White Paper

Changing the Equation on Big Data Spending

Big Data analytics can deliver new customer insights, provide competitive advantage, and drive business innovation. But complexity is holding back the return-on-investment for Big Data – it's too time-consuming, expensive, and resource-intensive to initiate and scale deployments for Hadoop and related technologies. The time is ripe for a new approach that will change the equation for 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 Gartner[1] found that more than 75 percent of enterprises are investing or planning to invest in Big Data in the next two years.

> "As Big Data becomes the new normal, information and analytics leaders are shifting focus from hype to finding value."
>
> *- Lisa Kart, research director at Gartner*

Yet despite all these planned commitments, relatively few enterprises have much to show for their initiatives in terms of concrete ROI from production Big Data deployments. The exceptions are companies like Yahoo and Google, which have nearly limitless resources and people to put towards their Big Data projects. Many other enterprises are struggling to turn commitment into results. In the Gartner survey, 43 percent of those planning to invest and 38 percent of those that have already invested in Big Data don't know if their ROI will be positive or negative.

Somewhere between intention and execution, 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.

The Risks of Ignoring Big Data

While some companies are 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/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

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 business can reap the benefits of market leadership.

[1] Source: "Survey Analysis: Practical Challenges Mount as Big Data Moves to Mainstream", Gartner, September 2015

Big Data On-Premises or in the Cloud?

Enterprises today are finding that their traditional database and data warehousing systems don't deal well with these new types of Big Data – especially the semi-structured and unstructured data types. Managing and analyzing Big Data require new approaches, using solutions like Hadoop, Spark, 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, these enterprises may choose between running their Big Data infrastructure on-premises (i.e. in their own data centers) or in the public cloud. In a public cloud service like Amazon Web Services' Elastic MapReduce (EMR), the cloud provider owns and manages the infrastructure, which it shares among multiple tenants. For on-premises deployments, the organization owns and runs its own infrastructure; most on-premises Hadoop deployments run on bare-metal physical servers with direct attached storage.

Many factors affect the on-premises versus public cloud decision. The choice of on-premises infrastructure is often driven by issues of *control, security, and governance.* With Hadoop-as-a-Service in the public cloud, organizations have the benefits of self-service and instant cluster provisioning. They can also avoid the complexity and challenges of an on-premises infrastructure deployment; the simplicity of the cloud model is compelling. In particular, many data scientists and analysts tend to use public cloud services for initial development and testing.

However, organizations have less control over the security or availability of their data in a public cloud service. They have no visibility into exactly where data resides, and there are often industry or government regulations that may preclude large organizations from putting data into a public cloud environment. If there's an outage, they do not know the reason, nor do they control the response. And they have little control over the ongoing costs – the organization may avoid upfront capital expenses for on-premises infrastructure, but the ongoing operating expenses can become significant over time.

And perhaps most importantly, much of the data that enterprises want to analyze is already on-premises. It can be time-consuming, expensive, and risky to move large volumes of data into a public cloud service. For these reasons, many businesses choose to run Big Data applications on their own on-premises data center infrastructure.

Big Data Infrastructure Complexity

Any on-premises Big Data deployment runs on physical *infrastructure:* the servers and storage at the lowest level of the diagram in Figure 1. Those physical resources are provisioned for specific Hadoop distributions as well as various analytical tools and applications. These *distribution* and *application* layers are where most of the Big Data attention is focused.

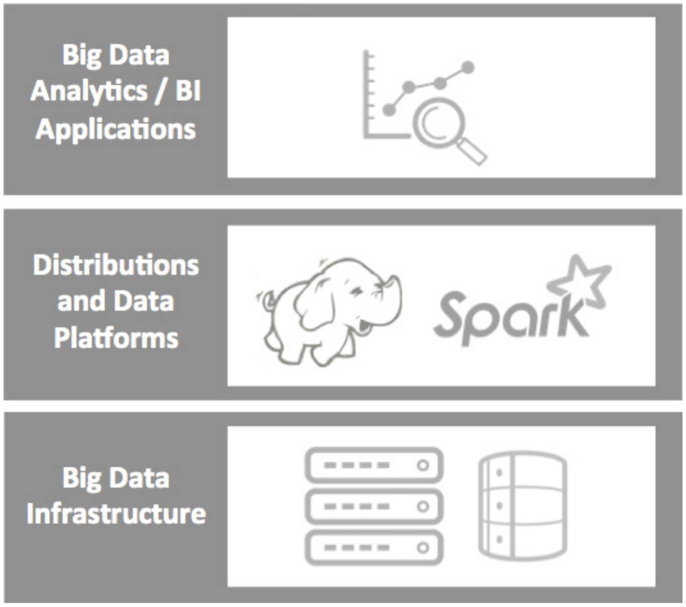
Big Data Analytics / BI Applications
Distributions and Data Platforms
Spark
Big Data Infrastructure

Figure 1: *Big Data Applications, Data Platforms, and Infrastructure*

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 Hadoop Distributed File System (HDFS) for the cluster. This process is known as the "ingest" of the data.

And for each new physical cluster, a new bare-metal server is required. The provisioning and configuration of these clusters may take weeks (sometimes months). Moreover, these 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 Big Data applications, a new infrastructure approach is required.

Infrastructure Constraints and Costs

The focus and investments on the top layers of the Big Data stack 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 on-premises Big Data deployments include the following:

- **Cluster sprawl:** As they provision multiple physical 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 (typically 30% or less).

- **Duplicate data stores:** Big Data applications typically need data in a dedicated file system like HDFS. Data that already exists elsewhere must be copied to the file system. And because different clusters cannot easily share the same file systems, administrators spend time copying 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 increasing costs and a shortage of 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 is spent on deployment and scaling issues, the less time Big Data experts have for higher-value projects.

IT Has Solved Similar Problems Before

If these problems sound familiar, there's a good reason. 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 physical 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 solved similar problems in the past. *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 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 Spark clusters on physical infrastructure. To date, that hasn't been possible due to concerns about data locality and I/O performance. But with today's technology advancements, those issues have been addressed. It's now possible to get the I/O performance of bare-metal physical server deployments for Big Data, with all the agility and efficiency benefits of virtualization.

BlueData: Big Data Infrastructure Made Easy

Founded by veterans from VMware, BlueData is the pioneer in Big Data virtualization. BlueData provides a software platform to streamline and simplify Big Data infrastructure — eliminating complexity as a barrier to adoption. BlueData's infrastructure software platform uses virtualization and container technology to make it easier, faster, and more cost-effective to deploy Big Data.

Using BlueData's EPIC™ software platform, enterprises can now deploy agile and secure Big Data environments 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 provide a cloud-like experience for their on-premises deployments – delivering Big-Data-as-a-Service while leveraging their own data center infrastructure.

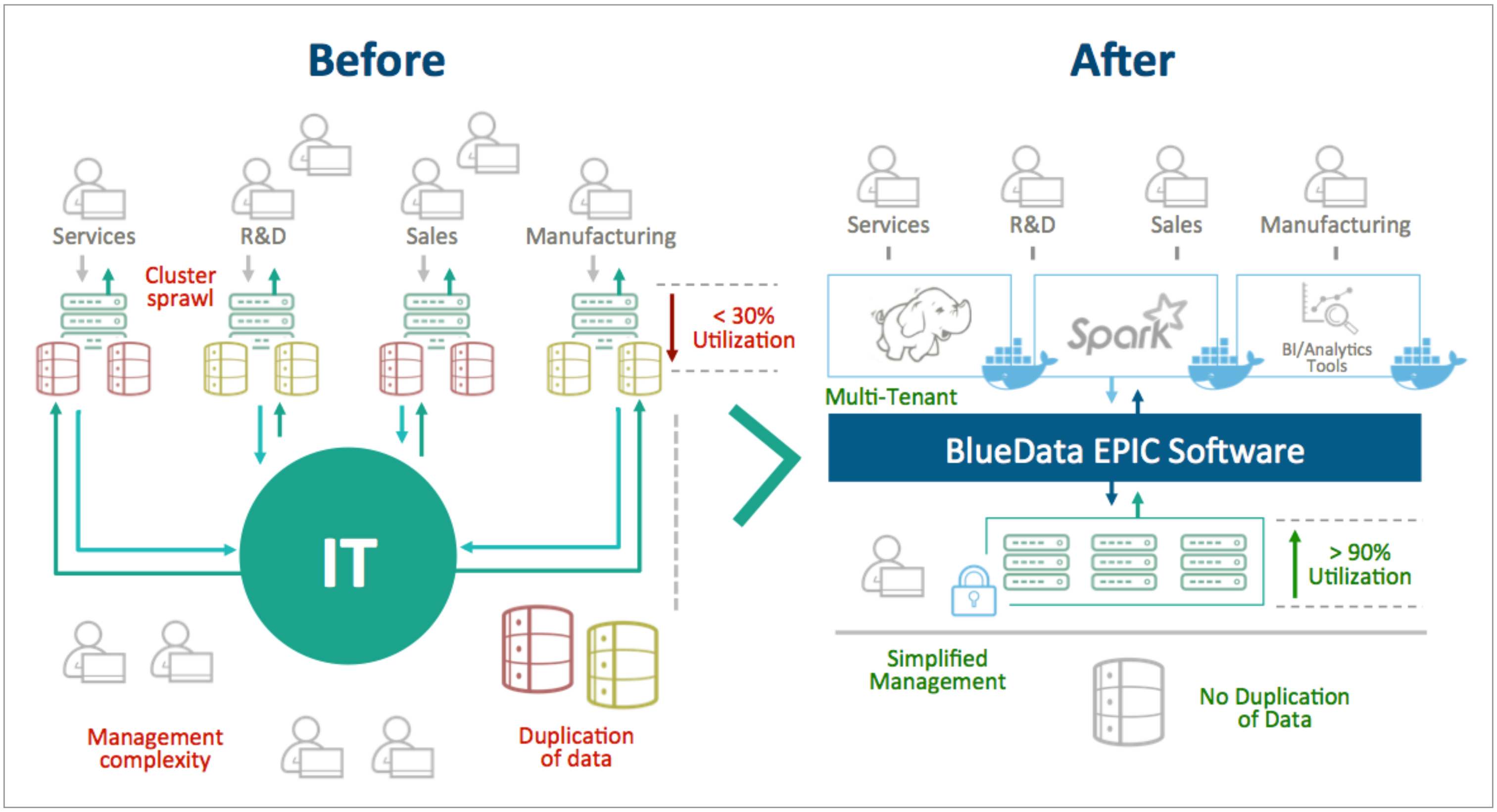

Figure 2: *A New Big Data Deployment Model*

BlueData offers essential features to simplify Big Data infrastructure:

- Instant, *self-service* provisioning of Hadoop and Spark clusters lets organizations create and launch jobs as needed, quickly and without deep expertise.
- With a *multi-tenant* architecture, different groups or applications can share the same physical infrastructure securely, with virtual isolation.
- By separating compute and storage, IT teams can *increase utilization* and scale infrastructure resources independently.
- Enterprises can run analytics using any *shared storage,* eliminating the need to duplicate and move data.

The BlueData EPIC platform incorporates many patent-pending innovations 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 *ElasticPlane™,* which provides an easy-to-use interface for spinning up instant Hadoop and Spark clusters in Docker containers; *IOBoost™,* which uses application-aware caching to deliver comparable I/O performance to that of bare-metal; and *DataTap™,* which accelerates time-to-results by eliminating delays in copying large volumes of data.

Changing the Economics of Big Data

The BlueData EPIC platform does for Big Data infrastructure what server virtualization did for data center infrastructure more than a decade ago – it reduces infrastructure and operational costs, while increasing speed and agility. Some of the key benefits include:

- **Eliminate Cluster Sprawl:** Create instant Hadoop and Spark clusters in a multi-tenant environment with smart resource utilization. When different clusters share the same physical infrastructure, overall utilization increases and cluster sprawl can be brought under control.
- **Keep Data in One Place:** With BlueData, data can stay in existing enterprise storage systems; the data can be shared and made accessible for multiple user groups and multiple applications. This eliminates the need for data duplication, reduces storage costs and minimizes data leakage risks.
- **Accelerate Time-to-Insights:** BlueData can create clusters immediately through self-service, without requiring specific Big Data or infrastructure expertise. Data scientists and developers can get faster time-to-insights, and they no longer need to wait for IT to spin up clusters for infrequent, transient, or one-off applications.

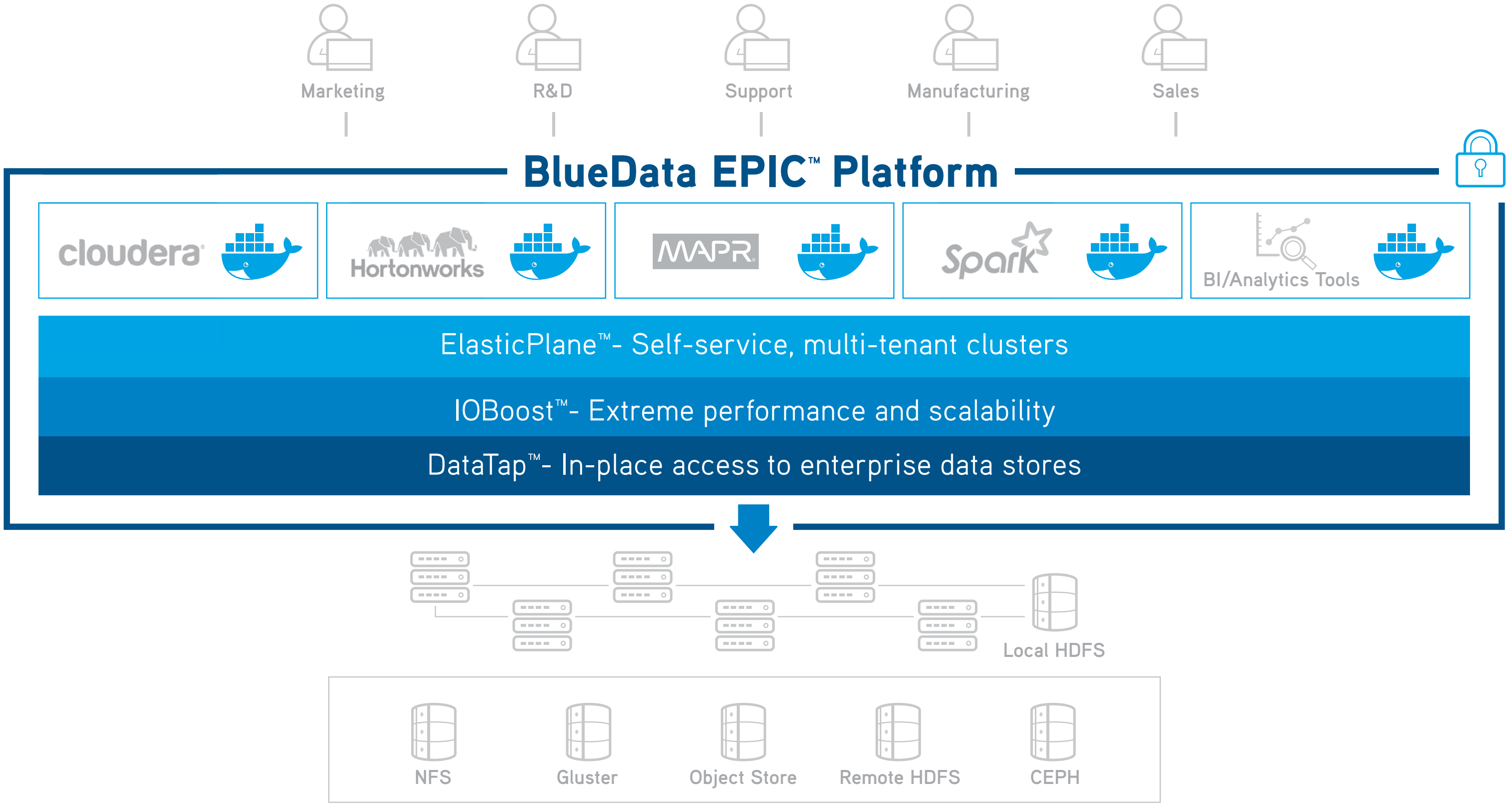

Figure 3: *The BlueData EPIC Software Platform*

By eliminating cluster sprawl and data duplication, BlueData allows organizations to reduce costs for Big Data infrastructure and operations. And because it works with any storage or server environment, IT can leveraging their existing investments in data center infrastructure.

By simplifying the provisioning and deployment of infrastructure, BlueData also reduces the time and expertise required to deploy Big Data on-premises. Big Data experts can focus instead on the data and insights that drive competitive advantage for the business.

Summary

BlueData has introduced a simplified deployment approach that directly addresses the budget and expertise constraints that limit today's Big Data initiatives. BlueData enables Big-Data-as-a-Service in an on-premises model, reducing both the capital expenditures (on servers and storage) and operating costs (on provisioning, deploying and managing the infrastructure) for Big Data deployments.

Enterprises need to spend more of their budgets on innovation and new business opportunities resulting from data insights – and less on hiring new Big Data specialists, deploying new infrastructure, or moving existing data.

With BlueData, IT organizations can save up to 75% on the infrastructure and operational costs for Big Data deployments by improving hardware utilization, reducing cluster sprawl, and eliminating data duplication. And they can accelerate the deployment time for their Big Data infrastructure and applications. With BlueData, enterprises can now change the equation for their Big Data spending – they can realize faster time-to-value and better overall return-on-investment.

To learn more, visit www.bluedata.com