High performance. Delivered.

High Performance IT Insights

Building the Foundation for Big Data

accenture

consulting | technology | outsourcing

For years, companies have been contending with a rapidly rising tide of data that needs to be captured, stored and used by the business. But the nature of that tide has been changing, and increasingly, it includes data from a variety sources, such as social media, sensors, machines and individual employees. This unstructured data now makes up a very significant portion of the data, and companies are rapidly exploring technologies for analyzing this kind of data to gain competitive advantage. (See Figure 1.)

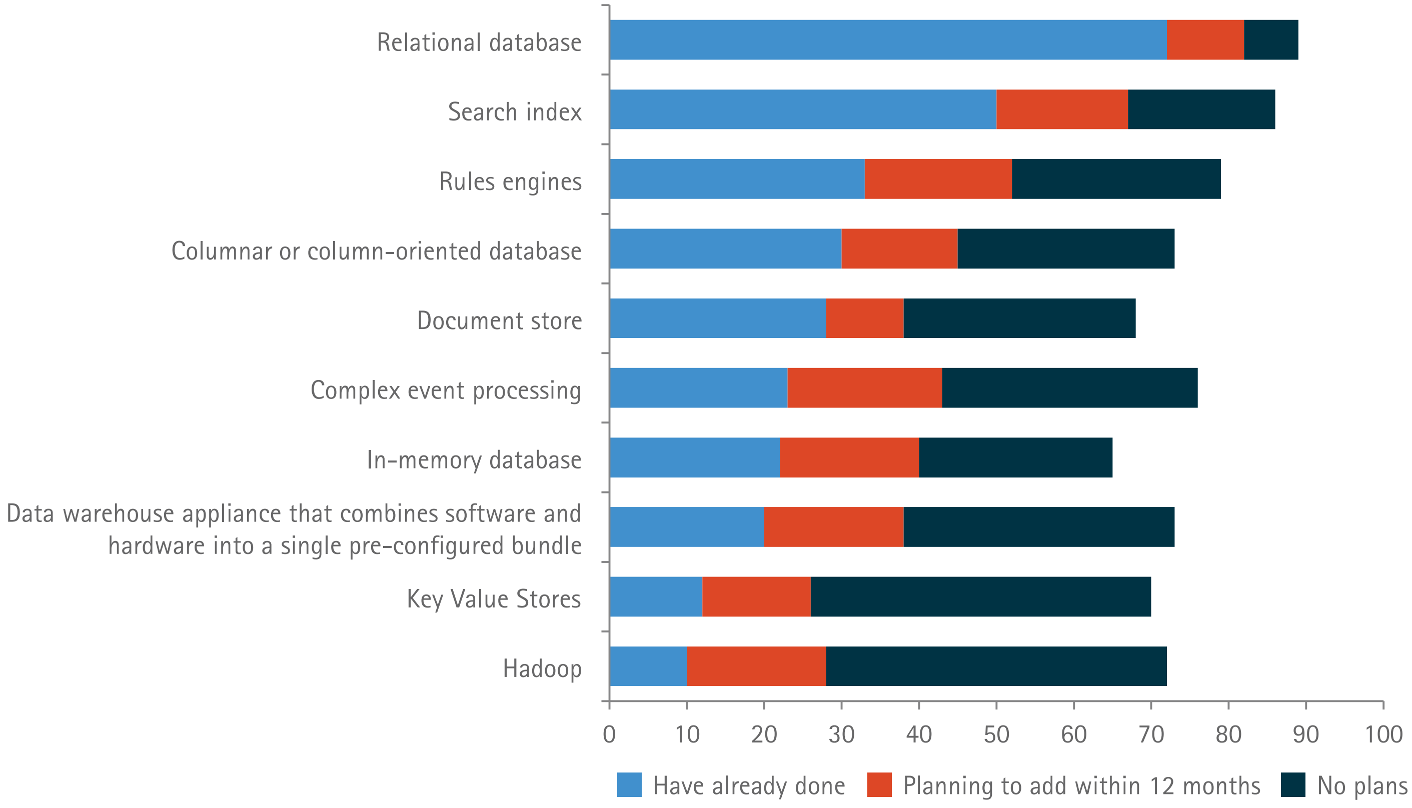
For many, unstructured data represents a powerful untapped resource—one that has the potential to provide deeper insights into customers and operations and ultimately help drive competitive advantage. But this data cannot be easily managed with traditional relational databases and business intelligence tools, which are designed for structured data. At the same time, the rapidly increasing volume and velocity of structured and unstructured data are only complicating that problem. All of this has led to the development of new distributed computing paradigms known collectively as Big Data, and analytics technologies such as Hadoop, NoSQL and others that handle unstructured data in its native state.

Big Data technologies have enabled companies to explore how to increase the efficiency, tame the total cost of ownership and flexibility of their underlying IT infrastructure. The continued growth of data has forced companies to seek new ways to acquire, organize, manage and analyze this data.

In reality, the advent of Big Data is bringing new, unprecedented workloads to the data center. Handling those workloads will require a distinct, separate infrastructure, and IT will need to find ways to simultaneously manage both the old and the new—and ultimately bring the two together.

Figure 1

Q; Does your organization use or have plans to deploy the following technologies?



Source: IDC and Computerworld BI and Analytics Survey Research Group IT Survey, 2012, n = 111

Understanding the Costs

The implementation of Big Data capabilities can have an impact on many aspects of the IT infrastructure. Before launching a Big Data initiative, companies should make sure that they have a clear idea of the total costs involved. To do so, they need to consider factors such as:

- Hardware costs, including servers, storage and networking.
- Software costs, including Big Data software (like Hadoop and its ecosystem) and connectors required for integration with traditional databases and business intelligence tools.
- Implementation costs, such as research, design and plan work; installation and configuration; integration with existing business intelligence applications; and post-installation development and testing.
- The cost of risk and quality problems, because Big Data implementations are not easy, and issues can slow progress and require costly re-work.
- IT opportunity costs, because the time spent on the installation and integration of Big Data reduces the time that IT can spend on activities that add value to the business.
- The costs of delayed business improvements, because the expected workforce productivity increases and delivery of business insights cannot be achieved until implementation is complete.

Understanding the Big Data Platform

The benefits of being able to take advantage of Big Data are clear and significant. But so too are the challenges that Big Data brings to the data center. IT infrastructure teams will have to deal not only with huge volumes of data, but also with the complexity of varied types of data and the increasing velocity with which that data needs to move. In addition, not all of this data has value, and IT must assist data scientists to sift through huge amounts to find the "needle in a haystack" needed to create business insights.

All in all, Big Data will require an infrastructure that can store, move and combine data with greater speed and agility—and traditional IT infrastructures are simply not engineered to meet this need. It is technically possible to translate unstructured data into structured form, and then use relational database management systems to manage it, of course. However, that translation process takes a great deal of time, driving up costs and delaying time to results. In general, the problem is not so much technological as financial; it is simply not economically feasible to use the traditional infrastructure to manage Big Data.

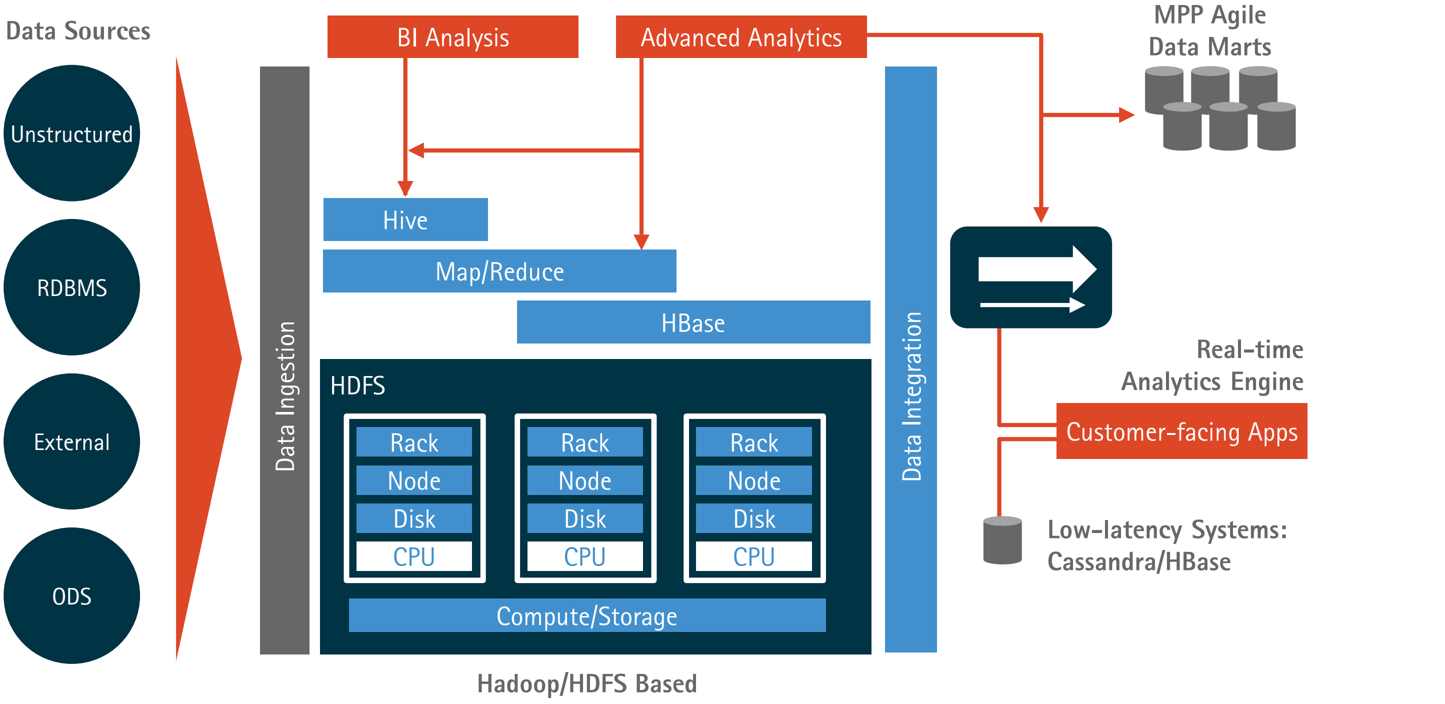
It's clear then that Big Data will require its own, more cost-effective approach to infrastructure—and in many cases, that approach will represent a shift from past practices. For several years, data centers have been focused on virtualization and consolidation, moving to smaller footprints based on relatively few large servers connected to large shared storage platforms. However, Accenture believes that Big Data may often require essentially the opposite approach based on a more decentralized model. In many situations, the right Big Data platform will consist of clusters of numerous smaller, commodity servers, rather than enterprise-class platforms. And storage will be handled locally at the individual server level, rather than centralized and shared. (Certainly, there are cases where pre-built Big Data engineered systems will be the most appropriate approach; these are discussed later.)

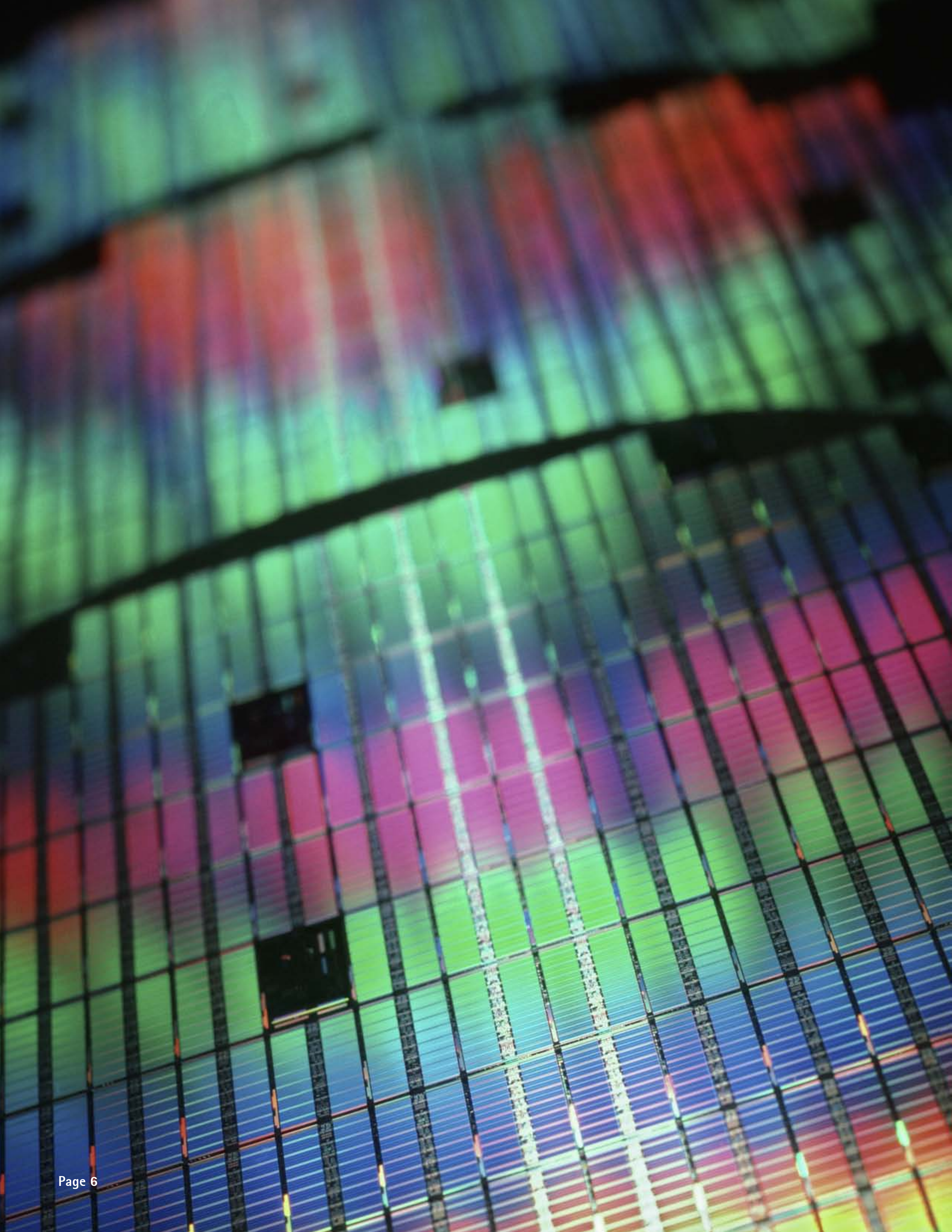
This decentralized approach to Big Data has several advantages. For example, it provides cost-effective flexibility, with the ability to scale out quickly to include thousands of relatively inexpensive servers, rather than going through the expensive upgrade of enterprise servers and storage equipment. And in terms of performance, the shared-nothing model eliminates the need to funnel data through a limited number of shared storage disks—thereby doing away with a huge bottleneck that could seriously affect performance when dealing with large amounts of data.

None of this is to say that this Big Data platform will replace the existing infrastructure, or eliminate the need for virtualization and consolidation in the traditional infrastructure. In looking at the Big Data and existing platforms, it is not a matter of "either-or," but rather of "and." To derive business value from the full spectrum of data, IT infrastructure teams will have to work with both infrastructure models, and operate two more or less distinct platforms—and then develop data architectures that encompass both.

With that in mind, Accenture foresees a rebalancing of the database landscape as data architects embrace the fact that relational databases are no longer the only tool in the toolkit. "Hybrid solution architectures" will mix old and new database forms, and advances pioneered in the new infrastructure will be applied to invigorate the older infrastructure. (See Figure 2.) In short, tomorrow's conversations about data architectures will center on rebalancing, coexistence and cross-pollination between the two infrastructures.

Figure 2: Converging Data Architecture





Big Data and the Data Center

The emergence of these Big Data platforms will have significant ramifications for the data center, and infrastructure professionals will have to address a number of new challenges. For example, data centers will need to manage large-scale Big Data platforms, with hundreds or thousands of new clustered servers being added to the infrastructure. They will also need to manage the provisioning and orchestration of services across these numerous nodes, and integrate the Big Data management suite with the traditional management suite.

Big Data will also put new demands on the network infrastructure, which will need to move terabyte-sized data sets. Even the basics of physical infrastructure will be affected, as the installation of large numbers of commodity servers will require adjustments in power supplies and heating/cooling, as well as floor space.

Similarly, the Big Data storage infrastructure will need to have multi-petabyte capacity and the ability to support, potentially, billions of objects. And because unstructured data represents an increasingly valuable business asset, companies will have to take steps to keep it protected and available. This, too, may require new approaches, because the volumes involved may be too large to back up and restore through conventional methods. Security features of Big Data technologies are continuously maturing and organizations will need to consider implementing adequate controls to address how to prevent data compromise and theft.

IT governance will also need to be adapted to support Big Data. As a rule, companies will have to make sure that governance processes are in place for everything from performance management to service chargeback, incident/problem management and service desk support for the Big Data platform.

Finally, IT will need to determine how to integrate the Big Data platform with the rest of the IT infrastructure. Companies will want to draw on both structured and unstructured data, so that the two together can provide a fuller picture of the business, and so Big Data can be understood in the context of other enterprise data. This integration will allow companies to leverage existing data warehouses and analytic tools, and enable decision makers to make widespread use of Big Data throughout the organization.

Planning the Infrastructure

IT groups will need to take a comprehensive, multidiscipline approach to create Big Data platforms. IT infrastructure teams should work with other IT professionals who can provide perspectives on analytics, risk and compliance, business applications and IT governance. These varied perspectives can help ensure that data center services are reengineered for the volume, velocity and complexity of Big Data, and that there is a path to bring the Big Data and traditional architectures together—with an ongoing focus on the economics involved. As per Accenture's research a "data-centric" design is more important than it has ever been.

It is also important to recognize that there is not a single "one-size-fits all" approach to the Big Data platform. Each company's situation will be different, which makes careful upfront planning critical. Infrastructure teams need to fully understand the impact that Big Data will have on the data center. That means analyzing data center capacity, storage and networking requirements. It means identifying potential data sources and gauging the data set sizes that will need to be managed. It means understanding the analytics workload in term of volume and velocity, as well as CPU and IO workloads. And it means determining the level of integration that will be needed between the Big Data platform and traditional business intelligence tools.

As discussed earlier, some companies are likely to find the decentralized, commodity-hardware "shared nothing" infrastructure to be appropriate. But there are many cases where other approaches make better sense. For example, the use of the commodity platform with shared storage may be right when smaller workloads are involved, and where concerns about storage bottlenecks impairing performance are minimal. This might include situations where the company is just beginning to explore Big Data, and workloads are limited. (See Figure 3.)

Figure 3: Infrastructure Solution Patterns

Solution Patterns	Use Cases	Details
Commodity platform local storage	1. High flexibility and large scale outs 2. Hadoop implementation skills easily accessible 3. Develop or have access to a reference architecture for Hadoop implementation	1. Commodity physical servers 2. Configured in PODs comprising *racks of commodity servers* 3. Direct attached storage ~12x3TB per node 4. Onsite disaster recovery backup and recovery 5. Infrastructure automation and orchestration 6. Plan for data center capacity
Commodity platform shared storage	1. Small—medium implementation 2. Hadoop implementation skill easily accessible 3. Develop or have access to a reference architecture for Hadoop implementation	1. Virtual servers running on hypervisors like VMWare ESXi 2. Configured in PODS comprising *n*ESX clusters with *n to 1 density* 3. Shared scale out NAS 4. Shared storage can be a potential bottleneck 5. Onsite backup and recovery 6. Offsite replication for disaster recovery 7. Infrastructure automation and orchestration 8. Plan for data center capacity
Big Data Appliances (Teradata, DCA, Oracle)	1. Fast time to delivery 2. Tight integration with existing BI Analytics platforms (Oracle, Greenplum, Teradata)	1. Bundled computer, storage, network and Big Data components 2. Engineered for high availability and fault tolerance 3. Simple and unified management 4. Hadoop management tools 5. System management tools 6. Single support
Cloud implementation (single tenant or multi-tenancy)	1. Fast time to delivery 2. Data center capacity issue 3. Want to experiment with Big Data 4. Already using could for infrastructure	1. Challenge to move data sets to cloud at start (and possible termination of service) 2. Attention to data security and privacy 3. Data ownership in cloud

In other cases, packaged, engineered systems might be appropriate—particularly when time-to-implement is critical. These solutions may involve more upfront hard costs than clusters of commodity servers. But because they bundle technology and software, it is possible to get them in place much more quickly, as well as avoid the complexities (and additional costs) of implementing Hadoop and connecting hardware, which can be significant. The engineered solution approach may also be useful in cases where they can facilitate integration with the existing infrastructure. (See "Simplifying Big Data Implementation.") For example, the Oracle Big Data appliance can streamline such integration for companies that are using Oracle databases and business intelligence tools to handle structured data.

Keeping the Focus on the Business

While Big Data and traditional infrastructures will differ in many ways, one fundamental principle holds true across both—the need to ensure that IT enables business results. That means companies need to carefully assess and monitor total cost of infrastructure ownership as they move forward with Big Data platforms.

At the same time, they need to look beyond costs and target infrastructure capabilities that support business agility and growth. Accenture research shows that high-performance businesses tend to emphasize a number of important factors, such as having adaptive, executable strategies in a changing environment, and creating competitive advantage through continual innovation. Big data technologies allow for much more flexibility in mobilizing data more quickly to meet the demands of business. The Big Data infrastructure can be key to enabling those approaches. To that end, companies need to ensure that the infrastructure lets IT continually cost-optimize operations, scale up and down against business demand and demonstrate value back to the business.

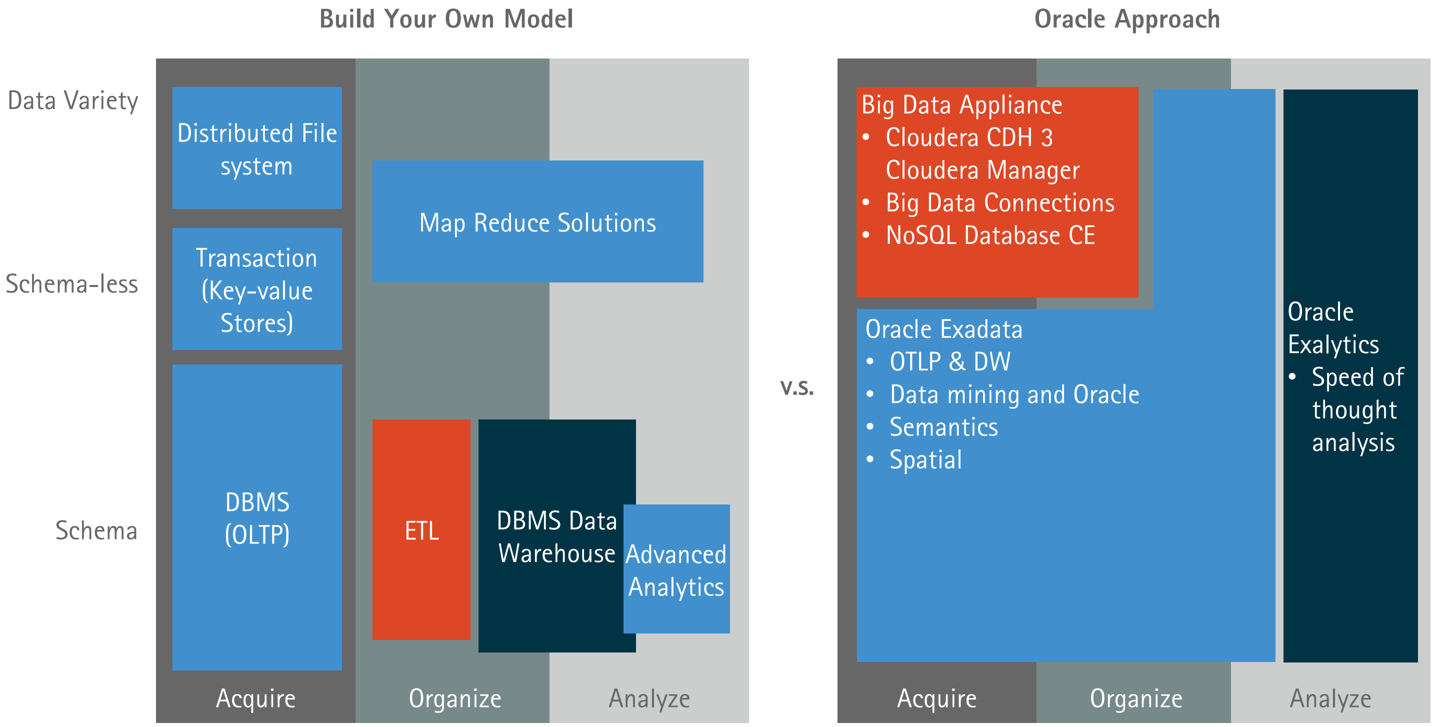
A well-planned approach to the Big Data infrastructure can deliver that type of infrastructure—and ultimately help both IT and the business succeed. Computing paradigms may change with the advent of Big Data, but the business' expectations for IT as an enabler of efficiency and innovation will not—and that will ultimately be the measure of success for the Big Data infrastructure.

Simplifying Big Data Implementation

A Big Data implementation, including the integration of various infrastructure components, can be a complex task that requires specialized skills. In addition, as Big Data plays an increasingly vital role in companies, it will become more important that the related infrastructure have the kind of performance, security and support seen in other critical business solutions. With these realities in mind, companies may want to consider packaged, engineered systems that provide "ready-made" Big Data platforms.

In essence, the potential value of these engineered solutions comes down to reduced set-up times and streamlined ongoing management—factors that can be vitally important in some situations. For example, Oracle's offering in this space—the Oracle Big Data Appliance—includes 648 terabytes of raw storage and 216 CPU processing cores in a single rack, optimized for Big Data. The appliance includes a complete set of Big Data software, such as Hadoop and NoSQL. (See Figure 4.) The entire pre-configured package is designed to provide the high levels of performance, availability and security required for enterprise systems.

Figure 4: Two Approaches: Build Your Own vs. Use Oracle Engineered Systems



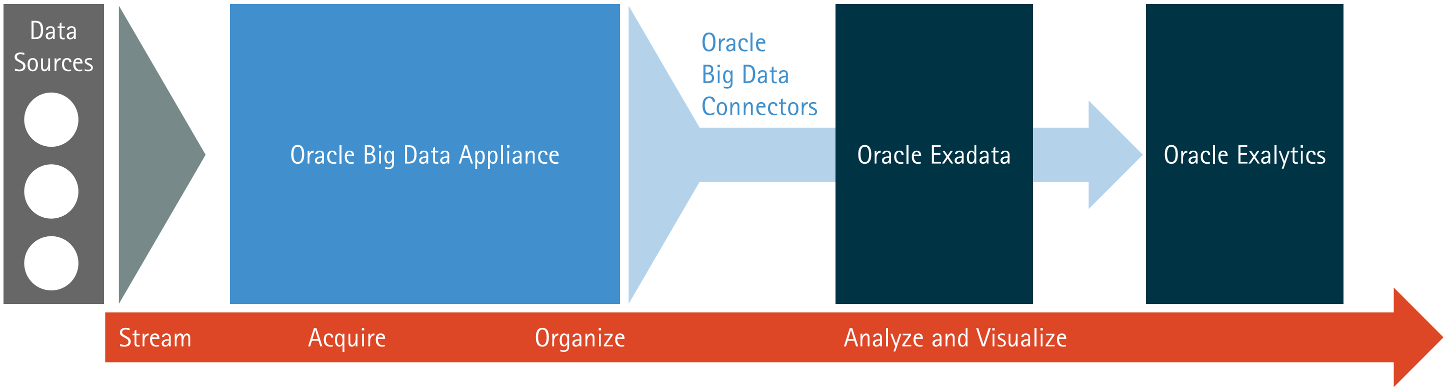
Reprinted with permission of Oracle Corporation.

This approach can cut implementation time significantly. For example, in one assessment of a large Big Data effort, Oracle estimated that for a 10-rack, 144-node system, hardware implementation alone would require nearly 1,800 cables and about 1,300 man-hours of work. With the appliance package, the same implementation would require just 48 cables and 38 man hours. And looking at the long run, the Oracle appliance provides a single point of support for Big Data hardware and software, reducing the complexity of dealing with multiple vendors.

The appliance approach can also facilitate integration of Big Data platforms with the rest of the infrastructure. For example, many companies use Oracle databases, and the Oracle Big Data Appliance offers special software "connectors" for integrating the appliance with the Oracle Database, as well as with the Oracle Exadata storage solution. (See Figure 5.) The Oracle Big Data Appliance uses high-speed InfiniBand links to connect with these other Oracle systems, enabling companies to create a very high-powered computing environment—again, with a single vendor supporting the entire environment. Overall, this approach can enable companies to weave Big Data into the overall analytics ecosystem, helping to simplify the process of acquiring, organizing and analyzing a broad range of data.

Each company needs to weigh a number of factors, including its own needs, to determine which approach to Big Data is right. But as Big Data tools and technologies evolve, the ready-made engineered solutions provided by vendors are likely to provide an appealing option for many.

Figure 5: Oracle Analytics – Powered by Engineered Systems



Reprinted with permission of Oracle Corporation.

If you would like to know more about how Big Data can help your organization achieve high performance, contact:

Arun Sondhi
arun.sondhi@accenture.com
+1 262 212 9496

Dr. Mark Gold
mark.a.gold@accenture.com
+1 703 947 3639

Patrick Sullivan
patrick.sullivan@accenture.com
+1 312 693 3411

Vincent U. Dell'Anno
vincent.u.dellanno@accenture.com
+1 719 244 6325

About Accenture

Accenture is a global management consulting, technology services and outsourcing company, with 257,000 people serving clients in more than 120 countries. Combining unparalleled experience, comprehensive capabilities across all industries and business functions, and extensive research on the world's most successful companies, Accenture collaborates with clients to help them become high-performance businesses and governments. The company generated net revenues of US$27.9 billion for the fiscal year ended Aug. 31, 2012. Its home page is www.accenture.com.