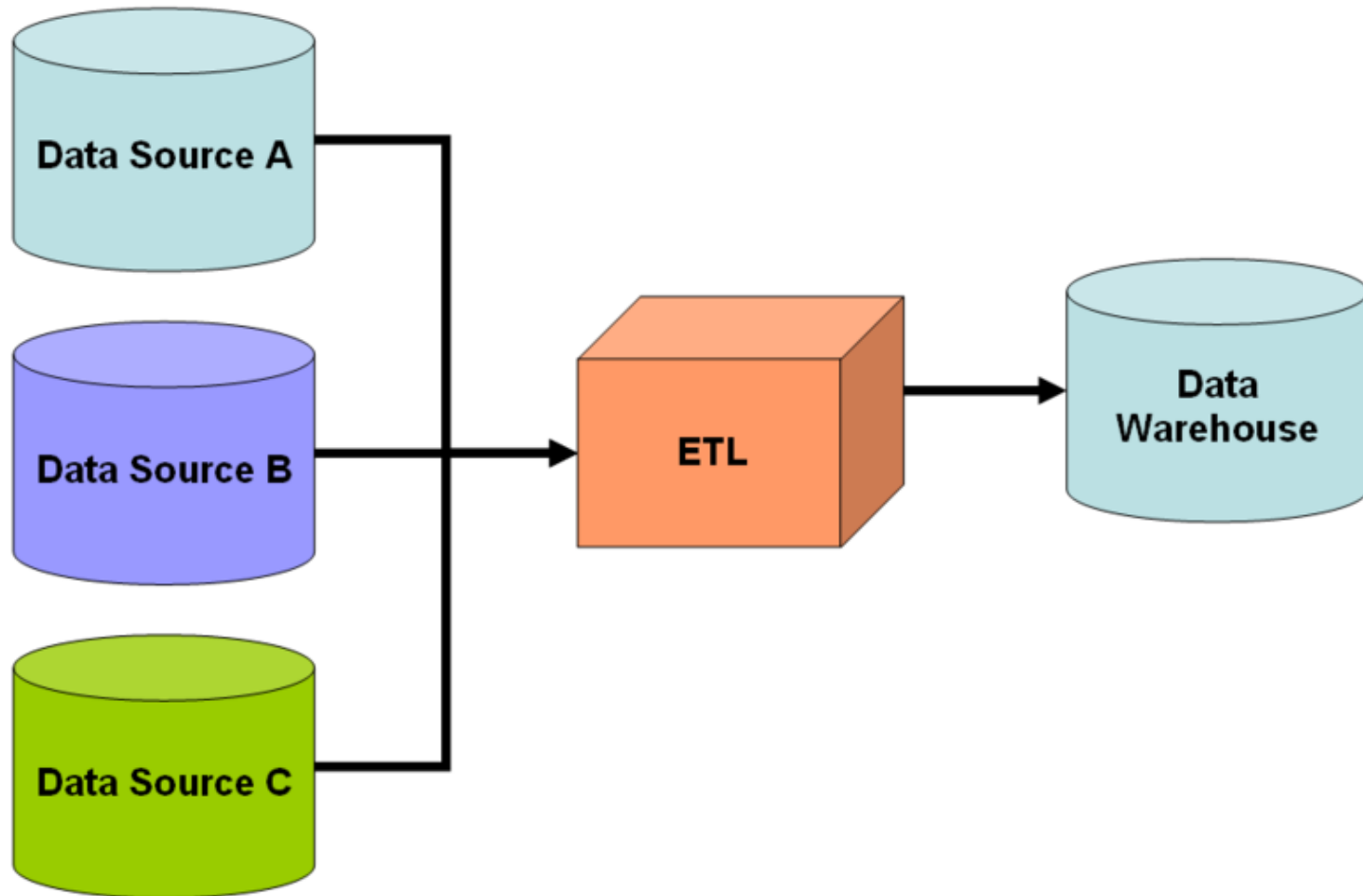
Integration at the Data Layer



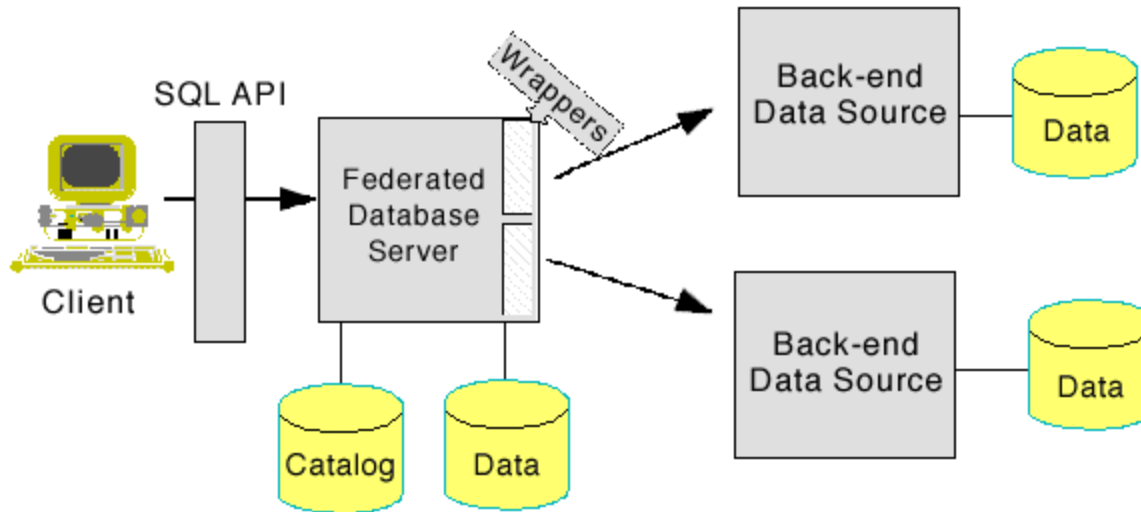
Integration Approaches

- Data warehouse
 - Centralized and replicated data
- Virtual/Federated database (Enterprise Information Integration)
 - Using wrapper or mediators

Data Warehouse Approach



Virtual/Federated Database



Dealing with Syntactic Difference

- Data format transformation/conversion
- Serialization and de-serialization
- XML is often used as the mediating format

Dealing with Structural Differences

- Using meta-data or schema
- Standard schema
 - Dublin Core
 - RDF
- Schema mapping/reconciliation
 - Mapping to a global schema
 - Direct mapping: XML transformation

Semantic Integration

- The most difficult problem
- Clustering based approaches
- Ontology based approaches