IBM InfoSphere Streams

Real-time analytics processing for actionable insight

Highlights

- Perform advanced real-time analytics on data in motion

- Rapidly ingest, correlate and continuously analyze a massive volume and variety of structured and unstructured streaming data as it arrives from thousands of real-time sources

- Make real-time predictions and discoveries as data enters the system

- Visualize data easily with drag-and-drop development tools

- Detect and respond to critical events immediately

- Learn and update models for future analysis and trend prediction with cognitive computing

Every day, consumers and businesses generate data at a pace that would have seemed unprecedented just a few years ago. For example, in an average day, AT&T transfers about 30 petabytes of data through its network.[1] Dublin City Centre analyzes 50 bus location updates per second.[2] Twitter handles 340 million tweets daily, mostly from mobile users.[3] By 2016, annual Internet traffic is expected to reach 1.3 zettabytes,[4] and 80 percent of that data will be in unstructured formats.

Faced with this growing volume of constantly changing data, today's organizations are challenged to make informed, real-time business decisions and stay ahead of the competition. However, savvy companies are fast realizing that they can extend the value of their existing systems and generate significant business advantages by updating their data mining methods to support untraditional, unstructured data sources such as audio, video and email. This enables them to react more quickly to changes in customer sentiment, uncover new market opportunities and introduce groundbreaking new products aligned with the latest trends.

As a key enabler for this new generation of analytic processing methods, IBM® InfoSphere® Streams provides a state-of-the-art computing platform that can help companies transform burgeoning data into actionable information and business insights. InfoSphere Streams is a critical component of the IBM Big Data Platform and delivers a highly scalable, agile software infrastructure to perform in-motion analytics on a wide variety of relational and non-relational data types entering the system at unprecedented volumes and speeds—and from thousands of real-time sources. With InfoSphere Streams, organizations can capture and act on key business data just in time, all the time.

A new paradigm for information processing

InfoSphere Streams is the result of pioneering work from IBM Research done in conjunction with the US government. The software provides a development platform and runtime environment that enables organizations to develop and execute applications that ingest, filter, analyze and correlate massive volumes of continuous data streams. These data streams can originate from structured or unstructured data sources and may include a huge variety of digital information, including:

- Text files, spreadsheets, images, video and audio recordings
- Email, chat and instant messages, web traffic, blogs and social networking sites
- Financial transactions, customer service records, telephone usage records, system and application logs
- Data from satellites, GPS tracking, smart devices and network traffic sensors

InfoSphere Streams fuses these heterogeneous data types onto a computing platform that supports powerful data analysis with exceptional performance and impressive response times.

Powerful, real-time analytic processing made simple

InfoSphere Streams addresses a major data challenge—analyzing large volumes of data in motion. The ability to analyze this continuously generated data is often critical for organizations that must react in real time to market alerts or events, or when organizations must filter huge amounts of data and identify the rich, high-value information before moving it into a data warehouse or Apache Hadoop system.

To cost-effectively analyze streaming data, InfoSphere Streams uses cutting-edge technology like its unique Streams Processing Language (SPL). InfoSphere Streams applications can be extended with C, C++ or Java applications to reuse existing logic and speed time-to-value. In addition, the Getting Started feature makes it easy to install, build, configure and manage application instances with just a few clicks. Drag-and-drop visual development helps reduce the learning curve and speed application development.

Key components of InfoSphere Streams

Comprehensive development tools

- Eclipse IDE
- Web console
- Drag-and-drop editor
- Instance graph
- InfoSphere Streams visualization
- InfoSphere Streams Debugger
- Java improvements
- Mapped operators

Scale-out architecture

- Clustered runtime for near-limitless capacity
- Large-scale deployment
- Support for Red Hat Enterprise Linux v5.3 and above, CentOS v6.1 and above
- Support for x86 and IBM Power® multi-core hardware
- Support for InfiniBand and Ethernet
- Support for SUSE Linux Enterprise Server 11.2 and above

Sophisticated analytics with toolkits and accelerator

- CEP, database, finance, IBM InfoSphere Data Explorer, IBM InfoSphere DataStage®, geospatial, Internet, messaging with JMS adapter, mining, R, IBM SPSS®, standard, text, Time Series toolkits
- Application accelerators: Telco, social data analytics

Enterprise integration: Extending the power of InfoSphere Streams

Out-of-the-box integration with other products in the IBM Big Data Platform and built-in XML support gives InfoSphere Streams access to all your data and systems.

- **IBM InfoSphere BigInsights™** lets you store streaming data in an enterprise-class Hadoop environment for additional analysis or historical retention. InfoSphere Streams and InfoSphere BigInsights use the same Advanced Text Analytics capabilities to simplify natural language processing applications for both data in motion and data at rest. In addition, InfoSphere BigInsights can be used to augment streaming sources with contextual information, and users can visualize InfoSphere Streams data in the InfoSphere BigInsights console.
- **IBM InfoSphere Data Explorer** enables users to visualize InfoSphere Streams data in the InfoSphere Data Explorer CXO dashboard and add streaming data to the InfoSphere Data Explorer index.
- **IBM InfoSphere DataStage** helps users perform deep analysis and gain additional insight using contextual and source data from other parts of the infrastructure.
- **Support for XML** allows developers to fuse a broader range of traditional and untraditional data.

Scale-out architecture

InfoSphere Streams software helps organizations extend their current IT investments without a massive infrastructure overhaul. It scales from a single server to a virtually unlimited number of nodes to process data of any volume—from terabytes to zettabytes. InfoSphere Streams provides a clustered runtime environment that can easily handle up to millions of events per second with microsecond latency. Actionable results can be achieved with near-zero latency.

The Advanced Compiler fuses parts of the application for increased performance and can distribute parts of the application to run across multiple hardware nodes. It also supports multiple high-speed transports, including Ethernet and InfiniBand. Existing applications can be dynamically extended with new applications to read the same streams of data, further leveraging current investments.

A web-based management console makes it easy to configure and manage the runtime and applications, including automatically placing features and deploying application components. Applications and their individual elements can be monitored for status and performance metrics to help ensure the company attains its service-level agreements.

Comprehensive tools for an agile development environment

With InfoSphere Streams, developers and administrators can easily build applications to analyze high-volume, high-velocity and high-variety data in real time.

InfoSphere Streams Studio, an Eclipse-based integrated development environment (IDE), supports rapid development with editors, wizards, application structure graphs and runtime monitoring to simplify the process of building and managing InfoSphere Streams applications. These tools are designed to help organizations quickly and easily extend their applications' capabilities with sophisticated real-time data analytics:

- **InfoSphere Streams Debugger** provides an interactive debugger for debugging SPL applications that focuses on examining the flow of data in and out of SPL operators.
- **The drag-and-drop graphical editor** allows users to build applications while automatically synching graphical and SPL source code views. This round-tripping process lets developers create an application via text or graphical editor, and update later using the other editor.
- **An instance graph**, available in both InfoSphere Streams Studio and the management console, provides visual monitoring of application health and metrics and lets users quickly identify issues using customizable views.
- **The latest InfoSphere Streams data visualization capabilities** allow users to dynamically add new views to running applications with charts provided out-of-the-box.

Sophisticated analytics with toolkits and accelerators

InfoSphere Streams features integrated toolkits and sample applications that facilitate the development of solutions for particular industries or functions. These include the Data Mining Toolkit, Financial Services Toolkit and Standard Toolkit, which encompass an array of the most commonly used operators for smooth integration with existing infrastructures and applications.

For example, a financial services organization can use InfoSphere Streams to combine incoming stock market quotes with industry data or information on interested traders that is held on an internal database. Adding context to stock market ticks helps the organization achieve more complex levels of analysis and customized alerting than ever before. In addition, InfoSphere Streams helps determine, in real time, which data should be stored in a database for future analysis—reducing storage and administration costs.

InfoSphere Streams: System requirements

Hardware requirements

- Intel/AMD x86 architecture (64-bit) or IBM POWER7® architecture systems (minimum 500 MB memory)

- 2 GB memory to run simple applications

Software requirements

- Red Hat Enterprise Linux (RHEL) Version 5.6 or later, or Version 6.1 or later for x86 architecture hardware

- RHEL Version 6.2 and above for IBM POWER7 architecture hardware

- CentOS Version 6.1 or later for x86 architecture hardware

- SUSE Linux Enterprise Server (SLES) V11.2 or above for x86 architecture hardware (SLES does not support Security Enhanced Linux)

- Eclipse Platform SDK v3.6.2, v3.7.2 or v3.8.0

- IBM Java SE v6.0-9.0 SDK for x86 or IBM POWER7 architecture hardware

- Oracle Java SE v6.0.x SDK for x86 architecture hardware

- Mozilla Firefox v10 and above

- Microsoft Internet Explorer 9.0 and above

Other toolkits in InfoSphere Streams include:

- **Complex Event Processing (CEP),** which offers high-performance and rich analytics by using patterns to detect composite events in streams of simple events. Existing applications are easily integrated and can be migrated into an InfoSphere Streams environment to take advantage of increased scalability and the ability to process up to 10 times more events per second on the same hardware.
- **InfoSphere Data Explorer,** which brings visualization and storage capabilities to InfoSphere Streams by enabling users to pass visualization widgets from InfoSphere Streams to InfoSphere Data Explorer, or pass data to InfoSphere Data Explorer for indexing and later retrieval.
- **InfoSphere DataStage** allows users to perform in-depth data analyses as part of the information integration flow and see rapid results. An InfoSphere Streams toolkit provides adapters that exchange data between InfoSphere Streams and InfoSphere DataStage for real-time analytics processing.
- **IBM SPSS** leverages the SPSS Modeler to develop and build predictive models, and then deploy them using the SPSS Scoring Operator. SPSS models can be dynamically refreshed in InfoSphere Streams without suspending InfoSphere Streams applications.
- **Geospatial** is often used in the transportation and telecommunications industries. It enables location-based marketing based on high-performance analysis and processing of geospatial data.
- **Messaging** allows InfoSphere Streams to receive data from or send data to IBM WebSphere® MQ.
- **Time series** is a rich set of tools that includes data generation (synthesizing or extracting), preprocessing (preparation and conditioning), analysis (statistics, correlations, decomposition and transformation), cognitive computing and modeling (prediction, regression and tracking).

InfoSphere Streams also includes application accelerators that help improve time-to-value and leverage best practices around a specific solution implementation or use case. They provide business logic, data processing and UI/visualization capabilities—analytic operators or sample applications that can be used as design templates for new applications.

The IBM accelerator for social data analytics provides rapid implementation to analyze large volumes of various types of social media data with real-time processing. This helps organizations gain vital insight for effective, highly targeted marketing campaigns and timely product and marketing decisions. It also facilitates insights that support competitive intelligence efforts, and building customer retention and new-customer acquisition programs.

The IBM accelerator for telecommunications event data analytics provides transformation and analytics for telephone companies' call and event-detail records, which can help with real-time revenue assurance and fraud detection. This helps telecommunications providers strengthen billing insights based on services, vendors and business lines. It also supports companies' efforts to differentiate services, strengthen customer loyalty, reduce churn, offer targeted services, personalize billing and improve the customer experience.

InfoSphere Streams V3.1: Features and enhancements

The latest version of InfoSphere Streams includes a wide range of new features and functionality enhancements to support better performance, development and analytics, administration, and integration.

Enhanced performance

Version 3.1 significantly increases the performance of Java programs running on InfoSphere Streams, yielding faster throughput and reduced resources for fused Java operators. Additional performance improvements for operators involving maps and lists enable 2 to 10 times runtime improvement.[5]

Enhanced development and analytics

InfoSphere Streams now supports R analytics, extending the family of analytic options. Developers can use R scoring in a native format, simultaneously leveraging the scalability, parallelism and partitioning of InfoSphere Streams to achieve exceptional throughput.

New operators broaden the capabilities of the Time Series toolkit, enabling real-time and predictive analysis of regularly generated data from numerous sources. Existing operator functionality has also been expanded, along with added usability improvements and the introduction of control data.

Enhanced administration

Several new features make InfoSphere Streams V3.1 easier to install, maintain and recover. Red Hat Package Manager support simplifies the installation process and helps reduce the work required to install InfoSphere Streams on multiple servers.

InfoSphere Streams V3.1 also simplifies instance recovery by checking failed services and automatically starting them in the correct order. Similarly, simplified restart of application components enables InfoSphere Streams administrators to move or restart multiple application components on a new node, helping to reduce the amount of time required to maintain an optimal runtime.

In addition, InfoSphere Streams now supports SUSE Linux Enterprise Server (SLES) on x86 hardware as an alternate platform for the runtime and development tools.

Enhanced integration

Version 3.1 features new integrations that allow for easier adoption and implementation. Java Message Service (JMS) support, a standard interface to send and receive messages between two or more systems, enables InfoSphere Streams to communicate with more systems. Representational State Transfer application programming interface (REST API), a public API for management and monitoring, allows callers to get state and metric information for the system and applications. This first release within InfoSphere Streams V3.1 implements a subset of REST APIs for the most popular services. Plus, Teradata and Asterdata support allows the exchange of data between InfoSphere Streams and Teradata and Asterdata databases and warehouses.

Changing the face of big data analytics

With InfoSphere Streams, organizations can see events and trends as they are happening and respond proactively. By bringing meaning to unwieldy, vast and fast-moving data streams, the software has already delivered real benefits to companies in a wide range of industries.

For example, a communications company is currently using InfoSphere Streams to mediate more than 200,000 call-detail records per second and simultaneously create summaries of that data. The summary process formerly took at least 12 hours—but now, with summaries available almost immediately, the company can make critical decisions based on real-time call information. This allows them to target high-value customers with personalized marketing offers based on their current location, when it is still possible to influence the business impact.

In the financial services sector, an InfoSphere Streams–based application analyzes and correlates over 5 million market messages per second to execute algorithmic option trades with an average latency of 30 microseconds—giving a business the agility to take advantage of market trends more quickly than the competition. And in healthcare, InfoSphere Streams is being used to analyze information from multiple sensors in a neonatal intensive care unit, with the goal of being able to detect changes in patients' medical conditions up to 24 hours earlier than before—a decided edge in treating illness successfully.

For more information

Any organization that requires immediate, accurate analysis and business decisions based on up-to-the-minute information can benefit from InfoSphere Streams. To learn more about InfoSphere Streams and how it can help your business gain a competitive advantage, visit: **ibm.com**/software/data/infosphere/streams

[1] Kahn, Mobeen. "The Big Data Boom: How Mobile and Fixed Devices Are Feeding Big Data and What It Means to Businesses," AT&T Networking Exchange blog, July 19, 2012. http://networkingexchangeblog.att.com/enterprise-business/the-big-data-boom/

[2] Dublin City Centre IBM Reference.

[3] "Twitter turns six," Twitter blog, March 21, 2012. http://blog.twitter.com/2012/03/twitter-turns-six.html

[4] Moscaritolo, Angela. "Internet Traffic to Reach 1.3 Zettabytes by 2016," PC Magazine, May 30, 2012. www.pcmag.com/article2/0,2817,2405038,00.asp

[5] Runtime performance increase for operations using bounded lists and maps compared to InfoSphere Streams 3.0.