

Data Pipeline Engineering

William McKnight

Jake Dolezal



Module 0: About the Course (3 min)

Module 1: Introduction to Data Pipelines (53 min)

- Module Overview
- Data Pipelines Defined
 - What is Data Pipeline?
 - Data Sources
 - Data Ingestion
 - Data Transformation
 - Data Targets
 - Orchestration
 - Analytics
 - Topic Summary
- Data Consumption
 - Consumption by Reports
 - Consumption by Dashboards
 - Consumption by Visualization
 - Consumption by Self-Service BI
 - Consumption by Mobile Apps
 - Consumption by Alerts & Notifications
 - Consumption by APIs
 - Consumption by Ad Hoc
 - Consumption by Spreadsheets
 - Consumption by NLP
 - Consumption that is Collaborative
 - Consumption by ML
 - Topic Summary
- Data Pipeline Engineering
 - What is Data Pipeline Engineering?
 - Data Pipeline Engineers
 - Overlap & Dependencies: Part 1
 - Overlap & Dependencies: Part 2
- Types of Data Pipelines
 - Types of Data Pipelines: Batch Data Quality Analytics
 - Types of Data Pipelines: Streaming
 - Types of Data Pipelines: ETL
 - Types of Data Pipelines: Streaming ETL
 - Types of Data Pipelines: Replication
 - Types of Data Pipelines: Event-driven
 - Types of Data Pipelines: Aggregation
 - Types of Data Pipelines: CDC
 - Types of Data Pipelines: Lakehouse
 - Types of Data Pipelines: ML



DE-01 Data Pipeline Engineering

- Types of Data Pipelines: Data Quality
- Types of Data Pipelines: Analytics
- Topic Summary
- Module Summary

Module 2: Data Pipeline Design Patterns (34 min)

- Anatomy of a Data Pipeline
- Module Overview
- Bulk Loading
 - Bulk Loading Architecture
 - Bulk Loading Key Steps
 - Bulk Loading Validation
 - Bulk Loading Transformation
 - Bulk Loading
 - Bulk Loading Cleanup
 - Bulk Loading Considerations
 - Bulk Loading Tools
 - Topic Summary
- ETL
 - Extract, Transform, Load (ETL) Architecture
 - Extract, Transform, Load (ETL) Key Steps
 - Extract, Transform, Load (ETL) Transform
 - Extract, Transform, Load (ETL) Load
 - Extract, Transform, Load (ETL) Monitor
 - Extract, Transform, Load (ETL) Considerations – Part 1
 - Extract, Transform, Load (ETL) Considerations – Part 2
 - Extract, Transform, Load (ETL) Examples
 - Topic Summary
- CDC
 - Change data Capture (CDC) Architecture
 - Change data Capture (CDC) Use
 - Change data Capture (CDC) Key Steps
 - Change data Capture (CDC) Considerations
 - Change data Capture (CDC) Examples
 - Topic Summary
- Data Stream Processing
 - Data Stream Processing Architecture
 - Data Stream Processing Key Steps
 - Data Stream Processing Considerations
 - Examples
- Module Summary



Module 3: Anatomy of a Data Pipeline (50 min)

- Module Overview
- Data Pipeline Architecture
- Components
- Sourcing and Processing Components
 - Data Sources: Legacy
 - Data Sources: Transactional
 - Data Sources: Web-Based
 - Data Sources: Third Party
 - Data Sources: Social Media
 - Data Sources: Machine
 - Data Sources: Geospatial
 - Processing: ETL/ELT
 - Processing: Event Flow
 - Processing: Change Data Capture
 - Workflow Components
 - Topic Summary
- Post-Processing Components
 - Data Stores/Storage: Temporary/Sandbox
 - Data Stores/Storage: Staging
 - Data Stores/Storage: Warehouse/Mart
 - Data Stores/Storage: Data Lakes/Data Lakehouses
 - Data Stores/Storage: Master Data Management Hub
 - Data Stores/Storage: Vector Databases
 - Destination Application: Reporting
 - Destination Application: OLAP
 - Destination Application: Scorecards
 - Destination Application: Dashboards
 - Destination Application: Exploration
 - Destination Application: Machine Learning
 - Topic Summary
- Module Summary

Module 4: Data Pipeline Development Process (33 min)

- Module Objectives
- Requirements Analysis
 - Requirements Analysis Business
 - Requirements Analysis Data Sources
 - Requirements Analysis Transformation
 - Requirements Analysis Technical
 - Requirements Analysis Operational
 - Topic Summary
- Data Analysis
 - Requirements Analysis Inventory
 - Requirements Analysis Volume and Velocity



DE-01 Data Pipeline Engineering

- Requirements Analysis Data Quality
- Requirements Analysis Access and Security
- Requirements Analysis Mapping
- Requirements Analysis Validation
- Topic Summary
- Design and Build
 - Design & Build Ingest
 - Design & Build Process
 - Design & Build Transform
 - Design & Build Target
 - Topic Summary
- Testing and Deployment
 - Testing & Deployment Testing
 - Testing & Deployment Performance
 - Testing & Deployment Data quality
 - Testing & Deployment Strategy
 - Testing & Deployment Pre-Production
 - Testing & Deployment Production
 - Development Lifecycles
 - Topic Summary
- Module Summary

Module 5: Data Pipeline Operations (16 min)

- Module Objectives
- Orchestration, Monitoring & Observability
 - Orchestration Scheduling and Automation
 - Orchestration Dependency Management
 - Orchestration Error Handling and Recovery
 - Orchestration Parallelism and Optimization
 - Orchestration Tools
 - Monitoring & Observability Key Metrics
 - Monitoring & Observability Data Quality
 - Monitoring & Observability Tools
 - Topic Summary
- Service Levels
 - Service Levels
 - Change Management
 - Change Management Benefits
- Module Summary

Module 6: Data Pipeline Best Practices (15 min)

- Module Objectives
- Best Practices in Design
 - Best Practice: Design



DE-01 Data Pipeline Engineering

- Best Practice: Hybrid
- Best Practice: Data Quality
- Best Practice: Robustness
- Topic Summary
- Best Practices in Operations
 - Best Practice: Tracking
 - Best Practice: Cost
 - Best Practice: CI/CD
 - Best Practice: Culture
 - Best Practice: Security
 - Topic Summary
- Module Summary
- Course Summary