Big Data in Cloud Era

Dong Bo，VMware, Inc.

© 2011 VMware Inc. All rights reserved

Intensive Data demands changing approach to managing data

Databases no longer just have traditional applications requirements where one size fits all…

Traditional Applications

Microsoft IBM SAP ORACLE

New Data Types

YouTube flickr

0 123456 789012

Ubiquity of Devices

Real-time Expectations

… new applications are putting additional pressure on the database

vmware

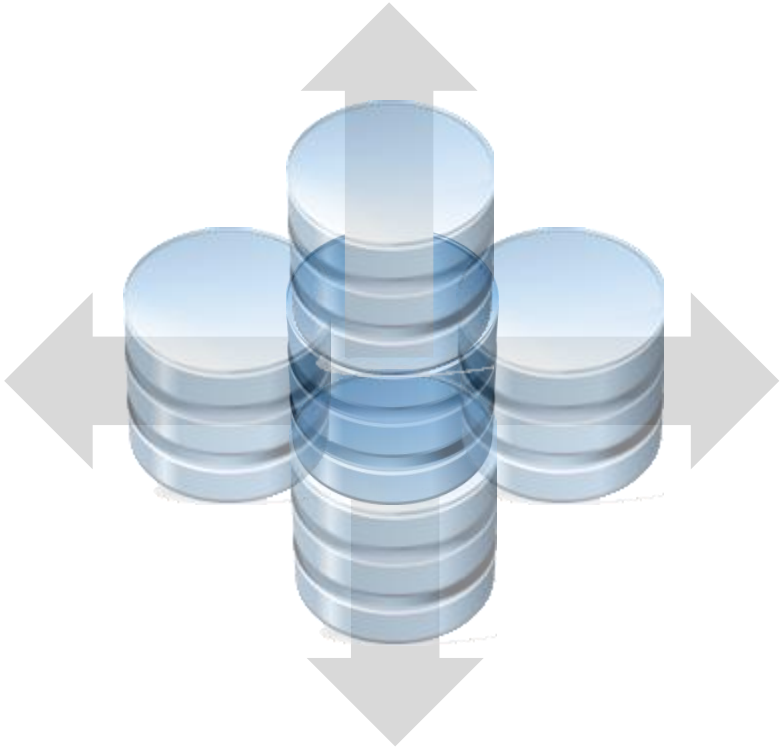
The Database is Being Stretched

Fast Data

- Low latency expectations
- Horizontal scale

Big Data

- Petabytes vs. Gigabytes
- Democratize BI



Flexible Data

- Unstructured data
- Developer productivity

Cloud Delivery

- Virtualized
- Offered “-as-a-Service”

Big, Fast and Flexible Data

Big

Big Data Processing

hadoop

Greenplum

Serengeti

Big Data Analytics

CETAS
INSTANT INTELLIGENCE

Fast

OLTP workloads

SQLFire

GemFire

Analytic workloads

SQLFire

Flexible

Object

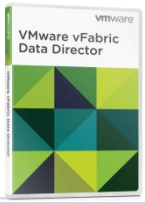
GemFire

Key / Value

redis

OSS Relational

vPostgres



Cloud Delivery Model

Data as a service for private and public clouds

CLOUD FOUNDRY™

vmware

Hadoop in Cloud

Hadoop Journey in Enterprises

Integrated

Standalone

Stage 3: Big Data Production

- ✓ Serve many departments
- ✓ Often part of mission critical workflow
- ✓ Integrated with other big data services like MPP DB, NoSQL,

Stage 2: Hadoop Production

- ✓ Serve a few departments
- ✓ A few more use cases
- ✓ Core Hadoop + components
- ✓ Typically in tens and hundreds of nodes

Stage1: Piloting

- ✓ Often start with line of business
- ✓ Try 1 or 2 use cases to explore the value of Hadoop
- ✓ Typically under 20 nodes

Scale

0 node

20

300

vmware

Stage 1: Piloting

Stage1: Piloting

- ✓ Often start with line of business
- ✓ Try 1 or 2 use cases to explore the value of Hadoop
- ✓ Typically under 20 nodes
- ✓ Led by either data team or infrastructure team

- **Requirements:**

✓ Make it quick
  - ✓ I don't want to wait for weeks or months
  - ✓ Get me a Hadoop cluster quick

✓ Make it easy
  - ✓ Make it easy for me to access the data
  - ✓ Make it easy for me to try different algorithms and data sets

Benefit of Virtualize at this stage

- Leverage existing vSphere infrastructure and hardware → No need to wait
- Apply spare machine cycle for these Hadoop piloting projects → No need to purchase additional hardware/software
- Use resource pools and DRS to contain the resources used for these Hadoop piloting projects → No impact to existing workloads
- The data is probably already in the shared storage → No need to move data around

→ It's a no brainer to leverage virtualization for this stage

Project Serengeti

- Open source project launched in June, 2012
- Toolkit that leverage virtualization to simplify Hadoop deployment and operations
- To learn more, projectserengeti.org

Serengeti

Deploy a Hadoop cluster in 10 Minutes

Customize Hadoop cluster

Use Your Favorite Hadoop Distribution

One stop command center

vmware

Stage 2: Hadoop Production

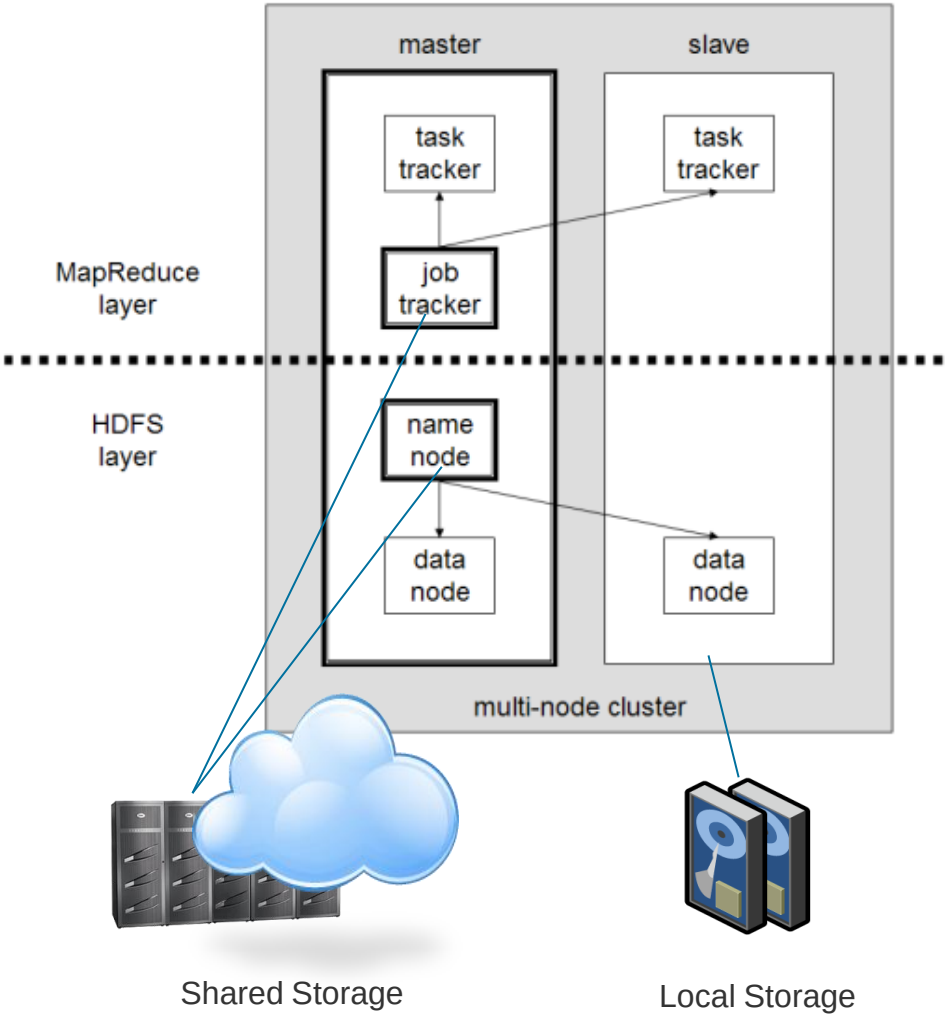
Stage 2: Hadoop Production

- ✓ Serve a few departments
- ✓ A few use cases
- ✓ Core Hadoop + some non-core components
- ✓ Dedicated Hadoop administrator

Requirements:

- ✓ High availability
  - ✓ We are in production and need SLA
  - ✓ High availability of the entire Hadoop Stack
- ✓ Agility
  - ✓ We are getting new Hadoop use requests all the time, make it easy for me to scale the cluster
  - ✓ We need to configure and reconfigure the clusters often
- ✓ Differentiated level of services
  - ✓ We have production Hadoop jobs, need to ensure high priority
  - ✓ We also have people trying “ad hoc” Hadoop jobs, need to satisfy their request too

Hybrid storage model to get the best of both worlds



- Master nodes:
  - Name node, job tracker etc. on shared storage
  - Leverage vSphere vMotion, HA and FT
- Slave nodes
  - Task tracker/data node on local storage
  - Lower cost, scalable bandwidth

Stage 3: Big Data Production

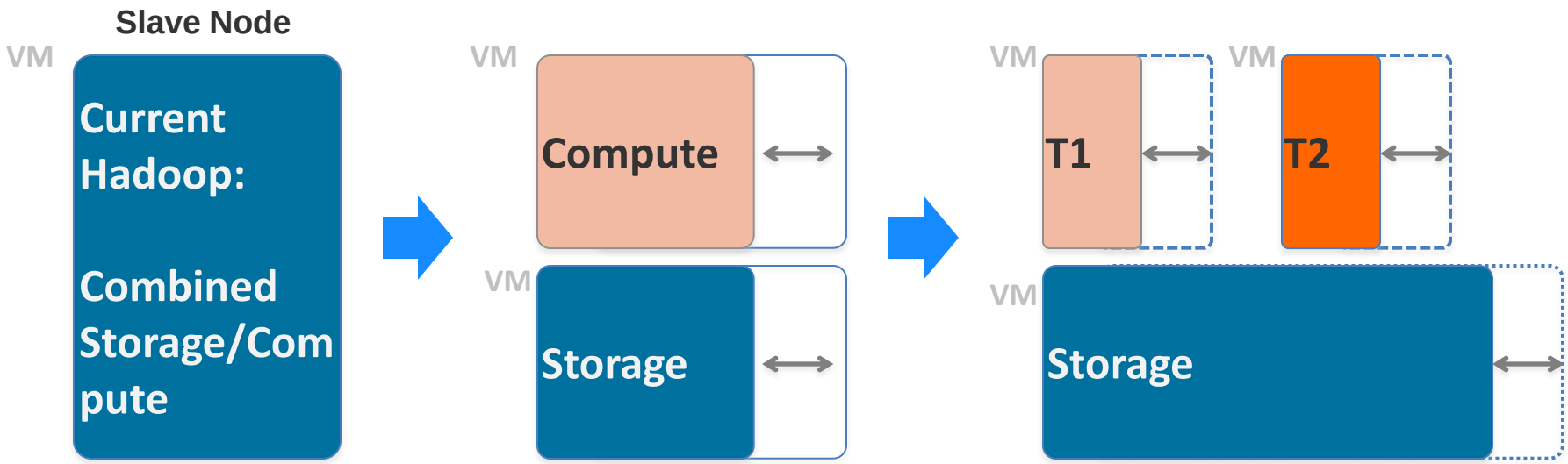
Stage 3: Big Data Production

- ✓ Serve many departments
- ✓ Often part of mission critical workflow
- ✓ Offer other big data services like MPP DB, NoSQL, more non-core components

Requirements:

- ✓ Multi-tenancy
  - ✓ We have many tenants on the cluster now, and need ensure resource isolation, configuration isolation between different tenants
- ✓ Elasticity
  - ✓ With more and more users and jobs on the system, we need to make sure the Hadoop cluster is elastic and adjust to changing demands
- ✓ Integrated big data production
  - ✓ It's not just about Hadoop anymore, Hadoop is now critical part of overall big data analytics workflow

Evolution of Hadoop on VMs – Improvement on Model 2



Hadoop in VM

- VM lifecycle determined by Datanode
- Limited elasticity
- Limited to Hadoop Multi-Tenancy

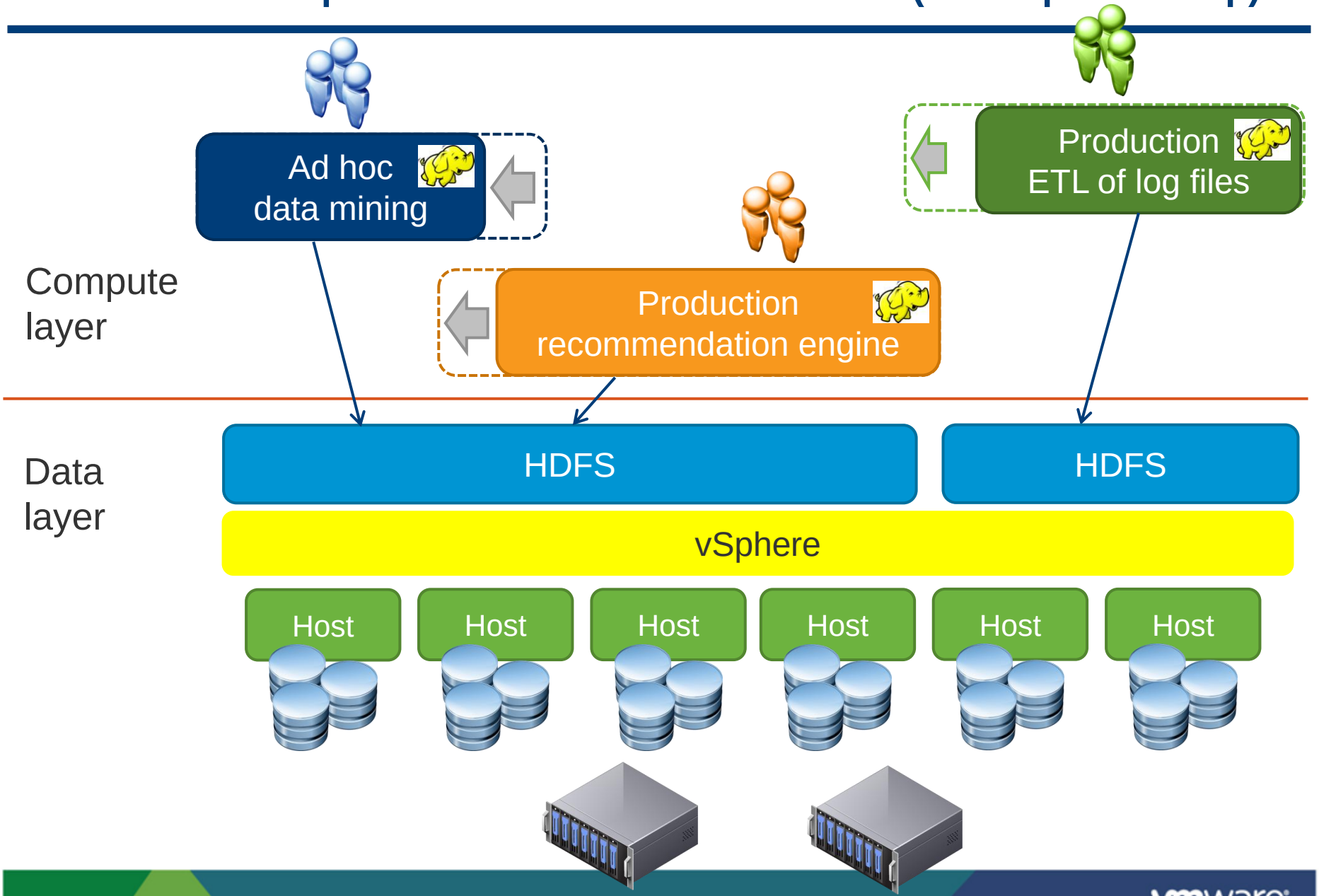
Separate Storage

- Separate compute from data
- Elastic compute
- Enable shared workloads
- Raise utilization

Separate Compute Clusters

- Separate virtual clusters per tenant
- Stronger VM-grade security and resource isolation
- Enable deployment of multiple Hadoop runtime versions

In-house Hadoop as a Service "In house EMR" – (Hadoop + Hadoop)



Thank you!