Big Data and the Data Lake

TDWI Denver: September 2015

Mac Moore
Solutions Engineering
Hortonworks
Data Lake Agenda

• What is a Data Lake?
• Hadoop is the perfect match
• The journey to data driven
• Real-world use cases
• Key data architecture capabilities
What is a Data Lake?
What is a Data Lake?

Architectural Pattern in the Data Center
Uses Hadoop to deliver deeper insight across a large, broad, diverse set of data efficiently

- **Multipurpose, Open PLATFORM** for Data (NOT a database)
- Land all data in a single place and interact with it in many ways
- Allows for the ecosystem to provide higher level services (SAS, SAP, Microsoft, MPP, In-memory, etc..)
- First class data management capabilities (metadata management, security, transformation pipelines, replication, retention, etc..)
Hadoop is the perfect fit
Traditional systems under pressure

1. Challenges
   - Constrains data to app
   - Can't manage new data
   - Costly to Scale

2. New Data

Business Value

2012
2.8 Zettabytes

2020
40 Zettabytes

INDUSTRY LEADERS

LAGGARDS
What is Hadoop?

Hadoop is an open data platform

• consisting of a collection of tools for solving problems at internet scale.
• with a centralized approach to governance, security and operations.
• that allows organizations to eliminate data silos and cost effectively bring more data under management.
A Blueprint for Enterprise Hadoop

**DATA MANAGEMENT**

- Load data and manage according to policy
- Access your data simultaneously in multiple ways (batch, interactive, real-time)
- Store and process all of your Corporate Data Assets

**PRESENTATION & APPLICATION**

- Enable both existing and new application to provide value to the organization

**ENTERPRISE MGMT & SECURITY**

- Empower existing operations and security tools to manage Hadoop
- Provide layered approach to security through Authentication, Authorization, Accounting, and Data Protection
- Deploy and effectively manage the platform

**DATA ACCESS**

- Data Operating System (Resource Mgmt)

**SECURITY**

- Provide deployment choice across physical, virtual, cloud

**OPERATIONS**

- Present and application
- Governance & integration
- Data access
- Data management
- Security
- Operations
The Journey to Data Driven
Hortonworks® customers leverage our technology to transform their businesses, either by achieving new business objectives or by reducing costs. The journey typically involves both of those goals in combination, across many use cases.
Business executives are driving transformational outcomes with next-generation applications that empower new uses of Big Data including: data discovery, a single view of the customer and predictive analytics.
IT executives are delivering substantial reductions in operating costs by modernizing their data architectures with Open Enterprise Hadoop. These cost saving innovations include active archive of cold data, offloading ETL processes and enriching existing data.
Customer Journeys

The Business Case Stories
Customer Journeys (Resources)

Hortonworks Customer Page:
http://hortonworks.com/customers/

Mercy

Merck:
http://hortonworks.com/blog/hdp-for-manufacturing-yield-optimization-in-pharma/

Neustar
http://hortonworks.com/customer/neustar/

Cardinal Health:
http://hortonworks.com/customer/cardinal-health/
Key Data Platform Capabilities
A Blueprint for Enterprise Hadoop

**PRESENTATION & APPLICATION**
Enable both existing and new application to provide value to the organization

**DATA ACCESS**
Access your data simultaneously in multiple ways (batch, interactive, real-time)

**YARN Data Operating System**
Store and process all of your Corporate Data Assets

**SECURITY**
Provide layered approach to security through Authentication, Authorization, Accounting, and Data Protection

**OPERATIONS**
Deploy and effectively manage the platform

**GOVERNANCE & INTEGRATION**
Load data and manage according to policy

**DATA MANAGEMENT**
Provide deployment choice across physical, virtual, cloud

**ENTERPRISE MGMT & SECURITY**
Empower existing operations and security tools to manage Hadoop

**DEPLOYMENT OPTIONS**
Open Enterprise Hadoop

- Open
- Central
- Interoperable
- Enterprise Ready
A Genuinely Open Data Platform

- **Eliminates Risk**
  - of vendor lock-in by delivering 100% Apache open source technology

- **Maximizes Community Innovation**
  - with hundreds of developers across hundreds of companies
  - **Integrates Seamlessly**
  - through committed co-engineering partnerships with other leading technologies
Centralized Platform for operations, governance and security

Diverse Applications run simultaneously on a single cluster

Maximum Data Ingest including existing and new sources, regardless of raw format

Shared Big Data Assets across business groups, functions and users
Provides Consistent Operations

Centralized
management and monitoring of Hadoop clusters

Automated Provisioning
either on-premises or in the cloud with the Cloudbreak API for clusters in minutes

Managed Services
for high availability and consistent lifecycle controls, with dashboards and alerts
Enables Trusted Governance

Data Management
along the entire data lifecycle

Modeling with Metadata
enables comprehensive data lineage through a hybrid approach

Interoperable Solutions
across the Hadoop ecosystem, through a common metadata store
Ensures Comprehensive Security

Comprehensive Security through a platform approach

Encryption of data at rest and in motion

Centralized Administration of security policies and user authentication

Fine-Grain Authorization for data access control
Synchronized with Industry Standards

Improves Ecosystem Interoperability
as part of the Open Data Platform (ODP) initiative, founded by Hortonworks

Unlocks Choice
for the customer to use components from multiple vendors integrated with HDP

Eliminates Wasteful Guesswork
for the architect who needs to coordinate system versions
Integrated with the Ecosystem
Goals for Hadoop Operations

Scale Operations
Provision, manage and monitor Hadoop clusters at scale

Integrate the Enterprise
Leverage a robust API for integration with existing enterprise systems

Extend the Ecosystem
Provide extensible platform, combining technologies using tools such as Stacks and Views

Coordinate Services
Schedule Hadoop jobs, maintain and synchronize configuration information
Key Players in Hadoop Operations

Lead Persona: Hadoop Operator

Supporting Personas

- System Administrator
- Information Security Officer
- Database Administrator
- Network Administrator
Apache Ambari

Provision, Manage and Monitor Hadoop Clusters
About Apache Ambari

100% Open Source
operational framework, developed in coordination with other Apache components

Ecosystem Awareness
for easy integration via REST APIs and visible through a single pane of glass

Intuitive User Interface
for ongoing, frequently refreshed insight into cluster performance
5 Core Activities Managed with Ambari

- **Cluster Management**: installation, upgrade and setup security
- **Configuration Management**: host groups, versioning, comparisons, reversion and recommendations
- **Extensibility**: with Stacks and Views
- **Monitoring**: dashboard, health checks and alerts
- **Service Management**: lifecycle controls, rolling restarts, expand and shrink cluster capacity
Cluster Provisioning

Install Clusters automatically, with configuration and health checks

Automated Upgrades for Ambari and HDP

Establish Security Kerberos setup, either automated or manual
Configuration Management in HDP

Manage Settings
set, revert, version and compare HDP settings

Recommended Settings
provide industry-recommended defaults

Configuration Groups
target configurations in a mixed host environment
New guided configurations makes it easier to manage settings.
Ambari Views Framework to Customize the UI

Extend the Ambari Web Interface by exposing custom UI features for Hadoop services.

Ambari Admins Assign Views to Ambari Web Users via an entitlement framework that controls access.
New user interface enables fast & easy SQL definition and execution.
Ambari Views Gallery

Built by Hortonworks, Community and Partners

Find Community Views:

Hortonworks Gallery. By Developers. For Developers

Key repos for Hadoop® development

The Hortonworks Gallery gathers the code, views, blueprints and tutorial you need in one place. Browse the repositories, find the code and examples you need, and jumpstart your development projects. And as a member of the community, you can take freely and contribute frequently to help move Hadoop forward.

Find Community Views:

Ambary Views, Developer, Featured
Ambary API Explorer
An Ambari View to explore its REST API

View (Pig)
View (Hive)
View (Tez)
…

Views Framework

Core to Ambari with Ambari 1.7+
Cloudbreak

Quickly Launch HDP in the Cloud
Launch on Any Major Cloud Platform with Blueprints

Cloudbreak
1. Pick a Blueprint
2. Choose a Cloud
3. Launch HDP

Example Ambari Blueprints
IoT Apps, BI / Analytics, Data Science, Dev / Test

Microsoft Azure
IoT Apps
(Storm, HBase, Hive)

Amazon Web Services
BI / Analytics
(Hive)

Google Cloud Platform
Data Science
(Spark)

Open Stack
Dev / Test
(all HDP services)

Microsoft Azure

Amazon Web Services

Google Cloud Platform

Open Stack
Security and Data Governance in Open Enterprise Hadoop
Security Challenges for a Hadoop Data Lake

Central Repository
of critical, sensitive data

Long-term Retention
of data stored for years or decades

Reliable Integration
always secure despite a fluctuating ecosystem

Dynamic Access
permits users to analyze data in new and different ways, always in flux
Our Comprehensive Approach To Security

**Administration**
- Centrally manage consistent security
- How do I set policy across the entire cluster?

**Authentication**
- Prove the identity of systems and users
- Who are you and how can you prove it?

**Authorization**
- Provide secure access to data
- What can you do once you’re authenticated?

**Audit**
- Maintain a record of data access events
- What did you do and when did you do it?

**Data Protection**
- Safeguard data at rest and in motion
- How can you encrypt the data?
Our Comprehensive Approach To Security

**Administration**
- Centrally manage consistent security: **APACHE RANGER**

**Authentication**
- Prove the identity of systems and users: **KERBEROS & APACHE KNOX**

**Authorization**
- Provide secure access to data: **APACHE RANGER**

**Audit**
- Maintain a record of data access events: **APACHE RANGER & APACHE ATLAS**

**Data Protection**
- Safeguard data at rest and in motion: **HDFS TDE with RANGER KMS**
Integrated Platform Security

Hortonworks Data Platform 2.3

**GOVERNANCE**
- Apache Atlas
- Apache Falcon
- Apache Kafka
- Apache Flume
- Apache Sqoop

**BATCH, INTERACTIVE & REAL-TIME DATA ACCESS**
- YARN: Data Operating System (Cluster Resource Management)
- Map Reduce
- Apache Hive
- Apache Pig
- Apache HBase
- Apache Accumulo
- Apache Solr
- Apache Spark
- Apache Storm
- YIS Engines
- HDFS (Hadoop Distributed File System)

**SECURITY**
- Apache Ranger
- Apache Knox
- Apache Atlas
- HDFS Encryption

**OPERATIONS**
- Apache Ambari
- Cloudbreak
- Apache ZooKeeper
- Apache Oozie

Security integrated into all platform components.
Integrated Platform Security

Security consistently applied across the data access engines
Build or retire applications without effecting security
Apache Ranger

Comprehensive security for Enterprise Hadoop
Ranger Centralizes Security for Deep Visibility

Centralized Platform
Consistently define, administer and manage security policies
Define a policy once and apply it to all the applicable components across the stack

Fine-grained Definitions
Administer security for:
• Database
• Table
• Column
• LDAP Groups
• Specific Users

Deep Visibility
Administrators have complete visibility into the security administration process
## Policy Details:

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy ID</td>
<td>18</td>
</tr>
<tr>
<td>Policy Name</td>
<td>Call_Details_Table</td>
</tr>
<tr>
<td>Hive Database</td>
<td>xademo</td>
</tr>
<tr>
<td>Hive Table</td>
<td>call_detail_records</td>
</tr>
<tr>
<td>Hive Column</td>
<td>phone_number</td>
</tr>
<tr>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>Audit Logging</td>
<td>YES</td>
</tr>
</tbody>
</table>

## User and Group Permissions:

<table>
<thead>
<tr>
<th>Permissions</th>
<th>Select Group</th>
<th>Select User</th>
<th>Permissions</th>
<th>Delegate Admin</th>
</tr>
</thead>
<tbody>
<tr>
<td>select</td>
<td>developer</td>
<td>Select User</td>
<td>select</td>
<td></td>
</tr>
</tbody>
</table>

[Select User]
### List of Policies: sandbox_hive

<table>
<thead>
<tr>
<th>Policy ID</th>
<th>Policy Name</th>
<th>Status</th>
<th>Audit Logging</th>
<th>Groups</th>
<th>Users</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>sandbox_hive-1-20150529142947</td>
<td>Enabled</td>
<td>Enabled</td>
<td>--</td>
<td>xapolymgr</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Hive Global Tables Allow</td>
<td>Disabled</td>
<td>Enabled</td>
<td>public</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Hive Global UDF Allow</td>
<td>Disabled</td>
<td>Enabled</td>
<td>public</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Call_Details_Table</td>
<td>Enabled</td>
<td>Enabled</td>
<td>developer</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Customer_Details_Table</td>
<td>Disabled</td>
<td>Enabled</td>
<td>Marketing</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Hive Demo Table Loader</td>
<td>Enabled</td>
<td>Enabled</td>
<td>--</td>
<td>hive</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Hive Demo UDF Loader</td>
<td>Enabled</td>
<td>Enabled</td>
<td>--</td>
<td>hive</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>admin policy</td>
<td>Enabled</td>
<td>Enabled</td>
<td>--</td>
<td>admin</td>
<td></td>
</tr>
</tbody>
</table>
Apache Knox

A single point of secure access for Hadoop clusters
# Apache Knox Provides API Security

## Single Access Point
- Kerberos encapsulation
- REST API hierarchy
- Consolidated API calls
- Multi-cluster support

## Central Controls
- Eliminates SSH “edge node”
- Central API management
- Central audit control
- Service level authorization

## Integrated with Existing Systems
- SSO integration – Siteminder and OAM
- LDAP and Active Directory integration
Governance in Open Enterprise Hadoop
Important Data Governance Terminology

Data
- HDFS files
- HCatalog definitions
- Falcon pipelines
- Ranger users

Metadata
- Title
- Description
- Author
- Subject
- Date created
- Date modified
- Data sensitivity

Taxonomy
- Business classification
- Customer/industry vocabulary
- Industry compliance standards

Governance Answers
- Who
- What
- Where
- When
- How
Apache Falcon

A framework for managing data lifecycles in the cluster
An Overview of Apache Falcon

Data Lifecycle Management
reusable data pipelines, central definitions, auto-generate process in Oozie

Business Continuity and Disaster Recovery
data replication and retention in HDFS and Hive, end-to-end pipeline monitoring

Audit and Compliance
visualize data pipeline lineage, track data pipeline audit logs & free form business labels
Data Replication with Falcon

Primary Hadoop Cluster

Staged Data → Cleansed Data → Conformed Data → Presented Data

Failover Hadoop Cluster

Staged Data → Present Data

Replication
Data Retention with Falcon

Staged Data -> Cleansed Data -> Conformed Data -> Presented Data

- Retain 5 Years
- Retain 3 Years
- Retain 3 Years
- Retain Last Copy Only

Retention Policy
Apache Atlas

Agile enterprise compliance through metadata exchange
Apache Atlas is Part of HDP

**Rest API for flexible access**
to Atlas services, HDP components and external tools

**Search with SQL-like domain specific language**
via key word, faceted and full-text searches

**Lineage for data and schema**
by capturing all SQL runtime activity on HiveServer2

**Exchange**
import existing metadata and export metadata to downstream systems
Name: sales_fact_monthly_mv
Description: sales fact monthly materialized view

sales_fact_monthly_mv  →  loadSalesMonthly  →  sales_fact_daily_mv  →  loadSalesDaily

sales_fact  →  time_dim
Governance-ready Certification Program

Engaged vendor partners to Atlas services, HDP components and external tools

Customers choose features to deploy a la carte

Low switching costs

Stability and interoperability with HDP at the core
Questions?

Thank You!
This presentation contains forward-looking statements involving risks and uncertainties. Such forward-looking statements in this presentation generally relate to future events, our ability to increase the number of support subscription customers, the growth in usage of the Hadoop framework, our ability to innovate and develop the various open source projects that will enhance the capabilities of the Hortonworks Data Platform, anticipated customer benefits and general business outlook. In some cases, you can identify forward-looking statements because they contain words such as “may,” “will,” “should,” “expects,” “plans,” “anticipates,” “could,” “intends,” “target,” “projects,” “contemplates,” “believes,” “estimates,” “predicts,” “potential” or “continue” or similar terms or expressions that concern our expectations, strategy, plans or intentions. You should not rely upon forward-looking statements as predictions of future events. We have based the forward-looking statements contained in this presentation primarily on our current expectations and projections about future events and trends that we believe may affect our business, financial condition and prospects. We cannot assure you that the results, events and circumstances reflected in the forward-looking statements will be achieved or occur, and actual results, events, or circumstances could differ materially from those described in the forward-looking statements.

The forward-looking statements made in this prospectus relate only to events as of the date on which the statements are made and we undertake no obligation to update any of the information in this presentation.

**Trademarks**

Hortonworks is a trademark of Hortonworks, Inc. in the United States and other jurisdictions. Other names used herein may be trademarks of their respective owners.