

# ETL Architecture Fundamentals

## Why Attend

This class provides an overview of 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-agnostic architecture class for the designer who must keep a broad perspective. The course is organized around the 34 necessary ETL subsystems which are discussed in detail throughout the course.

Bringing Kimball University onsite enables everyone on the team to get on the same page with a common vocabulary and understanding of core techniques; the result is more efficient education with lower travels cost and lost productivity, plus less downstream "tire spinning" within the team

## 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](http://www.kimballgroup.com):

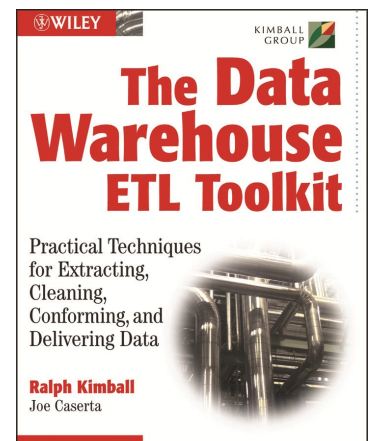
- Resist the Urge to Start Coding (Nov 2007)
- Set Your Boundaries (Dec 2007)
- Data Wrangling (Jan 2008)
- Dimensional Perspectives (Feb 2008)

## Instructors

Bob Becker, co-instructor of Kimball University's public ETL Architecture in Depth course

## Course Overview

- Day 1 • Surrounding the ETL Requirements  
• 34 ETL Subsystems: Data Profiling through Delivering Dimension Tables
- Day 2 • Continuation of 34 ETL Subsystems: Delivering Fact Tables through Operations  
• Modifying your ETL Architecture for Real Time Data Warehousing



## ETL Architecture Fundamentals: Course Details

---

### DAY 1

#### Surrounding the Requirements

- Business needs
- Data Profiling
- ETL tool versus coding

#### Introduction to 34 ETL Subsystems

- Subsystem 1–Data profiling
- Subsystem 2–Change data capture
- Subsystem 3–Extract
  - Extract window
  - Immediate transformations
  - Extract staging table designs
  - Retention
- Subsystems 4, 5 and 6–Cleaning
  - Data quality and architecture
  - Data quality screens
  - Error event fact table
  - Audit dimension
  - Compliance tracking
- Subsystem 28–Sorting
- Subsystem 7–De-duplication and survivorship
- Subsystem 8–Conforming
  - Conformed dimensions
  - Mapping incompatible structures
- Subsystems 9, 10, 11, 12 and 15–Delivering dimension tables
  - Time variance designs
  - Surrogate key generator
  - Bridge tables
  - Hierarchies
  - Special dimensions

### DAY 2

#### Continuation of the 34 ETL Subsystems

- Subsystem 13, 14 and 16–Delivering fact tables
  - Fact table builders for transaction, periodic and accumulating grains
  - Surrogate key pipeline
  - Graceful extensibility
  - Late arriving data
- Subsystems 17 and 18–Dimension manager and fact provider
  - Responsibilities and procedures
  - Real time complexities
  - Distributed, federated warehouses
  - Delivering remote dimensions, attributes and facts
- Subsystems 19 and 20–Delivering aggregates
  - Aggregations
  - Aggregate navigation
  - OLAP cubes
- Subsystems 21–Data propagation manager
  - Feeding data mining
  - Presentation layer extracts
  - 3rd party flat files
- Subsystem 22–Job scheduler
- Subsystem 23 and 24–Backup and recovery
- Subsystem 25 and 26–Version control
  - Version control
  - Migration and testing
- Subsystems 27, 29 and 30–Managing ETL workflow
  - Workflow monitor
  - Job scheduler
  - Lineage and dependency tracking
  - Problem escalation system
- Subsystems 31, 32, 33 and 34–Development and operations
  - Parallel processing and pipelining
  - Security
  - Compliance
  - Metadata

#### Modifying the ETL Architecture for Real Time Data Warehousing

- Hot partition
- Streaming versus batch ETL

