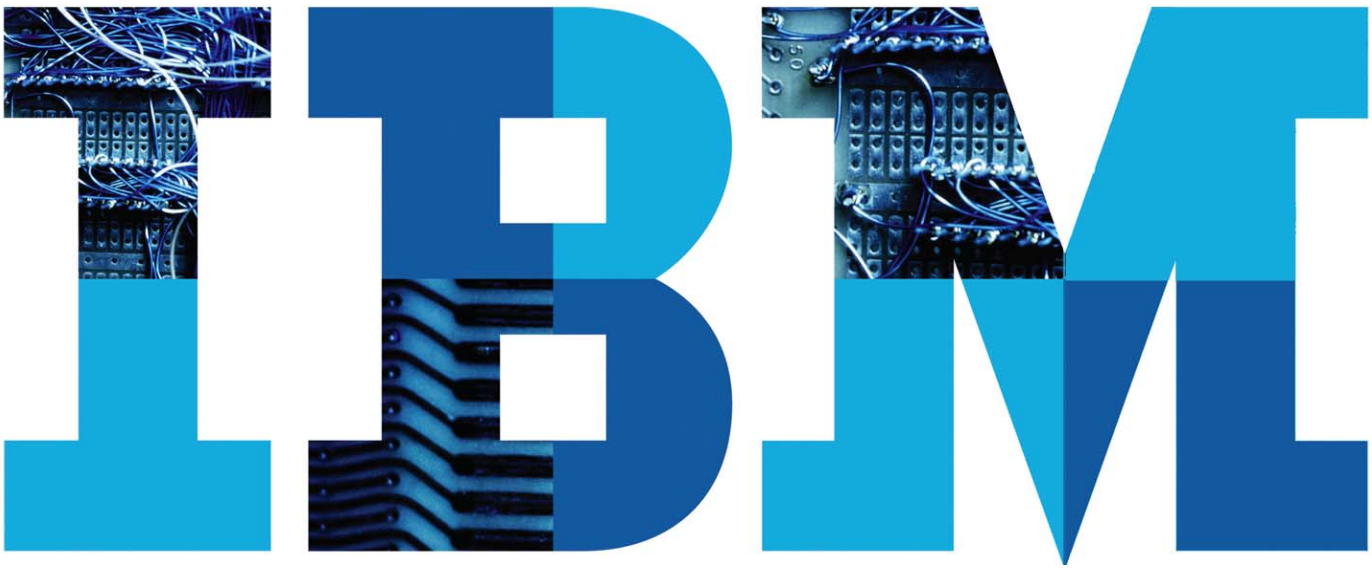
IBM Institute for Advanced Security

White Paper

December 2010

Cybersecurity Analytics for a Smarter Planet

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



Contents

Executive Summary

As the world's critical infrastructures become more instrumented, interconnected, and intelligent, the amount of cybersecurity-related data about those infrastructures has grown at a rate far exceeding the capacity of traditional cybersecurity management systems. This security data about network traffic, access control exceptions, and other unsafe or unsecure conditions was already stressing our systems' ability to use that data.

Facing exploding cybersecurity data volumes, organizations are struggling to make real-time decisions and continue to maintain the integrity and security of their systems. The conventional approach first requires data to be stored in a database and then queries are run after the fact to detect or determine the cause of problems. They are fast realizing that the time lost in this process results in a reactive security stance that is often far too late.

IBM InfoSphere™ Streams can address this challenge by providing a futuristic technology that can extract knowledge from cybersecurity data streams still in motion, allowing prompt protective actions.

Introduction

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

- Respond quickly to events and changing requirements
- Continuously analyze data at rates that are 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 some research and commercial initiatives try to address those technical challenges in isolation, IBM is seeking to address all of them together. The primary goal of stream processing is to break through a number of fundamental barriers to create a system designed to meet these challenges. The project, which began in IBM Research in 2003 as System S, was demonstrated in various application environments and is now available as an IBM offering. The technology is currently installed in dozens of sites across three continents.

Stream Computing

Stream computing is a new standard. 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 results in a single result set. With stream computing, one can execute 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. Stream processing goes further by allowing the continuous queries to be modified over time. A simple view of this distinction is as follows:

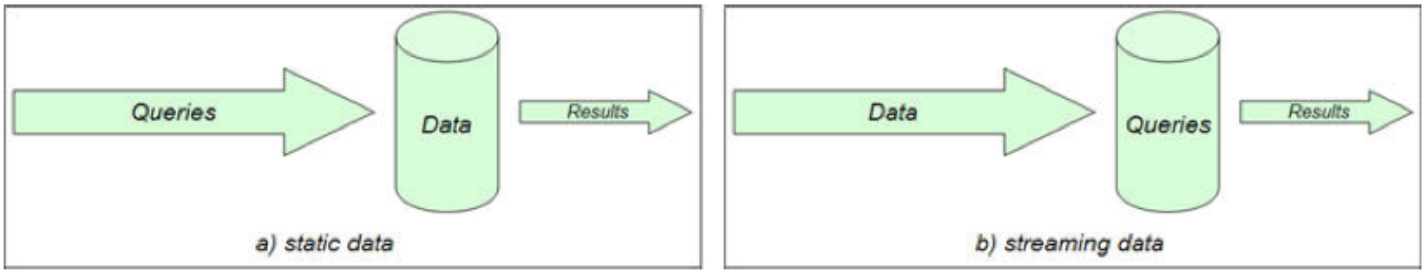


Figure 1: A conceptual overview of static data versus streaming data

While there are other systems that embrace the stream computing model, IBM takes a different approach for continuous processing and differentiates with its distributed runtime platform, programming model, and tools for developing continuous processing applications. These applications rapidly analyze information as it streams from thousands of real-time sources, including sensors, cameras, news feeds, network devices (firewalls or web servers), cybersecurity monitors, or various other sources, including traditional databases.

Architectural Overview

The 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, such as scheduling, load balancing, and high availability.

In stream processing, continuous applications are composed of individual operators, which interconnect and operate on multiple data streams. Data streams can come from outside the system or can be produced internally as part of an application. The following flow diagram shows an example of how multiple types of streaming data can be filtered, classified, transformed, correlated, and used to make an equities trade decision by using dynamic earnings calculations, adjusted according to news analyses, and real-time risk assessments such as the impact of impending hurricane damage.

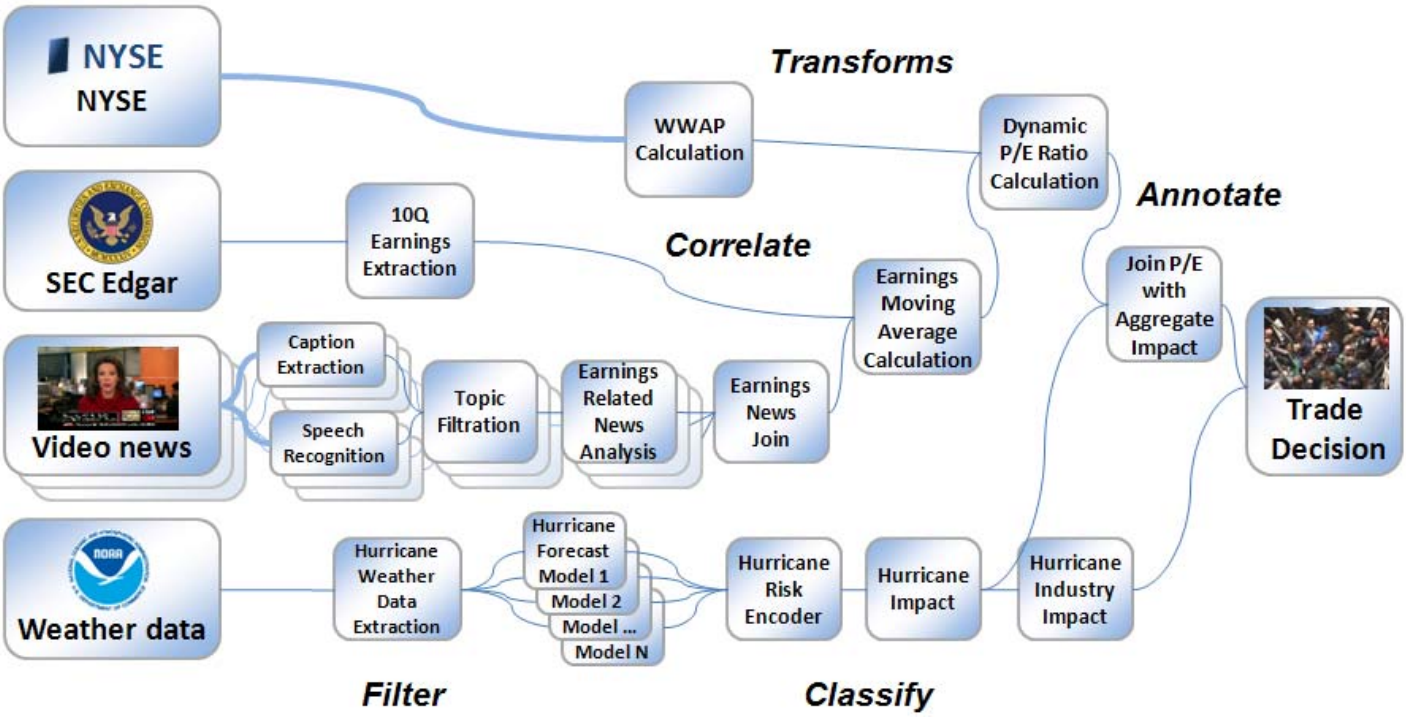


Figure 2: An example of making a trade decision based on continuous data streams

For the purposes of this overview, it is not necessary to understand the specifics of Figure 2; rather, its purpose is to demonstrate how diverse streaming data sources 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 different ways, including display within a dashboard, driving business actions, or storage in enterprise databases for further offline analysis.

Figure 3 illustrates the complete prototype infrastructure. 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.

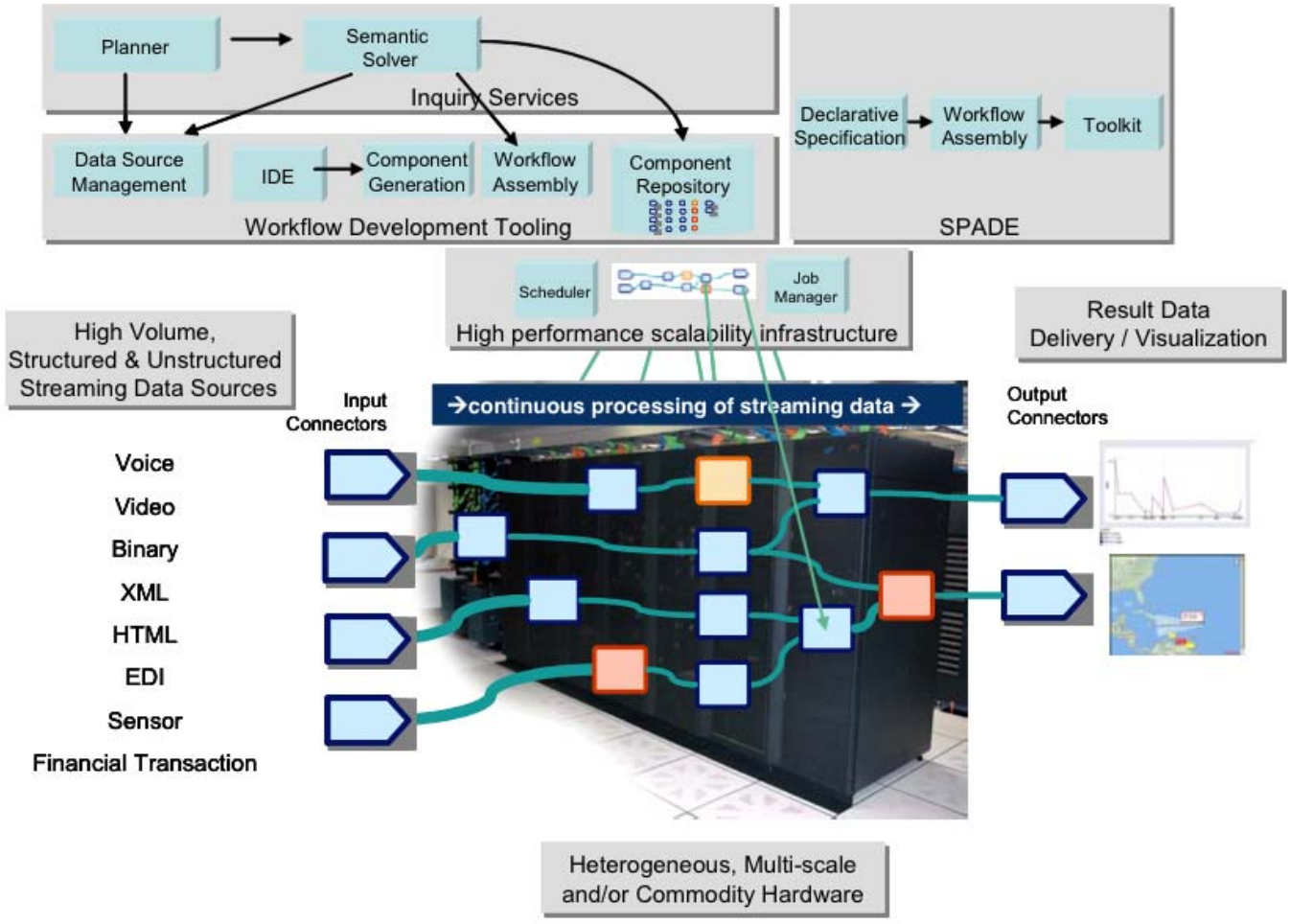


Figure 3: System overview

The system offers three methods for users to operate on streaming data:

- The Stream Processing Application Declarative Engine (SPADE) provides a language and runtime framework to support streaming applications. Users can create applications without having to understand the lower-level stream-specific operations. SPADE offers the ability to bring streams from outside the system and export results, and a facility to extend the underlying system with user-defined operators.
- Users can pose inquiries 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 by 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 operate directly on SPADE operators rather than the base, low-level applications components, but will still let new operators be developed.

All three of these methods are supported by the underlying runtime system. As new jobs are submitted, the scheduler determines how it might reorganize the system in order to meet the requirements of both newly submitted and already executing specifications, and the Job Manager automatically affects 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 the system. 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.

InfoSphere Streams Technology for Cybersecurity

In early 2010, IBM made this streams technology generally available under the name InfoSphere Streams. Many applications have been pursued, including financial services, energy trading services, health monitoring, and manufacturing. Most recently, the technology has been used to address the deluge of security data in critical cyber infrastructure systems.

As the systems that operate the critical infrastructures of the world become more intelligent and interconnected, they become more reliable and efficient. However, their exposure to cybersecurity vulnerabilities can increase, as these systems were not previously interconnected to each other or the Internet. Furthermore, the importance of these systems to everyday life has driven the incorporation of security monitoring and reporting features. While well intentioned, this has led to a deluge of real-time cybersecurity data that is so large that it is sent directly to storage (when possible) or ignored. With the increase of user devices added to the growth of sensors and actuators in these critical cyber infrastructure systems, the nature of the security data is changing:

- Data volumes and network bandwidth are expected to grow tenfold in the next three years.
- With the expansion of information, come large variances in the complexion of available data. Often the data is noisy with many errors and there is no opportunity to cleanse it in a world of real-time decision-making.
- One of the driving factors for the cyber infrastructure is the increased speed at which decisions have to be made. The market demands that businesses optimize decisions, take action based on good information, and use advanced predictive capabilities — all with speed and efficiency.

The powerful analytic capability provided by InfoSphere Streams is a good match for this challenge. Figure 4 shows how the cybersecurity decision processes can be enhanced with InfoSphere Streams.

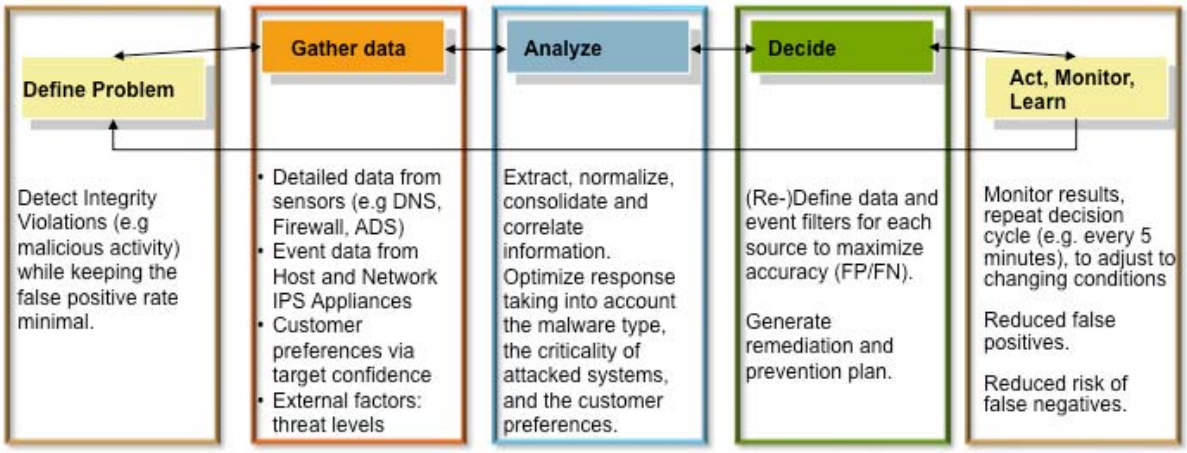


Figure 4: Realizing smarter security decisions using InfoSphere Streams

As a specific example of how InfoSphere Streams can be used for cybersecurity, consider the problem of botnets. Botnets are a network of compromised computers (called zombies), which are controlled by single host or group called the botmaster. The size of these botnets can range from hundreds to millions of hosts scattered around the world. Typical uses of a botnet include denial of service attacks, spam delivery, stealing banking credentials, and data theft, to name a few. The real owner of these zombie machines is typically unaware that a computer has been taken over. The botmaster uses a highly distributed command and control structure, often with standard functions, such as IRC, HTML, or IM to tell the zombie computers what to do.

The effective detection of botnet activity requires the aggregation of large data streams of differing formats and semantics from distributed sites across companies, jurisdictions, and countries.

This real-time data can come from several sources at differing speeds and qualities:

- Net flow data from enterprise networks, containing source/destination address/port, protocol, data volume, and router-specific summaries of point-to-point connections
- Intrusion Prevention System (IPS) logs
- Firewall logs
- DHCP logs
- DNS query logs
- Web server logs

The data rates can vary widely, from a few records per hour to more than 5,000 per second. This data is analyzed and key features that indicate botnet-like behavior are extracted. Finally, these features must be correlated across the different streams to confirm the botnet's presence.

For example, in one experiment, the collected live network flow data was analyzed and 32 hosts were found to be accessing known botnet command and control (C&C) servers. In another instance, the flow of DNS queries was analyzed and 12 hosts were found to be actively querying the known botnet C&C servers. Neither of these analyses was possible before due to the vast amount of real-time data that was required. InfoSphere technology provided the ability to quickly identify potential zombie systems in an ongoing, real-time manner.

Summary

The research that began in 2003 and eventually became IBM InfoSphere Streams has demonstrated many successes in a variety of commercial and scientific applications. It provides an infrastructure to support mission-critical data analysis with exceptional performance and interoperability.

Research around stream computing continues and is expected to further increase the scale and diversity of the IBM InfoSphere Streams' infrastructure, tools, support, and potential applications.

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

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

Produced in the United States of America
December 2010