About Eric Baldeschwieler

• Co-Founder and CTO of Hortonworks
• Prior to Hortonworks:
  – VP Hadoop Software Engineering for Yahoo!
  – Technology leader for Inktomi’s web service engine (acquired by Yahoo! in 2003)
  – Developed software for video games, video post production systems and 3D modeling systems
• Master’s degree in Computer Science
  University of California, Berkeley
• Bachelor’s degree in Mathematics and Computer Science
  Carnegie Mellon University
• Follow me on Twitter: @jeric14
Agenda

- Hortonworks Data Platform
  - Community roadmap process
  - Hortonworks development process & support model
  - 2012 roadmap highlights

- Observed Trends and Anticipated Investment Areas
Yahoo! embraced Apache Hadoop, an open source platform, to **crunch epic amounts of data** using an **army of dirt-cheap servers**

Yahoo! spun off 22+ engineers into Hortonworks, a company focused on **enabling Apache Hadoop to be next-generation data platform**
Balancing Innovation & Stability

- **Apache: Be aggressive - ship early and often**
  - Projects need to keep innovating and visibly improve
  - Aim for big improvements
  - Make early buggy releases

- **Hortonworks: Be predictable - ship when stable**
  - We need to ship stable, working releases
  - Make packaged binary releases available
  - We need to do regular sustaining engineering releases
  - HDP quarterly release trains sweep in stable Apache projects
    - Enables HDP to stay reasonably current and predictable while minimizing risk of thrashing that coordinating large # of Apache projects can cause
The Hortonworks Development Process

- **The Hortonworks Data Platform is 100% open source**
  - We plan our engineering in the open
    - We take feedback and use it to refine our plans
  - We develop enhancements to Hadoop in collaboration with the Apache community, via the Apache processes & infrastructure
    - Others contribute their own improvements and refine our work
    - We iterate and we decide as a community when a release is ready
This presentation is part of that process

- Apache Hadoop and related projects are owned by the Apache foundation and worked on by a host of volunteers
  - This presentation outlines what Hortonworks engineers currently plan to contribute to Apache projects
  - We share our plans regularly with
    - Apache contributors & the wider Apache Hadoop user community
    - our business partners & customers

- We listen and we refine the plan
  - Others share their plans & help us refine our thinking
  - Our partners express their needs & requirements

- The plan evolves
Challenge:
• Integrate, manage, and support changes across a wide range of open source projects that power the Hadoop platform; each with their own release schedules, versions, & dependencies.
  • Time intensive, Complex, Expensive

Solution: Hortonworks Data Platform
• Integrated, certified platform distributions
• 100% Open Source
• Extensive Q/A process
• Industry-leading Support with clear service levels for updates and patches
• Continuity via multi-year Support and Maintenance Policy

Hadoop Core Pig Zookeeper Hive HCatalog HBase

= New Version
Support & Distribution Model

Hortonworks Data Platform
- Fully supported, integrated, tested, maintained
- 100% Apache license, or compatible: BSD, MIT/X11, NCSA, W3C Software license, X.Net

Universe: Open Source Ecosystem
- Validated & interoperable with HDP
- Technical guidance support; work with OSS projects
- 100% OSI-compliant licenses
- Optionally installed

Multiverse: Commercial Ecosystem
- Validated & interoperable with HDP
- Technical guidance support; work with TSANet
- 3rd-party vendor licenses and support options
- Optionally installed

Model and terminology conceptually similar to Ubuntu’s model:
http://www.ubuntu.com/project/about-ubuntu/licensing
Hortonworks Data Platform (HDP)
Key Components of “Standard Hadoop” Open Source Stack

Hortonworks Data Platform

Zookeeper
(Cluster Coordination)

HBase
(Columnar NoSQL Store)

MapReduce
(Distributed Programing Framework)

HCatalog
(Table & Schema Management)

HDFS
(Hadoop Distributed File System)

Pig
(Data Flow)

Hive
(SQL)

Universe

Ambari &
Ganglia and Nagios…

Oozie
Workflow scheduling

Sqoop &
Other Ingest, ETL tools

Mahout &
Other libraries
Hadoop Now, Next, and Beyond

Apache community, including Hortonworks investing to improve Hadoop:
- Make Hadoop an Open, Extensible, and Enterprise Viable Platform
- Enable More Applications to Run on Apache Hadoop

“Hadoop.Beyond”
Future investments

“Hadoop.Next”
(Hadoop 0.23)
HDP2
- HA, Next-gen MapReduce
- Extension & Integration APIs
- HCatalog data APIs

“Hadoop.Now”
(Hadoop 1.0)
HDP1
- Most stable Hadoop ever
- HBase, security, WebHDFS

- Enable More Applications to Run on Apache Hadoop
Hortonworks Data Platform Timeline

Hortonworks Data Platform 1

Q1: 1.0 preview
Q2: 1.0
Q3: 1.1 preview
Q4: 1.1

Hortonworks Data Platform 2

Q1: 2.0 preview
Q2: 2.0
Q3: 2.1

36 Month support policy, from GA date
Hortonworks Data Platform 1

Consumable “Hadoop.Now” Platform

• Based on Hadoop 1.0 (a.k.a. 0.20.205)
  - The most stable release of Hadoop, ever!

• Differentiators:
  - Code straight from Apache
    - Apache release process restarted after hiatus since 2010!
  - First Apache line supporting Security, HBase, WebHDFS – and many stability fixes
  - Common table and schema management via HCatalog (M/R, Pig and Hive)
  - Capacity scheduler (very stable, high RAM job support, multi-tenant protections)

• Components:
  - “Standard Hadoop” stack: HDFS+M/R, Hive, Pig, HBase, HCatalog, ZooKeeper
  - “Universe” items: Sqoop, Oozie, Mahout, …
  - Packaging (.tar, RPM, DEB, single-box VM, single-box AMI, …)
  - Monitoring via Nagios and Ganglia
Continuing investment in 1.0 line

- Quarterly updates will include bug fixes and enhancements to all components, from Apache

- **Hortonworks plans to invest substantially in:**
  - Ambari and other management components
    - The Hadoop community has traditionally not provided good open source tooling in this area, we are committed to closing this gap
  - HCatalog
    - When fully realized, we think HCatalog will have a huge impact
    - Vastly simplifying the management of data on Hadoop
Management in HDP1.0 and beyond

- **Now:** Focus on Monitoring capability (most often used)
  - Package Nagios & Ganglia
    - The most common Open Source Hadoop Monitoring Tools
  - Web dashboard for unified Hadoop-specific view
    - System Health: NetSNMP
    - Hadoop Metrics: Ganglia plug-in, JMX, SNMP MIB (to add)
    - Alerting: Nagios
    - Collect & Display: Nagios, Ganglia, RRD

- **Next:** Deployment and Management
  - Provisioning & Configuration with Puppet (Chef + others to follow)
  - VM and cloud packaging
  - LDAP, Active directory integration
  - User metering and management API/GUI
HCatalog 0.3 and beyond
Common Table, Schema, Metadata Management

• In HDP 1.0
  - Enable Hive, Pig, and MR to use the same tables / metadata
  - Generalizes Hive’s Table/Metadata System
    - Pig/Hive/MR use same HCatalog file storage methods
    - Pig/Hive/MR use common code for their IO
  - Manages Data Format and Schema Changes
    - Allows columns to be appended to tables in new partitions
    - Allows storage format changes
  - CLI to create table w DDL, change default format, add columns, change data location, compact data, register data as a table
    - Templeton APIs layer on top

• Longer-term: A simple uniform API that supports
  - a rich set of data management tools
  - transparent migration of data between systems & formats
  - document models and other non-relational data gracefully
  - Easy lookup, addition & modification of objects / records
Related Hortonworks Webinars

• **HCatalog, Table Management for Hadoop**
  - Wednesday, February 22
  - 10:00am Pacific/1:00pm Eastern
  - http://hortonworks.com/webinars/

• **Other Topics Coming Soon**
  - Importing Data Into Hadoop
  - Monitoring and Managing Hadoop Clusters
  - HBase and Hive
  - http://hortonworks.com/webinars/
Based on Hadoop 0.23.*
- Next generation of Hadoop
- First release of a set of features under development since 2010

Highlights:
- Next Generation MapReduce architecture
  - Refactor to provide enhanced scalability and performance
  - Decouple MapReduce from resource management architecture
    - Enables MapReduce to evolve quickly
    - Enables new application types (streaming, graph, MPI, bulk sync, etc)
- HDFS Federation
  - Improved scalability and isolation
  - Extension API that will allow new storage services to share HDFS storage
- HDFS NameNode High Availability
  - Automatic failover
  - Multiple-options for failover (shared disk, Linux HA, Zookeeper)
- Include latest stable “standard Hadoop” components
Related Hortonworks Webinars

• What’s In Store For Hadoop.Next

• Hadoop HDFS High Availability: HA NameNode

• Other Topics Coming Soon
  - Extending Hadoop beyond Map-Reduce – Wednesday March 7th
  - HDFS Federation
  - http://hortonworks.com/webinars/
Observed Trends,  
Anticipated Investments
Trend: Agile Data, Hadoop as data hub

• The old way
  - Operational systems keep only current records, short history
  - Analytics systems keep only conformed / cleaned / digested data
  - Unstructured data locked away in operational silos
  - Archives offline
    - Inflexible, new questions require system redesigns

• The new trend
  - Keep all data in Hadoop for a long time (raw inputs & processed)
  - Able to produce a new analytics view on-demand
  - Keep a new copy of data that was previously only in silos
  - Can immediately do new reports, experiments in Hadoop
  - New products / services can be added very quickly
  - Agile outcome justifies new infrastructure
Traditional Enterprise Data Architecture

Data Silos

Serving Applications

- Web Serving
- NoSQL RDMS
- ...

Unstructured Systems

- Serving Logs
- Social Media
- Sensor Data
- Text Systems
- ...

Traditional Data Warehouses, BI & Analytics

- EDW
- Data Marts
- BI / Analytics

Traditional ETL & Message buses
Agile Data Architecture w/Hadoop

Connecting All of Your Big Data

Serving Applications

- Web Serving
- NoSQL RDMS
- ... (other serving applications)

Traditional Data Warehouses, BI & Analytics

- EDW
- Data Marts
- BI / Analytics

Unstructured Systems

- Serving Logs
- Social Media
- Sensor Data
- Text Systems
- ... (other unstructured systems)

ETL, Archive all data, Custom Analytics

Traditional ETL & Message buses
“Hadoop as data hub” implies…

- **ETL integration / data ingest**
  - Hadoop should work well with industry standard tools
    - [Importing data webinar]

- **Offsite backups / DR**
  - HDFS Snapshots, cloud backup, other tools

- **Object / event level storage APIs**

- **Open Data APIs**

- **Non-relational model data**
  - HCatalog (for all 3 above) [HCatalog webinar]

- **Efficient / low cost storage**
  - Compression, Raid / reed solomon

- **No storage limits**
  - No file limits, scale beyond 10,000 computers / cluster…
Trend: Data-centric Applications

• **Limited runtime logic driven by huge lookup tables**
  – Data flows from application into Hadoop for processing
  – Flows from Hadoop back into application for serving

• **Complex computations in Hadoop**
  – Machine learning, other expensive computation offline
  – Personalization, classification, fraud, value analysis…

• **Application development requires data science**
  – Huge amounts of actually observed data key to modern services
  – Hadoop used as the science platform
• Serving Maps
  • Users - Interests

• Five Minute Production

• Weekly Categorization models

» Machine learning to build ever better categorization models

» Identify user interests using Categorization models

Build customized home pages with latest data (thousands / second)
“Data-centric Applications” imply...

• **New programming frameworks** [YARN webinar]
  - In RAM models (MPI, Spark, Giraph…)
  - Services in Hadoop (HBase, Stream processing, Serving…)

• **Huge Lookup tables**
  - HDFS enhancements for HBase
  - HBase improvements, integrations with other distributed stores

• **Predictable latency**

• **Continuous availability**
  - Core investments [HA webinar]

• **Data Driven applications in Hadoop**
  - Investments in Oozie, integration with HCatalog data model …
Trend: Hadoop goes to the Enterprise

- Hadoop used in production processes
- Hadoop used in a multi-tenant environments
- Hadoop used in audited environments
- Hadoop integrated into many more environments
- Forwards and backwards compatibility a must
“Hadoop goes to the Enterprise” implies…

- **Authentication, Authorization & Security**
  - Encryption, Active directory integration
  - Simplified security deployment

- **Open administration**

- **Open provisioning**

- **Open monitoring**
  - [Monitoring and managing webinar]

- **Cloud support**
  - Amazon and Azure support, Whirr / JCloud
  - VM tuning and deployment support

- **Engineered for compatibility**
  - New protocol buffer based protocols
Hortonworks Focus and Value

• **Hortonworks is focused on:**
  - Technology Leadership within Apache Hadoop Community
  - Customer and Partner Enablement around Hadoop Platform

• **What value do we provide?**
  - 100% Open Source Hortonworks Data Platform
  - Expert Role-based Training
  - Full Lifecycle Support and Services

• **Next Webinar:** HCatalog, Table Management for Hadoop
  - Wednesday, February 22
  - 10:00am Pacific/1:00pm Eastern
  - [http://hortonworks.com/webinars/](http://hortonworks.com/webinars/)

Register Now
Thank You!  Questions?

Eric Baldeschwieler
@jeric14 @hortonworks