Using Big Data for Smarter Decision Making

Colin White, BI Research

July 2011
Sponsored by IBM

USING BIG DATA FOR SMARTER DECISION MAKING

"To increase competitiveness, 83% of CIOs have visionary plans that include business intelligence and analytics."[1]

"Global digital content created will increase some 30 times over the next ten years – to 35 zettabytes)."[2]

Becoming an analytics-driven organization helps companies reduce costs, increase revenues and improve competitiveness, and this is why business intelligence and analytics continue to be a top priority for CIOs. Many business decisions, however, are still not based on analytics, and CIOs are looking for ways to reduce time to value for deploying business intelligence solutions so that they can expand the use of analytics to a larger audience of users.

Big data sources are largely untapped by business intelligence applications

Companies are also interested in leveraging the value of information in so-called *big data* systems that handle data ranging from high-volume event data to social media textual data. This information is largely untapped by existing business intelligence systems, but organizations are beginning to recognize the value of extending the business intelligence and data warehousing environment to integrate, manage, govern and analyze this information.

This paper looks at new developments in business analytics and discusses the benefits analyzing big data bring to the business. It also examines different types of big data workloads and offers suggestions on how to optimize systems to handle different workloads and integrate them into a single infrastructure for making smarter and faster business decisions.

WHAT IS BIG DATA?

Big data is a popular buzzword, but it is important to realize that this data comes in many shapes and sizes. It also has many different uses – real-time fraud detection, web display advertising and competitive analysis, call center optimization, social media and sentiment analysis, intelligent traffic management and smart power grids, to name just a few. All of these analytical solutions involve significant (and growing) volumes of both multi-structured[3] and structured data.

Most of these analytical solutions were not possible previously because they were too costly to implement, or because analytical processing technologies were not capable of handling the large volumes of data involved in a timely manner. In some cases, the required data simply did not exist in an electronic form.

[1] "The Essential CIO." IBM CIO study involving 3,018 CIOs spanning 71 countries and 18 industries, 2011.
[2] "Digital Universe Study." IDC study sponsored by EMC, May 2010.
[3] This type of data has unknown, ill-formed, or overlapping schemas. The term unstructured is also used to describe this data, but this is misleading because most of the data in this category does have some structure.

New technologies enable the analysis of big data

New and evolving analytical processing technologies now make possible what was not possible before. Examples include:

- **New systems** that handle a wide variety of data from sensor data to web and social media data.
- **Improved analytical capabilities** (sometimes called *advanced analytics*) including event, predictive and text analytics.
- **Operational business intelligence** that improves business agility by enabling automated real-time actions and intraday decision making.
- **Faster hardware** ranging from faster multi-core processors and large memory spaces, to solid-state drives and virtual data storage for handling hot and cold data.
- **Cloud computing** including on-demand software-as-a-service (SaaS) analytical solutions in public clouds, and data platforms and virtualization in private clouds.

Supporting big data involves combining these technologies to enable new solutions that can bring significant benefits to the business.

Analyzing big data involves extreme workloads

Big data involves more than simply the ability to handle large volumes of data. Instead, it represents a wide range of new analytical technologies and business possibilities. The challenge is how to deploy these technologies and manage the many extreme analytical processing workloads involved, while at the same time providing faster time to value.

THE CHALLENGES OF BIG DATA AND EXTREME WORKLOADS

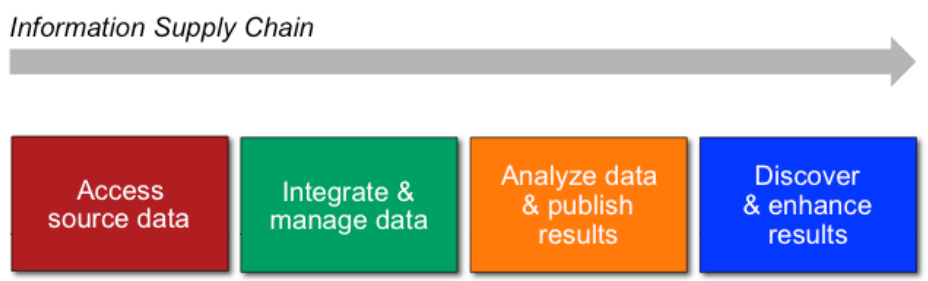
An enterprise data warehouse can be used to handle big data and extreme workloads, but often it is more efficient to preprocess the data before loading it into the warehouse. Event data from hardware sensors, for example, has more business value if it is filtered and aggregated before using it for analysis. Usually the raw event data is not required for historical purposes, and storage and processing costs are reduced if it is not kept in the warehouse.

Big data and extreme workloads require optimized hardware and software

The best way of handling big data and supporting the extreme processing involved is to deploy optimized hardware and software solutions for processing different types of big data workloads, and then combine these solutions with the existing enterprise data warehouse to create an integrated *information supply chain* (see Figure 1). The objectives of an information supply chain are to consume and integrate the many varieties of raw source data that exist in organizations, analyze that data, and then deliver the analytical results to business users. This information supply chain enables the data component of what IBM calls *Smarter Computing*.

Supporting extreme workloads is not a new challenge for the computing industry. Business transaction processing systems have supported extreme transaction workloads ever since the advent of data processing. As new business needs arose, organizations employed custom and optimized transaction processing systems to handle application workloads that pushed the boundaries beyond what could be handled by more generalized technologies. Airline reservation systems, bank ATM systems, retail point-of-sale terminals, financial trading systems and mobile phone systems, are all examples of these types of applications. More recently, applications for handling sensor networks that track items with RFID tags can be added to the list.

Figure 1. The information supply chain



In recent years there has been a similar trend toward analytical processing reaching the performance limitations of more generalized systems. Today, multi-terabyte data warehouses are no longer the exception and analytical processing workloads are becoming increasingly more complex. The result is that once again organizations require optimized systems to meet the challenges of extreme workloads. The industry response has been to offer packaged hardware and software solutions that are optimized for analytical processing.

Satisfying business agility requirements is another important factor in supporting analytical processing. In today's fast paced business environment, organizations need to make faster decisions, and this agility is important to business success. In the case of fraud detection, for example, real time action is required. Not all business decisions have to be made in real time, but for many organizations the ability to act in a few seconds or minutes, rather than hours or days, can be a significant financial and competitive advantage.

Managing and Analyzing Big Data

Data volume and variety, and workload complexity and agility are the challenges of big data

The main challenges of big data and extreme workloads are data variety and volume, and analytical workload complexity and agility. Each of these can be viewed from the perspective of the information supply chain shown in Figure 1 above.

Input to the information supply chain consists of the raw source data required for analysis. For the past two decades most business analytics have been created using structured data extracted from operational systems and consolidated into a data warehouse. Big data dramatically increases both the number of data sources and the variety and volume of data that is useful for analysis. A high percentage of this data is often described as *multi-structured* to distinguish it from the *structured* operational data used to populate a data warehouse. In most organizations, multi-structured data is growing at a considerably faster rate than structured data.

There are two main techniques for analyzing big data – the *store and analyze* approach, and the *analyze and store* approach.

Store and Analyze Approach

Traditional data warehousing is used to produce data analytics

The **store and analyze** approach integrates source data into a consolidated data store before it is analyzed. This approach is used by a traditional data warehousing system to create *data analytics*. In a data warehousing system, the consolidated data store is usually an enterprise data warehouse or data mart managed by a relational or multidimensional DBMS. The advantages of this approach are improved data integration and data quality management, plus the ability to maintain historical information. The disadvantages are additional data storage requirements and the latency introduced by the data integration task.

Two important big data trends for supporting the store and analyze approach are relational DBMS products optimized for analytical workloads (often called analytic RDBMSs, or ADBMSs) and non-relational systems (sometimes called NoSQL systems) for processing multi-structured data. A non-relational system can be used to produce analytics from big data, or to preprocess big data before it is consolidated into a data warehouse. Certain vendors in the search and content management marketplaces also use the store and analyze approach to create analytics from index and content data stores.

Analytic RDBMSs (ADBMSs)

Analytic RDBMSs improve price performance for creating data analytics

An analytic RDBMS is an integrated solution for managing data and generating analytics that offers improved price/performance, simpler management and administration, and time to value superior to more generalized RDBMS offerings. Performance improvements are achieved through the use of massively parallel processing, enhanced data structures, data compression, and the ability to *push* analytical processing into the DBMS.

ADBMSs can be categorized into three broad groups:[4] packaged hardware and software appliances, software-only platforms, and cloud-based solutions.

Packaged hardware and software appliances fall into two sub-groups: purpose-built appliances and optimized hardware/software platforms. The objective in both cases is to provide an integrated package that can be installed and maintained as a single system. Depending on the vendor, the dividing line between the two sub-groups is not always clear, and this is why in this article they are both categorized as appliances.

A *purpose-built appliance* is an integrated system built from the ground up to provide good price/performance for analytical workloads. This type of appliance enables the complete configuration, from the application workload to the storage system used to manage the data, to be optimized for analytical processing. It also allows the solution provider to deliver customized tools for installing, managing and administering the integrated hardware and software system.

Many of these products were developed initially by small vendors and targeted at specific high-volume business area projects that are independent of the enterprise data warehouse. As these appliances have matured and added workload management

[4] Some vendors simply call all of these *appliances*, but this is oversimplified and confusing. Although the three groups overlap, each of them has its own unique architecture, strengths and weaknesses, and analytical use cases.

capabilities, their use has expanded to handle mixed workloads and in some cases support smaller enterprise data warehouses. Large and well-established vendors have acquired several of these solutions. An example is the IBM Netezza TwinFin appliance.[5]

The success of these purpose-built appliances led to more traditional RDBMS vendors building packaged offerings by combining existing products. This involved improving the analytical processing capabilities of the software and then building integrated and optimized hardware and software solutions. These solutions consist of *optimized hardware/software platforms* designed for specific analytical workloads. The level of integration and optimization achieved varies by vendor. In some cases, the vendor may offer a choice of hardware platform. An example of this type of approach is the IBM Smart Analytics System.

IBM Smart Analytics System and IBM Netezza are examples of optimized systems

The IBM Smart Analytics System is intended for extending the performance and scalability of an enterprise data warehouse, whereas the IBM Netezza TwinFin appliance is used to maintain a separate data store in situations where it is unnecessary, impractical, or simply not cost effective to integrate the data into an enterprise data warehouse for processing. The IBM Netezza TwinFin is also used to analyze data extracted from an enterprise data warehouse, either to offload an extreme processing workload, or to create a *sandbox* for experimental use and/or complex ad hoc analysis.

A **software-only platform** is a set of integrated software components for handling analytical workloads. These platforms often make use of underlying open source software products and are designed for deployment on low-cost commodity hardware. The tradeoff for hardware portability is the inability of the product to exploit the performance and management capabilities of a specific hardware platform. Some software platforms are available as virtual images, which are useful for evaluation and development purposes, and also for use in cloud-based environments.

Cloud-based solutions offer a set of services for supporting data warehousing and analytical application processing. Some of these services are offered on public clouds, while others can be used in-house in private cloud environments. The underlying software and hardware environment for these cloud-based services may be custom built, employ a packaged hardware and software appliance, or use the capabilities of a software-only platform. The role of cloud computing for business intelligence and data warehousing is discussed in more detail later in this paper.

Non-Relational Systems

Non-relational systems are used to process multi-structured data

A single database model or technology cannot satisfy the needs of every organization or workload. Despite its success and universal adoption, this is also true for RDBMS technology. This is especially true when processing large amounts of multi-structured data and this is why several organizations with big data problems have developed their own non-relational systems to deal with extreme data volumes. Web-focused companies such as Google and Yahoo that have significant volumes of web information to index and analyze are examples of organizations that have built their

[5] IBM acquired Netezza Corporation in September, 2010.

own optimized solutions. Several of these companies have placed these systems into the public domain so that they can be made available as open source software.

Non-relational systems are useful for processing big data where most of the data is multi-structured. They are particularly popular with developers who prefer to use a procedural programming language, rather than a declarative language such as SQL, to process data.[6] These systems support several different types of data structures including document data, graphical information, and key-value pairs.

Hadoop with MapReduce is one of the leading non-relational systems

One leading non-relational system is the Hadoop distributed processing system from the open source Apache Software Foundation. Apache defines Hadoop as “a framework for running applications on a large hardware cluster built of commodity hardware.” This framework includes a distributed file system (HDFS) that can distribute and manage huge volumes of data across the nodes of a hardware cluster to provide high data throughput. Hadoop uses the MapReduce programming model to divide application processing into small fragments of work that can be executed on multiple nodes of the cluster to provide massively parallel processing. Hadoop also includes the Pig and Hive languages for developing and generating MapReduce programs. Hive includes HiveQL, which provides a subset of SQL.

Hadoop MapReduce is intended for the batch processing of large volumes of multi-structured data. It is not suitable for low-latency data processing, many small files, or the random updating of data. These latter capabilities are provided by database products such as HBase and Cassandra that run on top of Hadoop.

Several companies offer commercialized open source or *open core* versions of Hadoop for handling big data projects. IBM’s InfoSphere BigInsights product fits into this category.

Which DBMS To Use When?

Organizations will use both analytic RDBMSs and non-relational systems

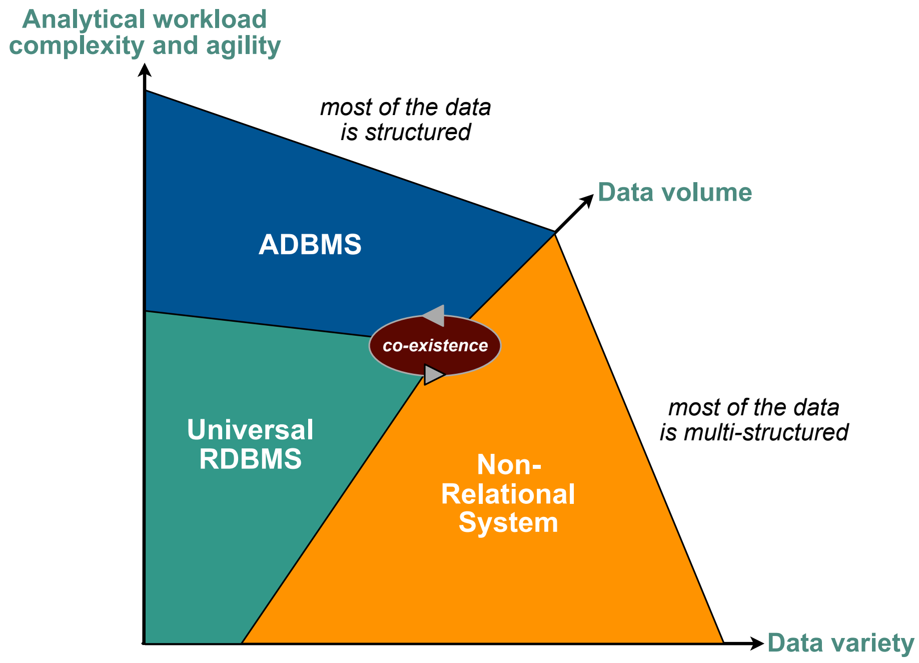
It is important to realize that generalized relational DBMSs, analytic RDBMSs, and non-relational data systems are not mutually exclusive. Each approach has its benefits, and it is likely that most organizations will employ some combination of all three of them.

When choosing a system for handling big data and extreme processing, several factors have to be considered: the volume of data, the variety of data, and the complexity and agility requirements of the analytical workloads. Figure 2 positions the different approaches with respect to these factors. To provide an integrated analytical infrastructure these approaches must coexist and interoperate with each other. This is why vendors such as IBM are delivering connectors that allow data to flow between the different systems.

The database technologies shown in Figure 2 are capable of supporting big data from the perspective of data volume, data variety and analytical workload complexity. To provide business agility, data must move along the information supply chain at a pace that matches the *action time* requirements of the business.

[6] The term *NoSQL* is frequently used to describe these systems, but this misleading because some of them support a subset of SQL. A *non-relational system* is a better term to use.

Figure 2. The three dimensions of big data


Analytical workload complexity and agility
most of the data is structured
Data volume
ADBMS
co-existence
Universal RDBMS
Non-Relational System
most of the data is multi-structured
Data variety

Action time is best explained in terms of the information supply chain (see Figure 1) where the input is raw source data and the output is business analytics. When processing data for business decision making the data has to be collected for analysis, analyzed, and the results delivered to the business user. The user then retrieves the results and decides if any action is required. There is always a time delay, or *latency*, between a business event occurring and the time the business user acts to resolve an issue or satisfy a requirement. This action time varies from application to application based on business needs. For a fraudulent credit card transaction it is the time taken from the credit card being used to it being rejected for fraud.

Reducing action time helps make organizations more agile. The business benefits obtained, however, must be balanced against the IT costs of supporting that level of service. In general, the closer the action time gets to real-time the higher the costs of supporting that action time.

Agility can be improved by analyzing data as it flows through systems

To reduce action time in the store and analyze approach, data has to be gathered and analyzed faster, and business users need to make faster decisions and take faster actions. In some cases this may not be possible for technology or cost reasons. Another way of reducing action time is to analyze the data as it flows through operational systems, rather than integrating it into a data store before analyzing it. This not only reduces action time, but also reduces storage, administration and security resources requirements. This can be thought of as an *analyze and store* approach. This approach analyzes data *in motion*, whereas the store and analyze approach analyzes data *at rest*.

Analyze and Store Approach

The **analyze and store** approach analyzes data as it flows through business processes, across networks, and between systems. The analytical results can then be published to interactive dashboards and/or published into a data store (such as a data warehouse) for user access, historical reporting and additional analysis. This approach can also be used to filter and aggregate big data before it is brought into a data warehouse.

There are two main ways of implementing the analyze and store approach:

Process analytics can be created by embedding analytical processing in business processes

- **Embedding the analytical processing in business processes.** This technique works well when implementing business process management and service-oriented technologies because the analytical processing can be called as a service from the process workflow. IBM supports this style of processing in its WebSphere product set. This technique is particularly useful for monitoring and analyzing business processes and activities in close to real-time – action times of a few seconds or minutes are possible here. The *process analytics* created can also be published to an operational dashboard or stored in a data warehouse for subsequent use.

Stream analytics can be created by analyzing data as it flows through networks and across systems

- **Analyzing streaming data** as it flows across networks and between systems. This technique is used to analyze data from a variety of different (possibly unrelated) data sources where the volumes are too high for the store and analyze approach, sub-second action times are required, and/or where there is a need to analyze the data streams for patterns and relationships. To date, many vendors have focused on analyzing event streams (from trading systems, for example) using the services of a complex event processing (CEP) engine, but this style of processing is evolving to support a wider variety of streaming technologies and data. IBM's InfoSphere Streams product, for example, supports a stream-processing engine that creates *stream analytics* from many types of streaming data such as event, video and GPS data.

The benefits of the analyze and store approach are fast action times and lower data storage overheads because the raw data does not have to be gathered and consolidated before it can be analyzed.

Figure 3 shows how systems for analyzing in-motion data can be combined with the existing business intelligence and data warehousing environment and data coming from non-relational systems such as Hadoop.

THE ROLE OF CLOUD COMPUTING

Cloud computing, like several of the other technologies discussed in this paper, covers a wide range of capabilities. One important and growing use of cloud computing is to deploy an integrated on-premises computing environment, or *private cloud.* This allows the many different workloads involved in both operational and analytical processing to co-exist and be consolidated into a single system. Some of these workloads may run as virtualized machines, whereas others may run as native systems.

A private cloud can be used to consolidate multiple workloads

The private cloud approach provides a consolidated, but flexible system that can scale to support multiple workloads. This solution, however, is not suited to all types of analytical processing. Not only does it take longer to implement a private cloud than a standalone appliance, but also it is often difficult to tune and optimize a cloud environment consisting of operational workloads as well as complex analytical workloads. These latter workloads are often best suited to an appliance approach, rather than a cloud-computing environment. It is therefore important to match business requirements and performance needs to the right type of technology solution and budget.

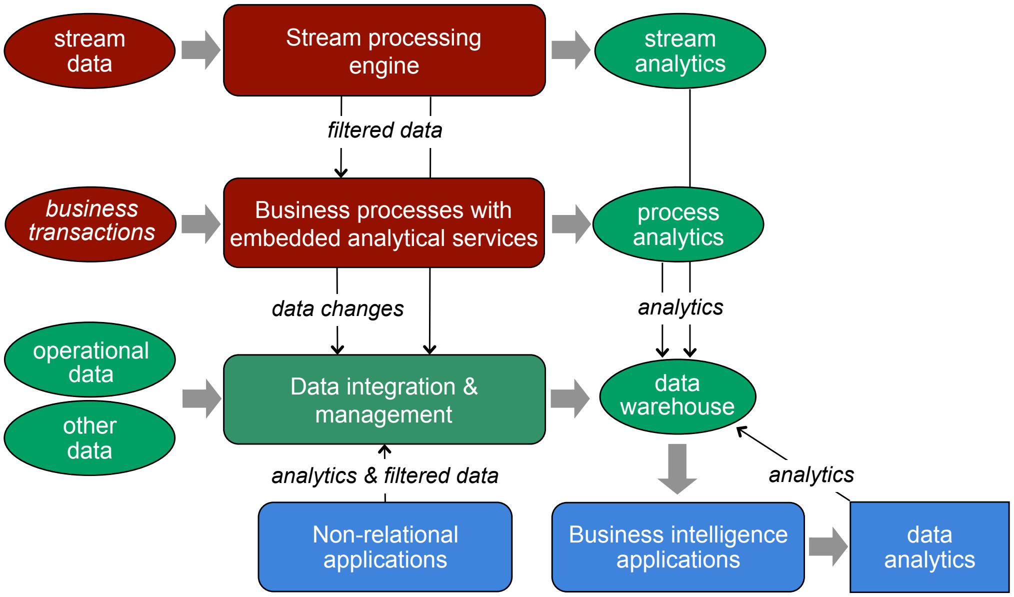


Figure 3. Integrating stream and process analytics into the traditional data warehousing environment

Big Data Storage Considerations

Several new technologies help companies maximize storage utilization

Many organizations are struggling to deal with increasing data volumes, and big data simply makes the problem worse. To solve this problem, organizations need to reduce the amount of data being stored and exploit new storage technologies that improve performance and storage utilization. From a big data perspective there are three important directions here:

- **Reducing data storage requirements** using data compression and new physical storage structures such as columnar storage. These technologies may be implemented in hardware and/or software. IBM, for example, has real-time hardware data compression on several of its disk systems that compress data before storage and decompress it on retrieval. This approach offloads the overheads of data compression from application processing systems. Note also that several non-relational products support data compression and columnar storage to reduce the overheads of storing big data.
- **Improving input/output (I/O) performance** using solid-state drives (SSDs). SSDs are especially useful for mixed analytical and enterprise data warehouse

workloads that involve large amounts of random data access. Sequential processing workloads gain less performance benefit from the use of SSDs.

- **Increasing storage utilization** by using tiered storage to store data on different types of devices based on usage. Frequently used *hot* data can be managed on fast devices such as SSDs, and less frequently accessed *cold* data can be maintained on slower and larger capacity hard disk drives (HDDs). This approach enables the storage utilization of the HDDs to be increased since they don't have to be underutilized to gain performance. System software moves the data between different storage types based on usage. The location of the data is transparent to applications. The use of SSDs and tiered storage can significantly improve both performance and throughput.

All of these technologies can be used to improve the price/performance of data storage in big data environments, and as organizations plan for big data they need to reevaluate their storage strategies to take advantage of these developments.

Conclusions

Big data adds several new high-volume data sources to the information supply chain. Several new and enhanced data management and data analysis approaches help the management of big data and the creation of analytics from that data. The actual approach used will depend on the volume of data, the variety of data, the complexity of the analytical processing workloads involved, and the responsiveness required by the business. It will also depend on the capabilities provided by vendors for managing, administering, and governing the enhanced environment. These capabilities are important selection criteria for product evaluation.

Big data involves more than just technology – senior management needs to understand the benefits of smarter decision making

An enhanced information supply chain, however, involves more than simply implementing new technologies. It requires senior management to understand the benefits of smarter and timelier decision making, and what IBM calls Smarter Computing. It also requires the business to make pragmatic decisions about the agility requirements for analyzing data and producing analytics given tight IT budgets. The good news is that many of the technologies outlined in this paper not only support smarter decision making, but also provide faster time to value.

Note that brand and product names mentioned in this paper may be the trademarks or registered trademarks of their respective owners.

About BI Research

BI Research is a research and consulting company whose goal is to help companies understand and exploit new developments in business intelligence, data management, and collaborative computing.