Enterprise Hadoop

Jeff Markham
Technical Director, APAC
jmarkham@hortonworks.com
Upcoming Announcements

**Hortonworks Data Platform 2.1**
A continued focus on innovation within the core of Enterprise Hadoop to enable an ecosystem to flourish and cement Hadoop's role in the data architectures of tomorrow

- **Interactive SQL Query**: Final phase of Stinger Delivered.
- **Comprehensive Features**: Governance, Security, Operations
- **Processing Versatility**: Storm, Search

**April 2**

**Concurrent Partnership**
Cascading is the proven application development platform for building data applications on Hadoop

Integrate and Deliver the Cascading SDK into HDP 2.1

- Collection of tools, documentation, libraries, tutorials and example projects
- Simplifies SQL integration and enables Scala development for Hadoop

**April 21**

**Hadoop Summit Europe 2014**
SOLD OUT, double exhibitors, double content, year over year.

**April 2**

**LucidWorks partnership**
A resell agreement has been inked with Lucidworks to provide tier 2 and tier 3 support for HDP Search

**April 3**
Hadoop within an emerging **Modern Data Architecture**

**Applications**
- Business Analytics
- Custom Applications
- Packaged Applications

**Data System**
- RDBMS
- EDW
- MPP
- Repositories

**Sources**
- OLTP, ERP, CRM Systems
- Documents, Emails
- Web Logs, Click Streams
- Social Networks
- Machine Generated
- Sensor Data
- Geolocation Data

**Dev & Data Tools**
- Build & Test

**Operations Tools**
- Provision, Manage & Monitor
### Core Capabilities of Enterprise Hadoop

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DATA MANAGEMENT</strong></td>
<td></td>
</tr>
<tr>
<td>Load data and manage according to policy</td>
<td></td>
</tr>
<tr>
<td>Access your data simultaneously in multiple ways (batch, interactive, real-time)</td>
<td></td>
</tr>
<tr>
<td>Store and process all of your Corporate Data Assets</td>
<td></td>
</tr>
<tr>
<td>Provide deployment choice across physical, virtual, cloud</td>
<td></td>
</tr>
<tr>
<td><strong>PRESENTATION &amp; APPLICATION</strong></td>
<td>Enable both existing and new application to provide value to the organization</td>
</tr>
<tr>
<td><strong>DATA ACCESS</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SECURITY</strong></td>
<td>Provide layered approach to security through Authentication, Authorization, Accounting, and Data Protection</td>
</tr>
<tr>
<td><strong>OPERATIONS</strong></td>
<td>Deploy and effectively manage the platform</td>
</tr>
</tbody>
</table>
...delivered in Open Source
HDP 2.1: Enterprise Hadoop

HDP 2.1
Hortonworks Data Platform

GOVERNANCE & INTEGRATION
Data Workflow, Lifecycle & Governance
- Falcon
- Sqoop
- Flume
- NFS
- WebHDFS

DATA ACCESS
Batch
- Map
- Reduce
Script
- Pig
SQL
- Hive
- Tez
- HCatalog
NoSQL
- HBase
- Accumulo
Stream
- Storm
Search
- Solr
Others
- In-Memory Analytics
- SV engines

YARN: Data Operating System

SECURITY
Authentication
Authorization
Accounting
Data Protection
Storage: HDFS
Resources: YARN
Access: Hive, ...
Pipeline: Falcon
Cluster: Knox

DATA MANAGEMENT
Linux
Windows
Deployment Choice
On-premise
Cloud

OPERATIONS
Provision,
Manage &
Monitor
- Ambari
- Zookeeper
Scheduling
- Oozie
HDP 2.1 Investment Themes

HDP 2.1 Represents a MAJOR step forward for Hadoop

*Delivery of Interactive Query via Stinger Initiative, Addition of Data Governance, more Security, Stream Processing and Search, Highlight Release*

Three Key Highlights of Release

1. **Stinger Initiative DELIVERED**: Interactive Query in *Apache Hive*

2. **NEW Capabilities** for Hadoop
   - **Governance**: delivered with *Apache Falcon*
   - **Security**: *Apache Knox* extends perimeter security for Hadoop

3. **NEW Engines** included in HDP
   - **Stream processing**: Apache Storm to analyze/process streams of data
   - **Search**: via *Apache Solr*
HDP 2.1: Reliable, Consistent & Current

HDP certifies most recent & stable community innovation
Interactive SQL IN Hadoop Delivered

Stinger Initiative – DELIVERED
Next generation SQL based interactive query in Hadoop

Speed
Improve Hive query performance has increased by 100X to allow for interactive query times (seconds)

Scale
The only SQL interface to Hadoop designed for queries that scale from TB to PB

SQL
Support broadest range of SQL semantics for analytic applications running against Hadoop

Apache Hive Contribution… an Open Community at its finest

1,672
Jira Tickets Closed

145
Developers

44
Companies

~390,000
Lines Of Code Added… (2x)

13
Months

Stinger Project

Stinger Phase 1:
- Base Optimizations
- SQL Types
- SQL Analytic Functions
- ORCFile Modern File Format

Stinger Phase 2:
- SQL Analytic Functions
- Advanced Optimizations
- Performance Boosts via YARN

Stinger Phase 3:
- Hive on Apache Tez
- Query Service (always on)
- Buffer Cache
- Cost Based Optimizer (Optiq)

HDFS (Hadoop Distributed File System)
New: Data Governance & Integration

Apache Falcon
Simplified Data Governance for Enterprise Hadoop

- First time included in HDP
- Provides key governance framework for:
  - Acquisition & processing of data sets
  - Replication & Retention of datasets
  - Redirect datasets to non-Hadoop extensions
  - Provides audit trail & lineage

Another great example of Open Community Innovation
Originally built and contributed to Apache by InMobi
- Fastest path to innovation is the open community
- 14 months in the making
- Tested in production
- Vibrant community of developers building

Investment Phases

Phase-1
- Incubate Apache Falcon
- Dataset replication & retention
- Falcon tech preview

Phase-2
- Basic dashboard for pipeline viewing
- Kerberos security support
- Ambari integration for management
- Hive/HCatalog integration

Phase-3
- Advanced Dashboard for pipeline definition & management
- Audit
- Lineage
- Data tagging
- File import SSH & SCP
New: Apache Knox for Perimeter Security

Important Note: Security for Hadoop must be addressed within every layer of the stack and integrated into existing frameworks.

For a full description of what is available in Enterprise Hadoop today across Authentication, Authorization, accountability and Encryption please visit our security labs page.

Apache Knox
Perimeter security for Hadoop

- A common place to perform authentication across Hadoop and all related projects
- Integrated to LDAP and AD
- Currently supports: WebHDFS, WebHCAT, Oozie, Hive & HBase
- Broad community effort, Incubated with Microsoft, broad set of developers involved

Security Investments

Phase 1
- Strong AuthN with Kerberos
- HBase, Hive, HDFS basic AuthZ
- Encryption with SSL for NN, JT, etc.
- Wire encryption with Shuffle, HDFS, JDBC

Security Phase 2:
- ACLs for HDFS
- Knox: Hadoop REST API Security
- SQL-style Hive AuthZ (GRANT, REVOKE)
- SSL support for Hive Server 2
- SSL for DN/NN UI & WebHDFS
- PAM support for Hive

Security Phase 3:
- Audit event correlation and Audit viewer
- Support Token-Based AuthN beyond kerb
- Data Encryption in HDFS, Hive & Hbase
- Knox for HDFS HA, Ambari & Falcon

HDP 2.1
New: Stream Processing with Apache Storm

Apache Storm
Real-time event processing for sensor and business activity monitoring

- Unlocks new business cases for Hadoop
- Key component of a data lake architecture
- Scale: Ingest millions of events per second. Fast query on petabytes of data
- Integrated with Ambari to manage

Investment Phases

**Phase-1**
- Install, Start, & Stop via Ambari
- Kafka, HBase, & HDFS Connectors
- Ganglia & Nagios based monitoring

**Phase-2**
- Storm-on-YARN
- Ingest & Notification for JMS
- Data persistence: EDWs, RDBMS, Cassandra

**Phase-3**
- High Availability mgmnt w/Ambari
- AD/LDAP plugin for authentication
- Declarative “wiring”
- Hive update support
- Advanced scheduler
New: Search for Hadoop

- **Apache Solr**
  Open source enterprise search for Hadoop and HDP
  - Open architecture: In the community, for the community
  - Simple, powerful UI for advanced search applications
  - High performance indexing & sub-second search times over billions of documents
  - Deep Integration Roadmap with HDP

- **Partnership with LucidWorks**
  - LucidWorks provides tier 3 & 4 support
  - Alignment w/ strategy of working within the community and with the core committers
  - 9 committers total (7 PMC)
Cascading SDK & HDP 2.1

**Cascading SDK**
Enables the rapid development of batch and interactive data-driven applications

**Integration Roadmap**
- **Step 1**: Integrate Cascading SDK for customers to use with HDP 2.1
- **Step 2**: Integration with Tez
Tech Preview: Apache Spark

In-memory processing is “HOT!”

…however, most of the world using for science and machine learning

In memory sandbox for iterative data analytics used by a handful of data scientists

Hortonworks provides guidance for initial applicability and scale

- Exploring key use cases with customers focused on iterative access & machine learning
- Experience thus far supports target deployments of no more than: 1 TB of data, 40 nodes, and 1-3 users
- Skill set required: Scala (Java-based API Framework)
Apache Ambari is the only 100% open source framework for provisioning, managing and monitoring Apache Hadoop clusters.

New in HDP 2.1
- Support new Data Access Engines
- Stack extensibility, Cluster Blueprints
- Rolling restarts
- Maintenance mode
- more...
HDP 2.1 Investment Themes

HDP 2.1 Represents a MAJOR step forward for Hadoop

Delivery of Interactive Query via Stinger Initiative, Addition of Data Governance, more Security, Stream Processing and Search, Highlight Release

Three Key Highlights of Release

1. **Stinger Initiative DELIVERED**: Interactive Query in **Apache Hive**

2. **NEW Capabilities** for Hadoop
   - **Governance**: delivered with **Apache Falcon**
   - **Security**: **Apache Knox** extends perimeter security for Hadoop

3. **NEW Engines** included in HDP
   - **Stream processing**: Apache Storm to analyze/process streams of data
   - **Search**: via Apache Solr

...AND the HDP Spark Tech Preview, Simultaneous Linux & Windows Release, COUNTLESS additional features
Thank You

Jeff Markham
Technical Director, APAC
jmarkham@hortonworks.com