Data Warehousing and ETL using Apache Spark
By Dr. Vishwanath Rao

Day 1

Introduction to Data Warehousing

- **Basic Concepts and Definitions**
    - Definition of the data warehouse (DW)
    - Overall architecture of a DW
    - DW processes
    - Categories of DW technology
    - DW project and initiative types
- **Project Management Deliverables**
    - DW strategy
    - DW project scope
    - DW project plan
    - Managing a DW project
    - The iterative release model
- **Introduction to the Dimensional Model**
    - Facts
    - Dimensions
    - Star schemas
    - Snowflakes
- **Architectural deliverables**
    - Requirements
    - Analysis
    - Design
    - Infrastructure
    - Implementation

Day 2

Implementation, Operation and Expansion

- **Implementation Deliverables**
    - Outcomes from analysis
    - Outcomes from design
    - Outcomes from construction
    - Outcomes from deployment
    - Operational Deliverables
    - Service level agreements
    - Outcomes of usage

    - DW monitoring
    - DW governance
- **Maintaining the DW**
    - Incremental DW releases
    - Follow-up to DW
    - On-going assessment
    - Post mortem and lessons learned
    - Managing consultants
    - Managing the vendor
    - Getting started with data warehousing

Data Warehouse Architectures

- Centralized DW
- Functional DW
- Federated DW
- Independent Data Marts
- Dependent Data Marts

Day 3

Data Warehouse Methodology

- Explanation of methodology steps
- Iterative nature of development

Information Gathering

- Facilitated sessions
- Interviews
- Information gathering techniques
    - Events
    - Objectives
    - Queries
    - Goals
    - Decisions
    - Problems

Data Store Layer

- Building the Data Warehouse Model
- Facts, dimensions
- Summarized data
- Levels of Data In the Enterprise

- Base grains
- Intermediate Summaries
- Specialized summaries

Day 4

Modeling Time and History

- Short term and long term view
- Four ways of handling time and date
- Time-series data
- Capturing business changes
- Importance of representing the business time dimension

ETL Layer

- Defining transformation requirements
- Defining transformation rules
- The transformation requirements spreadsheet
- Building transformation processes
- Enforcing controls in the ETL process
- Designing the transformation process
- Complete coverage transformation types
- Dealing with change data
- Supporting surrogate keys
- Near-real time transformation

- Working with Metadata
- Accessing sequential data
- Partitioning and collecting
- Combining data
- Sorting and aggregating data
- Transforming data
- Working with relational data

Day 5

- Job controls and pipelines
- Processing unstructured data
- Data masking
- Using data rules
- Processing XML data
- Runtime Column propagation

BI Layer

- Designing the BI interface
- Matching the BI interface to the user
- Types of BI technologies and design
- Types of reporting
- OLAP in all its forms:
    - MOLAP
    - HOLAP
    - DOLAP
    - ROLAP
- Data sparsity and density
- Data explosion due to calculations, rollups and summaries

Important Considerations and Issues

- System load
- Denormalization and performance
- Archiving and purging
- Data distribution and replication
- Change control
- Copy management
- Alternative Models For Copied Data

All major topic coverage comes with Apache Spark demos using streams, dataset, data frames and Spark Sql.