Databricks Quality Engineering Workshop

2 Enterprise Capstone Projects • 2-Day Strategy & Execution

TARGET AUDIENCE

- QE Leads & Architects
- Data Engineers
- Governance / GRC
- MLOps Engineers

KEY OUTCOMES

- Production-Grade DLT
- Unity Catalog Governance
- AI Agents for DQ
- RFP Pitch Ready

CAPSTONE PROJECT 01

Retail/CPG Lakehouse Quality

Implement an end-to-end QE strategy for product master & inventory. Includes DLT expectations, SLA-driven workflows, and AI-powered drift detection.

CAPSTONE PROJECT 02

FSI Claims & Compliance

Build a regulatory-ready claims engine. Focus on validation automation, complex anomaly detection, and explainable lineage for audits.

DataQualityOps

Updated: January 2026

Workshop Agenda

Modules 1–8: From Lakehouse Strategy to AI-Driven QE

Total Duration: 12 Hours

2 Days Intensive

DAY 1

8 Hours • Strategy & Architecture

M1: Enterprise Lakehouse QE Blueprint — 2h

QA Strategy (Bronze/Silver/Gold), Data Mesh QE, RFP scoring impact.

M2: Native Data Validation Framework — 2.5h

Delta Expectations, DLT Validation, SLA workflows & Breakpoints.

M3: Data Contracts & Governance — 2h

Unity Catalog schema enforcement, Lineage-based compliance, Scorecards.

M4: Intelligent Validation Automation — 1.5h

Rule automation, AI Anomaly Detection, Photon performance testing.

DAY 2

4 Hours • AI & Business Narrative

M5: AI & Agents for Data Quality — 1.5h

AI Agents for drift/RCA, Vector Search RAG, Auto Rule Generation.

M6: MLflow Governance for Models — 1h

Monitoring Freshness, Feature Drift, Model Explainability.

M7: Retail/CPG QE Case Study — 1h

Workshop: Product Master & Inventory validation, Business outcomes.

M8: RFP Narrative & Final Pitch — 30m

Differentiation storyline, Value prop templates, Next skills roadmap.

CAPSTONE PROJECT 01

Retail/CPG: Product Master & Inventory Quality

Complexity Level

PROJECT CONTEXT

A major retailer struggles with a fragmented product master and inconsistent supplier data, leading to store inventory stockouts and delayed merchandising insights. The goal is to unify these streams into a trusted Lakehouse foundation.

BUSINESS GOALS

Reduce DQ Escalations
Cut data quality incidents by 30–50% via proactive validation

Inventory Freshness
Ensure stock levels are updated < 4 hours from close of business

Accelerated Analytics
Enable faster category performance reporting for merchandising

KEY PERSONAS

- QE Lead / Architect
- Data Engineer
- Data Steward
- Merchandising Analyst
- Security / GRC

SUCCESS TARGETS

< 4 Hours

DATA LATENCY

Capstone 1: Retail/CPG Architecture

End-to-End Lakehouse QE Flow with Native Validation

Architecture

UNITY CATALOG — Access Controls (ACLs) · PII Tagging · Data Lineage · Discovery

Validates across Bronze → Silver → Gold

- Supplier Feeds
- POS Transactions
- Inventory Snapshots

Bronze — AUTO LOADER

- Raw Ingestion
- Append Only
- JSON/Parquet

QE Check:
Schema Validation, File Corruption

Silver — DLT CORE

- Clean & Dedupe
- Join Products + Sales
- Conformed Tables

QE Check (Expectations):
Nulls, Ref. Integrity, Constraints

Gold — PHOTON SQL

- Aggregations
- Availability KPIs
- Daily Snapshots

QE Check:
Business Logic, KPI Thresholds

- KPI Dashboard
- Demand Forecast
- Stockout Alerts

OBSERVABILITY — System Tables · DLT Event Logs · Alerting (PagerDuty) · DQ Scorecard

Monitors SLA & Data Quality

CI/CD via Databricks Asset Bundles · Serverless Compute

Bronze Layer (Ingest)

Raw & Rescued

Ingest raw feeds with schema evolution. Preserve all history. Quarantine corrupt files.

`cloud_files_state` `_rescue_data`

Silver Layer (Refined)

Clean & Trusted

Deduplicated, type-casted, and enriched. Enforce referential integrity.

`IS NOT NULL` `Product_ID Unique`

Gold Layer (Business)

Audit Ready

Aggregated KPIs for merchandising. Strict business logic validation.

`Margin > 0%` `Stock >= 0`

DLT Validation Patterns

```
expect("valid_ts", "ts > '2020-01-01'")
```

WARN

Log metric, keep processing. Good for monitoring.

```
expect_or_drop("valid_id", "id IS NOT NULL")
```

DROP

Remove row from target. Prevent downstream pollution.

```
expect_or_fail("critical_sku", "sku IS NOT NULL")
```

HALT

Stop pipeline immediately. Requires intervention.

Data Contracts (Shift Left)

Enforce schema and constraints at the source before ingestion.

```
{
  "column": "store_id",
  "type": "integer",
```

Governance + Observability

Unity Catalog, Lineage Compliance & Operational Scorecards

Module 3 & 6 Focus

Unity Catalog Controls

Schema Enforcement

Rejects writes that violate schema on ingest (Bronze)

Table & Row ACLs

Fine-grained access control for sensitive PII/PHI columns

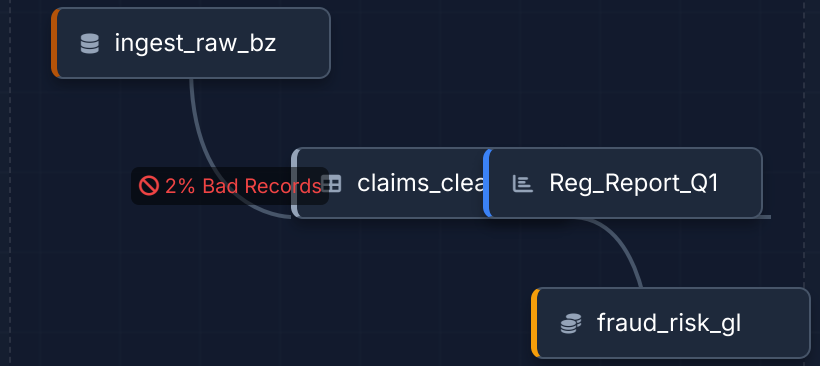
Automated PII Tagging

AI-based classification of credit cards, emails, SSNs

Time Travel & Audit

Immutable logs for every DML operation

Lineage & Compliance



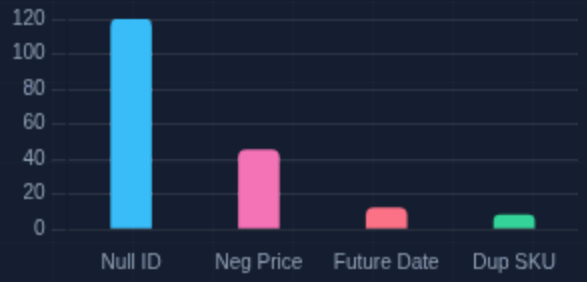
Audit Ready: Full lineage traceability from regulatory report back to raw source file.

Ops Scorecard

98.5%
OVERALL DQ SCORE

4h 12m
FRESHNESS SLA

Expectation Failures (Last 24h) — High Severity

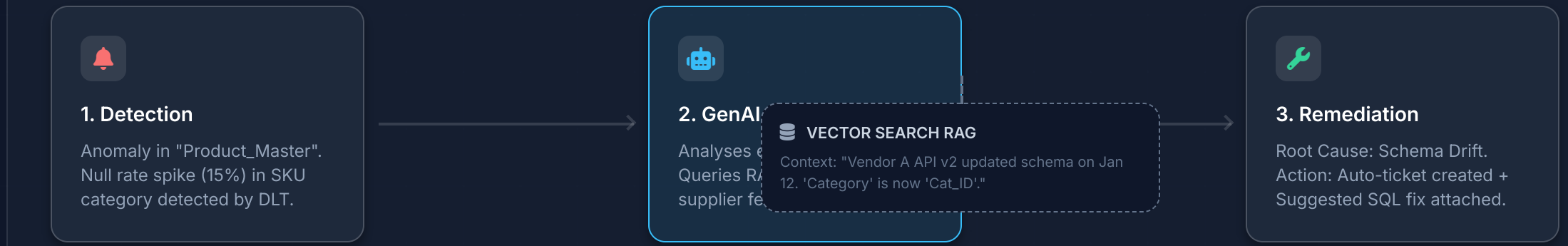


AI Agents & MLflow Governance

Capstone 1: Retail/CPG Intelligent Quality Automation

Module 5 & 6

AI-Driven Resolution Workflow



MLflow Governance & Monitoring

Feature Drift (PSI)

0.24

CRITICAL

Demand Forecasting Model
Drift in 'Price_Elasticity'

Data Freshness

< 2 min

HEALTHY

Inventory Silver Table
SLA Target: 15 min

Model Staging

v2.1

PENDING APPROVAL

Stockout Predictor
Requires: QA Sign-off

Capstone 1 Outcomes

KPIs, Acceptance Criteria & Final Deliverables

PROJECT COMPLETE

94/100

DATA QUALITY SCORE

↑ Above Target (90)

98.5%

SLA ATTAINMENT

✓ Inventory < 4h

< 2h

MTTR (FIX RATE)

↓ Reduced from 24h

-20%

VALIDATION COST

↓ Native vs External Tool

Acceptance Criteria (Definition of Done)

Silver Layer Schema Conformance

≥ 98% of records pass strict schema validation; invalid records quarantined to `_rescue` table.

Lineage-Verified Availability KPI

On-shelf availability metric computed with full Unity Catalog lineage for audit traceability.

SLA Latency Controls

Inventory updates delivered within 4 hours of POS close; Alert configured in PagerDuty.

Key Deliverables

DLT Pipeline Repo

Python/SQL DLT code with Expectations

UC Governance Policies

JSON definitions for Grants & Tags

DQ Scorecard Dashboard

Databricks SQL Dashboard (.json export)

AI Assistant Notebook

CAPSTONE PROJECT 02

FSI: Insurance Claims Lakehouse QE

Complexity Level

PROJECT CONTEXT

> Processing claims from multiple Third-Party Administrators (TPAs) is plagued by inconsistent schemas and PHI risks. This leads to payment leakage, unreliable fraud signals, and challenges in meeting strict regulatory reporting deadlines.

BUSINESS GOALS

Reduce Claims Leakage
Minimize overpayments by validating claim lines against policy limits

Fraud Signal Reliability
Ensure high-quality features for ML fraud detection models

Regulatory Audit Readiness
Automate lineage & evidence packs for compliance audits

KEY PERSONAS

- QE Lead / Architect
- Data Engineer
- Claims Operations
- Fraud Analytics
- Compliance / Legal

SUCCESS TARGETS

100%
AUDIT LINEAGE

Capstone 2: FSI/Insurance Architecture

Controls-First Lakehouse: Compliance, Lineage & Native Validation

Financial Services

UC GOVERNANCE — Row/Col Masking | Sensitive Data Tags (PII) | Audit Logs | Approval Workflow — Enforced Compliance & Security

Claims Files (TPA)
Policy Data
Provider Networks

Bronze — CLOUD FILES

- Raw Landing
- PII Tokenization
- Quarantine Bad Lines

QE Check:
Contract Enforcement, Malware Scan

Silver — DLT PIPELINES

- Entity Resolution
- Claim Type Std.
- Deduplication

QE Check (Strict):
Provider License, Policy Active

Gold — PHOTON SQL

- Loss Triangles
- Fraud Features
- Reg. Reporting

QE Check:
Cross-ref Consistency, Aggregates

Audit Evidence
Fraud Detection
Regulatory Reports

LINEAGE & OBS — E2E Lineage Graph | Evidence Packs | SLA Breach Alerts | DQ Dashboards — Audit-Ready Traceability

Photon Accelerated | CI/CD Pipeline Promotion | FinOps Tagging

Bronze (Raw Claims)

Untrusted

Ingest EDI/JSON from TPAs. Immediate PII/PHI tagging and quarantine of malformed files.

`tag_phi_columns` `quarantine_bad_fmt`

Silver (Standardized)

Conformed

Validate provider licenses and cross-reference policy coverage. Detect logical conflicts.

`claim_dt <= today` `valid_npi_lookup`

Gold (Regulatory)

Audit Grade

Loss triangles for reserving and fraud features. Zero-tolerance for data drift.

`reserve_amt >= 0` `fraud_score_valid`

Rule Automation & AI Functions

```
expect_all("claims_logic",
valid_dates_macro)
```

MACRO

Reusable logic across 50+ state pipelines.

```
expect("amt_ok", "ai_anomaly(amt) < 0.9")
```

AI FUNC

Detect statistical outliers in payment amounts.

```
on_fail("critical_breach") { halt_job() }
```

BREAK

Prevent bad data from entering Gold layer.

SLA Workflows & TCO

Native DLT validation eliminates data movement to external DQ tools, reducing latency and licensing costs.

↓ 30% Total Cost

-4hr Time to Insight

Capstone 2: Controls + Compliance

Evidence Packs, Explainability & MLflow Governance for Insurance

FINANCIAL SERVICES

Privacy & Masking

Dynamic Views

Real-time masking based on user group permissions.

```
CASE WHEN is_member('fraud_ops')
THEN ssn ELSE '***-**-****' END
```

Immutable Audit

System Tables

Complete traceability of who accessed what claim.

```
SELECT user_identity.email,
action_name
FROM system.access.audit
```

Retention Policy

Time Travel & Vacuum

Retain history for 7 years (Reg compliance), hard delete PII on request.

```
VACUUM claims_silver RETAIN 168
HOURS;
-- Plus archival into cold storage
```

Regulatory Evidence Pack Workflow

Automated Reporting Job

Explainability & Gov

TOP FRAUD FACTORS (SHAP)

MLFLOW REGISTRY

Claims_Fraud_Model v2.1

Winning Storyline

"Shift Left" validation + "Unity Catalog" governance delivers trust without the tool tax.

Move beyond simple validation. Pitch a unified data intelligence platform that prevents bad data before it enters the ecosystem.

Lower TCO & Complexity

Native DLT validation removes need for Informatica/Talend DQ. Serverless compute scales to zero.

Audit-Ready Confidence

End-to-end lineage (Bronze to Report) satisfies regulatory compliance (GDPR/CCPA) automatically.

AI-Driven Productivity

GenAI agents reduce "mean-time-to-resolution" for data incidents by 60% via auto-root cause analysis.

Proposal Section Template

1. Executive Summary — Page 1-2

- **Problem Framing:** Current cost of poor data quality (escalations, fines).
- **Vision:** The "Reliable Lakehouse" Outcome.
- **Success Metrics:** < 4h freshness, 99.9% availability.

2. Target Architecture — Diagrams

- **Bronze/Silver/Gold Flow:** Map data products to layers.
- **Control Plane:** Unity Catalog governance overlay.
- **Tech Stack:** DLT, Photon, MLflow, Databricks SQL.

3. QE Implementation Plan — Timeline

- **Phase 1 (Weeks 1-4):** Ingest & Bronze controls.
- **Phase 2 (Weeks 5-8):** Silver transformation & Business Rules.
- **Phase 3 (Weeks 9-12):** Observability Dashboards & Alerting.

4. Investment & ROI — Financials

- **Resource Plan:** 1 Architect, 2 Data Engineers.

Execution Plan & Roadmap

2-Day Delivery, Risk Management & Assessment

Workshop Close

2-Day Delivery Plan

D1 — Foundation & Native Validation (8h)

Modules 1–4: Architecture setup, DLT pipeline implementation, Unity Catalog governance policies, and initial observability scorecard.

D2 — Intelligent QE & Pitch (4h)

Modules 5–8: Integrating AI Agents, Vector Search RAG, MLflow governance models, and final RFP narrative construction.

Next Skills Roadmap

- Data Engineer Pro
- UC Admin Specialist
- GenAI Fundamentals
- MLflow Ops
- Advanced DLT

Capstone Scoring Rubric (100 Pts)

Architecture	25 pts
QE Automation	25 pts
Governance / UC	20 pts
Observability	15 pts
AI & MLflow	10 pts
RFP Narrative	5 pts

Key Risks & Mitigations

Upstream Schema Changes

Risk: Breaking ingestion pipelines

✓ Data Contracts & Schema Evolution

Missed SLAs

Risk: Stale dashboards for business

✓ Retry Policies & Auto-Scaling

Compute Cost Overruns