Complete Data Lineage Analysis

From Airbyte Ingestion to Business Intelligence

Lab 2 - Factory Pattern Implementation

MLOps Data Pipeline Analysis

Generated: September 28, 2025

Author: AI Data Analysis System

Executive Summary

This comprehensive analysis documents the complete data lineage from external API sources through Airbyte ingestion to business intelligence views. The system integrates 4+ external APIs integrated via Airbyte through 5 distinct processing stages to deliver Explainable data pipeline enabling self-service analytics.

Technical Achievements

- Airbyte-based ingestion handling complex nested JSON structures
- Comprehensive feature engineering with business context
- Weather-business correlation analysis and integration
- Self-documenting semantic views with explainable columns
- End-to-end data quality and audit trail maintenance

Business Impact Areas

- Weather Intelligence: Operational optimization based on weather conditions
- Customer Intelligence: Data-driven sales prioritization and targeting
- Operational Intelligence: Performance normalization and predictive analytics
- Integration Intelligence: Cross-domain insights and correlation analysis

Airbyte Data Ingestion

Weather API Integration

Source System	Weather API Service
API Endpoint	https://api.weather.service/v2/daily
Frequency	Daily at 06:00 UTC
Authentication	API Key based authentication
Data Format	JSON via REST API

Weather API Response Structure

weather_id: Unique identifier for weather record
date: ISO date format (YYYY-MM-DD)
location_id: Numeric location identifier
temperature: Temperature in Fahrenheit (float)
humidity: Relative humidity percentage (float)
precipitation: Precipitation amount in inches (float)
wind_speed: Wind speed in MPH (float)

Airbyte Transformation Pipeline

• 1. API Response Validation: Check data types and required fields
• 2. Duplicate Detection: Identify and handle duplicate records
• 3. Data Type Conversion: Ensure numeric fields are proper types
• 4. Timestamp Standardization: Convert to UTC timezone
• 5. Schema Validation: Verify against predefined schema
• 6. Data Quality Flags: Add quality indicators

Business Data Integration

Toast POS API Integration

The Toast Point of Sale system provides comprehensive transaction data through their REST API. Authentication uses OAuth 2.0 with refresh tokens, and data is ingested both real-time via webhooks and through hourly batch synchronization. The API returns complex nested JSON structures that require sophisticated flattening strategies to extract meaningful business metrics.

Toast API Object Structure

Order: Root order object with metadata

Check: Individual check within an order

Selection: Menu item selections with modifiers

Payment: Payment information and methods

Customer: Customer information when available

Feature Engineering Pipeline

Weather Feature Engineering

Temperature Feature Derivation

Source Transformation: temperature -> temperature_fahrenheit (direct)
Business Context: Temperature impacts customer comfort and outdoor activity preferences

Derived Temperature Features:

- `temperature_celsius: (temperature_fahrenheit - 32) * 5/9`
- `temperature_category: binning into Cold/Cool/Warm/Hot ranges`
- `temperature_deviation: difference from seasonal average`
- `temperature_trend: 7-day moving average calculation`

Composite Weather Intelligence

Outdoor Activity Score Algorithm

Formula: f(temperature, humidity, wind_speed, precipitation)
Business Use: Predict customer likelihood for outdoor dining and events

Calculation Logic:

- `Base score: 50`
- `Temperature adjustment: +20 for 60-80°F, +10 for 50-90°F, -20 otherwise`
- `Precipitation penalty: -15 if > 0`
- `Wind adjustment: -10 if > 15 MPH`
- `Humidity adjustment: +5 for 30-70%, -5 otherwise`
- `Final score: clipped to 0-100 range`

Business Feature Engineering

Revenue and Operational Metrics

Revenue Metrics:

- `order_value: sum of all check totals`
- `average_item_price: total_value / item_count`
- `party_size_adjusted_value: order_value / number_of_guests`
- `upsell_rate: (order_value - base_items) / base_items`

Operational Metrics:

- `service_duration_minutes: closed_date - opened_date`
- `order_complexity_score: function of item_count and modifications`
- `kitchen_efficiency: prep_time / item_count`
- `table_turnover_rate: orders_per_day / table_capacity`

Semantic Business Views

Weather Intelligence View

Purpose: Transform raw weather data into business intelligence
Business Context: Enable business users to understand weather impact on operations without needing meteorological expertise

Explainable Column Definitions

temperature_celsius: Fahrenheit temperature converted to Celsius using (F-32)*5/9 formula

temperature_category: Business-friendly temperature ranges: Freezing (<32°F), Cold (32-50°F), Cool (50-70°F), Warm (70-85°F), Hot (>85°F)

outdoor_activity_score: Composite score (0-100) indicating suitability for outdoor activities, calculated from temperature, humidity, wind, and precipitation

customer_comfort_index: Customer comfort level (0-100) for indoor/outdoor dining, optimized around 72°F temperature and 50% humidity

operational_impact_flag: Binary flag (0/1) indicating extreme weather conditions that may require operational adjustments

Customer Intelligence View

Purpose: Integrate customer and lead data for sales intelligence
Business Context: Provide sales teams with actionable intelligence for lead prioritization and outreach strategy

Customer Intelligence Column Definitions

priority_score: Weighted score (0-10) combining decision authority, company size, timing, and engagement signals

decision_maker_probability: ML model prediction (0-1) of whether contact has decision-making authority for purchases

engagement_likelihood: Historical pattern-based prediction (0-1) of positive response to outreach

revenue_potential: Estimated annual contract value based on industry benchmarks and company size

priority_category: Business-friendly prioritization: High (8.0+), Medium (6.0-7.9), Low (<6.0)

deal_size_category: Deal classification: Enterprise ($75K+), Mid-Market ($25K-$75K), SMB (<$25K)

Complete Data Lineage Flow

Purpose: Trace data from original API sources through to business intelligence views
Scope: Weather data, business transactions, customer intelligence, and integrated analytics
Methodology: Airbyte-based ingestion with feature engineering and semantic view layers

Five-Stage Data Pipeline

Stage 1: Ingestion

Description: Raw data ingestion from external APIs via Airbyte
Output: Raw data tables with basic validation and type conversion

Stage 2: Processing

Description: Data quality improvement and initial transformations
Output: Clean, validated data ready for feature engineering

Stage 3: Feature Engineering

Description: Business-relevant feature creation and enrichment
Output: Feature-rich datasets with business context

Stage 4: Semantic Views

Description: Business-friendly views with explainable columns
Output: Ready-to-use business intelligence views

Stage 5: Analytics

Description: Advanced analytics and machine learning applications
Output: Actionable business insights and recommendations

Data Quality Assurance

Quality Gates:

• API Response Validation: Ensure data completeness and type correctness
• Business Rule Validation: Apply domain-specific validation rules
• Referential Integrity: Maintain relationships between datasets
• Temporal Consistency: Ensure proper time-series continuity
• Statistical Validation: Detect outliers and anomalies

Quality Metrics:

• Completeness Score: Percentage of required fields populated

- Accuracy Score: Validation against known correct values
- Consistency Score: Internal consistency across related fields
- Timeliness Score: Data freshness and update frequency
- Validity Score: Conformance to business rules and constraints

Transformation Audit Trail

Weather Data Transformations

temperature_celsius: Source: API temperature_fahrenheit, Formula: (F-32)*5/9, Validation: Range check -50 to 50°C

outdoor_activity_score: Sources: temperature, humidity, wind, precipitation, Formula: Composite scoring algorithm, Range: 0-100

operational_impact_flag: Sources: temperature, precipitation, wind, Logic: Boolean OR of extreme conditions, Values: 0/1

Business Data Transformations

priority_score: Sources: decision_signals, company_size, industry, Formula: Weighted scoring algorithm, Range: 0-10

revenue_potential: Sources: industry, employee_count, company_revenue, Formula: Industry benchmark calculation, Currency: USD

weather_adjusted_revenue: Sources: daily_revenue, outdoor_activity_score, Formula: Revenue * weather_factor, Currency: USD

Data Governance Framework

The comprehensive data lineage system implements robust governance capabilities to ensure data quality, compliance, and business transparency. This framework supports enterprise-grade data management requirements while enabling self-service analytics for business users.

Lineage Tracking

Complete field-level lineage from source to consumption

Quality Assurance

Multi-stage validation with quality scoring

Audit Capability

Full transformation audit trail for compliance

Business Definitions

Domain-specific definitions for all derived metrics

Conclusion

This comprehensive data lineage analysis demonstrates a sophisticated approach to modern data engineering that bridges the gap between technical implementation and business value. The system successfully integrates multiple external data sources through Airbyte, applies intelligent feature engineering, and delivers business-friendly semantic views that enable self-service analytics. The five-stage pipeline (Ingestion → Processing → Feature Engineering → Semantic Views → Analytics) provides a scalable foundation for enterprise data operations. Quality gates and audit trails ensure data integrity, while explainable columns and business context make the system accessible to non-technical stakeholders. Key achievements include weather-business correlation analysis, customer intelligence scoring, and operational optimization capabilities. The complete field-level lineage documentation supports compliance requirements while enabling confident decision-making based on trusted data. This implementation serves as a template for modern data engineering practices that prioritize both technical excellence and business usability.

Report generated on September 28, 2025 at 09:09 PM