Maximizing the Value Provided by a Big Data Platform

June 2012
Maximizing the Value Provided by a Big Data Platform

“Big Data” is one of the key buzzwords for 2012; and for good reason. Thanks to our increasingly instrumented and interconnected world, the data flows generated by today’s devices and sensors can provide significant insight into customer interactions, manufacturing facilities, railway infrastructures, and so much more. Or they would, if we were able to capture, store, search, analyze, and visualize this massive amount of high-velocity information in an efficient, timely and meaningful manner.

Academic research suggests that organizations using data and business analytics to support the decision making process are more productive and experience higher returns on investments than seen by their counterparts. Specifically, the effective use of data and analytics has been correlated with a 5 to 6 percent improvement in productivity, as well as higher profitability and market value. In short, organizations can use controlled experiments to test hypotheses and more accurately analyze results (distinguishing between causation from mere correlation) to help further guide investment decisions and operational changes.

This whitepaper makes the assumption that the reader is familiar with the term Big Data, the increasing pool of data and data sources that are captured on a day to day basis, the speed at which this data is produced, and the fact that its value continues to remain untapped. Specifically, it focuses on how organizations can discover the value hidden within this data and how best to deliver that insight to its decision makers and management—while keeping tight reins on the monitoring and management of these changing topologies. While more familiar processes such as data warehousing and business intelligence are briefly acknowledged, the focus is on the more recent emergence of MapReduce and real-time analytics. Further, it also looks at how organizations can minimize the time to value and total cost of ownership, while simultaneously maximizing return on investment.

The Common Perspective on Big Data

Typical analytic processes and tools focus on using stored and most likely structured data. To accommodate structured storage, data often requires several time consuming and expensive preparation steps (e.g., manual processes, optical scanning, audio to text, cleansing, etc.) to prepare this “noisy” data. The associated expense and sheer scale are key contributing factors to why the value in this mass data remains hidden. Organizations would rather spend on areas where they know there will be an immediate return, compared to areas where they are unsure of the return, if any.

Big data technologies help address the difficulties working with these large unstructured, semi-structured and structured data sets not met through relational database (RDBMS) and desktop applications. To process this enormous amount of data in a timely fashion requires a collection of technologies including massively parallel processing (MPP) warehouses, data mining grids, distributed file systems, stream computing, information integration and governance, and more.

When exploring big data, the results of computations continue to evolve with the discovery and extraction of more information. As such, a relational database is not always the best destination for data even if it has been scrubbed—given the static nature of schemas. For example, data arriving within a structured format such as JSON, may contain a significant amount of unstructured information like text, video, and images. Similarly, while some vendors offer native XML document support and dedicated XML stores (e.g., IBM DB2), others “shred” the XML content to fit the database model.

Over the past few years, an alternative to relational databases (based on the SQL language) continues to garner a great deal of attention given its ability to accommodate semi-structured and unstructured data. Referred to as NoSQL, these types of databases provide enough structure to organize data but do not require an exact schema for data before storing it. Such databases have been used to allow Web-based businesses to manage the large quantities of data generated by their users.

While some pundits of NoSQL view these new technologies as a move away from traditional relational databases, others associate NoSQL with “Not only SQL,” meaning that these databases are not to be considered a replacement to existing solutions. It is true that NoSQL can be used to store structured data. And this is particularly useful when getting data into an SQL repository is not worth the added investment (e.g., short shelf life, non-repeatable, or quickly changing schemas). However, NoSQL is not a replacement for more traditional hierarchical databases (favoured for online transactional processing) or the use of relational databases for data warehousing and online analytic processing.

Maximizing the Value Provided by a Big Data Platform

June 2012

To address the additional requirements for processing this massive amount of stored data (distributed file systems, parallel processing, etc), Apache Hadoop has emerged as the de facto standard framework. Rather than a product in and of itself, this open source solution (inspired by Google MapReduce and Google File System) has become an ecosystem of multiple components. These include, but are not limited to: Hadoop Distributed File System (HDFS), providing unstructured storage across tens, hundreds or thousands of machines (replicated for redundancy); MapReduce for job scheduling and RAM processing; HBase for supporting real-time queries; Chukwa for displaying, monitoring and analyzing logs; Pig for data flows; Hive for batch SQL-like queries; and more.

Thanks to Hadoop’s open source status, many existing analytics vendors are moving to incorporate it and many of its key features into existing products. IBM, with arguably one of the best analytics portfolios in existence today, offers InfoSphere BigInsights, adding significant value to a Hadoop implementation, including but not limited to integration with other popular offerings such as IBM Netezza (providing in-database MapReduce) and IBM DB2.

Although big data solutions continue to increase in availability, even the largest organizations are struggling to find workers with the skills necessary to analyze and translate results into actionable business information that will positively impact the bottom line. Reports already note the shortage of analytical and managerial talent necessary to get the most from big data. “The United States alone faces a shortage of 140,000 to 190,000 people with deep analytic skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings.”2 “The availability of big data solutions that focus on consumability (i.e., customers’ end-to-end experience with technology solutions) and optimizing common tasks (e.g., accelerators, toolsets, declarative languages, etc.) can go a long way to help mitigate the associated complexities, allowing organizations to focus on the needs of the business.

The Value in Big Data

There are many papers that outline the value locked away in the plethora of big data that continues to accumulate. Within this data lie valuable patterns and information, previously hidden because of the amount of work required to extract them. Simply making this data more easily accessible to relevant stakeholders in a timely fashion can help unlock this tremendous value and any insights that might otherwise remain hidden; helping improve decision-making and minimizing risk.

Customer facing companies have long used collected data to help segment and target customers. However, big data provides organizations with an ability to use data from social media and other unstructured sources to compile larger and more accurate customer profiles. Where organizations in the past may have had to settle for assumptions and hunches using online analytic processing of click traffic, this additional information can be used to help portray customer sentiment around products, such as intent to buy.

While there is significant value in big data, depending on the industry, its value can quickly diminish over time. The rate of which this data is generated is not the issue. There are many examples of successful implementations that can write big data into bulk storage for later batch processing. The issue rather, is the speed of the feedback loop; from data input to final decision. This is akin to driving using the rear-view mirror or crossing a busy street with snapshots of the intersection from five minutes ago. There are times when waiting for a report to run, or a Hadoop job to complete is unacceptable.

Whether in a relational database or in a Hadoop Distributed File System, the key point is that this data is stored (often referred to as data “at rest”) before it can be analyzed. The ability to analyze this data “in motion”, before (or if ever) it is stored, can help maximize value, particularly where speed (versus depth of analysis) is key. For example, an online retailer could better track individual customers updating their preferences and behavior profiles in real-time. Ideally, this could be used to help recognize when a customer is nearing a purchase decision or is likely to abandon their online shopping cart, making it possible to offer a bundling of preferred products or discount to help nudge the transaction to completion.

Additional examples illustrating the power of real-time analysis could include: Telecommunications, where cable modems, set-top boxes and other Internet connected devices create IPDRs (Internet Protocol Detailed Records) enabling the collection of usage metrics to monitor service patterns and maintain service level agreements, quality of service, manage service consumption and more; and oil field operations, where wellheads, pipelines, and mechanical systems are instrumented to provide constant feedback of their condition. Using this information, real-time results can be used to adjust oil flows to optimize production and minimize downtimes. One major oil company has

2 McKinsey Global Institute, Big Data: The next frontier for innovation, competition, and productivity
done just this, with the net effect of reduced operating and staffing costs by 10 to 25 percent and increased production of about 5 percent.

While the focus of big data to date has primarily been on data at rest, the movement towards real-time analysis of data in motion continues to gain traction. However, very few vendors offer commercial solutions (e.g., IBM InfoSphere Streams), and alternatives are not yet polished.

Expanding on the Definition (What’s next for Hadoop?)

Big data continues to evolve as a business concept in tandem with the underlying technologies; NoSQL and the Hadoop platform were created to solve specific issues around big data that traditional databases could not address. As Hadoop gains in popularity, many predict that 2012 will see IT more rapidly adopt NoSQL. Hadoop is quickly becoming the open source de facto standard for batch processing analytics applications. Hadoop (and MapReduce in general) facilitates the parallel processing of massive amounts of collected data.

By its very nature, Hadoop operates in batch. Specifically, a typical usage pattern involves the loading of data (in large batches for efficiency) into HDFS, execution of the MapReduce operations, and the retrieval of results from HDFS. It is for this reason that Hadoop is best suited for analytical and non-interactive computing tasks. In short, the required batch processing simply does not allow for the continuous analysis of events as they stream in at a rate of millions of events per second per server. Organizations looking for real-time analytics to deliver “next click” results (i.e., by the time the page loads, the semantic implication is reflected in the page) or expect results with extremely low latency between data in and analytics out are required to look elsewhere. Evidence of this movement is already available through those organizations that were once considered the “poster children” for MapReduce; Google and Yahoo.

While Hadoop provides organizations with persistent data that supports deep analytics, there is an increasing requirement for real-time analytics. Instead of the large batch processing with latencies measured in minutes or hours, organizations such as Google, Yahoo, and Facebook are building solutions to support smaller individual updates. For example, Google has built its Percolator system to support incrementally processing updates to large data sets which allows it to process the same number of documents per day but reduces the average age of documents in Google’s search results by 50%. Yahoo itself is working on a couple of real-time analytics projects; S4 and MapReduce Online. While many other vendors are also pursuing the development of such projects (e.g., Appistry and Accenture Cloud MapReduce, and DataStax Brisk Hadoop), they have yet to match the capabilities of the commercially available IBM InfoSphere Streams.

The movement towards real-time processing of data in motion facilitates more immediate answers and responses to current information. It also allows for analytics when storing data for later analysis is too expensive or time consuming. However, real-time analytics are further complemented by stored (at rest) information, playing a key role in fuelling educated responses to real-time events. Statistical analysis can be run on a long history of data looking for patterns and insights that can be used to optimize real-time decision-making. In short, the insights discovered and learned from the increasing volume of data at rest, can be applied to the real-time analytics to provide more pertinent results.

The creation of solutions based on MapReduce and NoSQL were inspired by companies like Google, Yahoo, Facebook, and Amazon in an attempt to provide a new level of scalability before their infrastructures were crushed under the rapid growth of their success. While these organizations have budgets and teams dedicated to develop and govern this type of implementation, this is far from the norm. The related technologies are not yet productized to allow other enterprises (i.e., those without dedicated resources) to deploy similar solutions. While some work has been done to assemble distributions to address the many dependencies of these complicated systems, they do not yet meet the governance, high availability, security, and large streaming data set challenges of the more common enterprise. Vendors such as IBM are providing enterprise organizations with the necessary tools and added value to more easily deploy and govern these types of implementations, eliminating the requirement for large dedicated budgets and development teams.

IBM’s Big Data Vision

To address the many varied needs of its customers, typically requiring multiple technologies, IBM has developed a comprehensive and integrated platform for big data incorporating Hadoop, stream computing, data warehousing, and information integration and governance. Advanced software analytic tools and sophisticated mathematical models help identify patterns, correlate events, and more. With these technologies, organizations can better understand, anticipate, and make necessary changes with increased confidence. While many vendors are building solutions based on Hadoop or simply making it easier for organizations to implement a Hadoop infrastructure more quickly, IBM provides customers with an extensive analytics platform to address the entire big data ecosystem. IBM can also help customers uncover...
new data, combining it with existing information assets, and move from at-rest only analytics to incorporate a real-time analytical processing engine.

IBM is one of the few vendors (if not the only) that can address all the data related needs of an analytics driven organization. The IBM Big Data Platform incorporates a number of core capabilities including: IBM InfoSphere BigInsights based on Hadoop to provide distributed processing of large data sets across commodity servers; IBM InfoSphere Streams for real-time analytic processing (RTAP) of streaming data; IBM Vivisimo to automate the federated discovery of big data regardless of format or residence; IBM InfoSphere Warehouse along with the Smart Analytic Systems and Netezza for deep analytic insight with massive parallel processing in-database analytics; Netezza currently offers significantly more integrated and in-database analytics than any competitor on the market today; and IBM InfoSphere Information Server to provide high volume data integration and transformation delivering trusted information to critical business initiatives. The data warehouse remains an important part of the big data ecosystem and is further enhanced by the ability of Hadoop to consume new data types and sources for even better insight. Understanding that agility and maintaining SLAs are critical, IBM offers different types of solutions to meet the diverse analytic workloads. Together, customers can address multiple scenarios for high velocity and historical analysis, and implement better solutions for log analysis of critical systems, fraud detection, real-time personalization, and more.

In May of 2011, IBM committed a $100 million investment in big data analytics to help fulfill its encompassing vision. This was in addition to the over $14 billion invested in 28 analytic related acquisitions since 2005, including analytic intellectual properties (e.g., Vivisimo and Netezza), as well as the ongoing contributions it makes to the core Apache Hadoop open source project. While some research, open source, and commercial vendors have taken initiatives to package distributions that ease the installation of components of the Apache Hadoop ecosystem, IBM has made significant investments to further reduce the time required before organizations begin capitalizing on the value of their data, and reduce the total cost of ownership. Specifically, BigInsights builds upon Apache Hadoop adding unique technologies and management capabilities from across IBM to meet the needs of its enterprise customers.

What is important to note is that the technologies incorporated to make up IBM’s big data platform are not simply net new introductions, but are instead a culmination of many years of research and refinement in production environments. For example, technology used for advanced analytics has evolved through its inclusion with many IBM products (e.g., InfoSphere Warehouse, eDiscovery Analyzer and more). Similarly, the recently acquired Vivisimo represents over a decade of experience related to data navigation and visualization technologies for both structured and unstructured data.

To address the increasing need for supporting real-time data analysis of high-volume streaming data, IBM offers its InfoSphere Streams. Where other open source and commercial initiatives are only starting to address these technical challenges, IBM InfoSphere Streams was initiated back in 2002 as a collaboration between IBM and the United States Government. It has since been implemented by many organizations to build hundreds of applications in industries including government, telecommunications, financial markets, energy, healthcare and more. Generally available in 2009, “IBM InfoSphere Streams is built on a massively scalable clustered infrastructure (not dissimilar to Hadoop)” to provide customers with a data-in-motion analytics framework that manages multiple data streams on the fly and delivers nearly instantaneous analysis through advanced algorithms.

“ln motion” data insights or relevant facts can also be stored for further searching and discovery of additional insight. IBM InfoSphere Streams...
includes adapters to store in motion data into a Hadoop cluster (and traditional databases) and conversely let Streams applications read data from the cluster. This integration of IBM InfoSphere BigInsights and IBM InfoSphere Streams enables interaction between data at rest and data in motion; Integration with IBM Netezza is also supported to provide similar facilities for structured data. In short, it facilitates real-time analytic processing while using existing data to inform the intended response to the real-time information. This also allows observational models to adapt to changes in real-world conditions so streaming data can be analyzed more accurately.

For an example that helps illustrate the power IBM big data related technologies could have in real world applications, one needs look no further than IBM’s Watson. For those unfamiliar with Watson, in February 2011, Watson made history being the first computer to compete against humans on the popular television quiz show ‘Jeopardy’!. Further, Watson achieved a landslide win over champions Ken Jennings and Brad Rutter. What makes this achievement so fundamentally important is that Watson, receiving the questions at the same time as its human participants, needed to process and understand natural language, its subtle nuances and double entendres that are typical on Jeopardy!. Watson incorporated massively parallel analytic capabilities to understand the actual meaning behind the words, distinguish between relevant and irrelevant content, and then “buzz in” before its human counterparts through the same mechanical means. Watson was also able to learn from previously correct and incorrect decisions and applied risk analysis to future decisions. Specifically, if Watson wasn’t highly confident it had the correct answer, Watson wouldn’t buzz in.

At its core, Watson is a very sophisticated natural language Q&A system that incorporates data analytics and an insight engine designed to consult a wide variety of information sources to answer these questions. While Watson is not yet commercially available, the technologies used in its development are the same that have been used in corporate production environments. Further, Watson’s use of Hadoop, Lucene, and other technologies to store and pre-process its corpus of knowledge have culminated into the foundational capabilities now provided by IBM’s big data platform.

IBM can address all the data related needs of an analytics driven organization.
Introducing Vivisimo – Federated Discovery and Navigation for Big Data

IBM recently finalized its acquisition of Vivisimo and will incorporate these capabilities into IBM’s Big Data platform. Vivisimo solutions provide unique federated data discovery and navigation capabilities designed to accelerate deep insight from big data. Vivisimo has been solving big data challenges for nearly a decade, providing a robust system without a requirement for extensive custom development projects or large teams of support staff to monitor and manage deployments. Some of the most notable benefits of the Vivisimo solution include:

» Accuracy: Vivisimo provides more relevant results thanks to its unique position-based index technology and its ability to leverage usage and feedback information (e.g., tagging). The index is core to a federated discovery system and a key determinant of overall system performance. Specifically, in big data implementations, differences in index structure, size, management and other characteristics are magnified making any structural or design issues immediately apparent. Vivisimo’s indexing system is a “schema-less” framework that—by operating without reliance on a rigid schema—provides greater flexibility in how content is collected, stored and manipulated.

» Security: Vivisimo respects the native security profiles and access permissions of each application and underlying system. This ensures that users cannot view data that would not otherwise be visible to them had they signed directly into the target repositories. This security model extends to the field level allowing passages or fields within a document to be protected with their own permissions and updated without having to re-index a full document. Specifically, end users can see only the parts of a document for which they have access.

» Scalability: Vivisimo provides the highly elastic scalability necessary for big data deployments. It has been used to index trillions of records and petabytes of data, supporting a variety of strategies to accommodate large data volumes. Vivisimo's architecture enables deployment of automated agents for routine management, load balancing, failover and recovery. An implementation can contain any number of “collections” and can expand and contract with the changing demands of an application to allow optimal exploitation of data and resources. Graceful fail-over (without loss of data or performance) is built into the system using techniques such as master-master replication. Distributed deployments can offer a number of different deployment strategies and techniques, including, but not limited to: scaling “out” to a large number of servers with easy administration on commodity hardware; scaling “up” through parallelized data collection and indexing tasks, as well as index compression; and master-master replication allowing any member of a distributed server configuration to update indices.

» Integration: Vivisimo can provide customers with an extensive range of pre-built connectors for a wide range of sources, and a framework for building additional connectors when needed. With a unique ability to access enterprise applications as well as big data repositories such as Hadoop, Vivisimo facilitates the fusion of information from operational and knowledge management systems within corporate and government enterprises with any type of big data collection. This facilitates insights and analysis that would otherwise require manual merging or time-consuming duplication. This fusion of data can be performed “on the fly” at query time or through the creation of “virtual documents” that represent information from multiple sources. In either case, Vivisimo’s connectivity framework allows data to stay in the applications where it is managed, while data scientists (as well as researchers and business users) can focus on answering questions. Vivisimo also provides a federated framework that supports the ability to dispatch queries to outside the firewall, which cannot typically be indexed. Vivisimo has a long history in federation and provides deep capabilities for fusing indices from either big data or traditional data sources with federated search results.

Enterprise Value

Today, big data is not yet considered essential for many organizations; particularly for those not yet benefiting from the value tucked away in their continual data flows. However, as companies begin to capitalize on this information to drive better business insight and real-time decisions, its importance continues to encroach on the realm of critical. So too, big data increases the importance of enterprise level capabilities such as availability and scalability.

IBM has become synonymous with enterprise availability and scalability. Organizations have come to appreciate that solutions delivered by IBM meet these fundamental needs. Following in this vein, IBM InfoSphere BigInsights provides customers with a hardened Apache Hadoop solution that is thoroughly tested with each Hadoop release to meet these stringent demands. BigInsights Basic Edition is available at no charge, with the option to purchase support and services if desired. While IBM has added value on top of Hadoop, the open source components have not been altered, meaning there has been no “forking” of the Hadoop source code. IBM also provides support for third party Hadoop engine distributions (e.g., Cloudera) providing customers with the flexibility to select an implementation that best suits their needs, while still offering the opportunity to add the benefits provided by IBM InfoSphere BigInsights Enterprise Edition and the IBM Big Data platform.

IBM InfoSphere BigInsights Enterprise Edition adds extensive capabilities through additional components that complement Hadoop and significantly reduce the time to value of customer implementations while maintaining reduced IT staff requirements. More than integrated GUI-driven installation tools, IBM adds extensive value including text analytics and toolsets to speed development, customization, management and data visualization, all in an automated install.
IBM eases the governance of a big data implementation through a number of technologies. IBM InfoSphere BigInsights Enterprise Edition provides a web-based console for the installation and management of BigInsights. It facilitates the launching and publishing of applications, inspecting the status of jobs, navigating through the file system, and performing several other tasks. IBM helps improve overall availability with automated failure detection promoting cluster members as required, and administrators can perform maintenance on machines without shutting down entire applications. The addition of IBM InfoSphere Information Server facilitates comprehensive data integration that ensures data from BigInsights delivered to other systems is of the highest quality.

IBM InfoSphere BigInsight Enterprise Edition enables organizations to squeeze the most from existing Hadoop related hardware infrastructures, incorporating MapReduce optimization, intelligent job scheduling, support for data compression and more. For example, IBM provides its Intelligent Scheduler—in addition to the Hadoop included FIFO (first in first out), Fair, and Capacity based schedulers—to provide increased control with respect to cluster performance. This scheduler manipulates the Fair Scheduler to provide organizations with more control over scheduling and the customization of workloads to meet service level agreements. Workloads can be tuned to maximize resources for small jobs for faster completion, proportionally allocate resources (i.e., big jobs have a higher priority), or allocate resources based on a particular user.

During a MapReduce job, Hadoop divides the data into many pieces known as splits. Each split is assigned a mapper that is subsequently deployed for processing. What is important to note, is that while smaller splits can reduce the costs associated with a failure compared to larger splits, smaller splits (and hence more of them) introduces additional start-up overhead (i.e., more resource intensive) for each map task. IBM extends Hadoop with the concept of Adaptive MapReduce to make individual mappers self-aware and aware of other mappers, allowing a dynamic balance between smaller and larger splits to optimize performance.

When working with large volumes of data, it is often beneficial to compress data to help reduce storage related requirements and subsequently the amount of data travelling over the wire. Additionally, splitting this compressed data allows BigInsights to run jobs on multiple mappers. BigInsights includes support for a number of compression codecs including its own CMX compression. BigInsights CMX is based on the LZO lossless data compression algorithm for its noted speed and ability to be split, but eliminates the overhead management of an index file. While Hadoop does not natively support splittable text compression (meaning a text file must be decompressed as a whole and only a single mapper task can process it), the IBM supplied Jaql query language (for JavaScript Object Notation data interface format) natively addresses this issue creating multiple map tasks for processing a single file.

Many will note that with any compression and decompression, there will be a CPU expense. As such, organizations will need to find the appropriate balance between storage savings and CPU expense. If however, an environment is I/O bound, there should be enough cycles to spare for the compression/decompression process. The interesting side effect here, is that because of a decrease in the amount of data travelling over the wire, many organizations have seen a performance increase based on the increase in data flow (i.e., more data over the same wire given its compression).

In a standard Hadoop installation, clustered machines have unrestricted communication with one another. To address the security needs of the enterprise, IBM helps secure implementations out of the box by reducing the number of open ports, providing immutability, role based security, and more. For example, in a production environment, the cluster would be behind a firewall in a private network. A secured console is used (the installation wizard helps setup an LDAP credentials store), which includes a reverse proxy to retrieve resources on behalf of a client providing only a single access point (and port) to clusters behind the firewall.

End User Value

When dealing with big data, rarely do the varied sources fall neatly together in relational structures, perfectly ordered and ready for processing. Data sources can include text from social networks, image related data, sensor data, etc. The common use of big data is to take these unstructured (or semi-structured) sources and process them in a way to extract something meaningful. In short, this requires IT involvement, lengthening processing times and/or increasing learning curves. But what if Line of Business Managers could get access to the value hidden in this data without IT involvement, through familiar easy to learn interfaces?

IBM BigSheets provides non-technical users with a Web-based visualization tool similar in function to a standard spreadsheet type interface.
Maximizing the Value Provided by a Big Data Platform

June 2012

It facilitates the easy collection of data, extraction and analysis, and visualization through tag clouds, pie charts, heat maps, bar charts, and more, without a requirement for IT staff. Under the covers (i.e., hidden from the end user), Pig jobs are created which are compiled down to MapReduce jobs. Because there is no programming or technical assistance required, there are significant cost and time to value benefits for the organization and its users.

With IBM BigSheets, business managers can quickly see the big picture

Included in the IBM InfoSphere BigInsights Enterprise Edition, BigSheets facilitates the gathering of structured and unstructured data from web-based repositories (including a website crawler) and performs the detailed analysis to find its hidden value. Business managers can quickly see the entire picture, achieve a better understanding of its customers, research competitors, and be the first to discover relevant issues or trends.

Through IBM BigSheets, organizations now have the opportunity to mine the flood of information generated by consumers through various social networking facilities (e.g., Twitter) and websites. Using sentiment and semantic analysis, managers can quickly assess responses to new marketing campaigns in real-time, and adjust strategies as necessary. A real-life example of this technology is the USC Annenberg student created Film Forecaster. Collecting web-wide information (like comments or blog posts) about different movies, it was able to predict which films would be successful at the box office; it correctly predicted a clamour for “Hangover 2” that resulted in a $100 million opening over Memorial Day weekend. The tool has also been used to gauge sentiment around last year’s Arab uprisings, potential Republican presidential candidates, and to gauge overall sentiment amongst the MLB National League Championship Series.

While IBM BigSheets comes with a number of pre-built modules to jumpstart the identification of data sources, apply analytics and more, it still provides the necessary facilities to address specific scenarios. Custom development could include LanguageWare or IBM Classification Module Macros, Custom Extensions, and/or Insight Visualization development and more.

Developer Value

Apache Hadoop has become the de facto open source framework for dealing with big data at rest. MapReduce is the programming model implemented for processing these large data sets. What is important to note, is that any time a question needs to be asked in the Hadoop repository, a Java-based analytics application needs to be written. While IBM BigSheets does help to ease this burden for end users, BigInsights provides facilities for much deeper analysis. To help speed the development of custom applications and reduce the number of Java developers required to build Hadoop based solutions, IBM provides a number of toolkits, pre-built analytics, samples, IDE plug-ins and more.

To get developers implementing solutions more quickly, IBM provides a number of accelerators, toolkits and sample applications that cover common scenarios. For example, an Internet Toolkit for IBM InfoSphere Streams provides an operator that receives text-based data from remote sources using HTTP, generating an input stream from the content. A sample application illustrates a number of the parameters of the operator and an included Makefile shows how to compile the Internet toolkit with an application.

Mining data streams to extract relevant information or intelligence is a common criticality for the majority of stream processing applications. This typically involves applying models (learned from historical analysis) and complex algorithms in order to detect patterns of interest, critical in applications for fraud detection, intrusion detection, customer segmentation, and more. As such, IBM provides its InfoSphere Streams Mining Toolkit, integrating algorithms from the IBM InfoSphere Warehouse using the Predictive Model Markup Language (PMML). PMML is supported by a number of data mining software tools found in IBM’s big data platform (e.g., InfoSphere Warehouse, SAS Enterprise Miner, SPSS) and allows specifications for different mining models.

Industry related accelerators, such as the IBM InfoSphere Streams Financial Services Toolkit help reduce the time and effort required to develop applications, in this case Streams-based applications for the financial industry. The components of this financial toolkit provide pre-built solution models, pre-packaged access to popular function libraries, and annotated examples for leveraging existing customer assets. Including over 300 functions and operators and over 100 sample applications, components are available at multiple levels of abstraction to address different roles within an application. For example, input/output adapters are provided to translate data formats from other environments and protocols (e.g., Financial Information eXchange) and starter applications are available to further customize and enhance proprietary built analytics.

IBM also provides customers with toolkits for Advanced Text Analytics and more. The Advanced Text Analytics Toolkit, originally developed in 2004, has been included with many IBM products over the years (e.g., Lotus Notes, Cognos Consumer Insight, etc.) but remained hidden from end users. Today, it incorporates a number of areas (e.g., information extraction, natural language processing, and sentiment analysis), providing a text analytics language, processing engine, and pre-built text extractors (helps identify patterns such as a person’s name, address,
URL, phone number, etc.) allowing developers to directly leverage these powerful capabilities to distil valuable insight from unstructured and semi structured data. Both IBM InfoSphere BigInsights and IBM InfoSphere Streams share the same powerful text analytics capabilities helping minimize learning curves for those organizations that wish to analyze both data at rest and data in motion.

To help reduce the amount of coding and optimization developers face, IBM has long facilitated the use of declarative languages. In the realm of big data, IBM empowers customers through its Streams Processing Language, Annotation Query Language, and more, to automatically optimize application performance depending on the needs of the job (e.g., for CPU or for I/O). These declarative languages facilitate the development of “fit for purpose” applications without a requirement to understand the lower-level specific operators. When deployed, a compiler translates the high-level declarative language to executable code, making decisions to optimize its performance. This not only facilitates developer agility when solving challenges around programming for multiple cores, but ensures high performance and low latency processing no matter the developers skill level.

IBM also provides visual tools through an Eclipse-based IDE and related plug-ins to help speed development through syntax highlighting and automatic error detection. When using the Stream Processing Language, Streams Studio helps developers create, edit, visualize, test, debug and run applications. With respect to advanced analytics, the development environment makes it possible to create and deploy Unstructured Information Management Architecture (UIMA) annotators in just a few clicks, performing everything from simple dictionary lookups to more sophisticated syntactic and semantic text analysis. Finally, the Eclipse-based IDE also facilitates the creation of low-level application components in C++ and Java that can be incorporated and reused in applications.

Simply put, IBM provides organizations with a comprehensive toolset to significantly reduce developer effort. Support for high-level declarative languages helps reduce learning curves associated with lower-level operators, while the collection of assets provided through the IBM toolkits and the availability of sophisticated pre-built text analytics further reduces development times.

**Administration Value**

While many vendors are packaging Hadoop distributions to help address the challenge of installation and configuration, few are addressing the new challenges around monitoring, management, and deployment of, and to, these complex environments. Big data and its inherent infrastructure scale-out requirements combines a number of technologies including distributed file systems, distributed databases, distributed caches, and distributed virtual machines, forming an incredibly complex runtime environment.

Through an AppWorld or AppStore paradigm, IBM provides the facilities for third party developers to build big data applications and easily make them available for others to browse, purchase, and execute. Users can add the necessary input parameters to meet their specific requirements and monitor the status of their jobs.

Rather than deal with multiple tools and interfaces required to view and manage these various distributed technologies, a unified console and cluster-wide provisioning capabilities to simplify their management would better suit administrators. IBM provides organizations with easy to use Web-based consoles to administer big data infrastructures. This includes monitoring the health of individual nodes, job status (i.e., applications) and the content within the file system. For those administrators familiar and comfortable using the original individual open source software administrative interfaces, these are still made available. To provide further value, IBM provides access to these interfaces through a single console without a requirement to access each interface individually.

IBM InfoSphere Streams offers both command line and graphical interfaces to monitor and administer a Streams runtime and applications. As Stream applications are submitted, the runtime determines where to best deploy the operators in order to meet the resource requirements of both newly submitted and already executing specifications, continuously monitoring the state and utilization of these resources. The Web-based console provides granular health metrics to monitor the environment across a distributed collection of hosts. Operator capabilities support the creation, starting, and stopping of instances on nodes in a cluster and the ability to add or remove processor nodes to or from a cluster. This allows administrators to perform the necessary maintenance on the operating system or on data streams without shutting down the entire InfoSphere Streams application improving overall availability.
IBM solutions are well known for their enterprise availability and scalability achievements. IBM also has a long legacy of helping organizations integrate and capitalize on existing enterprise assets. When dealing with big data, using these assets as data sources is of fundamental importance. It is for this reason that InfoSphere BigInsights supports data exchange with a number of sources including IBM Netezza, IBM DB2 (for Linux, UNIX, and Windows), RDBMS via JDBC, InfoSphere Streams, InfoSphere Information Server (DataStage), R Statistical Analysis Apps and more. Through pre-built facilities and tools such as the DataStage ETL tool to push and pull data to and from BigInsights, administrators and developers can optimize insight value, capitalizing on the structured information already used to make so many important operational decisions. Similarly, InfoSphere Streams also provides several edge adapters that can connect to external data sources (e.g., IBM DB2, IBM SolidDB, and ODBC) for the purpose of consuming or storing data.

More than simply providing a distribution that eases the installation of a Hadoop cluster, IBM provides customers with an extensive set of user interfaces and self-regulating optimizations to ease administrative burden. Support for LDAP security eliminates requirements to administer multiple authorization repositories, while pre-built integration with existing data assets reduces custom integration efforts. Where many vendors are simply focused on the implementation of a Big Data infrastructure, few can provide such simplicity of deployment and monitoring of applications; natively built or purchased.

In Closing

2012 is expected to see a significant uptake in solutions that help organizations extract insights currently hidden in their data. It's no wonder given the research showing those that capitalize on this information see increases in productivity, profitability, and market value. Similarly, real-life applications are already improving focus of marketing campaigns, helping improve healthcare, optimizing production, or helping monitor and manage the sentiment behind a brand.

While there are many vendors putting together distributions to ease the installation of the Hadoop framework, few address many of the needs of an enterprise organization. Fewer still can provide the tools to help speed development, ease governance, and support an organizations multifaceted approach to discovering the insight and value locked away in historical, recently stored, and live in-motion data. As the value provided by this data continues to become critical, organizations need to be positioned with a vendor that will allow them to grow to meet new demands, in a manageable manner, helping keep administrative and other collateral costs to a minimum.

IBM's vision for big data not only goes beyond the installation of the Hadoop framework to meet the security, support, and availability and scalability needs of the enterprise, it provides an entire platform of integrated tools, solutions and support to help organizations achieve that competitive advantage while remaining budget conscience.
About Branham Group Inc.

Branham Group is a leading industry analyst and strategic consulting company servicing the global information and communication technology (ICT) marketplace. Branham Group assists ICT companies and related institutions in achieving market success through its custom consulting services (Planning, Marketing and Partnering), and through its multi-client research subscription programs (Digital Health, Outsourcing, and Green IT). Branham also produces an annual listing of the top ICT companies in Canada (www.branham300.com), tracks the Canadian Outsourcing industry (www.branhamoutsourcing.com) and monitors over 450 Healthcare ICT vendors.

For more information regarding Branham Group, please visit www.branhamgroup.com.

Branham Group Inc.
45 O’Connor Street, Suite 1150
Ottawa, ON
Canada K1P 1A4
Tel: 613.745.2282
Fax: 613.745.4990
www.branhamgroup.com

The views, analysis and opinion expressed in this document are those of Branham Group Inc. and are based upon information obtained from public sources that are believed to be reliable. No independent verification has been made as to accuracy and completeness of such publicly available information. To the best of Branham Group’s knowledge and belief, the information contained in this document is correct as on the date of this document. The opinion expressed in this document is Branham Group’s current opinion and is subject to change without notice.

This documentation is provided “as is” without any warranty whatsoever and to the maximum extent permitted, Branham Group disclaims all implied warranties, including without limitation the implied warranties of merchantability, non infringement and fitness for a particular purpose, with respect to the same. Branham shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. Branham Group shall not be responsible for any damages, including without limitation, direct, indirect, consequential or incidental damages, arising out of the use of, or otherwise related to, this documentation or any other documentation. Nothing in this documentation is intended to be a recommendation of a specific product or company or intended to constitute an offer to buy or sell, or the solicitation of an offer to buy or sell, any product.

Reproduction and/or distribution of this document in any form without the prior written authorization of Branham Group Inc. is strictly prohibited. Copyright © 2012 Branham Group Inc.

Copyright © 2012 Branham Group Inc. Branham and the Branham Group logo are trademarks or registered trademarks of Branham Group Inc. in Canada and other countries. Other product and company names mentioned herein may be trademarks of their respective owners.