



# 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**

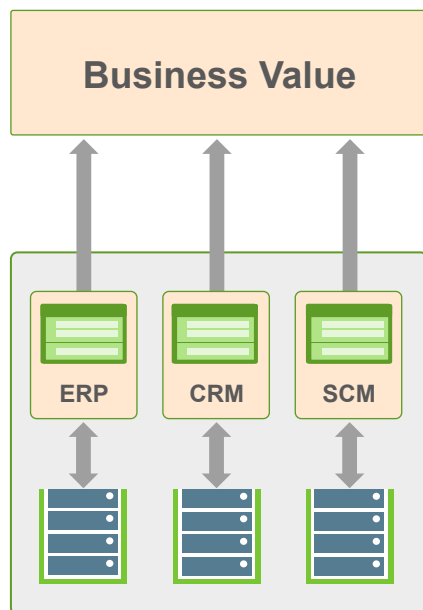
- **Multipurpose, Open PLATFORM** for Data (NOT a database)
- Land all data in a single place and interact with it in many ways
- Allows for the ecosystem to provide higher level services (SAS, SAP, Microsoft, MPP, In-memory, etc..)
- **First class data management capabilities** (metadata management,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

# Hadoop is the perfect fit

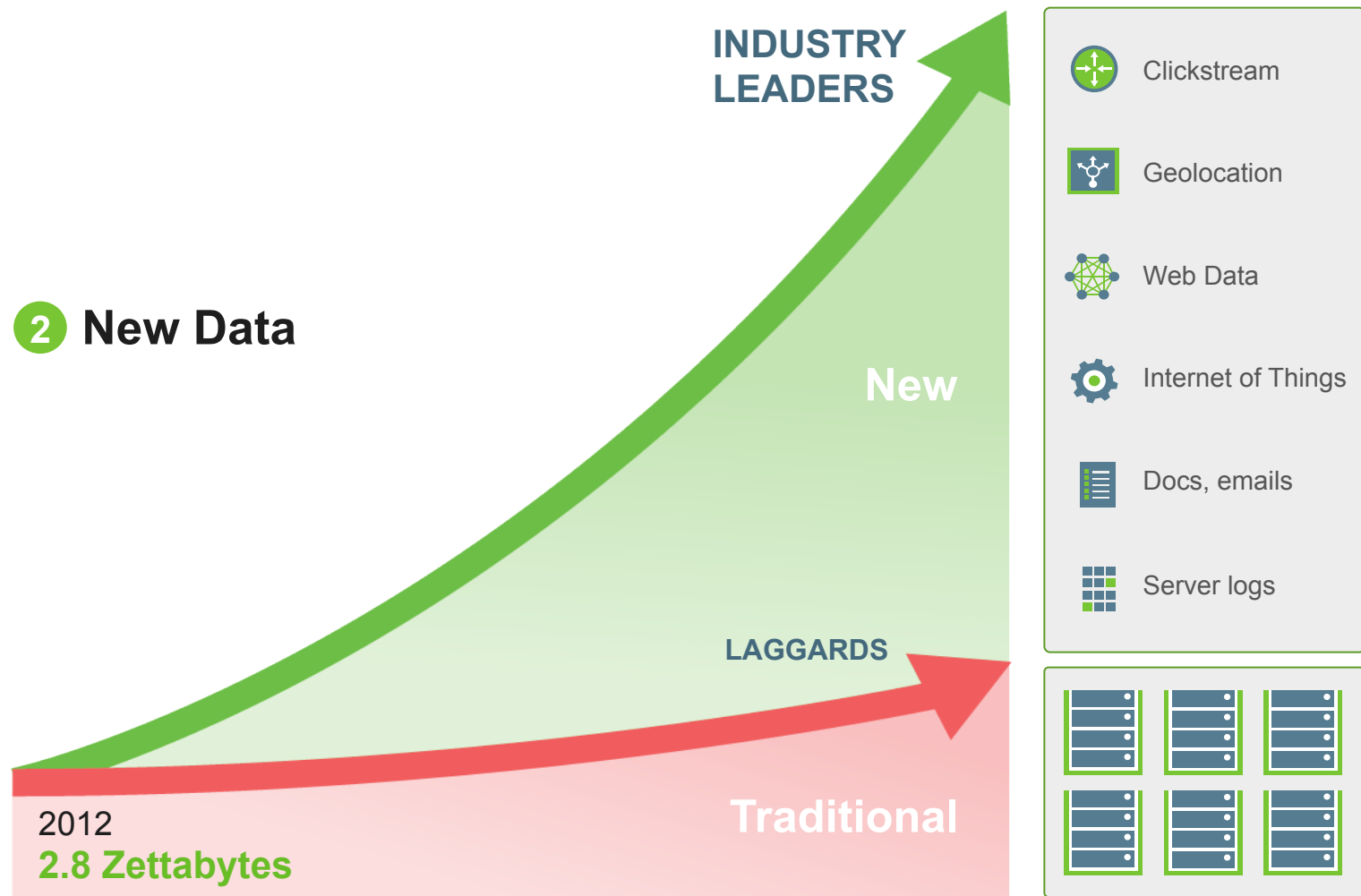
# Traditional systems under pressure

## 1 Challenges

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

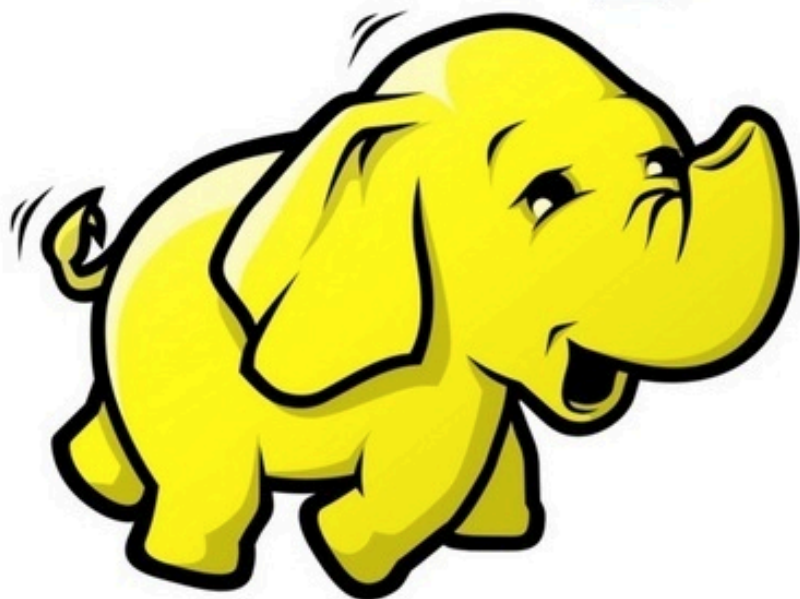


## 2 New Data



# What is Hadoop?

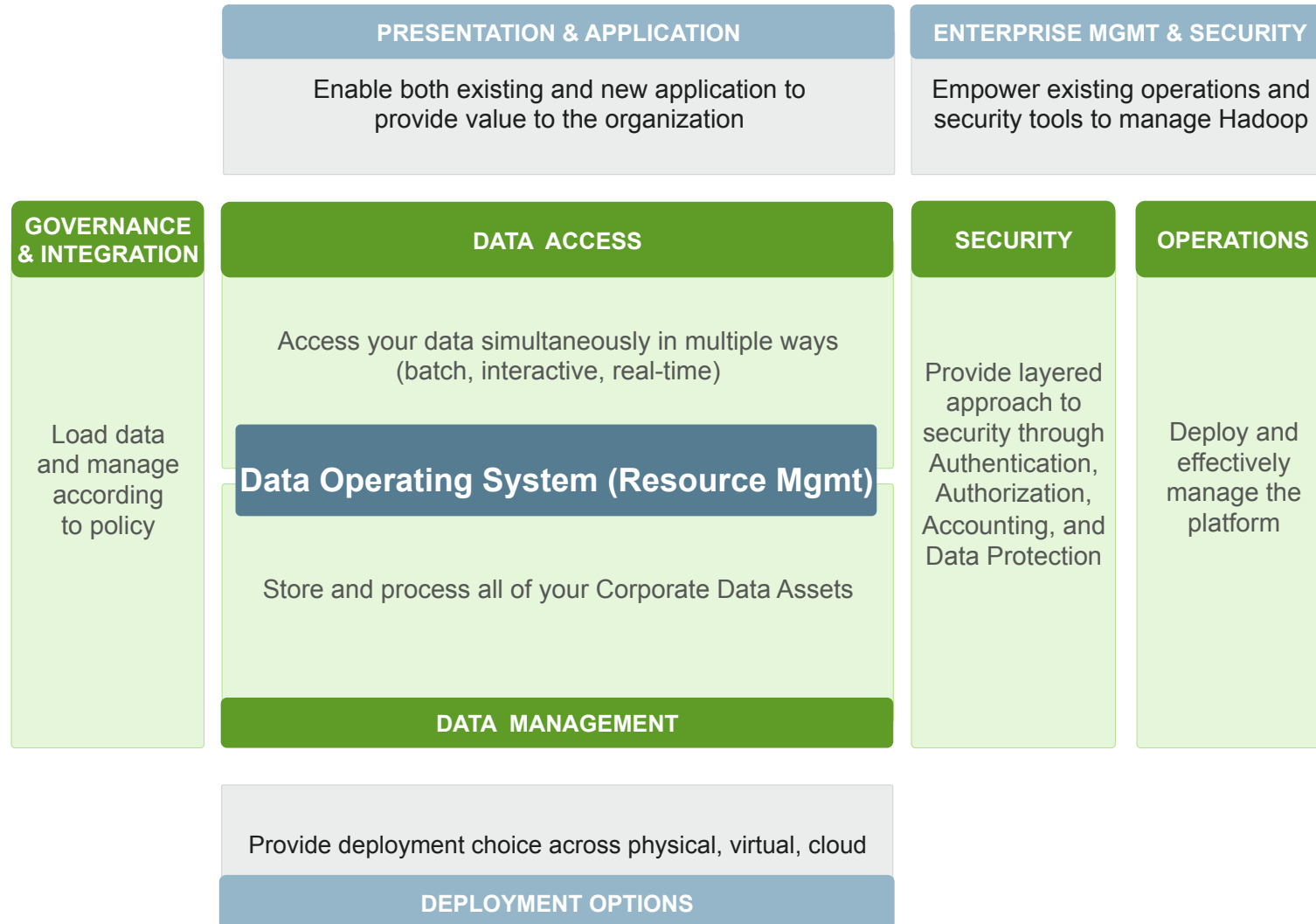
# *hadoop*



## 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.

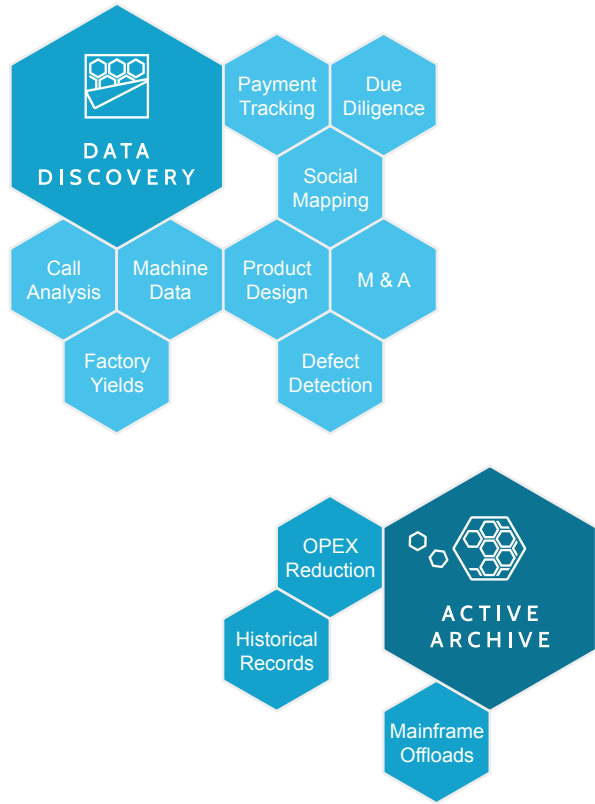
# A Blueprint for Enterprise Hadoop



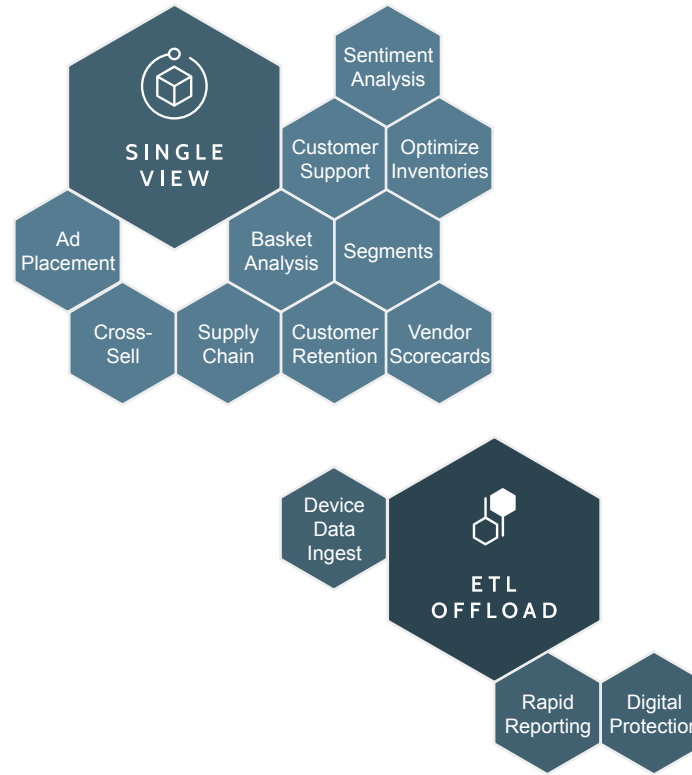


# The Journey to Data Driven

## EXPLORE



## OPTIMIZE

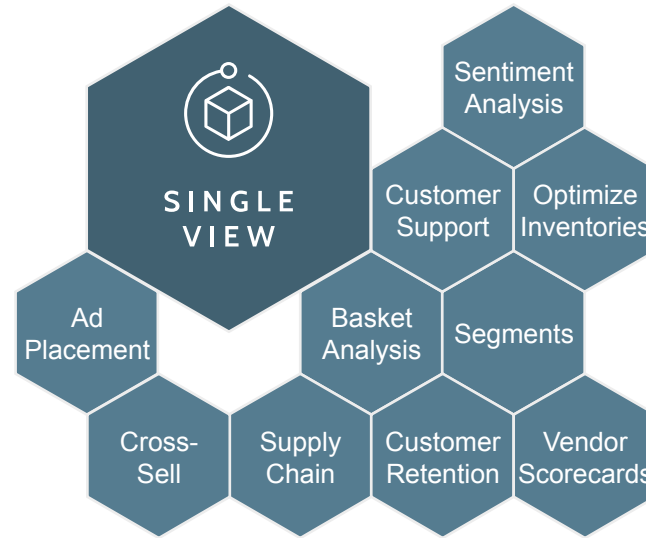
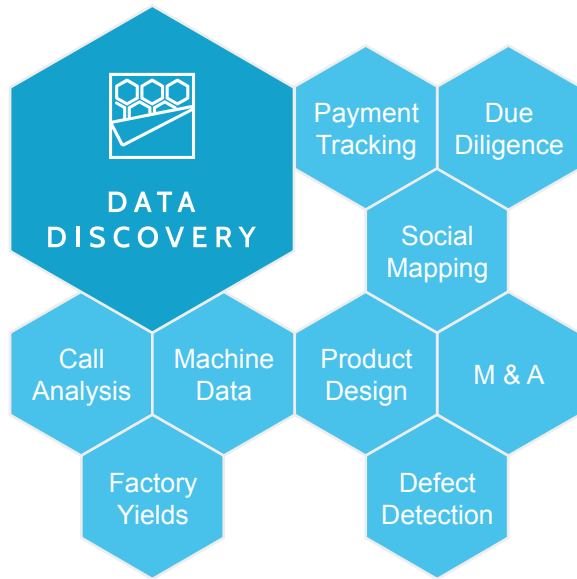


## TRANSFORM



# CUSTOMER JOURNEY

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.

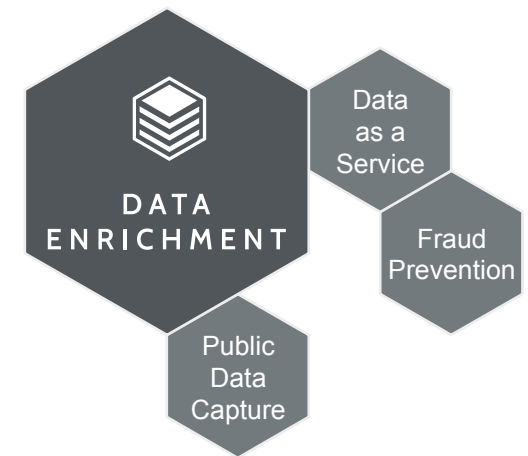
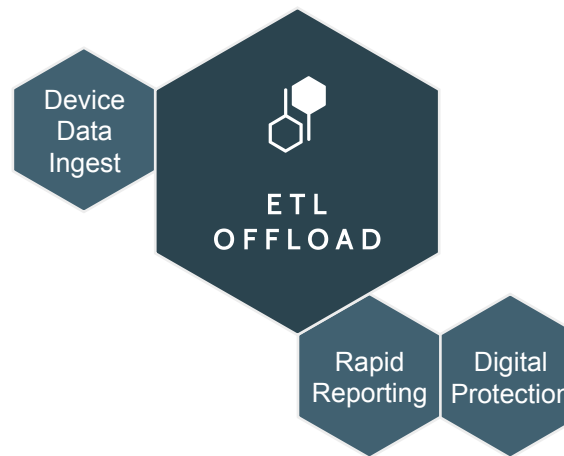
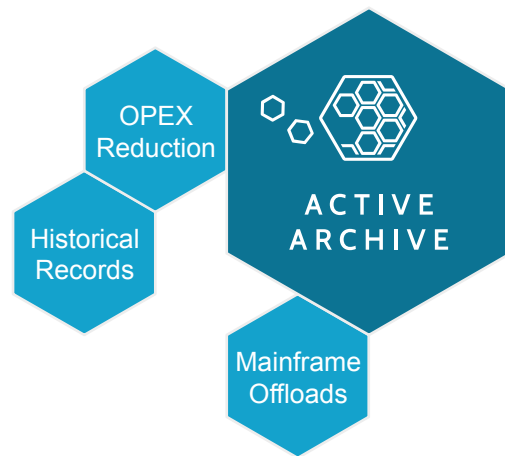


# BUSINESS OUTCOMES

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.

# COST SAVINGS

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

<http://hortonworks.com/blog/journey-to-a-health-care-data-lake-hadoop-at-mercy/>

## Merck:

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

## Neustar

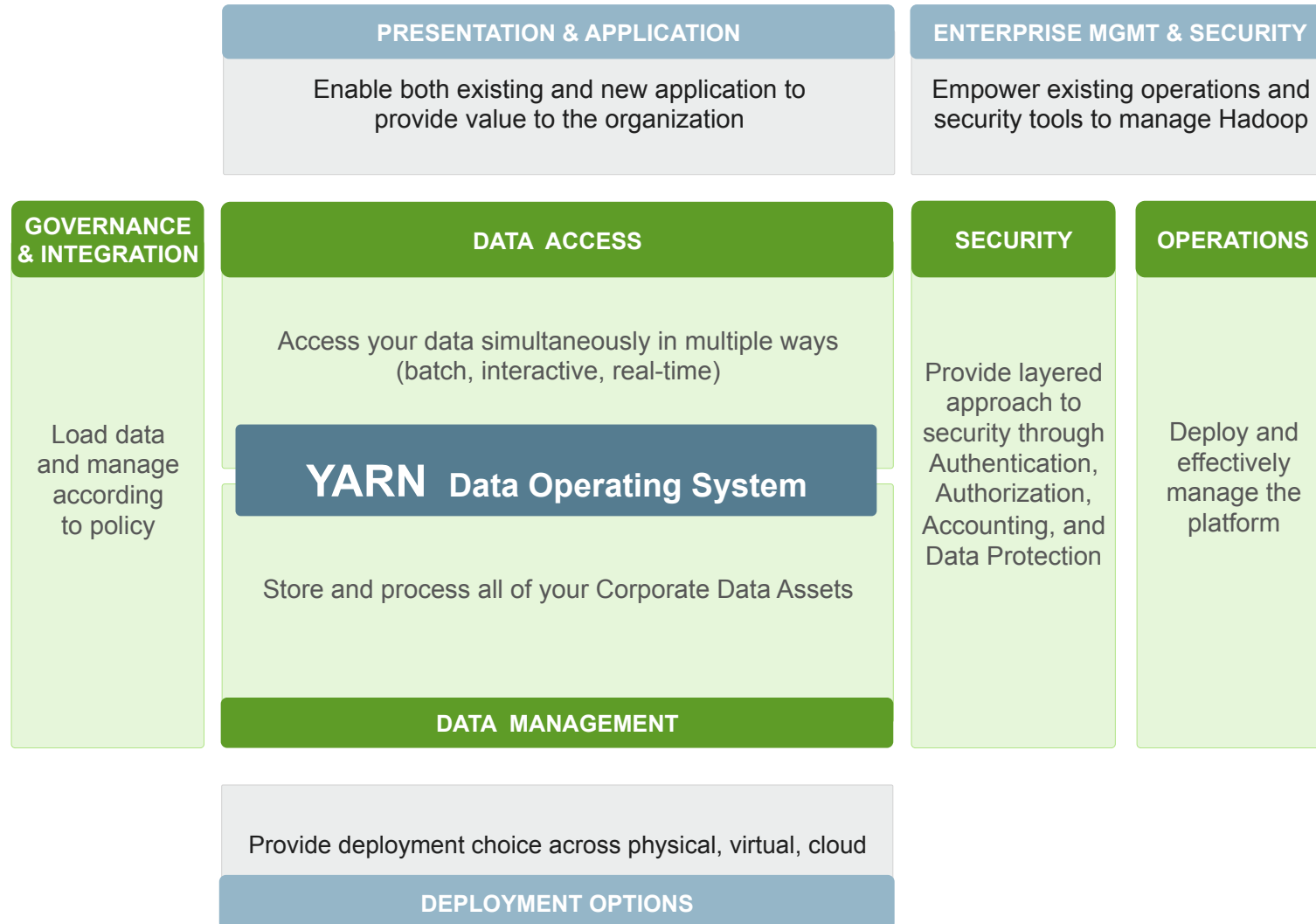
<http://hortonworks.com/customer/neustar/>

## Cardinal Health:

<http://hortonworks.com/customer/cardinal-health/>

# Key Data Platform Capabilities

# A Blueprint for Enterprise Hadoop





# Open Enterprise Hadoop



Open



Central

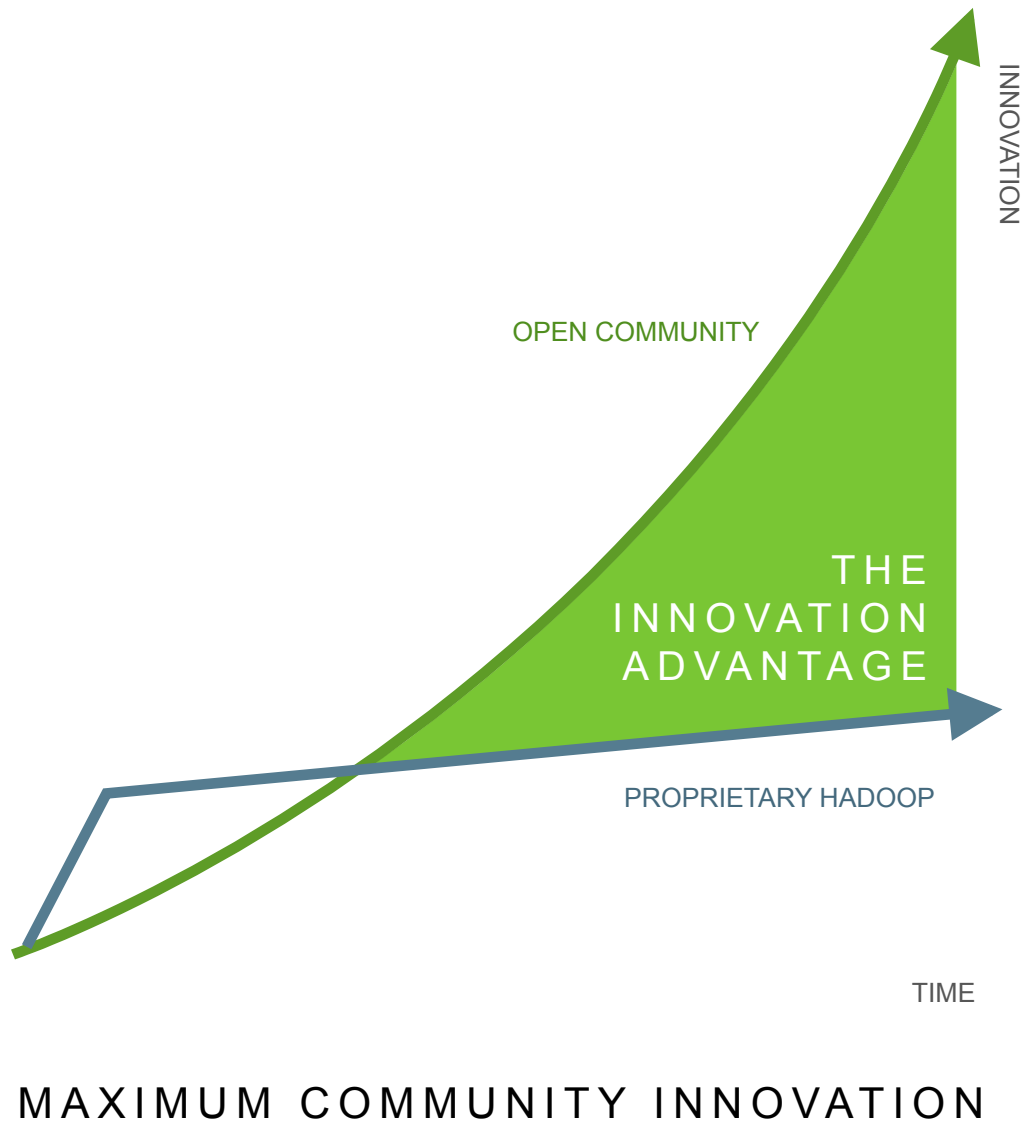


Interoperable



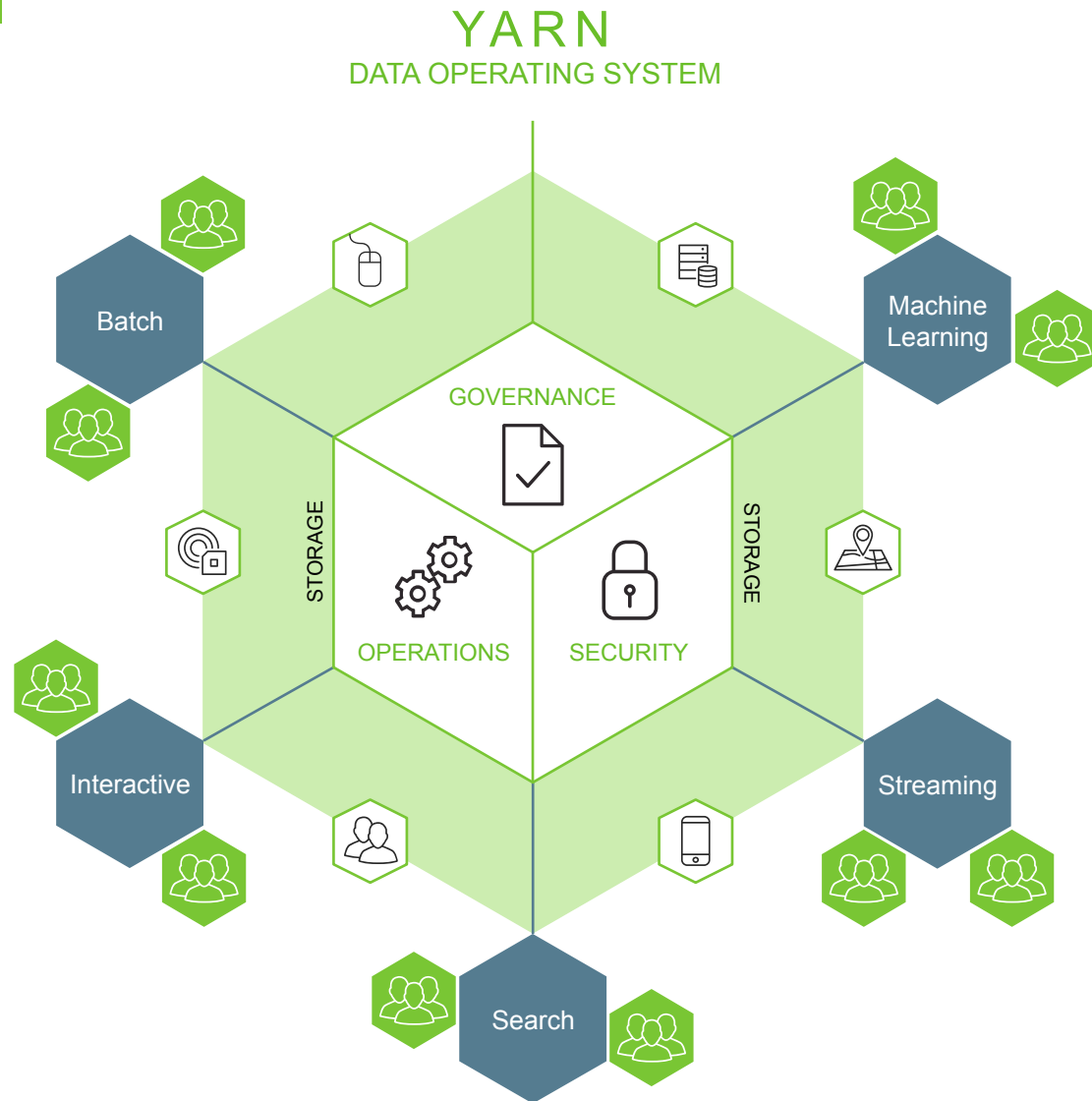
Enterprise Ready

# 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 with YARN-Based Architecture



## Centralized Platform

for operations, governance and security

## Diverse Applications

run simultaneously on a single cluster

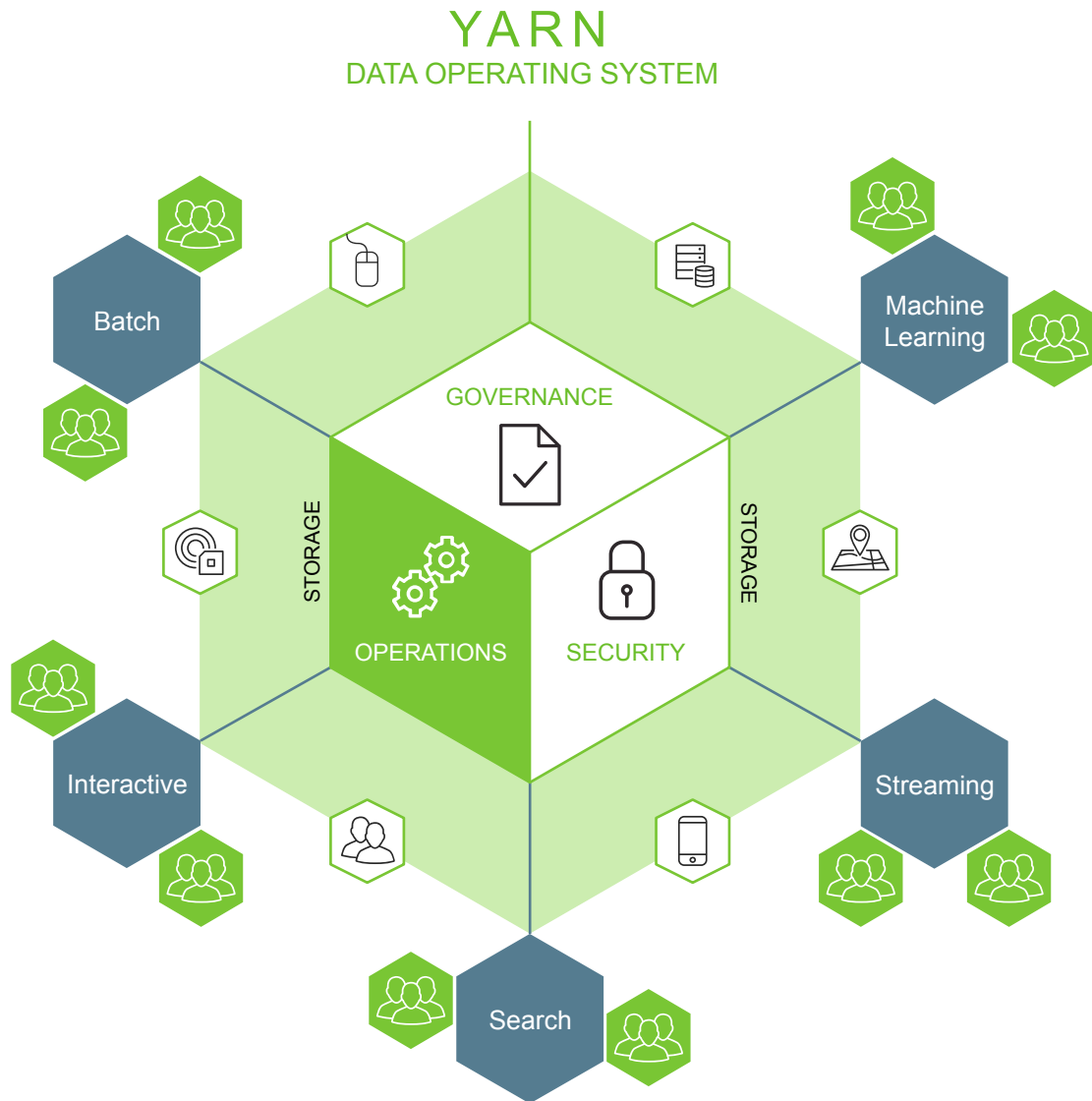
## Maximum Data Ingest

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

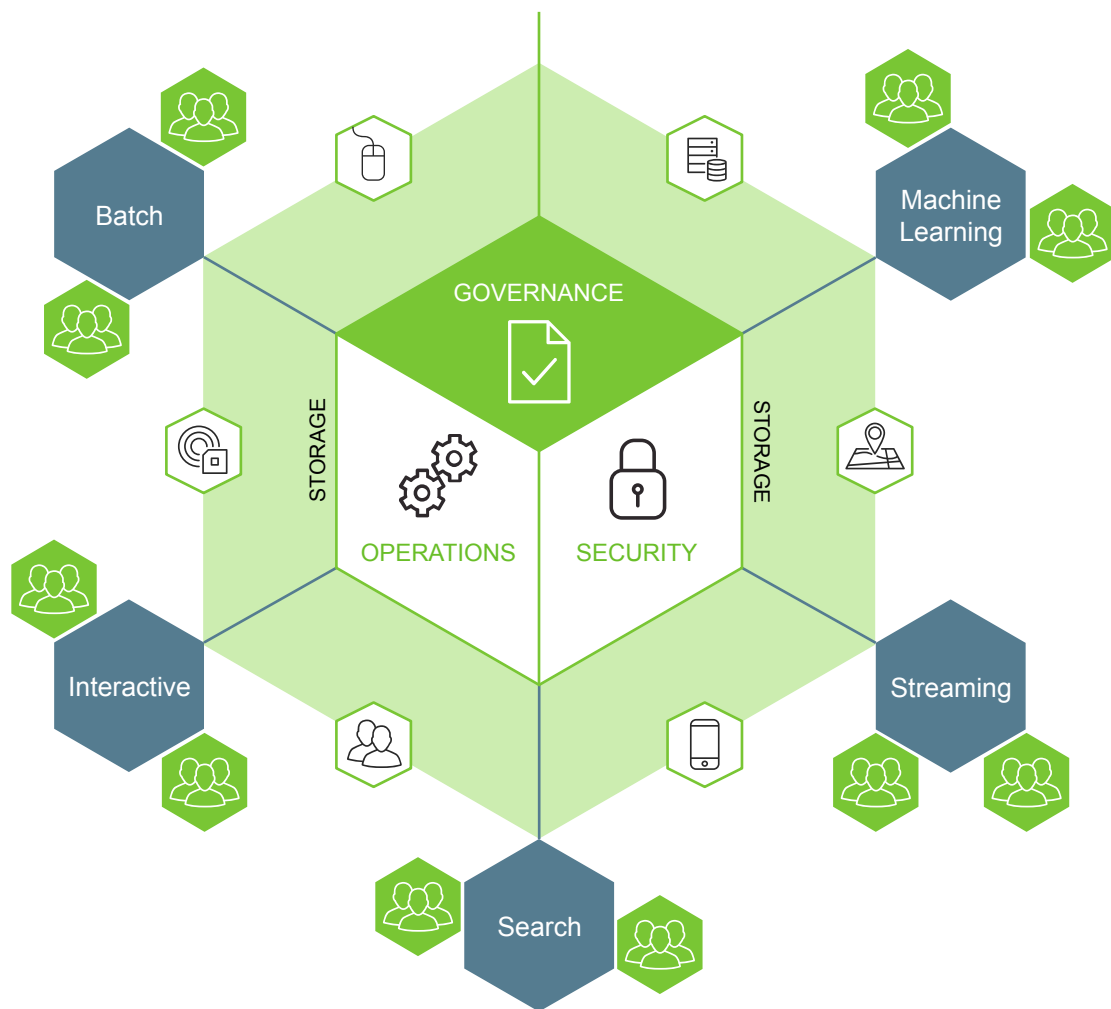
either on-premises or in the cloud with the Cloudbreak API for clusters in minutes

## Managed Services

for high availability and consistent lifecycle controls, with dashboards and alerts

# Enables Trusted Governance

## YARN DATA OPERATING SYSTEM



## 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

## YARN DATA OPERATING SYSTEM



### **Comprehensive Security**

through a platform approach

### **Encryption**

of data at rest and in motion

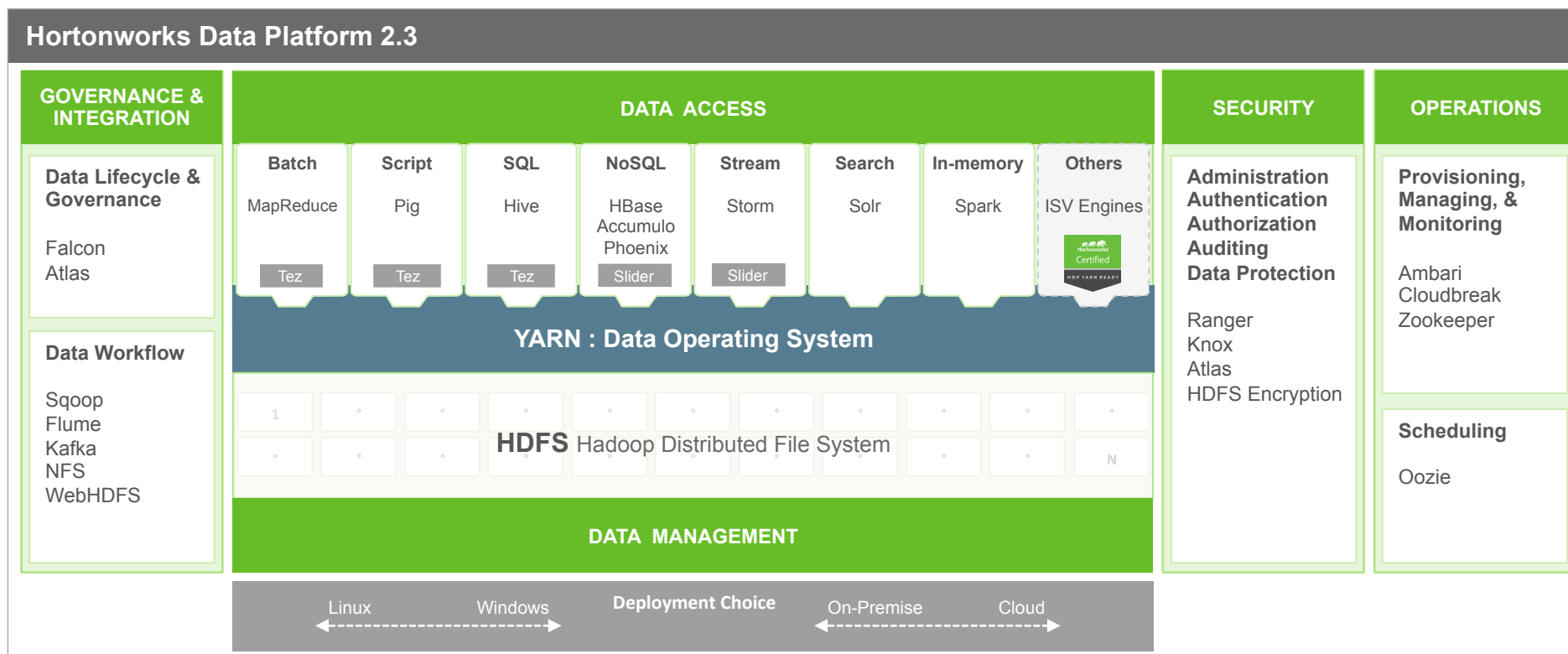
### **Centralized Administration**

of security policies and user authentication

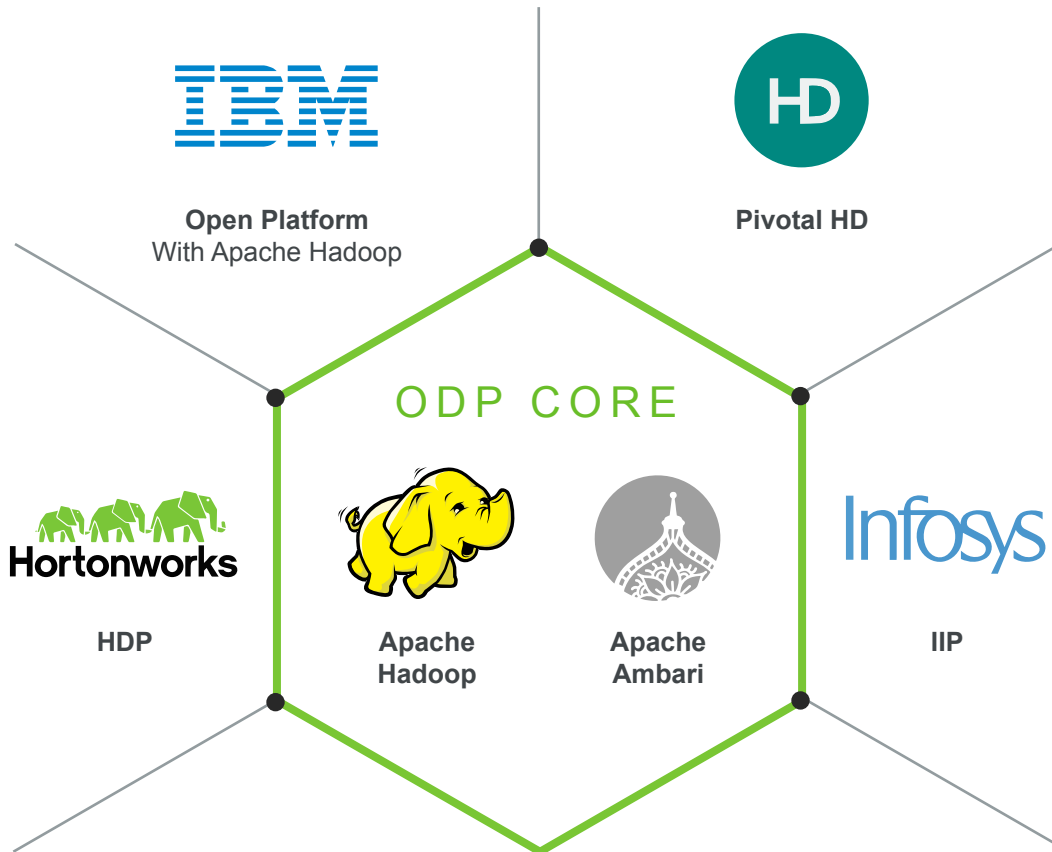
### **Fine-Grain Authorization**

for data access control

# Hadoop Distribution



# 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



HDP YARN READY





# Operating Hadoop at Scale

# 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**

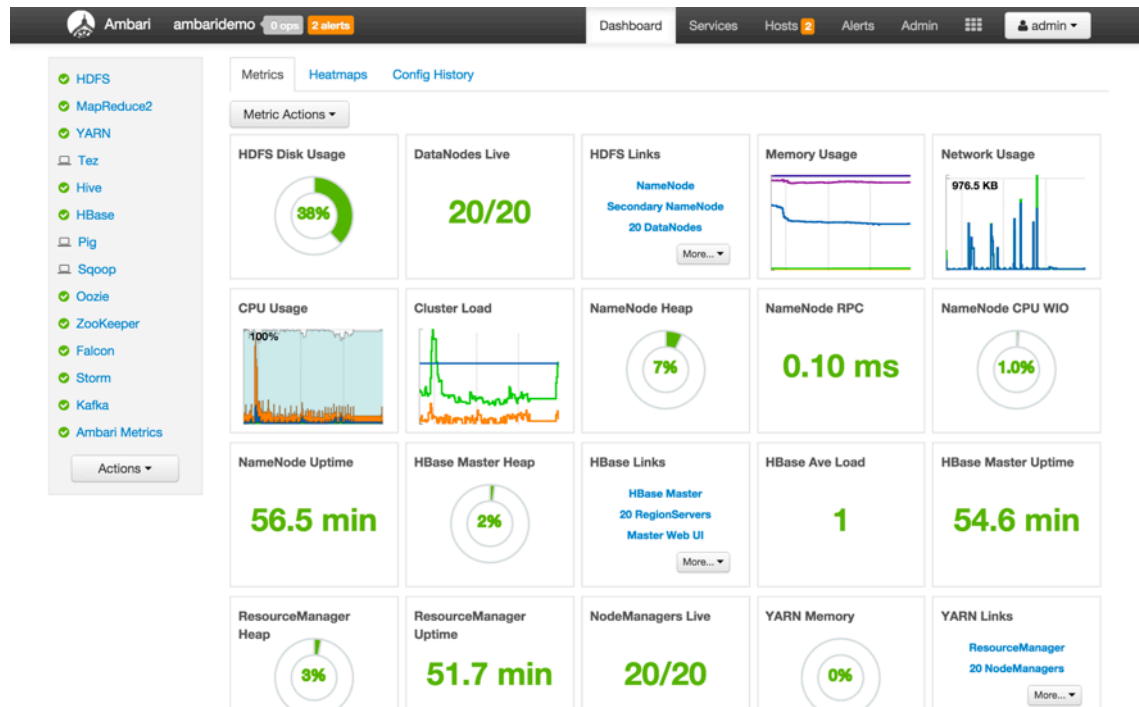
**Database  
Administrator**

**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

**Ambari**

CLUSTER INSTALL WIZARD

- Get Started
- Select Stack
- Install Options
- Confirm Hosts
- Choose Services
- Assign Masters
- Assign Slaves and Clients
- Customize Services
- Review
- Install, Start and Test**
- Summary

### Install, Start and Test

Please wait while the selected services are installed and started.

72 % overall

Show: All (20) | In Progress (20) | Warning (0) | Success (0) | Fail (0)

Host	Status	Message
ambari200-1.c.pramod-thangali.internal	64%	Preparing to start RegionServer
ambari200-2.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-3.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-4.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-5.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-6.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-7.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-8.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-9.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-10.c.pramod-thangali.internal	64%	Starting History Server
ambari200-11.c.pramod-thangali.internal	59%	Waiting to start Metrics Collector
ambari200-12.c.pramod-thangali.internal	74%	Preparing to start RegionServer
ambari200-13.c.pramod-thangali.internal	74%	Preparing to start RegionServer

## 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

The screenshot displays the HDP Configuration Management interface. At the top, there are tabs for 'Summary', 'Heatmaps', 'Configs', and 'Quick Links'. Below this, a 'Group' dropdown is set to 'YARN Default (1)' with a 'Manage Config Groups' link. A horizontal scrollable list shows configuration versions V7 through V10, each with a version number, user 'admin', and timestamp. Version V10 is highlighted with a green checkmark. Below the list, a dark bar shows the current version V10 and the author 'admin' with a timestamp 'Tue, Jul 14, 2015 19:37'. Underneath, there are 'Settings' and 'Advanced' tabs. The main content area is divided into two sections: 'Memory' and 'Container'. The 'Memory' section has a 'Node' sub-section with a slider for 'Memory allocated for all YARN containers on a node' set to 1024MB. The 'Container' section has two sliders: 'Minimum Container Size (Memory)' set to 170MB and 'Maximum Container Size (Memory)' set to 1024MB.

## 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

- ✓ HDFS
- ✓ MapReduce2
- ✓ YARN
- Tez
- Hive**
- Pig
- Sqoop
- ✓ ZooKeeper
- ✓ Ambari Metrics

Actions ▾

Summary **Configs**

Group **Hive Default (5)** Manage Config Groups

◀ ▶

V1 admin  
4 hours ago  
HDP-2.3

✓

↻ V1 ✓ admin authored on Fri, May 29, 2015 05:52

Discard Save

Settings **Advanced**

### ACID Transactions

ACID Transactions

Off

Run Compactor

False

Number of threads used by Compactor



### Interactive Query

Default query queues

default queue ▾

Start Tez session at Initialization

False

Session per queue



Max idle tez session length

◀ 0 ▶ ◀ 10 ▶

### Security

Choose Authorization

None ▾

Run as end user instead of Hive user

True

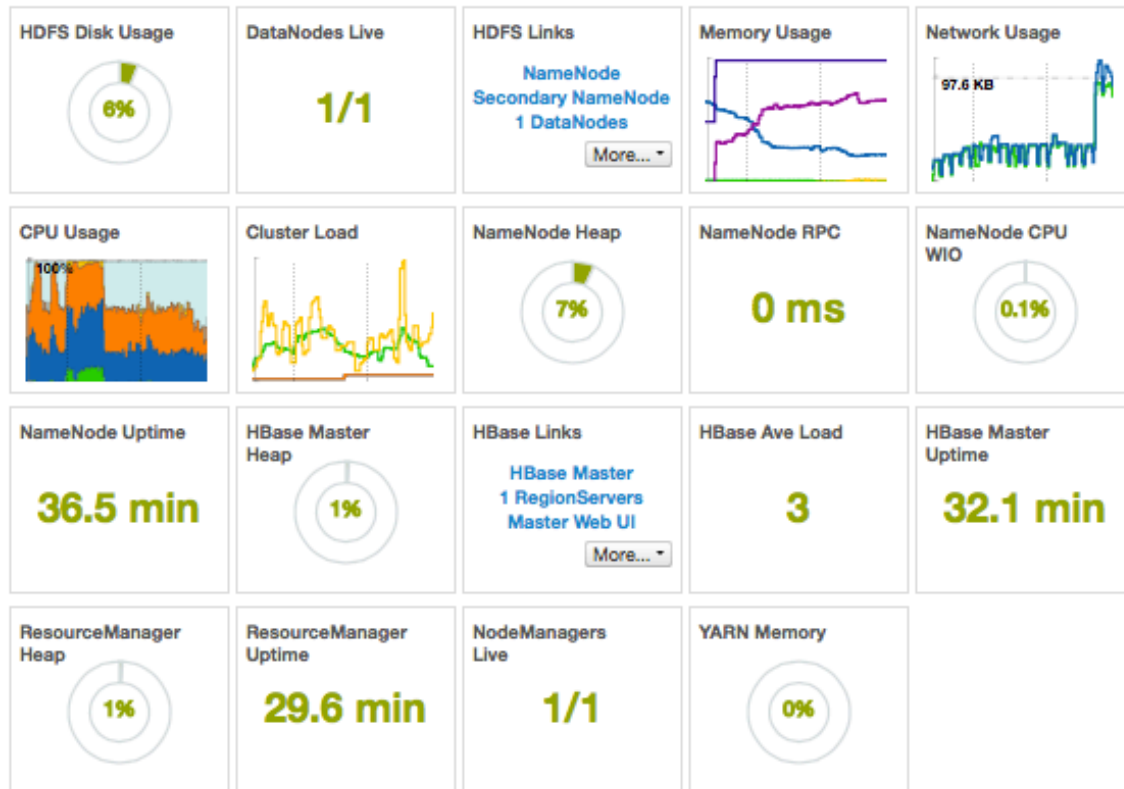
HiveServer2 Authentication

None ▾

Use SSL

False

# 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

New user interface enables fast & easy SQL definition and execution.

**Database Explorer**

consumption

Search tables...

Databases

- consumption
  - power
  - power2
    - adate STRING
    - atime STRING
    - global\_active\_power DOUBLE
    - voltage DOUBLE
    - global\_intensity DOUBLE
    - sub\_metering\_1 DOUBLE
    - sub\_metering\_2 DOUBLE
    - sub\_metering\_3 DOUBLE
  - power3
  - power4
  - sample\_03
  - sample\_04
  - default

**Query Editor**

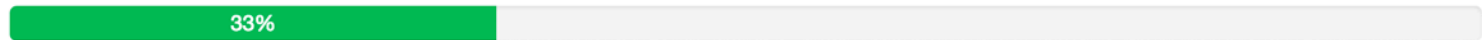
Worksheet

```

1 insert into table power4
2 select adate, sum(p.Global_active_power)
3 from power p
4 join power2 p2
5 on p.adate=p2.adate
6 group by p.adate;

```

Execute Explain Save as... New Worksheet



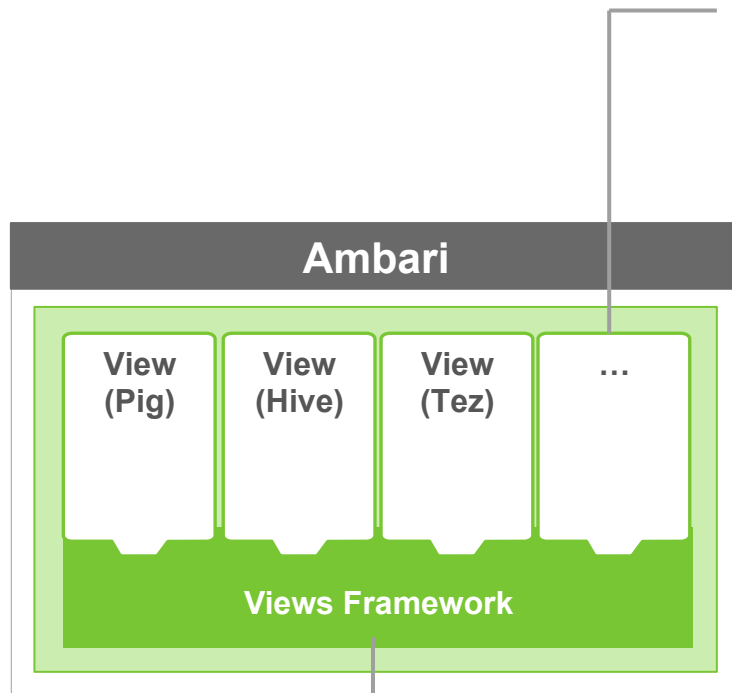
**Query Process Results (Status: RUNNING)**

Logs Results

INFO : Tez session hasn't been created yet. Opening session  
 INFO :  
 INFO : Status: Running (Executing on YARN cluster with App id application\_1432908792576\_0002)  
 INFO : Map 1: -/- Map 3: -/- Reducer 2: 0/4  
 INFO : Map 1: 0/1 Map 3: 0/1 Reducer 2: 0/4  
 INFO : Map 1: 0/1 Map 3: 0/1 Reducer 2: 0/4

TEZ

# Ambari Views Gallery



Built by Hortonworks,  
Community and Partners

Core to Ambari with  
Ambari 1.7+

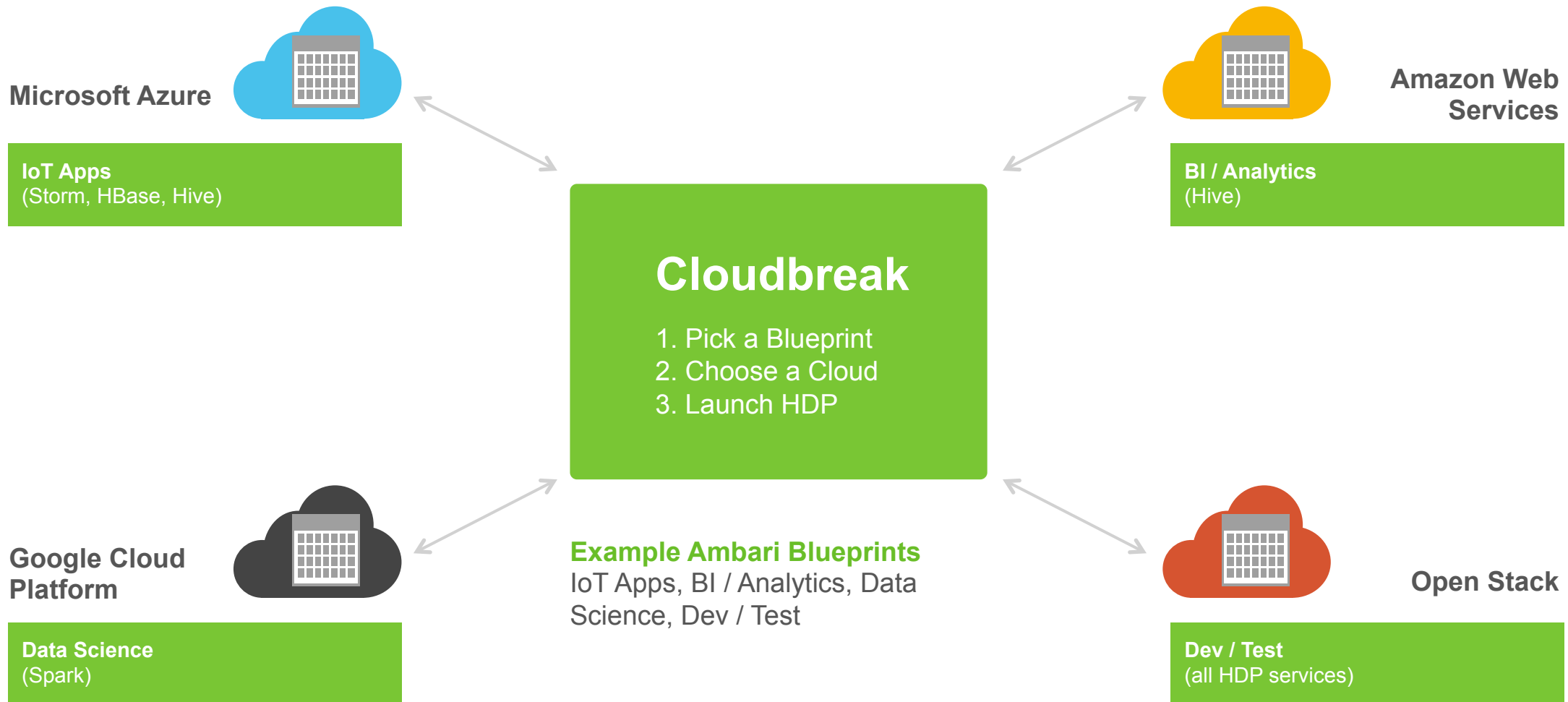
Find Community Views:

The screenshot shows the Hortonworks Gallery website. At the top is the Hortonworks logo. Below it is an orange banner with the text 'Hortonworks Gallery. By Developers. For Developers'. The main content area is titled 'Key repos for Hadoop® development' and includes a paragraph of introductory text. Below this is a navigation bar with tabs for 'FEATURED', 'Ambari Extensions', 'Ambari Views', 'Tutorials', 'Utilities', and 'Sample Apps'. The 'Ambari Views' tab is selected. Below the navigation bar, there are three repository cards. Each card has a title, a description, and a 'Repository' button. The first card is 'AMBARI VIEWS, DEVELOPER, FEATURED' with the description 'Ambari API Explorer An Ambari View to explore its REST API'. The second card is 'AMBARI VIEWS, ADMIN, FEATURED' with the description 'Ambari Store View An Ambari View that provides access to a...'. The third card is 'AMBARI VIEWS, VISUALIZATION, FEATURED' with the description 'HBase Metrics View View to graphically monitor Hbase metrics'. A green button with the text 'developer.hortonworks.com' is located below the navigation bar.

# Cloudbreak

Quickly Launch HDP in the Cloud

# Launch on Any Major Cloud Platform with Blueprints





# 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

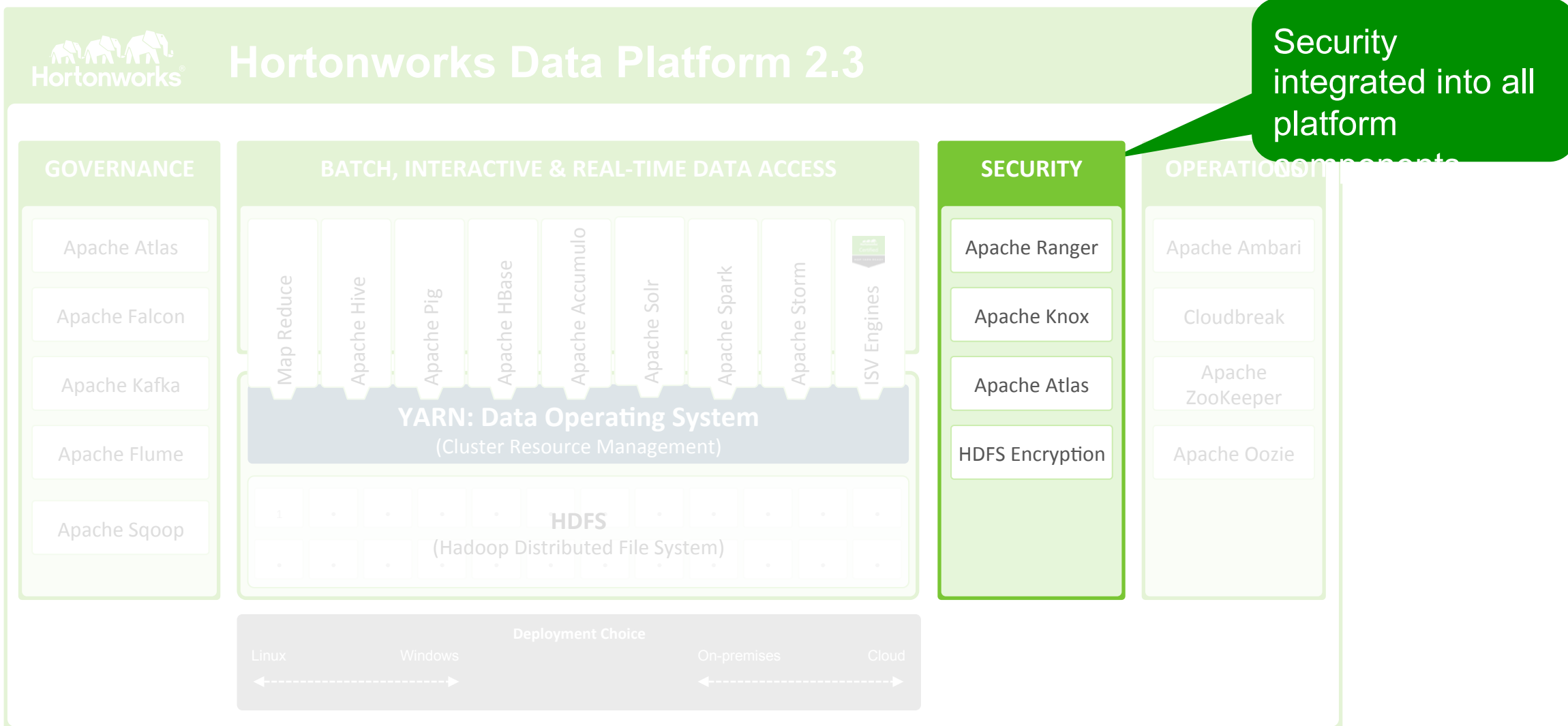
APACHE RANGER & APACHE ATLAS

## Data Protection

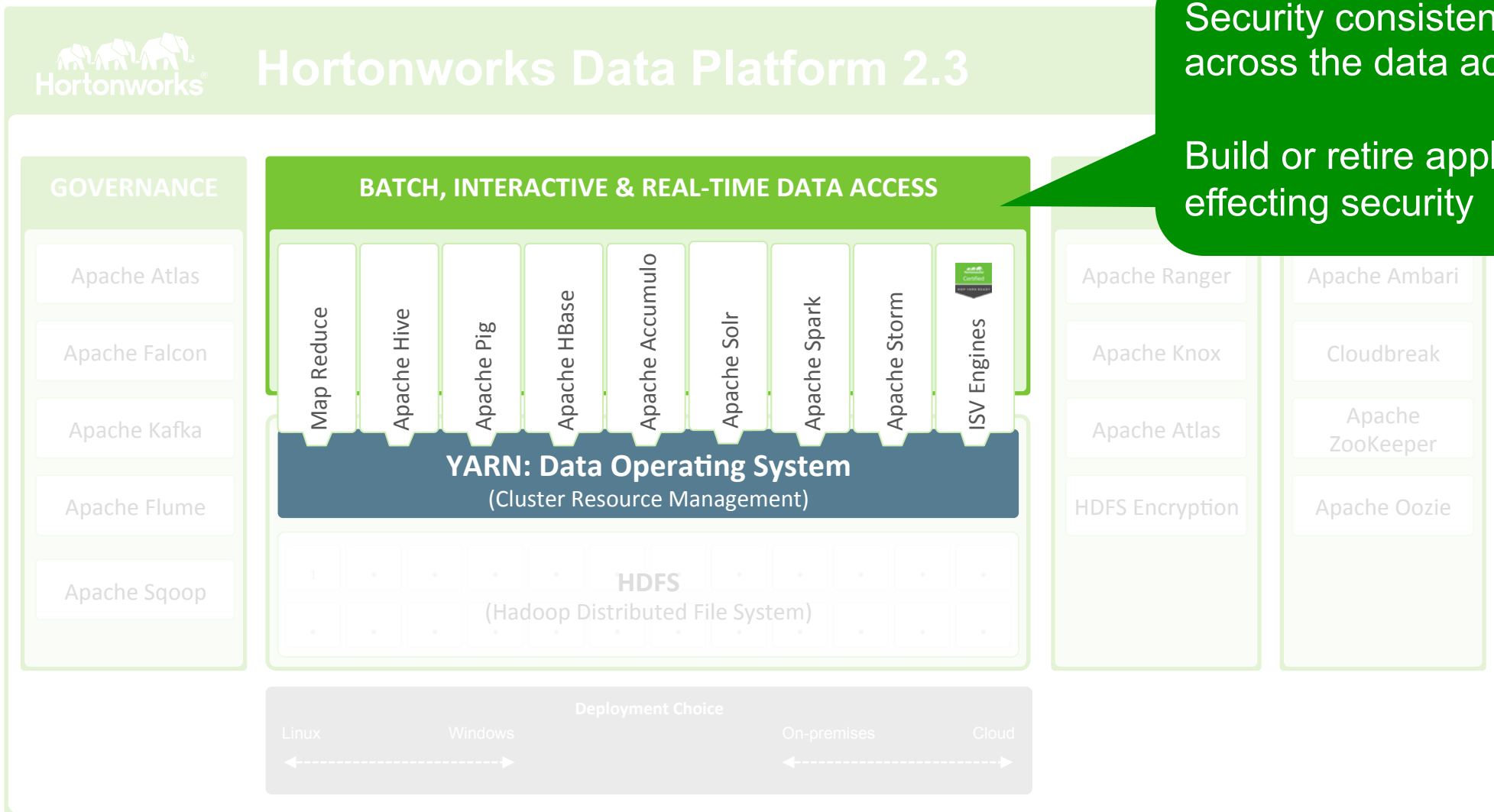
Safeguard data at rest and in motion

HDFS TDE with RANGER KMS

# Integrated Platform Security



# 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

## Service Manager

## Service Manager

## HDFS +

Hadoop\_Prod



Hadoop\_Dev



## HBASE +

HBase\_Prod



HBase\_Dev



## HIVE +

Hive\_Prod



Hive\_Dev



## YARN +

Yarn\_Prod



## KNOX +

Knox\_Prod



## STORM +

Storm\_Dev



## SOLR +

Solr\_Dev



## KAFKA +

Kafka\_Dev





## Edit Policy

## Policy Details :

Policy ID **18**

Policy Name \*

Call\_Details\_Table

enabled

Hive Database \*

xademo

include

table

x call\_detail\_records

include

Hive Column \*

x phone\_number

exclude

Description

Audit Logging

YES

## User and Group Permissions :

Permissions

Select Group

Select User

Permissions

Delegate Admin

x developer

Select User

















select

+

x

## List of Policies : sandbox\_hive

[Add New Policy](#)

Policy ID	Policy Name	Status	Audit Logging	Groups	Users	Action
3	sandbox_hive-1-20150529142947	Enabled	Enabled	--	xapolicymgr	 
4	Hive Global Tables Allow	Disabled	Enabled	public	--	 
5	Hive Global UDF Allow	Disabled	Enabled	public	--	 
18	Call_Details_Table	Enabled	Enabled	developer	--	 
19	Customer_Details_Table	Disabled	Enabled	Marketing	--	 
20	Hive Demo Table Loader	Enabled	Enabled	--	hive	 
21	Hive Demo UDF Loader	Enabled	Enabled	--	hive	 
29	admin policy	Enabled	Enabled	--	admin	 

# 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

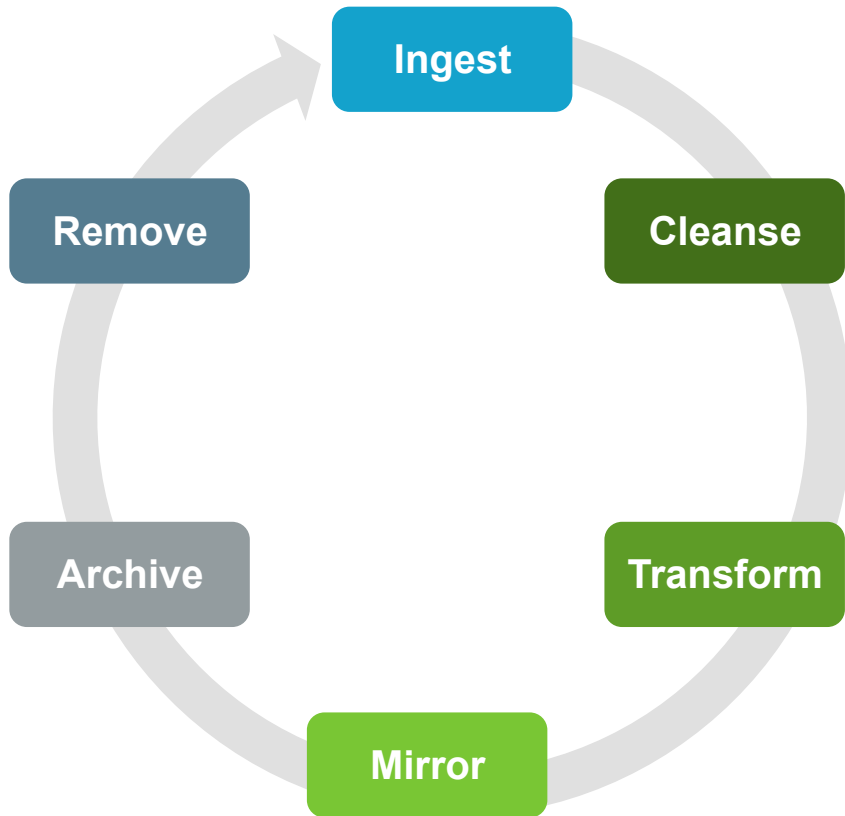
How



# 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

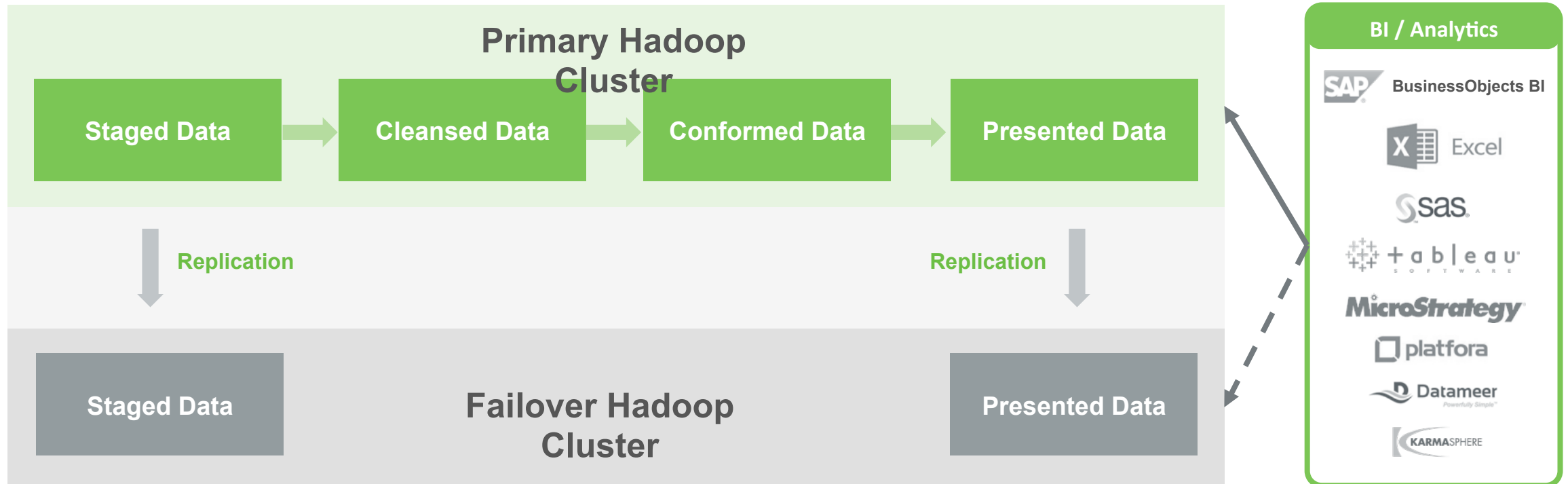
data replication and retention in HDFS and Hive, end-to-end pipeline monitoring

## Audit and Compliance

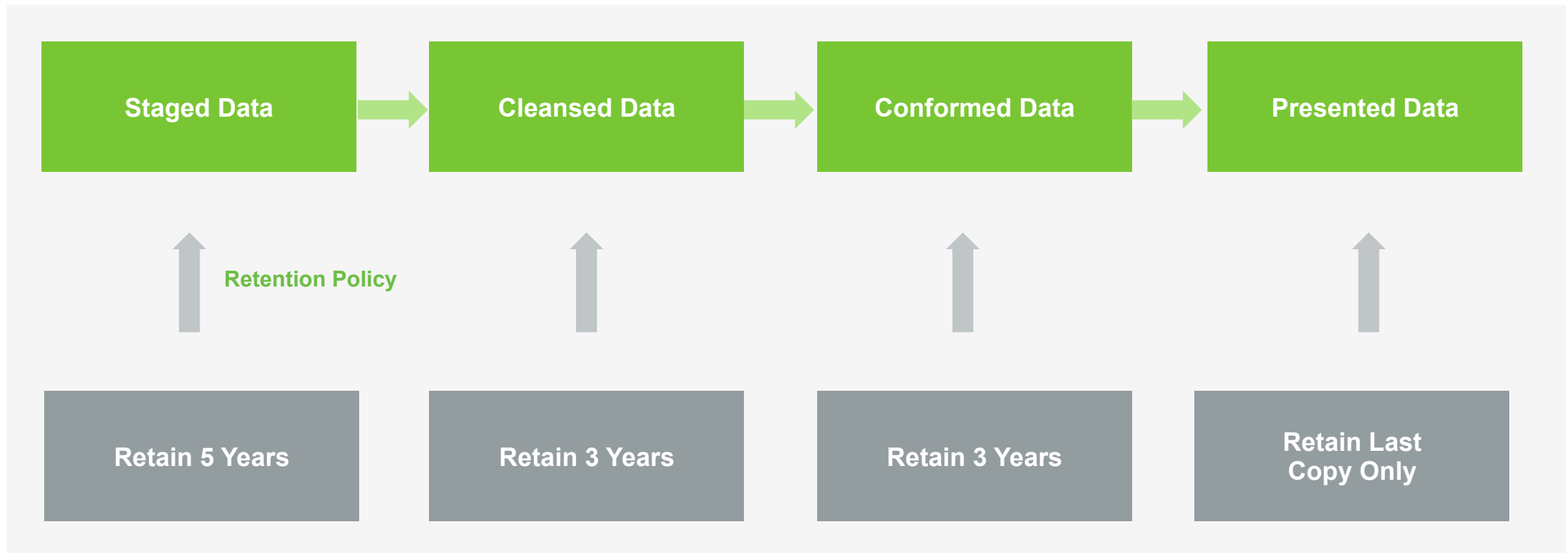
visualize data pipeline lineage, track data pipeline audit logs & free form business labels



# Data Replication with Falcon



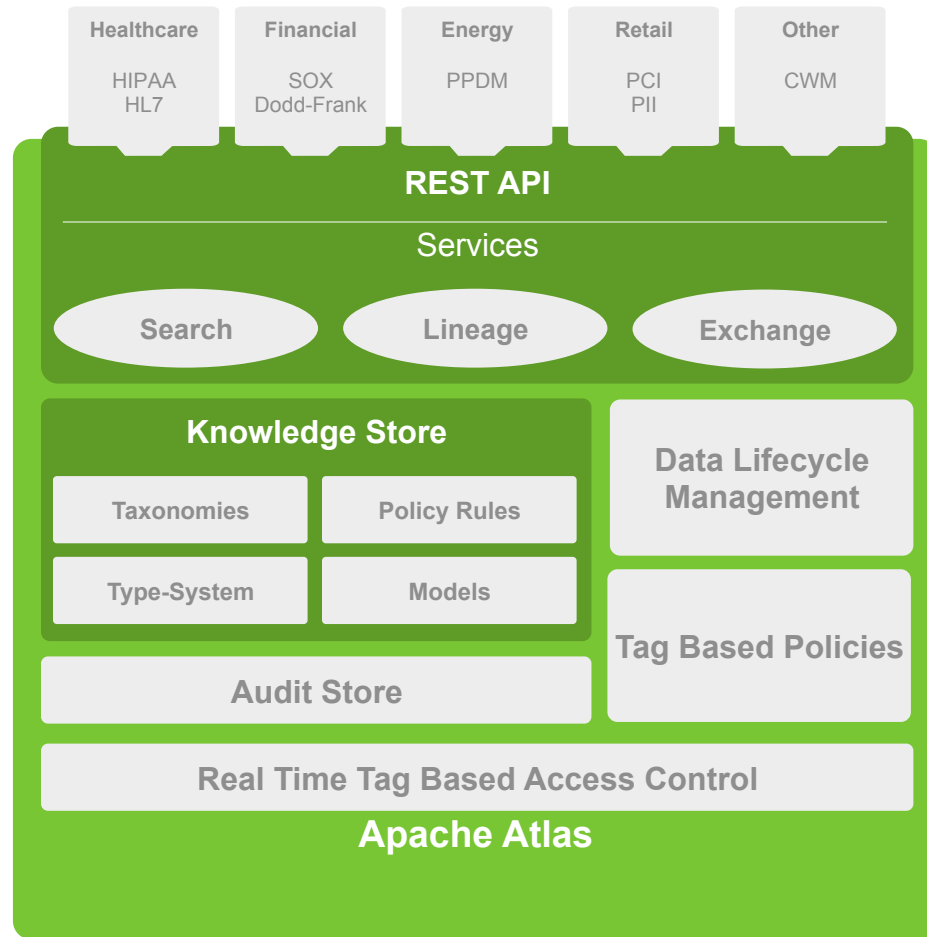
# Data Retention with Falcon



# 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

[Back To Result](#)

## Name: sales\_fact\_monthly\_mv

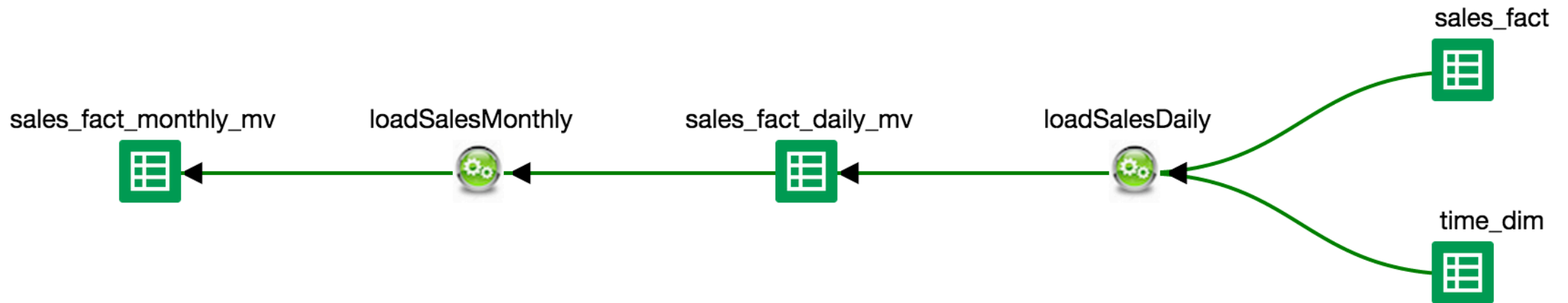
Description: sales fact monthly materialized view

[Details](#)

[Schema](#)

[Output](#)

[Input](#)



# 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.