

**Practical Data Quality**

**Course Number:** DATA-126WA  
**Duration:** 2 days

**Overview**

The integrity of data (or lack thereof) affects the overall success of any analytical work. This Practical Data Quality training course teaches attendees how to maintain high data quality standards to make sound tactical and strategic business decisions. Participants learn how to resolve errors and flaws in datasets, implement tactics for monitoring and building workflows, and more.

**Prerequisites**

All attendees must have data processing experience. Familiarity with the Python programming syntax is beneficial, but not required.

**Materials**

All Data Quality training attendees receive comprehensive courseware.

**Software Needed on Each Student PC**

* Computer with Internet connectivity
* Ability to install software on the computer
* Recent 64-bit OS, such as Windows 10, macOS, or Linux

**Objectives**

* Understand the factors that contribute to poor data quality
* Measure data quality
* Validate and normalize data
* Perform unit testing
* Implement best practices to ensure data quality

**Outline**

* Data Quality Introduction
  + Data Quality Defined
  + Data Quality Dimensions/Properties
  + Interpreting Data Quality Properties
  + The Typical Data Analytics (Machine Learning) Pipeline
  + Data Quality Assurance
  + Common Factors Contributing to Poor Data Quality
  + Is Bad Data Quality a Good or a Bad Thing?
  + Data Quality is a Shared Concern
  + Data Governance
  + Common Issues that can be Prevented Through Effective Governance
  + The Data Steward Role
  + Common Steps to Overcome Data Quality Issues
  + Data Observability
  + Application Performance Monitoring (APM) and Observability Magic Quadrant
  + Example of (Operational) Observability Dashboard
  + Data Quality and Data Observability Relationship
  + Example of an Observability-Enabling Service
  + A Glossary of Business Terms
  + Data Dictionaries
  + Example of a Data Dictionary
  + SLAs
  + SLAs and Non-Functional Requirements
  + The Great, Fast, and Cheap Quality Diagram
* Measuring the Quality of the Data
  + Examples of Data Quality Metrics
  + Measuring Data Quality
  + Common Corrective Measures for Data Quality Problems
  + Descriptive Statistics
  + Correlation
  + Normal Distribution and Z-Score
  + Non-uniformity of a Probability Distribution
  + Shannon Entropy
  + Gini Impurity
  + Example of Using Gini Impurity Formula
  + Confusion Matrix
  + The Binary Classification Confusion Matrix
  + A Binary Classification Confusion Matrix Visually
  + Example of a Confusion Matrix
* Methods and Techniques for Data Quality
  + Connecting to the Digital Realm
  + States of Digital Data
  + Maintenance
  + Automation
  + Workflow (Pipeline) Orchestration Systems
  + Example of a Workflow Orchestration System: Apache NiFi
  + NiFi Processor Types
  + Building a Simple Data Flow in the NiFi Designer
  + Logging Levels
  + Data Formats
  + Interoperable Data
  + Timeliness
  + Efficient Storage with Columnar Formats
  + Storage and Querying Efficiencies of the Parquet Columnar Storage Format
  + Assertions
  + The assert Expression in Python
  + Two Types of Errors
  + Runtime Errors/Exceptions
  + Life after an Exception
  + Assertions vs. Errors (Exceptions)
  + Data Validation
  + Data Normalization
  + DDL-based Data Validation
  + An SQL DDL Schema with Constraints Example
  + Apache Hive and Schema-on-Demand
  + An Example of Hive DDL
  + XML and JSON Schemas
  + The Schema Production and Consumption Diagram
  + Example of an XSD Schema Authoring Editor
  + Regular Expressions Elements
  + What is Unit Testing and Why Should I Care?
  + Unit Testing and Test-Driven Development
  + TDD Benefits
  + Testing for Failure
  + Logging and Monitoring
* Data Consistency
  + The Consistency Consensus
  + The Two-phase Commit (2PC) Protocol Diagram
  + The CAP Theorem
  + Mechanisms for Guaranteeing a Single CAP Property
  + The CAP Triangle
  + Eventual Consistency
  + Example of the Consistency vs. Availability Gap
  + How eBay Preempts Possible Database Corruption
  + The Saga Pattern
  + Saga Log and Execution Coordinator
  + The Saga Happy Path
  + A Saga Compensatory Requests Example
  + The Event Sourcing Pattern
  + Event Sourcing Example
  + Applying Efficiencies to Event Sourcing
  + Time Accuracy and Consistency
  + Network Time Protocol (NTP)
* Data Quality Best Practices
* Conclusion