# 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.

> "Through 2015, more than 85 percent of Fortune 500 organizations will fail to effectively exploit Big Data for competitive advantage."
>
> Source: Gartner

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/noSQL 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.

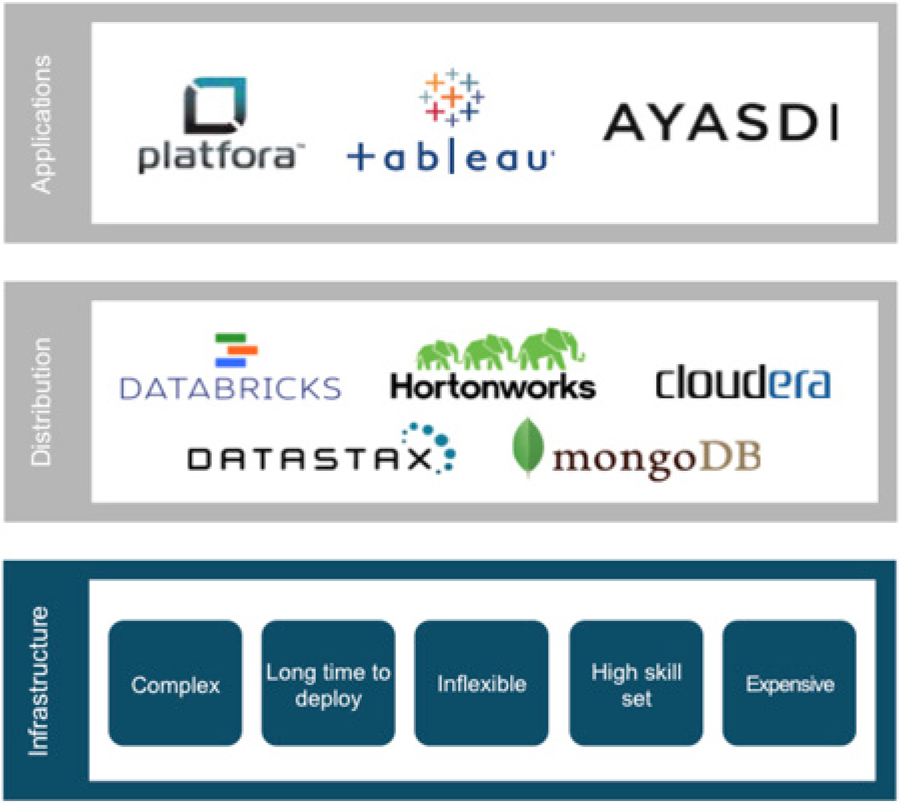
Big Data infrastructure is extremely complex, rigid and expensive. Today, it takes enterprises anywhere from six weeks to three months to stand up Hadoop applications after securing the necessary server hardware. The rigidity of this infrastructure model constrains the business and results in loss of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

## Rigid Infrastructure Holding Back Big Data

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 less than 30% and storage cannot be scaled independently.

This sort of rigid infrastructure stymies the organizations' abilities to provision, manage and run their Big Data jobs in a number of ways, including:

- » An 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, not only a costly and time-intensive process that slows the time to results but one that also 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 infrequent Big Data job.



## A Fresh Approach to Big Data Infrastructure

BlueData was founded on the premise that the bridging of the Big Data intention-deployment gap would require a fundamental transformation. The company is democratizing Big Data by streamlining and simplifying Big Data infrastructure and eliminating complexity as a barrier to adoption. With its 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.

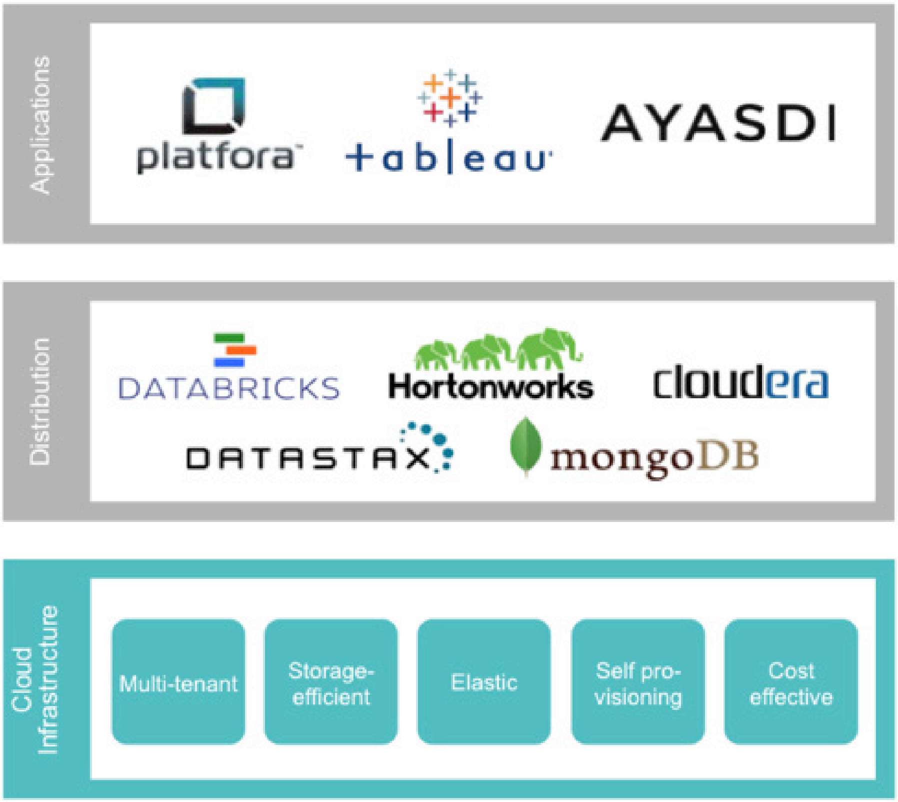
To ensure that enterprises are able to quickly, cost-effectively and consistently derive value from their growing data sets, we designed our infrastructure solutions with the following core principles as our guidance:

**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, a 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 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 an 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 infrastructures restrict an enterprise's choice of applications to those that are provided 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.

**Take advantage of a private cloud environment.** When private cloud virtualization is wedded with the performance of an adaptable Big Data framework, organizations find that the improved management and increased flexibility achieved by running Hadoop applications on virtual servers dwarfs any benefits that might come from running Hadoop on physical servers. This flexible private cloud virtualization model also enables an enterprise to:

- Experience enterprise-grade data security not available through an IaaS or PaaS 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 & Low 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 schemes and eliminate the Big Data intention-deployment gap. The re-envisioned infrastructure layer that serves as the foundation of 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 private cloud, vendor-agnostic Big Data platform that separates compute from storage. The result is an agile, cost-efficient infrastructure that gives an enterprise total control and flexibility in how it takes advantage of its Big Data.

Core features of the BlueData solution include:

- **Self-service provisioning** of Hadoop and NoSQL clusters to empower 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 to Hadoop analytics.

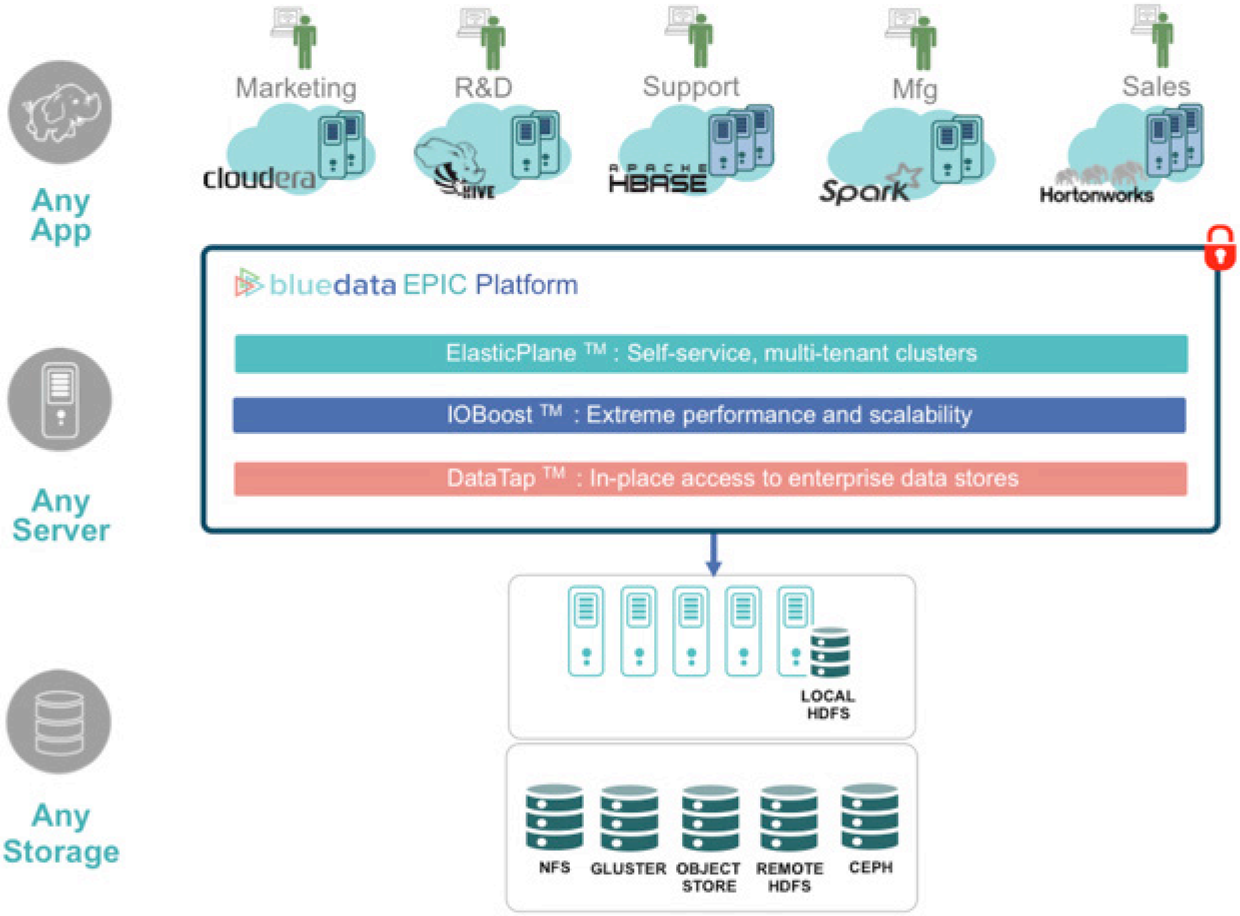
- **Elasticity** that permits the platform to dynamically adapt to changing workload requirements in the most cost-efficient manner.
- The ability to use **any distributed storage technology**.
- **Instant scalability** both up and down that lets the enterprise immediately respond to changing Big Data requirements.
- **Enterprise-grade security** for sensitive data because there is no need to copy and move it off the organization's servers.
- **I/O optimization** that provides all the benefits of running Hadoop on virtual servers 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 deploy and test applications in minutes**.
- An opportunity to **employ any Hadoop application**, not just a specific vendor's Hadoop distribution or a limited set of currently available ones.

## BlueData Platform

**Any Application**
Enterprises can run multiple Hadoop distributions and any application simultaneously on any piece of hardware.

**Any Server**
Vendor-agnostic hardware allows enterprises to select server hardware that best matches their needs as well as mix and match various server generations.

**Any Storage**
Enterprises can access Big Data from any storage device, including those running Gluster, HDFS, Swift and NFS.



# 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 fierce as it, there is no room for false starts or mistakes. Success in this new technology 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 the pioneer in big data private cloud. The company is democratizing big data by streamlining and simplifying big data infrastructure and eliminating complexity as a barrier to adoption. With its 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. Based in Mountain View, CA, BlueData is founded by a highly experienced team from VMware, Akamai, Intel, and SGI and backed by industry luminaries from Silicon Valley.

## A Hadoop Deployment Checklist

Enterprises have several infrastructure options from which to choose when it comes to running Big Data jobs. Those options include:

- » Physical in-house Hadoop cluster(s),
- » Hadoop in a public cloud as either:
  - – Hadoop Infrastructure as a Service (IaaS)
  - – Hadoop Platform as a Service (PaaS)
- » Virtual Hadoop clusters in an in-house private cloud with data anywhere, or
- » Some combination of the above.

Before making a Big Data infrastructure investment, CIOs 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?** Most infrastructures require you to copy data into a HDFS file system before your Big Data applications can access it. This time-consuming, expensive process results in the creation of multiple copies of data. Choose infrastructure 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 servers.
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 an infrastructure 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. Ensure your infrastructure platform simplifies the complexity of Big Data management by providing end users with self-service provisioning and managing of their Big Data jobs.
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. Ensure your infrastructure does not lock you into a specific vendor's Hadoop distribution or a limited set of currently available applications.

bluedata