White Paper

Solving the Big Data Intention-Deployment Gap

Big Data is on virtually every enterprise's to-do list these days. Recognizing both its potential and competitive advantage, companies are aligning a vast array of resources—human and technological—to access and analyze this strategic asset. And yet, despite best intentions and resources, the vast majority of these Big Data initiatives are either extremely slow in their implementation or are not yielding the results and benefits that enterprises expect.

Why this gap between intention and deployment? Because enterprises are working from a top-down assumption in which all they need to do is pick the right analytics tool, the right Hadoop distribution and train the right people. Once these elements are in place, the rest is smooth sailing.

Unfortunately, this assumption is only half correct. Enterprises also need a bottom-up approach, one that guarantees that their Big Data initiatives are built on an infrastructure that is designed for Big Data.

> "Through 2017, 60% of Big Data projects will fail to go beyond piloting and experimentation and will be abandoned."
>
> *Source: Gartner*

The traditional approach to Big Data infrastructure is extremely complex, rigid and expensive. Today, it takes enterprises anywhere from six weeks to three months to stand up Hadoop applications after securing the necessary server hardware. The rigidity of this infrastructure model constrains the business and results in loss of agility. Only when Big Data analytics are combined with flexible, easy-to-use, and simple-to-deploy infrastructure will organizations have the formula for success.

Big Data Infrastructure: Barrier to Adoption

In its early days, the BlueData team met with dozens of enterprises about their Big Data challenges. The result confirmed our belief that the majority of Big Data failures take place during the transition from the proof-of-concept stage to the production stage. Digging in deeper, we discovered that these difficulties usually could be traced to the complex, inflexible infrastructure that serves as the foundation of most Hadoop deployments, especially physical cluster deployments where compute utilization is often less than 30% and storage cannot be scaled independently.

We found that this sort of rigid infrastructure holds back these organizations' abilities to provision, manage and run their Big Data jobs in a number of ways, including:

- The inability to easily and cost-effectively manage multiple Hadoop clusters running at the same time when clusters have different priorities, Quality of Service/Service Level specifications, and/or data security requirements.
- The need to copy data into an HDFS file system before Big Data applications are allowed to access it. This is a costly and time-intensive process that slows the time to results, and increases the risk of data leakage.
- The inability to run applications written for different Hadoop distributions on the same cluster or to quickly introduce newly developed analytic tools such as Spark.
- An explosion in uncontrolled "cluster sprawl" as different departments build their own separate clusters to meet diverse demands across business units.
- The reluctance to repurpose an idle Hadoop cluster because so much time was spent installing and configuring it for an ad-hoc or infrequent Big Data job.

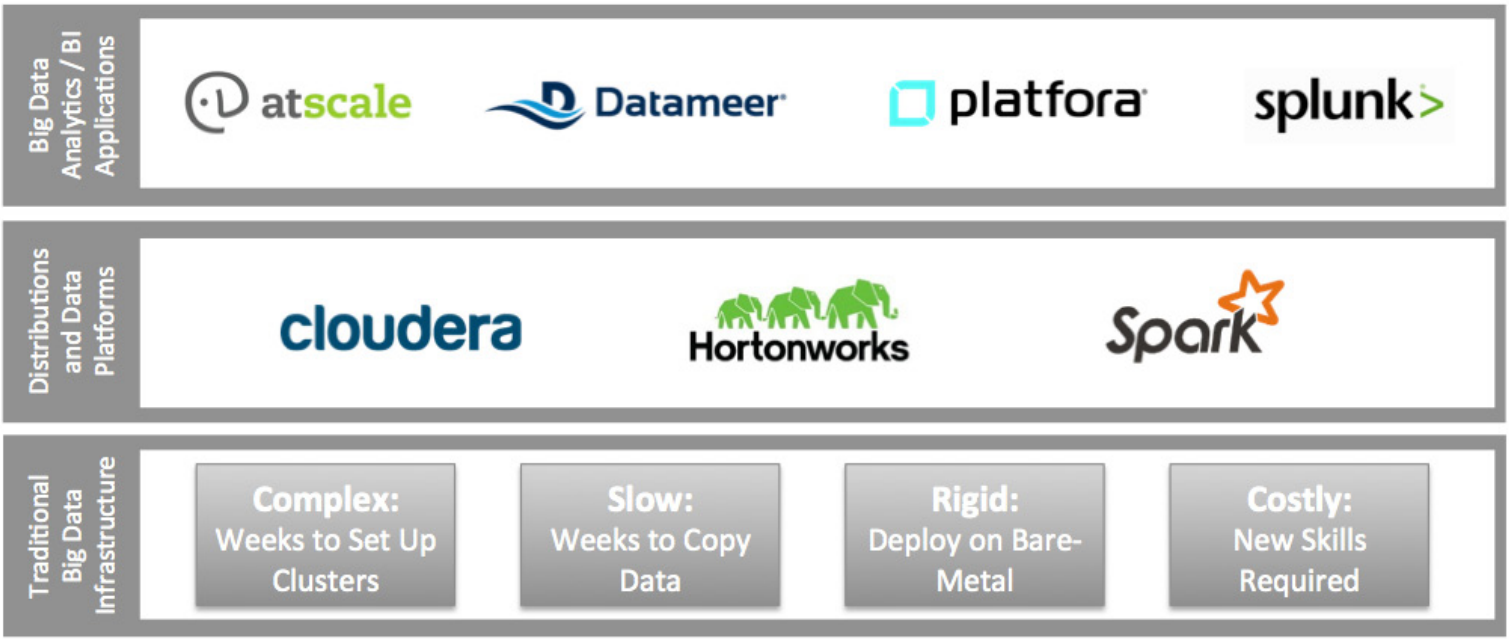

Figure 1: *The Traditional Deployment Model for Big Data Infrastructure is Complex, Slow, Rigid, and Costly*

A New Approach to Big Data Infrastructure

BlueData was founded on the premise that a fundamentally new approach is required to the bridge the gap between Big Data intentions and deployment success.

BlueData's mission is 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 for enterprises to deploy Big Data infrastructure on-premises.

With the BlueData EPIC software platform, enterprises can now deliver value from Big Data within days instead of months and at a cost savings of up to 75% compared to traditional approaches. With BlueData, enterprises can create a cloud-like, Big-Data-as-a-Service experience for their on-premise environments.

To ensure that enterprises are able to quickly, cost-effectively and consistently derive value from their growing data sets, we designed our infrastructure software platform with the following core principles as our guidance:

Make Big Data available to everyone. As more stakeholders within an organization realize the potential of Big Data and demand access to a company's resources, Big Data infrastructure must be flexible enough to accommodate all their needs. To do so, the foundation of any Big Data platform must:

- Simplify the complexity of provisioning Big Data jobs so that non-experts have the capacity to build and manage their own jobs without the cost and expertise of Big Data infrastructure specialists.
- Permit the simultaneous provisioning and running of multiple clusters so that users throughout the organization may create their own Big Data jobs as needed within minutes, instead of waiting weeks or months.

Separate compute and storage. Infrastructure must be flexible enough to allow an enterprise to fully disconnect analytical processing from data storage. This sort of separation allows an organization to:

- Take advantage of any distributed storage technology.
- Access data directly from the organization's enterprise storage systems and avoid the expensive and time-consuming step of copying data to HDFS prior to running any analytics.
- Keep sensitive data within an organization's secure enterprise storage systems.
- Independently scale compute (CPU) and storage on an as-needed basis.

Permit the use of any Hadoop application. Most Big Data infrastructure envionments restrict an enterprise's choice of applications to those that are supported by a specific vendor's Hadoop distribution, limiting an organization's ability to quickly bring the latest open source Big Data applications into use. A truly flexible infrastructure allows an organization to take advantage of any Big Data application available across the open source community.

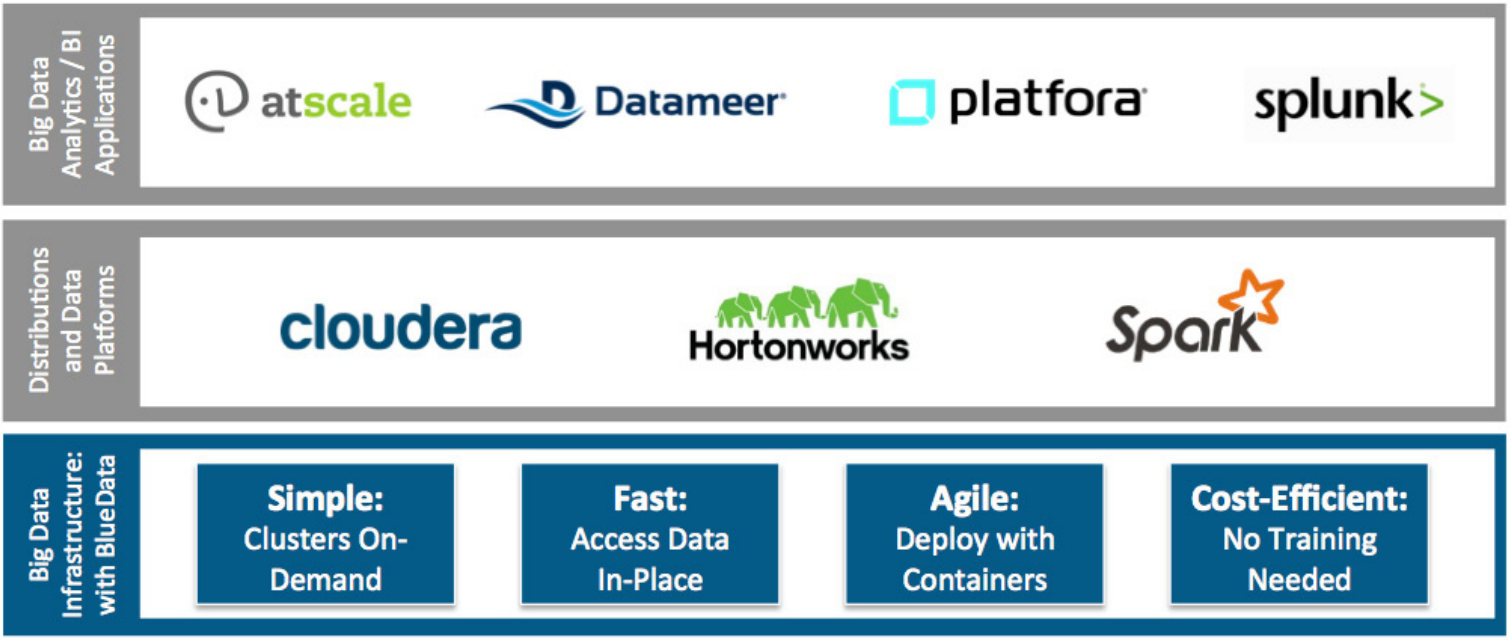

Figure 2: *BlueData Enables a New Deployment Model for Big Data Infrastructure*

Provide the benefits of Big-Data-as-a-Service. When virtualization and container technology is wedded with the performance of the BlueData software platform, organizations find that the self-service agility and increased flexibility achieved by running Big Data applications on virtual clusters outweighs any benefits that might come bare-metal deployments. This flexible, cloud-like model also enables an enterprise to:

- Avoid the challenges of moving data from an on-premises environment into the public cloud.
- Ensure enterprise-grade data security not available for Hadoop or Spark running in a public cloud.
- Quickly scale operations as needed to take advantage of dynamic market conditions.
- Realize significant capital expenditure savings.

The BlueData Solution: Agility, Control, and Lower Cost

The goal of the BlueData solution is to unleash the power of Big Data by giving enterprises a simple, cost-effective Big Data architecture that lets them move off their rigid infrastructure and eliminate the Big Data intention-deployment gap. The virtual infrastructure layer that serves as the foundation for the BlueData platform fundamentally changes how enterprises provision, run and manage their Big Data jobs.

This approach comes from the power of virtualization and a container-based, vendor-agnostic Big Data platform that separates compute from storage. The result is an agile and flexible infrastructure that gives enterprises total control over their Big Data environment.

Key capabilities of the BlueData EPIC software platform include:

- **Self-service provisioning** of Hadoop and Spark clusters, empowering diverse end users across the enterprise with the ability to set up and launch Big Data jobs tailored to their needs.
- **Multi-tenancy** that gives disparate stakeholders within the organization (e.g., marketing, R&D, sales, manufacturing) the ability to run simultaneous Big Data jobs.
- The ability to **access and run Big Data jobs directly from existing enterprise-class storage systems** without the requirement to copy and move data before it is accessible for Big Data analytics.
- **Elasticity** that permits the platform to dynamically adapt to changing workload requirements in the most cost-efficient manner.
- The ability to use **any distributed storage technology.**
- **Instant scalability** both up and down that lets the enterprise immediately respond to changing Big Data requirements.
- **Enterprise-grade security** for sensitive data because there is no need to copy and move it out of the data center.
- **I/O optimization** that provides all the benefits of running Hadoop in a virtualized environment while retaining the performance of a physical cluster.

- **Policy-based** automation and management that includes control over reservation of resources for different tenants and application-sensitive caching to maximize performance.
- Full **IT visibility and centralized control** of all clusters.
- The ability for IT to **spin up new clusters in minutes** - to evaluate, deploy, and test multiple Hadoop distributions, including multiple versions of those distributions, and Spark standalone.
- An opportunity to quickly deploy virtual clusters for **any Big Data application,** including analytics, business intelligence, data preparation, and search tools.

Big Data Infrastructure Software

The BlueData EPIC software platform leverages virtualization technology and patent-pending innovations to deliver self-service, speed, and efficiency for Big Data environments:

- **ElasticPlane**™ enables users to spin up virtual clusters on-demand in a secure, multi-tenant environment.
- **IOBoost**™ ensures performance on par with bare-metal, with the agility and simplicity of Docker containers.
- **DataTap**™ accelerates time-to-value for Big Data by eliminating time-consuming data movement.

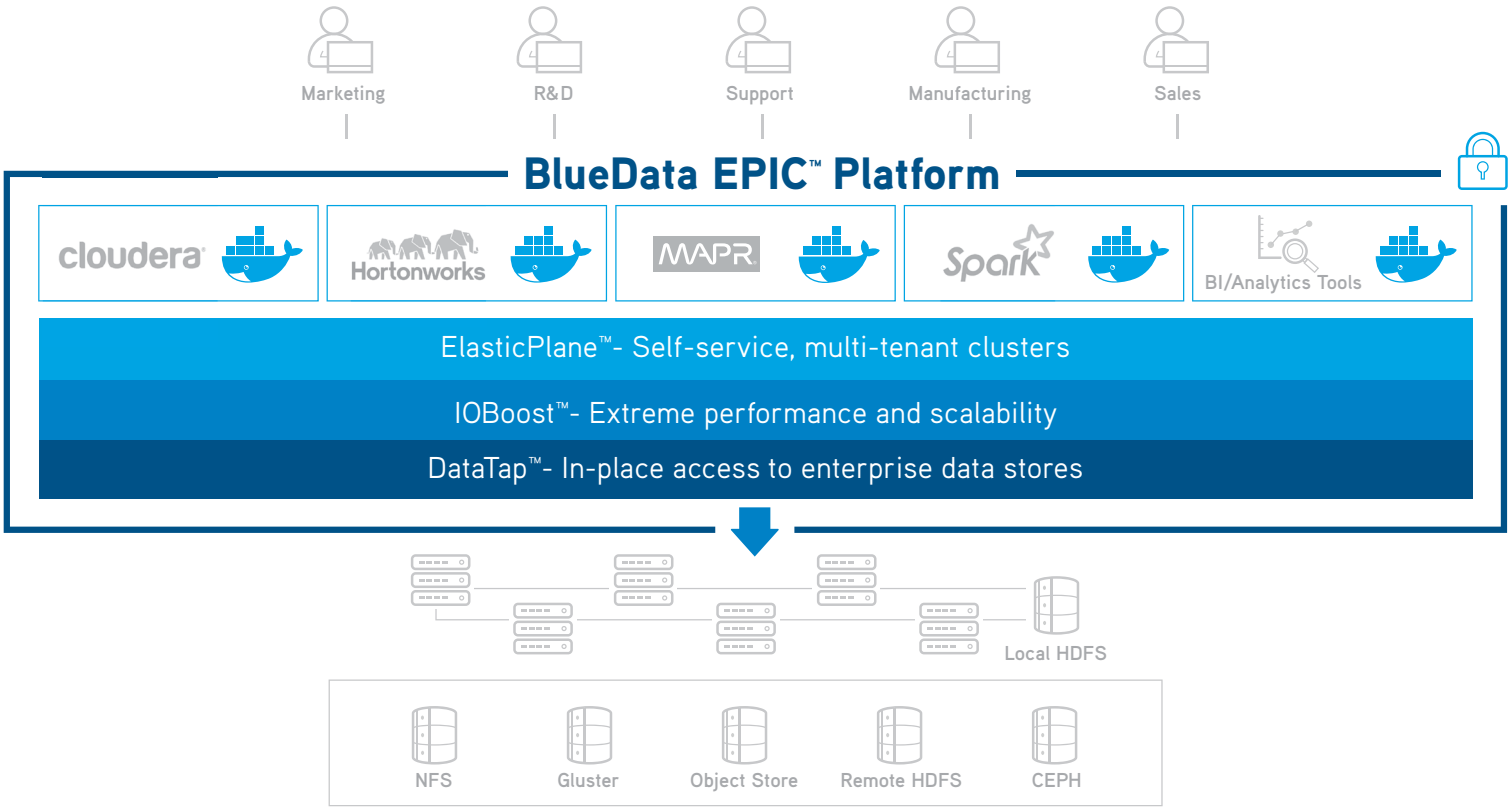

Figure 3: *The BlueData EPIC Software Platform*

Big Data: It Starts with a Solid Foundation

Turning today's Big Data into tomorrow's insights and economic benefits is the critical challenge for today's enterprises. And with the competitive landscape as fast-paced as it is, there is no room for false starts or mistakes.

Success with Big Data requires two components: 1) a solid understanding of the new analytics tools that can generate strategic insights; and 2) a more flexible infrastructure that fundamentally addresses and solves the complexities, costs and constraints that frustrate Big Data pioneers today. Only when these two components work in harmony will enterprises be able to turn their data into insights and translate those insights into effective frontline action.

About BlueData

BlueData is transforming how enterprises deploy their Big Data applications and infrastructure. The BlueData EPIC software platform uses virtualization technology to make it easier, faster, and more cost-effective for enterprises of all sizes to leverage Big Data – enabling Hadoop-as-a-Service in an on-premises deployment model. With BlueData, they can spin up virtual Hadoop or Spark clusters within minutes, providing data scientists with on-demand access to the applications, data and infrastructure they need. Based in Mountain View, California, BlueData was founded by VMware veterans and its investors include Amplify Partners, Atlantic Bridge, Ignition Partners, and Intel Capital.

To learn more, visit www.bluedata.com

A Hadoop Deployment Checklist

Enterprises have several infrastructure deployment options from which to choose when it comes to running Hadoop. Those options include:

- Physical Hadoop clusters deployed on-premises;
- Virtual Hadoop clusters deployed on-premises;
- Hadoop-as-a-Service in a public cloud;
- Some combination of the above.

Before making a Big Data infrastructure investment, IT executives and their teams need to ask themselves the following questions:

1. **How many distinct Hadoop clusters do I need to run?** In addition to needing separate clusters for development and production purposes, you'll also want to provide diverse stakeholders in your organization the opportunity to run data on their own clusters. Select infrastructure that gives you the capacity to run multiple Hadoop clusters simultaneously.
2. **How often will I run Hadoop jobs:** 24x7, weekly, monthly, or quarterly? When Hadoop jobs are not running, your expensive physical systems will be idle. Select an infrastructure that permits you to instantly scale up or down, as needed.
3. **How will I run Hadoop jobs on existing data?** The traditional Hadoop deployment model requires you to copy data into a HDFS file system before your Big Data applications can access it. This time-consuming and expensive process results in the creation of multiple copies of data. Choose a deployment model that gives you the option to access and run Big Data jobs directly from your enterprise storage.
4. **How will I implement data security?** If a Big Data job requires the use of sensitive data, be aware that some infrastructure options require you to copy and move that data from your existing enterprise-class storage systems into HDFS before it's accessible to Hadoop jobs. The result is multiple copies of data to secure and manage. Ensure your infrastructure does not require you to copy and move sensitive data off your organization's secure systems.
5. **What sort of speed do I require for my Big Data job?** The speed at which a Hadoop job runs depends on its ability to be broken into smaller data sets for simultaneous processing. The number of nodes in a cluster limits this "parallelism." For a highly parallelizable job, you want as many nodes as possible to quickly complete the job. However, when running a less parallelizable job, you don't want to pay for unused nodes. Choose a deployment model that allows you to easily adjust job compute resources for maximum cost efficiency.
6. **Do I have in-house Hadoop infrastructure expertise?** To achieve maximum performance, Hadoop clusters require careful, precise tuning tailored for specific applications. Keeping Hadoop software current with patch sets and functional improvements also requires specialized expertise. Your Big Data infrastructure should be simple to manage and maintain, and it should enable self-service provisioning of Hadoop clusters to make it easy for end users.
7. **Is my Hadoop infrastructure future proofed?** Applications written for Hadoop constantly change, and chances are the ones you run today will not be the ones you'll use a year from now. Your infrastructure should be flexible enough to adapt to these changes, with the ability to easily add new services and applications.