Simplifying the Process of Uploading and Extracting Data from Apache Hadoop

Rohit Bakhshi, Solution Architect, Hortonworks

Jim Walker, Director Product Marketing, Talend

About Us

Rohit Bakhshi

Solution Architect at Hortonworks

Experience

- Hadoop in enterprise architecture
- Building advanced analytical applications

Enjoys live jazz and drinking espresso

Jim Walker

talend*
*open integration solutions

Director Product Marketing at Talend

Experience

- 10 years as developer, 10 as marketer
- Computer Security, DQ, MDM… Big data

Is a bit of a foodie and enjoys baseball (White Sox)

Agenda

- **Introduction**
- **Impact Big Data in the Enterprise**
- **Hortonworks Data Platform**
- **Talend Overview**
- **Demo**
- **Q&A**

Hortonworks Vision

We believe that by the end of 2015, more than half the world's data will be processed by Apache Hadoop

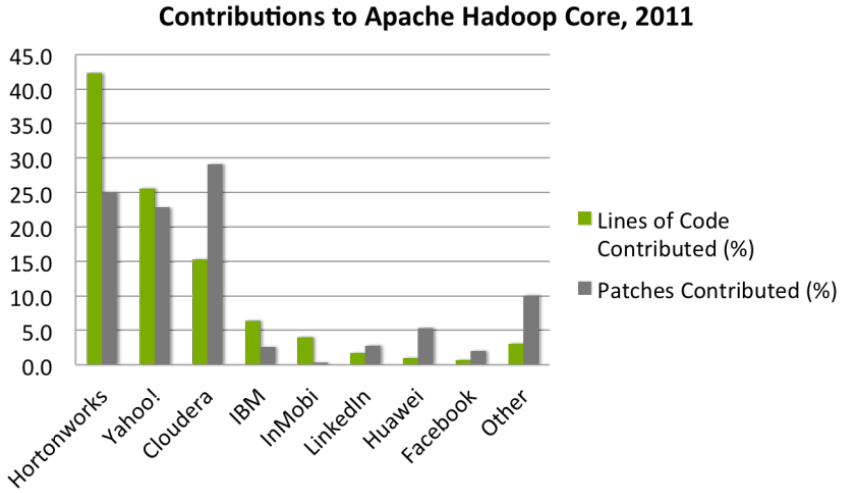
How to achieve that vision???

Enable ecosystem around enterprise-viable open source data platform.

Hortonworks Strategic Focus

Enable Hadoop to be next-generation enterprise data platform

- **Lead within Hadoop Community**
  - Engineering team that delivered every major Hadoop release since 0.1
  - Experience managing world's largest deployment
  - Ongoing access to Y!'s 1,000+ users and 40k+ nodes for testing, QA, etc.



- **Unify & Enable Hadoop Ecosystem**
  - Provide 100% open source product
  - Empower customers and partners overcome Hadoop knowledge gaps
  - Enable organizations successfully develop and deploy solutions based on Hadoop

Expert Role-based Training

Hortonworks CERTIFIED DEVELOPER

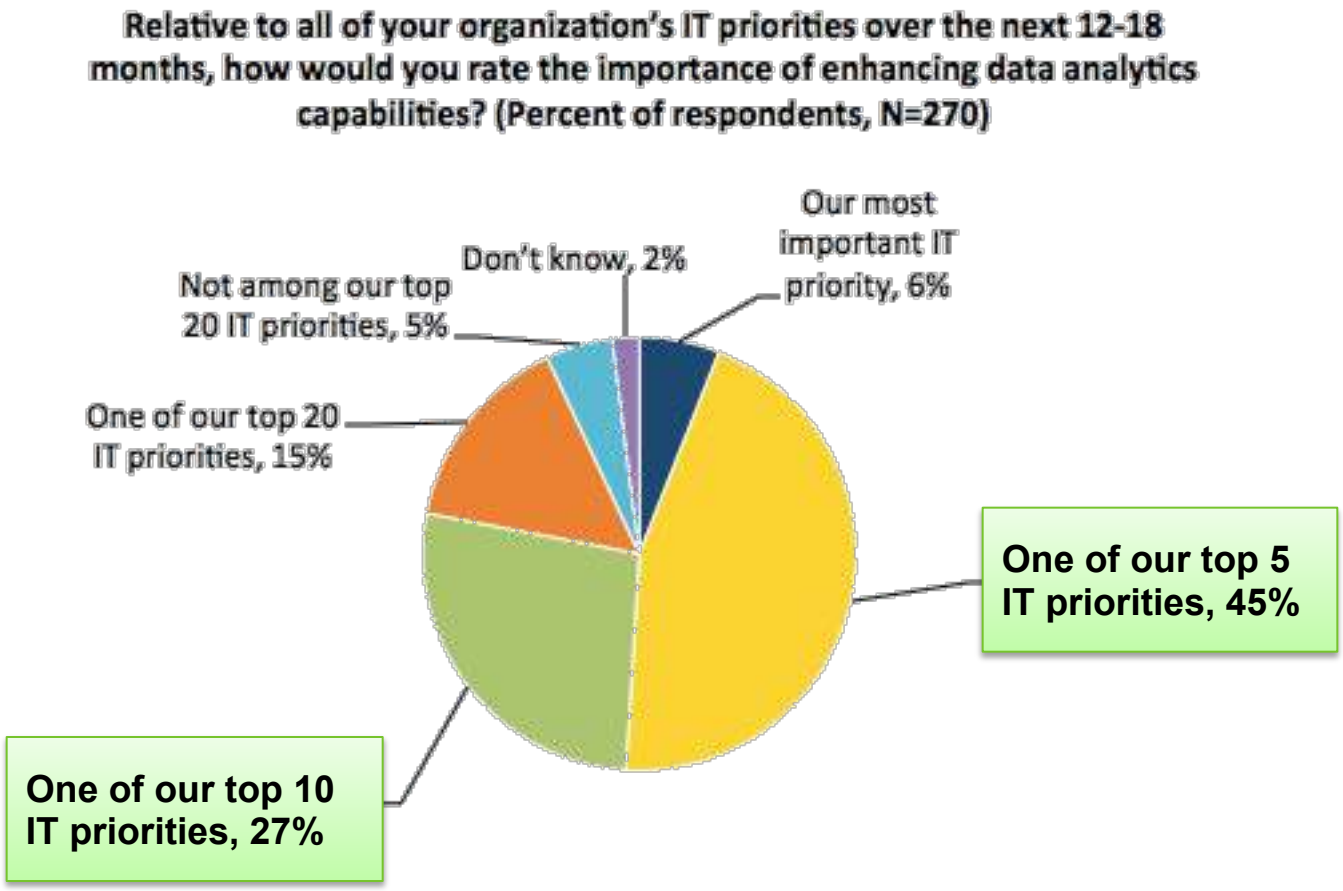
Hortonworks CERTIFIED ADMINISTRATOR

Hortonworks CERTIFIED DATA ANALYST

Full Lifecycle Support and Services

Evaluate → Pilot → Production

Impact of Big Data on Data Analytics



Source: Enterprise Strategy Group, 2012

Transactions – Interactions – Observations

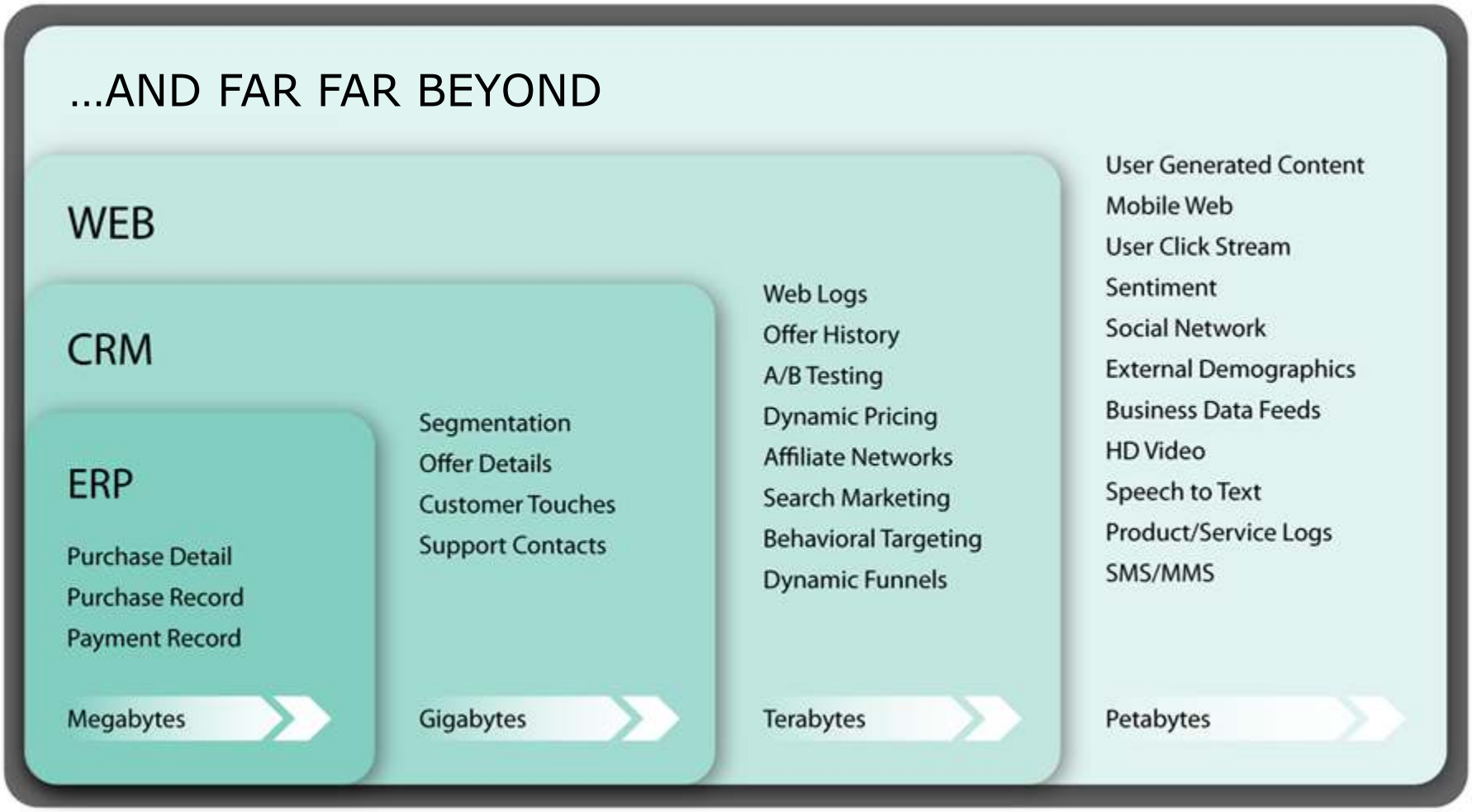


Chart content courtesy of Teradata, Inc.

Data-Driven Business

The days are over when you build a product once and it just works.

You have to take ideas, test them, iterate them, use data and analytics to understand what works and what doesn't in order to be successful.

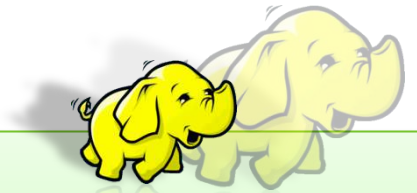
And that's how we use our big data infrastructure.

Aaron Batalion, CTO of LivingSocial
The Big Promise of Big Data, PCWorld, March 13, 2012

http://www.pcworld.com/businesscenter/article/251754/the_big_promise_of_big_data.html

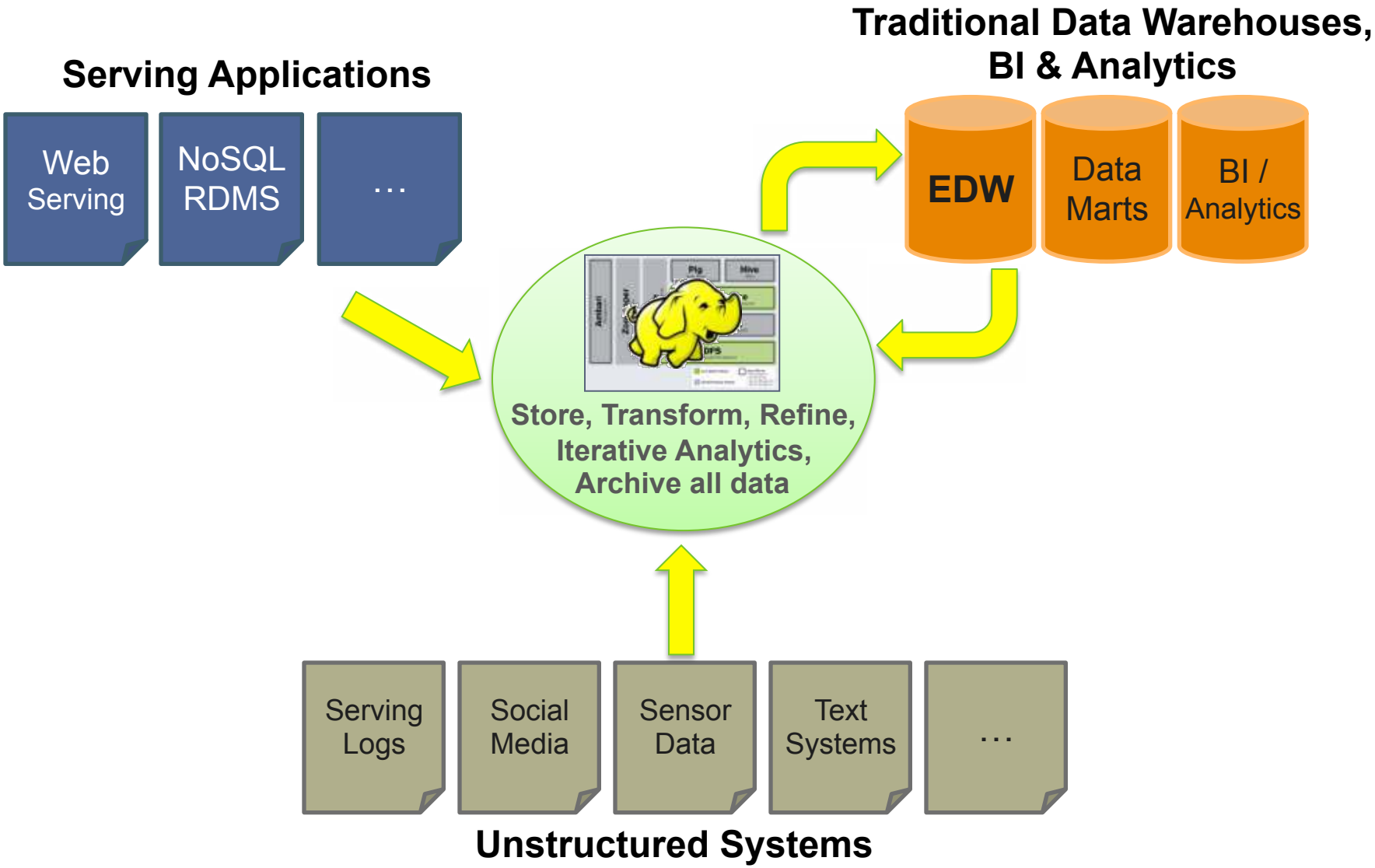
What is Apache Hadoop?

- **Solution for Big Data**
  - Designed for volume, velocity, variety & complexity of data
- **Data Platform Deployed on Commodity Hardware that**
  - Stores petabytes of data reliably
  - Runs highly distributed applications
  - Enables a rational economics model
- **Set of Open Source Projects**
  - Apache Software Foundation
  - Loosely coupled, ship early/often



One of the best examples of open source driving innovation and creating a market

Connecting All Of Your Big Data



Bridging Classic & Big Data Worlds

Integrating EDW & Hadoop

"Capture only what's needed"

Classic Method

Structured & Repeatable Analysis

?

Business determines what questions to ask

IT structures the data to answer those questions

EDW

Hadoop

"Capture in case it's needed"

?

IT delivers a platform for storing, refining, and analyzing **all data sources**

Big Data Method

Multi-structured & Iterative Analysis

Business explores data for questions worth answering

Bridging Classic & Big Data Worlds

Enabling Developers, Data Scientists, and Business Analysts

Java, C/C++, Pig, JavaScript, Python, R, SAS, SQL, Excel, Reporting, etc.

Ingest, Transform, Archive, Discover, Explore, Analyze, Report

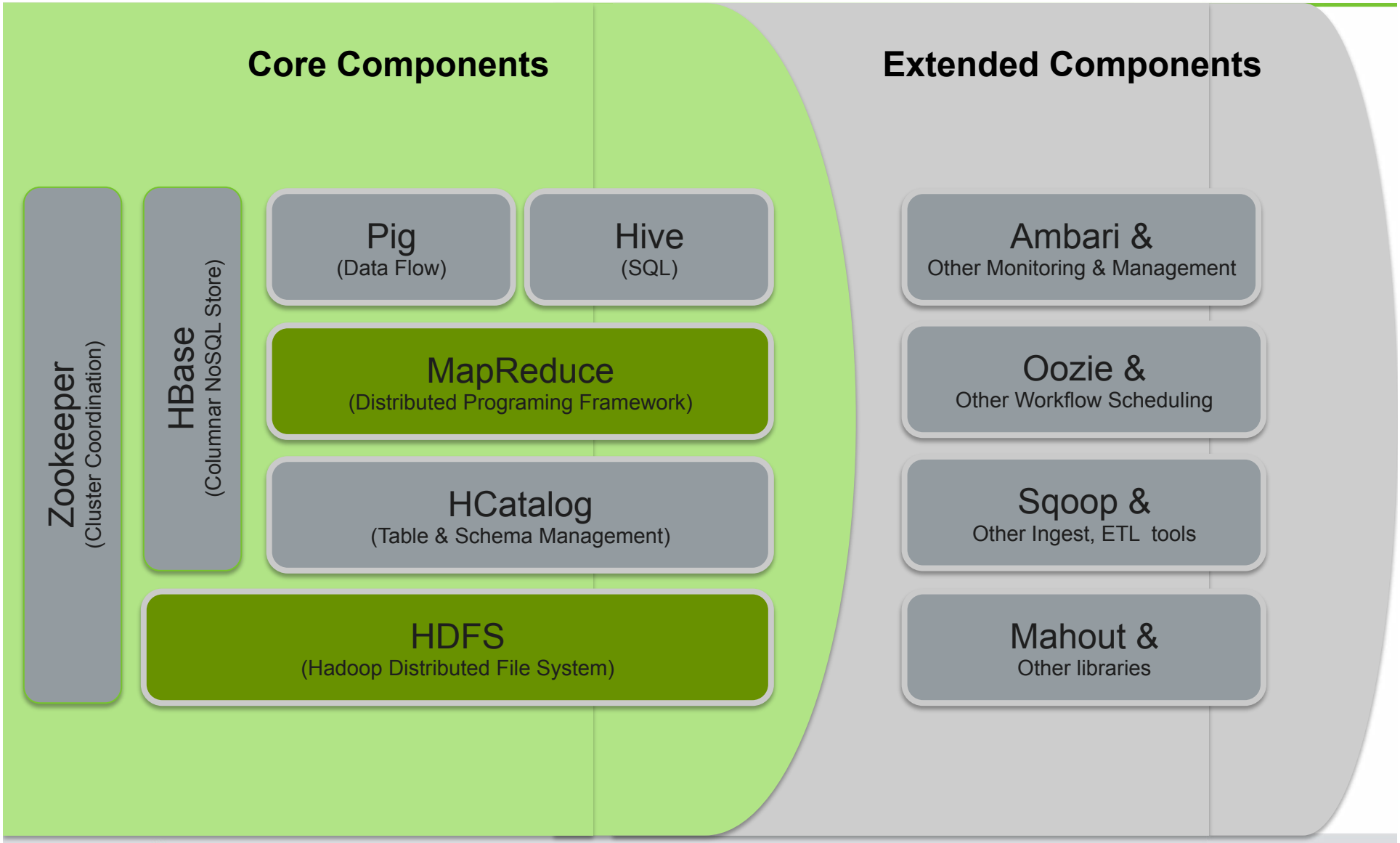
- Fast data loading
- ELT/ETL and refinement
- Iterative analysis
- Online archival

- Path & pattern analysis
- Graph analysis
- Text analysis
- Machine learning

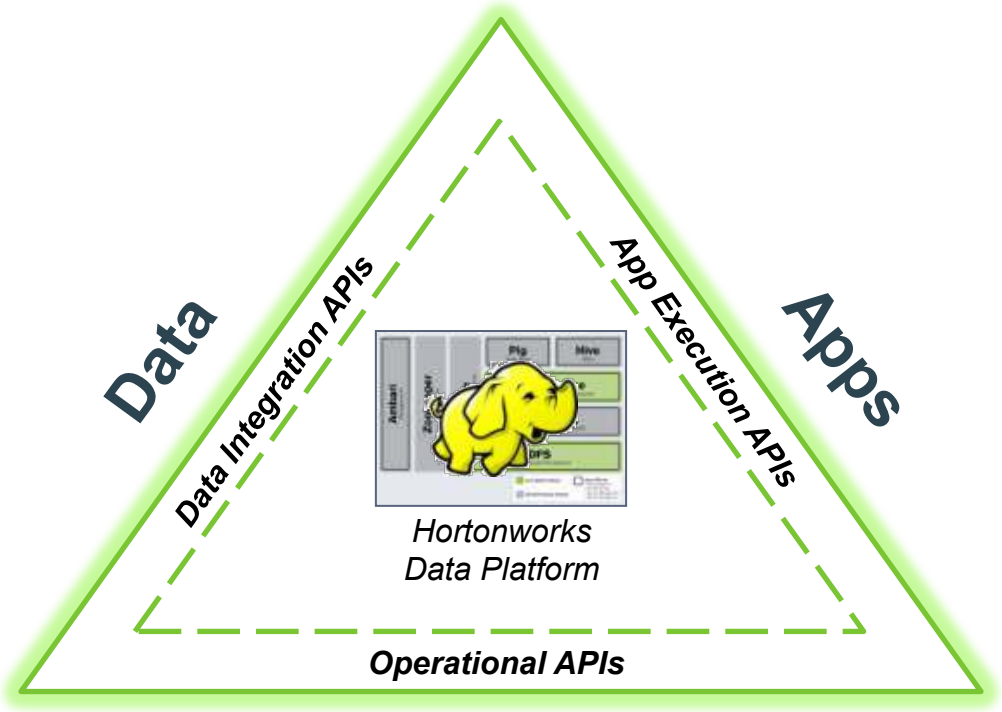
- Operational analysis
- Transactional analysis
- High volume ad-hoc
- Elastic data marts

Batch **Interactive** **Active**

Key Components of Hadoop Stack



Enable Ecosystem Around Platform

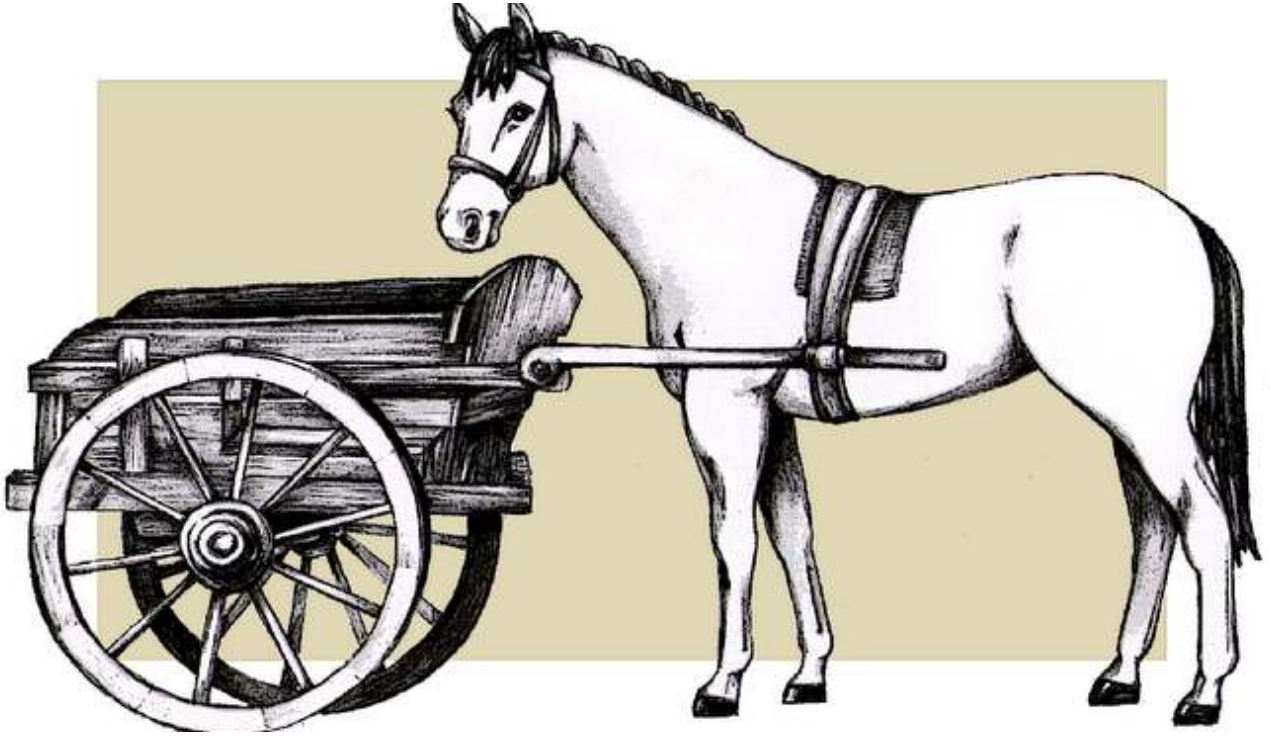


The market needs a platform that is Open across 3 facets:

- ***Data****: directly processes as well as coexists/integrates with any data flowing through a business*
- ***Apps****: delivers business value by enabling innovative new apps and enhancing existing apps*
- ***Operations****: integrates with operational models within the enterprise datacenter and the cloud*

Talend and big data…

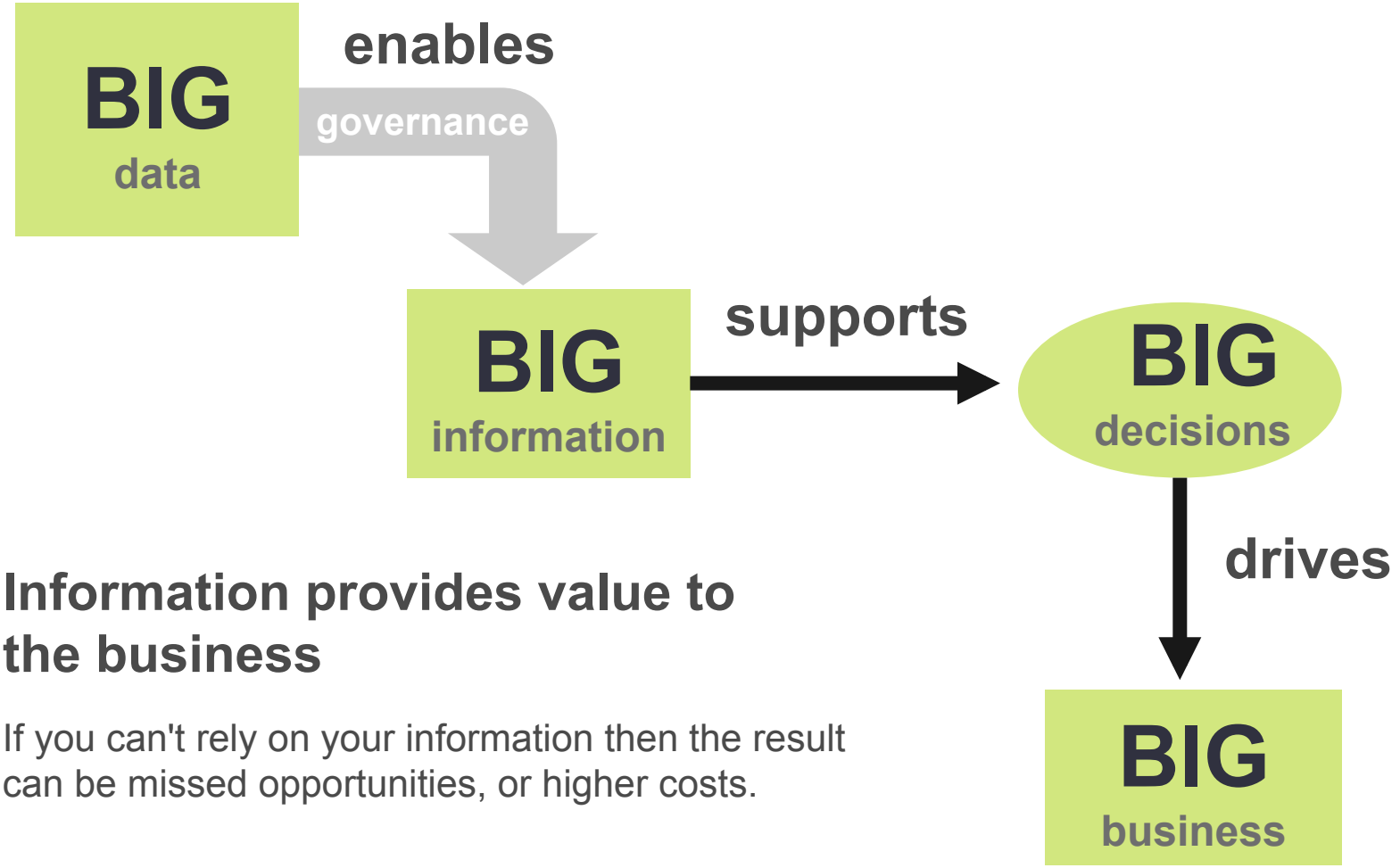
talend*
*open integration solutions



…everything old is new again!

talend
*open integration solutions

Challenge 1: Where project management?



Information provides value to the business

If you can't rely on your information then the result can be missed opportunities, or higher costs.

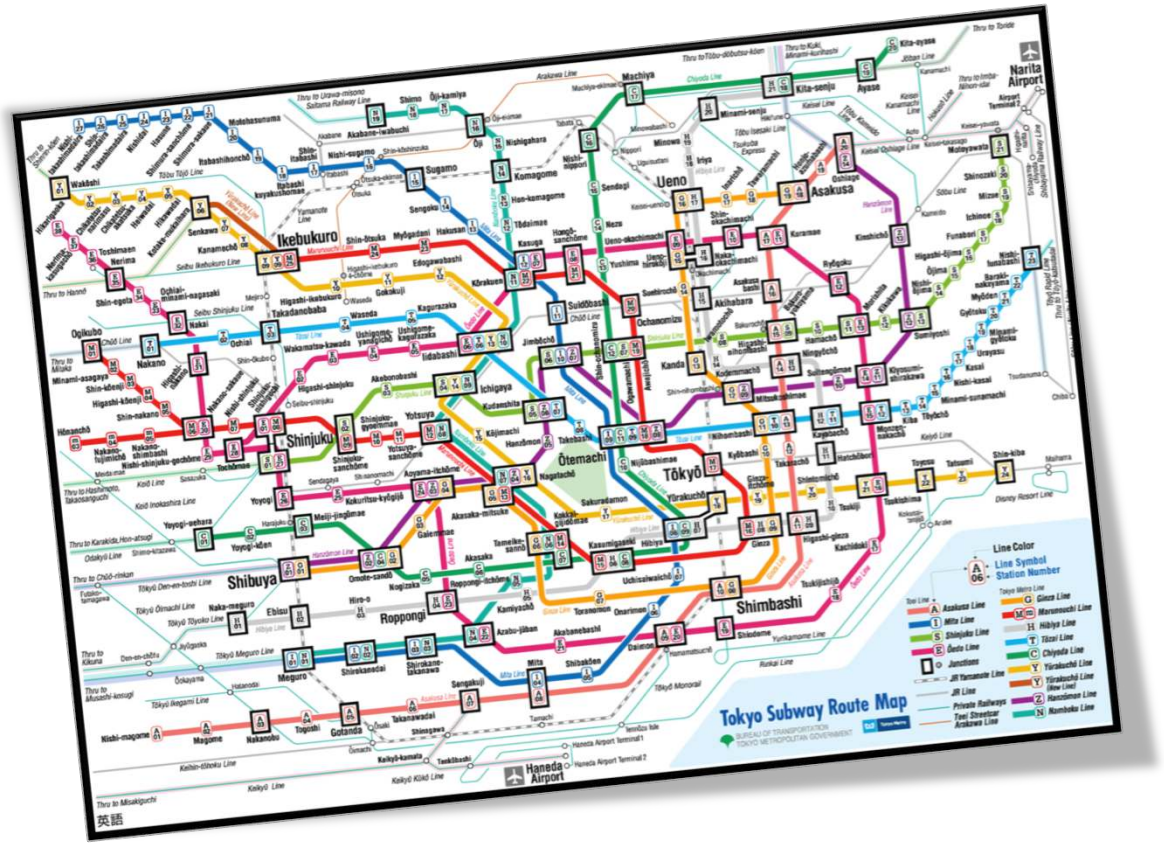
Matthew West and Julian Fowler (1999). Developing High Quality Data Models. The European Process Industries STEP Technical Liaison Executive (EPISTLE).

talend
*open integration solutions

Challenge 2: Data Quality

talend
*open integration solutions

Challenge 3: Complex technology, limited resources



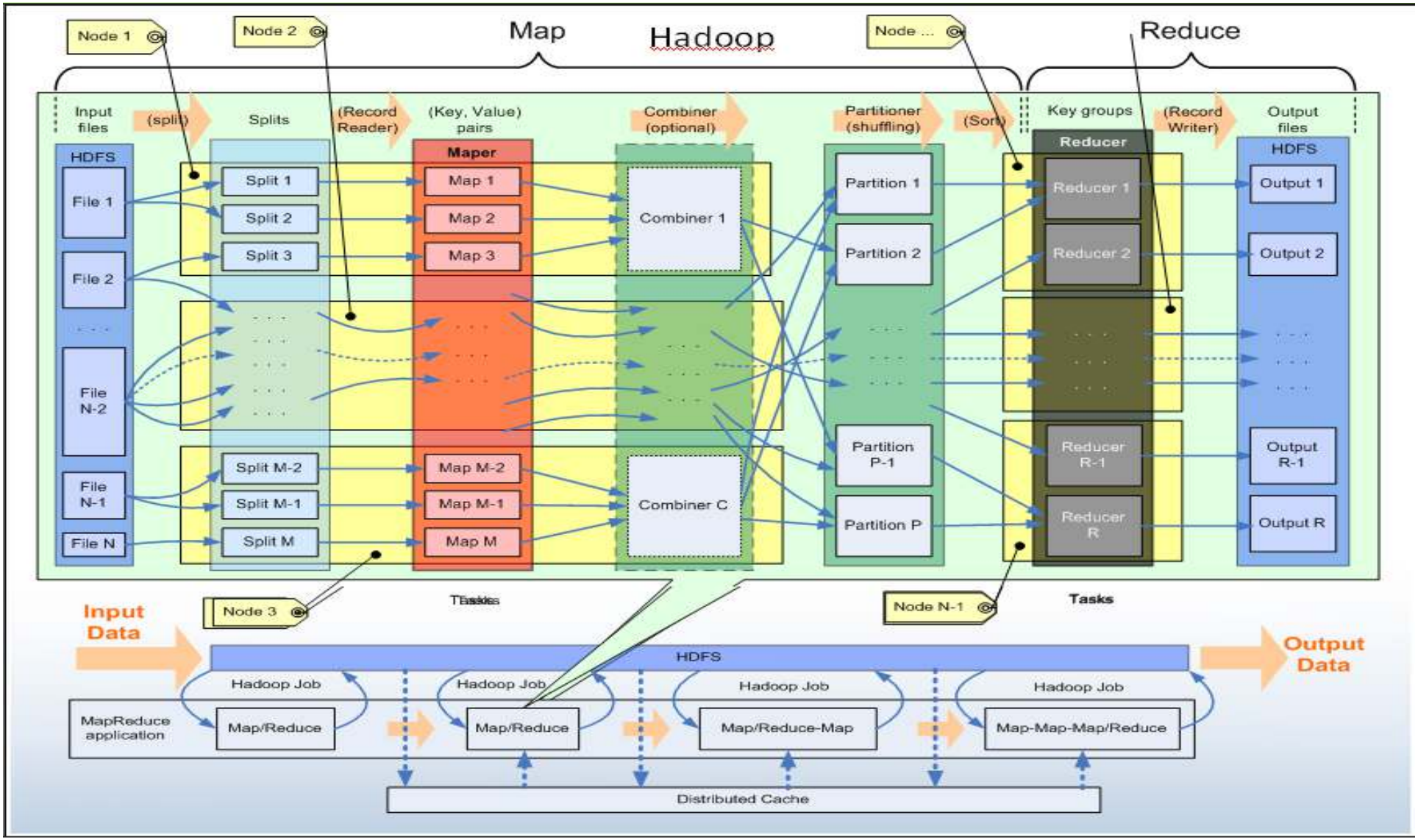
volume, velocity, variety… resources?

Talend Big Data Strategy

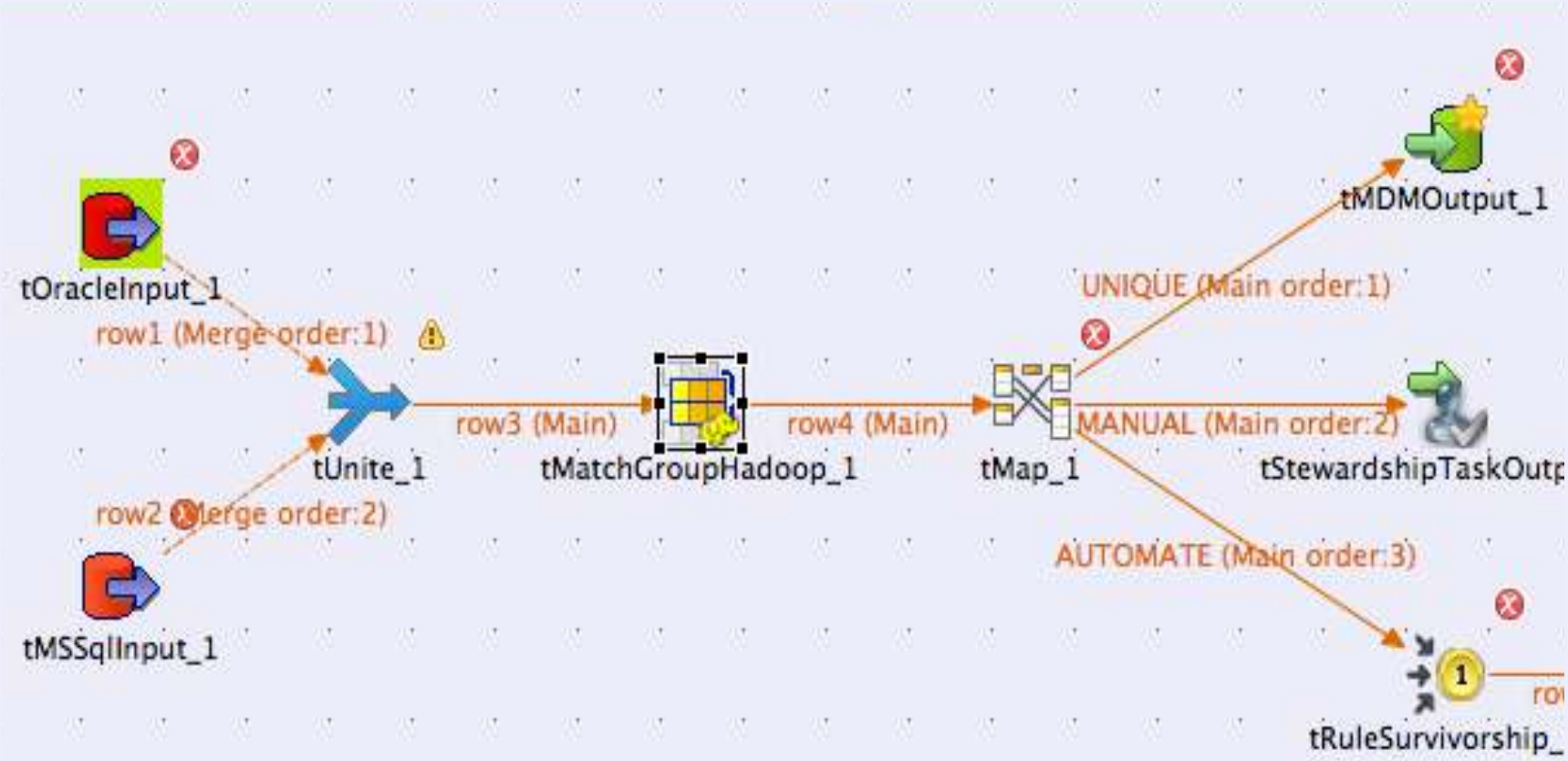
- **Big Data Integration**
  - Land data in a BD cluster without coding
  - Code generation for Hadoop HDFS, Hive, Sqoop
- **Big Data Manipulation**
  - Simplify manipulation, such as sort and filter
  - Pig components, HBase
- **Big Data Quality & Governance**
  - Identify linkages & duplicates, validate big data
  - Match component, execute basic quality features
- **Big Data Project Management**
  - Place frameworks around big data projects
  - Common Repository, scheduling, monitoring

talend
*open integration solutions

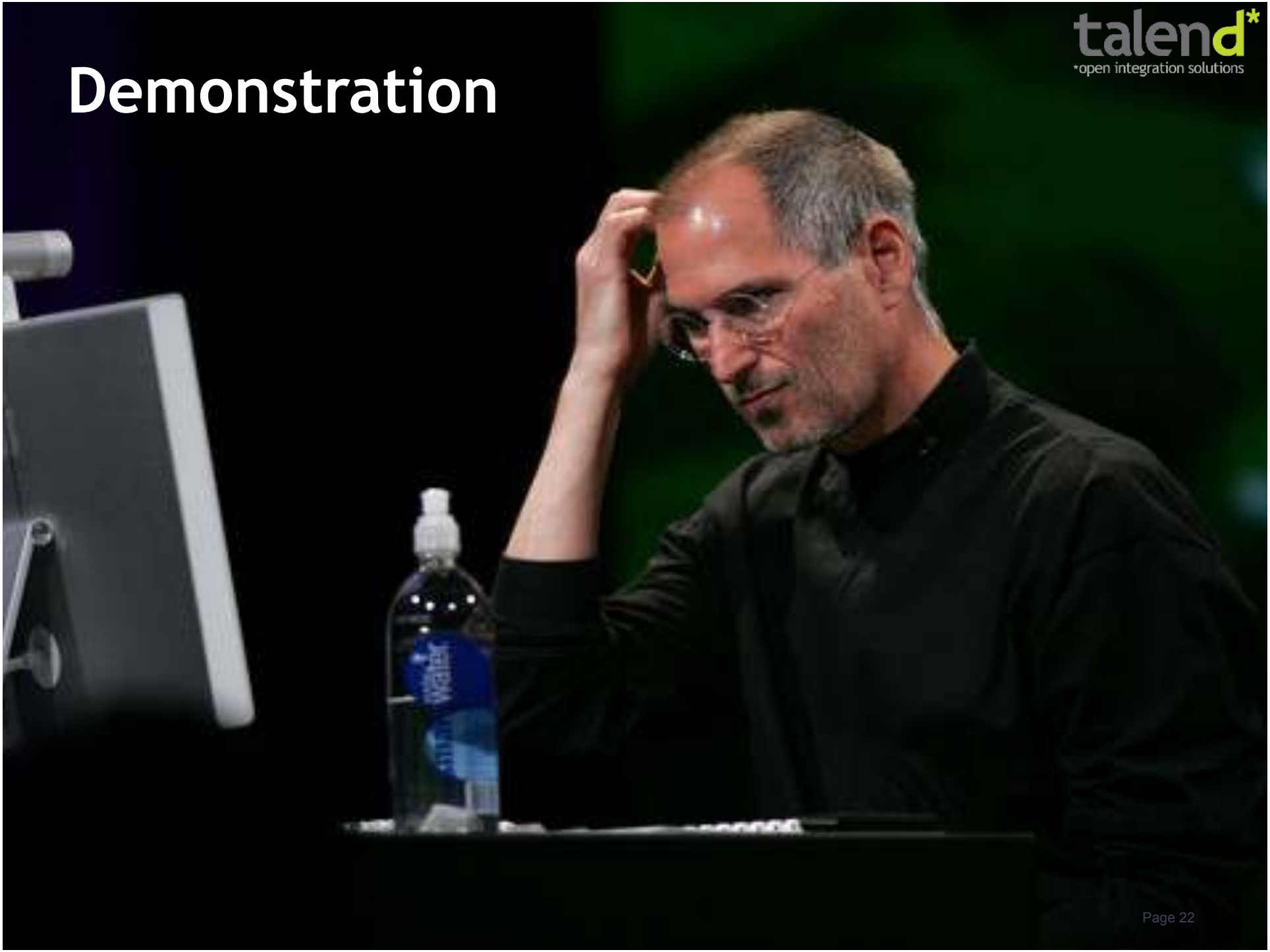
Why Talend…



Why Talend…



Demonstration



2012 Talend Roadmap

Computationally Intense Functions

- Matching (complete)
- Profiling
- Parsing
- Survivorship

… continue to monitor "winning" technologies & expand partnerships

Talend Open Studio for Big Data

…an open source ecosystem

Democratize Big Data

Talend Open Studio for Big Data

- Improves efficiency of big data job design with graphic interface
- Abstracts and generates code
- Run transforms inside Hadoop
- Native support for HDFS, Pig, Hbase, Sqoop and Hive

- Apache License
- Available at talend.com
- Embedded in HWx Data Platform

Other Resources

- **Next webinar: HDFS Federated**
  - April 18, 2012 @ 10am PST
  - Register now :http://hortonworks.com/webinars/

- **Hadoop Summit**
  - June 13-14
  - San Jose, California
  - www.Hadoopsummit.org

- **Hadoop Training and Certification**
  - Developing Solutions Using Apache Hadoop
  - Administering Apache Hadoop
  - http://hortonworks.com/training/

Thank You!

Questions?