Saving Millions through Data Warehouse Offloading to Hadoop

Jack Norris, CMO MapR Technologies
MapR Technologies Overview

- Open, enterprise-grade distribution for Hadoop
  - Easy, dependable and fast
  - Open source with standards-based extensions

- Background of team
  - Deep management bench with extensive analytic, storage, virtualization, and open source experience
  - Google, Microsoft, EMC, Network Appliance, Transarc, Apache

- MapR is deployed at 1000’s of companies
  - From small Internet startups to the world’s largest enterprises

- MapR customers analyze massive amounts of data:
  - Hundreds of billions of events daily
  - 90% of the world’s Internet population monthly
  - $1 trillion in retail purchases annually
Arrival of Big Data Impacts Data Warehouse

- **Volume**: Prohibitively expensive storage costs
- **Velocity**: Faster arrival and processing needs
- **Variety**: Inability to process unstructured formats

The impact on the Data Warehouse is illustrated by the interplay of these three dimensions.
Top Concern for Big Data

“Too many different types, sources & formats of critical data”

Multiple Data sources

Multiple technologies

Multiple copies of data
The Hadoop Advantage

- Fueling an industry revolution by providing infinite capability to store and process big data

- Expanding analytics across data types

- Compelling economics
  - 20 to 100X more cost effective than alternatives
Important Drivers for Hadoop

- Compute on data drives efficiencies
- You don’t need to know what questions to ask beforehand
- More data beats better models
- Powerful ability to analyze unstructured data

Pioneered at Google

©MapR Technologies . All rights reserved.
Hadoop Growth
How Do You Lower & Control Data Warehouse Costs?

Source Data

- Transactions, OLTP, OLAP
- Documents and Emails
- Social Media, Web Logs
- Machine Device, Scientific

Traditional Targets

- Databases and data warehouses are exceeding their capacity too quickly
- Batch windows hitting their limits putting SLAs at risk
- Raw data or infrequently used data consuming capacity

Batch ETL
Lower Data Management Costs

Source Data
- Transactions, OLTP, OLAP
- Documents and Emails
- Social Media, Web Logs
- Machine Device, Scientific

Informatica
- Archive
- Profile
- Parse
- Transform
- Cleanse
- Match

Traditional Targets
- Enterprise Data Warehouse
- RDBMS
- MDM

MapR Technologies
- Access & Ingest
- Parse & Prepare
- Discover & Profile
- Transform & Cleanse
- Extract & Deliver

©MapR Technologies . All rights reserved.
Leverage Familiar Tools With MapR Cluster

No-code visual development environment

Preview results at any point in the data flow
Maximize Your Return on Data

1. Lower Costs
   - Order of magnitude savings with MapR
   - Optimized end-to-end performance
   - Rich pre-built connectors, library of transforms for ETL, data quality, parsing, profiling

2. Increase Productivity
   - Up to 5x productivity gains with no-code visual development environment
   - No need for Hadoop expertise for data integration

3. Increase ROI
   - Ability to scale platform easily with demand
   - Leverage for additional analytics
Customer Case Study

Business Impact:
- Saved millions in TCO
- 10x faster, 100x cheaper
- Maintain the same SLAs
- Implemented the change without impacting users
Data Warehouse Offload: Cost Savings + Analytics

Benefits:
- Both structured and unstructured data
- Expanded analytics with MapReduce, NoSQL, etc.

<table>
<thead>
<tr>
<th>Solution</th>
<th>Cost / Terabyte</th>
<th>Hadoop Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop</td>
<td>$333</td>
<td></td>
</tr>
<tr>
<td>Teradata Warehouse Appliance</td>
<td>$16,500</td>
<td>50x savings</td>
</tr>
<tr>
<td>Oracle Exadata</td>
<td>$14,000</td>
<td>42x savings</td>
</tr>
<tr>
<td>IBM Netezza</td>
<td>$10,000</td>
<td>30x savings</td>
</tr>
</tbody>
</table>
What is the Best Way to Deploy Hadoop?

**Transitory Data Store**
- No long-term scale advantages
- Unprotected data
- ETL Tool focus

**Permanent Data Store**
- Highly available and fully protected data
- Works with existing tools
- Real-time ingestion and extraction
- Archive data from data warehouse

Enterprise Data Hub
An Enterprise Data Hub

- Combine different data sources
- Minimize data movement
- One platform for analytics
Key Elements of Enterprise Data Hub

Enterprise-grade data platform for the long-term

- Reliability to support stringent SLAs
- Protection from data loss and user or application errors
- Support business continuity and meet recovery objectives
High Availability and Dependability

Reliable Compute

- Automated stateful failover
- Automated re-replication
- Self-healing from HW and SW failures
- Load balancing
- Rolling upgrades
- No lost jobs or data
- 99999’s of uptime

Dependable Storage

- Business continuity with snapshots and mirrors
- Recover to a point in time
- End-to-end check summing
- Strong consistency
- Data safe
- Mirror across sites to meet Recovery Time Objectives
Data Protection with Snapshots

- Snapshots without data duplication
- Save space by sharing blocks
- Lightning fast
- Zero performance loss on writing to original
- Scheduled, or on-demand
- Easy recovery by user
Disaster Recovery with Mirroring

Business Continuity and Efficiency

Efficient design
- Differential deltas
- Compressed and check-summed

Easy to manage
- Scheduled or on-demand
- WAN, Remote Seeding
- Consistent point-in-time
Key Elements of Enterprise Data Hub

Enterprise-grade platform for the long-term

- Consistent performance with large horizontal scalability
- Integrate with existing applications and tools
- Support multiple users, groups & applications securely
World Record Performance and Scale

MinuteSort World Record

1.5 TB in 1 minute
2103 nodes
Direct Integration with Existing Applications

- 100% POSIX compliant
- Industry standard APIs - NFS, ODBC, LDAP, REST
- More 3rd party solutions
- Proprietary connectors unnecessary
- Language neutral
When Hadoop Looks Like Enterprise Storage

- Data ingestion is easy
  - Popular online gaming company changed data ingestion from a complex Flume cluster to a 17-line Python script

- Database bulk import/export with standard vendor tools
  - Large telco saved tens of millions on Teradata costs by leveraging MapR to pre-process data prior to loading into Teradata

- 1000s of applications/tools
  - Large credit card company uses MapR volumes as the user home directories on the Hadoop gateway servers

```
$ find . | grep log
$ cp
$ vi results.csv
$ scp
$ tail -f part-00000
```
Multi-tenant Capabilities to Share a Cluster Successfully

- **Isolation**
  - Data placement control
  - Label based job scheduling

- **Quotas**
  - Storage, CPU, Memory

- **Security and delegation**
  - ACLs
  - AD, LDAP, Linux PAM

- **Reporting**
  - About 70 resource usage metrics
  - REST API integration
Enterprise Data Hub Supports a Range of Applications

**Batch**
- 99.999% HA
- Self-Healing
- Instant recovery

**Interactive**
- Data Protection
- Snapshots for point in time recovery from user or application errors
- Disaster Recovery
- Mirroring across clusters and the WAN
- Scalability & Performance
- Unlimited files & tables
- Record setting performance

**Real-time**
- Enterprise Integration
- Direct data ingestion and access
- Fully compliant ODBC access and SQL-92 support
- Multi-tenancy
- Secure access to multiple users and groups
Expand Data For Existing Applications

- **Network security**: Network IDS with a 3-day window instead of a 10-minute window

- **Trade Surveillance**: Rogue trader detection on intraday instead of end-of-day market data

- **Insurance**: Calculate risk triangles for individual properties instead of neighborhoods

**Advantages:**

- 1T files and tables
- Real-time data ingestion with streaming writes
- 24x7 operations with automated failure recovery
- Better hardware utilization with 2x performance
Combine Different Data Sources

Streaming writes to Hadoop

Advantages:

- Exponential decrease in time to market
- Real-time data ingestion with streaming writes
- 1T files and tables
- 24x7 operations with automated failure recovery
New Analytics

• Enhanced search
• Real-time event processing
• MapReduce-enabled machine learning algorithms

Advantages

✓ Increased ROI with 2x performance
✓ High available, fully data protected environment
✓ Multiple users running different jobs on one cluster
One Platform with a Range of Functionality

Batch → Interactive → Real-time

- Map Reduce
- File-Based Applications
- SQL
- Database
- Search
- Stream Processing

99.999% HA
Data Protection
Disaster Recovery
Scalability & Performance
Enterprise Integration
Multi-tenancy
One Platform with a Range of Functionality

- Batch
- Interactive
- Real-time

- Map Reduce
- SQL
- Database
- Search
- Stream Processing

- 99.999% HA
- Data Protection
- Disaster Recovery
- Scalability & Performance
- Enterprise Integration
- Multi-tenancy
Apache Drill Overview

- Apache Drill Alpha now available
- Community-driven project
- Apache Drill supports Centralized and Self Describing
  - Centralized Schema Management
    - HCatalog - Hive metastore
    - Centralized data structure
    - DBA defines and maintains structure
  - Self Describing Data
    - Structure stored with data
    - Document database orientation
    - Like MongoDB

Project Contributors

MapR
Pentaho
Oracle
VMWare
Microsoft
Thoughtworks
UT Austin
UW Madison
RJMetrics
XingCloud
NoSQL Applications

Batch → Interactive → Real-time

- MapReduce
- File-Based Applications
- SQL
- Database
- Search
- Stream Processing
- 99.999% HA
- Data Protection
- Disaster Recovery
- Scalability & Performance
- Enterprise Integration
- Multi-tenancy
M7 – Enterprise-Grade NoSQL on Hadoop

- NoSQL Columnar Store
- HBase API
- Integrated with Hadoop
Tradeoffs with 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
High Performance with Consistent Low Latency

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

Source: 2TB (1K RowSize)

Read Latency ONLY: 10-sec Moving Average & y-Axis Cap=400msec
Search Applications

Batch → Interactive → Real-time

- Map Reduce
- File-Based Applications
- SQL
- Database
- Search
- Stream Processing

99.999% HA
Data Protection
Disaster Recovery
Scalability & Performance
Enterprise Integration
Multi-tenancy
The Need for Integrated Search

Hadoop Advantages

• Scale
• Distributed processing
• Extended analytics such as machine learning

Search Requirements for Hadoop

• Ability to randomly read and write indices over Hadoop
• Store and process large number of small files for indices
• High availability for data as well as indices
Integration Issues for Search

- **Support** for Steaming Writes
- **Capability to easily integrate with other Hadoop components**
- **Point-in-time** recovery from errors
- **Data Versioning for Modeling**
- **Store index directly on Hadoop** instead of local file system
- **Leverage scale, replication and data management**
- **Enterprise-grade Hadoop**
Streaming Applications

Batch → Interactive → Real-time

- Map Reduce
- File-Based Applications
- SQL
- Database
- Search
- Stream Processing

99.999% HA
Data Protection
Disaster Recovery
Scalability & Performance
Enterprise Integration
Multi-tenancy
Other Distributions Require Separate Clusters

Twitter

Twitter API

TwitterLogger

Kafka Cluster

Kafka

Kafka API

Storm

Flume

Hadoop

Flume

HDFS Data

Hadoop

Web Data

NAS

Web-server

http

©MapR Technologies . All rights reserved.
Enterprise Hub with MapR

Other Distributions
- 8+ separate Cluster
- 20-25 nodes
  - 3 Kafka nodes; >2 TwitterLogger; 5-10 Hadoop; >3 Storm; 3 zookeepers; NAS for web storage; >2 web servers

MapR – One Platform
- 5-10 nodes total, any node does any job
- Full HA included
- Snapshots included
- Disaster recovery included
Advantages for Enterprise Data Hub

• Enterprise Grade Platform
  • 99.999% HA
  • Full data protection
  • Disaster recovery
• Easiest Integration
  • Industry-standard interfaces: NFS, ODBC, LDAP, REST
  • Streaming writes
• Best ROI
  • Faster time to market
  • Eliminate risk
  • Reuse existing apps and tools
An Enterprise Data Hub

- Combine different data sources
- Minimize data movement
- One platform for analytics
MapR: Open Source + Innovation

Apache Hadoop

Apache Hadoop

Apache Hadoop

Management Innovations

Management Innovations

Infrastructure Innovations
Q & A

Engage with us!

@mapr
mapr-technologies
john@maprtech.com

maprtech
MapR
maprtech