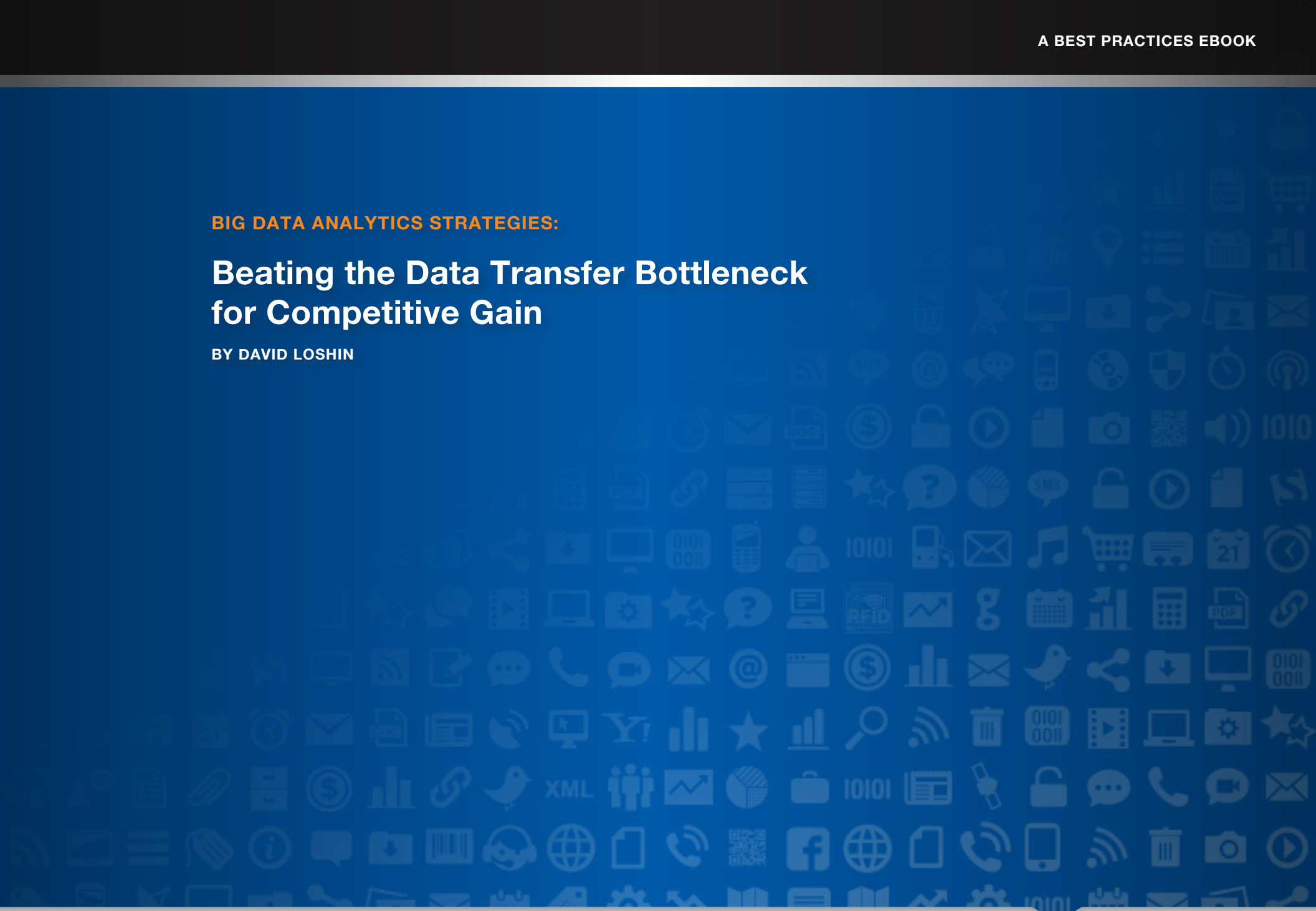
A BEST PRACTICES EBOOK

BIG DATA ANALYTICS STRATEGIES:

Beating the Data Transfer Bottleneck for Competitive Gain

BY DAVID LOSHIN



ATTUNITY

Please feel free to link to this eBook on your blog, social media pages or email it to whomever you believe would benefit from reading it. Thank you.

Table of Contents

Foreward: Information Availability. 4

Chapter 1: The State of the Market - Big Data 5
The Democratization of Analytics and the Demand for Data 5
Provisioning Data to Support Pervasive Analytics 5
Big Data Analytics and Information Availability 6
Data Demand, Bottlenecks, and the Need for High-Performance Data Replication 7

Chapter 2: Big Data and the Challenges of Information Delivery 8
Characteristics of Big Data Solutions 8
Operational Implications Resulting from the Information Delivery Bottleneck 9
Analytical Implications Resulting from the Information Delivery Bottleneck 10
Breaking the Information Delivery Bottleneck 11

Chapter 3: Technical Objectives for High-Performance Information Delivery 12
Three Key Characteristics of the Information Delivery Bottleneck 12
Reducing Data Latency 13
Broadening Data Accessibility 14
Maintaining Consistency 15

Chapter 4: Technical Componentry for High Performance Information Delivery 15
Technical Requirements for Big Data and Information Availability 15
Replication and Change Data Capture 16
Connectivity 17
Federation and Virtualization 18
Considerations 18

About the Author 20

About Attunity 20

Foreward: Information Availability

Anyone reading the popular technology-oriented press is bound to have been inundated with articles, papers, and webcasts touting the wonders of big data analytics: data scientists mining massive data volumes using analytics appliances or other high performance and scalable systems to deliver actionable intelligence to key decision-makers across the organization. Anyone with even a slight familiarity with the big data meme would come to an obvious conclusion: big data coupled with analytics is the wave of the future due to the competitive advantage it can provide. This seems to hold true whether your organization is large or even a small- or medium-sized business.

In this ebook, we systematically examine the state of the big data market to understand some of the real issues that have already been identified regarding information availability. We will also help the reader properly anticipate the data accessibility and utilization hurdles that tend to plague those who are unprepared to support the imminent kickoff of big data analytics projects. In chapter 1 we examine the state of the market for big data and big data analytics. Specifically, we focus on the coupled needs for more expansive availability of analytical results distributed to a variety of (both human and automated) consumers, and how that increasingly depends on big data analytics. We also highlight that the key to high-performance big data is ensuring high-speed accessibility is unconstrained by the typical data transfer bottlenecks.

Chapter 2 drills into the challenges of information delivery for big data. By understanding the operational and the analytical impacts of constrained data accessibility and high data latency, we can begin to consider the facets of the information delivery challenge. In the context of this information, chapter 3 then scopes the technical objectives for high-performance information delivery by looking at key characteristics of the data transfer bottleneck while discussing what needs to be addressed. This includes reducing data latency, broadening data accessibility, and maintaining consistency.

Chapter 4 looks at the types of tools and technologies that can be used to solve high-performance information availability challenges. The chapter examines the requirements, and then considers how techniques such as replication, change data capture (CDC), connectivity/accessibility, data federation, and data virtualization can all be used to contribute to a reasonable solution. Finally, we summarize with some best practice considerations for moving forward with a big data analytics initiative.

David Loshin
President
Knowledge Integrity Incorporated

Chapter 1: The State of the Market - Big Data

The Democratization of Analytics and the Demand for Data

When you review the body of work surrounding the big data phenomenon, there are some similar themes about the need for scalability that run through much of the articles and presentations. Some of these can be summarized as:

- The need for a big data strategy: The explosive expansion of data volumes is going to overwhelm your organization if you do not have a "big data" strategy.
- Data Sources: The expectation that organizations be capable of accessing and combining data pushed from many different sources with large amounts of "legacy data" collected and archived over many (possibly dozens) of years.
- Data Type Variety: The types of data that must be absorbed are varied in their structure, content, and organization, and go way beyond the typical sets of structured records generated by batch transaction processing systems.
- Delivery: The results of reporting and analytics inform both strategic and operational decision-makers across broad spectrums of use — up and down the management chain.
- Data Volume: The accelerating rates at which data is made available will confound organizations that cannot rapidly absorb and analyze this data and then react in real time.

These can all be more succinctly summarized to say that organizations must embrace business analytics of massive data volumes to remain competitive. At the same time, there is a single thread of **data criticality** that weaves itself across all of these themes and goes beyond abstract concepts implied by an enterprise information strategy. This introduces one of the core challenges to success: there is an increased expectation for rapid and accurate decision-making that is tightly coupled with dwindling tolerance for non-real-time responses. This highlights the importance of what we can refer to as *information availability.*

The concept of "democratization of analytics" suggests that there is an expanding user base for insightful knowledge in which the benefits of analysis are made available to a broad set of user constituencies and types of decision-makers. And this broader dependence on analytics results has exposed the need for high-performance approaches for making consistent views of information available and accessible to those user communities when that information is needed.

Provisioning Data to Support Pervasive Analytics

In the early days of the decision-support systems that preceded the modern data warehouse, the scope of reporting and analytics focused attention on reviews of historical transaction data. This was used to provide insights to senior-level managers to drive strategic decision-making. The high hardware costs, intense work effort, and computational performance associated with extracting data from transactional sources as well as moving that data into the decision-support system placed severe limitations on use. This restricted the delivery of reports to key staff members over an extended period of time.

However, flash forward twenty years to the point where most of the restrictions have been effectively lifted. For example, comparative hardware costs have plummeted: today's laptops are sporting the same computational power of 1990's-vintage supercomputers, while manufacturers are packaging orders of magnitude more storage space into ever-shrinking footprints. The significant effort associated with extraction, transformation, and loading is now simplified using a set of commodity products for data integration, of which a number are available as open-source tools. And the community of users has drastically increased, as operational business intelligence provides actionable insights to middle-managers. Reports can now be pushed out directly to individuals outside the enterprise. In effect, business intelligence has migrated away from the C-level set as a larger and more varied community of users can take advantage of results flowing out of data warehouses and analytical appliances.

In fact, the idea of pervasive or integrated analytics melds the use of business intelligence reporting and analysis with ongoing operational processes. In some cases, the users are not even aware that they are consuming the results of analyses, as notifications and alerts are engineered directly into operational applications in real time, such as real-time scripting for call center representatives that adapts in relation to customer responses to prompts.

While most of the barriers to usability have been lowered, one significant barrier remains: making information available in a timely-enough fashion to meet the needs of the business. Increased disk capacity, computational power, network bandwidth, and user demand has only whetted the appetite for more data. Because more data is flowed into analytical environments and more reports and information streams out of the business intelligence framework, data provisioning is the only aspect of the analytics process that has not scaled with the remaining aspect of the infrastructure. In essence, when it comes to integrating analytics across the enterprise, any inability to provide timely access to consistent and up-to-date data continues to be a significant bottleneck to productivity and accurate business decision-making. Addressing any gaps in timely data accessibility requires alternative approaches for ensuring high-performance information availability.

Big Data Analytics and Information Availability

Indeed, over the past twenty years, increased hardware performance coupled with innovations in data management and integration tools has created an environment in which the spectrum of business intelligence users has expanded. This, in turn, has only added to the demand for information availability. This chapter highlights an alternate set of drivers for information availability, namely the potential of harnessing and analyzing massive data sets for business value creation.

The appeal of big data analytics stems from three specific perceptions of benefit:

1. The expectation of being able to consume very large data sets and separate relevant business "signals" from the massive amount of "noise."
2. Scalable combination of commodity components lowers the barrier to entry.
3. Ease of implementation and use, especially in contrast to some of the effort associated with monolithic enterprise data warehouses.

The promise of big data analytics additionally reflects the “democratization of business intelligence” effect suggested previously in that it presents a methodology for rapidly conjuring up an environment that, once the massive amounts of data are loaded, can enable a wide spectrum of investigations and analyses in a way that is both scalable and flexible.

Of course, that one qualification: once the massive amounts of data are loaded: remains the kicker, as it did for enterprise data warehousing and integrated pervasive business intelligence. Big data platforms are configured to take advantage of massive parallelism, and their incorporation of commodity computational and storage componentry enables elasticity and scalability within the platform environment. However, the challenges that often accompany this part of the process include:

- Data sets to be analyzed generally do not originate within the analysis platform;
- The data comes in different sizes and formats; and
- Much of that data is bound to be completely absent of structure.

In addition, Internet bandwidth tends to be insufficient to satisfy timely delivery of massive data sets, only adding to the complexity and bottlenecks. The data sets will need to be collected, collated, and then brought into the big data server before any of the analyses can begin.

In other words, even with the promise of high-performance calculation and computation, the analysis processes are still limited by the challenges of making information available in ‘right’ time. Environments that have not considered approaches - such as data replication for improving the speed of data movement - will still be constrained by the data access bottleneck.

Data Demand, Bottlenecks, and the Need for High-Performance Data Replication

Today, there are several emerging business intelligence and analytics capabilities that are rapidly gaining mainstream acceptance. In particular, methods for pervasive, integrated business intelligence and big data analytics enable environments in which actionable insight can be seamlessly delivered in real time to a quickly-expanding community of business users.

The performance characteristics of new technologies enable greater acceptance, leading to increased demand for actionable intelligence. However, the increased demand for actionable insight simultaneously increases the requirement for the raw data to be absorbed into the analytical platforms. And as we have discussed, data access becomes the bottleneck for performance. In essence, the same market factors that drive the demand for high-performance information delivery simultaneously create the bottlenecks that impede its success! This is not an unexpected phenomenon. It is similar to the concept that building new highways leads to an increase in highway driving: the more capacity presented, the greater the likelihood that more people will use that capacity.

In this highway example, congestion will be increased as more cars look to enter the highway, and without engineering on-ramps that smoothly migrate cars onto the highway, you end up with a bottleneck. This is similar to the need for moving larger and larger data sets into the high-performance analytical platforms. That means that any organization seeking to evolve its information strategy to incorporate programmatic data analytics and BI must assess the potential demand for data availability and data access. They must also consider methods (such as high-performance data replication) for relieving the bottlenecks caused by data access and delivery latencies to satisfy the expectation of real-time (or near real-time) analytics performance.

Chapter 2: Big Data and the Challenges of Information Delivery

Characteristics of Big Data Solutions

The byproduct of the growing fascination with the cross-bred concepts of business analytics and big data is the expectation of "out-of-the-box" miracles resulting from the installation of a big data analytics platform. In reality, people are still trying to get their hands around the concepts of what "big data" really means from a practical perspective, especially in contrast to what is already in place in the organization.

Let's simplify the discussion by agreeing on a straightforward description of some of the key characteristics of a big data analytics platform and the types of applications that can use it:

- Elastic: the platform, built on commodity CPU and disk hardware, can dynamically expand and contract its use of available resources based on the data and computational demand.
- Scalable: the computational and storage capacity of the platform scales linearly with the amount of resources employed.
- Parallelizable: big data applications developed for analytics must be amenable to both task parallelism (in which specific jobs are allocated to the pool of computational resources) and data parallelism (in which the same of similar tasks execute at different computation nodes on chunks of distributed data).
- Large data volumes: of course, the platform must satisfy the application's needs for massive amounts of data from a variety of sources.

At the same time, developing big data applications requires understanding of the platform, whether at the conceptual level (for packaged software + hardware solutions) or at the granular level (for the technologists ready to download the popular open source platform Hadoop and deploy it to a configuration of their own making). And a review of much of the popular literature out there will leave it at that, or perhaps augment those ideas with a simplistic example of an application developed using the MapReduce programming model. This includes the seemingly universal "word counting" algorithm for scanning massive amounts of web pages as a way to generate a reasonable index.

There are two obvious conclusions that could be drawn from reading these articles. The first is that the installation of a big data platform is sufficient for these big data miracles to happen out of the box, and the second is that any application written using MapReduce is going to significantly reduce application runtimes. However, you would be mistaken if you were to presume either of these to be true across the board.

While it is true that implementing a scalable high-performance execution platform should lead to improved performance, there is one basic thing the pundits writing these articles tend to ignore: getting the massive amounts of data loaded into (and sometimes moving the data around) the big data platform. Yet the delays associated with the typical approaches to massive data delivery will drown out any performance benefit expected from the high-performance execution platform, and this will have severe impacts to any area of the business that anticipates big data benefits. After all, data that is outdated can be inaccurate. This could be a critical factor affecting effective decision-making and competitive edge.

Operational Implications to Business Resulting from the Information Delivery Bottleneck

If big data is expected to be a mainstay of the operational infrastructure for a broad range of companies (that is, both large companies as well as small/medium-sized businesses), that implies the need for acquiring and integrating those massive data sets that give big data its name. However, the lack of understanding that the bottleneck causes by latency in data delivery leads to missed opportunities for performance improvement. In fact, the problem of the bottleneck in delivering data into big data platforms is significantly exacerbated by the (somewhat ingenuous) expectation that massive data sets can be easily provisioned over the Internet. In some cases, the information delivery delays are so oppressive that express mailing hard drives loaded with data sets is preferable to access over the Web!

But in lieu of attempting to engineer a solution on the fly, it is valuable to take a step back and consider some of the business challenges that are manifested as a result of data access and delivery bottlenecks, both in terms of analytical and operational business application scenarios. Some of the implications affect day-to-day operations, such as:

- Longer durations of key cross-functional business processes: Organizations are increasingly recognizing the need for monitoring cross-functional process performance, with key metrics focusing on the end-to-end duration of the process. For example, one key measure for the order-to-cash process is the order-to-fulfill process time, with the goal to reduce that cycle time. Increased operational delays that are related to data availability (slowed by the data latency bottleneck) are going to increase the duration of those kinds of processes; your operational performance metric is going to be negatively impacted even though the root cause has nothing to do with your team's actual performance.

- Delayed accessibility to "warm" data: One of the promises of big data technology is the ability to create interim archival frame works using tools like the Hadoop Distributed File System (HDFS). HDFS enables users to store and manage large data sets that might be subject to accessibility for various types of business processes (a good example is e-discovery in the legal industry). These "warm data fields" must be loaded with the source data, and some of that data is expected to be sourced both from internal and external locations. Reliance on delivery over traditional channels such as the Internet will prove to be a source of congestion and therefore — indigestion.

- Scalability of business intelligence and analytics: When the demand for integrated, "right-time" analytics increases (especially as more business applications embed the use of analytical results within standard operating procedures), there is a corresponding increased need for increased data volumes. This is shared with a broader constituency of users, as well as

high-speed data accessibility and expectations of data currency. These are impacted by sluggishness in information availability.

The above two examples are specific cases of a more general operational failure of the Information Technology groups to meet agreed-to service levels across organizations as a result of inadvertent narrowing of the channels for data movement.

Analytical Implications to Business Resulting from the Information Delivery Bottleneck

To a large extent, the messages of the big data movement are clearly targeted to aspirants in the analytics space. In other words, there is an expectation that by engaging some data scientists to create a big data analytics platform, scraping together a handful of data sets from various sources, and getting started on some analysis projects, new business opportunities will magically appear.

While the task of creating those platforms is straightforward, the task of making it work right is not, especially because of the data delivery challenge. For example, consider the challenge of developing and debugging big data applications. This may include trying to run and rerun the program multiple times until believable results can be achieved in a big data environment. And this exercise may require multiple iterations of determining that the data was not set up right to begin with. This would necessitate many passes in loading and reloading the data. At the same time, the algorithms may need to be tinkered with as the data sources morph unpredictably.

As with the impacts to operational system performance, the information delivery bottleneck ultimately will impact analytical applications as well, with these kinds of examples:

- Delays or deferrals of decision-making: While "decision-making speed" is not a typical corporate performance indicator, those delays in presenting analytical results that propagate from delays in information delivery may prevent individuals (and automated systems) from taking action within the reasonable window of opportunity. This ties back to my previous point about right-time integrated analytics.
- Reliability of the decisions that are made: Business people making decisions in real time based on big data analytics expect the source data to be current and consistent. Latency in information delivery contributes to confusion regarding the currency of right-time intelligence.
- Inability of data scientists to create the number of models necessary to support the business: Similar to the challenge of the developing (and debugging) process, data scientists are going to be pressured to continually develop new models to support emerging consumer analytics demands. Each of these tasks will depend on accessibility to massive data sets, and that means that low-latency data movement is bound to be a critical success factor.

Actually, that last bullet item bears a little deeper consideration. In a typical analytics development life cycle, the data analyst must develop a model and train that model using a training data set. For a big data application, that development cycle will require training data sets of massive volume. And for each iteration, the results of the model must be evaluated to determine a level of confidence in its predictive capability. That also requires loading massive data sets.

In other words, almost all aspects of model development and testing depend on the availability of many huge data sets that need to be loaded onto the big data development platform. Information availability becomes the bottleneck of the development process. More to the point: delays in the ability to load massive data sets effectively strangle the data scientists' ability to develop new analytical applications in a timely manner.

In summary, we are left with one basic conclusion: increased latency in data movement will reduce or even eliminate the effectiveness of the expected benefits of big data and the growing expectation for right-time analytics integration.

Breaking the Information Delivery Bottleneck

In this chapter we have considered that in those organizations that are introducing big data technologies and using large (or massive) data sets and volumes, the key stakeholders must recognize that ensuring low-latency data transport and access is a critical success factor. To anyone who has been working in high-performance computing for any length of time, this will not come as a shock, since the processing delays associated with data access latency has always dwarfed the delays associated with the computational resources.

The appearance of "big data" in the pages of the New York Times and the Wall Street Journal signal the mainstreaming of high-performance computing. Yet the focus continues to be on the magic, not the mechanics, and clumsy attitudes about information availability will kill the "big data buzz" long before the program will gain any traction.

It has been suggested to me that the absence of attention to reducing latency is a form of incompetence, but I think that is too strong a condemnation. Rather, I suggest that there are existing techniques that have been in use for decades that can be adapted to today's need for high-speed data access, reduced transport latency, and improved data currency. To truly assess how existing techniques can solve the problem, we need to review the requirements in finer detail, and that is the focus of the next chapter.

Chapter 3: Technical Objectives for High-Performance Information Delivery

Three Key Characteristics of the Information Delivery Bottleneck

Having examined the business implications of a bottleneck in delivering information to the target locations for analysis, especially in the context of big data applications, it is worth reviewing how those business impacts are typically a direct result of three key characteristics of the information delivery bottleneck:

- High data access latency: The amount of time it takes to move massive volumes of data (both archived and streamed, as well as both structured and unstructured) becomes the delaying factor in the production flow for delivery and integration of right-time analytics into everyday decision-making scenarios.
- Data accessibility issues: The wide variety of data sources, systems and platforms from which data needs to be accessed poses a challenge to many business processes and situations.
- Questionable data consistency and synchronization: Integrating data about the same fundamental sets of real-world entities (such as customers or products) from different sources is bound to present consistency issues, especially when considering streamed social media transactions, real-time transactions, and more static profiles. The challenge of synchronization — ensuring that the combined sequences of transactions are properly ordered when viewed together — can introduce questions about data consistency and trustworthiness.

In short, any organization that desires to take advantage of big data techniques must put into place the technical solutions that address the information delivery bottleneck before attempting to design and develop any kind of large-scale applications. Simply put, that means acquiring and instituting the tools and techniques that alleviate the aforementioned pains. To provide some clarity and define the characteristics of a reasonable information delivery solution, it must, purely and simply:

- Reduce data latency
- Broaden data accessibility
- Maintain data consistency

As we will see in the next few sections, any framework supporting big data and big data analytics must seek to address these directives.

Reducing Data Latency

If delays in information delivery pose the bottleneck, then reducing data latency should ease that bottleneck. The implication, then, is that one first needs to determine the root causes of data latency and then identify ways to eliminate those root causes. For the most part, the root causes are relatively straightforward as can be seen during the process of moving data from the source to its target:

- Complexity in extracting or accumulating data from the source: The source system may be blamed as the culprit here, as the speed of execution of queries used to extract and then bundle result sets is presumed to be dependent on the underlying system's performance characteristics. However, it is important to recognize two other contributing factors: the complexity of the queries used for extraction and the level of performance tuning skill necessary for the person developing those queries.
- Load balancing: the source systems are already presumably being used for existing operational and transaction processing. Adding a complex query for extraction means more competition for CPU and disk resources along with the consequential delays added to all processing.
- Data volume: The expectation is that absorbing and analyzing large data sets will be more the norm than the exception. The time for extracting and preparing massive data sets is (obviously) related to the size of the data set.
- Constricted bandwidth for transporting data between source and target: This is key; the typical sets of direct I/O channels and network bandwidth are configured for general operations, not for massive data migration.
- Memory hierarchy delays at the disk level: data sets are often extracted, sent to a staging area, and then relayed to the target to be loaded. Storage to disk introduces delays for both writing and the subsequent reading in preparation for loading.
- Slowness of bulk data loading into the target: similar to the extraction side, bulk data loads may also drag down overall performance.

Reflecting on these root causes, the obvious approaches to reducing data latency would combine one or more of these notions:

- Reducing complexity in data extraction
- Avoiding a load conflict
- Effectively managing data volumes and reducing volume when possible
- Utilizing greater bandwidth when possible
- Avoiding writing data to disk
- Speeding bulk data loads

Broadening Data Accessibility

Reducing data access latency addresses a facet of the problem. In this section we consider the set of issues that center on data accessibility. There is a wide assortment of data sources, ranging from legacy difficult-to-extract sources all the way to unstructured data streams whose characteristics can change rapidly. Therefore, improving access capabilities to these different sources is critical.

Even without considering opportunities for absorbing and making sense of unstructured data, the variety of potential sources of structured data is quite wide. Organizations have been capturing and archiving data for decades, employing a variety of data storage and representation models and organization schemes, with ever-increasing complexity. This can range from simple flat file structures to more sophisticated indexed tables and relational views. Some common challenges include:

- Variety of sources: There are numerous legacy database management systems, many of which are particularly testy when it comes to exposing proper interfaces for direct data access (other than through programming layers).
- Difficulty in access: Even when applications employ standard commodity platforms (such as the usual suspects for relational database management systems), programmed applications often obfuscate or skew the underlying representations and complicate the ability to extract the right sets of records.
- Proprietary formats and models: Many business functions within vertical industries are easily abstracted, which has led to a reliance on externally-developed applications that are brought in-house, sometimes customized, and then deployed. These applications employ data models and data formats provided by the application vendor. Fundamental differences between the vendor's approaches to modeling and those used internally will complicate direct data access.
- Hosted solutions: An added complication to the use of vendor solutions occurs when the actual application platform is hosted outside corporate administrative boundaries. Accessing that data not only requires the capability to connect and extract from the underlying system, it also involves permissions, access rights, and network connectivity to gain access through the host's firewall.

The last issue, to some extent, goes beyond the scope of what can be distinctly automated, since it involves "negotiations" between the application host and your organization. The resulting bullet items are suggestions for how to broaden data accessibility:

- Expanding the realm of connectors to the widest variety of source systems
- Providing a semantic layering capability to remediate variances in the underlying representations, models, and data formats so that they "look" right to the application doing the extraction.

Maintaining Consistency

One of the thorniest issues associated with information delivery is maintaining the different aspects of consistency when accumulating data sets from different sources into a target for analysis. As an example, there may be different sources of data about prospective customers that is accumulated from internal marketing sources and from various externally-acquired sources. The same attributes associated with the prospect may exist in both sources, but which ones are more current and up-to-date? A change of address in one system may not have been captured in the other system, but that knowledge may be hidden from the information delivery and integration processes.

In other words, because information about the same entity concepts (such as customers, vendors, products, etc.) is scattered across different sources, and because most environments still want to ensure transaction consistency for all business applications touching shared data, there is a requirement for any information availability solution to ensure consistency across the usage scenarios.

Consistency depends on defining the rules for determining the currency of data sources, the periodicity for data set synchronization, and the methods by which incremental updates are propagated to the target system. In many scenarios, once the original movement of a large data set is completed, there is then a need to propagate only the incremental changes, so continual reloads of the same massive data set will not be required. That suggests that the most significant technical requirement for ensuring consistency involves methods for synchronizing changes to the source data sets that have already been migrated to the target on a continuous basis.

In this chapter, we have considered the technical objectives for high-performance information delivery. Our next and final chapter looks at the types of technical components that should be utilized to support these technical objectives.

Chapter 4: Technical Componentry for High Performance Information Delivery

Technical Requirements for Big Data and Information Availability

We can conclude from the points addressed in the previous chapters that to support the emerging information availability requirements for big data, the corresponding technical objectives for high-performance information delivery must be driven by three key goals: reducing latency, broadening data accessibility, and maintaining data consistency. In this final chapter, we examine the technologies that will help achieve those objectives by reducing or smoothing out data latency, enabling access to many data sources, and generally providing a seamless capability for provisioning large data sets to the right targets within a reasonable time frame. In particular, we'll consider four distinct methods:

- Replication: using a time-tested technique in new ways to enable consistent data sharing;
- Change Data Capture: a technique that is used to continuously update replicas and maintain consistency among the different copies across the enterprise;

- Connectivity: providing services ensuring broad access to many different sources; and
- Data federation and data virtualization: building on data access, replication, and virtual caching coupled with a data service access layer that enables a unified view to similar data concepts maintained in different sources.

Interestingly, none of these are revolutionary or new ideas or techniques. While federation and virtualization are still relatively recent innovations, they do build on a foundation of what used to be called enterprise information integration (EII) and enterprise application integration (EAI). Connectors for broad-based data access have been around for years, and most interestingly, replication and change data capture (CDC) are two techniques with a long pedigree whose benefits are rapidly being adapted to satisfy information delivery needs for big data.

Replication and Change Data Capture

At the most basic level, data replication is a technique for copying data from one location to another location. In the early days of replication, one of the market drivers was the need for performance improvement for database transactions occurring at physically disparate locations. For example, a business with global offices in the United States, Europe, Asia, and Australia would want to enable 24/7 application operation. However, the network speed for enabling real-time transactions deteriorates with distance. The proposed solution was to mirror the application and its underlying data subsystems in numerous locations.

The idea is that shortening the network distances reduces the latency of data movement, thereby allowing transactions to execute within the expected level of service. Mirroring the application is relatively straightforward by copying the program to the target machines. The challenge is the question of data consistency: first, to deploy each copy of the application, you need an up-to-date copy of the database, and second, once the replicas operate independently, the underlying databases will rapidly become unsynchronized and then inconsistent, creating a risk of incorrect executions.

That is where replication and change data capture come in. Replication provides the bulk copying of the underlying database. Replication techniques are designed to extract, move, and load data quickly in a way that preserves the data structure. Today, replication engines not only provide easy-to-use interfaces for describing sources and targets, but they also take advantage of additional optimizations ranging from data compression to utilization of parallel I/O channels for copying large-volume data sets.

And the incremental changes? That is where change data capture (CDC) comes in. CDC is a technique that is complementary to data replication in that it monitors changes to the database and communicates those changes to the other replicas. In best-of-breed solutions, this is done by observing database or system logs, seeing where updates are made, and batching and sending out the updates to the replicated systems.

Because both of these techniques are easily scalable, big data environments can be the beneficiaries of data replication and CDC technologies. As we have discussed, the most significant bottleneck for big data applications is moving the data into the environment. Replication provides simple, yet rapid methods for initial loads (thereby addressing our first technical objective of reducing data latency), while CDC ensures that the big data environment's view remains consistent with the source data systems (addressing our third technical objective — maintaining consistency). The conclusion is that by virtue of their ability to rapidly load and continuously update the data in an analytical environment, data replication and change data capture can be critical technical components for any big data environment.

Connectivity

Data latency issues can be remediated using replication and change data capture, addressing the need to speed information delivery to the analytical platform, especially in the world of big data. However, we still need to address a different facet of the process: broadening access to a variety of data sources that are to feed our large-scale analytics. This is a particularly thorny issue when considering the use of data embedded in older legacy systems whose architectures pre-date the development of relational database systems. In addition, the growing demand for streamed data or raw data sets stored in files implies an extra requirement for absorbing different types of source data.

Again, fortunately, there are proven methods for standardized accessibility to data. First, the Open Database Connectivity (ODBC) standard API (for languages like C and C++) opens access to many existing database management systems (DBMS), while a corresponding version of Java (called JDBC, or Java Database Connectivity) provides similar capabilities for Java programs. There are more standard data access methods and standards as well, such as Microsoft's OLE DB and ADO.NET, and some of the best-in-class solutions include proprietary optimizations to make targeted connectivity as seamless as possible.

Tools can be designed using standard drivers and interfaces to go beyond accessing data in existing database management systems, since drivers can be architected to use these standard APIs for accessing non-DBMS data. For example, it is straightforward to design a JDBC driver to access rows of raw data in a flat file. Specialized file structures can also be accommodated, and this means that a good software developer/vendor will incorporate many of these drivers in their tool suites to ensure the broadest coverage. In fact, drivers can be engineered to support incremental delivery of data instances flowing through data streams, enabling a "channeling" of streamed data using the same type of access methods.

When evaluating data access tools in preparation for a big data initiative, universality of source asset coverage is important, but don't forget to keep performance and ease-of-use criteria in mind, such as speed of access, the ability to take advantage of multiple I/O channels, configurability, and maintainability.

Federation and Virtualization

Replication and CDC address the speed of loading and synchronizing data, while enhanced methods of connectivity provide breadth of data access. Together, these technologies address the "accessibility endpoints" of the high-performance information availability framework necessary to support big data analytics. However, there is one remaining technical challenge to overcome that becomes especially acute in a big data environment. When there is a continuous need to share information among the different analytical platforms, speeding delivery of consistent information to the end users can be problematic.

Business analysts seeking to answer an iterative set of questions will probably need to have access to analytical results, source data sets, and profiles stored in a data warehouse at the same time. However, with different latencies associated with different systems, there are bound to be consistency and synchronization issues. This reflects yet another of the technical objectives for information delivery — maintaining consistency.

Data federation and data virtualization are two maturing technical approaches to smoothing out access times as well as standardizing access through a common, virtual data representation. Federation allows you to create a view representative of a shared concept (such as "customer" or "product") and use a single query mechanism to access multiple views of those concepts in different sources. At the same time, it can deliver a unified view of the query results. In other words, federation enables a homogenous view of data in heterogeneous platforms, all transparent to the user.

Federation and virtualization can also contribute to performance improvements for data access latency by introducing levels of software caching. Once the source data has been accessed, it can be cached locally. As long as the source has not changed, the data can be reused without going back through the network. The benefit of exploiting this data locality is both reduced network traffic and faster data access. And these techniques can be blended as well. For example, the cached copies of accessed data can be incrementally updated using change data capture to trickle-feed modifications.

Considerations

The last chapter reviewed a set of technical approaches that, when used together, address the technical objectives and performance demands for information availability. Replication and change data capture are used to significantly reduce data latency, and standards-based connectivity methods can be incorporated into a tool suite for expanding access to a wide variety of sources. Federation/virtualization improves speed while making access to heterogeneous sources transparent to the user. When used together, these techniques can alleviate information availability bottlenecks and provide a sound information delivery architecture to support the demands of the new wave of big data applications.

That being said, when seeking to launch a big data analytics project, be sure to consider the information availability characteristics and ensure that a scalable and flexible framework is in place to properly satisfy the information accessibility and availability demands:

- Assess the breadth and volumes of the data to be absorbed into the analytical framework and scale the hardware and software technologies for information intake, preparation, loading, and throughput;
- Make sure that there are the proper connectivity and access methods for all of the source data sets;
- Use data replication technology for massive bulk data transfers to rapidly load data into your big data platform;
- Maintain consistent updates through the use of change data capture tools; and
- Standardize access through the use of data federation and speed access using virtualization.

Look for vendors with experience in these markets as well as best-of-breed technologies to ensure the greatest degree of interoperability. Incorporating an information availability plan into your big data strategy is a key success factor for high-performance, large-scale data analytics for the future.

About the Author

David Loshin, President
Knowledge Integrity, Inc.

David Loshin, President of Knowledge Integrity, Inc., is a recognized thought leader and expert consultant on data management. Mr. Loshin specializes in the areas of data quality, master data management and business intelligence (BI). The author of numerous books, including “The Practitioner’s Guide to Data Quality Improvement” (2010), “Master Data Management” (2008), and “Business Intelligence - The Savvy Manager’s Guide” (2003), Mr. Loshin is also the creator of many courses, tutorials and web seminars on data management best practices.

About Attunity

Attunity is a leading provider of information availability solutions that enable access, sharing and distribution of data , including Big Data, across heterogeneous enterprise platforms, organizations, and the cloud. Our software solutions include data replication, change data capture (CDC), data connectivity, enterprise file replication (EFR), managed-file-transfer (MFT), and cloud data delivery. Using Attunity’s software solutions, our customers enjoy significant business benefits by enabling real-time access and availability of data and files where and when needed, across the maze of heterogeneous systems making up today’s IT environment.

Attunity has supplied innovative software solutions to its enterprise-class customers for nearly 20 years and has successful deployments at thousands of organizations worldwide. Attunity provides software directly and indirectly through a number of partners such as Microsoft, Oracle, IBM and HP. Headquartered in Boston, Attunity serves its customers via offices in North America, Europe, and Asia Pacific and through a network of local partners. For more information, visit http://www.attunity.com or our In Tune blog and join our community on Twitter, Facebook, LinkedIn and YouTube.

Over 2,000 companies — including half of the Fortune 500 — trust Attunity solutions.

Attunity exclusively focuses on engineering high-performance information availability solutions that are fast to deploy and easy to operate, empowering enterprises to simply and cost-effectively ensure business-critical information is accessible when, where and how it's needed to become a more agile, intelligent enterprise.

Contact Us

North America
T: +1 781 730-4070
866-288-8648
sales@attunity.com

Europe, Middle East & Africa
T: +44 (0) 1344 742 805
info-uk@attunity.com

Asia Pacific
T: + (852) 2756 9233
info-hk@attunity.com

Connect With Us

www.attunity.com

blog.attunity.com

ATTUNITY