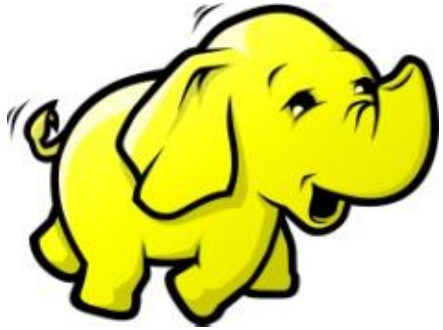
Big Data in the Cloud

Ronaldo Amá
VP, R&D, Data Services
VMware, Inc

The answer is 

What was your question?

Well, this is a Hadoop show after all

Big Data Landscape

Real Time Streams
(Social, sensors)

Machine Learning
(Mahout, etc…)

Real-Time Processing
(s4, storm, spark)

Data Visualization
(Excel, Tableau)

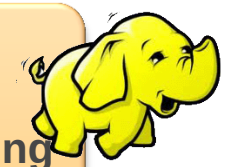
ETL

(Informatica, Talend, Spring Integration)

Real Time Database
(Shark, Gemfire, hBase, Cassandra)

Interactive Analytics
(Impala, Greenplum, AsterData, Netezza…)

Batch Processing
(Map-Reduce)



Structured and Unstructured Data
(HDFS, MAPR)

Cloud Infrastructure

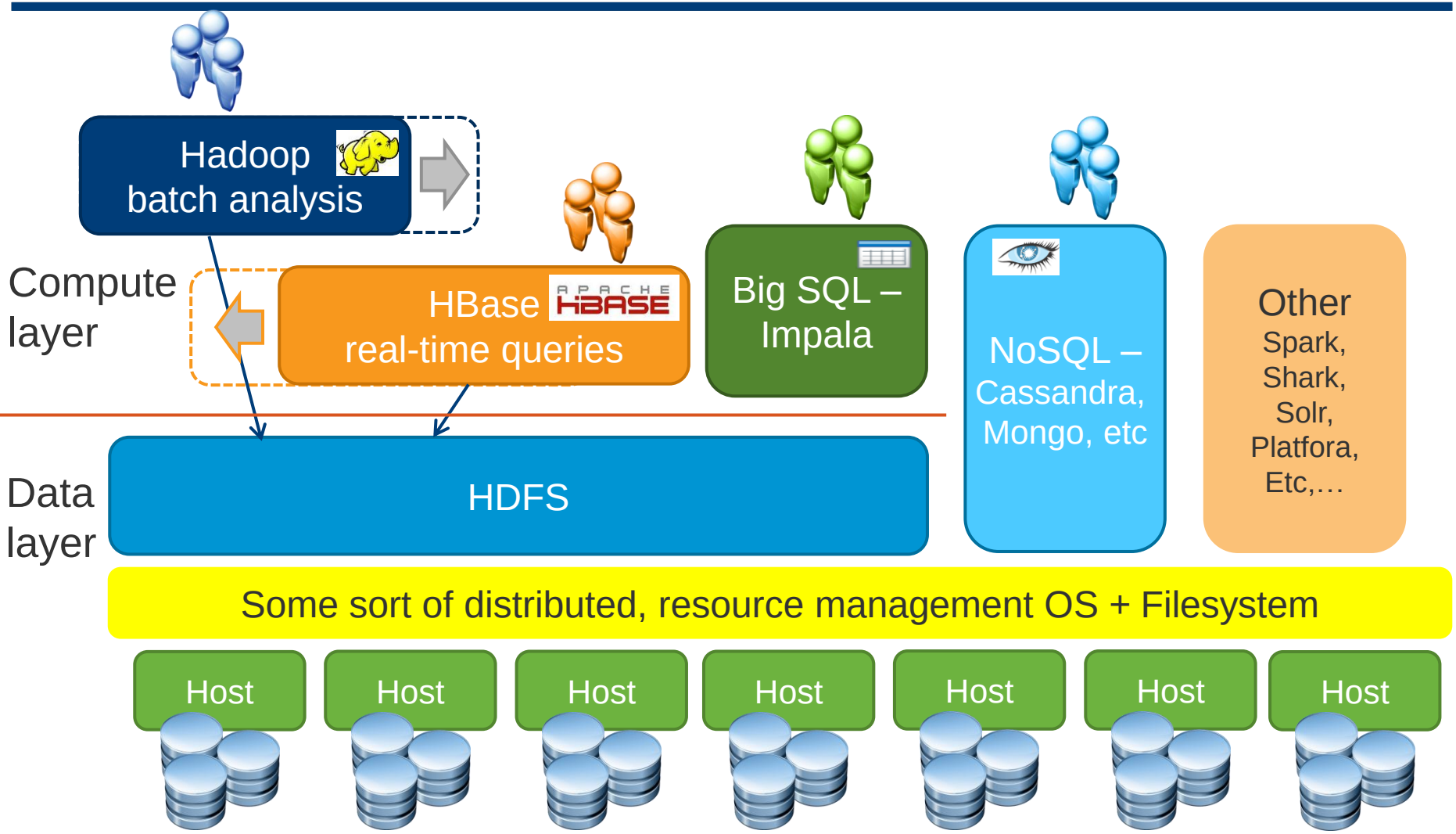
Compute

Storage

Networking

vmware

Technology Stack



Why Virtualize Hadoop

Operational Simplicity

- Rapid deployment, cloning
- Unified life-cycle management
- Easy to configure/reconfigure

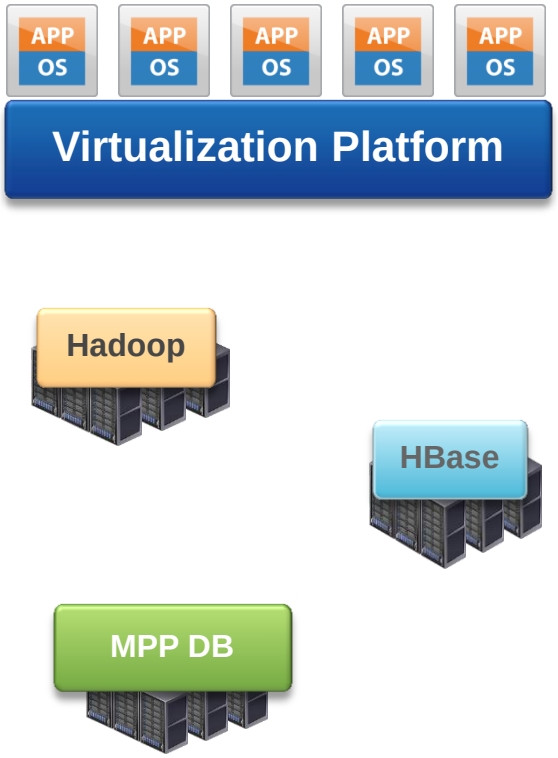
Highly Availability

- High availability for entire Hadoop stack
- One click to setup
- Proven solution

Elasticity & Multi-tenancy

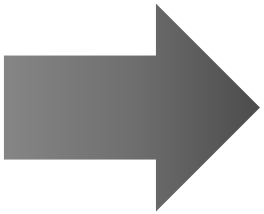
- Shrink and expand cluster on demand
- Independent scaling of Compute and data
- Strong multi-tenancy

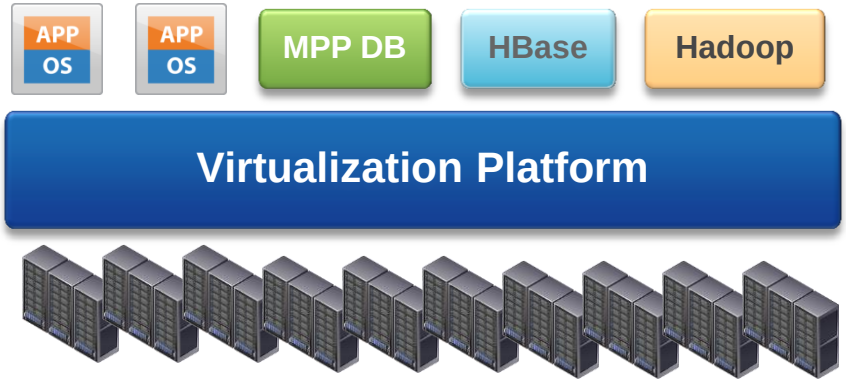
Common Infrastructure for Big Data



Cluster Sprawling

Single purpose clusters for various business applications lead to cluster sprawl.

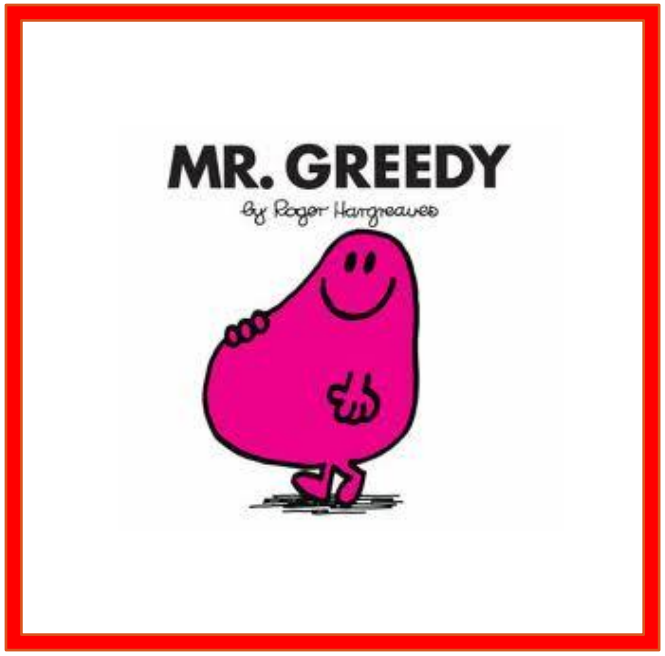




Cluster Consolidation

- **Simplify**
  - Single Hardware Infrastructure
  - Unified operations
- **Optimize**
  - Shared Resources = higher utilization
  - Elastic resources = faster on-demand access

Mixing Workloads: Three big types of Isolation are Required

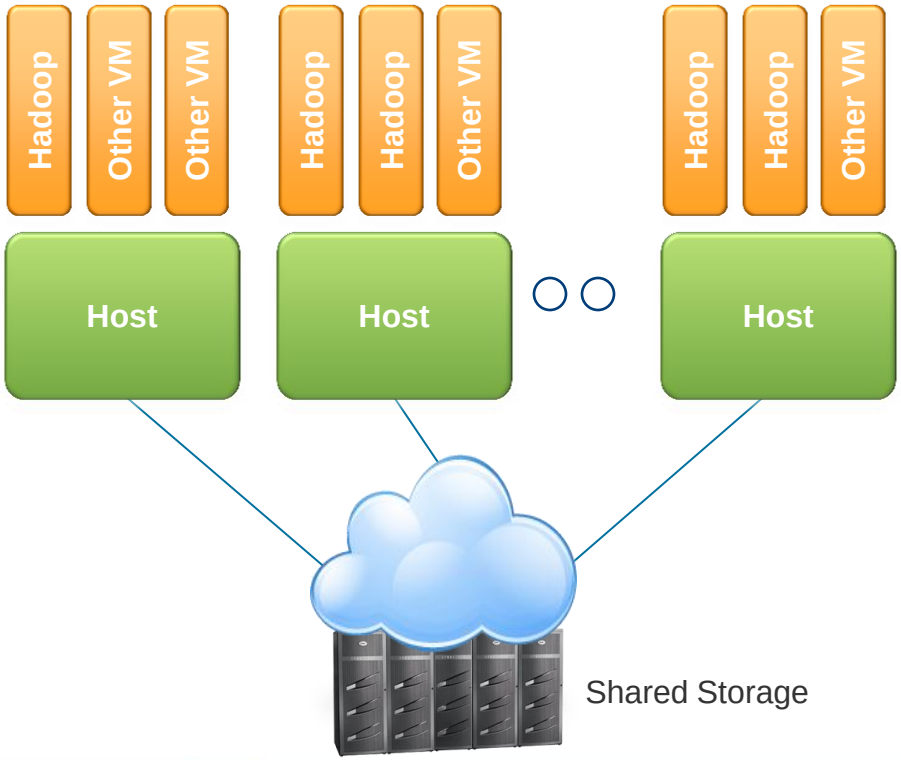


- **Resource Isolation**
  - Control the greedy noisy neighbor
  - Reserve resources to meet needs
- **Version Isolation**
  - Allow concurrent OS, App, Distro versions
- **Security Isolation**
  - Provide privacy between users/groups
  - Runtime and data privacy required

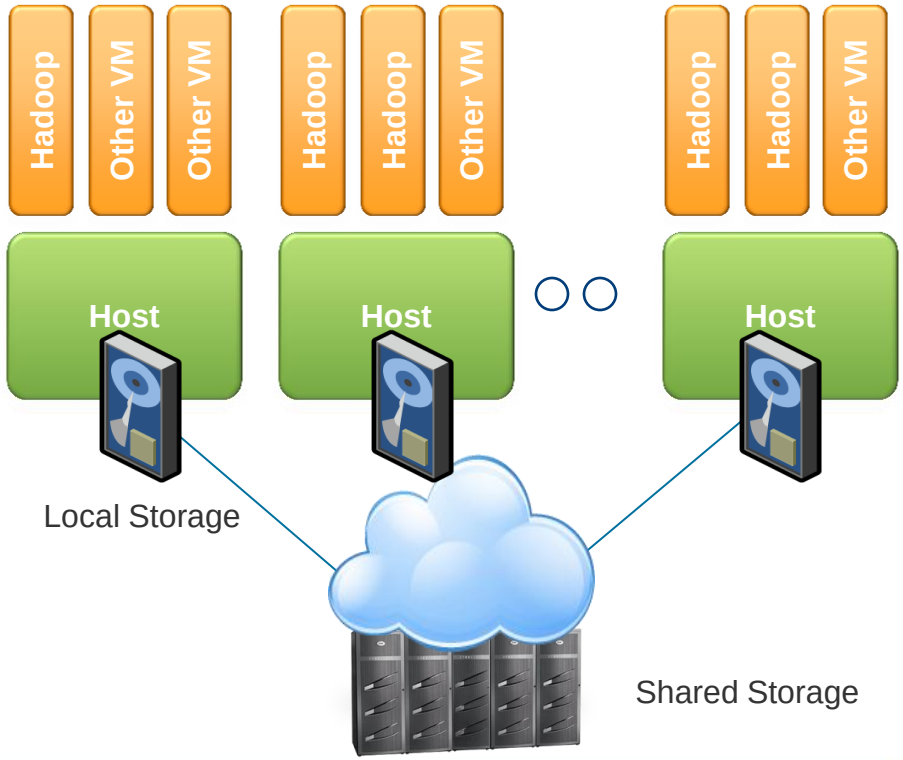
Some sort of distributed, resource management OS + Filesystem

Virtual Storage Architecture Include Local Disk

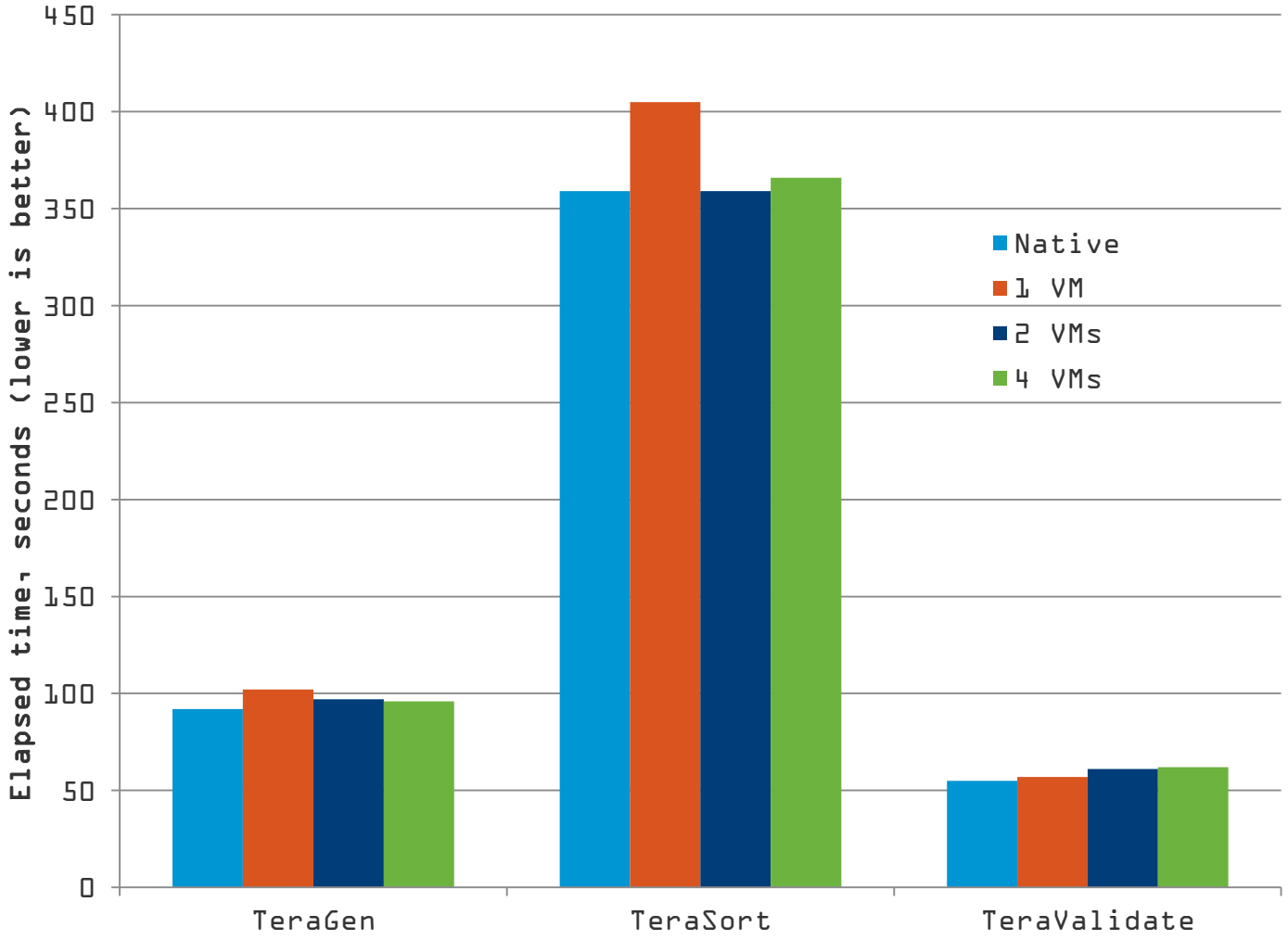
- **Shared Storage: SAN or NAS**
  - Easy to provision
  - Automated cluster rebalancing
  - Leverage vmotion/HA/FT



- **Local Storage: Local Disks**
  - Local disk for Hadoop
  - Scalable Bandwidth, lower cost/GB



Hadoop Runs Well on Virtualization



Source: http://www.vmware.com/files/pdf/techpaper/VMW-Hadoop-Performance-vSphere5.pdf

Project Serengeti

- Open source project launched in June 2012, meta-updates released on regular schedule (~3 Months intervals)
- Toolkit that leverage virtualization to simplify Hadoop deployment and operations
- Commercial support via Data Director

Deploy a Hadoop cluster in 10 Minutes

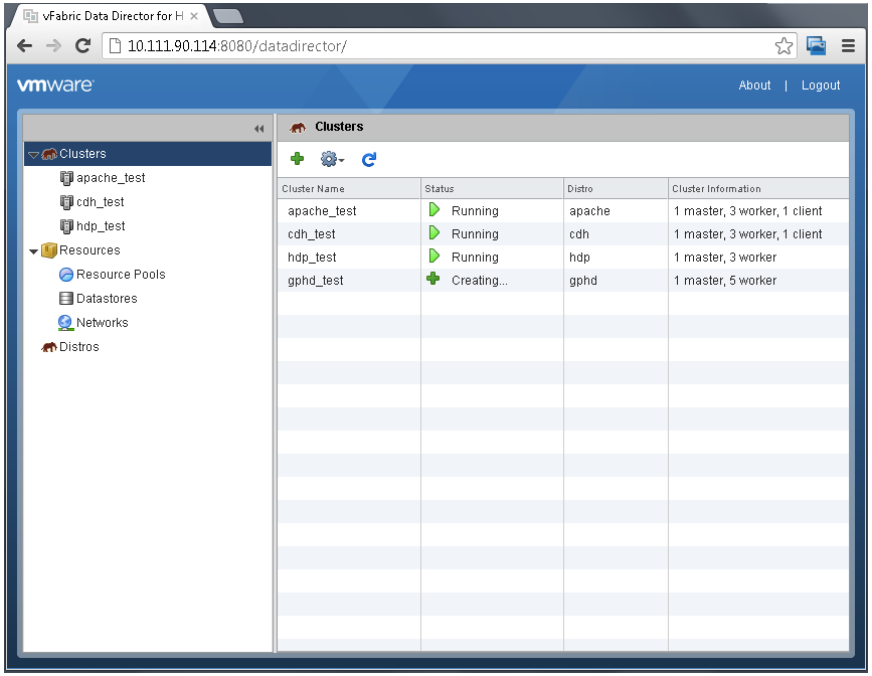
Customize Hadoop cluster

Use Your Favorite Hadoop Distribution

One stop command center

Hadoop Resources

- **Download and try Serengeti**
  - projectserengeti.org
- **Commercial support via Data Director**
  - vmware.com/products/application-platform/vfabric-data-director/overview.html



- **VMware Hadoop site**
  - vmware.com/hadoop

- **Hadoop performance on vSphere**
  - vmware.com/files/pdf/VMW-Hadoop-Performance-vSphere5.pdf
- **Hadoop High Availability solution**
  - vmware.com/files/pdf/Apache-Hadoop-VMware-HA-solution.pdf

THANK YOU!

Ronaldo Amá
VP, R&D, Data Services
VMware, Inc