OmniStream: Multi-source Data Engineering Pipeline

OmniStream Logo

Project Overview

OmniStream is a sophisticated data engineering platform designed to process, monitor, and analyze real-time data from multiple sources. The platform features automated data quality controls, anomaly detection, and comprehensive dashboards for monitoring pipeline performance.

Key Features

- **Real-time Data Processing**: Ingest and process data from multiple sources with low latency
- **Automated Quality Controls**: Continuously monitor data quality and detect anomalies
- **Interactive Dashboards**: Visualize pipeline performance and data insights
- **Multi-stage Processing**: Implement a complete data engineering workflow from ingestion to analysis
- **Scalable Architecture**: Designed to handle growing data volumes across diverse sources

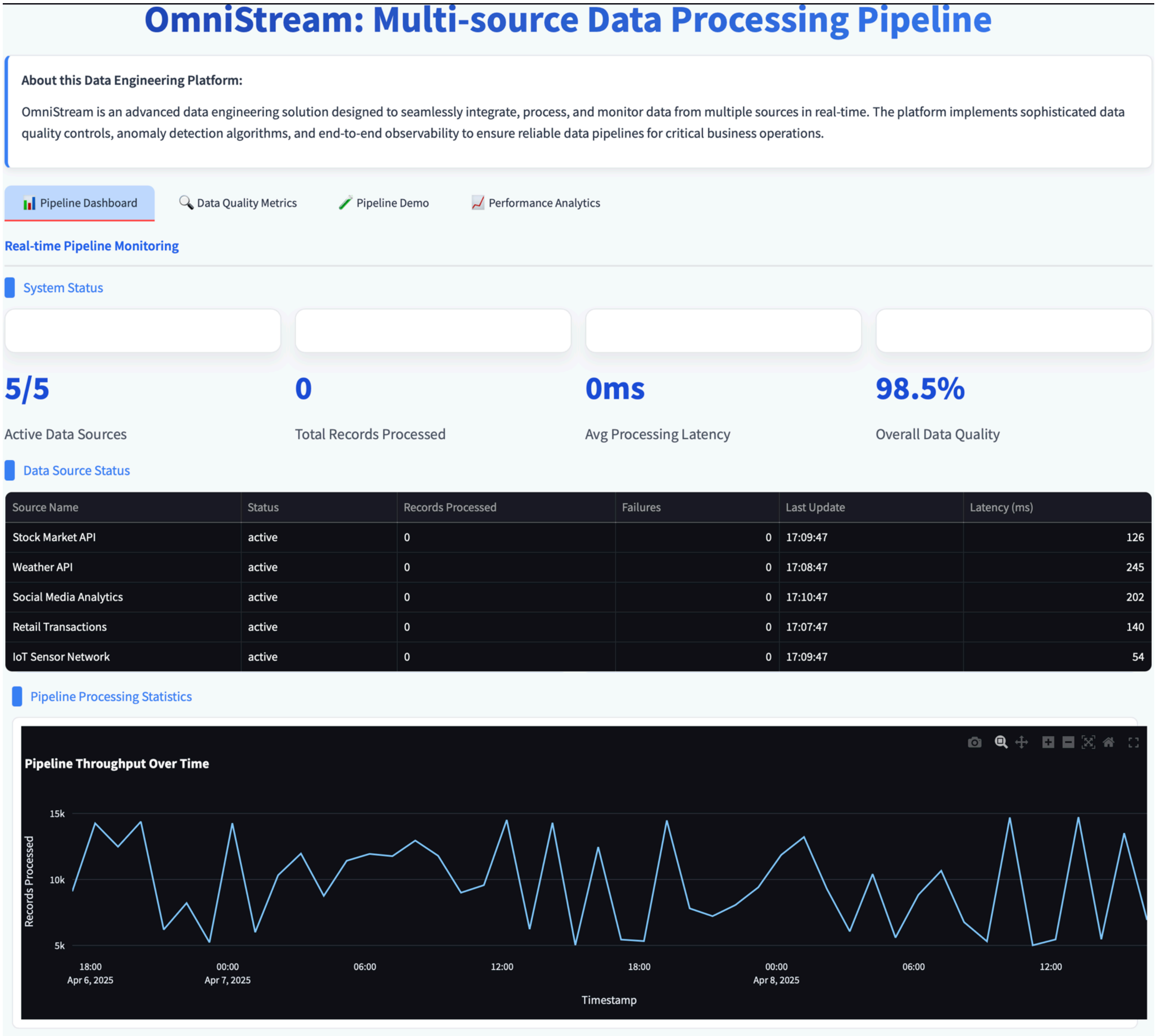
Dashboard Components

The application features four main dashboard tabs:

1. Pipeline Dashboard

The main monitoring interface that provides a real-time overview of the entire data pipeline, including:

- **System Status Cards**: Visual indicators of active data sources, records processed, processing latency, and data quality
- **Data Source Status Table**: Detailed status of each connected data source including processing metrics
- **Recent Alerts & Events**: Timeline of system activities and issues requiring attention
- **Throughput Chart**: Time-series visualization of data volume processed over time



OmniStream: Multi-source Data Processing Pipeline

About this Data Engineering Platform:

OmniStream is an advanced data engineering solution designed to seamlessly integrate, process, and monitor data from multiple sources in real-time. The platform implements sophisticated data quality controls, anomaly detection algorithms, and end-to-end observability to ensure reliable data pipelines for critical business operations.

Real-time Pipeline Monitoring

System Status

5/5
Active Data Sources

0
Total Records Processed

0ms
Avg Processing Latency

98.5%
Overall Data Quality

Data Source Status

Source Name	Status	Records Processed	Failures	Last Update	Latency (ms)
Stock Market API	active	0	0	17:09:47	126
Weather API	active	0	0	17:08:47	245
Social Media Analytics	active	0	0	17:10:47	202
Retail Transactions	active	0	0	17:07:47	140
IoT Sensor Network	active	0	0	17:09:47	54

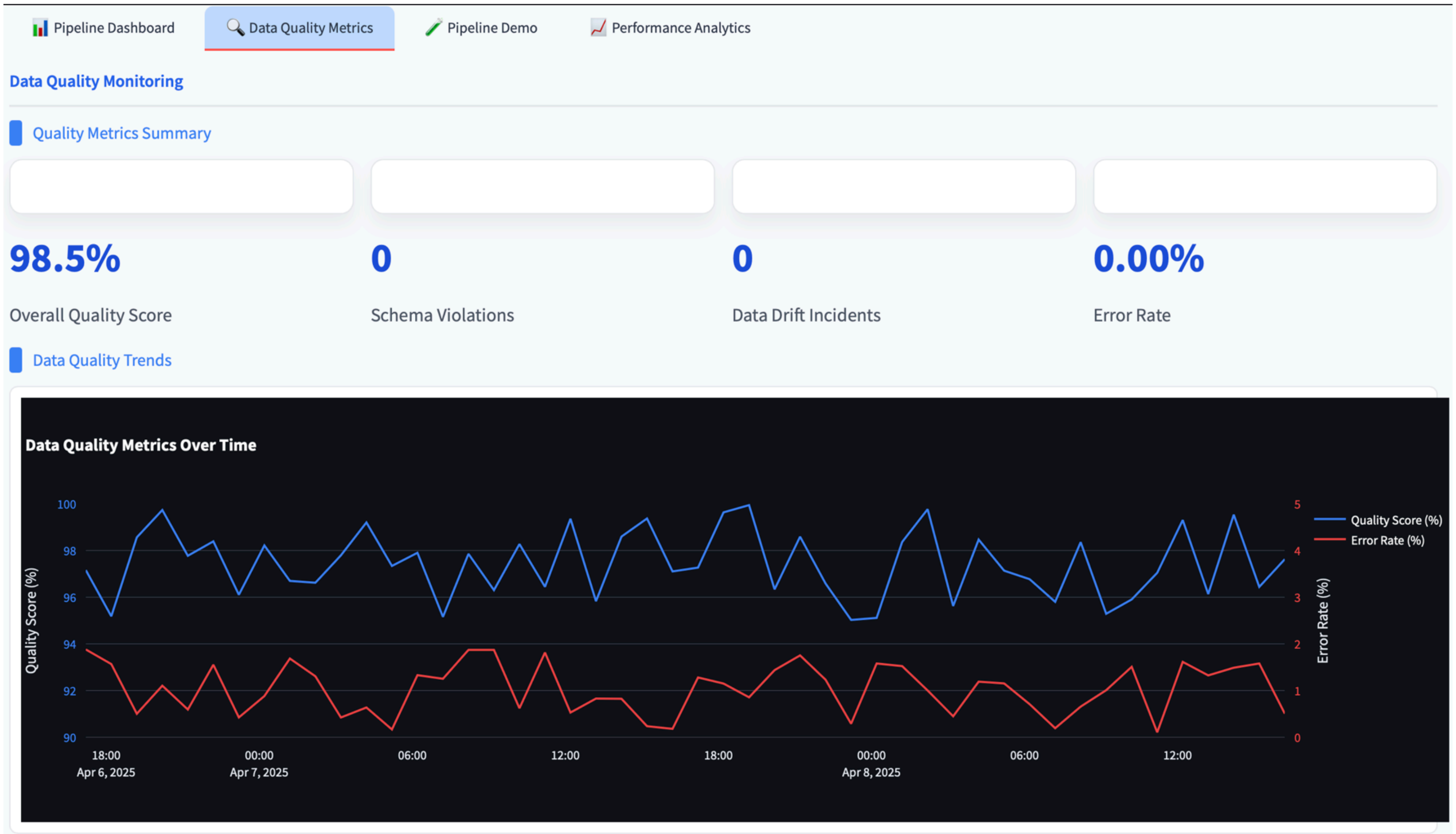
Pipeline Processing Statistics

Pipeline Throughput Over Time

2. Data Quality Metrics

Comprehensive visualization of data quality across the system:

- **Quality Score Metrics**: Overall quality metrics with breakdown of violations and incidents
- **Quality Trend Chart**: Dual-axis visualization showing quality score and error rate over time
- **Automated Data Quality Rules**: Table of configured quality validation rules with severity and status
- **Data Enrichment Processes**: Details of the enrichment processes applied to incoming data



Data Quality Monitoring

Quality Metrics Summary

98.5% Overall Quality Score

0 Schema Violations

0 Data Drift Incidents

0.00% Error Rate

Data Quality Trends

Data Quality

3. Pipeline Demo

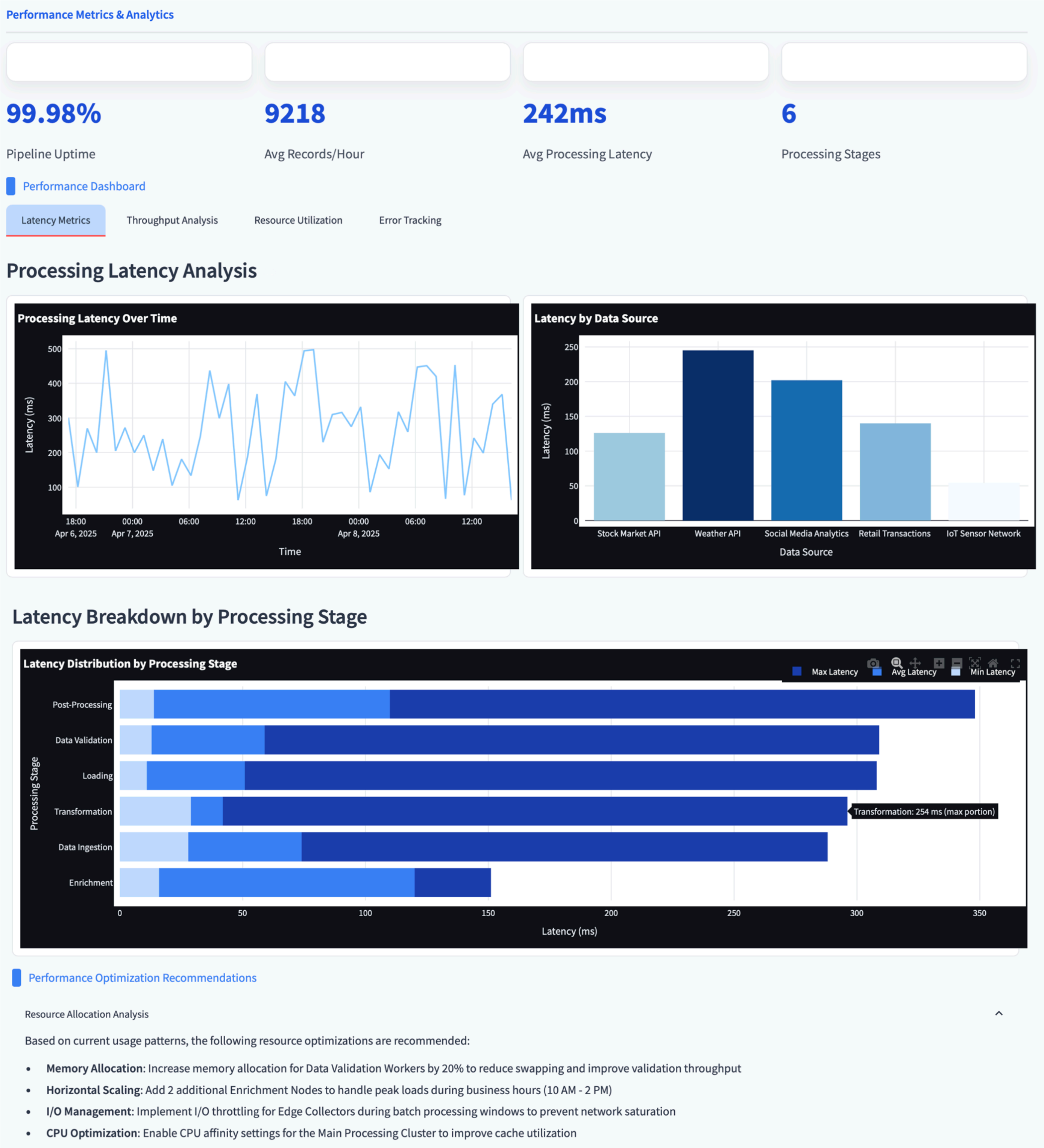
Interactive demonstration of the complete pipeline process:

- **Pipeline Architecture Diagram**: Visual representation of the data flow through the system
- **Step-by-step Execution**: Interactive walkthrough of each pipeline stage with real-time metrics
- **Technical Implementation Examples**: Code snippets showing how key components are implemented
- **Progress Visualization**: Real-time tracking of pipeline execution

4. Performance Analytics

Detailed performance metrics with sophisticated visualizations:

- **Enhanced Performance Dashboard**: Multi-tab interface for in-depth analysis
- **Latency Metrics**: Processing time analysis across different pipeline stages
- **Throughput Analysis**: Visualizations of data volume patterns and distribution
- **Resource Utilization**: System resource consumption monitoring
- **Error Tracking**: Detailed error rate analysis with breakdowns by type and source
- **Optimization Recommendations**: Actionable insights for improving pipeline performance



Performance Metrics & Analytics

99.98%

Pipeline Uptime

9218

Avg Records/Hour

242ms

Avg Processing Latency

6

Processing Stages

Performance Dashboard

Processing Latency Analysis

Latency Breakdown by Processing Stage

Performance Optimization Recommendations

Resource Allocation Analysis

Based on current usage patterns, the following resource optimizations are recommended:

- **Memory Allocation**: Increase memory allocation for Data Validation Workers by 20% to reduce swapping and improve validation throughput
- **Horizontal Scaling**: Add 2 additional Enrichment Nodes to handle peak loads during business hours (10 AM - 2 PM)
- **I/O Management**: Implement I/O throttling for Edge Collectors during batch processing windows to prevent network saturation
- **CPU Optimization**: Enable CPU affinity settings for the Main Processing Cluster to improve cache utilization

Visualization Types

The platform includes a variety of advanced data visualizations:

1. **Time-series Line Charts**: Track metrics like latency, throughput, and error rates over time
2. **Area Charts**: Visualize cumulative metrics like total records processed

3. **Bar Charts**: Compare metrics across different data sources or processing stages
4. **Donut Charts**: Show distribution of data volume or errors by category
5. **Stacked Bar Charts**: Display composite metrics with component breakdowns
6. **Heat Maps**: Visualize patterns in hourly or daily processing volumes
7. **Grouped Bar Charts**: Compare multiple metrics across different dimensions
8. **Scatter Plots**: Analyze relationships between different performance metrics
9. **Horizontal Bar Charts**: Compare metrics across different system components
10. **Multi-axis Charts**: Display related metrics with different scales on a single chart

Technical Implementation

OmniStream is built using a modern data engineering tech stack:

- **Front-end**: Streamlit for interactive dashboards and visualizations
- **Data Processing**: Simulated pipeline based on Apache Kafka, Spark, and Airflow patterns
- **Data Quality**: Implementation of Great Expectations patterns for quality monitoring
- **Monitoring**: Prometheus-style metrics collection and visualization
- **Database**: Connectivity with PostgreSQL for data persistence

Use Cases

This platform demonstrates advanced data engineering capabilities useful for:

1. **Enterprise Data Integration**: Combining data from multiple business systems
2. **IoT Data Processing**: Handling high-volume sensor data with quality controls
3. **Financial Data Analysis**: Processing market data feeds with strict quality requirements
4. **E-commerce Data Pipelines**: Managing customer, product, and transaction data flows
5. **Social Media Analytics**: Processing and analyzing engagement metrics in real-time

Getting Started

To run the application locally:

```
# Install dependencies
pip install streamlit pandas numpy plotly psycopg2-binary sqlalchemy

# Run the application
streamlit run app.py
```

Showcase

This project demonstrates advanced data engineering skills including:

- Data pipeline architecture design
- Real-time data processing
- Data quality monitoring and enforcement
- Performance optimization
- Advanced data visualization
- System observability implementation