TDWI RESEARCH

TDWI CHECKLIST REPORT

Using Streaming Analytics

for Continuous Operational Intelligence

By Philip Russom

Sponsored by:

VITRIA®

tdwi.org

tdwi

MAY 2014

TDWI CHECKLIST REPORT

Using Streaming Analytics

for Continuous Operational Intelligence

By Philip Russom

TABLE OF CONTENTS

555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD

According to TDWI's 2013 survey on managing big data, roughly half of user organizations surveyed are already managing and leveraging streaming data that's generated frequently or continuously by sensors, machines, geospatial devices, and Web servers.[1] However, most of these users are today merely capturing and storing streaming data for offline study, whereas they need to mature by using real-time practices and technologies. This would enable them to analyze streaming data as it arrives, then take immediate action for the highest business value.

For example, consider some of the use cases that the real-time, continuous analysis of streaming data is making a reality today:

- Monitor and maintain the availability, performance, and capacity of interconnected infrastructures such as utility grids, computer networks, and manufacturing facilities
- Understand customer behavior as seen across multiple channels so you can improve the customer experience as it's happening
- Identify compliance and security breaches, then halt and correct them immediately
- Spot and stop fraudulent activity even as fraud is being perpetrated
- Evaluate sales performance in real time and meet quotas through instant incentives such as discounts, bundles, free shipping, and easy payment terms

Compelling use cases such as these typically result from a "perfect storm" of desirable data types, software functions, and fast-paced business processes:

Streaming data. The swelling swarm of sensors worldwide (plus the extended "Internet of things") produces large volumes of streaming data that can be leveraged for business advantage. For example, robots have been in use for years in manufacturing; now they have additional sensors that can perform quality assurance, not just assembly. For decades, mechanical gauges have been common in many industries (chemicals, utilities); now the gauges are replaced by digital sensors and "smart meters" to provide real-time monitoring and analysis. GPS and RFID signals now emanate from mobile devices and assets ranging from smart phones to trucks to shipping pallets so all can be tracked in real time and controlled precisely.

Streaming analytics. The growing consensus is that analytics is the most direct path to business value drawn from new forms of big data, which includes streaming data. Existing analytic techniques—based on mining, statistics, predictive algorithms, queries, scoring, clustering, and so on—apply well to machine data once it's captured and stored. Luckily, newer vendor tools are reengineering these and creating new analytic methods so they can operate on data that streams continuously as well as stored data.

Continuous analytics. Most analytic operations are scheduled to run on a 24-hour or longer cycle. Getting the most out of streaming data, however, requires analytics that execute or update every few seconds or milliseconds to process each event, message, record, transaction, or log entry as it arrives in case the new data signals a business event that requires immediate attention. In other words, continuous analytics go hand-in-hand with streaming data. Imagine the results of a query incrementally updated with each new event without needing to rerun the query against all pertinent data. Likewise, continuous analytics may rescore an analytic model, recalculate a statistic, remap a cluster, and so on but as efficient, incremental updates, not execution from scratch.

Complex event processing (CEP). Event processing technology has been applied to streaming data for decades, and a recent TDWI Best Practices Survey shows that more than 20 percent of organizations surveyed are doing event processing today in their DW/BI solutions.[2] However, traditional event processing tends to be very simple, monitoring one stream of data at a time. The newer practice of CEP can monitor multiple streams at once while correlating across multiple streams, correlating streaming data with data of other vintages, and continuously analyzing the results.

Operational intelligence (OI). OI is a new form of business analytics that delivers visibility and insight into business operations and similar processes, as they are happening. This new class of enterprise software includes all the capabilities discussed above, but in a unified tool that empowers users to explore data streams, understand business processes (as seen via data), model processes, write rules for event-driven alerts and responses, and create full-blown business monitoring and surveillance applications. When these applications run and respond continuously in real time, you have *continuous operational intelligence*.

This TDWI Checklist Report examines the user best practices and vendor tool functions for analyzing streaming data, with a focus on those that enable new applications in continuous operational intelligence.

[1] See the discussion around Figure 13 in the TDWI Best Practices Report, *Managing Big Data*, available online at tdwi.org/bpreports.

[2] See the discussion around Figure 1 in the TDWI Best Practices Report, *Next Generation Data Integration*, available online at tdwi.org/bpreports.

NUMBER ONE

FOR GREATEST INSIGHTS, MAKE CORRELATIONS ACROSS MULTIPLE STREAMS AND DIVERSE DATA FORMATS

Data diversity is exploding. The current trend toward big data has us thinking about burgeoning data volumes. However, both big data and traditional enterprise data sources are also exploding in terms of the diversity of data models, schema, formats, and other structures. In addition, more streaming data sources are becoming available as organizations deploy sensors, tap the logs of existing servers, integrate with partnering companies, and subscribe to third-party data feeds. In a growing number of enterprises, these trends result in an eclectic mix of structured data (mostly relational data), unstructured data (human language text), semi-structured data (RFID, XML, JSON), and streaming data (from machines, sensors, Web applications, RSS feeds, and social media). This is a challenge for continuous operational intelligence because it must make complex correlations across multiple real-time streams and latent traditional sources, despite the tremendous diversity of data and the stringent requirement to operate in real time (or close to it).

Complex correlations depend on data of diverse types, structures, and sources. New insights are created when vastly different types of data are brought together. For example, consider a data stream emitted by a quality assurance sensor on an assembly robot in a manufacturing facility. Sensor data indicates that a specific part has been failing at a rate that's slightly higher than usual. Correlation with product quality metrics in a data warehouse reveals that the current supplier has a history of bad lots. A second correlation suggests that ejecting bad parts and replacing them with others might slow production, such that the daily service-level agreement (SLA) for production yield will be missed. Each data point in isolation is not really a problem, but correlation across all three conditions results in a risk calculation that triggers immediate action from managers.

Note that the three data points correlated in the above example come from three different systems, each with its own data schema, interfaces, and latency levels. This is typical of OI solutions. Therefore, a platform for OI must support a long list of interface types (to reach both old and new applications, databases, and other data sources), and schema (whether relational, hierarchical, proprietary, standard, or flexible).

The practices of OI assume that data is also diverse in terms of latency. For example, latent data (from a DW or similar database) provides a historic context for real-time events. This applies to many use cases, from tailoring a purchase recommendation (based on prior purchases) to detecting potential fraud (as when an insured motorist is involved in multiple similar losses over time).

Flexible schemas are the new norm. Many of the newer data types and sources have schema that are somewhat unpredictable. For example, parent-child relationships can vary in the hierarchies of XML documents, and just about any data structure may appear in a message from an RSS feed. Furthermore, data schema will naturally evolve, as best practices evolve concerning the data output from sensors, machinery, mobile devices, and so on. Similarly, data feeds from partnering companies and third-party data providers are infamous for unannounced schema changes.

Hence, for many of the new data types and sources seen in big data and streaming data, the new norm involves "flexible schemas," meaning data structures that are unpredictable or regularly evolving. Organizations wishing to practice continuous OI with such data sources should seek OI solutions that can adapt to schema changes gracefully as well as handle extremes of richly structured and loosely structured data. Likewise, organizations should seek OI solutions that handle big data and streaming formats (e.g., XML, JSON) in their native form without the need to transform and normalize the data into a standard schema. Because the data sources discussed here rarely have metadata, a tool that can deduce metadata more or less automatically is highly desirable.

NUMBER TWO

EXPLORE DATA STREAMS TO DISCOVER ACTIVITIES AND PROCESSES, NOT JUST DATA

Most analytic methodologies begin an analytic project with data exploration, and operational intelligence is no exception. With OI, exploration usually focuses on studying the streaming data delivered by one or more streams; and the study involves both data in motion (arriving via a stream) and data at rest (stored in a file or database).

Stream exploration and discovery is important to streaming analytics and OI in general:

Operational intelligence is concerned with pattern and process discovery, not just data or stream discovery. Seeing event patterns and business processes unfolding in one or more streams provides evidence for understanding online customer behavior and improving business processes. In fact, many OI users explore streams to discover what actually happens in a business process, which is knowledge at a level of accuracy they can't get elsewhere.

Exploring streams and related data can be inspirational. This is how OI developers get a sense of how processes work so they can build a dataset, model, and rules that yield a correlation advantageous to the organization. This is also how a developer discovers new sources and understands their applicability to specific use cases.

Exploration and discovery are replacing older practices in requirements gathering. Traditional requirements gathering takes months, generates documentation that rarely applies directly to a solution, and too often promises data that doesn't exist. Users need a more agile and better aligned method. Data exploration and discovery set accurate expectations because they work with available data, and the work resulting from discovery can be folded into prototypes directly and immediately. This is true across all kinds of advanced analytics, including OI and streaming analytics.

Continuous, automated discovery is an important goal for OI and stream analytics. With OI, it's best to use a tool that supports the automatic discovery of patterns and relationships among streams and other data. This applies to early stages, when you're planning a solution. However, it also applies later when you deploy the solution. A mature OI platform will continuously process streaming data to discover business activity patterns, exceptions, and bottlenecks; it then proactively responds based on discovered insights. Examples of activity patterns that can be quickly uncovered, analyzed, and acted upon include those related to financial transactions, orders, shipments, packages, vehicles, online customers, passengers, and people of interest.

NUMBER THREE

USE ADVANCED ANALYTICS TECHNIQUES TO PROVIDE DEEPER INSIGHT

Streaming analytics incorporates several analytic methods, all considered advanced because of the great diversity of data types and structures found in a stream, as well as the many real-world use cases that analytics may serve. Here are some of the advanced analytics techniques typically required of OI and streaming analytics:

High-performance data ingestion. Operational intelligence is designed to continuously ingest massive volumes of both streaming and stored data. The highly efficient complex event processing (CEP) engine within it continuously queries, filters, correlates, integrates, enriches, and analyzes this data to discover exceptions, patterns, and trends that are presented through live dashboards. By leveraging in-memory processing, the results are delivered with ultra-low latency.

Continuous data-stream mining. This form of mining uses filters to extract relevant information from the data elements that fly by in a stream. Each filter captures a specific type of information. It's common to have several filters per stream (multiplied by several streams) and to correlate filtered data with other filters and other data sources. Filters collect data elements into so-called windows or micro batches, which then become the basis of averages, other time-variant calculations, dimensions, and time-series slices.

Continuous predictive analytics. The goal of many streaming analytics applications is to predict the class or value of new instances in the data stream, given some knowledge about the class membership or values of previous instances in the data stream. A mature OI platform will support multiple approaches, including Bayesian techniques, the prediction of near-term opportunities and threats, and recommendations for the next best action, whether that is to guide a customer to a purchase, avoid a process bottleneck, or mitigate a threat. Streaming analytics enables predictions to be continuously rescored and re-evaluated to reflect the most recent data updates and changes in the business situation.

Machine learning. Continuous machine learning techniques can be used to learn predictive tasks in an automated way. Likewise, it can be instrumental in coping with online learning, as well as with structural changes. For example, when used with regression analysis, continuous machine learning techniques can discover sudden changes in underlying model parameters immediately, such as a sudden shift in customer demand or a new baseline for system loading. Such rapid detection can enable systems to dynamically adapt to change.

NUMBER FOUR

ANALYZE DATA AS IT ARRIVES AND RESPOND IMMEDIATELY AND INTELLIGENTLY

Business benefit via timely responses. Streaming analytics is frequently applied to business use cases that demand a timely response, such as real-time, one-to-one marketing; online customer experience management; and fraud detection. To maximize the business benefit, responses to opportunities and threats need to happen in an appropriate time frame; in the use cases cited, this involves a system response in real time or near real time. The term real time has many definitions and expectations, but most measure the time from event reception to system response in a few milliseconds to a few minutes.

Continuous analysis for low latency. Traditional analytic tools and methodologies are inherently latent (and, therefore, *not* real time) because they depend on the aggregation and persistence of data. Solutions for OI operate with little or no latency to deliver insights in minutes and automated responses in milliseconds. Operational intelligence is accelerated into true real time by high-performance complex event processing. In addition, OI's method of continuous analysis (which yields incremental updates of analytic results) streamlines the analytic method naturally for improving analytic speed and results delivery.

"Intelligent" process management. When an OI platform has process management capabilities integrated into it, developers can define software processes for automated responses (based on business rules and application logic), which are then automatically executed at run time (based on discovered insights). For example, an automated response from an OI solution may trigger a fraud investigation, a repair process, or a personalized marketing offer.

Moreover, response processes can continually monitor analytic results and adapt the behavior of the response as the situation evolves. The ability to dynamically adapt processes is essential for seizing one-to-one marketing opportunities and mitigating security incidents with rapidly escalating threat levels and reach.

Unified platform for seamless analytics and action. Achieving the benefits discussed here depends on using a unified OI platform, where continuous analytics interact seamlessly and in real time with automated responses based on intelligent process management.

NUMBER FIVE

TAKE A MODEL-DRIVEN APPROACH TO DEVELOPING EVENT-DRIVEN STREAMING ANALYTICS

Development of OI should be model driven. An OI solution of any complexity will involve multiple model types, including process models, data models, data integration models, dashboard models, rules, service definitions, and models for one or more event processing networks (EPNs). Multiple model types (and multiple instances of each type) come together to constitute an OI solution.

Event processing network (EPN). Support for EPNs is a kind of "secret sauce" for OI and streaming analytics. An EPN models complex event analytics and patterns as multi-stage flows, allowing event processing to be decomposed into simpler event processing steps. The resulting analytics and complex events from one EPN can feed other EPNs, thereby forming a network of EPNs. The network enables critical and innovative functionality, such as correlations across multiple streams and other data structures, parallel processing, incremental analytic updates, stream discovery, and continuous real-time operation.

Hence, EPN modeling coupled with streaming analytics is fundamental to developing event-driven solutions for OI. This is true for industries as diverse as telecommunications, financial services, retail, utilities, health care, manufacturing, oil and gas, and transportation. Event-driven OI solutions use stream processing technologies and analytics to identify meaningful patterns, relationships, and trends involving seemingly unrelated events and then trigger immediate response processes.

Model-driven rapid development. Taking a model-driven approach to OI development requires a unified modeling tool that supports the diverse types of models applied in OI solutions, but in a graphical and collaborative environment. This way, multiple developers can create multiple models—for streams, CEP queries, policies, processes, process networks, dashboards, and so on—as well as share these collaboratively with other developers and offer stewards and other business users visibility into development projects.

Model-driven life cycle management. Given that solutions for OI and streaming analytics consist of multiple models, the models and development artifacts should be managed in a common repository for easy creation, sharing, audits, updates, execution, and administration through all life cycle stages. The repository also allows multiple OI solutions to share resources such as common feeds, databases, and services.

NUMBER SIX

EMPOWER YOUR BUSINESS USERS WITH SELF-SERVICE ANALYTICS

For better outcomes, involve business people in streaming analytics. If you select an OI platform that has functionality appropriate to mildly technical business people, then a number of benefits can follow. Involving business people "hands-on" yields better IT/business alignment. Business people can more accurately perform data governance, data stewardship, and other forms of governance and compliance. The involvement of business people tends to increase the likelihood of success with sponsorship, funding, and the perception of ROI. The business people involved should be domain experts so they can provide valuable knowledge transfer.

Empower some business users via self-service OI tool functions. For example, many data analysts and other power users want to define and monitor key analytics for streaming data. Empowering business users enables the pervasive use of analytics, encourages experimentation by the domain experts, democratizes streaming analytics, and allows for the "crowdsourcing" of analytical ideas by all stakeholders. Most important, it removes bottlenecks that can cripple the widespread adoption of streaming analytics—such as a lack of available IT resources and the shortage of technicians skilled in programmatic analytics tools.

Give business people visibility into streams and processes via dashboards. Dashboards provide users with a picture of current performance, and dashboards visually highlight anomalies and exceptions. Users can drill into specific activities or transactions to get the context and take the appropriate action. Dashboards should be easy to compose and personalize (without programming) by business analysts and similar power users, assembled from diverse data sources, and listed in a registry that's user friendly. Time-sensitive data should be updated in dashboards in real time. Finally, dashboards should be HTML5 compliant for use on desktop and mobile devices.

Look for tool functions that automate user collaboration. This assumes a unified development environment for streaming analytics and OI, including functionality for both technical and non-technical users. This enables some classes of business users to view data and development artifacts, plus annotate these to mark streams and other data sources of interest for a particular project. Obviously, collaboration among technical personnel is enabled, too.

NUMBER SEVEN

DEMAND A SCALABLE ARCHITECTURE WITH FLEXIBLE DEPLOYMENT OPTIONS

We've already established that real-time performance is critical to OI and streaming analytics. Equally important is the ability to elastically scale performance commensurate with event volume, event frequency, and analytical complexity. This way, the system allocates server resources as they're needed by independent applications, jobs, and processes because these ramp up and later subside. Furthermore, elastic scaling is desirable because it contributes to overall system scalability, and it works well with standard commodity hardware servers and horizontal scaling techniques, whether in the clouds or on premise.

Operational intelligence needs a distributed architecture that scales elastically from small to big. To independently scale individual components of an OI solution requires a distributed architecture. Grids of pooled servers have become the preferred architecture for elastic scaling, as seen in Hadoop, clouds, and many other modern platforms, including those for OI, CEP, and streaming data. In such an architecture, components can be elastically scaled with minimal impact on running applications, and components can be distributed on commodity hardware, which helps maximize scale while minimizing cost. Furthermore, an elastic grid of this sort is conducive to deployments on clouds, on premises, or a combination of both.

Monitoring is key to elasticity. A good OI system will include resource monitors that report on (and act on) system resource consumption (e.g., CPU and memory) for various components of the OI runtime environment (e.g., input/output rates, last timestamp processed, and subscriber counts). Similarly, monitoring the elastic server pool supports high availability; failed solutions or components are detected immediately and can be automatically restarted.

Hadoop connectivity is critical. Hadoop is rising in importance as a massively scalable yet cost-effective platform for capturing, managing, and processing a wide range of data types, including streaming data. All OI platforms should include a connector that persists streaming data in Hadoop so the data can be queried to provide historical and baseline analysis. Likewise, data captured by Hadoop should also stream back out to an OI platform for instant analysis and timely detection of opportunities and threats.

NUMBER EIGHT

COMPLEMENT EXISTING BI/DW INFRASTRUCTURE WITH A PLATFORM FOR STREAMING ANALYTICS

The extreme real-time focus of OI and streaming analytics place them at the far end of the temporal spectrum compared to traditional business intelligence (BI) and data warehousing (DW) infrastructures, which are most often applied to latent data at rest. After all, the majority of BI/DW applications serve long-term strategic decision making, where real-time data and its processing simply isn't a requirement. Even when near real-time capabilities are retrofitted onto BI/DW technologies—as is common with the practices of operational BI and performance management—users usually expect responses in the range of two to four hours. Hence, in terms of information delivery speeds and user response times, plus actual business use cases, streaming analytics and traditional BI/DW functions are complementary with very little (if any) overlap.

Note that ample technology exists for "near real-time retrofits" for BI/DW infrastructure, yet almost none exists for adding true real time. For that reason—plus the fact that OI and BI/DW are complementary—user organizations needing true real-time analytics do not even attempt a retrofit to BI/DW. Instead, they choose to add a new platform type, typically in the form of operational intelligence, streaming analytics, complex event processing, or some variation or combination.

Organizations in that situation should seek OI and streaming analytics solutions that rapidly integrate with an organization's existing BI and DW infrastructure, as well as with similar data platforms such as NoSQL databases, big data frameworks, and Hadoop. OI products should offer pre-built and pluggable, non-intrusive connectors that make seamless the end-to-end process of exchanging information between BI/DW applications/sources and OI solutions and data streams.

Such integration between OI and BI/DW enables real-time intelligence to be correlated with historical trends, which can reveal whether current real-time observations are anomalous or expected. Such correlations can avoid both false positives (sales that increase due to seasonality instead of current marketing campaigns) and false negatives (cyber attacks that hide in the noise of peak business application and network activity).

ABOUT OUR SPONSOR

www.vitria.com

Vitria Technology, Inc., is a pioneer in streaming analytics and continuous operational intelligence (OI) software. Vitria OI has been deployed by customers globally to help them uncover, analyze and act on insights from streaming data—while it still counts. Examples of how enterprises benefit from using Vitria OI include being able to 1) engage in more targeted 1:1 marketing in real time to improve their customer experience; 2) track and monitor financial transactions, user behavior, and smart grids to proactively detect and prevent cyber-security attacks and fraud; 3) continuously track and trace shipments to drive better supply chain visibility; and 4) analyze weather patterns in real time to predict their impact on electrical grids. Vitria OI's unified modeling environment accelerates solution development, and its elastic architecture enables it to scale to handle extreme Big Data volumes.

ABOUT THE AUTHOR

Philip Russom is the research director for data management at The Data Warehousing Institute (TDWI), where he oversees many of TDWI's research-oriented publications, services, and events. He's been an industry analyst at Forrester Research and Giga Information Group, where he researched, wrote, spoke, and consulted about BI issues. Before that, Russom worked in technical and marketing positions for various database vendors. Over the years, Russom has produced over 500 publications and speeches. You can reach him at prussom@tdwi.org.

ABOUT THE TDWI CHECKLIST REPORT SERIES

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data warehousing solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide Membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.