Turning Data into Capital with Hadoop

Brian Marshall
VP, Corporate Development
Hortonworks
Big Data, Hadoop and the Modern Data Architecture

How do you define big data?

Will the term “big data” go away?

Are big data and Hadoop synonymous?

Hi, my name is Hadoop.
Quick Audience Poll

Which best describes how your org is using Hadoop?

A. Hadoop is fully integrated  
B. We’re in the process of getting Hadoop integrated  
C. We’re using Hadoop standalone, but we plan to integrate  
D. We’re using Hadoop standalone, and do not plan to integrate  
E. We don’t have Hadoop installed  
F. I don’t know
Traditional Systems under Pressure

1. Challenges
   - Constrains data to app
   - Can't manage new data
   - Costly to Scale

2. New Data

   Business Value

   ERP | CRM | SCM

   2012
   2.8 Zettabytes

   INDUSTRY LEADERS
   2020
   40 Zettabytes

   LAGGARDS

   Clickstream
   Geolocation
   Web Data
   Internet of Things
   Docs, emails
   Server logs
Hadoop Emerged as Foundation of New Data Architecture

Apache Hadoop is an open source data platform for managing large volumes of high velocity and variety of data

- Built by Yahoo! to be the heartbeat of its ad & search business
- Donated to Apache Software Foundation in 2005 with rapid adoption by large web properties & early adopter enterprises
- Incredibly disruptive to current platform economics

Traditional Hadoop Advantages

✓ Manages new data paradigm
✓ Handles data at scale
✓ Cost effective
✓ Open source

Traditional Hadoop Had Limitations

✗ Batch-only architecture
✗ Single purpose clusters, specific data sets
✗ Difficult to integrate with existing investments
✗ Not enterprise-grade
YARN is the Data Operating System of Hadoop

HDP is uniquely built around YARN serving as a data operating system that provides multi-tenant Resource Management, consistent Governance & Security and efficient Operations services across Hadoop applications.

Hortonworks Data Platform

YARN Data Operating System

- A centralized architecture of consistent enterprise services for resource management, security, operations, and governance.
- The versatility to support multiple applications and diverse workloads from batch to interactive to real-time, open source and commercial.

Key Benefits

- Multiple applications on a shared data set with consistent levels of service: a multitenant data platform.
- Provides a shared platform to enable new analytic applications.
- Delivers maximum cost efficiency for cluster resource management. Fewer servers fewer nodes.
Modern Data Architecture Unifies Data and Processing

Modern Data Architecture

- Enable applications to have access to all your enterprise data through an efficient centralized platform
- Supported with a centralized approach governance, security and operations
- Versatile to handle any all application and any data no matter the size or type
MDA Enables Move to Proactive Value Chains

- Retail: Mass branding → real-time personalization and 360° customer view
- Financial Services: Daily risk analysis → real-time trade surveillance & compliance analysis
- Healthcare: Mass treatment → proactive diagnostics and designer medicine
- Manufacturing: Break then fix → proactive maintenance
- Telco: Customer service silos → personalized quality of service & channel consolidation
Common Drivers of Hadoop Adoption

**Data Architecture Optimization**
- Low-value tasks and oft used data consume significant EDW resources.
- Offload to Hadoop for efficiently, freeing up your data warehouse to perform high-value functions like analytics and operations.

**Single View of Customer**
- HDP powers a single view of each customer, allowing organizations to provide targeted, personalized customer experiences.
- Single view reduces attrition, improves cross-sell and improves customer satisfaction.

**Predictive Analytics**
- HDP captures, stores and processes large volumes of data streaming from connected devices.
- Stream processing and data science help introduce new analytics for real-time and batch analysis.

**Data Discovery**
- HDP allows exploration of new data types and large data sets that were previously too big to capture, store & process.
- Unlock insights from data such as clickstream, geo-location, sensor, server log, social, text and video data.
Optimize Investments...with Dramatic Cost Savings

Current Reality

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

Augment w/ Hadoop

- 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

- Storage Costs/Compute Costs from $19/GB to $0.23/GB
- Fully-loaded Cost Per Raw TB of Data (Min–Max Cost)
# New Analytic Applications for New Types of Data

<table>
<thead>
<tr>
<th>INDUSTRY</th>
<th>USE CASE</th>
<th>Sentiment &amp; Web</th>
<th>Clickstream &amp; Behavior</th>
<th>Machine &amp; Sensor</th>
<th>Geographic</th>
<th>Server Logs</th>
<th>Structured &amp; Unstructured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial Services</td>
<td>New Account Risk Screens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trading Risk</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Insurance Underwriting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telecom</td>
<td>Call Detail Records (CDR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Infrastructure Investment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Real-time Bandwidth Allocation</td>
<td>☑</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retail</td>
<td>360° View of the Customer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Localized, Personalized Promotions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Website Optimization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manufacturing</td>
<td>Supply Chain and Logistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assembly Line Quality Assurance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crowd-sourced Quality Assurance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthcare</td>
<td>Use Genomic Data in Medical Trials</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monitor Patient Vitals in Real-Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharmaceuticals</td>
<td>Recruit and Retain Patients for Drug Trials</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improve Prescription Adherence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oil &amp; Gas</td>
<td>Unify Exploration &amp; Production Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monitor Rig Safety in Real-Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government</td>
<td>ETL Offload/Federal Budgetary Pressures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sentiment Analysis for Government Programs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Mainstream Enterprise Use Cases

<table>
<thead>
<tr>
<th>USE CASE</th>
<th>HOME IMPROVEMENT RETAILER</th>
<th>ECOMMERCE MARKETPLACE</th>
<th>AUTO &amp; PROPERTY INSURANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>USE CASE</strong></td>
<td>Enable a 360° customer view</td>
<td>Unify data for online recommendation engine</td>
<td>Speed analysis for usage-based insurance</td>
</tr>
<tr>
<td><strong>PROBLEM</strong></td>
<td>Unable to effectively engage customer across different channels</td>
<td>Siloed data sets hampering product recommendations</td>
<td>Incoming data pressuring existing systems &amp; processes</td>
</tr>
<tr>
<td></td>
<td>Siloed data repositories with increasing data storage costs</td>
<td>Data storage costs too high for non-transactional data</td>
<td>ETL captured 25% of the dataset and took 5-7 days to run</td>
</tr>
<tr>
<td><strong>SOLUTION</strong></td>
<td>Price optimization drives increased revenue</td>
<td>Increased sales through improved product recommendations</td>
<td>Faster time-to-insight with clickstream analytics</td>
</tr>
<tr>
<td></td>
<td>Data warehouse optimization lowers recurring expense</td>
<td>Optimized storage costs allows more data under management</td>
<td>ETL captures 100% of the dataset and takes 3 days or less to run</td>
</tr>
</tbody>
</table>
Hadoop Business Value: a Journey in Four Phases

Start small and grow over time...
End Game: Leaders are Data Driven

A mature Data Driven organization treats data like capital

Chief Financial Officer
- Managed for shareholder value
- Top-down mandate to allocate capital
- Cost control pushed to front-line
- Cost of capital
- Every dollar can be invested once
- Value compounds over time

Chief Data Officer
- Managed for shareholder value
- Top-down mandate for a data-driven organization
- Data-driven decision making pushed to front line
- Cost of data processing infrastructure and expertise
- Every data point can be processed many times
- Value compounds with use, as more use cases, sources, time periods join in a data lake
New Era Companies are Born Data Driven

12 Month Results at TrueCAR

- Six Production Hadoop Applications
- Sixty nodes/2PB data
- Storage Costs/Compute Costs from $19/GB to $0.23/GB

“We addressed our data platform capabilities strategically as a pre-cursor to IPO.”

12 months execution plan

June 2013
Begin Hadoop Execution

July 2013
Hortonworks Partnership

Aug 2013
Training & Dev Begins

Nov 2013
Production Cluster
60 Nodes
2 PB

Dec 2013
Three Production Apps
(3 total)

Jan 2014
40% Dev Staff Proficient

Feb 2014
Three More Production Apps
(6 total)
Hadoop is a Platform Decision

Adoption follows a consistent journey
Data architecture efficiencies, new analytic apps, and ultimately to a "data lake".

HDP subscription supports entire lifecycle
World class experience to ensure success from architecture to production to expansion.

HDP: A completely open data platform
Platforms are ultimately defined by open communities.

HDP: A centralized architecture built on YARN
Any application, any data, anywhere.
Early Majority Adoption is Happening Now
Hadoop Enables a Massive Market

GLOBAL HADOOP MARKET

- **Software**
- **Services**
- **Hardware**

$2 Billion to $50 Billion

58% CAGR

2013 to 2020

Source: AMR Research estimates
Questions

Brian Marshall
VP, Corporate Development
Hortonworks
bmarshall@hortonworks.com