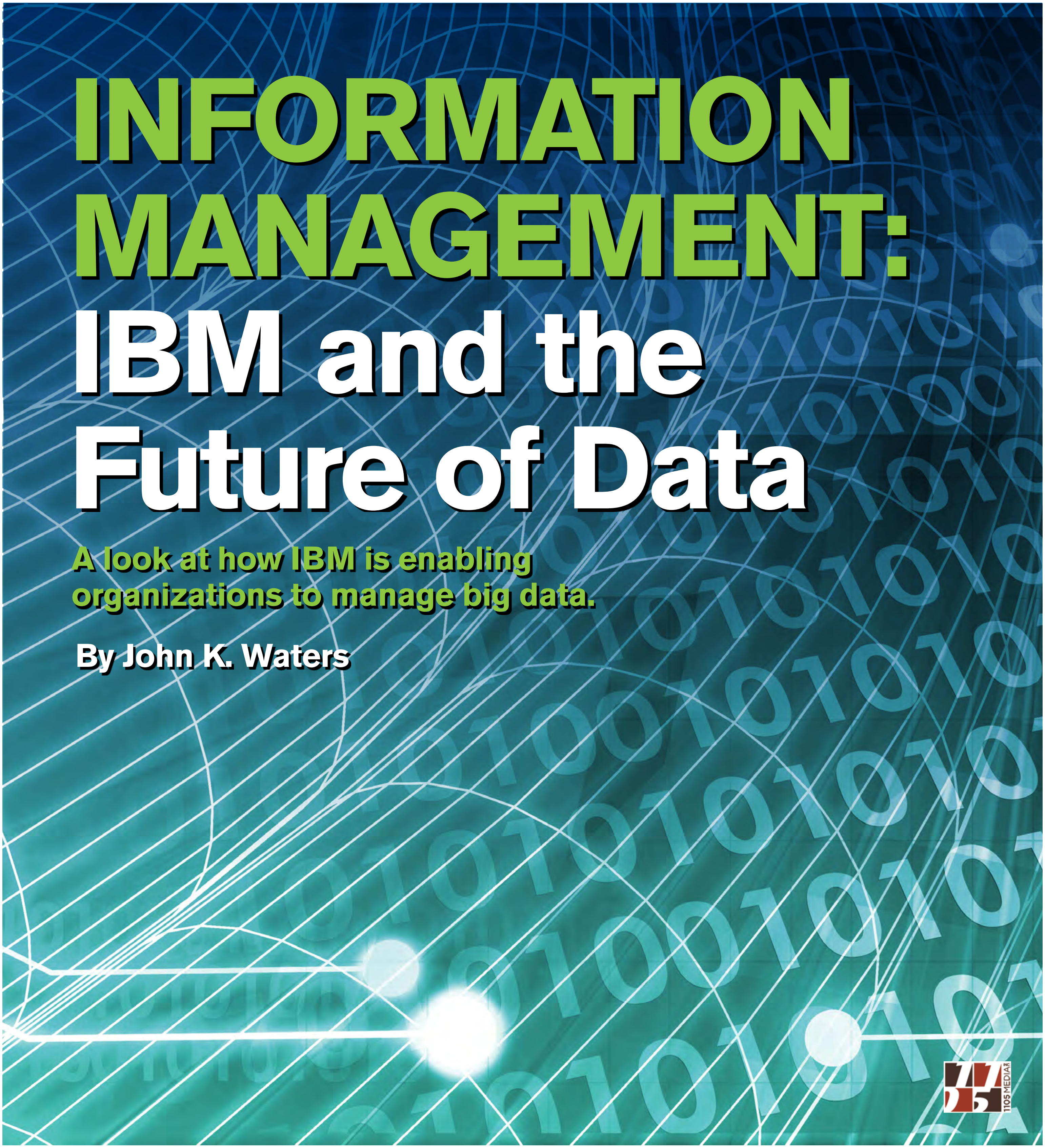
CONTENT BY

ADTMAG
APPLICATION DEVELOPMENT TRENDS

SPONSORED BY

IBM®



INFORMATION MANAGEMENT: IBM and the Future of Data

A look at how IBM is enabling organizations to manage big data.

By John K. Waters

1105 MEDIA

People and machines are generating data at a staggering rate–2.5 quintillion bytes daily, according to IBM's latest estimate. It's coming in fast, it's coming from an ever expanding array of sources, and it's redefining enterprise information management. "Big data," the umbrella term for these large, difficult-to-manage data sets, showed up in the enterprise lexicon only a few years ago, but it now represents as pressing a challenge as the enterprise has ever faced.

A number of vendors currently offer a range of tools and technologies that address varying aspects of this information management challenge. One of the broadest of the big data solution portfolios comes from IBM, which is no stranger to data management and distributed computing platforms. This paper looks at IBM's big data offerings within the context of recent trends in this rapidly evolving space.

One organization's petabyte is another's gigabyte.

THE TRENDS

Industry watchers point to a number of trends that led to the birth of big data and continue to support its growth. Three in particular were clearly essential:

Plummeting cost of storage

"Big" is a relative term, of course; one organization's petabyte is another's gigabyte. But the plummeting cost of storage has made it possible for organizations of all sizes to collect volumes of data previously unheard of. Price wars among storage service providers and increased adoption of cloud storage has made a place for big data in the enterprise. And techniques such as storage virtualization, de-duplication, thin provisioning, and automatic tiering have made storage management more efficient. It has simply never been cheaper to collect and store data. In fact, it's no longer a question of "what data should I keep?" because it's now economically feasible to simply keep everything, just in case.

"Electronification" of communication

Big data is about more than just volume. These enormous data sets also come from a variety of sources, and they're often being collected in near real time. (Thus the Three Vs–volume, variety, and velocity.) Most of the data now being generated is unstructured or semi-structured: email, log records, clickstreams, social media data, news feeds, electronic sensor output, recorded help desk calls, and even some transactional data. Thanks to what has been called the "electronification of communication," much (maybe most) of our interactions are now subject to electronic capture. Consequently, the ease of data acquisition of an enormous range of data types has increased by an order of magnitude.

New tools

Add to these trends the arrival of Apache Hadoop, the open-source platform for data-intensive distributed computing, and shortly thereafter, a rapidly evolving and innovative Hadoop ecosystem. Pioneered by Google and Yahoo, Hadoop enables the distributed processing of large data sets across clusters of commodity servers. It's designed to scale up from a single server to thousands of machines with a high degree of fault tolerance.

In a recently published report, Forrester analysts Mike Gualtieri and Noel Yuhanna declared that Hadoop's momentum as a key enabler of big data solutions and management is "unstoppable" ("The Forrester Wave: Big Data Hadoop Solutions, Q1 2014"). The roots of the technology are growing "wildly and deeply into enterprise data management architectures," they wrote, and it has become "an essential data management technology."

Ten years ago it simply would not have been possible to manage petabytes of information and do anything meaningful with them.

IBM Senior Product Marketing Manager Gord Sissons put it this way in a recent blog post: "When it comes to big data projects, time to value matters. Gathering and storing vast amounts of data, as tough as that is, is actually the easy part; what differentiates Hadoop distributions is the available tooling for manipulation and analysis of data."

Ten years ago it simply would not have been possible to manage petabytes of information and do anything meaningful with them. Hadoop makes it practical to do things with previously unimaginable amounts of data. It's fair to say that "Hadoop" has become synonymous with "big data."

IBM's ENTERPRISE-GRADE HADOOP

IBM is not generally perceived as a major supporter of the Hadoop community, and yet the company's approach to information management in the age of big data does, in fact, embrace the open-source framework in a very big way. IBM's InfoSphere BigInsights big data platform is based on standard Hadoop, and bundles familiar Hadoop tools alongside a number of IBM enhancements. The company offers a Quick Start Edition (free download, selected features), a Standard Edition (basic Hadoop capabilities), and an Enterprise Edition (the full complement of enhancements).

IBM really flexes its big-data-in-the-enterprise muscles with the enterprise version of InfoSphere BigInsights. The list of enhancements in this version, which the company describes as "Enterprise-Grade Hadoop," includes components designed to reach beyond basic Hadoop capabilities to address enterprise-specific big data demands:

Big SQL: This is an ANSI compliant massively parallel processing SQL engine that deploys directly to the physical Hadoop Distributed File System (HDFS) cluster.

BigSheets: This is a browser-based analytic tool that uses a familiar spreadsheet metaphor to enable business-oriented users to manipulate and visualize large data sets in Hadoop.

Text Analytics: IBM addresses one of the trickiest challenges associated with big data mining with an advanced text analytics component that can extract information from unstructured or semi-structured text.

Big R: Focused on the R statistically language, this library of functions provides end-to-end integration between R and the BigInsights platform.

Adaptive MapReduce: This a component addresses the challenge of improving the Hadoop framework's performance and scheduling agility by allowing the orchestration of distributed services on a shared grid in response to dynamically changing workloads.

GPFS: IBM solves one of the knottier Hadoop challenges with a high-performance clustered file system that is POSIX-compatible and also supports HDFS file access semantics.

As you peer into very large data sets, patterns emerge.

These components address three key challenges of big data management: extracting meaningful insights from the data, providing non-technical business users with easy access to those insights, and reducing the complexity of the process.

Insights from Raw Data

As you peer into very large data sets—the larger, the better—patterns emerge, nuances appear, and trends reveal themselves. Because of the size of the data sets involved, big data can yield otherwise unknowable harvests of insights. This is the fundamental goal of big data management, and it's how enterprises get to business value.

Surfacing those patterns and nuances requires sophisticated data analytics tools, and BigInsights comes with two strong analytics components: Text Analytics and Big R.

Text analytics in particular has become an essential big data management capability. It allows organizations to make sense of the ocean of unstructured and semi-structured data generated by things like customer service interactions, call-center conversations, and social-media activity. Designed to simplify text analytics and natural language processing, IBM provides text analytics tooling similar to that popularized by IBM's Watson artificially intelligent computer system (the Jeopardy winning computer).

Text analytics has become an essential big data management capability.

Big R integrates the R statistical language with Hadoop. R is a popular language among data scientists, but the data models are often memory constrained. Developers work around this limitation in Hadoop environments by writing Map and Reduce logic to distribute R-based code. Big R eliminates the need for this work-around by providing a comprehensive set of analytic functions callable using familiar R language semantics that auto-parallelize across the Hadoop cluster. Essentially, Big R insulates developers from the fact that their code is running on a parallel framework. Also, Big R works with existing open-source R tools and downloadable CRAN projects (Comprehensive R Archive Network) available from r-project.org.

BigInsights also employs what IBM calls "In-Hadoop Analytics." Sissons explained the process this way: "A key design principle of Hadoop is to minimize data movement by vectoring compute tasks to the nodes housing relevant data blocks. Moving analytics to data is hardly a new idea. So what are we talking about exactly? The trick is to be able to run higher-level analytic functions that parallelize easily and transparently, but respect data locality in a fashion that hides complexity from the developer. Big R is a framework for doing exactly this: pushing down R language analytics into the distributed dataset. Similar analytic functions are embedded in IBM Big SQL and other BigInsights facilities so that users can simply embed analytic functions in queries, and let BigInsights do the work."

Ease of Access

In a modern enterprise, data is no longer the exclusive province of the DBA. Multiple, non-technical users need access to the insights the organization is pulling from its big data investment. IBM addresses this need in BigInsights with two components: Big SQL and BigSheets.

There are a number of SQL-on-Hadoop implementations on the market today. IBM, which invented SQL, entered this market fairly recently with Big SQL, essentially optimizing its venerable DB2 product for Hadoop. The result is a fast, ANSI compliant implementation that runs natively over Hadoop data formats and supports sophisticated query optimization, memory management, and rich SQL analytic functions so that standard queries run without modification. SQL federation means that users can formulate queries that join data from multiple sources, including InfoSphere BigInsights, and other IBM and third party offerings, such as Teradata, Oracle, DB2, Netezza, and others.

There are a number of SQL-on-Hadoop implementations on the market today.

IBM wanted to demonstrate how the performance of Big SQL compared with its nearest competitors, but found that Hadoop did not support some of the requirements the leading big data benchmark, the Transaction Processing Council's TPC-DS. The TPC-DS was designed to test a traditional data warehouse, so IBM was essentially forced used the key parts of the TCP-DS to create its own benchmark: Hadoop-DS. Hadoop-DS, is not an official TPC-DS benchmark, but it uses the TPC-DS schema and data generation tools, uses all 99 queries, meets the multi-user requirement, and has been audited as a non-TPC benchmark by an approved TPC auditor.

IBM used the Hadoop-DS benchmark to compare Big SQL with Cloudera's Impala and Hortonworks' Hive. Big SQL was able to run the full query set (99), which represented a retail business workload, 12 with minor modifications allowed under TPC-DS. Impala was able to run 52 queries, 35 out-of-the-box and 17 with allowable minor modifications. Hive ran 58 queries, 32 out of the box, and 26 with allowable minor modifications. Of the queries that all three vendors could run (46), Big SQL was 3.6 times faster than Impala, and 5.4 times faster than Hive.

Reducing Complexity

How do you manage your big data in such a way that you don't need a team of PhDs to run the cluster? How, in other words, do you keep something so complex simple? The BigInsights platform addresses this challenge with two features: the GPFS file system and Adaptive MapReduce, a feature that makes it easier to manage MapReduce jobs.

At its core, Hadoop is a combination of Apache MapReduce and the Hadoop Distributed File System (HDFS). MapReduce is a programming model for processing and generating large data sets. It supports parallel computations on so-called unreliable computer clusters. HDFS is designed to scale to petabytes of storage and to run on top of the file systems of the underlying operating system. POSIX (portable operating system specification for Unix systems) is commonly considered a standard file system. Linux, Unix, even Windows NT file systems are POSIX-compliant. HDFS, however, is not, and developers have to code specifically for Hadoop if they want their programs to run on it.

At its core, Hadoop is a combination of Apache MapReduce and the Hadoop Distributed File System.

BigInsights smooths over this mess of complexity with GPFS FPO (General Parallel File System), a variant of IBM's high-performance clustered file system. GPFS is widely deployed in large supercomputing environments. In BigInsights, it fully implements HDFS interfaces and semantics, which means that programs can use the same file system for Hadoop and non-Hadoop data. GPFS FPO works under the covers, so it looks like standard HDFS to Hadoop applications.

Most big data apps rely on Hadoop's MapReduce core framework to enable parallelism. Adaptive MapReduce is IBM's approach to improving the framework's performance and scheduling agility. It's an optional feature that can be deployed as an alternative to the standard open-source Hadoop scheduler. It uses a low-latency scheduling algorithm to dispatch Map and Reduce tasks out to nodes, while minimizing the amount of data movement.

IBM has claimed that Adaptive MapReduce can accelerate many types of MapReduce applications, while maintaining full application capability. In 2013 the company called on technology research and testing tools company STAC to compare pure Apache Hadoop with Adaptive MapReduce. In the audited benchmark test, Adaptive MapReduce outperformed Apache MapReduce on average by a factor of four, running a large-scale social media workload.

BigInsights also comes with a set of pre-packaged "accelerators" for popular use cases, such as machine data analytics, social media analytics, and the extraction and analysis of text. The idea is to get to market faster by leveraging pre-written code supporting common use cases.

BigInsights Case Studies

IBM has a number of examples of customer implementations of BigInsights. Two in particular are illustrative of range of the platform's capabilities.

Vestas

Vestas Wind Systems A/S is a Denmark-based wind energy company that uses one of the world's largest supercomputers and big data modeling to process the truly vast amounts of data it needs to accurately select optimal sites for its wind turbines. The company has installed more than 43,000 land-based and offshore wind turbines in 66 countries on six continents. Today, Vestas installs an average of one wind turbine every three hours, 24 hours a day, the company says, and its turbines generate more than 90 million megawatt-hours of energy per year.

The idea is to get to market faster by leveraging pre-written code supporting common use cases.

To pick the best spots for these wind-harvesting machines, Vestas gathers data from 35,000 meteorological stations around the world, and from its own in-place turbines. The result is a "wind library" that also helps forecast wind and power production for Vestas' customers. The company's engineers use this library to create grid patterns to establish exact wind flows in particular locations. Using computational fluid dynamics models, the engineers have been able to tighten the grid resolution from 17x17 miles to about 32x32 feet. Tightening a grid makes it more accurate, but smaller grids means more data. This company expects its wind library to grow to more than 10 fold to include a larger range of weather data over a longer period of time–to between 18 and 24 petabytes.

Growing that library was easy compared with the challenge of extracting useful, accurate information from it, said Vestas VP Lars Christian Christensen. To extract that knowledge, Vestas used IBM InfoSphere BigInsights software running on IBM System x iDataPlex system servers as its core infrastructure. "Before, it could take us three weeks to get a response to some of our questions simply because we had to process a lot of data. We expect that we can get answers for the same questions now in 15 minutes," Christensen said.

Data currently stored in the Vestas wind library comprises nearly 2.8 petabytes and includes more than 178 parameters, such as temperature, barometric pressure, humidity, precipitation, wind direction, and wind velocity from the ground level up to 300 feet—not to mention the company's own recorded historical data. The company expects to use this data in the future to predict global deforestation metrics, satellite images, historical metrics, geospatial data, and data on phases of the moon and tides.

Independence Blue Cross

Independence Blue Cross (IBX) is a leading health insurer in southeastern Pennsylvania serving via its affiliates more than 7.5 million people in 19 states, including 2.2 million locally. The company does business with a wide range of health care providers, most of whom are switching from paper to electronic health records. However, the level of technical readiness among those organizations varies, and IBX receives medial documents in differing formats, including text files, XML files, faxes, PDFs, and scanned documents.

IBX discussed this challenge in a public session at the annual IBM Insight Conference in a presentation on the company's use of the BigInsights Hadoop infrastructure to mine these large data sets for actionable intelligence. The text analytics capabilities of the platform were of particular interest to the company, because the medical records it hoped to mine are largely unstructured and semi-structured.

Text Analytics comes with a set of pre-defined extractors based on IBM's Annotation Query Language.

Text Analytics comes with a set of pre-defined extractors based on IBM's Annotation Query Language (AQL), a declarative language used to identify and extract textual information. It's part of a text analytics engine developed by the Watson team. AQL defines domain-specific language—such as medical terminology—and provides users with extractors for information based on syntax. The ability of AQL to express jargon has made it possible to parse meaning from a mountain of medial records.

The BigInsights platform is running Apache OpenNLP, a machine-learning-based toolkit for processing natural language text; the open source Lucene information retrieval software library and its Solr enterprise search platform; Gate, one of the leading toolkits for text mining and information extraction; and the Weka collection of machine learning algorithms for data mining.

IBX is using the text analytics capabilities of BigInsights to score these records and prioritize them for human review. Among other things, the company is using text mining to cross reference warranty recalls from manufacturers of implants and prosthetics and identify surgery involving recalled items. The company uses the system to schedule report alerts with potential identified members that match the recall manufacturers. And it uses BigSheet to present the data. ESJ

IBX is using the text analytics capabilities of BigInsights to score records and prioritize them for human review.

Journalist and author ***John K. Waters*** *has been covering the information technology beat from Silicon Valley and the San Francisco Bay Area for a range of print and online media for more than 20 years. As Editor-at-Large for Application Development Trends he writes the WatersWorks blog and contributes news and feature stories about enterprise software development, management, and security. He also writes software reviews for Law Technology News and contributes regularly to The Technology Horizons in Education Journal and Campus Technology. His work has appeared in Redmond Developer News, Virtualization Review, Microsoft Certified Professional, and many others.*

John has written more than a dozen books, including a new entry in the Adams Media "Everything" series: The Everything Guide to Social Media. Early in his career he wrote a number of computer game guides, including the original strategy guide for the blockbuster "Diablo." He also co-scripted the documentary film, Silicon Valley: A 100 Year Renaissance, which was narrated by the late Walter Cronkite.