BOSTON, MA
JUNE 23-26, 2015

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

Big Data Explosion

IDC Analyze the Future

85% from new data types

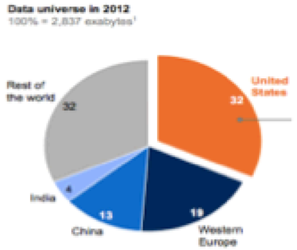
40 ZB digital universe by 2020

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*

McKinsey Global Institute

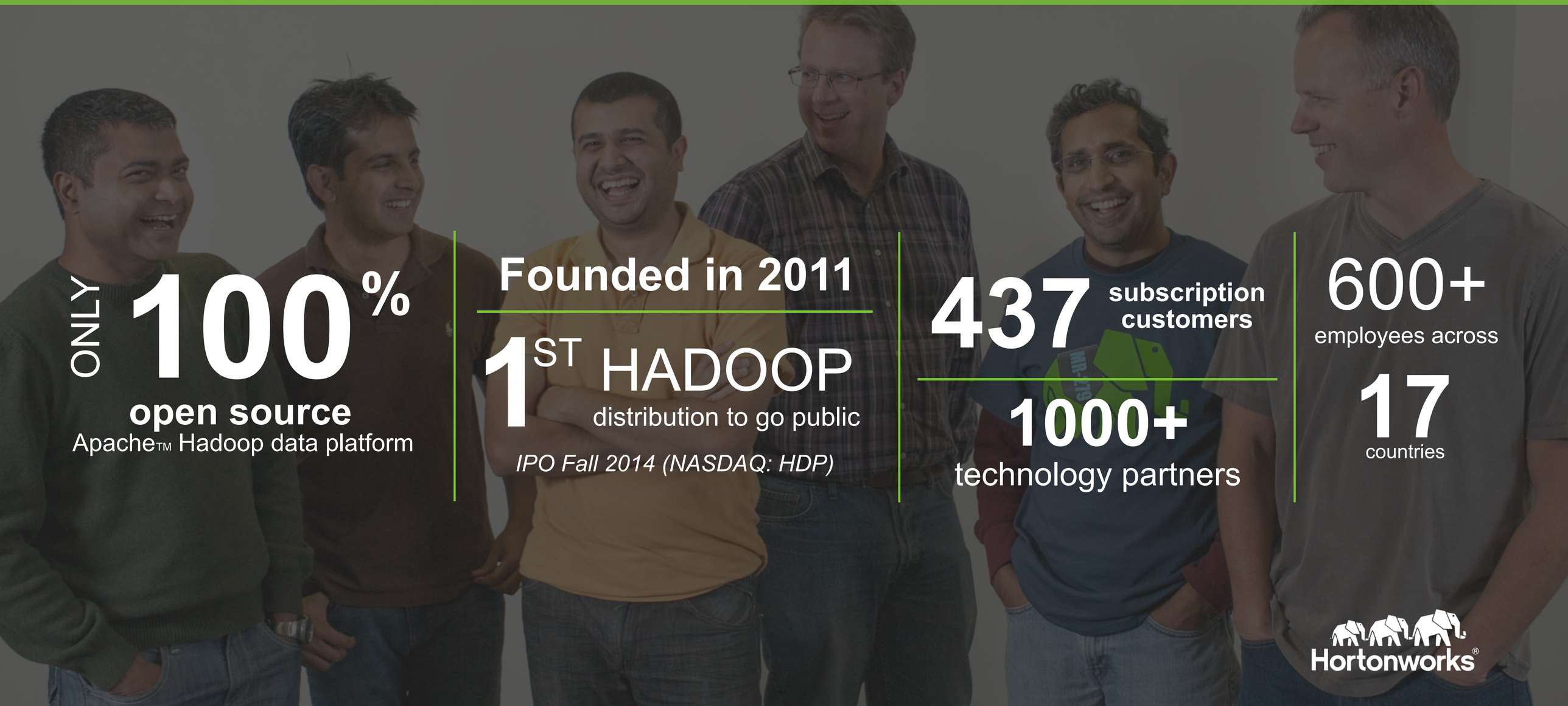
65% analytic apps with Hadoop inside by 2015

⇧ Hadoop enabled DBMS's

Gartner

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

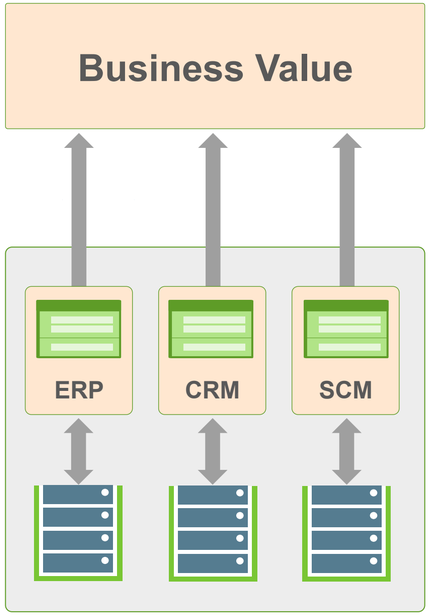
Hortonworks®

Traditional Data Systems are Under Pressure…

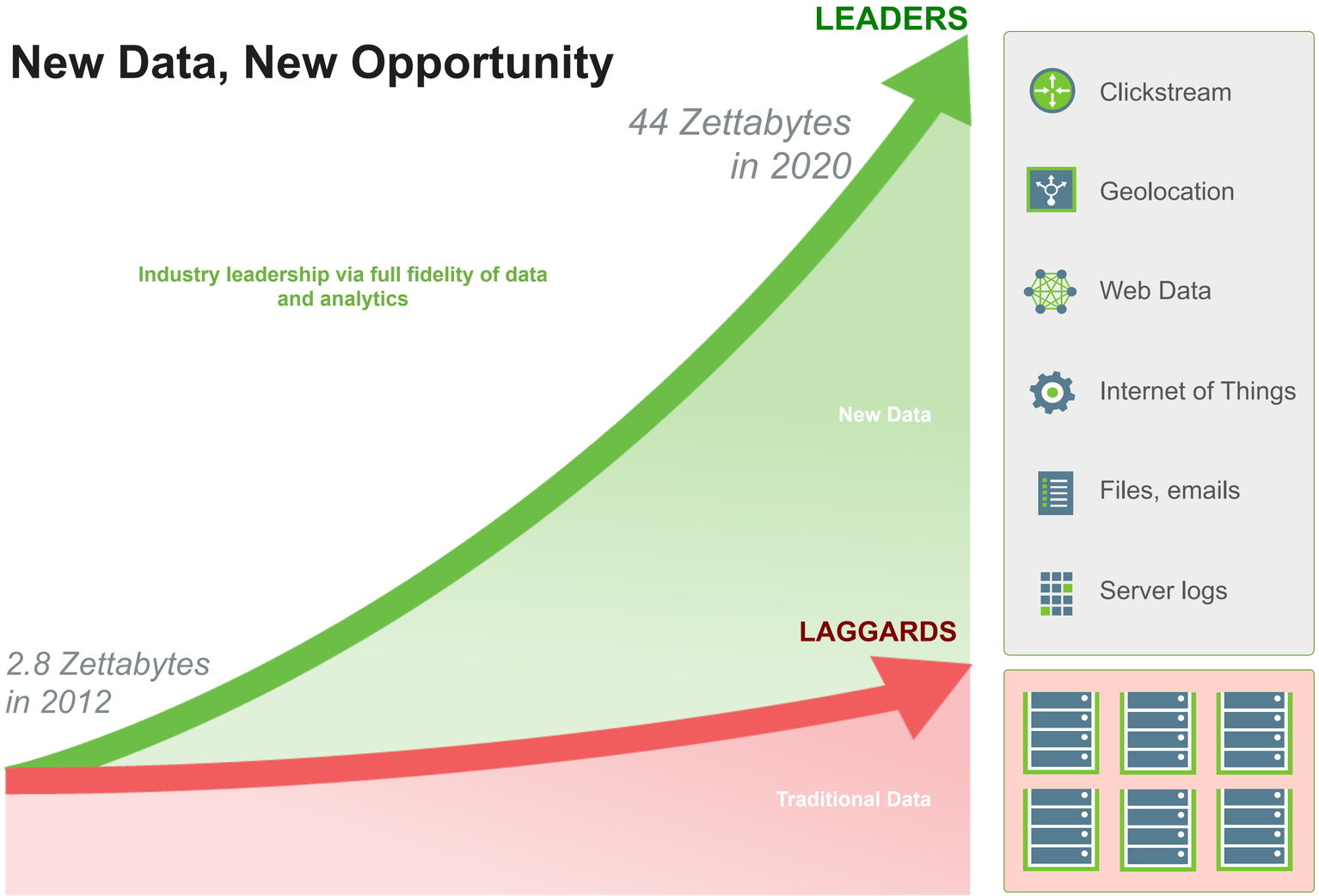
Traditional Systems

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

Limited ability to innovate



New Data, New Opportunity



1 Zettabyte (ZB) = 1 million Petabytes (PB); Sources: IDC, IDG Enterprise, and AMR Research

Hortonworks®

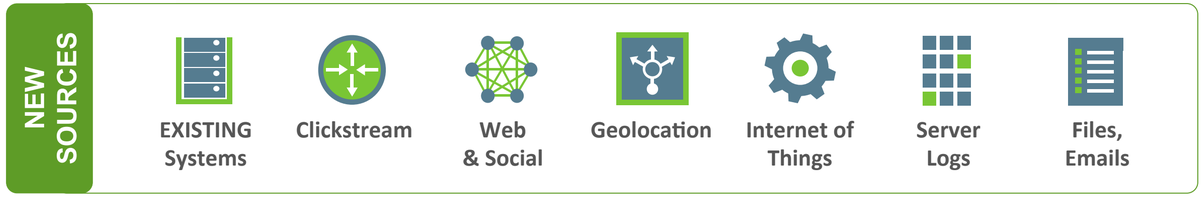
A New Approach Is Needed

The goal:

Turn data into value

The problem:

Data architectures don't scale



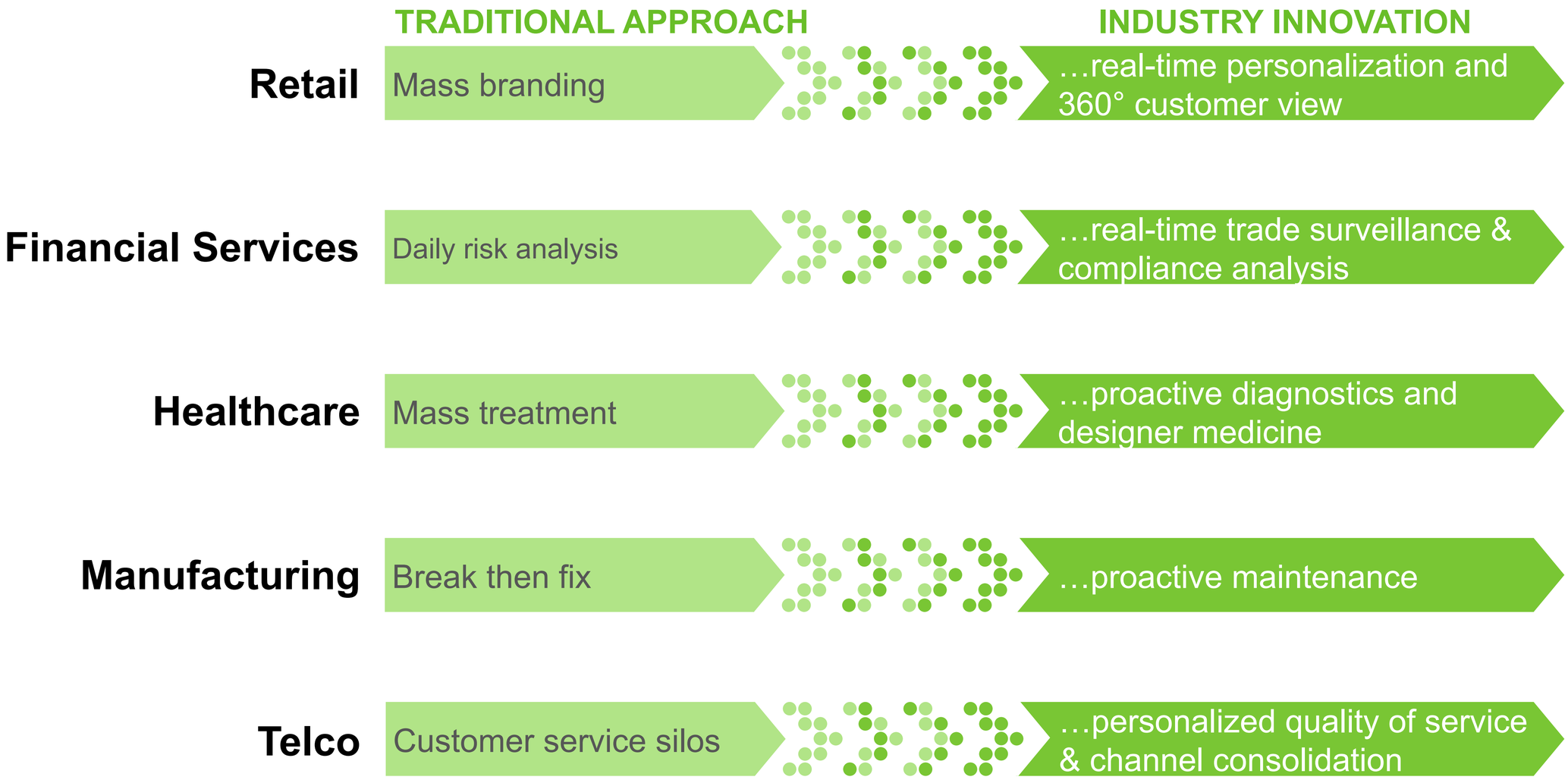
Costs

Silos

Data Structure

Hortonworks®

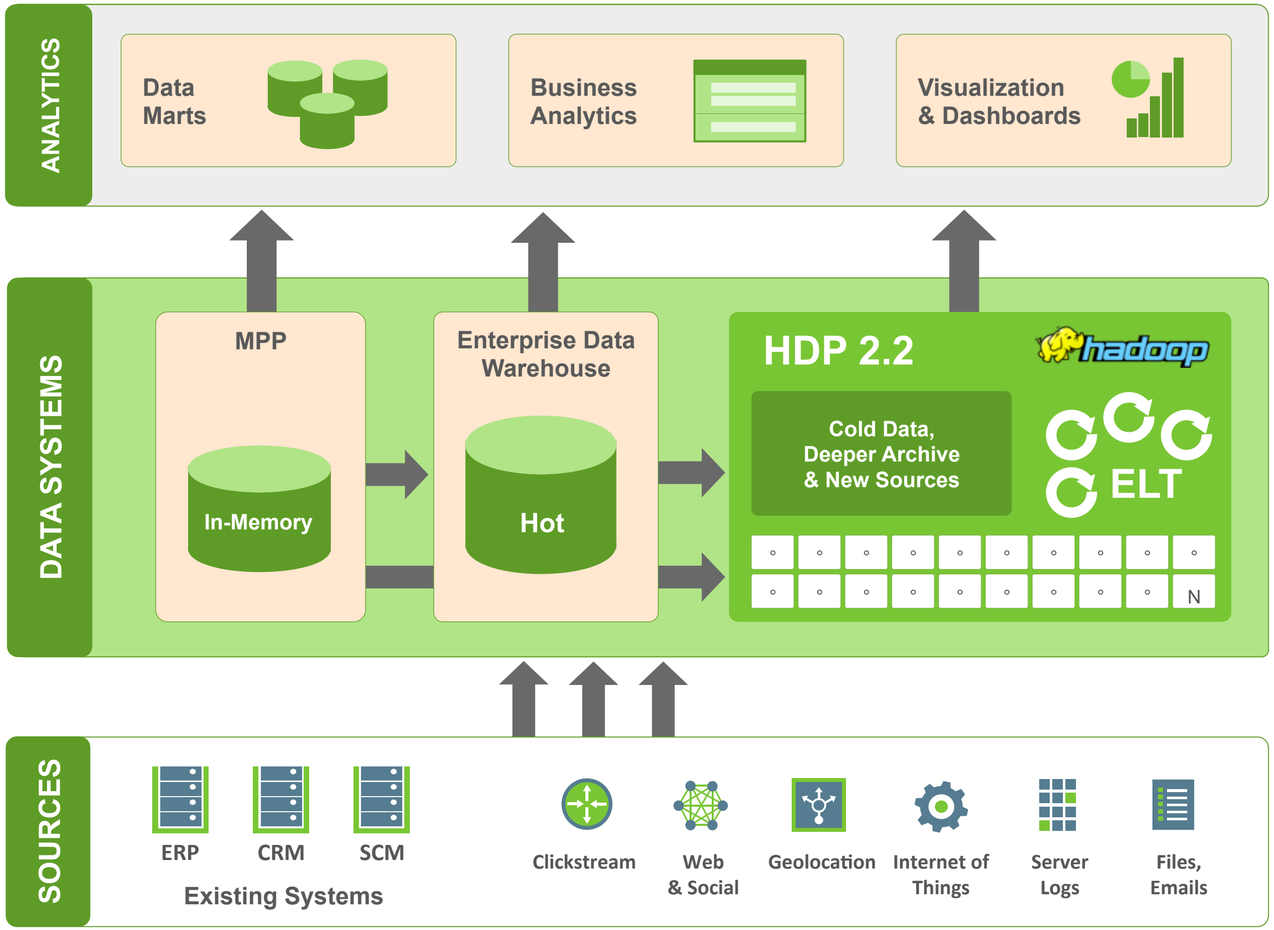
Big Data: From Reactive to Proactive Value Chains



Hortonworks®

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

Hortonworks®

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

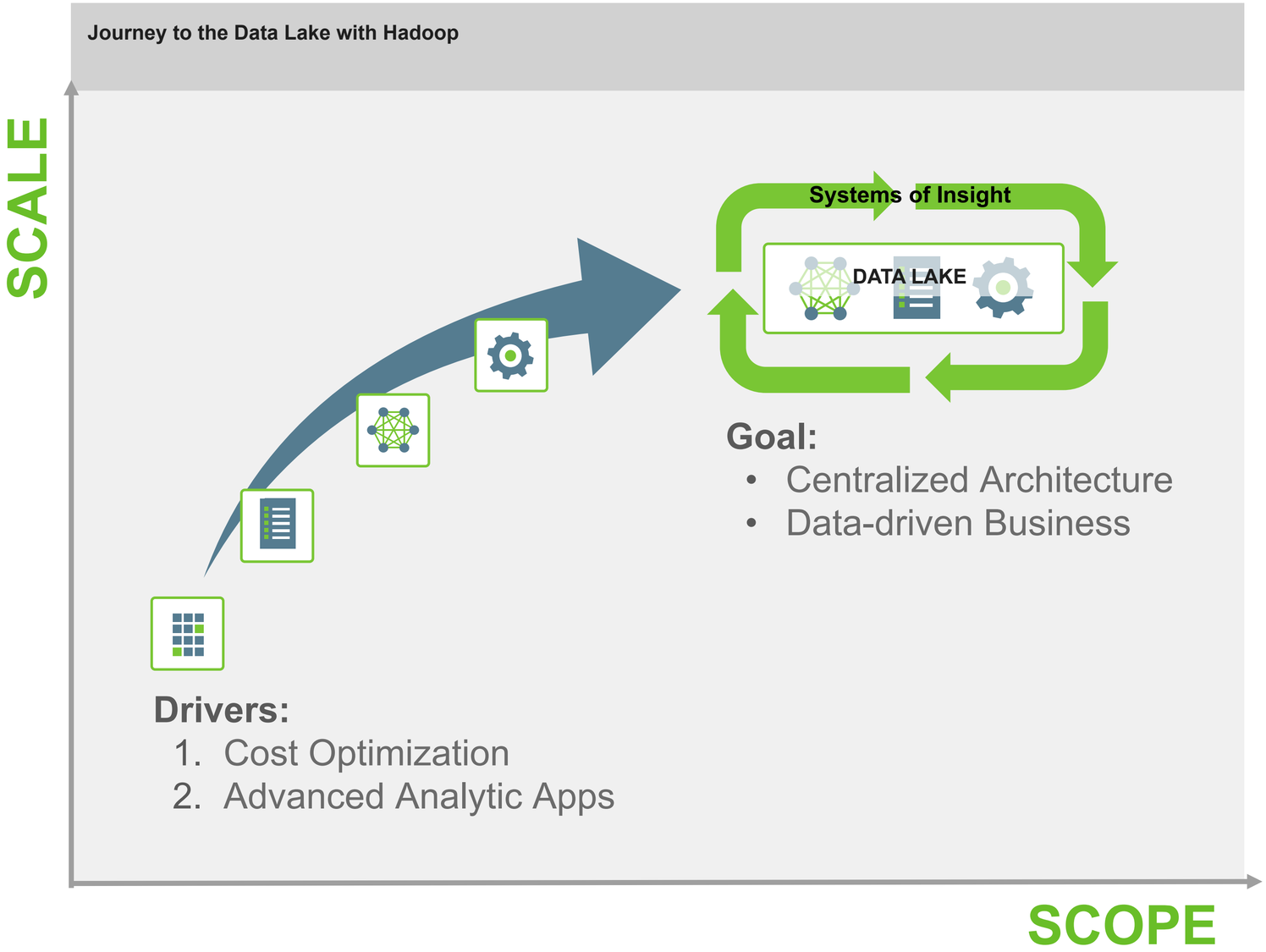
- HDP captures, stores and processes large volumes of data streaming from connected devices
- Stream processing and data science help introduce new analytics for real-time and batch analysis

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.

Hortonworks®

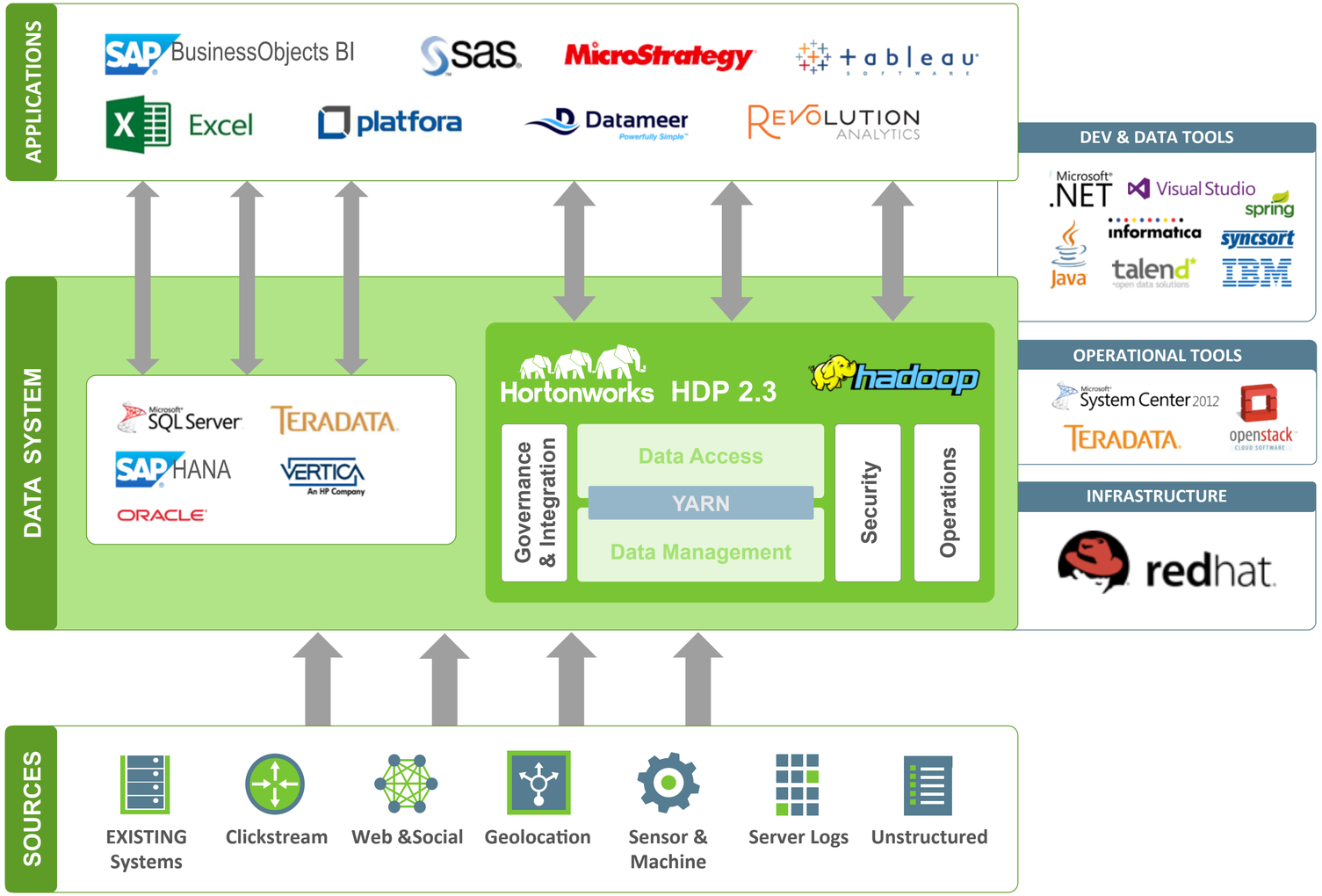
Hadoop Drivers and the Journey to a Data Lake



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.

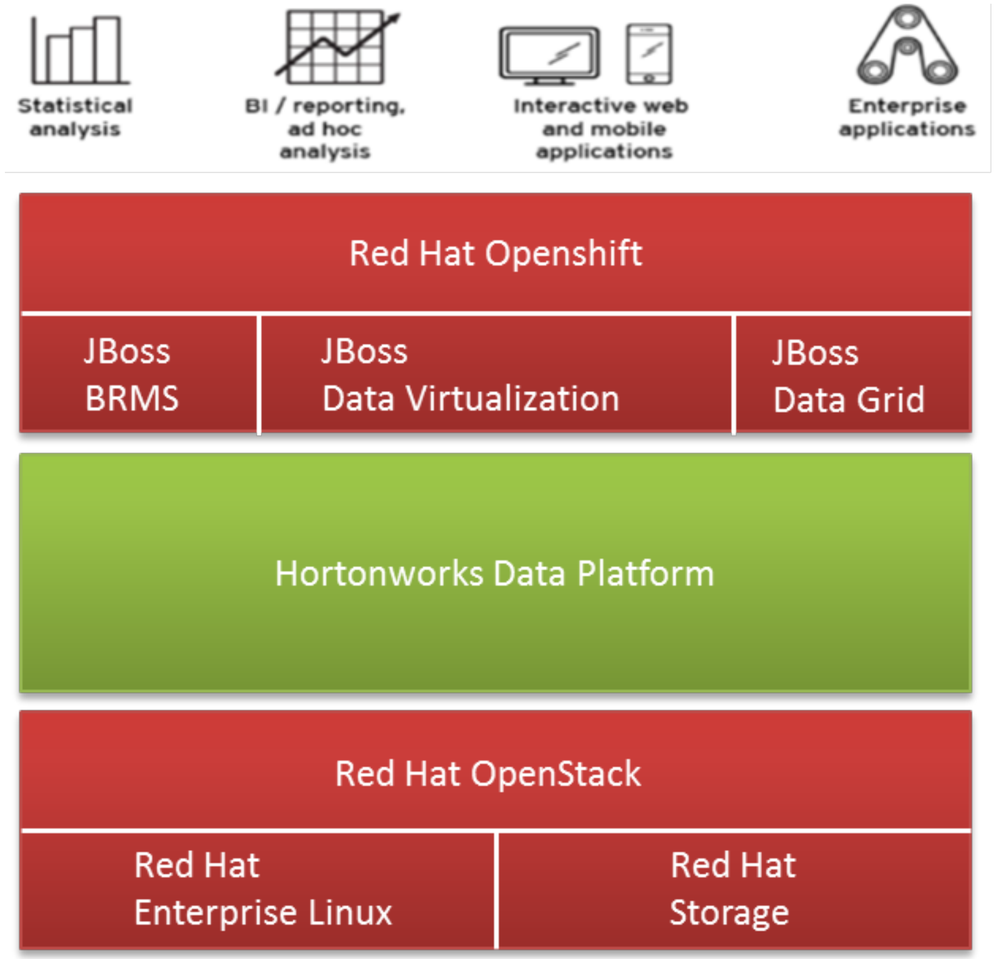
HDP is deeply integrated in the data center



- 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

Hortonworks®

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

Hortonworks®

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

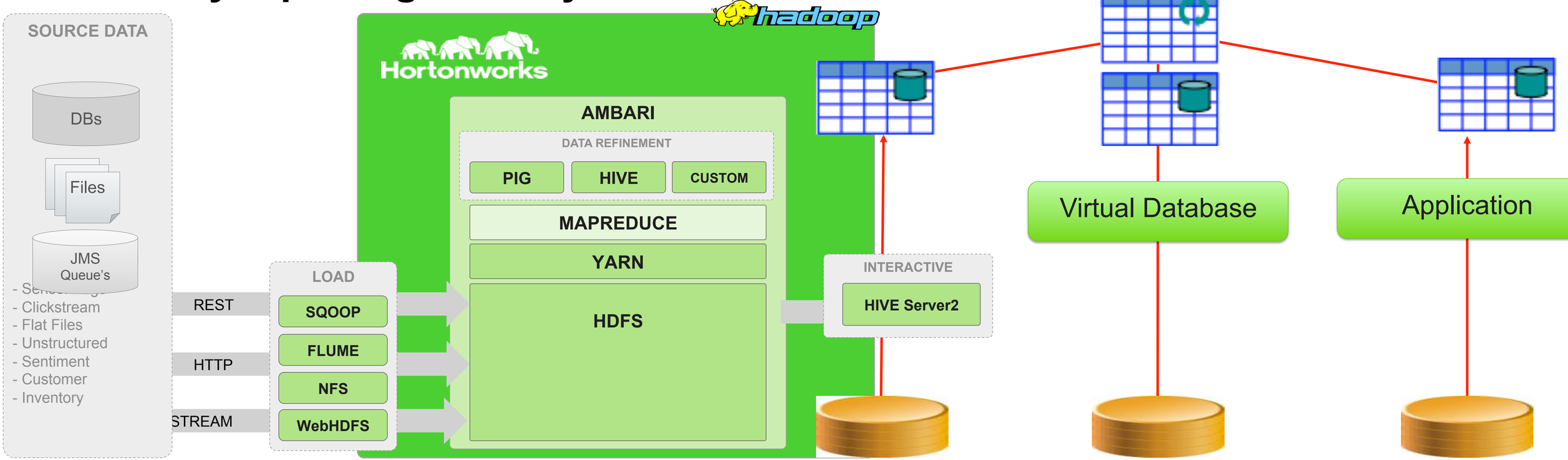
- HDP 2.1 on Red Hat Storage 3.0.2
- Hadoop Plug-in Refresh Release and Ambari
- Red Hat JBoss Data Virtualization with HDP
- HDP 2.2 on Red Hat Enterprise Linux with OpenJDK

Hortonworks

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**

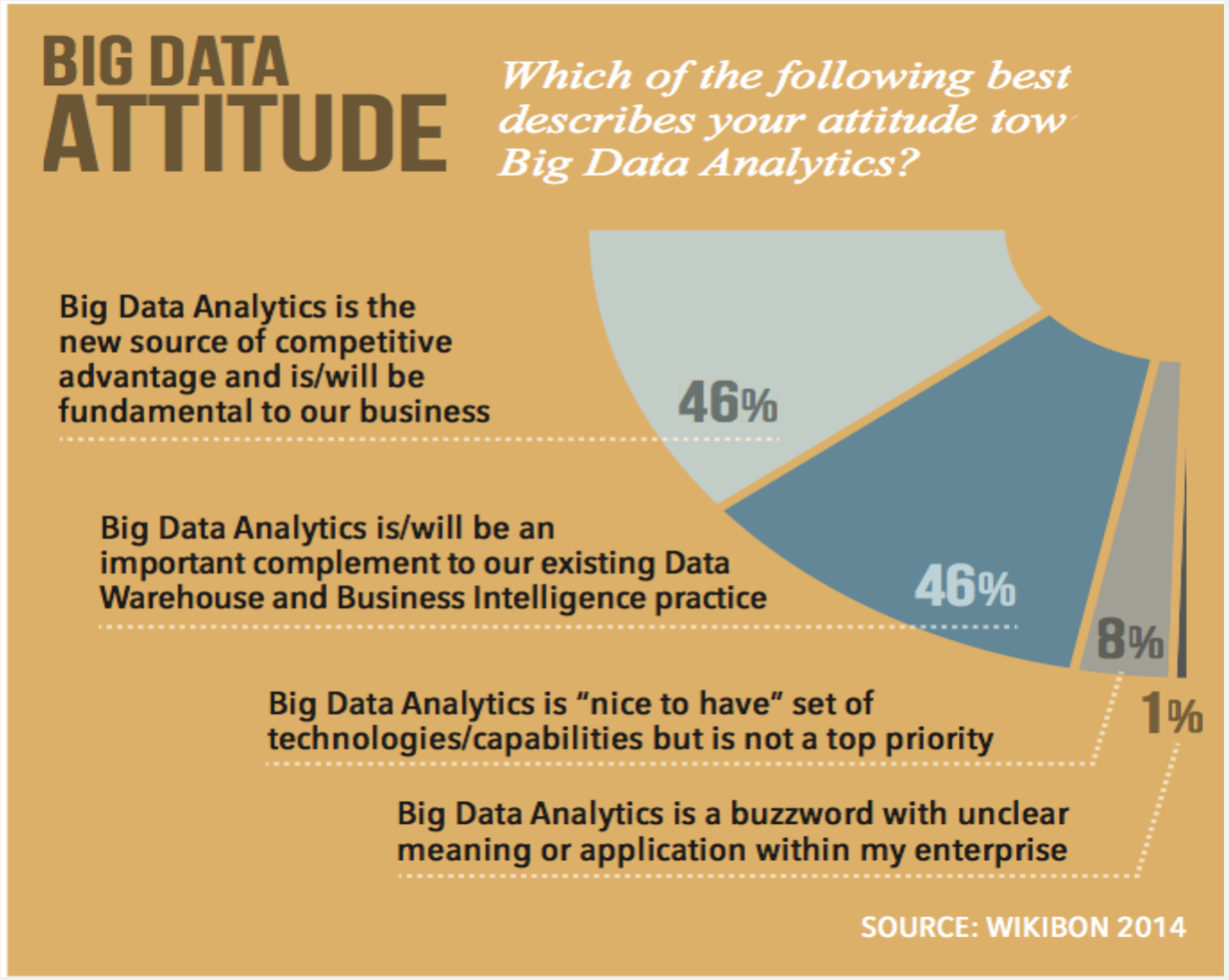


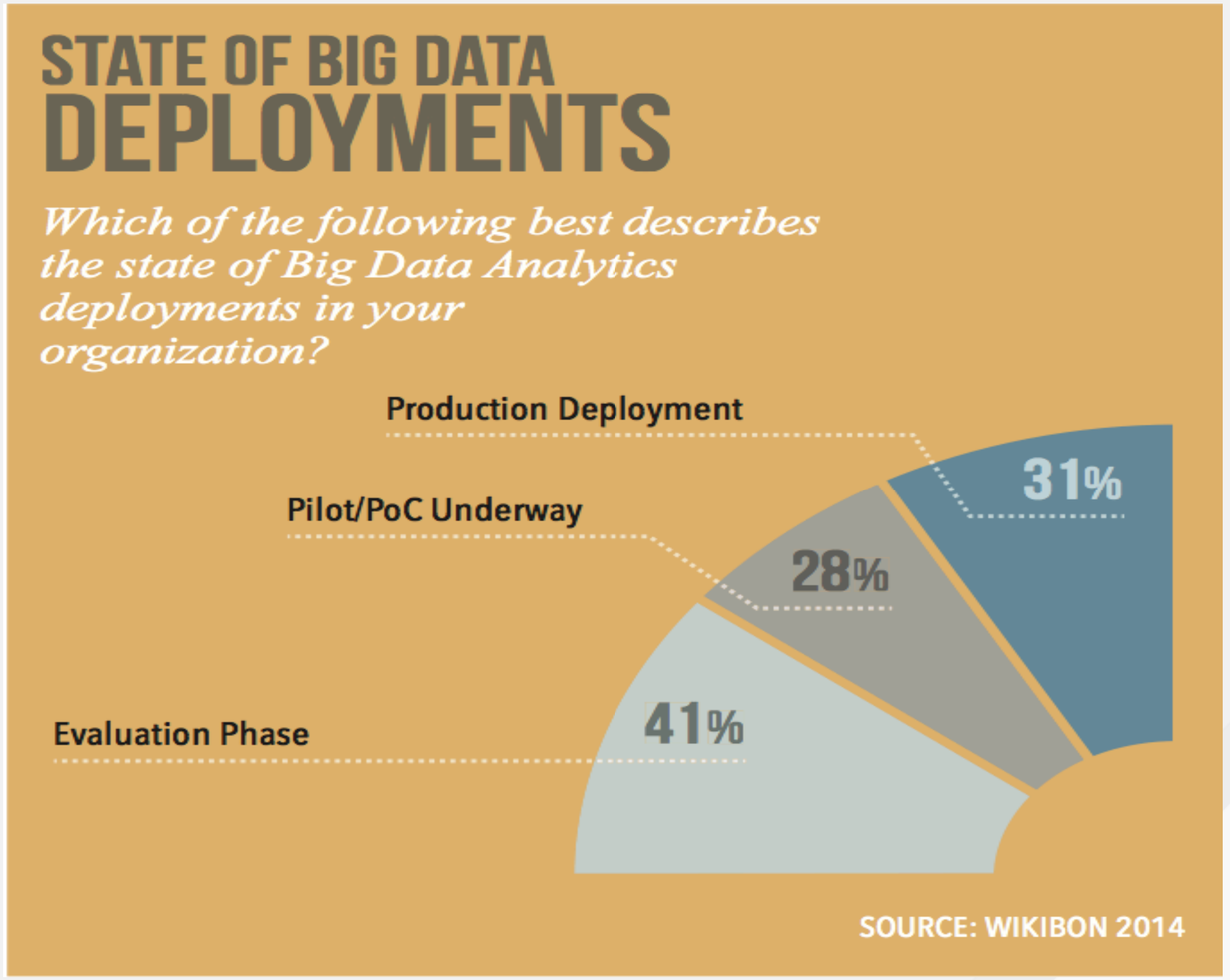
Hortonworks

Red Hat JBoss Data Virtualization and Hortonworks HDP

Kimberly Palko

Current state of big data deployments





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

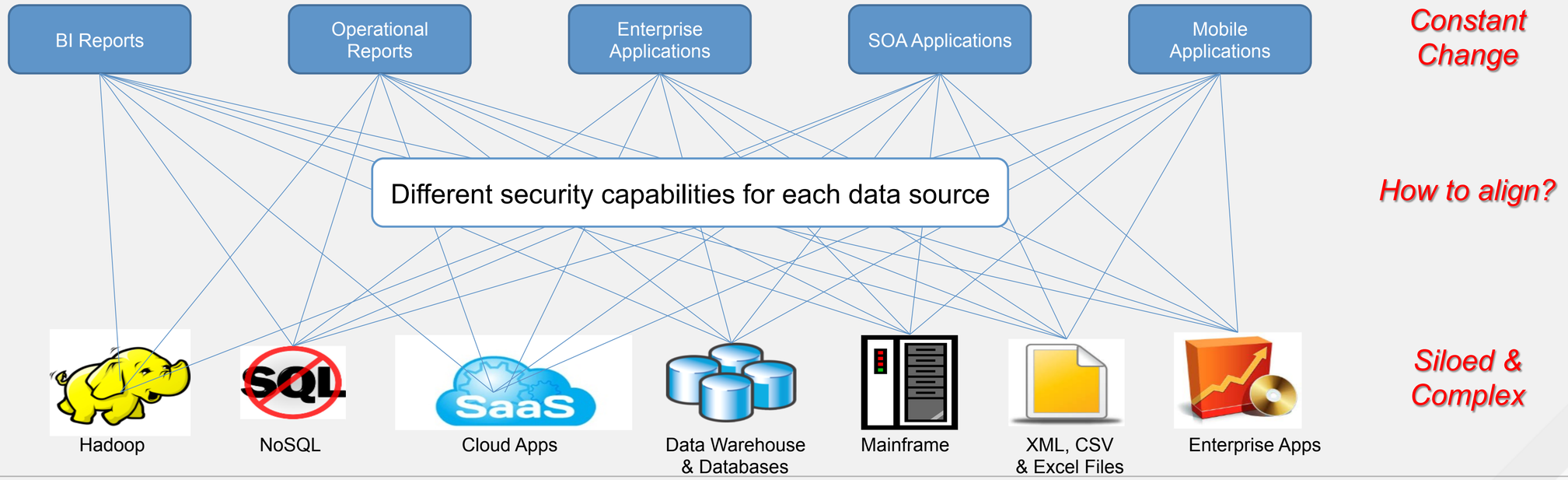
http://blogs.forrester.com/category/big_data

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



Constant Change

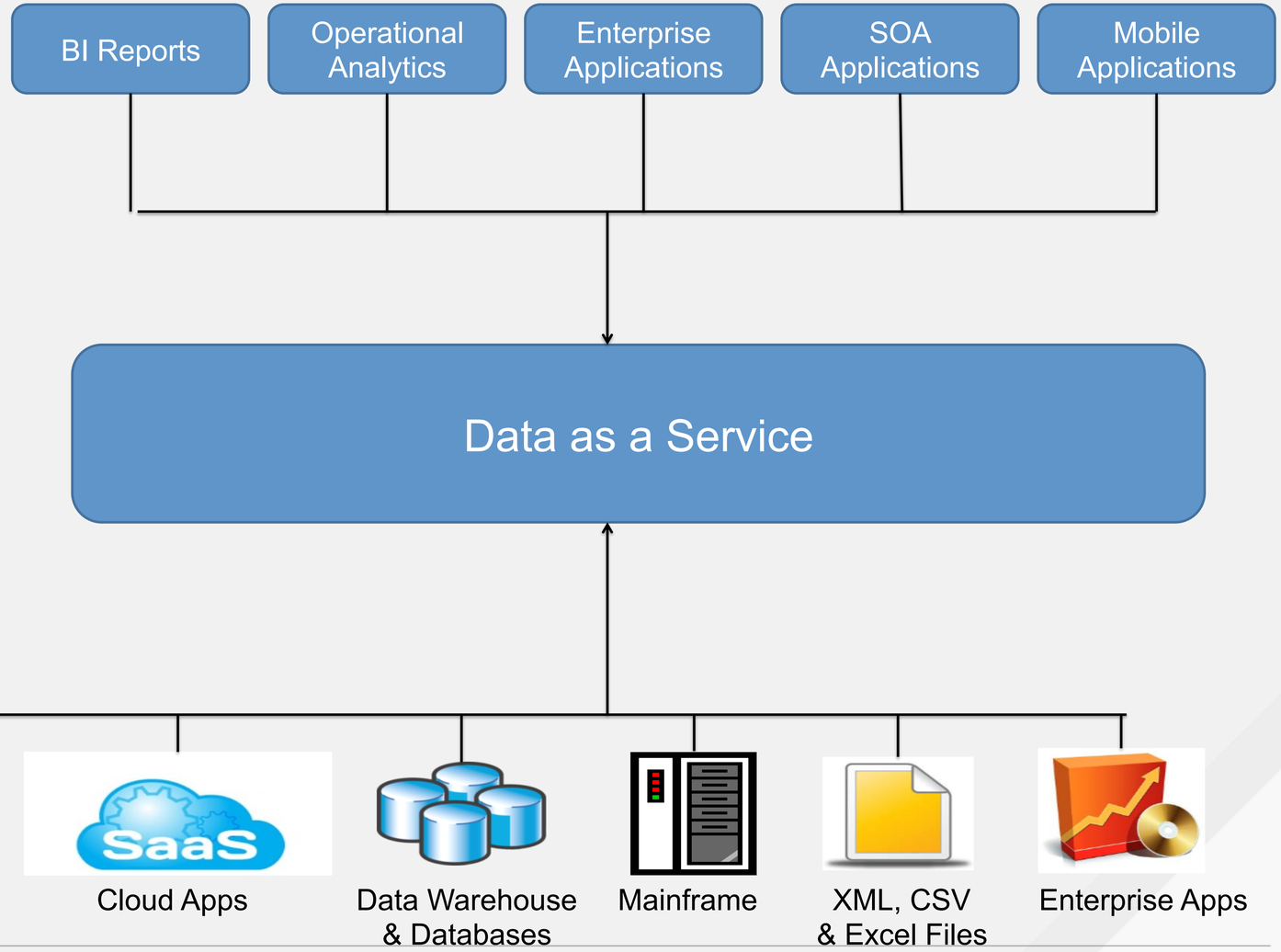
How to align?

Siloed & Complex

DESIRED STATE

Data as a Service

- Standard based interface
- Single view of disperate source data
- Single point of access / integration
- Reuse of Data



But you cannot achieve this by writing more application code…

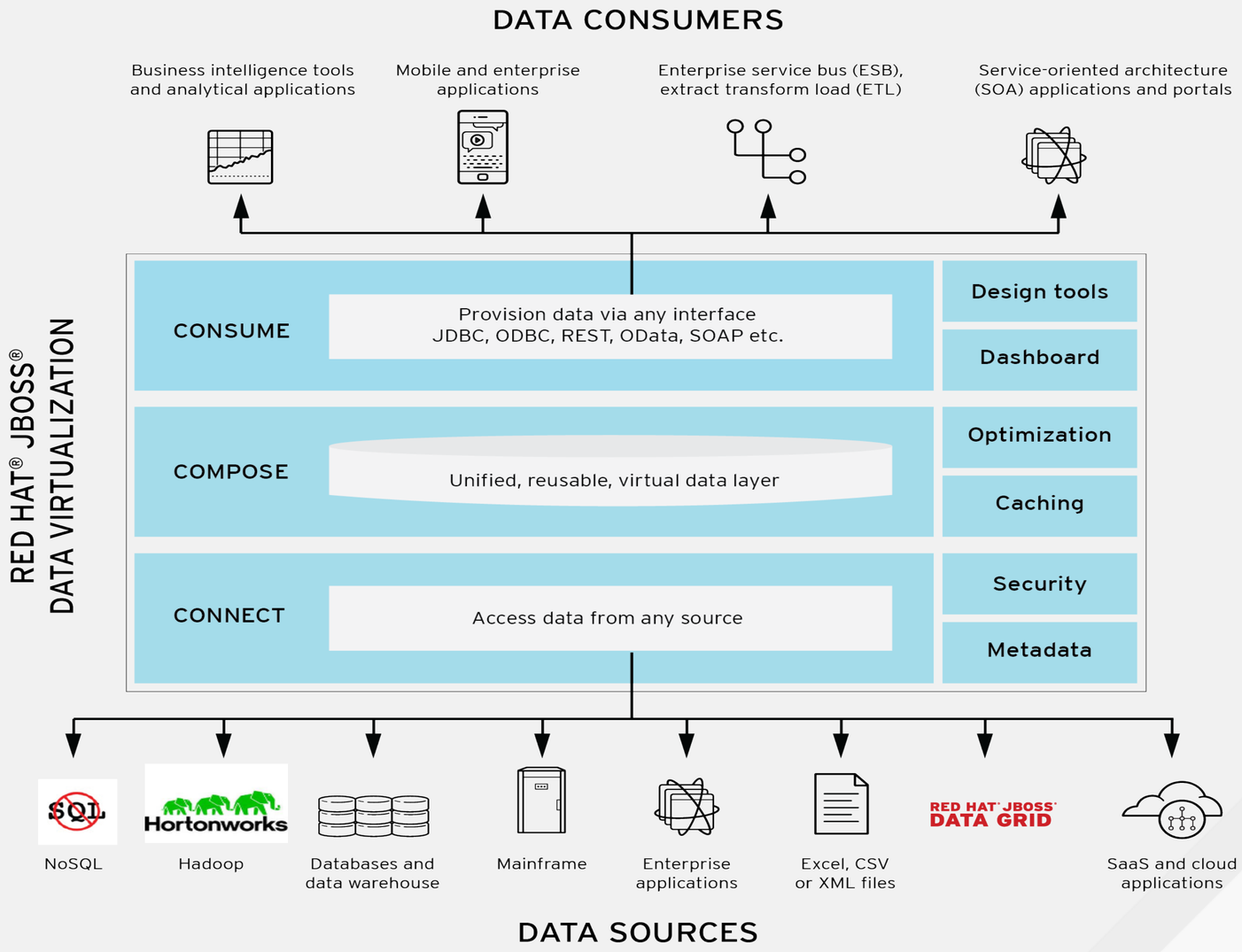
Data Sources
Siloed & Complex

Data Supply and Integration Solution

Data Virtualization sits in front of multiple data sources and

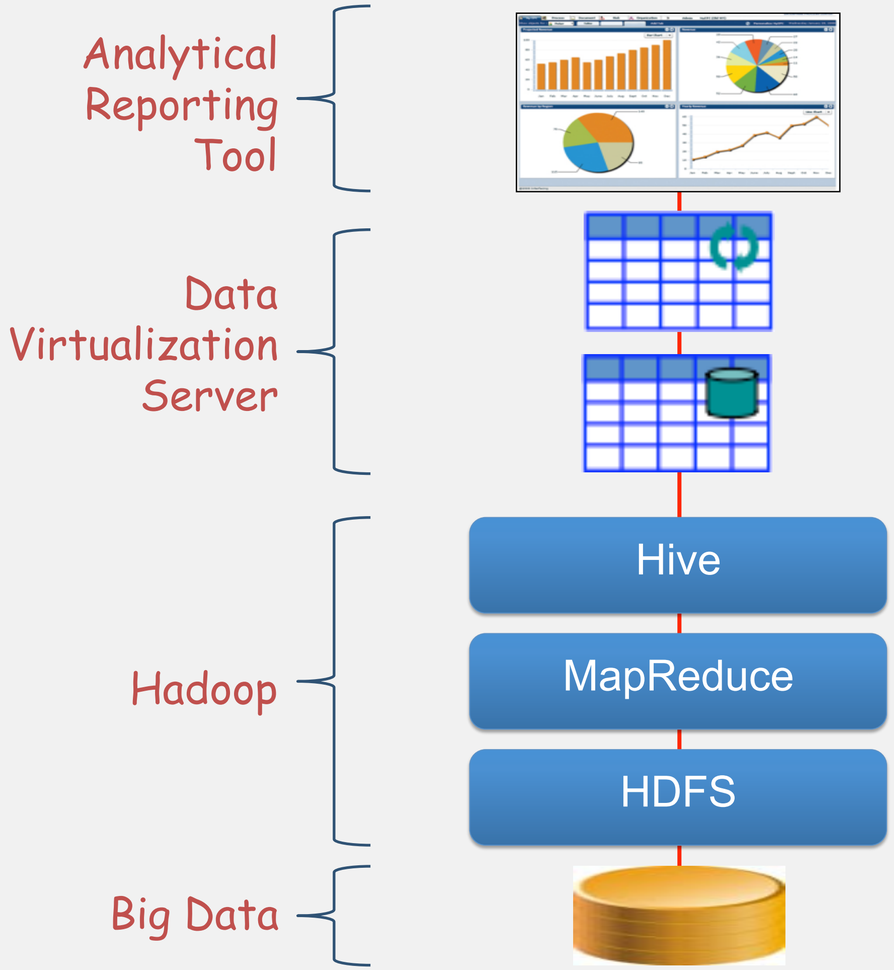
- ✓ allows them to be treated a single source
- ✓ delivering the desired data
- ✓ in the required form
- ✓ at the right time
- ✓ to any application and/or user.

THINK VIRTUAL MACHINE FOR DATA



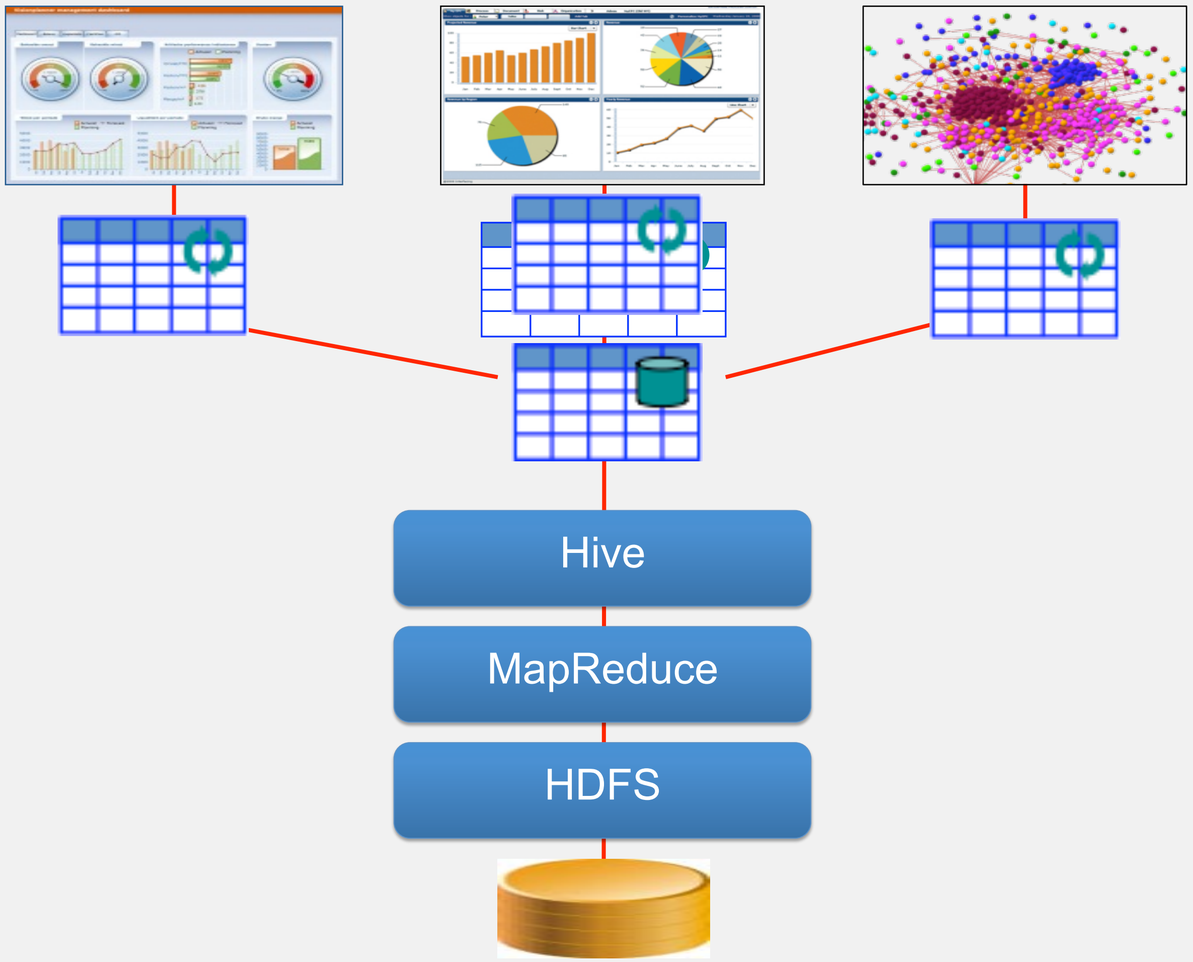
JB0041-2

Easy Access to Big Data



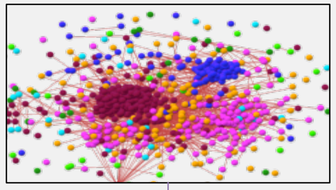
- Reporting tool accesses the data virtualization server via *rich* SQL dialect
- The data virtualization server translates rich SQL dialect to HiveQL
- Hive translates HiveQL to MapReduce
- MapReduce runs MR job on big data

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

Caching For Faster Performance – Virtual View



Query 1

Query 2

Virtual Database (VDB)

View 1

Cached
View 1

Hortonworks

Hadoop

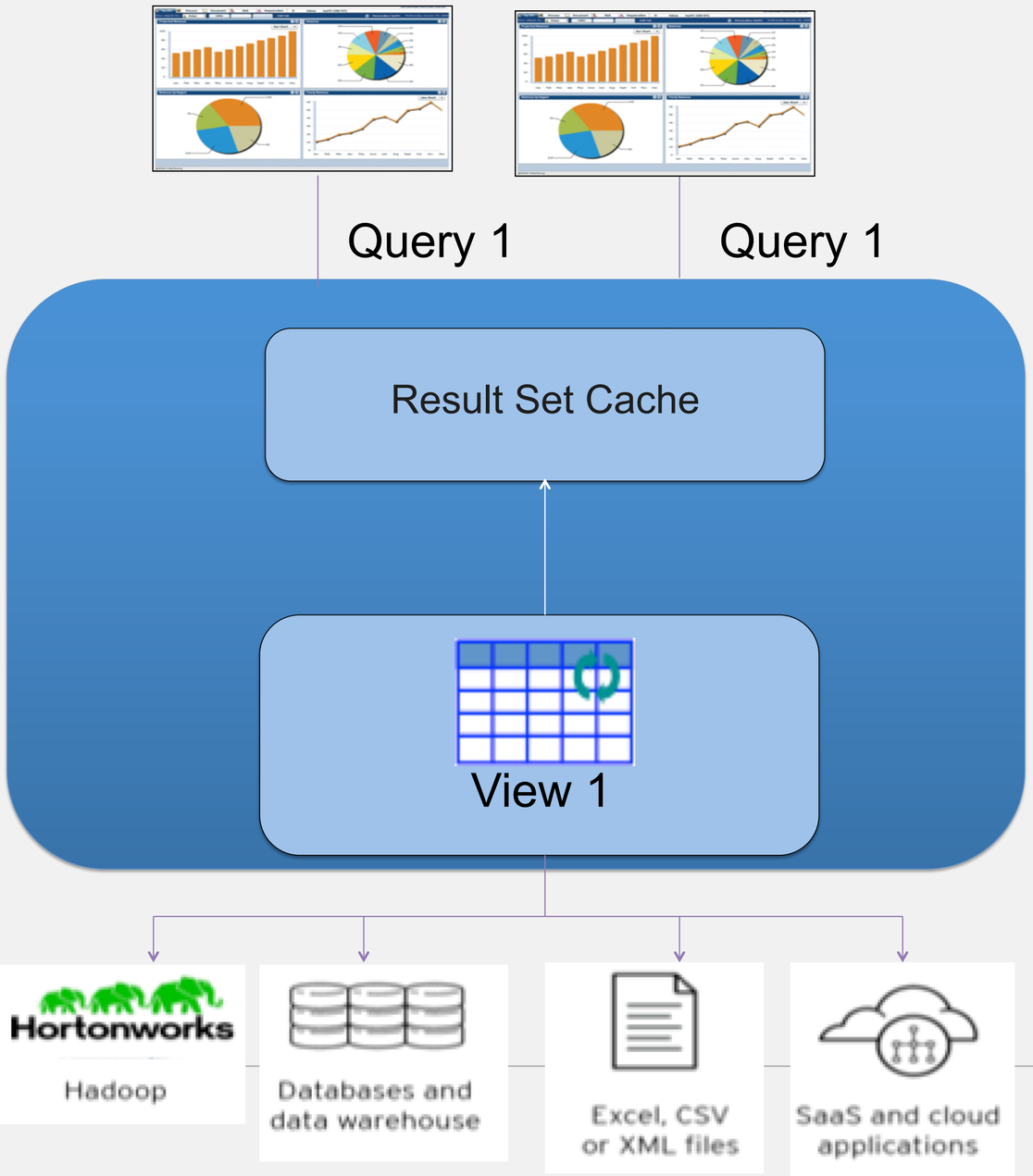
Databases and data warehouse

Excel, CSV or XML files

SaaS and cloud applications

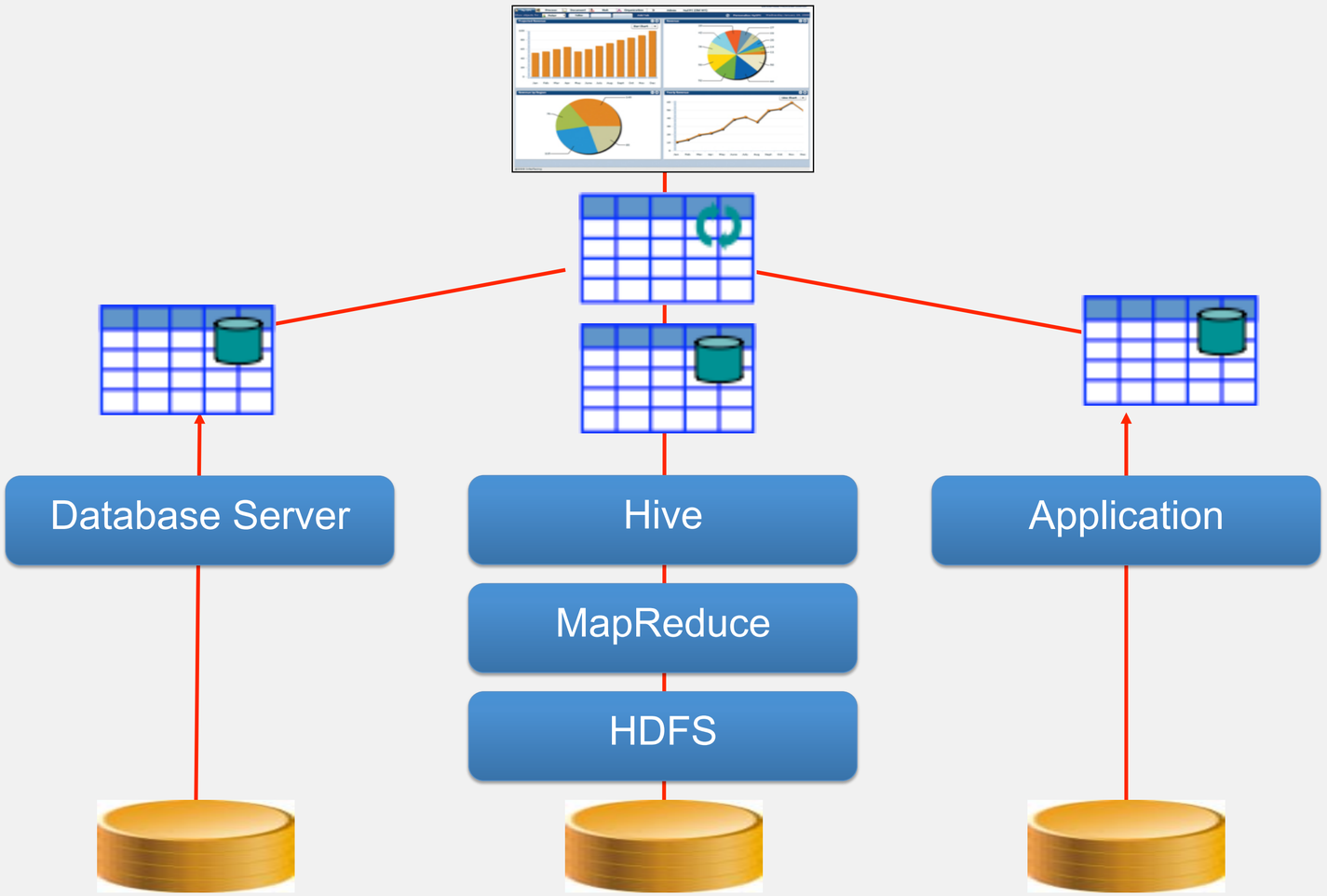
- Same cached view for multiple queries
- Refreshed automatically or manually
- Cache repository can be any supported data source

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

Use Case 1 – Combining sentiment data with existing enterprise data

Objective:

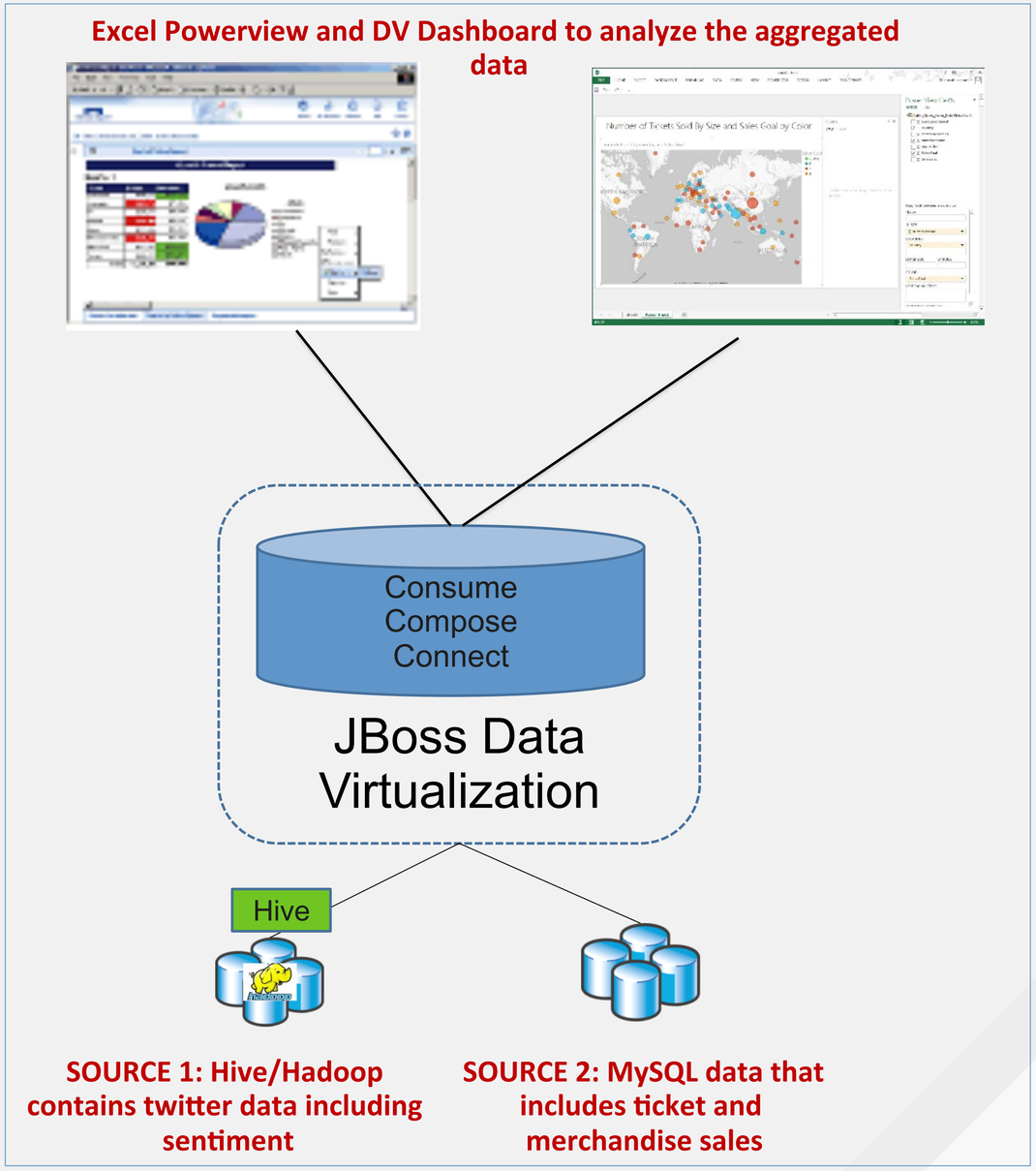
-Determine if sentiment data from the first week of the Iron Man 3 movie is a predictor of sales

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.



Use Case 1 - Resources

- GUIDE

https://drive.google.com/folderview?id=0B5kKwcd4kOq9RUlHcVBMVjJuX2c&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

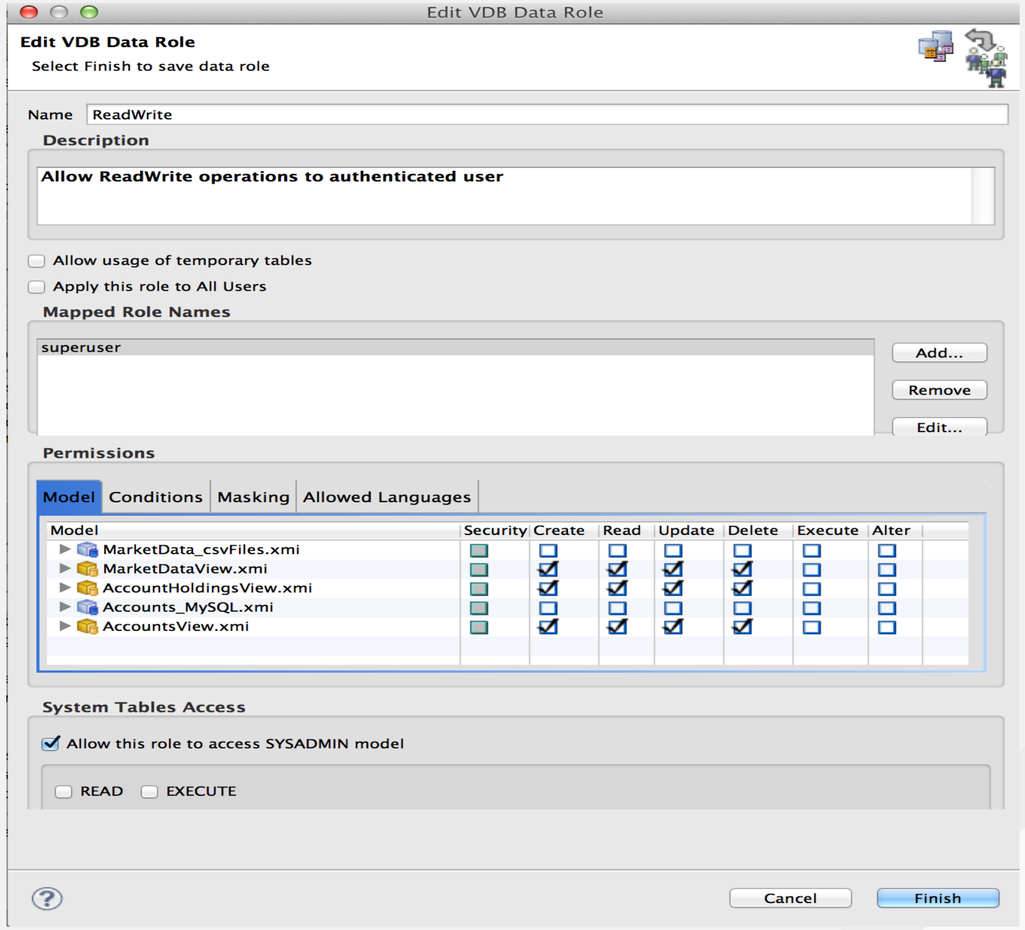
- Define roles based on organization hierarchy

Users

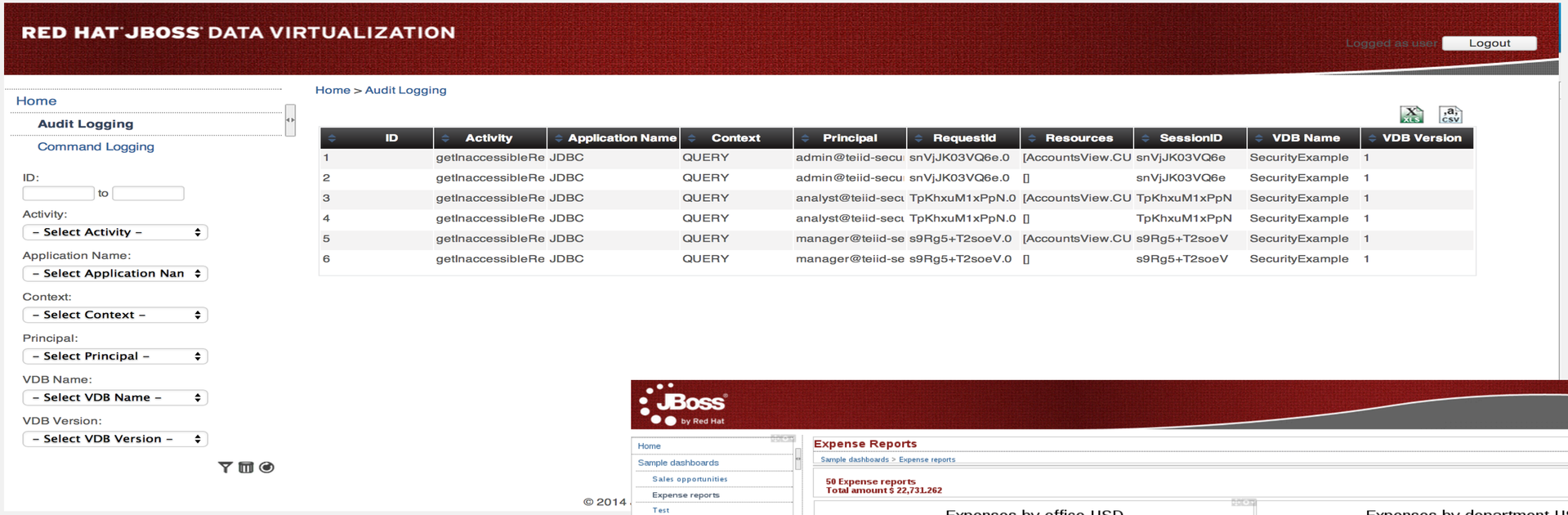
- External authentication via Kerberos, LDAP, etc.

VDB

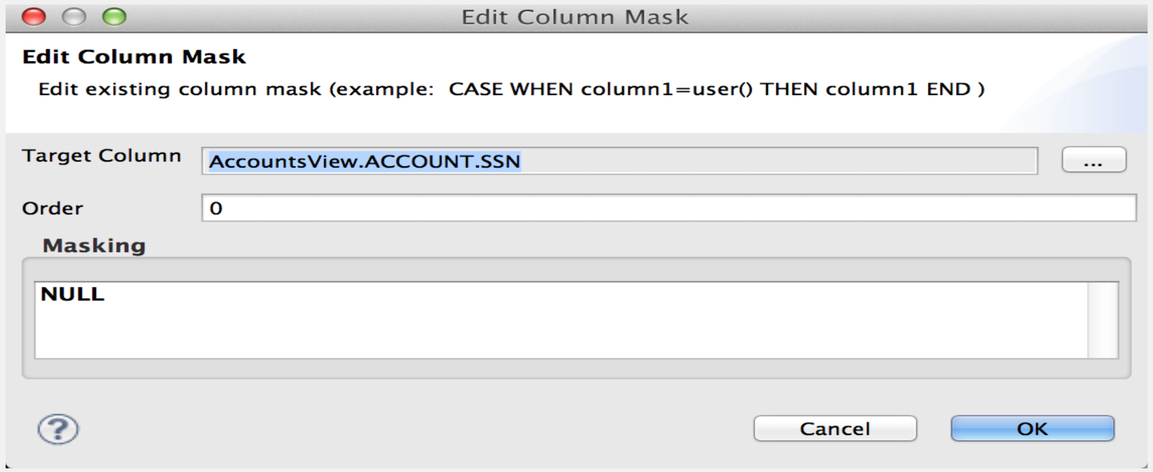
- Assign users and groups to a virtual data base



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:

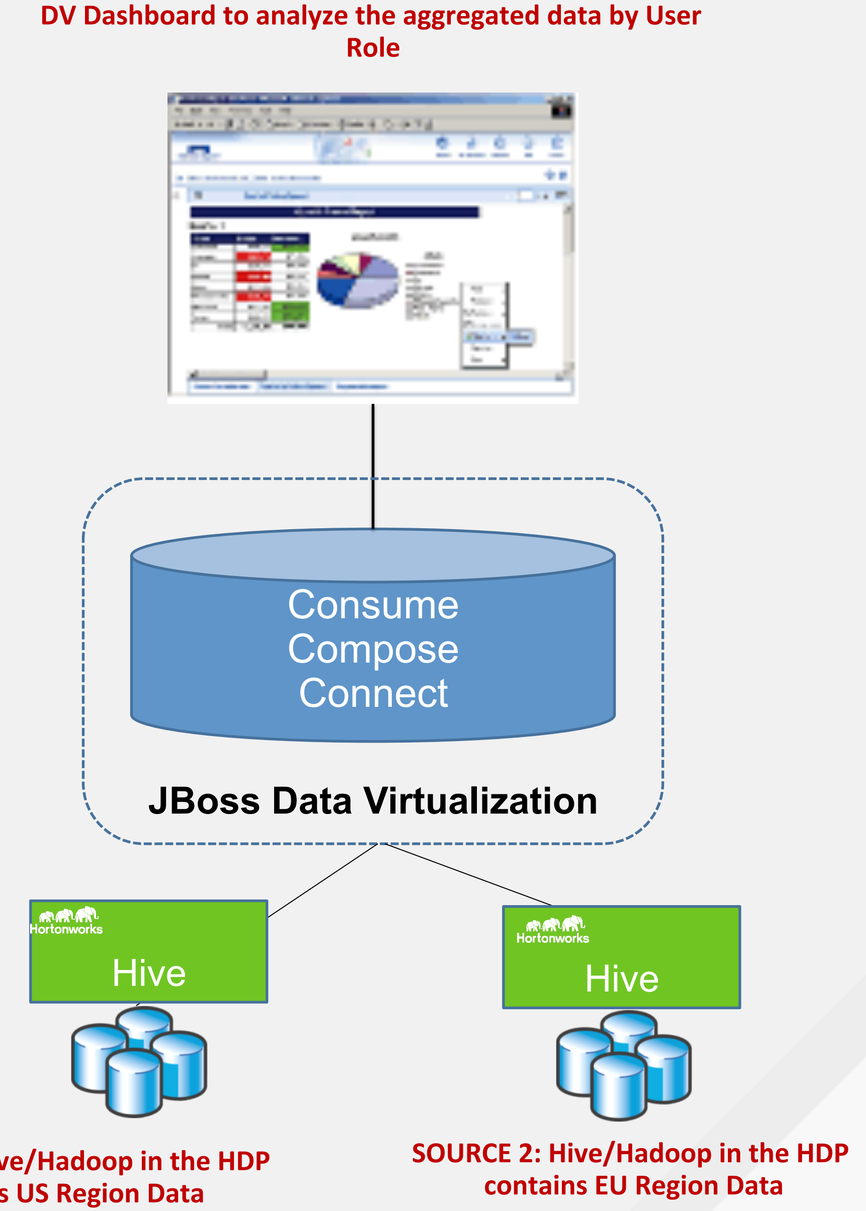
Secure data according to Role for row level security and Column Masking

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=0B5kKwcd4kOq9RUlHcVBMVjJuX2c&usp=sharing

- VIDEOS:

http://vimeo.com/user16928011/hortonworksusecase2short

http://vimeo.com/user16928011/hortonworksusecase2short

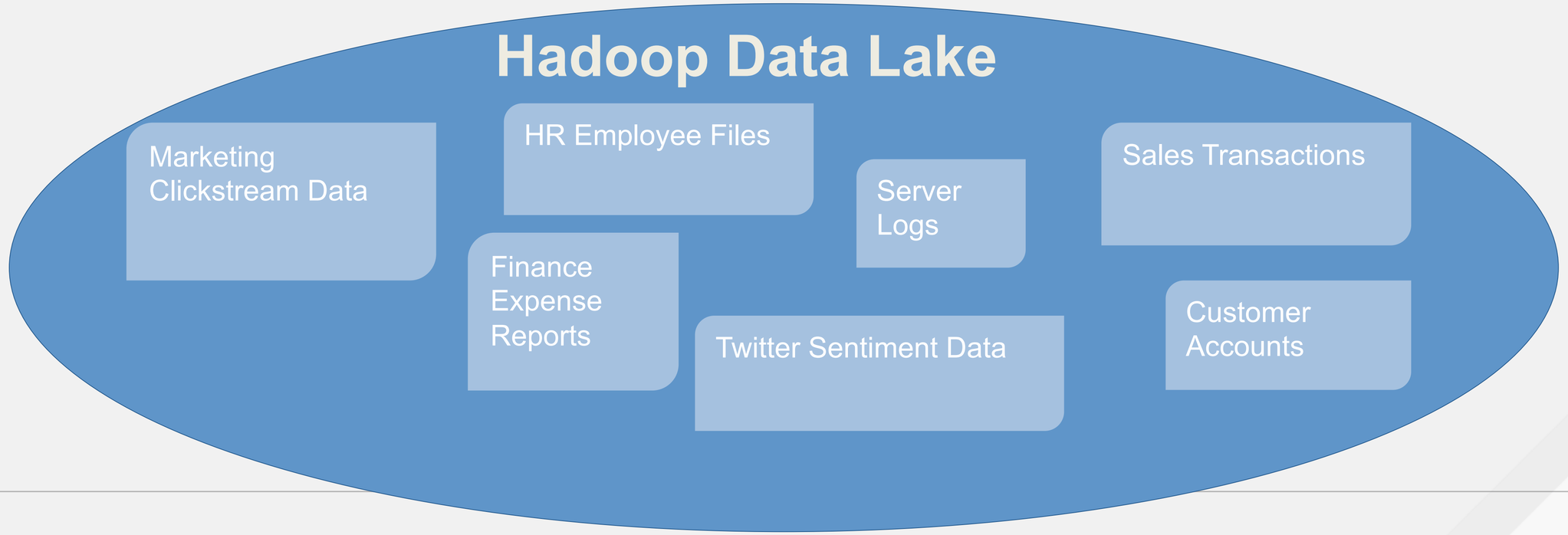
- SOURCE:

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?

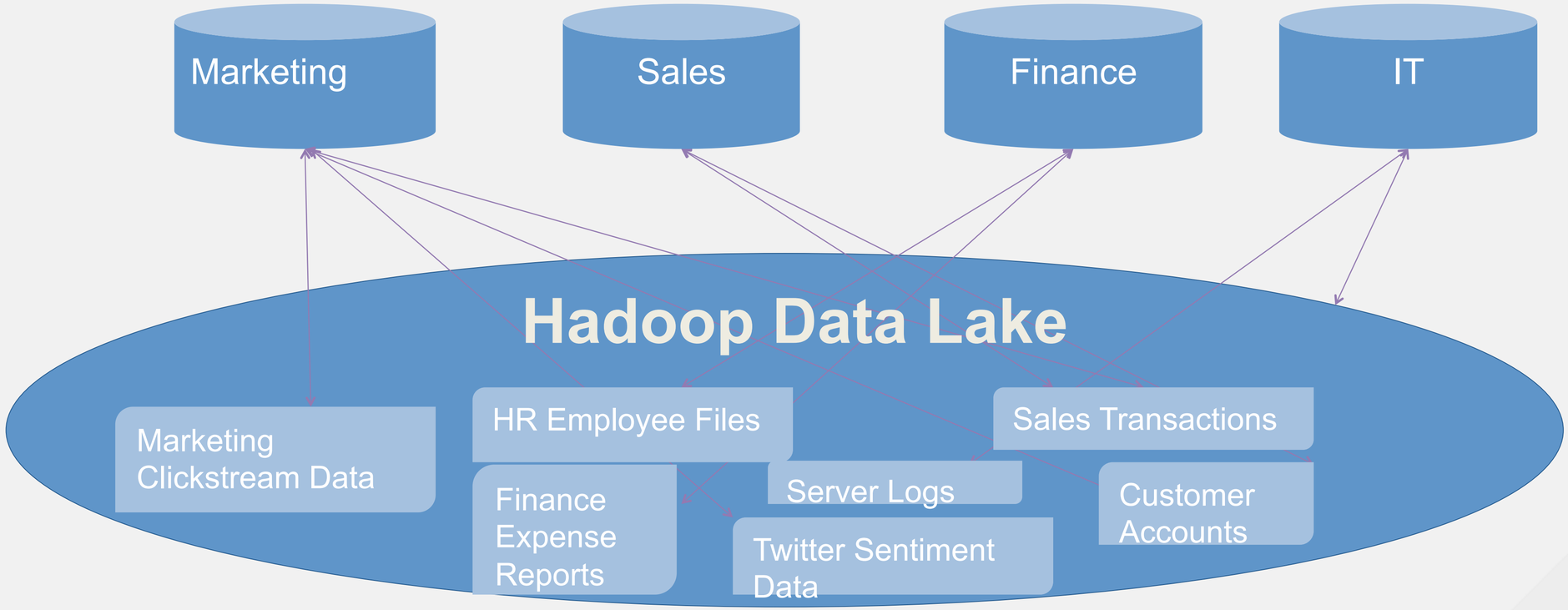
- Does every line of business have access to everyone's data?
- How do business users get access to the data they need in a simple (even self-service) way?



Secure, Self-Service Virtual Data Marts for Hadoop

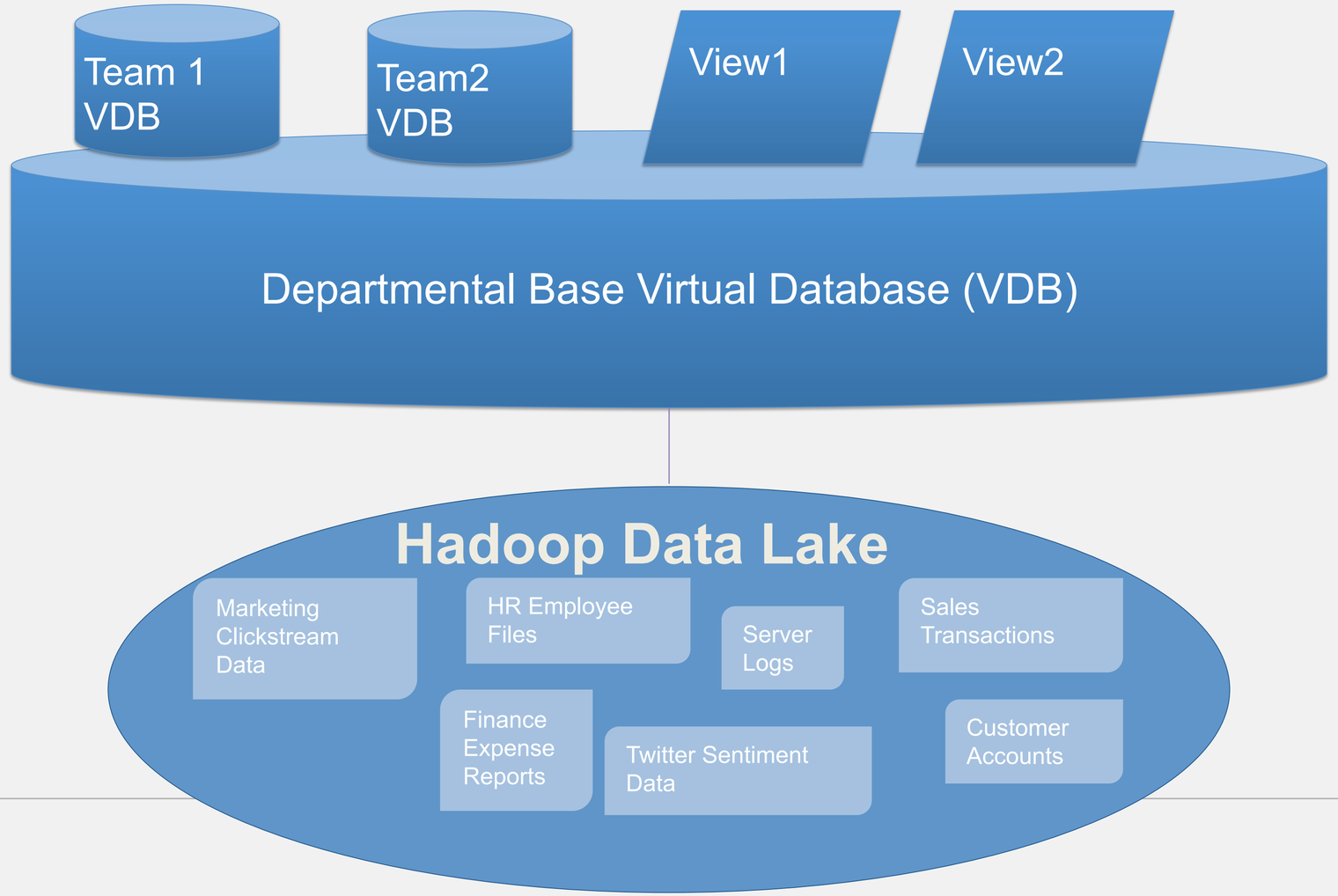
Solution: Use JBoss Data Virtualization to create virtual data marts on top of a Hadoop cluster

- 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

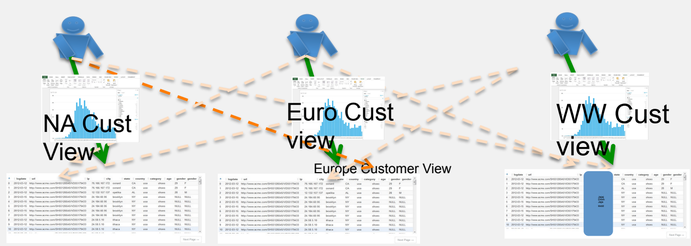
Cojan van Ballegooijen

Use Case 3 – Virtual data marts with Hadoop Data Lake

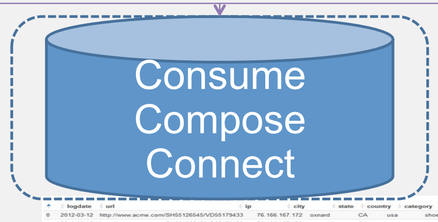
Objective:
–Purpose oriented data views for functional teams over a rich variety of semi-structured and structured data

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

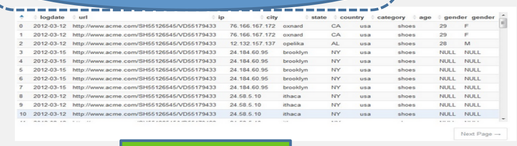
Solution:
–Leverage HDP to mashup Clickstream analysis data with product and customer data on HDP to answer
- Leverage Jboss Data Virt to provide Virtual data marts for each of Marketing and Product teams



JBoss Data Virtualization



Worldwide Customers



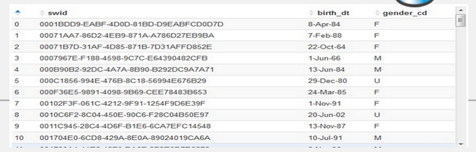
Fine-grained role-based access to geo specific data

Joins rich variety/volume of structured/ unstructured data

Omniture Clickstream logs

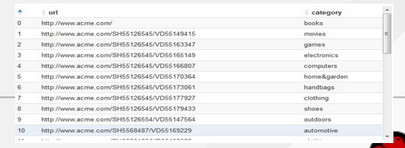
Cdjdkdd djjj djdjdjdjdj djjdjdjd djdjjd
Jjdjd jdjdjd djjdjd djjd
Jdjdj ddjjd djdjjd jdjdj
Jdjdjdj djdjdjdjdj djdjdj djjdjdjd

Users



	swid	birth_dt	gender_cd
0	0001BDD9-EA8F-4D0D-81BD-D9EA8FCD0D7D	8-Apr-84	F
1	00071AA7-86D2-4EB9-871A-A786D27EB9BA	7-Feb-88	F
2	00071B7D-31AF-4D85-871B-7D31AFFD852E	22-Oct-64	F
3	0007967E-F188-4598-9C7C-E64390482CFB	1-Jun-66	M
4	000B90E2-92DC-4A7A-8B90-B292DC9A7A71	13-Jun-84	M
5	000C1856-994E-476B-8C18-56994E676B29	29-Dec-80	U
6	000F36E5-9891-4098-9B69-CEE78483B653	24-Mar-85	F
7	00102F3F-061C-4212-9F91-1254F9D6E39F	1-Nov-91	F
8	0010C6F2-8C04-450E-90C6-F28C04B50E97	20-Jun-02	U
9	0011C945-28C4-4D6F-81E6-6CA7EFC14548	13-Nov-87	F
10	001704E0-6CD8-429A-8E0A-89024019CA6A	10-Jul-91	M

Products



	url	category
0	http://www.acme.com/	books
1	http://www.acme.com/SH55126545/VD55149415	movies
2	http://www.acme.com/SH55126545/VD55163347	games
3	http://www.acme.com/SH55126545/VD55165149	electronics
4	http://www.acme.com/SH55126545/VD55166807	computers
5	http://www.acme.com/SH55126545/VD55170364	home&garden
6	http://www.acme.com/SH55126545/VD55173061	handbags
7	http://www.acme.com/SH55126545/VD55177927	clothing
8	http://www.acme.com/SH55126545/VD55179433	shoes
9	http://www.acme.com/SH55126554/VD55147564	outdoors
10	http://www.acme.com/SH5568487/VD55169229	automotive

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.

FORRESTER®

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.

RHT Summit Sessions of interest

Connect to multiple data sources without writing code

Room	Date	Time	Technical difficulty
201	Wednesday, June 24	7:00 pm - 8:00 pm	3

Integration with Red Hat JBoss Fuse and Red Hat JBoss Data Virtualization

Room	Date	Time	Technical difficulty
207	Wednesday, June 24	4:50 pm - 5:50 pm	4

iPaaS & beyond: Red Hat's integration roadmap

Room	Date	Time	Technical difficulty
312	Friday, June 26	9:45 am - 10:45 am	

Red Hat Storage Server as a hybrid storage solution for Splunk Enterprise

Room	Date	Time	Technical difficulty
209	Wednesday, June 24	4:50 pm - 5:50 pm	2

Building a big data, risk-management solution for financial services

Room	Date	Time	Technical difficulty
207	Thursday, June 25	3:40 pm - 4:40 pm	2

Make a data-driven investment decision about big data

Room	Date	Time	Technical difficulty
309	Friday, June 26	9:45 am - 10:45 am	2

Using Apache Spark to build analytical applications in the Cloud

Room	Date	Time	Technical difficulty
208	Wednesday, June 24	2:30 pm - 3:30 pm	

Big data on the open private cloud

Room	Date	Time	Technical difficulty
310	Thursday, June 25	2:30 pm - 3:30 pm	3

Building a big data, risk-management solution for financial services

Room	Date	Time	Technical difficulty
207	Thursday, June 25	3:40 pm - 4:40 pm	2

Drinking from the firehose with duct-tape-free reactive Java applications

Room	Date	Time	Technical difficulty
206	Tuesday, June 23	10:30 am - 11:30 am	3

OpenStack nirvana: Big data & elastic infrastructure together at last

Room	Date	Time	Technical difficulty
302	Wednesday, June 24	2:30 pm - 3:30 pm	2

RED HAT
SUMMIT

LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.