Software & Services Solution Brief

Big Data Lab Accelerator

Accelerate Hadoop and Spark deployment in a multi-tenant lab environment for dev/test/QA, evaluation of multiple Big Data distributions and tools, and other use cases. BlueData™ provides a turnkey solution with software and services to get up and running in two weeks.

SOLUTION HIGHLIGHTS

- Accelerate the deployment of your lab for Big Data analytics, with a turnkey multi-tenant environment for development, testing, and quality assurance.
- In two weeks, you'll have a ready-to-run environment for up to 30 virtual nodes of your preferred Hadoop distribution(s) and/or Spark.
- Increase business agility by empowering data scientists and analysts to spin up new clusters in a matter of minutes, with just a few mouse clicks.
- Deliver faster time-to-results with the ability to quickly and easily evaluate multiple distributions, versions, services / components, and BI / analytical tools.
- Enable multiple Hadoop or Spark clusters to share a single set of files, thereby eliminating data duplication and / or data movement.
- Limited time offer: discounted 1 year subscription of BlueData EPIC Enterprise software (5 server license) + professional services.

Big Data technologies like Hadoop and Spark are complex; there are multiple components, systems, and infrastructure resources required. These components are available as free open source software, or packaged and distributed by several commercial vendors. Either way, it can be time-consuming and challenging to evaluate and get these new environments deployed and operational – even in a lab for initial development, testing, and quality assurance.

If your organization is looking to set up a new Hadoop or Spark lab environment (e.g. for dev/test/QA, evaluation of multiple Big Data tools and technologies), there is a better way.

Get Started with a Big Data Lab for Dev/Test/QA

BlueData's mission is to make Big Data infrastructure easy. BlueData has developed patent-pending software innovations that are fundamentally changing the deployment model for Big Data – leveraging containers and virtualized infrastructure. The BlueData EPIC software platform is purpose-built to simplify and accelerate the infrastructure deployment for Hadoop, Spark, and related tools for Big Data analytics. We partner with the leading distribution, application, and infrastructure vendors in the Big Data market to make it easier, faster, and more cost-effective to get started.

Our new **Big Data Lab Accelerator** solution provides the software and professional services you need to accelerate the deployment of an on-premises multi-tenant Big Data lab environment in two weeks. As part of this deployment, we also work with you to implement a few key use cases for your new Big Data lab.

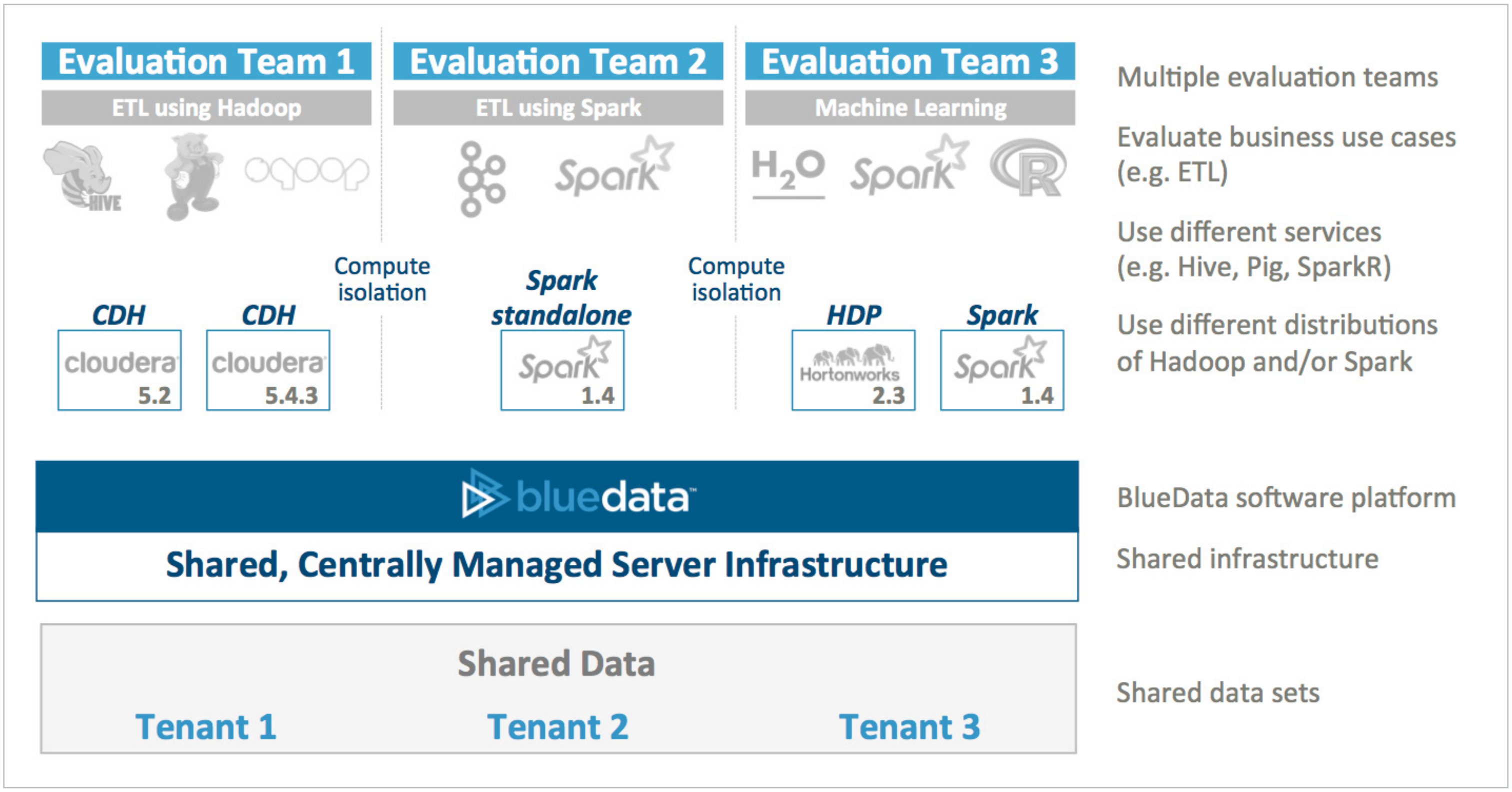
Target Audience

- Organizations looking to get started with Big Data, setting up a lab for Hadoop and/or Spark
- Big data architects, enterprise architects, data engineers, data analysts, data scientists, and IT infrastructure teams

Solution Architecture

With the Big Data Lab Accelerator, your organization will have a ready-to-run lab environment to evaluate multiple distributions, services, and tools on a shared, cost-effective infrastructure for multiple tenants.

The figure below illustrates an example environment for multiple different teams (tenants) – each with different use cases, distributions, services, and tools to evaluate – running on the BlueData EPIC software platform and shared infrastructure.



As shown in the above illustration, each tenant leverages shared virtualized infrastructure, with the ability to tap into shared data sets. Yet each team can run their own independent evaluation with different use cases and tools; the architecture provides secure and logical separation with compute isolation between each tenant. Some of the benefits include:

- **Self-service agility.** Users can spin up or spin down instant virtual clusters of Hadoop or Spark (with different Hadoop distributions and / or versions) – on-demand, within minutes.
- **Lower cost.** Your organization can save up to 70% on infrastructure, with the ability to run up to 30 virtual nodes on five shared physical servers or virtual machines.
- **No need to copy data.** Enable multiple Hadoop or Spark clusters to share a single set of files, thereby eliminating data duplication and / or data movement.
- **Faster time to results.** Within 2 weeks, you'll have a shared, multi-tenant DevOps lab environment for Big Data teams – promoting faster development / testing / QA.

To learn more about the BlueData EPIC software platform, visit www.bluedata.com

Scope and Methodology

Deployment Services (2 days)

- Deployment of BlueData EPIC software on up to 5 physical servers or 5 virtual machines (1 day)
- Spin up multiple Hadoop and Spark clusters (e.g. up to 30 virtual nodes) of Hortonworks, Cloudera, and/or Spark stand-alone (1 day). Implementation may also include one or more related business intelligence / analytics / ETL tools (e.g. AtScale, Platfora, Splunk / Hunk, Tableau)
- Hands-on exercises

Hadoop / Spark Key Concepts (1 day)

- Focus on hands-on-exercises for the three key elements of a data pipeline
- Data ingestion (Sqoop, Flume, Kafka)
- Data processing (MapReduce, Spark, Pig)
- Data serving (e.g. Hue, SQL, HBase)

Use Case Discovery Workshop (2 days)

- Choose two candidate use cases from the sample list on the following page or create a scoped use case during workshop (1 day)
- Identify and copy data to shared storage system. Alternatively, leverage public data sets associated with each use case (1 day)

Use Case Implementation (5 days)

- Divide into teams or tenants as necessary
- Create data pipeline for use case on a specific Hadoop or Spark cluster
- Document the use case implementation and learnings

Total elapsed time: 10 working days

For pricing questions or additional information, contact sales@bluedata.com

Sample Hadoop Use Cases

Offload ETL processing from existing RDBMS (e.g. Oracle, Teradata)

- Sqoop the data from RDBMS
- Write a M/R or Pig job for data transformation
- Create a Hive or Impala table
- Use Tableau or other business intelligence tool to query the data

Storage and processing of semi-structured application or machine logs

- Use Flume to collect application logs from log files
- Write data to HDFS or NFS
- Write a data transformation job to split the log data
- Aggregate the log data to generate counts of errors, warnings etc. (e.g. by time of day)

Create a secure data lake for long term storage of structured and unstructured data, with the ability to search/index data

- Use Sqoop and Flume to copy different types of data (structured, semi-structured)
- Deploy open source tool (e.g. Solr) or commercial tool (e.g. Splunk) to search the data
- Enforce security at the compute (Solr, Hive) and storage (HDFS) layers, using Active Directory and Kerberos

Create a single source of truth for customer data in an enterprise

- Ingest customer data from ERP customer master on an on-going basis using Sqoop
- Ingest customer data from Sales and Marketing on an on-going basis using Sqoop
- Ingest support customer data
- Run a de-dup workflow periodically using machine learning algorithms
- Create golden records

Sample Spark Use Cases

Real-time and batch processing of streaming data

- Deploy Kafka, Spark clusters
- Read data from Kafka queues in parallel
- Process using Spark Streaming
- Persist in HDFS
- Run queries on data in HDFS

Advanced analytics and machine learning on data sets

- Read data from one or more data sources into Spark
- Join and transform as needed for a machine learning pipeline and run ML algorithms
- Use the resulting model for predictive analytics