System Integration

Mini Case Studies

Data Integration

Shawn A. Butler, Ph.D.
Senior Lecturer, Executive Education Program
Institute for Software Research
Carnegie Mellon University

Objectives

- Understand why data integration is so challenging
- Techniques for data integration
- Understand performance tradeoffs in data integration

Data and Reality, William Kent, 1st Books Library, 2000

Agenda

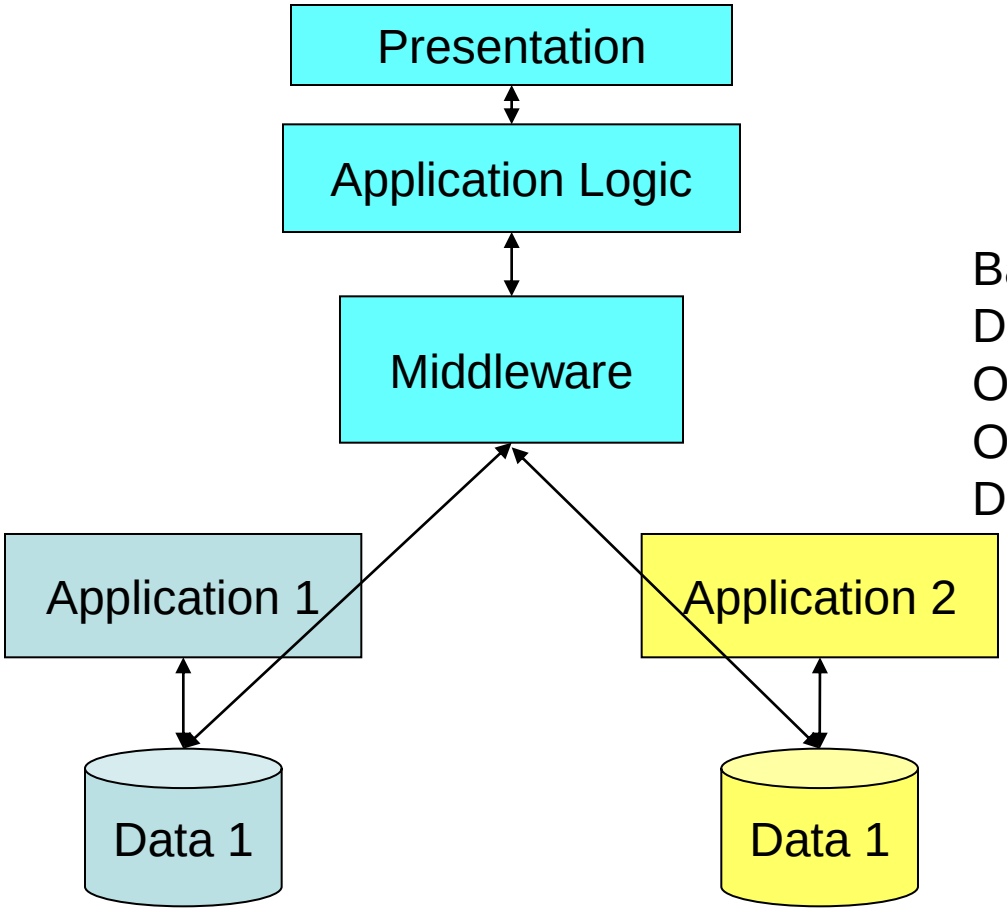
- A quick review of purpose
- Data Integration Architectures
- What does *It* mean?
- Relationships
- Data normalization

"Entities are a state of mind.
No two people agree what the real world view is."
[A. Metaxides]

Two Architectures For Data Integration

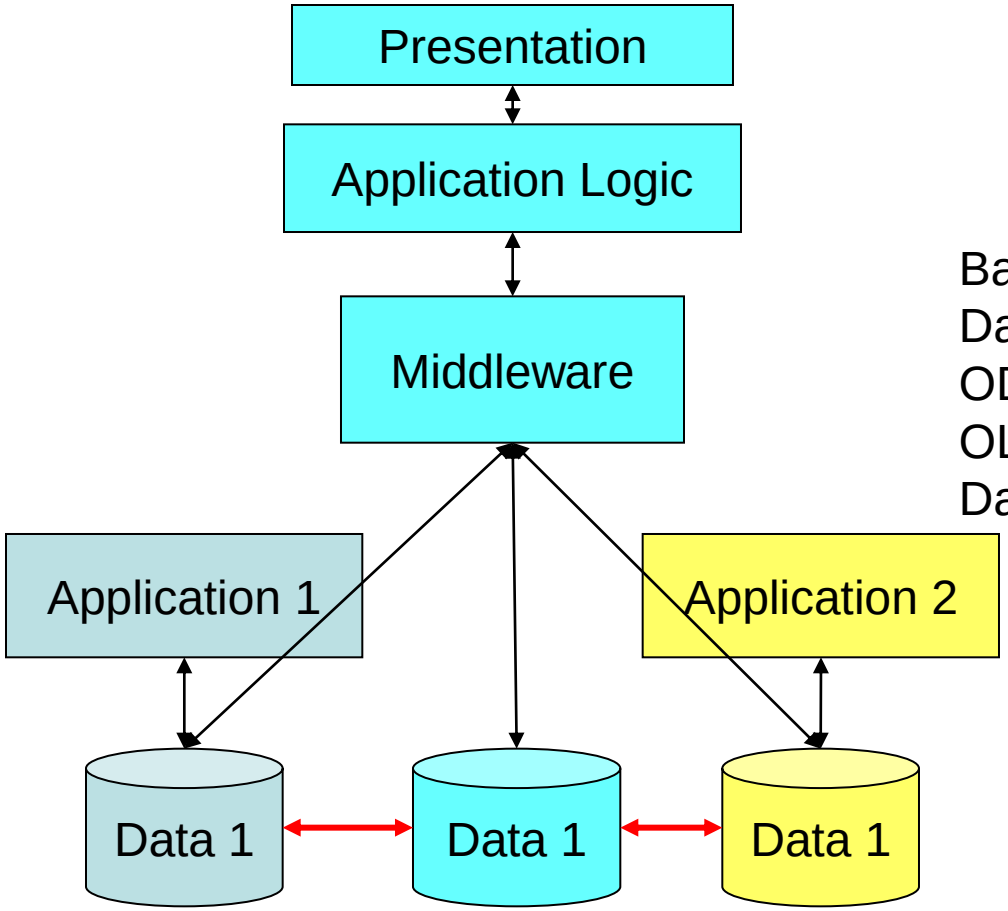
- Problem – Develop an application that aggregates information from several applications

Data Integration Model I



Batch File Transfer
Database Gateway
ODBC
OLAP
Data Transformation

Data Integration Model II



Batch File Transfer
Database Gateway
ODBC
OLAP
Data Transformation

2 Architecture-Many Choices

- Get all the records and aggregate manually
- Aggregate field in a special top level record, application making changes update the field
- System updates the aggregate field in system business logic
- The new application computes the aggregate on retrieval
- Special application “query processor” aggregates the information when needed
- Interface provides this as part of the system

Data Integration Common Mistakes

- Creating yet another database
- Waiting for the data analyst to finish developing the perfect schema
- Implementing the perfectly normalized schema
- Assuming the data exists as described in the documentation
- Testing without a sufficient set of real data
- Assuming that one site is a good representation of data at all sites

What is One Thing?

- We exist in a world of ambiguity
- The system cannot tolerate ambiguity

  - Oneness
  - Sameness – When are two things the same?
  - What is it? – In what categories does it exist?

Meaning

- What is a warehouse?
  - A location within a building?
  - One physical building?
  - Several physical buildings at a single location?
  - A logical concept of where things are stored?

All definitions may be correct, but different among applications.

Meaning

- What is a street?
  - Segments along the physical road may have different names
  - Different streets may have the same name
  - Some roads have discontinuous segments
  - Is a street terminated by a city, county, state?
  - Does street imply motor traffic?
  - Does it also mean freeways, highways, expressways, toll ways, circles, etc.

Meaning

- How do we think of skills?
  - What we know how to do is usually quite varied
  - Categorization arbitrarily limits how that information is defined
  - Categorization arbitrarily limits the number of skills that can be described

How Many Things is *It*?

- A person is:
  - An employee
  - Spouse
  - Shift supervisor
  - Stockholder
- A warehouse is:
  - Physical building
  - A place where parts are stored

When does change make *It* different?

- Car
  - Does a different color make it a different car?
  - Does a new engine make it a different car?
  - Slowly replace the parts of a car, at what point is it a different car?
  - Is the essence of change the same in two different systems?
- Parts
  - When does it matter when the component parts are different?
- Versions
  - At what point does a version move to a different product?

Change

- Sometimes our perception changes
  - When do two things become one?

Categories, Attributes, Relationships

- Categories (aka types)
  - Require arbitrary decisions
  - Categories often have subsets
  - Overlap with other categories
    - Plaintiffs are people, corporations, government agencies…)
  - Once categories are established real things are assigned
    - Employee (part time, fulltime, managers, etc…)
  - Items change in a category, does it still fit the category

Categories, Attributes, Relationships

- Attributes may be part of the category or they may be an essential part of the categorization
  - Cars, by the way, have wheels
  - Not all cars have wheels
- Relationships
  - Relationship is not meaningful between things within a category
  - Relationship is meaningful, but cannot be related to itself
  - Relationship is meaningful, and things can be related to themselves

Relationships

- Optional versus Mandatory
- Transitivity
- Symmetry and Anti-symmetry
- Implication
- Subset consistency
- Constraints
- Attributes of the relationship
- Names
  - No name
  - One name
  - Multiple names

Relational Databases

- Database normalization rules are designed to prevent anomalies and inconsistencies in databases
- Database normalization rules, strictly applied, may introduce inefficiencies in database design
- All databases have various schemas that work well for their application, but don’t combine well into an efficient schema for all applications
- Database design is always a tradeoff among application performance optimization

Normalization of Databases

- First normal form – All record types must contain the same number of fields

Transportation	John Smith	Anne Harbor	Connie Redwood	
Accounting	Mark Johnson	Mary Ebner		
Logistics	Sally Worth	Chris Waters	Tracy Elmore	Judd Heron

Normalization of Databases

- Second Normal Form – Non-key field is not related to a key field

Employee	Department	Manager	Dept_Address

1. The Department address is repeated in every record that refers to an employee in that Department.
2. If the address of the Department changes, every record referring to a part stored in that Department must be updated.
3. Because of the redundancy, the data might become inconsistent, with different records showing different addresses for the same Department.
4. If at some point in time there are no employees in a Department, there may be no record in which to keep the Department's address.

Normalization of Databases

- Third normal form – Non-key field is not a fact about another non-key field

Employee	Department	Manager	Department Address

Employee	Department	Manager

Department	Department Address

Summary

- Data integration will most likely be the most difficult challenge of a system integration project
- Data integration is difficult because each data source will have their own views of what *It* means
- There will always be a balance between performance and normalization of the data model
- Don’t let the data tail wag the system dog