FREQUENTLY ASKED QUESTIONS

VMware vSphere Big Data Extensions

Extending the vSphere Platform to Support Big Data and Hadoop

Q: What is VMware vSphere® Big Data Extensions?

A: VMware vSphere® Big Data Extensions, or BDE, is a feature within vSphere to support Big Data and Apache Hadoop workloads. BDE provides an integrated set of management tools to help enterprises deploy, run and manage Hadoop on a common virtual infrastructure.

Q: Which Hadoop distributions does vSphere® Big Data Extensions support?

A: VMware has built BDE to support all major Hadoop distributions including Apache Hadoop, Cloudera, Pivotal, Hortonworks and MapR. Associated Apache Hadoop projects such as Pig, Hive and HBase are also supported.

Q: Is it possible to deploy multiple Hadoop distributions and/or versions with vSphere® Big Data Extensions?

A: Yes, customers can easily upload supported distributions of their choice and configure Big Data Extensions to deploy their preferred distributions and/or versions.

Q: Does vSphere® Big Data Extensions support features uniformly across Hadoop distributions?

Big Data Extensions provides differing levels of feature support depending on the Hadoop distribution and version you configure for use. Please refer to the VMware vSphere® Big Data Extensions Administrators and Users Guide for a detailed feature support matrix by distribution.

Q: What is Project Serengeti? How is it related to vSphere® Big Data Extensions?

A: Project Serengeti is an open source project initiated by VMware to automate the deployment and management of Apache Hadoop and HBase on virtual environments such as vSphere. vSphere® Big Data Extensions runs on top of Project Serengeti and is the commercial/supported version of the open source project; it also provides additional features like a graphical user interface integrated with the vSphere® Web Client, runtime automation, and elastic scaling. For additional details, please refer to the VMware vSphere® Big Data Extensions Administrators and Users Guide.

Q: What is the core value proposition of vSphere® Big Data Extensions?

A: vSphere® Big Data Extensions has three core value propositions:

Achieve Operational Simplicity with Performance: BDE automates the configuration and deployment of Hadoop clusters; IT departments can provide self-service tools (with vCloud Automation Center) that empower data scientists and analysts. Also, benchmark results show that virtualized Hadoop performs comparably with respect to physical configuration.

Maximize Resource Utilization on New or Existing Hardware: Big Data Extensions allow IT departments to lower the total cost of ownership by maximizing resource utilization on new or existing infrastructure. Elastic scaling, multi-tenancy and mixed workloads are the primary drivers of improved efficiency.

Architect Scalable and Flexible Big Data Platforms: Big Data Extensions is designed to support multiple Hadoop distributions and hardware architectures. BDE enables IT administrators to define a Big Data platform that can grow and scale effectively within the enterprise.

Q: How can I get vSphere® Big Data Extensions? Does VMware support it?

A: vSphere® Big Data Extensions is available as a download-able virtual appliance with a plug-in to the VMware vCenter Server™. Customers with current entitlements to vSphere Enterprise or Enterprise Plus licenses will receive commercial support for BDE.

Q: Is vSphere BDE a standalone product?

A: vSphere BDE is not currently offered as a separate product.

Q: Are there any pre-requisites/dependencies for vSphere® Big Data Extensions?

A: vSphere® Big Data Extensions works with vSphere 5.0 (or later) Enterprise or Enterprise Plus. The Big Data Extensions graphical user interface is only supported when using vSphere Web Client 5.1 and later. If you install Big Data Extensions on vSphere 5.0, all administrative tasks must be performed using the command-line interface.

Q: Are there performance limitations for running Hadoop workloads on vSphere vs. physical?

A: The performance of Hadoop workloads running on vSphere is close to that of native when using a single virtual machine per host. Improvements in elapsed time of up to 13% can be achieved by partitioning each host into two or four virtual machines, resulting in competitive or even better than native performance. For details, please refer to the Virtualized

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Copyright © 2010 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at http://www.vmware.com/go/patents. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.

VMware vSphere Big Data Extensions

Extending the vSphere Platform to Support Big Data and Hadoop

Hadoop Performance with VMware vSphere® 5.1 white paper.

Q: Does vSphere® Big Data Extensions work with other Hadoop Distributor Management Tools?

A: Yes, vSphere® Big Data Extensions is integrated with Cloudera Manager and Hortonworks Ambari, enabling these tools to perform installation and subsequent monitoring.

Q: Does vSphere® Big Data Extensions allow creation of HBase only cluster?

A: Yes, vSphere® Big Data Extensions allows customers to create HBase only clusters using existing Apache based Hadoop HDFS and Isilon.

Q: Does YARN or Hadoop 2.0 replace some of the functionality of vSphere® Big Data Extensions?

A: No. Hadoop 2.0, also known as Yet Another Resource Negotiator (YARN) is the newest generation of the Hadoop technology that is in popular use today for highly distributed processing and management of big data. YARN is now shipped by the Hadoop distributors as part of their Hadoop 2.x distributions. YARN changes the architecture that was inherent in Hadoop 1.0 in order to allow the system to scale to new levels and to assign responsibilities more clearly to different components. Looking deeper into the functionality that YARN offers, it is clear that there are many good reasons for virtualizing it. YARN and vSphere technologies are complementary and they serve mutually beneficial purposes in building big data clusters.

The architecture underlying Hadoop 2.0 has changed from the original MapReduce-centric one. This was done in order to allow further scalability as well as to accommodate different application designs. Those applications range from those that are interactive (Tez), in-memory (Spark), streaming (Storm), graphing (Giraph), and others to the familiar batch-style MapReduce applications where Hadoop started. The need for virtualizing these applications and the Hadoop infrastructure has not changed. Different communities require different Hadoop clusters to co-exist on the same sets of hardware.

SECTION I – Product

Q: What are the main features supported by vSphere® Big Data Extensions?

A: Main features include:

vCenter integration to rapidly deploy Hadoop clusters

- Deploy clusters with HDFS, MapReduce, HBase, Pig, and Hive Server
- Automate the deployment and scaling of Hadoop clusters

Hadoop-as-a-Service

- Enable self-service provision of Hadoop clusters in the private cloud with vCloud Automation Center
- Remove dependency and potential bottleneck associated with IT infrastructure management

Management Tool Integration

- Use BDE to setup infrastructure and leverage Hadoop distributor management tools to perform installation and subsequent monitoring
- Integration with Cloudera Managera and Hortonworks Ambari

Elastic Scaling and True Multi-tenancy

- Elastically scale compute and data separately
- Preserve data locality to improve performance

Architectural Flexibility

- Gain platform flexibility with support from major Hadoop distributions
- Select from hybrid, local storage, and shared storage options

Support for Major Hadoop Distributions: Big Data Extensions includes support for Apache Hadoop, Cloudera, , Hortonworks, MapR and Pivotal. HBase, Pig, and Hive are also supported. BDE also supports Hadoop 2.0 and YARN.

Distribution	Supported Versions
Apache Hadoop	1.2.1
Apache BigTop	0.8
Cloudera	5.1, 5.0, 4.7
Hortonworks	2.1, 2.0, 1.3
MapR	3.1, 3.0
Pivotal	2.1, 2.0

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Copyright © 2010 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at http://www.vmware.com/go/patents. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.

VMware vSphere Big Data Extensions

Extending the vSphere Platform to Support Big Data and Hadoop

Q: Is vSphere® Big Data Extensions a software solution or a service operated by VMware?

A: vSphere® Big Data Extensions is not a service. It is a software solution in the form of a downloadable virtual appliance that extends the benefits of virtualization to Hadoop.

Q: How do I get support for vSphere® Big Data Extensions?

A: vSphere® Big Data Extensions is GA and is fully supported by VMware. All support requests should be made directly to VMware Global Support.

Q: When will customers be able to try vSphere® Big Data Extensions?

A: vSphere® Big Data Extensions is GA and readily available for download. Customers with entitlements to vSphere Enterprise or Enterprise Plus licenses will receive commercial support for BDE.

SECTION II – Relationship with Other Products

Q: Is it possible to provide Hadoop-as-a-Service using vSphere® Big Data Extensions and vCloud Automation Center?

A: Yes, customers can use vSphere® Big Data Extensions and vCloud Automation Center 6.0 to provide Hadoop-as-a-Service for end users. An integrated VMware vCloud Automation Center and BDE solution is available on the VMware Solutions Exchange, along with technical resources and user guide.

Q: Does vSphere® Big Data Extensions require vSphere?

A: Yes, Big Data Extensions requires an Enterprise or Enterprise Plus edition of vSphere 5.0 or later.

Q: Can vSphere® Big Data Extensions be used in conjunction with vSphere Storage DRS?

A: No. Do not use BDE in conjunction with vSphere Storage DRS as it will disrupt the placement policies of your Big Data cluster virtual machines.

SECTION III – Technology

Q: Does vSphere® Big Data Extensions offer multi-tenancy?

A: Yes, vSphere® Big Data Extensions offers multi-tenancy with customers able to deploy separate compute clusters for different tenants sharing HDFS. VM level resource and configuration isolation keeps data persistent and safe, and allows users to run mixed workloads simultaneously on a single physical host.

Q: Does vSphere® Big Data Extensions support HA for the Apache Hadoop stack?

A: Yes. BDE allows one-click HA for the full Apache Hadoop stack including NameNode and JobTracker nodes, and for Apache Hadoop tools Hive and HBase.

Please refer to Apache Hadoop 1.0 High Availability Solution on VMware vSphere™ for details/reference architecture.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Copyright © 2010 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at http://www.vmware.com/go/patents. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.