Paper 4140-2016

Listening for the Right Signals – Using Event Stream Processing for Enterprise Data

Tho Nguyen, Teradata Corporation
Fiona McNeill, SAS Institute Inc.

ABSTRACT

With the big data throughputs generated by event streams, organizations can opportunistically respond with low-latency effectiveness. Having the ability to permeate identified patterns of interest throughout the enterprise requires deep integration between event stream processing and foundational enterprise data management applications. This paper describes the innovative ability to consolidate real-time data ingestion with controlled and disciplined universal data access - from SAS® and Teradata™.

INTRODUCTION

In today's big data world, there are great challenges and many opportunities. Organizations need to have the ability to make the right decisions with precision, accuracy and speed - in order to enhance the competitive advantage in a global economy of constant change. As a result, the state of the business impacted by dynamic conditions requires continuous monitoring and evaluation by separating the right signals from the noise. These events of interest are only apparent when they are understood and heard by the dependent parts of the organization. This requires event processing that follows through the organization in contextually and relevant data-driven actions. The ability to ingest data and process streams of events effectively identifies patterns and correlations of importance, focusing organizational activity to react and even proactively drive the results they seek and respond to in real time. Instead of collecting, analyzing and storing the data in the traditional method, data can now be analyzed constantly, as it is occurs, empowering organizations to adjust situational intelligence as new events transpire.

With the emergence of the Internet of Things (IoT), ingesting streams of data and analyzing events in real-time become even more critical. The interconnectivity of IoT from web and mobile applications provides organizations with even richer contextual data and more profound volumes to decipher in order to harness insights. These insights can uncover greater business value to better understand customer habits and behavior, enhance operational efficiencies and expand product and services offerings. Capturing all of the internal and external data streams is the first step to enable listening for the important signals that customers are emitting, based on their event activity. When you hear what they want from the data they generate, the right and data-driven actions can happen, and now more rapidly than ever, positively impacting bottom line profitability.

Of course, one obvious challenge is deploying a reliable, scalable and persistent streaming environment. This environment needs to provide the necessary self-service capabilities for data administrators, application developers and data scientists alike, so they can rapidly configure new and different combinations of data streams and continuous queries for insights. Some organizations have explored and implemented open source technologies for real time streaming. However, many have already come to realize the inherent challenges of scaling across multiple event streams, building a dynamic and yet stable environment that is flexible for adaptation to business dynamics and one that is supportive of enterprise goals, ongoing needs and timelines.

As such, innovative organizations are moving beyond constructing enterprise environments that require extensive manual coding from the ground up, to ones that take advantage of pre-built capabilities that are readily available and integrated with existing organizational assets to drive automated, intelligent streaming insights. Together, SAS® and Teradata provide an integrated pre-built environment for exploiting enterprise data that listens for the right streaming signals – improving data-driven decisions for the entire organization.

SAS® EVENT STREAM PROCESSING

Event stream processing (ESP), is designed to connect and analyze real-time event-driven information. ESP processes event streams with the mission of identifying meaningful patterns and correlations as they occur. Doing more than pipeline transport, the ability to enrich data by correlating events, identifying naturally occurring clusters of events, event hierarchies, event probabilities and other aspects such as contextual meaning, membership and timing – event stream processing , delivers deep insights to real-time activity for a new, fast data infrastructure.

SAS® Event Stream Processing is a comprehensive technology that delivers fast data insights based on a publish and subscribe framework that ingests event streams, executes continuous queries using a suite of pre-built and interchangeable window types and operators and delivers insights and instructions for automated actions to dependent systems, applications and big data warehouses. In the traditional data infrastructure approach, data is amassed, stored and then analyzed. Instead of storing data and then running queries against this data at rest, SAS Event Stream Processing stores queries to continuously enrich streaming data while it is in motion. As such, event streams are examined as they are received, in real-time, and can incrementally update with new intelligence as new events happen. Focusing on enriching data while events are still in motion demands a highly scaled and optimized process to address the hundreds of thousands of events per second common to event streams. SAS Event Stream Processing has the ability to enrich and filter event, differentiating and analyzing text and structured streaming data with embeddable analytics that instantly translate to real-time insights for event-driven actions. SAS® Event Stream Processing Studio is the visual data flow interface, simplifying the construction of event stream continuous queries, and saving time and efficiency of application developers, data scientists and IT architects.

Given event stream data is never clean data, even when generated by machine sensors, SAS Event Stream Processing includes pre-built data quality routines to aggregates, normalize, standardize, extract and correct, enrich and filter event data before it is stored in a data platform. By eliminating data quality issues upfront, countless resources and computing hours are saved, big data stores avoid unnecessary pollution and IT and data scientists are more productive. Not only does productivity improve with this traditional data cleansing now happening on data in motion, it takes care of the necessary data preparation needed for successful in-stream analytics. Furthermore, by filtering the data to what is cleansed and relevant, unnecessary storage of irrelevant event noise helps focus all other activity on what is relevant.

The ability to listen for events and ingest, consolidate streams of data is critical to real-time actions, ones that impact transitory event opportunities and avoid impeding threats. Low latency response for real-time actions, with millisecond and sub-millisecond response times, not only demands high performance processing but also requires tightly integrated data communication access to event stream sources and delivery to streaming insight consumers. SAS Event Stream Processing comes with a suite of prebuilt connectors and adapters (such as Teradata) to consume structured and semi-structured data streams. Connectors and adapters operate through the publish/subscribe layer (as illustrated in Figure 1), and can also be custom built as APIs in C, Java, and Python. Supporting authentication and encryption, they publish data from any source into the continuous query and publish data out to any subscribed source. In addition, they include communication protocols across different streams for enterprise level use of streaming insights from a range of messaging bus and data transport protocols. Creating a robust ecosystem with both pre-built, editable and open APIs to ingest consolidate and manage multiple event streams mitigates the risk of limiting insights and relieves the need to write code by specialized programmers for ongoing support and maintenance.

Continuous queries are at the heart of driving new, enriched insights from streaming data (depicted in Figure 1). SAS Event Stream Processing enables a comprehensive suite of advanced analytics to event streams, like forecasting, data mining, and machine learning algorithms for governed, streaming decisions (McNeill et al., 2016). Data governance is key to addressing not only the dynamic nature of streaming data, it also ensures fully documented and readily understood event stream processing application – empowering agility to make and understand the impact of changes necessitated by the dynamic nature of business.

Customizable alerts, notifications and updates directly issued from SAS Event Stream Processing provide precise and accurate situational awareness so that actions are relevant and informed as to what's happening and what's likely to happen. These actions are fueled by continuous, accurate, and secured event pattern detection from SAS Event Stream Processing patented 1+N-Way failover, guaranteed delivery (without persistence), full access to event stream model metadata, live stream queries, dynamic streaming model updates, along with deep analytic capabilities.

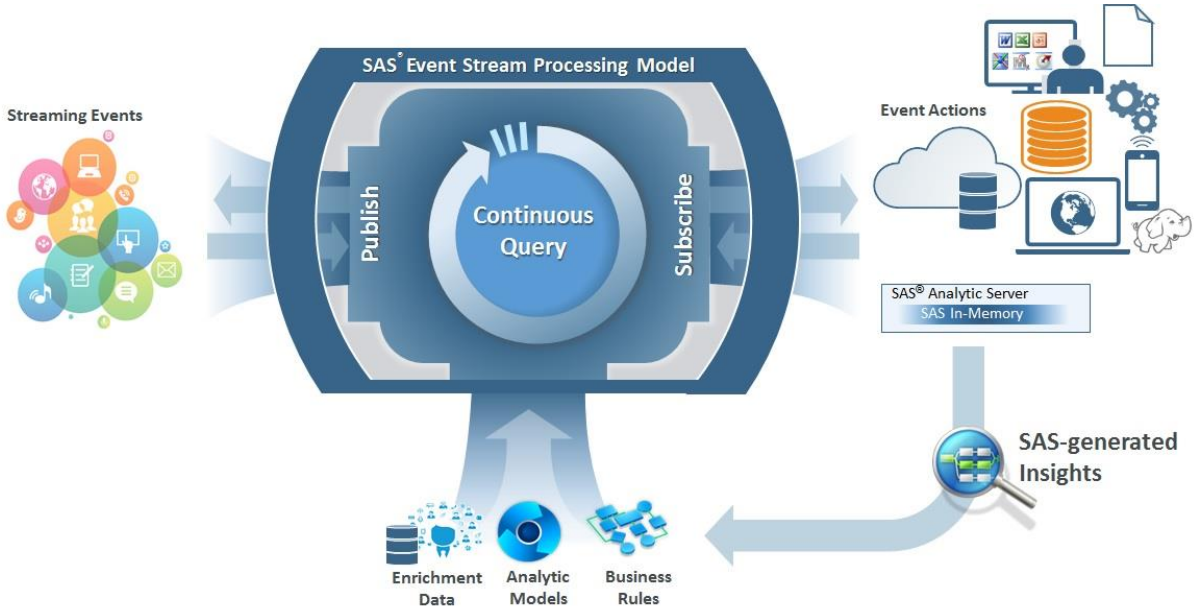

Figure 1: SAS® Event Stream Processing conceptual architecture

SAS Event Stream Processing captures true business value otherwise lost through information lag. Businesses are able to analyze events as they happen and seize new opportunities through producing data-driven actionable intelligence with no latencies. It enables new analysis and processing of models to be developed and modified quickly to meet the changing needs of the business and the competitive landscape.

SAS EVENT STREAM PROCESSING WITH TERADATA

Traditionally data has been stored in a database. Once the data is captured, it goes through a rigorous ETL (Extraction, Transformation, Load) process to integrate the data stored in the data warehouse. The ETL process can take days or even weeks to complete, depending on the size of the data. Data analysts, business analysts and automated reports are gleaned from queries that run against the trusted and vetted data warehouse. However, this traditional processing paradigm isn't well suited to driving insights from events that are happening in near real-time. Figure 2 illustrates the comparison of traditional relationship database (RDBMS) processing with that of event stream processing.

By integrating SAS Event Stream Processing with Teradata, organizations now have a new, modernized approach to percolate current events and streams of data to existing reporting and insight-driven applications. Enabled by the SAS Event Stream Processing connector that leverages the Teradata Parallel Transporter (TPT) API supports subscribe operations against the Teradata server.

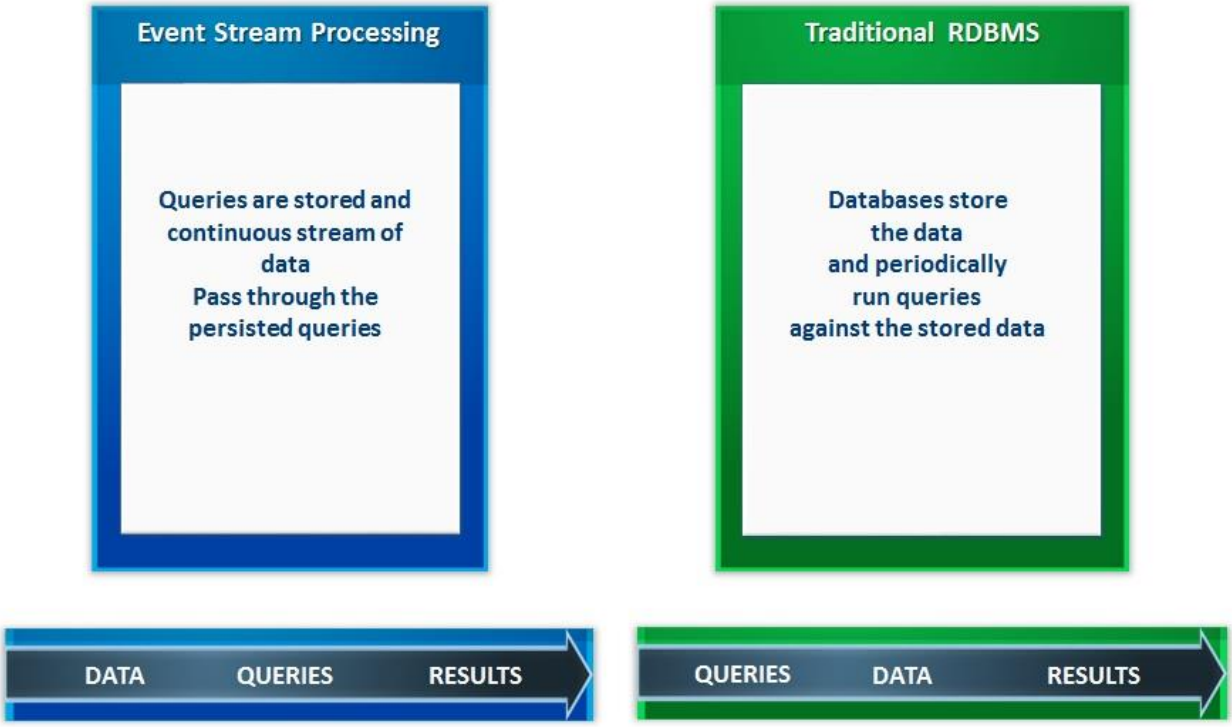

Figure 2: ESP and Database processing

From SAS Event Stream Processing, the Teradata server subscribes to the following operations:

- Stream – operating similarly to a standard event stream processing database subscriber, but with improved derived from TPT. Supports insert, update, and deletion of events. As events are received from the subscribed window, it writes them to the pre-defined target table. If (the required) tdatainsertonly configuration parameter is set to “false”, serialization is automatically enabled in TPT to maintain correct ordering of row data over multiple sessions.
- Update - Supports insert/update/delete events, but writes them to the target table in batch mode. The batch period is a required configuration parameter. At the cost of higher latency, this operator provides better throughput with longer batch periods (for example minutes instead of seconds).
- Load - Supports insert events. Requires an empty target table. Provides the most optimized throughput. Staggers data through a pair of intermediate staging tables. These table names and connectivity parameters are additional configuration parameter specification requirements. Writing from a staging table to the ultimate target table uses the generic ODBC driver used by the database connector. Thus, the associated connect string configuration and odbc.ini file specification is required. The staging tables are automatically created by the connector. If the staging tables and related error and log tables already exist when the connector starts, it automatically drops them at start-up.

Having integrated connectors is certainly a good start. New innovations expand upon this to facilitate even faster processing and reduced latency. The new Teradata Listener™ is an integrated offering that delivers a unified solution to handle the endless torrent of digital information streams. With the constant flood of digital information exponentially growing by all estimates, the complexities to integrate streaming insights across the enterprise will correspondingly become more important and complex. Integrating the Teradata Listener with SAS Event Stream Processing provides a new frontier for analyzing all big data in a massively parallel processing environment delivery new, timely and current fact-based insights to all in the enterprise.

TERADATA LISTENER™ AND SAS® EVENT STREAM PROCESSING

Ingesting streams of data is the key design element of the Teradata Listener. As an intelligent, self-service software solution that ingests and distributes exceedingly fast moving data streams throughout the enterprise analytical ecosystem. Listener™ collects data from multiple, high volume, real time streams from sources such as social media feeds, web clickstreams, mobile events and IoT (server logs, sensors and telematics). As mentioned, as a subscribed source, Listener can also ingest streaming analytic insights defined in SAS Event Stream Processing.

The key value of Listener is to allow developers and data administrators to build real time processing capabilities. It handles large volumes logs and event data streams, and reliably handles mission critical data streams ensuring data delivering without loss. The Teradata Listener offers a self-service capability to ingest streams of data without coding. And with no manual coding, it accelerates time to deeper insights as a streamlined and traceable process. It simplifies the IT process, maintenance and cost of custom-built systems. It can act as a centralizing system that can scale to the complete organization, operate with hundreds of applications built by silo teams, all of which can be plugged into the same, consistent system.

STREAMING SIMPLIFIED WITH SELF SERVICE

Teradata Listener streamlines the data ingestion process through a self-service dashboard, which can be accessed by multiple users (developers, administrators, and data scientists) throughout the enterprise. The intuitive Listener dashboard makes configuration of data sources and targets an easy task, eliminating the need for programming. Technical users can easily add, remove, or edit sources and targets to create streaming data pipelines. There is no need to request for access or change or add IT tickets. And there is no waiting for a programming team to develop and test another interface to a home grown streaming capture module.

The Teradata Listener ingestion services are invoked from RESTful interfaces through the very popular http transport protocol— a universally accepted protocol for modern-day applications. Any developer can easily invoke the Listener's ingestion services to send continuous data streams to a data warehouse, analytical platform, or Hadoop or any other big data platform.

Additionally, APIs (such as developed with SAS Event Stream Processing) provide more flexibility to developers to access the data flowing through Listener. And in the case of connection with SAS Event Stream Processing, the Teradata Listener is receiving streams that have been vetted, cleansed, filtered and enriched, improving the content of the streaming pipeline sourced by Listener, as per Figure 3.

Output from Teradata Listener is used to inform existing reporting work streams, updating custom dashboards and integrating other processing engines for additional transformations. Moreover, (and as depicted in Figure 3), the Listener output can stream back into SAS applications, other data repositories (aka. data at rest) and reporting systems and even back into SAS Event Stream Processing.

INGEST CONTINOUS STREAMS

Sources of event streaming data proliferate whether it is from web events, email, sensors, social media, machine data, IoT, SAS Event Streaming output, and others as shown in Figure 3. Teradata Listener brings together the big data ingestion process by collecting multiple, high volume data streams continuously from a variety of sources, and storing them into one or more of the data stores that comprise the enterprise data ecosystem. Listener has capabilities to write to a variety of target stores – in integrated data warehouse, analytical platform, or Hadoop. Listener can also write results back into SAS Event Stream Processing. Now enriched with more data, new insights, and even directions from end-users – Listener output can be analyzed further, and actions as streaming decisions to devices, and objects at aggregation points in-stream, and even to the edges of the IoT.

Listener is agnostic to data variety, working effectively with both structured and semi-structured data. A Teradata Listener cluster of servers scales horizontally to meet the growing demands of multiple data streams in the enterprise.

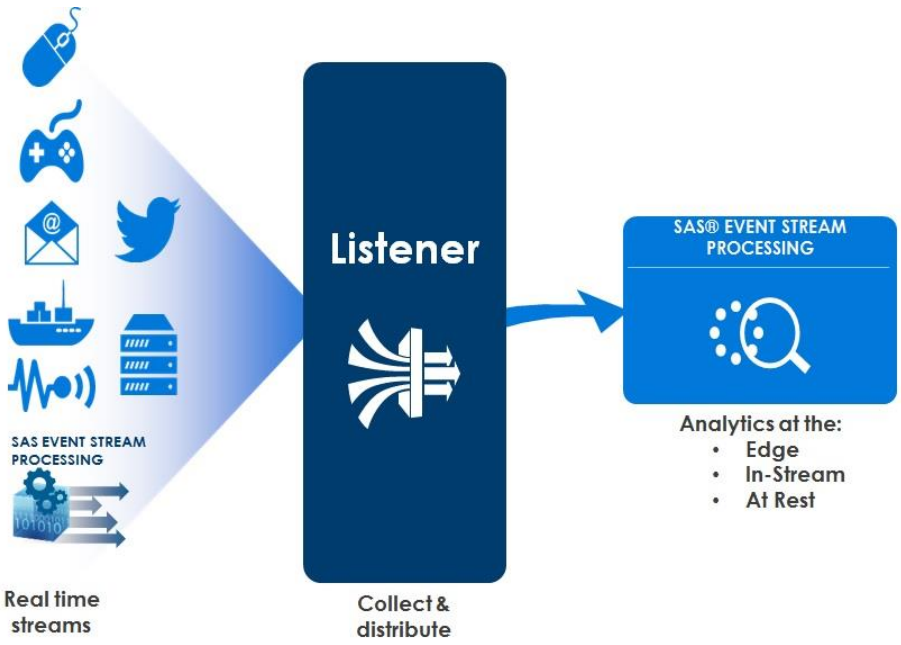

Figure 3: Teradata Listener and SAS Event Stream Processing

DATA-DRIVEN INTELLIGENCE

Listener continuously monitors incoming data streams automatically, gathering critical information exploited from the graphical user interface and dashboards to serve a deep understanding of the data. Various metrics on this dashboard help end-users understand current activity both in and out of Listener's ingest and distribution processes. Users can intuitively discover when a stream has stopped or when a target halts accepting the data output.

Teradata Listener's micro services architecture enables decoupling of ingestion process of incoming data streams with the outgoing distribution processes. Listener buffers the distribution output intelligently when the target systems are full, activating the distribution later when target system allows - all without any manual intervention.

CONCLUSION

As business conditions evolve, the need to continuously monitor and measure streaming events of interests is imperative. Machine-driven with human-guided curation of event streams, enriched with analytic intelligence and focused to relevant events that are heard throughout the enterprise is unique value that SAS and Teradata provide. Instead of the traditional “stream, score and store” process, data can now be analyzed immediately as it is ingested or received and adjusting situational intelligence as new events happen using Teradata Listener and SAS Event Stream Processing.

Applicable across a wide range of industries, the ability to process streaming data once, and persist to stores and applications and other streams, across the enterprise is a foundational benefit for analytical workloads. Efficient and well-managed processing is paramount to low latency, real-time responsiveness – and when time matters, the ability to complete the full analytical lifecycle to drive better decisions becomes critical. Whether that be the need to re-optimize mobile dispatch units based on live location streams, preventing hazardous events by prioritizing maintenance needs based on current weather predictions, or recognizing the need for new streams of data to improve projected operational effectiveness, listening for the right signals provides focus.

With SAS and Teradata, the combined and integrated technology offers a scalable and reliable solution to ingest data and process streams of events, leveraging the embeddable streaming analytics of SAS so that organizations can pro-actively respond to even the most complex issues.

REFERENCES

McNeill, F., D. Duling, S. Sparano. 2016. “Paper SAS6367 Streaming Decisions: How SAS Puts Streaming Data to Work” *Proceedings of SAS Global Forum 2016*, Los Vegas, NV

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Tho Nguyen
Teradata
tho.nguyen@teradata.com

Fiona McNeill
SAS
fiona.mcneill@sas.com