Use InfoSphere Streams and BigInsights for real-time Hadoop analytics at scale

Timothy Landers

October 01, 2013

Configuring effective and efficient big data analytics at scale has long been a challenge. Moreover, storing huge amounts of unstructured data presents numerous obstacles to analyzing the data on a real-time or near-real-time basis. InfoSphere® BigInsights™ and InfoSphere Streams — based upon Apache Hadoop — provide a long-awaited solution to producing an optimal, decoupled computation framework for analyzing big data in real time at scale.

To help businesses use data effectively, InfoSphere BigInsights and InfoSphere Streams pair up with Apache Hadoop to tame big data analytics in real time. IT professionals supporting web analytics, social network content managers performing new or existing product research, marketing professionals promoting product lines, subscription research providers, and executives will find IBM's new analytics solution one of interest. A classic application of the software solution is InfoSphere Streams monitoring Twitter messages for sentiment related to a particular brand of phone.

“Big data will join mobile and cloud as the next 'must-have' competency as the volume of digital content grows to 2.7ZB (1ZB = 1 billion terabytes [TB]) in 2012, up 48% from 2011, rocketing toward 8ZB by 2015. There will be lots of Big Data.”

IDC: Top 10 Predictions 2012

This article provides procedures and configurations for accomplishing real-time analytics at scale using InfoSphere BigInsights, InfoSphere Streams, and Hadoop, and describes considerations for building, configuring, and deploying big data analytics to target new revenue streams.

Hadoop analytics

Hadoop is at the core of big data innovation. It enables the distributed parallel processing of huge amounts of data, and includes important data management and support capabilities. As a data
source, Hadoop uses mapping to store and process large structured and unstructured data sets by reducing computations and using algorithms and extensible programming models while interfacing with systems such as R programming.

The Hadoop process is to first divide big data into multiple parts, or segments, for simultaneous processing and analysis. Next, Hadoop loads the segments of data (structured or unstructured) into a multi-node file store referred to as the Hadoop Distributed File System (HDFS). HDFS records the data location and type as it distributes the segments to ensure fail-over by one or more file store nodes. Then Hadoop's MapReduce framework performs a query for requested data from each node simultaneously. The key benefit to InfoSphere BigInsights and InfoSphere Streams is that Hadoop can aggregate search results from multiple nodes in parallel, alleviating the performance overhead involved in querying from a central location. The set of search results then populates the InfoSphere BigInsights and InfoSphere Streams environments, where analytics are produced.

Using real-time Hadoop analytics at scale

One of the top 10 predictions by International Data Corporation (IDC) in 2012 is that big data will join popular demands for mobile and cloud technologies as the volume of digital content soars to 8 zettabytes (8 ZB) — where 1 ZB equals 1 billion terabytes (1 billion TB) by 2015 (see Related topics). Together, InfoSphere Streams and InfoSphere BigInsights accomplish — in real time, as well as batch — big data analytics, accommodating thousands of computers and storage devices, while simultaneously performing the parallel processing of large data sets.

The classic example of big data analytics at work is the sentiment analysis bar chart. Sentiment analysis discovers and tallies the number of positive and negative opinions about a product or concept. In the analytics bar chart shown in Figure 1, each bar's color represents the moving average for the last 60 seconds.

**Figure 1. Analytics bar chart**

The green bar above the zero line reflects positive feedback. All other bars, which are below the zero line, reflect negative feedback but are categorized by cause (InfoSphere BigInsights does this work for you):
• The yellow bar represents negative feedback based on a technical problem.
• The blue bar represents negative feedback based on a functional issue.
• The gold bar represents negative feedback based on an unknown issue (where the cause is unknown).

To get to this point, Hadoop has queried its distributed data segments on multiple nodes and returned a results set, for which InfoSphere BigInsights has analyzed the data and created categories. From this point, InfoSphere Streams monitors the streaming data for thresholds (see Figure 2). When a threshold is exceeded, InfoSphere BigInsights performs additional analyses to associate a meaning with the trend or pattern of the category’s data. InfoSphere BigInsights proves resourceful in using related data to identify the meaning of the trends, turning unknown causes into known categories.

Figure 2. Monitoring streaming data in InfoSphere Streams

Managing big data's volume, variety, and velocity

The big data types are evolving — from computer to web to mobile devices and more. Now that the data no longer has to be centralized, managing big data's volume, variety, and velocity are nearly unrestrained and in real time:

• **Volume**— Exponential growth in data has created the need for complex data-management solutions.
• **Variety**— New types of data are evolving as well, such as data from the social media revolution.
• **Velocity**— The turnover rate for developing new data types is decreasing, such that new data is being developed at increasing rates, requiring real-time analytics just to keep up with the demand to understand how to use the data to increase profitability.

**Environment**

**Download InfoSphere BigInsights**

InfoSphere BigInsights Quick Start Edition is a complimentary, downloadable version of InfoSphere BigInsights, IBM’s Hadoop-based offering. Try out the features that IBM has built to extend the value of open source Hadoop, like Big SQL, text analytics, and BigSheets. [Download BigInsights Quick Start Edition now.](#)

**Download InfoSphere Streams**

InfoSphere Streams Quick Start Edition is a complimentary, downloadable, non-production version of InfoSphere Streams, a high-performance computing platform that enables user-developed applications to rapidly ingest, analyze, and correlate information as it arrives from thousands of real-time sources. [Download Streams Quick Start Edition now.](#)

An important precondition of companies wanting to transform their business environment to be big data analytics-centric is to implement and deploy a data interoperability system. Doing so allows for growth in the volume of data beyond current system capacity. When multiple nodes are formally given the right to back up and store Hadoop-segmented data, InfoSphere BigInsights and InfoSphere Streams can begin managing data volumes to scale. Such systems can be used in many business types:

• **Consumer-facing enterprises (point-of-sale [POS], retail)**— Harnessing big data analytics means companies can analyze huge volumes of transactions and other data sources that conventional business intelligence (BI) programs could not handle, thus making better business decisions. Businesses derive the greatest value from the dynamic accessibility of data at POS systems and retail venues, such as inventory verification, cross-referenced identification checks, and credit checks.

• **Service providers (financial institutions, service management)**— Of key importance to financial services and insurance companies today is the ability to abstract valuable information from various data sources quickly and efficiently. Competition hinges on being "in the know."

• **Platform (developers, manufacturers)**— IBM’s mostly open source InfoSphere BigInsights distribution also provides the unique feature of allowing companies to perform rapid application development to generate prototypes or models of custom algorithms. Developers and manufacturers alike can use the IBM big data analytics platform to provide real-time analytics that optimize business processes on a whim using shell scripting but with high-capacity results.

• **Standardized bodies**— Standardization and industry collaborations influence big data evolution by preparing existing technologies for integration with big data. The 3rd Generation Partnership Project (3GPP), for example, is preparing to integrate big data with network infrastructures and devices. The IPSO Alliance is preparing to integrate big data with IP-networked devices, as well as integrating big data unified gateway interfaces. The European
Telecommunications Standards Institute (ETSI) is preparing to integrate big data with various architectures that broker services. Such changes will standardize on the use of big data in various industries, using regulatory change and policies.

**Embracing the technology**

Big data is embracing the technology revolution. Big data analytics are fueling the push toward cross-industry standardizations for the purpose of conforming big data technologies to current industry best practices.

**Technology drivers**

Hadoop is an open source software project that enables the distributed processing and handling of failures at the application layer. For businesses, a complementary technology is the HDFS, which takes on the following tasks:

- **Increasing the number of mobile telephone subscribers**— Data sources turn into insight sources.
- **Growing the adoption of mobile Internet financial services**— Transportation, logistics, retail, utilities, and telecommunications sensor data is being generated at an accelerating rate from fleet global positioning system transceivers, radio-frequency identification tag readers, smart meters, and mobile phones (call data records). That data is used to optimize operations and drive operational business intelligence to realize immediate business opportunities.
- **Potentially expanding a customer base in mobile network organizations**— Performing analytics over social media means dynamically processing and analyzing streaming data in large-volume environments. The potential for vertical integration and market expansion is enormous. As social networks, blogs, and forums continue to generate large volumes of data, the economy is positively affected and opportunities arise to supply new demands.
- **Providing faster transaction times, lower operating costs, enhanced customer relationships, and innovative loyalty services, such as e-coupons and rewards**— Almost 30 percent of the data in an organization is from logs. Multiple applications, servers, and networks generate tremendous amounts of data, and these log files can grow to be research programs.
- **Capitalizing on existing payment infrastructures to immediately offer near-field communication (NFC) mobile payments (m-payments), increasing return on investment**— The use of intelligent meters in "smart grid" energy systems that shift from a monthly meter reading to an "every 15 minutes" meter reading can translate into a multi-thousandfold increase in data generated.

The amount of data analyzed to decide which ads should be displayed on your favorite website is staggering. Large advertising networks also have to collect data on millions of users. Business can leverage existing smart card reader infrastructures to support the deployment of NFC-enabled mobile services.

The finance industry is another segment that generates large volumes of data. Hadoop can help solve financial problems by analyzing large sets of transactions or stock prices.
Big data isn’t just for business, however. It also offers numerous ways to help individuals make better decisions by employing predictive analytics to forecast otherwise-unknown events. From how we spend to how we vote to how we keep apace of trends and protect our identities, data mining has fueled the advent of predictive analytics, which generates meaning from interrelating various types of information, including models, trends, measurements, and indicators.

The big data analytics services developed by IBM discovers, visualizes, predicts, and models data, giving companies and individuals the freedom to explore and contemplate data in multiple dimensions, such as by location, by product or service, and by customer preference. In addition, InfoSphere BigInsights and InfoSphere Streams, which leverage Hadoop technology, could prove especially promising in the area of financial and credit management services. Where data is the primary product for services rendered, IBM’s big data solution handles billions of transactions, practically alleviating the need to carry cash and multiple credit cards. Financial and credit management services companies gain their competitive advantages from the availability of big data analytics for their industries.

Another aspect of big data benefits to individuals is record-keeping for all transactions, a service that protects the integrity of an individual’s transactions and ensures the transparency of the individual’s legal rights. Big data provides individuals with the confidence of knowing that their actions are justified in the eyes of both consumer creditors and the law.

**Impediments to analytics**

Finding highly qualified big data professionals to build and maintain big data systems has become a primary impediment to implementing these analytics solutions. More and more, organizations are adopting big data technologies only to find that human resources are scarce in this area.

**Impediments to analytics**

- Finding qualified data professionals
- Lack of an agreed-upon business model for near-field communication
- Lack of adequate investment in R&D by mobile network operators
- Extra requirements for transparency and rationalization for outsourced security, networking, and Wi-Fi communications products

Similarly, the scale and scope of companies’ access to data are undeniably changing the way companies do business. To date, the best NFC business model has not been determined. Similarly, a lack of agreement exists as to how the best NFC business model should accomplish its capabilities. But the general agreement is on a model in which big data enables monetization of data, innovates new means of generating revenues, and uses related data sources to create new perspectives on the same information.

Big data solutions are a major contributor to mobile network operators (MNOs) and their fast-rising competencies. To shape the commercialized use of big data among NFC services, MNOs will have to play a leading role in research and development, investing for the industry to keep pace with demand. Without MNO investments in this area, development and growth may shift to a different industry — one that supports the necessity of R&D investments to sustain competitive advantages. One of the latest trends in big data is the use of mobile wallets. Looking for new
revenue sources, MNOs are now combining the smart phone with a credit card, effectively making MNOs the financial transaction acquirer and processor.

Manufacturers are discovering the value of big data in providing timely responses to decision-making, productivity, and realizing value. For example, manufacturers outsourcing 100 percent of their manufacturing of security, networking, and Wi-Fi communications products are finding that they need big data for transparency and rationalization. To help manufacturers adapt to the massive increases in data resulting from vertical integration and outsourcing, real-time and partial real-time data reflecting product performance can help optimize the management of manufacturing demand and supply during peak usage periods. Without the dynamic availability of streaming data and the analytics necessary to quickly translate the data, manufacturers will lose out to more knowledgeable and agile competitors.

Managing all this big data requires the necessary technological functionality for prompt delivery of big data analytics. The IBM solution offers value in producing analysis, data detection and protection, natural language processing, pattern recognition, sentiment analysis, algorithms, and predictive modeling. In addition, big data technologies can provide increased security and assurance in the following areas:

- **Man-in-the-middle attacks**— Big data has proved useful in preventing man-in-the-middle attacks by combating the growing threat of online theft and fraud. Shaping big data into intelligent mapping algorithms can serve to identify companies and pre-empt these attacks on web servers sooner than was previously possible.
- **Eavesdropping**— From home security to the U.S. Department of Homeland Security, up-to-date information is key to pre-empting and preventing breaches in security. Moreover, security companies are discovering that big data generates new revenue streams for the tech-savvy. What is being deemed as “predictive analysis” is now being attributed to the protection of information and the generator of big data revenues. From computers controlling home security and electrical functions to ensuring border control and screening at major airports and other travel ports, revenues will follow the proven use of big data to facilitate the application of meaningful information to our daily lives.
- **Data corruption and manipulation**— Parallel distributed processing on multiple nodes effectively reduces bottlenecks and optimizes CPU performance. Notwithstanding, when a node fails, the fail-over features of parallel distributed processing quickly recover data for uninterrupted processing, but with multiple nodes sharing a single query, the failure of a single node could result in data corruption. IBM offers various tools for the manipulation of data, including Jaql, the query language for JavaScript Object Notation (JSON), which offers the filtering, joining, and grouping of data to accommodate data query customizations.
- **Denial of service (DoS)**— With big data analytics, technology now has the capability to quickly identify a DoS attack while it is forming, allowing enough time to prevent an attack in addition to network security software services.

**Conclusion**

InfoSphere is multi-dimensional in its use of data, offering structured and unstructured, real-time data processing, and analysis. The key is that IBM is using Hadoop to highlight InfoSphere
BigInsights and InfoSphere Streams capabilities, a combination of functionality that is unique in the industry. Where InfoSphere BigInsights and InfoSphere Streams process streaming data in real time to turn large, complex data volumes into insights, Hadoop adds distributed parallel processing, which effectively allows InfoSphere BigInsights and InfoSphere Streams to focus on analytics and target specific data streams. IBM has innovated essential data processing and gone a step further to revolutionize data analytics, touching on a new level of business intelligence.
Related topics

• For the latest news about Apache Hadoop and other elements of its ecosystem, visit Apache.org.
• Learn more about InfoSphere Streams in the developerWorks article "An introduction to InfoSphere Streams."
• Sentiment analysis is considered the ideal use case for Hadoop.
• Discover all the technologies and uses available from the IBM big data platform.
• The IPSO Alliance is most closely associated with defining the "Internet of Things," but the consortium of leading technology organizations is also proving important in the big data world.
• ETSI is a standards body for the European telecommunications industry and as such is poised to be a major factor in the big data world.
• Discover Jaql, the JSON query language.
• The IBM Global Business Services Business Analytics and Optimization Executive Report Analytics: The new path to value illustrates the power of big data analytics, including its drivers and the organizations that are benefiting from its use.
• Watch developerWorks on-demand demos ranging from product installation and setup demos for beginners to advanced functionality for experienced developers: Informix Warehouse InfoSphere products | Cognos products
• Download InfoSphere BigInsights Quick Start Edition, a complimentary, downloadable version of InfoSphere BigInsights, IBM's Hadoop-based offering.
• Download InfoSphere Streams Quick Start Edition, is a complimentary, downloadable, non-production version of InfoSphere Streams, a high-performance computing platform that enables user-developed applications to rapidly ingest, analyze, and correlate information as it arrives from thousands of real-time sources.
• Check out the entire list of ">big data and business analytics products available for trial download.

© Copyright IBM Corporation 2013
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)