Service Flyer
Big Data & Analytics

# Parallel Streaming Transformation Loader Service

Accelerate business insight

## Executive Summary

In a dynamic, fast-paced digital world, being able to make business and operational decisions in near real time is a huge competitive advantage. Near real-time decision making relies on having accurate, up-to date insight, and many organizations have turned to Big Data solutions for this kind of insight. Yet many struggle to make this a reality. The sheer volume of data, the large number of data sources, and the disparity in data types and formats have meant that for most organizations, 'near real-time insight' remains just beyond reach.

One of the main hurdles has been the inability to ingest and transform large amounts of data from multiple sources in real time. While most organizations employ data analysts and scientists, the reality is that they spend most of their time—up to 80%, according to some generally accepted estimates—in data preparation: collecting data sets and cleaning and organizing data.

Parallel Streaming Transformation Loader (PSTL) from Micro Focus® software Services is a Big Data solution that dramatically reduce both the time and latency involved in real time data collection, loading, and transformation.

## The Data Deluge Bottleneck

Organizations have historically focused on collecting and analyzing historical data using solutions like Business Intelligence (BI) and Enterprise Data Warehouse (EDW). But over the last decade, they have had to deal with an explosion of data volumes, a shift from structured to unstructured data, and an increasing need from the business for near real-time insight. Whether it is data from IoT devices, sensors, databases, web logs, data center logs, or customer sentiments on social media, organizations are drowning in data and have turned to Big Data solutions such as Micro Focus Vertica and Hadoop to help them turn all of this data into insight.

Engineering a robust data pipeline for data that continuously streams in, however, is far from trivial. Data is generated by numerous sources and stored in disconnected 'islands.' Having a scalable, robust, raw data pipeline for data ingestion is just the first step. You still need to transform it: process, cleanse, enrich, and format it, and ensure that it is compliant with whatever regulation your company is subject to. Only then can you feed it to your Big Data analytics engine.

This often requires large investments of time, money, and resources to engineer, and with limited budgets, it is left to the data analysts/scientists to manually perform these tasks rather than doing actual data science work such as building and refining algorithms or mining data for patterns. This not only wastes their skills on nonproductive work, but also creates bottlenecks that stand in the way of near real-time business insight.

## Parallel Streaming Transformation Loader

PSTL removes these bottlenecks. It is the result of multiple generations of streaming solutions we have built, coupled with our collective experience in open- and closed-source software with many customers over the years. We built it to solve problems we've faced in the real world with data pipelines that included all of the big V's (Volume, Velocity, Variety, and Veracity). PSTL is a highly scalable, fault-tolerant, extensible, self-service application framework that enables data analysts/scientists to write SQL over dynamic streaming data sources. It provides all of the necessary components for a complete end-to-end Big Data pipeline that would otherwise require custom coding and take many person-months of architecture, design, development, and testing.

### Features

- A Spark application with out-of-the-box integration from Kafka to Vertica and Hadoop; integration to other data systems via no-code configurations
- No-ETL, no-ELT, no-code required SQL streaming solution
- Single set of semantics for multiple sinks (Vertica, Kafka, Hive Tables, Opentsdb, or Spark Datasets)
- Out-of-the-box support for Confluent Kafka Sources
- Processes semi-structured JSON, Avro, Protobuf, Delimited, and CSV data into optimized data at rest
- Advanced job management of Spark Streaming Jobs
- A no-code approach for Change Data Capture, Slowly Changing Dimensions, Streaming Table Mappings, from external JDBC connectors
- A simple extensibility model for data validation and transformations

Contact us at:
**www.vertica.com**

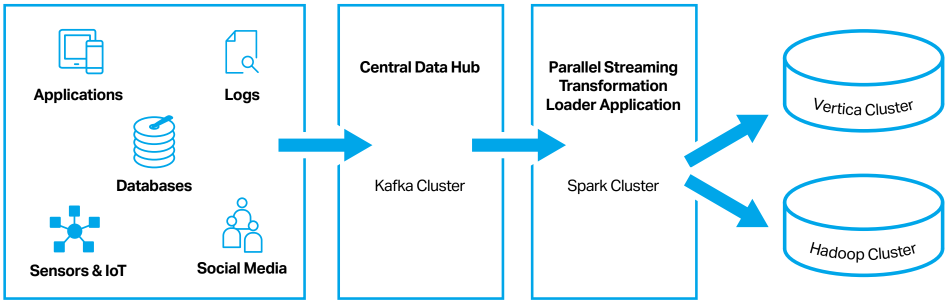

**Figure 1.** PSTL schematic architecture

## Service Overview

### Platform Deployment

This is the first stage of delivery, where we install and configure the PSTL application. We deploy onto your infrastructure either in your data center or in the cloud. After deployment, we connect PSTL to the in-scope data sources and Big Data analytics engines (Vertica or Hadoop) that you have identified prior to the commencement of delivery, and check that the data collection is functioning.

### Design and Development

This is the second stage of delivery, where we work with your Big Data team to define, prioritize, develop, configure, and test the SQL queries required to perform the data transformations you need. During this phase, we work side by side with your staff so that we can transfer knowledge and help them become self-sufficient.

### Solution Management

Once the PSTL is fully functioning, we transition ongoing management to our offshore solution management team. They continue to monitor and maintain the solution, freeing you to continue to connect more data sources and develop more data transformations..

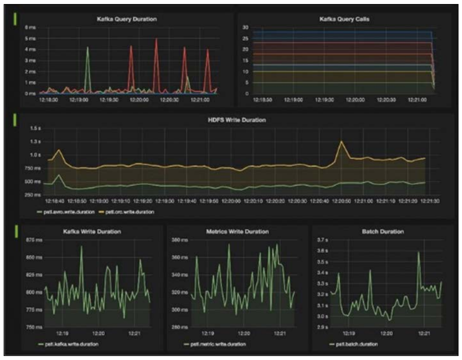

**Figure 2.** PSTL monitoring console

## Benefits

### Faster Time to Insight

Our experience tells us that building a complete end-to-end Big Data pipeline from scratch can take as long as one to two years. With PSTL, you can be up and running in as little as two months.

### Reduce Cost

The reality is that if you want low-latency analytics, your choice is to either build or buy a streaming data solution. Building your own end-to-end Big Data pipeline is expensive. We have seen projects exceeding $2M. PSTL (depending on scope) represents a fraction of this cost.

### Increase Productivity

PSTL frees your data analysts/scientists to actually do what they are meant to be doing: building and refining algorithms or mining data for patterns, thus making them more productive.

## The Micro Focus Service Difference

Micro Focus Services provides unmatched capabilities with a comprehensive set of Big Data consulting services. We offer:

- Fast time-to-value: We help you rapidly realize business value by leveraging our deep expertise in Big Data solutions and our structured, focused implementation approach
- Proven Big Data solution implementation track record of helping large, complex, global organizations realize value from their Big Data investments
- Rich intellectual property and unparalleled reach into product engineering

Micro Focus services brings together consulting expertise and the industry-leading software to help you perform better.