

Introduction to Data Warehousing

Course Summary

Description

This course will give the student a fundamental understanding of data warehouse systems. The students will be presented with a real world problem. As the concepts of data warehousing are presented throughout the course of the class, the students will act as a team of consultants and will have regular meetings/discussions that will allow them to apply their newly learned skills to the problem that they were presented. By the end of the course, the students will have acted as a team to create a conceptual data warehouse. Students will leave this class with the practical skills necessary to design an end to end data warehouse.

Objectives

At the end of this course, students will be able to:

- Understand and be familiar with the terminology relating to Data Warehouses
- Make sound decisions regarding the implementation of a Data Warehouse

Topics

- Introduction to data warehousing
- Introduction to dimensional modeling
- Designing the dimensional model
- Designing the database
- Intro to ETL
- Developing ETL

Audience

This course is intended for IT personnel with some relational database experience.

Prerequisites

The student should be familiar with basic database concepts.

Duration

Five days

Introduction to Data Warehousing

Course Outline

- I. Introduction to data warehousing**
 - A. What is a data warehouse?
 - B. Benefits of a data warehouse
 - C. Data warehouse characteristics
- II. Introduction to dimensional modeling**
 - A. What is dimensional modeling?
 - B. Dimensional modeling vs. normalized modeling
 - C. Fact Tables
 - D. Dimension Tables
 - E. Four-Step dimensional design process
 - F. Degenerate dimensions
 - G. Slowly changing dimensions
 - H. Type 1, Type 2 and Type 3 approach
 - I. Role-Playing dimensions
 - J. Junk dimensions
 - K. Handling Hierarchies
 - L. Many-Valued dimensions
 - M. Three grains of fact tables
- III. Designing the dimensional model**
 - A. Identify design participants
 - B. Revisit the requirements
 - C. Establish naming conventions
 - D. Research and profile source data
 - E. Apply the four-step modeling process
 - F. Build the high level dimensional model
 - G. Develop the detailed dimensional model
 - H. Review and validate the model
 - I. Finalize the design documentation
- IV. Designing the database**
 - A. Follow naming conventions
 - B. Decide on nullable columns
 - C. Build staging tables
 - D. Develop file naming and location standards
 - E. Primary keys
 - F. Foreign keys
 - G. Design the physical data structure
 - H. Finalize the S2T map
 - I. Star vs. Snowflake
- V. Intro to ETL**
 - A. Data quality
 - B. Data integration
 - C. Data latency
 - D. Archiving
 - E. Data profiling
 - F. Capture change data
 - G. Extract system
 - H. Data cleansing
 - I. Error handling
 - J. Auditing
 - K. De-duplication
 - L. Conforming
 - M. Managing slowly changing dimensions
 - N. Surrogate keys
 - O. Managing hierarchy
 - P. Managing special dimensions
 - Q. Building fact tables
 - R. Surrogate key pipeline
 - S. Building bridge tables
 - T. Handling late arriving data
 - U. Managing dimensions
 - V. Building aggregates
 - W. Workflow Monitor
 - X. Lineage and dependency analyzer
 - Y. Pipelining
- VI. Developing ETL**
 - A. Develop a high level plan
 - B. Choose a tool
 - C. Decide on default strategies
 - D. Drill down table by table
 - E. One time historic load processing
 - F. Populate dimension tables with historic data
 - G. Perform the fact table historic load
 - H. Develop an incremental process for dimension tables
 - I. Develop an incremental process for fact tables
 - J. Develop a process to load aggregate tables
 - K. Operation and automation