Hortonworks

HDP Overview: Essentials

A Technical Understanding for Business Users and Decision Makers

Hortonworks. We do Hadoop.

A Hortonworks University Training Course

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

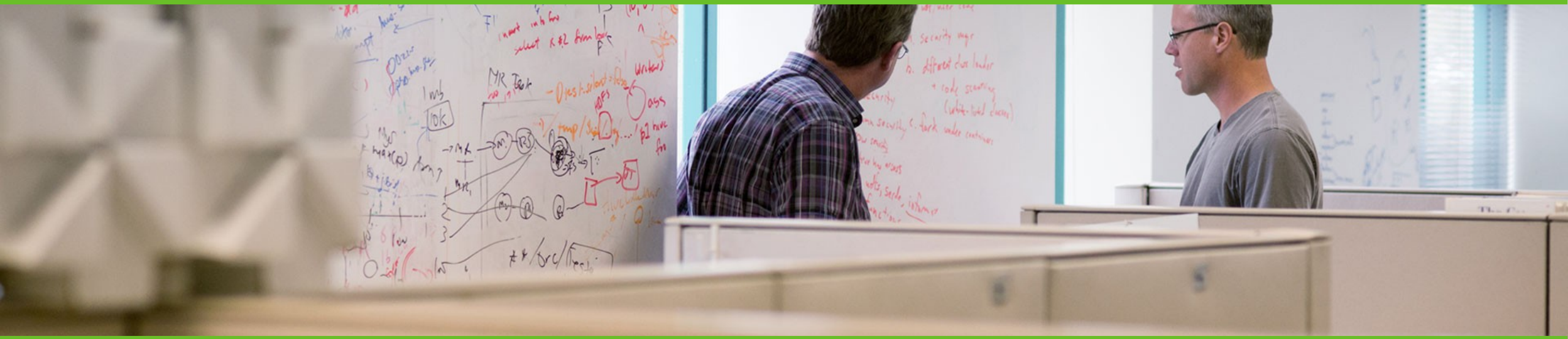
Course Outline

Unit 1 Understanding Big Data
Unit 2 Understanding Hadoop
Unit 3 Data Integration
Unit 4 The Hadoop Ecosystem
Unit 5 Adoption

Hortonworks

Unit 1 – Understanding Big Data

Data, data, and more data



Hortonworks

Who am I?

izhar@abyres.net

Who is Hortonworks?

100% Open Source – Democratized Access to Data

The leaders of Hadoop's development

Hortonworks do Hadoop.

Drive Innovation in the platform – We lead the roadmap

Community driven, Enterprise Focused

Hortonworks

Hortonworks do Hadoop *successfully*.

Contribute to Apache

Support Hadoop adopters

Provide Consulting Services

Deliver Training


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

Who is ABYRES

Hortonworks

Local support and service providers for products from Open Source principals such as Hortonworks and Red Hat

Support and service providers for community Open Source softwares

We do Open Source Solutions

Help our customers move from proprietary lock-in into Open Source

The only Hortonworks System Integrator and Training Partner in South East Asia

Hortonworks

Our Clients

Hortonworks

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

What is Hadoop?
What is everyone talking about?

Hortonworks

Big Data

Hortonworks

“Big Data” is the marketing term of the decade in IT

Hortonworks

What lurks behind the hype is the democratization of Data.

Hortonworks

You need data.

But what do you do with your data now?

Hortonworks

We are obsessive compulsive about collecting and structuring our data.

Hortonworks

Put it away, delete it, tweet it, compress it, shred it, wikileak-it, put it in a database, put it in SAN/NAS, put it in the cloud, hide it in tape…

Hortonworks

You need value from your data. You need to make decisions from your data.

Hortonworks

So what are the problems with Big Data?

Hortonworks

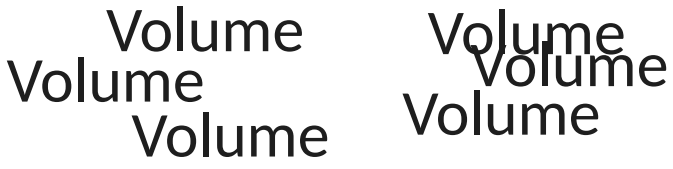
Volume

Volume

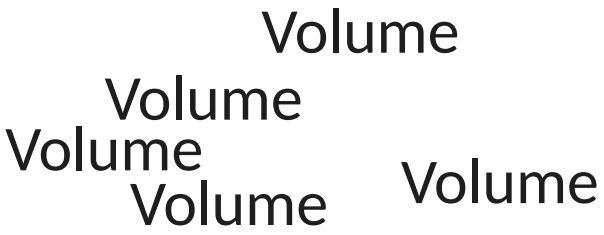
Volume

Volume

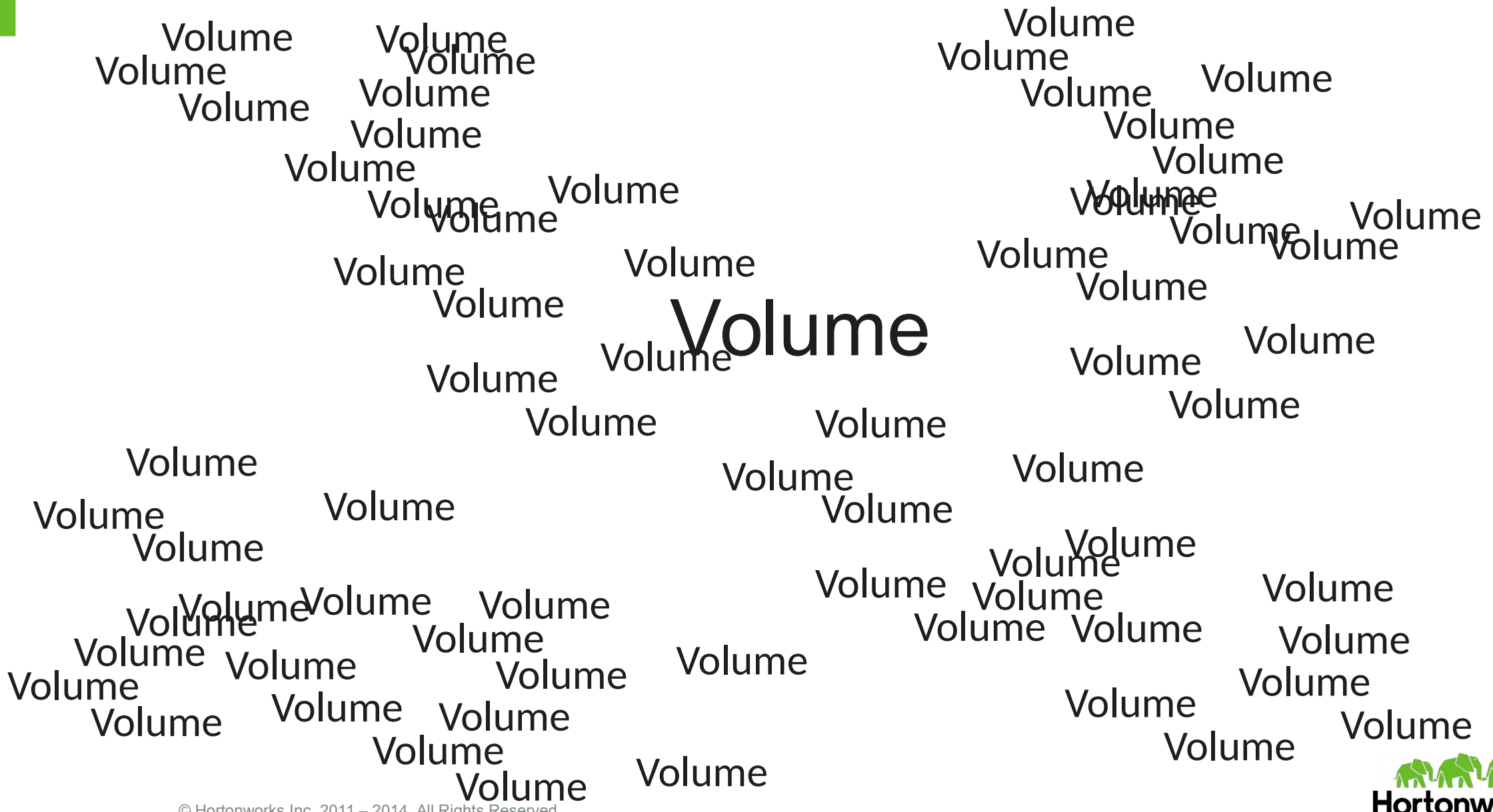
Hortonworks



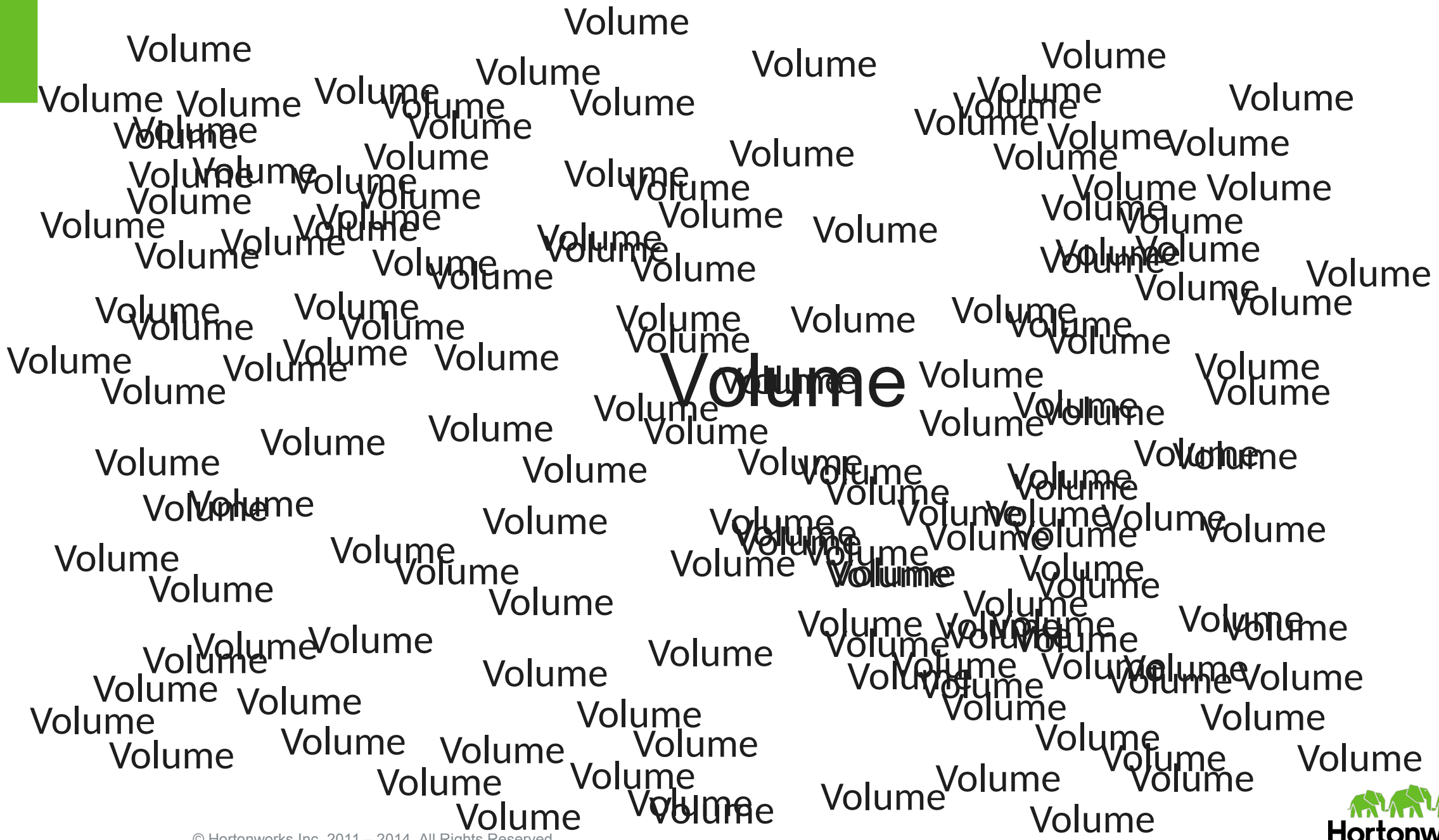
Volume



Hortonworks



Hortonworks



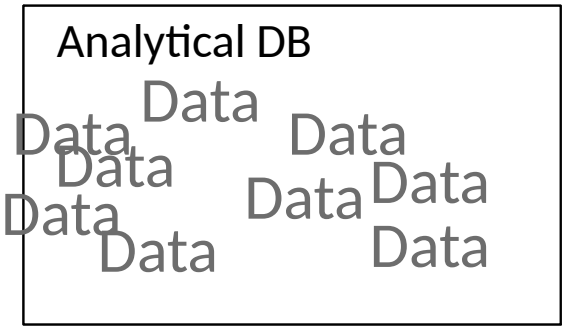
Hortonworks

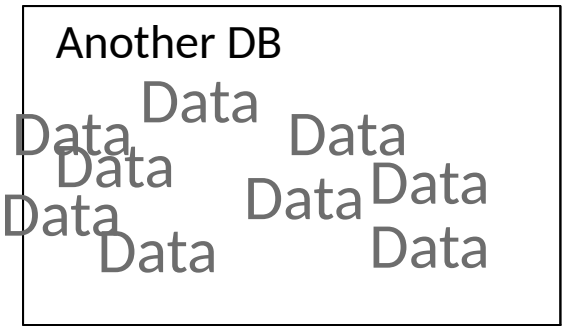
Storage, Management, Processing all become challenges with Data at Volume

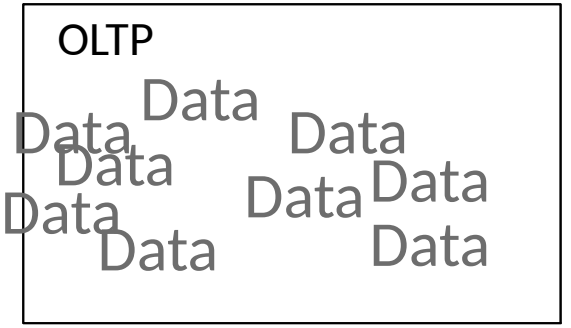
Hortonworks

Traditional technologies adopt a divide, drop, and conquer approach

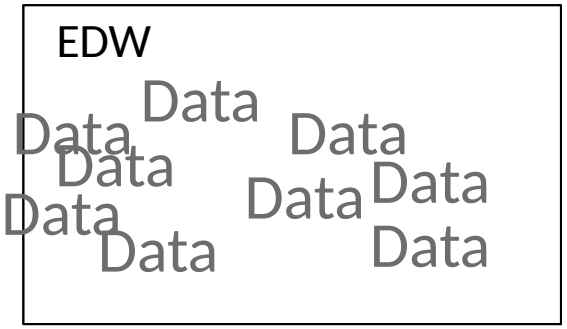
Hortonworks

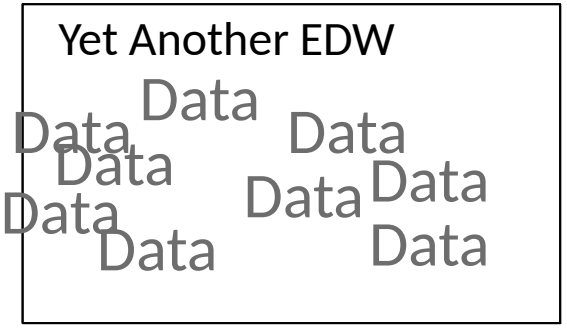




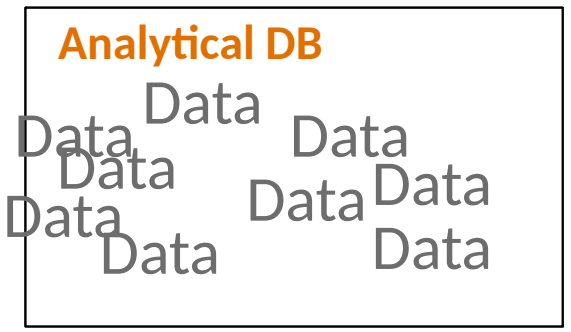


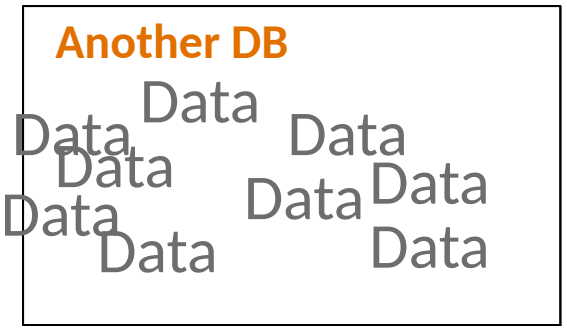
The solution?

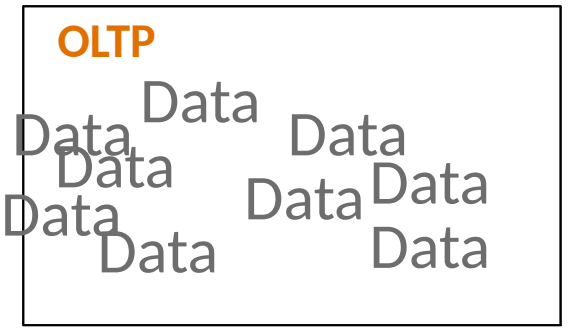




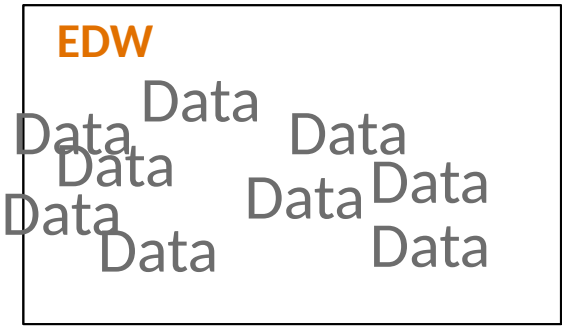
Hortonworks

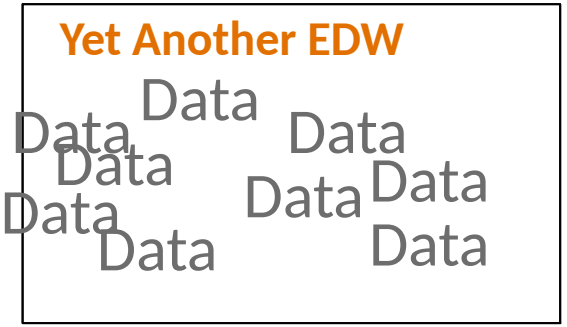






Did you forget something?





Hortonworks

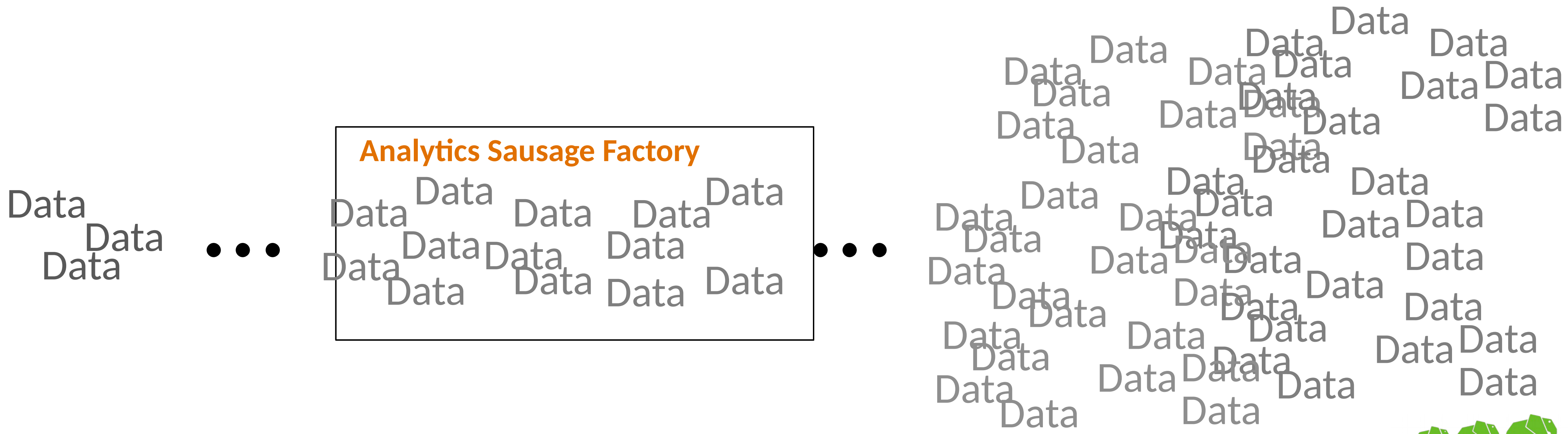
Analyzing the data usually raises more interesting questions…

Hortonworks

…which leads to more data

Hortonworks

Wait, you've seen this before.



Data begets Data.

Hortonworks

What keeps us from our Data?

Hortonworks

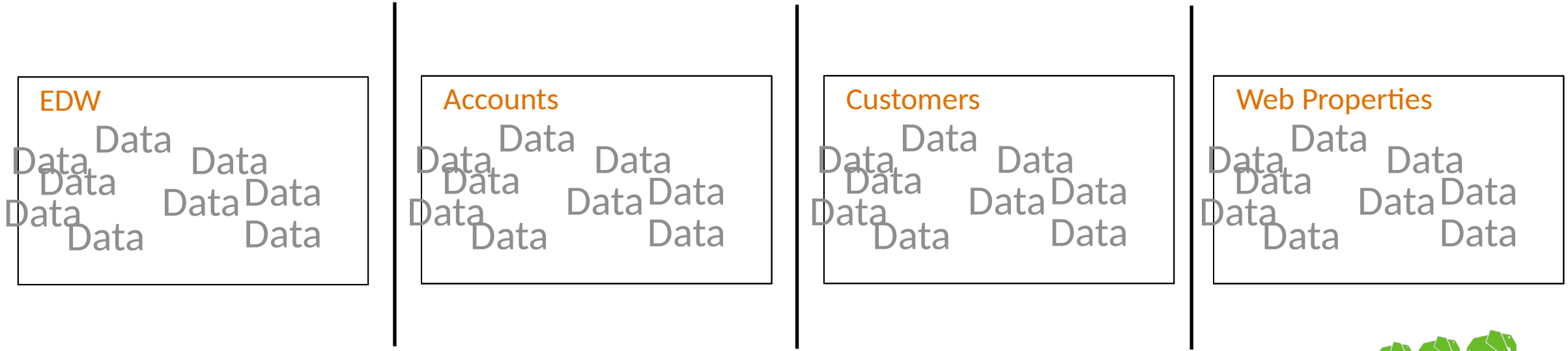
“Prices, Stupid passwords, and Boring Statistics.”

- Hans Rosling

http://www.youtube.com/watch?v=hVimVzgtD6w

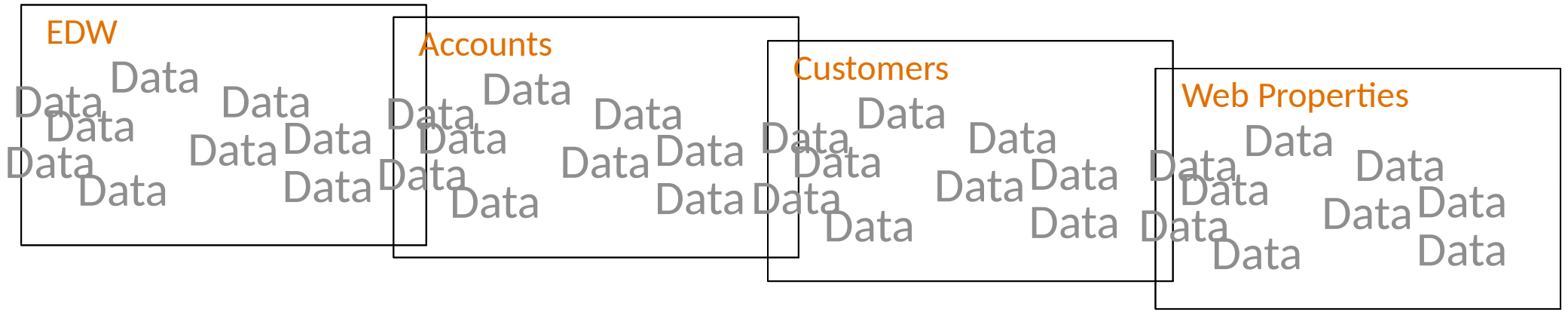
Hortonworks

Your data silos are lonely places.

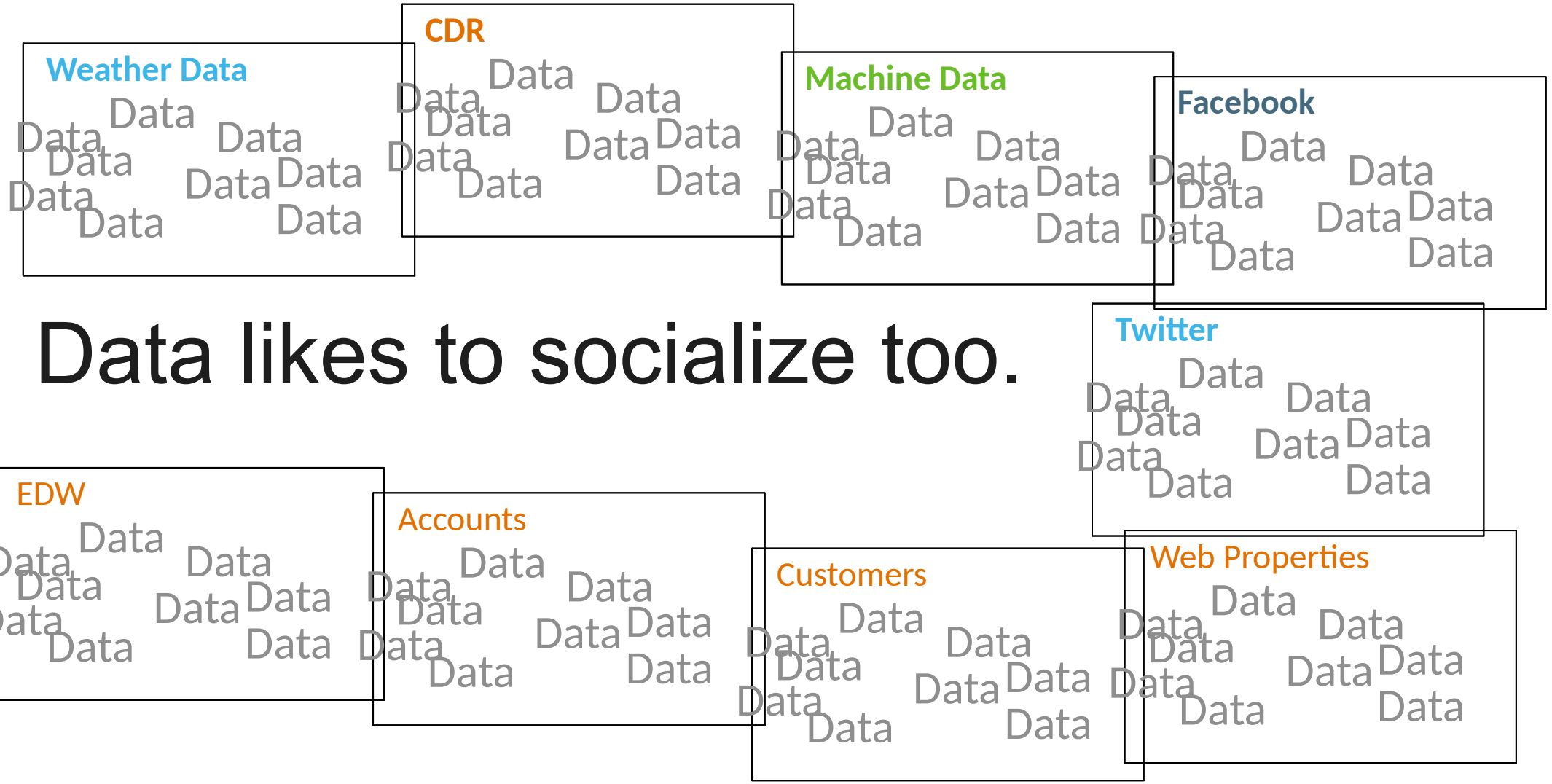


Hortonworks

… Data likes to be together.

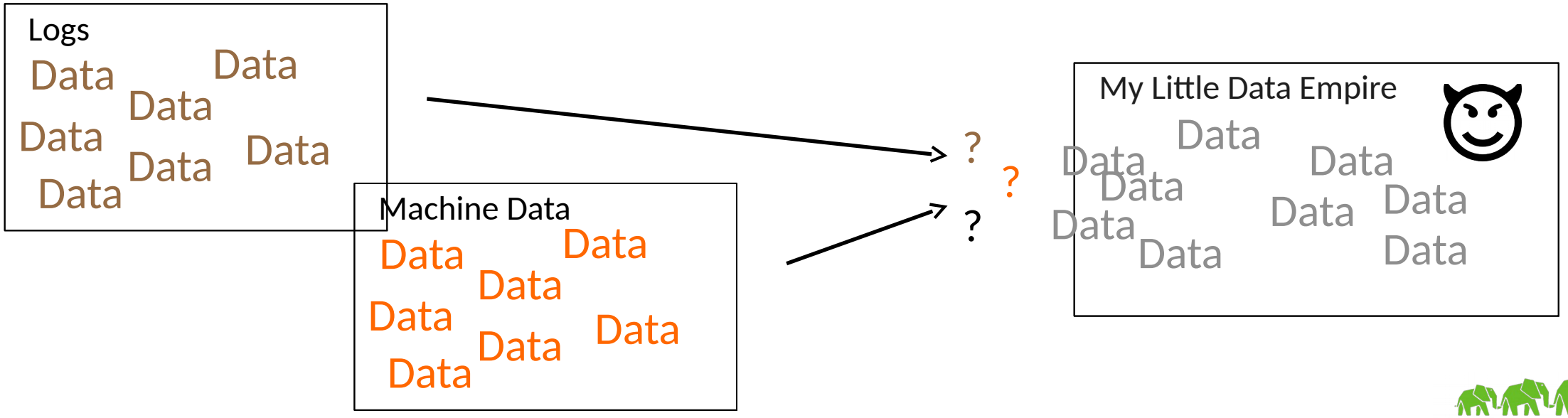


Hortonworks



Hortonworks

New types of data don't quite fit into our pristine view of the world.

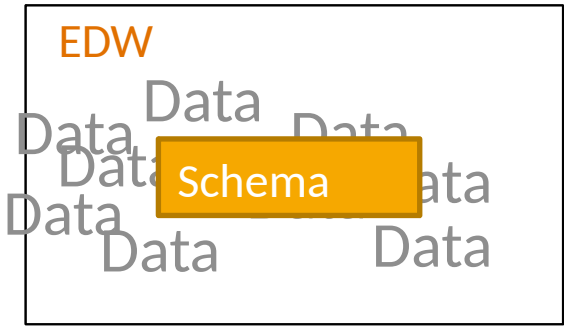


Hortonworks

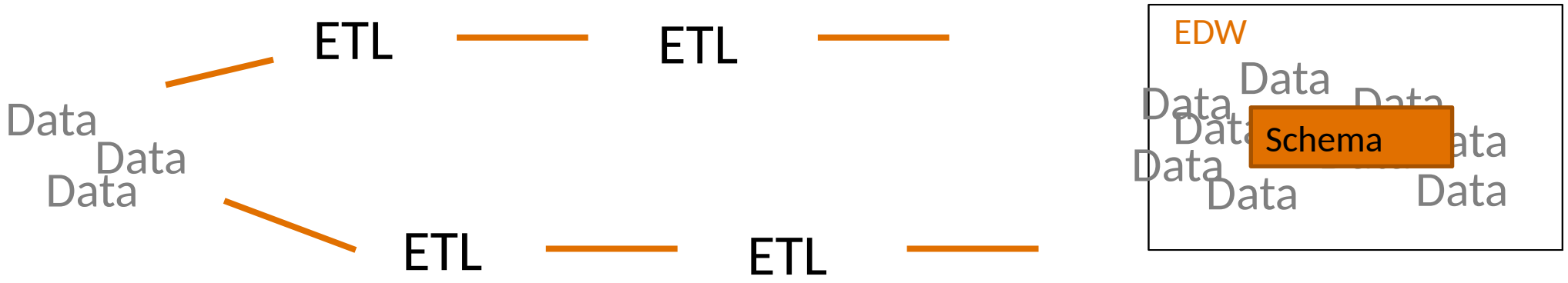
To resolve this, some of us have learned a lot from Lord Of The Rings...

Hortonworks

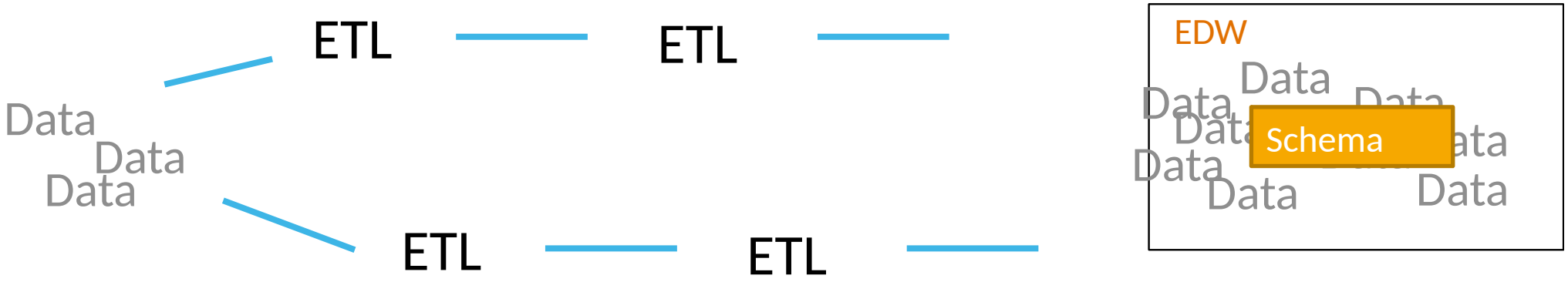
…to create One-Schema-To-Rule-Them-All…



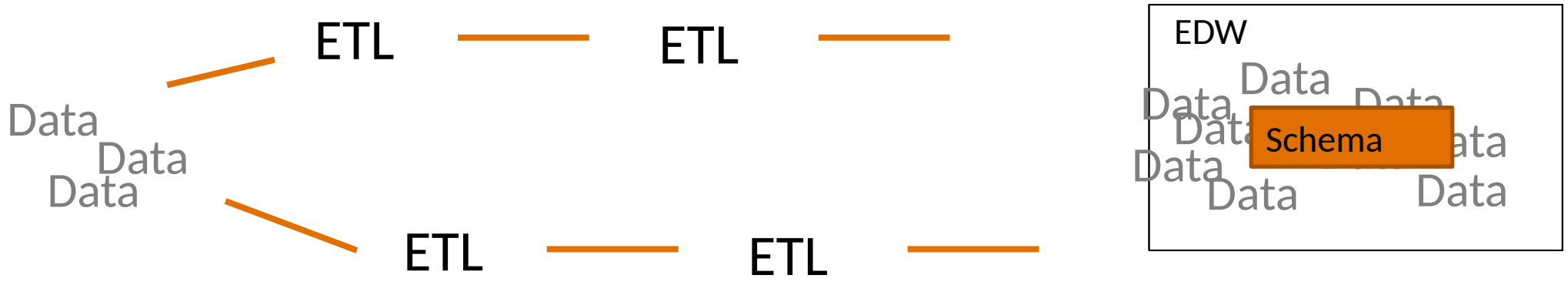
Hortonworks



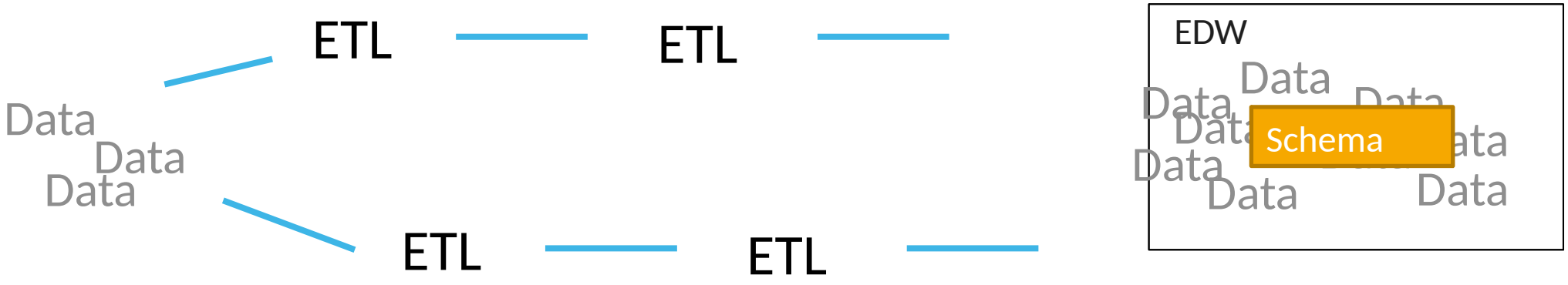
…but that has its problems too.



Hortonworks



Fragile workflows make supporting the analytical models you want expensive and time-consuming.



Hortonworks

What do you want to do with data?

Hortonworks

Marketing Analytics needs data.

Work with the population, not just a sample.

Hortonworks

Town/City

Middle Income Band

Your segmentation today.

Age: 25-30

Female

Product Category Preferences

Male

Hortonworks

Spent 25 minutes looking at tea cozies

GPS coordinates

Walking into Starbucks right now…

Looking to start a business

Unhappy with his cell phone plan

$65-68k per year

Product recommendations

Your segmentation with better data.

A depressed Chicago Blackhawks Fan

Hippie

High tea party

Pregnant

Female

Age: 27 but feels old

Male

Thinking about a new house

Gene Expression for Risk Taker

Products left in basket indicate drunk Amazon shopper

Hortonworks

Pick up all of that data that was prohibitively expensive to store and use.

Hortonworks

Why do viewer surveys…

Hortonworks

…when raw data can tell you what button on the remote was pressed during what commercial for the entire viewer population?

Hortonworks

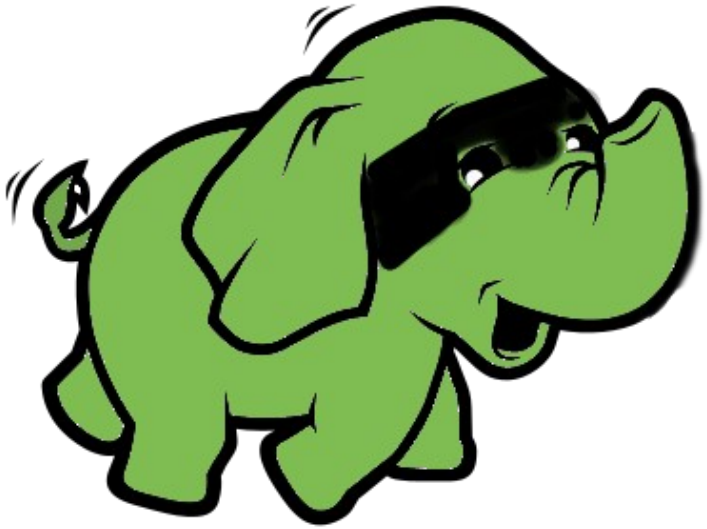
To approach these use cases you need an affordable platform that stores, processes, and analyzes the data.

Hortonworks

So what is the answer?

Hortonworks

Enter the Hadoop.



Hortonworks

Hadoop was created because traditional technologies never cut it for the Internet properties like Google, Yahoo, Facebook, Twitter, and LinkedIn

Hortonworks

Traditional architecture didn't scale enough…

App App App App

SAN

App App App App

SAN

App App App App

SAN

Hortonworks

Databases can become bloated and create problems rather than help solve them.

Hortonworks

$upercomputing

Traditional architectures cost too much at that volume…

$pecial Hardware

$/TB

Hortonworks

So what is the answer?

Hortonworks

If you could design a system that would handle this, what would it look like?

Hortonworks

It would probably need a highly resilient, self-healing, cost-efficient, distributed file system.

Storage **Storage** **Storage**

Storage **Storage** **Storage**

Storage **Storage** **Storage**

Hortonworks

It would probably need a completely parallel processing framework that took ***tasks*** to the ***data***…

Processing Storage	**Processing Storage**	**Processing Storage**
Processing Storage	**Processing Storage**	**Processing Storage**
Processing Storage	**Processing Storage**	**Processing Storage**

Hortonworks

It would probably run on commodity hardware, virtualized machines, and common OS platforms

Processing Storage	**Processing Storage**	**Processing Storage**
Processing Storage	**Processing Storage**	**Processing Storage**
Processing Storage	**Processing Storage**	**Processing Storage**

Hortonworks

It would probably be ***open source*** so innovation could happen as quickly as possible

Hortonworks

It would need a critical mass of users

Hortonworks

Hadoop 2 hit the ground recently: Introducing **YARN**

Hortonworks

{Processing + Storage}

=

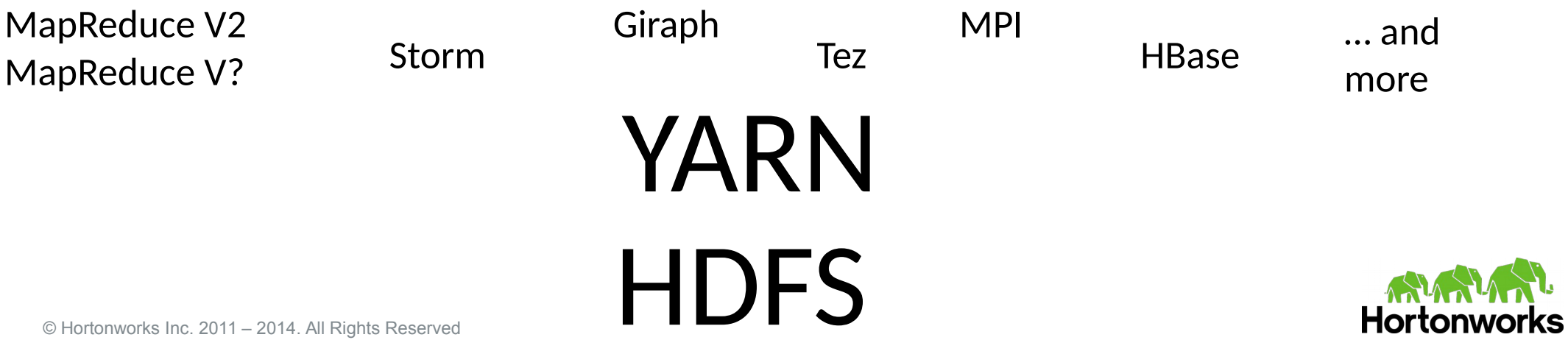
{MapReduce/YARN + HDFS}

=

{Hadoop}

Hortonworks

YARN lets you run more data apps than ever before



Hortonworks

YARN turns Hadoop into a smart phone: An App Ecosystem

hortonworks.com/yarn/

Hortonworks

YARN:

Yeah, we did that too.

hortonworks.com/yarn/

Hortonworks

Hortonworks Data Platform

Apache Ambari
http://incubator.apache.org/ambari

KNOX

APACHE
HBASE

Apache
Solr

Spark

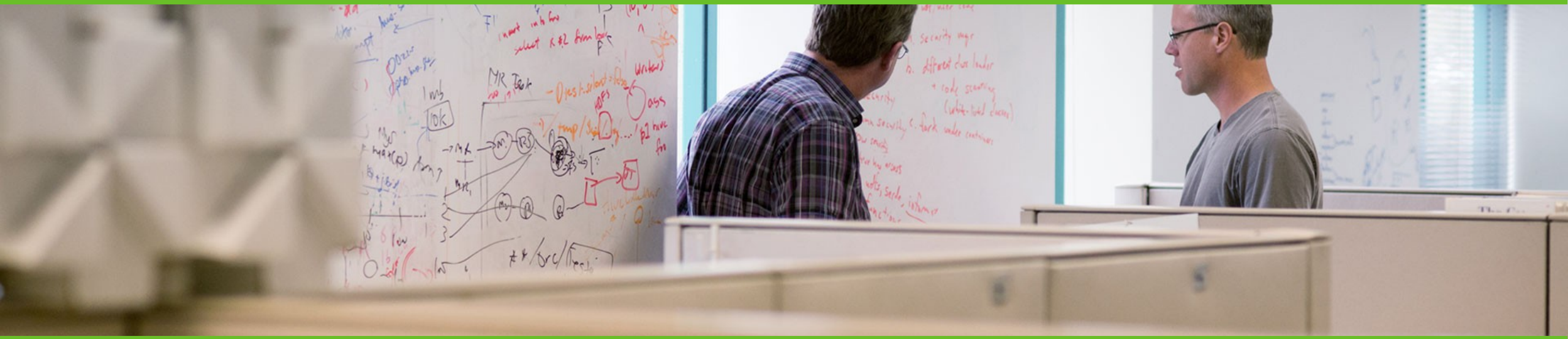
Phoenix

accumulo™

runs on

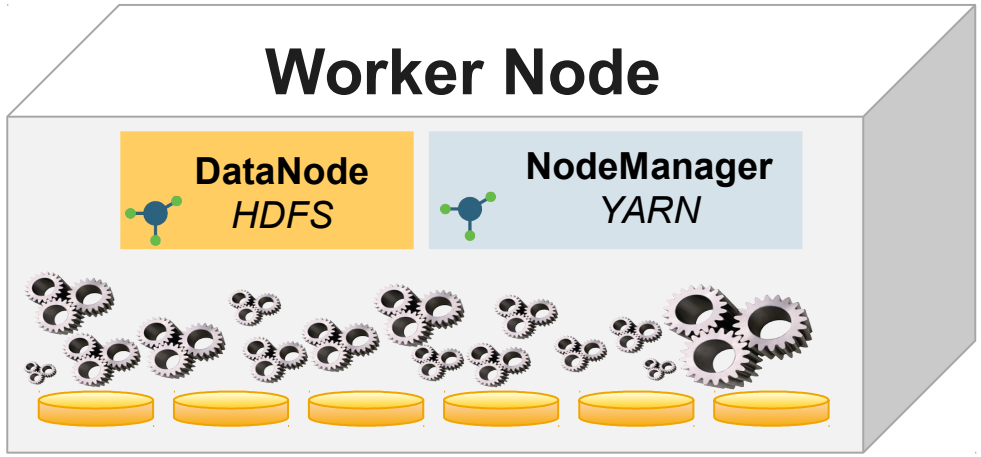
Unit 2 – Understanding Hadoop

Storage and Processing



Hortonworks

Core hadoop



Disk, CPU, Memory

+

NameNode
HDFS

/directory/structure/in/memory.txt

ResourceManager
YARN

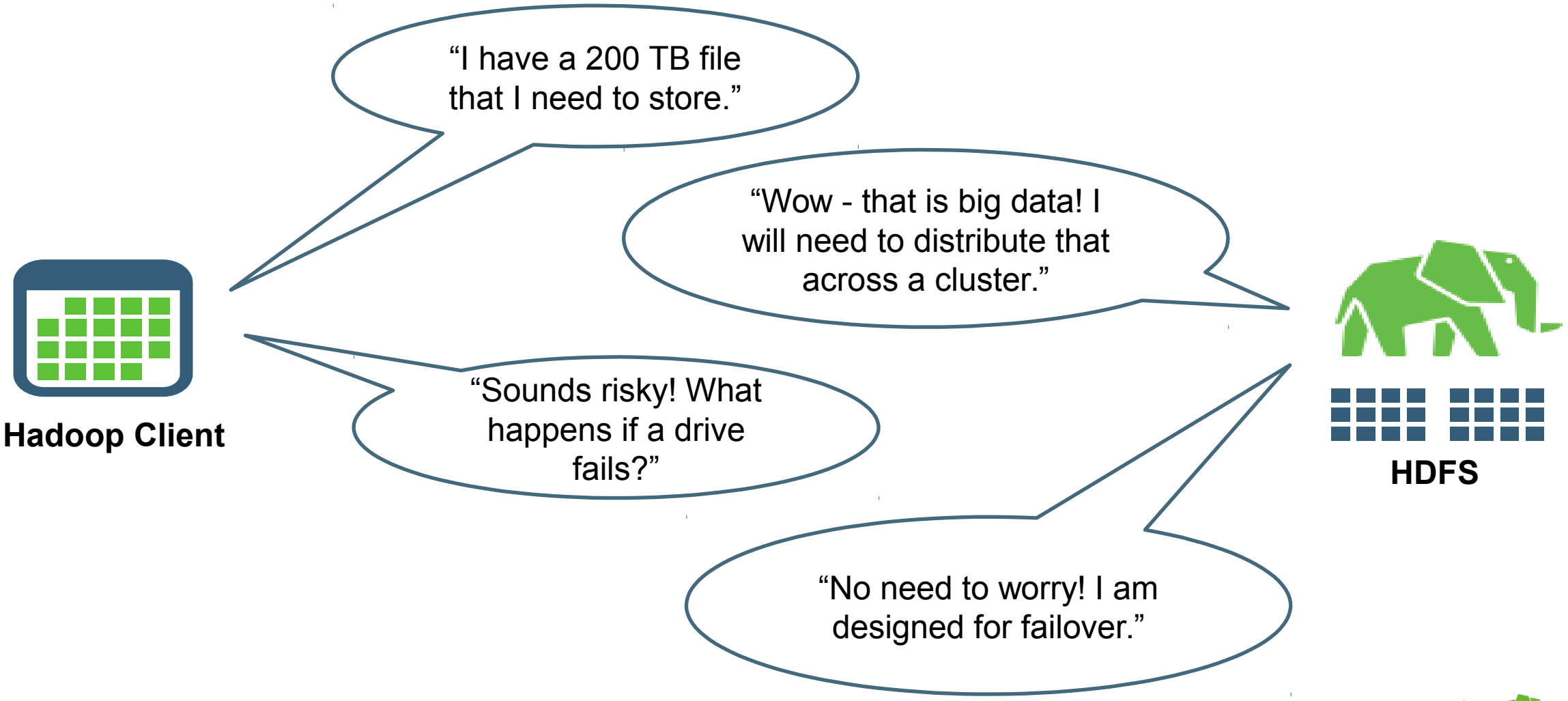
Resource management + scheduling

Hadoop daemon

User application

Hortonworks

What is HDFS?

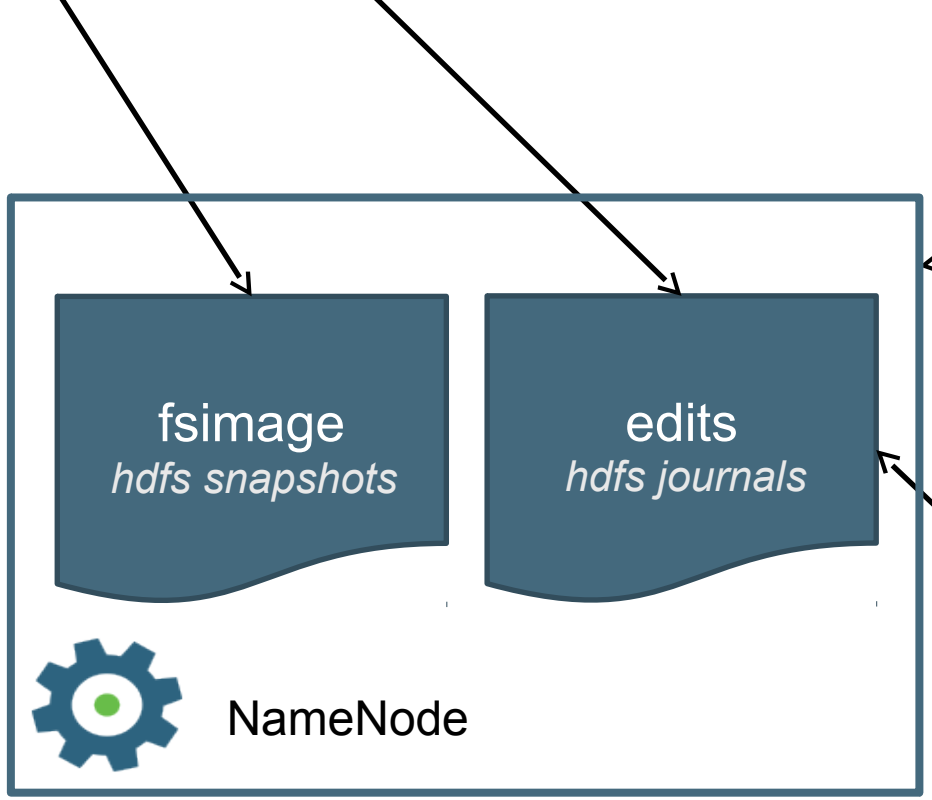


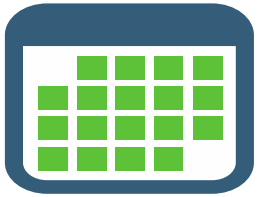
Hortonworks

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.





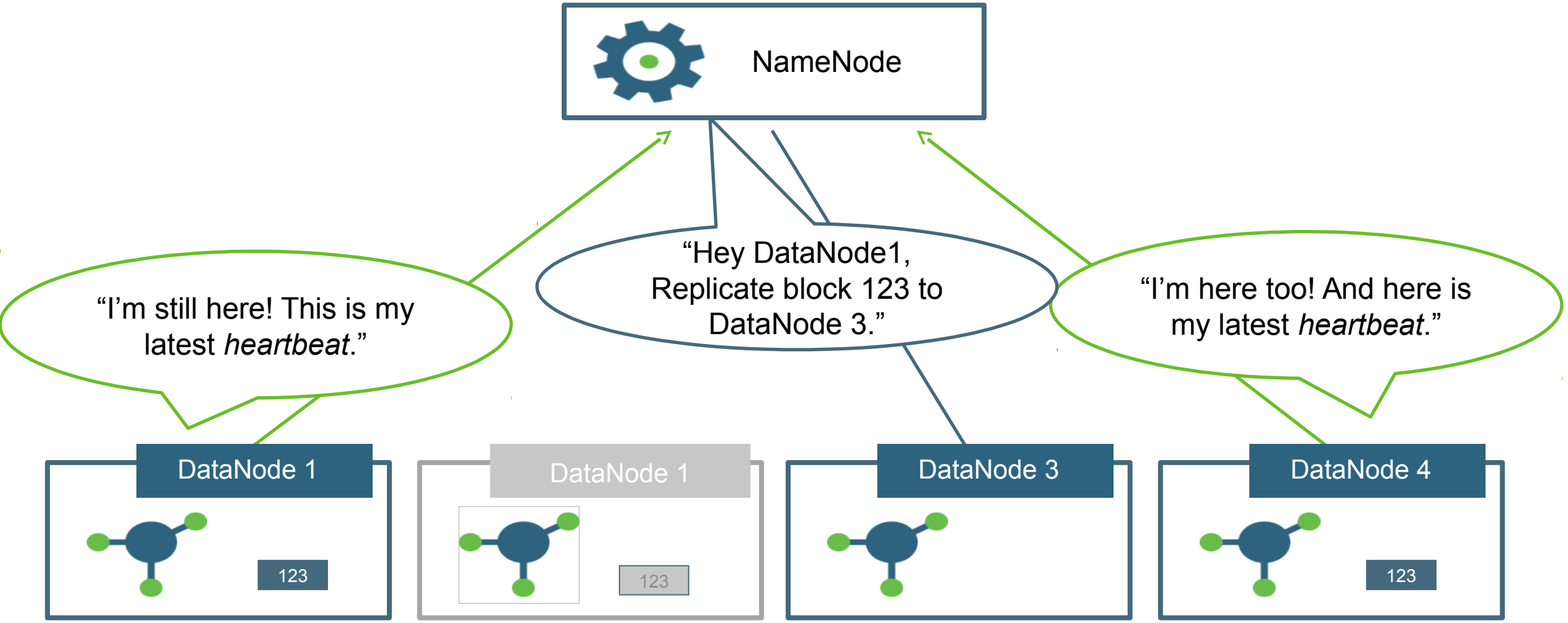
3. A client application creates a new file in HDFS.

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

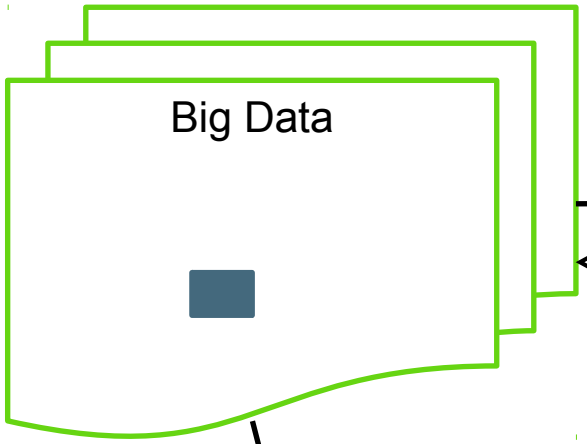
4. The NameNode logs that transaction in the **edits** file.

Hortonworks

The DataNodes



Hortonworks

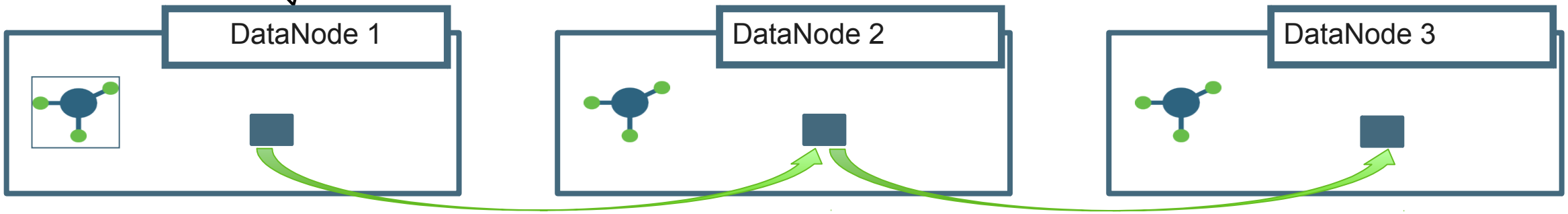


1. Client sends a request to the NameNode to add a file to HDFS.

1. NameNode gives client a *lease* to the file path.

1. For every block, the client will request the NameNode to provide a new *blockid* and a list of *destination* DataNodes.

2. The client will write the block directly to the first DataNode in the list.



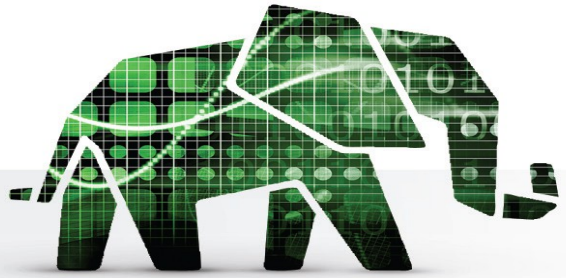
1. The first DataNode pipelines the replication to the next DataNode in the list.

Hortonworks

Hortonworks UNIVERSITY

Demonstration

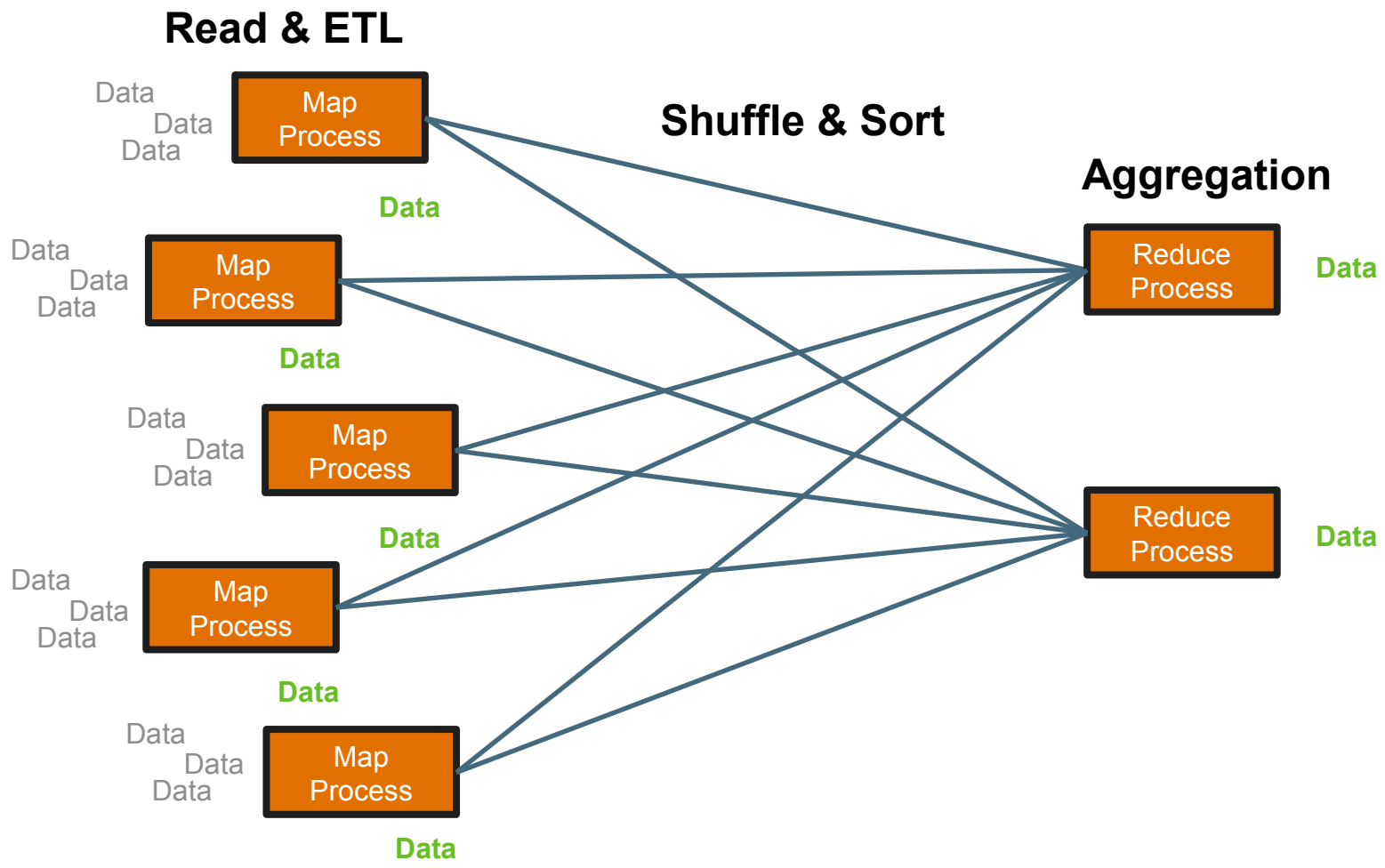
Understanding HDFS & Block Storage



Hortonworks

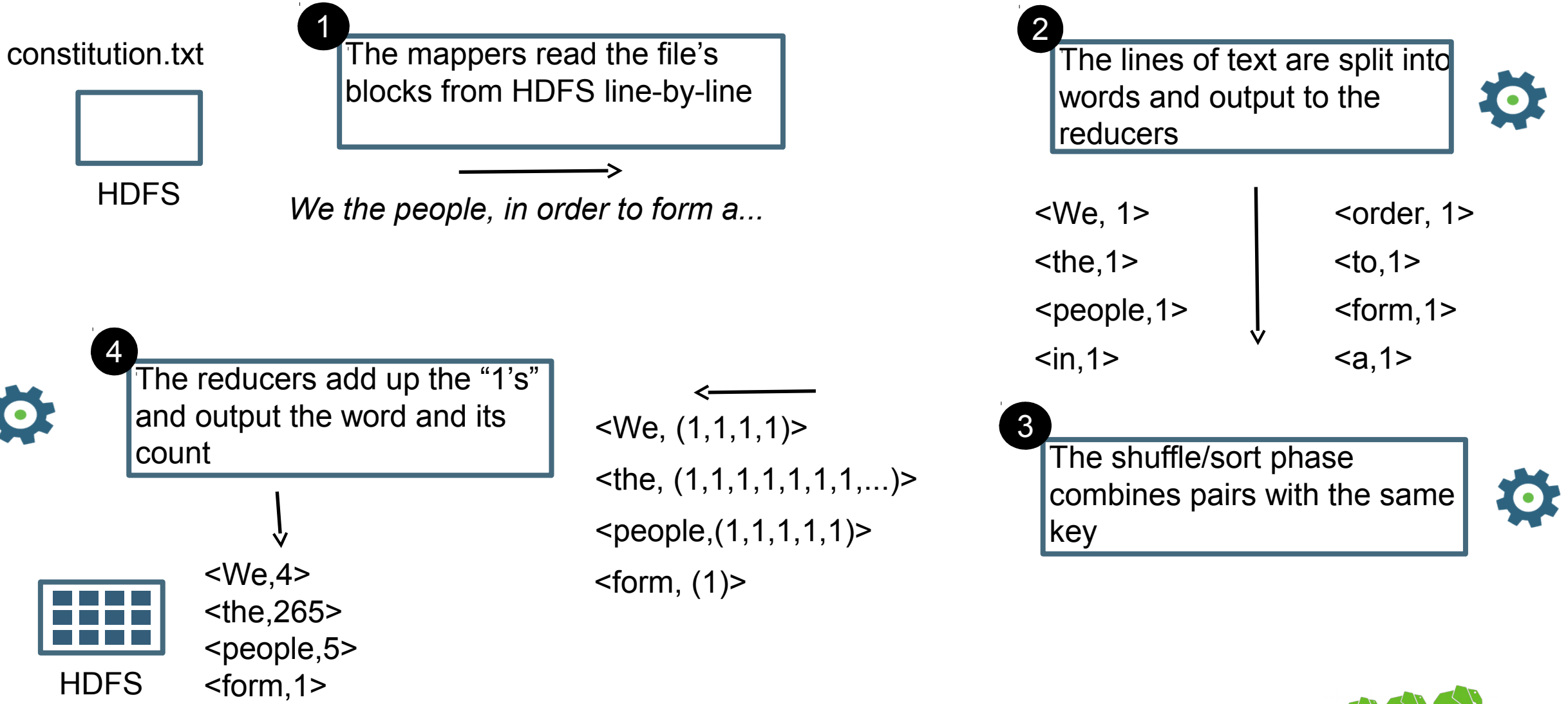
What is MapReduce?

Break a large problem into sub-solutions



Hortonworks

WordCount in MapReduce



Hortonworks

Map Phase

Shuffle/Sort

Reduce Phase

```
SELECT word,
COUNT(*)
FROM constitution
WHERE….
```

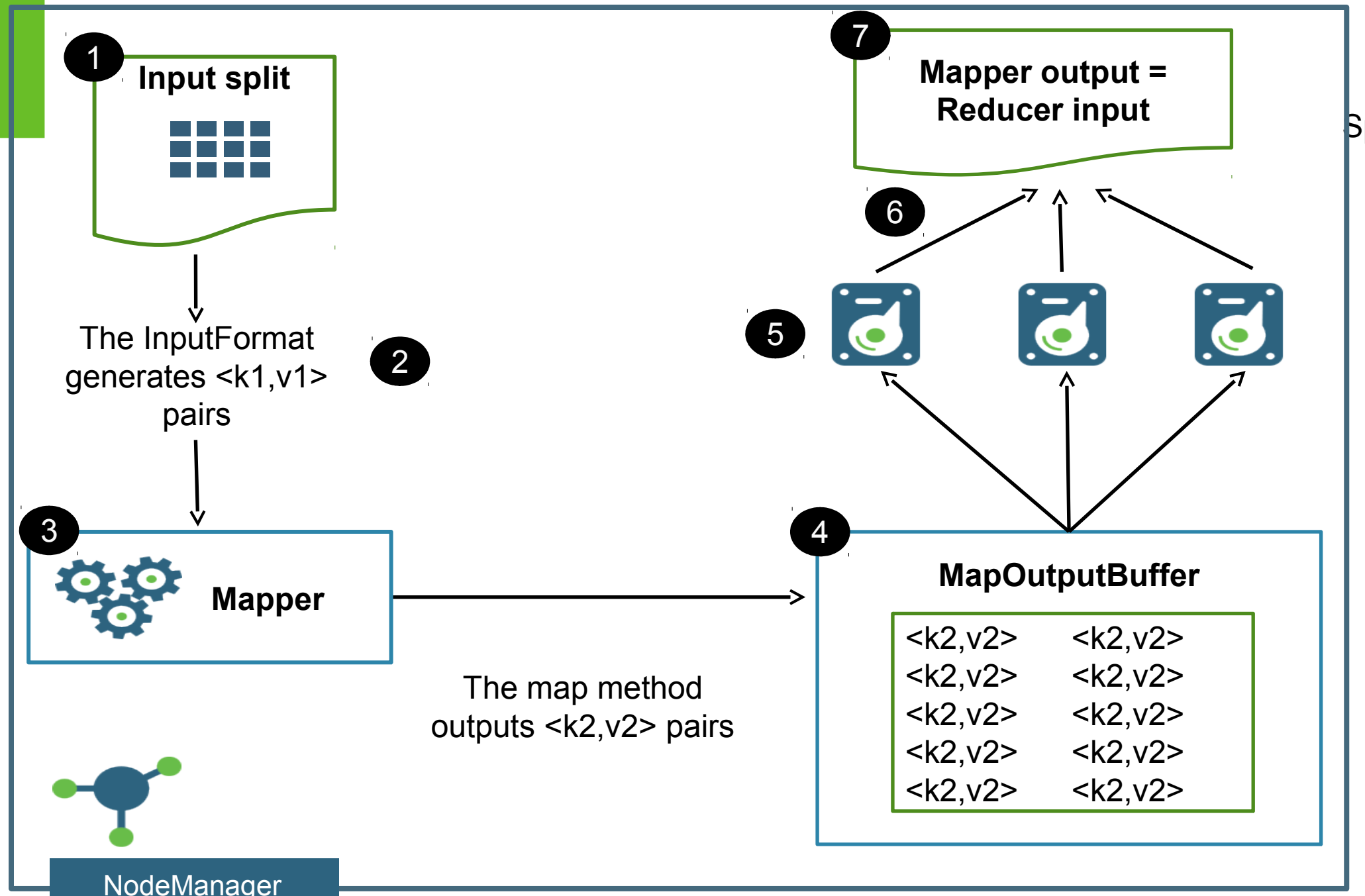
```
GROUP BY
```

Hive/Pig compile as Reduce side function

```
ORDER BY
JOIN
DISTINCT
```

Examples of more Reduce side functions

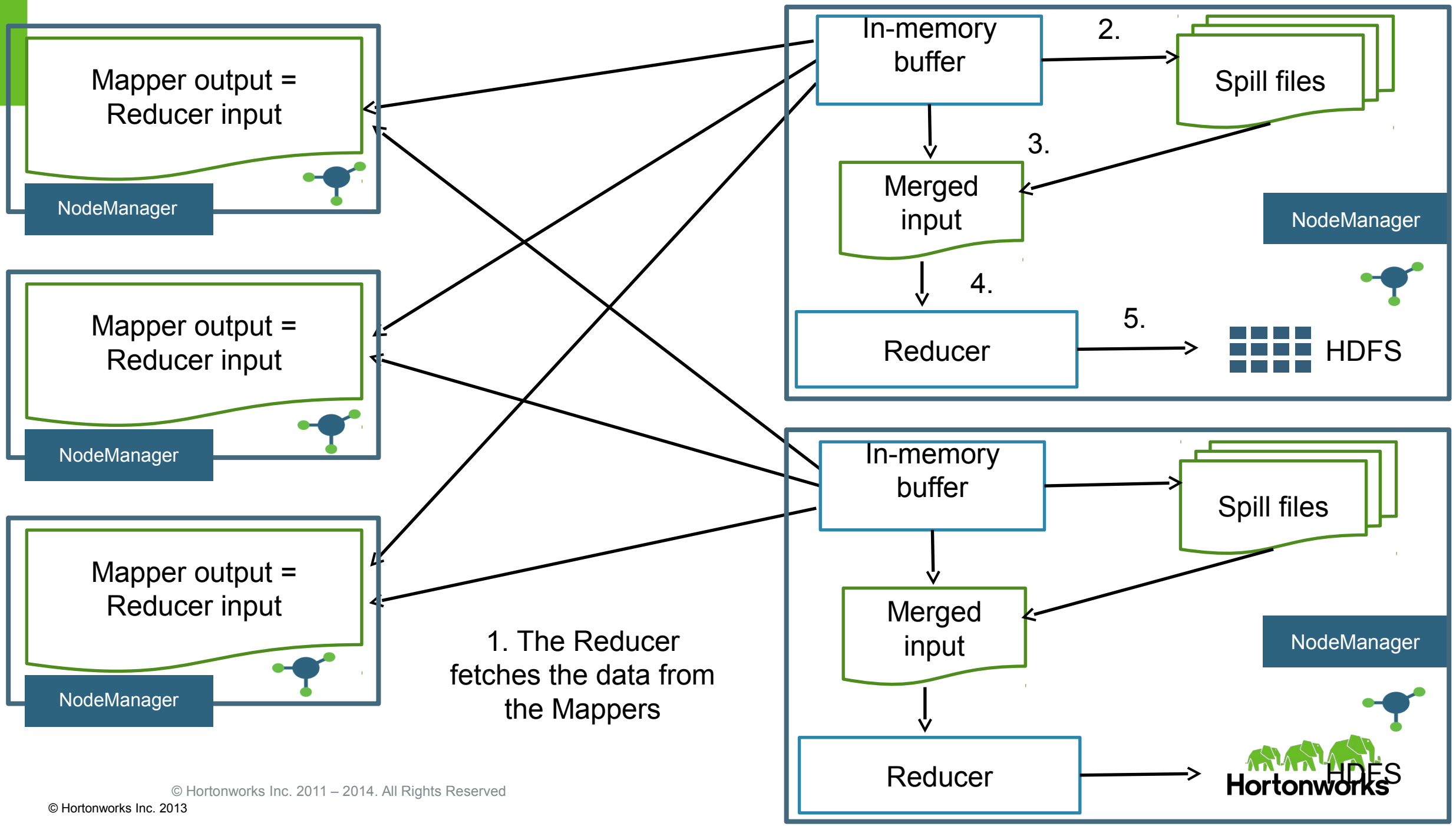
Hortonworks



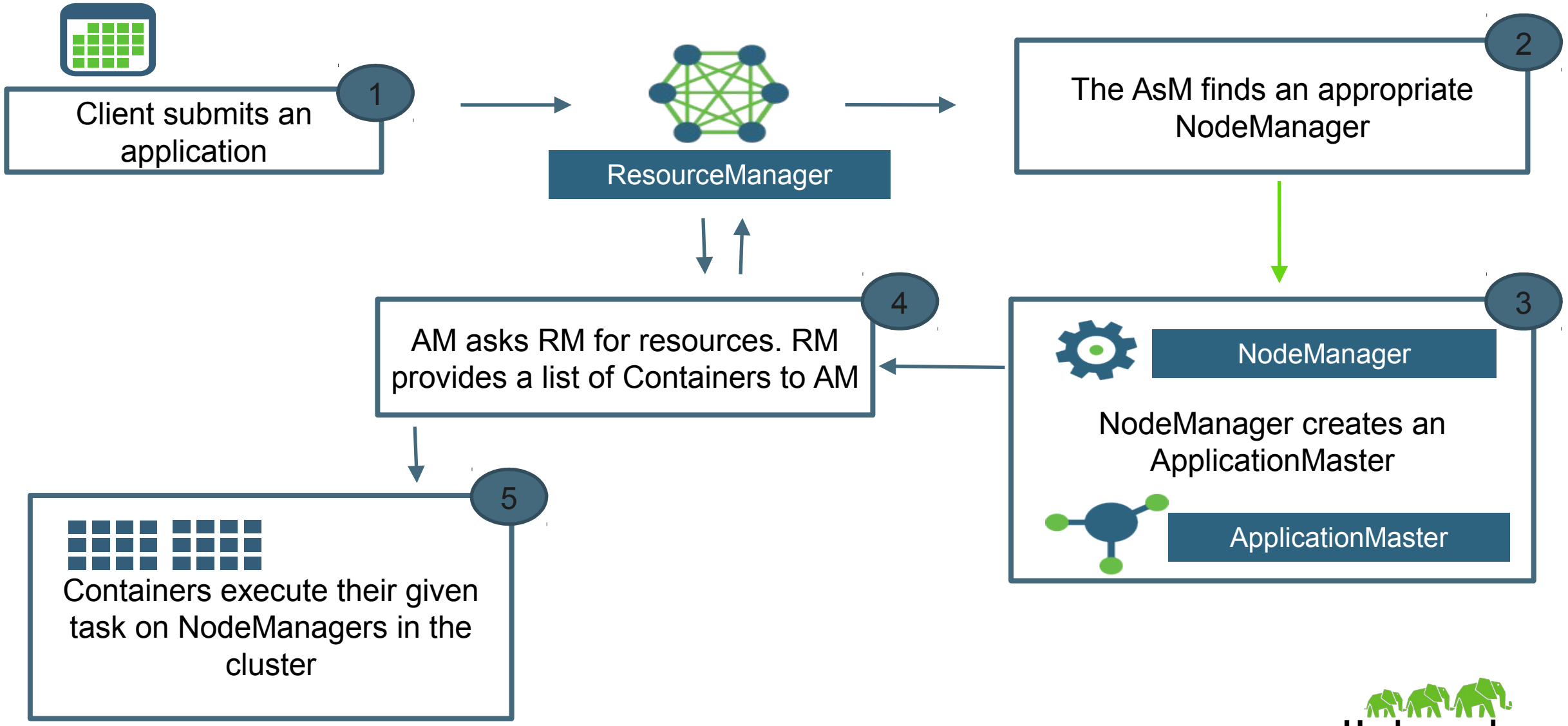
Spill files are merged into a single file

Records are sorted and spilled to disk when the buffer reaches a threshold

Hortonworks

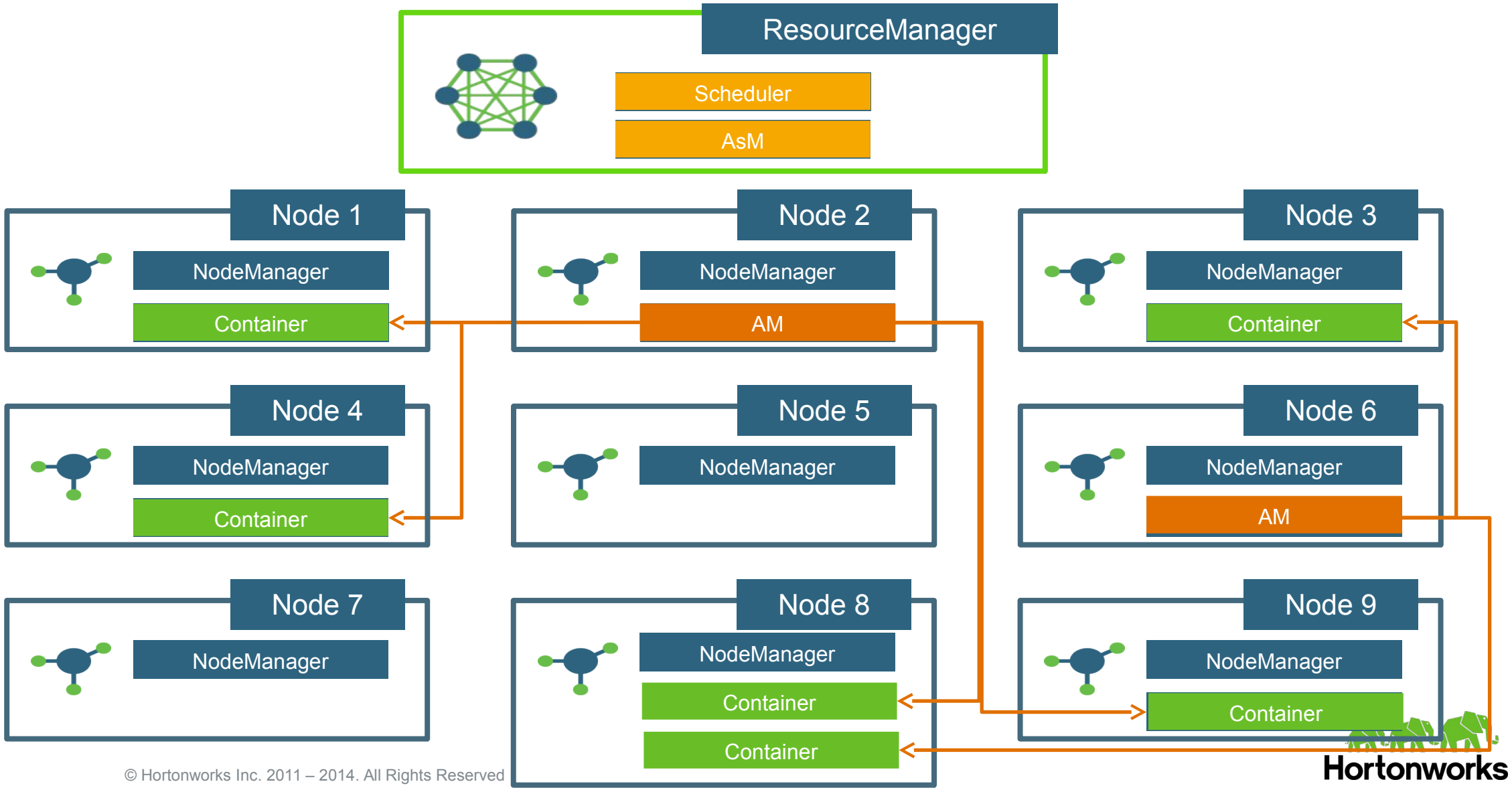


Lifecycle of a YARN Application



Hortonworks

A Cluster View Example



Hortonworks

Hortonworks UNIVERSITY

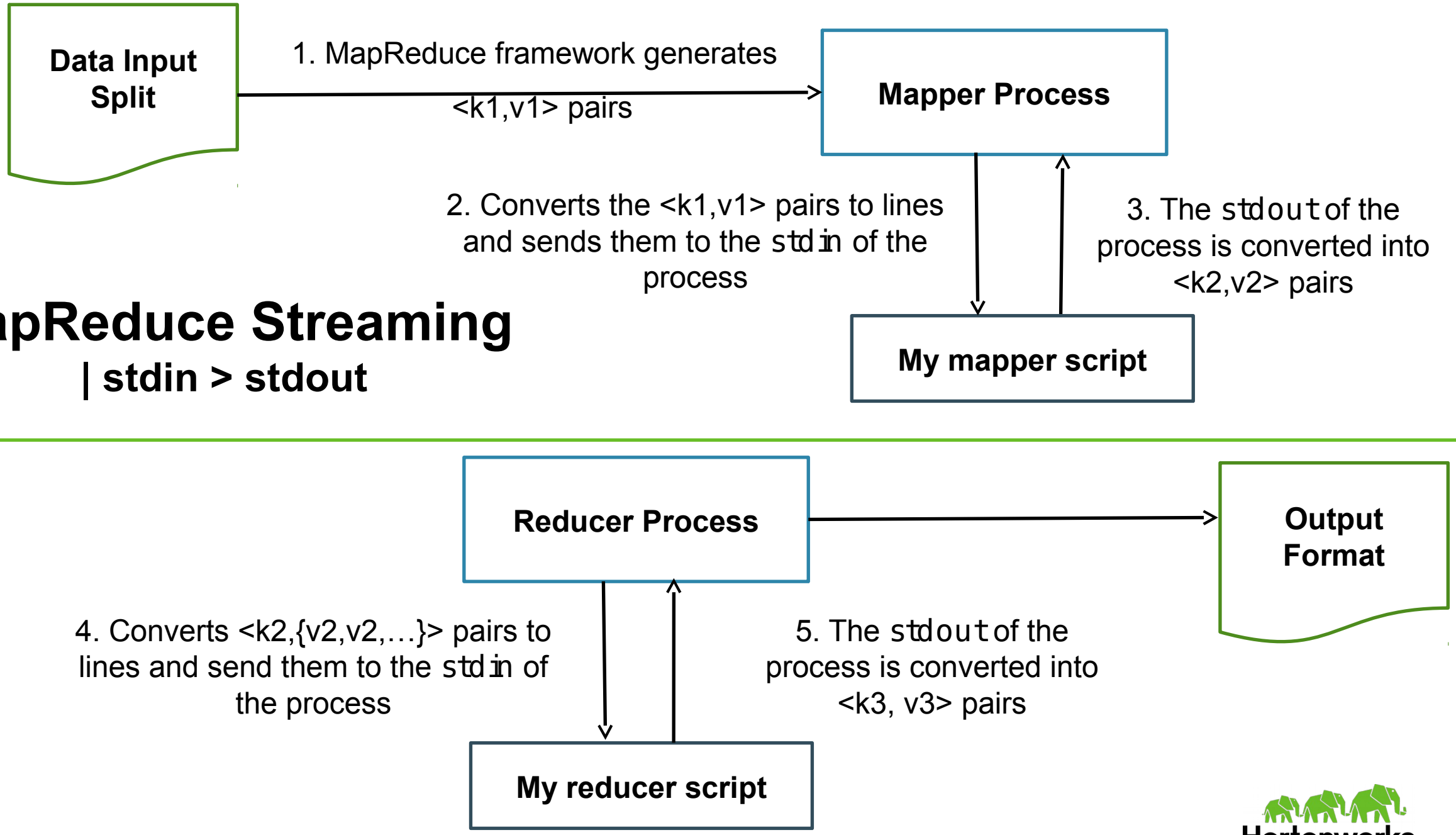
Demonstration

WordCount MapReduce

Hortonworks

MapReduce Streaming

| stdin > stdout



Hortonworks

Hortonworks

HDP 2.1
Sandbox

Welcome to the Hortonworks Data Platform (HDP) 2.1 Sandbox

HDP 2.1 is a major release that delivers required enterprise functionality for data management, data access, data governance, integration, security and operations developed and delivered completely in the open. Incorporating the very latest community innovations across all Apache Hadoop projects, HDP 2.1 provides the foundational platform for organizations looking to incorporate Hadoop in a modern data architecture.

Get started with Hadoop

Try New Features

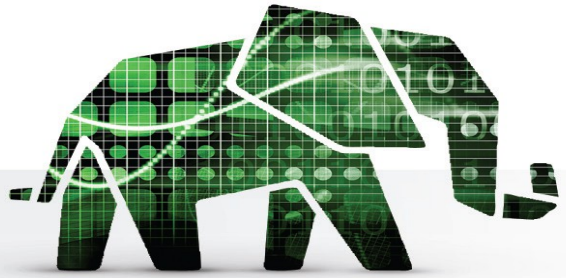
Dive right in

Hortonworks

Hortonworks UNIVERSITY

Demonstration

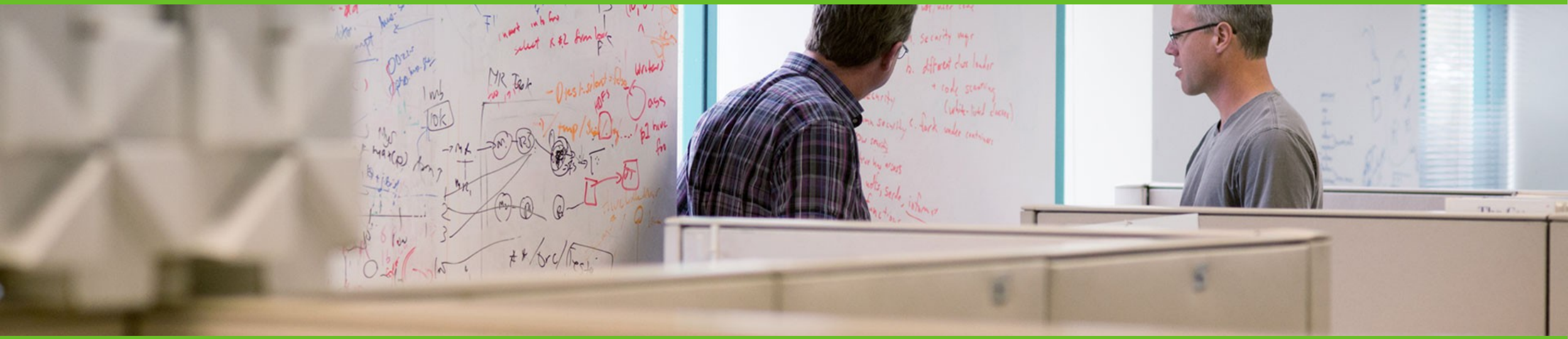
The Hortonworks Sandbox



Hortonworks

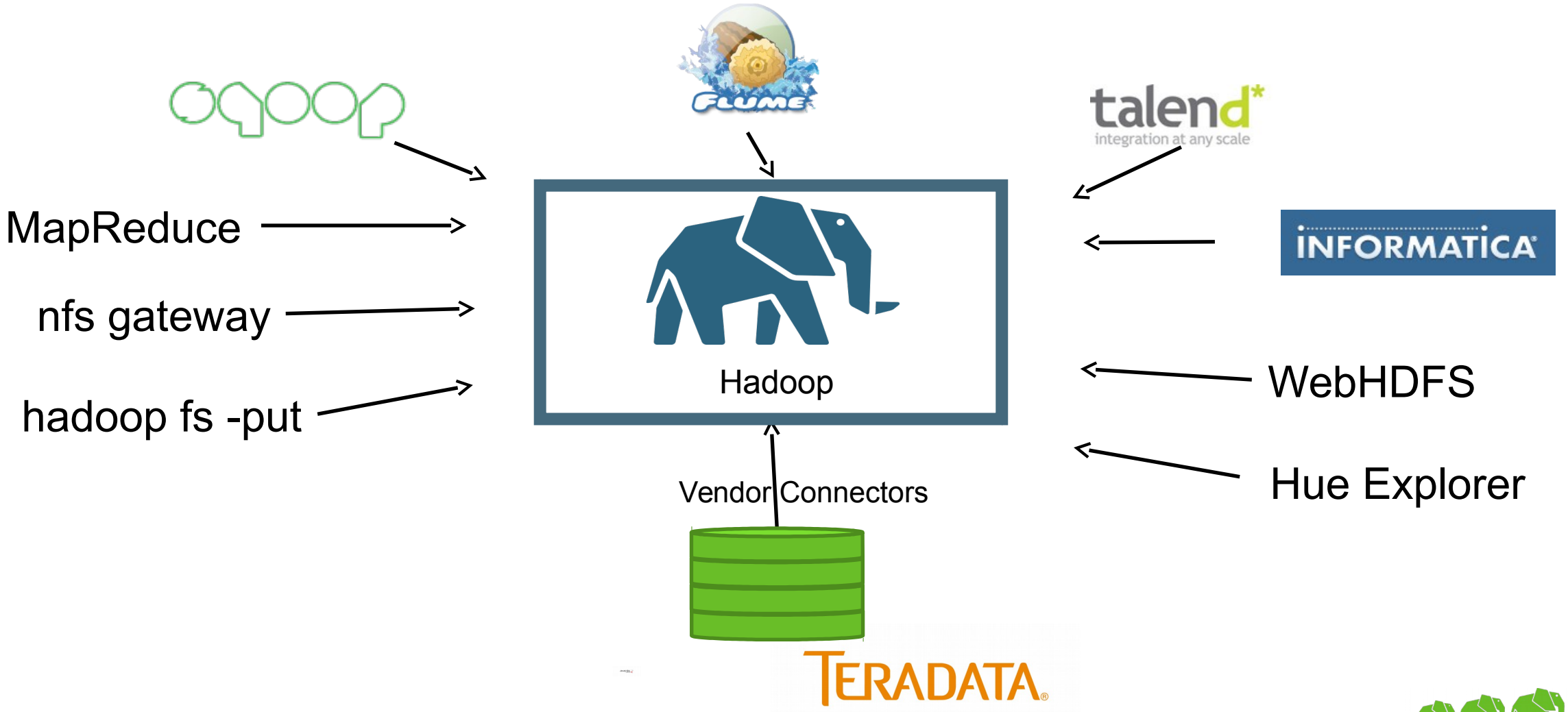
Unit 3 – Data Integration

Loading data into Hadoop



Hortonworks

Options for Data Input



Hortonworks

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
```

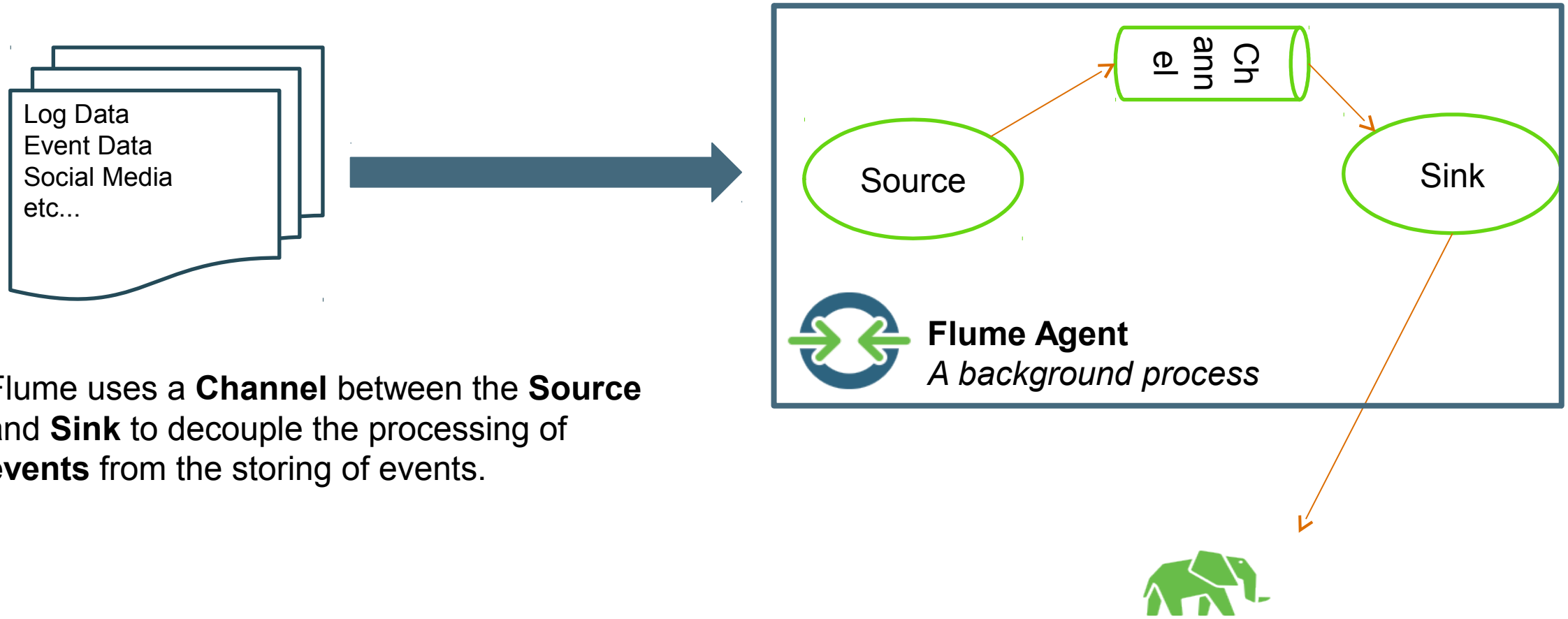

Hortonworks

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

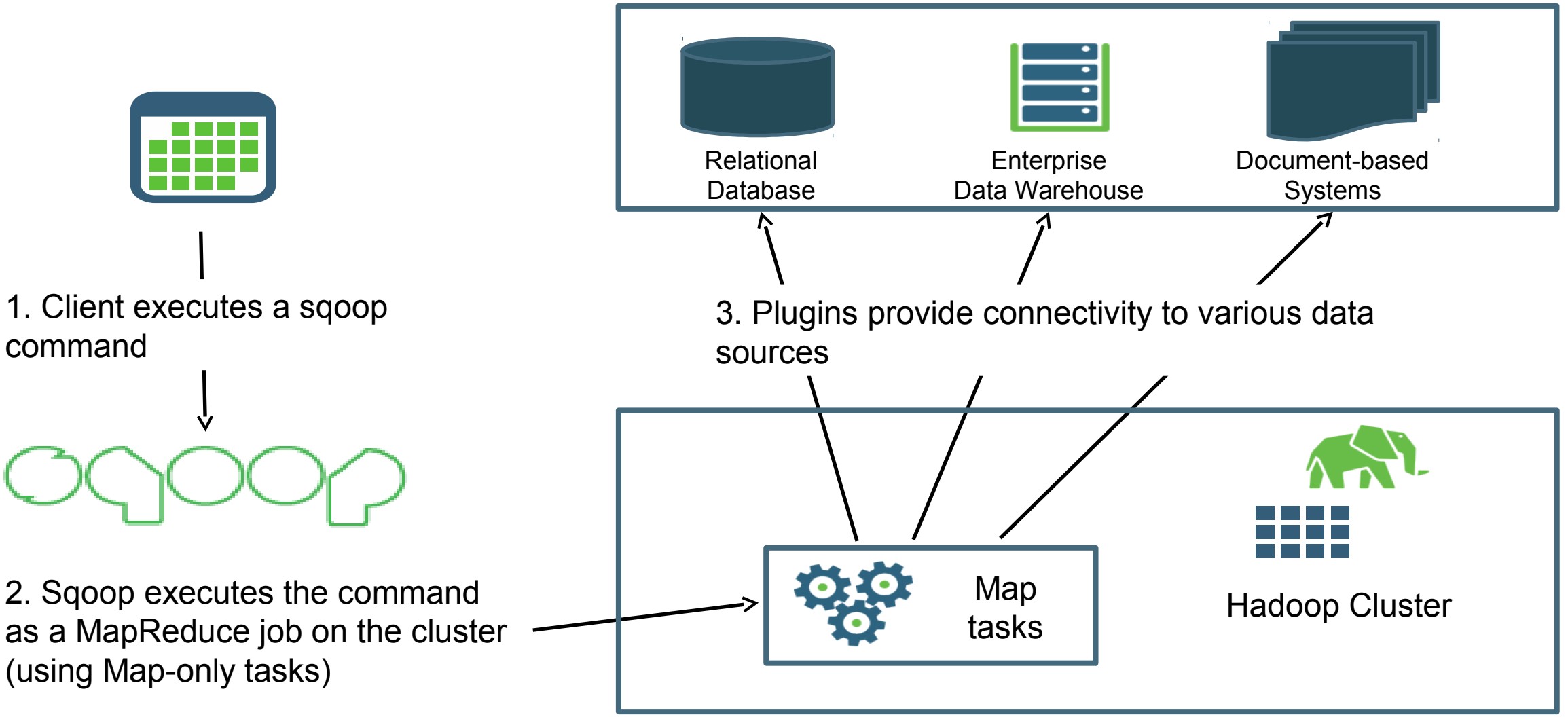
Overview of Flume: Data Streaming



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

Hortonworks

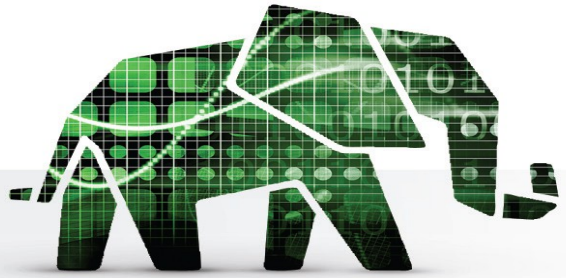
Overview of Sqoop: Database Import/Export



Hortonworks

Hortonworks UNIVERSITY

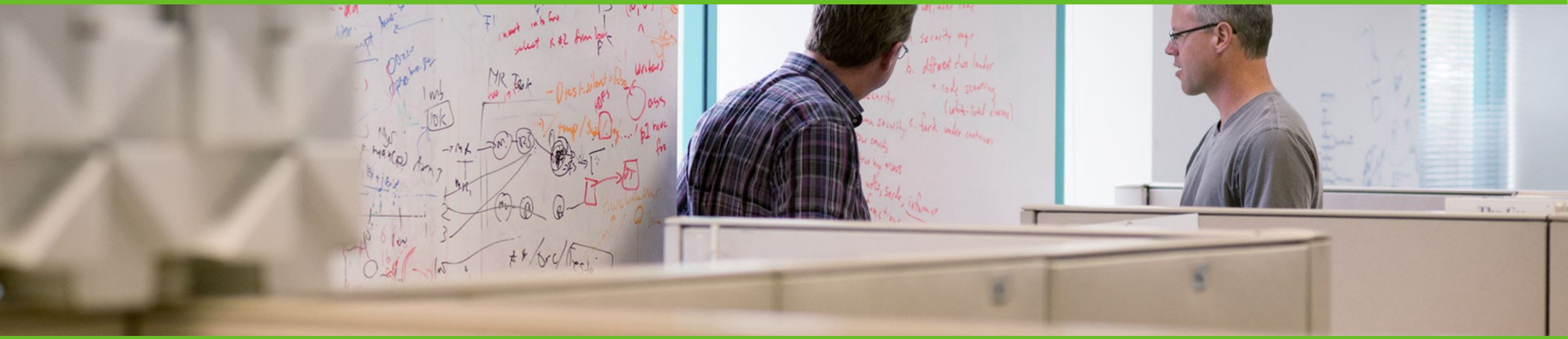
Demonstration
Using Sqoop



Hortonworks

Unit 4 – The Hadoop Ecosystem

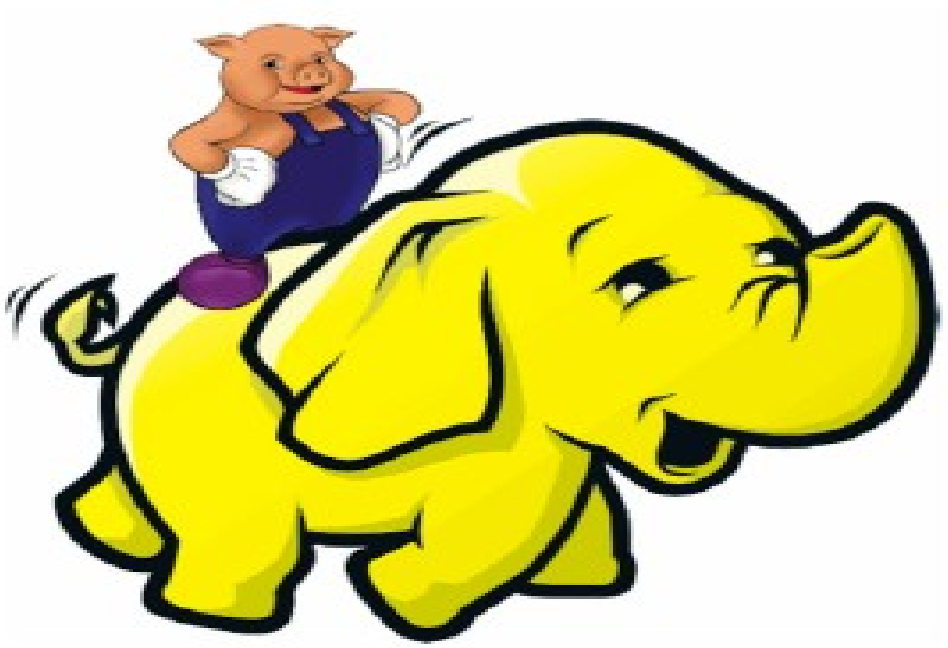
Tour of the Hadoop Ecosystem



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

```
1  users = LOAD 'input/users' USING PigStorage(',')
2          AS (name:chararray, age:int);
3
4  filtrd = FILTER users BY age >= 18 and age <= 25;
5
6  pages = LOAD 'input/pages' USING PigStorage(',')
7          AS (user:chararray, url:chararray);
8
9  jnd = JOIN filtrd BY name, pages BY user;
10
11 grpd = GROUP jnd BY url;
12
13 smmd = FOREACH grpd GENERATE group, COUNT(jnd) AS clicks;
14
15 srtd = ORDER smmd BY clicks DESC;
16
17 top5 = LIMIT srtd 5;
18
19 STORE Top5 INTO 'output/top5sites' USING PigStorage(',');
```


Hortonworks

Hadoop Ecosystem: **Hive**

Use existing SQL tools and existing SQL processes

Store and query all data in Hive

Weblog

Sensor

Mobile

Operational/MPP

Hortonworks

What is Hive?

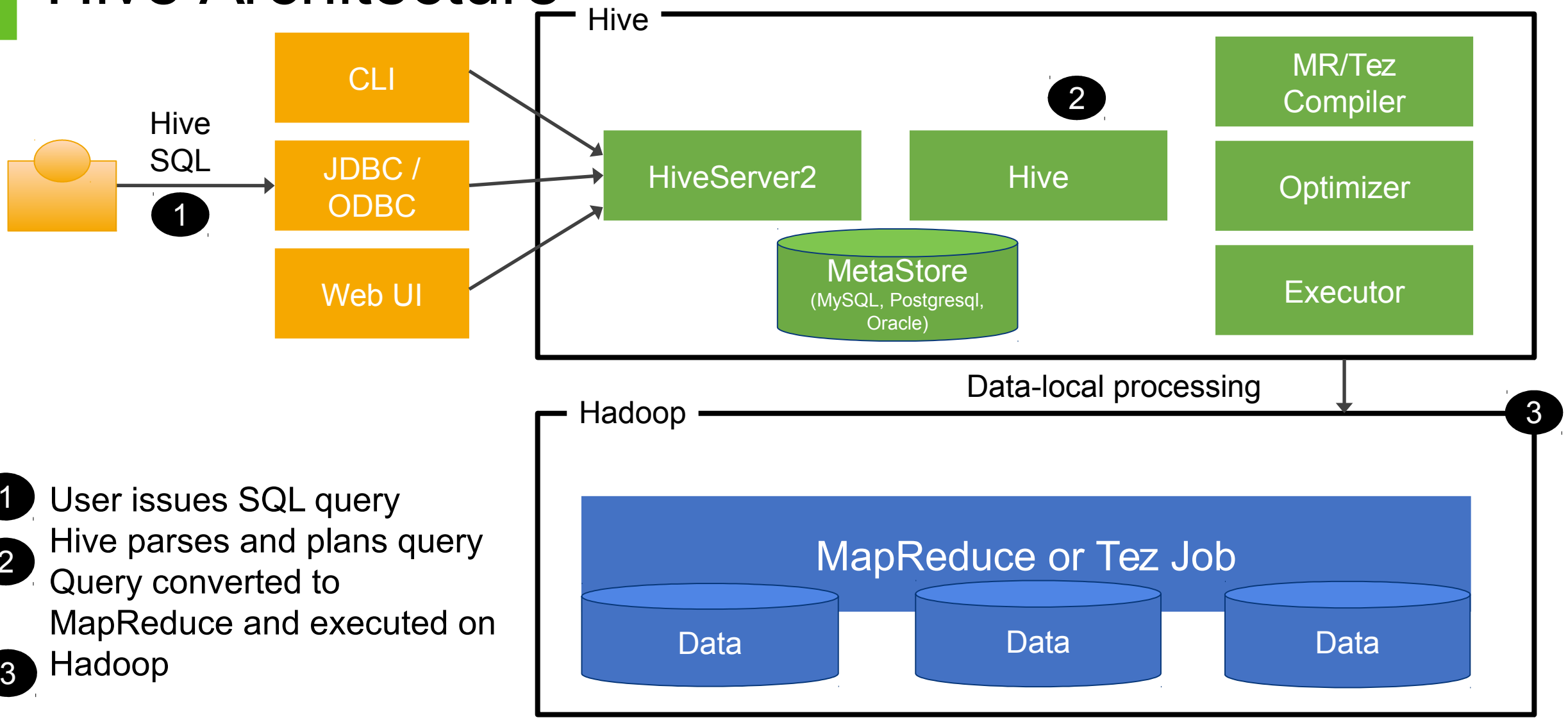
- Data warehouse system for Hadoop
- Create schemas/table definitions that *point to data* in Hadoop
- Treat your data in Hadoop as tables
- SQL 92
- Interactive queries at scale

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

Hive Architecture

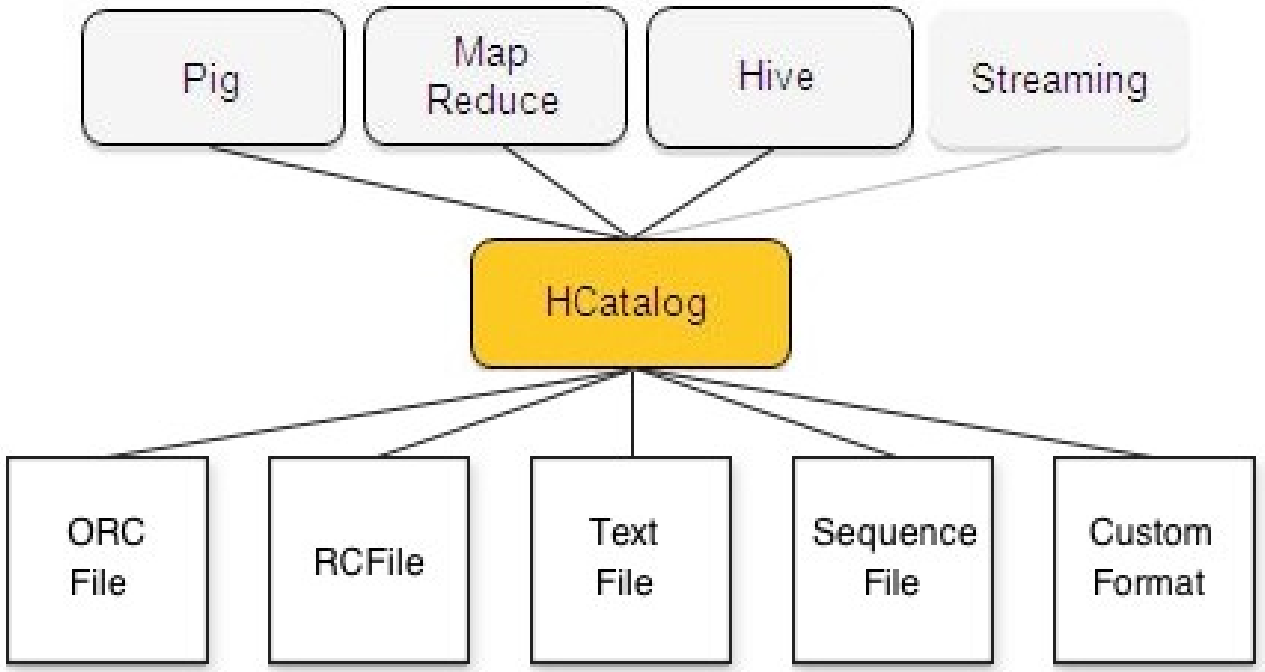


1. User issues SQL query
2. Hive parses and plans query
   Query converted to
   MapReduce and executed on
3. Hadoop

Hortonworks

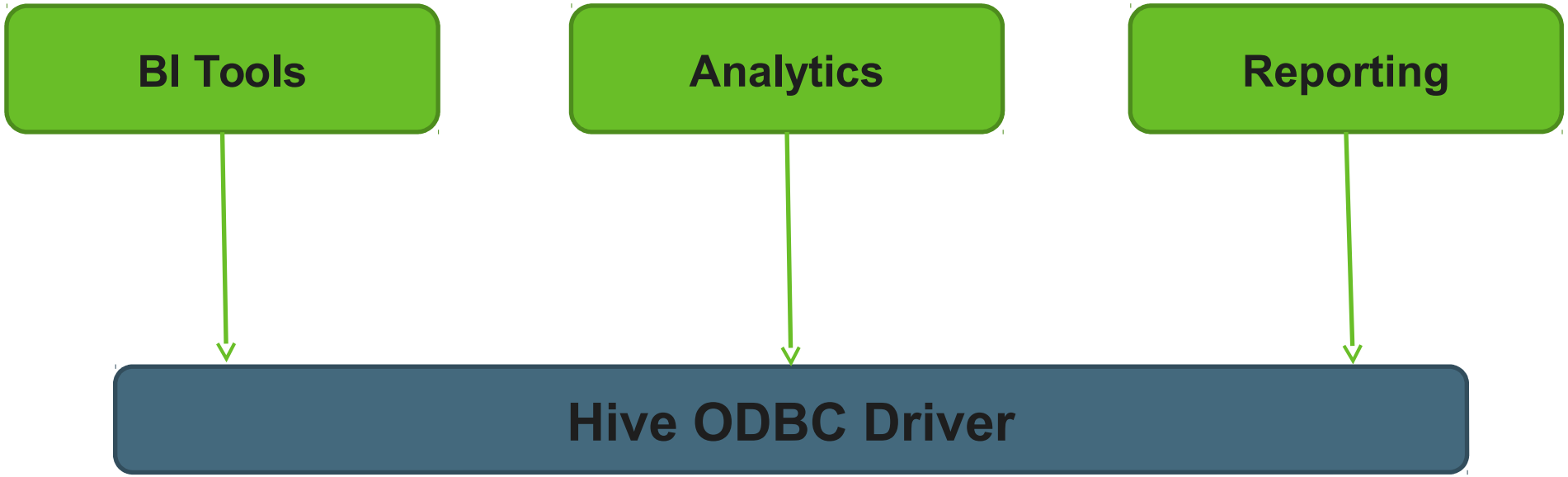
Hadoop Ecosystem: **HCatalog**

- 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

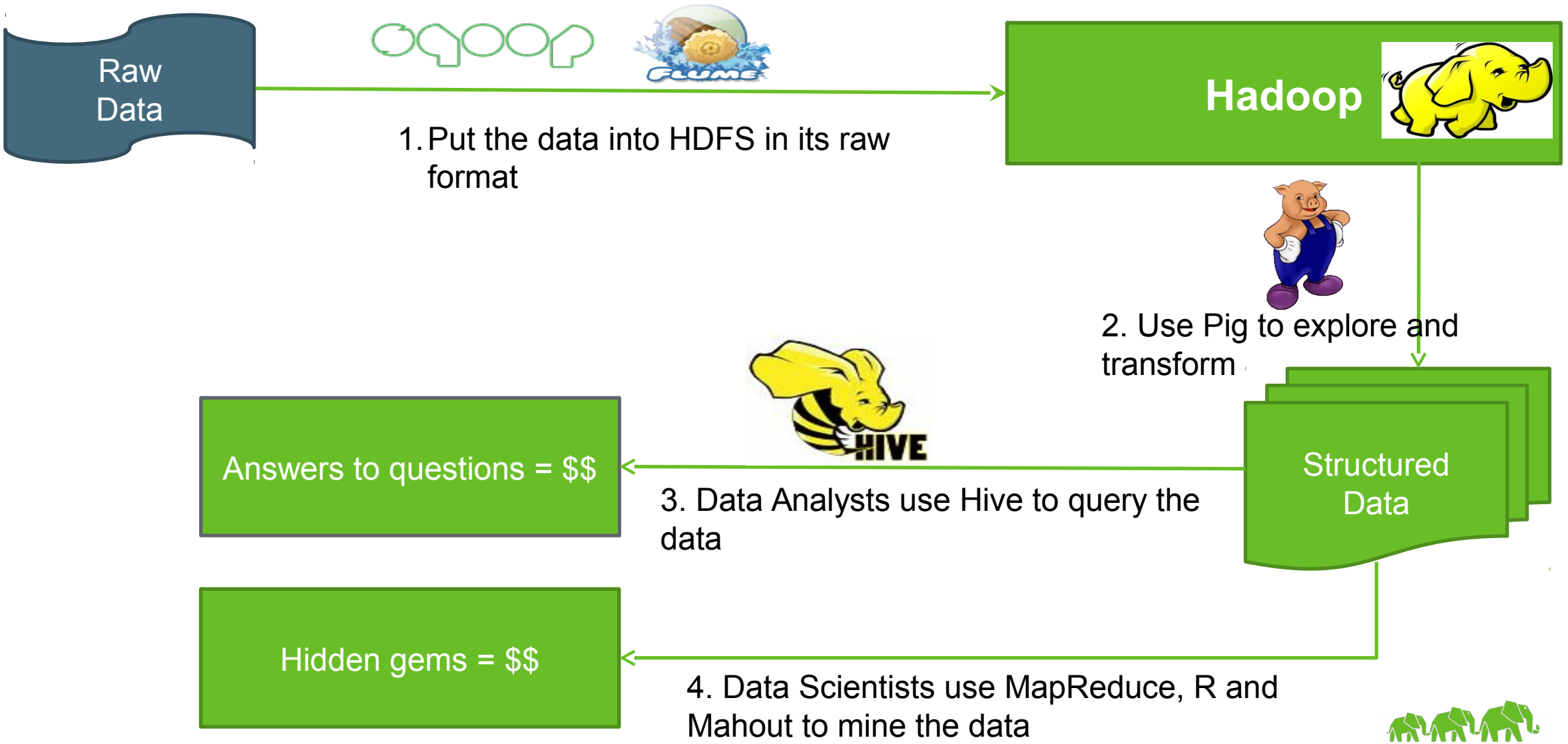
Overview of the Hive ODBC Driver



http://hortonworks.com/hdp/addons

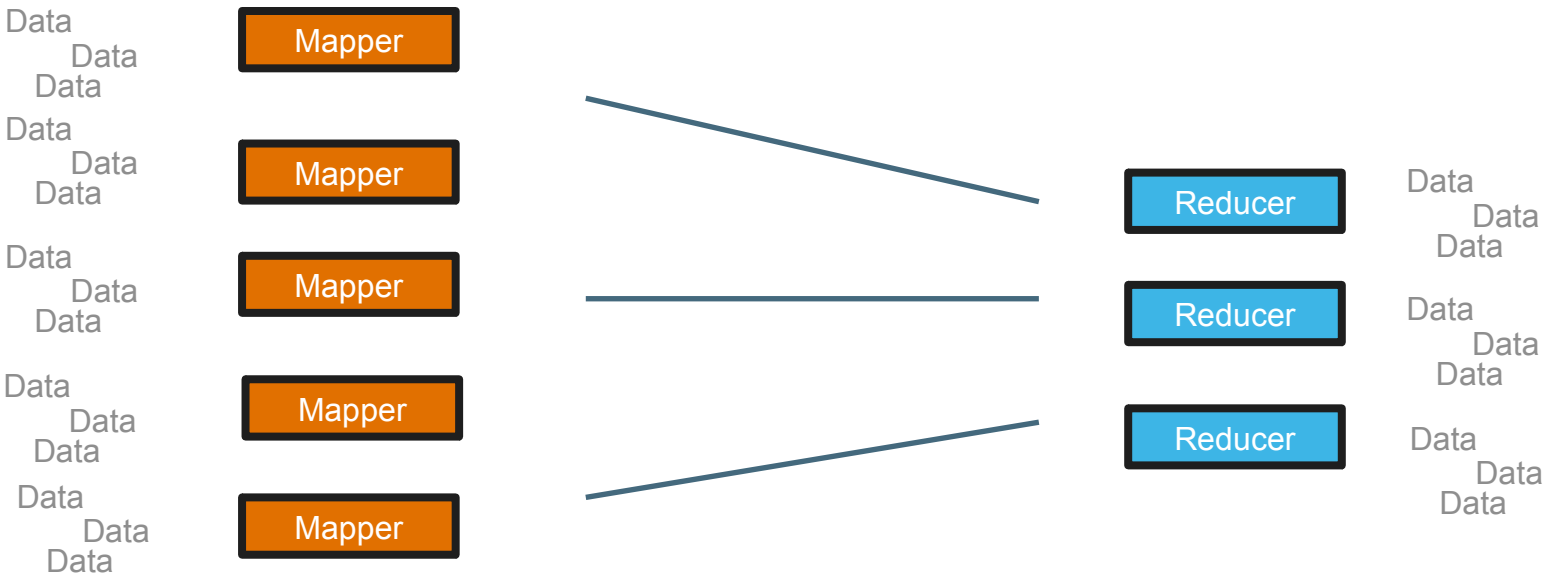
Hortonworks

ROI from Your Cluster



Hortonworks

MapReduce applies to *a lot* of data processing problems

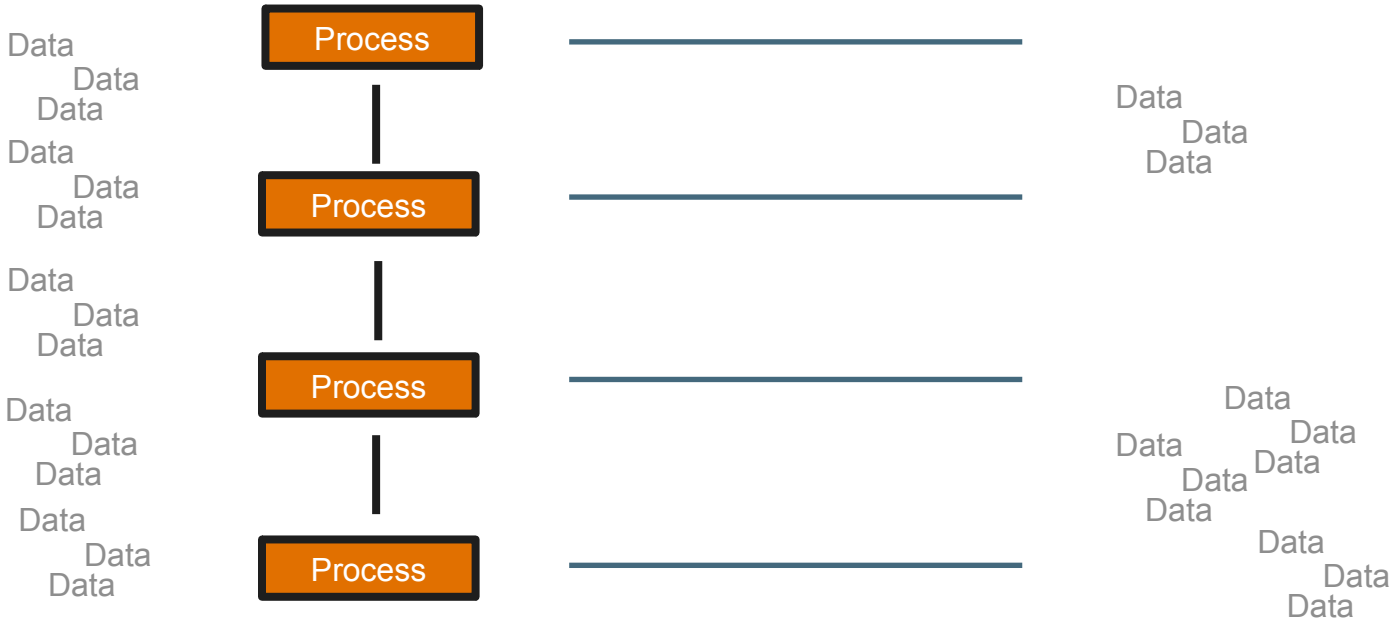


Hortonworks

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

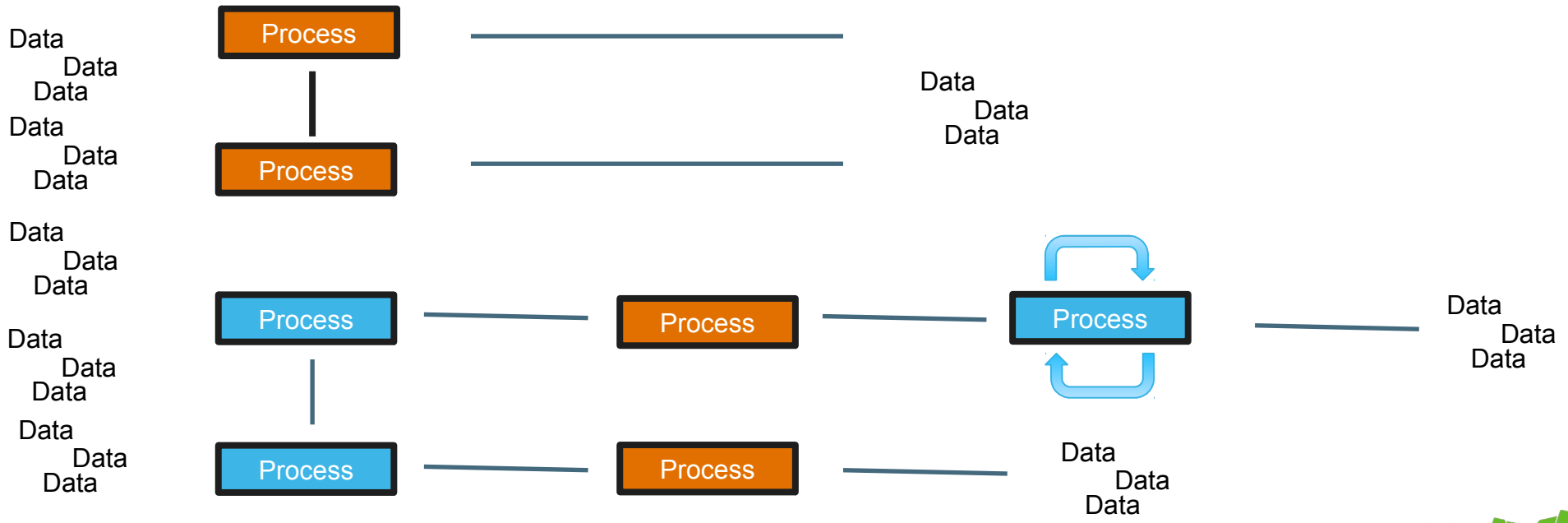


Hortonworks

…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

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

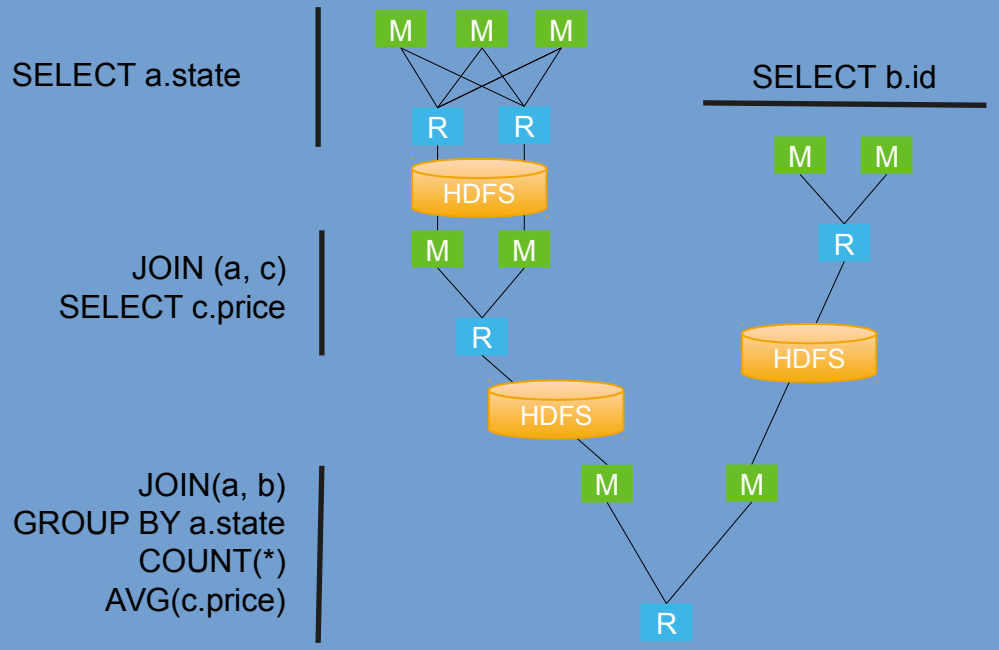
Hortonworks

Tez

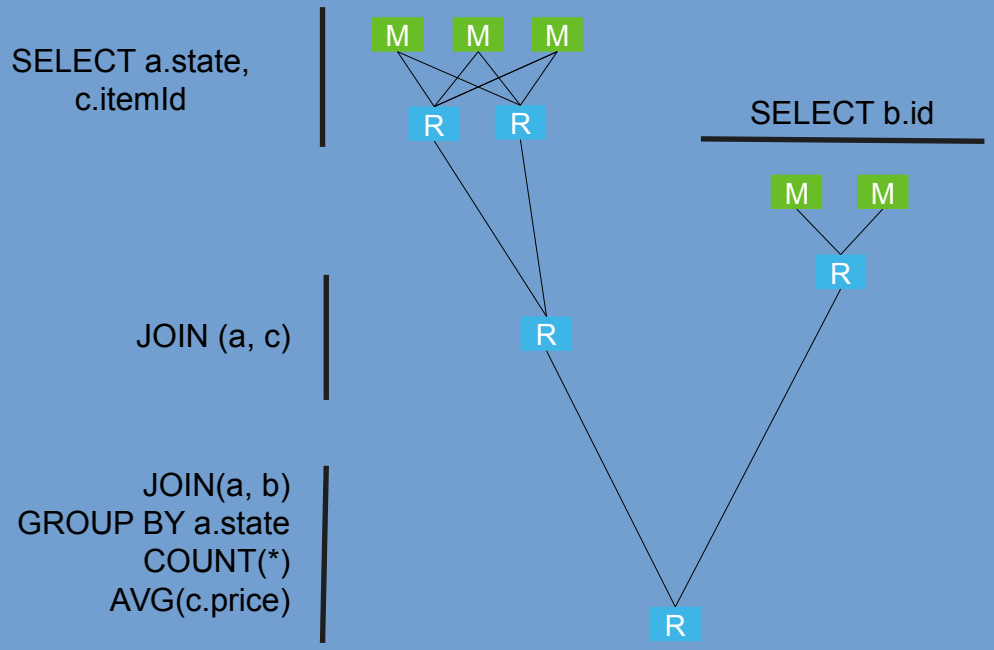
```
SELECT a.state, COUNT(*), AVG(c.price)
FROM a
JOIN b ON (a.id = b.id)
JOIN c ON (a.itemId = c.itemId)
GROUP BY a.state
```

Tez avoids unneeded writes to HDFS

Hive – MapReduce



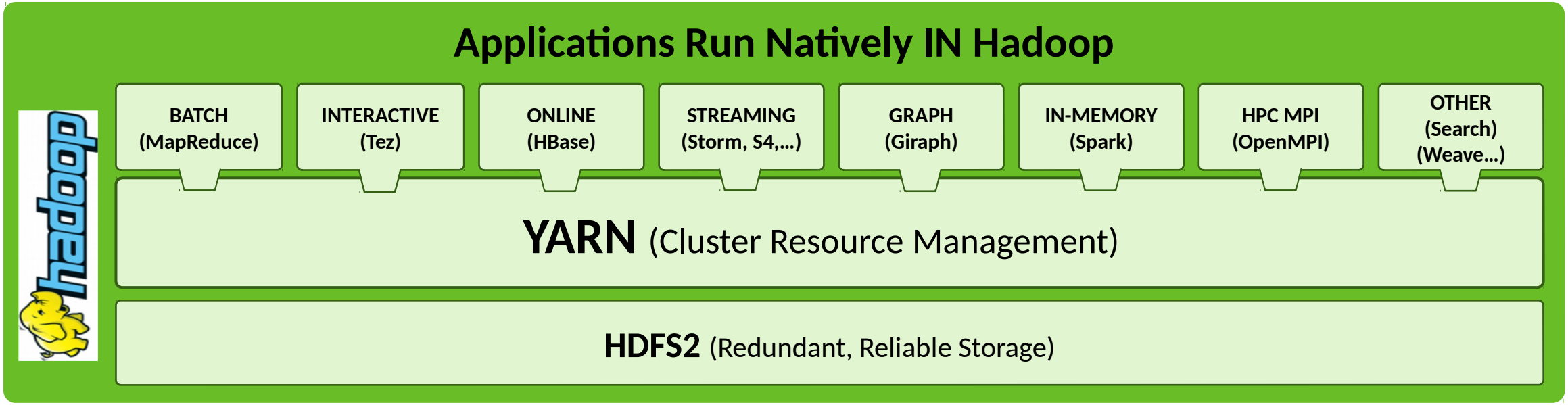
Hive – Tez



Hortonworks

Back to YARN: Taking Hadoop Beyond Batch

- With YARN, applications run natively ***in*** Hadoop



Hortonworks

Hadoop Security

Kerberos Authentication — All services can be principals

HBase, Hive & HDFS Authorization — ACLs for each

Wire encryption — HDFS, Shuffle, JDBC

Basic Auditing — Logging of events

Knox: Hadoop REST Gateway — Perimeter security for Hadoop cluster

Centralized Security Administration

Centralized Audit Reporting

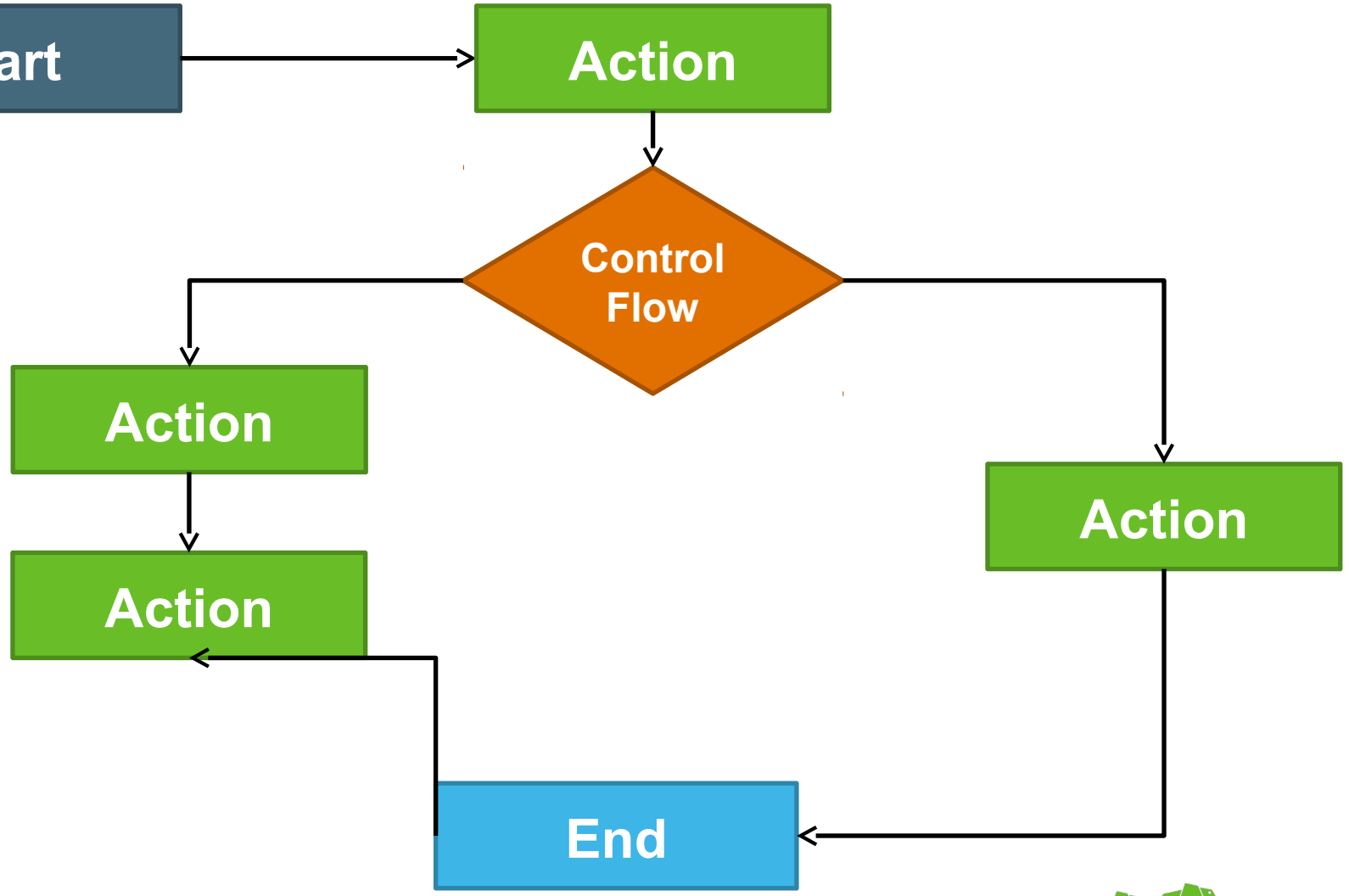
Delegated Policy Administration

} XA Secure

Hortonworks

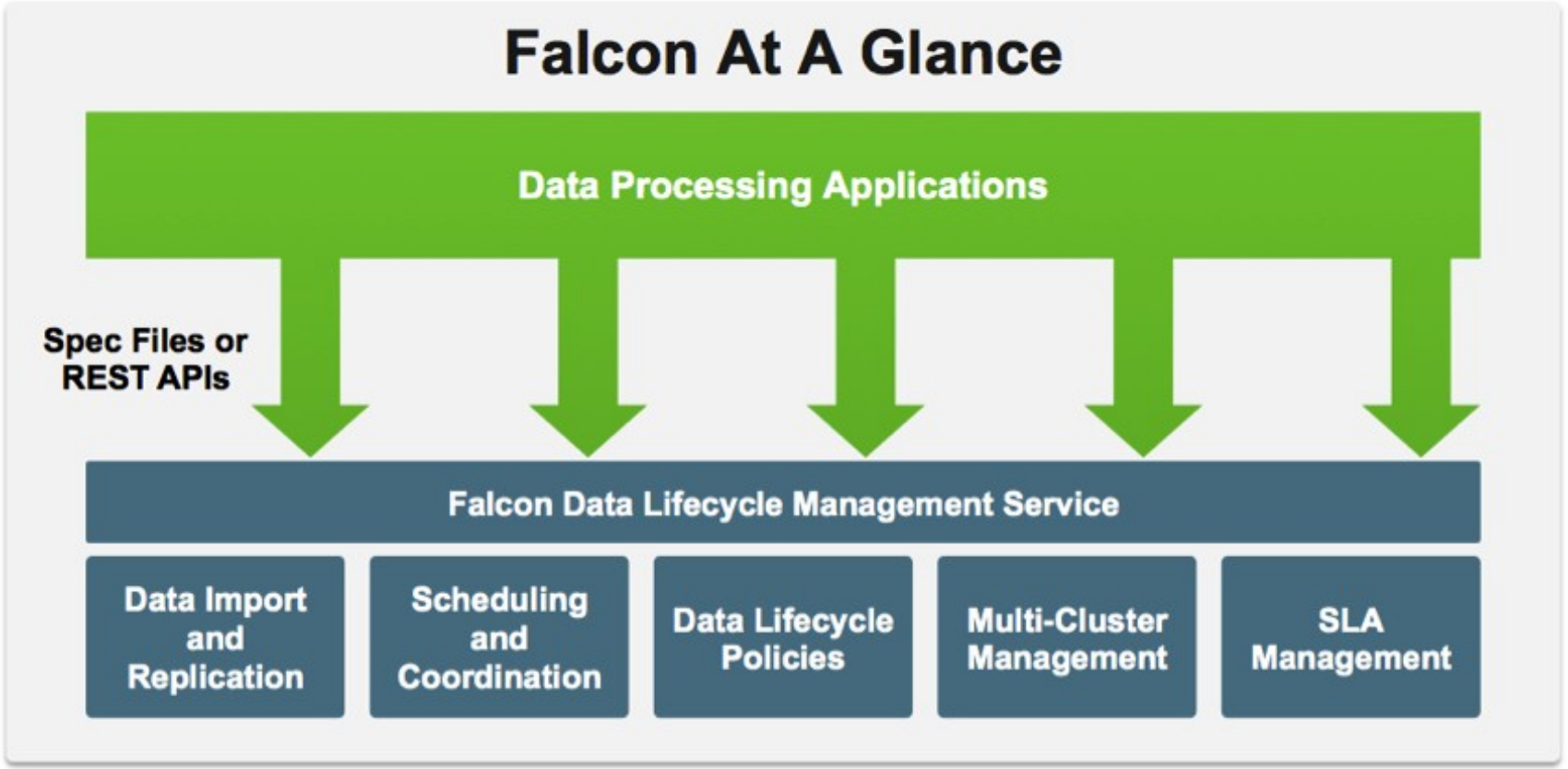
Defining an Oozie Workflow

- 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

Hortonworks Data Platform

Apache Ambari
http://incubator.apache.org/ambari
Cluster System Operations

ETL

RDBMS Import/Export

KNOX
Secure Gateway

Streaming Data Ingestion

Distributed Storage & Processing Framework

Workflow Management

SQL

Data Management

NoSQL DB

Search & Indexing

SQL on HBase

Distributed Registry

Even Faster Data Processing

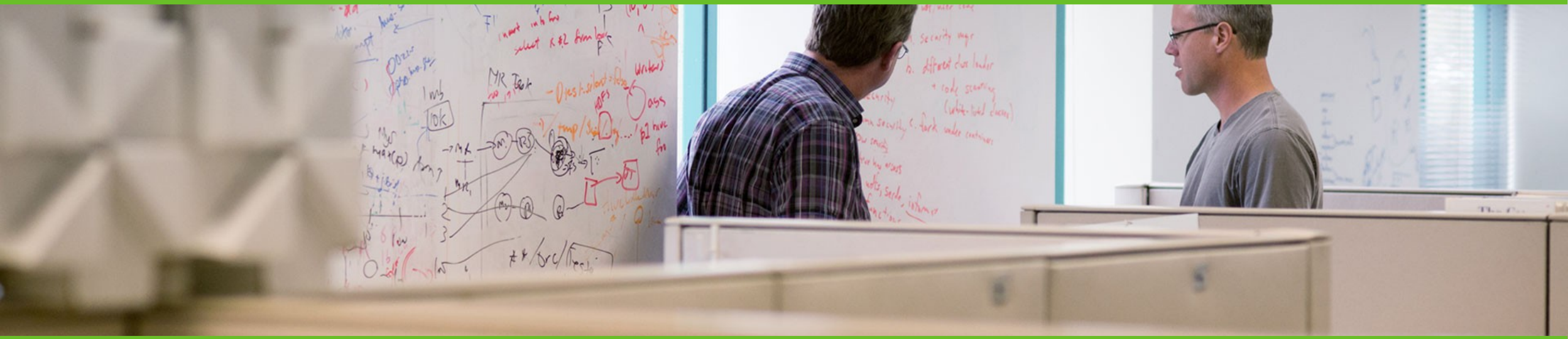
Machine Learning

Secure NoSQL DB

runs on

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**

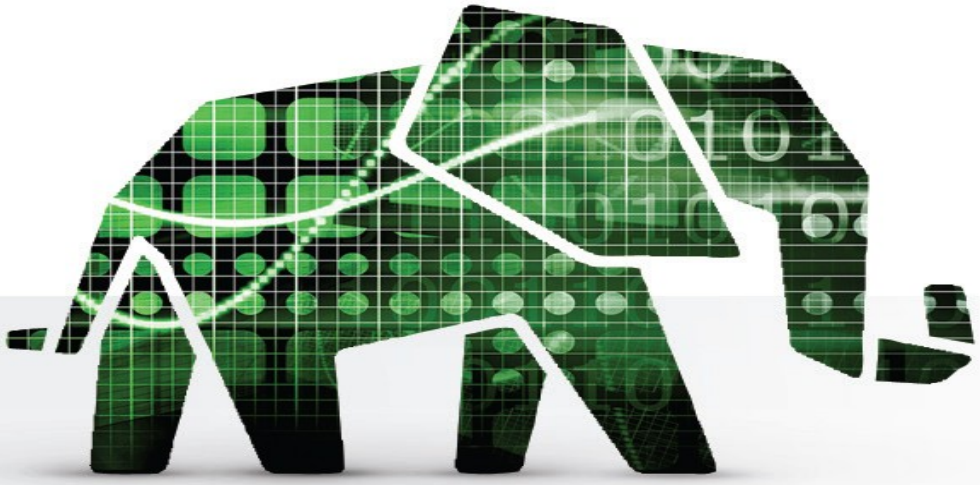
Interested in:
SQL & Scripting Languages
Large Scale Data Sets
Creating Value & Opportunity

Data Classes:

Applying Data Science with Hadoop ›
2 Days

Data Analysis with Hive & Pig ›
4 Days

Thank you!



hwuniversity@hortonworks.com

izhar@abyres.net

Hortonworks