Apache Hadoop and the Big Data Opportunity in Banking

A webcast from Hortonworks & Tresata

© Hortonworks Inc. 2011

© tresata all rights reserved

Presenters

- Arun C. Murthy
  - Co-founder of Hortonworks
  - Lead of NextGen MapReduce in Apache Hadoop
  - Long-time contributor and committer to Apache Hadoop

- Abhishek (Abhi) Mehta
  - Co-founder of Tresata
  - Creator of first Hadoop-powered big data & analytics platform for financial industry data

Agenda

- What is Apache Hadoop?
- Creating Value from Big Data
- Tresata and Hadoop for Banking
- Future of Hadoop

What is Apache Hadoop?

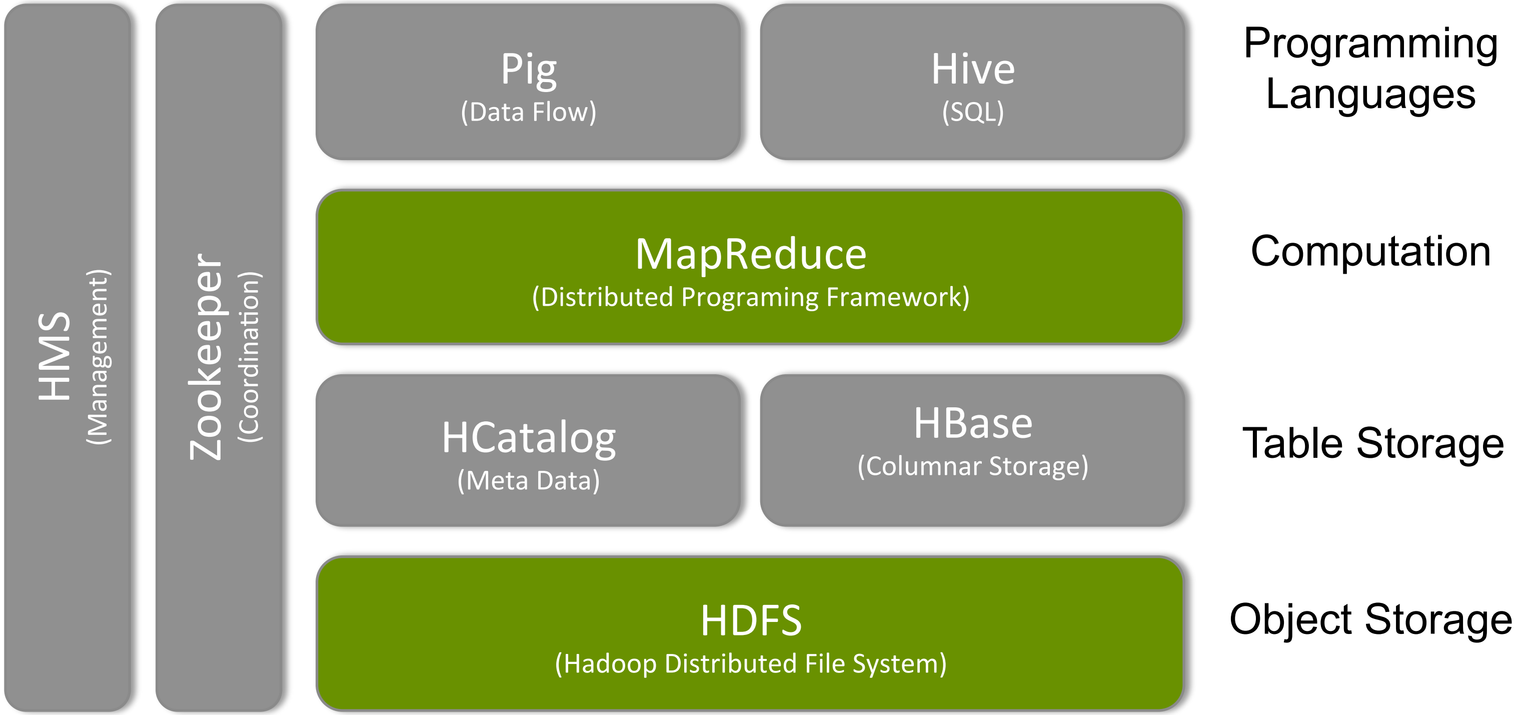
A set of **open source** projects owned by the Apache Foundation that transforms **commodity computers** and network into a **distributed service**

- HDFS – Stores petabytes of data reliably
- MapReduce – Allows huge distributed computations

Key Attributes

- **Reliable and redundant** – Doesn't slow down or loose data even as hardware fails
- **Very powerful** – Harnesses huge clusters, supports best of breed analytics
- **Scalable** – scales linearly to handle "big data" volumes
- **Cost-effective** – runs on commodity machines & network
- **Simple and flexible APIs** – enabling a large ecosystem of solution providers

Core Apache Hadoop Projects



HMS
(Management)

Zookeeper
(Coordination)

Pig
(Data Flow)

Hive
(SQL)

Programming Languages

MapReduce
(Distributed Programing Framework)

Computation

HCatalog
(Meta Data)

HBase
(Columnar Storage)

Table Storage

HDFS
(Hadoop Distributed File System)

Object Storage

🟩 Core Apache Hadoop ▨ Related Apache Projects

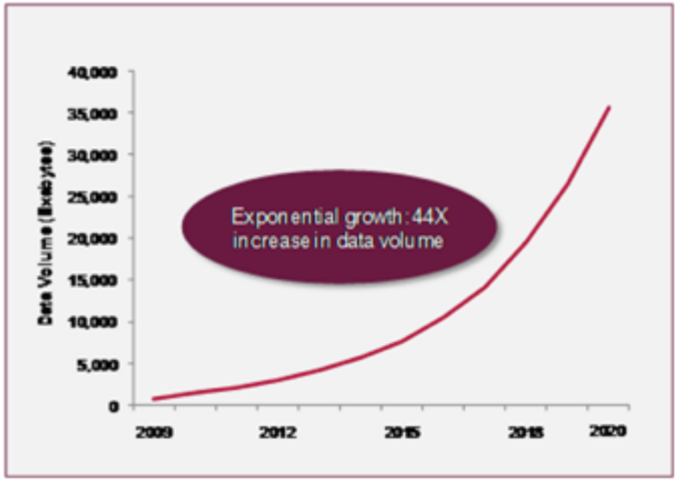
tresata

hortonworks

Apache Hadoop: Why is it Transformational?

Data Deluge (growth faster than Moore's law)

- Economist: Only 5% of generated data is structured
- Gartner: Data growth is the biggest data center hardware infrastructure challenge for large enterprises
- Forrester: Four Vs - Volume, Velocity, Variety, Variability
- Hundreds of ***exabytes*** of data per year!



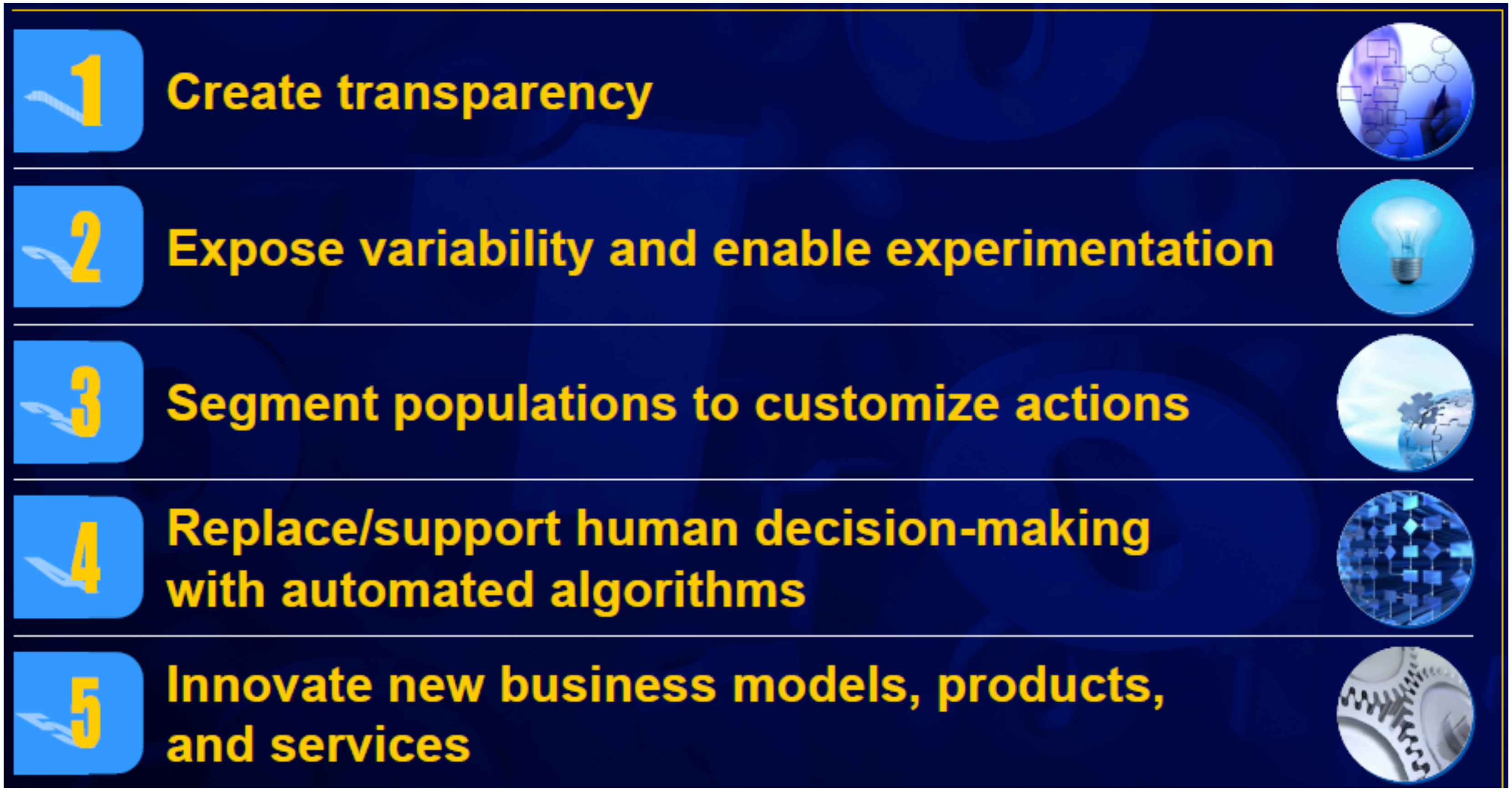
Source: IDC Digital Universe Study, May 2011

Apache Hadoop: Why is it Transformational?

New way of thinking about your data:

- Current
  - What data do I keep?
  - What reports do I run?
  - Sample – Store – Extrapolate → *What-If* scenarios → Variable *insights*
- New dawn
  - Store everything (viable and economical) – Process whenever!
  - Test every what-if situation, prove every hypothesis… do it in a timely manner!
  - No more down-sampled data

5 Ways to Create Value from Big Data

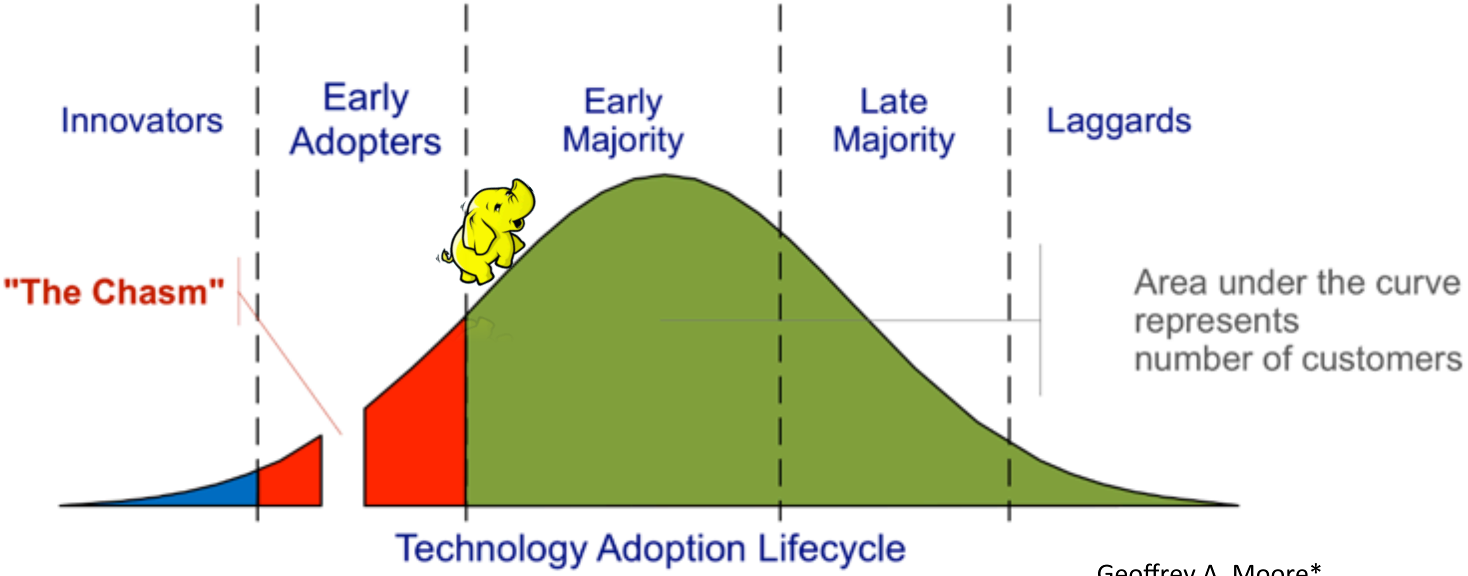


1. Create transparency
2. Expose variability and enable experimentation
3. Segment populations to customize actions
4. Replace/support human decision-making with automated algorithms
5. Innovate new business models, products, and services

Source: McKinsey & Company report. Big data: The next frontier for innovation, competition, and productivity. May 2011.

Crossing the Chasm

Disruption: Data


Innovators
Early Adopters
Early Majority
Late Majority
Laggards
"The Chasm"
Area under the curve represents number of customers
Technology Adoption Lifecycle

Geoffrey A. Moore*

Typical Applications & Early Adopters

Y!

Google

facebook

analyzing web logs

data analytics

advertising optimization

machine learning

mail anti-spam

text mining

web search

content optimization

customer trend analysis

ad selection

video & audio processing

data mining

user interest prediction

social media

Powerset

ebay

IBM

Linked in

Bai du 百度

AOL

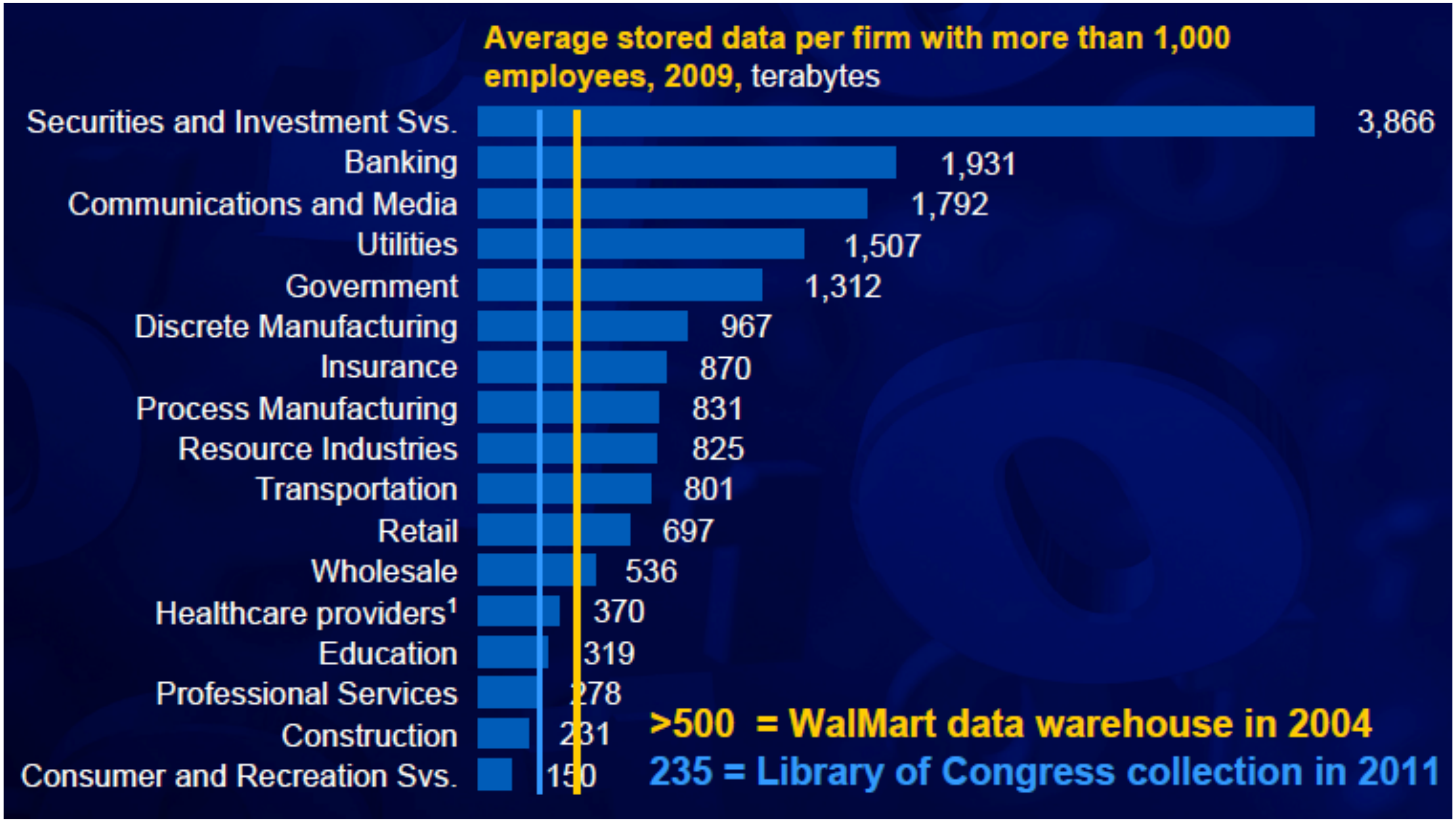
last.fm

tresata

hortonworks

Big Data Has Reached Every Market Sector

Digital data is personal, everywhere, increasingly accessible, and will continue to grow exponentially



Average stored data per firm with more than 1,000 employees, 2009, terabytes

>500 = WalMart data warehouse in 2004

235 = Library of Congress collection in 2011

Source: McKinsey & Company report. Big data: The next frontier for innovation, competition, and productivity. May 2011.

Big Data Value Creation Opportunities

Financial Services

- Detect fraud
- Model and manage risk
- Improve debt recovery rates
- Personalize banking/insurance products

Healthcare

- Optimal treatment pathways
- Remote patient monitoring
- Predictive modeling for new drugs
- Personalized medicine

Retail

- In-store behavior analysis
- Cross selling
- Optimize pricing, placement, design
- Optimize inventory and distribution

Web / Social / Mobile

- Location-based marketing
- Social segmentation
- Sentiment analysis
- Price comparison services

Manufacturing

- Design to value
- Crowd-sourcing
- “Digital factory” for lean manufacturing
- Improve service via product sensor data

Government

- Reduce fraud
- Segment populations, customize action
- Support open data initiatives
- Automate decision making

Why can you bank on Hortonworks?

- Architects of Hadoop since big bang, circa 2006
- Real world experience supporting the largest Hadoop install in the world:
  - 50,000 node footprint
  - Over 200PB of data
  - 24x7 service
  - Billions of ad dollars
- ***We've taken the 3am calls to fix stuff when it breaks!***

Tresata = Hadoop for Banking

What We Believe

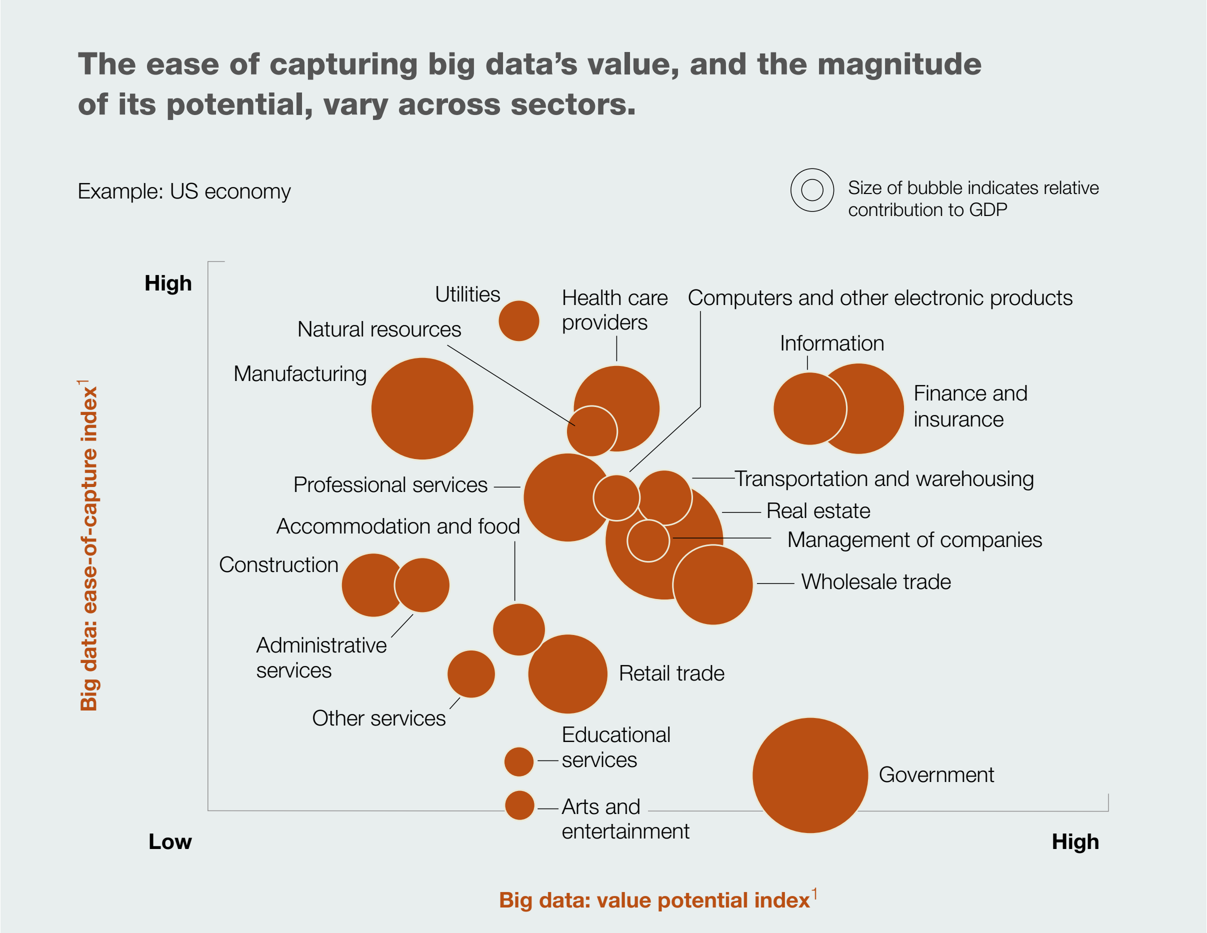
1. **Data will reboot** the financial service industry

2. Data growth is **viral**…existing tools can't keep up

3. The **economics** to store, process, analyze and visualize **all** of your data makes it a **'no-brainer'** to do so

4. **Big Data capabilities** are needed to address **business problems**

and What McKinsey Said



Source: mcksinsey global institute october 2011

The Opportunity

1. **1-5% of data** in a financial institution **is analyzed**

2. **Not all data is stored**

3. **Top-down macros** cannot be implemented or acted on

4. **New approaches to data & analytics** are essential

Source: tresata research

The Business Problems

1. **Storage & retrieval –** archive data on disk

2. **Consumer behavior analysis –** as it happens

3. **Modeling & Analytics** – model entire data sets…

4. **Single View Of Customer** – structured & unstructured data

Hadoop in Banking Today

1. **Early adopter**, like any other technology trend
   a. Interest is **global**, not just limited to the US
   b. **Approved for use** by most technology teams & architects
   c. **Proof of concepts** ongoing at most financial services institutions

2. **Broad agreement** that:
   a. **Hadoop** will be the **Big Data operating system**
   b. A hadoop powered **Data processing & analytics platform** will spur rapid adoption
   c. Need to apply to **business problems** with revenue/profit impact

Elephants in the Room

1. Resources

a. **Talent** – how many MapReduce developers can we get

b. **Applications** – data, analytics and business problems are the same, why should each institution build their own hadoop applications to run the same processes

c. **Essentials** – how to manage security, provisioning, & performance

2. Implementation

a. **Integration** – fit with existing technology infrastructures & business processes

b. **Support** – experience with managing thousands of nodes/ servers

c. **Open Source** – 'out of the box' applicability

How Tresata Changes the Game…

1. **Our Application**
   a. **FS for Hadoop** – fully built on hadoop. Store, process, analyze and visualize leveraging the full power of hadoop
   b. **Processing Pipeline** – automated ingestion, cleaning, de-duping, matching engine built for financial data
   c. **Analytics** – massively parallel scoring and algorithm containers codified by business problem

2. **Tame the elephants** (making it work at scale)

Tresata Case Study

A. Client Business Problem

i. Problem – Process data and score for >**30 MM client** applications

ii. Data Sources – **23** separate data sources, **multiple time series**

iii. Raw Variables – **100 variables** per client per data source

iv. Current state - **expensive** legacy platform, algorithms developed on **sub-samples**, **unable to scale** algorithms to full data set

B. Tresata Solution

i. Data Engine – **automated** data import, cleaning, matching, scoring

ii. Compute Engine – algorithms process & score >30MM in **minutes**

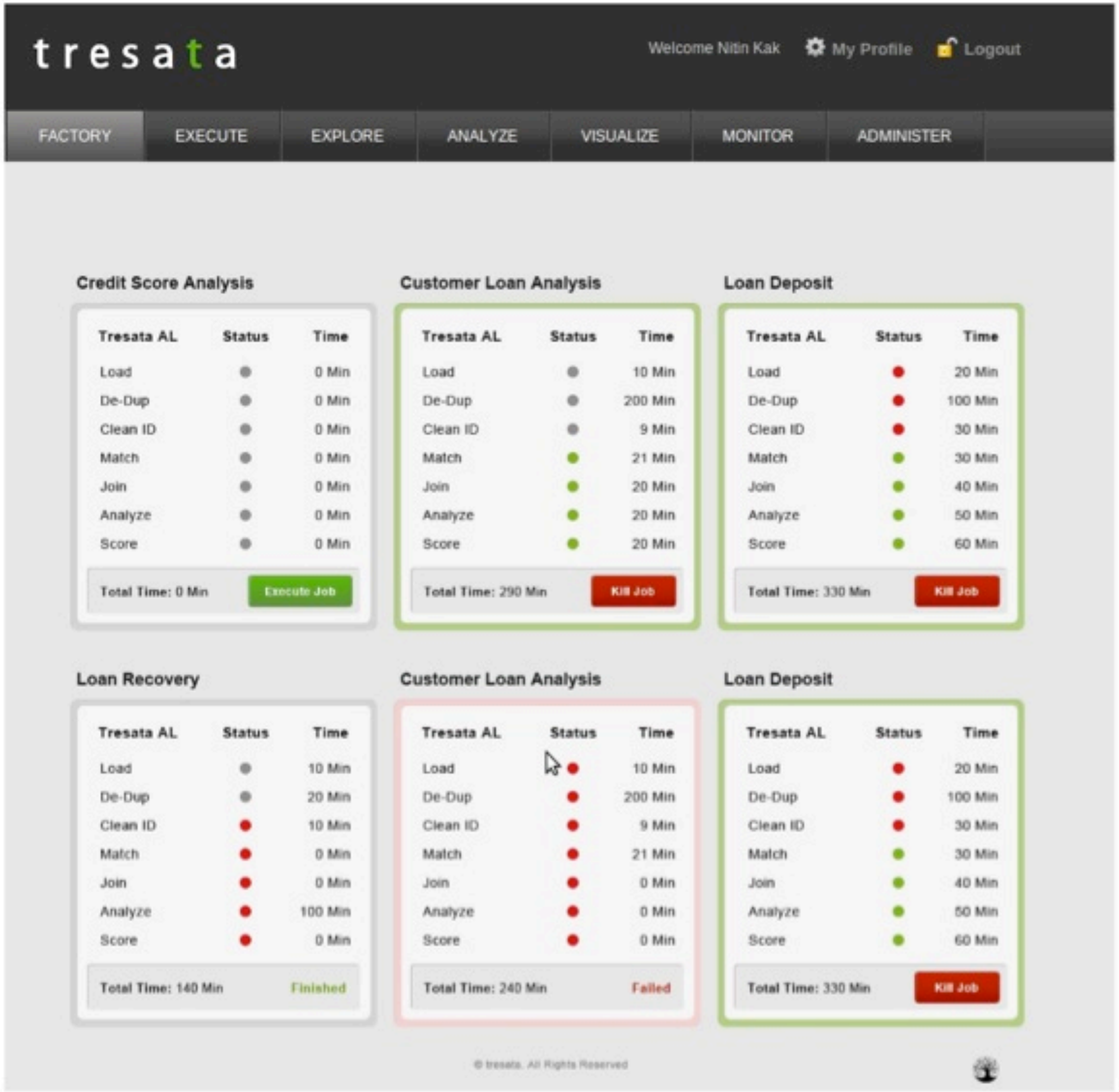
iii. Integration – work with existing tools and processes

iv. Scalable deployment – **Big Data as a Service** delivery

tresata

hortonworks

Tresata Big Data Application



Credit Score Analysis

Tresata AL	Status	Time
Load	●	0 Min
De-Dup	●	0 Min
Clean ID	●	0 Min
Match	●	0 Min
Join	●	0 Min
Analyze	●	0 Min
Score	●	0 Min

Total Time: 0 Min — Execute Job

Customer Loan Analysis

Tresata AL	Status	Time
Load	●	10 Min
De-Dup	●	200 Min
Clean ID	●	9 Min
Match	🟢	21 Min
Join	🟢	20 Min
Analyze	🟢	20 Min
Score	🟢	20 Min

Total Time: 290 Min — Kill Job

Loan Deposit

Tresata AL	Status	Time
Load	🔴	20 Min
De-Dup	🔴	100 Min
Clean ID	🔴	30 Min
Match	🟢	30 Min
Join	🟢	40 Min
Analyze	🟢	50 Min
Score	🟢	60 Min

Total Time: 330 Min — Kill Job

Loan Recovery

Tresata AL	Status	Time
Load	●	10 Min
De-Dup	●	20 Min
Clean ID	🔴	10 Min
Match	🔴	0 Min
Join	🔴	0 Min
Analyze	🔴	100 Min
Score	🔴	0 Min

Total Time: 140 Min — Finished

Customer Loan Analysis

Tresata AL	Status	Time
Load	🔴	10 Min
De-Dup	🔴	200 Min
Clean ID	🔴	9 Min
Match	🔴	21 Min
Join	🔴	0 Min
Analyze	🔴	0 Min
Score	🔴	0 Min

Total Time: 240 Min — Failed

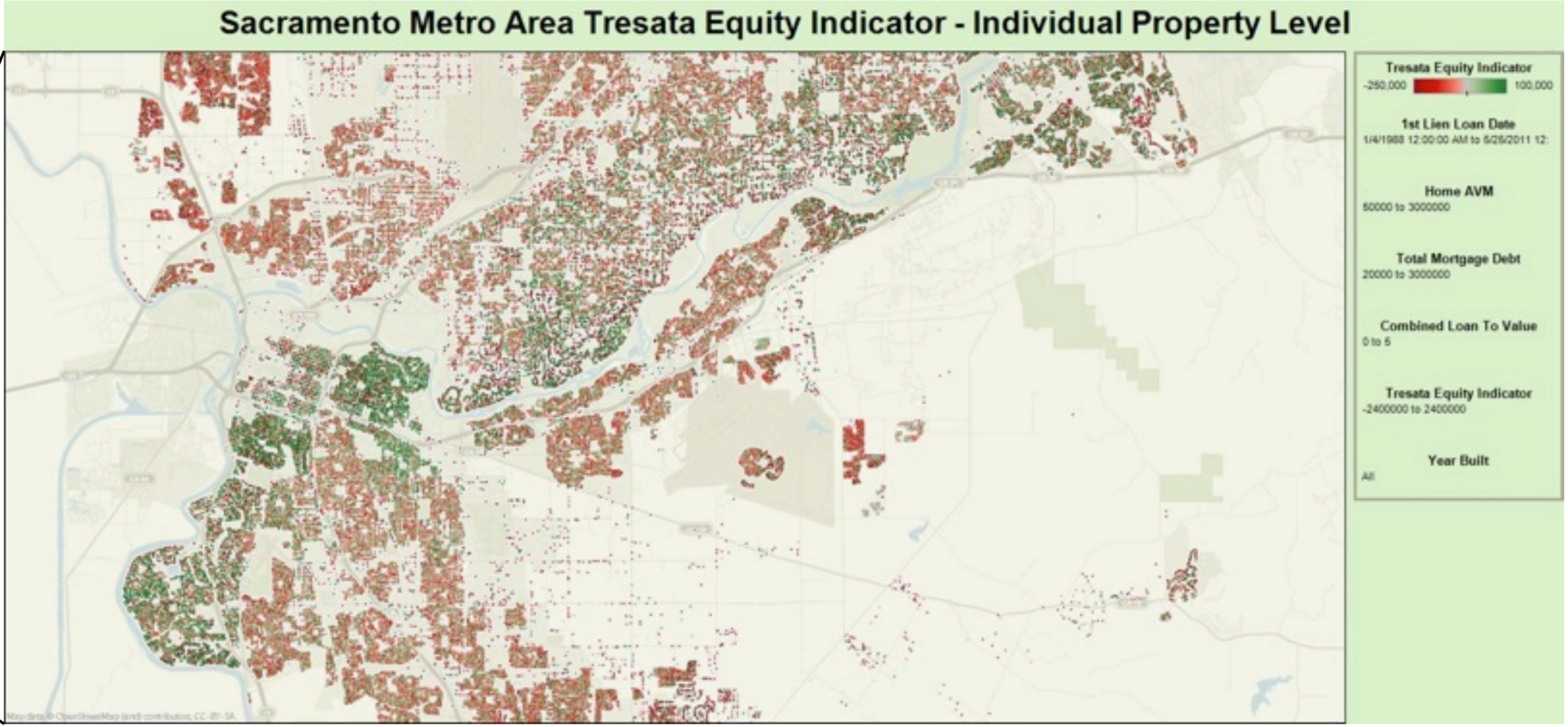
Loan Deposit

Tresata AL	Status	Time
Load	🔴	20 Min
De-Dup	🔴	100 Min
Clean ID	🔴	30 Min
Match	🟢	30 Min
Join	🟢	40 Min
Analyze	🟢	50 Min
Score	🟢	60 Min

Total Time: 330 Min — Kill Job

© tresata. All Rights Reserved

Tresata Analytics



Sacramento Metro Area Tresata Equity Indicator - Individual Property Level

Tresata Equity Indicator
-250,000 🟥🟩 100,000

1st Lien Loan Date
1/4/1988 12:00:00 AM to 6/26/2011 12:

Home AVM
50000 to 3000000

Total Mortgage Debt
20000 to 3000000

Combined Loan To Value
0 to 5

Tresata Equity Indicator
-2400000 to 2400000

Year Built
All

Tresata + Hortonworks

1. Commitment to train

a. Series of webinars

b. Hadoop training tailored for financial services

c. Dedicated training programs (tailored for client needs)

2. Commitment to support

a. Production scale support model (Tresata certified for Financial Svcs)

b. Meet and/or exceed industry standards on distro

hortonworks

tresata

Where Hadoop is Going

© Hortonworks Inc. 2011

© tresata   all rights reserved

The Future

- Hadoop is going mainstream
  - Amazon, Microsoft, Oracle, EMC, NetApp, etc.
- Apache Hadoop is covering new ground (Hadoop .Next)
  - Much of the development led by Hortonworks
  - More scale, more performance
  - Other paradigms than MapReduce for data processing
  - Enhanced operability and management (Ambari)
  - Metadata management (Hcatalog)

Technology Roadmap

Hadoop.Now – Making Apache Hadoop Accessible • Release the most widely deployed version of Hadoop ever (0.20.205) • Release directly usable code via Apache (RPMs, .debs…) • Frequent sustaining releases off of the stable branches	**Q4 2011**
Hadoop.Next – Next Generation Apache Hadoop • Address key product gaps (HBase support, HA, Management…) • Enable community & partner innovation via modular architecture & open APIs • Work with community to define integrated stack	**2012** (Alphas starting late 2011)

Next Steps

- Engage with Hortonworks & Tresata
  - www.hortonworks.com
  - www.tresata.com
- Additional Webcast Series
  - Reference Architecture for Hadoop in Financial Services - Nov 15
  - Hadoop in Financial Services DEEP DIVE - Dec 6
  - www.hortonworks.com/webcasts
- Hortonworks party @ Hadoop World
  - Tuesday Nov 8 at 7pm
  - Inc Lounge at the Time Hotel
  - RSVP: hortonworks.eventbrite.com

Questions?

© Hortonworks Inc. 2011

© tresata all rights reserved