

**DataOps for IT Professionals**

**Course Number:** DATA-128WA
**Duration:** 1 day

**Overview**

This DataOps for IT Professionals training course teaches attendees how to elevate the quality of their data, increasing the effectiveness of the analytical work based on this data that supports organizational decisions. Participants learn how to incorporate practical plans and technical assistance throughout the entire data lifecycle, including data acquisition, storage, processing, and consumption.

**Prerequisites**

All attendees must have practical work experience in data processing.

**Materials**

All DataOps 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 10/11, macOS, or Linux

**Objectives**

* Understand the enterprise data processing and IT systems challenges
* Correct "bad" input data
* Perform data cleansing
* Deal with missing and duplicate data
* Enforce data consistency
* Implement data governance

**Outline**

* DataOps Introduction
	+ DataOps Enterprise Data Technologies
	+ Enterprise Data Processing Challenges and IT Systems' Woes:
		- Data Quality
		- What Makes Information Systems Cluttered and Myopic
		- Fragmented Data Sources
		- Different Data Formats
		- System Interoperability
		- Maintenance Issues
	+ Data-Related Roles
	+ Data Engineering
	+ What is DataOps?
	+ The DataOps Technology and Methodology Stack
	+ The DataOps Manifesto
	+ Agile Development
	+ DevOps
	+ The Lean Manufacturing Methodology
	+ Key Components of a DataOps Platform
	+ Overview of DataOps Tools and Services
	+ Overview of DataOps Platforms
* Data Quality
	+ Data Quality Definitions
	+ Dimensions of Data Quality
	+ Defining "Bad" Data
		- Missing Data
		- Wrong/Incorrect Data or Data Format
		- Inconsistent Data
		- Outdated (Stale) Information
		- Unverifiable Data
		- Withheld Data
	+ Common Causes for “Bad" Data
		- Human Factor
		- Infrastructure- and Network-Related Issues
		- Software Defects
		- Using the Wrong Tool for the Job
		- Using Untrusted Data
		- Aggregation of Data from Disparate Data Sources that have Impedance Mismatch
		- Wrong QoS Settings of Queueing Systems
		- Wrong Caching System Settings, e.g. TTL
		- Not Using the "Ground Truth" Data
		- Differently Configured Development/UAT/Production Systems
		- Confusing Big-Endian and Little-Endian Byte Order
	+ Ensuring Data Quality
		- Ensuring Integrity of Datasets
	+ Dealing with "Bad" Input Data
		- DDL-enforced Schema & Schema-on-Demand (-on-Read)
		- SQL Constraints as Rules for Column-Level and Table-Wide Data
		- XML Schema Definition (XSD) for XML Documents
		- Validating JSON Documents
		- Regular Expressions
		- Data Cleansing of Data at Rest
		- Controlling Integrity of Data-in-Transit
		- Database Normalization
		- Using Assertions in Applications
		- Operationalizing Input Data Validation
	+ Data Consistency and Availability
	+ Dealing with Duplicate Data
	+ Dealing with Missing (NaN) Data
	+ Master (Authoritative) Data Management
	+ Enforcing Data Consistency with the scikit-learn LabelEncoder Class
	+ Data Provenance
	+ The Event Sourcing Pattern
	+ Adopting the Culture of Automation
	+ On-going Auditing
	+ Monitoring and Alerting
	+ UiPath
	+ Workflow (Pipeline) Orchestration Systems
* How to Lead with Data
	+ Enterprise Architecture Components
		- Business Architecture
		- Information Architecture
		- Application Architecture
		- Technology Architecture
	+ DataOps Functional Architecture
	+ The Snowflake Data Cloud
	+ Cloud Design for System Resiliency
	+ New Data Architecture:
		- Data Ownership
		- Shared Environment Security Controls
* Data Governance (Optional)
	+ The Need for Data Governance
	+ Controlling the Decision-Making Process
	+ Controlling "Agile IT"
	+ Types of Requirements
		- Product
		- Process
	+ Scoping Requirements
	+ Governance Gotchas
	+ Governance Best Practices
* Conclusion