bluedata™

Real-Time Pipeline Accelerator

Get started with Spark Streaming, Kafka, and Cassandra for real-time data analytics. BlueData™ provides a turnkey solution with software and services to accelerate your deployment and build end-to-end real-time data pipelines within minutes.

SOLUTION HIGHLIGHTS

- Accelerate the deployment of your lab for real-time analytics with Spark Streaming, Kafka, and Cassandra.
- Build real-time data pipelines with a turnkey environment for rapid prototyping, development, testing, and quality assurance.
- Provide a standardized user experience for creating consistent and repeatable pipelines, with support for various stages of the application lifecycle.
- Improve agility through self-service, empowering data scientists to spin up new clusters in a matter of minutes—with just a few mouse clicks.
- Increase developer productivity with a multi-tenant sandbox for real-time analytics, including Zeppelin notebooks and other JDBC-supported tools.
- 1 year subscription of BlueData EPIC Enterprise software for up to 60 physical cores (approximately 5 nodes) + professional services.

Speed is a key driver for an increasing number of analytics use cases, ranging from fraud detection for financial transactions to Internet of Things (IoT) monitoring with sensor-generated data. Immediate analysis of these new data streams in real-time can bring tremendous value—whether in delivering competitive business advantage, averting potential crises, or creating new revenue opportunities.

If your organization wants to enable rapid prototyping and development for real-time analytics, there is now an on-premises solution to help you get started quickly and easily.

Spark Streaming + Kafka + Cassandra

BlueData makes it easy to deploy Spark infrastructure and applications on-premises. The BlueData EPIC™ software platform is purpose-built to simplify and accelerate the deployment of Spark, Hadoop, and other tools for Big Data analytics—leveraging Docker containers and virtualized infrastructure.

Our new **Real-Time Pipeline Accelerator** solution provides the software and professional services you need for building data pipelines in a multi-tenant environment for Spark Streaming, Kafka, and Cassandra. With this solution, your data scientists will be able to create integrated pipelines to capture, analyze (model/score), and store real-time data streams within a matter of minutes.

Now your data scientists and developers can focus on their use cases and pipelines, without worrying about the infrastructure complexities of technologies like Spark, Kafka, and Cassandra. And as your uses cases mature and expand over time, you can use the BlueData EPIC platform to extend your pipelines with complementary applications and frameworks.

Target Audience

- Organizations looking to get started with real-time data pipelines
- Organizations with existing data pipelines using Spark Streaming, Kafka, and Cassandra that need a sandbox environment for Dev/QA/UAT
- Data scientists, analysts, engineers, architects, and IT infrastructure teams

What is a Real-Time Data Pipeline?

Today, there are continuous streams of data in large volumes being generated from financial markets, sensors, machine logs, social media, mobile applications, and many other sources. Rather than staging this data and analyzing it after it's been stored, it's becoming increasingly important to analyze the data in real-time as events are happening—to make instant decisions and take immediate action. This data is often perishable and may lose its operational value in a very short time frame. Speed is of the essence.

In the general sense, a data pipeline consists of all the steps necessary to capture, prepare, and process your data for analysis. Building data pipelines for real-time analysis requires specific technologies and methodologies:

- Real-time means you can't afford to wait for days, hours, or even minutes before running your analysis and generating insights–the tools need to support this fundamental requirement.
- The technologies and frameworks for real-time analytics are different from existing enterprise systems and batch-oriented data processing frameworks.
- There are multiple components (both software and infrastructure) needed to capture and hold streaming data, process and analyze multiple sets of streams in small batches, and store these valuable results for continuous analysis.
- It is complex and time-consuming to assemble all the infrastructure and software required, and most organizations lack the skills to deploy and wire together all of these components.

Implementing the Real-Time Trinity

For data scientists and developers working with real-time data pipelines, the stack of Spark-Kafka-Cassandra has quickly emerged as the best place to start. This new trinity delivers on key requirements for real-time data analytics:

- **Spark:** a fast in-memory data processing engine, and the fastest growing Apache open source technology. Spark Streaming is an extension of the core Spark API; it allows integration of real-time data from disparate event streams.
- **Kafka:** a messaging system to capture and publish streams of data. With Spark you can ingest data from Kafka, filter that stream down to a smaller data set, augment the data, and then push that refined data set to a persistent data store.
- **Cassandra:** this data needs to be written to a scalable and resilient operational database like Cassandra for persistence, easy application development, and real-time analytics.

These open source frameworks are powerful and well-suited for the requirements of building real-time data pipelines. However, it may take weeks and even months for your team to get ramped up and started. For example, you will likely need to train at least one or two team members on Spark, Kafka, and Cassandra (typically one week of training or more, for each framework). You will need to build pipeline integrations between these different frameworks and test them internally on your infrastructure (requiring at least another couple weeks or more).

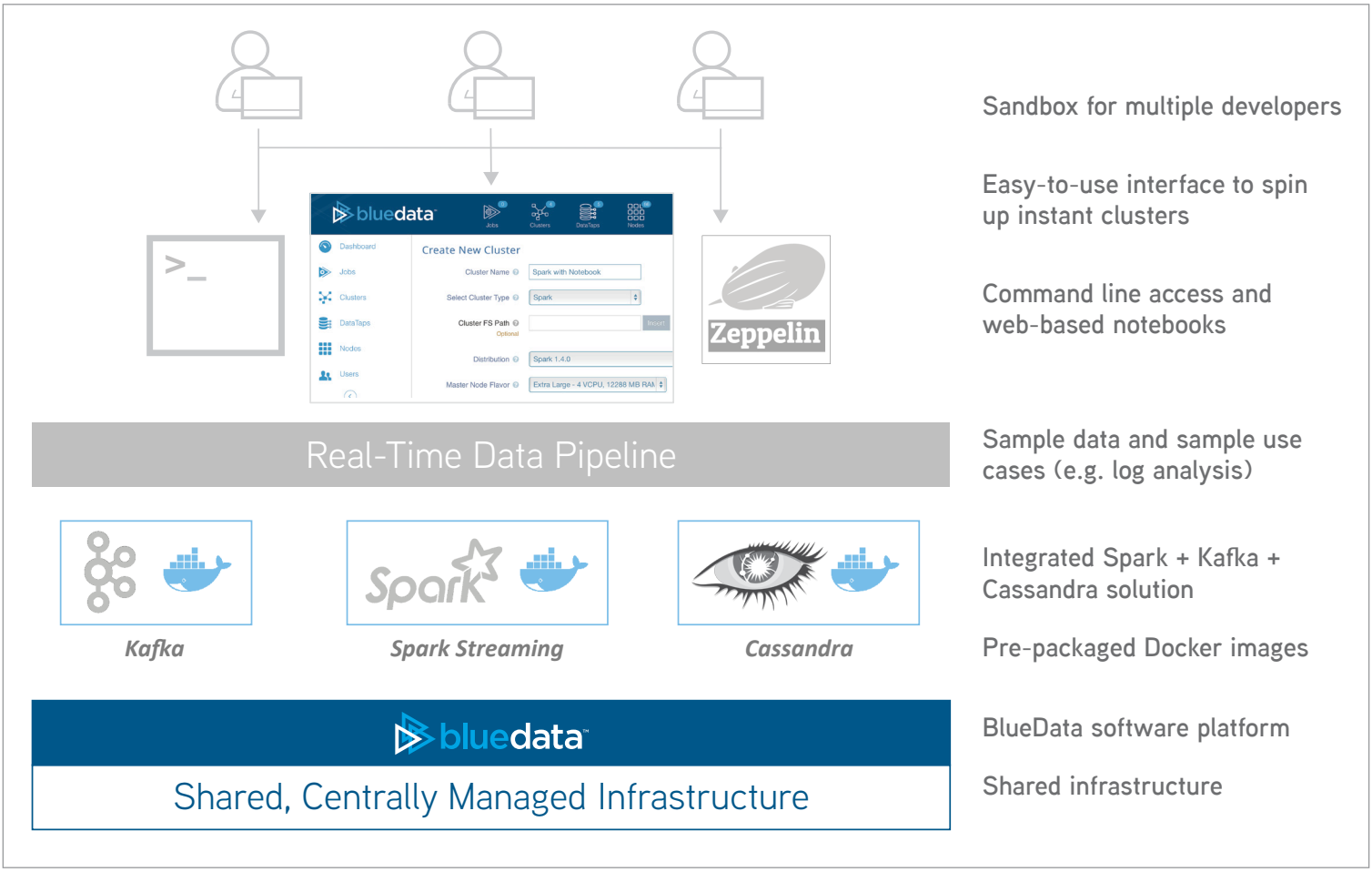
Creating a Spark-Kafka-Cassandra lab for multiple data scientists and developers—to support rapid prototyping, development, testing, and quality assurance—can be a challenging and expensive initiative. And as you begin to add more use cases and users, you will need to expand the infrastructure, provision more hardware, and integrate more tools.

BlueData's **Real-Time Pipeline Accelerator** solution is designed to address these challenges—making it simple and easy to get up and running with this new stack for real-time analytics.

Real-Time Pipeline Accelerator

With the new **Real-Time Pipeline Accelerator** solution from BlueData, your organization will have a ready-to-run, multi-tenant lab environment for Spark Streaming, Kafka, and Cassandra.

The figure below illustrates an example sandbox environment for multiple data scientists and developers (tenants) with the BlueData EPIC software platform running on shared, cost-effective infrastructure. BlueData provides an easy-to-use interface and out-of-the-box support for Zeppelin notebooks, command line access, and other JDBC tools–along with sample data and use cases for real-time pipelines. Developers and data scientists can self-provision the key components for their real-time data pipelines in a matter of minutes–using pre-packaged Docker images for Spark, Kafka, and Cassandra.



The key components of the Real-Time Pipeline Accelerator solution include:

- 1 year subscription license for BlueData EPIC Enterprise up to 60 physical cores (approximately 5 nodes)
- Pre-configured Docker images for Spark, Kafka, and Cassandra
- UI-based cluster manager to build real-time data pipelines with Spark Streaming, Kafka and Cassandra
- Quickstart examples and sample documentation for two end-to-end data pipelines (e.g. log analysis)
- 5 days of remote professional services and consulting from BlueData experts

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

Solution Benefits

BlueData's **Real-Time Pipeline Accelerator** solution is designed to accelerate the deployment of data pipelines for real-time analytics. Some of the benefits include:

- **Self-service agility.** Users can spin up or spin down instant virtual clusters of Spark, Kafka, and Cassandra–on-demand, within minutes.
- **Improved productivity.** Developers can start being productive immediately and focus on their use cases. With web-based Zeppelin notebooks, developers can easily share their datasets and analysis with other users.
- **Consistent and reproducible pipelines.** A standardized user experience enables the creation of consistent and repeatable data pipelines, with support for various stages of the application lifecycle.
- **Sample data pipelines.** Sample datasets and use cases are available for immediate use. BlueData experts will help deploy the Spark-Kafka-Cassandra stack and implement sample data pipelines.
- **Lower cost.** Your organization can save up to 75% on server and storage infrastructure, with the ability to run multiple virtual nodes on shared infrastructure.
- **Scalable.** With the solution's multi-tenant architecture, it's easy to scale your lab environment and add more users or infrastructure resources as your deployment grows.
- **Extensible.** As your use cases mature and expand beyond the Spark-Kafka-Cassandra stack, BlueData supports complementary applications and frameworks to extend your pipelines and integrate additional tools.
- **No lock-in.** The solution stitches together open source components in a loosely coupled fashion, so there is no vendor lock-in. The BlueData EPIC Platform is highly configurable, and designed to support a wide variety of different Big Data use cases and applications.

With this solution, your organization will have a ready-to-run lab environment for rapid prototyping, development, testing, and quality assurance with Spark Streaming, Kafka, and Cassandra. With help from the BlueData team, you'll also have two end-to-end real-time data pipelines as a starting point.

And with your Enterprise license of the BlueData EPIC software platform, you'll have a multi-tenant infrastructure platform that can be easily extended to additional Big Data uses cases and applications—for both data in motion and data at rest—with support for Spark and Hadoop as well as leading business intelligence, analytics, visualization, and data preparation tools.

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