Reference Architecture and Best Practices for Virtualizing Hadoop Workloads

Justin Murray
VMware
Agenda

• The Hadoop Journey
• Why Virtualize Hadoop?
• Elasticity and Scalability
• Performance Tests
• Storage Reference Architectures
• Isilon Architecture and Benefits
• vSphere Big Data Extensions
• Conclusion and Q&amp;A
The Customer Journey with Hadoop
The Hadoop Journey

Stage 1: Piloting
- Often start with line of business
- Try 1 or 2 use cases to explore the value of Hadoop

Stage 2: Hadoop in Production
- Serve a few departments
- A few more use cases
- Core Hadoop + components

Stage 3: Cloud Analytics Platform
- Serve many departments
- Often part of mission critical workflow
- Integrated with other big data services

Integrated
Why Virtualize Hadoop?
Customer Example: Enterprise Adoption of Hadoop

Translation of Diagram:

Production
Test
Experimentation

Dept A: recommendation engine
Dept B: ad targeting

Log files

Issues:
1. Multiple clusters to manage
2. Redundant common data in separate clusters
3. Peak compute and I/O resource is limited to number of nodes in each independent cluster

SLA: Jobs complete in 15 minutes
Bandwidth limited to 30 nodes at peak

Social data
Transaction data
Historical cust behavior
What if you could…

One physical platform to support multiple virtual big data clusters

Without Virtualization
- Multiple copies of common data (e.g. historical data, log data etc.) in separate Hadoop clusters

Consolidate and virtualize
- Single copy of common data results in less storage requirements while maintaining good isolation between different MapReduce clusters
Big Data Extensions Value Propositions

- Rapid Deployment
- Self service tools
- Performance

- Elastic scaling
- Avoid dedicated hardware
- VM-based isolation
- Increase resource utilization
- True multi-tenancy

- Deployment choice
- Maintain management flexibility at scale
- Control Costs
- Leverage vSphere features
Hadoop 2.0 – Yet Another Resource Negotiator
A Virtualized Hadoop 2.0 Cluster
vSphere Big Data Extensions
- Deploy Hadoop Clusters in Minutes

Server preparation → OS installation → Network Configuration → Hadoop Installation and Configuration

From a manual process

To fully automated, using the GUI
Elastic, Multi-Tenant Hadoop with Virtualization

Hadoop Node

- Combined Compute and Storage

Unmodified Hadoop node in a VM
- VM lifecycle determined by Datanode
- Limited elasticity

Separate Compute from Storage
- Separate compute from data
- Stateless compute
- Elastic compute

Separate Virtual Compute Clusters per tenant
- Separate virtual compute
- Compute cluster per tenant
- Stronger VM-grade security and resource isolation
Performance and Reference Architectures
Native vs. Virtual, 32 hosts, 16 disks per host

Source: http://www.vmware.com/resources/techresources/10360
Reference Architecture: 32-Server Performance Test

Up to four VMs per server
vCPUs per VM fit within socket size (e.g. 4 VMs x 4 vCPUs, 2 X 8)
Memory per VM - fit within NUMA node size

2013 Tests done using Hadoop 1.0
I/O Profile of a Hadoop MapReduce Job
(TeraSort example application)

75% of Disk Bandwidth

Map Task
Map Task
Map Task
Map Task

Reduce
Reduce
Sort
Spills
DFS Input Data
DFS Output Data

Map Output file.out
Shuffle Map_*.out
Combine Intermediate.out

Spills & Logs spill*.out

75% of Disk Bandwidth

12% of Bandwidth

12% of Bandwidth

HDFS
The Combined Model – Standard Deployment

Virtualization Host

Shared storage SAN/NAS

Local disks

OS Image – VMDK

VMDK

VMDK

VMDK

VMDK

VMDK

VMDK

VMDK

NodeManager

Hadoop Virtual Node

Datanode

Ext4

Ext4

Ext4

Ext4

Ext4

Ext4
Combined Model – Two Virtual Machines on a Host Server

Virtualization Host

Hadoop Virtual Node 1

Virtualization Host

Hadoop Virtual Node 2

OS Image – VMDK

VMDK

VMDK

VMDK

VMDK

Shared storage SAN/NAS

Local disks

Combined Model – Two Virtual Machines on a Host Server

Hadoop Virtual Node 1

NodeManager

DataNode

Ext4

Ext4

Ext4

Ext4

Hadoop Virtual Node 2

NodeManager

DataNode

Ext4

Ext4

Ext4

Ext4
The Data-Compute Separation Deployment Model

Hadoop Virtual Node 1
- NodeManager
- Ext4
- Ext4
- Ext4
- Ext4
- Ext4
- Ext4
- Ext4

Hadoop Virtual Node 2
- DataNode
- Ext4
- Ext4
- Ext4
- Ext4

Virtualization Host
- OS Image – VMDK
- VMDK
- VMDK
- VMDK
- VMDK
- VMDK
- VMDK
- VMDK
- VMDK
- VMDK
- VMDK
- VMDK

Shared storage SAN/NAS
Local disks

2014 SNIA Analytics and Big Data Summit. © Insert Your Company Name. All Rights Reserved.
Data Paths: Combined vs Data-Compute Separation

**Combined Model**
- Hadoop Virtual Node
  - DataNode
  - NodeManager

**Separated Model**
- Hadoop Virtual Node 1
  - DataNode
  - NodeManager
- Hadoop Virtual Node 2

Virtualization Host
- Virtual Switch

2014 SNIA Analytics and Big Data Summit. © Insert Your Company Name. All Rights Reserved.
Alternative Storage for Data/Compute Separation

- Hadoop Virtual Node 1
  - DataNode
  - Ext4
  - Ext4
- Hadoop Virtual Node 2
  - NodeManager
  - Ext4
  - Ext4
  - Ext4
  - Ext4
  - Ext4
  - Ext4
  - Ext4
  - Ext4

- Virtualization Host
  - OS Image – VMDK
  - VMDK
  - VMDK
  - VMDK
  - VMDK
  - VMDK
  - VMDK
  - VMDK
  - VMDK

- Shared storage
  - SAN/NAS

- Local disks
VMDK Isolation for Performance

NodeManager

Hadoop Virtual Node 1

Ext4 Ext4 Ext4 Ext4

OS Image – VMDK

Virtualization Host

Shared storage SAN/NAS

Local disks for Temp Data

JBOD

Hadoop Virtual Node 2

Ext4 Ext4 Ext4

OS Image – VMDK

Local disks for HDFS Data

JBOD

Datanode

VMDK VMDK VMDK

SAN/NAS

Local disks for Temp Data

JBOD
Data/Compute Separation With Isilon

ResourceManager

Hadoop Virtual Node 1

Ext4

ResourceManager

Hadoop Virtual Node 2

Ext4

NodeManager

Hadoop Virtual Node 3

Ext4

Virtualization Host

OS Image – OS Image – OS Image – VMDK

Virtualization Host

VMDK

VMDK

Temp

Temp

Shared storage SAN/NAS

2014 SNIA Analytics and Big Data Summit. © Insert Your Company Name. All Rights Reserved.
Larger Architecture with Data Compute Separated

- Dell Force 10 S60 1GbE
- Stacked
- Dell PowerEdge R720
  - (8) Management Nodes
    - NameNodes
    - Client/Edge Nodes
    - JobTrackers
    - HBase Master

- Dell Force 10 S4810 10GbE
- (2) 10 GbE Data Traffic
- Dell PowerEdge R720XD
  - (100) Worker Nodes
    - DataNodes
    - TaskTrackers
    - HBase Region Servers

- Dell Force 10 Z9000 40GbE
- 40 GbE
- Top-of-Rack

- 10 GbE
- Management

2014 SNIA Analytics and Big Data Summit. © Insert Your Company Name. All Rights Reserved.
Hybrid storage model - the best of both worlds

- **Master nodes**
  - NameNode, ResourceManager, ZooKeeper etc. on shared storage
    - Leverage vSphere vMotion, HA and FT

- **Worker nodes**
  - NodeManager/DataNode on local storage
  - Lower cost, scalable bandwidth
  - Temp data is written to local storage for best performance
  - NFS storage for HDFS data is a very good alternative to local
vSphere Big Data Extensions and Project Serengeti
Big Data Extensions - Highlights

- Open source project
- Tool to simplify virtualized Hadoop deployment & operations

- Virtualization changes for core Hadoop
- Contributed back to Apache Hadoop
- Advanced resource management on vSphere

Serengeti

Hadoop Virtualization Extensions (HVE)

Virtual Hadoop Manager (VHM)
Introducing vSphere Big Data Extensions (BDE)

vCenter Plugin

Hadoop as a Service with vCloud Automation Center
Brief Tour of Big Data Extensions
One Click to Scale out the Cluster on the Fly
BDE Allows Flexible Configurations

```json
{
  "name": "master",
  "roles": [
    "hadoop_namenode",
    "hadoop_resourcemanager"
  ],
  "instanceNum": 1,
  "instanceType": "LARGE",
  "cpuNum": 2,
  "memCapacityMB": 4096,
  "storage": {
    "type": "LOCAL",
    "sizeGB": 50
  },
  "haFlag": "on",
  "rpNames": [
    "rp1"
  ]
}
```

Storage configuration Choice of shared or local

```json
{
  "name": "data",
  "roles": [
    "hadoop_datanode"
  ],
  "instanceNum": 3,
  "instanceType": "MEDIUM",
  "cpuNum": 2,
  "memCapacityMB": 2048,
  "storage": {
    "type": "LOCAL",
    "sizeGB": 50
  },
  "placementPolicies": {
    "instancePerHost": 1,
    "groupRacks": {
      "type": "ROUNDROBIN",
      "racks": ["rack1", "rack2", "rack3"]
    }
  }
}
```

Number of nodes and resource configuration

High Availability option

VM placement policies
External HDFS: Simple to Set Up

Cluster name: Tier2_demo
Hadoop distro: Pivotall-D
Deployment type: Compute-only Hadoop Cluster

DataMaster
DataMaster URL: hdfs://client1.vmware.com

ComputeMaster Node Group
Number of nodes: 1
Resource template: Medium
2 vCPU, 7500 MB RAM, 50 GB storage on Shared datastore

Worker Node Group
Number of nodes: 16
Resource template: Customize...
1 vCPU, 3748 MB RAM, 25 GB storage on Shared datastore

Client Node Group (Optional)
Number of nodes: 1
Resource template: Email
1 vCPU, 3748 MB RAM, 25 GB storage on Shared datastore

Network: defaultNetwork
How BDE works
Provision the virtual machines at the right size
- Reserve 6% of physical memory on the ESXi Server for vSphere usage
- Avoid over-commitment
- Enable NUMA and keep the virtual machine memory and cpu size within the NUMA node

NUMA scheduler is important for virtualized Hadoop performance
- Poor configuration can result in performance degradation
- Data VM preferably should be distributed across NUMA nodes
VMware vSphere BDE and Hadoop Resources

- VMware vSphere BDE web site
  - [http://www.vmware.com/bde](http://www.vmware.com/bde)

- Virtualized Hadoop Performance with VMware vSphere 5.1
  - [http://www.vmware.com/resources/techresources/10220](http://www.vmware.com/resources/techresources/10220)

- Benchmarking Case Study of Virtualized Hadoop Performance on vSphere 5

- Hadoop Virtualization Extensions (HVE)

- Apache Hadoop High Availability Solution on VMware vSphere 5.1
Conclusions

- Hadoop workloads work very well on VMware vSphere
  - Various performance studies have shown that any difference between virtualized performance and native performance is minimal
  - Follow the general best practice guidelines that VMware has published

- vSphere Big Data Extensions enhances your Hadoop experience on the VMware virtualization platform
  - Rapid provisioning tool for deployment of Hadoop components in virtual machines
  - Algorithms for best layout of your Hadoop data and cluster components are built into the BDE HVE components
  - Design patterns such as data-compute separation can be used to provide elasticity of your Hadoop cluster on demand.
  - User self service available with Hadoop using tools such as vCloud Automation Center integrated with BDE
Thank You

jmurray@vmware.com
Justin Murray
Backup Slides
Today’s Challenges on Hadoop Infrastructure

• Fixed compute and storage coupling leads to low utilization and inflexibility
  • Compute and storage linked together with fixed ratio based on the hardware specification
  • Not all jobs are created equal (data vs. compute intensive)
  • Inflexible infrastructure leads to waste
    • Too little compute power → slow processing
    • Too much compute power → sitting idle
• Problem compounds with larger clusters
  • So what happens?
Getting more out of your infrastructure

- Decouple the linkage between compute and storage
- Stateless compute can grow and shrink elastically
- Data locality is preserved, place the compute where data resides
- Extra compute capacity can be used for other workloads
Elasticity and Scalability
Elastic Scalability & Multiple Workloads

- Deploy separate compute clusters for different tenants sharing HDFS.
- Commission/decommission compute nodes according to priority and available resources.

Diagram:

- Compute layer
  - Compute VMs
  - Resource Manager

- Dynamic resource pool
  - Production

- Data layer
  - VMware vSphere + Big Data Extensions