bluedata

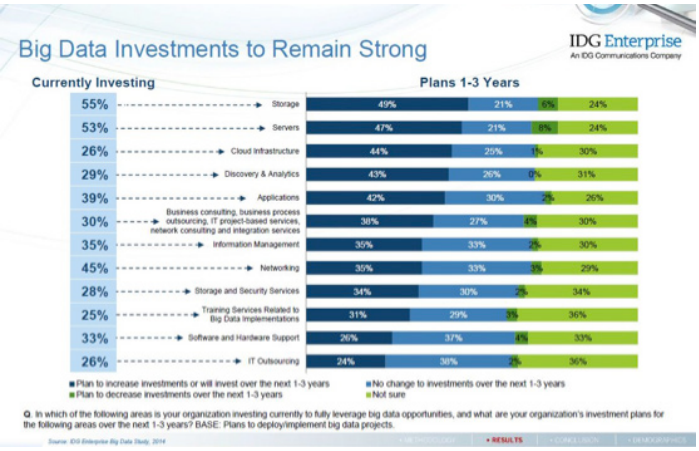
Changing the Equation on Big Data Spending

The Big Data Dilemma

Over the past few years "Big Data" has evolved from an interesting technology topic into a source of major competitive advantage. As a result, many businesses are making significant investments in Big Data. A survey of enterprises by IDG[1] found that Big Data initiatives were a high priority for 60% of enterprises, and that they planned to spend on average $8 million on those initiatives in 2014.

Despite all the planned financial and IT commitments, few enterprises have much to show for their initiatives in either production applications or competitive information. The exceptions are companies like Amazon and Google, which have nearly limitless resources and people to put towards their Big Data projects. Almost everyone else is struggling to turn commitment into results. Gartner predicts that 85% of the Fortune 500 will fail to see a competitive advantage from their Big Data initiatives by the end of 2015.[2]

Somewhere between intention/investment and execution/production, Big Data initiatives are falling into a gap. Before investing more resources into Big Data projects, companies need to understand exactly how and where things are going wrong.



Enterprises are increasing investments in servers, storage and cloud infrastructure [IDG Enterprise Big Data Study, 2015]

The Risks of Ignoring Big Data

While some companies are joining the "Big Data backlash" or simply staying on the sidelines, they run the significant business risk of missing the benefits and advantages that accrue from successful Big Data applications.

Nearly every business is already generating and relying on many new types of data – including geolocation data, social graphs, click-streams, etc. With the ever-expanding Internet of Things, sensor-generated data is fueling innovative services and applications. As businesses exploit ways to gain a competitive edge through data, ignoring this source of business insight is a risky strategy.

What makes Big Data big?

Gartner first defined Big Data more than a decade ago, using the "3 V's" – data with high *volume*, high *velocity* and high *variety*. Its current definition is as follows: "***Big data*** is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

At the least, taking advantage of the Big Data already within the business is necessary to keep even with the industry. Those companies that creatively apply Big Data to transform their markets can reap the benefits of market leadership.

Over the past few years "Big Data" has evolved from an interesting technology topic into a source of major competitive advantage. As a result, many businesses are making significant investments in Big Data. A survey of enterprises by IDG found that Big Data initiatives were a high priority for 60% of enterprises, and that they planned to spend on average $8 million on those initiatives in 2014.

[1] "CEOs Call for Big Data and IT Continues to Lead Investment Decisions," IDG Enterprise, January 7, 2014. http://www.idg.com/www/pr.nsf/ByID/MYAR-9F5Q7P
[2] "Gartner Reveals Top Predictions for IT Organizations and Users for 2012 and Beyond," Gartner Inc., December 1, 2011. http://www.gartner.com/newsroom/id/1862714

Despite all the planned financial and IT commitments, few enterprises have much to show for their initiatives in either production applications or competitive information. The exceptions are companies like Amazon and Google, which have nearly limitless resources and people to put towards their Big Data projects. Almost everyone else is struggling to turn commitment into results. Gartner predicts that 85% of the Fortune 500 will fail to see a competitive advantage from their Big Data initiatives by the end of 2015.

The Cloud Decision: Public or Private?

Enterprises today are finding that their traditional database systems don't deal well with this new type of data – especially the semi-structured and unstructured data types. Managing and analyzing Big Data require new approaches, using solutions like Hadoop, NoSQL, and others. Businesses must invest in people with the right skill sets for running and managing these applications, as well as the environment to run it.

When making that investment, businesses can choose between *public* and *private* cloud infrastructure. In the public cloud (Amazon AWS and Google Cloud), the cloud provider owns and manages the infrastructure, which it shares among multiple tenants. In a private cloud, the organization owns and runs its own cloud infrastructure, shared among different internal clients.

Many factors affect the public or private cloud decision. The choice of private cloud is often driven by the issue of *control*. With a public cloud infrastructure, organizations have no real control over the security or availability of their data. They have no visibility into exactly where data resides, and whether the NSA is collecting data from the public cloud provider. If there's an outage, they do not know the reason, nor do they control the response. And they have little control over the cost of running the applications – for large or growing applications, the cost of using public cloud infrastructure can become very high.

For these reasons, many businesses choose to run Big Data applications on private cloud infrastructure.

The Private Cloud for Big Data

Any private cloud is built on physical infrastructure – the servers and storage at the lowest level of the diagram below. Those physical resources are provisioned to specific applications in a cloud infrastructure. Above that are the *distribution* and *application* layers, where most of the Big Data attention is focused.

Most companies, seeking business value and competitive differentiation, are investing heavily in the top layers. But decisions made at this level can create problems in the infrastructure layer that can only be addressed with more hardware or more specialist/expert time and effort.

For example, when using physical Hadoop clusters, before any analysis can start, the data must be copied into the HDFS file system for the cluster. This process is known as the "ingest" of the data.

Physical Hadoop clusters do not share or prioritize resources across jobs well. High priority jobs may easily be delayed by lower priority jobs on the same cluster.

To achieve the differentiating benefits you want from the Big Data applications, you must start with an infrastructure that is truly elastic, secure, and easy to manage.

Infrastructure Constraints and Costs

The focus and investments related to the top layers often have unintended consequences in the infrastructure. The scope of these problems may not be clear until it's time to scale out and into production. Common problems plaguing current private cloud Big Data deployments include the following:

Cluster sprawl: As they provision multiple Hadoop clusters to handle applications with different Quality of Service levels, priorities or security requirements, businesses end up managing many diverse clusters with low overall utilization (30% or less).

Duplicate data stores: Big Data applications typically need data in a dedicated file system. Data that already exists elsewhere must be copied to the cluster file system. And because different compute clusters cannot easily share the same file systems, administrators spend time copying massive amounts of data, increasing both storage costs and the risks of data leakage.

Deployment delays: Applications written for different Hadoop distributions cannot easily run on the same cluster, so each new application requires time to spin up another cluster. The longer it takes to handle the infrastructure issues, the longer it takes to realize value from Big Data investments.

Each of these problems contributes to the shortage of capital and skills. For example, cluster sprawl and data duplication consume capital costs for hardware, as well as the operating costs necessary to run the poorly utilized equipment. And the more time spent on deployment and scaling issues, the less time Big Data experts have for higher-value projects.

To achieve the differentiating benefits you want from the Big Data applications, you must start with an infrastructure that is truly elastic, secure, and easy to manage.

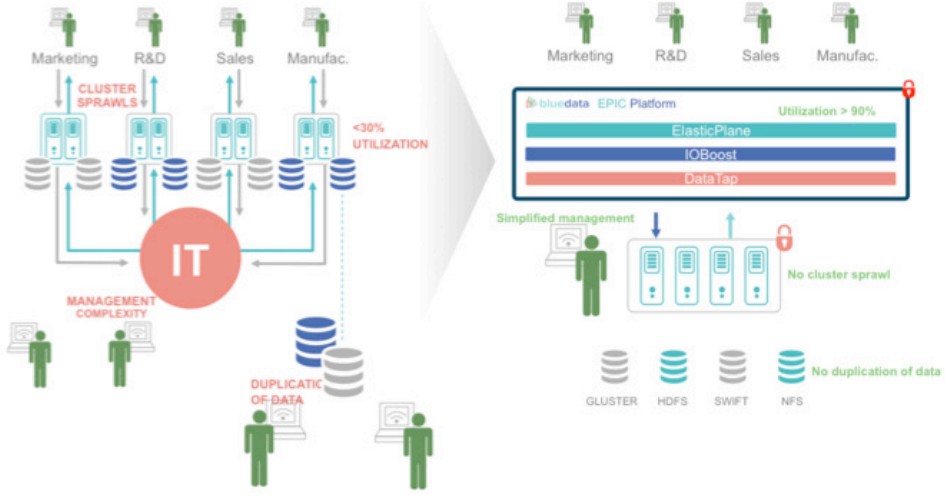
IT Has Solved Similar Problems Before

If these infrastructure problems sound familiar, there's a good reason for it. IT organizations have faced similar problems in the past:

- *Cluster sprawl* is closely related to *server sprawl* – the proliferation of servers in an IT environment with servers dedicated to specific applications.
- *Duplicate data stores* look very much like the storage silo problems when storage was directly attached to all of those sprawling servers.

Modern data centers have basically solved those problems. *Server virtualization* (such as that provided by VMware) simplifies the provisioning of applications on the underlying physical infrastructure while significantly increasing resource utilization. *Storage virtualization* technologies (including Storage Area Network and Network Attached Storage) reduce the need to maintain separate islands of storage for each application.

The industry needs a similar approach to simplifying the creation of virtual Hadoop or NoSQL clusters on the physical infrastructure underneath the private cloud.



BlueData Platform: Private Cloud Infrastructure for Big Data Apps

Founded by veterans from VMware, BlueData is the pioneer in big data virtualization. The company is democratizing big data by streamlining and simplifying big data infrastructure and eliminating complexity as a barrier to adoption. Using BlueData's EPIC software platform, enterprises can now build agile, secure and cost-effective big data deployments that deliver value in days instead of months and at a cost savings of 50-75% compared to traditional approaches. With BlueData, enterprises of all sizes can create a public cloud-like experience from their on-premise environments and get the same value out of their big data as companies like Google, Facebook and Yahoo, at a fraction of the cost and with far few resources.

Within the private cloud, BlueData offers essential features to simplify provisioning and management:

- Instant, self-service provisioning of Hadoop and NoSQL clusters lets organizations create and launch jobs as needed, quickly and without deep expertise.
- With a secure *multi-tenant* architecture, different groups or applications can share the same physical infrastructure securely, with virtual isolation.
- Intelligent resource management capabilities increase utilization, supporting resource allocation according to business and application needs.
- Policy layers establish and apply service levels for enterprise applications.

The BlueData platform incorporates many patent-pending virtualization enhancements for distributed data workloads, addressing issues that were previously barriers to the use of virtualization. BlueData delivers self service, speed and scale through innovations such as ***IOBoost***™, an application-aware caching service to streamline I/O and network performance; ***DataTap***™ that accelerates time to results by eliminating delays in copying large volumes of data and an ***ElasticPlane***™ that leverages next-generation hypervisor and container technologies for policy-based cluster management services.

For more information about the BlueData architecture, read the paper *Solving the Big Data Intention-Deployment Gap*.

Changing the Economics of Big Data

The BlueData EPIC platform does for the Big Data infrastructure what server virtualization did for the data center infrastructure – reducing both the capital cost and administrative costs of managing the infrastructure running the business applications.

- ***Eliminate Cluster Sprawl:*** Create instant personalized clusters in a multi-tenant, private cloud environment with smart resource utilization. When different clusters share the same core physical infrastructure in a securely, overall utilization increases and cluster sprawl ends.
- ***Keep Data in One Place:*** With BlueData, data can stay in existing enterprise storage and be accessible to multiple applications, reducing data duplication, storage costs and data leakage risks.
- ***Accelerate Time to Value:*** BlueData can create personalized Hadoop and NoSQL clusters immediately, without requiring time or expertise. Developers can create and then repurpose clusters for infrequent or one-off applications.

By eliminating cluster sprawl and data duplication, BlueData reduces spending on servers, storage and cloud infrastructure. Because it works with any storage or server, businesses can repurpose existing equipment.

By simplifying the provisioning and deployment of infrastructure for cloud applications, BlueData reduces the time and expertise devoted to the lower levels of the application stack. Big Data experts can focus instead on the data that drives competitive advantage.

Summary

Using BlueData directly addresses the budget and expertise constraints that limit today's Big Data initiatives. BlueData creates an elastic, enterprise-grade cloud infrastructure that reduces both capital expenditures (on servers and storage) and operating costs (on provisioning, deploying and managing the infrastructure).

Using BlueData, businesses can get straight to the high-value application work, realizing a faster time-to-value for Big Data investments.

To learn more, visit www.bluedata.com

bluedata