Enable your Modern Data Architecture by delivering Enterprise Apache Hadoop

**Open Leadership**
Drive innovation in the open exclusively via the Apache community-driven open source process

**Enterprise Rigor**
Engineer, test and certify Apache Hadoop with the enterprise in mind

**Ecosystem Endorsement**
Focus on deep integration with existing data center technologies and skills

Headquartered in Palo Alto, CA; 300+ employees and growing

Reseller Partners:
A data architecture under pressure from new data

- **APPLICATIONS**
  - Business Analytics
  - Custom Applications
  - Packaged Applications

- **DATA SYSTEM**
  - RDBMS
  - EDW
  - MPP

- **REPOSITIES**
  - Existing Sources (CRM, ERP, Clickstream, Logs)

- **SOURCES**
  - OLTP, ERP, CRM Systems
  - Unstructured documents, emails
  - Server logs
  - Sentiment, Web Data
  - Sensor, Machine Data
  - Geolocation
  - Clickstream

- **Statistics**
  - 2.8 ZB in 2012
  - 85% from New Data Types
  - 15x Machine Data by 2020
  - 40 ZB by 2020

Source: IDC
Hadoop within an emerging Modern Data Architecture
Hadoop: typically used for new analytic applications…
and incrementally delivers a ‘Data Lake’

Data Lake
An architectural shift in the data center that uses Hadoop to deliver deeper insight across a large, broad, diverse set of data at efficient scale
Hadoop Value: New types of data

**Sentiment**
Understand how your customers feel about your brand and products – right now

**Clickstream**
Capture and analyze website visitors' data trails and optimize your website

**Sensors**
Discover patterns in data streaming automatically from remote sensors and machines

**Geographic**
Analyze location-based data to manage operations where they occur

**Server Logs**
Research logs to diagnose process failures and prevent security breaches

**Unstructured**
Understand patterns in files across millions of web pages, emails, and documents
Hadoop unlocks a new approach: Iterative Analytics

**Current Reality**
Apply schema on write
Dependent on IT

**Augment w/ Hadoop**
Apply schema on read
Support range of access patterns to data stored in HDFS: polymorphic access

**SQL**
Single Query Engine
Repeatable Linear Process

- Determine list of questions
- Design solutions
- Collect structured data
- Ask questions from list
- Detect additional questions

**Hadoop**
Multiple Query Engines
Iterative Process: Explore, Transform, Analyze

- Batch
- Interactive
- Real-time
- Streaming
New analytic applications for new types of data

**Financial Services**
- New Account Risk Screens
- Fraud Prevention
- Trading Risk
- Maximize Deposit Spread
- Insurance Underwriting
- Accelerate Loan Processing

**Retail**
- 360° View of the Customer
- Analyze Brand Sentiment
- Localized, Personalized Promotions
- Website Optimization
- Optimal Store Layout

**Telecom**
- Call Detail Records (CDRs)
- Infrastructure Investment
- Next Product to Buy (NPTB)
- Real-time Bandwidth Allocation
- New Product Development

**Manufacturing**
- Supplier Consolidation
- Supply Chain and Logistics
- Assembly Line Quality Assurance
- Proactive Maintenance
- Crowdsourced Quality Assurance

**Healthcare**
- Genomic data for medical trials
- Monitor patient vitals
- Reduce re-admittance rates
- Store medical research data
- Recruit cohorts for pharmaceutical trials

**Utilities, Oil & Gas**
- Smart meter stream analysis
- Slow oil well decline curves
- Optimize lease bidding
- Compliance reporting
- Proactive equipment repair
- Seismic image processing

**Public Sector**
- Analyze public sentiment
- Protect critical networks
- Prevent fraud and waste
- Crowdsource reporting for repairs to infrastructure
- Fulfill open records requests
Hadoop delivers compelling economics

**EDW Optimization**

Current Reality + Augment w/ Hadoop

- EDW at capacity: some usage from low value workloads
- Older data archived, unavailable for ongoing exploration
- Source data often discarded

Free up EDW resources from low value tasks
- Keep 100% of source data and historical data for ongoing exploration
- Mine data for value after loading it because of schema-on-read

**Commodity Compute & Storage**

Hadoop Enables Scalable Compute & Storage at a Compelling Cost Structure

<table>
<thead>
<tr>
<th>Cloud Storage</th>
<th>HADOOP</th>
<th>Fully-loaded Cost Per Raw TB of Data (Min–Max Cost)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud Storage</td>
<td>HADOOP</td>
<td></td>
</tr>
<tr>
<td>Cloud Storage</td>
<td>HADOOP</td>
<td>NAS</td>
</tr>
</tbody>
</table>
Delivering the Core Capabilities of Enterprise Hadoop
Core Capabilities of Enterprise Hadoop

<table>
<thead>
<tr>
<th>PRESENTATION &amp; APPLICATION</th>
<th>ENTERPRISE MGMT &amp; SECURITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enable both existing and new application to provide value to the organization</td>
<td>Empower existing operations and security tools to manage Hadoop</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GOVERNANCE &amp; INTEGRATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load data and manage according to policy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DATA ACCESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access your data simultaneously in multiple ways (batch, interactive, real-time)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DATA MANAGEMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store and process all of your Corporate Data Assets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SECURITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provide layered approach to security through Authentication, Authorization, Accounting, and Data Protection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deploy and effectively manage the platform</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DEPLOYMENT OPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provide deployment choice across physical, virtual, cloud</td>
</tr>
</tbody>
</table>
Core Technologies of Enterprise Hadoop

GOVERNANCE & INTEGRATION
- Data Workflow, Lifecycle & Governance
  - Falcon
  - Sqoop
  - Flume
  - NFS
  - WebHDFS

DATA ACCESS
- Batch: Map Reduce
- Script: Pig
- SQL: Hive/Tez, HCatalog
- NoSQL: HBase, Accumulo
- Stream: Storm
- Search: Solr
- Others: In-Memory Analytics, ISV engines

YARN: Data Operating System

SECURITY
- Authentication
- Authorization
- Accounting
- Data Protection
- Storage: HDFS
- Resources: YARN
- Access: Hive, ...
- Pipeline: Falcon
- Cluster: Knox

DATA MANAGEMENT

HDFS
(Hadoop Distributed File System)

OPERATIONS
- Provision, Manage & Monitor
  - Ambari
  - Zookeeper

- Scheduling
  - Oozie
Hortonworks provides leadership to Hadoop

<table>
<thead>
<tr>
<th>Apache Project</th>
<th>Committers</th>
<th>PMC Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>Tez</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Hive</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>HBase</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Pig</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Sqoop</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ambari</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>Knox</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Falcon</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Oozie</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Zookeeper</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Flume</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Accumulo</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Storm</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>103</strong></td>
<td><strong>49</strong></td>
</tr>
</tbody>
</table>

Total Net Lines Contributed to Apache Hadoop

- Hadoop: 147,933
- Tez: 3
- Hive: 6
- HBase: 3
- Pig: 5
- Sqoop: 0
- Ambari: 12
- Knox: 2
- Falcon: 2
- Oozie: 2
- Zookeeper: 1
- Flume: 0
- Accumulo: 2
- Storm: 0

Total Number of Committers to Apache Hadoop

- End Users: 449,768
- 25
- 10 Others

- 7 Cloudera
- 5 Facebook
- 3 LinkedIn
- 3 IBM
- 10 Yahoo
The Partners You Rely On, Rely On Hortonworks for Hadoop
Hadoop is wholly integrated into the data center

DEPTH
Hortonworks engages in deep engineered relationships with the leaders in the data center, such as Microsoft, Teradata, Redhat & SAP

BREADTH
Hundreds of partners work with us to certify their applications to work with Hadoop so they can extend big data to their users
Financial Services High Level Use Cases

[Diagram showing data sources, data repos, analysis & visualization, and associated technologies]
Telco High Level Use Cases

DATA REPOS
- EDW
- MPP

DATA SOURCES
- CRM
- ERP
- Billing Data
- Subscriber Data
- Product Catalog
- Network Data
- Clickstream
- Online Chat
- Sensor Data
- Social Media
- Server Logs
- Call Detail Records
- Merchant Listings
- DMP

ANALYSIS
- Operational dashboards
- Customer scorecards
- CDR analysis
- Proactive maintenance
- Infrastructure investment
- Bandwidth allocation
- Product development

Hortonworks
- Ingest
  - Sqoop
  - Flume
  - HDFS Put
  - Web HDFS

HDP Hadoop Cluster
- Metadata Management: HCatalog
- Multitenant Processing: YARN
- (Hadoop Operating System)
- Linear Scale
  - Compute & HDFS Storage

Governance, tag, filter & process

Operations (Ambari)

Security (Knox, Hive, HDFS, etc...)

THE POWER TO KNOW.
HDP 2.1: Reliable, Consistent & Current

HDP certifies most recent & stable community innovation
Hortonworks process for Enterprise Hadoop

Upstream Community Projects

- Apache Hadoop
- Apache HBase
- Apache Hive
- Apache Pig
- Apache Falcon
- Apache Knox
- Apache Storm

Downstream Enterprise Product

Certified at scale using the most advanced Hadoop test bed on the planet

- 1000’s of production nodes at Yahoo!
- Over 1500 unit & system tests

Virtuous cycle when development & fixed issues done upstream & stable project releases flow downstream
Working in the community for the enterprise

**Hortonworks de-risks your investment...**

All support and implementation backed by the largest and most experienced Hadoop team on the planet

**A Two Way Street**

- Gain access to the expertise and knowledge of the Hadoop community
- All issues or feature requests represented in the community
Hortonworks: A Leader In Hadoop

The Forrester Wave™: Big Data Hadoop Solutions, Q1 2014

“Hortonworks loves and lives open source innovation”

Vision & Execution for Enterprise Hadoop.
Hortonworks leads with a strong strategy and roadmap for open source innovation with Hadoop and a strong delivery of that innovation in Hortonworks Data Platform.

World Class Support and Services.
Hortonworks’ Customer Support received a maximum score and was significantly higher than both Cloudera and MapR.

Key Strategic Partnerships.
Hortonworks unique strategic partnerships with Microsoft, SAP, Teradata and others are a key strength as part of its overall strategy of ecosystem partnership to accelerate Hadoop adoption in the enterprise.
The value of Open...

Connect With the Community
We employ a large number of Apache project committers & innovators so that you are represented in the open source community.

Avoid Vendor Lock In
Hortonworks Data Platform is close to the open source trunk as possible and is developed 100% in the open so you are never locked in.

Partnerships that Matter
We work with partners to deeply integrate Hadoop with data center technologies so you can leverage existing skills and investments.

Certified for the Enterprise
We engineer, test and certify the Hortonworks Data Platform at scale to ensure reliability and stability you require for enterprise use.

Support from the Experts
We provide the highest quality of support for deploying at scale. You are supported by hundreds of years of Hadoop experience.
Thank You

Jeff Markham
Technical Director, APAC
jmarkham@hortonworks.com