BIGDATA ANALYTICS

Big Data plays another role in today's businesses: Large organizations increasingly face the need to maintain massive amounts of structured and unstructured data—from transaction information in data warehouses to employee tweets, from supplier records to regulatory filings—to comply with government regulations. That need has been driven even more by recent court cases that have encouraged companies to keep large quantities of documents, e-mail messages, and other electronic communications, such as instant messaging and Internet provider telephony, that may be required for e-discovery if they face litigation.

WHERE IS THE VALUE?

Extracting value is much more easily said than done. Big Data is full of challenges, ranging from the technical to the conceptual to the operational, any of which can derail the ability to discover value and leverage what Big Data is all about.

Perhaps it is best to think of Big Data in multidimensional terms, in which four dimensions relate to the primary aspects of Big Data. These dimensions can be defined as follows:

1. **Volume.** Big Data comes in one huge size. Large Enterprises are awash with data, easily amassing terabytes and even petabytes of information.

2. Variety. Big Data extends beyond structured data to include unstructured data of all varieties: text, audio, video, click streams, log files, and more.

3. Veracity. The massive amounts of data collected for Big Data purposes can lead to statistical errors and misinterpretation of the collected information. Purity of the information is critical for value.

4. Velocity. Often time sensitive, Big Data must be used as it is streaming into the enterprise in order to maximize its value to the business, but it must also still be available from the archival sources as well.

Many of those technologies or concepts are not new but have come to fall under the umbrella of Big Data. Best defined as analysis categories, these technologies and concepts include the following:

Traditional business intelligence (BI). This consists of a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data. BI delivers actionable information, which helps enterprise users make better business decisions using fact-based support systems. BI works by using an in-depth analysis of detailed business data, provided by databases, application data, and other tangible data sources. In some circles, BI can provide historical, current, and predictive views of business operations.

Data mining. This is a process in which data are analyzed from different perspectives and then turned into summary data that are deemed useful. Data mining is normally used with data at rest or with archival data. Data mining techniques focus on modeling and knowledge discovery for predictive, rather than purely descriptive, purposes—an ideal process for uncovering new patterns from large data sets.

Statistical applications. These look at data using algorithms based on statistical principles and normally concentrate on datasets related to polls, census, and other static data sets. Statistical applications ideally deliver sample observations that can be used to study populated data sets for the purpose of estimating testing, and predictive analysis. Empirical data, such as surveys

and experimental reporting, are the primary sources for analyzable information.

Predictive analysis. This is a subset of statistical applications in which data sets are examined to come up with predictions, based on trends and information gleaned from databases. Predictive analysis tends to be big in the financial and scientific worlds, where trending tends to drive predictions, once external elements are added to the data set. One of the main goals of predictive analysis is to identify the risks and opportunities for business process, markets, and manufacturing.

Data modeling. This is a conceptual application of analytics in which multiple "what-if" scenarios can be applied via algorithms to multiple data sets. Ideally, the modeled information changes based on the information made available to the algorithms, which then provide insight to the effects of the change on the data sets. Data modeling works hand in hand with data visualization, in which uncovering information can help with a particular business endeavor.

Sources of Big Data

The biggest challenges for most organizations is finding data sources to use as part of their analytics processes. As the name implies, Big Data is large, but size is not the only concern. There are several other considerations when deciding how to locate and parse Big Data sets. The first step is to identify usable data. While that may be obvious, it is anything but simple. Locating the appropriate data to push through an analytics platform can be complex and frustrating. The source must be considered to determine whether the data set is appropriate for use. That translates into detective work or investigative reporting.
\
Considerations should include the following:

- Structure of the data (structured, unstructured, semistructured, table based, proprietary)

- Source of the data (internal, external, private, public)
- Value of the data (generic, unique, specialized)
- Quality of the data (verified, static, streaming)
- Storage of the data (remotely accessed, shared, dedicated platforms, portable)
- Relationship of the data (superset, subset, correlated)

All of those elements and many others can affect the selection process and can have a dramatic effect on how the raw data are prepared ("scrubbed") before the analytics process takes place.

BIG DATA SOURCES

The list can include many other internally tracked elements; however, it is critical to be aware of diminishing returns on investment with the data sourced. In other words, some log information may not be worth the effort to gather, because it will not affect the analytics outcome.
Finally, external data must be taken into account. There is a vast wealth of external information that can be used to calculate everything from customer sentiments to geopolitical issues. The data that make up the public portion of the analytics process can come from government entities, research companies, social networking sites, and a multitude of other sources.

- For example, a business may decide to mine Twitter, Facebook, the U.S. census, weather information, traffic pattern information, and news archives to build a complex source of rich data. Some controls need to be in place, and that may even include scrubbing the data before processing (i.e., removing spurious information or invalid elements).
- The richness of the data is the basis for predictive analytics. A company looking to increase sales may compare population trends, along with social sentiment, to customer feedback and satisfaction to identify where the sales process could be improved. The data warehouse can be used for much more after the initial processing, and realtime data could also be integrated to identify trends as they arise.

Many industries fall under the umbrella of new data creation and digitization of existing data, and most are becoming appropriate sources for Big Data resources. Those industries include the following:

Transportation, logistics, retail, utilities, and telecommunications :
Sensor data are being generated at an accelerating rate from fleet GPS transceivers, RFID (radiofrequency identification) tag readers, smart meters, and cell phones (call data records); these data are used to optimize operations and drive operational BI to realize immediate business opportunities.

Health care. The health care industry is quickly moving to electronic medical records and images, which it wants to use for short-term public health monitoring and long-term epidemiological research programs.

Government. Many government agencies are digitizing public records, such as census information, energy usage, budgets Freedom of Information Act documents, electoral data, and law enforcement reporting.

Entertainment media. The entertainment industry has moved to digital recording, production, and delivery in the past five years and is now collecting large amounts of rich content and user viewing behaviors.

Life sciences. Low-cost gene sequencing (less than $1,000) cangenerate tens of terabytes of information that must be analyzed to look for genetic variations and potential treatment effectiveness.

Video surveillance. Video surveillance is still transitioning from closed-caption television to Internet protocol television cameras and recording systems that organizations want to analyze for behavioral patterns (security and service enhancement).

Financial transactions. Thanks to the consolidation of global trading environments and the increased use of programmed trading, the volume of transactions being collected and analyzed is doubling or tripling. Transaction volumes also fluctuate much faster, much wider, and much more unpredictably. Competition among firms is creating more data, simply because sampling for trading decisions is occurring more frequently and at faster intervals.

Smart instrumentation. The use of smart meters in energy grid systems, which shifts meter readings from monthly to every 15 minutes, can translate into a multi thousand fold increase in data generated. Smart meter technology extends beyond just power usage and can measure heating, cooling, and other loads, which can be used as an indicator of household size at any given moment.

Mobile telephony. With the advances in smart phones and connected PDAs, the primary data generated from these devices have grown beyond caller, receiver, and call length. Additional data are now being harvested at exponential rates, including elements such as geographic location, text messages, browsing history, and even motions, as well as social network posts and application use.

The Nuts and Bolts of Big Data

Assembling a Big Data solution is sort of like putting together an erector set. There are various pieces and elements that must be put together in the proper fashion to make sure everything works adequately, and there are almost endless combinations of configurations that can be made with the components at hand. With Big Data, the components include platform pieces, servers, virtualization solutions, storage arrays, applications, sensors, and routing equipment. The right pieces must be picked and integrated in a fashion that offers the best performance, high efficiency, affordability, ease of management and use, and scalability.

THE STORAGE DILEMMA

Big Data consists of data sets that are too large to be acquired, handled, analyzed, or stored in an appropriate time frame using the traditional infrastructures. Big is a term relative to the size of the organization and, more important, to the scope of the IT infrastructure that's in place. The scale of Big Data directly affects the storage platform that must be put in place, and those deploying storage solutions have to understand that Big Data uses storage resources differently than the typical enterprise application does.

These factors can make provisioning storage a complex endeavor, especially when one considers that Big Data also includes analysis; this is driven by the expectation that there will be value in all of the information a business is accumulating and a way to draw that value out. Originally driven by the concept that storage capacity is inexpensive and constantly dropping in price, businesses have been compelled to save more data, with the hope that business intelligence (BI) can leverage the mountains of new data created every day.

Organizations are also saving data that have already been analyzed, which can potentially be used for marking trends in relation to future data collections. Aside from the ability to store more data than ever before, businesses also have access to more types of data. These data sources include Internet transactions, social networking activity, automated sensors, mobile devices, scientific instrumentation, voice over Internet protocol, and video elements. In addition to creating static data points, transactions can create a certain velocity to this data growth. For example, the extraordinary growth of social media is generating new transactions and records. But the availability of ever-expanding data sets doesn't guarantee success in the search for business value.

As data sets continue to grow with both structured and unstructured data and data analysis becomes more diverse, traditional enterprise storage system designs are becoming less able to meet the needs of Big Data. This situation has driven storage vendors to design new storage platforms that incorporate block- and file-based systems to meet the needs of Big Data and associated analytics.

Meeting the challenges posed by Big Data means focusing on some key storage ideologies and understanding how those storage design elements interact with Big Data demands, including the following:

- **Capacity:**

  Big Data can mean petabytes of data. Big Data storage systems must therefore be able to quickly and easily change scale to meet the growth of data collections. These storage systems will need to add capacity in modules or arrays that are transparent to users, without taking systems down. Most Big Data environments are turning to scale-out storage (the ability to increase storage performance as capacity increases)technologies to meet that criterion. The clustered architecture of scale-out storage solutions features nodes of storage capacity with embedded processing power and connectivity that can grow seamlessly, avoiding the silos of storage that traditional systems can create. Big Data also means many large and small files. Managing the accumulation of metadata for file systems with multiple large and small files can reduce scalability and impact performance, a situation that can be a problem for traditional network-attached storage systems. Object-based storage architectures, in contrast, can allow Big Data storage systems to expand file counts into the billions without suffering the overhead problems that traditional file systems encounter. Object-based storage systems can also scale geographically, enabling large infrastructures to be spread across multiple locations.

- **Security**:

  Many types of data carry security standards that are driven by compliance laws and regulations. The data may be financial, medical, or government intelligence and may be part of an analytics set yet still be protected. While those data may not be different from what current IT managers must accommodate, Big Data analytics may need to cross-reference data that have not been commingled in the past, and this can create some new security considerations. In turn, IT managers should consider the security footing of the data stored in an array used for Big Data analytics and the people who will access the data.

- **Latency:**

  In many cases, Big Data employs a real-time component, especially in use scenarios involving Web transactions or financial transactions. An example is tailoring Web advertising to each user's browsing history, which demands real-time analytics to function. Storage systems must be able to grow rapidly and still maintain performance. Latency produces "stale" data. That is another case in which scale-out architectures solve problems. The technology enables the cluster of storage nodes to increase in processing power and connectivity as they grow in capacity. Object-based storage systems can parallel data streams, further improving output. Most Big Data environments need to provide high input output operations per second (IOPS) performance, especially those used in high-performance computing environments. Virtualization of server resources, which is a common methodology used to expand compute resources without the purchase of new hardware, drives high IOPS requirements, just as it does in traditional IT environments. Those high IOPS performance requirements can be met with solid-state storage devices, which can be implemented in many different formats, including simple server-based cache to all-flash-based scalable storage systems.

- **Access:**

  As businesses get a better understanding of the potential of Big Data analysis, the need to compare different data sets increases, and with it, more people are bought into the data sharing loop. The quest to create business value drives businesses to look at more ways to cross-reference different data objects fro various platforms. Storage infrastructures that include global file systems can address this issue, since they allow multiple users on multiple hosts to access files from many different back-end storage systems in multiple locations.

- **Flexibility:**

  Big Data storage infrastructures can grow very large and that should be considered as part of the design challenge dictating that care should be taken in the design and allowing the storage infrastructure to grow and evolve along with the analytics component of the mission. Big Data storage infrastructures also need to account for data migration

challenges, at least during the start-up phase. Ideally, data migration will become something that is no longer needed in the world of Big Data, simply because the data are distributed in multiple locations.

- **Persistence:**

  Big Data applications often involve regulatory compliance requirements, which dictate that data must be saved for years or decades. Examples are medical information, which is often saved for the life of the patient, and financial information, which is typically saved for seven years. However, Big Data users are often saving data longer because they are part of a historical record or are used for time-based analysis. The requirement for longevity means that storage manufacturers need to include ongoing integrity checks and other long-term reliability features as well as address the need for data-in-place upgrades.

- **Cost:**

  Big Data can be expensive. Given the scale at which many organizations are operating their Big Data environments, cost containment is imperative. That means more efficiency as well as less expensive components. Storage de duplication has already entered the primary storage market and, depending on the data types involved, could bring some value for Big Data storage systems. The ability to reduce capacity consumption even by a few percentage points provides a significant return on investment as data sets grow. Other Big Data storage technologies that can improve efficiencies are thin provisioning, snapshots, and cloning.

- **Application awareness:**

  Initially, Big Data implementations were designed around application-specific infrastructures, such as custom systems developed for government projects or the white-box systems engineered by large Internet service companies. Application awareness is becoming common in mainstream storage systems and should improve efficiency or performance, which fits right into the needs of a Big Data environment.

- **Small and medium business:**
  The value of Big Data and the associated analytics is trickling down to smaller organizations, which creates another challenge for those building Big Data storage infrastructures: creating smaller initial implementation that can scale yet fit into the budgets of smaller organizations.

BUILDING A PLATFORM

Like any application platform, a Big Data application platform must support all of the functionality required for any application platform, including elements such as scalability, security, availability, and continuity. Yet Big Data Application platforms are unique; they need to be able to handle massive amounts of data across multiple data stores and initiate concurrent processing to save time. This means that a Big Data platform should include built-in support for technologies such as MapReduce, integration with external Not only SQL (NoSQL) databases, parallel processing capabilities, and distributed data services. It should also make use of the new integration targets, at least from development perspective. Consequently, there are specific characteristics and features that a Big Data platform should offer to work effectively with Big Data analytics processes:

- **Support for batch and real-time analytics:** Most of the existing platforms for processing data were designed for handling transactional Web applications and have little support for business analytics applications. That situation has driven Hadoop to become the de facto standard for handling batch processing. However, real-time analytics is altogether different, requiring something more than Hadoop can offer. An event processing framework needs to be in place as well. Fortunately, several technologies and processing alternatives exist on the market that can bring real-time analytics into Big Data platforms, and many major vendors, such as Oracle, HP, and IBM, are offering the hardware and software to bring real-time processing to the forefront. However, for the smaller business that may not be a viable option because of the cost. For now, real time processing remains a function that is provided as a service via the cloud for smaller businesses.

- **Alternative approaches**:

  Transforming Big Data application development into something more mainstream may be the best way to leverage what is offered by Big Data. This means creating a built-in stack that integrates with Big Data databases from the NoSQL world and creating MapReduce frameworks such as Hadoop and distributed processing. Development should account for the existing transaction-processing and event-processing semantics that come with the handling of the real-time analytics that fit into the Big Data world. Creating Big Data applications is very different from writing a typical "CRUD application" (create, retrieve, update, delete) for a centralized relational database. The primary difference is with the design of the data domain model, as well as the API and Query semantics that will be used to access and process that data. Mapping is an effective approach in Big Data, hence the success of MapReduce, in which there is an impedance mismatch between different data models and sources. An appropriate example is the use of object and relational mapping tools like Hibernate for building a bridge between the impedance mismatches.

- **Available Big Data mapping tools:** Batch-processing projects are being serviced with frameworks such as Hive, which provide an SQL-like facade for handling complex batch processing with Hadoop. However, other tools are starting to show promise. An example is JPA, which provides a more standardized JEE abstraction that fits into real-time Big Data applications. The Google app Engine uses Data Nucleus along with Bigtable to achieve the same goal, while GigaSpaces uses OpenJPA's JPA abstraction combined with an in-memory data grid. Red Hat takes a different approach and leverages Hibernate object-grid mapping to map Big Data.

- **Big Data abstraction tools:** There are several choices available to abstract data, ranging from open source tools to commercial distributions of specialized products. One to pay attention to is Spring Data from SpringSource, which is a high-level abstraction tool that offers the ability to map different data stores of all kinds into one common abstraction through annotation and a plug-in approach. Of course, one of the primary capabilities

offered by abstraction tools is the ability to normalize and interpret the data into a uniform structure, which can be further worked with. The key here is to make sure that whatever abstraction technology is employed deals with current and future data sets efficiently.

- **Business logic:** A critical component of the Big Data analytics process is logic, especially business logic, which is responsible for processing the data. Currently, MapReduce reigns supreme in the realm of Big Data business logic. MapReduce was designed to handle the processing of massive amounts of data through moving the processing logic to the data and distributing the logic in parallel to all nodes. Another factor that adds to the appeal of MapReduce is that developing parallel processing code is very complex. When designing a custom Big Data application platform, it is critical to make MapReduce and parallel execution simple. That can be accomplished by mapping the semantics into existing programming models. An example is to extend an existing model, such as Session Bean, to support the needed semantics. This makes parallel processing look like a standard invocation of single-job execution.

- **Moving away from SQL.** SQL is a great query language. However, it is limited, at least in the realm of Big Data. The problem lies in the fact that SQL relies on a schema to work properly, and Big Data, especially when it is unstructured, does not work well with schema-based queries. It is the dynamic data structure of Big Data that confounds the SQL schema-based processes. Here Big Data platforms must be able to support schema-less semantics, which in turn means that the data mapping layer would need to be extended to support document semantics. Examples are MongoDB, CouchBase, Cassandra, and the GigaSpaces document API. The key here is to make sure that Big Data application platforms support more relaxed versions of those semantics, with a focus on providing flexibility in consistency, scalability, and performance.

- **In-memory processing:** If the goal is to deliver the best performance and reduce latency, then one must consider using RAM-based devices and perform processing in-memory.

However, for that to work effectively, Big Data platforms need to provide a seamless integration between RAM and disk-base devices in which data that are written in RAM would be synched into the disk asynchronously. Also, the platforms need to provide common abstractions that allow users the same data access API for both devices and thus make it easier to choose the right tool for the job without changing the application code.

- **Built-in support for event-driven data distribution.** Big Data applications (and platforms) must also be able to work with event-driven processes. With Big Data, this means there must be data awareness incorporated, which makes it easy to route messages based on data affinity and the content of the message. There also have to be controls that allow the creation of fine-grained semantics for triggering events based on data operations (such as add, delete, and update) and content, a with complex event processing.

- **Support for public, private, and hybrid clouds:** Big Data applications consume large amounts of computer and storage resources. This has led to the use of the cloud and its elastic capabilities for running Big Data applications, which in turn can offer a more economical approach to processing Big Data jobs. To take advantage of those economics, Big Data application platforms must include built-in support for public, private, and hybrid clouds that will include seamless transitions between the various cloud platforms through integration with the available frameworks. Examples abound, such as JClouds and Cloud Bursting, which provides a hybrid model for using cloud resources as spare capacity to handle load.

- **Consistent management:** The typical Big Data application stack incorporates several layers, including the database itself, the Web tier, the processing tier, caching layer, the data synchronization and distribution layer, and reporting tools. A major disadvantage for those managing Big Data applications is that each of those layers comes with different management, provisioning, monitoring, and troubleshooting tools. Add to that the inherent complexity of Big Data applications, and effective management, along with the associated maintenance, becomes difficult. With that in mind, it becomes critical to

choose a Big Data application platform that integrates the management stack with the application stack. An integrated management capability is one of the best productivity elements that can be incorporated into a Big Data platform. Building a Big Data platform is no easy chore, especially when one considers that there may be a multitude of right ways and wrong ways to do it. This is further complicated by the plethora of tools, technologies, and methodologies available. However, there is a bright side that stresses flexibility, and since Big Data is constantly evolving, flexibility will rule in building a custom platform or choosing one off the shelf.