Analytics - Fundamentals of DataOps

Course Number: WA3219

Duration: 1 days

Course Description

This course introduces you to DataOps basics, including its origins, components, real-life applications, and ways to implement it.

Skills Gained

- Understand the fundamentals of DataOps and how people collaborate to deliver data for specific purposes.
- Learn the three standard DataOps pipelines (Production, Development, and Environment) and how to orchestrate the necessary teams, tools, and processes
- Test, measure, and iteratively improve DataOps production pipelines
- Structure the development pipeline to fit the development lifecycle and use it to achieve fast deployments
- Build environment pipelines, manage components, and adapt the environment to different use cases
- Apply Lean DataOps to improve your organization’s data operations

Who Can Benefit

- Data and Business Analysts
- Information Architects
- Technical Managers

Prerequisites

General knowledge of programming and data processing.

Course Details

Outline

DataOps Introduction

- Data Analytics On the Run
- Impediments to the Data Analytics Cycle Time
- Finding a Solution ...
- What is DataOps?
- Agile Development ...
- DevOps
- The DataOps Technology and Methodology Stack
- The DataOps and Data Science Relationship
- DataOps Relationships with Other Data Management Disciplines and Concerns
- Standing Up a DataOps Practice
- The Lean Manufacturing Methodology
- Statistical Process Control
- What is Six Sigma?
- DataOps Enterprise Data Technologies
- The DataOps Manifesto
- Problems that DataOps Solves
- DataOps Leadership Principles

The DataOps Problem Domain

- Connecting to the Digital Realm ...
- Data is King
- Actionable Insights
- Snowflake Environments
- Data Observability
- Cloud Resource Monitoring Dashboards
- Fragmented Data Sources
- Data Formats
- Interoperable Data

- The Data-Related Roles
- What is Data Engineering
- The Typical Data Analytics (Machine Learning) Pipeline
- IT Systems' Woes
- Types of Architecture
- How to Lead with Data (the "Fidelity Way" *)
- How to Lead with Data: Ownership
- How to Lead with Data: Shared Environment Security Controls
- How to Lead with Data: the Current Trends
- DataOps Functional Architecture
- Key Components of a DataOps Platform
- Automation
- Maintenance
- DataOps Data Pipelines
- Building Pipelines: Aggregating System DAGs
- Distributed Data Flow Challenges
- Promoting Teamwork
- The Tragedy of the (Unmanaged) Commons
- Tests in Data Analytics
- Test Types
- The Netflix Simian Army Test Suite
- Input Data "Irregularities"
- Dealing with Missing Data in Python

DataOps Technology and Tools

- Data Storage System Types
- The CAP Theorem
- The CAP Triangle - Which Storage System to Choose
- Mechanisms to Guarantee a Single CAP Property
- Data Physics (a.k.a Distributed Data Economics)
- Hadoop: Example of Collocating Data and Computation
- An Example of Hive DDL
- Efficient Storage with Columnar Formats
- Example: AWS Athena Storage and Processing Cost Savings

- Example: Converting the CSV Data Format into Parquet Using HiveQL CTAS Statement
- The Cloud: Value Proposition
- Lessons from the Field
- Design for System Resiliency
- How eBay Preempts Possible Database Corruption
- Cloud Data Services
- The Cloud Strategy
- Virtualization
- Virtualization Benefits
- What is Docker
- What is Kubernetes
- Computing Services in the Cloud
- Get Educated ...
- "Good/Not so Good" Use Cases for the Cloud
- Infrastructure as Code (IaC)
- Example of Provisioning and Running a PostgreSQL Database in Docker
- IoC Systems and Tools
- Workflow (Pipeline) Orchestration Systems
- Example of a Workflow Orchestration System: Apache NiFi
- NiFi Processor Types
- Building a Simple Data Flow in the NiFi Designer
- An Annotated Example of Using scikit-learn Python Machine Learning (ML) Pipeline Class
- Version Control Systems
- Branching and Merging Visually
- Some Popular Version Control Systems
- Overview of DataOps Tools and Services

IT Governance

- IT Governance
- Data Governance
- Controlling the Decision-Making Process
- Enterprise IT Governance Models
- Key Artifacts

- Agile IT
- Types of System Requirements
- Scoping Requirements
- Requirements Gathering ...
- Data Governance Overview
- Data Governance Roles and Responsibilities
- Roles and Responsibilities in DataOps
- Example of Assigning Responsibilities (AWS Shared Responsibility Model)
- Example of a Governance-Enabling Service
- Governance Best Practices
- Governance Gotchas
- The Goldilocks Principle

Lab Exercises

- Lab 1. Data Availability and Consistency