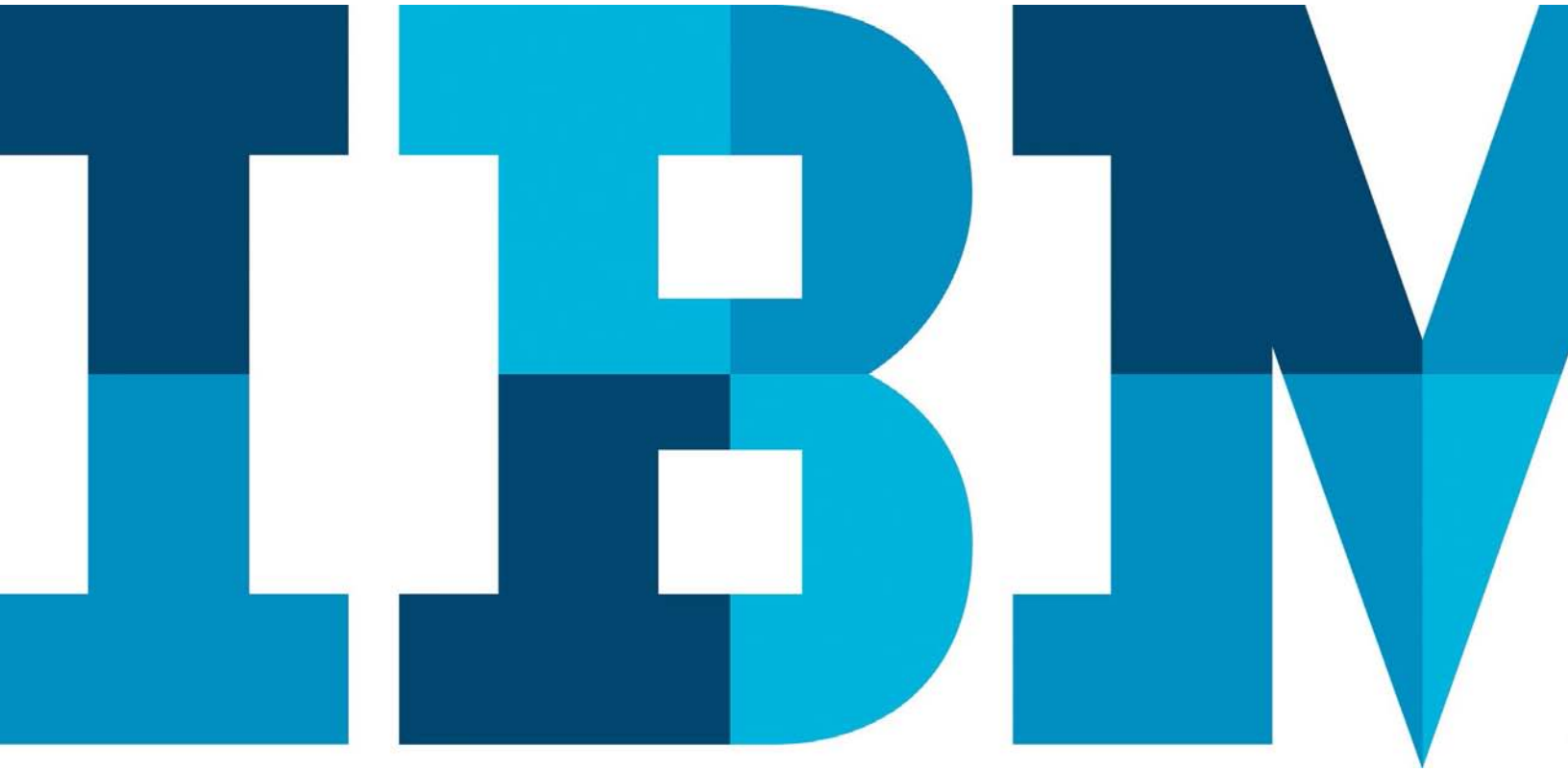
November 2016

An integrated, enterprise-grade solution for deploying Apache Spark

Simplify deployment and boost performance with IBM Spectrum Conductor with Spark



IBM

Highlights

- Improved time to results through efficient resource scheduling and shared infrastructure that maximizes resource utilization
- Spark multitenancy allows running of several instances and different versions of Apache Spark and other applications
- Improved performance and efficiency with a proven solution for granular and dynamic resource allocation
- Simplified management through a consolidated framework for Spark deployment, monitoring and reporting
- A complete, solution for deploying Spark in an organizational environment with enterprise-class security

Ever-increasing competitive and regulatory pressures mean capitalizing quickly on all your information is more important than ever. That's why accelerating time to results is a key driver for the increased adoption of Apache Spark as a cluster computing system for businesses that rely on big data analytics.

Apache Spark offers higher-performance processing compared to MapReduce. It can run in memory, allowing organizations to make critical decisions faster, as they process larger amounts of data. Spark is also developer-friendly, with libraries, simplified application programming interfaces (APIs) for a variety of programming languages and a rich set of high-level tools to support big data analytics. However, adopting Spark in a production environment can present significant challenges.

Overcome the challenges of Spark deployment and management

Organizations already invested in frameworks, such as MapReduce, face additional investment in new expertise, tools and workflows, as they integrate Spark into their existing environment.

Additionally, different groups or departments in larger organizations may set up separate compute clusters to host individual Spark environments. This ad hoc, siloed growth presents problems for the enterprise, such as inefficient resource use, management challenges and security issues that impede the move to a production environment. The rapid pace of Spark advancement can also put pressure on IT resources. Updates to open source Apache Spark are quite frequent and can result in various groups running different versions of the framework.

IBM® Spectrum Conductor with Spark is designed to address these multiple challenges.

At a glance: IBM Spectrum Conductor with Spark

- Integrated solution with Spark distribution and resource management
- Consolidated framework for simplified deployment and monitoring
- Highly efficient resource scheduling for improved time to results
- Increased resource utilization, resulting in better cost containment
- Elimination of resource silos tied to different instances and versions of Spark, and other frameworks
- Graphics processing unit (GPU) floating and vector processing support for compute-intensive tasks
- Sharing of cached and persisted Resilient Distributed Datasets (RDDs) across users to avoid reloading or recomputing previous results

Simplify deployment and boost performance for Spark

IBM Spectrum Conductor with Spark is a multitenant enterprise solution for Apache Spark. It also supports integration with other frameworks and allows organizations to deploy Spark efficiently and effectively. The solution supports multiple instances and different versions of Spark and other applications, increases performance and scale, and maximizes resource usage. In addition, it eliminates silos of resources that would otherwise be inefficiently tied to separate implementations.

Organizations are looking to move to solutions that optimize storage, analysis and protection of their information assets. IBM Spectrum Conductor with Spark is designed specifically to help users deploy Apache Spark in a production environment.

Unlike other offerings that require piecemeal assembly of components, it is an integrated solution for data analysis. It incorporates a Spark distribution, augmented by technology for granular and dynamic resource allocation that has been proven in many demanding customer environments to improve performance and efficiency.

Apache Spark requires a separate resource manager. Accordingly, IBM Spectrum Conductor with Spark includes a framework for workload management, monitoring and reporting. The solution incorporates dynamic resource allocation on demand as well as policy-based control for application service-level agreements (SLAs) and infrastructure efficiency.

For storage management, IBM Spectrum Conductor with Spark can be combined with IBM Spectrum Scale™. Unlike the open source Hadoop Distributed File System (HDFS), IBM Spectrum Scale is Portable Operating System Interface (POSIX)-compliant and provides significant advantages in storage efficiency. Users may optionally use HDFS if they prefer. The entire solution is backed by IBM services and support.

Improve time to results and cut costs

By providing advanced service orchestration and sophisticated resource-sharing capabilities for Apache Spark and other frameworks, IBM Spectrum Conductor with Spark enables individual applications to take full advantage of available resources. A proven, efficient resource scheduler offers fine-grain dynamic resource allocation, helping to deliver superior application performance and faster response to business demands. Running multiple Spark instances on shared infrastructure helps maximize resource utilization and contain infrastructure costs.

IBM Spectrum Conductor with Spark also provides GPU support to utilize the full power of GPU floating and vector processing in compute-intensive tasks. In addition, cached or persisted RDDs can be shared across users to avoid reloading or recomputing previous results. These elements combine to provide the fastest possible time to results, while minimizing expenditure on computing infrastructure.

Increase management efficiency and consistently meet SLAs

The ability to simultaneously deploy different versions of Spark in a shared environment also helps users manage Spark lifecycles in the face of frequent updates. Different groups can run various versions of Spark, and it's not necessary for all Spark instances to be upgraded in lockstep. Assets can be more efficiently monitored to consistently meet service-level objectives, and users can quickly bring new resources online as needed.

Deliver enterprise-class security with role-based access control

IBM Spectrum Conductor with Spark also provides role-based access control (RBAC) to help protect Spark environments. Most organizations require different users to manage diverse system administrative duties. RBAC enables organizations to assign permissions to perform specific tasks according to the role of each user, minimizing the opportunity for any one user to cause accidental or malicious damage to the system.

Gain key advantages versus open source schedulers

For organizations looking to realize faster time to insight for big data analytics, IBM Spectrum Conductor with Spark provides advantages over open source schedulers, such as YARN and Mesos.

Advantages compared to Spark on YARN

- Time-based scheduling policies
- Centralized graphical user interface (GUI) for managing resource plans
- Dynamic updates of resource plans
- Preemption with sophisticated reclaim logic
- Does not require Hadoop distribution complexity
- Proven multitenancy with SLA management and quality of service
- Faster time to results through 41 percent greater throughput performance[1]

Advantages compared to Spark on Mesos

- Technology maturity with proven results at scale
- Preemption with sophisticated reclaim logic
- Time-based scheduling policies
- Centralized GUI for managing resource plans
- Dynamic updates of resource plans
- Proven multitenancy with SLA management and quality of service
- Advanced scheduling policies
- Faster time to results through 58 percent greater throughput performance[1]
- Superior predictability of job completion times[2]

Why IBM?

IBM Spectrum Computing offers a comprehensive portfolio of software-defined infrastructure solutions designed to help your organization deliver IT services in the most efficient way possible. These offerings optimize resource utilization to speed time to results and reduce costs. They also help maximize the potential of your infrastructure to accelerate your analytics, high-performance computing (HPC), Apache Hadoop, Spark and cloud-native applications at any scale. Additionally, they extract insight from your data and get higher-quality products to market faster.

Whether deployed in a data center or on the cloud, IBM Spectrum Computing solutions fuel product development, critical business decisions and breakthrough insights in a number of industries. Industries that can benefit include financial services, manufacturing, digital media, oil and gas, life sciences, government, research and education. From designing Formula One race cars to credit risk analysis, organizations in a wide variety of industries are using IBM Spectrum Computing as a foundation for software defined infrastructure solutions for big data, analytics, HPC and cloud to improve business results.

For more information

To learn more about IBM Spectrum Conductor with Spark, contact your IBM representative or IBM Business Partner, or visit:

- **ibm.com**/systems/platformcomputing/products/conductor/
- **ibm.com**/software-defined-infrastructure

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

IBM Systems
Route 100
Somers, NY 10589

Produced in the United States of America
November 2016

[1] "Spark Resource Manager Comparison of IBM Platform Conductor for Spark, Apache YARN and Apache Mesos – Phase 1," *STAC*. https://stacresearch.com/IBM160229

[2] In independently audited benchmark tests, the longest Spark on Mesos job took 45 times longer to complete than the longest IBM Spectrum Conductor with Spark job: 4178 seconds versus 92 seconds. https://stacresearch.com/IBM160229

Please Recycle