Hadoop Essentials

A Technical Understanding for Business Users and Decision Makers

Hortonworks University. We Know Hadoop.

Lester Martin
lmartin@hortonworks.com

A Hortonworks University Training Course

Connection before Content

Lester Martin – Hortonworks Professional Services

lmartin@hortonworks.com

http://lester.website

(links to blog, twitter, github, LI, FB, etc)

Hortonworks

Course Objectives

- The case for Hadoop
- What is Hadoop and its ecosystem
- Moving data in and out of Hadoop
- Processing data in Hadoop
- The Hortonworks Data Platform (HDP)
- Preparing your Enterprise for Hadoop
- Key roles in a Hadoop environment

Hortonworks

Course Outline

Unit 1 Understanding Big Data

Unit 2 Understanding Hadoop

Unit 3 Data Integration

Unit 4 The Hadoop Ecosystem

Unit 5 Adoption

Hortonworks

Class Logistics

- Schedule

  9am ~ 4pm
- Restrooms
- Lunch @ Noon-*ish*
- Q & A <u>welcomed early and often</u>

Hortonworks

Unit 1 – Understanding Big Data

Hortonworks

What disrupted the data center?

hadoop ?

Data?

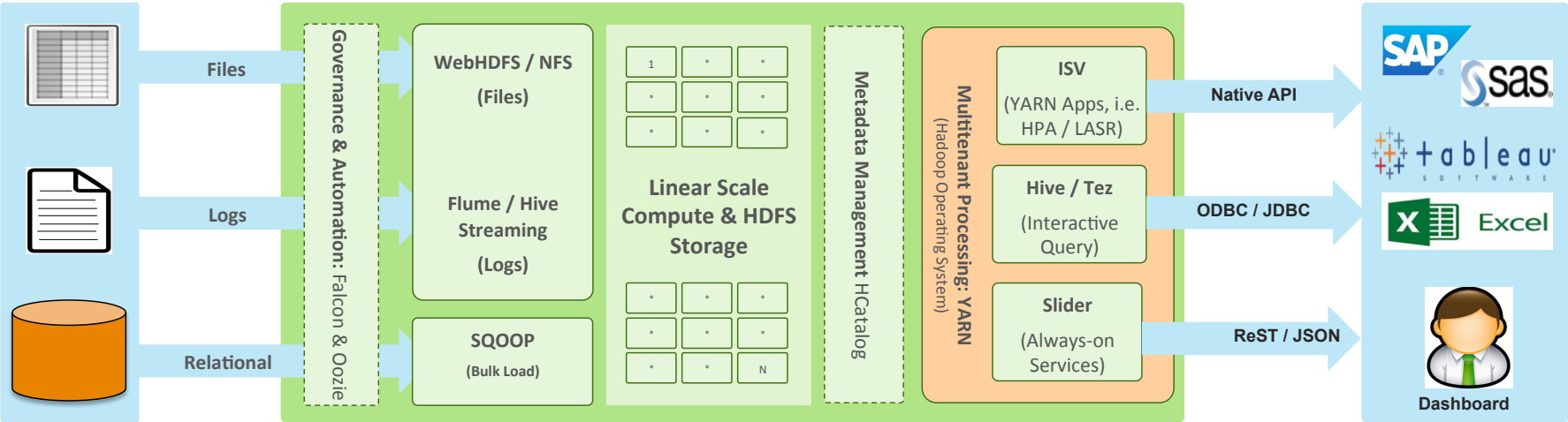
Hortonworks

Hortonworks Data Platform

Let's Demystify!

runs on

Data Lake ELT: Reference Architecture

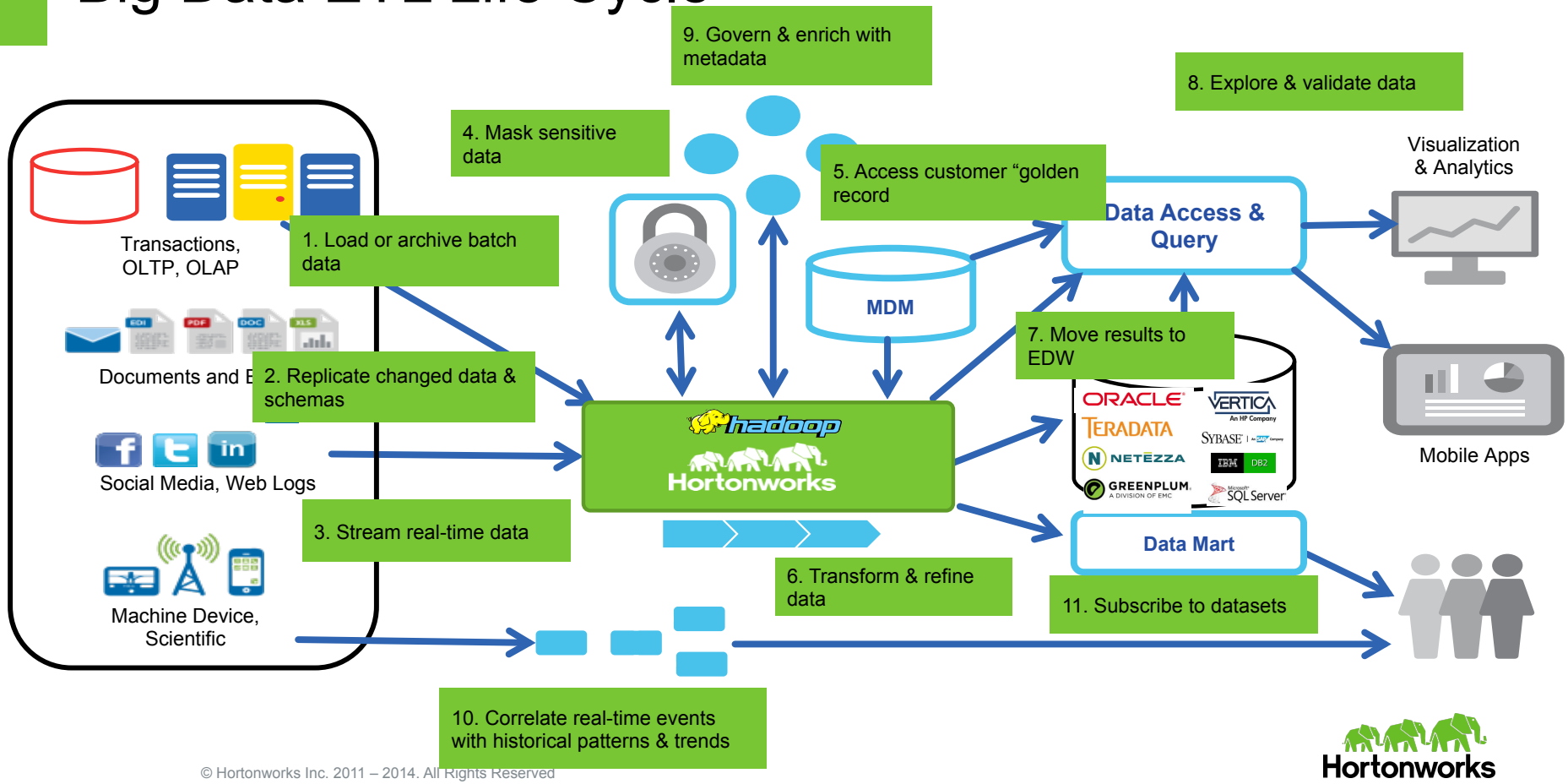


Batch / ELT

- Files are ingested through WebHDFS (Rest API) or NFS directly into HDFS
- Log data is ingested by Flume (w/ Hive Streaming API, Flume can store directly to ORC)
- Relational data is ingested using Sqoop Import (Sqoop can import directly to ORC)
- Native Data Processing within Hadoop can be performed using SQL (Hive) or Scripting (Pig)
- Various 3rd party tools for Data Processing are certified for use with HDP, including Cascading, Talend, Informatica, Syncsort, and others
- Data is available via native API's for certified downstream analytics apps, such as SAS. Hive / Tez provides ad hoc, interactive SQL access to downstream BI / Visualization tools.
- Custom web services, hosted by Apache Slider, can be used to provide data to downstream Web Applications

Hortonworks

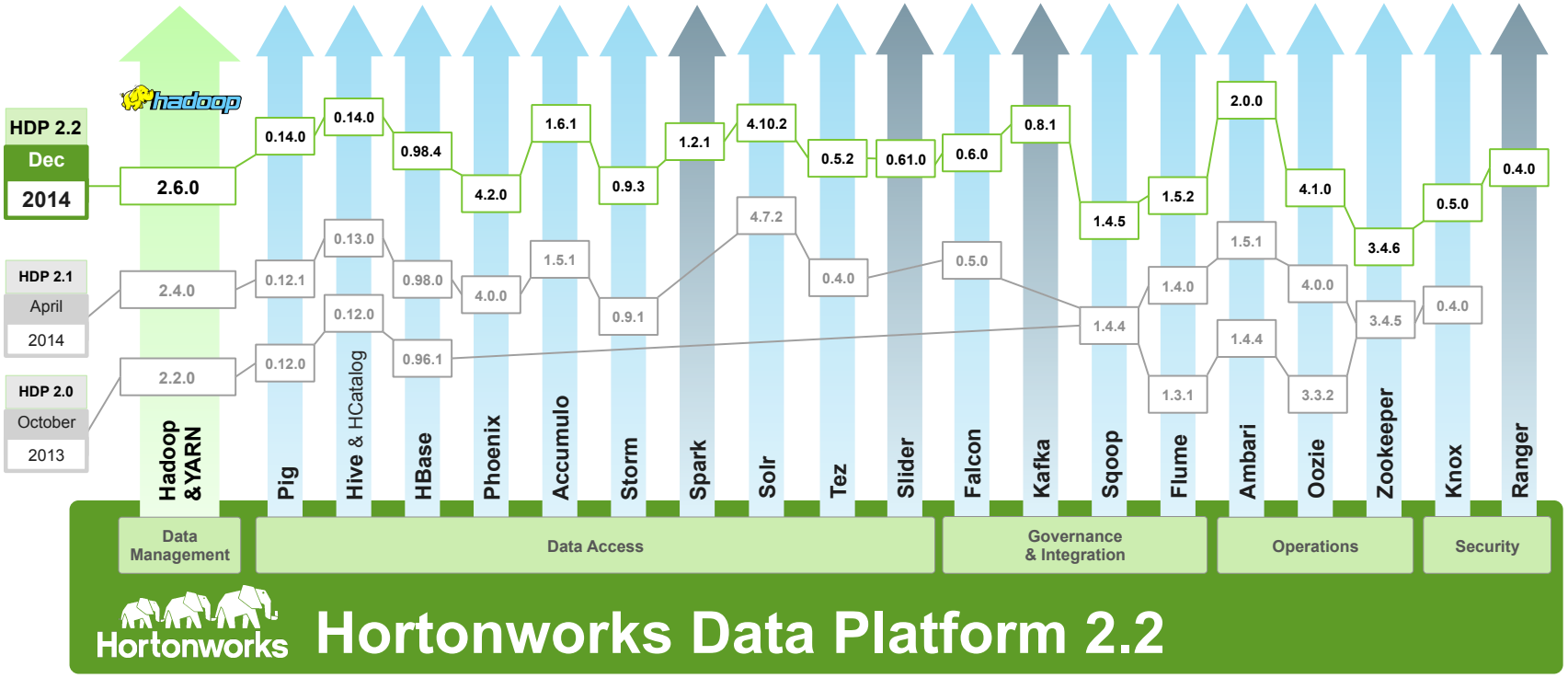
Big Data ETL Life Cycle



Hortonworks

HDP 2.2 IS Apache Hadoop

There is ONE Enterprise Hadoop: everything else is a vendor derivation

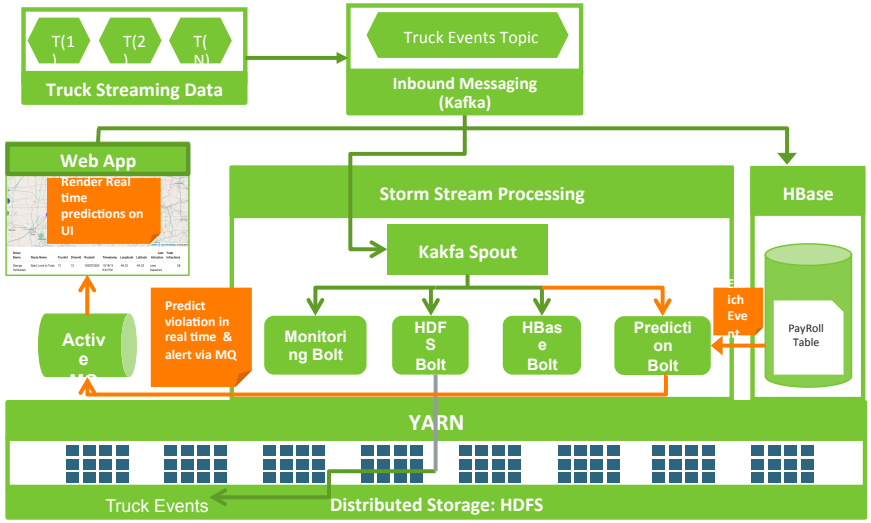


* version numbers are targets and subject to change at time of general availability in accordance with ASF release process

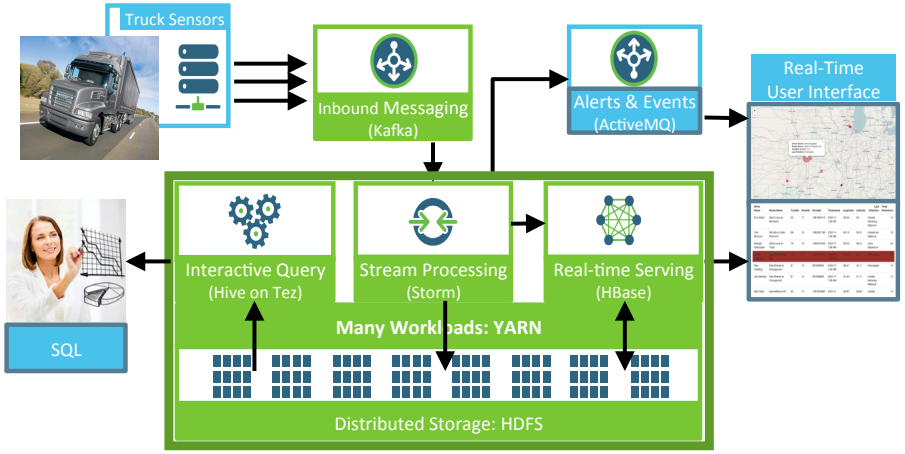
Blueprint (SE Assemblies)

- Set of Hadoop Components to address a particular Use Case

Predictive Analytics Blueprint



IoT Blueprint

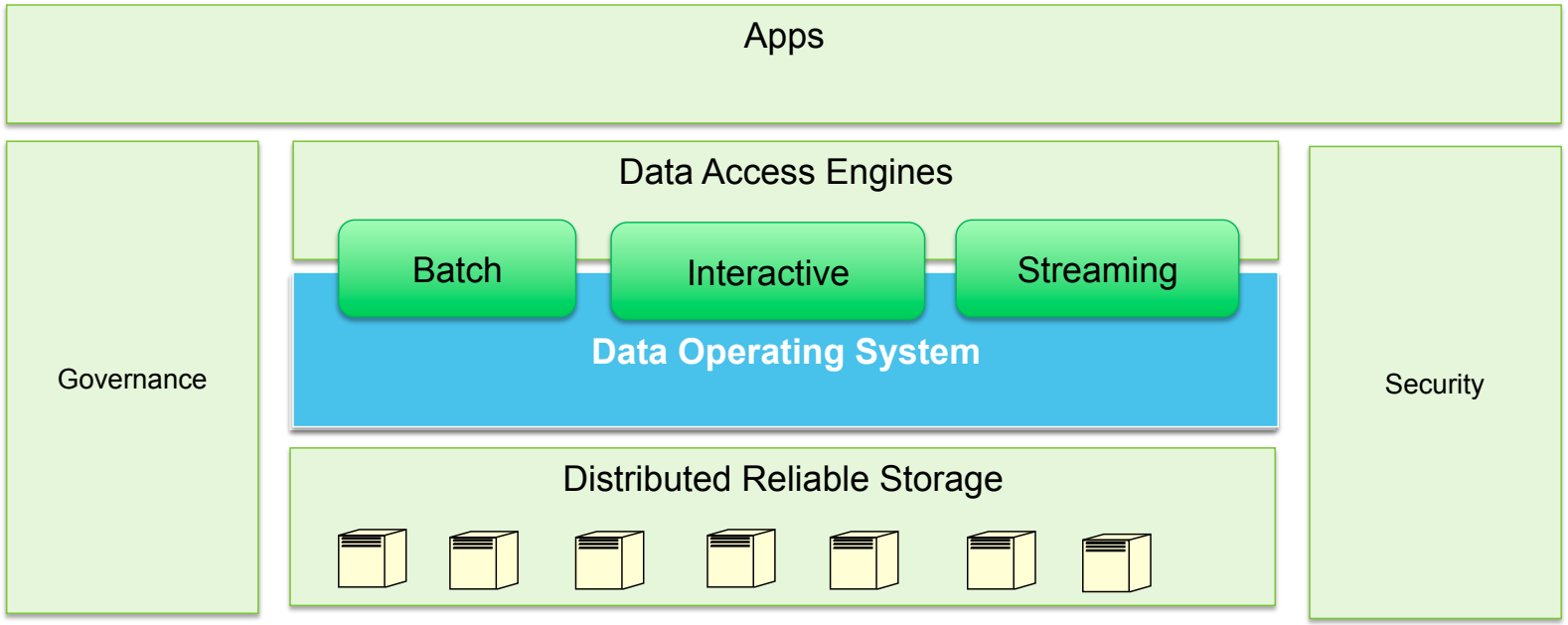


Hortonworks

What is the key behind Hadoop architecture?

Hortonworks

Hadoop Architecture



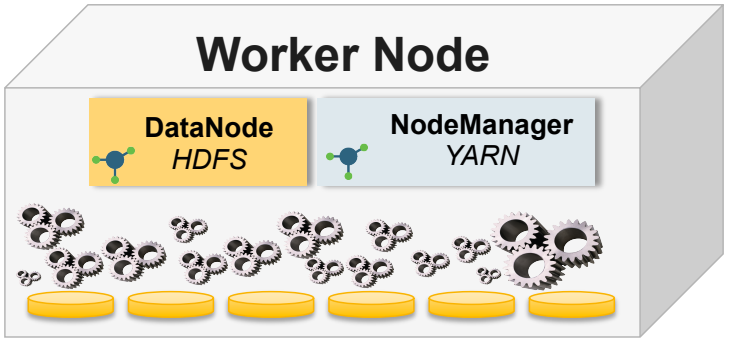
Hortonworks

Unit 2 – Understanding Hadoop

Storage and Processing

Hortonworks

Core hadoop



Disk, CPU, Memory

+

NameNode
HDFS

`/directory/structure/in/memory.txt`

ResourceManager
YARN

Resource management + scheduling

Hadoop daemon

User application

Hortonworks

Joys of Real Hardware (Jeff Dean)

Typical first year for a new cluster:
~0.5 overheating (power down most machines in <5 mins, ~1-2 days to recover)
~1 PDU failure (~500-1000 machines suddenly disappear, ~6 hours to come back)
~1 rack-move (plenty of warning, ~500-1000 machines powered down, ~6 hours)
~1 network rewiring (rolling ~5% of machines down over 2-day span)
~20 rack failures (40-80 machines instantly disappear, 1-6 hours to get back)
~5 racks go wonky (40-80 machines see 50% packetloss)
~8 network maintenances (4 might cause ~30-minute random connectivity losses)
~12 router reloads (takes out DNS and external vips for a couple minutes)
~3 router failures (have to immediately pull traffic for an hour)
~dozens of minor 30-second blips for dns
~1000 individual machine failures
~thousands of hard drive failures
slow disks, bad memory, misconfigured machines, flaky machines, etc

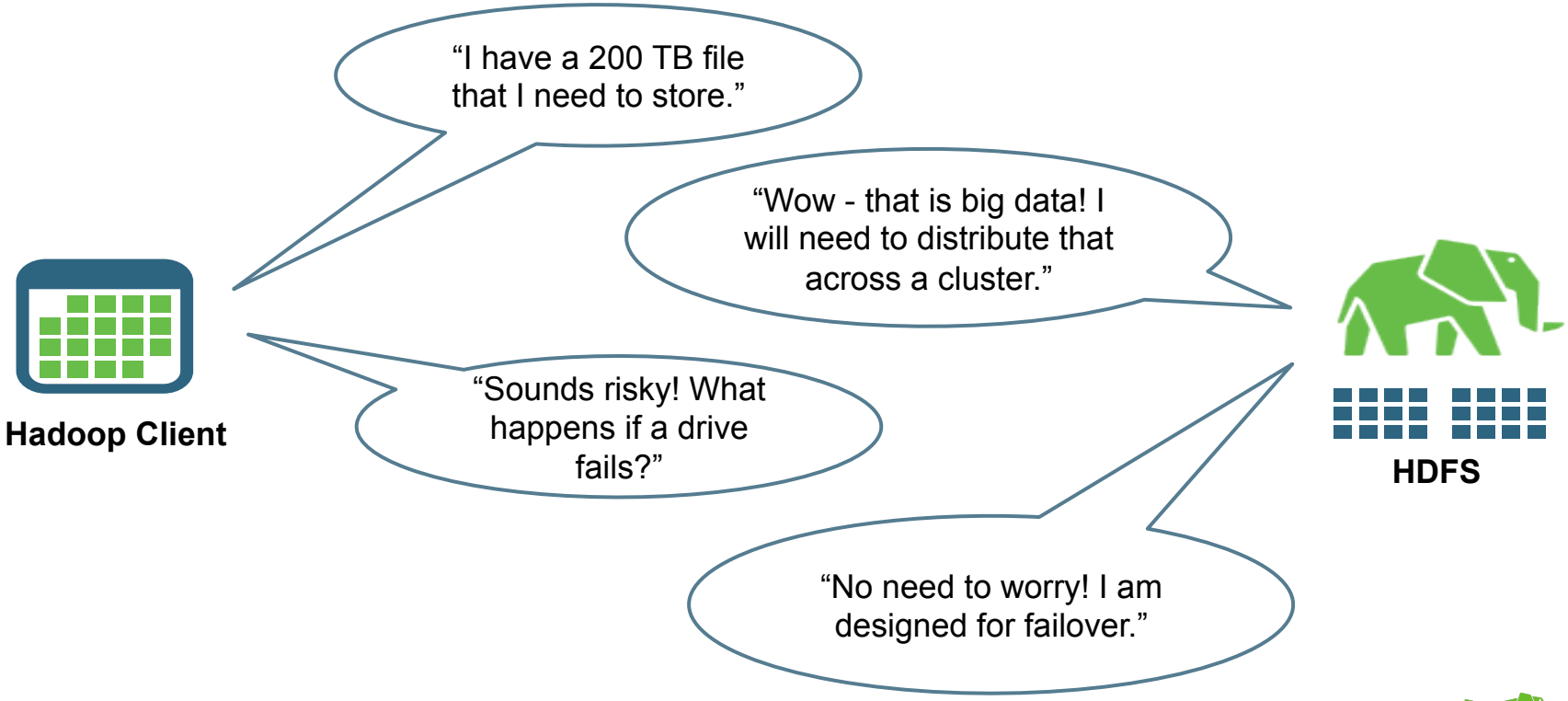
Hortonworks

HDFS

Hadoop Distributed File System

Hortonworks

What is HDFS?

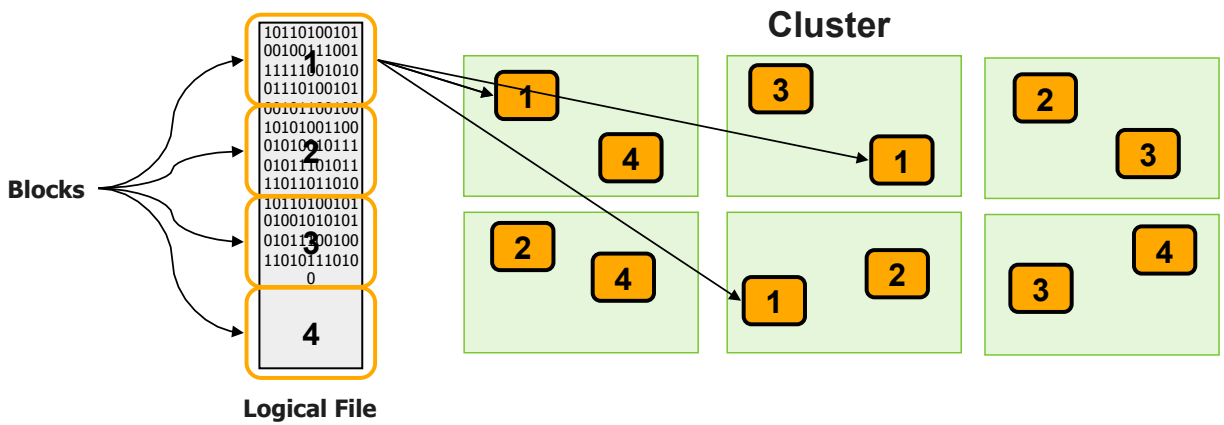


Hortonworks

Hadoop Distributed File System (HDFS)

Key ideas

- Divide files into big blocks and distribute **randomly** across the cluster
- Programs can ask "**where do the pieces of my file live?**"

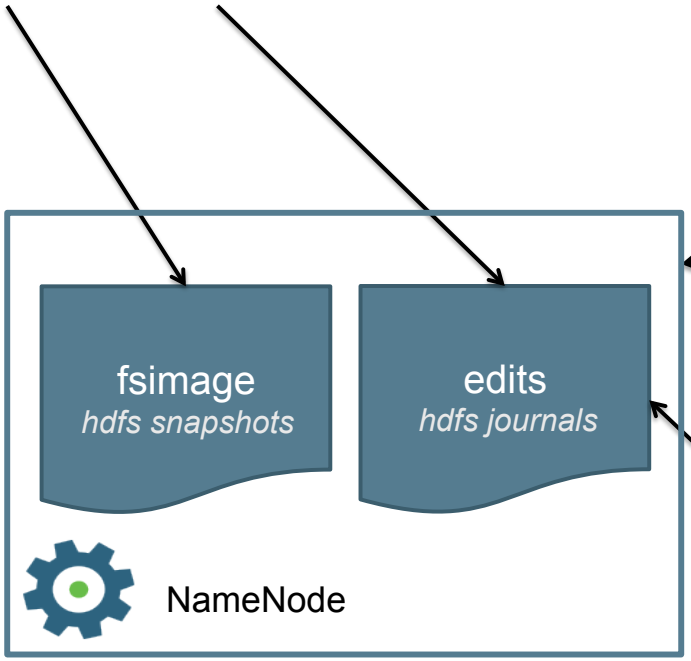


Hortonworks

The NameNode

1. When the NameNode starts, it reads **fsimage** and **edits** files from disk.

2. The transactions in **edits** are merged with **fsimage**, and **edits** is emptied.



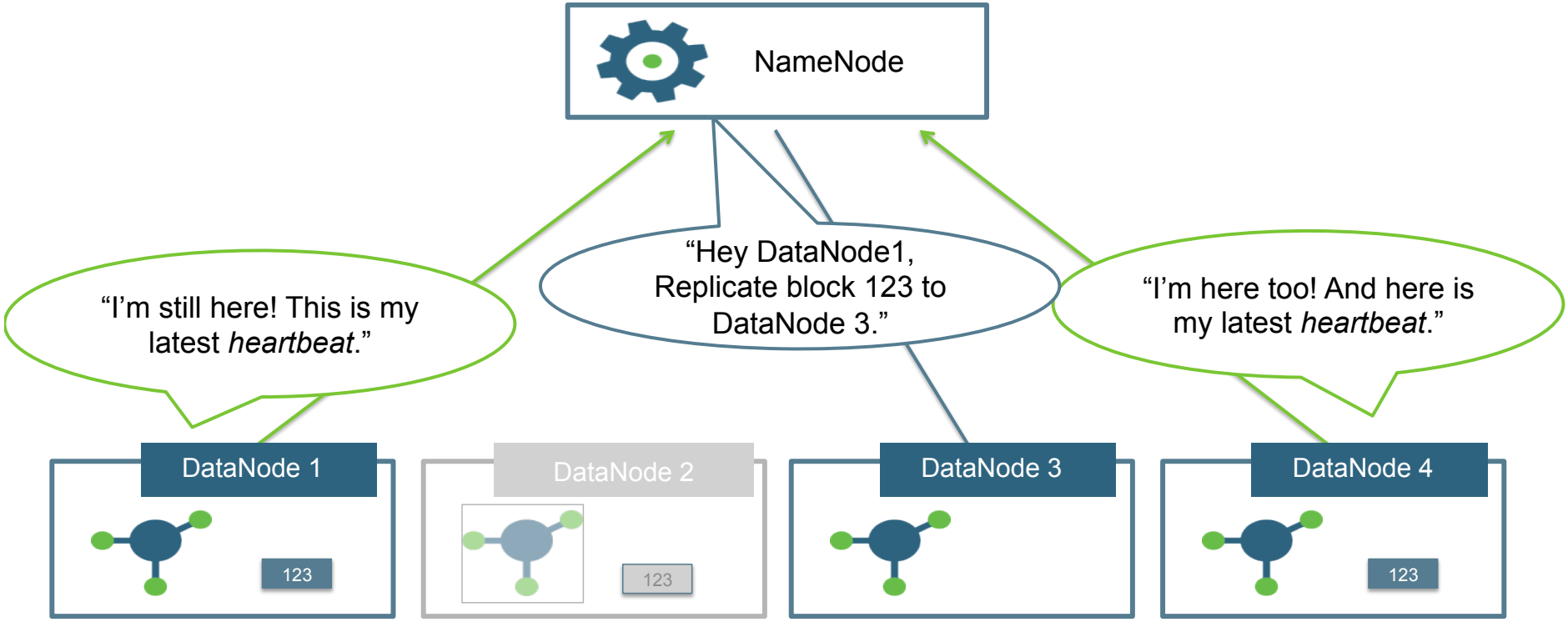
3. A client application creates a new file in HDFS.

```
hadoop fs -put foo.log bar/foo.log
```

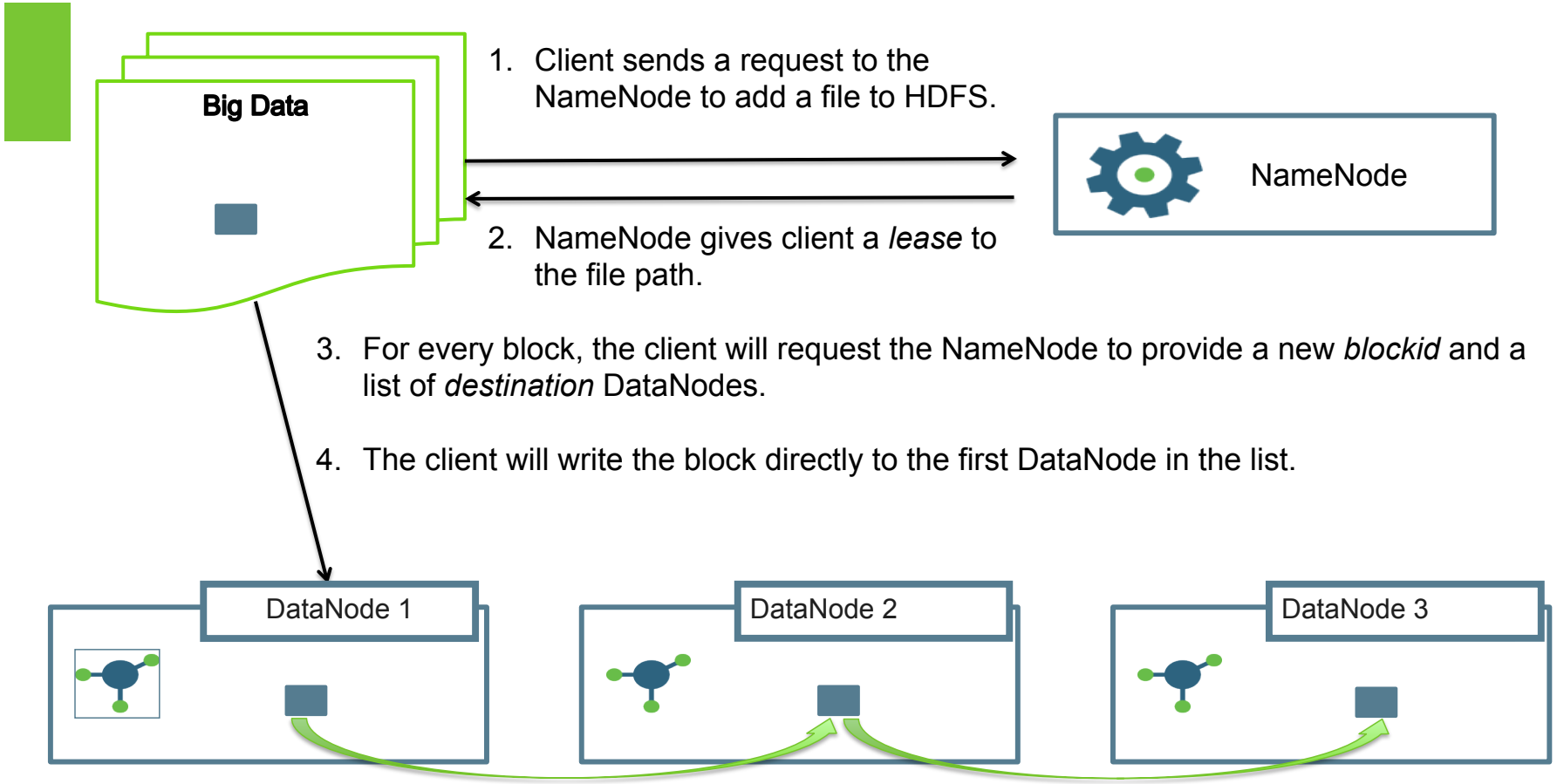
4. The NameNode logs that transaction in the **edits** file.

Hortonworks

The DataNodes



Hortonworks



5. The first DataNode pipelines the replication to the next DataNode in the list.

Hortonworks

Demonstration on HDP Sandbox 2.3

Understanding HDFS via GUIs

Heterogeneous Storage

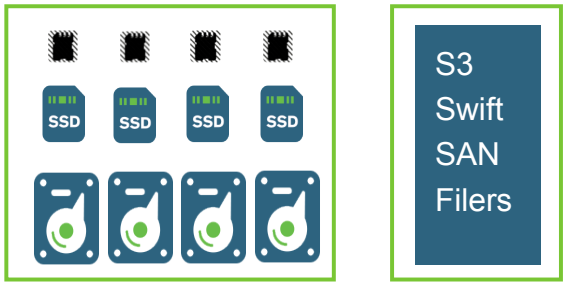
Before

- DataNode is a single storage
- Storage is uniform - Only storage type Disk
- Storage types hidden from the file system

All disks as a single storage

New Architecture

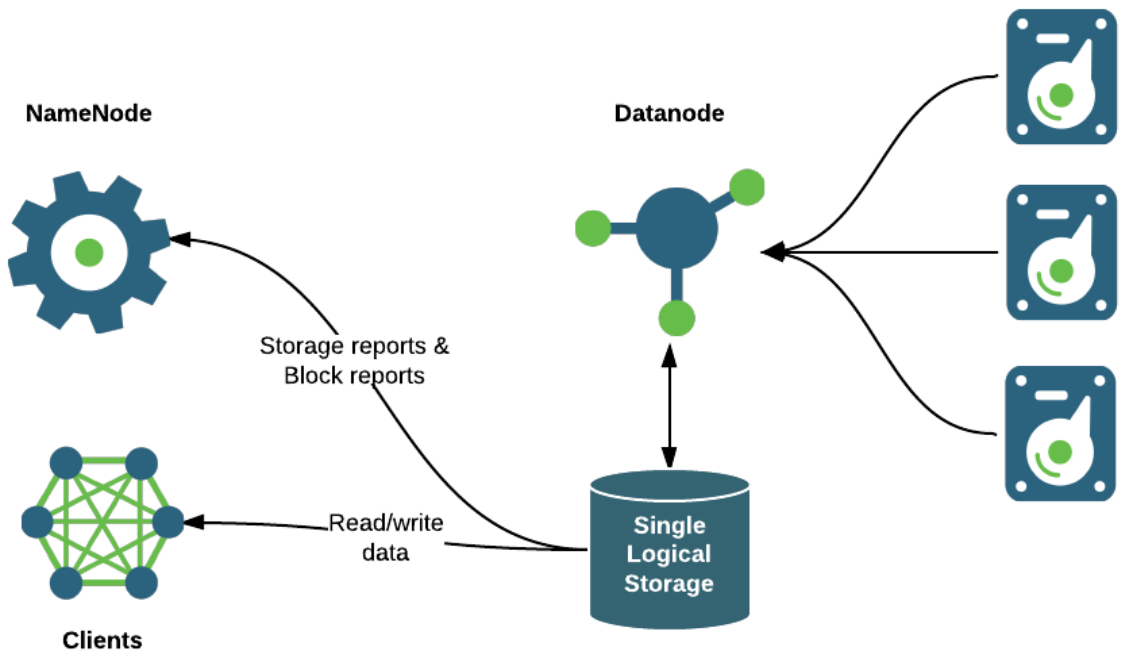
- DataNode is a collection of storages
- Support different types of storages
  - Disk, SSDs, Memory



Collection of tiered storages

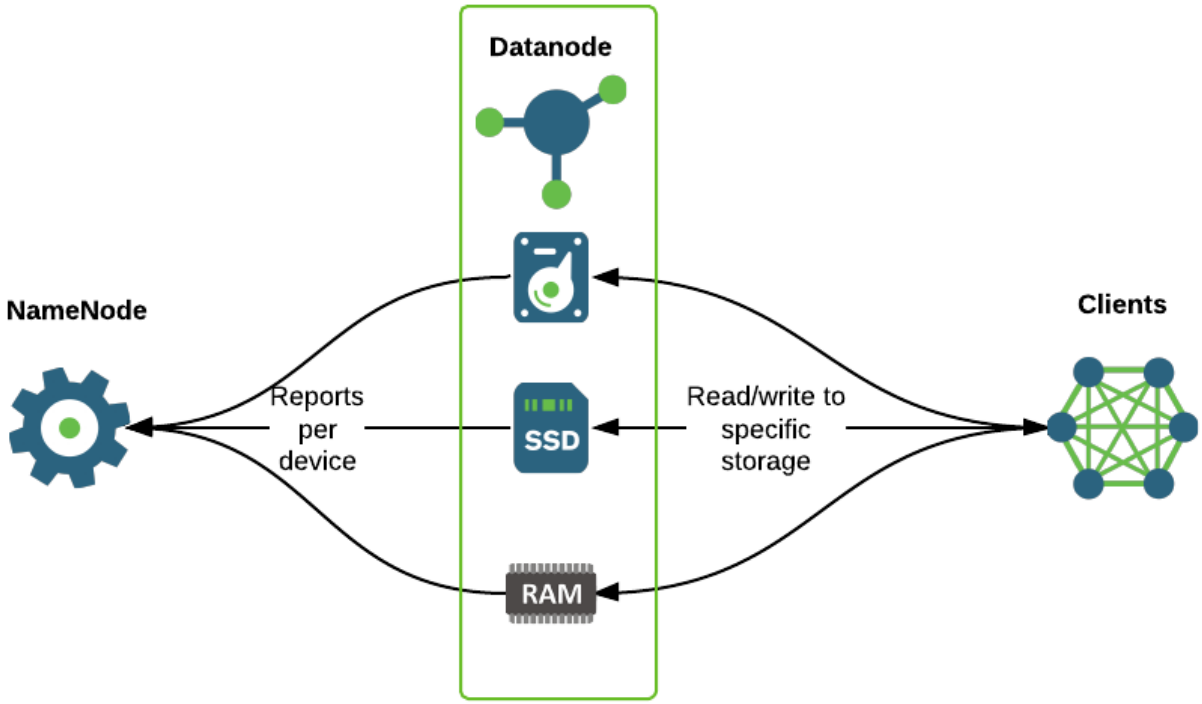
Hortonworks

HDFS Storage Architecture - Before



Hortonworks

HDFS Storage Architecture - Now



Hortonworks

YARN & Multi-Tenancy

Architecture

Hortonworks

What Benefits Does Yarn Provide You..

Better Utilization of existing clusters

- 60% – 150% improvement on node utilization

MR2 on YARN Performance Gains

Enable next-generation Vendor Integration

- YARN is an application framework. (e.g: SAS, R, SAP)

The ability to run Next-generation Workloads

- Native Tez YARN (interactive SQL), Native Storm YARN, Native HBase and Accumulo on YARN

YARN in Production

- Yahoo: ~40,000 nodes, multiple clusters running YARN across over 365PB of data
- Spotify, Progressive, Kohls, UHG, Sprint, JPMC, Target, AIG, Samsung

Hortonworks

Multi-Tenancy Requirements

Multi-Tenancy in one shared cluster

- Multiple Business Units
- Multiple Application Applications/Jobs

Requirements

- Shared Processing Capacity
- Shared Storage Capacity
- Data Access Security

Hortonworks

Concepts

Application

- Application is a *temporal job* or a *service* submitted YARN
- Examples
  - Map Reduce Job (job)
  - Hbase Cluster (service)

Container

- Basic unit of allocation
- Fine-grained resource allocation of resources (RAM, CPU, disk, network, GPU, etc.)
  - container_0 = 2GB, 1CPU
  - container_1 = 1GB, 6 CPU
- Replaces the fixed map/reduce slots

Hortonworks

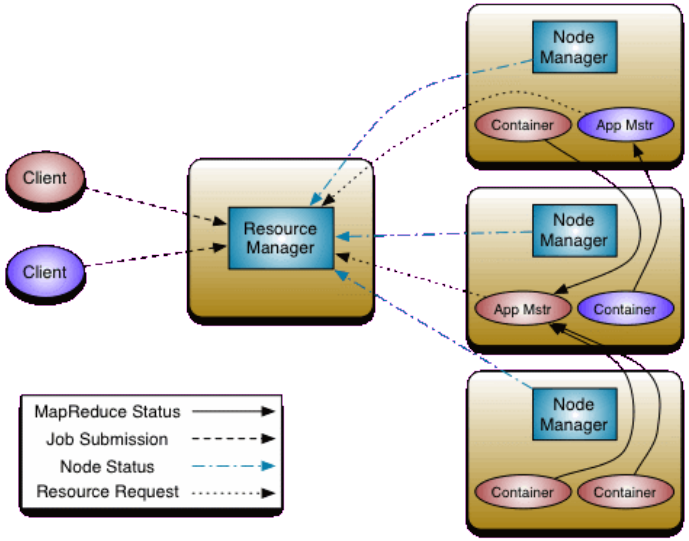
YARN Architectural Components

Resource Manager

- Global resource scheduler
- Hierarchical queues

Node Manager

- Per-machine agent
- Manages the life-cycle of container
- Container resource monitoring

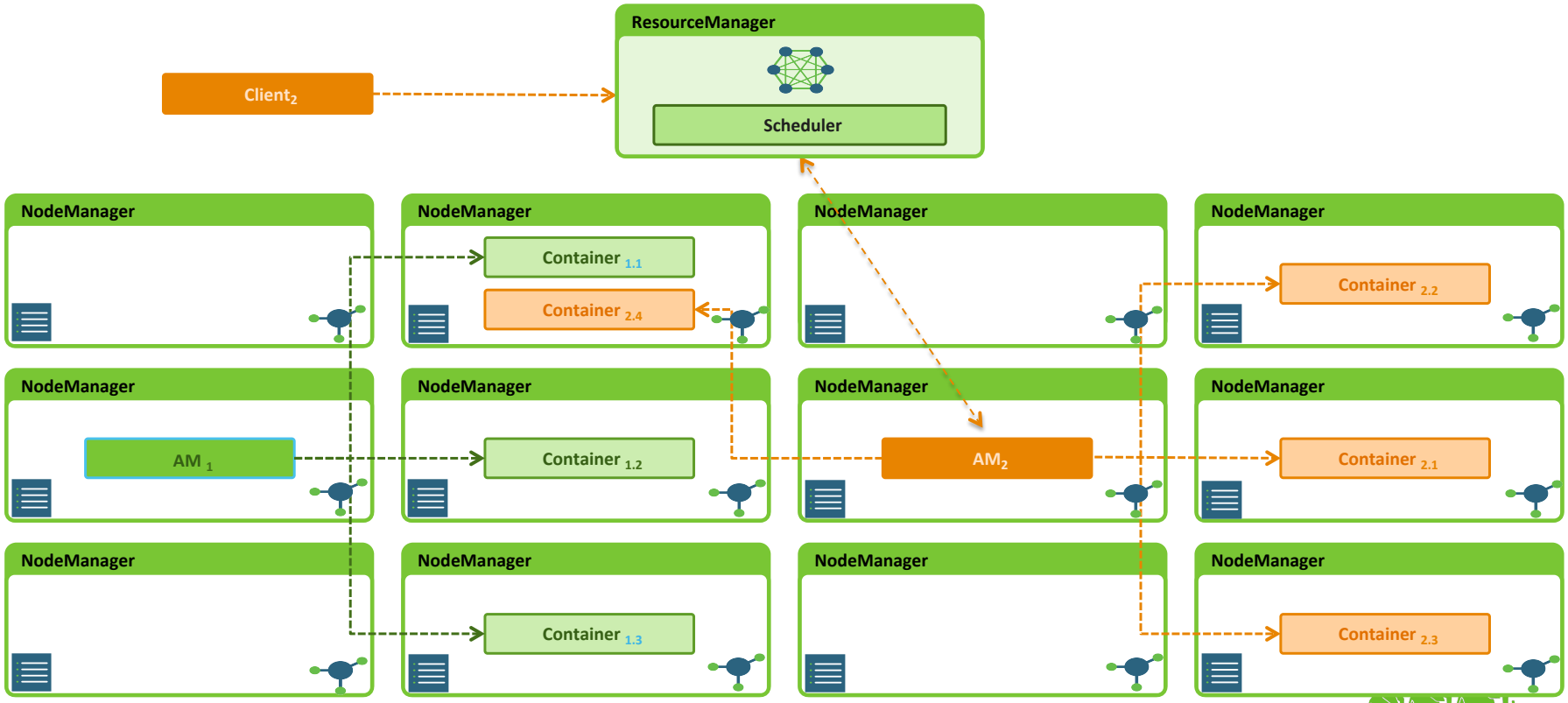


Application Master

- Per-application
- Manages application scheduling and task execution
- E.g. MapReduce Application Master

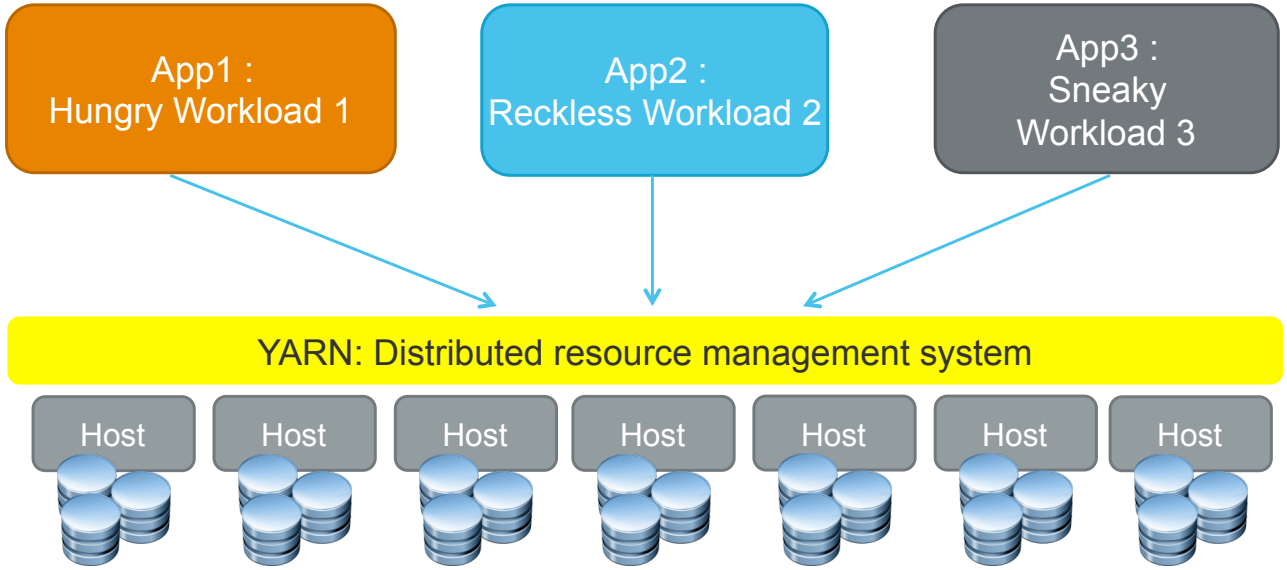
Hortonworks

YARN Architecture - Walkthrough



Hortonworks

Shared Infrastructure Challenges and Solution



YARN containers provides required Isolation,
Resource Management and Security

Hortonworks

Resource Request

- Fine-grained resource *ask* to the ResourceManager
- Ask for a specific amount of resources (memory, CPU, etc.) on a specific machine or rack
- Use special value of * for resource name for *any* machine

ResourceRequest
priority
resourceName
capability
numContainers

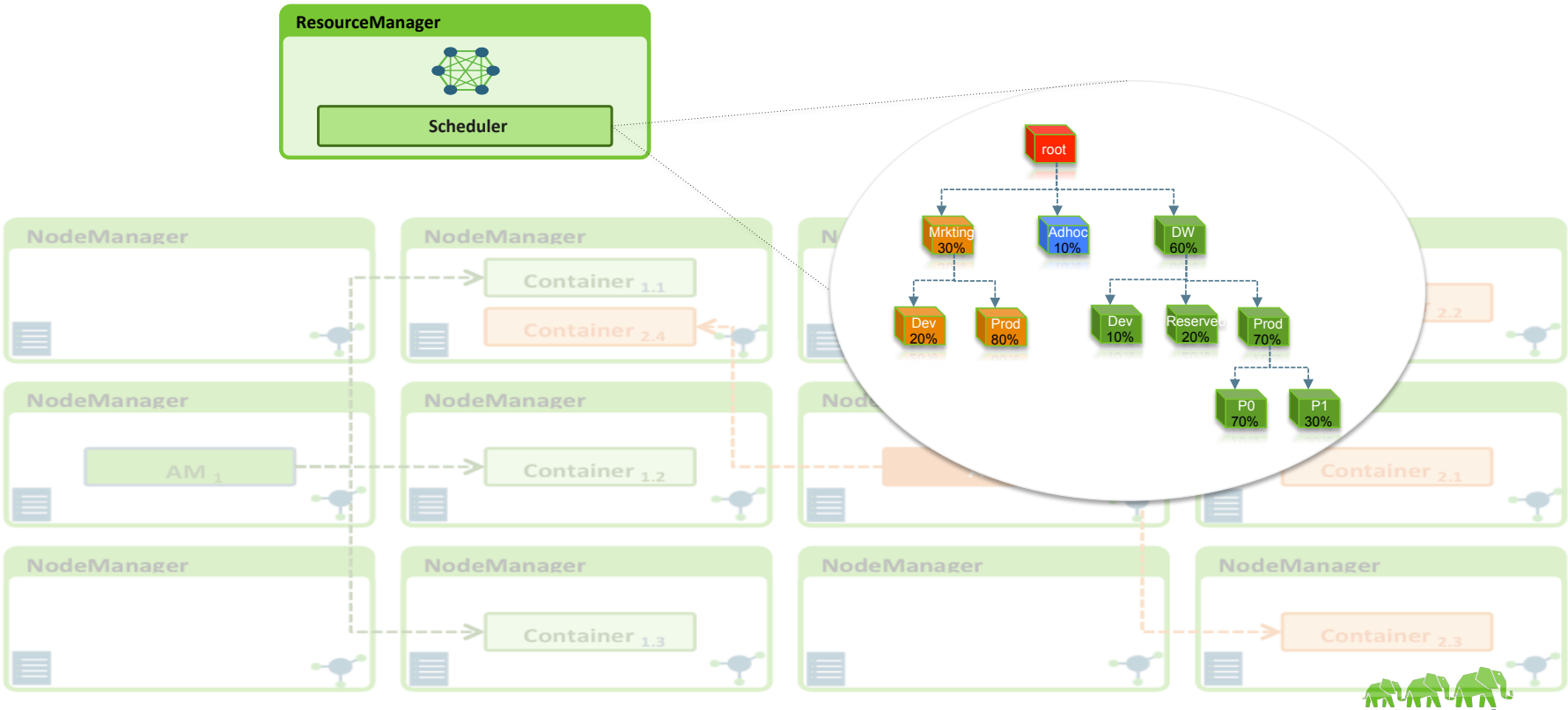
Hortonworks

Application Resource Request

priority	capability	resourceName	numContainers
0	<2gb, 1 core>	host01	1
		rack0	1
		*	1
1	<4gb, 1 core>	*	1

Hortonworks

Getting a Fair Share: Capacity Scheduler



Hortonworks

Multi-Tenancy with Capacity Scheduler

Queues

Economics as *queue-capacity*

- Hierarchical Queues

SLAs

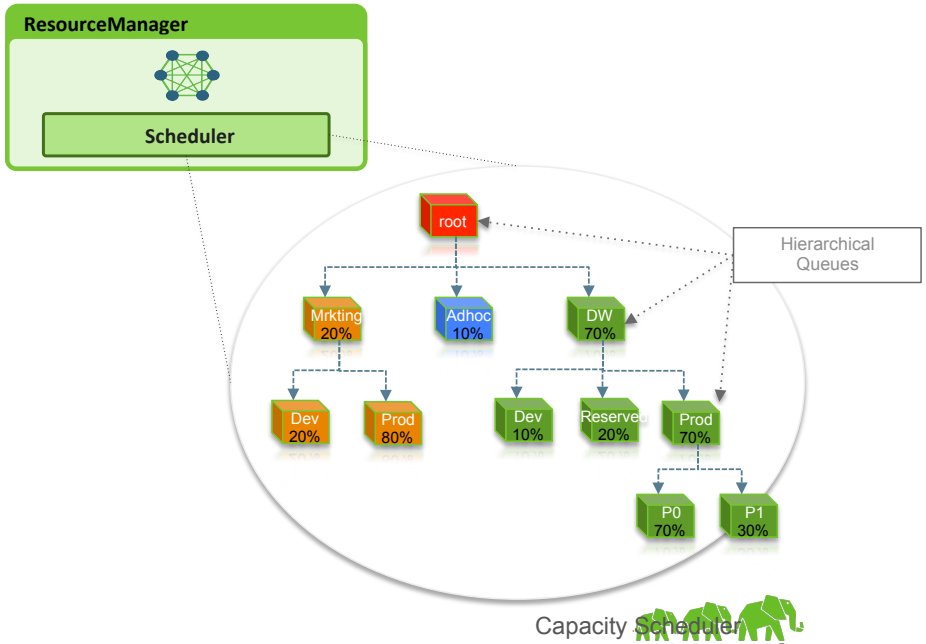
- Preemption

Resource Isolation

- Linux: CGroups
- MS Windows: Job Control

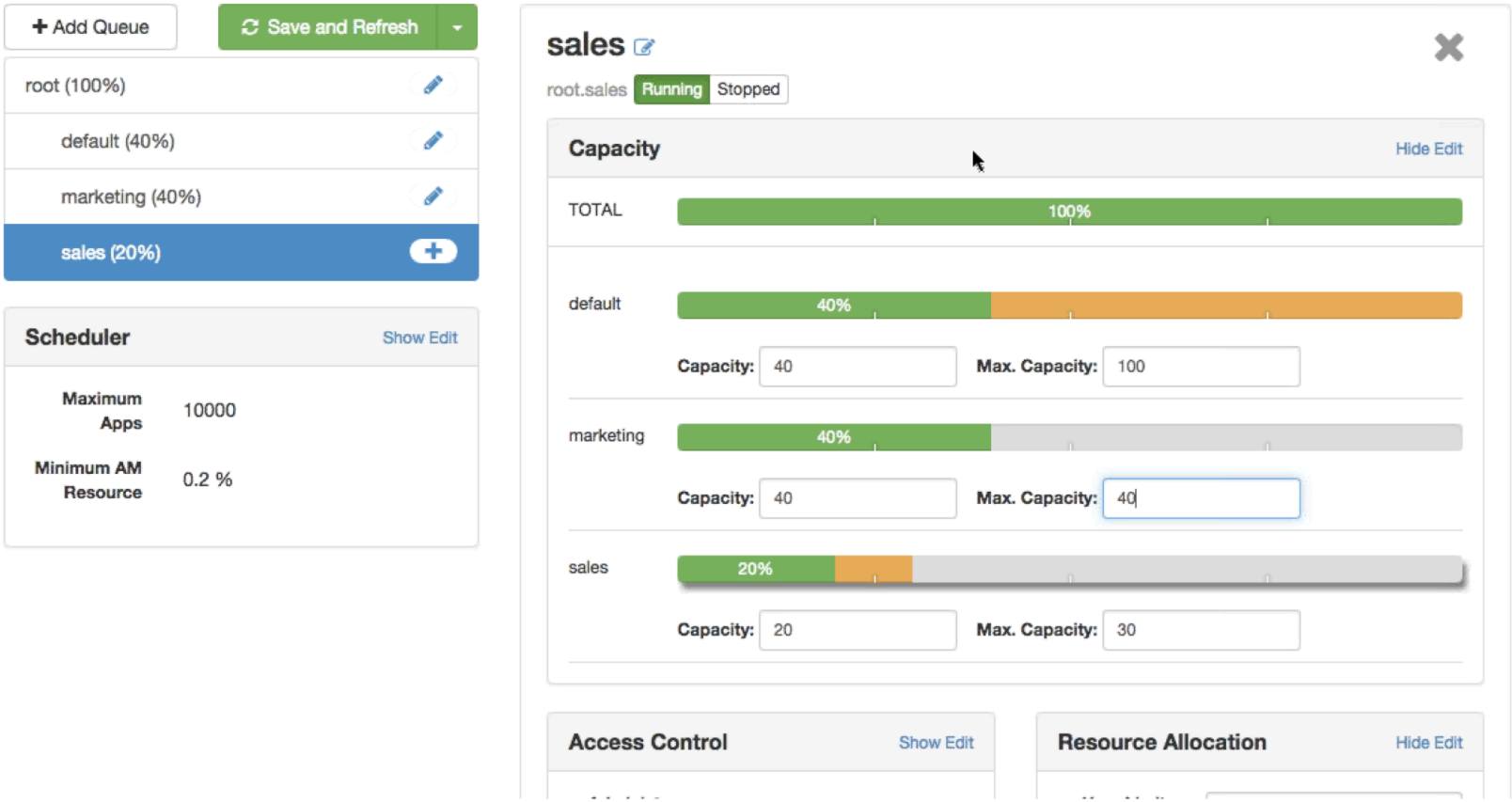
Administration

- Queue ACLs
- Run-time re-configuration for queues
- Charge-back



Capacity Scheduler

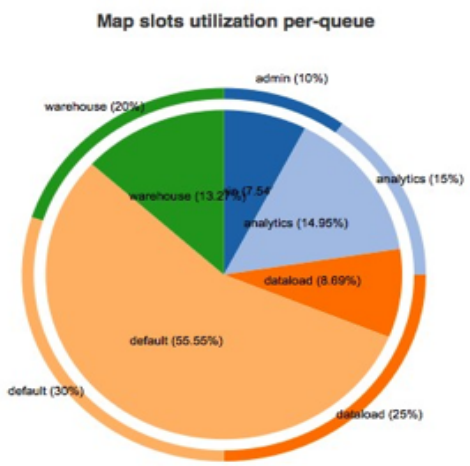
Manage Queue Limits with Ambari

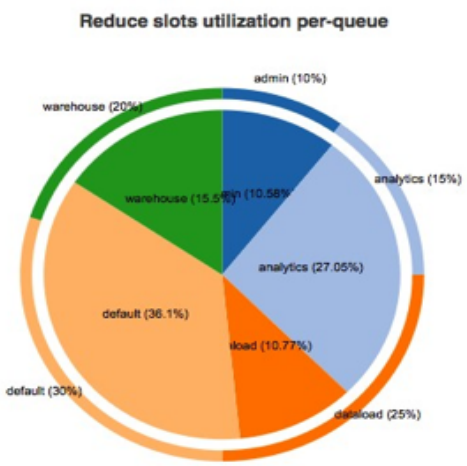


Capacity Scheduler Controls

- **Configure queues and manage queue ACLs**

- **Usage Reporting**
  - Report on queue usage per user
  - Report on various other stats on user jobs
  - Jobs per day, data processed per day, data usage





Hortonworks

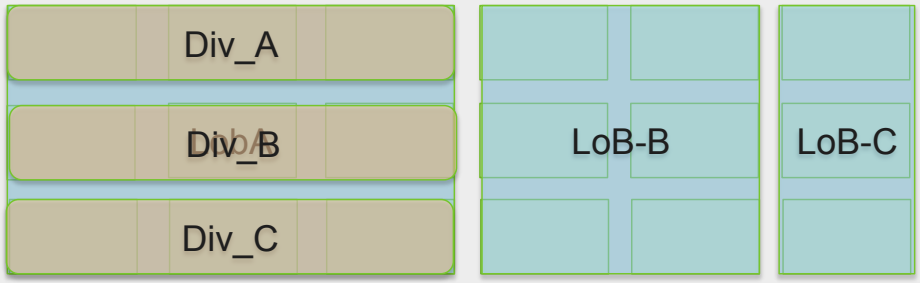
Summary: Policy-based use of computing resources

Scheduler Queues

- Capacity Scheduler allows for multiple tenants to share resources
- Queues limit access to resources
- Sub-queues are possible allowing capacity to be shared within a tenant
- Each queue has ACLs associated with users and groups
- Capacity guarantees can be set to provide minimum resource allocations
- Soft and hard limits can be placed on queues

Hadoop Cluster

"Root" queues can be setup for each tenant, and sub-queues for logical division within tenants. Each queue is allocated a % of total capacity. Queues at the root level divide the actual resources.



Queues do not allocate specific resources, only a % of the total.

Tuning of queues and limits will minimize idle resources.

Sub-Queues sub-divide the % of resources allocated to them.

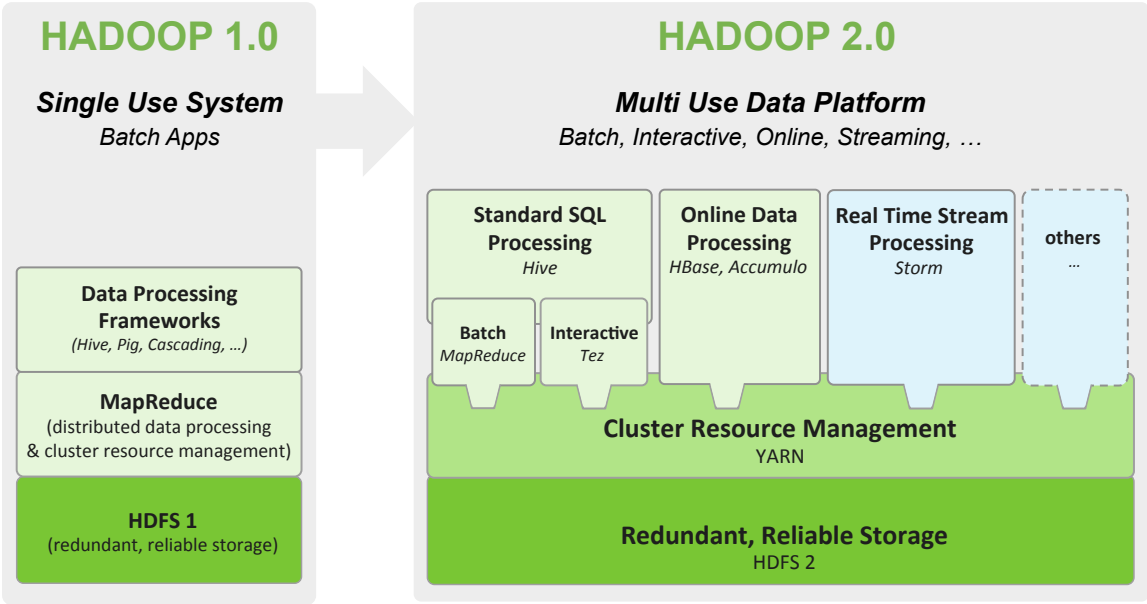
Hortonworks

YARN Application Examples

MapReduce & Tez

Hortonworks

Powering the Data Lake

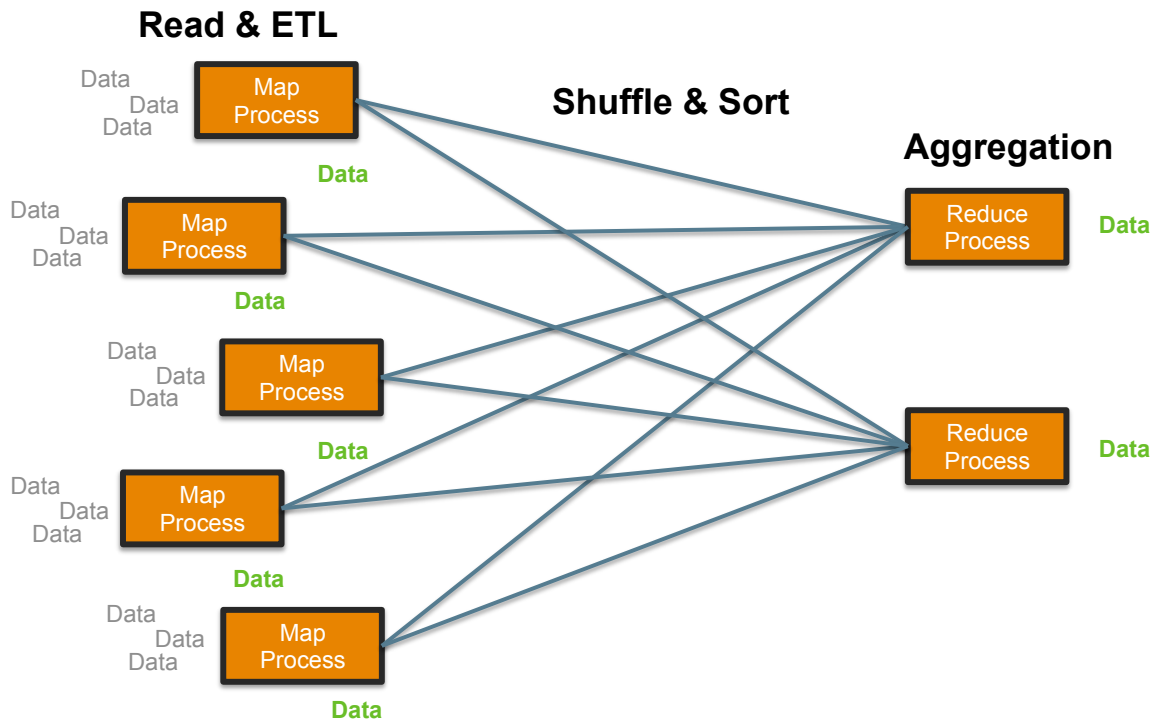


Interact with all data in multiple ways simultaneously

Hortonworks

What is MapReduce?

Break a large problem into sub-solutions



Hortonworks

Simple Algorithm

1. **Review stack of quarters**
2. **Count each year that ends in an even number**

Processing at Scale

Distributed Algorithm – MapReduce

Map
(total number of quarters)

Reduce
(sum each person's total)

WordCount in MapReduce

constitution.txt

HDFS

1 The mappers read the file's blocks from HDFS line-by-line

We the people, in order to form a...

2 The lines of text are split into words and output to the reducers

<We, 1>
<the,1>
<people,1>
<in,1>

<order, 1>
<to,1>
<form,1>
<a,1>

3 The shuffle/sort phase combines pairs with the same key

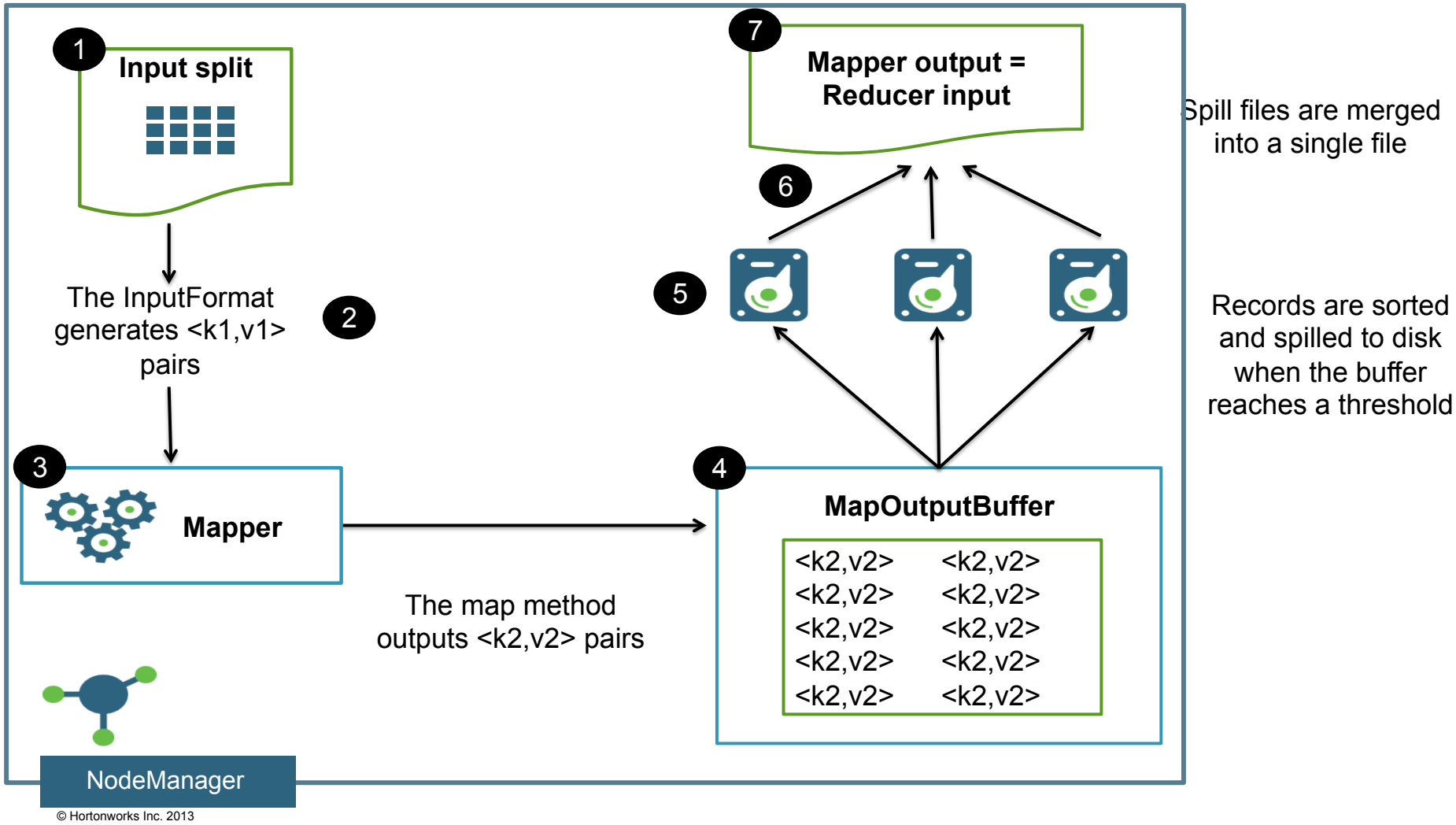
<We, (1,1,1,1)>
<the, (1,1,1,1,1,1,1,...)>
<people,(1,1,1,1,1)>
<form, (1)>

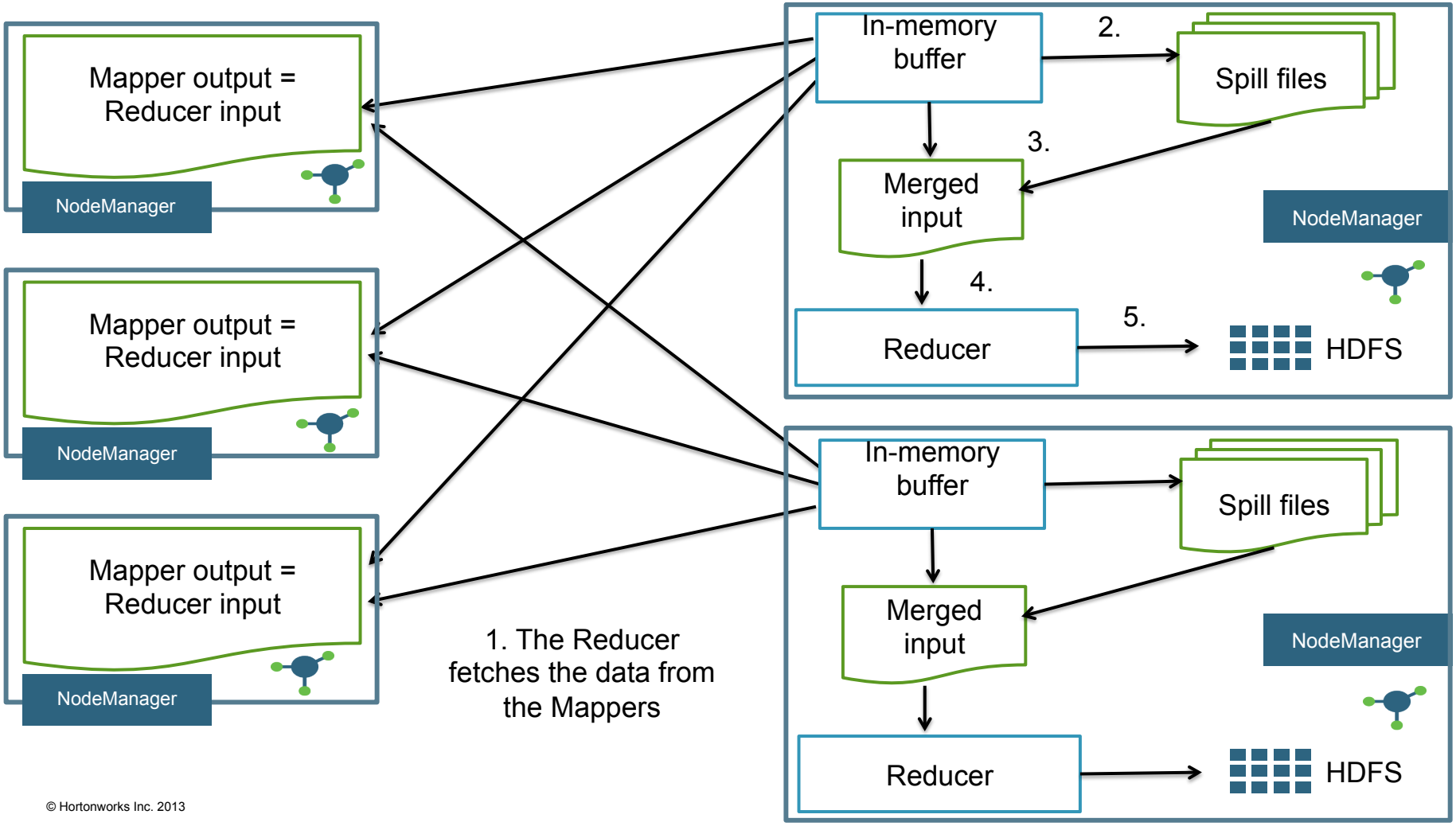
4 The reducers add up the "1's" and output the word and its count

HDFS

<We,4>
<the,265>
<people,5>
<form,1>

Hortonworks





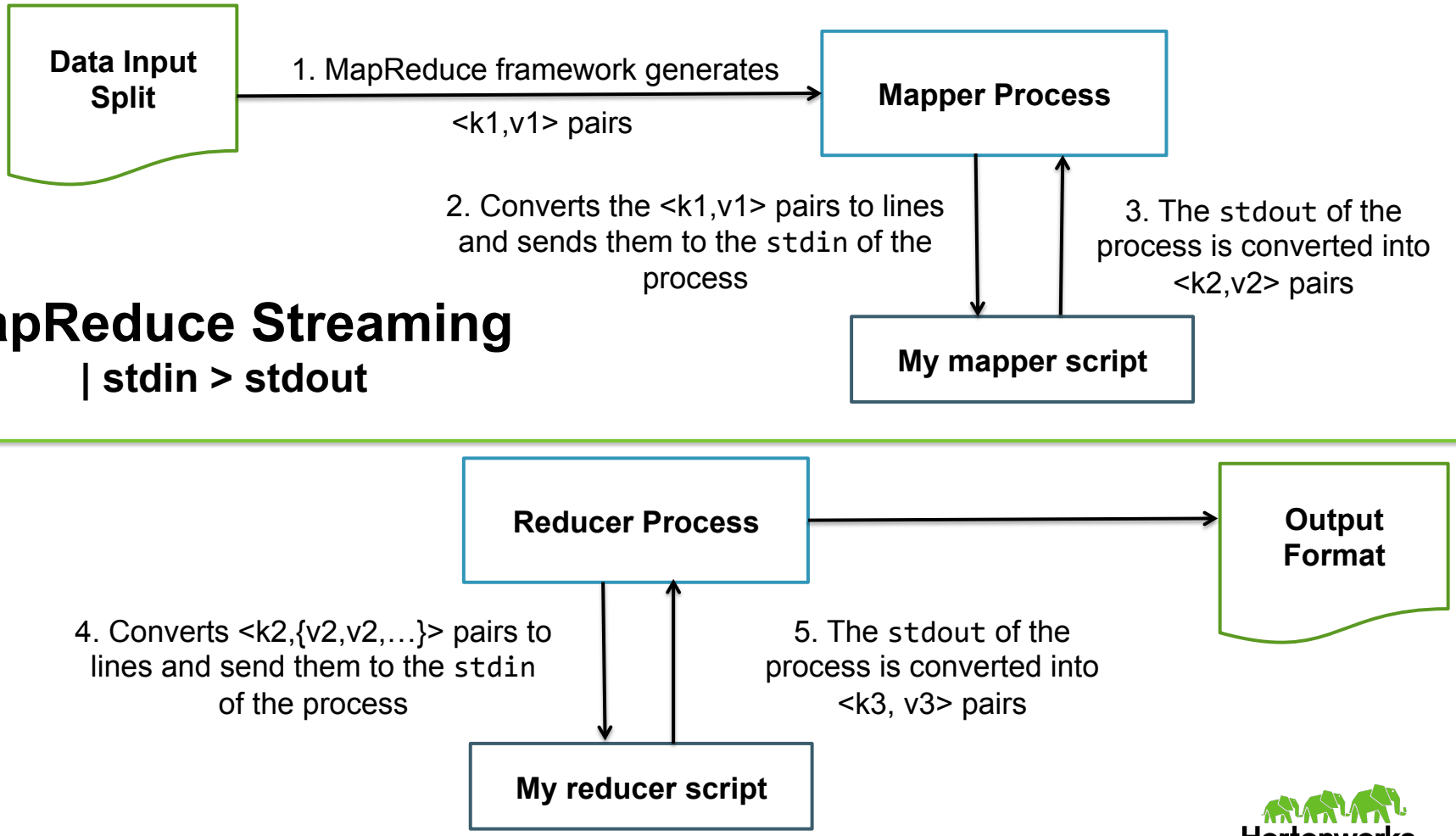
Hortonworks UNIVERSITY

Demonstration

Java MapReduce Jobs

MapReduce Streaming

| stdin > stdout

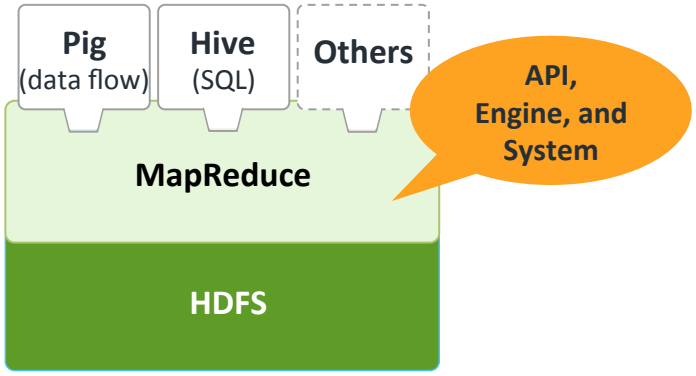


Hortonworks

Tez Provides Modern Execution Engine

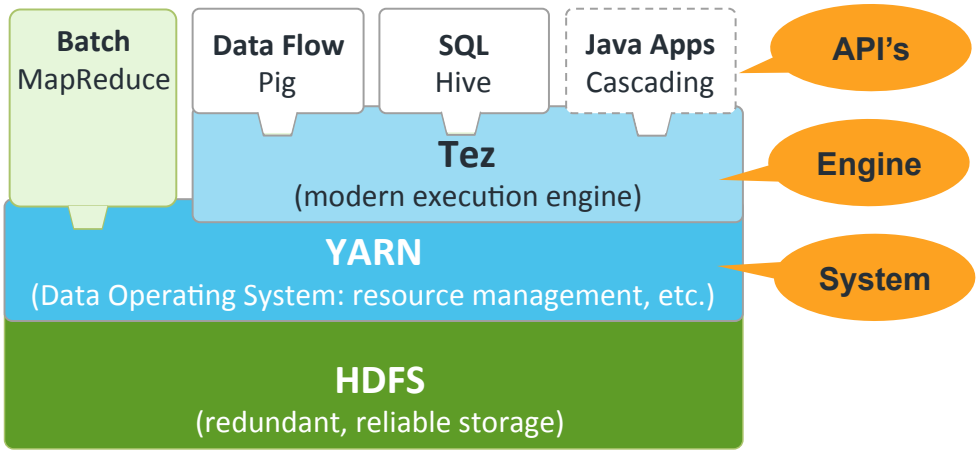
Hadoop 1

MapReduce as the Base



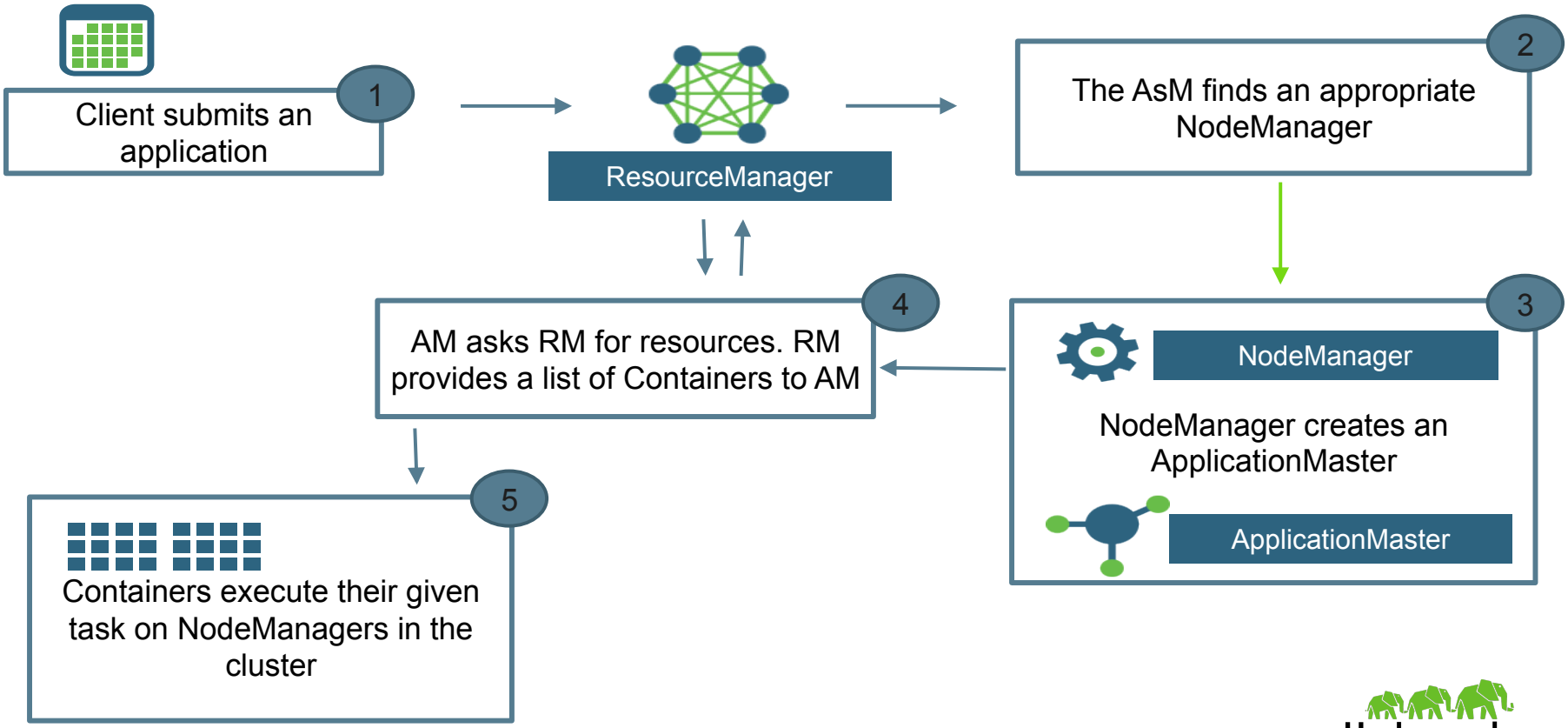
Hadoop 2

Apache Tez as a Base



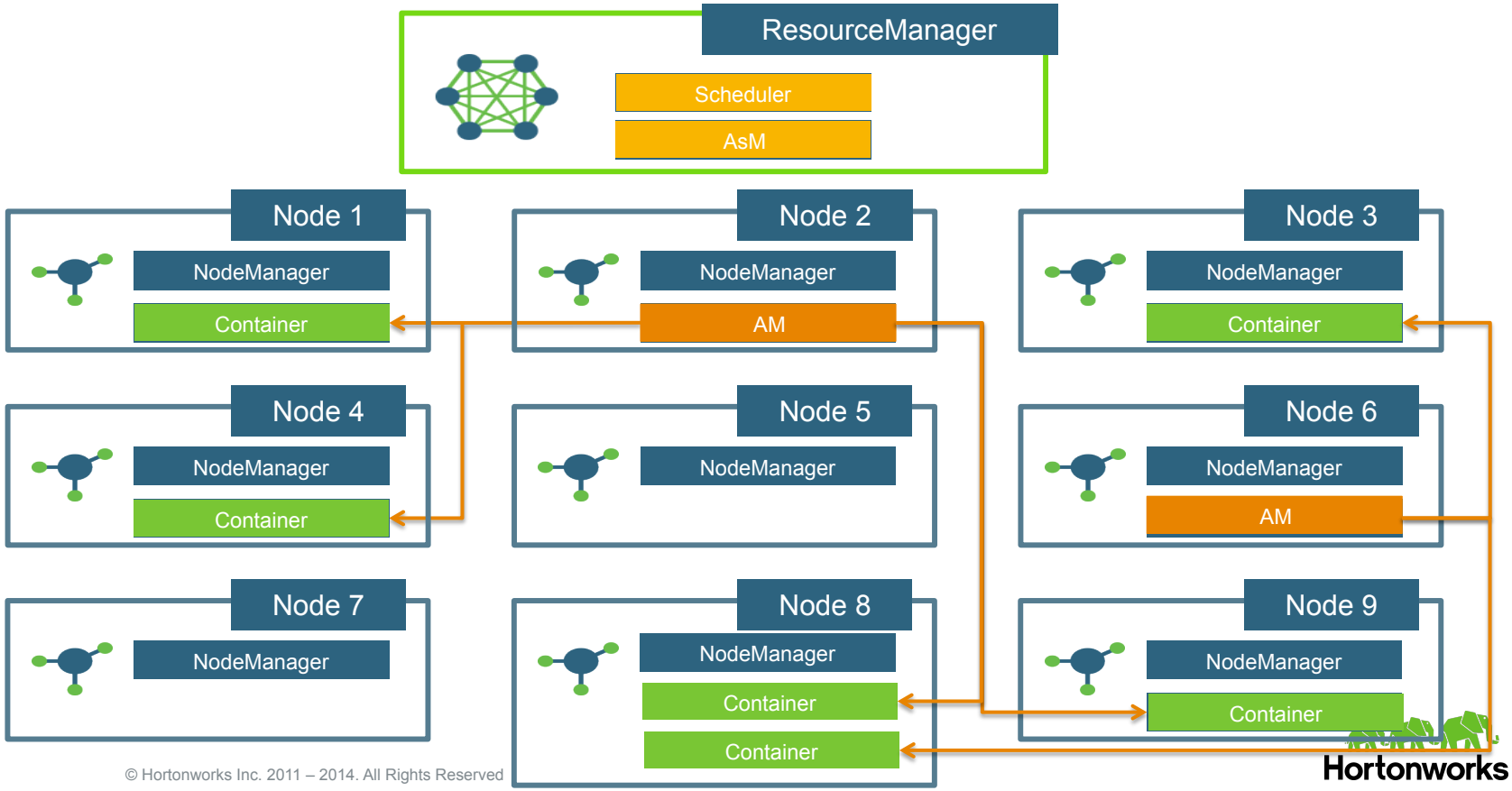
Hortonworks

Lifecycle of a YARN Application



Hortonworks

A Cluster View Example



Hortonworks

Hortonworks Sandbox

The easiest way to get started with Enterprise Hadoop

Download & Install

The Hortonworks Sandbox provides an easy way to get started to learn and develop with the Hortonworks Data Platform (HDP) anywhere. You can either run it in the cloud or your personal machine.

Hortonworks Sandbox on a VM

Developers | Administrators | Data Scientists & Analysts | Partner Tutorials

Develop with Hadoop

Start developing with Hadoop. These tutorials are designed to ease your way into developing with Hadoop:

Contact us

Apache Spark on HDP

1. **Hands-on Tour of Apache Spark in 5 Minutes**
   Introduction Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs in Scala, Java, and Python...

2. **A short primer on Scala**
   Scala is relatively new language based on the JVM. The main difference between other “Object Oriented Languages” and Scala is that everything...

3. **Interacting with Data on HDP using Scala and Apache Spark**
   In this section we are going to walk through the process of using Scala and Apache Spark to interactively analyze data on a Apache Hadoop Cluster....

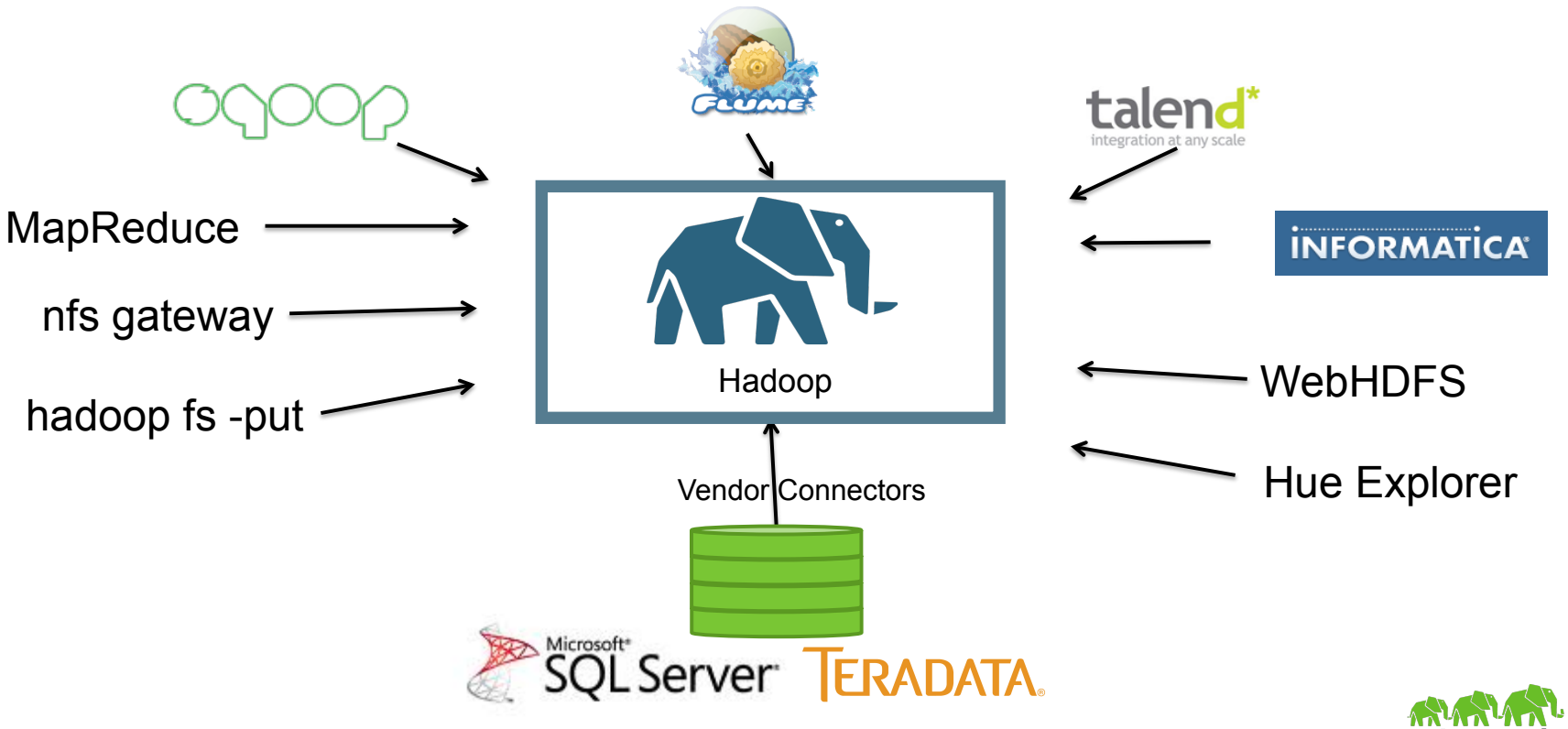
4. **Using IPython Notebook with Apache Spark**
   In this tutorial we are going to configure IPython notebook with Apache Spark on YARN in a few steps. IPython notebook is an interactive Python shell...

Unit 3 – Data Integration

Loading data into Hadoop

Hortonworks

Options for Data Input



The Hadoop Client

- The **put** command to uploading data to HDFS
- Perfect for inputting local files into HDFS
  - Useful in batch scripts
- Usage:

```
hadoop fs –put mylocalfile /some/hdfs/path
```

- POSIX utility commands such as `ls, mv, cp, touch, cat, mkdir` are also supported
- Full list of commands

```
hadoop fs
```

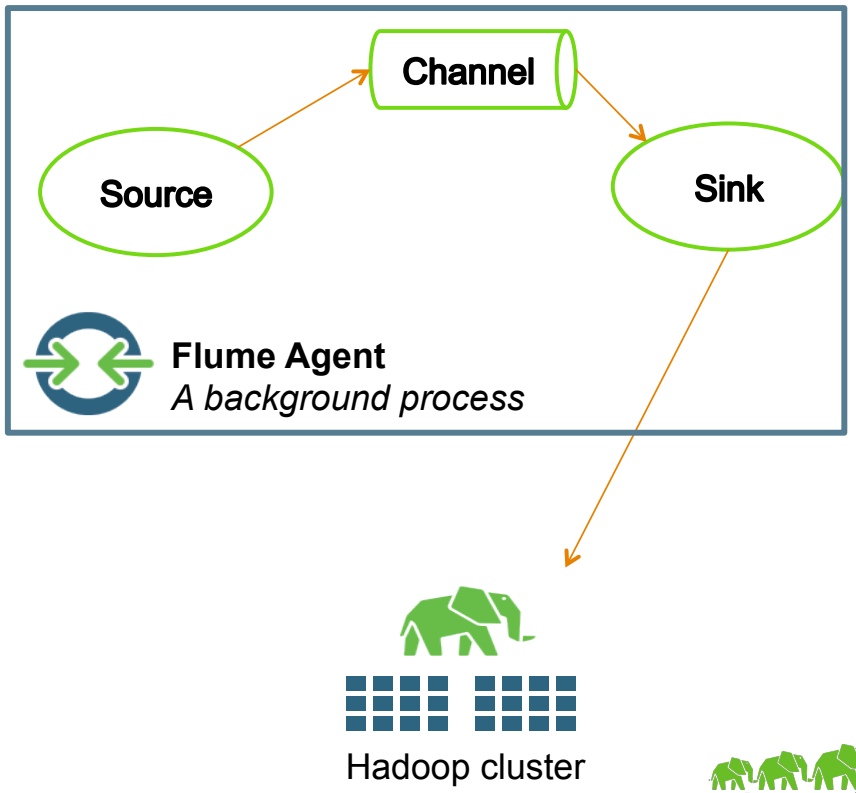

Hortonworks

WebHDFS

- REST API for accessing all of the HDFS file system interfaces:
  - http://*host*:*port*/webhdfs/v1/test/mydata.txt?op=OPEN
  - http://*host*:*port*/webhdfs/v1/user/train/data?op=MKDIRS
  - http://host:port/webhdfs/v1/test/mydata.txt?op=APPEND

Hortonworks

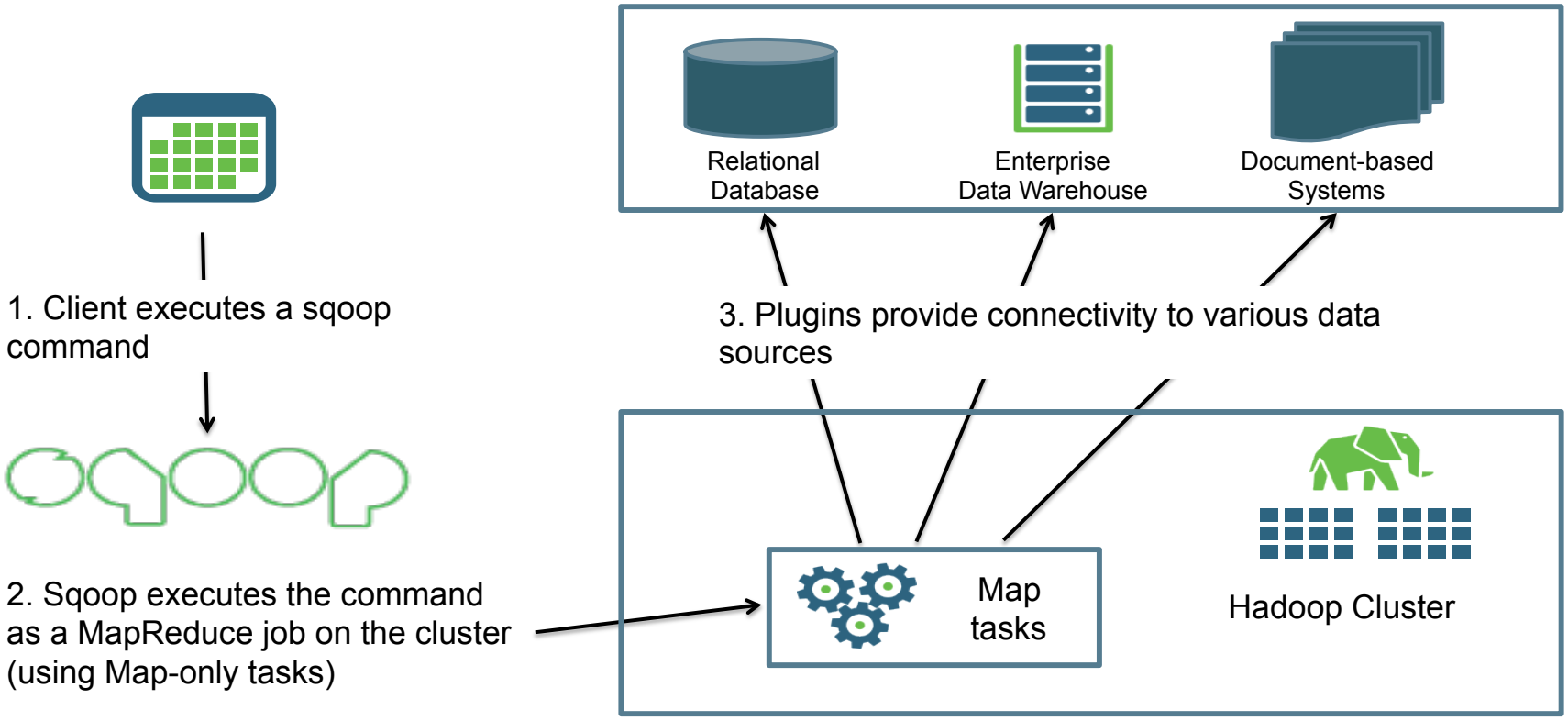
Overview of Flume: Data Streaming



Flume uses a **Channel** between the **Source** and **Sink** to decouple the processing of **events** from the storing of events.

Hortonworks

Overview of Sqoop: Database Import/Export



Hortonworks

Demonstration

Using Sqoop

Unit 4 – The Hadoop Ecosystem

Tour of the Hadoop Ecosystem

Hortonworks

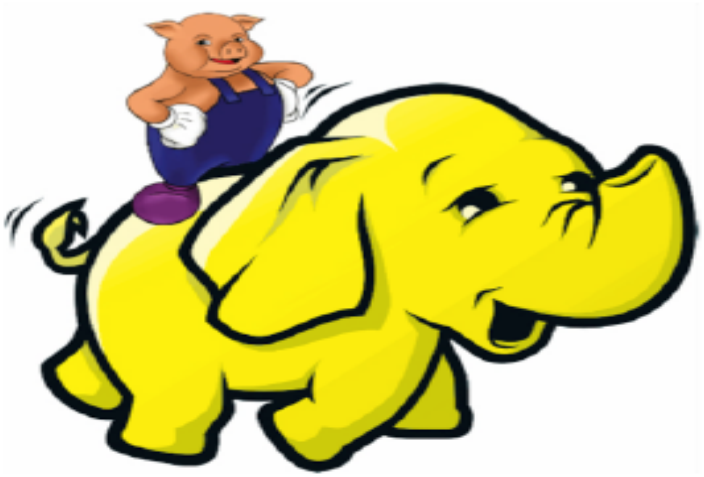
Pig

Hadoop ETL

Hortonworks

Hadoop Ecosystem: **Pig**

- An engine for executing programs on top of Hadoop
- It provides a language, Pig Latin, to specify these programs



Hortonworks

Why use Pig?

- Maybe we want to join two datasets, from different sources, on a common value, and want to filter, and sort, and get top 5

```
users = LOAD 'input/users' USING PigStorage(',')
        AS (name:chararray, age:int);

filtrd = FILTER users BY age >= 18 and age <= 25;

pages = LOAD 'input/pages' USING PigStorage(',')
        AS (user:chararray, url:chararray);

jnd = JOIN filtrd BY name, pages BY user;

grpd = GROUP jnd BY url;

smmd = FOREACH grpd GENERATE group, COUNT(jnd) AS clicks;

srtd = ORDER smmd BY clicks DESC;

top5 = LIMIT srtd 5;

STORE Top5 INTO 'output/top5sites' USING PigStorage(',');
```


Hortonworks

Hortonworks UNIVERSITY

Demonstration

Apache Pig

Hive

Hadoop SQL Data Warehouse

Hortonworks

What is Hive?

- Data warehouse system for Hadoop
- Create schemas/table definitions that *point to data* in Hadoop
- Treat your data in Hadoop as tables
- SQL 92
- Interactive queries at scale

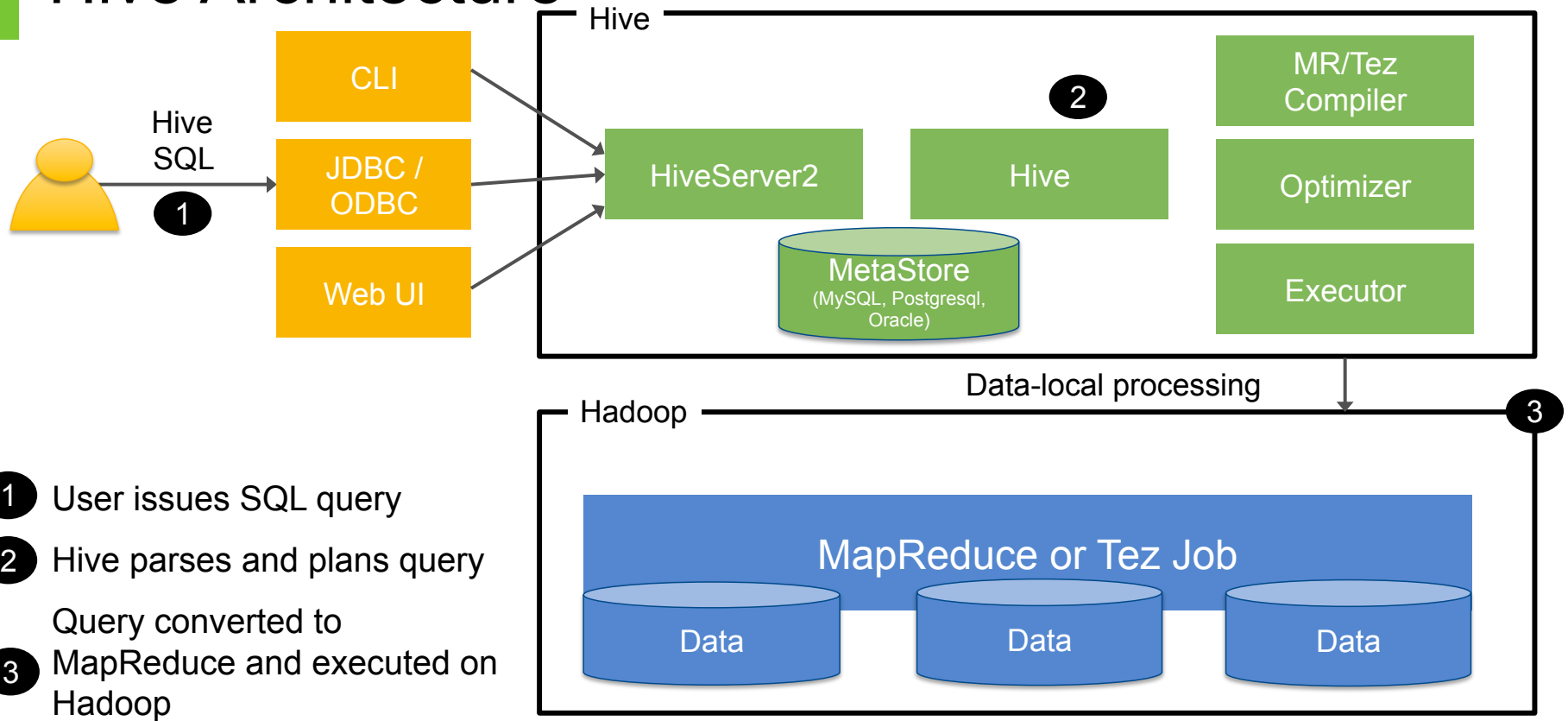
Hortonworks

HiveQL

SQL Datatypes	SQL Semantics
INT	SELECT, LOAD, INSERT from query
TINYINT/SMALLINT/BIGINT	Expressions in WHERE and HAVING
BOOLEAN	GROUP BY, ORDER BY, SORT BY
FLOAT	CLUSTER BY, DISTRIBUTE BY
DOUBLE	Sub-queries in FROM clause
STRING	GROUP BY, ORDER BY
BINARY	ROLLUP and CUBE
TIMESTAMP	UNION
ARRAY, MAP, STRUCT, UNION	LEFT, RIGHT and FULL INNER/OUTER JOIN
DECIMAL	CROSS JOIN, LEFT SEMI JOIN
CHAR	Windowing functions (OVER, RANK, etc.)
VARCHAR	Sub-queries for IN/NOT IN, HAVING
DATE	EXISTS / NOT EXISTS
	INTERSECT, EXCEPT

Hortonworks

Hive Architecture



1. User issues SQL query
2. Hive parses and plans query
3. Query converted to MapReduce and executed on Hadoop

Hortonworks

Submitting Hive Queries

- Hive CLI
  - Traditional Hive client that connects to a HiveServer instance
  - `$ hive -h hostname`
    `hive>`
- Beeline
  - A new command line client that connects to a HiveServer2 instance
  - `$ beeline`
    `Hive version 0.11.0-SNAPSHOT by Apache`
    `beeline> !connect`
    `jdbc:hive2://hostname:10000 username password`
    `org.apache.hive.jdbc.HiveDriver`

Hortonworks

Defining a Hive-Managed Table

```
CREATE TABLE customer (
     customerID INT,
     firstName STRING,
     lastName STRING,
     birthday TIMESTAMP,
  ) ROW FORMAT DELIMITED
    FIELDS TERMINATED BY ',';
```


Hortonworks

Defining an External Table

```
CREATE EXTERNAL TABLE salaries (
   gender string,
   age int,
   salary double,
   zip int
 ) ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ',';
```


Hortonworks

Defining a Table LOCATION

```
CREATE EXTERNAL TABLE SALARIES (
   gender string,
   age int,
   salary double,
   zip int
 ) ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ','
   LOCATION '/user/train/salaries/';
```


Hortonworks

Loading Data into Hive

```
LOAD DATA LOCAL INPATH ’/tmp/customers.csv' OVERWRITE
INTO TABLE customers;
```

```
LOAD DATA INPATH '/user/train/customers.csv' OVERWRITE INTO
TABLE customers;
```

```
INSERT INTO birthdays
    SELECT firstName, lastName, birthday
    FROM customers
    WHERE birthday IS NOT NULL;
```


Hortonworks

Performing Queries

```
SELECT * FROM customers;
```

```
FROM customers
    SELECT firstName, lastName, address, zip
    WHERE orderID > 0
    GROUP BY zip;
```

```
SELECT customers.*, orders.*
    FROM customers
    JOIN orders ON
    (customers.customerID = orders.customerID);
```


Hortonworks

Map Phase

Shuffle/Sort

Reduce Phase

```
SELECT c.zip,
COUNT(*)
FROM customers c
WHERE….
```

Data is shuffled across the network and sorted

```
JOIN orders o ON
c.cid = o.cid
GROUP BY (c.zip)
```

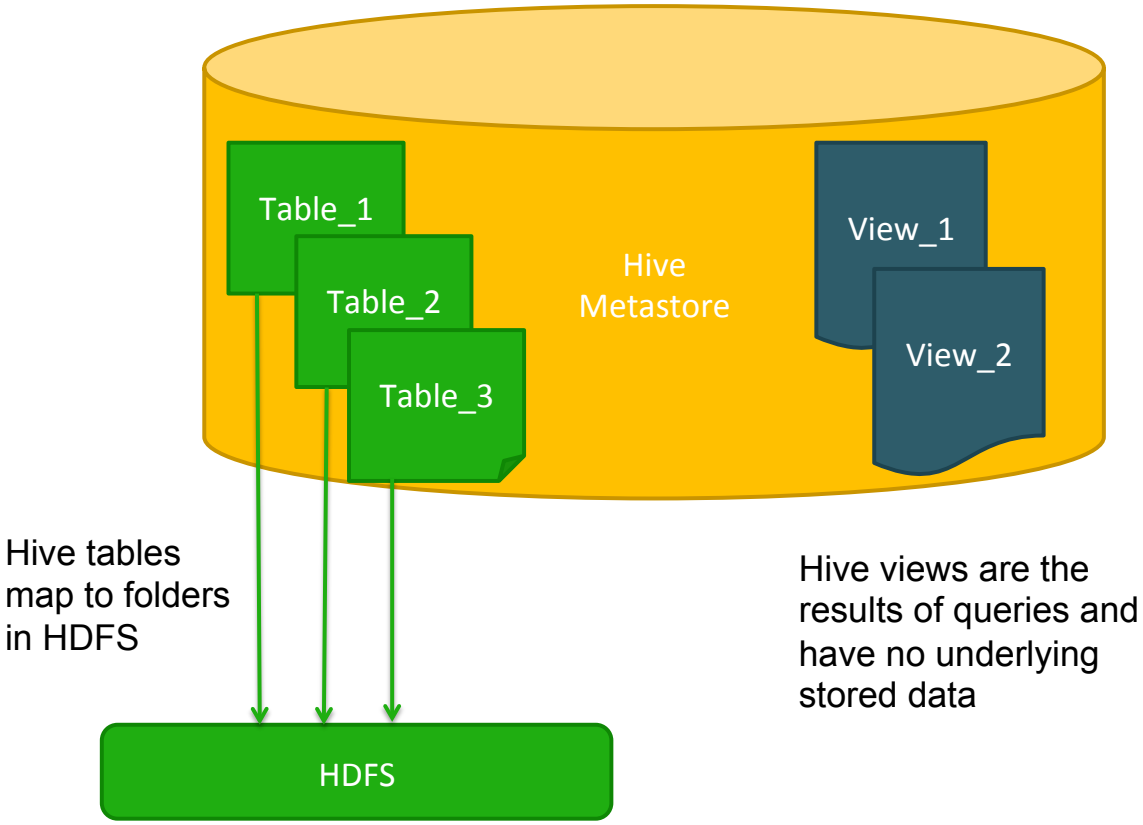
Hive/Pig compile as Reduce side function

```
ORDER BY
DISTINCT
```

Examples of more Reduce side functions

Hortonworks

Understanding Views



Hortonworks

Defining Views

```
CREATE VIEW 2010_visitors AS
  SELECT fname, lname,
     time_of_arrival, info_comment
  FROM wh_visits
  WHERE
  cast(substring(time_of_arrival,6,4)
AS int) >= 2010
  AND
  cast(substring(time_of_arrival,6,4)
AS int) < 2011;
```


Hortonworks

Using Views

You use a view just like a table:

```
from 2010_visitors
  select *
  where info_comment like "%CONGRESS%"
  order by lname;
```


Hortonworks

Interactive Hive: Overview of Stinger

Performance Optimizations

100X+ Faster Time to Insight

Deeper Analytical Capabilities

Base Optimizations

Generate simplified DAGs
In-memory Hash Joins

YARN

Next-gen Hadoop data processing framework

+

Tez

Express tasks more simply
Eliminate disk writes
Pre-warmed Containers

ORCFile

Column Store
High Compression
Predicate / Filter Pushdowns

+

Vector Query Engine

Optimized for modern processor architectures

Query Planner

Intelligent Cost-Based Optimizer

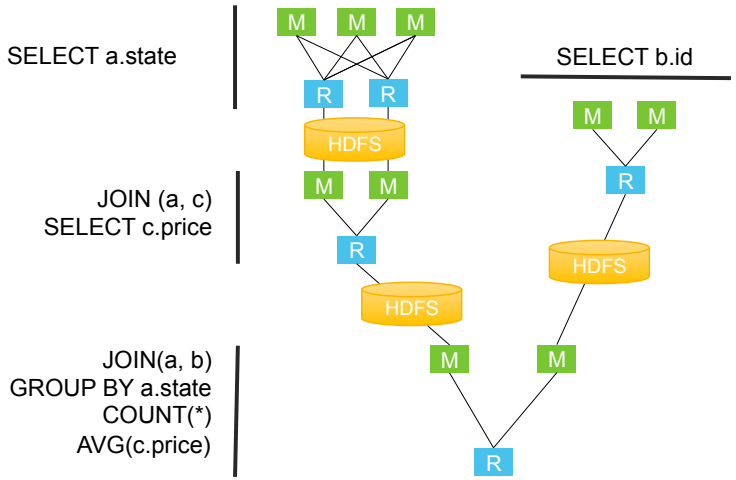
Hortonworks

Tez

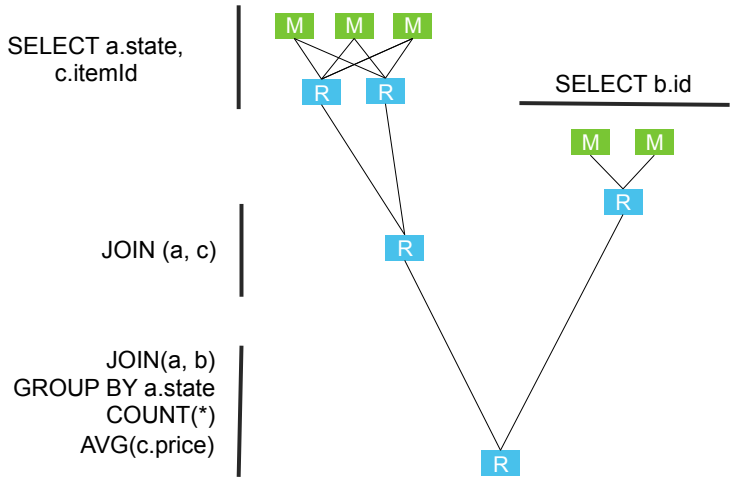
```
SELECT a.state, COUNT(*), AVG(c.price)
FROM a
JOIN b ON (a.id = b.id)
JOIN c ON (a.itemId = c.itemId)
GROUP BY a.state
```

Tez avoids unneeded writes to HDFS

Hive – MapReduce



Hive – Tez



Hortonworks

Hortonworks UNIVERSITY

Demonstration

Apache Hive

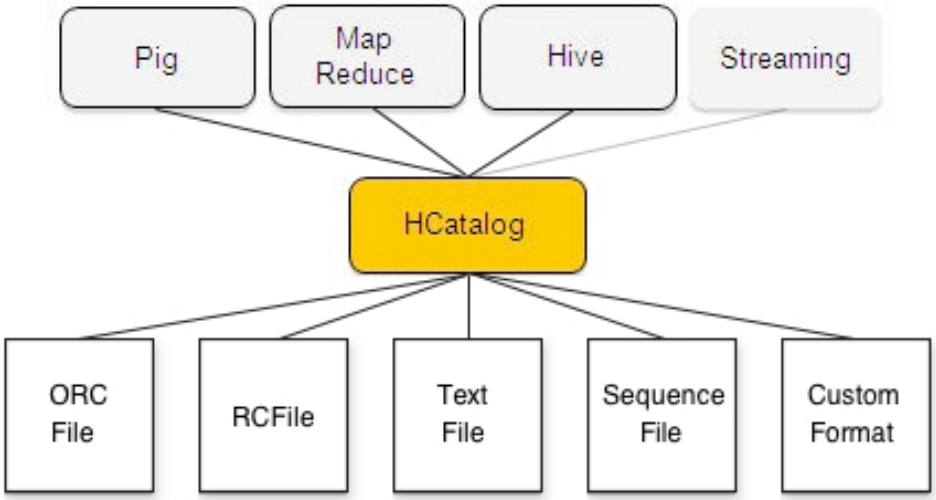
HCatalog

Hive Metastore Access

Hortonworks

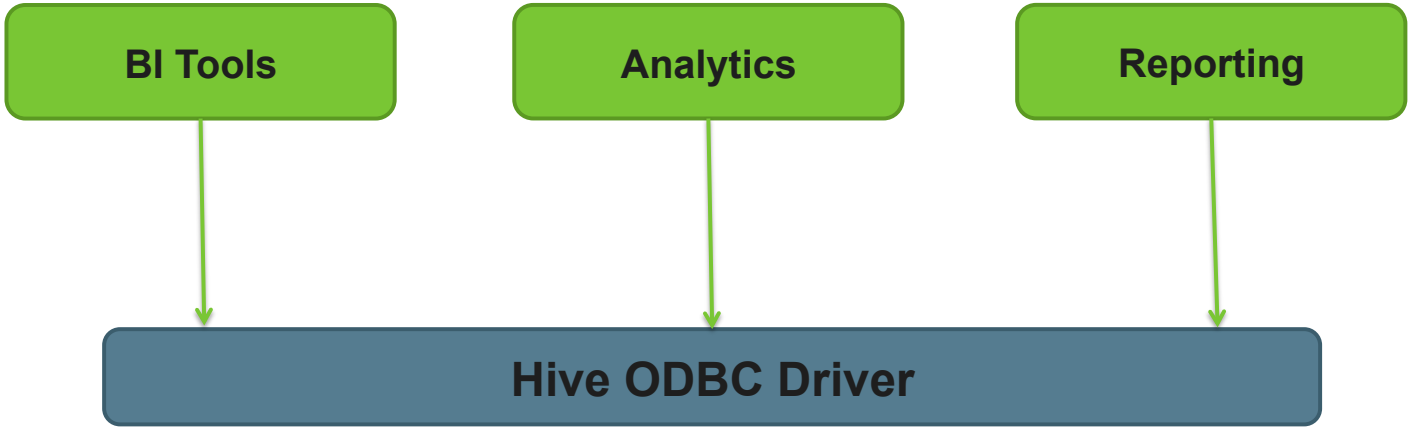
Hadoop Ecosystem: **HCatalog**

- Hive component
- Glue between Pig & Hive
  - Schema visibility to Pig Scripts & MapReduce
- REST API to
  - Access Hive schemas
  - Submit DDL
  - Launch Hive queries
  - Launch Pig jobs
  - Launch MR
  - Notifications to message broker



Hortonworks

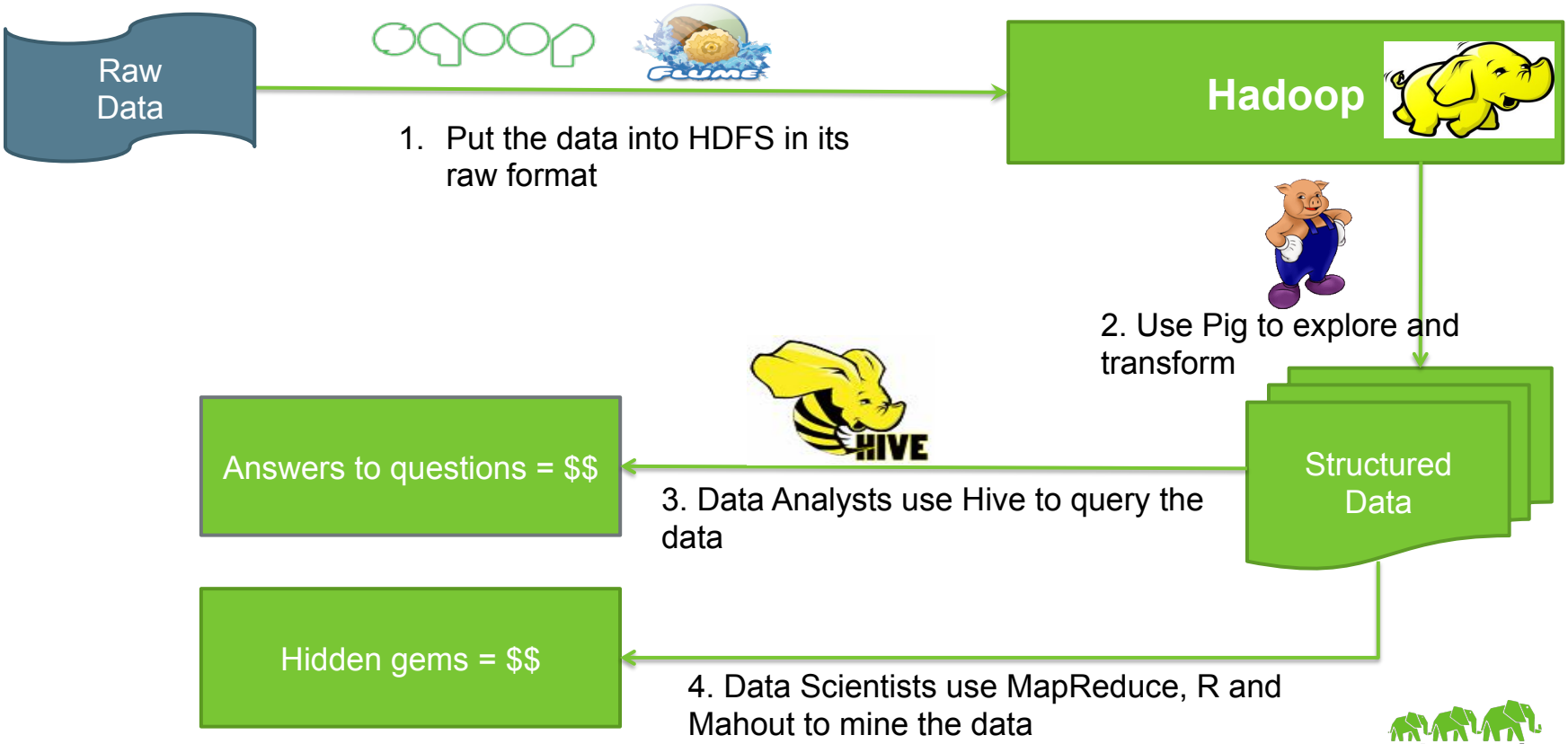
Overview of the Hive ODBC Driver



http://www.hortonworks.com/hdp/addons

Hortonworks

ROI from Your Cluster



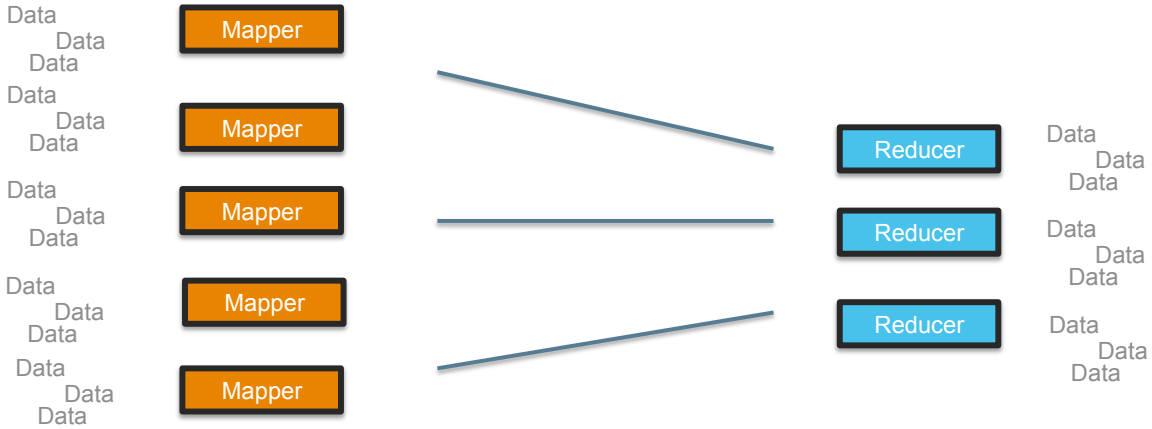
Hortonworks

Hadoop Ecosystem

Additional Processing Frameworks

Hortonworks

MapReduce applies to *a lot* of data processing problems

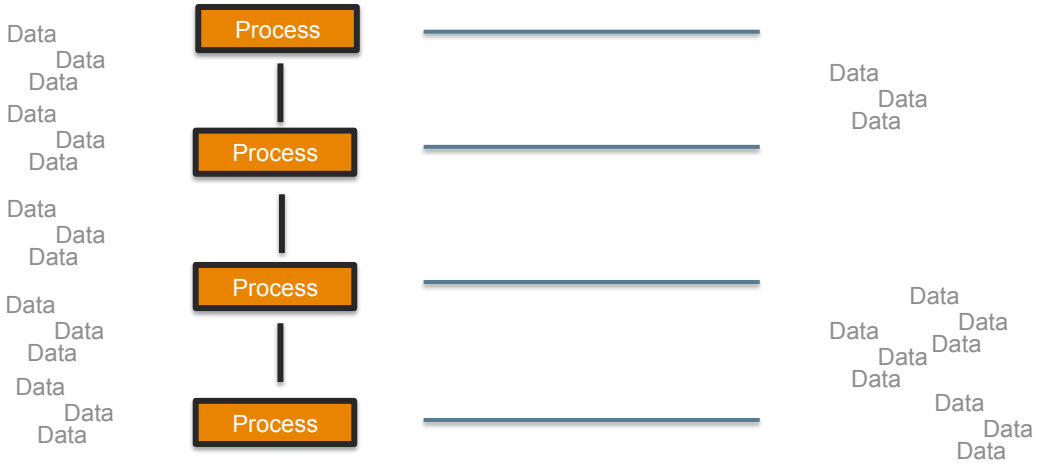


Hortonworks

MapReduce goes a long way, but not all data processing and analytics are solved the same way

Hortonworks

Sometimes your data application needs parallel processing ***and*** inter-process communication

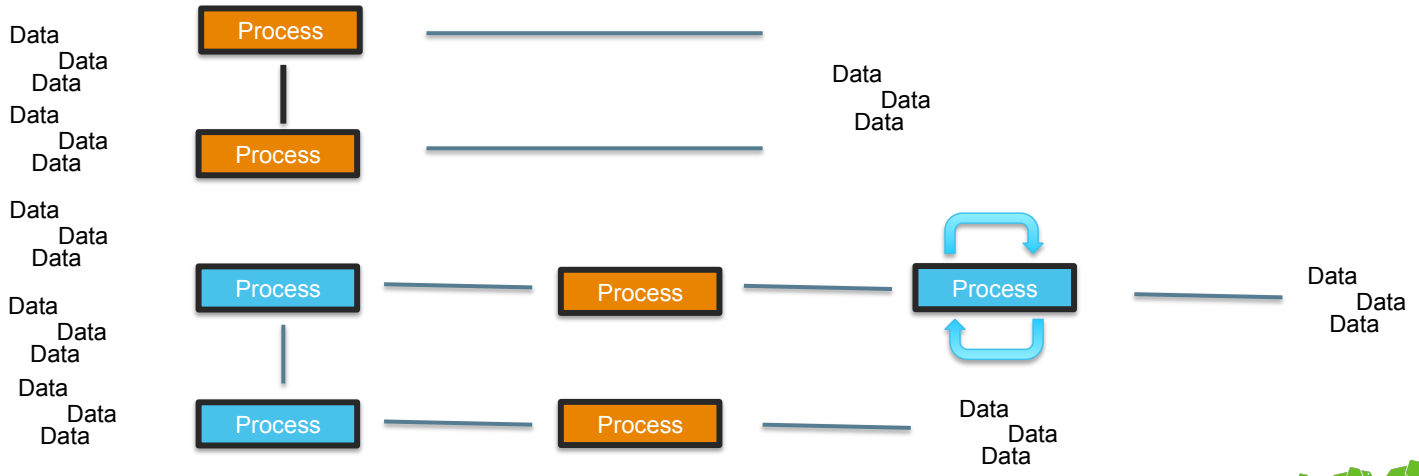


Hortonworks

…like complex event processing in
Apache Storm

Hortonworks

Sometimes your machine learning data application needs to process in memory and iterate



Hortonworks

…like in Machine Learning in **Spark**

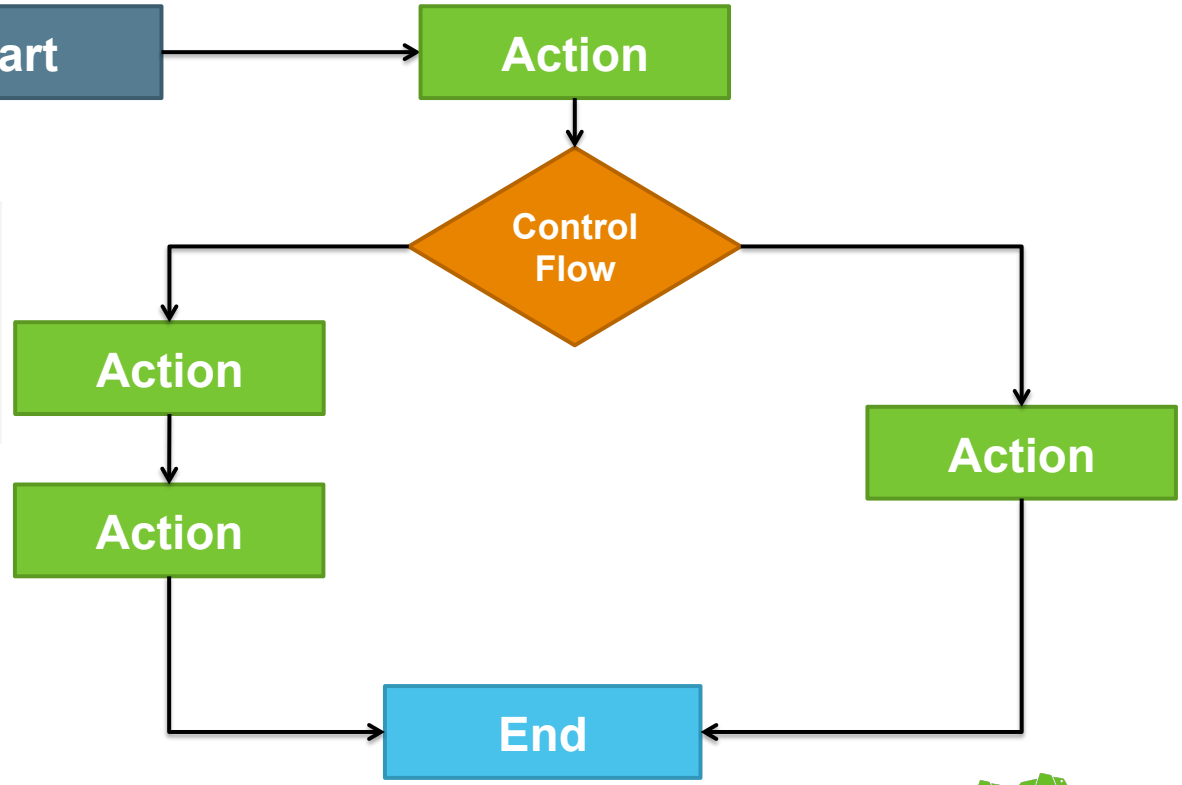
Hortonworks

Workflows & Governance

Hortonworks

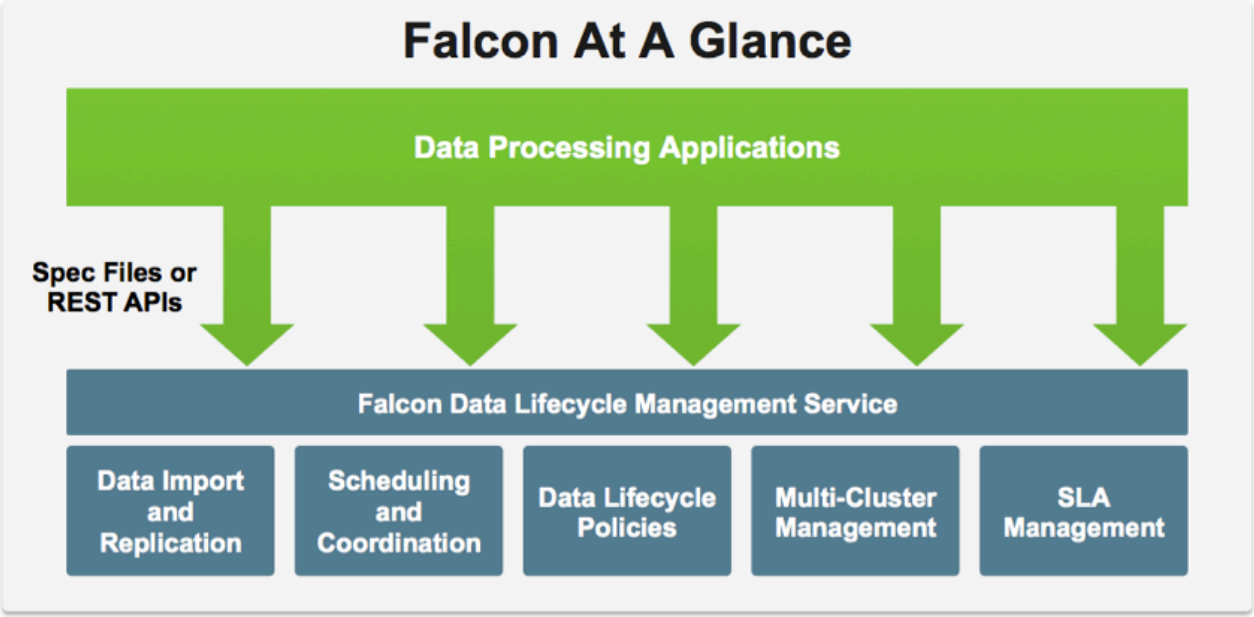
Defining an Oozie Workflow

- Expressed in XML
- Can also create workflows in a web UI



Hortonworks

Falcon: Data Lifecycle Management



> Falcon provides the key services data processing applications need.
> Complex data processing logic handled by Falcon instead of hard-coded in apps.
> Faster development and higher quality for ETL, reporting and other data processing apps on Hadoop.

Hortonworks

Security

Hortonworks

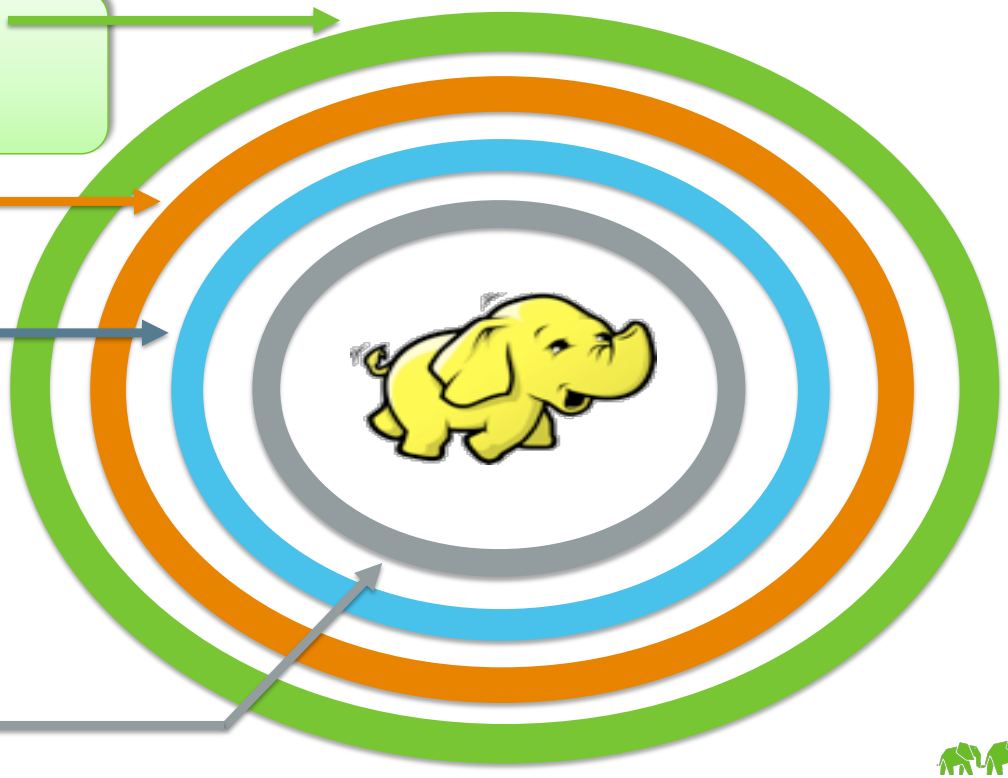
Security: Rings of Defense

Perimeter Level Security
- Network Security (i.e. Firewalls)
- Apache Knox (i.e. Gateways)

Authentication
- Kerberos

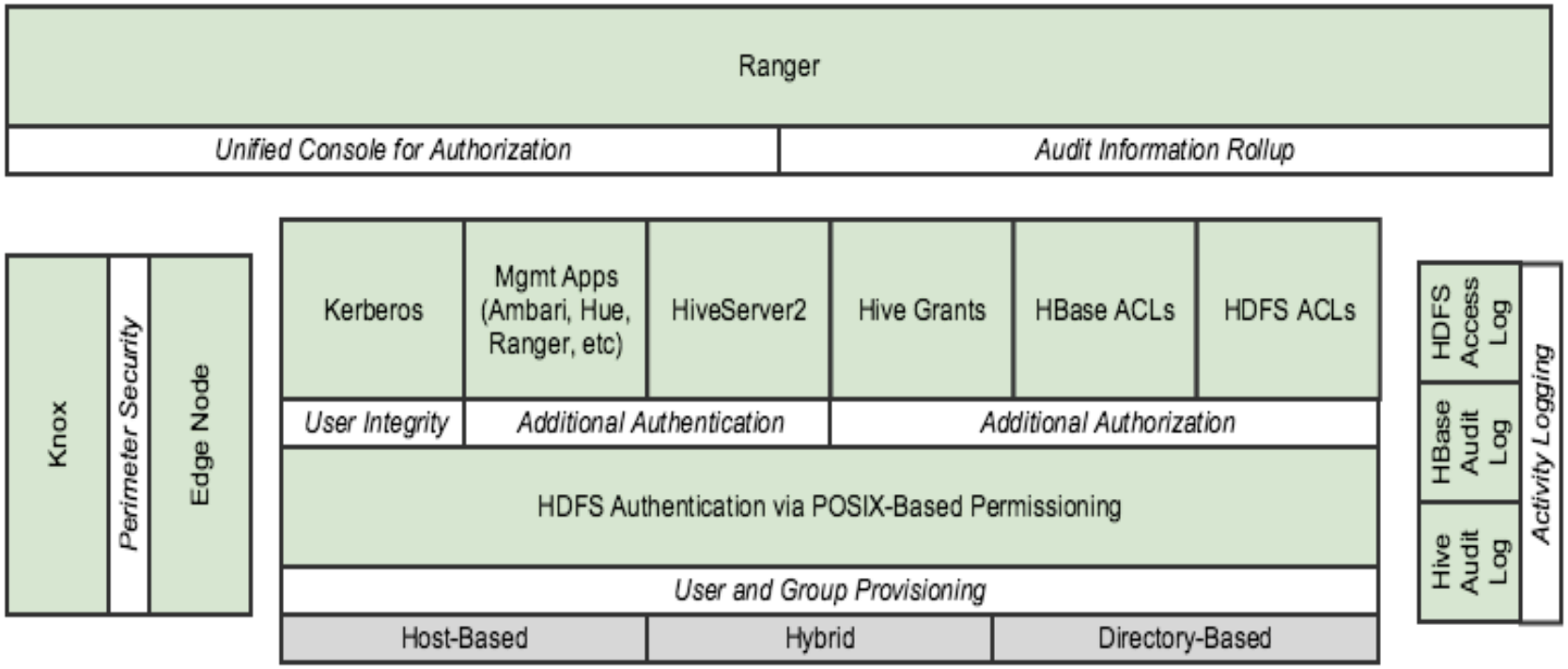
Authorization
- MR ACLs
- HDFS Permissions
- HDFS ACLs
- HiveATZ-NG
- HBase ACLs
- Accumulo Label Security

OS Security



Hortonworks

Security: Visualized as Layers



Hortonworks

Security in Hadoop with HDP + Apache Ranger

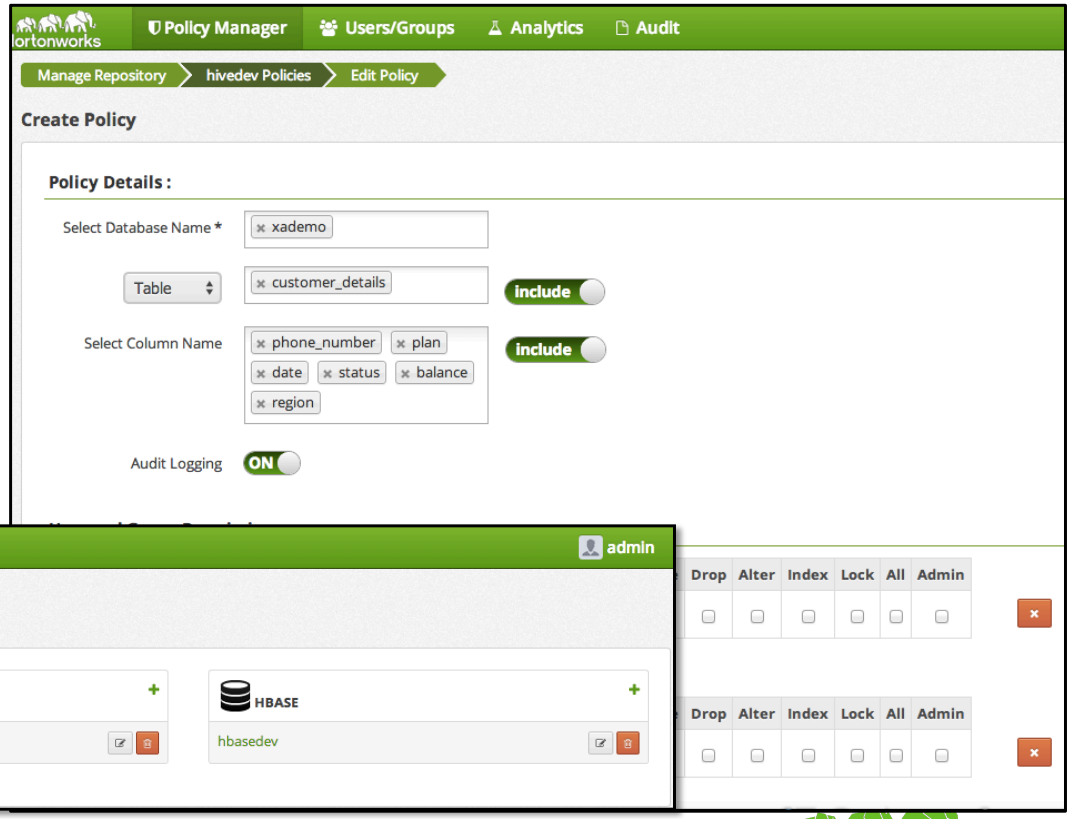
Centralized Security Administration

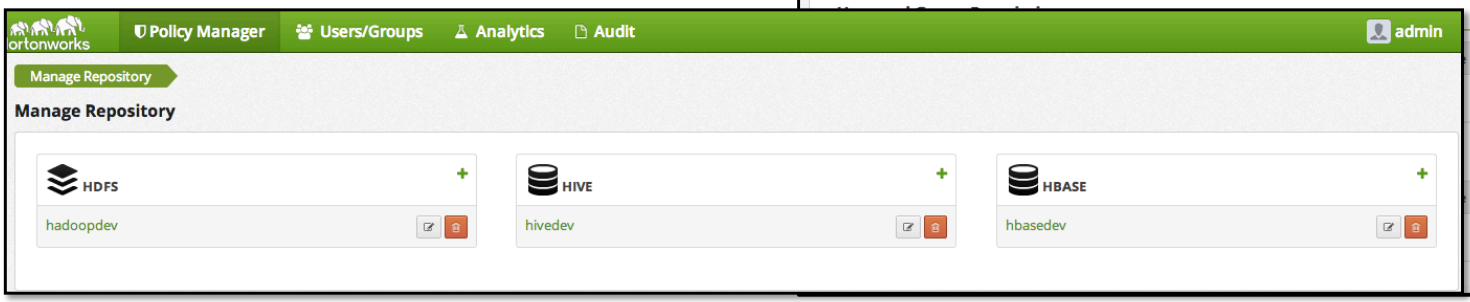
	Authentication Who am I/prove it?	**Authorization** Restrict access to explicit data	**Audit** Understand who did what	**Data Protection** Encrypt data at rest & in motion
HDP 2.2	• Kerberos in native Apache Hadoop • HTTP/REST API Secured with Apache Knox Gateway	• HDFS Permissions, HDFS ACL, • Audit logs in with HDFS & MR • Hive ATZ-NG		• Wire + at rest encryption avail • Open Source Initiatives • Partner Solutions
+				
Ranger	• *As-Is, works with current authentication methods*	• HDFS, Hive, HBase, Storm, Knox • Fine grain access control • RBAC	• Centralized audit reporting • Policy and access history	• *Future Integration*

Central Security Administration

Apache Ranger

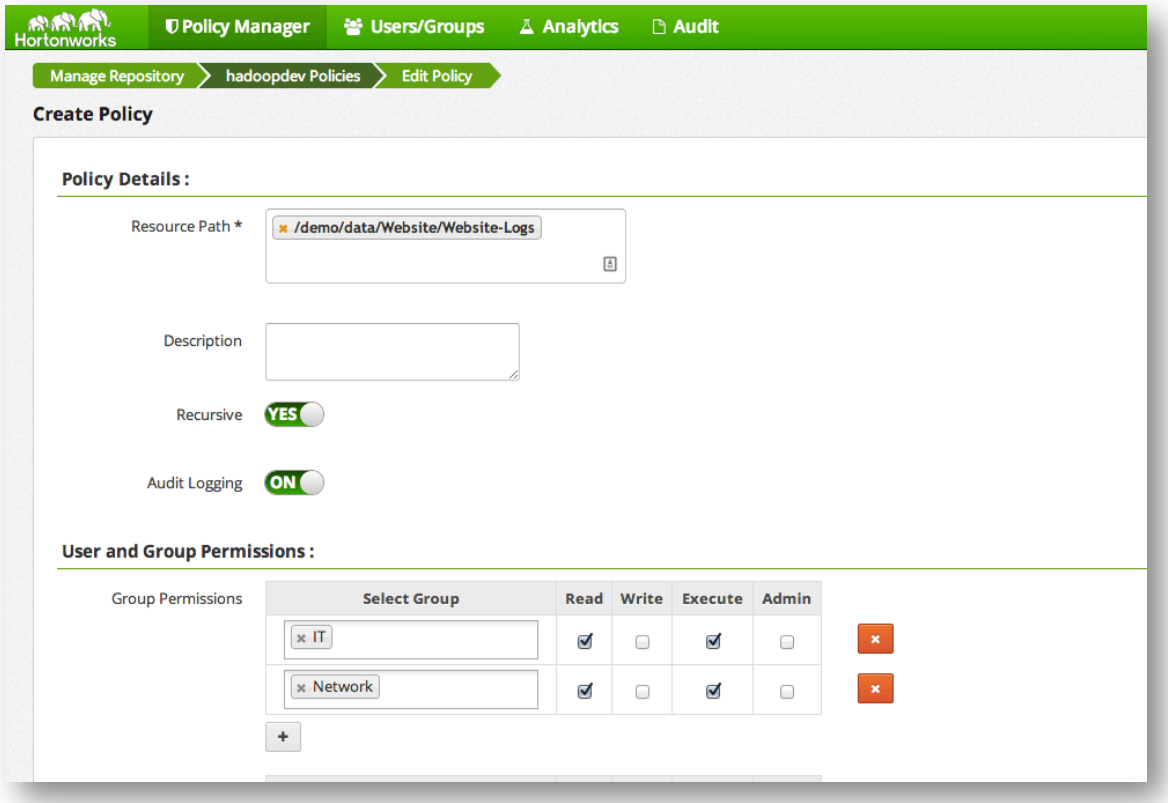
- Delivers a ‘single pane of glass’ for the security administrator
- Centralizes administration of security policy
- Ensures consistent coverage across the entire Hadoop stack





Hortonworks

Setup Authorization Policies



File level access control, flexible definition

Control permissions

Monitor through Auditing

Hortonworks | Policy Manager | Users/Groups | Analytics | Audit | admin

Big Data | Admin | Login Sessions | Agents

START DATE: 06/20/2014

Last Updated Time : 06/20/2014 05:32:12 PM

Event Time	User	Repository Name / Type	Resource Name	Access Type	Result	Access Enforcer	Client IP
06/20/2014 05:31:53 PM	mapred	hadoopdev HDFS	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	/10.0.2.15
06/20/2014 05:31:30 PM	mapred	hadoopdev HDFS	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	/10.0.2.15
06/20/2014 05:30:30 PM	mapred	hadoopdev HDFS	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	/10.0.2.15
06/20/2014 05:29:30 PM	mapred	hadoopdev HDFS	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	/10.0.2.15
06/20/2014 05:28:53 PM	mapred	hadoopdev HDFS	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	/10.0.2.15
06/20/2014 05:28:30 PM	mapred	hadoopdev HDFS	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	/10.0.2.15
06/20/2014 05:27:30 PM	mapred	hadoopdev HDFS	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	/10.0.2.15
06/20/2014 05:26:31 PM	mapred	hadoopdev HDFS	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	/10.0.2.15

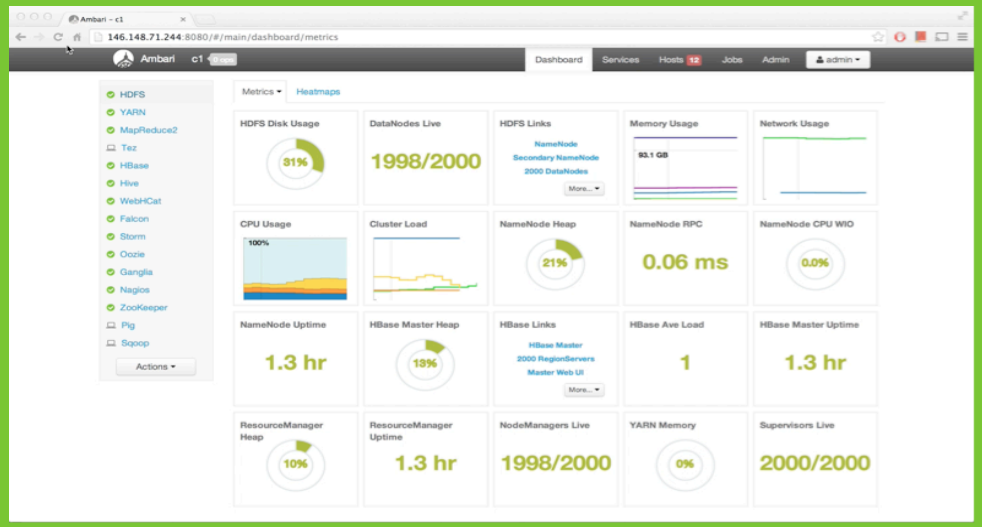
Ambari

Manage, Monitor & Provision Hadoop Clusters

Hortonworks

How do you Operate a Hadoop Cluster?

Apache Ambari is a platform to provision, manage and monitor Hadoop clusters



Hortonworks

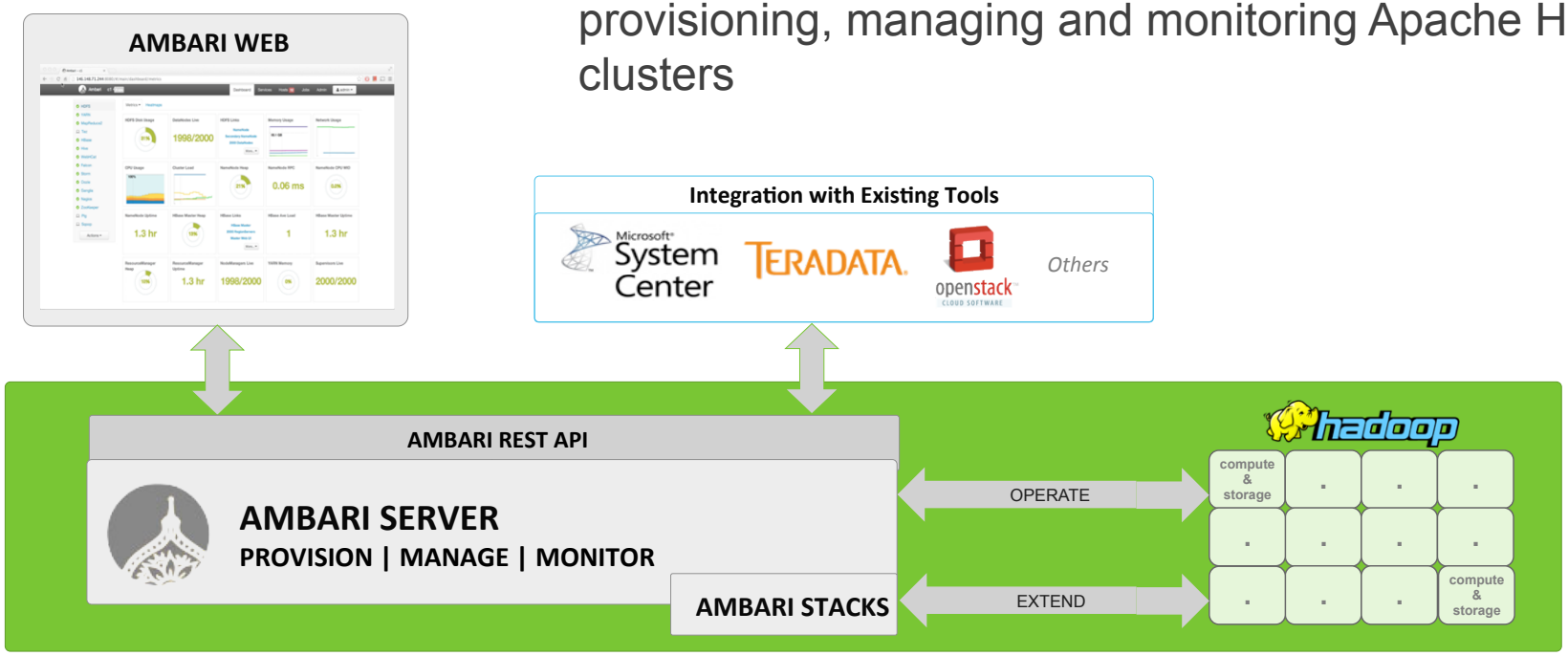
Apache Ambari Themes

Operate Hadoop at Scale	Deliver the **core** operational capabilities to **provision**, **manage** and **monitor** Hadoop clusters at **scale**.
Integrate with the Enterprise	**Robust API** for integration with **existing enterprise systems**, such as **Teradata Viewpoint** and **Microsoft SCOM**.
Extend for the Ecosystem	Provide **extensible platform** for Customers, Partners and the Community to, such as **Stacks** and **Views**.

Hortonworks

Hadoop Operations Platform

Apache Ambari is a 100% open source *platform* for provisioning, managing and monitoring Apache Hadoop clusters



Hortonworks

Extensibility Features

Goal: *Extend* Ambari without hard-coding *in* Ambari

Stacks	• To add new Services (ISV or otherwise) beyond HDP Stack • To customize a Stack for customer specific environments
Blueprints	• To use Ambari for automating cluster installations • To share best practices on layout and cluster configuration
Views	• To extend and customize the Ambari Web UI • Add new capabilities, customize existing capabilities

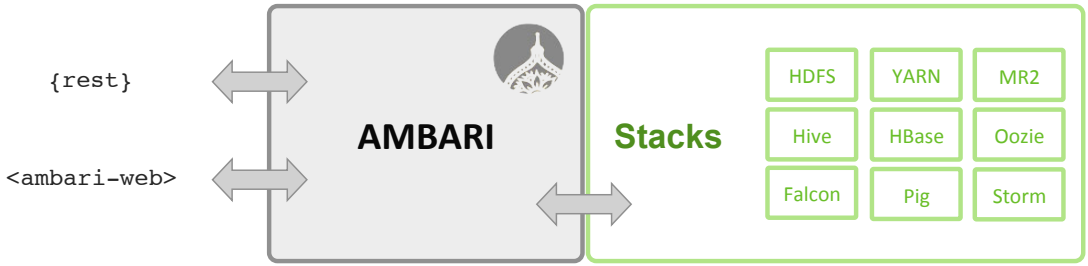
Hortonworks

Ambari Stacks

Defines a consistent Stack lifecycle interface that can be extended

Encapsulates Stack Versions, Services, Components, Dependencies, Cardinality, Configurations, Commands

Dynamically add Stack + Service definitions



Hortonworks

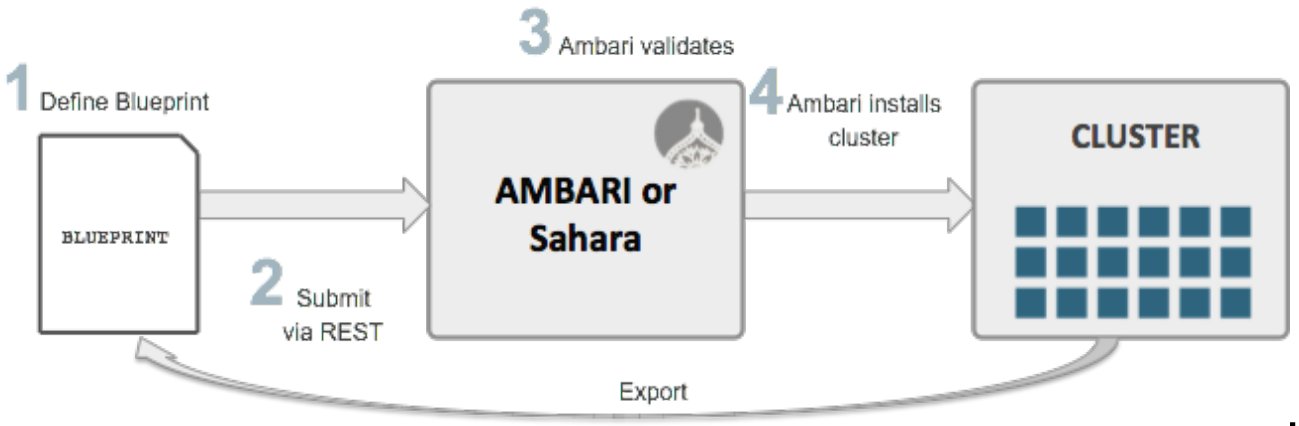
Ambari Blueprints

Two primary goals of Ambari Blueprints:

- **Provide API-driven deployments based on a self-contained cluster description**
- **Ability to export a complete cluster description of a running cluster**

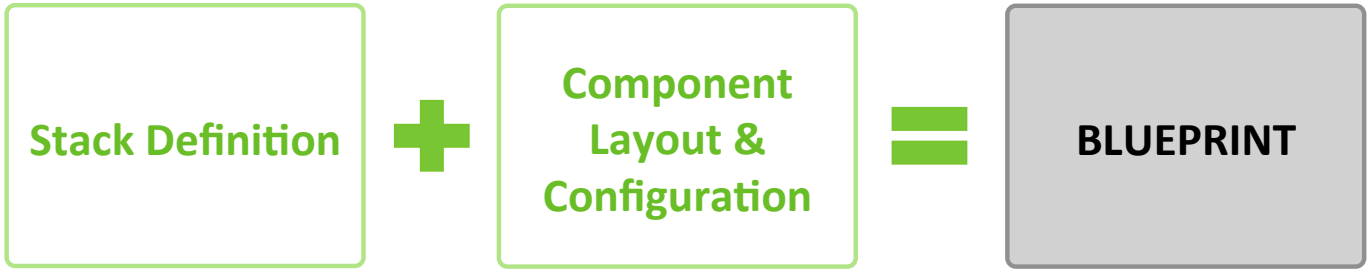
Blueprints contain cluster topology and configuration information

Enables interesting use cases between physical and virtual environments



Hortonworks

Ambari Stacks + Blueprints Together



Hortonworks

Ambari Views Framework

Goal: enable the delivery of custom UI experiences in Ambari Web

Developers can extend the Ambari Web interface

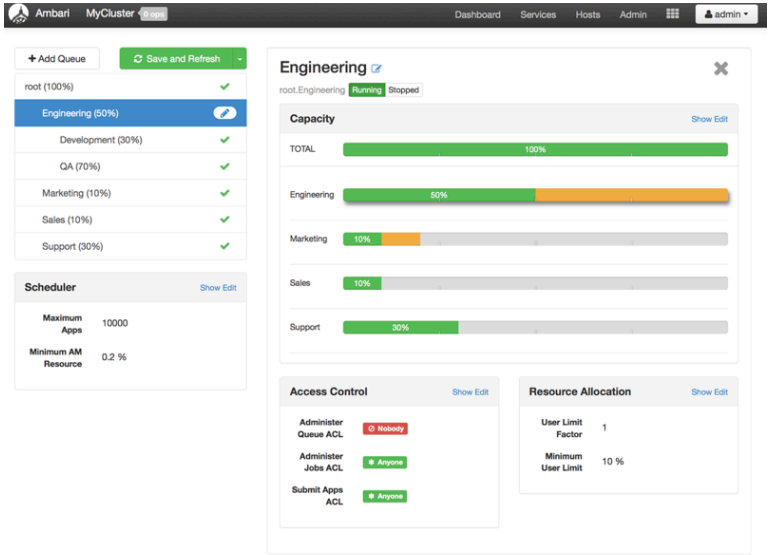
- Views expose custom UI features for Hadoop Services

Ambari Admins can entitle Views to Ambari Web users

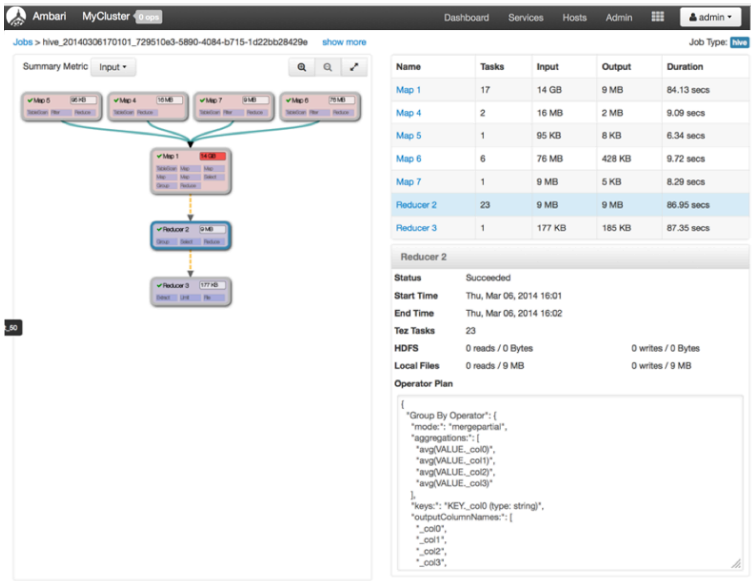
- Entitlements framework for controlling access to Views

Hortonworks

Example Views



Capacity Scheduler
"Queue Manager" View



Hive Tez
"Jobs" View

Hortonworks

Example Ambari Views

Operators

- Capacity Scheduler Queue Management
- YARN Resource Utilization
- Heatmaps
- HDFS / Hive Mirroring

Data Workers

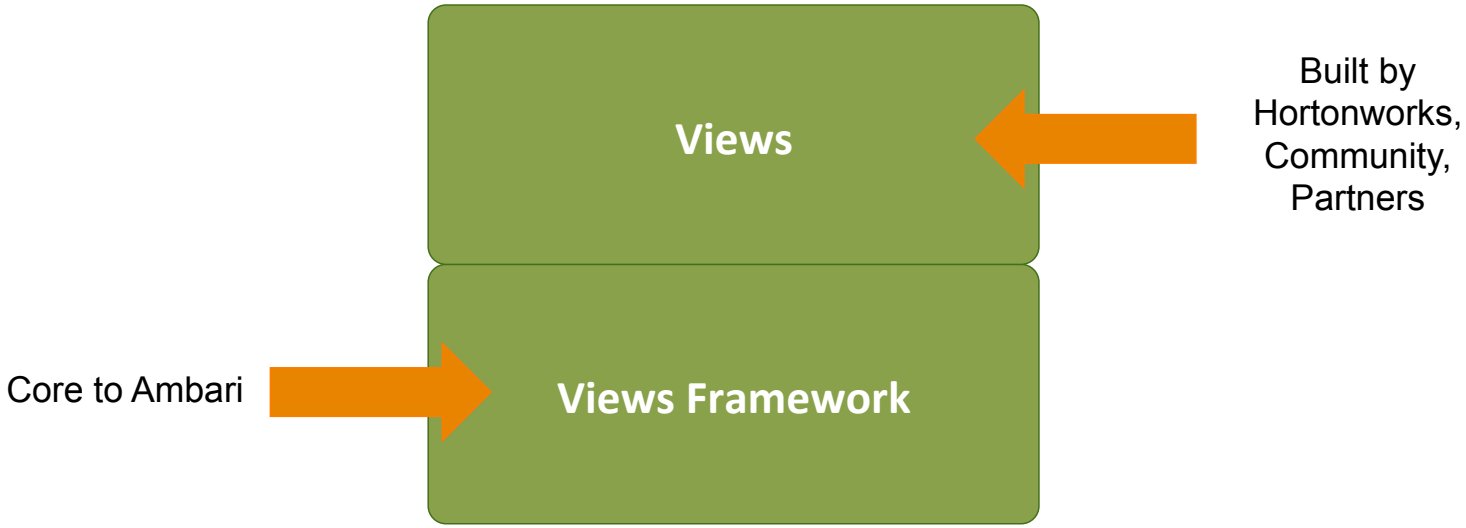
- Pig Query Editor
- Hive Query Editor
- Workflow Design
- HDFS File Browser
- Job Visualization

Application Developers

- Job Visualization
- Streaming Topology Visualization

Hortonworks

Views Framework vs. Views



Hortonworks

Choice of Deployment Model

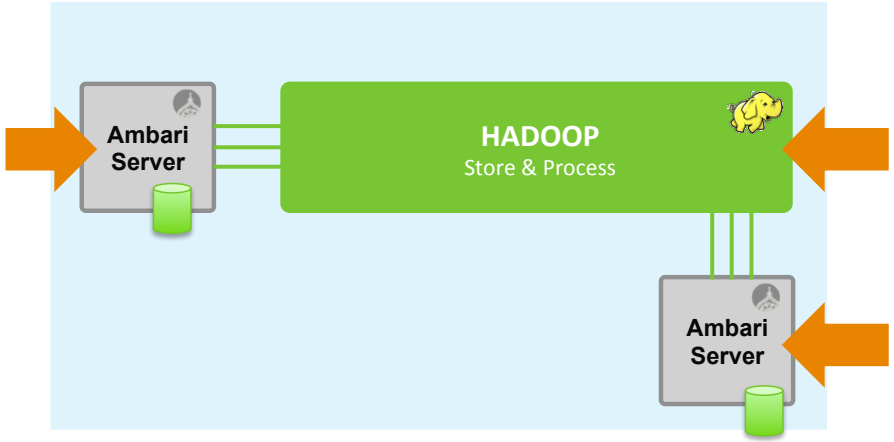
- **For Hadoop Operators:**

  Deploy Views in an Ambari Server that is managing a Hadoop cluster

- **For Data Workers:**

  Run Views in a “standalone” Ambari Server

Operators manage the cluster, **may have Views** deployed



Data Workers use the cluster and use a “standalone” Ambari Server **for Views**

Hortonworks

Unit 5 – Hadoop Adoption

Charting a path to adopt Hadoop

Hortonworks

Hadoop Adoption Jumpstart

- **Download & run the Sandbox**

- **http://www.youtube.com/hortonworks**

  Sandbox tutorials have related videos on YouTube

- **POC on Sandbox**

  Easily transfer sample data onto Sandbox

Hortonworks

Hadoop Proof of Concept

- **Start small**
  - For example, a 4-node cluster is a good start
- **A POC cluster can be scaled and productionized**
- **Empower yourself & team with knowledge**
- **What more do I need to learn:**
  - To *use* the Hadoop Ecosystem?
  - To *manage* the Hadoop Ecosystem?

Hortonworks

Use Cases

Run tutorials on Sandbox

Videos of tutorials

youtube.com/hortonworks

Hortonworks

WHAT'S YOUR PATH?

I'm a **Developer**

Interested in:
Architecture & Fundementals
MapReduce Programming
Real-time Analytics

Developer Classes:

Developing Apps with Java ›
4 Days

Data Analysis with Hive & Pig ›
4 Days

Developing Solutions on Windows ›
4 Days

Developing Custom YARN Applications ›
2 Days

I'm a **System Admin**

Interested in:
Cluster Monitoring
Security & Governance
Certification

Admin Classes:

Operations Management ›
4 Days

I'm a **Data Analyst**

Interested in:
SQL & Scripting Languages
Large Scale Data Sets
Creating Value & Opportunity

Data Classes:

Applying Data Science with Hadoop ›
2 Days

Data Analysis with Hive & Pig ›
4 Days

http://www.hortonworks.com/training

We do Hadoop *successfully*.

Contribute to Apache

Support Hadoop adopters

Provide Consulting Services

Deliver Training

We do Hadoop *successfully everywhere.*

Spotify
ebay
KABAM
at&t
verizon
T · ·Mobile·
neustar
KOHL'S
PayPal
Comcast
SAMSUNG
YAHOO!
ROGERS
UNITEDHEALTH GROUP
Johnson Controls
TELUS
XING
salesforce.com
match.com
GAP
WELLS FARGO
warner music group
DISCOVER
Bank of America
Georgia-Pacific
SOCIAL SECURITY USA ADMINISTRATION
PROGRESSIVE
TRUECar
FLS connect
Target Defined. Message Delivered.
U.S. BUREAU OF LABOR STATISTICS
GE Healthcare
BARNES & NOBLE BOOKSELLERS
CardinalHealth
cisco
UNION PACIFIC
Trustwave
Security begins with Trust
OVERSEE.NET
inmar
think FINANCE
compete
ENTRAVISION COMMUNICATIONS corporation
ZirMed
move
24/7 MEDIA
TAGGED
amdocs
WD Western Digital
O.C. TANNER appreciate.
CDW

We do Hadoop *successfully, everywhere, with partners.*

Thank you!

lmartin@hortonworks.com