The Data Warehouse ETL Toolkit

Course Summary

Description

This course provides students with the skills necessary to plan, design, build, and run the ETL processes which are needed to build and maintain a data warehouse. It is based on the Ralph Kimball and Joe Caserta book *The Data Warehouse ETL Toolkit* published in 2004 by Wiley Publishing, Inc, ISBN: 0-7645-6757-8. The course also uses the book *Mastering Data Warehouse Aggregates* published in 2006 by Wiley Publishing, ISBN: 0-471-77709-9. If the students will be using SQL Server 2005 Integration Services (SSIS), the following book may be substituted the Ralph Kimball book: *The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset* by Joy Mundy, Warren Thornthwaite, and Ralph Kimball published by Wiley, February 13, 2006, ISBN 0471267155.

Topics

- Surrounding the Requirements
- ETL Data Structures
- Extracting
- Cleaning and Conforming
- Delivering Dimension Tables
- Delivering Fact Tables
- Development
- Operations
- Metadata
- Responsibilities
- Real-Time ETL Systems
- Conclusions

Audience

This course is targeted at technical staff, team leaders and project managers who need to understand how to plan, design, build, and run the Extract, Transformation, and Load (ETL) processes which are necessary to build and maintain a data warehouse.

Prerequisites

Students should have good knowledge of how to design a data base through at least third normal form (3NF) as well as good knowledge of how to design a star schema.

Duration

Three or four days

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

The Data Warehouse ETL Toolkit

Course Outline

I. Surrounding the Requirements
- A. Requirements
  1. Business Needs
  2. Compliance Requirements
  3. Data Profiling
  4. Security Requirements
  5. Data Integration
  6. Data Latency
  7. Archiving and Lineage
  8. End User Delivery Interfaces
  9. Available Skills
  10. Legacy Licenses
- B. Architecture
  1. ETL Tool Versus Hand Coding (Buy a Tool Suite or Roll Your Own)
  2. The Back Room—Preparing the Data
  3. The Front Room—Data Access
- C. The Mission of the Data Warehouse
  1. What the Data Warehouse Is
  2. What the Data Warehouse Is Not
  3. Industry Terms Not Used Consistently
  4. Resolving Architectural Conflict: A Hybrid Approach
  5. How the Data warehouse Is Changing
- D. The Mission of the ETL Team

II. ETL Data Structures
- A. To Stage or Not to Stage
- B. Designing the Staging Area
- C. Data Structures in the ETL System
  1. Flat Files
  2. XML Data Sets
  3. Relational Tables
  4. Independent DBMS Working Tables
  5. Third Normal Form Entity/Relation Models
  6. Nonrelational Data Sources
  7. Dimensional Data Models: The Handoff from the Back Room to the Front Room
  8. Fact Tables
  9. Dimension Tables
  10. Atomic and Aggregate Fact Tables
  11. Surrogate Key Mapping Tables
- D. Planning and Design Standards
  1. Impact Analysis
  2. Metadata Capture
  3. Naming Conventions
  4. Auditing Data Transformation Steps
- E. Summary

III. Extracting
- A. The Logical Data Map
  1. Designing Logical Before Physical
- B. Inside the Logical Data Map
  1. Components of the Logical Data Map
  2. Using Tools for the Logical Data Map
- C. Building the Logical Data Map
  1. Data Discovery Phase
  2. Data Content Analysis
  3. Collecting Business Rules in the ETL Process
- D. Integrating Heterogeneous Data Sources
  1. The Challenge of Extracting from Disparate Platforms
  2. Connecting to Diverse Sources Through ODBC
- E. Mainframe Sources
  1. Working with COBOL Copybooks
  2. EBCDIC Character Set
  3. Converting EBCDIC to ASCII
  4. Tranferring Data Between Platforms
  5. Handling Mainframe Numeric Data
  6. Using PICtures
  7. Unpacking Packed Decimal
  8. Working with Redefined Fields
  9. Multiple OCCURS
  10. Managing Multiple Mainframe Record Type Files
  11. Handling Mainframe Variable Record Lengths
- F. Flat Files
  1. Processing Fixed Length Flat Files
  2. Processing Delimited Flat Files
- G. XML Sources
  1. Character Sets
  2. XML Meta Data

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

The Data Warehouse ETL Toolkit

Course Outline (cont'd)

H. Web Log Sources
   1. W3C Common and Extended Formats
   2. Name Value Pairs in Web Logs
I. ERP System Sources
J. Extracting Changed Data
   1. Detecting Changes
   2. Extraction Tips
   3. Detecting Deleted or Overwritten Fact Records at the Source
K. Summary

IV. Cleaning and Conforming
A. Defining Data Quality
B. Assumptions
C. Part 1: Design Objectives
   1. Understand Your Key Constituencies
   2. Competing Factors
   3. Balancing Conflicting Priorities
   4. Formulate a Policy
D. Part 2: Cleaning Deliverables
   1. Data Profiling Deliverable
   2. Cleaning Deliverable #1: Error Event Table
   3. Cleaning Deliverable #2: Audit Dimension
   4. Audit Dimension Fine Points
E. Part 3: Screens and Their Measurements
   1. Anomaly Detection Phase
   2. Types of Enforcement
   3. Column Property Enforcement
   4. Structure Enforcement
   5. Data and Value Rule Enforcement
   6. Measurements Driving Screen Design
   7. Overall Process Flow
   8. The Show Must Go On—Usually
   9. Screens
   10. Known Table Row Counts
   11. Column Nullity
   12. Column Numeric and Date Ranges
   13. Column Length Restriction
   14. Column Explicit Valid Values
   15. Column Explicit Invalid Values
   16. Checking Table Row Count Reasonability
   17. Checking Column Distribution Reasonability
   18. General Data and Value Rule Reasonability
F. Part 4: Conforming Deliverables
   1. Conformed Dimensions
   2. Designing the Conformed Dimensions
   3. Taking the Pledge
   4. Permissible Variations of Conformed Dimensions
   5. Conformed Facts
   6. The Fact Table Provider
   7. The Dimension Manager: Publishing Conformed Dimensions to Affected Fact Tables
   8. Detailed Delivery Steps for Conformed Dimensions
   9. Implementing the Conforming Modules
   10. Matching Drives Deduplication
   11. Surviving: Final Step of Conforming
   12. Delivering
G. Summary

V. Delivering Dimension Tables
A. The Basic Structure of a Dimension
B. The Grain of a Dimension
C. The Basic Load Plan for a Dimension
D. Flat Dimensions and Snowflaked Dimensions
E. Date and Time Dimensions
F. Big Dimensions
G. Small Dimensions
H. One Dimension or Two
I. Dimensional Roles
J. Dimensions as Subdimensions of Another Dimension
K. Degenerate Dimensions
L. Slowly Changing Dimensions
M. Type 1 Slowly Changing Dimension (Overwrite)
N. Type 2 Slowly Changing Dimension (Partitioning History)
O. Precise Time Stamping of a Type 2 Slowly Changing Dimension
P. Type 3 Slowly Changing Dimension (Alternate Realities)
Q. Hybrid Slowly Changing Dimensions
R. Late-Arriving Dimension Records and Correcting Bad Data
S. Multivalued Dimensions and Bridge Tables

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

The Data Warehouse ETL Toolkit

Course Outline (cont'd)

T. Ragged Hierarchies and Bridge Tables
U. Populating Hierarchy Bridge Tables
V. Using Positional Attributes in a Dimension to Represent Text Facts
W. Summary

VI. Delivering Fact Tables
A. The Basic Structure of a Fact Table
B. Guaranteeing Referential Integrity
C. Surrogate Key Pipeline
1. Using the Dimension Instead of a Lookup Table
D. Fundamental Grains
1. Transaction Grain Fact Tables
2. Periodic Snapshot Fact Tables
3. Accumulating Snapshot Fact Tables
E. Preparing for Loading Fact Tables
1. Managing Indexes
2. Managing Partions
3. Outwitting the Rollback Log
4. Loading the Data
5. Incremental Loading
6. Inserting Facts
7. Updating and Correcting Facts
8. Negating Facts
9. Updating Facts
10. Deleting Facts
11. Physically Deleting Facts
12. Logically Deleting Facts
F. Factless Fact Tables
G. Augmenting a Type 1 Fact Table with Type 2 History
H. Graceful Modifications
I. Multiple Units of Measure in a Fact Table
J. Collecting Revenue in Multiple Currencies
K. Late Arriving Facts
L. Aggregations
1. Design Requirements #1 Through #4
2. Administering Aggregations, Including Materialized Views
M. Delivering Dimensional Data to OLAP Cubes
1. Cube Data Sources
2. Processing Dimensions
3. Changes in Dimension Data
4. Processing Facts
5. Integrating OLAP Processing into the ETL System
6. OLAP Wrapup
N. Summary

VII. Development
A. Current Marketplace ETL Tool Suite Offerings
B. Current Scripting Languages
C. Time Is of the Essence
1. Push Me or Pull Me
2. Ensuring Transfers with Sentinels
3. Sorting Data During a Preload
4. Sorting on Mainframe Systems
5. Sorting on UNIX and Windows Systems
6. Trimming the Fat (Filtering)
7. Extracting a Subset of the Source File Records on Mainframe Systems
8. Extracting a Subset of the Source File Fields
9. Extracting a Subset of the Source File Records on UNIX and Windows Systems
10. Extracting a Subset of the Source File Fields
11. Creating Aggregated Extracts on Mainframe Systems
12. Creating Aggregated Extracts on UNIX and Windows Systems
D. Using Database Bulk Loader Utilities to Speed Inserts
1. Preparing for Bulk Load
E. Managing Database Features to Improve Performance
1. The Order of Things
2. The Effect of Aggregates and Group Bys on Performance
3. Performance Impact of Using Scalar Functions
4. Avoiding Triggers
5. Overcoming ODBC Bottlenecks
6. Benefiting from Parallel Processing
F. Troubleshooting Performance Problems
G. Increasing ETL Throughput
1. Reducing Input/Output Contention
2. Eliminating Database Reads/Writes
3. Filtering as Soon as Possible
4. Partiioning and Parellelizing
5. Updating Aggregates Incrementally
6. Taking Only What You Need

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

The Data Warehouse ETL Toolkit

Course Outline (cont'd)

7. Bulk Loading/Eliminating Logging
8. Dropping Database Constraints and Indexes
9. Eliminating Network Traffic
10. Letting the ETL Engine Do the Work

H. Summary

VIII. Operations

A. Scheduling and Support
1. Reliability, Availability, Manageability Analysis for ETL
2. ETL Scheduling 101
3. Scheduling Tools
4. Load Dependencies
5. Metadata

B. Migrating to Production
1. Operational Support for the Data Warehouse
2. Bundling Version Releases
3. Supporting the ETL System in Production

C. Achieving Optimal ETL Performance
1. Estimating Load Time
2. Vulnerabilities of Long-Running ETL Processes
3. Minimizing the Risk of Load Failures

D. Purging Historic Data

E. Monitoring the ETL System
1. Measuring ETL Specific Performance Indicators
2. Measuring Infrastructure Performance Indicators
3. Measuring Data Warehouse Usage to Help Manage ETL Processes

F. Tuning ETL Processes
1. Explaining Database Overhead

G. ETL System Security
1. Securing the Development Environment
2. Securing the Production Environment

H. Short-Term Archiving and Recovery

I. Long-Term Archiving and Recovery
1. Media, Formats, Software, and Hardware
2. Obsolete Formats and Archaic Formats
3. Hard Copy, Standards, and Museums
4. Refreshing, Migrating, Emulating, and Encapsulating

J. Summary

IX. Metadata

A. Defining Metadata
1. Metadata—What Is It?
2. Source System Metadata
3. Data-Staging Metadata
4. DBMS Metadata
5. Front Room Metadata

B. Business Metadata
1. Business Definitions
2. Source System Information
3. Data Warehouse Data Dictionary
4. Logical Data Maps

C. Technical Metadata
1. System Inventory
2. Data Models
3. Data Definitions
4. Business Rules

D. ETL-Generated Metadata
1. ETL Job Metadata
2. Transformation Metadata
3. Batch Metadata
4. Data Quality Error Event Metadata
5. Process Execution Metadata

E. Metadata Standards and Practices
1. Establishing Rudimentary Standards
2. Naming Conventions

F. Impact Analysis

G. Summary

X. Responsibilities

A. Planning and Leadership
1. Having Dedicated Leadership
2. Planning Large, Building Small
3. Hiring Qualified Developers
4. Building Teams with Database Expertise
5. Don't Try to Save the World
6. Enforcing Standardization
7. Monitoring, Auditing, and Publishing Statistics
8. Maintaining Documentation
9. Providing and Utilizing Metadata
10. Keeping It Simple
11. Optimizing Throughput

B. Managing the Project
1. Responsibility of the ETL Team
2. Defining the Project
3. Planning the Project

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

The Data Warehouse ETL Toolkit

Course Outline (cont'd)

4. Determining the Tool Set
5. Staffing Your Project
6. Project Plan Guidelines
7. Managing Scope

C. Summary

XI. Real-Time ETL Systems

A. Why Real-Time ETL?
B. Defining Real-Time ETL
C. Challenges and Opportunities of Real-Time Data Warehousing
D. Real-Time Data Warehousing Review
   1. Generation 1—The Operational Data Store
   2. Generation 2—The Real-Time Partition
   3. Recent CRM Trends
   4. The Strategic Role of the Dimension Manager
E. Categorizing the Requirement
   1. Data Freshness and Historical Needs
   2. Reporting Only or Integration, Too?
   3. Just the Facts or Dimension Changes, Too?
   4. Alerts, Continuous Polling, or Nonevents?
   5. Data Integration or Application Integration?
   6. Point-to-Point Versus Hub-and-Spoke
   7. Customer Data Cleanup Considerations
F. Real-Time ETL Approaches
   1. Microbatch ETL
   2. Enterprise Application Integration
   3. Capture, Transform, and Flow
   4. Enterprise Information Integration
   5. The Real-Time Dimension Manager
   6. Microbatch Processing
   7. Choosing an Approach—A decision Guide
G. Summary

XII. Conclusions

A. Deepening the Definition of ETL
B. The Future of Data Warehousing and ETL in Particular
   1. Ongoing Evolution of ETL Systems

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically