Virtualizing Apache Hadoop

June, 2012
Virtualizing Apache Hadoop

Table of Contents

EXECUTIVE SUMMARY ........................................................................................................................................3
INTRODUCTION ..................................................................................................................................................3
VIRTUALIZING APACHE HADOOP ....................................................................................................................4
  INTRODUCTION TO vSPHERE™ ....................................................................................................................... 4
  USE CASES AND ADVANTAGES OF VIRTUALIZING HADOOP ...................................................................... 4
MYTHS ABOUT RUNNING HADOOP IN A VIRTUALIZED ENVIRONMENT ......................................................... 6
CONCLUSION ......................................................................................................................................................8
REFERENCES ....................................................................................................................................................8
Executive Summary

Key business and technology trends are disrupting the traditional data management and processing landscape. Big data analytics is increasingly being viewed as a competitive advantage and businesses are embracing Big data technologies to gain significant insight into their business for continued success. Apache Hadoop is emerging as one of the leading application in the big data space and is being used by enterprises across verticals for Big data analytics to help make better business decisions based on large data sets.

This document introduces the benefits and use cases for virtualizing Hadoop and dispels some common myths. It also describes some of the initiatives being taken by VMware in support of an optimal virtualized platform for Apache Hadoop.

Introduction

The amount of digital data being generated and stored has exploded in recent years. 7 exabytes of digital data was added in the enterprise in the US last year alone [1]. Data is increasing in complexity as enterprises look to exploit the value locked up in a variety of data to get insight into its business for continued growth and success. Conventional BI systems, data warehouses, and database systems are simply not able to meet the ever increasing demands of this new situation for several reasons. The amount of data is far too large to store in relational database systems efficiently and maintain the desired level of performance. Further the data is often in unstructured format making it unsuitable for systems that only support structured schemas. Finally, the hardware required for traditional BI and Data Warehousing applications is too costly at large scale, making analytics effectively inaccessible to IT.

Apache Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It has grown to be one of the leading Big data applications to address several of the issues discussed above in a cost effective manner, making it a natural fit as an analytics, transformation (ETL) and integration platform. These capabilities of Hadoop along with unstructured data explosion are causing CIOs to reconsider Enterprise data strategy.

Figure 1: Industry Trends (Source: Forrester survey of 60 CIO's, September 2011)
Virtualizing Apache Hadoop

Introduction to vSphere™

VMware’s vSphere™ 5.0, being a cloud operating system, virtualizes the entire IT infrastructure such as servers, storage, and networks. It groups these heterogeneous resources and transforms the rigid, inflexible infrastructure into a simple and unified manageable set of elements in the virtualized environment.

Broadly, vSphere™ offers two sets of services:
- Infrastructure Services: Virtualize and Aggregate Hardware Resources
- Application Services: Built-in Service Level Controls for Applications

Figure 2: vSphere™ 5.0 services

Use Cases and advantages of virtualizing Hadoop

Apache Hadoop is emerging as the de facto standard for big data processing, however, deployment and operational complexity, the need for dedicated hardware, and concerns about security and service level assurance prevent many enterprises from leveraging the power of Hadoop. By decoupling Hadoop nodes from the underlying physical infrastructure, VMware can bring the benefits of cloud infrastructure – rapid deployment, high-availability, optimal resource utilization, elasticity, and secure multi-tenancy – to Hadoop.

Discussed below are some of the advantage and use cases for running Apache Hadoop on a virtualized infrastructure.

- **Rapid Provisioning:** Using various tools and virtualization capabilities such as cloning, using templates, and resource allocation, significantly increases the speed of deployment of Hadoop. This is especially applicable for workloads like Hadoop that need to deploy and configure multiple nodes. On demand Hadoop instances, which are started on-demand, and shut down when not necessary are possible. VMware just launched a new open source project, Serengeti, to enable enterprises to quickly deploy, manage and scale Apache Hadoop in virtual and cloud environments. [4]
• **High Availability (HA) and Fault Tolerance (FT):** Although Hadoop is known to provide reliability via replication for storing data, there are several major components that are single points of failure in the system. Examples include the 'namenode', the 'jobtracker' and other supporting components such as Pig, Hive, Zookeeper, HBase, etc. Virtualizing Hadoop can address the high availability needs of all these components in a generic way with vSphere™ vMotion™, High Availability (HA) and Fault Tolerance (FT) features and keeping the system running with minimal or no downtime. For example, vSphere™ HA and vMotion™ technology can reduce downtime when nodes need to be brought down for planned upgrades and maintenance.

• **Datacenter efficiency:** Virtualizing Hadoop can increase datacenter efficiency by increasing the types of mixed workloads that can be run on a virtualized infrastructure. This includes running different versions of Hadoop itself on the same cluster, or running Hadoop along side other applications forming an elastic environment. Shared resources lead to higher consolidation ratios that leads to requirement of less hardware, software, and infrastructure to run the customer’s required set of business apps, thereby reducing the CapEx.

Figure 3: Virtualized infrastructure leads to data center consolidation

• **Efficient Resource Utilization:** Co-locating Hadoop VMs and other kinds of workloads on the same hosts and applying resource controls based on priority often allows better overall utilization by consolidating applications that use different kinds of resources.

• **Multi-tenancy:** Hadoop is a multi-tenant application. Running it on a virtualized environment can improve the Quality of Service (QoS) and offered SLA’s to the tenants by virtue of instance isolation and VMware resource pools. Also, in a virtualized environment, different tenants can run mixed workloads other than Hadoop on the same physical cluster, addressing yet another variance of multi-tenancy.

• **Security:** A virtualized environment provides organizational boundaries to secure the data and isolate it amongst users. An entire cluster can be run in an isolated group of virtual machines, providing full data isolation and security, while sharing the same underlying physical hardware.

• **Time sharing:** Taking advantage of unused capacity is simplified in a virtualized environment by running jobs during periods of low hardware usage by spinning up and down virtual machines easily.

• **Easy maintenance and movement of environment:** A cluster of Hadoop nodes running in a virtualized environment can be easily replicated or moved from one environment to another. This includes use cases such as moving the VM's
Virtualizing Apache Hadoop

from staging to production, from one cluster to another within a data center or even deploying Hadoop in a Hybrid Cloud model.

- **Hadoop-as-a-service:** VMware platform enables Hadoop to run in a Cloud environment. VMware vCloud™ director can be configured to offer a full Hadoop-as-a-Service solution in a private or public Cloud in order to offer an agile, controlled, elastic, cost-effective, secure, and a multi-tenant service, while benefiting from the management, deployment, and provisioning tools included with it. vCenter Chargeback can account for resource usage by multiple tenants of the cluster who can then be billed back accordingly.

**Myths about running Hadoop in a virtualized environment**

This section dispels some of the myths around virtualization as a platform for Hadoop.

- **Performance:** VMware and partners have done considerable amount of work on evaluating Hadoop performance in a virtualized environment. Results show that Hadoop works quite well on vSphere™, and in fact does better than native under certain configurations. Running 2 or 4 smaller VMs per physical machine usually resulted in better performance, often exceeding native performance. For further details, refer to [2].

- **SAN, NAS or Local Disk** – vSphere™ supports local disks and Hadoop can be configured to use local disk with same performance and functionality as native for HDFS. Local disks are recommended for cost and performance reasons and large scale. Hadoop also runs well in a shared SAN environment for small to medium sized clusters but has different performance and cost metrics. With advent of high bandwidth networks, such as 10 GB Ethernet, FoE, iSCSI etc., accessing data over the network is becoming less of a concern.

- **Total Cost of Ownership (TCO)** - Another concern among users is that virtualization increases the TCO of running Hadoop clusters due to acquisition cost of hardware and additional licensing costs (i.e. CAPEX). However, datacenter efficiency and hardware consolidation resulting from a virtualized infrastructure can reduce the physical hardware footprint, and bring CAPEX in line with purely commodity hardware. Further, virtualized infrastructure reduces OPEX through enabling automation, higher utilization, more efficient management and provisioning of hardware, configuration, turning etc. [3] Virtualization can minimize any potential lost revenue associated with downtime, outages, and failures resulting in reduced TCO and increased ROI.
VMware’s support for virtualized Apache Hadoop for enterprises

Apache Hadoop has the potential to transform business by allowing enterprises to harness very large amounts of data for competitive advantage. VMware is working with the Hadoop community to allow enterprise IT to deploy and manage Hadoop easily in their virtual and cloud environments and make VMware vSphere™ the best platform for scalable, highly available Enterprise Hadoop.

- **Project Serengeti**: VMware has recently launched Project Serengeti to enable enterprises to quickly deploy, manage, and scale Apache Hadoop in virtual and cloud environments. [4] Available for free download under the Apache 2.0 license, Serengeti, is a “one-click” deployment toolkit, that allows enterprises to leverage VMware vSphere™ platform to deploy a highly available Hadoop cluster in minutes, including common Hadoop components such as HDFS, MapReduce, Pig, and Hive on a virtual platform. By using Serengeti to run Hadoop on VMware vSphere™, enterprises can easily leverage the high-availability, fault tolerance, and live migration capabilities of the world’s most trusted and widely deployed virtualization platform to ensure the availability and manageability of Hadoop clusters. Serengeti supports multiple Hadoop based distributions from a range of vendors including: Apache Hadoop, Cloudera Distribution, Greenplum HD, and Hortonworks Data Platform. Serengeti’s open architecture makes it easy to rapidly add support for additional distributions.

![Figure 4: Overview of Project Serengeti](image)

- To further simplify and speed the enterprise use of Apache Hadoop, VMware is working with the Apache Hadoop community to contribute changes to enhance the support for failure and locality topologies by making Hadoop “virtualization-aware”. The topology changes help to achieve optimal data placement on a virtual infrastructure, thereby improving performance and reliability. This enables the enterprises to achieve a truly elastic and secure Hadoop cluster. Hadoop Virtualization Extensions work with multiple hypervisors. [5]

- VMware has also updated Spring for Apache Hadoop, an open source project first launched in February 2012 to make it easy for enterprise developers to build distributed processing solutions with Apache Hadoop. These applications range from small standalone applications to integration and workflow applications based on the Spring Integration and Batch projects. [6] The current release of Spring for Apache Hadoop enables developers to create, configure, and execute all types of Hadoop jobs including Map-Reduce, Streaming, Hive, Pig, and Cascading. The newly announced updates allow Spring developers to easily build enterprise applications that integrate with the HBase database, the Cascading library, and Hadoop security. Spring for Apache Hadoop is free to download and available now under the
open source Apache 2.0 license.

- Java workloads run well on vSphere™. VMware has published Java best practices guidelines and these also apply to Hadoop running on a virtualized infrastructure. [7]

Together, these projects and contributions will help accelerate Hadoop adoption and enable enterprises to leverage Big data analytics applications, such as Cetas Software, to obtain real-time and intelligent insight into large quantities of data. VMware acquired Cetas [8] in April 2012 and the Cetas analytics service is currently available at http://www.cetas.net/

Conclusion

In conclusion, infrastructure virtualization brings several benefits to Hadoop deployments that include:
- Rapid provisioning
- HA solution
- Hardware consolidation
- Multi-tenancy and security through isolation of resources
- Automation

References

3. VMware ROI/TCO calculator: http://roitco.vmware.com/vmw/
4. Project Serengeti: http://projectserengeti.org
7. Java Best practices on VMware: http://www.vmware.com/resources/techresources/1087