Buyer’s Guide to Data Analytics and Visualization
# Contents

Introduction: What Do We Need from Big Data Analytics?  
 Extracting Value from Big Data  
 Why Data Integration Is Key to Big Data Analytics  
 The Implications of Visualization for Big Data Analytics  
 Features of Successful Analytics Systems—The Power of Visualization  
 What’s Next?
Introduction: What Do We Need from Big Data Analytics?

Big data is on the minds of all CIOs and business leaders these days—and for good reason. Big data, in vast quantities and frequently in new formats, has the potential to transform how companies operate internally to support new and better ways to serve customers. With so much new data available, companies are feeling pressure and competition to analyze and make sense of it all. Often the first investment is in the infrastructure needed to store and process big data.

But in many cases, the storage of big data becomes an obsession of sorts. The repository becomes bigger and bigger, but it seems that the data goes in easily but doesn't come out without tremendous effort. To avoid this trap, it is important to have a plan both for acquiring and storing big data as well as for analyzing and using it.

Businesses face a number of challenges with big data analytics:

- How to integrate big data with their existing data streams and repositories
- How to manage the sheer volume of big data
- How to manage complexity of technologies required to analyze big data
- How to make all this data speak intelligently to business analysts
- How to extract the maximum value from big data
- How to explore big data through visualization
- How to mesh existing and future analytics so companies can adapt to change

CITO Research sees tremendous opportunities in big data analytics for businesses. These opportunities can be realized with the right analytics solution. This guide will speak to what is needed from big data analytics, explain how essential data integration is to a successful big data approach, and describe the ideal set up for big data visualization.
Extracting Value from Big Data

Too much is made of the volume of big data. While big data does place a number of new demands on companies, for most organizations, there is value in it if it is managed and analyzed in the right way.

The most important step in creating valuable big data analytics comes from data integration. A business can easily run far afield if users overemphasize the uniqueness of big data and attempt to view it in isolation as a singular phenomenon. To find value in big data, it must be integrated and blended with existing data. Current BI and data integration approaches should not be thrown away. They will be a part of any big data solution because at its core, big data is as much an operational challenge as a technological one. Value in big data comes from data integration with all data warehouses so they speak coherently to one another. Big data matters within the context of the individual business and value emerges as big data is combined with existing data sources.

Big Data Is New, Except for all the Ways It’s Not

To paraphrase a Yogi Berra-ism, big data is entirely new except for all the ways it’s not. Big data quality varies greatly and companies must be careful not to mistake quantity for quality. Big data can often overwhelm through its sound and fury, but not signal anything worth pursuing. As with any data, big data will not provide an analytics story to guide a business’s decisions on its own. The data must be scrubbed, enhanced by data from existing data warehouses, and then refined and explored by content experts and BI professionals who can identify patterns and distinguish distortion from truth.

Companies will experience the greatest ROI when their analytics solutions allow big data to be added to their current BI and application infrastructure so that all data can be blended together for more effective analytics. Data integration therefore should be driven by the need for analytics. Fortunately, thanks to current technology solutions, this is now easier than ever before.

Once data is integrated, a business is ready to delve into analytics. A company should adopt an analytics solution that is highly customizable and scalable to its unique needs. More than one type of analytics may be required. But successful businesses are dynamic enterprises and business needs change over time. High value analytics solutions adjust and operate seamlessly with existing technology as well as technological advances not yet created. Businesses should avoid being boxed in by their existing BI solution and instead invest in products and tool sets that allow for greater growth and easier adaptation.
Interactive and Appealing Analytics Really Matter

And though it may be obvious to experienced BI professionals, analytics work best when reporting is both visually appealing and highly interactive. Analytics is not just about the questions a user wants to ask at the beginning. Analytics is about being able to ask the questions that spring from data exploration. Being able to interact and investigate in this way is crucial to maximizing big data’s potential.

Why Data Integration Is Key to Big Data Analytics

Data integration and big data analytics go hand-in-hand. Yet, despite its importance to analytics, data integration and a company’s use of big data cannot occur overnight. It must be an evolution. Businesses must thoroughly think through how big data will impact their operations and BI before embarking on any data implementation plan because analytics and BI must be a united front.

Businesses should also steer towards big data integration solutions that support them through the entire evolutionary process. Look for solutions that don't just solve the problem of data integration, but also offer analytics and visualization, covering every step of the process, and thereby producing the most value for organizations. The solution should be a one-stop-shop for all of the businesses’ analytics needs.

Like the Goldilocks principle, data integration works best when it hits that “just right” sweet spot, in which all data systems operate harmoniously and speak the same language. Without integration, the value of data of any form is highly limited.

Businesses should also adopt data integration technology that is fluid, flexible, and highly adaptable. Heavily siloed data systems are the antithesis of this type of agility. They promote unnecessary clans within BI and lead to a frayed analytics conception.

With true data integration, in which big data is layered on top of existing data streams and then accessed in the same manner as any other source information, and in relation to other data, a company ensures collaboration among IT, BI professionals, and analysts, establishing internal partnerships and cooperation. With a data integration system that enables analytics, users can access the power of big data anywhere, at anytime, without barriers.
The Implications of Visualization for Big Data Analytics

From the standpoint of the analytics process, big data analytics is not inherently different. To frame the value of visualization to the analytics process, we’ll use the Cross-Industry Standard Process for Data Mining (CRISP). CRISP consists of six phases:

- **Business understanding.** All analytics start with business needs. In order to make analytics useful, you must first understand the business problem and domain.

- **Data understanding.** The next step is to see what data you have (or can acquire), determine what the data can tell you, and understand how it applies to the problem.

- **Data preparation.** The third step is to prepare the data for analysis, which includes blending data sources in meaningful ways. Integration is an important part of this process.

- **Modeling.** This step involves modeling your data to represent a potential solution for the business problem you are working on.

- **Evaluation.** You then evaluate whether the model is working. Most likely, you will iterate, potentially bringing in more data sources, blending them, changing the model, and testing another hypothesis.

- **Deployment.** Once the model is ready, you take the final step and deploy it.

These steps represent a cycle: once the model is deployed and knowledge is gained, it provides new business understanding and insights. At this point, the cycle repeats itself as the company further refines its data strategy.
Visualization for the Data Understanding Phase

If you have a small data set, much of this process can be done manually, inspecting the data to understand it. But the bigger your data is, and the more data sources you apply to the business problem, the more you need a different approach to understanding the data.

Sampling is one way that statisticians deal with large data sets. But sampling requires using a technique to ensure that the sample is statistically significant. But statistically significant on what basis? You don’t know what the basis is until your models are done, so sampling data at the beginning of the process is not a viable solution. If the tools are sampling data before you do your analysis on it, you’re losing valuable information, and not only may your models be inaccurate, they may be totally wrong.

A better way to deal with an overwhelming amount of data is to visualize it. A good toolset for big data analytics offers visualization at all relevant stages of the analytic process. The more data sources you have, and the bigger those data sources are, the more you need visualization to help you with understanding exactly what data you have and how it can help you solve the business problem.

Visualization for Data Preparation

Blending data sources is another key part of the analytics process. By visualizing this process in a tool that allows you to verify that the way you’re blending data accurately reflects the meaning of the data sources (see callout page 6: The Importance of an Integrated Solution), enriching traditional data sources with relevant big data sources becomes an exciting visual experience rather than a tedious task. Furthermore, effective analytics systems allow you not only to blend data sources but to allow data to be dynamically blended to feed the model. In other words, you’re not creating a static dataset as a result of the blending, but a recipe that will be blended anew with the latest information available when you run the model.
The Importance of an Integrated Solution

Here is one real-world example of how an ineffectual analytics system without data integration can negatively impact the way in which a company utilizes big data. Experiences just like this have affected countless businesses as they attempt to adapt to big data.

A business adopted a big data system that claimed to integrate big data but only did so partially. As a result, no attempt was made to ensure uniformity of categorizations across data types. A BI user looked at two fields, both labeled “revenue,” from a big data source and from one of the company’s existing data sources. The big data revenue category represented monthly revenue while the existing data category showed daily revenue.

Assuming the big data field was a daily figure, it appeared that certain customers were spending far more than they actually were. The business consequently targeted a campaign of increased offers and incentives to a cohort of customers it believed were high-purchasers, but who in actuality, seldom frequented the business.

As a result, the company not only went after the wrong consumers, likely ignoring truly profitable customers, but it also gave undeserving customers significant discounts. Those customers had less investment in the company and therefore would likely use the discounts but not provide repeat business—if they used them at all. The result is a business applying its limited resources to the wrong customers and undermining its profit margin.

These types of costly mistakes are common in the emerging world of big data. To avoid them, businesses need analytics solutions that speak across data types and synchronize sources.

Visualization for the Modeling Phase

In the modeling phase, you test the data to see if your working hypothesis is supported by the data or not. You explore the data using simple measures at first (univariate statistics). Visualization is extremely important during the modeling phase, particularly because the model will in most cases need to be revised as you iterate, adding data sources, summarizing data in new ways, pivoting it to see new aspects of your data.
Visualization for the Deployment Phase

Many toolsets offer visualization during deployment only. Visualization plays a critical role here, where analytics are embedded in application where they can drive value for various types of end users and inform their work. Although the end product, the deployment phase, is a critical use of visualization, as we’ve seen, effective big data analytics requires visualization throughout the analytic process.

Features of Successful Analytics Systems—The Power of Visualization

The necessity of data integration for successful big data analytics is clear. But what are the key features of successful analytics systems?

Companies should look for solutions that cover the spectrum of analytics, not just one or two specialized functions. A solution should also cover the entire data analytics supply chain, from obtaining data to delivery of the analytics. A system should include both:

- operational reporting and traditional BI dashboards
- emerging technologies such as visualization and predictive analytics

This isn’t a case of trying to have your cake and eat it too. It makes sense to have a comprehensive platform solution, as it allows users to explore every type of data in one place, without having to toggle back and forth between tools. Taking that one step further, the solution should allow analytics to be embedded where they make the most sense, informing users in the context of the applications where they do their work rather than forcing them to open a separate application for analytics.

A full spectrum tool should provide analytics from all integrated data, allowing interaction between big data and other data sources. This blend shouldn’t be stale and prepared but generated in real-time, pulled from the most recent input. Analytics and visualization should access and highlight data from multiple sources easily and quickly. Factoring in data from multiple sources results in an enriched and immersive analytics environment.
Navigate Big Data with Pattern Analysis

To navigate big data, companies also need tools that avoid dealing with big data on a granular level and that provide relational capabilities to identify patterns. Because of the volume of big data, any attempt to engage with big data on a micro level would be about as productive and efficient as attempting to count all the drops of water in the ocean. Instead, analytic tools must summarize data in concrete and accessible ways. When analysts find particular items they want to pursue on a more specific basis, the tools must also empower them to drill down.

Adaptability is also essential. In order to support big data analytics, companies need a solution that allows users to access data through any data source, such as traditional databases, data warehouses, Hadoop, or other systems. A system should also have the ability to be deployed either on-premise or in the cloud. A company may prefer on-premise at present but in the future want to transition to the cloud. A data integration and analytics solution should support this type of adaptability.

Customizable and Flexible Analytics

An adaptable system should be customizable, flexible, and mobile-accessible. Users should be able to install only the components that are applicable to their needs, giving them as much or as little analytical power as they require, and then scale accordingly at any given time. They should also be able to embed the service in other products and applications, allowing existing data architecture and systems to be incorporated into the new solution. Finally, they should be able to access the system on any device, whether a laptop, an iPad, or a smartphone.

Customization is critical for visualization. The solution should offer a visualization library with expansive options. However, the system should also give users the ability to generate new visualizations as they see fit. This enables a business to truly make the system its own.

It’s also crucial to keep in mind that analytics, despite recent trends, isn’t just about identifying where a business is going. Equally, it’s about establishing where a business is today. Thus, any solution has to be able to do operational analytics as well as predictive analytics.

Users of every skill level should be able to use the system without having to be retrained, allowing for implementation without disruption of ongoing work or existing employees. The system should be intuitive and responsive to each user—that means intuitive for every user, not just for experienced IT or analytics professionals.
Part of this ease of accessibility for every user comes through having an interactive dashboard with strong visualization capabilities. The visualizations must be captivating, yet simple to generate. Visualizations of this kind make big data more understandable and applicable to a wider audience. Visualizations must also be instant and interactive, available on all mobile devices, so that reporting can occur anywhere, at any time, for every employee. And users must be able to create and hone these new analyses without having to turn to IT at every step of the process. This depletes resources and efficiency.

**Future-Proof: An Extensible and Flexible Platform**

A solution should not be built just for today. Technology is just as susceptible to fads as fashion. The hot system today may not be around in five years. Simultaneously, technology is constantly advancing, so any system must incorporate new features without sending a company back to square one. A customizable, extensible, and flexible platform provides businesses with this future-proof security.

Look for a solution that is:

- **Technology agnostic.** Although Hadoop is currently hot, an analytics solution that can be adapted to work with any underlying platform is a better long-term investment. As the underlying landscape evolves technically, it’s important to have a solution that can grow with that changing landscape and not become outdated.

- **Using open standards.** Once a platform espouses proprietary formats, it becomes easy to get locked into a single vendor. Proprietary formats of data interchange between the stages of analysis is one red flag that a solution is designed to lock businesses into a single vendor.

- **Transparent, not a black box.** There are no magic bullets. If the solution you’re evaluating claims to be able to do everything but you don’t have visibility into how it does that, it’s a bad sign.

- **Supported.** Lots of big data types means that you will be incorporating new kinds of data, and, as a result, the first run may not be smooth. Ongoing support helps you work through the kinks as you blend big data with traditional data sources.
What’s Next?

Companies should not approach big data analytics with either fear or too much reverence. With proper big data integration, big data can generate promising analytics to help guide a company’s decision-making. This data integration should be gradual and incorporate existing data so that all analytics are viewed within the context of the business and the business problem. A big data analytics solution should be inclusive of employees of all skill sets and therefore be customizable, adaptable, and ready for the future.

Most importantly, businesses should not buy multiple products to meet their big data analytics needs when they can buy one product that provides all the services they need in a single platform.

CITO Research recommends the Pentaho approach, which supports the entire big data analytics process—from data integration, to data exploration and visualization, to predictive analytics—and will insulate you from much of the big data risk involved in a disruptive, changing market.

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

This paper was created by CITO Research and sponsored by Pentaho