Apache Hadoop Today & Tomorrow

Eric Baldeschwieler, CEO

Hortonworks, Inc.

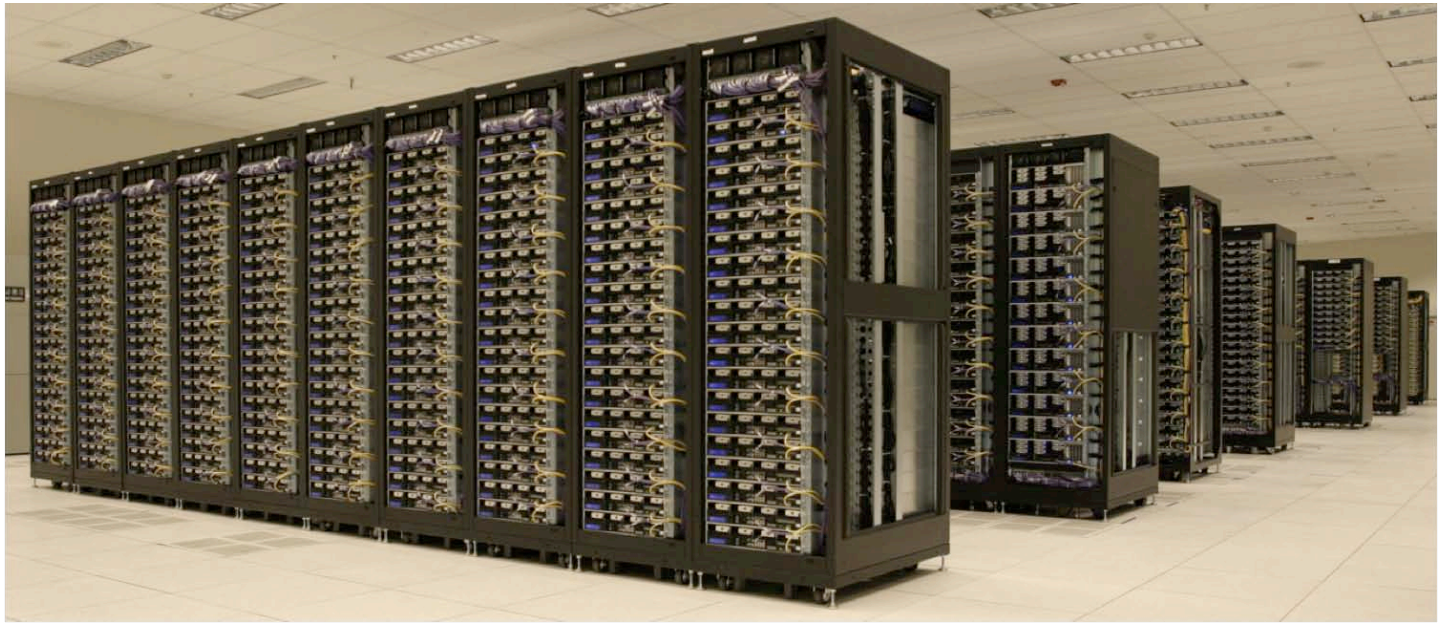
twitter: @jeric14 (@hortonworks)

www.hortonworks.com

Agenda

- Brief Overview of Apache Hadoop
- Where Apache Hadoop is Used
- Apache Hadoop Core
  - Hadoop Distributed File System (HDFS)
  - Map/Reduce
- Where Apache Hadoop Is Going
- Q&A

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
- Map-Reduce – Allows huge distributed computations

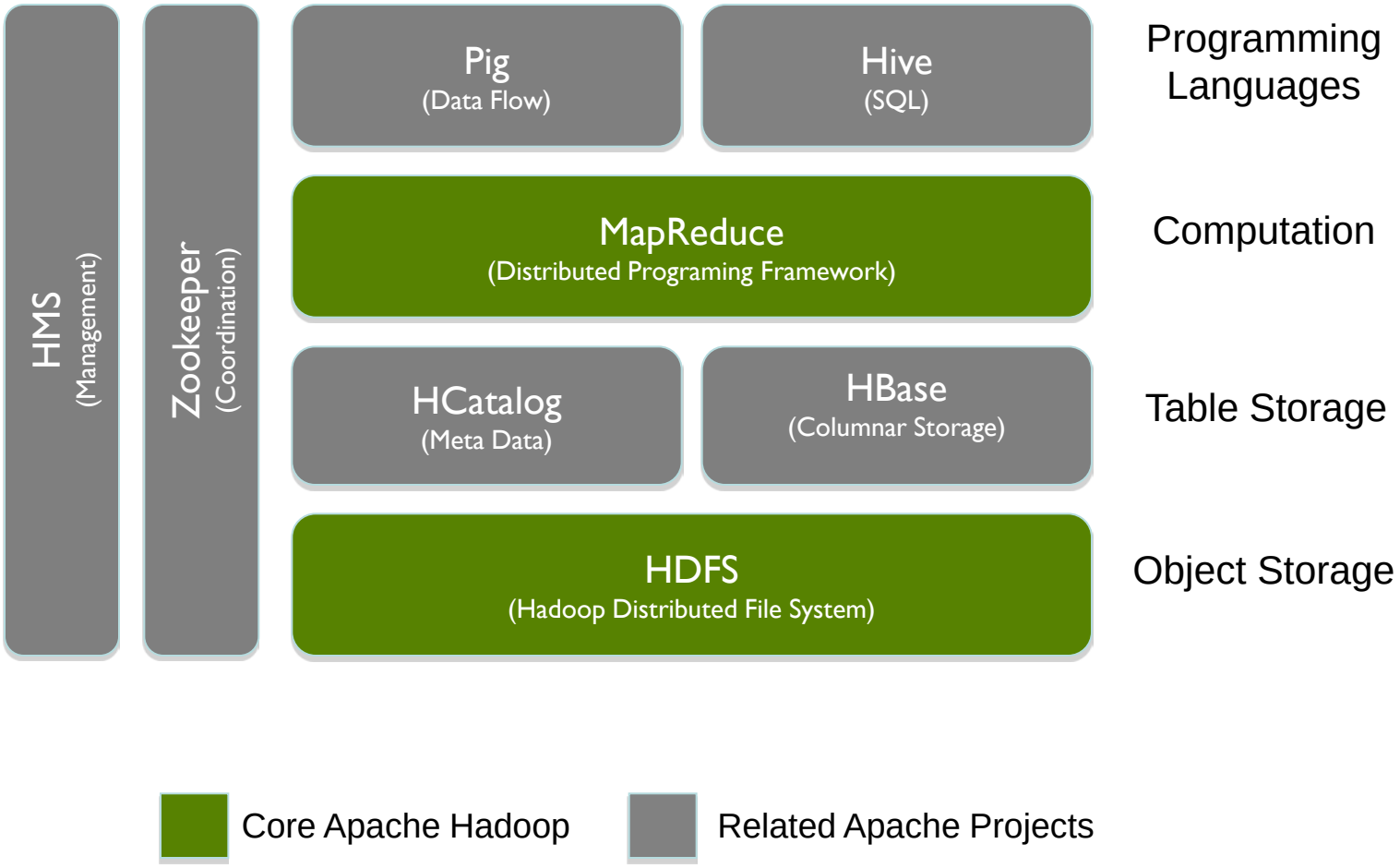
Key Attributes

- Reliable and Redundant – Doesn't slow down or loose data even as hardware fails
- Simple and Flexible APIs – Our rocket scientists use it directly!
- Very powerful – Harnesses huge clusters, supports best of breed analytics
- Batch processing centric – Hence its great simplicity and speed, not a fit for all use cases

What is it used for?

- Internet scale data
  - Web logs – Years of logs at many TB/day
  - Web Search – All the web pages on earth
  - Social data – All message traffic on facebook
- Cutting edge analytics
  - Machine learning, data mining…
- Enterprise apps
  - Network instrumentation, Mobil logs
  - Video and Audio processing
  - Text mining
- And lots more!

Apache Hadoop Projects



Where Hadoop is Used

Everywhere!

2006

2007

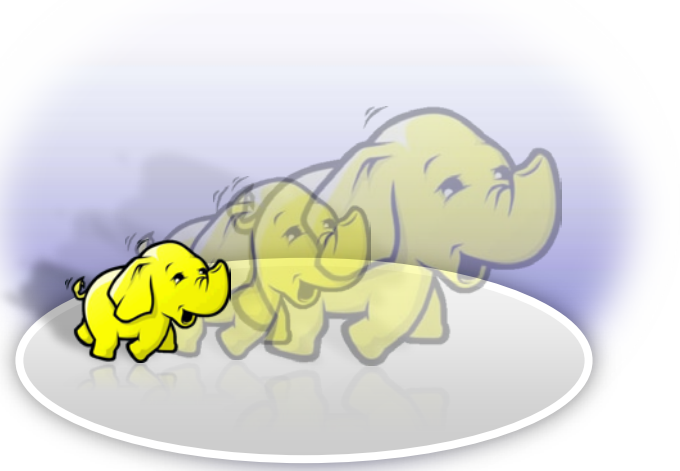
2008

2009

2010

The Datagraph Blog

HADOOP @ YAHOO!



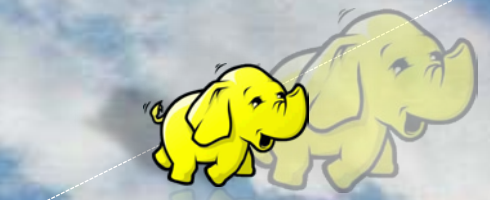
40K+ Servers

170 PB Storage

5M+ Monthly Jobs

1000+ Active users

CASE STUDY
YAHOO! HOMEPAGE



Personalized

for each visitor

Result:

twice the engagement

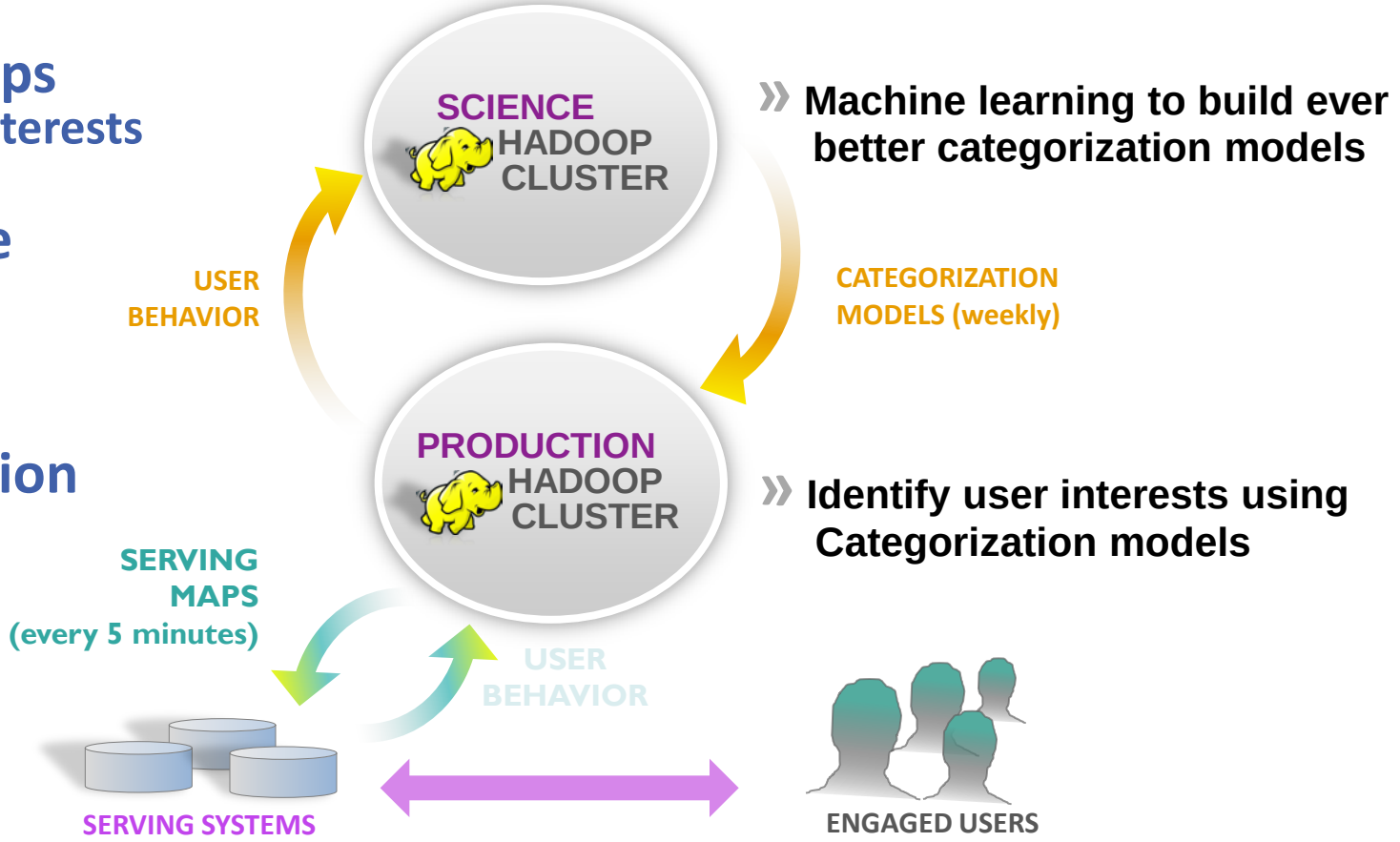
+79% clicks
vs. randomly selected

+160% clicks
vs. one size fits all

+43% clicks
vs. editor selected

CASE STUDY
YAHOO! HOMEPAGE

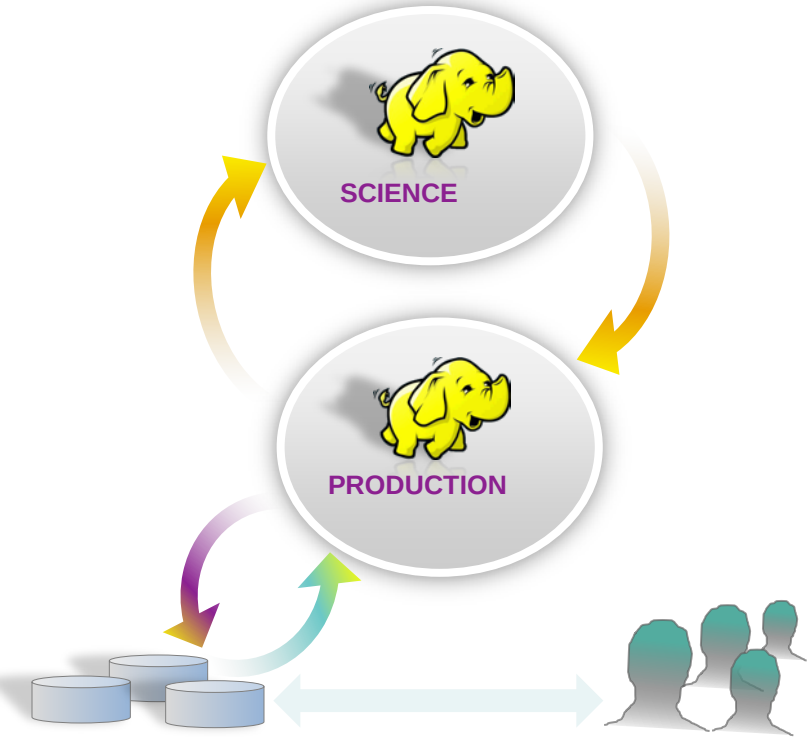
- Serving Maps
  - Users - Interests
- Five Minute Production
- Weekly Categorization models



Build customized home pages with latest data (thousands / second)

CASE STUDY
YAHOO! MAIL

Enabling quick response in the spam arms race

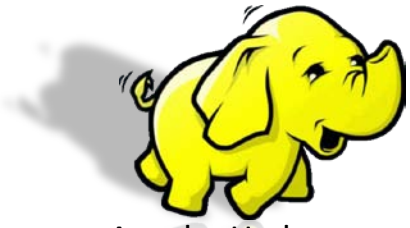


- 450M mail boxes
- 5B+ deliveries/day

- Antispam models retrained every few hours on Hadoop

> "40% less spam than Hotmail and 55% less spam than Gmail"

A Brief History



YAHOO!, early adopters — 2006 – present

Scale and productize Hadoop

Other Internet Companies — 2008 – present

Add tools / frameworks, enhance Hadoop

facebook twitter LinkedIn ...

Service Providers — 2010 – present

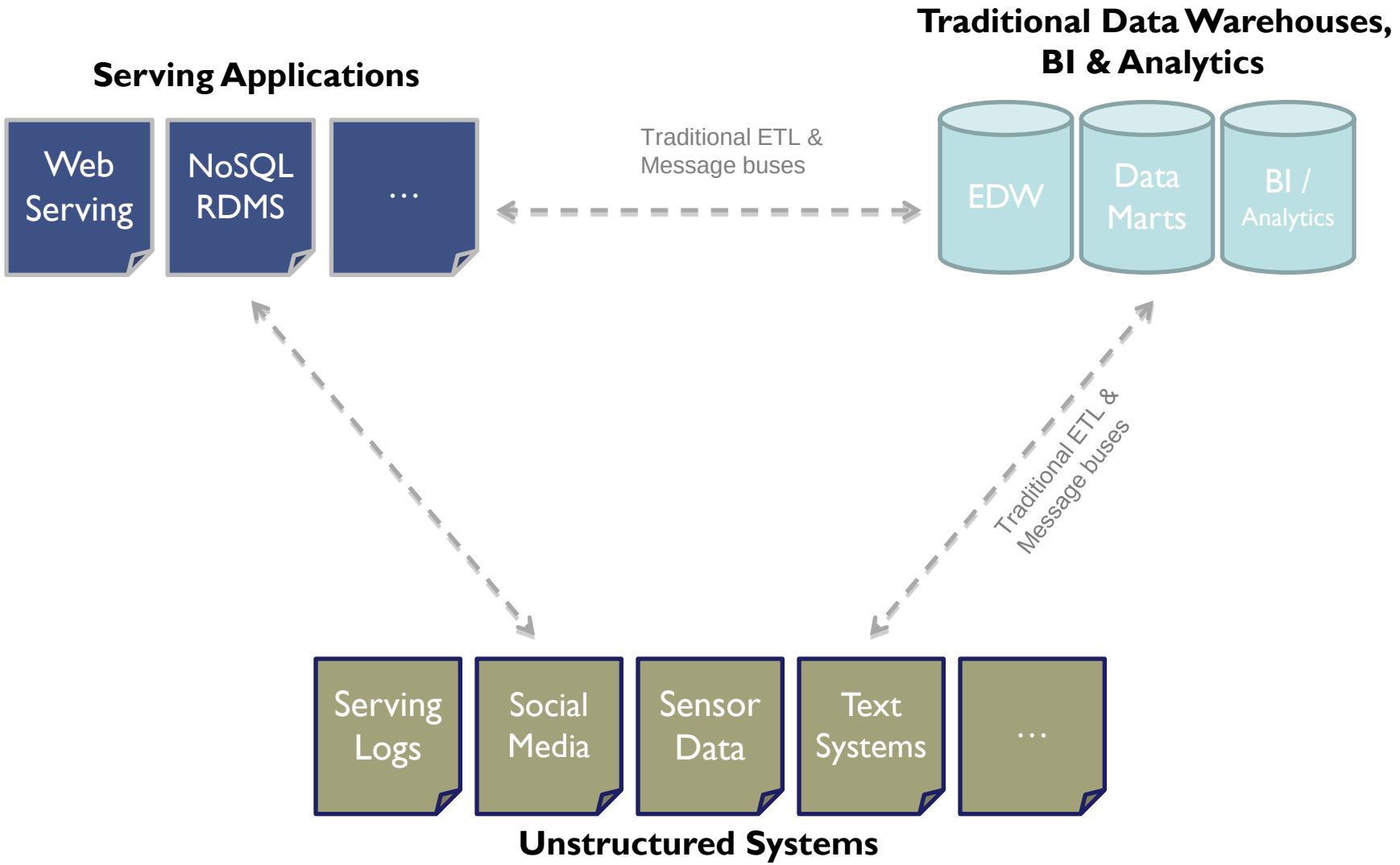
Provide training, support, hosting

hortonworks amazon web services IBM ...

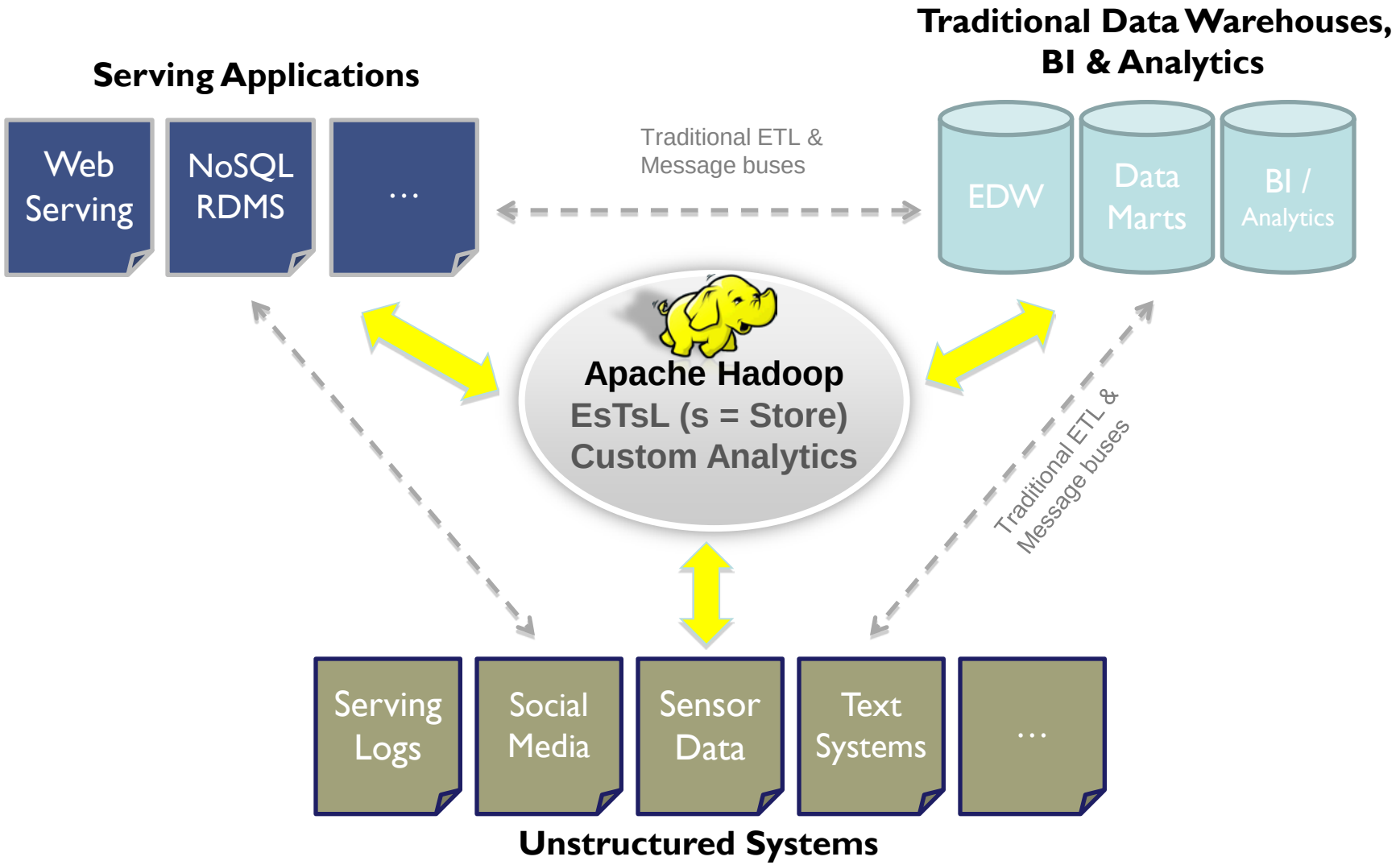
Wide Enterprise Adoption — Nascent / 2011

Funds further development, enhancements

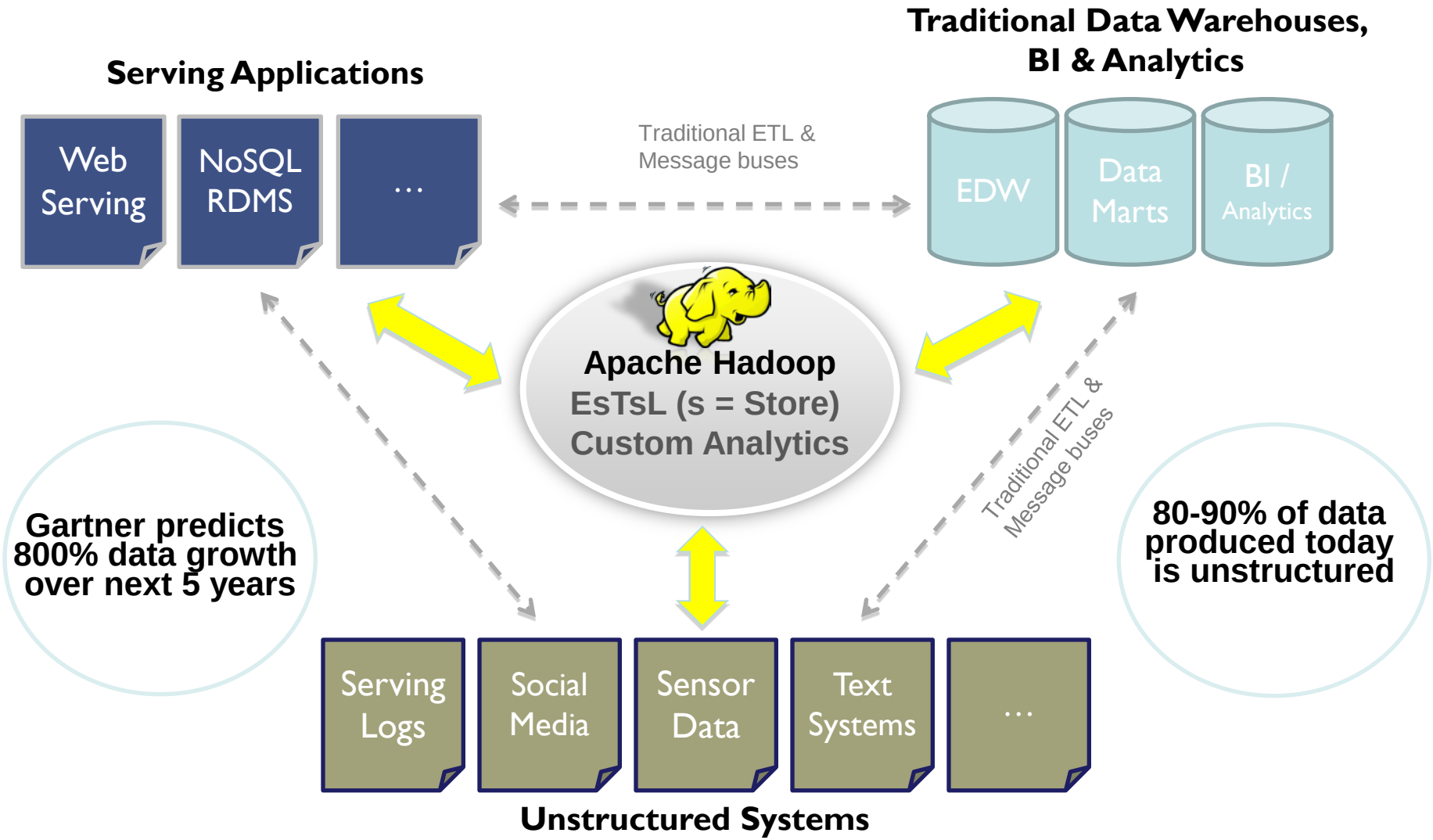
Traditional Enterprise Architecture
Data Silos + ETL



Hadoop Enterprise Architecture Connecting All of Your Big Data



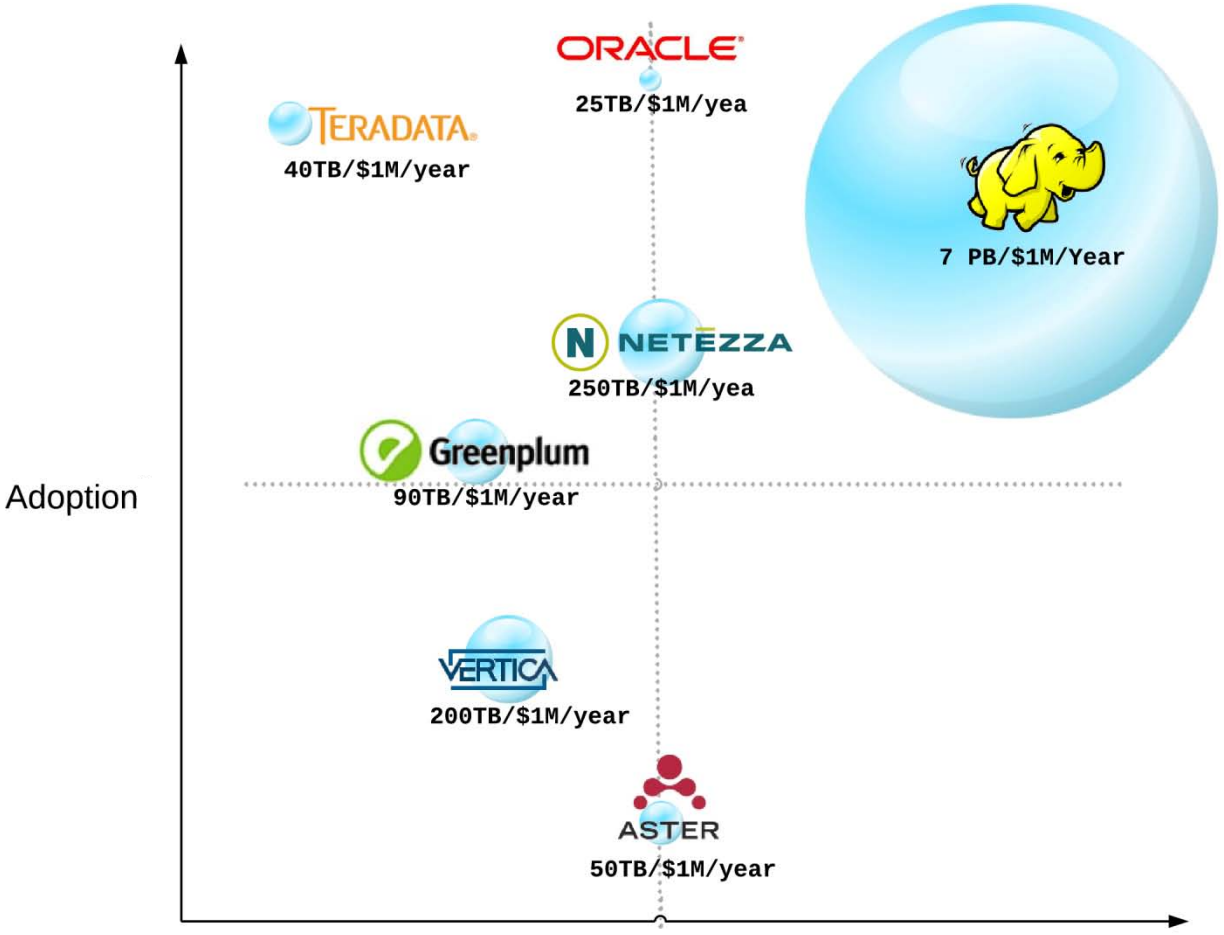
Hadoop Enterprise Architecture Connecting All of Your Big Data



What is Driving Adoption?

- **Business drivers**
  - Identified high value projects that require use of more data
  - Belief that there is great ROI in mastering big data
- **Financial drivers**
  - Growing cost of data systems as proportion of IT spend
  - Cost advantage of commodity hardware + open source
    - Enables departmental-level big data strategies
- **Technical drivers**
  - Existing solutions failing under growing requirements
    - 3Vs - Volume, velocity, variety
  - Proliferation of unstructured data

Big Data Platforms
Cost per TB, Adoption



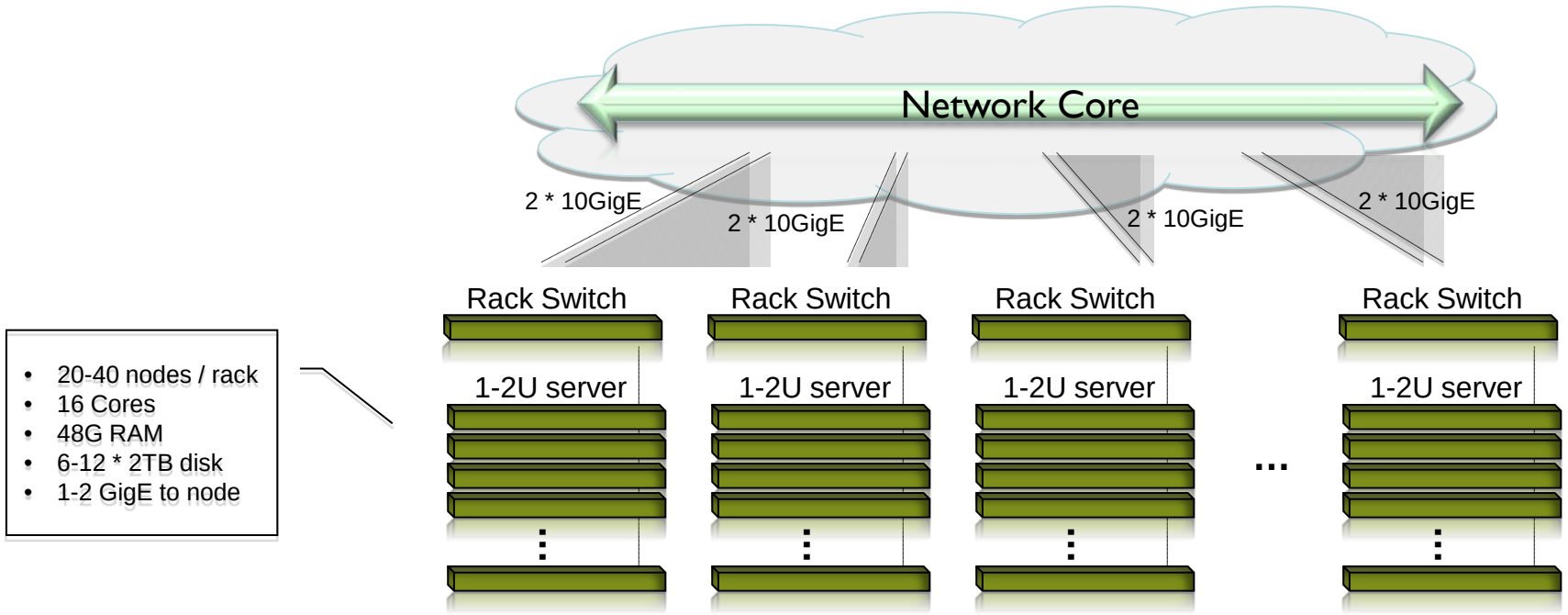
Size of bubble = cost effectiveness of solution

Source:
THiNK BIG ANALYTICS

Apache Hadoop Core

Overview

- Frameworks share *commodity* hardware
  - Storage - HDFS
  - Processing - MapReduce



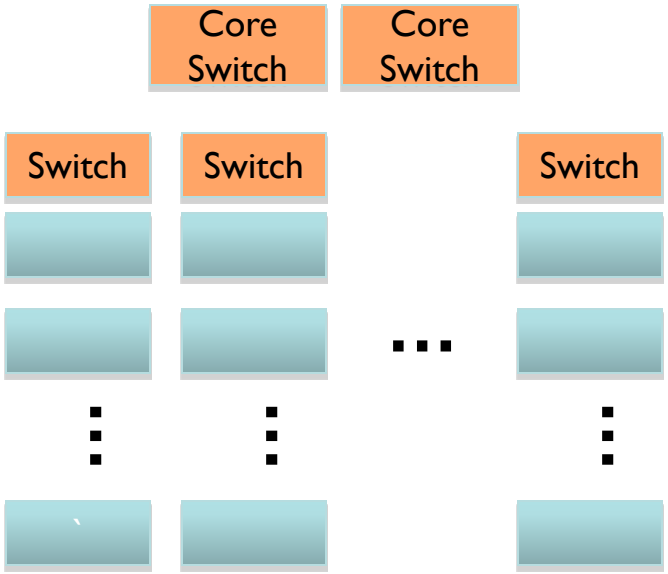
Map/Reduce

- Map/Reduce is a distributed computing programming model
- It works like a Unix pipeline:
  - cat input | grep | sort | uniq -c > output
  - Input | **Map** | Shuffle & Sort | **Reduce** | Output
- Strengths:
  - Easy to use! Developer just writes a couple of functions
  - Moves compute to data
    - Schedules work on HDFS node with data if possible
  - Scans through data, reducing seeks
  - Automatic reliability and re-execution on failure

HDFS: Scalable, Reliable, Managable

Scale IO, Storage, CPU

- Add ***commodity*** servers & ***JBODs***
- 4K nodes in cluster, 80

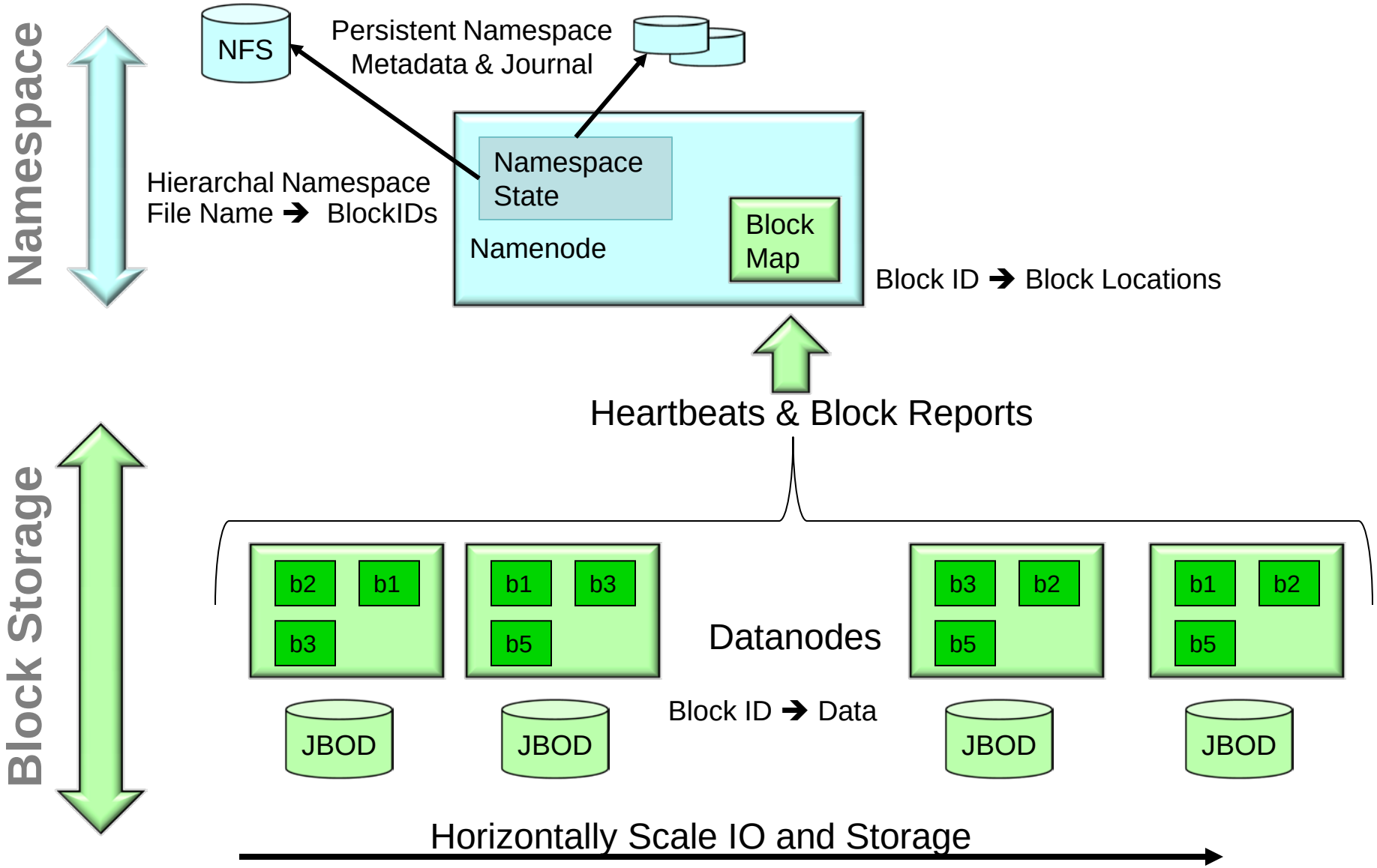


- Fault Tolerant & Easy management
  - Built in redundancy
  - Tolerate disk and node failures
  - Automatically manage addition/removal of nodes
  - One operator per 8K nodes!!
- Storage server used for computation
  - Move computation to data
- Not a SAN
  - But high-bandwidth network access to data via Ethernet
- Immutable file system
  - Read, Write, sync/flush
    - No random writes

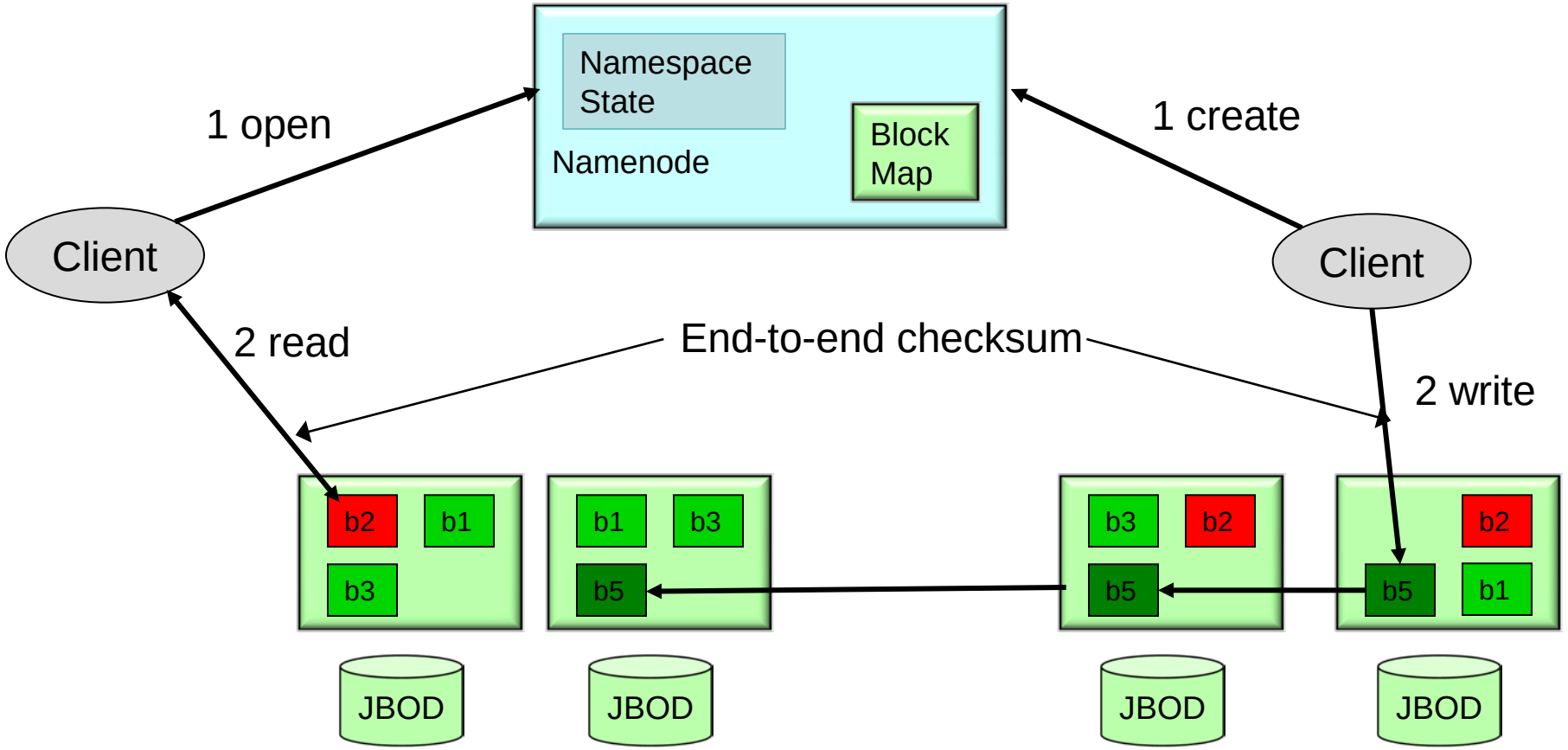
HDFS Use Cases

- Petabytes of unstructured data for parallel, distributed analytics processing using commodity hardware
- Solve problems that cannot be solved using traditional systems at a cheaper price
  - Large storage capacity (>100PB raw)
  - Large IO/Computation bandwidth (>4K servers)
    - > 4 Terabit bandwidth to disk! (conservatively)
  - Scale by adding commodity hardware
  - Cost per GB ~= $1.5, includes MapReduce cluster

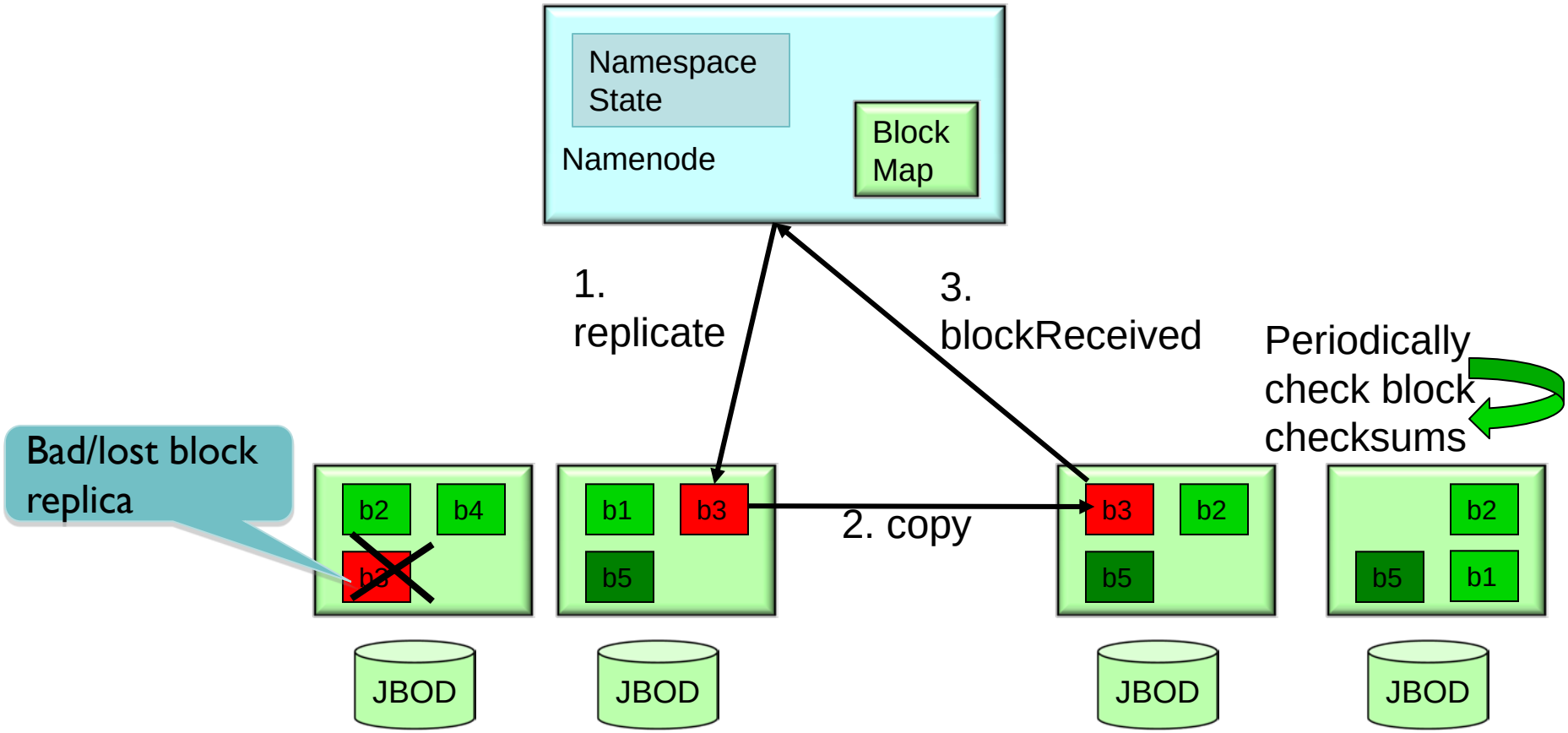
HDFS Architecture



Client Read & Write *Directly* from *Closest* Server



Actively maintain data reliability



HBase

- Hadoop ecosystem Database, based on Google BigTable
- Goal: Hosting of very large tables (billions of rows X millions of columns) on commodity hardware.
  - Multidimensional sorted Map
    - Table => Row => Column => Version => Value
  - Distributed column-oriented store
  - Scale – Sharding etc. done automatically
    - No SQL, CRUD etc.

What's Next

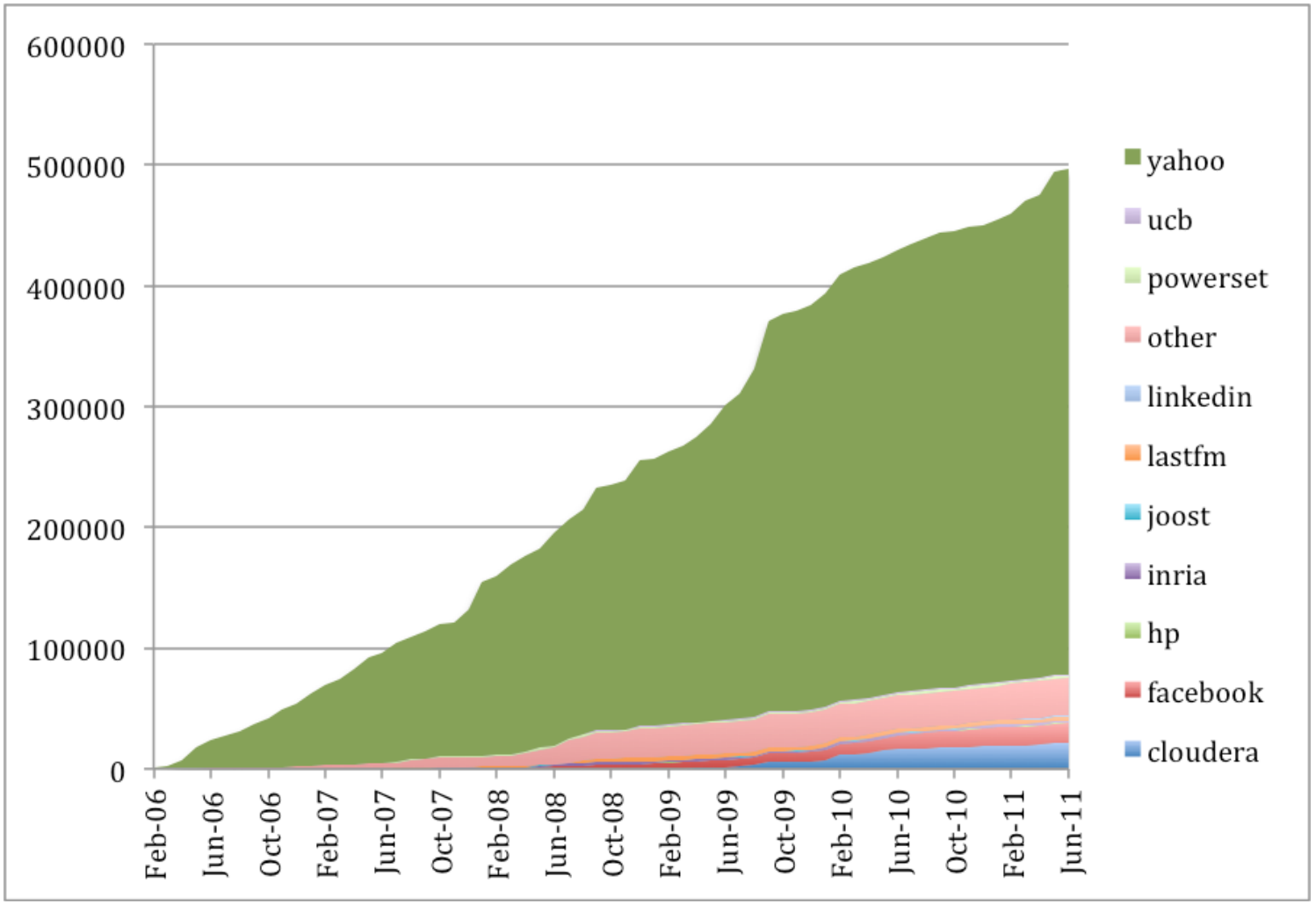
About Hortonworks – Basics

- **Founded** – July 1st, 2011
  - 22 Architects & committers from Yahoo!
- **Mission** – Architect the future of Big Data
  - Revolutionize and commoditize the storage and processing of Big Data via open source
- **Vision** – Half of the worlds data will be stored in Hadoop within five years

Game Plan

- Support the growth of a huge Apache Hadoop ecosystem
  - Invest in ease of use, management, and other enterprise features
  - Define APIs for ISVs, OEMs and others to integrate with Apache Hadoop
  - Continue to invest in advancing the Hadoop core, remain the experts
  - Contribute all of our work to Apache
- Profit by providing training & support to the Hadoop community

Lines of Code Contributed to Apache Hadoop



Apache Hadoop Roadmap

Phase 1 – Making Apache Hadoop Accessible **2011**

- Release the most stable version of Hadoop ever (Hadoop 0.20.205)
  - Frequent sustaining releases
- Release directly usable code via Apache (RPMs, .debs…)
- Improve project integration (HBase support)

Phase 2 – Next-Generation Apache Hadoop **2012**

(Alphas in Q4 2011)

- Address key product gaps (HA, Management…)
- Enable partner innovation via open APIs
- Enable community innovation via modular architecture

Next-Generation Hadoop

- **Core**
  - HDFS Federation – Scale out and innovation via new APIs
    - Will run on 6000 node clusters with 24TB disk / node = 144PB in next release
  - Next Gen MapReduce – Support for MPI and many other programing models
  - HA (no SPOF) and Wire compatibility
- **Data - HCatalog 0.3**
  - Pig, Hive, MapReduce and Streaming as clients
  - HDFS and HBase as storage systems
  - Performance and storage improvements
- **Management & Ease of use**
  - Ambari – A Apache Hadoop Management & Monitoring System
  - Stack installation and centralized config management
  - REST and GUI for user & administrator tasks

Thank You!

Questions

Twitter: @jeric14 (@hortonworks)
www.hortonworks.com