ETPL – Extract, Transform, Predict and Load

An Oracle White Paper
March 2006

ORACLE

ETPL – Extract, Transform, Predict and Load.

EXECUTIVE SUMMARY

Everyone wants the their data warehouse environment to be fully integrated. This means using standards based approaches to extraction, transformation, loading and application development. Most of this is now in place depending on your vendor, or vendors, of choice. This integrated approach to data warehousing has made many organizations re-consider how they provide their end users with access to geo-spatial, multi-dimensional and predictive analytic features.

While most users are demanding integration across these features IT departments are still treating these as separate stand-alone features that can have little or no influence on each other. This is largely because each feature requires separate proprietary database/server engines and corresponding non-standard programming languages.

Predictive analytics is proving to be the next big concepts that many forward thinking data warehouse teams are looking to utilize. Whereas, most software vendors are focusing on simply moving the analytic features to the business user community, there are exciting opportunities for traditional ETL developers to embed predicative analytic features directly within their extract, transform and load processes to create a new model: ETPL – extract, transform, predict and load.

Using predictive analytics ETL developers can enhance their deployments by using feature such as:

- Predicting missing values
- Predicting values for new attributes
- Determining most important attributes using "*Explain*" features
- Defining bins or clusters for flat data sources
- Creation of new hierarchies using market basked analysis

The ability to quickly and easily add these new features into existing data warehouse environments is dependent on ETL and predictive processes sharing a common language and preferably the same database engine. Oracle Warehouse Builder automatically generates deployment modules using PL/SQL. Oracle Data Mining is also able to generate PL/SQL, as well as Java if required. Since both products share a common language, it is also possible to extend the Warehouse Builder

transformation library to include predictive analytic functions and models, which can then be seamlessly embedded into existing ETL processes.

Adding this additional processing into an already crowded batch window will only work if the predictive analytics are built directly into the same database engine that hosts both the source and target schemas. With an integrated single multi-talented instance significant business value can be quickly and easily designed into any target schema that directly benefits business users.

The Oracle 10g database is the only truly integrated multi-talented database engine that allows companies to:

- KNOW MORE leverage any piece of data and discover valuable new information and insights that were previously hidden
- DO MORE build applications that automate the extraction and dissemination of new information and insights
- SPEND LESS Oracle's multi-talented database as a solution is significantly less expensive compared to other best-of-breed approaches and, as a component of your investment in Oracle technology, significantly reduces your total cost of ownership.

WHY EXTRACT, TRANSFORM, PREDICT AND LOAD?

Most companies today are gathering increasing quantities of data and moving that data using traditional extract, transform and load processes into a corporate data warehouse. This traditional ETL approach is currently failing many end users because their analysis of this ever-growing warehouse is constrained by a number of key issues:

- Missing information
- Lack of relevant business attributes
- Lack of focus on the most relevant dimensions, hierarchies and levels
- Ability to locate critical data in a timely fashion

All businesses are continually trying to uncover new trends and patterns of customer and/or product behavior that were previously hidden from view. Once these patterns have been exposed there is huge benefits to be gained from being able to use those results to drive further analytical analysis. This level of integration and cross-pollination results is now a *must have* feature for every company's data warehouse environment.

Data mining technology can generate a huge variety of new opportunities. Since predictive analytics automates the process of finding new values or associations within in a large database there is no need to perform the traditional types of manual analysis. Predictive analytics can be used to predict replacement values where nulls exist in the source data, and it can be used to discover new insights hidden in source data structures. It can help businesses target their best customers, find and prevent fraud, discover the most influential attributes that affect Key Performance Indicators (KPIs), and find valuable new information hidden in the data.

There are two ways to incorporate predictive analytics within a data warehouse. The first is to follow the current linear approach to data analysis, enforced by most software vendors, where each analytic operation (predict, explain, multi-dimensional model etc) works in isolation and information is locked into a specialized engine. This linear approach ensures that cross-fertilization of analysis is difficult if not impossible to achieve. Most software vendors, and hence customers, give little thought or consideration for automatically seeding this type of analysis directly within a data warehouse schema itself. As a result, the concept of predictive analytics for the masses for many organizations is a long way in the future.

To successfully implement predictive analytics and make it available to all data warehouse users, customers must push the processing back down the analysis chain. This implies extending the normal extract, load, transform model to create a new more powerful model: ETPL – extract, transform, *predict* and load.

BASIC REQUIREMENTS FOR SUCCESSFUL ETPL

Technical professionals should be able to use predictive analytics to help find patterns within source data, identify key attributes, discover new clusters and associations, and uncover valuable insights and seed these for generic access across the whole data warehouse community.

By embedding predictive analytics within the process of creating a data warehouse schema a number of key business issues can be quickly and easily resolved:

- Predict values for missing data points
- Predict new attribute values
- Determine most important attributes using "*Explain*" features
- Define bins or clusters for flat data sources
- Create new hierarchies using features such as market basket analysis

A typical example of where predictive analytics can be used to automate a predictive problem is targeted marketing campaigns. By using data on past promotional mailings, predictive analytics can identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events. The predicted attributes need to be incorporated directly into the data warehouse schema. This allows users to analyze trends on these attributes over time, discovering growth or contraction on month-by-month or year-by-year comparisons.

Predictive analytics also automates the discovery of previously unknown patterns: functions can sweep through tables and identify previously hidden patterns. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors

For customers to successfully extend their existing ETL processes to include a new *predict* stage, the following are key requirements:

- Integrated database engine
- Industry standard APIs

INTEGRATED DATABASE ENGINES

Many companies have implemented a best of breed approach for their data warehouse environment. While this has many perceived benefits, it also severely restricts a customer's ability to maximize their investment in those analytic engines. Each engine requires its own ETL process, which overloads the network as data is constantly shipped around various servers. To share these results requires companies to implement a continuous process of ETLE (extract, transform, load and extract) for each and every analytic engine within their data warehouse environment. At some point in time all the data is then re-loaded back into the data warehouse, which requires even more network activity and assumes the target data warehouse schema can fully support the results from each engine. Frequently the target database cannot store the resulting data objects, which not only limits end user access but also restricts the scope of analysis.

Most software vendors enforce this linear approach to data analysis and ETLE model, where each analytic operation (predict, explain, multi-dimensional model etc) works in isolation and information is locked into a specialized engine. This linear approach ensures that cross-fertilization of analysis is difficult if not impossible to achieve.

The Oracle 10g database is the only integrated database engine that supports both the storage and analysis of spatial, multi-dimensional and predictive analytics. Cross-fertilization of analysis comes as standard and because these objects are stored directly within the database all the processing associated with those features such as predictive analytics is performed directly against the source data without the need to first perform an extract step.

Doing More in a Shrinking Batch Window

Every data warehouse is constrained by very tight batch windows for extracting data from source systems, transforming the data and then loading the results into a target schema for access by various reporting tools.

The thought of doing more processing to predict missing values and also add additional attributes must seem an impossible dream for many data warehouse teams. Trying to extend existing processes to build a circular ETLE model to support predictive analytics may not be possible for many customers that have chosen to implement a best-of-breed approach.

However, by rejecting the best-of-breed linear approach and adopting the alternative model based on a single integrated database engine, such as Oracle 10g, it is possible to incorporate more processing within a single window and successfully implement ETPL. Only Oracle10g customers can safely implement ETPL today, without stretching their existing batch windows to breaking point.

INDUSTRY STANDARD APIs

Software companies are always pushing the latest technology and features and trying to ensure customers adopt these as quickly as possible. However, most IT departments have limited time and limited budgets. New concepts and features can only be adopted if they are wrapped within a familiar framework. For this to succeed everyone must be able to come together and talk the same language to allow exchange of features and concepts across data sources and data targets. If everyone talks a different language communication becomes complicated and data has to be continually moved and translated. Once a common language is established developers can build their models and processes and share them with other developers quickly and easily whilst focusing on their own area of expertise.

The industry standard for interrogating and analyzing data stored in an RDBMS is SQL. By providing a SQL interface millions of developers and business users can immediately start to use a feature within their favorite query and reporting tools.

Many developers prefer to extend their use of SQL by wrapping it within PL/SQL. By using PL/SQL developers can derive significant benefits. PL/SQL is Oracle's procedural extension to industry-standard SQL. PL/SQL naturally, efficiently, and safely extends SQL. Its primary strength is in providing a server-side, stored procedural language that is easy-to-use, seamless with SQL, robust, portable, and secure. Thus, it offers a platform for robust, high-performing enterprise applications; such as data warehouse environments that rely on extract, transform and load processes.

The most successful data warehousing customer's design, manage and deploy their data warehouse environments using Oracle Warehouse Builder. This is an enterprise level tool for designing, deploying and managing a data warehouse. ETL developers can use Warehouse Builders interface to visually design the extract, transform and load scripts. At generation time these designs are transformed into PL/SQL, which follows the Oracle model of recommendation that PL/SQL be used wherever possible for processing jobs within the database.

Oracle's new Data Mining features represent a breakthrough in business intelligence. Oracle Data Mining moves the analytical functions into the database—placing them directly alongside the data. Traditional alternatives force you to extract the data out of the database to separate, unsecured and costly dedicated statistical, analytical or mining servers.

Oracle Data Mining's PL/SQL API facilitates the complete Automation of data mining tasks. Application programmers can control all aspects of data mining process. They can expose complex settings for advanced users or completely automate the process for both other IT developers and/or business users. Programmatic control extends from data preparation and model building to on-

demand scoring of single records and batch scoring of large data sets. Batch scores may be stored in relational tables for access by other down-stream applications.

From an ETL developer perspective the predictive model is simply a PL/SQL package that generates an output table containing results of a prediction. The output table becomes a source table that can be joined to any other source and used to populate a target table in the data warehouse schema.

As both Warehouse Builder and Oracle Data Mining provide automatic generation of PL/SQL models, ETL developers can quickly and easily embed predictive analytic functions directly within their existing ETL process. Therefore, customers can quickly and easily extend their current ETL implements to incorporate predictive analytics and move beyond old style ETL to a much more powerful model – ETPL: Extract, Transform, Predict and Load.

CONCLUSION

Implementing an "Extract, Transform, Predict and Load" process can generate a huge variety of new opportunities. Since predictive analytics automates the process of finding new values or associations within in a large database there is no need to perform the traditional types of manual analysis.

This new model is set to have a huge impact on the development and deployment of corporate data warehouses. The days of separate teams of business analysts focusing on the use of predictive analytics and using their results in isolation are over.

Oracle is the only company to acknowledge the importance of this new model. The 10g database comes fully ETPL enabled, with built in support for functions that enable high-speed data extraction, transformation, predictive analytics, and target loading. The key data warehouse tools, such as Warehouse Builder, provide a comprehensive GUI that sits directly on top of these features. They support the direct generation of ETPL models via the registration of predictive analytic functions and the automatic generation of PL/SQL packages that can be deployed directly within the 10g database.

The Oracle 10g database is the only truly integrated multi-talented database engine that allows companies to:

- KNOW MORE leverage any piece of data and discover valuable new information and insights that were previously hidden
- DO MORE build applications that automate the extraction and dissemination of new information and insights
- SPEND LESS Oracle's multi-talented database as a solution is significantly less expensive compared to other best-of-breed approaches and, as a component of your investment in Oracle technology, significantly reduces your total cost of ownership.

ETPL – Etract, Transform, Predict and Load
March 2006
Author: Keith Laker

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
www.oracle.com