Harnessing big data with Hortonworks Data Platform and Red Hat JBoss Data Virtualization

Kimberly Palko, Product Manager
Red Hat JBoss

Doug Reid, Director Partner Product Management
Hortonworks

Cojan van Ballegooijen, Solutions Architect
Red Hat
Big Data Market Trends

85% from new data types

40 ZB digital universe by 2020

15x growth rate of machine generated data by 2020

50x data growth 2010 to 2020

2.8 ZB data created & replicated in 2012

1 Zettabyte (ZB) = 1 Billion TBs

The US has 1/3 of the world’s data

Big Data is 1 of 5 US GDP Game Changers $325 billion incremental annual GDP from big data analytics in retail and manufacturing by 2020

65% analytic apps with Hadoop inside by 2015

Hadoop enabled DBMS’s

20% % by which org’s leveraging modern info management systems outperform peers by 2015
Hortonworks Profile

- ONLY 100% open source Apache™ Hadoop data platform
- Founded in 2011
- 1ST HADOOP distribution to go public
- IPO Fall 2014 (NASDAQ: HDP)
- 437 subscription customers
- 1000+ technology partners
- 600+ employees across 17 countries
Traditional Data Systems are Under Pressure…

Traditional Systems
- Data constrained to apps
- Can’t manage new data
- Costly to scale

Business Value

New Data, New Opportunity
- Industry leadership via full fidelity of data and analytics

2.8 Zettabytes in 2012

44 Zettabytes in 2020

Traditional Data Systems

Limited ability to innovate

New Data
- Clickstream
- Geolocation
- Web Data
- Internet of Things
- Files, emails
- Server logs

2012

2020

1 Zettabyte (ZB) = 1 million Petabytes (PB); Sources: IDC, IDG Enterprise, and AMR Research
A New Approach Is Needed

The goal: Turn data into value

The problem: Data architectures don’t scale

EXISTING Systems  Clickstream  Web & Social  Geolocation  Internet of Things  Server Logs  Files, Emails

NEW SOURCES

NEW VALUE

Costs  Silos  Data Structure
Big Data: From Reactive to Proactive Value Chains

<table>
<thead>
<tr>
<th>TRADITIONAL APPROACH</th>
<th>INDUSTRY INNOVATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Retail</strong></td>
<td></td>
</tr>
<tr>
<td>Mass branding</td>
<td>…real-time personalization and 360° customer view</td>
</tr>
<tr>
<td><strong>Financial Services</strong></td>
<td></td>
</tr>
<tr>
<td>Daily risk analysis</td>
<td>…real-time trade surveillance &amp; compliance analysis</td>
</tr>
<tr>
<td><strong>Healthcare</strong></td>
<td></td>
</tr>
<tr>
<td>Mass treatment</td>
<td>…proactive diagnostics and designer medicine</td>
</tr>
<tr>
<td><strong>Manufacturing</strong></td>
<td></td>
</tr>
<tr>
<td>Break then fix</td>
<td>…proactive maintenance</td>
</tr>
<tr>
<td><strong>Telco</strong></td>
<td></td>
</tr>
<tr>
<td>Customer service silos</td>
<td>…personalized quality of service &amp; channel consolidation</td>
</tr>
</tbody>
</table>
Hadoop Driver: Cost Optimization

HDP helps you reduce costs and optimize the value associated with your EDW

**Archive Data off EDW**
Move rarely used data to Hadoop as active archive, store more data longer

**Offload costly ETL process**
Free your EDW to perform high-value functions like analytics & operations, not ETL

**Enrich the value of your EDW**
Use Hadoop to refine new data sources, such as web and machine data for new analytical context
Hadoop Driver: Advanced Analytic Applications

**Single View:**
Improve acquisition & retention

- HDP enables 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:**
Identify next best action

- 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:**
Uncover new findings

- 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.
Hadoop Drivers and the Journey to a Data Lake

Data Lake Definition

- **Centralized Architecture**
  Multiple applications on a shared data set with consistent levels of service

- **Any App, Any Data**
  Multiple applications accessing all data affording new insights and opportunities.

- **Unlocks ‘Systems of Insight’**
  Advanced algorithms and applications used to derive new value and optimize existing value.

**Drivers:**
1. Cost Optimization
2. Advanced Analytic Apps

**Goal:**
- Centralized Architecture
- Data-driven Business

**SCOPE**

**SCALE**

Journey to the Data Lake with Hadoop
HDP is deeply integrated in the data center

- Enables millions of JBoss developers to quickly build applications with Hadoop
- Simplifies deployment of Hadoop on OpenStack
- Develops and deploys Apache Hadoop as integrated components of the open modern data architecture
Red Hat and Hortonworks

- Enables millions of JBoss developers to quickly build applications with Hadoop
- Simplifies deployment of Hadoop on OpenStack
- Develops and deploys Apache Hadoop as integrated components of the open modern data architecture
Red Hat + Hortonworks
Deliver Open Source Modern Data Architecture

A deeper strategic alliance
– Engineer solutions for seamless customer experience
– Joint go to market activities
– Integrated customer support

Available now
– HDP 2.1 on Red Hat Storage 3.0.2
– Hadoop Plug-in Refresh Release and Ambari
– Red Hat JBoss Data Virtualization with HDP
– HDP 2.2 on Red Hat Enterprise Linux with OpenJDK
Modern Data Architecture + Red Hat Data Virtualization
Extract and Refine

• Easily combine data from multiple sources without moving or copying data
• Use any reporting or analytical tool
Red Hat JBoss Data Virtualization and Hortonworks HDP

Kimberly Palko
Current state of big data deployments

**Big Data Attitude**

Which of the following best describes your attitude toward Big Data Analytics?

- Big Data Analytics is the new source of competitive advantage and is/will be fundamental to our business: 46%
- Big Data Analytics is/will be an important complement to our existing Data Warehouse and Business Intelligence practice: 46%
- Big Data Analytics is "nice to have" set of technologies/capabilities but is not a top priority: 8%
- Big Data Analytics is a buzzword with unclear meaning or application within my enterprise: 1%

**Source: Wikibon 2014**

---

**State of Big Data Deployments**

Which of the following best describes the state of Big Data Analytics deployments in your organization?

- Production Deployment: 31%
- Pilot/PoC Underway: 28%
- Evaluation Phase: 41%

**Source: Wikibon 2014**
Integration Challenges

Vast majority of respondents believe Big Data Analytics is critical to the success of their respective enterprises.

SOURCE: WIKIBON Big Data Analytics Survey 2014

We strongly believe that success for many organizations hinges on your ability to close the gap between available data and actionable insight.
--Forrester

http://blogs.forrester.com/category/big_data

Data integration, data transformation and integrating with existing infrastructure are the biggest technology barriers to success.

SOURCE: WIKIBON Big Data Analytics Survey 2014
Data Control Challenges Getting Bigger with Big Data, Cloud, and Mobile

• Security capabilities are tightly coupled to data sources
• Extracting and moving data adds risk
• Every project solves data access and integration in a different way
• Inconsistent and decentralized control of data

How to align?

Constant Change

Siloed & Complex
DESIRED STATE

Data as a Service

● Standard based interface
● Single view of disparate source data
● Single point of access / integration
● Reuse of Data

But you cannot achieve this by writing more application code...

Data Sources
Siloed & Complex

Hadoop
NoSQL
Cloud Apps
Data Warehouse & Databases
Mainframe
XML, CSV & Excel Files
Enterprise Apps
Data Supply and Integration Solution

Data Virtualization sits in front of multiple data sources and:

✅ allows them to be treated a single source

✅ delivering the desired data

✅ in the required form

✅ at the right time

✅ to any application and/or user.

THINK VIRTUAL MACHINE FOR DATA
Easy Access to Big Data

- Reporting tool accesses the data virtualization server via rich SQL dialect
- The data virtualization server translates rich SQL dialect to HiveQL
- Hive translates HiveQL to MapReduce
- MapReduce runs MR job on big data
Different Users Different Views of Big Data

- Logical tables with different forms of aggregation
- Logical tables containing extra derived data
- Logical tables with filtered data
- All reports/users share the same specifications

Diagram:
- Hive
- MapReduce
- HDFS
Caching For Faster Performance – Virtual View

- Same cached view for multiple queries
- Refreshed automatically or manually
- Cache repository can be any supported data source

Virtual Database (VDB)

View 1

Query 1

Query 2

Cached View 1
Caching for Faster Performance – Result Set

• Results for a single query are cached after first execution
• Each unique query has its own cache
Integration of Big Data with existing data

• Integrating existing data with big data is easy
• Integration specifications can be shared or be developed for individual reports

Database Server

Hive

MapReduce

HDFS

Application
Use Case 1 – Combining sentiment data with existing enterprise data

Objective:
- Determine if sentiment data from the first week of the Iron Man 3 movie is a predictor of sales

Problem:
- Cannot utilize social data and sentiment analysis with sales management system

Solution:
- Leverage JBoss Data Virtualization to mashup Sentiment analysis data with ticket and merchandise sales data on MySQL into a single view of the data.

SOURCE 1: Hive/Hadoop contains twitter data including sentiment
SOURCE 2: MySQL data that includes ticket and merchandise sales

Excel Powerview and DV Dashboard to analyze the aggregated data
Use Case 1 - Resources

• GUIDE
  https://drive.google.com/folderview?id=0B5kKwcd4kOq9RUIHcVBMVjJuX2c&usp=sharing

• VIDEOS:
  http://vimeo.com/user16928011/hortonworksusecase1short
  http://vimeo.com/user16928011/hortonworksusecase2short

• SOURCE:
  https://github.com/DataVirtualizationByExample/HortonworksUseCase1
JBoss Data Virtualization Security and Hortonworks HDP
Role based access control

**Roles**
- Define roles based on organization hierarchy

**Users**
- External authentication via Kerberos, LDAP, etc.

**VDB**
- Assign users and groups to a virtual data base
Audit Logging via Dashboard
Row and Column Masking

Row based masking
   Ex: keyed off geographic marker

Column masking to a constant, null, or a SQL statement

Example: change all but the Last 4 digits in a credit card number to stars

concat('****', substring(column, length(column)-4))
Use Case 2 - Federation/Securing
Enterprise Data By Role

Objective:
Secure data according to Role for row level security and Column Masking

Problem:
Cannot hide region data such as customer data from region specific users

Solution:
Leverage JBoss Data Virtualization to provide Row Level Security and Masking of columns
Use Case 2 - Resources

- GUIDE
https://drive.google.com/folderview?id=0B5kKwcd4kOq9RUIHcVBMVjJuX2c&usp=sharing

- VIDEOS:
http://vimeo.com/user16928011/hortonworkseusecase2short
http://vimeo.com/user16928011/hortonworkseusecase2short

- SOURCE:
https://github.com/DataVirtualizationByExample/HortonworksUseCase2
Data for entire organization in Hadoop Data Lake

Problem: How does IT control access and give business users just the data they need?
- Does every line of business have access to everyone’s data?
- How do business users get access to the data they need in a simple (even self-service) way?
Secure, Self-Service Virtual Data Marts for Hadoop

Solution: Use JBoss Data Virtualization to create virtual data marts on top of a Hadoop cluster
- Lines of Business get access to the data they need in a simple manner
- IT maintains the process and control it needs
- All data remains in the data lake, nothing is copied or moved
Optional hierarchical data architectures with virtual data mart can be combined with security features like user role access and row and column masking.
Demonstration Virtual Data Marts with Hadoop Data Lake

Cojan van Ballegooijen
Use Case 3 – Virtual data marts with Hadoop Data Lake

Objective:
– Purpose oriented data views for functional teams over a rich variety of semi-structured and structured data

Problem:
– Data Lakes have large volumes of consolidated clickstream data, product and customer data that need to be constrained for multi-departmental use.

Solution:
– Leverage HDP to mashup Clickstream analysis data with product and customer data on HDP to answer
  - Leverage Jboss Data Virt to provide Virtual data marts for each of Marketing and Product teams

Use Case 3: Virtual data marts with Hadoop Data Lake.
Use Case 3 - Resources

• GUIDE
How to guide: https://github.com/DataVirtualizationByExample/HortonworksUseCase3
Tutorial: Available soon

• VIDEOS:
http://vimeo.com/user16928011/hwxuc3configuration
http://vimeo.com/user16928011/hwxuc3run
http://vimeo.com/user16928011/hwxuc3overview

• SOURCE:
https://github.com/DataVirtualizationByExample/HortonworksUseCase3
Benefits of Data Virtualization on Big Data

• Enterprise democratization of big data
• Any reporting or analytical tool can be used
• Easy access to big data
• Seamless integration of big data and existing enterprise data
• Sharing of integration specifications
• Collaborative development on big data
• Fine-grained security of big data
• Speedy delivery of reports on big data

You Need A Data Virtualization Strategy To Avoid Falling Behind

“Without a data virtualization strategy, you risk knowing less about your customer, delivering fewer real-time business insights, losing competitive advantage, and spending more to address data challenges.”

Information Fabric 3.0
August 8, 2013
SUMMIT BY DAY
PARTY BY NIGHT

JOIN OUR JBOSS, OPENSHIFT, AND MOBILE TEAMS ON WED. JUNE 24 FOR A NIGHT OF GAMES, DANCING, AND OPEN CONTAINERS

Visit the Red Hat booth in Hall D for location and invitation.

An invitation doesn't guarantee entrance. Admission determined by city of Boston fire code.
<table>
<thead>
<tr>
<th>Session Title</th>
<th>Room</th>
<th>Date</th>
<th>Time</th>
<th>Technical difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connect to multiple data sources without writing code</td>
<td>201</td>
<td>Wednesday, June 24</td>
<td>7:00 pm - 8:00 pm</td>
<td>3</td>
</tr>
<tr>
<td>Integration with Red Hat JBoss Fuse and Red Hat JBoss Data Virtualization</td>
<td>207</td>
<td>Wednesday, June 24</td>
<td>4:50 pm - 5:50 pm</td>
<td>4</td>
</tr>
<tr>
<td>iPaaS &amp; beyond: Red Hat’s integration roadmap</td>
<td>312</td>
<td>Friday, June 26</td>
<td>9:45 am - 10:45 am</td>
<td></td>
</tr>
<tr>
<td>Red Hat Storage Server as a hybrid storage solution for Splunk Enterprise</td>
<td>209</td>
<td>Wednesday, June 24</td>
<td>4:30 pm - 5:30 pm</td>
<td></td>
</tr>
<tr>
<td>Make a data-driven investment decision about big data</td>
<td>309</td>
<td>Friday, June 26</td>
<td>9:45 am - 10:45 am</td>
<td>2</td>
</tr>
<tr>
<td>Using Apache Spark to build analytical applications in the Cloud</td>
<td>208</td>
<td>Wednesday, June 24</td>
<td>2:30 pm - 3:30 pm</td>
<td></td>
</tr>
<tr>
<td>Big data on the open private cloud</td>
<td>310</td>
<td>Thursday, June 25</td>
<td>2:30 pm - 3:30 pm</td>
<td>3</td>
</tr>
<tr>
<td>Building a big data, risk-management solution for financial services</td>
<td>207</td>
<td>Thursday, June 25</td>
<td>3:40 pm - 4:40 pm</td>
<td>2</td>
</tr>
<tr>
<td>Drinking from the firehose with duct-tape-free reactive Java applications</td>
<td>206</td>
<td>Tuesday, June 23</td>
<td>10:30 am - 11:30 am</td>
<td>3</td>
</tr>
<tr>
<td>OpenStack nirvana: Big data &amp; elastic infrastructure together at last</td>
<td>302</td>
<td>Wednesday, June 24</td>
<td>2:30 pm - 3:30 pm</td>
<td>2</td>
</tr>
</tbody>
</table>

#redhat #rhsummit
LEARN. NETWORK. EXPERIENCE OPEN SOURCE.