KIMBALL UNIVERSITY

ETL Architecture in Depth

Why Attend

This class helps you understand all the factors necessary for effectively designing the back room ETL system of your DW/BI environment. It tries to guarantee that critical processes within the ETL system are not overlooked. Even if you don't have an immediate qualified need for every ETL subsystem on our list, it is likely that you will over time. By the end of this course, you will understand how your data warehouse ETL system can be built to anticipate these potential requirements.

This is not a microscopic code-oriented implementation class; it is a vendor-neutral architecture class for the designer who must keep a broad perspective. The course is organized around the 34 necessary ETL subsystems which are developed in detail throughout the course progresses. During class, each student builds (on paper) a comprehensive ETL system based on a realistically complex example, starting with the first steps of extraction through the final steps of data delivery to the presentation area for your BI tool.

Who Should Attend

This course is designed for those responsible for building the back room ETL system of a data warehouse environment, including ETL architects, ETL designers and developers, and data warehouse operational staff.

Since dimensional models are the ultimate ETL deliverables, some familiarity with the basic principles of dimensional modeling is necessary. Students can gain this knowledge by reading the following *Data Management Review* articles found at www.kimballgroup.com:

- Resist the Urge to Start Coding (Nov 2007)
- Set Your Boundaries (Dec 2007)
- Data Wrangling (Jan 2008)
- Dimensional Perspectives (Feb 2008)

Instructors

Ralph Kimball, co-author of The Data Warehouse ETL Toolkit, and Bob Becker

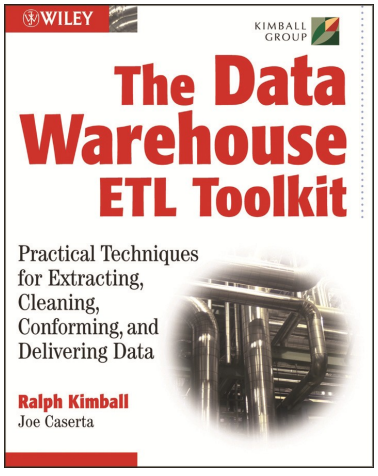
Course Overview

Day 1
- Surrounding the ETL Requirements: Developing the Essential Design Perspectives
- Data Profiling, Change Data Capture, and Extraction
- Job Scheduling, Backup, Recovery, and Restart

Day 2
- Cleaning: The Architecture of Data Quality
- Version Control, System Migration, Testing, Lineage Analysis, Problem Escalation
- Real Time Data Warehousing
- Big Data Predictive Analytics
- Conforming: The Architecture Data Integration

Day 3
- Building the ETL System
- Delivering Dimension Tables

Day 4
- Delivering Fact Tables
- Development and Operations



KIMBALL UNIVERSITY

ETL Architecture in Depth: Course Details

DAY 1

Surrounding the Requirements

Note: Augmented by class input – students/instructor propose requirements for a comprehensive ETL system design

- Business needs
- Compliance
- Data Profiling
- Security
- Integration needs
- Latency (daily, hourly, seconds, instantaneous)
- Archiving (recent history, very long term)
- Lineage and impact
- User profiles (developers, business users, analysts)
- Existing IT skills (traditional EDW, new Big Data systems)
- Existing technology licenses
- Hand coding vs. ETL tool choice
- Class roundtable exercise: Challenges in students' environments

Extract Steps: Bringing the Data to the Back Room

- Data types used in ETL systems
- (1) Data profiling
- Source to target map
- Access methods, source types (including new Big Data)
- Software, techniques
- (2) Change data capture
- (3) Extract window
- (3) Immediate transformations
- (3) Extract staging table designs, table types, retention, backup
- (3) Technical extraction tips
- (3) Traditional mainframe sources
- (3) XML sources and persistence of structures in back room
- ERP system sources
- Example vendors: Microsoft SSIS, Pentaho, Kettle
- Service oriented architectures, WSDL, and SOAP
- Big data sources
- (22) Job scheduler
- (22) Exception handling architecture
- 23) Backup
- Short term and long term recovery, archiving, sunsetting
- (24) Recovery, (24) Restart

DAY 2

Architecture of Data Cleansing

- (4) Data quality architecture
- (4) Data quality screens (column, structure, and business rule)
- (4) Business rule screens from statistical forecasts
- (4) Column and structure screens from data profiling
- (5) Error event fact tables
- Fact table surrogate keys
- Building the audit dimension and exposing in BI tool
- (4, 5, 6) Implementing data quality architecture in agile environment
- (7) De-duplication and survivorship

Real Time Data Warehousing

- Hot partition
- Streaming versus batch ETL
- Streaming delivery, query, reporting, dashboards, notifications
- Enterprise application integration (EAI) architecture
- Micro-batch ETL (MBETL) architecture
- Enterprise information integration (EII) architecture

Big Data Predictive Analytics

- Big Data use cases
- Four V's: volume, variety, velocity, value
- MapReduce, Hadoop, Pig, Hive, Hbase
- When to export to conventional RDBMS

DAY 2 CONTINUED

Architecture of Data Integration

- (8) Conforming dimensions, definition, impact on BI
- (8) Centralized and distributed responsibilities using conformed dimensions
- (8) Implementing conformed dimensions in agile environment
- (8) Example vendors: Pentaho, Microsoft SSIS, Informatica, Zend Studio
- (28) Sorting
- (25) Version control
- (26) System and version migration, testing and regression
- (27) Workflow monitor
- (27) Example vendors: Microsoft SSIS, IB Tivoli, Informatica
- (23) Job scheduler
- (29) Lineage and dependency analyzer
- (30) Problem escalation system

DAY 3

Delivering Dimension Tables

- (9) Time variance designs for slowly changing dimensions
- (10) Surrogate key generator
- (15) Multi-valued dimensions, bridge tables
- (11) Hierarchical dimensions
  - -Fixed
  - -Variable
  - -Ragged
  - -Bridge tables revisited
- (12) Special dimensions
  - -Date / Time dimensions
  - -Junk dimensions
  - -Mini-dimensions
  - -Small dimensions
  - -User maintained dimensions
  - -Shrunken dimensions
  - -Outrigger dimensions
  - -Behavior tags
  - -Step dimensions
  - -Super type / Sub type dimensions
  - -Study groups
  - -Special cases: extreme dimensionality and dimension width, incompatible members

DAY 4

Delivering Fact Tables

- (13) Fact table builder
  - -Transaction
  - -Periodic snapshot
  - -Accumulating snapshot
  - -Consolidated
- (14) Surrogate key pipeline
- Referential integrity
- Graceful extensibility
  - -Add attributes and facts
  - -Add dimensions to existing schemas
- (16) Late arriving dimension and fact data
- (17) Dimension manager
  - -Responsibilities and procedures
  - -Real time complexities
- (18) Fact provider
  - -Responsibilities and procedures
  - -Real time complexities
- (19) Aggregations
- (20) Feeding OLAP cubes