Information Management software

IBM InfoSphere Streams

Enabling complex analytics with ultra-low latencies on data in motion

Roger Rea, IBM Software Group
Krishna Mamidipaka, IBM Software Group

Contents

Executive Summary

Data volumes are expected to double every two years over the next decade. The global economic slowdown is resulting in organizations seeking to become more nimble with their operations and more innovative with their decisions. In the face of exploding data volumes and shrinking batch time windows, these organizations are struggling to make 'truly' real time decisions and gain competitive advantage. Existing tools and technologies that aid decision making first require data to be recorded on storage device and run queries after the fact to detect actionable insights. Savvy organizations are fast realizing that the time lost in this process leads to missed opportunities that might be the difference between success and failure.

InfoSphere Streams addresses this gap effectively by providing a futuristic technology that can detect insights within data streams still in motion.

Introduction

The goal of the IBM InfoSphere Streams is to provide breakthrough technologies that enable aggressive production and management of information and knowledge from relevant data, which must be extracted from enormous volumes of potentially unimportant data. Specifically, the goal of InfoSphere Streams is to radically extend the state of the art in information processing by simultaneously addressing several technical challenges, including:

- Respond quickly to events and changing requirements
- Continuous analysis of data at rates that are orders of magnitude greater than existing systems
- Adapt rapidly to changing data forms and types
- Manage high availability, heterogeneity, and distribution for the new stream paradigm
- Provide security and information confidentiality for shared information

While certain research and commercial initiatives endeavor to address the above technical challenges in isolation, no program – outside of InfoSphere Streams – attempts to simultaneously address all of them. The primary goal of InfoSphere Streams is to break through a number of fundamental barriers to enable the creation of a system designed to meet these challenges. The project, which began in IBM Research in 2003, has now reached a level of maturity that has permitted it to be demonstrated in a variety of application environments and to embark on a path to be made available as an IBM offering. InfoSphere Streams is currently installed in over ten sites across 3 continents.

Stream Computing

Stream computing is a new paradigm. In "traditional" processing, one can think of running queries against relatively static data: for instance - List all personnel residing within 50 miles of New Orleans, which will result in a single result set. With stream computing, one can execute a process similar to a "continuous query" that identifies personnel who are currently within 50 miles of New Orleans, but get continuous, updated results as location information from GPS data is refreshed over time. In the first case, questions are asked of static data, in the second case, data is continuously evaluated by static questions. InfoSphere Streams goes further by allowing the continuous queries to be modified over time. A simple view of this distinction is as follows:

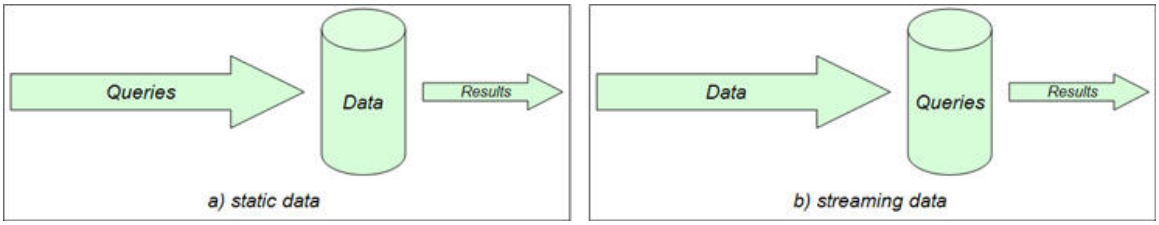

Figure 1: Static data versus streaming data: conceptual overview.

While there are other systems that embrace the stream computing paradigm, InfoSphere Streams takes a fundamentally different approach for continuous processing and differentiates with its distributed runtime platform, programming model, and tools for developing continuous processing applications. The data streams consumable by InfoSphere Streams can originate from sensors, cameras, news feeds, stock tickers, or a variety of other sources, including traditional databases.

Emerging Use Cases

As InfoSphere Streams becomes a generally available offering, a number of applications are being pursued. The following provides a summary of the pilots conducted by IBM, highlighting the types of usage that can be supported by InfoSphere Streams.

Radio Astronomy: A key strength of InfoSphere Streams lies in its ability to perform analytics on data-intensive streams to identify the few items that merit deeper investigation. One example of this use case is in the domain of radio astronomy. There are a number of projects globally that receive continuous streams of telemetry from radio telescopes. For example, these radio telescopes might have thousands or tens of thousands of antennae, all routing data streams to a central supercomputer to survey a location in the universe. The InfoSphere Streams middleware running on that supercomputer can provide a more flexible approach to processing these streams of data. We are working with the low frequency radio astronomy group of Uppsala University and the LOFAR Outrigger In Scandinavia (LOIS) project to develop analytics that identify anomalous and transient behavior such as high energy cosmic ray bursts. We are investigating expansion of this work to a similar effort with the Square Kilometre Array , with total data rates in the range of terabits per second.

Energy Trading Services (ETS): The ETS pilot demonstrates how InfoSphere Streams can support energy trading. The demonstrated system provides energy traders with real-time analysis and correlation of

events affecting energy markets, and allows them to make informed decisions faster than before. Analysis supporting energy traders include various heat maps, energy demand models, technical analysis of energy futures (Bollinger Band, Volume Weighted Average Price, etc.), news feed analysis to identify and evaluate energy-relevant events, and a map view of the predicted impact of a hurricane on the assets of oil companies. The traders can leverage shared computing infrastructure to obtain information quickly and at a low cost. The system also provides context-sensitive guidance that helps the traders select the best available sources and analytics for the task. The pilot uses MARIO (Mashup Automation with Runtime Invocation and Orchestration) to dynamically assemble applications needed by energy traders, deploying and operating the stream processing parts of these applications in a InfoSphere Streams cluster. The set of 250 independent analytics, data sources and configuration descriptions that were built for the ETS pilot are dynamically composed and parameterized in different combinations to create thousands of applications that analyze and present data relevant to energy trading. The demonstrated applications analyze real-time and/or previously recorded data obtained from external sources such as the National Oceanic and Atmospheric Administration (NOAA) and the New York Mercantile Exchange (NYMEX).

Financial Services: Many segments of the financial services industry rely on rapidly analyzing large volumes of data in order to make near-real time 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. 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 analysis; sources range from advanced weather prediction models to broadcast news. IBM and TD Bank Financial Group developed an InfoSphere Streams based trading prototype running on a Blue Gene/P super computer that could host scalable trading applications capable of processing OPRA data feeds sped up 21 times the recorded rate.

Health monitoring: Stream computing can be used to better perform medical analysis with reduced workload on doctors. Privacy-protected streams of medical device data can be analyzed to detect early signs of disease, correlations among multiple patients, and efficacy of treatments. There is a strong emphasis on data provenance in this domain, in tracking how data are derived as they flow through the system. A “First of a Kind” collaboration between IBM and the University of Ontario Institute of Technology will use InfoSphere Streams to monitor premature babies in a neonatal unit.

Manufacturing: IBM is working on a pilot within IBM’s Fishkill semiconductor chip fabrication line, in which InfoSphere Streams performs multivariate monitoring for real-time process fault detection and

classification. In this fashion, when process errors cause defects in manufactured chips, these errors can be detected within minutes rather than days or weeks. The defective wafers can then be potentially reworked prior to ensuing process steps which might render the wafers unusable, and more importantly, adjustments can be made before processing subsequent wafers.

In addition, other use cases of InfoSphere Streams are fast emerging in domains such as environment monitoring and control (wildfire detection, water flow monitoring etc), Smart traffic management, fraud prevention etc.,

Architectural Overview

The InfoSphere Streams architecture represents a significant change in computing system organization and capability. Even though it has some similarity to Complex Event Processing (CEP) systems, it is built to support higher data rates and a broader spectrum of input data modalities. It also provides infrastructure support to address the needs for scalability and dynamic adaptability, like scheduling, load balancing, and high availability.

In InfoSphere Streams continuous applications are composed of individual operators, which interconnect and operate on multiple data streams. Data streams can come from outside the system or be produced internally as part of an application. The following flow diagram shows how multiple sources of varying types of streaming data can be filtered, classified, transformed, correlated, and/or fused to inform equities trade decisions, using dynamic earnings calculations, adjusted according to earnings-related news analyses, and real-time risk assessments such as the impact of impending hurricane damage:

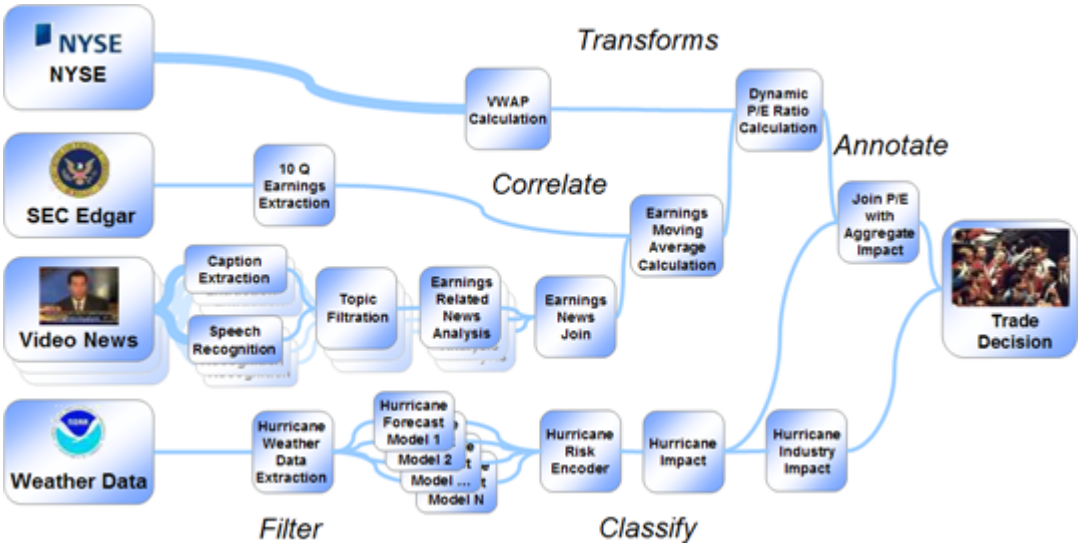

Figure 2: Trading Example.

For the purposes of this overview it is not necessary to understand the specifics of Figure 2 rather, its purpose is to demonstrate how streaming data sources from outside InfoSphere Streams can make their way into the core of the system, be analyzed in different fashions by different pieces of the application, flow through the system, and produce results. These results can be used in a variety of ways, including display within a dashboard, driving business actions, or storage in enterprise databases for further offline analysis.

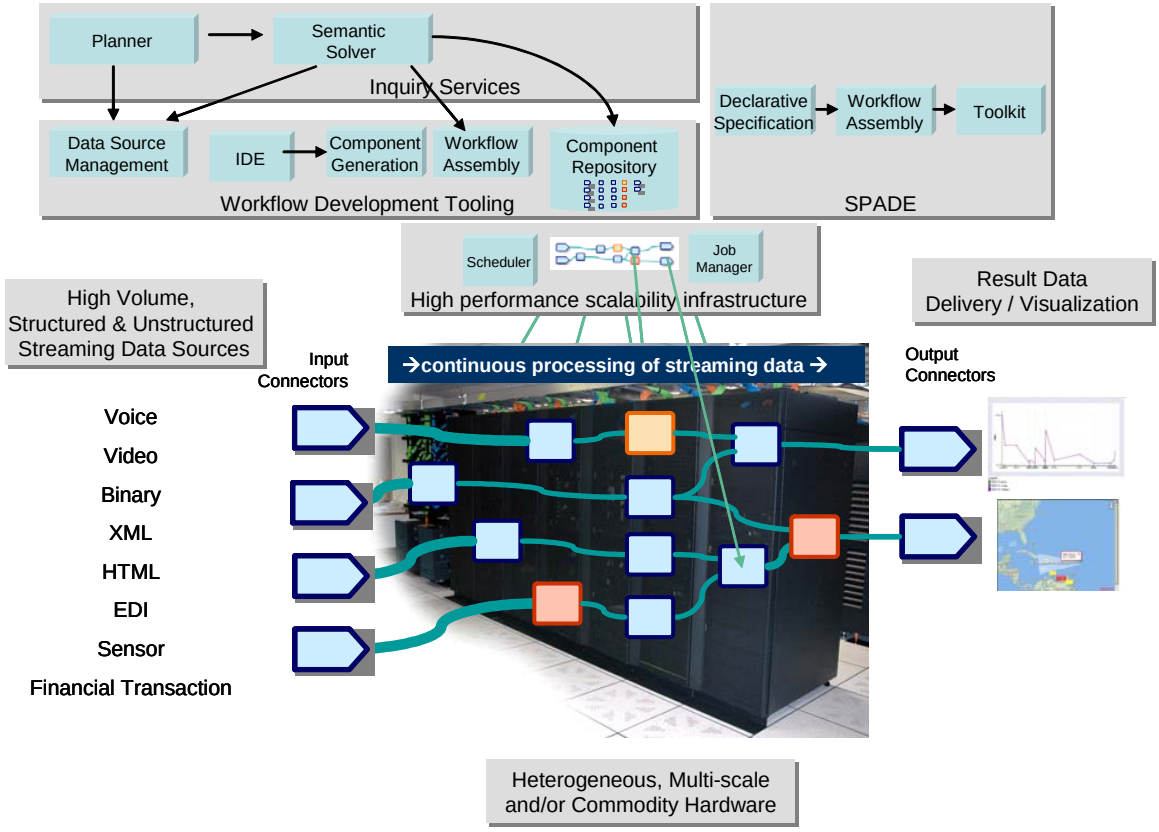

Figure 3: System overview.

Figure 3 illustrates the complete prototype infrastructure. As shown, data from input data streams representing a myriad of data types and modalities flow into the system. The layout of the operations performed on that streaming data is determined by high-level system components that translate user requirements into running applications. InfoSphere Streams offers three methods for end-users to operate on streaming data, as follows:

- The Stream Processing Application Declarative Engine (SPADE) provides a language and run-time framework to support streaming applications. Users can create applications without needing to understand the lower-level stream-specific operations. SPADE provides some built-in operators, the ability to bring streams from outside InfoSphere Streams and export results outside the system, and a facility to extend the underlying system with user-defined operators.

- Users may pose inquires to the system to express their information needs and interests. These inquiries are translated by a Semantic Solver into a specification of how potentially available raw data and existing information can be transformed to satisfy user objectives. The runtime environment accepts these specifications, considers the library of available application components, and assembles a job specification to run the required set of components.

- Users can develop applications through an Eclipse-based Workflow Development Tool Environment, which includes an Integrated Development Environment (IDE). These users can program low-level application components that can be interconnected via streams, and specify the nature of those connections. Each component is "typed" so that other components can later reuse or create a particular stream. This development model will evolve over time to directly operate on SPADE operators rather than the base, low-level applications components, but will still allow new operators to be developed.

All three of these methods are supported by the underlying runtime system. As new jobs are submitted, the InfoSphere Streams scheduler determines how it might reorganize the system in order to best meet the requirements of both newly submitted and already executing specifications, and the Job Manager automatically effects the changes required. The runtime continually monitors and adapts to the state and utilization of its computing resources, as well as the information needs expressed by the users and the availability of data to meet those needs.

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

Summary

In the 6 years since System S first began as an IBM research project that has resulted in InfoSphere Streams, it has demonstrated initial successes with a number of commercial and scientific applications. It provides an infrastructure to support mission-critical data analysis with exceptional performance and interoperability with existing application infrastructures. The anticipated adoption of technologies from IBM Research System S Stream Computing System into the IBM InfoSphere Streams is expected to further increase the scale and diversity of its infrastructure, tools, support, and potential applications.

For more information about InfoSphere Streams, please contact your IBM Marketing Representative or Authorized IBM Business Partner.

For more information

To learn more about IBM InfoSphere Streams and associated products to build them, visit:

http://www.ibm.com/software/data/infosphere/streams/

IBM Corporation
Software Group
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
05-09