ETL Tools

- ETL = Extract – Transform – Load
- Typically: data integration software for building data warehouse
- Pull large volumes of data from different sources, in different formats, restructure them and load into a warehouse
- A variety of tools:
  - major database vendors (IBM, Microsoft, Oracle)
  - independent companies (Informatica)
  - Open source (e.g. Clover ETL)
- Significant demand: $1.5B/year, with >15% annual growth rate

ETL tools cont'd

Emphasis on:

- data quality (in particular cleaning and profiling tools)
- transformations between specific formats
- latency requirements (towards real-time)

Much less (currently) emphasis on:

- nontrivial transformations
- proper query answering

Market:

- leaders: Informatica, IBM
- catching up: Microsoft, Oracle

IBM

- Product name: InfoSphere DataStage
- Main claims:
  - variety of data sources (almost any database, text, XML, web services)
  - capable of handling data arriving in real-time
  - scalability
- Unix (Linux) and Windows Platforms

InfoSphere DataStage cont'd

- InfoSphere – product line that includes software from WebSphere and Information Server lines.
- Includes lots of other things
  - application integration and transformation
  - online marketing tools
  - mobile, speech middleware
  - business process management
  - change data capture
  - information analyzer
  - data quality tools

InfoSphere Federation Server

- Federated (virtual) integration: “Access and integrate diverse data and content sources as if they were a single resource - regardless of where the information resides.”
- Integration across different relational products (db2, Oracle, SQL server)
- Integrity and accuracy guarantees
- Distributed query optimizer
- XML support
- Security strategies
- These are expensive products (>US$60K license)

IBM's view of data integration

- Key tasks, with associated products
- Tasks:
  - Connect to information (products: information server; data publisher)
  - Understand information (data architect, models for ... (banking, insurance, retail, telecom))
  - Cleanse information (QualityStage: matching engine, cleaning rules etc)
  - Transform information (DataStage)
  - Deliver information (Federation Server, DataStage)

IBM: data exchange

- Clio Project (IBM Almaden Research Center).
- Includes:
  - a semi-automatic schema mapping generation tool
  - universal solutions are the semantics of data exchange
  - they are generated by extended SQL queries
  - Extension: Skolem functions
- Part of IBM Product “Rational® Data Architect”
- Other features:
  - discovery of attribute correspondence; interactive construction of mappings
  - Extended schemas (not full XML but more than relations)

Microsoft

- Integration Services – part of SQL Server (SSIS)
- Supports multiple formats; converts everything into tabular format
- Transformations:
  - join, union
  - sort
  - aggregate
  - lookup
  - convert
- Has a data quality tool
- Goes beyond traditional ETL: e.g., data and text mining tools

Oracle

- Oracle Warehouse Builder (OWB)
- Data integration and metadata management tasks:
  - Extraction, transformation, and loading (ETL) for data warehouses
  - Migrating data from legacy systems
  - Designing and managing corporate metadata
  - Data profiling
  - Data cleaning
- Included in the Oracle database product.

Oracle: transformations

- Scalar value transformations (plenty of predefined ones):
  - Characters
  - Conversions
  - Dates
  - Numbers
  - Spatial objects
  - XML transformations (from very simple – select nodes by XPath expressions – to very complex, such as applying XSLT style sheet)
- Also user-defined (functions, procedures, packages)

Informatica

- Market leader – Informatica PowerCenter
- Provides support for
  - migration
  - synchronization
  - warehousing
  - cross-enterprise integration
- Works with multiple data formats
- Provides support for metadata management
- Real-time capabilities

Informatica: Transformation language

- Main orientation: scalar value transformations
- Functions: change data in a mapping
- Operators: create transformation expressions
- Syntax is SQL-based
- Part of it is essentially a programming language in a Java-like syntax for manipulating values.
- Roughly: looks at a portion of the source data, modifies it, and changes the target data accordingly.

Informatica: Transformation language cont'd

- `DD_DELETE` and `DD_INSERT` specify what to do with data items.
- E.g., `IIF(job='CEO', DD_DELETE, DD_INSERT)` says: items with job being CEO are marked for deleting, others for insertion.
- Operators:
  - Arithmetic
  - String
  - Comparisons
  - Logical
  - (almost) everything you can imagine
- Many functions for dealing with dates in different formats.

Informatica: Transformation language cont'd

- Large number of functions
- Aggregates: AVG, COUNT, MIN, MAX, MEDIAN, PERCENTILE, STDDEV, SUM, etc.
- Character functions: CONCAT, LENGTH, TRIM, etc
- Conversion functions (e.g., TO_CHAR for Date, TO_DECIMAL, TO_FLOAT, TO_DATE)
- Date functions: ADD_TO_DATE, DATE_DIFF, DATE_COMPARE, etc
- Numerical: the usual suspects.
- Scientific: SIN, COS, TAN, etc
- Search for a value in the source: LOOKUP
- This was quick; full manual – almost 250 pages.

Summary

- Complex tools; very good at transforming data values, and at working with specific formats (MS Word, Excel, PDF, UN/EDIFACT, RosettaNet, etc) and for specific industries (finance, insurance, health)
- Much better these days at getting real-time data; very good at bulk loading, supporting multiple formats
- Not so good:
  - virtual integration
  - complex structural transformation
  - query answering
  - metadata management
- A lot of effort will be put there over the coming years