Data Profiling

by Arkady Maydanchik
Module 0. About the Course (7 min)

Module 1. Introduction to Data Profiling (44 min)
- What is Data Profiling?
- Myth and Reality of Data Profiling
- Profiling Techniques
  - Column Profiling
  - Profiling Relational Data Models
  - Profiling Time-Dependent Data
  - Subject Profiling
  - Profiling State-Transition Models
  - Attribute Dependency Profiling
  - Dynamic Data Profiling
- Profiling Challenges
  - What to Profile?
  - How to Profile?
  - How to Organize the Results?
  - How to Analyze the Results?
- Role of Profiling
  - Profiling and Data Quality Management
  - Profiling and Master Data Management
  - Profiling and Data Migration
  - Profiling and Data Consolidation
  - Profiling and Data Integration Interfaces
  - Profiling and Data Warehouses
- People and Technology
  - Tools and Technology
  - Role of IT
  - Role of Business
  - Data Profiling Profession

Module 2. Column Profiling (89 min)
- Introduction
  - What is Column Profile?
  - The Objective of Column Profiling
  - Types of Columns
  - Elements of Column Profile
  - Implications of Field Types
  - Column Profiling Technology
- Basic Counts
  - Count of All Values
  - Count of Defective Values
  - Count of NULL or Missing Values
  - Count of Invalid Values
  - Overall Summary Report
  - Automating Result Analysis
  - Basic Column Summary Report
  - Basic Drill-Downs
  - Count of Distinct Values
- Value Frequency Charts
Module 3. Profiling Time-Dependent Data (58 min)

- **Introduction**
  - What are Time-Dependent Data?
  - Why Special Profiling Techniques?
  - Types of Time-Dependent Data
  - Profiling Technology
  - Profiling Techniques

- **Timeline Profiling**
  - Entity Timeline Profiling
  - Event Timeline Profiling
  - Code Timeline Profiling
  - Value Timeline Profiling
  - Granularity Profiling

- **Timestamp Pattern Profiling**
  - Timestamp Pattern Example
  - Profiling by Calendar Month
  - Profiling by Calendar Year
  - Types of Timestamp Profiling Intervals

- **Multi-Dimensional Profiling**
  - Two-Dimensional Timestamp Profiling
  - More Two-Dimensional Profiling
  - True Multi-Dimensional Profiling

- **Event Dependency Profiling**
Module 4. Profiling State-Transition Models (49 min)

- **Introduction**
  - State-Dependent Objects
  - State-Transition Models
  - State-Transition Model Features
  - Why Profile State-Transition Models?
  - Profiling Techniques

- **Data Structures for State-Dependent Data**
  - State-Dependent Entities
  - Preparation to Profiling

- **Profiling Techniques**
  - Starting Terminator Profiling
  - Ending Terminator Profiling
  - State-Transition Profiling
  - Action-State Profiling
  - State Duration Profiling
  - Other Profiling Techniques

Module 5. Other Profiling Techniques (65 min)

- **Subject Profiling**
  - What are Subjects?
  - Subject De-duplication
  - Subject Matching
  - Profiling by Source
  - Profiling by Entity
  - Further Drill-Downs

- **Relational Integrity Profiling**
  - Relational Data Models
  - Identity Profiling
  - Referential Integrity Profiling
  - Cardinality Profiling

- **Attribute Dependency Profiling**
  - Value Affinity
  - Conditional Optionality
  - Advanced Profiling Techniques

- **Dynamic Data Profiling**
  - What Profiles to Consider?
  - Which Data to Profile?
  - Manual Approach
  - Semi-Automatic Approach
  - Statistical Change Monitors