Enterprise Data Quality, Metadata and Data Lineage Management

Duration: 1 Week

Day 1: Data Quality Foundations & Assessment

Module 1: Enterprise Data Quality Fundamentals

- Data quality dimensions: completeness, accuracy, consistency, timeliness, uniqueness, validity
- Cost of poor data quality: financial, operational, and regulatory impact
- Data quality maturity model: reactive to proactive to predictive
- Data quality roles: data owners, stewards, custodians, and quality analysts
- Data quality tools landscape: Informatica DQ, Talend, IBM InfoSphere, Microsoft Purview

Lab Exercises:

1. Assess a sample dataset across 6 data quality dimensions; score each and identify the top 3 failure areas
2. Calculate the estimated cost of poor data quality for a given business scenario using a structured framework

Module 2: Data Profiling & Quality Assessment

- Data profiling: column analysis, pattern analysis, and cross-table analysis
- Null analysis, cardinality, frequency distributions, and outlier detection
- Referential integrity checks: orphaned records and dangling foreign keys
- Cross-system data consistency checks
- Profiling tools: Python (pandas-profiling, ydata-profiling), Great Expectations, SQL

Lab Exercises:

3. Profile a multi-table dataset using ydata-profiling; document nulls, cardinality, and anomalies for each column
4. Run cross-table referential integrity checks using SQL; identify and log all orphaned and mismatched records

Day 2: Data Quality Rules, Monitoring & Remediation

Module 3: Data Quality Rules & Validation Framework

- Rule categories: completeness rules, format rules, range rules, referential rules, business rules
- Writing declarative data quality rules in SQL and Python
- Great Expectations: expectations, suites, checkpoints, and data docs
- Rule severity levels: blocking, warning, and informational
- Reject and quarantine patterns for failed records

Lab Exercises:

5. Define and implement 10 data quality rules in Great Expectations for a sample dataset; generate a data docs report
6. Build a reject table pattern: route failed records to a quarantine table with error code, rule name, and timestamp

Module 4: Data Quality Monitoring & Scorecards

- Continuous data quality monitoring: batch and near-real-time
- Data quality KPIs: pass rate, error rate, trend over time
- Data quality scorecards: dimension-level scoring per domain
- Alerting on quality threshold breaches
- Data quality dashboards in Power BI and Grafana

Lab Exercises:

7. Build a data quality scorecard tracking 5 dimensions across 3 data domains with weekly trend charts
8. Configure threshold-based alerts for a critical dataset; trigger and verify alert notification on breach

Day 3: Metadata Management

Module 5: Metadata Types & Business Glossary

- Metadata types: business, technical, operational, and process metadata
- Business glossary: terms, definitions, owners, synonyms, and examples
- Data dictionary: column-level metadata, data types, nullability, and descriptions
- Metadata governance: ownership, review cycles, and approval workflows
- Metadata tools: Collibra, Alation, Microsoft Purview, Apache Atlas

Lab Exercises:

9. Build a business glossary for 25 core enterprise data terms with owners, definitions, related terms, and examples
10. Create a data dictionary for a 5-table schema: document every column with type, nullability, and business description

Module 6: Technical Metadata & Metadata Catalogue

- Technical metadata: schema, tables, columns, data types, indexes, partitions
- Operational metadata: job run times, row counts, load durations, error counts
- Metadata catalogue architecture: harvesting, storage, search, and stewardship
- Automated metadata harvesting from databases, pipelines, and BI tools
- Metadata search and discovery: enabling self-service data access

Lab Exercises:

11. Set up an Apache Atlas or Microsoft Purview metadata catalogue; harvest metadata from a PostgreSQL database
12. Register operational pipeline metadata: log job name, run time, row count, and status for every pipeline execution

Day 4: Data Lineage Management

Module 7: Data Lineage Concepts & Architecture

- Data lineage types: column-level, table-level, system-level, and end-to-end
- Lineage capture methods: manual, automated parsing, API-based, and log mining
- Lineage use cases: impact analysis, root cause, regulatory traceability, data trust
- Lineage tools: OpenLineage, Marquez, Microsoft Purview, Collibra, Informatica
- OpenLineage standard: events, run metadata, dataset facets, and job facets

Lab Exercises:

13. Map end-to-end lineage for a KPI metric from source system through transformations to the BI report
14. Implement OpenLineage event emission for a Python ETL pipeline; view lineage in Marquez UI

Module 8: Column-Level Lineage & Impact Analysis

- Column-level lineage: tracking field transformations across pipeline stages
- Lineage for derived columns: aggregations, joins, and business logic transformations
- Impact analysis: identifying all downstream consumers of a changed dataset or column
- Root cause analysis using lineage: tracing data errors back to their source
- Lineage in dbt: ref(), source(), and the dbt lineage DAG

Lab Exercises:

15. Trace column-level lineage for 5 derived KPI fields through a multi-stage pipeline; document each transformation

16. Run an impact analysis for a proposed schema change; identify and list all affected downstream tables, reports, and models

Day 5: Integration, Governance & Capstone

Module 9: Integrating DQ, Metadata & Lineage into Governance

- Data governance platform: integrating quality, metadata, and lineage in one view
- Trusted data asset certification: combining quality score, lineage, and metadata completeness
- Stewardship workflows: assigning, escalating, and resolving data issues
- Regulatory use cases: GDPR data flow mapping, BCBS 239 data lineage for banks
- Reporting to executives: data trust index and governance dashboard

Lab Exercises:

17. Build a data trust scorecard combining quality score, metadata completeness, and lineage coverage for 3 datasets
18. Map data flows for a GDPR compliance report: source, processing purpose, lineage path, and retention policy

Module 10: Capstone Project – End-to-End Data Quality & Lineage

- Capstone scenario: multi-source enterprise dataset with quality and lineage requirements
- Data profiling and quality rule implementation
- Metadata catalogue population: business glossary, data dictionary, technical metadata
- End-to-end lineage implementation using OpenLineage or dbt
- Presentation: quality scorecard, lineage diagram, governance recommendations

Lab Exercises:

19. Complete end-to-end: profile dataset, define and run quality rules, populate metadata catalogue, and implement lineage
20. Present to cohort: quality scorecard, lineage diagram, stewardship actions taken, and governance improvement roadmap