Breaking Down Big Data Barriers with BlueData*

Accelerate deployment of big data clusters and get business results quickly with BlueData's EPIC* platform

This solution brief describes how to solve business challenges through investment in innovative technologies.

If you are responsible for...

• **Business strategy:** You will better understand how a Big-Data-as-a-Service solution will enable you to successfully meet your business outcomes.

• **Technology decisions:** You will learn how a Big-Data-as-a-Service solution works to deliver IT and business value.

**Executive Summary**

Big data analytics aids business insight. Enterprise deployment of big data ecosystem components including Apache Hadoop* and Spark* are typically done on physical infrastructure, and it commonly takes weeks or months to deploy and optimize a new physical cluster. BlueData's EPIC* software platform eliminates conventional big data bottlenecks using Docker* container technology and is optimized to run on Intel® architecture. BlueData EPIC provides a simple web-based interface that assists with cluster provisioning and policy management for multiple tenants using a shared infrastructure. EPIC also separates compute and storage resources for independent scaling, whether those resources are on-premises or in the public cloud (see Figure 2). Data science teams can initiate their own Hadoop and Spark clusters in minutes, freeing up IT for other tasks.

Close collaboration between BlueData and Intel sped EPIC product development and has resulted in performance levels comparable to or better than bare-metal clusters. The two companies worked together on software optimizations for overall Intel® silicon use, data packet handling, and the Hadoop YARN* resource manager.

**BlueData EPIC* Platform for On-Demand Big Data Clusters**

Figure 1. BlueData's EPIC* platform sits atop a shared bed of compute and storage resources, either local or in the public cloud, and helps speed up the process of deploying big data clusters on that infrastructure.
**Business Challenge: Big Data is Too Hard!**

Big data continues to attract enterprises with the promise of turning large volumes of structured and unstructured data into actionable business intelligence. However, Gartner states that only 15 percent of businesses deploying big data projects see those projects reach production. A 2017 survey from NewVantage Partners found that about 30 percent of companies cite “technology resistance or lack of understanding” as an impediment to big data adoption. Finding IT staff with sufficient expertise to navigate complex Apache Hadoop® implementations remains challenging, as does the typical time required to deploy a new workgroup cluster. Most big data deployments use at least one Hadoop® distribution, and each variety requires more IT knowledge. Also, new big data technologies like Spark® and Kafka® require additional expertise.

Other challenges also stand in the way of companies realizing big data business value:

- **Compatibility.** Most big data deployments do not start from scratch. Enterprises need to meld new analysis systems into legacy investments, raising interoperability issues.

- **Security.** Centralizing data and increasing accessibility requires safeguards such as skilled IT and tools to help enable security in big data environments.

- **Financial.** In 2013, Wikibon found that most enterprises realize a return of only 55 cents for every dollar spent on big data. Again, a “lack of skilled big data practitioners” was found to be one of the top reasons. To improve on this, enterprises must find either more expert staff or tools able to lower technical barriers to entry.

- **Efficient utilization.** Often, each workgroup sets up its own big data cluster. This results in organization-wide cluster sprawl, rampant data copying and storage waste between clusters, and low per-cluster resource utilization. Meanwhile, sprawl incurs higher infrastructure and IT costs.

There's no question that big data is big business—IDC expects the big data technology and services market to grow at a compounded annual growth rate of 26.4 percent—six times that of the overall IT market. However, to maximize the value of big data investments, production deployment must get faster and scale more easily.

**Solution Value: Simple, Fast, Cost-Effective**

BlueData’s EPIC platform simplifies Hadoop, Spark, and related big data application deployments with speed, agility, low cost, and comprehensive security. The platform encompasses on-premises and cloud-based resources to harness any storage and any server, anywhere. Because EPIC places each virtual cluster within its own set of Docker® containers, the platform renders Hadoop, Cloudera®, and Spark into on-demand services that can be utilized almost instantly.

BlueData’s EPIC platform can support a wide variety of big data configurations and use cases. Current EPIC users include finance, insurance, government, high-tech, healthcare, education, and retail. The ways in which organizations employ EPIC can be equally varied. A financial group might find great value in aggregating and analyzing disparate databases without bringing all that data into a centralized Hadoop Distributed File System® (HDFS). With EPIC, that source data can remain in their security-enabled, governed locations, saving time and lowering solution complexity. A retail group may need to act on real-time customer data for competitive advantage, using big data technologies like Spark and Kafka for streaming analytics with a NoSQL® database like Cassandra®.

With the recent inclusion of public cloud support, BlueData EPIC deployment models fall into four main categories:

- **Departmental on-premises.** This model is for smaller deployments and frequently applies in non-production situations, such as for development, testing, and quality assurance. By using EPIC to quickly create a sandbox environment for a new big data technology, organizations can explore new use cases or technology approaches.

- **On-premises multi-tenant.** This model applies to production deployments in a large enterprise that involves different groups of users and multiple types of data storage that are locally centralized within the organization. Overall, this model emphasizes operational efficiency, reliability, security, and scalability.

- **Public cloud multi-tenant.** In situations where data already resides in the public cloud (such as Amazon Web Services®), this model focuses on speed, agility, and security enablement when provisioning big data workloads for multiple user groups. With EPIC, users also have greater flexibility to use their big data tools of choice in the public cloud.

- **Hybrid: public cloud and on-premises.** Many enterprises want to maintain control of sensitive data on-site while placing non-sensitive data in the cloud.
EPIC decouples compute resources from storage. Using on-premises, industry-standard servers and parallelization optimizations, EPIC can deliver the agility and speed of a bare-metal deployment, even under multi-tenancy conditions. This was recently borne out in a benchmarking study between BlueData and Intel.1

Testing used 10, 20, and 50 Hadoop compute nodes with 10 TB of HDFS data. Both groups used identical hardware configurations, with Intel® Xeon® processor E5-2699 v3-based servers and Intel® Solid State Drives. Employing the industry-standard BigBench* benchmarking software, BlueData simulated 30 queries through an unmodified Hadoop ecosystem and found the performance in both configurations to be comparable. In fact, when testing 50 Hadoop nodes with 10 TB of data, EPIC outperformed bare metal by 2.33 percent.1

EPIC packs the complexity of cluster creation and maintenance into a surprisingly simple management interface. This simplicity extends to the public cloud, as well. With Amazon Web Services, for example, EPIC absorbs most of the service’s details, so there is less for users to learn and navigate. This lowers development costs and accelerates deployment while still giving administrators control over resource quotas, start/stop controls, and cost reporting integrations.

With BlueData EPIC, organizations have a quick, affordable way to deploy big data analytics for new groups and new initiatives. Groups can self-service their big data implementations, which takes strain off IT. The container-based architecture leveraging shared infrastructure reduces the need for data duplication and helps enable workgroups to scale deployments as needed without incurring unnecessary additional costs.

**Solution Architecture: Layers of Abstraction**

The BlueData EPIC software platform accelerates big data deployments using Docker container technology and shared infrastructure (see Figure 2). Each of these containers can be made available to different analysts and data science teams. The result is effectively Big-Data-as-a-Service running on shared infrastructure resources that can bridge to public cloud resources as needed.

The EPIC software stack features three innovative components:

- **ElasticPlane** helps administrators and end users to quickly create and customize clusters in a security-enabled multi-tenant architecture. With just a few mouse clicks they can deploy their choice of different big data frameworks, analytical tools, and applications. The ElasticPlane portal includes a policy engine for service-level agreement management and is the key ingredient in making self-service cluster provisioning simple.

- **IOBoost** is a memory caching and tiering system designed to improve application speed for big data workloads. It does this based on a set of assumptions about how big data applications access data. The acceleration realized by this caching counterbalances the latencies introduced through containerization and remote storage access, allowing EPIC to deliver performance on par with bare-metal environments where compute and storage are collocated.1

- **DataTap** allows clusters to access HDFS, Network File System* (NFS), and other on-premises enterprise storage as well as Amazon S3*. The service abstracts the HDFS protocol so that any Hadoop-based application can run unmodified and still gain access to data from non-HDFS storage systems.
**Conclusion**

The potential value of big data analytics lies in unlocking vast amounts of business intelligence. To date, enterprises have been able to realize some benefit from big data, but most feel they could do better. BlueData’s EPIC software platform is designed to deliver the tools and optimizations businesses need to achieve their big data goals. EPIC removes much of the expertise needed to launch Hadoop and Spark clusters while simultaneously reducing the amount of time needed to do so. Combining EPIC’s advantages with Intel® hardware designed to accelerate analytics operations gives enterprises fresh hope for their big data futures.

Find the solution that is right for your organization. Contact your Intel representative or visit intel.com/bigdata.

---

**Amazing Results from Big Data with Intel® Technology**

Intel® Xeon® processor-based platforms and other Intel® technology are designed to accelerate big data workloads. BlueData® proved this to be the case in its BigBench® testing,

1. Support for multi-terabyte storage workloads. The Intel® Xeon® processor E5 family specifically aids extra-large storage workloads with outsized execution resources, large memory capacity, advanced reliability features, and performance-accelerating optimizations such as Intel® Advanced Encryption Standard - New Instructions (Intel® AES-NI) and Intel® Advanced Vector Extensions (Intel® AVX).

2. High storage throughput. Today’s analytics platforms should look beyond SATA solid state drives (SSDs). The Intel® SSD Data Center Family for PCIe® with Non-Volatile Memory Express (NVMe) reduces signal- and code-based latencies experienced with legacy disk storage to help process massive data sets from multiple sources.

3. Next-generation network performance. Modern analytics conquers data silos, but insufficient network bandwidth will essentially turn an analytics server into its own island. The Intel® Ethernet Converged Network Adapter XL710 keeps data flowing between network nodes.

---

**Learn More**

You may also find the following resources useful:

- BlueData web site
- Intel Developer Forum
- Intel White Paper on BlueData Benchmark Testing