Data Stream Ingestion & Complex Event Processing Systems for Data Driven Decisions

HTC global services

Introduction

Business events significantly influence all aspects in enterprises. Data arriving continuously in huge quantity is termed a “stream”. The arriving data can be in different types and formats. Most often the streaming data provides information about the process that generated it; this information may be called messages or events. The events include data from new sources, such as sensor data, as well as clickstreams from Web servers, machine data, and data from devices, events, transactions, and customer interactions. Counts of small events add up to massive amounts of streaming data. The types of data streaming into the enterprise are increasing every day. (More information on events is available in Event Processing Systems-Way to Improved & Quicker Decision-Making white paper).

Data ingestion is the process of obtaining, importing, and processing data for later use or storage in a database. It has become a significant challenge for enterprises not only to ingest the relentless feeds of data, but also to capture value from them. As data is ingested, it is used by analytic and decision engines to accomplish specific tasks on streaming data. Organizations that can process and ingest streaming data in real time can improve efficiencies and differentiate themselves in the market. Varied examples of real-time streaming analytics can be found across industries: personalized, real-time stock-trading analysis and alerts offered by financial services companies, real-time fraud detection, data and identity protection services, reliable ingestion and analysis of data generated by sensors and actuators embedded in physical objects (Internet of Things), web clickstream analytics, and Customer Relationship Management (CRM) applications issuing alerts when customer experience within a time frame is degraded. The business world is seeking highly flexible, reliable, and cost-effective way for ingesting and performing real-time event-stream data analysis to succeed in the highly competitive business environment.

Businesses today track and process streams of data about events that may impact their fortunes in the near- or long-term horizon. Streamlining all the data streams by integrating them into one coherent and manageable body of data is the key to streamlined data processing and analysis.

What are Event Patterns?

Complex event combinations create repeatable event patterns. Event patterns arise across the horizontal and vertical business data streams. Such patterns are witnessed as a sequence of events, which are considered as critical success factors for better Return on Investment (ROI).

Industry 4.0, commits its support for event patterns. As data driven organizations across the world are transforming their business structure to support this evolution, HTC has spent time in understanding the extremely useful aspects of event patterns in energizing business returns.

Event patterns are managed by event processing engines that come with a framework for creating agents or adapters to transform external data into a format understood by the engine. For example, event pattern engine detects a pattern condition and generates alerts like alarm on high temperature. These alerts can be based on a simple value or more complex conditions such as rate of increase, and so on.

Another example: A service queue stores a series of notifications or requests in first-in, first-out order. Sending a notification enqueues the request. The request processor then processes items from the queue. Requests can be handled directly or routed to interested parties. This decouples the sender from the receiver both statically and in time. The retail segment has a lot of data movement especially with POS (Point of Sales) terminals and have several combinations of moderate data streams. HTC's data analysts understood the complex data streams to enable Retail customers with custom designed event pattern based solutions to generate deeper insights and excellent visualizations to support the corporate, regional, market, and store levels requirements. Generating data insights based on event data streams, enabled the clients to attempt data driven decisions and optimize their operations and revenue.

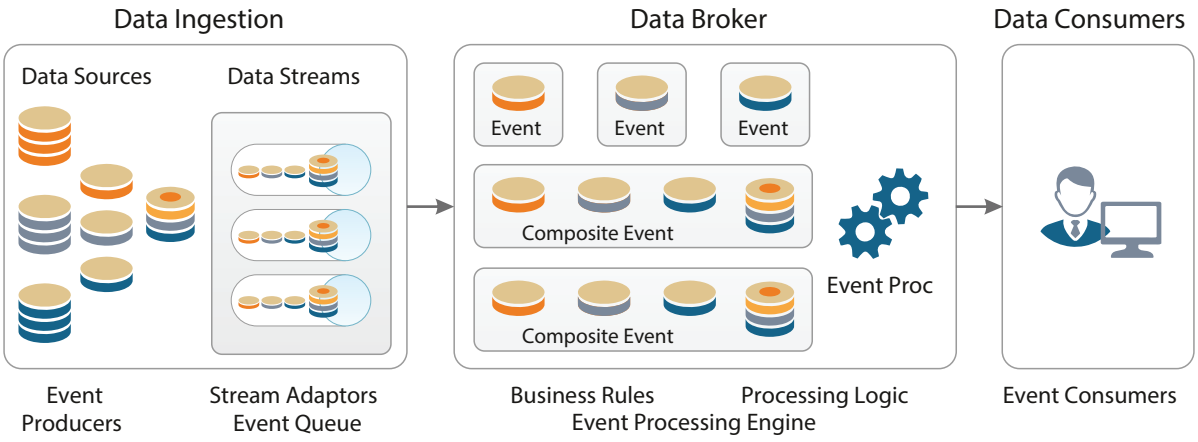
"Dynamic business environment requires operational analytics demanding answers before a question becomes obsolete. The sooner you decide, the greater you stand to reap".

- Practice Head - Big Data, HTC Global Services Inc.

Stream Processing

Most organizations work on data event streams; however, event stream processing demands higher level understanding of business and data streams. The shortcomings and drawbacks of conventional offline batch-oriented data processing was recognized by the big data community. The big data community was moving towards real time data management to ensure business course corrections at an early stage. They identified real-time query processing and instream processing as the immediate needs in many practical business application areas. A lot of traction and a whole bunch of solutions like Twitter's Storm, Yahoo's S4, Cloudera's Impala, Apache Spark, and Apache Tez have appeared and joined the army of big data and NoSQL systems to address the outlined needs.

The term “stream processing” is used to describe data processing before storing it to a system. Stream processing is composed of defined data processing components. A processing unit in a stream engine is generally called a Processing Element (PE). Every PE gets input from the input queues, performs computation on the input and produces output to their output queues. All the components of the stream processing engine work together as a solution framework. The below image depicts the method of stream processing:



What is an Event Stream Processing Engine?

Event Stream Processing (ESP) has picked up a lot of interest across industries as the way forward. It is a software solution built on a range of technologies that are used to create and deploy solutions to process large volumes of incoming events, analyze those complex events, and respond to conditions of business interest in real-time. Technological dependencies are captured as a framework with program models built around event patterns. The model describes developers' approach to build their applications. It includes the set of abstractions, specific language features, and APIs that developers need to use and apply to work with an ESP Engine. Stream processing technologies like Apache Samza and Apache Storm have received attention for large scale streaming analytics. HTC understands the interdependencies of solution framework and helps its clients to reap benefits from big data implementations.

What are the Types of Stream Processing Engines?

The solutions in the market are classified on certain aspects of business returns. Stream processing is addressed centrally and in a distributed fashion. This white paper classifies stream processing engine into two types. Each type has its own advantages and challenges. A good data consultant identifies the right deployment based on business conditions, expected returns, and the futuristic plans of the organization.

First type:

- There is a central server that orchestrates the distribution of PEs into physical nodes
- The physical node is responsible for load balancing tasks amongst the nodes and schedule the tasks to run on different nodes

Second type:

- There is no central server to orchestrate the work among the nodes
- The nodes work as fully distributed peer-to-peer system, sharing the work equally

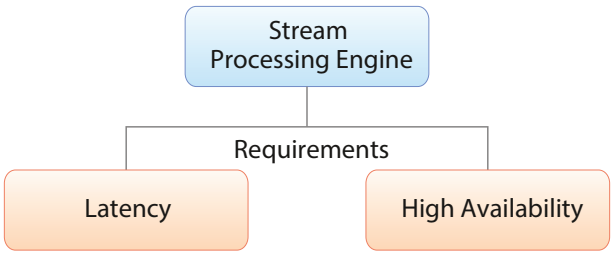
"High-performance analytics can churn data including big data asset into quicker, better business decisions, and ultimately generate competitive advantage".

- Practice Head - Big Data, HTC Global Services Inc.

What are the Requirements of a Stream Processing Engine?

The success of stream processing solutions depend on explicit data driven understanding of business constructs. The requirements of a good stream processing framework revolve around two important attributes - latency and high availability of the system.

Systems call for lowered latency with increased high availability. Both (latency and high availability) are competing requirements which rely on each other. It is important to understand that large scale batch processing systems such as Hadoop have high latency for operations. Therefore, delaying a computation when an event fails is non-critical in this environment. Hadoop demands for correct calculations without any loss of information amidst the failing components.



Latency is of greater importance for a streaming framework and must recover fast so that normal processing can continue with minimal effect to the overall system. The recovery methods for stream processing engines vary between recoveries with no effect to the system without any information loss to recoveryto recovery with some information loss. Recovery without any information loss and recovery with minimum latency are the two extremes of recovery methods. This is a critical reasoning and the data streaming solutions have addressed this as a part of solution framework.

What is Real-time Event Stream Processing?

Real-time Event Stream Processing (RT-ESP) is used to discuss a class of applications which manage and coordinate multiple streams of event data that have timeliness requirements. For example, we are discussing data streams routed from sensors (such as, surveillance cameras, temperature probes, etc.), as well as other types of monitors (such as, online stock trade feeds) considered as a part of real time data streams. It is important to understand the difference between continuously generated data streams which differ from streaming the contents of a data file (such as a fixed-size movie), since the end of RT-ESP data is not known.

Event Use Cases

All solutions in the market are geared to address specific use cases and business pain areas. Real-time analytics is built around streaming data which provides solutions to a number of business challenges that are specific to realtime data. Data analytic techniques are applied on the data streams before they are moved out and stored in the data repository. Organizations gain the benefits of real-time insights and deep data understanding as business events occur, eading to data driven decisions. They store information in a robust repository for historical analysis and later needs. HTC builds solutions with clarity when approaching data problems and data driven solution design to ensure maximum benefits to the customer. Some interesting data driven solutions are outlined in the following sub-sections.

Stock Trading

The stock market is an important part of a country's economy. It plays crucial role in the growth of industry and commerce, and also affects the economy. Therefore the government, industry, and the country's central bank keep a close watch on the happenings of the stock market. The stock market is important from both the industry's point of view as well as the investor's point of view.

Stock trading is critical from various aspects. Stock management is centrally controlled by government agencies. Spotting a trading opportunity in the stock market could involve computing the average price of a stock over the last hour and then comparing this aggregate event with the stock price over the last week. HTC is working with various financial services clients for several years and has gained good experience in this domain. A financial services organization that actively monitors financial markets, individual trader activity, and customer accounts will have an application running on a trader's desktop to track the moving average of value in the investment portfolio. This moving average needs to be updated continuously as stock updates arrive and trades are confirmed. A second application extracts events from live news feeds and correlates these events with market indicators to infer market sentiment, impacting automated stock trading programs.

HTC developed a custom data administration solution for a leading financial planning advisor company to track the progress of portfolio management for their clients. The data solution helped querying the patterns of events correlated

across time and data values, where each event has a short “shelf life”. The main features of the solution was to track client's assets, maintain activity log, data confidentiality, provide notification, news alerts, and notify on maturing investor certificates. The designed data solution opened up visibility on stock data streams and enabled clients and dependent customers to identify trading opportunities.

Device Telemetry

Telemetry is an emerging area of development opening up new opportunities of business and an excellent data insight. It is used by manufacturers to gain insight on various vehicle performance parameters. Telemetry with big data streaming solution framework provides insights creating new business streams.

By enabling real-time analytics for connected vehicles, manufacturers enable new levels of visibility into the running performance of a given vehicle, than waiting to download telemetry data during scheduled vehicle service. This in turn enables additional scenarios such as predictive maintenance whereby the health of the vehicle is consistently validated against proven machine learning models to detect anomalies in the vehicle. For example, insurance companies' rely on telemetry data streams for redesigning their insurance products. Insurance companies track the driving patterns of vehicle owners and adjust the insurance prices accordingly. HTC is working in this sphere of data streaming solution design to enable end customer benefits. This includes the combination of surveillance cameras, motion detection sensors, and audio sensors enabling security personal to monitor suspicious activity in real time environment.

Website Analytics

The website serves as the online cyber face of an organization. With its global reach, the cyber home of an organization is open for customers and is mandated to answer the queries of visitors. Websites generate huge volume of data based on customers' interactions and style attributes. It is evident that the customers leave an excellent cyber foot print for analysis, segmentation, and grouping. Web analytics applied on streaming data with supporting data models opens the various aspects of customer behavior and customer sentiments to fine tune the solutions. Such insights and web metrics are considered as gold mine by the CMO (Chief Marketing Officer) and the Marketing team.

By understanding the web metrics, web managers refine web layouts and web marketing content to suit visitors' trends. Real-time web metrics are useful for understanding the impact of advertising investment that drive people to access a site. By leveraging real-time telemetry and analysis of site actions, developers can perform highly interactive testing of site layouts to continually iterate changes and optimizations. This can be particularly effective in sites that experience heavy usage.

HTC has built its strength around text mining solutions and has spent time to understand the various aspects of sentiments and consumer polarity exhibited through many marketing channels. With clarity on multi-channel and omnichannel customer interactions and tag management solutions to track customer behavior, HTC has identified potential solutions to address the pain areas of web analytics.

Event Processing Systems

With dynamic growth in technology, coupled with Industry 4.0 standard and developments in big data technology, event processing platforms have become more prevalent in diverse domains of industry, such as capital markets, telecom, and healthcare. Event Processing System (EPS) address the data processing needs of organizations with high velocity data. The event processing market today is very heterogeneous with several competing products.

EPS provides information for understanding the status of business / infrastructure / information system. EPS information from various environment provides insight on the current state for decision-making. The architecture for real-time event processing is based on the pattern of data ingestion, data processing, and ultimately consumption. EPS facilitates complex analysis of high throughput live data feeds.

Complex Event Processing

Complex Event Processing (CEP) technology has gained popularity for efficiently detecting event patterns in real-time. They capture critical exceptions, threats, or opportunities that occur in a given domain area spread in space and time. To allow CEP technology to be an end-to-end solution, it is essential to promptly address the risks and opportunities detected by pattern queries in real-time.

CEP systems apply the “pattern matching” query technique where a set of event streams are matched against a complex pattern specifying constraints on the extent, order, values, and quantification of matching events. Pattern matching identifies the sequence of events occurring in order and correlates them based on the values of their attributes, providing the essential framework to react in near real time. Collection of patterns emerge as the process knowledge which drives the decision making of an organization.

Event Processing System Models

Event Processing (EP) systems adopt a processing model different from the conventional data management systems. The exhibited divergence is a natural consequence of varying requirements catalogued by event-driven applications in comparison with regular transaction processing and analytical applications. While databases deal with data that need to be stored for posterior access, the usefulness of events are limited to a short time after their occurrence. Therefore, EP systems manipulate transient data, which is kept most of the time in main memory to ensure quick answers.

There is another reason attributed for the emergence of EP systems. The Conventional Database systems adopted a pull-based approach that requires applications to issue a query to retrieve data from the database. On the other hand, EP systems are designed to deal with applications that generate answers based on the arrival of new data streams (push-based model). The two approaches are considered to be orthogonal.

What are Event-Driven Applications?

Having understood the nature of event streams, organizations can be benefitted by building applications around event streams to cater to the needs of the organization. Such applications are architected and designed with lot of planning and care to meet the dynamic hypermarket driven data analytics. HTC with varied domain specializations had spent time on different blue print designs and its impact to the hypermarket.

The process of extracting information from event streams is usually addressed with event stream filters. This often starts by selecting portions of the incoming records which meet certain criteria for further processing. Such selections are addressed with the help of data filters. Quite often it is necessary to filter and mix events emanating from different data streams to generate meaning out of it. For example, to detect non-ideal environmental conditions in a factory it may be necessary to merge readings from multiple sensor types. Such requirements are addressed through event driven applications.

Data reduction process is addressed either horizontally by discarding the entire events and keeping only those that satisfy a given predicate, or vertically by removing some attributes of each event. Such operations are equivalent to the relational operations selection and projection. HTC has spent time on the drawing board and the research labs to validate the application prototypes. One of the most fundamental features offered by event processing platforms is detecting sequences of events that together represent a situation of interest. Applications come in a spectrum of formats, and their functional and nonfunctional requirements tend to vary significantly from one domain to another. Event-driven applications need some form of integration with existing systems to receive events from external sources and data feeds and output results to consumers. To address this need, most event processing platforms are bundled with a set of input and output adapters that allow them to communicate through diverse technologies and protocols.

What are Event Queries?

Querying events of interest to satisfy the needs of the proposed application appeals for understanding the capability of query engines which can be deployed. There are a variety of query engines available in the market. Event query engines are built around query rules module defining a variety of event patterns which defines the limitations of the engine. Event query rules are used to detect complex event patterns from streams of raw events.

An event pattern query is generally a statement specifying a set of dependent events, relative order, and time spans within which the events need to happen. Instead of storing data once and executing queries multiple times over it, EP system queries are registered once and data is matched against them producing a continuous flow of answers. For instance, an intrusion detection application may look for a sequence of five consecutive failed login attempts emanating from the same remote terminal within an interval of one minute. Variation of the concept would look for negative patterns addressing the non-occurrence of events. For example, a fleet management event management software may need to emit a warning if a vehicle is known to have departed but no corresponding notification informing of its arrival at the destination is received within its expected travel time.

Esper is one of the leading open source CEP engines in which events are modelled as object instances exposing event properties through Java Bean-style getter methods. More information on Esper is available in Managing Real-time Data Streaming Effectively with CEP Solutions white paper. Event specification languages provide constructs to define relations between event streams.

Conclusion

Event processing systems play a vital role in data driven decisions. As discussed in the earlier sections of this white paper, EP applications have the capability to support and automate the critical jobs within an organization. EP systems when integrated with Industry.4.0, can run continuously for hours or even days without interruption. EP systems enable an organization to drive the quality systems and it is very likely that the conditions change during their execution. The primary benefit of using event processing systems lies in taking actions promptly when they are more effective. HTC had recognized the potential of data driven solutions at an early stage and groomed data consultants with specialized skills to address the needs of organizations. With HTC as a partner to help you design your Industry 4.0 ready event stream solutions, it would be easier to move your attention to core business zones. With HTC as an enabler, reap the advantage of better-informed decisions as EP systems are designed to rapidly extract and filter information from massive amounts of high velocity data that are impossible for humans to analyze.

Acronyms

The acronyms used in this white paper and their expansion are provided below:

Acronym	Expansion
CEP	Complex Event Processing
CRM	Customer Relationship Management
EP	Event Processing
EPS	Event Processing System
ESP	Event Stream Processing
IoT	Internet of Things
PE	Processing Element
POS	Point of Sales
ROI	Return on Investment
RT- ESP	Real-time Event Stream Processing

References

1. Internet of Things
http://www.cisco.com/web/solutions/trends/iot/portfolio.html

2. Scaled Management System
http://www.google.co.in/patents/US8055649

3. Real-Time Event Processing with Microsoft Azure Stream Analytics
https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CB0QFjAAahUKEwju8sbqrOfHAhVKno4KHSxZAfo&url=http%3A%2F%2Fdownload.microsoft.com%2Fdownload%2F6%2F2%2F3%2F623924DE-B083-4561-9624-C1AB62B5F82B%2Freal-time-event-processing-with-microsoft-azure-streamanalytics.pdf&usg=AFQjCNHL_MNZWppOGuUPe313BMRj1BlBQ&bvm=bv.102022582,d.c2E

4. Dataflow Networks for Event Stream Processing
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.8473&rep=rep1&type=pdf

5. Towards a REST-ful Visualization of Complex Event Streams and Patterns
http://worldcomp-proceedings.com/proc/p2014/EEE3246.pdf

6. Performance Evaluation and Benchmarking of Event Processing Systems
https://estudogeral.sib.uc.pt/jspui/bitstream/10316/24296/2/Performance%20Evaluation%20and%20Benchmarking%20of%20Event%20Processing%20Systems.pdf

7. Event Queue
http://gameprogrammingpatterns.com/event-queue.html

8. In-Stream Big Data Processing
https://highlyscalable.wordpress.com/2013/08/20/in-stream-big-data-processing/

9. Evaluating Transport Protocols for Real-time Event Stream Processing Middleware and Applications
http://www.dre.vanderbilt.edu/~jhoffert/FLEXMAT/adamant-doa09-sv.pdf

10. Semantic Processing of Sensor Event Stream by Using External Knowledge Base
http://ceur-ws.org/Vol-904/paper8.pdf

11. Survey of Distributed Stream Processing for Large Stream Source
http://grids.ucs.indiana.edu/ptliupages/publications/survey_stream_processing.pdf

About HTC's Big Data CoE

HTC's Center of Excellence for Big Data Management and Analytics brings in mature technologies and thought leadership. Our dedicated R&D team develops highly customized and cost-effective cutting edge solutions to enable clients manage and understand big data for improved and quicker decision makaing.

This white paper was developed by HTC's Big Data CoE.

HTC global services

Reaching out… through IT ®

World Headquarters

3270 West Big Beaver Road
Troy, MI 48084, U.S.A
Phone: 248.786.2500
Fax: 248.786.2515
www.htcinc.com