SCS5201 -BIG DATA ANALYTICS

Course material

UNIT 1

INTRODUCTION

Perhaps nothing will have as large an impact on advanced analytics in the coming years as the ongoing explosion of new and powerful data sources. When analyzing customers, for example, the day s of relying exclusively on demographics and sales history are past. Virtually every industry has at least one completely new data source coming online soon, if it isn't here already. Some of the data sources apply widely across industries; others are primarily relevant to a very small number of industries or niches. Many of these data sources fall under a new term that is receiving a lot of buzz: big data. Big data is sprouting up everywhere and using it appropriately will drive competitive advantage. Ignoring big data will put an organization at risk and cause it to fall behind the competition. To stay competitive, it is imperative that organizations aggressively pursue capturing and analyzing these new data sources to gain the insights that they offer. Analytic professionals have a lot of work to do! It won't be easy to incorporate big data alongside all the other data that has been used for analysis for years.

What exactly is Big Data?

At first glance, the term seems rather vague, referring to something that is large and full of information. That description does indeed fit the bill, yet it provides no information on what Big Data really is. Big Data is often described as extremely large data sets that have grown beyond the ability to manage and analyze them with traditional data processing tools. Searching the Web for clues reveals an almost universal definition, shared by the majority of those promoting the ideology of Big Data, that can be condensed into something like this: Big Data defines a situation in which data sets have grown to such enormous sizes that conventional information technologies can no longer effectively handle either the size of the data set or the scale and growth of the data set. In other words, the data set has grown so large that it is difficult to manage and even harder to garner value out of it. The primary difficulties are the acquisition, storage, searching, sharing, analytics, and visualization of data. There is much more to be said about what Big Data actually is. The concept has evolved to include not only the size of the data set but also the processes involved in leveraging the data. Big Data has even become synonymous with other business concepts, such as business intelligence, analytics, and data mining. Paradoxically, Big Data is not that new. Although massive data sets have been created in just the last two years, Big Data has its roots in the scientific and medical communities, where the complex analysis of massive amounts of data has been done for drug development, physics modeling, and other forms of research, all of which involve large data sets. Yet it is these very roots of the concept that have changed what Big Data has come to be.

THE ARRIVAL OF ANALYTICS

As analytics and research were applied to large data sets, scientists came to the conclusion that more is better—in this case, more data, more analysis, and more results. Researchers started to incorporate related data sets, unstructured data, archival data, and real-time data into the process, which in turn gave birth to what we now call Big Data. In the business world, Big Data is all about opportunity. According to IBM, every day we create 2.5 quintillion (2.5 3 1018) bytes of data, so much that 90 percent of the data in the world today has been created in the last two years. These data come from everywhere:

sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and cell phone GPS signals, to name just a few. That is catalyst for Big Data, along with the more important fact that all of these data have intrinsic value that can be extrapolated using analytics, algorithms, and other techniques.

Big Data has already proved its importance and value in several areas. Organizations such as the National Oceanic and Atmospheric Administration (NOAA), the National Aeronautics and Space Administration (NASA), several pharmaceutical companies, and numerous energy companies have amassed huge amounts of data and now leverage Big Data technologies on a daily basis to extract value from them.

NOAA uses Big Data approaches to aid in climate, ecosystem, weather, and commercial research, while NASA uses Big Data for aeronautical and other research. Pharmaceutical companies and energy companies have leveraged Big Data for more tangible results, such as drug testing and geophysical analysis. The New York Times has used Big Data tools for text analysis and Web mining, while the Walt Disney Company uses them to correlate and understand customer behavior in all of its stores, theme parks, and Web properties.

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.

BIGDATASOURCES

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.

Security Compliance Auditing and Protection

PROTECTING BIG DATA ANALYTICS

It is sad to report that protecting data is an often forgotten inclination in the data center, an afterthought that falls behind current needs. The launch of Big Data initiatives is no exception in the data center, and protection is too often an afterthought. Big Data offers more of a challenge than most other data center technologies, making it the perfect storm for a data protection disaster.

The real cause of concern is the fact that Big Data contains all of the things you don’t want to see when you are trying to protect data. Big Data can contain very unique sample sets—for example, data from devices that monitor physical elements (e.g., traffic, movement, soil pH, rain, wind) on a frequent schedule, surveillance cameras, or any other type of data that are accumulated frequently and in real time. All of the data are unique to the moment, and if they are lost, they are impossible to recreate.

That uniqueness also means you cannot leverage time-saving backup preparation and security technologies, such as deduplication; this greatly increases the capacity requirements for backup subsystems, slows down security scanning, makes it harder to detect data corruption, and complicates archiving. There is also the issue of the large size and number of files often found in

Big Data analytic environments. In order for a backup application and associated appliances or hardware to churn through a large number of files, bandwidth to the backup systems and/or the backup appliance must be large, and the receiving devices must be able to ingest data at the rate that the data can be delivered, which means that significant CPU processing power is necessary to churn through billions of files.

There is more to backup than just processing files. Big Data normally includes a database component, which cannot be overlooked. Analytic information is often processed into an Oracle, NoSQL, or Hadoop environment of some type, so real-time (or live) protection of that environment may be required. A database component shifts the backup ideology from a massive number of small files to be backed up to a small number of massive files to be backed up. That changes the dynamics of how backups need to be processed.

BIG DATA AND COMPLIANCE

Compliance issues are becoming a big concern in the data center, and these issues have a major effect on how Big Data is protected, stored, accessed, and archived. Whether Big Data is going to reside in the data warehouse or in some other more scalable data store remains unresolved for most of the industry; it is an evolving paradigm. However, one thing is certain: Big Data is not easily handled by the relational databases that the typical database administrator is used to working with in the traditional enterprise database server environment. This means it is harder to understand how compliance affects the data.

Big Data is transforming the storage and access paradigms to an emerging new world of horizontally scaling, unstructured databases, which are better at solving some old business problems through analytics. More important, this new world of file types and data is prompting analysis professionals to think of new problems to solve, some of which have never been attempted before. With that in mind, it becomes easy to see that a rebalancing of the database landscape is about to commence, and data architects will finally embrace the fact that relational databases are no longer the only tool in the tool kit. This has everything to do with compliance. New data types and methodologies are still expected to meet the legislative requirements placed on businesses by compliance laws. There will be no excuses accepted and no passes given if a new data methodology breaks the law.

The lessons learned to show that there is away to keep Big Data secure and in compliance. A combination of technologies has been assembled to meet four important goals:

1. **Control access by process, not job function**:

Server and network administrators, cloud administrators, and other employees often have access to more information than their jobs require because the systems simply lack the appropriate access controls. Just because a user has operating system–level access to a specific server does not mean that he or she needs, or should have, access to the Big Data stored on that server.

2. **Secure the data at rest:** Most consumers today would not conduct an online transaction without seeing the familiar padlock symbol or at least a certification notice designating that particular transaction as encrypted and secure. So why wouldn't you require the same data to be protected at rest in a Big Data store? All Big Data, especially sensitive information, should remain encrypted, whether it is stored on a disk, on a server, or in the cloud and regardless of whether the cloud is inside or outside the walls of your organization.

3. **Protect the cryptographic keys and store them separately from the data**.

Cryptographic keys are the gateway to the encrypted data. If the keys are left unprotected, the data are easily compromised. Organizations—often those that have cobbled together their own encryption and key management solution—will sometimes leave the key exposed within the configuration file or on the very server that stores the encrypted data. This leads to the frightening reality that any user with access to the server, authorized or not, can access the key and the data. In addition, that key may be used for any number of other servers. Storing the cryptographic keys on a separate, hardened server, either on the premises or in the cloud, is the best practice for keeping data safe and an important step in regulatory compliance. The bottom line is to treat key security with as much, if not greater, rigor than the data set itself.

4. **Create trusted applications and stacks to protect data from rogue users**.

You may encrypt your data to control access, but what about the user who has access to the configuration files that define the access controls to those data? Encrypting more than just the data and hardening the security of your overall environment—including applications, services, and configurations—gives you peace of mind that your sensitive information is protected from malicious users and rogue employees. There is still time to create and deploy appropriate security rules and compliance objectives. The health care industry has helped to lay some of the groundwork. However, the slow development of laws and regulations works in favor of those trying to get ahead on Big Data. Currently, many of the laws and regulations have not addressed the unique challenges of data warehousing. Many of the regulations do not address the rules for protecting data from different customers at different levels. Similarly, social media applications that are collecting tons of unregulated yet potentially sensitive data may not yet be a compliance concern. But they are still a security problem that if not properly addressed now may be regulated in the future. Social networks are accumulating massive amounts of unstructured data—a primary fuel for Big Data, but they are not yet regulated, so this is not a compliance concern but remains as a security concern.

There are still some very basic rules that should be used to enable security while not derailing the value of Big Data:

- **Ensure that security does not impede performance or availability**.

Big Data is all about handling volume while providing results, being able to deal with the velocity and variety of data, and allowing organizations to capture, analyze, store, or move data in real time. Security controls that limit any of these processes are a nonstarter for organizations serious about Big Data.

- **Pick the right encryption scheme**.

Some data security solutions encrypt at the file level or lower, such as including specific data values, documents, or rows and columns. Those methodologies can be cumbersome, especially for key management. File level or internal file encryption can also render data unusable because many applications cannot analyze encrypted data. Likewise, encryption at the operating system level, but without advanced key management and process based access controls, can leave Big Data woefully insecure. To maintain the high levels of performance required to analyze Big Data, consider a transparent data encryption solution optimized for Big Data.

- **Ensure that the security solution can evolve with your changing requirements.**

Vendor lock-in is becoming a major concern for many enterprises. Organizations do not want to be held captive to a sole source for security, whether it is a single server vendor, a network vendor, a cloud provider, or a platform. The flexibility to migrate between cloud providers and models based on changing business needs is a requirement, and this is no different with Big Data technologies. When evaluating security, you should consider a solution that is platform-agnostic and can work with any Big Data file system or database, including Hadoop, Cassandra, and MongoDB..

Best Practices for Big Data Analytics

The evolutionary aspect of Big Data tends to affect best practices, so what may be best today may not necessarily be best tomorrow. That said, there are still some core proven techniques that can be applied to Big Data analytics and that should withstand the test of time. With new terms, new skill sets, new products, and new providers, the world of Big Data analytics can seem unfamiliar, but tried and- true data management best practices do hold up well in this still emerging discipline. As with any business intelligence (BI) and/or data warehouse initiative, it is critical to have a clear understanding of an organization's data management requirements and a well-defined strategy before venturing too far down the Big Data analytics path. Big Data analytics is widely hyped, and companies in all sectors are being flooded with new data sources and ever larger amounts of information. Yet making a big investment to attack the Big Data problem without first figuring out how doing so can really add value to the business is one of the most serious missteps for would-be users.

START SMALL WITH BIG DATA

When analyzing Big Data, it makes sense to define small, high-value opportunities and use those as a starting point. Ideally, those smaller tasks will build the expertise needed to deal with the larger questions an organization may have for the analytics process. As companies expand the data sources and types of information they are looking to analyze, and as they start to create the all-important analytical models that can help them uncover patterns and correlations in both structured and unstructured data, they need to be vigilant about homing in on the findings that are most important to their stated business objectives. It is critical to avoid situations in which you end up with a process that identifies news patterns and data relationships that offer little value to the business process. That creates a dead spot in an analytics matrix where patterns, though new, may not be relevant to the questions being asked.

Successful Big Data projects tend to start with very targeted goals and focus on smaller data sets. Only then can that success be built upon to create a true Big Data analytics methodology that starts small and grows after the practice has served the enterprise rather well, allowing value to be created with little upfront investment while preparing the company for the potential windfall of information that can be derived from analytics. That can be accomplished by starting with "small bites" (i.e., taking individual data flows and migrating those into different systems for converged processing). Over time, those small bites will turn into big bites, and Big Data will be born. The ability to scale will prove important—as data collection increases, the scale of the system will need to grow to accommodate the data.

THINKING BIG

Leveraging open source Hadoop technologies and emerging packaged analytics tools makes an open source environment more familiar to business analysts trained in using SQL. Ultimately, scale will become the primary factor when mapping out a Big Data analytics road map, and business analysts will need to eschew the ways of SQL to grasp the concept of distributed platforms that run on nodes and clusters. It is critical to consider what the buildup will look like. It can be accomplished by determining how much data will need to be gathered six months from now and calculating how many more servers may be needed to handle it. You will also have to make sure that the software is up to the task of scaling. One big mistake is to be ignorant about the potential growth of the solution and the potential popularity of the solution once it is rolled into production.

AVOIDING WORST PRACTICES

There are many potential reasons that Big Data analytics projects fall short of their goals and expectations, and in some cases it is better to know what not to do rather than knowing what to do. This leads us to the idea of identifying "worst practices," so that you can avoid making the same mistakes that others have made in the past. It is better to learn from the errors of others than to make your own. Some worst practices to look out for are the following:

- **Thinking "If we build it, they will come."**

Many organizations make the mistake of assuming that simply deploying a data warehousing or BI system will solve critical business problems and deliver value. However, IT as well as BI and analytics program managers get sold on the technology hype and forget that business value is their first priority; data analysis technology is just a tool used to generate that value. Instead of blindly adopting and deploying something, Big Data analytics proponents first need to determine the business purposes that would be served by the technology in order to establish a business case—and only then choose and implement the right analytics tools for the job at hand. Without a solid understanding of business requirements, the danger is that project teams will end up creating a Big Data disk farm that really isn't worth anything to the organization, earning the teams an unwanted spot in the "data doghouse."

- **Assuming that the software will have all of the answers**.

Building an analytics system, especially one involving Big Data, is complex and resource-intensive. As a result, many organizations hope the software they deploy will be a magic bullet that instantly does it all for them. People should know better, of course, but they still have hope. Software does help, sometimes dramatically. But Big Data analytics is only as good as the data being analyzed and the analytical skills of those using the tools.

- **Not understanding that you need to think differently.**

Insanity is often defined as repeating a task and expecting different results, and there is some modicum of insanity in the world of analytics. People forget that trying what has worked for them in the past, even when they are confronted with a different situation, leads to failure. In the case of Big Data, some organizations assume that big just means more transactions an large data volumes. It may, but many Big Data analytics initiatives involve unstructured and semi structured information that needs to be managed and analyzed in fundamentally different ways than is the case with the structured data in enterprise applications and data warehouses. As a result, new methods and tools might be required to capture, cleanse, store, integrate, and access at least some of your Big Data.

- **Forgetting all of the lessons of the past.**

Sometimes enterprises go to the other extreme and think that everything is different with Big Data and that they have to start from scratch. This mistake can be even more fatal to a Big Data analytics project's success than thinking that nothing is different. Just because the data you are looking to analyze are structured differently doesn't mean the fundamental laws of data management have been rewritten.

- **Not having the requisite business and analytical expertise**.

A corollary to the misconception that the technology can do it all is the belief that all you need are IT staffers to implement Big Data analytics software. First, in keeping with the theme mentioned earlier of generating business value, an effective Big Data analytics program has to incorporate extensive business and industry knowledge into both the system design stage and ongoing operations. Second, many organizations underestimate the extent of the analytical skills that are needed. If Big Data analysis is only about building reports and dashboards, enterprises can probably just leverage their existing BI expertise. However, Big Data analytics typically involves more advanced processes, such as data mining and predictive analytics. That requires analytics professionals with statistical, actuarial, and other sophisticated skills, which might mean new hiring for organizations that are making their first forays into advanced analytics.

- **Treating the project like a science experiment**.

Too often, companies measure the success of Big Data analytics programs merely by the fact that data are being collected and then analyzed. In reality, collecting and analyzing the data is just the beginning. Analytics only produces business value if it is incorporated into business processes, enabling business managers and users to act on the findings to improve organizational performance and results. To be truly effective, an analytics program also needs to include a feedback loop for communicating the success of actions taken as a result of analytical findings, followed by a refinement of the analytical models based on the business results.

- **Promising and trying to do too much.**

Many Big Data analytics projects fall into a big trap: Proponents oversell how fast they can deploy the systems and how significant the business benefits will be. Overpromising and under delivering is the surest way to get the business to walk away from any technology, and it often sets back the use of the particular technology within an organization for a long time—even if many other enterprises are achieving success. In addition, when you set expectations that the benefits will come easily and quickly, business executives have a tendency to underestimate the required level of involvement and commitment. And when a sufficient resource commitment isn't there, the expected benefits usually don't come easily or quickly—and the project is labeled a failure.

BABY STEPS

It is said that every journey begins with the first step, and the journey toward creating an effective Big Data analytics holds true to that axiom. However, it takes more than one step to reach a destination of success. Organizations embarking on Big Data analytics programs require a strong implementation plan to make sure that the analytics process works for them. Choosing the technology that will be used is only half the battle when preparing for a Big Data initiative. Once a company identifies the right database software and analytics tools and begins to put the technology infrastructure in place, it's ready to move to the next level and develop a real strategy for success. The importance of effective project management processes to creating a successful Big Data analytics program also cannot be overstated. The following tips offer advice on steps that businesses should take to help ensure a smooth deployment:

- **Decide what data to include and what to leave out**.

By their very nature, Big Data analytics projects involve large data sets. But that doesn't mean that all of a company's data sources, or all of the information within a relevant data source, will need to be analyzed. Organizations need to identify the strategic data that will lead to valuable analytical insights. For instance, what combination of information can pinpoint key customer-retention factors? Or what data are required to uncover hidden patterns in stock market transactions? Focusing on a project's business goals in the planning stages can help an organization home in on the exact analytics that are required, after which it can—and should—look at the data needed to meet those business goals. In some cases, this will indeed mean including everything. In other cases, though, it means using only a subset of the Big Data on hand.

- **Build effective business rules and then work through the complexity they create.**

Coping with complexity is the key aspect of most Big Data analytics initiatives. In order to get the right analytical outputs, it is essential to include business focus data owners in the process to make sure that all of the necessary business rules are identified in advance. Once the rules are documented, technical staffers can assess how much complexity they create and the work required to turn the data inputs into relevant and valuable findings. That leads into the next phase of the implementation.

- **Translate business rules into relevant analytics in a collaborative fashion**.

Business rules are just the first step in developing effective Big Data analytics applications. Next, IT or analytics professionals need to create the analytical queries and algorithms required to generate the desired outputs. But that shouldn't be done in a vacuum. The better and more accurate that queries are in the first place, the less redevelopment will be required. Many projects require continual reiterations because of a lack of communication between the project team and business departments. Ongoing communication and collaboration lead to a much smoother analytics development process.

- **Have a maintenance plan**.

A successful Big Data analytics initiative requires ongoing attention and updates in addition to the initial development work. Regular query maintenance and keeping on top of changes in business requirements are important, but they represent only one aspect of managing an analytics program. As data volumes continue to increase and business users become more familiar with the analytics process, more questions will inevitably arise. The analytics team must be able to keep up with the additional requests in a timely fashion. Also, one of the requirements when evaluating Big Data analytics hardware and software options is assessing their ability to support iterative development processes in dynamic business environments. An analytics system will retain its value over time if it can adapt to changing requirements.

- **Keep your users in mind—all of them**.

With interest growing in self-service BI capabilities, it shouldn't be shocking that a focus on end users is a key factor in Big Data analytics programs. Having a robust IT infrastructure that can handle large data sets and both structured and unstructured information is important, of course. But so is developing a system that is usable and easy to interact with, and doing so means taking the various needs of users into account. Different types of people— from senior executives to operational workers, business analysts, and statisticians—will be accessing Big Data analytics applications in one way or another, and their adoption of the tools will help to ensure overall project success. That requires different levels of interactivity that match user expectations and the amount of experience they have with analytics tools—for instance, building dashboards and data visualizations to present findings in an easy-to-understand way to business managers and workers who aren't inclined to run their own Big Data analytics queries. There's no one way to ensure Big Data analytics success. But following a set of frameworks and best practices, including the tips outlined here, can help organizations to keep their Big Data initiatives on track. The technical details of a Big Data installation are quite intensive and need to be looked at and considered in an in-depth manner. That isn’t enough, though: Both the technical aspects and the business factors must be taken into account to make sure that organizations get the desired outcomes from their Big Data analytics investments.