ReSight: Building the Future of Pet Industry Analytics

ETL Infrastructure Strategic Overview

Nathan Lunceford

2025-01-21

Table of contents

1 Overview

ReSight is positioning itself to become the authoritative source of truth and insights for the U.S. pet industry. This transformation requires a robust, scalable ETL infrastructure capable of processing comprehensive industry data at scale.

1.1 Current State Analysis (2024)

Our current ETL infrastructure demonstrates significant daily processing capacity with notable variability in workload:

1.1.1 Daily Processing Statistics

Metric	Typical Day (Median)	Peak Day	Average (Mean)
Loads Processed	40	303	59.2
Rows Processed	70,588	824,719	105,218
Processing Window	Flexible	Flexible	Flexible

1.1.2 Load Distribution Analysis

- **Daily Load Range**: 1-303 loads per day
- **Typical Range (Q1-Q3)**: 24-65 loads per day
- **Standard Deviation**: 54.8 loads, indicating high variability
- **Processing Reliability**: 361 days of consistent operation with no outages

1.1.3 Data Volume Patterns

- **Daily Row Range**: 28-824,719 rows
- **Typical Range (Q1-Q3)**: 15,606-159,225 rows
- **Volume Variability**: Standard deviation of 115,282 rows
- **Processing Success Rate**: 100% (no missing days)

Metric	Value	Growth Factor

1.2 Target State (2026)

Metric	Value	Growth Factor
Daily Loads	400+/day	10x current median
Data Volume	700K+ rows/day typical	10x current median
Data Sources	1000+ integrated sources	10x current scale
Complexity	High (ML pipelines)	Significant increase
Processing Window	Near real-time requirements (if needed)	Minutes vs. flexible

1.3 Growth Requirements

Our next-generation ETL pipeline must support:

1. **Scalable Data Integration**
    - Handle 10x increase in daily load frequency
    - Process 10x current data volumes
    - Support 10x growth in data source connections
2. **Advanced Processing Capabilities**
    - Predictive analytics pipelines
    - Machine learning model integration
    - Market insight generation
3. **Enterprise-Scale Operations**
    - Consistent high-volume processing
    - 24/7 operation with high availability
    - Industry-leading security controls

2 Infrastructure Strategy

2.1 Core Requirements

2.1.1 Scalability

- Support for 400+ daily loads (10x current median)

- Peak capacity of 8M+ rows per day
- Elastic resource allocation
- Horizontal scaling support

2.1.2 Advanced Analytics

- ML pipeline integration
- Complex data transformations
- Data science toolkit support
- Predictive modeling capability

2.1.3 Reliability

- Zero downtime (matching current 100% reliability)
- Automated failover
- Comprehensive monitoring
- Proactive scaling

2.2 Key Performance Indicators

Metric	Current (2024)	Target (2026)
Daily Loads (Median)	40	400+
Peak Daily Loads	303	3000+
Daily Rows (Median)	70,588	700K+
Peak Daily Rows	824,719	8M+
Processing Latency	Hours	Minutes
Data Sources	~100	1000+

3 Implementation Roadmap

3.1 Phase 1: Foundation (Q1 2025)

- Scale current infrastructure to handle 2x current peak load
- Implement comprehensive monitoring
- Deploy new stream processing architecture

3.2 Phase 2: Scaling (Q2-Q3 2025)

- Expand data source integration capacity
- Implement ML pipeline framework

3.3 Phase 3: Optimization (Q4 2025)

- Scale to 5x current capacity
- Deploy advanced analytics capabilities

3.4 Phase 4: Enterprise Scale (2026)

- Achieve full target state capabilities
- Deploy full ML/AI integration

4 Conclusion

This ETL infrastructure strategy outlines our path from current state to future vision, supporting ReSight's goal of becoming the authoritative analytics platform for the pet industry. Our implementation roadmap ensures a methodical progression toward our 2026 targets while maintaining our current high standards of reliability and data quality.

Key success factors include:

- Maintaining 100% reliability while scaling 10x in load frequency
- Supporting 10x growth in data sources
- Enabling ML integration