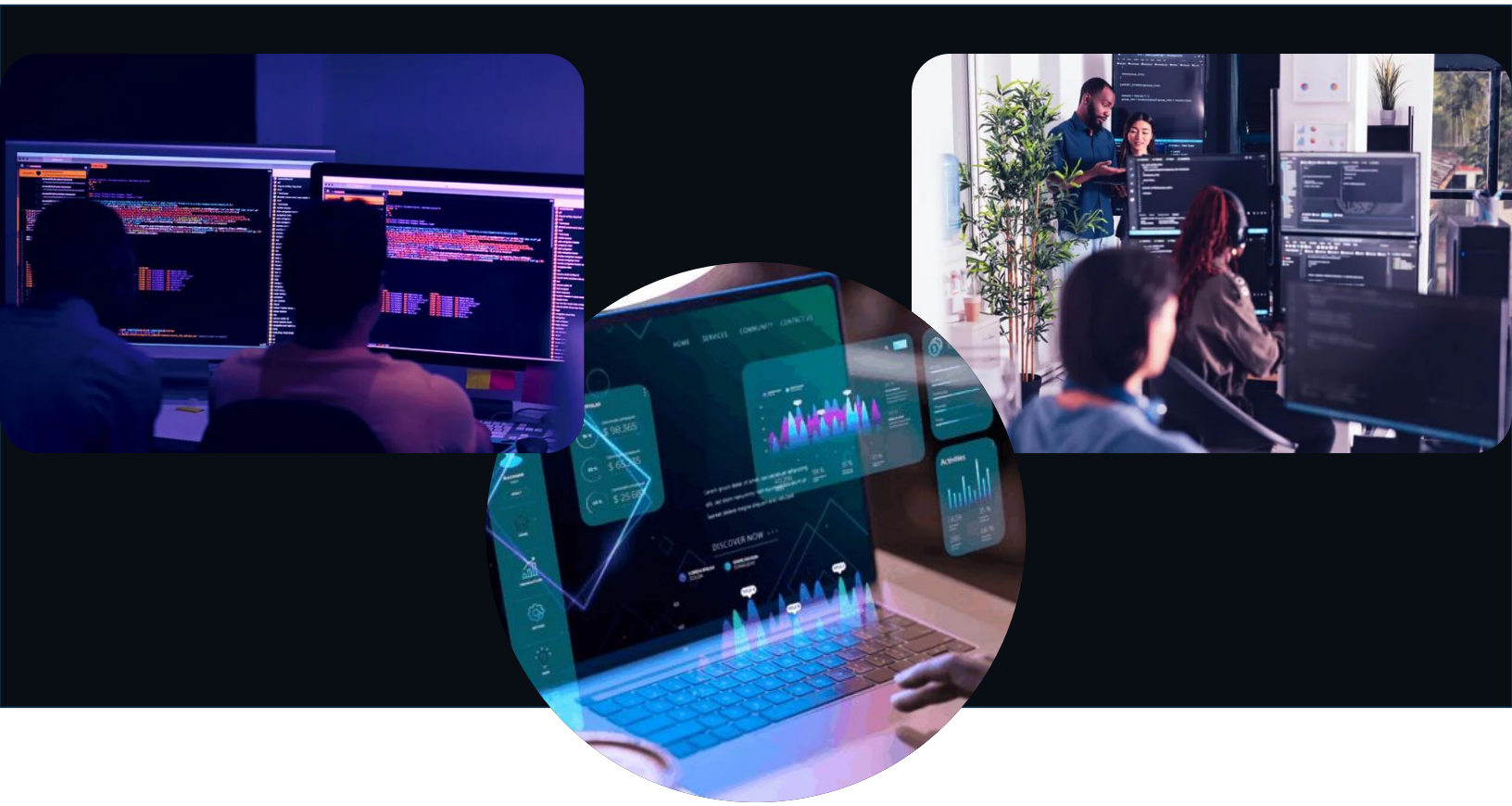
CodeRoad



# Data reliability & automation

A scalable data validation and reprocessing framework built for speed and accuracy.

Presented by The CodeRoad Team | January, 2026

# Executive Summary

A a leading solutions provider for performance marketing needs partnered with CodeRoad to resolve a critical data reliability issue that was slowing delivery, impacting reporting accuracy, and creating operational overhead for the Data Engineering team. The platform relied on third-party data providers that occasionally delivered late-arriving historical data, resulting in inconsistencies between raw and processed datasets and incorrect downstream client reports.

CodeRoad worked alongside the company's Data Engineering team to design and implement an automated data validation and reprocessing solution using PySpark within Databricks. The new system reconciles processed data against source data, detects discrepancies automatically, and triggers reprocessing of affected reporting pipelines—without requiring manual intervention. Since implementation, they have fully eliminated this class of production issues and established a standardized data quality mechanism that accelerates response time, improves reliability, and supports long-term development velocity.

## About the Project

As a leader in marketing technology and managed services built for local media organizations and agencies,  they offer a single-solution platform that brings together sales enablement, order management, fulfillment, and analytics into one consolidated system designed to simplify digital advertising operations and improve profitability.

The platform integrates with more than 100 leading martech solutions through robust API connections, enabling seamless workflows from proposal through campaign fulfillment and reporting. By unifying complex advertising operations into a single, scalable platform,  they helps media teams and agencies operate more efficiently while delivering better outcomes for advertisers.

Their reporting pipelines ingest large volumes of historical and real-time data from external providers and transform it into client-ready reports across multiple downstream systems. As they scaled their reporting footprint, maintaining data consistency across pipelines became mission-critical.

# Looking for Consistency to Scale

As a data-driven SaaS platform that delivers analytics, insights, and reporting for clients who depend on accurate, timely, and reliable data to make business decisions; their reporting pipelines ingest large volumes of historical and real-time data from external providers and transform it into client-ready reports across multiple downstream systems.

As they continued to scale their reporting footprint, maintaining data consistency across pipelines became mission-critical.

## Our Velocity Playbook in Action

To address the issue at its root, CodeRoad partnered with the company's Data Engineering team to embed data quality directly into the pipeline architecture. The focus was on automation, standardization, and resilience—ensuring that late-arriving data could be handled cleanly and consistently across all reporting workflows.

1. **Data Quality & Validation Framework**
   - Identify failure points caused by late-arriving historical data
   - Design validation logic to reconcile processed datasets against raw sources
   - Establish consistent data quality checks across reporting pipelines

2. **Automation Engineering**
   - Implement a PySpark-based validation and reconciliation solution
   - Deploy within the Databricks environment for scalable execution
   - Automate detection of discrepancies between source and destination datasets
   - Trigger automated reprocessing of affected reporting pipelines

3. **Standardization for Scale**
   - Apply the solution across all client reporting pipelines
   - Remove dependency on manual intervention by developers
   - Create a repeatable pattern for future data quality enforcement

# Shipping Fast, Building Right

## Reduced Operational Overhead

Engineers no longer need to intervene to resolve recurring data issues.

## Reliable Client Reporting

All downstream reports now reflect corrected and validated data consistently.

## Faster Response to Data Changes

Automated reprocessing eliminates delays caused by manual fixes.

## Reduced Operational Overhead

Engineers no longer need to intervene to resolve recurring data issues.

## Standardized Data Quality

One validation mechanism supports every reporting pipeline.

## Scalable System

The solution scales as data volume and client demand increase.

# A Velocity Roadmap to Scale Rapidly.

The new data validation and reprocessing framework established a reliable foundation for the company's reporting ecosystem. By embedding automated checks directly into the data pipeline, the system continuously compares processed datasets against raw source data, ensuring that inconsistencies caused by late-arriving records are identified as soon as they occur. This proactive approach prevents errors from propagating downstream and eliminates the need for reactive, manual investigations.

Once discrepancies are detected, the framework automatically triggers targeted reprocessing of the affected reporting pipelines. Built using PySpark within the Databricks environment, the solution scales efficiently across large datasets while maintaining performance and reliability. Corrections flow cleanly through all downstream processes, ensuring that client-facing reports always reflect the most accurate and up-to-date data without developer intervention.

As a result, they now operate with a standardized, resilient data quality mechanism that supports both current operations and future growth. The platform can confidently accommodate changes in data providers, increased data volume, and evolving reporting requirements. By making data validation and remediation an integral part of the pipeline, The company has strengthened trust in its analytics, reduced operational risk, and created a foundation that enables faster delivery of new data-driven capabilities.

## Technology in action

### PySpark

for scalable data validation and reconciliation

### Databricks

for distributed processing and pipeline orchestration

### Automated Reprocessing

for late-arriving data correction

### Standard Pipelines

for consistent reporting outcomes

# Talent + Acceleration + Confidence

To eliminate the manual overhead and reporting delays caused by late-arriving historical data, CodeRoad engineered an automated validation and reprocessing framework. We transformed a fragile data pipeline into a resilient, self-correcting system:

- **PySpark-Driven Automation:** Designed and implemented a robust reprocessing solution using *PySpark within Databricks* to automatically reconcile processed datasets against raw source data.
- **Automated Discrepancy Detection:** Built an observability layer that monitors data integrity in real-time, instantly triggering reprocessing for affected reporting pipelines without manual intervention.
- **Systemic Data Hardening:** Established a standardized quality mechanism that ensures downstream client reports remain the single source of truth; regardless of third-party data latency.
- **Operational Decoupling:** Freed the internal data engineering team from the firefighting cycle of manual data fixes, allowing them to focus on high-velocity feature development.

# Engineered Momentum

## CodeRoad's Business Impact by the numbers

**100%** data accuracy. We successfully deployed the automated framework with complete system availability and zero disruption to reporting cycles.

**ZERO** manual intervention. We automated the reconciliation of late-arriving data, removing a significant source of operational friction and developer burnout.

**100%** release confidence. We streamlined the data workflow, drastically reducing post-release production incidents and accelerating delivery tempo.

# How CodeRoad can help you accelerate

**CodeRoad delivers Velocity-as-a-Service (VaaS) by engineering self-healing data systems that eliminate manual reconciliation. We partner with you to build automated validation frameworks using PySpark and Databricks, ensuring late-arriving third-party data is processed with 100% accuracy. By replacing fragile manual fixes with a resilient, automated reprocessing engine, we restored full client trust and freed internal teams for roadmap innovation. CodeRoad's systems are engineered to turn complex data integrity challenges into a seamless, high-velocity reporting advantage.**

## Strategic Technology Solutions

- Data Pipeline Reliability & Automation
- Analytics & Reporting Platforms
- API & Integration Engineering
- Data Engineering & Processing
- AI-First Delivery
- Large-scale data transformations
- Data engineering & integrations

## The business impact of VaaS

Learn how velocity-as-a-service (VaaS) redefines outcome-based execution and engineers momentum for the future of technology. Get the latest on:

- Engineered Momentum
- Coordination Tax
- Outcome-Based Delivery
- Nearshore Transformation
- Roadmap Acceleration
- Predictable Scaling
- Predictable ROI

---

## Ready to accelerate your outcomes?

**Book Assessment Call**

1-954-866-3473
contactus@coderoad.com
CodeRoad.com

CodeRoad