ETL

Data Warehousing

23/11/2010

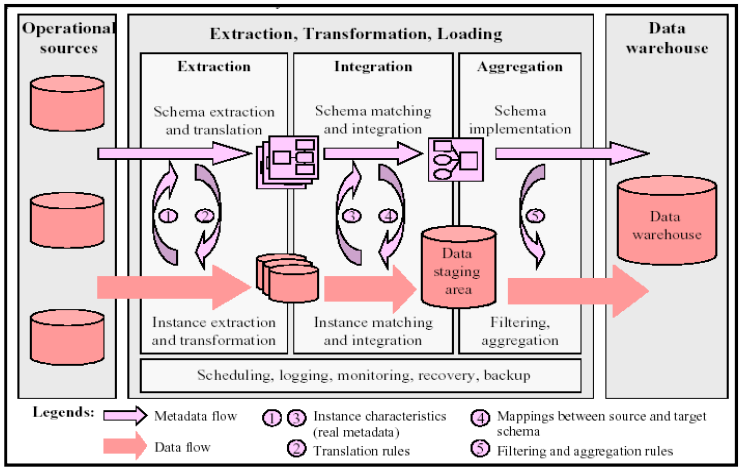
Part 1

- The process of ETL
- Data Transformation
- Schema Matching and Integration
- Some Formal Definitions
- Schema Matching Approaches
- Schema-Level Approaches
  - Granularity of match (element-level vs. structure-level)
  - Match cardinality
  - Linguistic approaches
  - Constraint-based approaches
- Combing Matchers

Part [1] Based on

- Rahm, E., and P. A. Bernstein, "A Survey of Approaches to Automatic Schema Matching," VLDB Journal 10, 4 (Dec. 2001), pp. 334-350
- Erhard Rahm and Hong Hai Do, "Data Cleaning: Problems and Current Approaches"

The Process of Extract – Transform – Load



Data Transformation

- ETL is essentially a process of acquiring data from OLTP.
- Technically, ETL is a Data Transformation process.
- Where Data Transformation is required?
  - Migrate legacy systems to modern applications
  - Optimize queries
  - Translate from one data model to another
  - Integrate heterogeneous systems into federated databases or warehouses
  - Perform data cleansing or scrubbing
  - Evolve a schema and its associated database as driven by changing user requirements
  - Construct user-customized web sites
  - Achieve enterprise-wide integration

The Process of Extract – Transform – Load

- Solutions in ETL
  - Schema Integration and Matching
  - Data Cleansing
  - Data Loading
- A number of strategies for each of these solution.
- Let's start with Schema Integration & Matching

Schema Integration & Matching

- Fundamental Schema Matching Operator - *Match*
  - Input: Multiple, Heterogeneous Schemas
  - Output: Mappings
- Application domain
  - Schema Integration: Structures and Terminological relationships
  - Data warehouses: Source-to-warehouse Transformation
  - E-commerce: Message Translation
  - Semantic query processing: A Run-time Scenario

Some Formal Definitions

- A schema is a set of elements connected by some structure represented by a particular physical model.
- A mapping is a set of *mapping elements*, each of which indicates that certain elements of a schema, say S1, are mapped to certain elements in the other schema, say S2.
- Each mapping element can have a *mapping expression* which specifies how the S1 and S2 elements are related.

Some Formal Definitions (contd.)

S1 Elements	S2 Elements
Table: Cust	**Table: Customer**
C#	**CustID**
CName	**Company**
First Name	**Contact**
Last Name	**Phone**

- Mapping Example
  - Mapping element relating Cust.C# to Customer.CustID
  - Mapping expression Cust.C# = Customer.CustID
- Match operation is a function that takes two schemas S1 and S2 as input and returns a mapping between those two schemas, called the *match result*.

Schema Matching Approaches

- Schema Level Approaches
  - Consider schema-level information only.
  - Information includes the usual properties of schema elements, such as name, description, data type, relationship types (part-of, is-a, etc.), constraints, and schema structure
  - Heavy Metadata usage
- Instance Level Approaches
  - Matching approaches that consider instance data (i.e., data contents).
  - Especially useful when schema information is limited, as is often the case for semi structured data.

Schema-level Approaches

- Granularity of match (element-level vs. structure-level)
- Match cardinality
- Linguistic approaches
- Constraint-based approaches