This Symposium brought to you by

www.ttcus.com

Linkedin/Group: Technology Training Corporation

@Techtrain

Technology Training Corporation

www.ttcus.com
Architecture Matters: One Platform for Enterprise Hadoop
MapR Distribution for Apache Hadoop

- Complete Hadoop distribution
- Comprehensive management suite
- Industry-standard interfaces
- Combines open source packages with enterprise-grade dependability
- Higher performance
MapR: The Enterprise-Grade Distribution
The Cloud Leaders Pick MapR

Amazon EMR is the largest Hadoop provider in revenue and # of clusters

Google chose MapR to provide Hadoop on Google Compute Engine

Deploying OpenStack? MapR is partnering with Canonical and Mirantis on OpenStack support.
One Platform for Big Data

Batch
Log file Analysis
Data Warehouse Offload
Fraud Detection
Clickstream Analytics

Interactive
Forensic Analysis
Analytic Modeling
BI User Focus

Real-Time
Sensor Analysis
“Twitterscraping”
Telematics
Process Optimization

Map Reduce
File-Based Applications
SQL
Database
Search
Stream Processing

99.999% HA
Data Protection
Disaster Recovery
Scalability & Performance
Enterprise Integration
Multi-tenancy
Real-time learning at major credit-card

- Machine learning is normally batch based
  - limits responsiveness
  - also limits how fast we can detect change points

- Current on-line tools do not integrate well with batch analysis systems

- 8+ separate clusters, 55-60 nodes
  - Web servers
  - Storm nodes
  - Kafka nodes
  - 8-10 Hadoop
  - NAS for web storage
  - ...

Unprocessed Data

Fully processed

Latest full period

Hadoop job takes this long for this data

©MapR Technologies - Confidential
Major CC plugs hole in Hadoop with MapR

- Real-time data with limited state
- Exactly what Storm does
- And what Hadoop does not

- With MapR – One Platform
  - 15-20 nodes total
  - Any node does any job
  - Full HA, backups
  - scales indefinitely
Interactive SQL Initiatives for Hadoop

- **Spark**: Shark/Spark Interactive queries, HiveQL
- **HADAPT**: SQL based analytics
- **APACHE DRILL**: Interactive queries, ANSI SQL 2003
- **Impala**: Interactive queries, HiveQL
- **HIVE**: SQL conversion to MapReduce
M7 – Enterprise-Grade NoSQL on Hadoop

- NoSQL Columnar Store
- HBase API
- Integrated with Hadoop

HBase
JVM
HDFS
JVM
ext3/ext4
Disks

Other Distros

MapR M7
Disks

©MapR Technologies - Confidential
Tradeoffs with Other NoSQL Solutions

24x7 applications with strong data consistency

Reliability

Performance
Continuous low latency with horizontal scaling

Easy Administration
Easy day-to-day management with minimal learning curve
## Bullet-proof NoSQL with Zero Administration

### Performance
- Over 1 million ops/sec with 10 node cluster

### Reliability
- No I/O storms, No compactions

### Easy Administration
- Zero Administration
  - No processes to manage, automated splits, self-tuning
- 24x7 Applications
  - Instant recovery, online schema modification, snapshots, mirroring
- High Scalability
  - 1 trillion tables
- Low TCO
  - Files and tables on one platform
M7: Tables for Developers

• Users can create and manage their own tables
  – Unlimited # of tables
  – First copy local

• Tables can be created in any directory
  – Tables count towards volume and user quotas

• No admin intervention needed
  – Perform tasks on the fly
  – No stopping/restarting of servers

• Automatic data protection and disaster recovery
  – Users can recover from snapshots/mirrors on their own
M7: Volume Based Data Management

/user/john
/user/dave
/project/ads

Quota
Snapshot

Mirror
Mirror
Case Study – 24x7 Applications

Web 2.0 company that optimizes email, mobile & social campaigns

Apache HBase

Service Disruptions
- 42 hours of compaction every weekend
- Long cold-starts
- “Engineering resources stuck fixing HBase problems”

24x7 Uptime
- No compactions
- Instant recovery
- Easy development and administration

©MapR Technologies - Confidential
Case Study – Cost Effective Scalability

Managed Security Services company that provides solutions for data security and compliance

- Limited analytical tools
- No machine learning capability

Expensive Data Store

- Extended analytical ecosystem: Machine Learning, Solr and Hive
- Similar Reliability

Cost Effective Scalability
# MapR M7 Accelerates HBase Applications

## MapR speedup with HDDs: 4x-7.2x

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>MapR 3.0.1 (M7)</th>
<th>CDH 4.3.0 (HBase)</th>
<th>MapR Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% read, 50% update</td>
<td>7086</td>
<td>1695</td>
<td>4x</td>
</tr>
<tr>
<td>95% read, 5% update</td>
<td>3716</td>
<td>602</td>
<td>6x</td>
</tr>
<tr>
<td>Reads</td>
<td>5520</td>
<td>764</td>
<td>7.2x</td>
</tr>
<tr>
<td>Scans (50 rows)</td>
<td>922</td>
<td>161</td>
<td>5.7x</td>
</tr>
</tbody>
</table>

**CPU:** 2 x Intel Xeon CPU E5645 2.40GHz 12 cores  
**RAM:** 48GB  
**Disk:** 12 x 3TB (7200 RPM)  
**Record size:** 1KB  
**Data size:** 2TB  
**OS:** CentOS Release 6.2 (Final)

## MapR speedup with SSD: 5x-11.3x

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>MapR 3.0.1 (M7)</th>
<th>CDH 4.3.0 (HBase)</th>
<th>MapR Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% read, 50% update</td>
<td>20856</td>
<td>2547</td>
<td>8.1x</td>
</tr>
<tr>
<td>95% read, 5% update</td>
<td>13455</td>
<td>2660</td>
<td>5x</td>
</tr>
<tr>
<td>Reads</td>
<td>18206</td>
<td>1605</td>
<td>11.3x</td>
</tr>
<tr>
<td>Scans (50 rows)</td>
<td>1135</td>
<td>116</td>
<td>9.8x</td>
</tr>
</tbody>
</table>

**CPU:** 2 x Intel Xeon CPU E5620 2.40GHz 8 cores  
**RAM:** 24GB  
**Disk:** 1 x 1.2TB Fusion I/O ioDrive2  
**Record size:** 1KB  
**Data size:** 600GB  
**OS:** CentOS Release 6.3 (Final)
MapR M7 vs. CDH – Mixed Load (50-50)

YCSB Mixed (50% Update-50% Read) Test (10Nodes)

Source: 2TB (1K RowSize)

Read Latency ONLY: 10-sec Moving Average & y-Axis Cap=400msec
# MapR Advantages

<table>
<thead>
<tr>
<th>Advantage</th>
<th>M7</th>
<th>OTHERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.999% High Availability</td>
<td>✔</td>
<td>☒</td>
</tr>
<tr>
<td>Instant Recovery from Failures</td>
<td>✔</td>
<td>☒</td>
</tr>
<tr>
<td>Continuous Low Latency (No Compactions)</td>
<td>✔</td>
<td>☒</td>
</tr>
<tr>
<td>Zero Administration (No Processes to Manage, Self-tuning)</td>
<td>✔</td>
<td>☒</td>
</tr>
<tr>
<td>Online Data Protection (Snapshots, Mirroring)</td>
<td>✔</td>
<td>☒</td>
</tr>
<tr>
<td>Scalability (Number of Tables Supported)</td>
<td>Trillion</td>
<td>Hundreds</td>
</tr>
</tbody>
</table>

©MapR Technologies - Confidential
Thank You!