Progress® DataDirect

ADDRESSING FIVE EMERGING CHALLENGES OF BIG DATA

David Loshin, President of Knowledge Integrity, Inc.

Table of Contents

Introduction - Big Data Challenges

Progress DataDirect includes ODBC, JDBC and ADO.NET connectors, providing broad market coverage and addressing many data connectivity challenges seen today. ISVs can easily embed DataDirect's connectivity technology for immediate and substantial benefits.

Big data technologies are maturing to a point in which more organizations are prepared to pilot and adopt big data as a core component of the information management and analytics infrastructure. Big data, as a compendium of emerging disruptive tools and technologies, is positioned as the next great step in enabling integrated analytics in many common business scenarios.

As big data wends its inextricable way into the enterprise, information technology (IT) practitioners and business sponsors alike will bump up against a number of challenges that must be addressed before any big data program can be successful. Five of those challenges are:

1. **Uncertainty of the Data Management Landscape** – There are many competing technologies, and within each technical area there are numerous rivals. Our first challenge is making the best choices while not introducing additional unknowns and risk to big data adoption.

2. **The Big Data Talent Gap** – The excitement around big data applications seems to imply that there is a broad community of experts available to help in implementation. However, this is not yet the case, and the talent gap poses our second challenge.

3. **Getting Data into the Big Data Platform** – The scale and variety of data to be absorbed into a big data environment can overwhelm the unprepared data practitioner, making data accessibility and integration our third challenge.

4. **Synchronization Across the Data Sources** – As more data sets from diverse sources are incorporated into an analytical platform, the potential for time lags to impact data currency and consistency becomes our fourth challenge.

5. **Getting Useful Information out of the Big Data Platform** – Lastly, using big data for different purposes ranging from storage augmentation to enabling high-performance analytics is impeded if the information cannot be adequately provisioned back within the other components of the enterprise information architecture, making big data syndication our fifth challenge.

In this paper, we examine these challenges and consider the requirements for tools to help address them. First, we discuss each of the challenges in greater detail, and then we look at understanding and quantifying the risks of not addressing these issues. Finally, we explore how a strategy for data integration can be crafted to manage those risks.

Challenge 1: Uncertainty of the Data Management Landscape

One disruptive facet of big data is the use of a variety of innovative data management frameworks whose designs are intended to support both operational and to a greater extent, analytical processing. These approaches are generally lumped into a category referred to as NoSQL (that is, "not only SQL") frameworks that are differentiated from the conventional relational database management system paradigm in terms of storage model, data access methodology, and are largely designed to meet performance demands for big data applications (such as managing massive amounts of data and rapid response times).

The wide variety of NoSQL tools, developers and the status of the market are creating uncertainty within the data management landscape.

There are a number of different NoSQL approaches. Some employ the paradigm of a document store that maintains a hierarchical object representation (using standard encoding methods such as XML, JSON, or BSON) associated with each managed data object or entity. Others are based on the concept of a key-value store that allows applications to associate values associated with varying attributes (as named "keys") to be associated with each managed object in the data set, basically enabling a schema-less model. Graph databases maintain the interconnected relationships among different objects, simplifying social network analyses. And other paradigms are continuing to evolve.

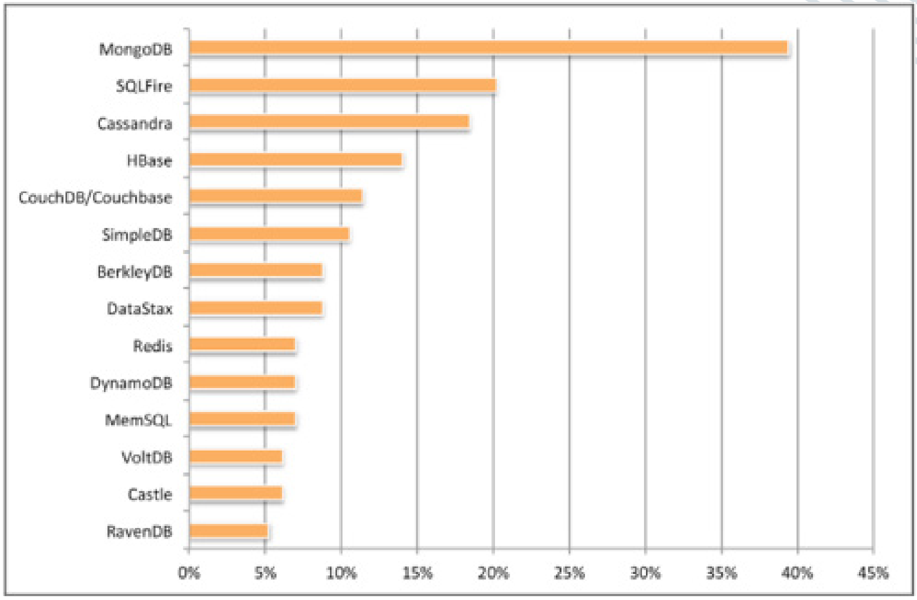

Figure 1
NoSQL and other innovative data management options are predicted to grow in 2014 and beyond.[1]

We are still in the relatively early stages of this evolution, with many competing approaches and companies. In fact, within each of these NoSQL categories, there are dozens of models being developed by a wide contingent of organizations, both commercial and non-commercial. Each approach is suited differently to key performance dimensions—some models provide great flexibility, others are eminently scalable in terms of performance while others support a wider range of functionality.

In other words, the wide variety of NoSQL tools and developers, and the status of the market lend a great degree of uncertainty to the data management landscape. Choosing a NoSQL tool can be difficult, but committing to the wrong core data management technology can prove to be a costly error if the selected vendor's tool does not live up to expectations, the vendor company fails, or if third-party application development tends to adopt different data management schemes.

For any organization seeking to institute big data, the challenge is to propose a means for your organization to select NoSQL alternatives while mitigating the technology risk.

[1] "2014 Data Connectivity Outlook paper"; Progress, January 2014

Challenge 2: The Big Data Talent Gap

It is difficult to peruse the analyst and high-tech media without being bombarded with content touting the value of big data analytics and corresponding reliance on a wide variety of disruptive technologies. These new tools range from traditional relational database tools with alternative data layouts designed to increased access speed while decreasing the storage footprint, in-memory analytics, NoSQL data management frameworks, as well as the broad Hadoop ecosystem.

The big data talent gap is real. Consider this statistic: "By 2018, the US alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."

There is a growing community of application developers who are increasing their knowledge of tools like those comprising the Hadoop ecosystem. That being said, despite the promotion of these big data technologies, the reality is that there is not a wealth of skills in the market. The typical expert, though, has gained experience through tool implementation and its use as a programming model, rather than the data management aspects. That suggests that many big data tools experts remain somewhat naïve when it comes to the practical aspects of data modeling, data architecture, and data integration. And in turn, this can lead to less-then-successful implementations whose performance is negatively impacted by issues related to data accessibility.

And the talent gap is real—consider these statistics: According to analyst firm McKinsey & Company, "By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."[2] And in a report from 2012, "Gartner analysts predicted that by 2015, 4.4 million IT jobs globally will be created to support big data with 1.9 million of those jobs in the United States. ... However, while the jobs will be created, there is no assurance that there will be employees to fill those positions."[3]

There is no doubt that as more data practitioners become engaged, the talent gap will eventually close. But when developers are not adept at addressing these fundamental data architecture and data management challenges, the ability to achieve and maintain a competitive edge through technology adoption will be severely impaired. In essence, for an organization seeking to deploy a big data framework, the challenge lies in ensuring a level of the usability for the big data ecosystem as the proper expertise is brought on board.

[2] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, "Big data: The next frontier for innovation, competition, and productivity," May 2011, viewed March 10, 2014

[3] Eric Lundquist, "Gartner: 2013 Tech Spending To Hit $3.7 Trillion" October 23, 2012, viewed March 10, 2014

Challenge 3: Getting Data into the Big Data Platform

It might seem obvious that the intent of a big data program involves processing or analyzing massive amounts of data. Yet while many people have raised expectations regarding analyzing massive data sets sitting in a big data platform, they may not be aware of the complexity of facilitating the access, transmission, and delivery of data from the numerous sources and then loading those various data sets into the big data platform.

The impulse toward establishing the ability to manage and analyze data sets of potentially gargantuan size can overshadow the practical steps needed to seamlessly provision data to the big data environment. The intricate aspects of data access, movement, and loading are only part of the challenge. The need to navigate extraction and transformation is not limited to structured conventional relational data sets. Analysts increasingly want to import older mainframe data sets (in VSAM files or IMS structures, for example) and at the same time want to absorb meaningful representations of objects and concepts refined out of different types of unstructured data sources such as emails, texts, tweets, images, graphics, audio files, and videos, all accompanied by their corresponding metadata.

Many people may not be aware of the complexity of facilitating the access, transmission, and delivery of data from the numerous sources and then loading those data sets into the big data platform.

An additional challenge is navigating the response time expectations for loading data into the platform. Trying to squeeze massive data volumes through "data pipes" of limited bandwidth will both degrade performance and may even impact data currency. This actually implies two challenges for any organization starting a big data program. The first involves both cataloging the numerous data source types expected to be incorporated into the analytical framework and ensuring that there are methods for universal data accessibility, while the second is to understand the performance expectations and ensure that the tools and infrastructure can handle the volume transfers in a timely manner.

Challenge 4: Synchronization Across the Data Sources

Once you have figured out how to get data into the big data platform, you begin to realize that data copies migrated from different sources on different schedules and at different rates can rapidly get out of synchronization with the originating systems. There are different aspects of synchrony. From a data currency perspective, synchrony implies that the data coming from one source is not out of date with data coming from another source. From a semantics perspective, synchronization implies commonality of data concepts, definitions, metadata, and the like.

The inability to ensure synchrony for big data poses the risk of analyses that use inconsistent or potentially even invalid information.

With conventional data marts and data warehouses, sequences of data extractions, transformations, and migrations all provide situations in which there is a risk for information to become unsynchronized. But as the data volumes explode and the speed at which updates are expected to be made, ensuring the level of governance typically applied for conventional data management environments becomes much more difficult.

The inability to ensure synchrony for big data poses the risk of analyses that use inconsistent or potentially even invalid information. If inconsistent data in a conventional data warehouse poses a risk of forwarding faulty analytical results to downstream information consumers, allowing more rampant inconsistencies and asynchrony in a big data environment can have a much more disastrous effect.

Challenge 5: Getting Useful Information of the Big Data Platform

Most of the most practical uses cases for big data involve data availability: augmenting existing data storage as well as providing access to end users employing business intelligence tools for the purpose of data discovery. These BI tools not only must be able to connect to one or more big data platforms, they must provide transparency to the data consumers to reduce or eliminate the need for custom coding. At the same time, as the number of data consumers grows, we can anticipate a need to support a rapidly expanding collection of many simultaneous user accesses. That demand may spike at different times of the day or in reaction to different aspects of business process cycles. Ensuring right-time data availability to the community of data consumers becomes a critical success factor.

This frames our fifth and final challenge: enabling a means of making data accessible to the different types of downstream applications in a way that is seamless and transparent to the consuming applications while elastically supporting demand.

BI tools must be able to connect to one or more big data platforms, provide transparency to the data consumers and reduce or eliminate the need for custom coding. At the same time, we anticipate the need to support an expanding collection of many simultaneous user accesses.

Considerations: What Risks Do These Challenges Really Pose?

Considering the business impacts of these challenges suggests some serious risks to successfully deploying a big data program. In Table 1, we reflect on the impacts of our challenges and corresponding risks to success.

Challenge	Impact	Risk
Uncertainty of the market landscape	Difficulty in choosing technology components Vendor lock-in	Committing to failing product or failing vendor
Big data talent gap	Steep learning curve. Extended time for design, development, and implementation.	Delayed time to value
Big data loading	Increased cycle time for analytical platform data population	Inability to actualize the program due to unmanageable data latencies
Synchronization	Data that is inconsistent or out of date	Flawed decisions based on flawed data
Big data accessibility	Increased complexity in syndicating data to end-user discovery tools	Inability to appropriately satisfy the growing community of data consumers

Table 1: Risks associated with our five big data challenges

Conclusion: Addressing the Challenge with Big Data Integration Strategy

In retrospect, all of our challenges reflect different facets of a more fundamental issue: the absence of a strategy for integrating big data into the enterprise environment. For example, the difficulty of the technology selection is related to the need to ensure the sustainability of the platform that leverages existing enterprise resources while enabling a path for evolving increased data management performance. Likewise for the talent gap and the associated struggle to find people with experience in production-quality big data analytics applications that support existing business processes.

In turn, the last three challenges more acutely demonstrate the big data integration risks. Migrating data sets from a wide variety of sources into the big data platform, ensuring process synchronization and data coherence, and exposing a virtual interface allowing uniform access to the information residing on the big data platform are all examples of capabilities typically provided by data integration tools and techniques.

Examining the challenges and risks suggests that devising a strategy and program plan for big data integration can help mitigate those risks. The strategy must accommodate all aspects of data availability and accessibility while reducing the need for knowledge of dozens of specialized data management schemes. This strategy must focus on providing a universal means for accessing data from multiple sources, moving that data to multiple targets, and instituting a level of abstraction for layering general data accessibility.

This can be achieved through the development of informed processes that take advantage of best-of-breed data integration technologies that address your evolving challenges and eliminate the correlated risks. Some key characteristics of these technologies include:

- Accessing data stored in a variety of standard configurations (including XML, JSON, and BSON objects)
- Relying on standard relational data access methods (such as ODBC/JDBC)
- Enabling canonical means for virtualizing data accesses to consumer applications
- Employ push-down capabilities of a wide variety of data management systems (ranging from conventional RDBMS data stores to newer NoSQL approaches) to optimize data access
- Rapid application of data transformations as data sets are migrated from sources to the big data target platforms

To mitigate these challenges, a comprehensive strategy and program plan for big data integration is required.

Efficiently managing today's big data challenges requires a robust data integration strategy backed by leading-edge data technologies and services that lets you easily connect to and access your data, wherever it resides.

About the Author

David Loshin, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of analytics, big data, data governance, data quality, master data management, and business intelligence. Along with consulting on numerous data management projects over the past 15 years, David is also a prolific author regarding business intelligence best practices, as the author of numerous books and papers on data management, including the recently published "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph," the second edition of "Business Intelligence – The Savvy Manager's Guide," as well as other books and articles on data quality, master data management, and data governance. David is a frequent invited speaker at conferences, web seminars, and sponsored web sites and channels including www.b-eye-network.com, and shares additional content at his notes and articles at www.dataqualitybook.com. David can be reached at loshin@knowledge-integrity.com, or at (301) 754-6350.

For more information or to talk to a Progress DataDirect representative, visit:
www.progress.com/datadirect-connectors.

To speak with a Progress DataDirect representative, call:
1-800-876-3101

About Progress

Progress (NASDAQ: PRGS) is a global leader in application development, empowering the digital transformation organizations need to create and sustain engaging user experiences in today's evolving marketplace. With offerings spanning web, mobile and data for on-premise and cloud environments, Progress powers startups and industry titans worldwide, promoting success one customer at a time. Learn about Progress at www.progress.com or 1-781-280-4000.

Rev 2016/11

Progress®