



inside**BIGDATA**

InsideBIGDATA Guide to
Streaming Analytics

by Daniel D. Gutierrez

BROUGHT TO YOU BY



IMPETUS
gathr

Streaming Analytics – An Overview

Streaming analytics platforms provide businesses a method for extracting strategic value from *data-in-motion* in a manner similar to how traditional analytics tools operate on *data-at rest*. Instead of historical analysis, the goal with streaming analytics is to enable near real-time decision making by letting companies inspect, correlate and analyze data even as it flows into applications and databases from numerous different sources. Streaming analytics allows companies to do event processing against massive volumes of data streaming into the enterprise at high velocity.

Instead of thousands or tens of thousands of events per second, a streaming analytics platform can process millions and even tens of millions of events per second.

Streaming analytics technologies enable action based on an analysis of a series of events that have just happened. Further, modern streaming analytics tools provide support for large data volumes and sophisticated query processing. Instead of thousands or tens of thousands of events per second, a streaming analytics platform

can process millions and even tens of millions of events per second. Because data in a streaming analytics environment is processed before it lands in a database, the technology supports much faster decision making than possible with traditional data analytics technologies. With traditional analytics you gather information, store it and do analytics on it later. This is called “at-rest analytics.” With streaming technologies the analysis is done as the data arrive.

Common use cases for streaming analytics abound. For example, dashboards and visualization software integrated on top of streaming analytics platforms can help enterprises visualize and monitor their business in real-time. Such tools can be used to monitor changing customer attitudes through use of social media sentiment analysis. Similarly, streaming analytics capabilities can be used to enable real-time alerts or leverage new business opportunities — like making promotional offers to customers based on their geographical location at a specific time. Streaming analytics capabilities are also vital in the security-monitoring context because they give organizations a way to quickly correlate seemingly disparate events to detect threat patterns and evaluate risks. Government agencies have used these capabilities to do security monitoring of both network and physical assets.

Contents

Streaming Analytics – An Overview	2	StreamAnalytix by Impetus	9
Genesis of Streaming Analytics.....	3	Open source technology.....	10
Streaming Analytics Tools	3	Spark streaming.....	10
The Business Value of Streaming to the Enterprise	4	Versatility and comprehensiveness	10
Selecting a Streaming Architecture	6	Abstraction layer driving simplicity.....	10
Case Studies: How are Enterprises Using Streaming Analytics	8	Compatibility	10
Case Study: Call center monitoring.....	8	“Low latency” capability and flexible scalability	11
Case Study: Real-time multilingual classification and sentiment analysis of text	9	Intricate robust analytics	11
Case Study: Estimating TSA security queue wait time.....	9	Detailed data visualization.....	11
		Summary	11

Genesis of Streaming Analytics

Streaming analytics technology grew out of demand by enterprises that experienced a strong upward trajectory of data volume, velocity and variety as well as a need to ingest and evaluate this data to quickly make strategic business decisions.

STREAM PROCESSING CAPABILITIES

Stream processing requires two specific technology capabilities:

- ▶ First, in order to do stream processing, organizations need to have a way to ingest data from multiple sources. Often, the data types and sources can be highly varied. Any technology that is used for stream processing needs to be able to consume different data types, at very high volumes, and from varying sources.
- ▶ The second capability takes the form of an analytics engine capable of filtering, aggregating and correlating streaming data in order to find useful patterns. Sometimes, in order to detect patterns and enable useful insights, there may be a need to enrich the data stream with data from an existing database.

Many companies combine stream processing with batch processing to derive optimal value from their data. In these situations, once queries have been run against the streaming data the data is stored in a platform like Hadoop or other database for later retrieval and analysis. Some vendors have begun using the term Lambda architecture to describe this sort of a hybrid streaming analytics and batch analytics data environment.

The manner in which streaming data is used after the initial ingest can vary to some degree. One tactic might be to land all the data in Hadoop or a data lake to do subsequent analysis. Another tactic might be to discard a lot of the data and just keep an aggregated form.

Streaming Analytics Tools

There are many technology options for streaming analytics today and the ecosystem is evolving fast. The big enterprise players in this space include SAP, IBM, Informatica, Software AG, Oracle and TIBCO. Open source streaming analytics projects such as Apache Storm and Spark Streaming also have generated a lot of attention recently. Early adopters of these technologies have included major Internet companies like Twitter, Groupon, Spotify, Yelp, Uber, Pinterest and the Weather Channel, as well as other large enterprises including Cisco, Bosch, Rockwell Automation, Schneider Electric, Emory University Hospital, UCLA Department of Neurosurgery and others. In addition there are the pure-play technology vendors like DataTorrent with Apache Apex and also Cask (formerly Continuuity) which has teamed with AT&T Labs on an open source project called Tigon, a real-time stream processing framework built on top of Hadoop and Hbase. In addition, Gathr, another prominent member of this ecosystem, is a unique streaming analytics platform based on a best-of-breed open source technology stack.

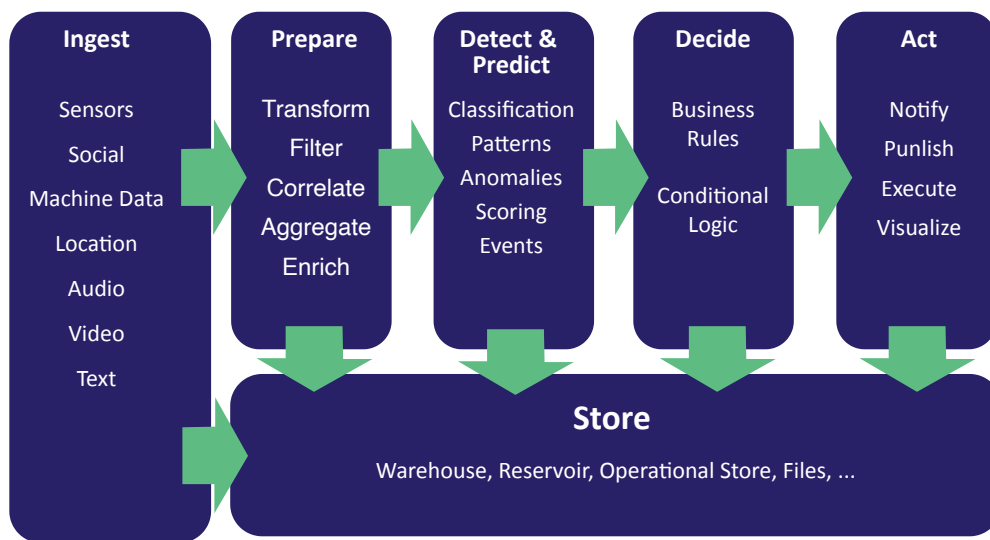
Many of these tools are configured for the use and support of query processing capabilities out-of-the box and offer relatively easy to use and intuitive visual interfaces for running queries.

In addition, many vendors offer hosted streaming analytics services that are good for companies with cloud applications. Amazon's Kinesis, Google's Data Flow and Microsoft's Azure, for instance, all support real-time processing and data analytics. Others include: Apache Samza from LinkedIn and Twitter's Heron which might become open source soon.

Further, the leading Hadoop distributions have all embraced streaming analytics for their architectures including: MapR, Hortonworks, and Cloudera.

(See graphic, page 4.)

Streaming Analytics Architecture



Source: Forbes Tech, July 30, 2015

The Business Value of Streaming to the Enterprise

We are experiencing a significant upward trajectory in terms of uptake of streaming capability across industries like financial services, retail, healthcare, telecommunications, oil & gas, ad tech and many more. Streaming analytics helps enterprises by visualizing the business in real-time, cutting preventable losses, detecting urgent situations and automating immediate actions. In recognition of these benefits, many important verticals are going through a committed proof-of-concept process.

The centralized architecture teams of an increasing number of enterprises are evaluating different technologies for streaming. As a result, it's a good time for business decision making in terms of allocating budget. Creating business use cases and detailing business value is going to be required as part of the due diligence process.

The business value for streaming analytics can be outlined as follows:

- Routine business operations (real time systems)
 - Manufacturing control systems
 - IT systems and network monitoring
 - Field assets monitoring and alerting, e.g. trucks, oil rigs, vending machines, radio towers
 - Financial transactions processing, e.g. authentications, validations, fraud
- Cutting preventable losses
 - Loss of lives and assets
 - Manufacturing defects
 - Major security breaches in retail
 - Stock exchange meltdown
 - Brokerage - fraudulent or risky trades
 - Medical/clinical – complex analytics in ICU
 - Disaster warning systems
 - Preventative maintenance – machine, plants
 - Customer churn
 - Brand reputation on social media
- Finding and monetizing missed opportunities – increased revenue, cost savings
 - Listening and learning from customers (social)
 - Context sensitive inventory, products, ads
 - Recommend, up-sell, cross-sell
 - Network optimization for cost, quality of service
 - Dynamic capacity management
 - Dynamic re-routing of traffic, cargo
 - Insurance adjudication, drone image analysis
- Creating new opportunities – new business models, products, services, revenue
 - Tractors are becoming soil sensors
 - Telecom giants selling data and insights

In order to derive business value from streaming analytics, organizations need to think a little bit differently about how they analyze data. Traditional analytical tools are optimized for request and response from static data. With streaming analytics, the data is pouring in continuously and you don't know what's in that data. Application developers need to stop thinking about request and response and start thinking about detecting interesting events as they come in. Streaming analytics offers organizations an opportunity to ingest and glean instant insights from real-time data coming in via transactions, cloud applications, web interactions, mobile devices, and machine sensors.

The emerging Internet-of-Things (IoT) will also fuel demand for streaming analytics capabilities in the near term. Sensor data from thousands of Internet connected devices can give companies valuable insights on the health of a network, system or application.

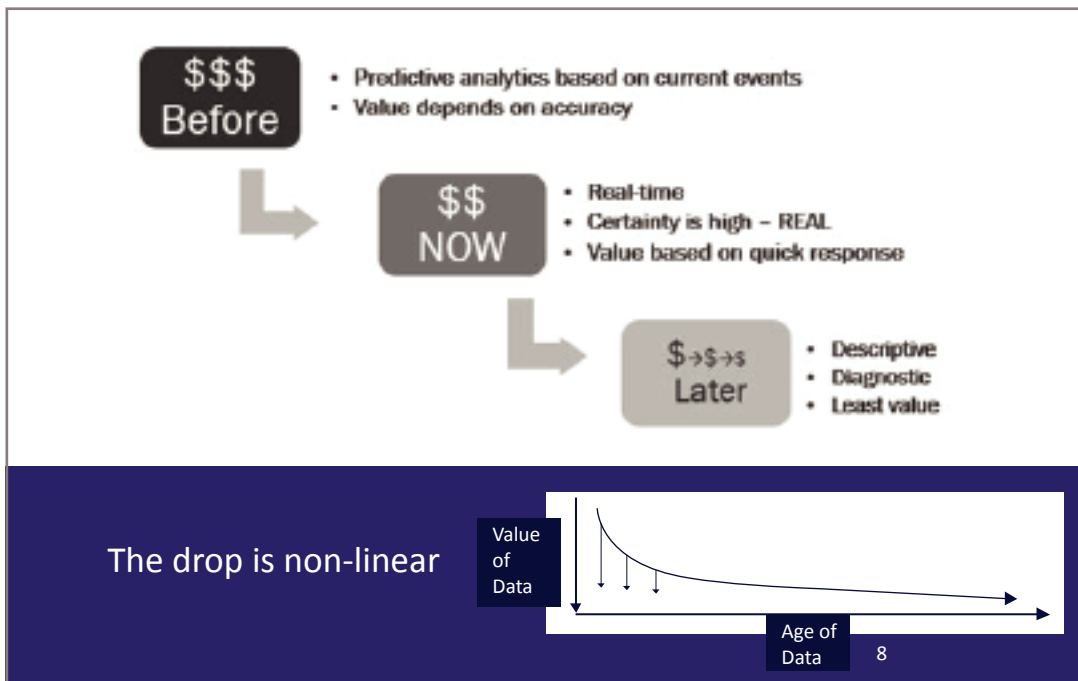
The best way to look for use cases in support of business value is to consider the most challenging business processes and walk through them at each

One common characteristic of streaming data value is that it decreases non-linearly over time. The key is to ingest data, compute actionable insights and react to them in real-time.

stage to identify situations where additional data might help. Ask yourself if there are data sources available that would provide information in real-time to make the process more efficient. On the customer side, walk through the customer journey and see if additional data can improve or detect something in real-time.

One common characteristic of streaming data value is that it decreases non-linearly over time. The key is to ingest data, compute actionable insights and react to them in real-time. The longer you wait, the less business value.

Streaming Analytics Business Value Diminishes in a non-linear manner with the age of data



Selecting a Streaming Architecture

APPROACHES TO SELECTION

There are several approaches an enterprise can use toward selecting a streaming analytics architecture that's right in terms of options and components.

▶ First, there is the “proprietary” platform where you purchase a license for using the software. Points of consideration for this approach include: there's no leverage of open source, you'll experience vendor lock-in and then there is a potentially high cost.

▶ Next there is the “do it yourself” approach. Although this option may be the most flexible, there are a number of considerations including: dealing with native open source, there will be no vendor support, you may experience integration and maintenance challenges and there could be significant delays in time-to-market.

▶ The third and arguably the most attractive is a “hybrid” approach that mitigates the disadvantages of the other two approaches and offers the benefits of both worlds to enterprises for streaming analytics:

- **Enterprise class** – reliable, manageable, vendor support and professional services
- **Open source** – community innovation, no vendor lock-in, low cost and future proof

We're now seeing critical mass for Hadoop deployments where the foundation of Big Data, the Hadoop platform, becomes solidly established across enterprises. Hadoop platforms have streaming as part of their portfolio — it is becoming easier for enterprises to add on this core capability as part of their business process. Enterprises have been asking for streaming and as a result the principal Hadoop distributions like MapR, Hortonworks, and Cloudera all have included some sort of streaming capability in their platforms. Enterprise architects are taking feature-sets of these platforms and preparing typical questions to ask to clarify his/her position before making a recommendation for procuring a streaming technology.

The pure open source solution is not necessarily enterprise ready. An enterprise might need a whole engineering team with specific skill-sets and not every company has this level of talent and even if it does, it might not be able to deploy a team fast enough.

For the enterprise that's decided they are going to move to streaming technology, decision makers will find the number of options in the marketplace to be quite daunting. Many organizations prefer an open source solution if there is a viable and credible option available. Open source technology is getting serious consideration by the largest enterprises.

This direction, however, requires some level of assistance because it's not an easy journey:

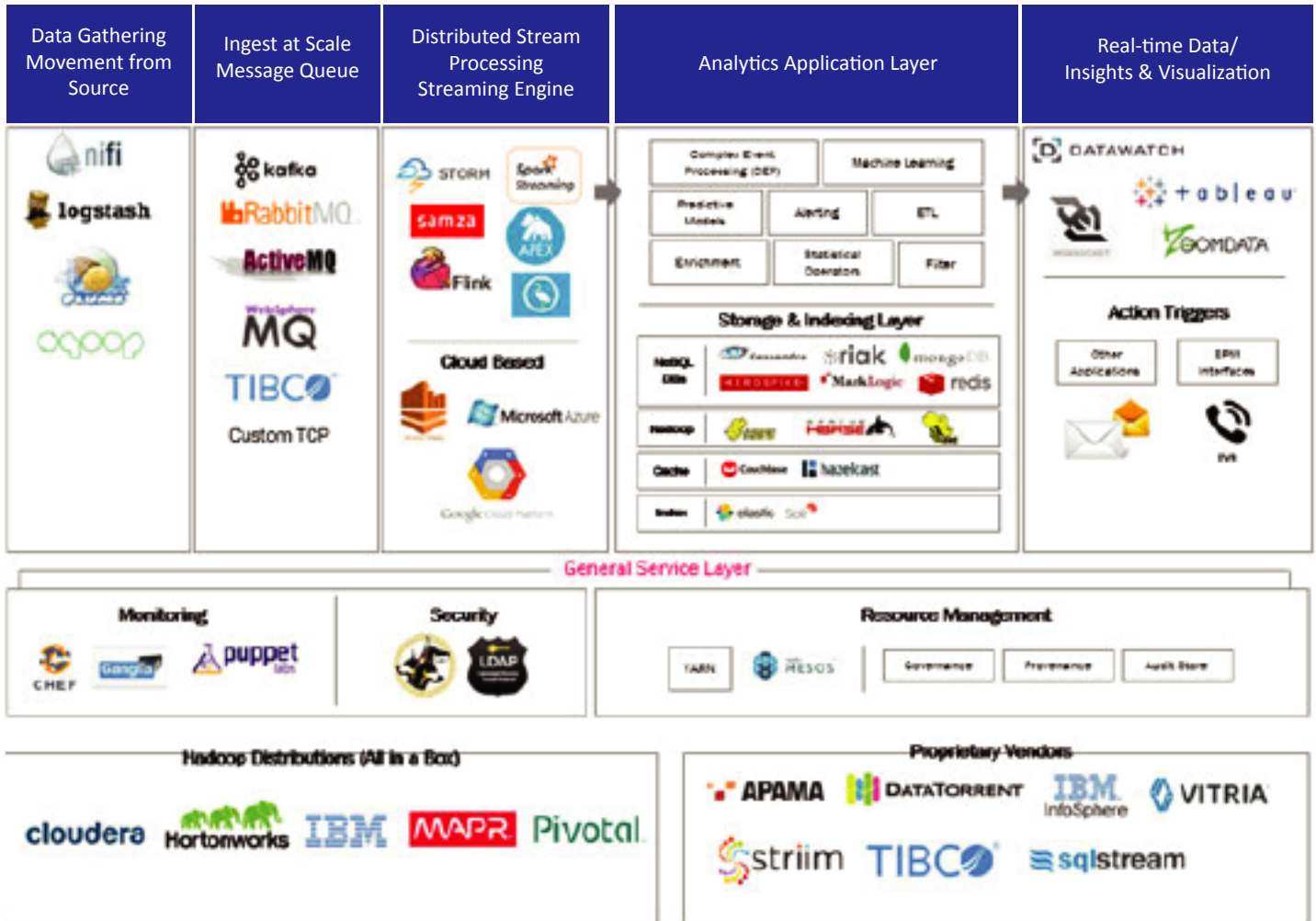
1. There are a lot of components to integrate and maintain
2. There are skills to develop
3. It's not intuitive
4. In many cases there is no UI at all

So for these reasons, the pure open source solution is not necessarily enterprise ready. An enterprise might need a whole engineering

team with specific skill-sets and not every company has this level of talent and even if it does, it might not be able to deploy a team fast enough. Moreover, the desire would be to invest the engineering resources on core business applications — not the underlying platform.

Gathr helping enterprises leverage best-of-breed open source technology and associated tools for building streaming analytics applications quickly and easily so they can get to market much faster.

Streaming Technology Ecosystem



Case Studies: How are Enterprises Using Streaming Analytics

As the technology grows in popularity, we see an increasing number of use case examples for how streaming analytics plays a significant role in cultivating competitive advantage. At a high level, there are a growing number of application areas such as IoT, mobile app analytics and call center monitoring and analytics. There are also a variety of horizontal applications coming to the forefront: customer experience, clickstream analytics, context-sensitive offers and recommendations, IT log analytics and security. In addition, we see an emergent list of verticals including:

- ▶ Call center analytics
- ▶ Predictive maintenance
- ▶ Clinical care and patient management
- ▶ Sensor data analytics
- ▶ Fleet operations
- ▶ Security
- ▶ Fraud and anomaly detection
- ▶ Gaming analytics
- ▶ Churn analytics
- ▶ Network traffic analysis and optimization
- ▶ Manufacturing
- ▶ Oil & Gas
- ▶ Healthcare – Clinical
- ▶ IoT
- ▶ Transportation/logistics
- ▶ Financial Services
- ▶ Retail
- ▶ E-commerce
- ▶ Log analytics
- ▶ Online advertising
- ▶ Banking
- ▶ Credit-line management
- ▶ Insurance claim validation
- ▶ Telecom

One particular use case that is growing in importance is *business activity monitoring (BAM)*. Many large organizations maintain extensive, complex processes with the need to make sure the systems are all saying the same thing. Reconciliation and auditing systems absorb data as it moves from various systems and there's a need to monitor and audit potentially hundreds of different parameters across

the entire business process flow. This is a good example of applications built on top of a streaming platform to achieve real-time reconciliation. Such a system avoids having to find out hours or days in the future that the systems were incongruent.

▶ CASE STUDY Call center monitoring

Call center monitoring with real-time customer interactions, rich context sensitive customer interactions is a good example of customer application building on top of the streaming platform. Call center monitoring proved the immediate and quantifiable business benefits from real-time streaming analytics for the telecom/VOIP/Call-Center industry. Some of these benefits include: numerous person-months of productivity gain, customer complaint resolution speed, customer satisfaction index and higher customer retention rates.

The successful streaming analytics solution allowed the call centers to process millions of minutes of calls per day across vast distributed networks around the globe, and also provided an infrastructure monitoring platform that allowed a unified view and analysis of events in real-time.

The solution included:

- **SLA Alerts:** Service level alerts in real-time allow managers to escalate issues and resolve them as they are happening
- **Sentiment Analysis:** The system performs real-time, multi-lingual classification and sentiment analysis of text data, including the ability to generate alerts on email and conversations happening in real-time
- **Predictive Analytics:** A reporting tool provides the ability to generate historical reports for future pricing models and requirement identification. The reports can be viewed on the UI for analysis and enabling business decisions

CASE STUDY

Real-time multilingual classification and sentiment analysis of text

A major telecom company providing nationwide telecom services wanted a system that performs real-time, multi-lingual classification and sentiment analysis of text data. The solution was required to allow storing, indexing, and querying petabytes (PBs) of data with a very high throughput. Some of the critical requirements were: ingest and parse high volume of data (15 TB per day) of varied types (e.g. weblogs, email, chat, and files), apply real-time multi-lingual classification and sentiment analysis with very high accuracy (four nines), store metadata and raw binary data for querying, with query response of 5s on cold data.

The successful streaming analytics solution included:

- Rapid and accurate real-time text categorization and sentiment analysis
- Adjustable text categorization for domain-specific classes
- Multi-lingual support
- Enhanced sentiment analysis to focus on feature-specific opinion mining
- Linear scalability to increase the number of nodes in the cluster
- Provision to add custom component for added functionalities

CASE STUDY

Estimating TSA security queue wait time

Another compelling use case for streaming analytics is estimating the TSA security queue wait time for passengers before they reach the airport by providing notifications such as “there will be a 45 minute wait” through a mobile app that’s distributing alert signals for hundreds of passengers.

Gathr by Impetus

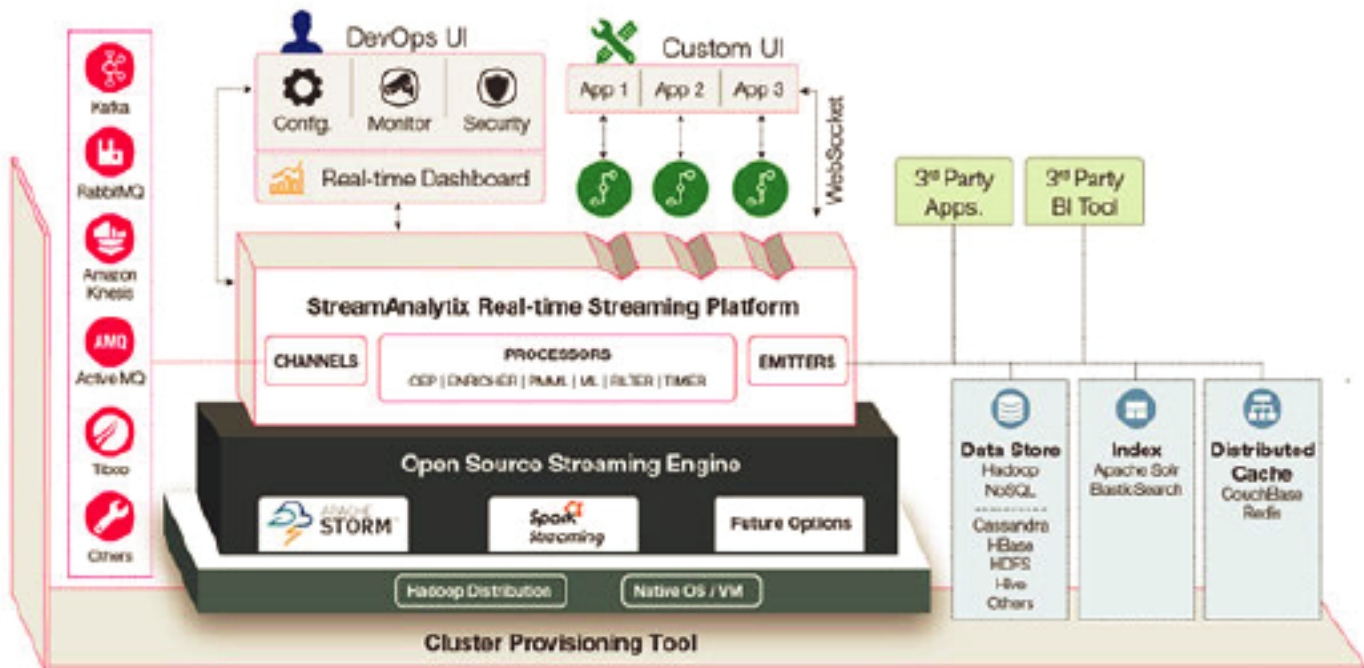
Gathr is a state-of-the-art streaming analytics platform based on a best-of-breed open source technology stack. Gathr is a horizontal product for comprehensive data-ingestion across industry verticals. It is developed on an enterprise-grade scale with open source components including Apache Kafka, Apache Storm and Apache Spark while also incorporating the popular Hadoop and NoSQL platforms into its structure. The solution provides all required components for streaming app-development not normally found in one place, all brought together under this platform combined with an extremely friendly UI. (See graphic, page 10.)

A key benefit of Gathr is the multi-engine abstracted architecture which enables alternative streaming engines underneath — supporting Spark Streaming for rapid and easy development of real-time streaming analytics applications in addition to

Being able to choose among multiple streaming engines means you can take the risk out of being constrained with a single engine.

original support for Apache Storm. Being able to choose among multiple streaming engines means you can take the risk out of being constrained with a single engine. With a multi-engine streaming analytics platform, you can do Storm streaming pipelines and Spark streaming pipelines and interconnect them — using the best engine for the best use case based on the optimal architecture. When new engines become widely accepted in the future they can be rolled into this multi-engine platform.

Gathr Functional Block Diagram A multi-engine streaming analytics platform



The following is an overview of the product and its enterprise-grade, multi-engine open source based platform:

Open source technology

Gathr is built on Apache Storm and Apache Spark (open source distributed real-time computation systems) and is therefore able to leverage the numerous upgrades, improvements and flow of innovation that are foundational to the global Open Source movement.

Spark streaming

Spark streaming includes a rich array of drag-and-drop Spark data transformations, Spark SQL support, and built-in operators for predictive models with inline model-test feature.

Versatility and comprehensiveness

Gathr is a “horizontal” product for comprehensive high-speed data-ingestion across industry verticals.

Its IDE development environment offers a palette of applications based on customer requirements. Multiple components can be dragged and dropped into a smart dash-board in order to create a customized work-sphere. The visual pipeline designer can be used to create, configure and administer complex real-time data pipelines.

Abstraction layer driving simplicity

The platform’s architecture incorporates an abstraction layer beneath the application definition interface. This innovative setup enables automatic selection of the ideal streaming engine while also allowing concurrent use of several engines.

Compatibility

Built on Apache Storm, Apache Spark, Kafka and Hadoop, the StreamAnalytix platform is seamlessly compatible with all Hadoop distributions and vendors. This enables easy ingestion, processing, analysis, storage and visualization of streaming data from any input data source, proactively boosting split-second decision making.

“Low latency” capability and flexible scalability

The platform’s ability to ingest high-speed streaming data with very low, sub-second latencies makes it ideal for use cases which warrant split-second response, such as flight-alerts or critical control of risk factors prevalent in complex manufacturing environments. Any fast-ingest data store can be used.

Intricate robust analytics

Gathr offers a wide collection of built-in data-processing operators. These operators enable high-speed data ingestion and processing in terms of complex correlations, multiple aggregation functions, statistical models and window aggregates. For rapid application development, it is possible to port predictive analytics and machine learning models built in SAS or R via PMML onto real-time data.

Detailed data visualization












Gathr provides comprehensive support for 360-degree real-time data visualization. This means the system delivers incoming data streams instantaneously in the form of appropriate charts and dashboards.

Summary

Many enterprises find themselves at a key inflection point in the big data timeline with respect to streaming analytics technology. There is a huge opportunity for direct financial and market growth for enterprises by leveraging streaming analytics. Streaming analytics deployments are being engaged by companies in a broad variety of different use-cases. The vendor and technology landscape is complex and numerous open source options are mushrooming. It’s important to choose a platform that will supply a proven and pre-integrated, performance-tuned stack, ease of use, enterprise-class reliability and flexibility to protect the enterprise from rapid technology changes.

Gathr Key Features

Differentiating value propositions and beneficial outcomes

	Visual Application Development and Monitoring
	Real-time Dashboards
	Multi-tenancy Support
	High-speed Data Ingestion
	Elastic Scaling
	Flexible Data Parsing
	Rule-based Alerts
	Pluggable Workflow Management
	Real-time Index and Search
	High Fault Tolerance and Data Integrity
	High Performance Optimization

It’s important to choose a platform that will supply a proven and pre-integrated, performance-tuned stack, ease of use, enterprise-class reliability and flexibility to protect the enterprise from rapid technology changes.

Maybe the most important reason to evaluate this technology now is that a company’s competitors are very likely implementing enterprise-wide real-time streaming analytics right now and may soon gain significant advantages in customer perception & market-share.