A Big Data Storage Architecture for the “Second Wave”

David “Sunny” Sundstrom
Principle Product Director, Storage
Oracle
Growth in Data Diversity and Usage
1.8 Zettabytes of Data in 2011, 20x Growth by 2020

- Enterprise
  45%/year growth in database data

- Analytics
  Customers top 50PB. Keep all your data.

- Cloud
  Public and private. IaaS, Paas, SaaS.

- Regulation
  300 Exabytes in archives by 2015.

- Mobile, POS, sensors
  Mobile #1 Internet access device in 2013

- Social Media and Commerce
  $30B/year in commerce by 2015

- Cloud Applications
  80% of new applications
Agenda

- Big Data Using Hadoop, The First Wave
- Big Data Today, The Second Wave
- A Storage Model for Big Data
- Implementations
- Plan for action
- Summary
Unstructured Data Analytics Matures

Hadoop

- The engine of choice for unstructured data
- By architecture an independent data store
Hadoop Ecosystem and Technologies

Vertical Apps
- Predictive Policing
- Bloomreach
- Myrrix

Log Data Apps
- Splunk
- Loggly
- Sumologic

Ad/Media Apps
- RocketFuel
- Collective
- Bluefin
- Recorded Future

Business Intelligence
- Oracle
- Hyperion
- SAP
- BusinessObjects
- IBM
- Cognos
- MicroStrategy

Analytics and Visualization
- Tableau
- Palantir
- Metamarks
- Tealdata
- Aster

Data As A Service
- Factual
- Gnip
- Bitly
- INRIX
- LexisNexis

Analytics Infrastructure
- Hortonworks
- Cloudera
- EMC
- Greenplum

Operational Infrastructure
- Couchbase
- 10gen
- Teradata
- TeraData
- MarkLogic

Infrastructure As A Service
- Amazon
- Windows Azure
- Google BigQuery

Structured Databases
- Oracle
- MySQL
- IBM DB2
- Sybase

Technologies
- Hadoop
- Hadoop MapReduce
- Apache
- Cassandra

Copyright © 2012 Dave Feinleib
dave@rcdave.com
blogs.forbes.com/davefeinleib

2012 SNIA Analytics and Big Data Summit. © Oracle. All Rights Reserved.
But Hadoop is not “The Matrix”
Agenda

- Big Data Using Hadoop, The First Wave
- Big Data Today, The Second Wave
- A Storage Model for Big Data
- Implementations
- Plan for Action
- Summary
### Data Types Used In Analytics

Defined by the 3 Vs – Variety, volume, and velocity

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Hadoop</th>
<th>Non Hadoop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer/member data</td>
<td>59%</td>
<td>68%</td>
</tr>
<tr>
<td>Transactional data from applications</td>
<td>44%</td>
<td>68%</td>
</tr>
<tr>
<td>Application Logs</td>
<td>17%</td>
<td>69%</td>
</tr>
<tr>
<td>Other Types of Event Data</td>
<td>23%</td>
<td>64%</td>
</tr>
<tr>
<td>Network Monitoring/Network Traffic</td>
<td>33%</td>
<td>41%</td>
</tr>
<tr>
<td>Online Retail Transactions</td>
<td>33%</td>
<td>34%</td>
</tr>
<tr>
<td>Other Log Files</td>
<td>26%</td>
<td>51%</td>
</tr>
<tr>
<td>Call Data Records</td>
<td>28%</td>
<td>32%</td>
</tr>
<tr>
<td>Web Logs</td>
<td>21%</td>
<td>46%</td>
</tr>
<tr>
<td>Text data from social media and online</td>
<td>15%</td>
<td>36%</td>
</tr>
<tr>
<td>Search logs</td>
<td>11%</td>
<td>36%</td>
</tr>
<tr>
<td>Trade/quote data</td>
<td>18%</td>
<td>15%</td>
</tr>
<tr>
<td>Intelligence/defense data</td>
<td>11%</td>
<td>18%</td>
</tr>
<tr>
<td>Multimedia (audio/video/images)</td>
<td>9%</td>
<td>21%</td>
</tr>
<tr>
<td>Weather</td>
<td>3%</td>
<td>8%</td>
</tr>
<tr>
<td>Smartmeter data</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>Other (please specify)</td>
<td>3%</td>
<td>5%</td>
</tr>
</tbody>
</table>

© 2012 Storage Networking Industry Association. All Rights Reserved.
Optimizing Big Data
Multiple Data Models with Multiple Strengths

- Explicit data model
- Large scale
- Structured data
- Power of relational model
- Real time and batch
- Robust security

- No/minimal data model
- Extreme scale
- Unknown data structures
- Flexibility of no schema
- Batch and near-real time
- Basic to no security
Aligning Cost and Business Value

- Keeping most (or ALL) of your data
  - Regression analysis
  - Trends
  - Compliance
- Cost versus value
  - Value varies widely
  - Rising volume and use exceeds declining cost
Other Key Big Data Storage Factors

- Scalability with performance
- Security
- Data integrity
Agenda

- Big Data Using Hadoop, The First Wave
- Big Data Today, The Second Wave
- A Storage Model for Big Data
- Implementations
- Plan for Action
- Summary
“With higher-end systems, there is a lot of data coming from all of the different business processes, from managing inventories to analyzing data for trends for future products. So in these systems, there is a real need to integrate a lot of different applications - a lot of different usages of the same massive amount of data - and that requires someone to think about how all of the pieces go together.”

Rich Partridge
Senior Analyst, Ideas International
Big Data Storage Strategy

Reapplying What We Have Already Learned to Big Data……..

- Multiple storage silos are generally costly and inefficient
  - Hadoop storage is a silo by definition
- One centralized storage solution for all needs does not optimize either data or computation
- Storage architecture generally needs to evolve
  - Green field implementations have few constraints but….
  - Most data centers have existing, comprehensive network and storage architectures
- Store in concert with value
  - Storing petabytes on high performance disks is not cost-effective
Big Data Functional Model

Analytics is a Multi-function Task

Data bus

Ingest

Consolidation

Oracle DB
- Transactions and other structured data

Hadoop
- Unstructured data

Other sources
- NoSQL, etc.
- Media

Analytics and Visualization

Value Retention and Reuse
Big Data Storage Architecture

- **Data focused** architecture
- **Integrated, end-to-end**
- **Heterogeneous, purpose-aligned** storage
- **Optimized** data management
  - Data aligned
    - Processing engines aligned with data types and sources
  - Maximized efficiency
  - Minimized movement and copy
  - Leveraged reuse
  - Leveraged data transport (network)
Big Data Storage Architecture

- Shared storage
- Hadoop or NoSQL Cluster
- Data management software
- RDBMS (Transaction and data warehouse)
- Legacy servers and clusters
- Archival storage (Tape)
- High performance datacenter backbone
- Data management software
## Big Data Integration Software

### The Magic for Optimizing the Movement of Data

<table>
<thead>
<tr>
<th>Function</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured data ingest</td>
<td></td>
</tr>
<tr>
<td>Structured data ingest (ELT)</td>
<td></td>
</tr>
<tr>
<td>RDBMS to HDFS</td>
<td></td>
</tr>
<tr>
<td>HDFS to RDBMS</td>
<td></td>
</tr>
<tr>
<td>HDFS to/from shared storage</td>
<td></td>
</tr>
<tr>
<td>RDBMS to/from shared storage</td>
<td></td>
</tr>
<tr>
<td>Tiered storage management</td>
<td></td>
</tr>
</tbody>
</table>
Aligning Cost and Business Value

Automation becomes key
Agenda

- Big Data Using Hadoop, The First Wave
- Big Data Today, The Second Wave
- A Storage Model for Big Data
- Implementations
- Plan for Action
- Summary
Big Data Storage Implementations
Unstructured-data dominated, greenfield solution/POC

Internal system backbone

Hadoop Cluster

Apache Flume
Big Data Storage Implementations

Existing Hadoop Cluster with NAS Hosted Replication

- Hadoop Cluster
- Sun ZFS 7420
- 10-40GigE datacenter backbone
- Apache Sqoop
- RDBMS (Transaction and data warehouse)
Big Data Storage Implementations
Existing Fibre Channel Datacenter Solution

- Hadoop Cluster
- RDBMS (Transaction and data warehouse)
- Legacy servers
- Fibre Channel SAN
- StorageTek SL8500
- Oracle Hadoop Connectors
- Oracle SAM
- Fibre Channel datacenter backbone

2012 SNIA Analytics and Big Data Summit. © Oracle. All Rights Reserved.
Agenda

- Big Data Using Hadoop, The First Wave
- Big Data Today, The Second Wave
- A Storage Model for Big Data
- Implementations
- Plan for Action
- Summary
Call to Action I
Implementing a Big Data Storage Architecture

- Start a Big Data Proof of Concept (POC)
  - Small Hadoop cluster using Cloudera CDH5
  - R for statistical computing
  - Build base expertise through the community
Call to Action II
Implementing a Big Data Storage Architecture

- Complete a current data inventory
  - Type and location and profile (value)
  - Organic growth projection
- Project ingest requirements from contemporary sources
  - The three V’s: variety, volume, velocity
- Project your data retention needs
  - Consider keeping ALL your data
- Assess need for specialized analytics and visualization
Call to Action III
Implementing a Big Data Storage Architecture

- Complete a data flow analysis
  - Quantify data movement
  - Map to datacenter assets to identify requirements
    - Storage and network
  - Optimize flow and cost across processing engines
    - Identify and test software components
- Align data value with cost
Agenda

- Big Data Using Hadoop, The First Wave
- Big Data Today, The Second Wave
- A Storage Model for Big Data
- Implementations
- Call to Action
- Summary
Big Data Storage Implementations
Greenfield, Oracle Purpose-Aligned Integrated Solution

- Oracle Big Data Appliance
- Oracle Exadata
- Oracle Exalytics
- Pillar Axiom
- Sun ZFS
- StorageTek

Oracle Big Data Connectors
Oracle SAM
InfiniBand datacenter backbone

2012 SNIA Analytics and Big Data Summit. © Oracle. All Rights Reserved.
Big Data Storage Implementations
Virtualized, Network Infrastructure

Shared storage

Hadoop or NoSQL Cluster

Xsigo Fabric

Data management software

RDBMS (Transaction and data warehouse)

Legacy servers and clusters

Archival storage (Tape)

Data management software
Big Data Storage Implementations
Moving the Processing to the Data

Shared storage with embedded Hadoop engines

Data management software

RDBMS (Transaction and data warehouse)

Legacy servers and clusters

Archival storage (Tape)
<table>
<thead>
<tr>
<th>Function</th>
<th>Oracle Component</th>
<th>Open source POC</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured data ingest</td>
<td>Apache Flume (Big Data Appliance)</td>
<td>Apache Flume</td>
<td>Hadoop startups</td>
</tr>
<tr>
<td>Structured data ingest (ELT)</td>
<td>Oracle Data Integrator</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDBMS to HDFS</td>
<td>Apache Sqoop (Big Data Appliance)</td>
<td>Apache Sqoop</td>
<td>Hadoop startup Custom DBs</td>
</tr>
<tr>
<td>HDFS to RDBMS</td>
<td>Oracle Hadoop Connectors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDFS to/from shared storage</td>
<td>Custom</td>
<td>Custom</td>
<td>Custom/Vendor</td>
</tr>
<tr>
<td>RDBMS to/from shared storage</td>
<td>Hybrid columnar compression (HCC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tiered storage management</td>
<td>Oracle SAM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

A Big Data Storage Model

- Assess the value of your data
- Design for an integrated whole...
  … with consideration for each function…
  ….ingest, consolidation, analytics and viz, value retention
- …across a high performance network.
- Optimize data usage and movement.
- Incorporate existing data center architecture
- Store in concert with value
Thank You

Dave “Sunny” Sundstrom
Sunny.sundstrom@oracle.com
Oracle’s Storage Strategy
Key Elements in Big Data Implementations

- Quality-of-Service for multi-application consolidation
- End-to-end data transport across IB networks without copy
- Automated, policy based tiering across flash, SSD and tape
- Integration with Oracle applications and management
- Easy, in-place upgradeability from small to large storage systems
- Critical-path data access optimization
- Advanced fault prevention, detection and correction
- Maximum archival media life and automated data verification
- Data compression to disk and tape

..............