Hortonworks

Architecting the Future of Big Data

Eric Baldeschwieler – CEO

twitter: @jeric14 (@hortonworks)

Formerly VP Hadoop Engineering @Yahoo!

8 Years at Yahoo!

June 29, 2011

About Hortonworks

- **Mission:** Revolutionize and commoditize the storage and processing of big data via open source

- **Vision:** Half of the world's data will be stored in Apache Hadoop within five years

- **Strategy:** Grow the Apache Hadoop Ecosystem by making Apache Hadoop easier to consume, profit by providing training, support and certification
  - ✓ An independent company
  - ✓ Focused on making Apache Hadoop great
  - ✓ Hold nothing back, Apache Hadoop will be complete

hortonworks

Credentials

- **Technical:** key architects and committers from Yahoo! Hadoop engineering team
  - Highest concentration of Apache Hadoop committers
  - Contributed >70% of the code in Hadoop, Pig and ZooKeeper
  - Delivered every major/stable Apache Hadoop release since 0.1
  - History of driving innovation across entire Apache Hadoop stack
  - Experience managing world's largest deployment
- **Business operations:** team of highly successful open source veterans
  - Led by Rob Bearden, former COO of SpringSource & JBoss
- **Investors:** backed by Benchmark Capital and Yahoo!
  - Benchmark was key investor in Red Hat, MySQL, SpringSource, Twitter & eBay

hortonworks

Hortonworks and Yahoo!

- Yahoo! is a **development partner**
  - Leverage large Yahoo! development, testing & operations team
    - ✓ More than 1,000 active & sophisticated users of Apache Hadoop
    - ✓ Access to the Yahoo! grid for testing large workloads
    - ✓ Only organization that has delivered a stable release of Apache Hadoop
  - Yahoo will continue to contribute Apache Hadoop code too!
- Yahoo! is a **customer**
  - Hortonworks provides level 3 support and training to Yahoo!
  - Yahoo deploys Apache Hadoop releases across its 42,000 grid
- Yahoo! is an **investor**

Current State of Adoption

Enterprise Adoption

- Early adopters
- Technology is hard to install, manage & use
- Technology lacks enterprise robustness
- Requires significant investment in technical staff or consulting
- Hard to find & hire experienced developer & operations talent

Technology & Knowledge Gaps Prevent Apache Hadoop from Reaching Full Potential

Customers are asking their vendors for help with Hadoop!

"We're seeing Hadoop in all of our fortune 2000 data accounts"

Vendor Ecosystem Adoption

- Early in vendor adoption lifecycle
- Hadoop is hard to integrate and extend
- Hard to find & hire experienced developer & operations talent

hortonworks

Hortonworks Role & Opportunity

Bridge the Gap!
Grow Market

hortonworks

Sell training and support via Partners

Fundamental shift in enterprise data architecture strategy

- Apache Hadoop becomes standard for managing new types & scale of data
- New applications & solutions will be created to leverage data in Apache Hadoop
- Creates massive big data technology and services opportunity for ecosystem

hortonworks

Hortonworks Objectives

- **Make Apache Hadoop projects easier to install, manage & use**
  - Regular sustaining releases
  - Compiled code for each project (e.g. RPMs)
  - Testing at scale
- **Make Apache Hadoop more robust**
  - Performance gains
  - High availability
  - Administration & monitoring
- **Make Apache Hadoop easier to integrate & extend**
  - Open APIs for extension & experimentation

All done within Apache Hadoop community

- Develop collaboratively with community
- Complete transparency
- All code contributed back to Apache

Anyone should be able to easily deploy the Hadoop projects directly from Apache

Technology Roadmap

Phase 1 – Making Apache Hadoop Accessible • Release the most stable version of Hadoop ever • Release directly usable code via Apache (RPMs, .debs…) • Frequent sustaining releases off of the stable branches	**2011**
Phase 2 – 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 Oct 2011)

Phase 2 - Next Generation Apache Hadoop

- **Core**
  - HDFS Federation
  - Next Gen MapReduce
  - New Write Pipeline (HBase support)
  - 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**
  - All components fully tested and deployable as a stack
  - Stack installation and centralized config management
  - REST and GUI for user tasks

hortonworks

Hortonworks @ Hadoop Summit

- 11am: "**Crossing the Chasm: Hadoop for the Enterprise**"
  - Tech talk by Sanjay Radia
- 1:45pm**:** "**Next Generation Apache Hadoop MapReduce**"
  - Community track by Arun Murthy
- 2:15pm: "**Introducing HCatalog (Hadoop Table Manager)**"
  - Community track by Alan Gates
- 4:00pm: "**Large Scale Math with Hadoop MapReduce**"
  - Applications and Research Track by Tsz-Wo Sze
- 4:30pm: "**HDFS Federation and Other Features**"
  - Community track by Suresh Srinivas

Thank You.