InfoWorld

SPECIAL REPORT

Big Data Analytics *Deep Dive*

Deriving Meaning From the Data Explosion

Sponsored by IBM

Making sense of big data

New analysis tools and abundant processing power unlock critical insights from unfathomable volumes of corporate and external data

By David S. Linthicum

THE ABILITY TO DERIVE MEANING quickly from huge quantities of structured and unstructured data has been an objective of enterprise IT since the inception of databases. In the past, this has been a distant dream — or at least a cost-prohibitive one.

Today we have a growing number of technologies that make that dream a reality. Cloud computing provides per-drink access to thousands of processor cores and massive amounts of on-demand data storage. Emerging technologies apply a divide-and-conquer approach to big data computing problems, using distributed processing to return results almost instantaneously from huge data sets.

New analytics tools take advantage of all this horsepower. Advanced data visualization technology makes large and complex data sets understandable and enables domain experts to spot underlying trends and patterns. Some tools even recognize patterns and alert users about issues that need attention.

Big data is riding a sharp growth curve. IDC predicts big data technology and services will grow worldwide from $3.2 billion in 2010 to $16.9 billion in 2015. This represents a compound annual growth rate of 40 percent — about seven times that of the overall information and communications technology market.

MAKING THE BUSINESS CASE

The business benefits are clear. On the one hand, we can derive meaning from data that once merely took up space — website clickstream data, system event logs, and so on — and use that information to improve a broad range of systems. A whole new world of vertical applications opens up.

What sort of applications? How about sensor data collected on jet engine performance that could, when analyzed over time, reduce the incidence of failure. Data on the vital signs of infants in hospital nurseries could be plumbed for patterns that reveal hazards or new opportunities for progress. The use cases are limited only by the imagination:

- **A bank** visualizes patterns within data and documents that determine the likelihood of fraud and can take corrective action before the business is impacted.
- **Doctors** determine patterns of treatment of diseases that provide the most desirable outcomes using 20 years of historical data from more than 30 different data sources.
- **Auto manufacturers** see the core cause of production delays, and perhaps attach these analytics to business processes to automatically take corrective action, such as leveraging different logistical approaches.
- **The utility industry** creates predictive models of consumer markets by deploying technologies such as a smart grid.

With the proper data visualization on top, big data can shortcut the usual process of business analytics, where stakeholders pass requirements to analytics professionals who create static reports and hand them back over the transom. Big data visualization — even a simple tag cloud — enables those with domain expertise to develop a more direct relationship with the data and spot patterns that an analytics professional might miss.

Despite all the promise, big data technology and applications are still very much at an early stage. Hadoop has become the preferred platform for processing large quantities of unstructured data, but it's a distributed processing framework and development environment that typically requires specialized application development skills to use effectively. And that's just for unstructured

or semi-structured data. Big data can also refer to quantities of structured data so large they can't be processed in a reasonable length of time.

The larger question is how big data analytics fit into companies that already have a business analytics infrastructure. In many cases, exploring large unstructured data sets is a precursor to drilling deep with more conventional tools.

BULKING UP BUSINESS INTELLIGENCE

Gartner reports that 45 percent of sales management teams identify sales analytics as a priority to help them understand sales performance, market conditions, and opportunities. In short, the better that information is — and the more routinely it is put in the hands of those who need it — the greater likelihood productivity and revenue will increase.

A few key trends within BI relate to the emerging use of big data. They include:

- The ability to leverage both structured and unstructured data, and visualize that data together as a single, logical set of information.
- The ability to imply structure at the time of analysis, and thus provide flexibility by decoupling the underlying structure from structured or unstructured physical data.
- The ability to leverage current, or near-real-time data, allowing critical applications, business processes, and humans to view the up-to-the-minute state of the data.
- The ability to access outside data sources in the cloud, thus allowing BI analysts to mash up data from outside of the enterprise to enhance or refine the data analysis process.
- The ability to bind data analytics to business processes and applications, and thus allow them to automatically deal with issues and opportunities without human intervention.

Traditional BI is based on structured data, but these insights often fall short in predictive and indicative analytics, due to data sets that are too old or limited in scope. Structured data is only a small portion of stored business data. Indeed, many analysts estimate that structured data accounts for only 5 percent of total enterprise data.

Big data analytics enables better analytical insights by integrating massive amounts of data of varying complexity, formats, and timeliness into one structured output. Combining text, voice, streaming data, and unstructured data analytics into one structure allows businesses to harness the different views of related information into dynamic analytical models. These models support multidimensional metrics that can be leveraged with traditional analytics.

BI tools are still evolving to support this approach to big data analytics. They provide better data visualization capabilities that take into account the use of near-real-time information and the widening range of structured and unstructured data. To put it simply, if it exists in any electronic format, chances are it can be analyzed.

REACHING OUT TO THE CLOUD

One of the more exciting areas of big data involves data sources that are related to your business but were not collected or stored by the business itself. On a simple level, this might entail mashing up your existing sales trends with key economic data — or, as is in vogue now, data about trending topics on social networking sites.

Indeed, social content is the fastest-growing category of new content in the enterprise and will eventually attain 20 percent market penetration. Gartner defines social content as unstructured data created, edited, and published on corporate blogs and communication and collaboration platforms, along with such commercial platforms as Facebook, LinkedIn, Twitter, YouTube, and more.

Companies now specialize in providing this data on demand for use within enterprise BI, including the larger IaaS and PaaS cloud computing providers such as Google and Amazon Web Services. Indeed, the growth of data providers offering big data analytics services should only increase as BI professionals and their stakeholders understand the value it offers.

Finally, there's the ability to close the loop. More and more automated systems will be able to bind big data analytics to business processes, allowing operational systems to react to various thresholds in near-real time. This is known as embedded analytics, and may be created programmatically, or configured and exposed from within tools that support such services. Business examples include analyzing real-time delivery metrics and rerouting orders to suppliers that have better track

records, or automatically adjusting production schedules based upon predictive sales trends that leverage known correlations with key predictive data.

MORE DATA SOURCES, MORE POSSIBILITIES

A key challenge for big data analytics is the sheer proliferation of data sources that may or may not have structure. We aggregate these data sources around an implied structure created for the purpose of the data query and then present this structure to either a data analytics API or service, or to a BI tool for the purposes of data visualization or other types of interactive analytics (see Figure 1).

Remember that the prevailing trend is to study a mix of structured and unstructured data. The unstructured data may come from a variety of sources, including:

- Web pages
- Video and sound files
- Documents
- Social media APIs or services that provide trending data
- External data sources, such as data services provided by IaaS and PaaS cloud computing players
- Legacy unstructured data, such as older text-based databases

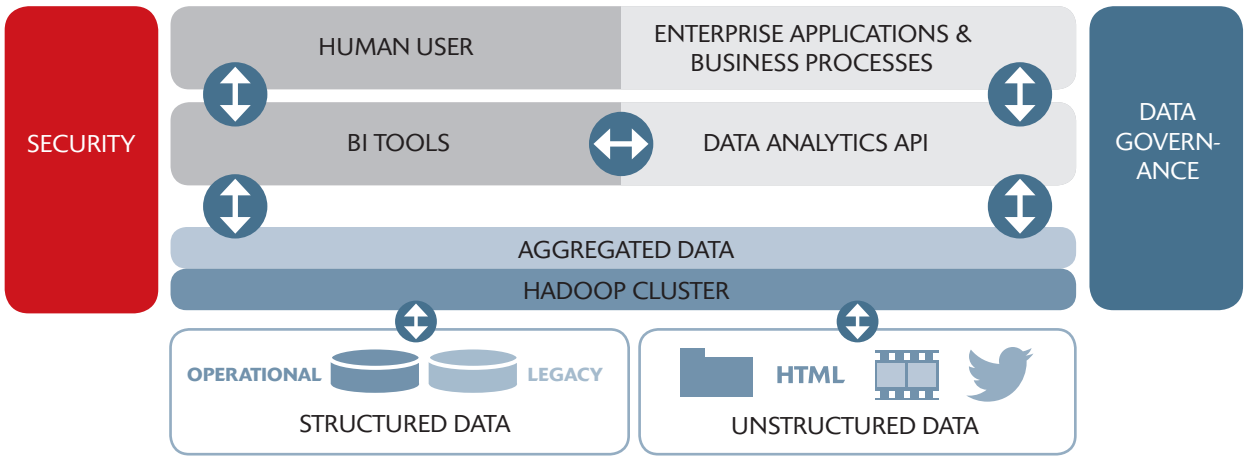
So how does this work? The unstructured and structured data is gathered into a file system (e.g., Hadoop Distributed File System, or HDFS). The data is stored in blocks on the various nodes in the Hadoop cluster (see Figure 1). The file system creates many replications of the data blocks, distributing them in the cluster in reliable ways that can be retrieved faster. Block sizes may vary, but a typical HDFS block size is 128MB, and is replicated to multiple nodes across the cluster.

We deal only with files, which means the content does not adhere to a structure before it exists in the file system. Data maps are then applied over the unstructured content to define the core metadata for that content. They may be mapped and remapped any number of times to support the changing metadata requirements of the analytical tools or others who leverage the data.

In some instances, Hadoop Hive is employed. Hive is a data warehouse system that provides data summarization, ad-hoc queries, and analysis of large datasets stored in the Hadoop cluster. Hive provides a mechanism to project structure onto this data and to query the data using a SQL-like language called HiveQL. The interface depends on your requirements and the data integration

FIG. 1. MAKING SENSE OF DATA

Big data analytics uses both structured and unstructured data and returns results to the BI tool in a matter of seconds. You can employ BI tools to analyze the data visually or as embedded analytics in an enterprise application or business process using a data analytics API or service.



capabilities of your BI tool.

Another option is Apache Pig. Pig is a high-level platform for creating MapReduce programs used with Hadoop. It abstracts the programming from the MapReduce engine. Like Hive, Pig uses its own language to interact with the data.

Generally speaking, when you execute a query from a BI tool, the following occurs:

1. The BI tool will reach out to the cluster to get the file metadata information. Typically, the BI tool will deal directly with a data structure that may exist specifically for the analytics use case or model (see Figure 2). You should consider this structure an abstract representation of the underlying structured or unstructured physical data.
2. From there, the system will reach out to the data nodes to get the real data blocks and bring those back into the structure. There may be any number of physical and logical nodes, depending upon the requirements of the system and how it's architected.
3. The MapReduce parallel programming model gathers the data from the Hadoop cluster. This system deals with the operational details, managing the processing load across available server resources.
4. The requested result set is returned to the BI tool for visualization or other types of processing, typically bound to specific data structures.
5. The BI tool can layer this data into defined models, including loading the data from the result sets directly into a dimensional model for complex analytical processing or into graphic representations of the data.
6. The data can be refreshed at any increment by repeating this process.

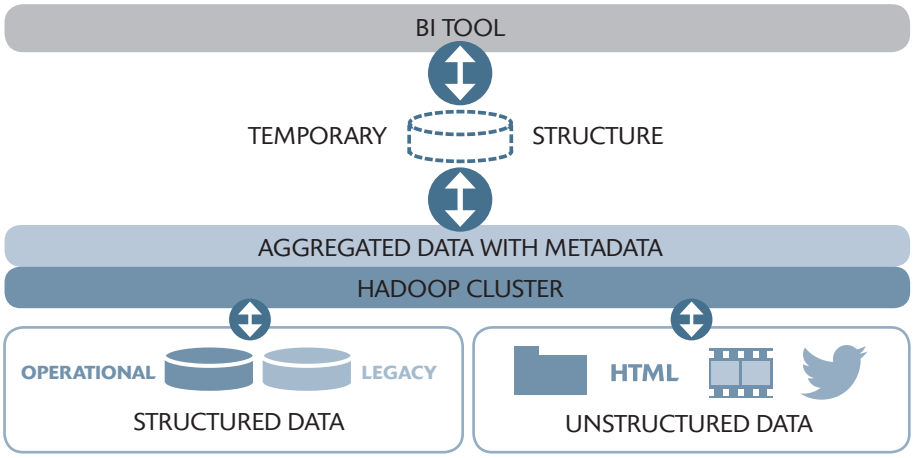
This is a very general scenario, and the BI tools you select may have a different approach. Many BI tools use mappers that make the data appear as if it were stored in a traditional relational database. Others take advantage of the native features of big data technology, including the ability to treat structured and unstructured data differently within the analytical models.

Some BI tools load the summarized or aggregated data into a temporary, multidimensional "cube" structure (see Figure 3). This allows the analyst to visualize the data coming from the big data systems in ways that are most useful.

What's different about this model is that both structured and unstructured data can now be visualized. Also,

FIG. 2. STRUCTURE ON-THE-FLY

The BI tools use a structure that can be created just for the purpose of data analytics. The information exists in the file system cluster and metadata is mapped onto the content as required to support the use case. This provides a more dynamic and flexible way to approach BI.



new and expanded analysis may be creatively derived from the availability of this data, such as:

- Reporting or descriptive analytics
- Modeling or predictive analytics
- Clustering
- Affinity grouping

What's most important about big data analytics is the new mindset that is emerging. We now consider most data as something to be explored by anyone who needs to explore it. We're moving away from a limited view of our own business data. Moreover, our analytical models, such as predictive modeling, provide better results because the data is more complete.

BIG DATA VISUALIZATION AND ANALYSIS USE CASES

Within vertical industries, interest in big data is high, but the degree of knowledge and investment varies widely (see Figure 4). Health care, transportation, and education lead the pack.

USE CASE: BUSINESS PROCESS IMPROVEMENT

Big data analytics enable companies to examine the state of their business with greater detail and accuracy, including the productivity of their business processes. Analytics coupled with data visualization shine a bright light on areas where business processes fall short.

For example, using data visualization, a business can zoom in on the process of recording a sale and processing it for shipping and explore the relationship between that business process and customer satisfaction. Optimizing that process seldom fails to increase repeat business.

USE CASE: BUSINESS-CRITICAL APPLICATION AUGMENTATION

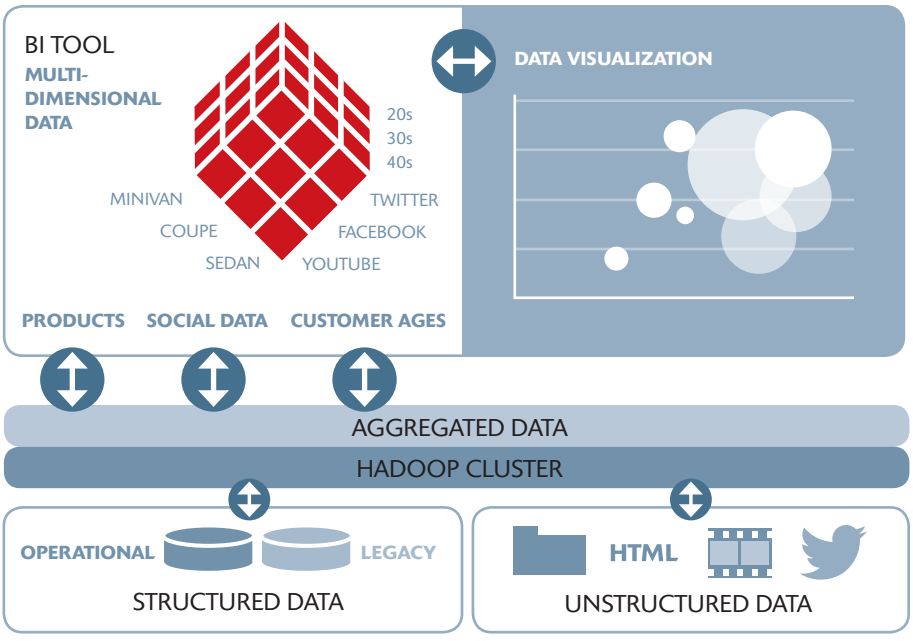
Embedded big data analytics coupled with operational enterprise applications can deliver enormous business value. For example, a company could augment a shipping application with analytical information about on-time delivery records culled from perhaps terabytes of PDF shipping records gathered over the years. That data might also be mashed up with information from outside sources such as complaints recorded in social media or blogs.

USE CASE: IMPROVING HEALTH CARE TREATMENT AND OUTCOMES

There is much low-hanging fruit when it comes to the business value of big data analytics in health care;

FIG. 3. DIMENSIONS OF DATA

BI tools employ a range of analytical models and structures to analyze big data. In this case the data is loaded into a multidimensional temporary model where it may be visualized in any number of ways.



however, the best use case is in patient diagnosis and treatment. The health care system stores our information in different places using different formats, which has made it difficult or impossible to analyze this data as a single cluster of information.

With big data analytics, we now can gather all structured and unstructured health care data and place this content in a single cluster for analysis by a BI tool. This improves health care professionals' ability to determine patterns of treatment success by analyzing holistic patient data and treatments that lead to the desired outcome.

USE CASE: IMPROVING RETAIL BUSINESS PERFORMANCE

Retail businesses depend on an in-depth understanding of specific markets and customers to gain competitive advantage. Here, big data analytics has huge potential value, allowing those who drive the BI tools to create models that determine the success of a product based upon predictive data points gathered from huge amounts of unstructured data.

This data may include demographics information around the existing customer base, as compared to the number of times the product is mentioned within social media systems, as compared to patterns around the success of products sold in the past, as compared to weather patterns that may affect the use of the product (say, a down coat in a very cold winter). The idea is to provide those who make critical decisions in the retail space with the means to slice and dice all the relevant data to determine which products to advertise, discount, or display adjacent to other products.

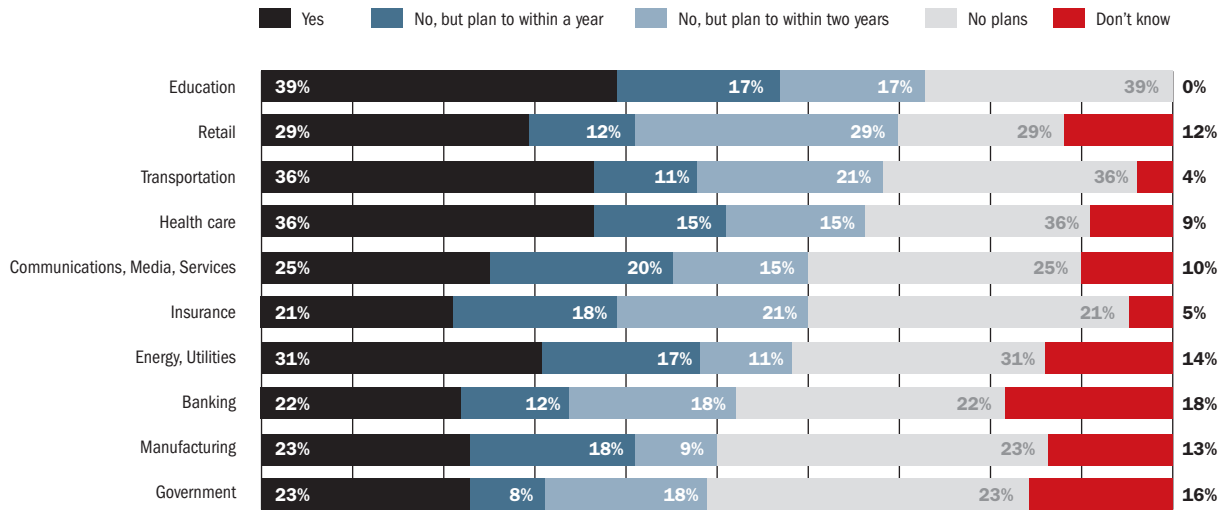
USE CASE: IMPROVING TRANSPORTATION SYSTEMS

Transportation systems depend on efficiencies. For example, airlines need to select the best and most profitable routes. Using big data analytics, they can predict the profitability of those routes by leveraging historical data visualized with key predictive metrics applied to data gathered from external sources.

Big data analytics allows airlines to gather years and years of flight data from the government, including point of origin, number of passengers, on-time records, etc.. They can then look at this data with known pricing information from other carriers. Add in predictive data, the number of times the destination appeared in Web searches over the past several years, and the number of

FIG. 4. Has your organization already invested in technology specifically designed to address the big data challenge?

According to Gartner, almost all verticals are investing in big data analytics, with transportation, education, and health care leading the way.



times it's been mentioned on social networking sites as well. By modeling this data within a BI tool, the airline would have a good idea as to the past use and profitability of the route, as well as the future possibility that tickets would sell, and at what price.

MAPPING A PATH FOR YOUR BUSINESS

To fully capitalize on big data analytics, you need to free yourself from the limitations of traditional BI. Unfortunately, those who created BI often attempt to force traditional BI approaches (square peg) into the new world of big data (round hole). As a result, they miss the full potential of this technology or fail altogether.

Here's a rough outline of how to implement big data analytics in your own business:

1. Understand your core business requirements for this technology and create a business case.
2. Define your data sources. Where are they? What are they? What's the best way to reach and replicate them when required?.
3. Define known use cases, including the analytical models you will need to understand aspects of your data.
4. Create a proof-of-concept to both learn about the technology and better understand the complexities of bringing the technology into the enterprise.
5. Consider performance, security, and data governance. These issues are often overlooked, but ultimately are required for a successful implementation.
6. Spend the time and money to evaluate BI technology in terms of features and functions. BI and data visualization provide your window into the data, and any limitations will limit the value of the data.
7. Define the metrics of success. Evaluate what's working and what's not after a year of living with this technology. Adjustments can be made with very little disruption.
8. Finally, make sure to create a road map for this technology. Include how it will be leveraged today, and into short- and long-term business planning.

The value of this technology is very clear to those in business. The ability to work with good information has always been the big challenge for IT. Now we have the tools to address this issue, and it's up to you to make it work.

David S. Linthicum is the CTO and founder of Blue Mountain Labs and an internationally recognized industry expert and thought leader. Dave has authored 13 books on computing, the latest of which is "Cloud Computing and SOA Convergence in Your Enterprise, a Step-by-Step Approach." Dave's industry experience includes positions as CTO and CEO of several successful software companies, and upper-level management positions in Fortune 100 companies. He keynotes leading technology conferences on cloud computing, SOA, enterprise application integration, and enterprise architecture.