Logistics Databricks Data Engineering Capstone Project

Objective

Build a complete end-to-end data engineering pipeline for ABC Express using Databricks Lakehouse Platform to transform raw operational data into actionable business intelligence, predictive analytics, and operational optimization. The solution must implement enterprise-grade data governance through Unity Catalog while processing real-time streaming data and providing comprehensive operational dashboards.

Dataset Schema

Unity Catalog Structure

- **Catalog:** `logistics`
- **Schema:** `workshop`
- **Volume:** `/Volumes/logistics/workshop/datasets/`

Data Sources

Streaming Data (Real-time JSON streams)

- **Package Scans** (`package_scans.jsonl`): 50+ events/sec package tracking
    - Fields: package_id, scan_timestamp, facility_id, scan_type, status
- **Vehicle Telemetry** (`vehicle_telemetry.jsonl`): 25+ events/sec fleet data
    - Fields: vehicle_id, timestamp, lat, lon, speed, fuel_level, engine_status
- **Facility Events** (`facility_events.jsonl`): 15+ events/sec operations
    - Fields: facility_id, event_timestamp, event_type, capacity_utilization, temperature

Historical Batch Data (Parquet files)

- **Historical Packages** (`historical_packages.parquet`): 500K package records
- **Weather Data** (`weather_data.parquet`): 365 days × 5K facilities
- **Route Performance** (`route_performance.parquet`): 100K route metrics

Reference Dimensions (Parquet files)

- **Facilities** (`facilities.parquet`): 5K global facilities
- **Customers** (`customers.parquet`): 100K customer records
- **Vehicles** (`vehicles.parquet`): 180K vehicle fleet
- **Routes** (`routes.parquet`): 50K route definitions

Deliverables

Technical Implementation

- **Unity Catalog Implementation** (Bronze-Silver-Gold Architecture)
- **Real-time Streaming Pipeline** using Auto Loader
- **Delta Live Tables Pipeline** with data quality expectations
- **Interactive Dashboards** using Databricks SQL
- **Monitoring & Alerting Framework**

Notebook

Complete Databricks notebooks covering:

- Data ingestion and preprocessing
- Streaming pipeline implementation
- Data quality checks and transformations
- Analytics and visualization
- Performance optimization techniques

Presentation Deck

1. Introduction & Business Context (1 min)

- ABC Express global logistics operations challenge
- Need for real-time operational intelligence and predictive analytics
- Impact of data-driven decision-making on logistics efficiency

2. Objective & Success Criteria (1 min)

- Build an enterprise-grade data platform on Databricks
- Process 100+ events/second with <30 second latency
- Achieve 99.9% pipeline uptime and 99.5% data accuracy
- Enable real-time KPI monitoring and operational optimization

3. Data & Approach (2 min)

- **Data Sources:** streaming sources, batch datasets, reference tables
- **Architecture:** Bronze-Silver-Gold medallion architecture using Unity Catalog
- **Technology Stack:** Delta Live Tables, Auto Loader, Databricks SQL
- **Challenges:** Schema evolution, late-arriving data, peak load handling

4. Key Insights & Findings (2-3 min)

- **Operational Efficiency:** Real-time facility capacity optimization
- **Delivery Performance:** SLA compliance tracking and exception handling
- **Predictive Analytics:** Weather impact correlation with delivery delays
- **Cost Optimization:** Route performance and vehicle utilization insights
- **Visual Dashboards:** Executive KPIs, operational metrics, data quality monitoring

5. Solution & Results (2 min)

- **Performance Metrics:** Sub-30 second latency, 100+ events/sec throughput
- **Data Quality:** Automated expectations with 99.5% accuracy
- **Business Impact:** Real-time visibility into global operations
- **Scalability:** Handles 300% Black Friday surge capacity

6. Business Impact & Recommendations (1 min)

- **Cost Reduction:** Optimized routing and resource allocation
- **Customer Satisfaction:** Proactive exception handling and accurate ETAs
- **Operational Excellence:** Real-time monitoring and automated alerting
- **Next Steps:**

Day-wise Plan

Day 1

10:00 AM - Team Allocation

- Form teams of 4-5 members
- Assign team leads and roles (Data Engineer, Analytics Expert, Dashboard Developer)
- Review project requirements and success criteria

10:-30 AM - Introduction to Problem Statement & Business Context

- Deep dive into ABC Express logistics operations
- Understanding dataset relationships and business rules
- Technical architecture overview and Unity Catalog setup

11:00 AM - Teams Start Working

- Environment setup and data exploration
- Unity Catalog structure creation
- Initial data ingestion and quality assessment

Checkpoint 1 – 2:00 PM

- **Review Progress:** Data ingestion completion, Unity Catalog setup
- **Design & Architecture:** Bronze-Silver-Gold layer design validation
- **Technical Discussion:** Auto Loader configuration and streaming setup
- **Next Steps:** Begin Delta Live Tables pipeline development

Checkpoint 2 – 4:30 PM

- **Progress Review:** Streaming pipeline implementation status
- **Technical Challenges:** Schema evolution, watermarking, error handling
- **Dashboard Planning:** KPI identification and visualization strategy
- **Wrap-up:** 5:00 PM with Day 2 objectives and deliverable priorities

Day 2

Checkpoint 3 – 11:00 AM

- **Progress Review:** Delta Live Tables pipeline completion
- **Dashboard Development:** Executive and operational dashboards
- **Data Quality Monitoring:** Expectations implementation and alerting setup
- **Presentation Preparation:** Begin slide deck creation and notebook documentation

Checkpoint 4 – 2:00 PM

- **Final Technical Review:** End-to-end pipeline testing and performance validation
- **Presentation Finalization:** Slide deck completion and demo preparation
- **Quality Assurance:** Code review, documentation, and troubleshooting

Final Presentations – 3:00 PM

Each team presents for **~15 minutes:**

Presentation Deck (10 minutes):

- Business context and technical approach
- Architecture and implementation highlights
- Key insights and business impact
- Performance metrics and optimization results

Notebook Walkthrough (5 minutes):

- Live demonstration of streaming pipeline
- Dashboard functionality and real-time updates
- Data quality monitoring and alerting
- Advanced features implementation

Feedback Session:

- Technical implementation assessment
- Business impact evaluation
- Recommendations for production deployment
- Best practices and optimization opportunities

Wrap-up by 5:00 PM

Success Metrics & Evaluation Criteria

Technical Excellence (45%)

- **Performance:** <30 sec latency, 100+ events/sec throughput, 99.9% uptime
- **Data Quality:** 99.5% accuracy with automated monitoring
- **Architecture:** Complete Bronze-Silver-Gold implementation
- **Advanced Features:** Liquid clustering, Photon acceleration, Delta Sharing

Business Impact (35%)

- **Operational Dashboards:** Real-time KPI monitoring and alerting
- **Executive Reporting:** Strategic insights and performance analytics
- **Use Case Implementation:** Black Friday surge, weather disruption scenarios
- **ROI Analysis:** Quantified cost savings and efficiency improvements

Innovation & Optimization (20%)

- **Creative Solutions:** Advanced Databricks features utilization
- **Performance Tuning:** Query optimization and resource efficiency
- **Monitoring Framework:** Comprehensive alerting and incident response
- **Documentation:** Technical and business implementation guides

Getting Started

Environment Setup

1. **Databricks Workspace:** Unity Catalog-enabled environment
2. **Dataset Access:** Import provided datasets to Unity Catalog volume
3. **Compute Configuration:** Appropriate clusters for streaming and batch workloads
4. **Initial Setup:** Run `/Volumes/logistics/workshop/datasets/config/workshop_setup.py`

Implementation Sequence

1. **Unity Catalog Structure:** Create catalog, schema, and table hierarchy
2. **Data Ingestion:** Configure Auto Loader for streaming sources
3. **Delta Live Tables:** Implement Bronze-Silver-Gold pipeline
4. **Dashboard Creation:** Build executive and operational dashboards
5. **Monitoring Setup:** Configure data quality expectations and alerting
6. **Performance Optimization:** Implement advanced features and tuning