PROGRESS® DataDirect®

WHITE PAPER

# TRANSITIONING TO BIG DATA:

## A Checklist for Operational Readiness

Moving to a Big Data platform:
Key recommendations to ensure operational readiness

# Overview

Many factors can drive the decision to augment existing IT environments with the capabilities of a modern system platform. Big Data platforms offer an appealing option with their ability to provide elasticity and scalability for larger data volumes at presumably lower cost. These platforms are able to offer this kind of flexibility through the use of technologies like scalable high-speed storage, high-performance networks, ecosystems like Hadoop, and NoSQL data management systems (among others). Unfortunately, the decision to introduce new technologies is often made with little more than marketing hype and gut feelings to drive the choice.

**Forrester defines Big Data as: The practices, processes, and technologies that close the gap between the data available and the ability to turn it into business insights.**[1]

When considering whether to adopt Big Data, decision makers within the organization must specify clear performance objectives and metrics to ensure that the desired systemic and environmental benefits would be derived from the transition. At the same time, those decision makers need to assess the potential risks and barriers to a successful deployment and understand how to mitigate those risks.

**Any transition to Big Data should not be done in a vacuum.**

This means that any transition to Big Data should not be done in a vacuum. It is important to formulate the organization's strategic direction for its future technology vision as well as legacy renovation, then align the technology choices with that direction, and develop an evolutionary program for migrating to Big Data at the point in time when it makes sense. In this paper, we'll discuss the key concepts to consider prior to making a decision to move to a big data platform. These recommendations will serve as your checklist for operational readiness to help you and your stakeholders make a responsible decision to renovate.

[1] Forrester Big Data Market Overview; Dec 2014; p2

# Getting Started: Identify the Key Challenges

The use of most, if not all, Big Data technologies presume immediate access to numerous different data sets to facilitate the types of querying and analytics that show the most promise. However, there are a number of hurdles associated with data access and movement that need to be overcome, such as:

- **Data distribution** – Accessing data sources that are managed locally within the organization as well as data sets that are managed by host or cloud providers may be a requirement. It may be challenging to ensure continuous and immediate access to that data so that it can be integrated into the Big Data platform's storage environment.

- **Volume** – Enabling the end users' needs for analysis and data accessibility in a real-time manner to support rapid decision-making requires the ability to quickly access large volumes of data held in Hadoop file systems.

- **Diversity** – Streamlining the manner for accessing data in different types of environments and encompassing any newer selected technologies in a way that is standardized for predictable performance.

- **Performance** – Delivering performance, the overriding value proposition for these new technologies, means ensuring that data accessibility meets predefined levels of service for on-demand, real-time responses.

It is important to recognize that these (and many other) high-level challenges can be addressed through a process of preparation for operational readiness. This process, described in the following checklist of best practices, helps assess the current and evolving organizational needs for transitioning to Big Data and provides a model that supports the decision process and the planning for the transition.

**Many high-level challenges can be addressed through a process of preparation for operational readiness.**

# Operational Readiness Checklist

## 1. Understand the Business Process the Application Performs

Imposing the structure of an enterprise architecture on top of applications is a relatively recent phenomenon. In most organizations, applications are designed and developed to meet specific transactional or operational needs. But once an application is moved into production, technologists have a tendency to dissociate the implementation of the application from its original intent. In some cases, the perception is that the specific implementation of the application becomes fused with the business process it was intended to support.

As the application ages, concern over its presumed archaic implementation may drive the desire to either move the existing application to a new platform or to redesign and redeploy the application. However, because of the perceived coalescence of the business process with a legacy application's implementation, an attempt to re-implement using new technology often ends up focusing on how the application in its current incarnation can be migrated, but not whether that approach is best suited to addressing the current, and more importantly, the future business needs.

The evolution of the enterprise technology framework provides an excellent opportunity to reengineer the implementation of applications to meet current and future business process needs. Considering a decision to adopt Big Data technologies provides the opportunity to reassess how well current application implementations support the business users. It also helps identify where the benefits of faster computation, greater data storage volume, and reduced capital costs can provide dramatic new capabilities to enable processes that are currently hampered by various performance constraints.

It's important to distinguish between the process and the application – understand the current requirements for the business process and the future expectations for growth and functional expansion. Examine the ways the various Big Data techniques better address those needs and how they enable new capabilities. Not only will this provide a means for evaluating the suitability of each choice, it will also help you establish a viable business justification for migration instead of blindly moving the existing implementation.

**It's important to distinguish between the process and the application – understand the current requirements for the business process and the future expectations for growth and functional expansion.**

## 2. Determine the Sources of Your Data

Understanding the business process provides a foundation for assessing the requirements for data volume and processing scalability. Recognizing that a key aspect of Big Data technologies is the ability to manage and process large data volumes, the next recommendation in our checklist requires conducting an inventory of data sources and their corresponding characteristics, including scale, aspects of data

accessibility and availability. It also means determining the alternative sources that must be acquired and incorporated to address the analytics needs to enhance the decision-making capability of the application.

As an example, a customer analytics application might blend sales transaction histories, customer service interactions, social media comments, and analytical customer profiles with demographic data acquired from external sources. This might require combining data pulled from an assortment of internal data warehouses, various streams of unstructured comment data from a variety of "fire hose" (that is, Twitter or Facebook aggregator) sites, as well as different sets of demographic data from various aggregators.

**[An effective]... strategy should access all available data and prioritize its usage to allow the organization to become more agile while operationalizing big data to inform business decisions in real time.[2]**

Delineating the data sources and the requirements for optimal data accessibility will help identify the technical perspectives for assessing platform characteristics, such as:

- Which data sets are required
- Where those data sets are persisted
- How those data sets need to be accessed
- How large the data sets are
- Which technologies are needed to facilitate data accessibility and availability

**You need to conduct an inventory of data sources and their corresponding characteristics, determine alternative sources that must be acquired and incorporated, and decide which technologies will facilitate optimal data accessibility and availability.**

## 3. Clearly Articulate Performance Criteria

Big Data is all about improving performance, whether it is faster computing, greater data volumes, increased analyst load, or a variety of other performance dimensions. But there are many other facets of performance that have different relevance in different contexts. For example, one organization might want to employ new technologies to increase the volume sizes for capturing and managing data sets, another might want to reduce the time it takes for certain analyses to be performed, while a third might want to increase the number of real-time streams being monitored for certain keywords.

In each of these scenarios, it is important to optimize a different facet of performance. The following list highlights several key dimensions of performance that should be considered as motivators for transitioning to Big Data technologies:

- **Data storage volume** – Enabling increasing volumes of data to be stored
- **File copy speed** – Decreasing the time to move data from one location to another
- **Data stream load** – Increasing the number of simultaneous data streams that can be monitored

[2] Teradata Magazine Blog, by David Kelly (guest blogger), Feb 2015

- **Data stream speed** – Increasing the speed at which a fixed number of data feeds are monitored
- **Interactive load** – Increasing the number of simultaneous users with access to data
- **Batch volume** – Increasing the amount of data that can be processed
- **Batch processing time** – Decreasing the time it takes to execute batch processes
- **Query response time** – Reducing the response time for interactive queries

These are just a sampling of dimensions for performance, but one thing is clear: when considering a transition to a new technology, first clarify which dimensions of performance are important, and then prioritize and set levels of acceptability to be met by any proposed future system.

**When considering a transition to a new technology, it is important to define, prioritize and set levels of acceptable performance to be met by a proposed future system.**

## 4. Project the Need for Scalability

The concept of performance is tightly coupled with the concept of scalability, or the ability to maintain relative performance by adding computing or storage resources as the sizes of the problems increase. The performance scales linearly if the ratio of resource to the performance remains the same (e.g., doubling the amount of resources corresponds to doubling the performance).

When considering performance, it is also important to review how the prioritized measures of performance relate to the need for scalability. For example, if the critical dimension of performance is increased processing speed for CPU-bound calculations, you need to address computational scalability, and correspondingly, the cost factors associated with increasing the number of computational units. The same is true for the other dimensions of performance, such as facilitating greater data volumes or supporting an increasing number of users.

You also need to project the growth along each of the selected dimensions and document how that growth will impact system sizing. Doing so will enable you to specify the key architectural components that need to possess linear scalability, which will help in forecasting the costs associated with continued growth.

**When considering performance, it is important to review how the prioritized measures of performance relate to the need for scalability and handling greater data volumes.**

## 5. Develop a Cost Model

There is a presumption that open source tools layered on top of commodity hardware is always going to be more cost-effective than proprietary systems. Realize, though, that when acquiring proprietary technology, a large part of the effort has already been subsumed by the system provider. When deploying open source tools, the costs for technology acquisition may be lower but there will be costs for customization, integration, and maintenance that must be undertaken.

Exercise fiduciary responsibility and develop a systemic cost model that takes all aspects of system management into account. Those aspects include:

- Hardware costs such as processing units, storage capacity, network infrastructure, and physical enclosures
- Facilities costs such as floor space, power, and cooling
- Development costs for customization, integration, and deployment
- Operations & Maintenance costs, such as system monitoring, continued management, and maintenance charges associated with the underlying platform

Developing a cost model for adoption of new technology provides an "apples to apples" comparison between the new systems and any existing platform investments, and establishes a foundation for a cost/performance business justification for replatforming.

## 6. Assess the Needs for Data Integration

**It is critical to exercise fiduciary responsibility and develop a systemic cost model that takes all aspects of system management into account.**

The desire for performance and scalability in the presence of expanding data volumes, broader data distributions, and more diverse types of inputs implies increased demands for data acquisition, movement, ingestion, integration, and processing. (See Figure 1 for detailed big data integration framework.) And despite the transition to a Big Data platform for analytics and reporting, it may take some time before much else of the existing data management infrastructure changes.

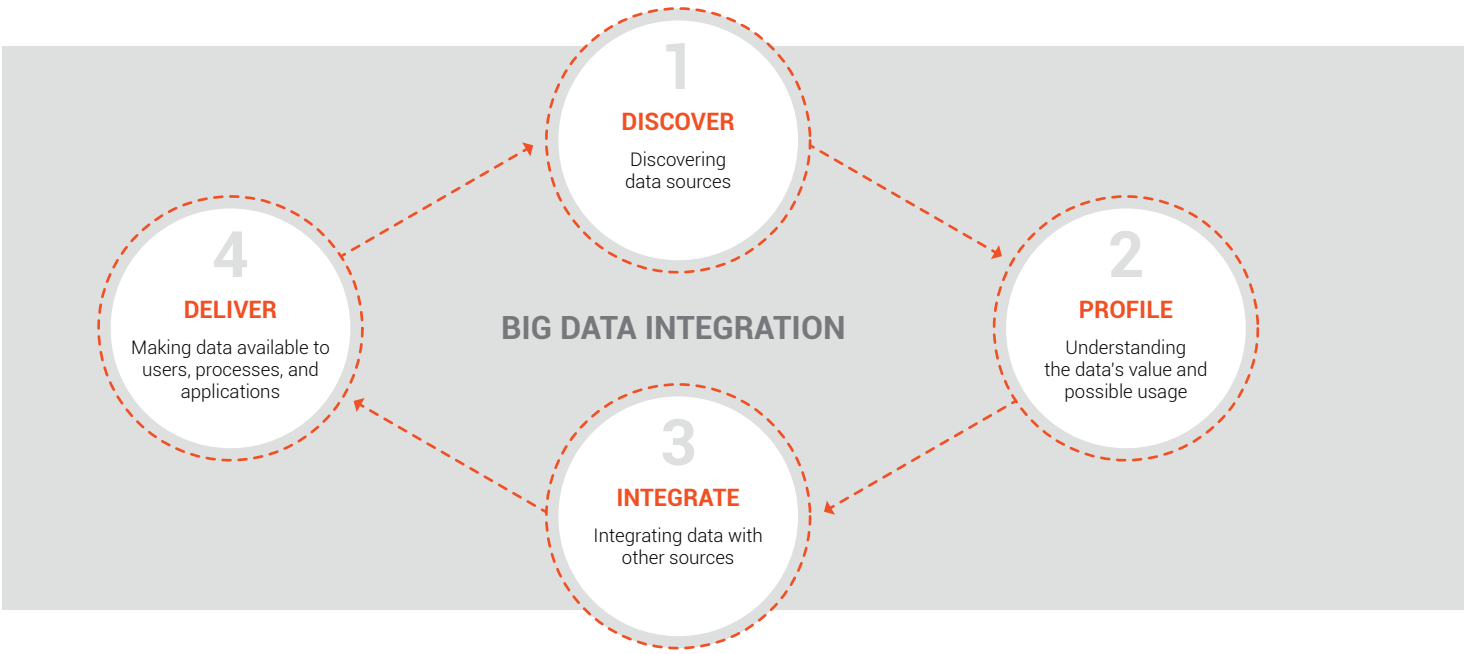


*Figure 1 - The Big Data Integration Framework*

In other words, as each selected application is implemented on the new platform, it will still need to be connected to a number of other existing platforms so that the required data can be accessed. Not only that, organizations whose systems blend internally managed data systems (including transaction processing systems, operational data stores, data warehouses, and data marts) as well as those with data managed in proprietary hosted systems or cloud-based applications will find a need to formulate a plan for ensuring that any Big Data application can access the data it needs in real time.

**As each selected application is implemented on the new platform, it will still need to be connected to a number of other existing platforms so that the required data can be accessed and accessed in real time.**

Perhaps just as important, your existing applications will probably be informed using the results of Big Data analytics. That implies a corresponding need for standardized accessibility to the data sitting in a Big Data platform, be it a NoSQL database or a Hadoop platform. Assess your needs for data integration, determine where potential gaps are, and devise a plan for embracing supporting tools for data accessibility and integration.

**Key recommendation from Forrester: Leverage big data integration for small to medium-size projects before tackling bigger ones, establishing the process, approach, and architecture.**[3]

## 7. Develop an Enterprise Integration Roadmap

No matter how agile your environment is, it is unlikely that there will be an overnight switchover from an established set of platforms to new technologies. That means that much of the existing environment will remain in production over an extended period of time.

You need to create a transition strategy that incrementally replaces heritage components with the appropriate new technologies. Clarify what the acceptance criteria are for newly developed and deployed Big Data applications, and assemble a plan for integrating new technologies into the existing enterprise. An enterprise integration roadmap should specify how new technologies are introduced in ways that do not disrupt ongoing operations yet provide the means for dual operation while acceptance testing goes on.

**Create a transition strategy that incrementally replaces heritage components with the appropriate new technologies.**

## 8. Assemble a Plan for Training

One of the biggest challenges in transitioning to Big Data is the skills gap. There are many data practitioners who are well-versed in the knowledge of mainframe data files or relational databases, but the complexities of many Big Data tools make them somewhat opaque. At the same time, those technologists with a firm grasp of the Hadoop world are more likely to have a programming background than to have a foundational understanding of database management, let alone the fundamentals of metadata necessary for an enterprise information architecture.

[3] Forrester Market Overview: Big Data Integration by Noel Yuhanna, December 5, 2014

One can finesse this roadblock by creating a bridge for employees for transitioning from legacy systems (COBOL, mainframes) to a new world of NoSQL, graph databases, in-memory computing, Hadoop, Hive, YARN, Impala, Spark, Tez, along with many other new and emerging Big Data technologies. Put together a training plan for staff members, and align those training sessions with the adoption plan so that they will be proficient in the technologies as they are embraced within the enterprise.

**Put together a training plan for employees, and align those training sessions with the adoption plan so that they will be proficient in the technologies as they are embraced within the enterprise.**

# Summary: Devise and Socialize the Big Data Adoption Strategy

When you review our checklist, it is clear that the decision to adopt Big Data technologies such as Hadoop cannot be made without an understanding of the impacts to the enterprise and the need for a strategy for business process assessment and evolution. That strategy must consider how technologies are best suited for the existing and future business processes, must enable the decision makers to be informed about the decisions they make, and ensure that those decisions can be justified in terms of cost, effort, and management of risks.

Understand the real motivating factors for transitioning to Big Data, and clearly articulate the usability and performance objectives, such as:

- **Performance**, such as query performance, system scalability, data load speeds, data volume capacity, and the ability to manage mixed workloads
- **Capabilities**, including support for advanced analytics, simplicity in data accessibility, reporting, support for SQL, support for structured and unstructured data models, and high availability
- **Platform independence**, allowing for deployment on-premises, in the cloud, or a hybrid model
- **Costs**, such as costs by data volume, costs to scale, operating costs, and the costs associated with configuration, management, and ongoing maintenance
- **Strategy**, including objectives for legacy phase-out, training, and managing the learning curve for team members over the transition period

It is essential to recognize the criticality of data integration and accessibility as a key component to your Big Data vision. Assess the data integration challenges and consider the use of tools that can provide unified access method for BI/analytics applications to gain access to a broad spectrum of data sources, regardless of where those data sources might reside, whether that is in the cloud, within the enterprise, or even behind a firewall. At the same time, make sure that bi-directional data accessibility is supported. Institute methods for allowing heritage applications to access to NoSQL and Hadoop storage environments in ways that are scalable from a performance perspective yet provide high availability.

In summary, remember: Making informed decisions about the adoption of Big Data within the context of an overall enterprise application renovation strategy will position those technology choices in the right way for successful incorporation and deployment.

For More Information

## Learn more about operational readiness and the role of data accessibility.

progress.com/solutions/data-connectivity

## OPERATIONAL READINESS CHECKLIST RECAP

1. Understand the Business Process the Application Performs
2. Determine the Sources of Your Data
3. Clearly Articulate Performance Criteria
4. Project the Need for Scalability
5. Develop a Cost Model
6. Assess the Needs for Data Integration
7. Develop an Enterprise Integration Roadmap
8. Assemble a Plan for Training

**PROGRESS**

Progress (NASDAQ: PRGS) is a global software company that simplifies the development, deployment and management of business applications on-premise or in the cloud, on any platform or device, to any data source, with enhanced performance, minimal IT complexity and low total cost of ownership.

**WORLDWIDE HEADQUARTERS**

Progress Software Corporation, 14 Oak Park, Bedford, MA 01730 USA Tel: +1 781 280-4000 Fax: +1 781 280-4095 On the Web at: www.progress.com
Find us on facebook.com/progresssw twitter.com/progresssw youtube.com/progresssw
For regional international office locations and contact information, please go to www.progress.com/worldwide

Progress and DataDirect are trademarks or registered trademarks of Progress Software Corporation or one of its affiliates or subsidiaries in the U.S. and other countries. Any other marks contained herein may be trademarks of their respective owners. Specifications subject to change without notice.
© 2015 Progress Software Corporation and/or its subsidiaries or affiliates. All rights reserved.
Rev 12/16 | 151130-0171