IBM Analytics
White Paper

October 2015

Top industry use cases for stream computing

Enable rapid actions in context with IBM Streams

Introduction

It's no secret: data is growing, fast-moving and varied. Cloud, mobile and social events—not just transactions—are energizing organizations across multiple industries and present an enormous opportunity to make businesses more agile, more efficient and more competitive—if they can act fast enough. Now is the time to move beyond managing data to acting on continuous data streams.

Many line-of-business decision-making tools require data to be recorded on a storage device before analytic queries can be run. Even advanced trickle-feed warehouse updates take between five minutes and two hours before data is available to users.

IBM® Streams effectively closes this gap with data stream analytics for your unique use case and industry. Using Streams, organizations can spot risks and opportunities in high-velocity data. These insights often can only be detected and acted on at a moment's notice. High-velocity flows of data from sources such as market data, the Internet of Things, mobile, sensors and clickstream remain largely un-navigated. It's time to unlock this data to optimize decisions, improve business insights and accelerate responses to critical events.

IBM delivers context-aware stream computing to continuously integrate and analyze data in motion and understand the context of everything from people to machines. Organizations leverage this insight to enhance and create more accurate analytical models and fuel cognitive systems.

In short, Streams delivers the continuous, complete, and connected solution needed for the current competitive environment and the future:

Continuous

- Analyzes data streams continuously
- Brings the analytics/algorithms to the data stream

Complete

- Can build, deploy and score models on Streams
- Integrates data streams into existing data and analytic environments

Connected

- Connects to any data source on the planet
- Scales up and down as data streams change
- Highly available and reliable

Streams capabilities

Streams, which began as a collaboration between the US government and IBM, is designed to deliver breakthrough capabilities that enable aggressive analysis and management of information and knowledge. Streams is highly efficient, using 14.2 times fewer hardware resources and delivering 12.3 times more throughput compared to open source offerings.[1]

A key component of the IBM Analytics Platform, Streams processes huge volumes and varieties of data from diverse sources while achieving extremely low latency. This enables decision makers to extract relevant information for timely analysis. Streams radically extends the state of the art in information processing by helping organizations to:

- Continuously analyze data in motion across multiple sources to deliver actionable insight
- Connect to any data stream to make predictions and discoveries as data arrives to enhance and improve analytic models and cognitive systems
- Deploy a complete set of streaming analytics such as natural language processing, geospatial, predictive and more to satisfy unique, industry-specific requirements and use cases
- Speed time to value with open source technologies through the use of Java APIs and visualize data easily with drag-and-drop development tools that support faster time to deployment and effective production management
- Detect and respond to critical events 10 times faster compared to traditional techniques to stay ahead of the competition, analyzing millions of interactions or events per second
- Administer with both web-based and command-line interfaces for optimal enterprise management strategies
- Integrate with Spark and Hadoop to capture data streams efficiently and use them alongside data at rest within the organization
- Gain confidence in data stream insights through integrated data lineage and governance

Choose cloud or on-premises deployment to help manage costs and further speed deployment. While certain research, open-source and commercial initiatives try to address these technical challenges individually, Streams is designed to simultaneously address all of them. This integrated approach enables Streams to break through a number of fundamental barriers. In fact, Forrester named IBM a leader in the Forrester Wave for Big Data Analytics Platforms.

Continuous, high-speed analysis through stream processing

Stream processing extends traditional approaches to information processing, such as transactional or complex event processing (CEP) systems. Traditional processing involves running analytic queries against historic data. For example, if an organization provides an app that calculates the distance users have walked each month using global positioning system (GPS) location data, the traditional approach would make one calculation using all the data transmitted from the user's mobile device and stored over the past month. In contrast, stream processing performs continuous analytics to keep running totals that are updated moment by moment, as GPS data is refreshed.

In the first case, questions are answered using historic data; in the second case, answers are continuously updated using streaming data (see Figure 1). Stream processing differs from a simple in-memory database in that traditional analytics systems must first load all the data—even if in memory—and then run a query. Streams augments the traditional analytics approach by enabling continuous analysis to be modified over time. For example, it can use machine-learning techniques to determine the most popular walking routes or detect deviations from a user's standard walking patterns.

To support this fundamentally different approach to continuous analysis, Streams offers a distributed runtime platform, a programming model and tools for developing continuous analytic applications. The data streams that Streams consumes may originate from satellites, sensors, cameras, news feeds, stock tickers or a variety of other sources, including traditional databases as well as Apache Hadoop systems.

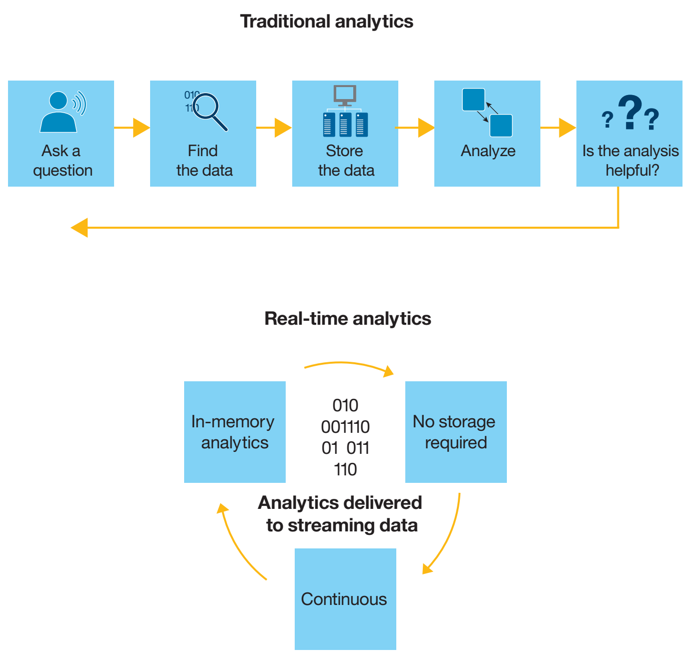

Figure 1: Comparison of analysis based on historic data versus streaming data

The latest Streams update, v4.1, focuses on developers. It delivers faster streaming application delivery by allowing the creation of streaming applications in Java. A developer with no prior Streams knowledge can create Streams applications in under an hour using Java APIs for streaming analytic libraries such as natural language processing, spatial, temporal, acoustic, image recognition and more. It also allows for more intelligent application through integration with open source technologies. Data streams can be captured efficiently and used alongside other data at rest within Hadoop. In addition, Streams which works well for event-driven low-latency apps, and can be used as a dynamic pair with Spark, which works well for data at rest. Finally, confidence in data stream insights can be increased through the creation of data lineage and flexible schema use to help with the ingestion of data. Moreover, integration with the IBM InfoSphere® Data Governance catalog allows for automatic schema discovery and mapping.

Use cases for stream processing

Over the past several years, organizations have developed hundreds of applications for Streams. These applications are used in diverse industries such as telecommunications, government, financial services, healthcare, energy and utilities, insurance and automotive, among many others (see Figure 2).

Telecommunications

The challenge of closing technology and business gaps has been especially critical for cellular service providers. They must deal with the increasing volumes of data generated by embedded chips in cell phones, which enable email, texting, pictures, videos and information sharing using social sites such as Facebook. For each of these transactions, cellular phone switches emit two call detail records (CDRs) to avoid data loss. The CDRs must be deduplicated before processing by the billing support systems. These rising data volumes have made mediation of CDRs difficult to perform in a timely manner.

Telephony
- CDR processing
- Social analysis
- Churn prediction
- Geomapping

Transportation
- Intelligent traffic management
- Automotive telematics

Energy and utilities
- Transactive control
- Phasor measurement unit
- Downhole sensor monitoring

Health and life sciences
- ICU monitoring
- Epidemic early warning system
- Remote healthcare monitoring

Natural systems
- Wildlife management
- Water management

Law enforcement, defense and cybersecurity
- Real-time multimodal surveillance
- Situational awareness
- Cybersecurity detection

Stock market
- Impact of weather on securities prices
- Market data analysis at ultra-low latencies
- Momentum calculator

Fraud prevention
- Multi-party fraud detection
- Real-time fraud prevention

eScience
- Space weather prediction
- Transient event detection
- Synchrotron atomic research
- Genomic research

Other
- Manufacturing
- Text analysis
- ERP for commodities

Figure 2: Streaming analytics have a wide range of benefits in all types of industries.

At the same time, number portability allows subscribers to move to a competitor at any time; some sophisticated users even switch between multiple providers at different times of the day to take advantage of promotions. To prevent customer churn, providers must not only reduce the processing window for CDRs to near-instantaneous, but also concurrently perform analytics to predict which customers may leave. With immediate insight into customer behavior, providers can take action to keep customers before they switch to a competitor.

An agile programming model based on high-level, declarative statements enables the Streams platform to generate complex machine code with a few simple commands. For example, developers would need to write roughly seven lines of Streams Processing Language (SPL) to read in-video frames, detect faces, parse images down to the faces and store only the faces. This approach enables telecom providers to handle their huge volume of CDRs with very low latency while analyzing the data at the same time. One IBM client uses Streams to process CDRs at a peak rate of over 1,500,000 records per second, with end-to-end latency of under one minute to persist the CDRs into a warehouse. Each CDR is checked against billions of existing CDRs to eliminate duplicates rapidly, effectively cutting in half the amount of data stored in databases.

This example illustrates a key Streams use case for applications in a wide range of industries: simultaneous processing, filtering and analysis of continuous data streams. High availability, automated fault tolerance and recovery—along with dashboard summaries—enhance IT operations. Instantaneous analysis of CDRs for revenue assurance and churn prediction enables improved business operations. Other providers use key insights from data stream analysis to feed campaign management systems to increase retention and average revenue per user.

Government

A key strength of Streams lies in its ability to perform analytics on data-intensive streams to identify the few items or patterns that merit deeper investigation. Cybersecurity is one example of this use case.

Botnets represent a significant threat to Internet-connected devices. A botnet is essentially a network of software agents, or bots, that run autonomously and automatically and can be overseen by command-and-control (C&C) servers. An underground economy has sprung up to provide per-pay access to botnets for criminal activity. These botnets are evolving rapidly through techniques such as fast fluxing and encryption, making them very difficult to detect. The Shadowserver Foundation, which tracks known bots, estimates that there are thousands of C&C servers and tens of thousands of bots.[3]

In the proof-of-concept model for a large US government agency, Streams analyzed over 100 MB per second of IP traffic and over 10 million Domain Name System (DNS) queries per hour to generate fast-fluxing botnet alerts. To create these alerts, Streams leveraged not only machine-learning models, but also models built using IBM SPSS® Modeler from historic data stored in IBM PureData® System warehouses. When attack patterns change, Streams issued requests to update the detection models that can be used to mine and analyze historical data. In this way, Streams helps organizations not only analyze data as it streams in, but also use what they learn to extract more intelligence from stored data.

Of course, analytics on data-intensive streams isn't just important for cybersecurity. Streams is being used by the University of Alberta to analyze over 10,000 data points per second collected by 500 sensors located throughout the forests. Some of the six variables being looked at are even captured up to 20 times per second. The result is that a better understanding of even subtle changes in ecosystem health can be detected and patterns in environmental response to climate change can be investigated.

Financial services

Many segments of the financial services industry rely on rapid analysis of large data volumes to make near-instantaneous business and trading decisions. Today, these organizations routinely consume market data at rates exceeding one million messages per second—twice the peak rates they experienced only a year ago.[4] This dramatic growth in market data is expected to continue for the foreseeable future, outpacing the capabilities of many current technologies. Industry leaders are extending and refining their strategies by including other types of data in their automated analyses; sources range from advanced weather prediction models to broadcast news.

IBM developed a Streams-based trading prototype running on 11x86 blade computers. The prototype is designed to host scalable trading applications capable of processing Options Price Reporting Authority (OPRA) data feeds at 21 times the recorded rate.[5]

Healthcare

Stream processing can help improve medical analysis while reducing the workload on nurses and doctors. Healthcare providers can analyze privacy-protected streams of medical-device data to detect early signs of disease, identify correlations among multiple patients and measure efficacy of treatments. There is a strong emphasis on data provenance in this domain—tracking how data is derived and how it changes as it flows through the system.

IBM and the University of Ontario Institute of Technology collaborated on a first-of-a-kind project that uses Streams to monitor premature babies in a neonatal unit, enabling early warnings that allow caregivers to proactively attend to potential complications. For example, data collected from a hospital in Toronto, Canada, helped caregivers detect infections in premature babies up to 24 hours before they exhibited symptoms. In the US, remote telemetry from a hospital has been operational using the same analytic routines.

Other hospitals, such as UCLA Medical Center and the Emory University Hospital, are using Streams to reduce the number of false-positive alarms in intensive-care units. Moreover, the University of Montana is looking at seizure prediction and the Irish Centre for Fetal and Neonatal Translational Research (INFANT) is using machine learning algorithms on streaming EEG waveform and vital-sign data to detect health events in infants. The Southern Ontario Smart Computing Innovation Platform (SOSCIP) is even using Streams to help Dr. Carolyn McGregor's Artemis Project analyze the roughly 2,000 breaths and 7,000 heartbeats infants have each hour to predict complications and intervene.

Energy and utilities

Traditional business models for the utilities industry are losing relevance. The energy production and delivery industry is placing more smart sensors and meters along production, transmission and distribution systems to get granular data streams about the current state of faults and load. Powerful analytics on this data, when combined with other sources such as Outage and Distribution Management Systems (OMS/DMS), weather data, third-party event monitoring systems and Meter Data Management Systems (MDMS) can help utilities take necessary actions to avoid electric grid failures, improve security and optimize capacity. Streams enables these predictive analytics, allowing energy and utility providers to capture and analyze data—all the time, just in time.

Streams supports smart grids with applications for outage detection and prediction, condition-based maintenance and load shedding. Energy and utility providers can monitor grid/plant elements, smart meters and networks to rapidly predict and analyze data. Extremely powerful analytics can use both structured data streams from smart meters and unstructured data—for example, satellite imagery feeds or weather forecasts—and apply it to a variety of uses such as price fluctuation forecasting, energy trading insights and more. Many energy and utility providers have experienced the benefits of optimized energy usage and reduced outages. For example, the Pacific Northwest Smart Grid project services 60,000 homes across five states. It enables towns to avoid power outages using a two-way advanced meter system. The solution also empowers consumers to make educated choices about how and when to use electricity. Another provider, Consolidated Communications Holdings, Inc. uses predictive insights to save USD 300,000 per year.[6]

Insurance

Changes facing insurance providers such as deregulation, increased competition, advances in technology and globalization combine to exert substantial pressure on insurers, brokers, asset managers and reinsurers, and on their ability to respond to these changes. Streams turns these pressures into opportunity and enables predictive analytics of data in motion for rapid decisions.

Streams provides telematic data stream analysis and dashboards of behaviors such as car speed and locations to automatically adjust risk scores. It speeds fraud detection so insurers will receive incident reports as they happen and immediately feed into claims processes.

With the ability to predict accidents or disasters rapidly, dynamically update risk models and ensure informed underwriting, Streams enhances cargo protection. Lastly, with call-center optimization, clients get a better experience, while insurers can automate next best actions and increase automated responses.

Streams helps increase services for insurance clients and decreases cost and fraud for providers. For example, one client is able to model risk of natural disasters throughout the day instead of quarterly. An international port on the Pacific can identify illegal cargo quickly. And an insurer delivers customized services based on simple “utterances” from clients, rather than full sentences or specific commands.

Automotive

Factors such as increased globalization, consumer demand for more innovative and sustainable vehicles, self-driving and connected cars, and growing regulatory and environmental requirements are putting unprecedented pressure on existing business and manufacturing models. One result is a massive increase in data. In fact, some plug-in hybrid vehicles generate 25 GB of data in just one hour.[7] Streams helps organizations transform the industry by enabling predictive analytics of data in motion for rapid decisions, allowing the automotive industry and its ecosystem to capture and analyze data as it arrives.

For example, Streams helps create a more profitable aftermarket for services and products with targeted offers based on driving preferences such as sound systems, child safety equipment and entertainment. It also enables a more interactive and safer driving experience with connected cars, which can deploy brakes automatically, operate windshield wipers dynamically, deploy airbags based on passenger weight or send offers for nearby businesses. Streams can also be used to create a more integrated vehicle data experience, securely sharing data across third parties such as insurance companies, retailers and emergency medical services. Because Streams can detect problems sooner, it can help predict breakdowns and ensure parts are in stock to keep customers satisfied.

Many companies such as Ford, Peugeot and General Motors are using streaming data to optimize operations, improve the driving experience and create safer roadways.

Streaming analytics in other industries

Environmental monitoring and control: Wildfire detection, water flow monitoring

Energy and utilities: Synchrophasor monitoring of smart grids, prediction of wind-power generation

eScience: X-ray diffraction using synchrotrons

Insurance, banking, social services and other government agencies: Fraud prevention, deep-packet inspection for law enforcement

Architectural overview of Streams

The Streams architecture represents a new paradigm in computing system organization and capability. While it has some similarity to CEP systems, it supports higher data analysis rates and more data types. Streams outperforms Apache Storm by 2.6 to 12.3 times in terms of throughput while simultaneously consuming 5.5 to 14.2 times less CPU time.[8] Furthermore, the throughput and CPU time gaps widen as data volume, degree of parallelism and/or number of processing nodes grows—the tests in this study went up to only four nodes and eight-way parallelism. Streams also provides infrastructure support to address scalability and dynamic adaptability through features such as scheduling, load balancing and high availability. Many deployment models exist; organizations can choose on-premises, cloud, or a mixture of both.

A compiler converts Streams applications, written in high-level declarative SPL code or through Java APIs, into machine language. The compiler also detects stateless and stateful operators, enabling the use of multiple threads on multi-core computers. Moreover, it can incorporate performance information from prior executions to make enhanced decisions on which operators to fuse into processing elements. These advanced capabilities not only facilitate developer agility to solve the many-core programming challenge, but also enable outstanding performance and very low-latency processing.

Individual operators interconnect and operate on one or more data streams, which normally come from outside the system or can be produced internally as part of an application. Figure 3 shows how multiple sources and varieties of streaming data related to smart meters—including electricity use, customer contracts and current weather—can be cleansed, filtered, analyzed, modeled, classified or fused.

Figure 3 demonstrates how streaming data sources from outside Streams can make their way into the core of the system, be analyzed in various ways by different pieces of the application, flow through the system and produce results. These results can be used in a variety of ways—for example, to display on a dashboard, drive business actions, or store in enterprise databases or warehouses for later historic analysis.

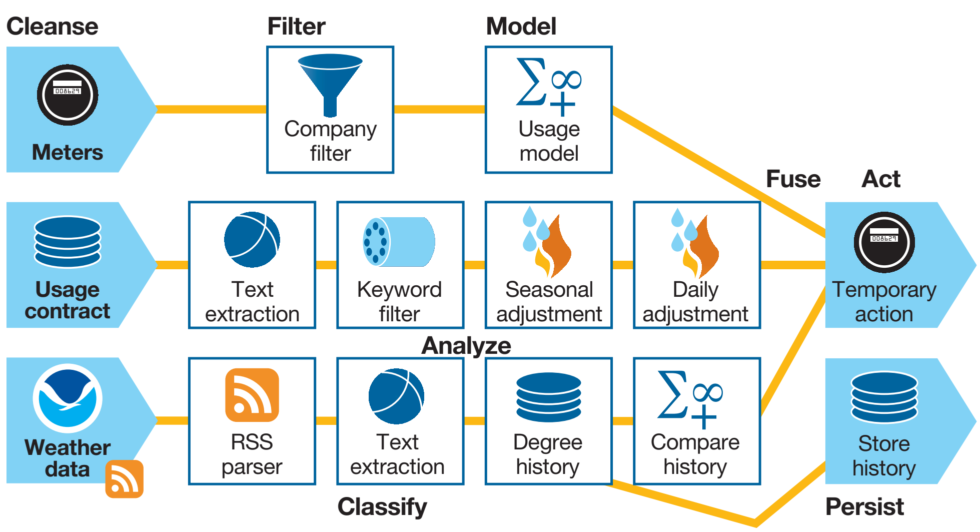

Figure 3: Example of streaming data sources associated with smart meters.

The Streams infrastructure supports one or more computing nodes based on the x86 or IBM POWER® architecture. As shown in Figure 4, data from input streams representing diverse types and modalities flow into the system. Management services communicate over a physical transport layer using Ethernet or InfiniBand networks. The management services continually monitor performance of each operator, processing element, job and node during runtime to optimize job deployment. Saved performance information is especially important during the application development phase, when it is used by the compiler to optimize operator fusion.

Figure 4 also shows that multiple jobs can be added dynamically to Streams during runtime; in this case, another smart meter job has been added. Data streams can be imported or exported between concurrently executing jobs. This flexibility creates tremendous agility for organizations, allowing the addition of jobs to analyze data in new ways. To meet growing capacities, new input streams, output streams and nodes can be dynamically added or removed from the Streams runtime environment without restarting the system.

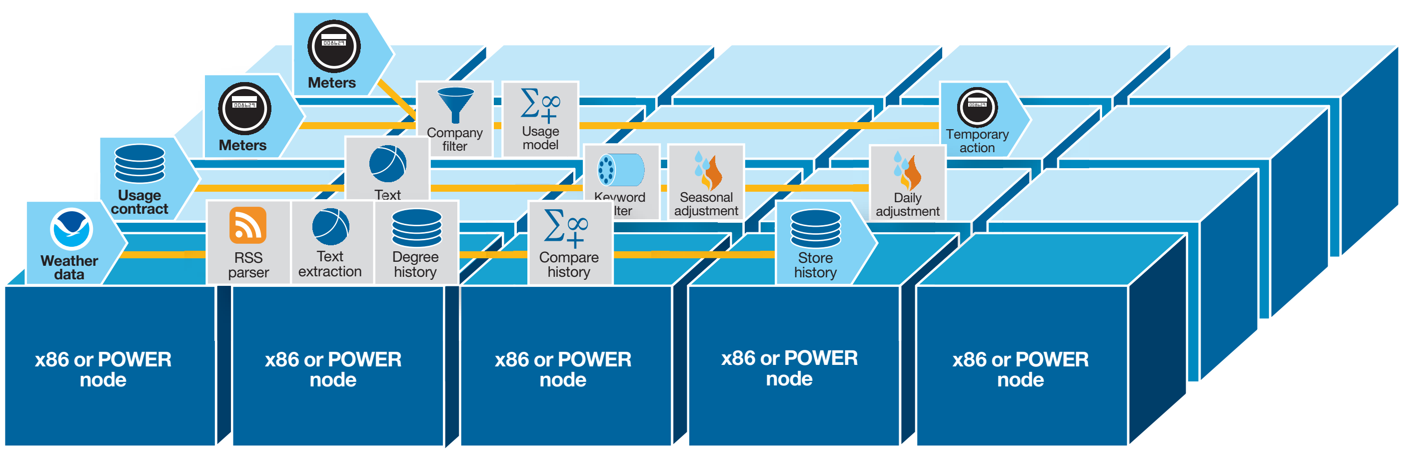

Figure 4: Typical Streams runtime deployment of a streaming application

Streams offers multiple methods for operating on the streaming data:

- Developers can create applications using Streams Studio, an Eclipse-based integrated development environment (IDE). Visual development with pop-up windows for completing parameters speeds learning and simplifies maintenance. End users can also program low-level application components that can be interconnected through streams, as well as specify the nature of those connections. Each component is typed so that other components can later reuse or create a particular stream. Existing Java or C++ analytic routines can be incorporated into Streams applications to reuse existing algorithms and potentially shorten development time.
- End users can deploy existing data-mining scoring models in Streams applications for data stream insights, as opposed to running those models on persistent, or stored, data. In addition, Streams applications can detect the need for a new model, trigger the generation of an updated model and deploy the new model in a running application, as highlighted in the cybersecurity use case.
- End users can log onto Streams using a Microsoft Excel extension, Streams for Microsoft Excel. Streaming data then flows into Excel, allowing the full function of Excel to analyze the data. Macros, mathematics, statistics and even charting can be performed against the streaming data.

These features are supported by the underlying runtime system. As jobs are submitted, the Streams scheduler determines how to best meet the requirements of both newly submitted and already executing specifications, and the job manager automatically effects the necessary changes. During runtime, it continually monitors and adapts to the state and utilization of its computing resources, as well as users' information needs and the availability of data to meet those needs.

Results from the running applications are acted upon by processes, such as web servers, that run outside of Streams. For example, an external application might use TCP connections to receive an ongoing data stream for visualization on a map, or it might alert an administrator to anomalous events.

Actionable insights from data stream sources

Streams has demonstrated far-reaching successes with hundreds of installations worldwide. Some installations report that they have been processing applications continually without a single reboot or failure for over two years. IBM continues to heavily invest in this growth area and make this technology more powerful, scalable, secure, relevant, and interoperable with existing information infrastructures.

For more information

To learn more about Streams, please contact your IBM marketing representative or IBM Business Partner, or visit: **ibm.com**/streams

Additionally, IBM Global Financing can help you acquire the software capabilities that your business needs in the most cost-effective and strategic way possible. We'll partner with credit-qualified clients to customize a financing solution to suit your business and development goals, enable effective cash management, and improve your total cost of ownership. Fund your critical IT investment and propel your business forward with IBM Global Financing. For more information, visit: **ibm.com**/financing

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
October 2015

1 “Of Streams and Storms—A Direct Comparison of IBM InfoSphere Streams and Apache Storm in a Real World Use Case.” IBM white paper (**ibm.com/**services/forms/signup.do?source=sw-infomgt&S_PKG=ov24826&S_CMP=is_bdwp36)

2 The Forrester Wave: Big Data Streaming Analytics Platforms, Q3 2014. Forrester Research. (**ibm.com/**services/forms/signup.do?source=sw-infomgt&S_PKG=ov25865&S_CMP=is_bdanalyst29)

3 Shadowserver Foundation. (www.shadowserver.org/wiki/pmwiki.php/Stats/Statistics)

4 “Stream computing for financial services.” (**ibm.com/**smarterplanet/ie/en/banking_technology/nextsteps/solution/S198762Q80085L60.html)

5 Park, Yoonho; King, Richard; Nathan, Senthil; Most, Wesley; Andrade, Henrique. IBM T.J. Watson Research Center. *“Evaluation of a High-Volume, Low-Latency Market Data Processing System Implemented with IBM Middleware.”* 2010.

6 Consolidated Communications Holdings. IBM case study. (**ibm.com/**common/ssi/cgi-bin/ssialias?subtype=AB&infotype=PM&appname=SWGE_IM_EZ_USEN&htmlfid=IMC14842USEN&attachment=IMC14842USEN.PDF)

7 Gardner, Dana. *“Ford scours for more big data to bolster quality, improve manufacturing, streamline processes.”* BriefingsDirect on ZDNet.com. Jan. 9, 2013. (www.zdnet.com/ford-scours-for-more-big-data-to-bolster-quality-improve-manufacturing-streamline-processes-700001045)

8 “Of Streams and Storms—A Direct Comparison of IBM InfoSphere Streams and Apache Storm in a Real World Use Case.” IBM white paper (**ibm.com/**services/forms/signup.do?source=sw-infomgt&S_PKG=ov24826&S_CMP=is_bdwp36)

Please Recycle

IMW14704-USEN-02