CITO Research

Advancing the craft of technology leadership

Buyer's Guide to Big Data Integration

SPONSORED BY

pentaho

A Hitachi Group Company

CONTENTS

Introduction

The arrival of new data types in amazing volumes, the phenomenon known as big data, continues to cause CIOs and business leaders to rethink their technology portfolios. Few companies build their own infrastructure. Most will buy it. But what should they buy? And how can they put the pieces together into a coherent whole?

CIOs and business leaders can speed progress by integrating the new world of big data with the existing world of business intelligence

The first challenge of big data is that it requires new technology. On the other hand, the arrival of big data has not rendered all other types of data and technology obsolete. Hadoop, NoSQL databases, analytic databases, and data warehouses live side by side. Analysts don't care where the data comes from: they will crunch it from any source.

The second challenge is data integration. How can the new technology for processing big data use all of the data and technology now in place? How can existing technology and data be improved by adding big data? How can new forms of analytics and applications use both the old and the new?

CITO Research believes that CIOs and business leaders can speed progress by focusing on the task of integrating the new world of big data with the existing world of business intelligence (BI). This buyer's guide describes how to think about purchasing technology for big data integration.

Challenges of Big Data Integration: New and Old

Aficionados of big data will be familiar with the ways that big data is different from previous generations of data. It's often described in terms of 3 V's—volume, variety, and velocity—a handy construct for thinking about big data introduced by Gartner analyst Doug Laney.

The challenge is to find a repository that can handle huge volumes of data. A related problem is analyzing streams of data that come from machines, servers, mobile devices, and sensors from the Internet of Things (IoT). The Hadoop ecosystem has emerged to handle the volume, velocity, and variety of data.

Another challenge is dealing with the fact that big data often requires new techniques for exploration and analysis. Big data is typically unstructured or semi-structured. In addition, raw text documents and video are often included as data types. Machine learning, text and video analytics, and many other techniques applied to the data in Hadoop, NoSQL databases, and analytics databases help messy data become meaningful.

Once you have met these challenges, the tasks related to using big data start to feel very much like those involved in processing existing data (see "Challenges Big Data and Existing Data Share").

Challenges Big Data and Existing Data Share

- Merging data from different sources
- Supporting exploration
- Creating a single reusable version of the truth
- Architecting blended data sets for more complete analytics
- Expanding the use of data
- Creating advanced analytical environments
- Supporting applications
- Controlling access to data
- Managing the analytics lifecycle
- Ensuring compliance

The equation for handling big data looks something like this:

(Repository for big data storage and processing) + **(New techniques for big data analysis)** + **(Existing BI)** = **Integrated big data environment**

While big data may change many things about the way BI is done, it will not make BI obsolete. This means that the right path to big data integration is likely to come through existing data integration solutions that have been adapted to incorporate big data.

In addition, there is a difference between performing proof of concepts and operationalizing big data. A big data integration technology should not only enable a science experiment but should also support the beginning, middle, and end of the journey toward making full use of big data in conjunction with existing applications and systems for BI.

From Hub and Spoke to Data Supply Chain

The blending of big data and existing BI brings about a large conceptual change. The data warehouse is no longer the center of the universe. Many special purpose repositories can support applications or new forms of analysis. In addition, data will increasingly come from outside the company through APIs. Instead of a hub and spoke paradigm with the data warehouse at the center, the data processing infrastructure more often resembles a distributed supply chain.

Big data is the primary driver of this new paradigm, and big data integration provides the plumbing to make it work. CIOs and business leaders who want to move fast to exploit big data and existing BI should focus on acquiring capabilities that form the backbone of a dynamic and responsive data supply chain.

What You Need for Big Data Integration

To make the right choice about assembling a system for big data integration, consider what you will need. Most organizations will need the following capabilities.

Connect, Transport, and Transform

Accessing, moving, and transforming data have been at the heart of several generations of data integration technology. Big data integration adds some new twists.

While many capabilities for accessing, moving, and transforming data exist in the current generation of integration technology, big data adds new requirements.

The ability to handle complex data onboarding at scale from many data sources is critical. Today, enterprises face the challenge of ingesting hundreds of data sources with many different formats and formats that change over time. Additional data sources are added regularly. Solutions should leverage templates and automation to reduce the manual work for creating jobs and transformations to onboard data into Hadoop. The data onboarding process should be flexible, regular, reliable, and automated as much as is feasible.

Access to data through Hadoop, NoSQL databases, and analytic databases must be supported, as well as connectivity to a variety of data formats, including JSON, XML, various log formats, emerging IoT standards, and so on. The ability to define or discover a schema is crucial.

Modern data integration technology must be **deployed in both cloud and on-premise.**

Synchronization of data between repositories is required as the data supply chain becomes more complex. The transport mechanisms of data integration technology need to be more sophisticated to handle the traffic. The insights from big data analysis must be delivered to applications to support more detailed, high-resolution models of reality.

The ability to transform big data is crucial. Tools should make designing and implementing transformations as easy as possible. Analysts need to blend and distill data from many sources to perform an analysis. Most of this work takes place in the data integration layer. Transformations must be reusable and sharable.

Additionally, it matters where and how data transformation is executed. **Transformation should be able to tap into the full power of Hadoop**. Data integration tools must be able to run natively on the Hadoop cluster to make use of its processing power and scalability. Transformation represents the bulk of the work involved in data integration, and with hundreds of data sources in play, transformation should take advantage of Hadoop's ability to perform processing in parallel across the cluster.

Because some jobs are more critical than others, it's important to have the ability to leverage YARN on Hadoop in order to efficiently allocate cluster resources to data integration jobs and transformations, with the ability to spin resources up and down as needed. Processing terabytes of big data on a daily basis will not yield the required performance unless the full power of modern Hadoop is leveraged.

Tools should also eliminate human productivity bottlenecks. Jobs should not have to be manually coded in Pig scripts or Java; data integration tools should allow analysts to visually design and run native MapReduce jobs without the need for coding.

Big data integration also requires processing **real time streams of data** from messaging systems, enterprise service buses, server log files, APIs, and other streaming data sources.

Integration and Canonical Forms

How does big data change things?

Here's what won't happen: All of your data and applications won't be based on big data and use technology for big data as their main repository. All of the data in BI and the data warehouses you've built won't become instantly useless.

Here's another thing that won't happen. All important business questions won't be answered by big data alone.

What does this mean? Simply that much of the time the right answer comes from blending big data with master data and transactional data stored in data warehouses.

In order to make the most of big data, it is vital to be able to combine it with existing data. This sort of data integration is crucial at all levels of analysis, from cleaning data to creating special purpose repositories to supporting advanced visualizations. **It is therefore vital that data integration technology combine both big data and existing forms of data**, most often stored in SQL repositories.

In other words, the key is to choose technology that speaks both the native language of big data sources like Hadoop, NoSQL databases, and analytic databases as well as traditional SQL. **Don't make big data a silo** by creating a separate infrastructure, team, and skill set.

To combine big data with existing data demands c**reating canonical forms of various kinds of information**. A customer master record that offers a 360-degree view has long been a goal of BI systems. In the era of big data, customer records can be supplemented with social media activity, mobile app data, website usage, and so on.

It is also important to manage canonical definitions of data in a lifecycle to create a shared understanding of data across the organization and so that changes to the standard forms of data can be controlled.

When evaluating big data integration technology, be sure that big data and existing data can be easily integrated and stored in canonical form.

Data Exploration

When companies make use of data, it is vital that everyone—analysts, end-users, developers, and anyone else who is interested—is able to explore the data and ask questions. This need for a hands-on way to examine and play with the data is required at all levels of the system.

It doesn't matter whether the data resides in a Hadoop cluster, a NoSQL database, a special purpose repository, an in-memory analytics environment, or an application. **The best results will come when anyone with a question can bang away and see if the data can answer it.**

Big data integration technology should support exploration at all levels of the data supply chain with automatic schema discovery and visualization.

For big data, this usually means that some sort of exploratory environment will be used in conjunction with the repositories, which typically only allow data access through writing programs or using complicated query mechanisms.

But when big data is combined with other data, exploration must also be supported.

One of the biggest challenges in creating exploratory environments that work in conjunction with big data is that much of the time the data is not structured into rows and tables. Each record may have many different parts. Several records may form a group that represents an object. The time each record was created could play a major role in the grouping.

Big data integration technology must support fast exploration of data with a flexible structure by creating schema on the fly that attempt to identify fields and patterns.

To support analytics needs, many organizations use a hybrid architecture in which an analytic database supports interactive visualizations while Hadoop is leveraged for extreme scale processing and refinement of diverse data.

The rationale for a hybrid architecture is that analytics solutions that run directly on Hadoop are still evolving and do not necessarily support the full breadth of production use cases. In particular, many SQL-like tools on Hadoop perform well in certain use cases, but don't always deliver the highly interactive analysis performance that the market is used to with relational data sources.

Taking a solution approach that combines the best of Hadoop (extreme scale processing and refinement of diverse data) with the best of analytic databases (speed of thought analysis on large volumes of relational data) often makes more sense.

In such a solution approach, it's important to be able to deliver data sets and analytics to the business in an on-demand fashion. This can be helped by automating data modeling processes and using parameterized data integration workflows that can adapt to the ever-changing business questions that analysts are asking. The goal is to create a process or framework once and avoid repeated requests that result in manual work and lengthen time to decision.

Analytics Support

It is well known among data analysts in any domain that as much as 80 percent of the work to get an answer or to create an analytical application is done up front to clean and prepare the data. Data integration technology has long been the workhorse of analysts who seek to accelerate the process of cleaning and massaging data.

In the realm of big data, this means that all of the capabilities mentioned so far must be present: easy to use mechanisms for defining transformations, the ability to capture and reuse transformations, the ability to create and manage canonical data stores, and the ability to execute queries. Of course, all of this needs to be present for big data repositories as well as those that combine all forms of data.

By supporting analysts in cleaning and distilling data using machine learning and sharing the results, the process of answering questions, building apps, and supporting visualizations is accelerated.

But analysts will face other problems unique to big data. As we pointed out earlier, big data is often dirty and noisy. Machine learning is needed to ferret out the signal. But machine learning techniques are often difficult to use.

The best big data integration technology will offer a guided experience in which machine learning suggests and then is moved in the right direction by analysts.

This guided approach is required because so many machine learning and advanced analytical techniques are available for many different types of data. The machine learning techniques used to create predictive models from streaming data are far different from those used for categorizing unstructured text.

Once an analyst has created a useful, clean data set, the value of the work can be amplified by allowing sharing and reuse. Right now, environments to support sharing and collaboration are emerging. Some environments support architected blending of big data at the source to enable easier use and optimal storage of big data. Big data integration technology should support such environments.

Preferred Technology Architecture

The ideal system for big data integration is different at every company. The most data intensive firms will likely need every capability mentioned. Most companies will need quite a few of them and more as time goes on.

The best way to provision the capabilities for big data integration is to acquire as few systems as possible that have the needed features. Most of the capabilities mentioned are stronger when they are built to work together.

The ideal big data integration technology should simplify complexity, be future proof through abstractions, and invite as many people and systems as possible to make use of data.

A fact of life in the world of data analysis is that everything is going to change. The best technology will insulate you as much as possible from changes. It should be the vendor's responsibility to create easy to use, powerful abstractions and maintain them going forward. The fact that big data technologies are evolving should not be your problem. Neither should the inevitable shakeout that will occur as various forms of technology and vendors fade away. Does this represent a form of lock-in? Of course, but in the end, it is better to be married to a higher level of abstraction than a lower level one.

Open source is and has been leading the way in big data innovation. A large part of the innovation in Hadoop and other big data ecosystem components has come via open source projects, not proprietary or closed approaches. Open source leads to a virtuous cycle of greater technology adoption and community-driven improvements. As such, it is key to look for data integration tools that embrace open source innovation and look to align with its capabilities. At the same time, open technologies tend to be more flexible and extensible than proprietary products. In an immature big data integration and analytics landscape where no one vendor can provide a complete out of box solution to meet all anticipated needs, support for the flexibility provided by open standards, open APIs, and well-developed SDKs is paramount.

By choosing technology that supports visual data modeling, it is possible to avoid a skill bottleneck. Programming knowledge should not be required for transforming, modeling, and blending data sources. Simplified environments allow more people to interact with data directly and in turn accelerate progress.

One key financial factor in choosing the right technology is the license model. Depending on how your software is deployed and the internal skill set for supporting software, there can be vast differences in the cost to acquire various capabilities. It is important to understand the benefits and drawbacks of traditional licenses, open source software licenses, and various hybrid offerings.

Select solutions based on real-world use cases, not hypothetical applications. Look for integration vendors that have helped customers achieve success specific to big data use cases, most importantly with Hadoop or NoSQL data. While the majority of vendors claim to work with big data, the reality is that many are new to the market or are older vendors that have had success with traditional use cases, but not big data use cases. Another thing to look for is deep services offerings and expertise. Solving major business problems with big data requires best practice architectures, proven project plans, hands-on training, and expert support.

Finally, the best systems for big data integration are built to be embedded into business processes and workflows. The simplified forms of transformation should be able to be pointed at big data sources or at SQL repositories and used inside MapReduce or applications. Data integration tools should enable transformed big data to be accessed through familiar BI tools and used to feed web pages, mobile apps, or enterprise applications.

The Rewards of Getting Big Data Integration Right

Data does no good unless it is presented to a human who can somehow benefit from it or unless it is used in an automated system that a human designed. The point of big data integration is to make it as easy as possible to access, understand, and make use of data.

The rewards of getting big data integration right are the benefits that come from greatly expanded and timely use of all available data. Reducing delays, eliminating skill bottlenecks, and getting fresh data to analysts and applications means that an organization can move faster and more effectively.

By purchasing components and systems that are part of a coherent vision while at the same time leveraging ongoing open source innovation, it is possible to minimize cost and avoid compromising on needed capabilities.

The questions we started with should now be easier to answer:

What to buy? As few systems as possible that provide the capabilities you need now and in the future in a way that is easy to use and future proof.

What is the coherent whole? A vision of big data integration that incorporates existing forms and sources of data into a new system that supports all phases of a responsive, dynamic data supply chain.

Solving Big Data Integration Challenges With Pentaho

Pentaho's big data integration and analytics platform provides broad connectivity to any type or source of data, with native support for Hadoop, NoSQL, and analytic databases. Pentaho's complete visual big data integration tools eliminate coding in SQL or writing MapReduce Java functions, and empowers you to architect big data blends at the source for more complete and accurate analytics. Learn more at www.pentaho.com.

This paper was created by CITO Research and sponsored by Pentaho

CITO Research

CITO Research is a source of news, analysis, research and knowledge for CIOs, CTOs and other IT and business professionals. CITO Research engages in a dialogue with its audience to capture technology trends that are harvested, analyzed and communicated in a sophisticated way to help practitioners solve difficult business problems.

Visit us at http://www.citoresearch.com