INSIGHT AND ADVICE ON BI, ANALYTICS AND DATA WAREHOUSING

MARCH 2012, VOLUME I, ISSUE 3

TBI

Trends + Strategies

BIG DATA ANALYTICS, OPERATIONAL BI MAY REQUIRE ARCHITECTURAL CHANGES

FOR MANY business intelligence and data management professionals, it has been both exciting and somewhat intimidating to watch the evolution of BI and analytics. The challenge this evolution presents is how to best use data to achieve business objectives. At the same time, organizations are trying to sort out the various options for enabling operational BI and advanced analytics and exploiting "big data" and other forms of unstructured information—options that include cloud-based BI implementations, columnar databases, Hadoop, NoSQL data stores, data virtualization and more.

There are strategic decisions that every organization must make so that their data architectures can handle today's and, more important, tomorrow's needs—from both a business and IT perspective. The questions being answered by data are now tied to strategic business goals, requiring an approach that marries the evolving needs of the business with the right IT and BI capabilities. In our lead article, BeyeNETWORK expert William McKnight provides advice and questions to ask to help future-proof an information management infrastructure.

Also in this issue, Mark Brunelli, SearchDataManagement.com's senior news editor, reports on the connectors being rolled out by various vendors for moving data between relational databases and Hadoop clusters. It's all part of the effort to make it easier to incorporate big data technologies, such as Hadoop, into the IT mainstream.

The "fuel" for analytics is data, and the quality of that data plays a key role in the effectiveness of the results. BeyeNETWORK expert David Loshin looks at how data flaws can affect BI results and provides five steps that organizations can take to improve the quality of analytical data. ■

JEAN SCHAUER
Editor in Chief, BeyeNETWORK

NEW BI DEMANDS PUSHING DATA ARCHITECTURE LIMITS

Big data and operational BI place new demands on information architectures. What worked in the past may not be the best choice for the advanced analytics that are poised to provide substantial business value. By William McKnight

WE ALL KNOW change is constant in IT. But when we are responding to demands to extract value from new forms of the most important asset in modern competitive business—information—and are busily managing the resulting increased data volumes, change is both constant and fast-paced. Don't get left behind: With "big data" increasingly taking hold, the five-year future is going to bring major transformations in the way that information systems are built to support business intelligence (BI) and analytics applications.

It's going to be more about the net additive effect of new possibilities than about dismantling current technology investments. It's going to be about seriously using information for BI and analytics through the exploitation of all possible data, including data that is larger than anything you've experienced to date, providing less value "per byte" and requiring new data management methods. It's going to be about exploiting information sooner in its lifecycle—as soon as possible.

Maturity in business intelligence is correlated to maturity and success in business. BI maturity is a worthy progression to understand, and working to increase maturity levels is a recommended strategy for every BI team.

Most companies begin their BI journeys because of a need for reporting. In many cases, reporting requirements can overwhelm, either technically or politically, the operational systems executing business transactions. That often leads

to the development of additional structures—data warehouses and marts—to house transactional data for reporting and analysis. Integrating that data, regardless of where it comes from, becomes critical, as does improving the overall quality of the information.

DEVELOPING A MORE EXPANSIVE VIEW OF BI

Eventually, monthly, weekly and even day-after reporting becomes insufficient, and accessing the valuable information being consolidated in data warehouses becomes a requirement on an intraday basis. An organization can then go full circle to doing whatever it can to support BI operationally.

With post-operational warehousing systems being bound by the extract, transform and load cycles that need to be performed there, real-time BI brings data analysis directly into operational systems. Business intelligence previously was viewed as "whatever we do to data in the data warehouse," but now BI is being expanded into the discipline of data usage and exploitation, wherever that data may reside.

For more and more companies, another aspect of BI maturity involves coping with new data types, which usually means harnessing the "big three" forms of unstructured or semi-structured big data: sensor, social media and Web activity data. It also can mean an organization availing itself of the syndicated data marketplace and bringing external information into its BI systems.

Before making big investments, assess your existing information management architecture to determine if it will meet the new BI demands of the enterprise. Suggested questions to ask as part of that assessment include the following:

- Is all company data, including unstructured and semi-structured data, under management?
- Is there a solid data quality program in place to ensure the distribution of high-quality information?
- Are all BI queries processed in post-operational batch systems now? Could they be done, and decisions made, in operational systems instead?
- Do new query requests automatically require using a particular system or type of BI technology, such as multidimensional cubes?
- Is the current usage of BI systems going to continue to scale? Is it time to consider alternatives?
- Can data marts, data warehouses and cubes with overlapping functionality be consolidated? And have the potential savings of doing so been calculated?

BI CHALLENGES—AND WAYS AROUND THEM

Maintaining sufficient performance levels is one of the top challenges in BI today, especially as data volumes grow. Alternative technologies for managing structured data—such as columnar databases, in-memory databases and data warehouse appliances (or a combination of them)—can provide additional performance on specific workloads beyond what conventional relational databases are capable of delivering.

Data integration also remains a major challenge. For many organizations, the answer may be the "perpetual short-term" solution of data virtualization, which can bring together data from separate, technically distinct data stores without requiring the information to be consolidated in a data warehouse.

Interest in the cloud is high as a place to store the data behind BI systems and manage the integra-

INFORMATION ARCHITECTURE: PREPARING FOR NEW REQUIREMENTS

KEY QUESTIONS to ask when evaluating whether your information management architecture is ready to handle emerging needs include the following:

- **Can business decisions that data warehouse and business intelligence (BI) systems support be made earlier than when the data comes together in them?**
- **For "column-selective" queries that need to access only a small number of the columns in database tables, can columnar databases be supported?**
- **Are NoSQL and "big data" technologies being implemented or considered within the organization? If so, is there a plan for accommodating and managing them?**
- **Are cloud BI and data management options on the table?**
- **Has the syndicated data marketplace been surveyed for information that aligns with company needs and could be pulled into internal systems?**
- **Has data virtualization been adopted for integration needs between different systems? Is it being used in cases where physically moving data would provide better performance? ■**

tion and data access components. Security fears and integration with on-premises data remain big concerns for many prospective users. Nonetheless, cloud and virtual deployments undoubtedly will house upwards of one-third of corporate databases within the next five years.

Centralizing management of master data is potentially valuable to BI initiatives as well, because it replaces individual efforts and creates a governed process for master data management (MDM) along with the infrastructure required to support distribution of the master data to different systems. But there are architectural considerations to address when evaluating an MDM program. For example, does each system need to build its own master data? Is one system's master data worthy of being used more broadly? Are the owners of master data ready for the responsibilities of providing it beyond their individual systems?

While few companies have a budget for "innovation," the good news is that advanced business intelligence is not really about innovation. It's about business, and when done well it provides a high return on investment even in the short periods that are demanded of IT investments these days. But there's the rub—it must be done well.

By asking some vital questions (see "Information Architecture: Preparing for New Requirements," page 5) before investing in information management technologies to support BI activities, organizations should be well prepared to manage their information assets effectively. Coupled with a focus on enabling better business outcomes as a result of efforts to increase a company's level of BI maturity, the resulting information management architecture will be able to support the BI data demands of the foreseeable future. ■

You can find more articles by William McKnight on information management topics in his BeyeNETWORK expert channel.

William McKnight, president of McKnight Consulting Group, is a strategist and information architect specializing in information management, business intelligence, data warehousing and master data management. McKnight has authored hundreds of articles and white papers and speaks regularly at conferences and seminars. He can be reached at wmcknight@mcknightcg.com.

HADOOP CONNECTORS HITCH DATABASES TO BIG DATA CLUSTERS

Various software vendors have begun offering connectors designed to let users easily transfer data between Hadoop clusters and relational databases. What can they best be used for? By Mark Brunelli

SOFTWARE VENDORS have gotten the message that Hadoop is hot—and many are responding by releasing Hadoop connectors that are designed to make it easier for users to transfer information between traditional relational databases and the open source distributed processing system.

Oracle, Microsoft and IBM are among the vendors that have begun offering Hadoop connector software as part of their overall "big data" management strategies. But it isn't just the relational database management system (RDBMS) market leaders that are getting in on the act. Data warehouse and analytical database vendors such as Teradata and Hewlett-Packard's Vertica unit have also built connectors for linking Hadoop to SQL databases, as have data integration vendors like Informatica and Talend. Vendors of Hadoop distributions, including Cloudera and MapR Technologies, are in the connector camp as well.

Organizations mulling the possibility of using connectors to link conventional database systems to Hadoop clusters should think about "where the best place is to analyze or search or sort or whatever it is you're trying to do with your data," said Rod Cope, an experienced Hadoop user who is chief technology officer at OpenLogic Inc. in Broomfield, Colo.

CONNECTOR USES, ISSUES

OpenLogic uses Hadoop in combination with HBase, a column-oriented NoSQL database that is part of the Hadoop framework, to keep track of open source software projects

around the world. It's all part of the company's flagship service, which helps corporate customers audit software applications to verify that the use of embedded open source code complies with relevant licenses. OpenLogic has yet to deploy any connectors, but Cope has looked closely at the technology—for example, as a possible means of moving infrequently accessed data from a relational database to HBase for archiving.

The connectors don't magically solve all of the issues involved in such pairings, according to Cope. He cautioned that prospective users should be aware of just how long it can take to load data from a database into Hadoop. "It's easy for people to forget when you have truly big data that anything you do with it takes a very long time," Cope said. Typically, he added, "it's not the Hadoop side that's slow; it's wherever you're trying to load it from."

David Menninger, an analyst at Ventana Research in San Ramon, Calif., said the Hadoop Distributed File System and specialized databases built on top of it are good at providing users with a place to manage and analyze information that doesn't fit neatly into a traditional RDBMS or data warehouse. That might include machine-generated forms of big data, such as application, search and website event logs, plus social media information, mobile phone call detail records and other "stuff that just simply wouldn't normally be thought of as structured relational information," Menninger said.

One of the most common use cases for a Hadoop connector, he said, involves an organization

WHAT IS HADOOP?

HADOOP IS A Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Developed by the Apache Software Foundation as an open source project, Hadoop was originally based on Google's MapReduce programming model, which lets developers break applications down into numerous small tasks that can be run in parallel on different computing nodes in clustered systems. Hadoop makes it possible to run applications on systems with thousands of nodes and terabytes of data; the Hadoop Distributed File System manages storage, facilitates data transfers among the nodes and enables the cluster to continue operating uninterrupted if individual nodes fail. —WHATIS.COM

using a Hadoop system to extract a small amount of structured analytical information from a much larger amount of unstructured data, then transferring that information to an RDBMS for further analysis and reporting with business intelligence tools.

EVERYTHING IN ITS RIGHT PLACE

"The reason you put it into a relational database is because you can't easily report on Hadoop data sources today," Menninger said. "We have a whole industry of tools that has evolved for reporting on and analyzing relational data."

Such data transfers don't have to be a one-time deal. "Maybe you were counting occurrences of a certain event and later decide that you want to count the number of times that two events occurred together," he said. "You go back to the source files and process the information again. That's why people don't throw the [unstructured] data away. They leave it in Hadoop."

In addition, Hadoop provides a much better environment for some advanced analytics and data mining applications than a SQL-based relational database does, Menninger said. One example he cited involves analyzing customer service call logs in combination with postings on Twitter, Facebook and other social media sites to try to identify customers who are likely to stop using a particular product or service.

"Those are hard things to express in SQL," Menninger said. But, he added, the analytical results can then be sent via a Hadoop connector to a relational database or data warehouse for further analysis and reporting and to drive follow-up actions aimed at keeping customers from defecting.

> "THE REASON YOU PUT [UNSTRUCTURED DATA] INTO A RELATIONAL DATABASE IS BECAUSE YOU CAN'T EASILY REPORT ON HADOOP DATA SOURCES TODAY."
>
> —DAVID MENNINGER
> analyst, Ventana Research

Cameron Befus, vice president of engineering at Tynt Multimedia Inc., a Web analytics company in Sausalito, Calif., that was acquired by 33Across Inc. in January, said his organization uses Hadoop to provide analytics services for more than 500,000 publishing websites. In addition, Tynt runs Oracle's open source MySQL database to power its back-office operations.

Thus far, Befus hasn't seen the need to install connector software to integrate the two environments. "We do move data around a little bit, but

it's usually pretty straightforward," he said, adding that the company directly loads files from Hadoop into MySQL. "A connector might make it slightly easier, but that just hasn't been a problem for us."

> "WHEN YOU'RE LOOKING AT [BIG DATA], WHAT YOU'RE LOOKING FOR IS, 'WHAT IS THIS DATA TELLING ME ABOUT SOME CRITICAL ISSUE?'"
>
> —JUDITH HURWITZ
> president and CEO,
> Hurwitz and Associates

Nonetheless, IT analysts such as Menninger and Judith Hurwitz, president and CEO of Hurwitz and Associates in Needham, Mass., expect demand for connectors to gradually increase as more organizations become Hadoop users.

Like Menninger, Hurwitz thinks interest in the technology will be driven by companies looking to put the results of Hadoop-based analyses into a greater business context.

"When you're looking at [big data], what you're looking for is, 'What is this data telling me about some critical issue?' " Hurwitz said. "[Users will] want to build bridges between this unstructured, streaming, 'get a sense of things' data and the very structured data that may include the details about how your company may be addressing those issues." ■

Mark Brunelli is senior news editor for SearchDataManagement.com. He covers topics and technologies such as databases, data warehousing, data integration, data quality, data governance and master data management. He can be contacted at mbrunelli@techtarget.com.

RELIABLE BI DATA REQUIRES COLLABORATIVE APPROACH

What constitutes data quality when analyzing millions of daily transactions? Does trustworthy data mean "perfect" data? You might be surprised at the answers. By David Loshin

BUSINESS competitiveness and agility are increasingly dependent on decisions that are informed and fueled by business intelligence (BI), reporting and analytics. For example, in an emerging "age of the algorithm," operational applications and processes are often enhanced as a result of business analytics. Meanwhile, power-user analysts explore various business scenarios by combining multiple large data sets, in many cases containing both structured and unstructured information.

As this dependence on BI grows, it should not be a surprise that business analytics users must have an implicit trust in their decision-making processes, which implies a reliance on having trustworthy data available to them.

Data quality is especially critical as the size of data volumes and the number of data sources grow, but what is meant by "high-quality data"? Data management professionals typically define data quality in terms of "fitness for use," but that concept rapidly becomes obsolete as we consider the numerous ways that the same data sets are repurposed and ultimately reused.

ASSESSING BI DATA QUALITY

From the analytics standpoint, data quality is best defined in terms of the degree to which data flaws impair the analytical results. Within an organization, that can be assessed using the following dimensions:

- **Completeness,** which measures whether a data set contains the full

number of records or instances that it should, as well as the degree to which each data instance has a full set of values for its mandatory data elements. Incomplete data can have a detrimental effect on analysis, particularly in the context of aggregations (such as sums and averages) that are skewed by missing data values.

- **Accuracy**, for checking data values against their real-world counterparts—for example, confirming that telephone numbers entered into a system match the actual numbers. A small number of inaccuracies in a large data set might not have statistical relevance; but as with an incomplete data set, a larger number of inaccuracies will

ACTION PLAN: FIVE STEPS TO HIGH-QUALITY BI DATA

HERE ARE SOME concrete steps that IT, data management and BI professionals can take to ensure the availability of high-quality data for business analytics.

1. **Solicit requirements: Engage the different kinds of business users within your organization and assess their information requirements and data quality expectations.**
2. **Create rules and metrics: Translate business user expectations into defined data quality rules and incorporate tools and techniques such as data profiling and parsing to develop data quality measurement services.**
3. **Assess data sets: Identify candidate data sources and use your quality measurement services to assess each data set's suitability for inclusion in the analytical environment.**
4. **Enable data transformation and cleansing: Deploy a customized set of tools and methodologies for data standardization and cleansing that can be employed by data owners and business users across the organization.**
5. **Build a data governance framework: Introduce the concepts of data stewardship and oversight to help maintain a consistent approach to managing data quality assurance. ■**

skew results. In addition, incorrect values can expose your organization to business impacts such as missed opportunities for revenue generation, increased costs and heightened risks.

- **Currency**, which focuses on how up to date the data sets being analyzed are. It is inadvisable to make critical business decisions based on stale data, so ensuring that your analytical data is current is vital.

- **Consistency**, which considers the degree to which the information in different data sets can be corroborated, as well as value agreement within and across sets of records. For example, in a record representing the terms of a contract, a begin date that is later in time than the contract end date would be a glaring inconsistency. Data sets that are inconsistent pose integration and merging issues, leading to duplicated and potentially inaccurate information.

ERRORS: NO HARM, NO FOUL

How can it be determined if source data is suitable for its many potential uses? The answer is simplified by correlating data errors and issues to the potential downstream business impact. The quality of a data set typically is acceptable as long as any errors do not affect business outcomes. As a result, organizations should use a collaborative approach to define measures, methods of scoring and levels of acceptability for all analytical usage scenarios.

This view of data quality can be illustrated by an example: an online mega-retailer might analyze millions of daily transactions to look for emerging patterns that create opportunities for product bundling and cross-selling. Because of the volume of records and the expected outcome, a small number of data errors can be tolerated. However, the retailer might not tolerate any data flaws when using the same data sets in responding to specific customer support questions.

HOW CAN IT BE DETERMINED IF SOURCE DATA IS SUITABLE FOR ITS MANY POTENTIAL USES? THE ANSWER IS SIMPLIFIED BY CORRELATING DATA ERRORS AND ISSUES TO THE POTENTIAL DOWNSTREAM BUSINESS IMPACT.

In other words, data quality requirements are directly related to the way data is used in individual business applications, including BI and analytics. Establishing the necessary level of trust in analytical data

involves engaging business users and understanding what they will be doing with the data. With specific knowledge of how data errors can affect decision making, controls can then be incorporated into data lifecycle management processes to ensure compliance with data quality rules, monitor quality against the agreed-to levels of acceptability and alert data stewards about quality issues while potentially triggering automated cleansing of data to meet downstream needs.

When done properly, the payoffs are enormous: Implementing effective data quality management and control procedures as part of your BI program will help lead to the data consistency, predictability and trustworthiness that are so critical to successful business analytics initiatives. ■

David Loshin is president of Knowledge Integrity Inc., a consulting, training and development company that focuses on information management, data quality and business intelligence. Loshin also is the author of four books, including *The Practitioner's Guide to Data Quality Improvement* and *Master Data Management*. He can be reached at loshin@knowledge-integrity.com.

Visit David Loshin's BeyeNETWORK expert channel for more articles on data quality and integration plus his blog.

TechTarget

BI Trends + Strategies is a joint e-publication of BeyeNETWORK, SearchBusinessAnalytics.com and SearchDataManagement.com.

Hannah Smalltree
Editorial Director

Jason Sparapani
Managing Editor, E-Publications

Jean Schauer
Editor in Chief, BeyeNETWORK

Craig Stedman
Executive Editor,
SearchBusinessAnalytics.com
and SearchDataManagement.com

Linda Koury
Director of Online Design

Mike Bolduc
Publisher
mbolduc@techtarget.com

Ed Laplante
Director of Sales
elaplante@techtarget.com

TechTarget
275 Grove Street, Newton, MA 02466
www.techtarget.com