Big Data: The Fourth Data Management Generation

Hugh J. Watson and Olivera Marjanovic

Hugh J. Watson is a C. Herman and Mary Virginia Terry Chair of Business Administration in the Terry College of Business at the University of Georgia. He is a TDWI Fellow and senior editor of the *Business Intelligence Journal*.
hwatson@uga.edu

Olivera Marjanovic is a senior lecturer in the discipline of business information systems in the business school at the University of Sydney.
olivera.marjanovic@sydney.edu.au

Companies such as Yahoo! and Google are at the "big data" forefront and have played major roles in developing technologies such as Hadoop and MapReduce for storing and analyzing big data. Others are implementing a variety of big data platforms and technologies, such as appliances, columnar databases, and cloud-based solutions, although they don't receive the same attention. Most companies, however, are just starting to think about or experiment with the potential uses of big data and how to integrate big data into their existing BI architectures (some implementations are just standalone solutions). Big data solutions do not eliminate the need for data warehouses and marts; rather, each data platform has its own role and value proposition.

Although big data is relatively new, the need for data to support decision making has existed since the early 1970s with decision support systems (DSSes). DSSes can be thought of as the first generation of decision support data management. Over time, additional generations have evolved. More specifically, around 1990, enterprise data warehouses (the second generation) appeared. Next, real-time data warehousing (the third generation) came on the scene in the early 2000s, and finally, starting in 2010, big data is creating a new, fourth generation. Each generation both replaced and built on previous generations.

These generations are characterized by scope, focus, decisions supported, users, volume, velocity, variety, data sources, architecture complexity, and value. As we will see, each generation was driven by business need, fueled by technological advances, and faced many implementation challenges.

The Early Generations

Understanding each of the first three generations provides perspective about what is currently taking place in the fourth generation.

The First Generation: The DSS

A decision support system was, as its name implies, focused on making decisions. The most popular characterization was of a single decision maker using data and analytic aids to generate information to support decision making. Sometimes a Gary Cooper in *High Noon* analogy was used. In the movie's climactic scene, the town marshal (the decision maker) faces the gang of killers (the decision) alone. Fortunately for DSS, Cooper survives the shootout and saves the town. To some extent, "spreadmarts" also fit the analogy.

The data used by the DSS was sourced from a single or just a few operational systems, was low in volume, updated in batch mode, and highly structured. The user of the system, either the decision maker or an "intermediary" who operated the system for the decision maker, employed the DSS to support tactical and strategic decision-making. These systems did create value when they were used, but usage was not high. The technology for building and operating a DSS was primitive—think of command language interfaces and no relational databases—and most users had very limited computer skills. In fairness, people who thought about the potential of DSSes envisioned a more comprehensive kind of system (e.g., internal and external data sources, support for operational decision making), but the available technology did not make it feasible.

The Second Generation: Enterprise Data Warehouses

By the 1990s, there was a need to supply a variety of BI applications (e.g., reporting, executive information systems) with data. Having separate databases (what we now call independent data marts) for each application was costly, resulted in data inconsistencies across applications, and failed to support enterprisewide applications. The outcome was the emergence of *enterprise data warehouses* (EDWs).

EDWs represent a data-focused approach to data management. When users and applications need data, that is where they go to get it, and over time, usage grows. The early warehouses were challenging to build because the tools were primitive (e.g., ETL) and there were few experienced professionals. Most of the data in the warehouse was structured and updated in batch mode. Some organizations built operational data stores (ODSes) to provide more real-time data to support operational decision making. Data was sourced from a large number of systems, data volumes grew, and the user base started to extend outside of the organization (such as to customers and suppliers).

The Third Generation: Real-Time Data Warehousing

The next generation was real-time data warehousing. Technology improvements by 2000 made it possible to capture data in real time and trickle feed it into the data warehouse. The significance of this evolution is that it changed the paradigm for what kinds of decisions could be supported. Real-time data could support operational decisions and processes. For example, retail websites could make product recommendations "on the fly" and customer-facing employees could better know their profitability and serve customers.

With real-time data, data volumes soared, though most data continued to be structured. The user base grew as BI became pervasive. The idea of the BI-based organization grew as more companies became dependent on BI in order to successfully compete in the marketplace.

The Big Data Generation

Big data is characterized by extremely high volume, velocity, and variety (commonly referred to as the "3 Vs"). It also exceeds the capabilities of most relational database management systems and has spawned a host of new technologies, platforms, and approaches. The data sources are relatively new (e.g., Web logs, machines, and social media), at least in terms of storing and analyzing the data for decision support purposes. When analyzed, this data can be used in a variety of new ways, such as understanding customers better or predicting when a part is likely to fail. It can also be combined with traditional data to provide greater context for decision making, such as sales

CHARACTERISTICS	FIRST GENERATION	SECOND GENERATION	THIRD GENERATION	FOURTH GENERATION
Scope	Individual/departmental	Enterprise	Enterprise	Extended enterprise
Focus	Application	Data	Application/ data	Application/data
Decisions supported	Tactical/strategic	Tactical/strategic	Operational/tactical/ strategic	Operational/tactical/ strategic
Users	Single	Multiple	Enterprise	Enterprise
Volume	Low	High	Very high	Extreme
Velocity	Batch	Batch/ODS	Real time	Real time
Variety	Structured	Structured	Structured	Structured/ unstructured
Data sources	Single internal source	Multiple internal sources	Multiple internal sources	Multiple internal/external
Architectural complexity	Low	Medium	High	Extreme
Value	Low	Medium	High	Potentially very high

Table 1: The characteristics of the four generations of decision support data management.

figures and sentiment analysis about what customers are saying about a product. Although the value propositions for big data are still emerging, it is clear that the potential is extremely great. Table 1 summarizes the characteristics of the data management generations.

Opportunities and Challenges

Let's consider some of the greatest opportunities and challenges associated with big data. They include identifying appropriate opportunities to use and benefit from big data, putting the data on the appropriate platform, integrating these platforms, providing governance of big data, and getting people with the right skills to do the work.

Identifying appropriate opportunities. In our conversations with executives, we find they are aware of big data (it's hard to miss with the unrelenting media attention) and may know of some specific uses (such as sentiment analysis). However, they are unsure of how it can be used in a larger way in their companies and what is required to be successful. In this regard, big data is both an opportunity and a challenge.

It is problematic when executives want to do something with big data but don't know exactly what, because, as always, a lack of clearly defined business requirements means significant risk. Like any IT project, big data projects need to be business driven. Once given the business requirements, the needed infrastructure, data sources, and analytics can be determined. It all starts, however, with the business requirements.

Many companies are using big data analytics to focus on customer-centric issues, such as improving the customer experience and better understanding customer preferences and behavior. The Starbucks story (see sidebar, page 8) is an example of this. We find it interesting that most big data projects involve internal data, such as in the smart meter example (see sidebar). This makes sense because the data is available and well understood.

AUTOMOBILE INSURANCE	MANUFACTURING, DISTRIBUTION, AND RETAIL	TRANSPORTATION AND LOGISTICS	TELECOMMUNI-CATIONS	UTILITIES	LAW ENFORCEMENT	GAMING
insurance pricing	Track shelf availability	Real-time management of truck fleets	Analysis of patterns of services across social networks	Smart-grid data to determine variable pricing models	Identify people linked to known trouble groups	Capture players' actions to provide comprehensive feedback to game producers
Better client risk analysis	Assess the impact of promotional displays	RFID data for asset tracking	Determine profitability of customers' social networks	Smart meters to better forecast energy demand	Determine the location of individuals and groups	Analysis of game plays to determine the best opportunities for in-game offers
Fraud detection	Assess the effectiveness of promotional campaigns		Churn minimization	Customized rate plans for customers		
Faster claims processing	Inventory management					
	Pricing					

Figure 1: Big data opportunities in different industries.

Bill Franks' recent book, *Taming the Big Data Tidal Wave*, provides great examples of the use of big data in different industries and is the basis for many of the opportunities shown in Figure 1. It is helpful in sorting out possible starting points for big data analytics.

Selecting the right platform. The alternatives for storing and analyzing big data have exploded over the past few years. Although there is no "formula" for choosing the right platform, the most important considerations include the volume, velocity, and variety of data; the applications that will use the platform; who the users are; and whether the required processing is batch or real time. The final choices ultimately come down to where the required work can be done at the lowest cost.

Integrating the platforms. BI directors have worked hard to integrate their BI architectures. For example, many independent data marts have been integrated into centralized data warehouses. This high level of integration is being challenged as organizations implement new big data platforms. Some of the platforms will be standalone solutions, but there are strong arguments for integrating them. When this is done, there should be fast, seamless interaction and collaboration among the component parts.

Consider these examples of how the platforms should work together. A report run on one platform should mirror a report run on another (the data needs to be synchronized). An analysis run on a specialized platform should be able to access information from the data warehouse (a source system for the specialized platform). If needed, the results of an analysis run on a specialized platform should be stored in the data warehouse (the specialized platform is a source system for the warehouse).

Vendors recognize the importance of integrating the component parts and are including software solutions for doing this. For example, Teradata is currently emphasizing its Unified Data Architecture that ties its family of products together.

Providing governance. Much as integrating the different platforms is a challenge, so, too, is putting governance in place. It might even be more difficult than in the past because of the aggressive way many business units (such as marketing and finance) are adopting big data platforms for their own purposes. To illustrate, there are forecasts that the marketing department's spending on IT will exceed IT's in a few years. These business units have their own objectives and do not always fully appreciate the importance of having, for example, data stewards, the importance of consistent data definitions, metadata management, and data life cycle management. Governance should be put around big data as soon as possible.

Getting the right people. If you might need to increase the human side of your big data analytic capabilities, a good starting point is a gap analysis of what skills you currently have and what you will likely need. If you find a gap, there

are several options—hiring new people, relying on outside professional services, or upgrading the skills of your current staff. These options are not mutually exclusive.

Hiring new professionals with the right skills mix is currently challenging, and projections indicate that it will get worse. However, it might not be as bad as you think. In the last year, universities have dramatically increased their offerings (including courses, concentrations, and degrees) in analytics and big data. If students take advantage of these opportunities, the shortage may be less severe and of shorter duration than anticipated. There is the question, however, of how many students will actually take advantage of the opportunities. Big data analytics is a tough field of study and many students are technology averse.

Consulting firms and vendors offer professional services in big data analytics. This may be a good way to kick-start a big data analytics initiative, but be sure to include knowledge transfer as part of the contract. This option can get expensive quickly and you will most likely want to have your own in-house capabilities for the longer term. Also, talk with firms about tying their fees to demonstrated benefits.

STARBUCKS

Starbucks introduced a new coffee product and was concerned that its taste might be too strong. On the day the coffee was rolled out, Starbucks monitored and analyzed Tweets, coffee blogs, and forums to learn what customers were saying (i.e., sentiment analysis). There were few complaints about the coffee's taste but a significant number of people thought that its price was too high. By the end of the day, Starbucks had lowered the price and sentiment analysis showed that customers were happy with the new product.

SMART METERS

Some utility companies have installed smart meters in homes and businesses, and these meters provide a constant flow of data about energy consumption. One way this information is used is to provide customers with information about their energy usage, and when combined with appropriate pricing plans, customers are motivated to move energy consumption to non-peak hours, such as washing clothes at night. This approach helps customers save money and allows utility companies to operate with lower peak demand capacity.

There is considerable talk about data scientists, the "high priests" of analytics, often with Ph.D.s in statistics or computer science, who are knowledgeable about a wide variety of sophisticated data analysis methods. There is no doubt that they are valuable resources for some organizations and applications, but not all big data analytics requires their skills. If you have business analysts who are bright, inquisitive, and analytics oriented, their skills can be upgraded through university courses, training programs, and conferences. Have them study Java, R, SAS Enterprise Miner, IBM SPSS Modeler, or Hadoop and MapReduce. Analytical workbenches such as SAS Enterprise Miner and IBM SPSS Model have integrated capabilities that help automate the analytics process and reduce the need for data scientists to do all the heavy-lifting analytics work.

Conclusion

How BI and IT groups handle the opportunities provided by big data can either confirm or change how senior management perceives them. It's potentially that important, and it can show that BI and IT are important enablers of the business strategy, deliver value to the business units, and deserve a seat at the senior management table. To make this a reality, however, BI and IT will have to work closely with management and business units to address the challenges posed by big data. For firms without this kind of relationship, BI is likely to be limited to descriptive analytics (e.g., reporting, dashboards, and scorecards), while IT manages the data infrastructure and some of the business units maintain their own big data platforms. ■

References

Franks, Bill [2012]. *Taming the Big Data Tidal Wave*, Wiley.

Acknowledgment

The authors would like to thank the University of Sydney Business School for providing the funding that made this collaboration possible.