Data Extraction, cleanup, and Transformation Tools

Tools Requirements

The tools that enable sourcing of the proper data contents and formats from operational and external data stores into the data warehouse to perform a number of important tasks that include

- Data transformation from one format to another on the basis of possible differences between the source and the target platform.
- Data transformation and calculation based on the application of the business rules that force certain transformations. Examples are calculating age from the date of birth, or replacing a possible numeric gender code with a more meaningful “male” and “female” and data cleanup, repair, and enrichment, which may include corrections to the address field based on the value of the postal zip code.

- Data conversations and integration, which may include combining several source records into a single record to be loaded into the warehouse.
- Metadata synchronization and management, which includes storing and/or updating metadata definitions about source data files, transformation actions, loading formats, and events etc.,
- When implementing a data warehouse, several selection criteria that affect the tools ability to transform, consolidate, integrate, and repair the data should be considered.
- The ability to identify data in the source environments that can be read by the conversion tool is important.
  - Support for flat files, indexed files
  - The capability to merge data from multiple data stores
  - The ability to read information from data dictionaries or important information from repository products is desired

- The code generated by the tool should be completely maintainable from within the development environment.
- Selective data extraction of both data elements and records enables uses to extract only the required data.
- A field-level data examination for the transformation of data into information is needed.
- The ability to perform data-type and character-set translation is a requirement when moving data between incompatible systems.
- The capability to create summarization, aggregation, and derivation records and fields is very important.
- The data warehouse database management system should be able to perform the load directly from the tool, using the native API available with the RDBMS.
- Vendor stability and support for the product are items that must be carefully evaluated.

Vendor Approaches

The integrated solutions can fall into one of the categories described below

Code generators

Database data replication tools

Rule-driven dynamic transformation engines capture data from source systems at user-defined intervals, transform the data, and then send and load the results into a target environment, typically a data mart

Access to Legacy Data

Many organizations develop middleware solutions that can manage the interaction between the new applications and growing data warehouses on one hand and back-end legacy systems in the other hand.

A three architecture that defines how applications are partitioned to meet both near-term integration and long-term migration objectives.

- The data layer provides data access and transaction services for management of corporate data assets.
- The process layer provides services to manage automation and support for current business process.
- The user layer manages user interaction with process and /or data layer services.

Vendor Solutions

Prism Solutions

Provides a comprehensive solution of data warehousing by mapping source data to a target database management system to be used as warehouse.

Warehouse Manager generates code to extract and integrate data, create and manage metadata, and build a subject-oriented, historical base.

Prism Warehouse Manager can extract data from multiple source environments including DB2, IDMS, IMS, VSAM, RMS and sequential files under UNIX or MVS. Target databases include ORACLE SYBASE, and INFIRMIX

SAS Institute

SAS tools to serve all data warehousing functions.

Its data repository function can act to build the informational database.

SAS Data Access Engine serve as extraction tools to combine common variables, transform data representation forms for consistency, consolidate redundant data, and use business rules to produce computed values in the warehouse.

SAS engines can work with hierarchical and relational databases and sequential files

Carleton Corporation's PASSPORT and MetaCenter.

PASSPORT.

PASSPORT is sophisticated metadata-driven, data-mapping and data-migration facility.

PASSPORT Workbench runs as a client on various PC platforms in the three-tiered environment, including OS/2 and Windows.

The product consists of two components.

The first, which is mainframe-based, collects the file, record, or table layouts for the required inputs and outputs and converts them to the Passport Data Language (PDL).

Overall, PASSPORT offers

- A metadata dictionary at the core of the process.
- Robust data conversion, migration, analysis, and auditing facilities.
- The PASSPORT Workbench that enables project development on a workstations, with uploading of the generated application to the source data platform.
- Native interfaces to existing data files and RDBMS, helping users to leverage existing legacy applications and data.
- A comprehensive fourth-generation specification language and the full power of COBOL.

The MetaCenter.

The MetaCenter, developed by Carleton Corporation in partnership with Intellidex System, Inc., is and integrated tool suite that is designed to put users in control of the data warehouse.

It is used to manage

- Data extraction
- Data transformation
- Metadata capture
- Metadata browsing
- Data mart subscription
- Warehouse control center functionality
- Event control and notification

Vality Corporation

Vality Corporation's Integrity data reengineering tool is used to investigate, standardize, transform, and integrate data from multiple operational systems and external sources.

- Data audits
- Data warehouse and decision support systems
- Customer information files and house holding applications
- Client/server business applications such as SAP, Oracle, and Hogan
- System consolidations
- Rewrites of existing operational systems

Transformation Engines

Informatica

Informatica's product, the PowerMart suite, captures technical and business metadata on the back-end that can be integrated with the metadata in front-end partner's products. PowerMart creates and maintains the metadata repository automatically.

It consists of the following components

PowerMart Designer is made up of three integrated modules- Source Analyzer, Warehouse Designer, and Transformation Designer

PowerMart Server runs on a UNIX or Windows NT platform.

The *Information Server Manager* is responsible for configuring, scheduling, and monitoring the Information Server.

The *Information Repository* is the metadata integration hub of the Informatica PowerMart Suite.

Informatica PowerCapture allows a data mart to be incrementally refreshed with changes occurring in the operational syste, either as they occur or on a scheduled basis.

Constellar

The Constellar Hub is designed to handle the movement and transformation of data for both data migration and data distribution in an operational system, and for capturing operational data for loading a data warehouse.

Constellar employs a hub and spoke architecture to manage the flow of data between source and target systems.

Hubs that perform data transformation based on rules defined and developed using Migration Manager

Each of the spokes represents a data path between a transformation hub and a data source or target.

A hub and its associated sources and targets can be installed on the same machine, or may run on separate networked computers.