IBM Software

IBM InfoSphere BigInsights

Enabling new, cost-effective solutions to turn complex information into business insight

Executive summary

Companies are hyper-connected to their customers and partners every minute of every day. With that connectedness comes exploding volumes of data, of greater variety and at a greater velocity. These flows are so large that they define a new category: big data. They offer tremendous potential for deep insights that support smarter decisions and increased revenue.

How can businesses benefit from these information flows? How can they harness big data to generate new insights quickly and cost-effectively, while building off their existing information management approaches and strategies?

Imagine if you were able to:

- Build sophisticated predictive models from the combination of existing information and big data information flows, providing a level of depth that only analytics applied at a large scale can offer
- Broadly and automatically perform consumer sentiment and brand perception analysis on data gathered from across the Internet, at a scale previously impossible using partially or fully manual methods
- Analyze system logs from a variety of disparate systems to lower operational risk and optimize advertising content targeting
- Leverage existing systems and customer knowledge in new ways that were previously ruled out as infeasible due to cost or scale

To be innovative, companies are looking for new ways to grow by finding value in this data and challenging traditional business models to provide greater efficiencies, increase revenue, create new value-add services, and in some cases transform how they do business. IBM® InfoSphere™ BigInsights enables a new class of solutions associated with big data challenges to help organizations optimize their business. InfoSphere BigInsights is an analytics platform that delivers unique IBM Research, emerging technologies and capabilities on top of Apache Hadoop open-source framework, enabling new solutions on a business-ready platform fully supported by IBM.

The big data opportunity

Many companies have benefited from the insights gained from structured data, such as point-of-sale data, that then comes to represent an important company asset. Like sales data, there are opportunities and competitive advantages that lie within big data. What if you could effectively analyze data in the enormous volumes—and variety of sources and formats—that can appear on the Internet? What if you could analyze system log files in combination with existing warehouse data? What if you could gain a unified understanding of customer behavior in your physical stores and online activities?

Sheer size, changing types and evolving patterns of Internet-scale information

There is simply too much information in too many formats coming from too many sources to manage it effectively with traditional tools. Information is growing every hour. During the course of a year, companies can amass gigabytes, terabytes or even petabytes of information. Some enterprises generate terabytes of new information every hour. Further complicating the challenge is that by many estimates as much as 80 percent of this data is semi-structured or unstructured. This includes web pages and web log files, clickstreams, search indexes, social media forums, instant messages, text messages, email, documents, consumer demographics and lifestyle segmentation, sensor data from active and passive systems, and more. This is the core of the big data challenge: companies need to be able to understand and analyze this Internet-scale information just as easily as smaller volumes of structured information.

Sifting through existing and new information types

New information types and sources do not mean that existing ones are no longer relevant. New insights cannot be achieved at the expense of forgetting what you already know. Internet-scale information, with its diversity and frequent changes, needs to be related to existing insights about a company's customers and business operations. Furthermore, new technologies cannot be embraced at the expense of upsetting existing systems, retraining existing users or disrupting current business. Companies are looking for guidance on how to use these new types of information in context with existing information management technologies and strategies.

Methodology gap

Big data requires new methodologies based on both the size and diversity of the information to be managed. Proven methodologies designed for structured data cannot practically be applied to most big data. Structured data practices require a well-defined database schema, and that the data adheres to absolute consistency and reflects a final, material state that allows aggregation to take a day-forward approach.

In contrast, absolute consistency is not the primary requirement for the majority of web application data. Instead, responsiveness outweighs consistency. With tens of thousands or even millions of interactive users, efforts to integrate information collected from the web by making it conform to an existing database definition are not only expensive, but take too much time. Insights need to be found quickly to maximize business value. Also, because web information changes rapidly, a data refresh often requires reaggregating the entire set from scratch.

Big data is the dominant form of dormant information today. Effectively finding value in non-traditional information types requires a new approach to handle the sheer volume and new, unique approaches to process this information cost-effectively.

InfoSphere BigInsights

InfoSphere BigInsights enables companies to turn complex information sets into insight and to do so at Internet scale. InfoSphere BigInsights is an analytics platform that provides unique capabilities from IBM emerging technologies, IBM research technologies and IBM software built on top of an Apache Hadoop open-source framework to deliver a platform that is business-ready to accelerate the time to value.

In addition to core capabilities for installation, configuration and management, InfoSphere BigInsights includes advanced analytics and user interfaces for the non-developer business analyst. Flexible enough to be used for unstructured or semi-structured information, the solution does not require schema definitions or data preprocessing and allows for structure and associations to be added on the fly across information types. The platform runs on commonly available, low-cost hardware in parallel, supporting linear scalability; as information grows, you simply add more commodity hardware.

Complements existing solutions

InfoSphere BigInsights is complementary to IBM InfoSphere Data Warehouse, IBM InfoSphere Information Server and IBM InfoSphere Streams. It does not replace existing solutions; in fact, the most successful big data strategies actively leverage and integrate existing solutions to bring context to new insights. The following descriptions of these solutions show how InfoSphere BigInsights fits as either a stand-alone or integrated option. An example is provided to show how each solution addresses different parts of a business problem.

Note: A key area of difference is the information types that the solutions support. These information types are broadly categorized as structured vs. unstructured.

InfoSphere Data Warehouse

Structured information from business automation systems such as ERP and CRM systems are consolidated into a data warehouse and used with business intelligence (BI) systems for operational planning, financial reporting and trend analysis. Many organizations also use advanced analytics on transactions for items stored in a warehouse to develop predictive models such as fraud models.

InfoSphere Streams

InfoSphere Streams is designed to uncover meaningful patterns from information-in-motion or data flows during a window of minutes to hours. An example is flagging potential fraud activity as transactions are occurring such as detecting abnormal buying sprees or exorbitant purchase amounts.

InfoSphere Information Server

InfoSphere Information Server plays a critical role in integrating enterprise data for trusted information. Organizations that are upgrading or consolidating enterprise applications, migrating data from legacy applications, building a data warehouse, integrating data for master data management or updating the data warehouse with patterns found from InfoSphere Streams and InfoSphere BigInsights need these capabilities to ensure information is trusted wherever it is needed.

Extends insight

InfoSphere BigInsights extends the insight that can be garnered by providing a level of depth that only analytics applied at a large scale can offer. Unlike InfoSphere Data Warehouse and InfoSphere Streams, InfoSphere BigInsights enables solutions to deliver insights on complex unstructured information sets that accumulated over a long period of time—and provide those results quickly. An example is consumer sentiment and brand perception analysis across social media forums. Another example is finding patterns in large, complex systems such as city traffic, or in distribution systems such as package delivery or the electric grid.

Through integration with InfoSphere Streams and InfoSphere Data Warehouse, analysis can be extended to encompass information-in-motion, unstructured and structured information accumulated over a long period of time. An example is the development of fraud models that continuously evolve based on changing fraudulent behaviors. Another example is finding insights in consumer purchasing patterns by associating consumers' online behavior with their purchasing history and known demographic data to predict when and where they will make future purchases.

In summary, InfoSphere BigInsights extends existing technologies to move beyond the information types for which they were designed to encompass a broader, more comprehensive set of information types, enabling a more complete view of the business. InfoSphere BigInsights supports open development standards and works with existing systems, including InfoSphere Warehouse, InfoSphere Streams, IBM DB2®, IBM SPSS, IBM Cognos®, IBM Enterprise Content Management (ECM) solutions, IBM Tivoli® and more.

Emerging uses

IBM has conducted early engagements and pilots with several companies to understand usage scenarios and requirements so that the software solution meets business needs. Based on these early pilots, several usage patterns emerged:

- Predictive modeling
- Consumer sentiment insight
- Deep self-service capabilities
- Research and business development

Common to these usage patterns is the ability to derive business value from complex, unstructured information in a timely, cost-effective way. The following is a brief overview of these patterns, customer examples and the value these customers garnered.

Predictive modeling

Predictive modeling refers to uncovering patterns to help companies make business predictions such as forecasting where there may be a propensity for fraud or how pricing affects holiday online candy sales. Finding patterns has been a mainstay for many companies for many years. However, traditional methods create models based on transaction or sales data sources; modeling across both traditional data and non-traditional sources of information (for example, unstructured information) is emerging.

Banking: Credit card fraud can cost credit card issuers as much as 7 percent of sales transactions, amounting to billions of dollars per year. To help mitigate risk by preventing these crimes, issuers create models to uncover patterns of events within a customer's life—such as divorce or foreclosure—that correlate to fraud. With traditional methods, fraud models can take as long as 20 days to develop. But understanding fraud patterns nearly a month after an incident occurs is only partially helpful as fraud schemes constantly change. Organizations need a better way to detect existing fraud patterns and stop new ones before they incur significant losses.

Financial services: A large brokerage firm is currently piloting InfoSphere BigInsights to gain a new level of comprehensive understanding of its business. The company is primarily an online business and has a variety of systems used by customers, internal employees and advisors. The systems come from prior business acquisitions—some are based on off-the-shelf software, but most were internally developed. There is a lot of interplay between these systems, but there was no comprehensive way of logging or reporting all system activity. With no unified way of understanding what happens online, the brokerage firm was unable to understand the customer experience on these systems in order to improve its services. Furthermore, IT was not able to find patterns leading up to system outages.

During the pilot, log files were loaded from eight different systems. InfoSphere BigInsights was used to structure and parse the log files on the fly and then perform predictive analytics. IBM BigSheets, a tool that comes with BigInsights, was used to visualize and interact with the results while enabling a non-developer, business analyst to help drive the analytics.

The results helped the company in several ways. First, IT was able to get a better understanding of site and system health and forecast system failures. From a business point of view, the company can now understand revenue and spend based on the online experiences of consumers and financial advisors, and determine how both groups are affected by different service levels. The results helped the company better understand operational risk and optimize both ad targeting and service levels.

The difficulty of analyzing web server and system logs is a common theme across large companies, including those that generate terabytes of data per hour during peak operations. This is an area of undiscovered value.

Insurance: A healthcare insurance provider performs analysis on more than 400 million insurance claims to determine the potential dangers of individual drug-to-drug interaction. By cross-referencing a list of prescriptions over the course of a patient's medical history and an external service for known drug-to-drug interaction problems, the service can flag potential conflicts of drug-drug, drug-disease and drug-allergy interactions for healthcare providers, improving the quality of patient care and lowering the cost.

The challenge with the insurance provider's internally developed system was that it was common for an analysis of a large set of patient data to take over 100 hours. The patient reference and treatment data is very complex because the data is intricately nested. With InfoSphere BigInsights, the insurance provider was able to reduce the analysis time from more than 100 hours to just 10 hours across more than six terabytes of data.

Retail: A large retailer wanted to better understand what customers do in physical stores and what customers do online in order to act in the most relevant way at the point of decision. Gaining insights required lots of exploration and "wandering" across large amounts of web logs and data representing physical stores found in its data warehouse. Once patterns were found, the retailer wanted to promote the information into a data warehouse to leverage its existing BI infrastructure, which is very mature and highly relied upon.

InfoSphere BigInsights was used to parse varying formats of web log files and match this up with data warehouse information to tie buying behavior online with that in physical stores. Predictive analytics was used to find patterns across these dimensions. IBM BigSheets was used to visualize and interact with the results to determine which elements should be tracked. Once determined, an ongoing process was implemented to cleanse the data with extract, transform and load (ETL) and load it into the data warehouse—enabling existing BI users to act upon information about buyer behavior in physical and online stores. The results enabled the retailer to increase sales with more targeted product suggestions online and product placement in physical stores to meet customer needs and preferences.

Consumer sentiment insight

Developing consumer insight involves uncovering consumer sentiments for brand, campaign and promotions management.

Consumer packaged goods: A large company specializing in drinks spends a substantial portion of its budget to market its brand to ensure household appeal and sales. The company piloted InfoSphere BigInsights to better understand consumer sentiments and brand perception in online social media forums. This included tracking general topics connected with its brands to answer questions such as:

- What amount of discussion around known topics in connection with the brand is high or low over a period of time?
- Within these discussions, how much of the commentary is favorable or unfavorable?
- What is the amount of discussion around a known spokesperson of the brand and what amount is favorable or unfavorable?
- Who in the "blog world" are the biggest influencers?
- What topics are emerging around the brand that the company doesn't know about?

Knowing the answers to those questions helps the company better target its marketing campaigns and promotions, resulting in an increased return on its marketing investment.

Currently, the company uses a third-party tool to provide a partial view, but has concerns about the validity of the results as the tool does not analyze a large enough set of information. In addition, the software cannot understand inherent meanings or focus on relevant topics without being too literal with misspellings, syntax or jargon.

With InfoSphere BigInsights, the company was able to aggregate large amounts of information from social media sites such as blogs, message boards, news feeds and so on. The organization used text analytics capabilities in InfoSphere BigInsights to sift through the information and find relevant discussions. For example, the analysis included identifying discussions that were in context and connected to the brand. Further analysis determined which discussions were favorable or not favorable and whether the conversation was about the company spokesperson or about another person with the same last name. Having the right granularity and relevance from a large sample can mean the difference between results that are useful or not useful at all. In the end, the company found answers to its questions with contextual integrity across a large amount of information in a timely way.

Deep self-service capability

Financial services: Brokerage companies provide their clients with monthly statements that show how their financial portfolios performed. The statements are available both in print and online. In addition, clients and advisors can track gains and losses online; however, to extend the tracking timeframe beyond two years, an advisor must go to multiple sources or copy statement information by hand. As most portfolio strategies are long term, this two-year view of gains and losses does not meet the needs of advisors and clients.

To enable clients and advisors to better search and track gains and losses across a longer timeframe, the firm needed to access old statements and extract appropriate content. However, because the statements were generated by a mainframe for the purpose of being printed on high-speed printers, the electronic statements—while viewable online like the printed form—were not in a searchable format.

Converting large backlogs of statements to text is not a new idea, but the company had considered it infeasible due to the cost of processing so much information in a timely way. The brokerage firm is now working on a pilot with InfoSphere BigInsights to use existing conversion programs to convert large backlogs of statement files to text quickly and cost-effectively. This way, clients and advisors can search old statements and extract gains and losses for further analysis. Once in production, and after the initial heavy backlog conversion is done, the system can be configured according to the size of ongoing conversion needs.

Research and business development

Government: The British Library has a very large and rapidly growing web archive portal that allows researchers and historians to explore preserved web content. The challenge is to preserve the digital culture of the nation as websites are published, modified or removed daily.

With initial manual methods, pages of 5,000 .uk websites were classified by 30 research analysts. However, the cost to manually archive the entire .uk web domain—which comprises 4 million websites and is growing daily—would be exorbitant.

In a recent pilot, the British Library used InfoSphere BigInsights and a classification module built by IBM to electronically classify and tag web content and enable/create visualizations across numerous commodity PCs in parallel, dramatically reducing the cost of archiving. The British Library can now archive and preserve massive numbers of web pages and allow patrons to explore and generate new data insights.

Financial services: A large credit card company piloted InfoSphere BigInsights for business-development purposes to improve intellectual property portfolio analysis for mergers and acquisitions. The goal of the project was to automate the process of comprehensively measuring the value of potential acquisitions. This required both public and private information, including:

- U.S. Securities and Exchange Commission filings, such as annual and quarterly reports
- U.S. Patent and Trademark Office patents, assignments and trademarks
- Company press releases
- Other M&A and inventor information from feeds and web pages

In addition, the company needed to collect information such as corporate genealogies, IP ownership and patents ranked by citation. This information was combined with items affecting IP value, inventor affiliation and citations ranked by time. As mergers and acquisitions do not happen quickly, the company needed ongoing tracking capabilities as well as the ability to present the information in a graphical format that business-development professionals could easily comprehend.

In the pilot, 1.4 million U.S. patents from 2002 through 2009 were gathered from the Internet, along with hundreds of company patents. Next, 6.1 million citations were extracted and correlated with the patent records. The correlations were the basis for a ranking value measurement—the more a patent is referenced, the higher its value or usefulness. The analysis was completed in hours, compared with the weeks that would have been required if manual methods were used. Ongoing tracking and refreshing allow the business-development professionals to review updates as changes occur.

Putting big data to work with InfoSphere BigInsights

Innovative companies look for new ways to grow by finding value in big data and challenging traditional business models to provide greater efficiencies, value-added services and a greater opportunity for transformation. IBM has the technology and experience, gained through numerous successful pilot programs, to bring InfoSphere BigInsights analytics solutions to companies worldwide. InfoSphere BigInsights enables new solutions for problems that were previously too large and complex to solve cost-effectively.

How can you use IBM InfoSphere BigInsights to change the game at your company?

For more information

Interested in piloting InfoSphere BigInsights? Please contact your IBM sales representative today, or learn more by visiting: **ibm.com**/software/data/infosphere/hadoop

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

Produced in the United States of America
October 2010
All Rights Reserved