INTRODUCTION

Data warehousing can be defined as:

A copy of data specifically structured for querying and reporting.

In most cases, the data is transactional data pulled from a number of different applications in the enterprise, but it can be any type of data. The goal of a data warehouse is to provide an enterprise with high quality information that they can use to make tactical and strategic decisions.

Most often the data is extracted from a variety of systems and applications to a central relational repository. It is not unusual for a corporation’s applications to reside on different platforms such as mainframes, UNIX, AS/400, and Windows. These applications store data in different databases or non-relational data formats. The data most often is stored in different formats using different codes to mean the same thing. For example, in application A gender may be stored as “Male”, but in application B gender is stored as “M”, and in application C it’s stored as “01”. A data warehouse converts the data to a uniform format and provides the ability to view, compare, and correlate information at an enterprise level.

Data Warehouses can be used as application feeds. Often a data ware house provides content for information displayed on portal pages. Additionally, data warehouses can be the back end to custom developed EIS (enterprise information systems).

THE PROCESS

There are five major steps in a data warehouse process. The first three steps consist of the building and subsequent loading of the data warehouse. This is commonly referred to as Extraction, Transformation, and Loading (ETL). The last two steps consist of creating data marts and reporting.

1. Extracting data from applications.
   In this step the data is extracted from the application’s data structure and most often placed in a temporary data structure or occasionally left in a file.
2. Transforming, or staging, the data (getting it in the proper format, often referred to as “transforming”).
   In the transformation step, the different data formats are reconciled. For example, all data formats for gender are transformed to either "Male", "Female", "Other".
3. Loading the transformed data into the data warehouse.

4. Creating data marts from the data warehouse.
5. Reporting (turning the data into information)
   The steps are discussed in greater detail later in this paper under the section titled Background.

THE PROBLEM

Running a data warehouse requires coordinating many operations across many applications, databases, and systems. In a large company, this might be as many as 20 discreet operations that must be performed in the correct sequence, at the correct time, and under the correct conditions. Typical data warehouse operations deal with extremely large amounts of data. It is not unusual for a large organization to load gigabytes of data nightly. Often data warehouses grow to sizes larger than one terabyte. Missing one step in the process, or executing a step at the wrong time, can result in a significant amount of wasted processing time, or in the worst case scenario, bad data.

THE SOLUTION

UC4 Workload Automation Suite is tailor made for automating the data warehousing process.

Each step in the process becomes a module in one or more chains.

The chains can be run on a set schedule, and individual steps in the chains can be executed based on a wide range of criteria such as the number of transactions or the number of records in the data warehouse.

UC4 Workload Automation Suite is application aware and can ensure that application processes are run prior to data extractions.

Data extracts, often long processes, can be checked periodically to ensure that the data being extracted is in the correct format.

Notification of nightly failures can be automated.

Complex recovery sequences can be automated.

Reports based on the data can be run and distributed automatically.

Workload on the system can be dynamically leveled using queue management.

UC4 Workload Automation Suite can interface with ETL (Extract, Transform, Load) tools such as Informatica ensuring that ETL transactions occur at the appropriate times. ETL tool processes can be scheduled with knowledge of application processes and backups.

The entire data warehouse environment, including all applications, scripts, and tools can be fully automated and managed from a central point of control.

BENEFITS

There are many benefits to using UC4 Workload Automation Suite to control a data warehouse, including ensuring the proper execution se-
quence, resulting in fewer errors, fewer recoveries, and higher data warehouse availability.

Scripts currently being used to manage the E-T-L process can be replaced, reducing coding and maintenance time. Fewer scripts means easier maintenance and less dependence on complex scripts that may have been written by one person who has since moved on to another position.

The data warehouse runs more efficiently because product initiates actions based on the state of the application and data warehouse databases, not on set times. Lag time between steps is eliminated.

End users get reports online within minutes of the data being available because UC4 Workload Automation Suite scheduled the generation of the data and knows when it can run the reports.

System resources are maximized because all applications used in the data warehouse environment are scheduled efficiently, leveling the load on the system, and preventing the system from being over loaded.

These benefits hold for new, as well as existing, data warehousing projects.

BACKGROUND
A data warehouse is a very large database containing data from many sources in an enterprise. The data warehouse is at the center of an entire data collection and distribution process. This process:

- extracts the data from enterprise sources such as ERP, CRM, and financial systems. These sources often reside on different platforms such as UNIX, Windows, mainframe, and AS400.
- transforms and cleanses the data so that data coming from the different sources is standardized or "normalized". In the transformation step, the different data formats are reconciled. For example, all data formats for gender are transformed to either "Male", "Female", "Other".
- places the normalized data in a staging area.
- loads the data into the actual data warehouse database and operational data store. The data warehouse is usually a relational database running on a UNIX platform.
- loads the data via OLAP tools into small databases called data marts for fast data access by specific groups within the enterprise. The data is usually stored in data cubes. A data cube presents data based on two or more attributes. For example, it might summarize sales by region (two attributes) or sales by region by product (three attributes). Products such as Cognos PowerPlay create optimized cubes of data for distribution. Data marts are often distributed to remote sites for faster access. Reports are usually generated from the data marts rather than the data warehouse.
- turns the data into information using various reporting tools.

The process is illustrated in the figure below.

Data Warehouse Environment

**Figure A.** The data warehouse environment includes data sources, a staging area that aggregates and transforms the data, a data warehouse that stores the transformed data, data marts that provide quick access to subsets of the data, and reporting tools for turning the data into information.

**AUTOMATING THE DATA WAREHOUSE WITH UC4 WORKLOAD AUTOMATION SUITE**

UC4 Workload Automation Suite can be used to schedule all software used in the data warehouse process:

Enterprise applications such as ERP (Enterprise Resource Planning), CRM (Customer Resource Management), Financials, and other internal applications:

- Backup tools
- E-T-L tools
- OLAP tools
- Reporting tools

In addition, efficiency in the data warehouse process by can be increased by:

- Eliminating the lag time between processing steps.
• Balancing the load placed on the servers by the enterprise applications, E-T-L tools, and reporting tools.
• Eliminating the scripts used to archive data cubes in the data marts.
• Distributing output generated by the reporting tools.

**BENEFITS TO THE DATA WAREHOUSE TEAM**

Building a data warehouse generally requires a team of experts. The different roles and the benefits UC4 Workload Automation Suite brings to each role are described below. One person on the team may function in more than one role.

**Project Manager.** This person will oversee the progress and be responsible for the success of the data-warehousing project.

A script-less approach reduces implementation times and provides the framework for promotion to the production environment. During the maintenance phase, UC4 Workload Automation Suite assists the Project Manager in managing the processes and notifications of failure. The Graphical Analysis Package can be used to analyze the impact of new data source additions. The project manager often reports to the CIO or the applications manager.

**DBA (Database Analyst).** This person is responsible for keeping the database running smoothly. Additional tasks are to plan and execute a backup/recovery plan, as well as performance tuning.

The data warehouse can be enterprise aware, including backups. The DBA can get notifications of failed processes, automate recovery, and proactively check the health of database structures prior to loading.

**Technical Architect.** This person is responsible for developing and implementing the overall technical architecture of the data warehouse, from the backend hardware/software to the client desktop configurations.

The Technical Architect benefits from using a single point of control for enterprise processes including the data warehouse processes.

**ETL Developer.** This person is responsible for planning, developing, and deploying the extraction, transformation, and loading routine for the data warehouse.

ETL tools provide their own scheduling system. The ETL developers may be resistant to moving control of ETL processes to an enterprise scheduler. However, the benefits they receive are: offloading the production support to an operations group, notifications of failures, and automation of recovery procedures. E-T-L is often done with scripts. Developers are usually more interested in developing the scripts than maintaining them. UC4 Workload Automation Suite eliminates the need for custom scripting, doing away with the maintenance issue.

**Front End Developer.** This person is responsible for developing the front-end, whether it be client-server or over the Web.

Front-end developers can kick off processes from their front-end application.

**OLAP Developer.** This person is responsible for the development of OLAP cubes.

OLAP tools provide their own scheduling system. The OLAP developers may be resistant to moving control of OLAP processes to an enterprise scheduler. However, the benefits they receive is: offloading the production support to an operations group, notifications of failures, automation of recovery procedures, and automation of the cube distribution.

**Data Modeler.** This person is responsible for taking the data structure that exists in the enterprise and modeling it into a schema that is suitable for OLAP analysis. They architect the transformation steps.

**DEFINITIONS**

Following are definitions for the most common terms used when discussing data warehouses. The definitions are extracted from a variety of Web sites.

**Data Warehousing.** An enterprise-wide system that extracts data from different servers and platforms and collects it in a single database. A data warehouse effectively consolidates data from multiple sources.

**Data Warehouse.** A very large database with special tools used to extract and cleanse data from operational systems and to analyze data. Data from various online transaction processing (OLTP) applications and other sources is selectively extracted and organized on the data warehouse database for use by analytical applications and user queries. Data warehousing emphasizes the capture of data from diverse sources for useful analysis and access, but does not generally start from the point-of-view of the end user or knowledge worker who may need access to specialized, sometimes local databases.

**Data Mart.** A focused subset of a data warehouse that deals with a single area of data and is organized for quick analysis. The data is usually stored in data cubes. The emphasis of a data mart is on meeting the specific demands of a particular group of knowledge users in terms of analysis, content, presentation, and ease-of-use. Users of a data mart can expect to have data presented in terms that are familiar.

**Data Cube.** A data cube presents data based on two or more attributes. For example, it might...
summarize sales by region (two attributes) or sales by region by product (three attributes). Information in data marts is usually stored in data cubes. Data cubes are usually archived for historical reporting. The archiving process is almost always scripted. UC4 Workload Automation Suite can replace the archiving scripts.

**Data Mining.** Sorting through data to identify patterns and establish relationships. Data mining parameters include:

- Association - looking for patterns where one event is connected to another event
- Sequence or path analysis - looking for patterns where one event leads to another later event
- Classification - looking for new patterns (May result in a change in the way the data is organized but that's ok)
- Clustering - finding and visually documenting groups of facts not previously known
- Forecasting - discovering patterns in data that can lead to reasonable predictions about the future

**Decision Support System (DSS).** An application that analyzes business data and presents it so users can make business decisions more easily. It is an "informational application" (in distinction to an "operational application" that collects the data in the course of normal business operation). Typical information that a decision support application might gather and present would be:

- Comparative sales figures between one week and the next
- Projected revenue figures based on new product sales assumptions
- The consequences of different decision alternatives, given past experience in a context that is described

**ETL.** Stands for extract, transform, and load.

- The extract function reads data from a specified source database and extracts a desired subset of data.
- The transform function works with the acquired data - using rules or lookup tables, or creating combinations with other data, to convert it to the desired state.
- The load function is used to write the resulting data (either all of the subset or just the changes) to a target database, which may or may not previously exist.

ETL can be used to acquire a temporary subset of data for reports or other purposes. A more permanent data set may be acquired for other purposes such as:

- the population of a data mart or data warehouse conversion from one database type to another
- migration of data from one database or platform to another

**Metadata.** Metadata is data that describes other data. Specifically, it is files or databases with information about another database’s attributes, structure, processing, or changes. It can describe any characteristic of the data such as the content or its quality. The normalized data created in the staging database can be considered metadata.

A lower level technology example of metadata is columns in a table. A DBA may create a person table:

<table>
<thead>
<tr>
<th>Data</th>
<th>Metadata (Column Name and physical layout)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Debbie Hamel</td>
<td>Name, Character 20</td>
</tr>
<tr>
<td>Metadata Evangelist</td>
<td>Title, Character 24</td>
</tr>
<tr>
<td>VA</td>
<td>State Code, Character 2</td>
</tr>
<tr>
<td>11111</td>
<td>Employee Id, Numeric</td>
</tr>
</tbody>
</table>

**OLAP** (online analytical processing). In data warehousing, OLAP is used to generate the data marts. OLAP makes it easy for a user to selectively extract and view data from different points of view. In short, it provides custom reporting. For example, a user can request a report that shows all of a company’s beach ball products sold in Florida in the month of July, compares revenue figures with those for the same products in September, and then compares other product sales in Florida in the same time period.

To facilitate this kind of analysis, OLAP data is stored in a multidimensional database called a cube. Whereas a relational database can be thought of as two-dimensional, a multidimensional database considers each data attribute (such as product, geographic sales region, and time period) as a separate "dimension." OLAP software can locate the intersection of dimensions (all products sold in the Eastern region above a certain price during a certain time period) and display them. Attributes such as time periods can be broken down into sub-attributes.

**Time Fencing.** The practice of scheduling data extraction based on estimated completion times of prerequisite processes. For example, a financials application may run from 1 A.M. to 2:30 A.M. every night. A process to extract data from this application may be scheduled for 3 A.M. to ensure the financials application process is complete. This often can result in the extraction process being run against bad data, wasting time and resources. UC4 Workload Automation Suite can eliminate time fencing by making the extraction contingent on the successful completion of the financials process.
ABOUT UC4 SOFTWARE

UC4 Software is a leading provider of workload automation and IT process optimization solutions that ensure core business processes and enterprise information systems run faster, more accurately and without interruption. More than 1,500 companies worldwide have successfully enhanced application processing performance and improved IT efficiency using UC4’s business acceleration solutions. Customers include American Suzuki Motor Corporation, Cadbury Schweppes, eBay, Eastman Kodak, General Electric, Mattel, McGraw Hill, Panasonic, Robert Bosch, Sun Microsystems, Symantec, T-Systems and Verizon. For more information, please visit WWW.UC4.COM.

CONTACT UC4 SOFTWARE

In the US at (877) 464-7300 (toll-free)
In Europe at +43 2233 77880

www.uc4.com
info@uc4.com