Tamkang University
淡江大學

Tamkang University

# Social Computing and Big Data Analytics
# 社群運算與大數據分析

# Fundamental Big Data:
## MapReduce Paradigm, Hadoop and Spark Ecosystem
## (大數據基礎：MapReduce典範、Hadoop與Spark生態系統)

1042SCBDA03
MIS MBA (M2226) (8628)
Wed, 8,9, (15:10-17:00) (Q201)


Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系
http://mail. tku.edu.tw/myday/


2016-03-02

# 課程大綱 (Syllabus)

| 週次 (Week) | 日期 (Date) | 內容 (Subject/Topics) |
|---|---|---|
| 1 | 2016/02/17 | Course Orientation for Social Computing and Big Data Analytics<br>(社群運算與大數據分析課程介紹) |
| 2 | 2016/02/24 | Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data<br>(資料科學與大數據分析：探索、分析、視覺化與呈現資料) |
| 3 | 2016/03/02 | Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem<br>(大數據基礎：MapReduce典範、Hadoop與Spark生態系統) |

# 課程大綱 (Syllabus)

| 週次 (Week) | 日期 (Date) | 內容 (Subject/Topics) |
|---|---|---|
| 4 | 2016/03/09 | Big Data Processing Platforms with SMACK: Spark, Mesos, Akka, Cassandra and Kafka (大數據處理平台SMACK：Spark, Mesos, Akka, Cassandra, Kafka) |
| 5 | 2016/03/16 | Big Data Analytics with Numpy in Python (Python Numpy 大數據分析) |
| 6 | 2016/03/23 | Finance Big Data Analytics with Pandas in Python (Python Pandas 財務大數據分析) |
| 7 | 2016/03/30 | Text Mining Techniques and Natural Language Processing (文字探勘分析技術與自然語言處理) |
| 8 | 2016/04/06 | Off-campus study (教學行政觀摩日) |

# 課程大綱 (Syllabus)

| 週次 (Week) | 日期 (Date) | 內容 (Subject/Topics) |
|---|---|---|
| 9 | 2016/04/13 | Social Media Marketing Analytics (社群媒體行銷分析) |
| 10 | 2016/04/20 | 期中報告 (Midterm Project Report) |
| 11 | 2016/04/27 | Deep Learning with Theano and Keras in Python (Python Theano 和 Keras 深度學習) |
| 12 | 2016/05/04 | Deep Learning with Google TensorFlow (Google TensorFlow 深度學習) |
| 13 | 2016/05/11 | Sentiment Analysis on Social Media with Deep Learning (深度學習社群媒體情感分析) |

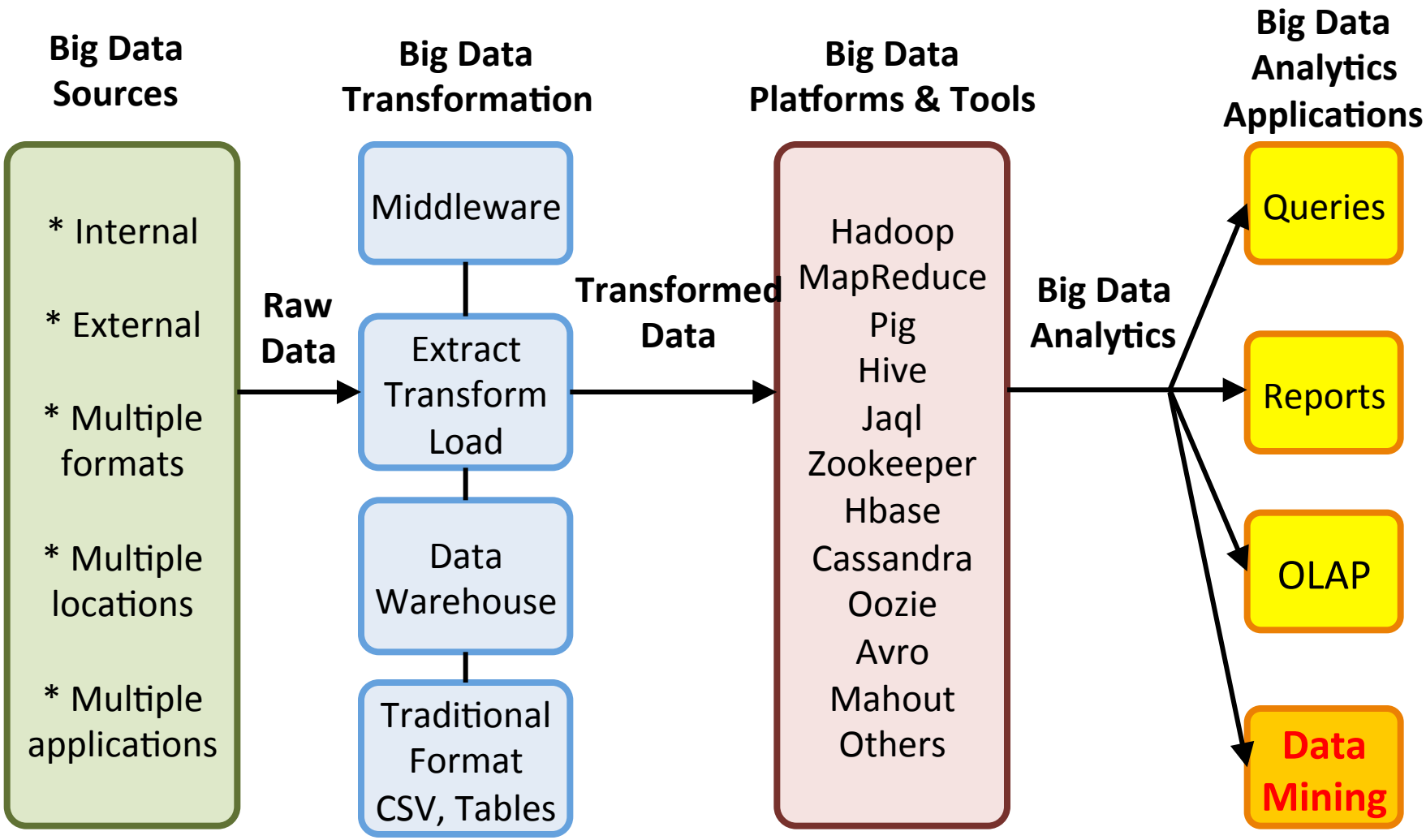
# 課程大綱 (Syllabus)

| 週次 (Week) | 日期 (Date) | 內容 (Subject/Topics) |
|---|---|---|
| 14 | 2016/05/18 | Social Network Analysis (社會網絡分析) |
| 15 | 2016/05/25 | Measurements of Social Network (社會網絡量測) |
| 16 | 2016/06/01 | Tools of Social Network Analysis (社會網絡分析工具) |
| 17 | 2016/06/08 | Final Project Presentation I (期末報告 I) |
| 18 | 2016/06/15 | Final Project Presentation II (期末報告 II) |

# 2016/03/02

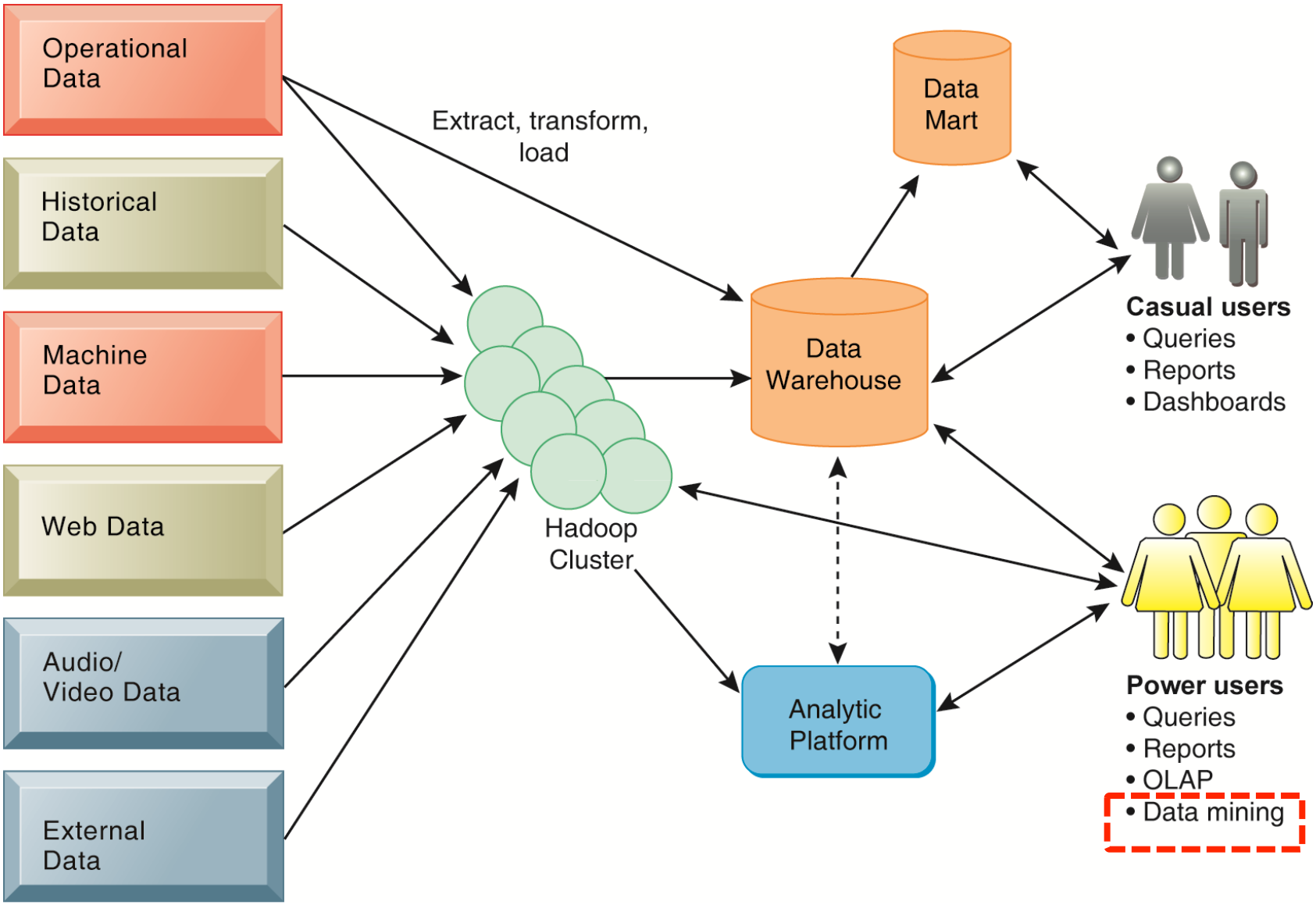
# Fundamental Big Data:
# MapReduce Paradigm, Hadoop and Spark Ecosystem
# (大數據基礎：MapReduce典範、Hadoop與Spark生態系統)

# Architecture of Big Data Analytics




Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

# Business Intelligence (BI) Infrastructure




Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

# Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem

National Security | Cyber security | Maritime security | Smarter Transport | ...

## VISUAL ANALYTICS

**DYNAMIC & INTERACTIVE**
Dashboard Graph
Map

**ENHANCE**
Understanding Investigation
User Experience


## BIG ANALYTICS

**QUERY & FILTER**
Complex queries
R²I²

**DETECT**
Anomalies
Communities
Typologies

**PREDICT**
Tending
Real-time
Prediction

**DECIDE**
Simulation
Optimization


## BIG DATA – Batch


## BIG DATA – Real Time


## DATA

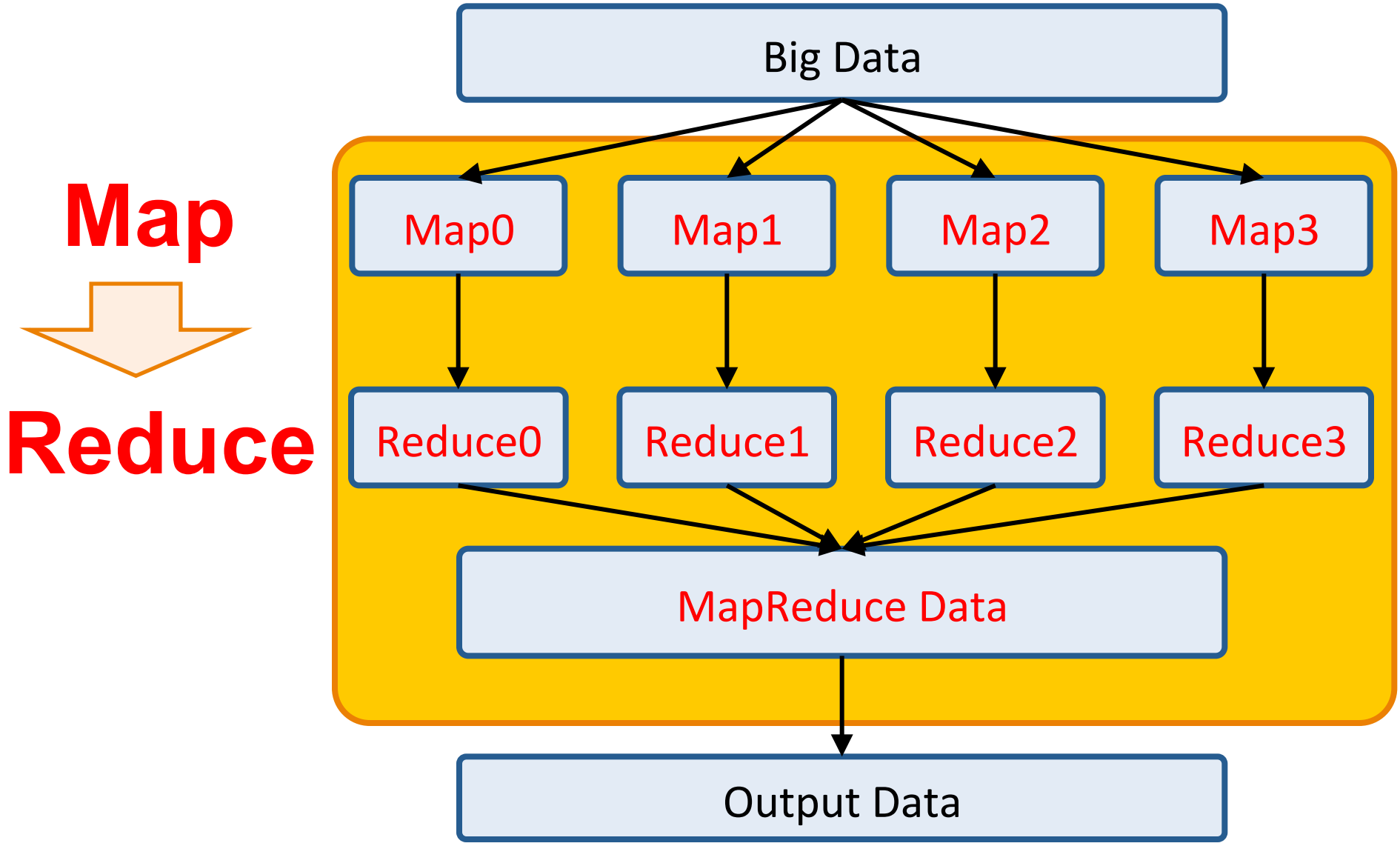
**Complex by nature**

**Complex by structure**

Source: https://www.thalesgroup.com/en/worldwide/big-data/big-data-big-analytics-visual-analytics-what-does-it-all-mean

# MapReduce Paradigm

# MapReduce Paradigm



# Hadoop Ecosystem


**The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.**

Source: http://hadoop.apache.org/

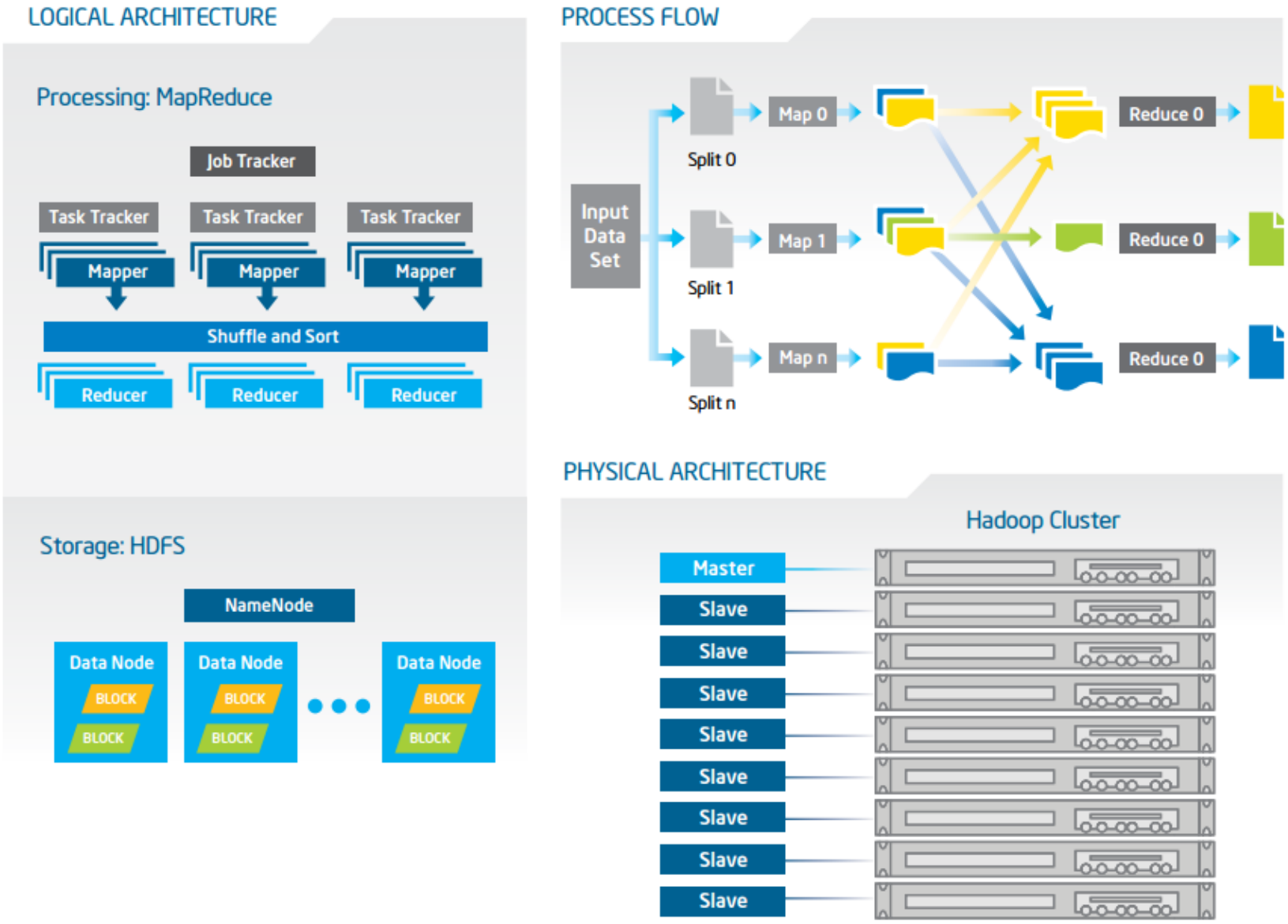

**MapReduce** **Processing**

+

**HDFS** **Storage**

Source: http://hadoop.apache.org/

# Big Data with Hadoop Architecture

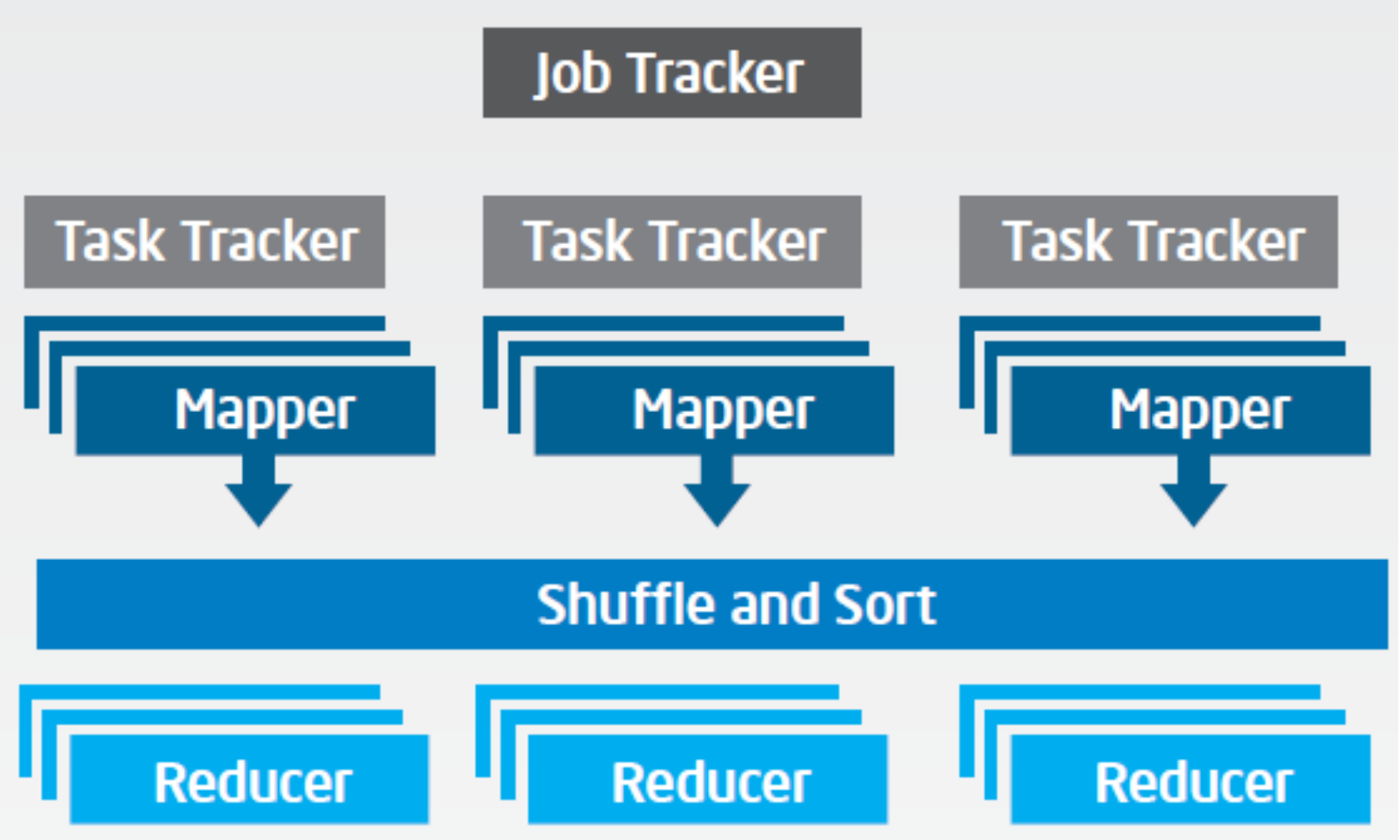



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

# Big Data with Hadoop Architecture

## Logical Architecture

### Processing: MapReduce

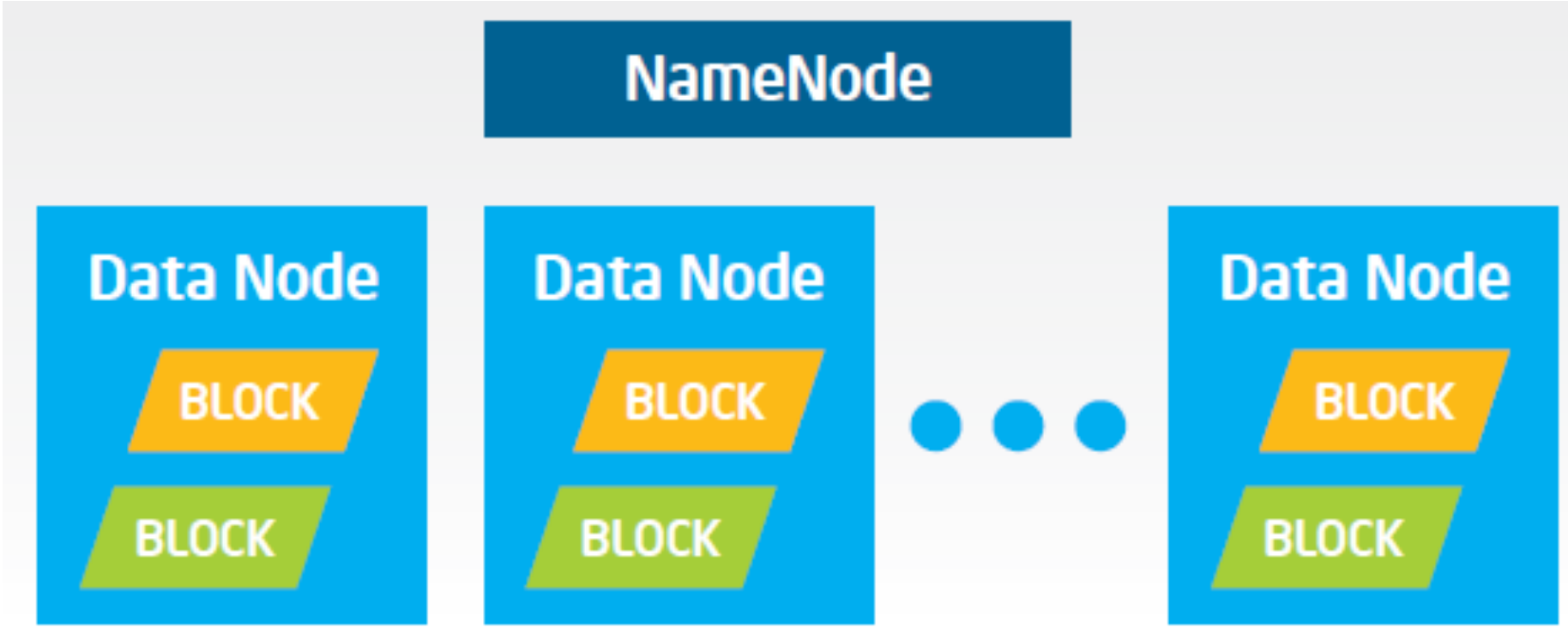



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

# Big Data with Hadoop Architecture

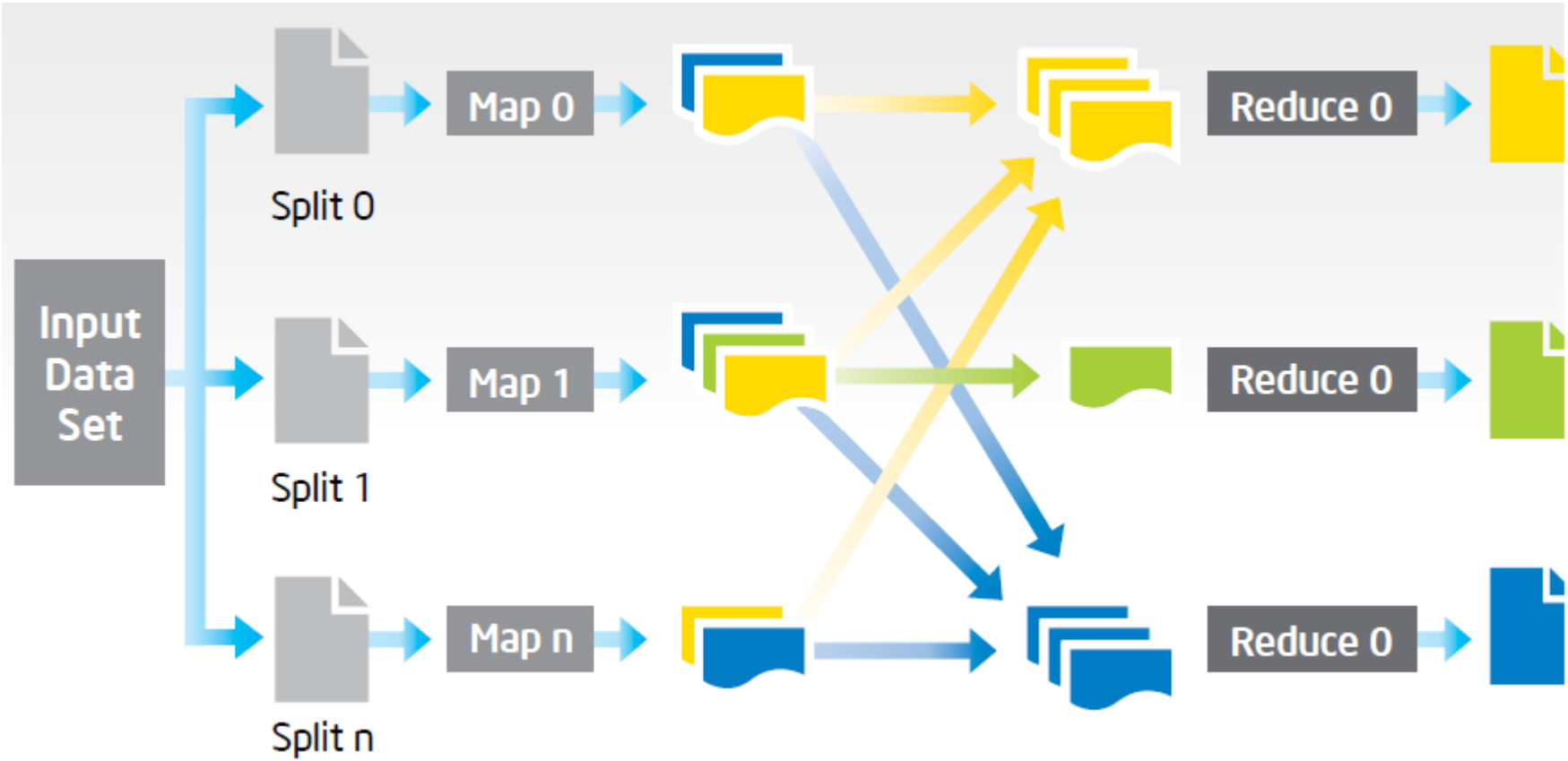
## Logical Architecture

### Storage: HDFS




Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

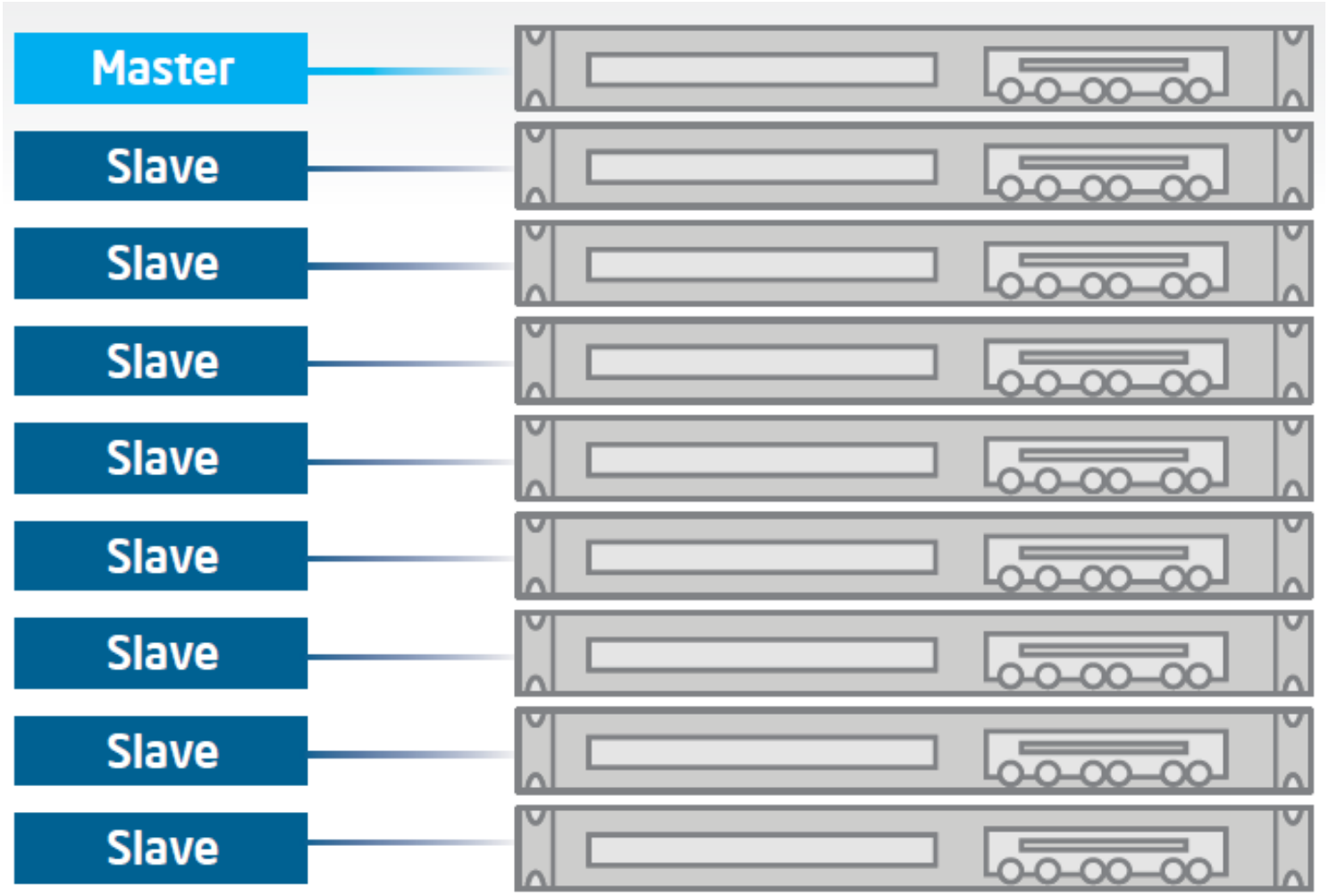
# Big Data with Hadoop Architecture
## Process Flow




Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

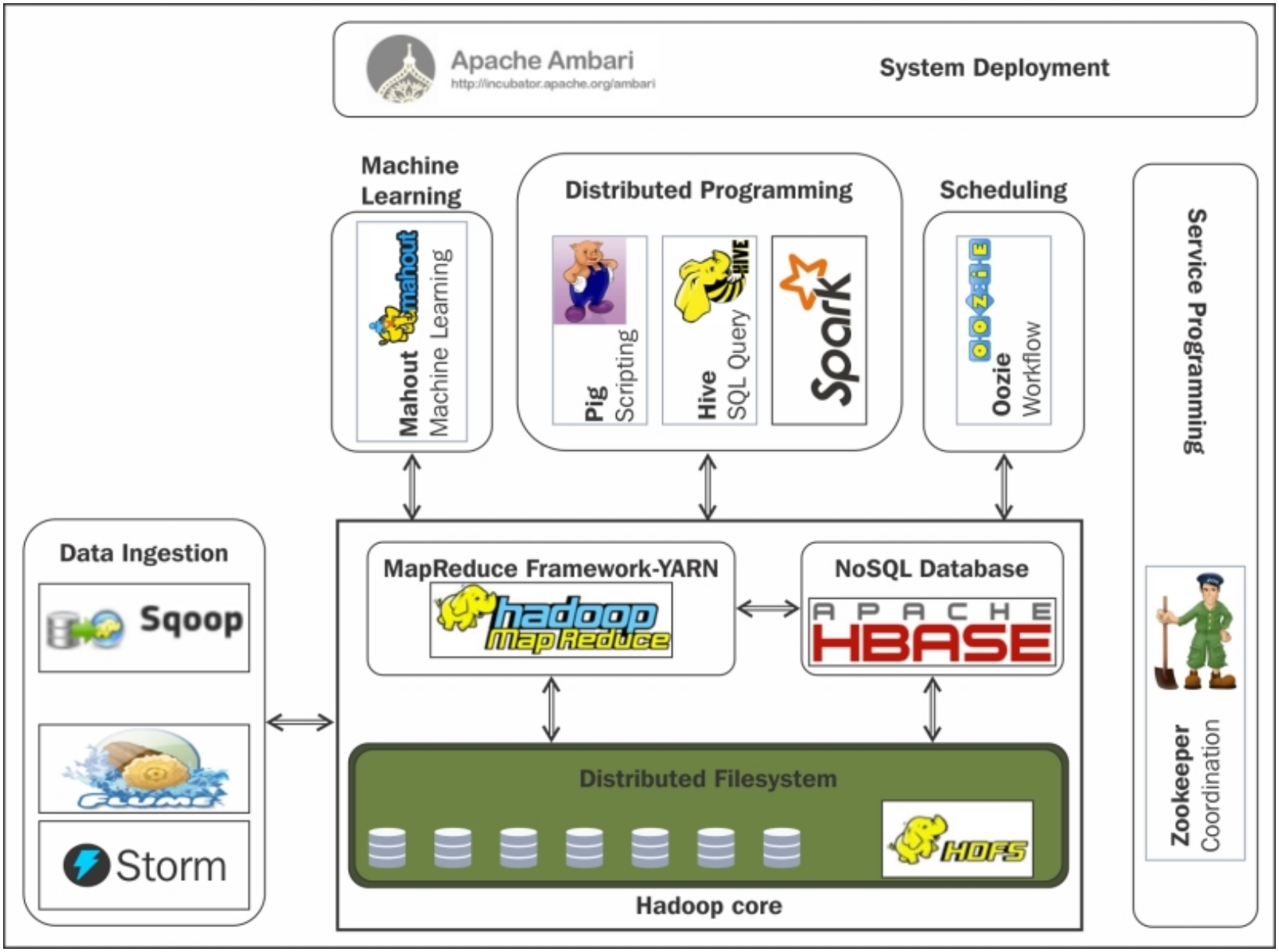
# Big Data with Hadoop Architecture

## Hadoop Cluster




Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

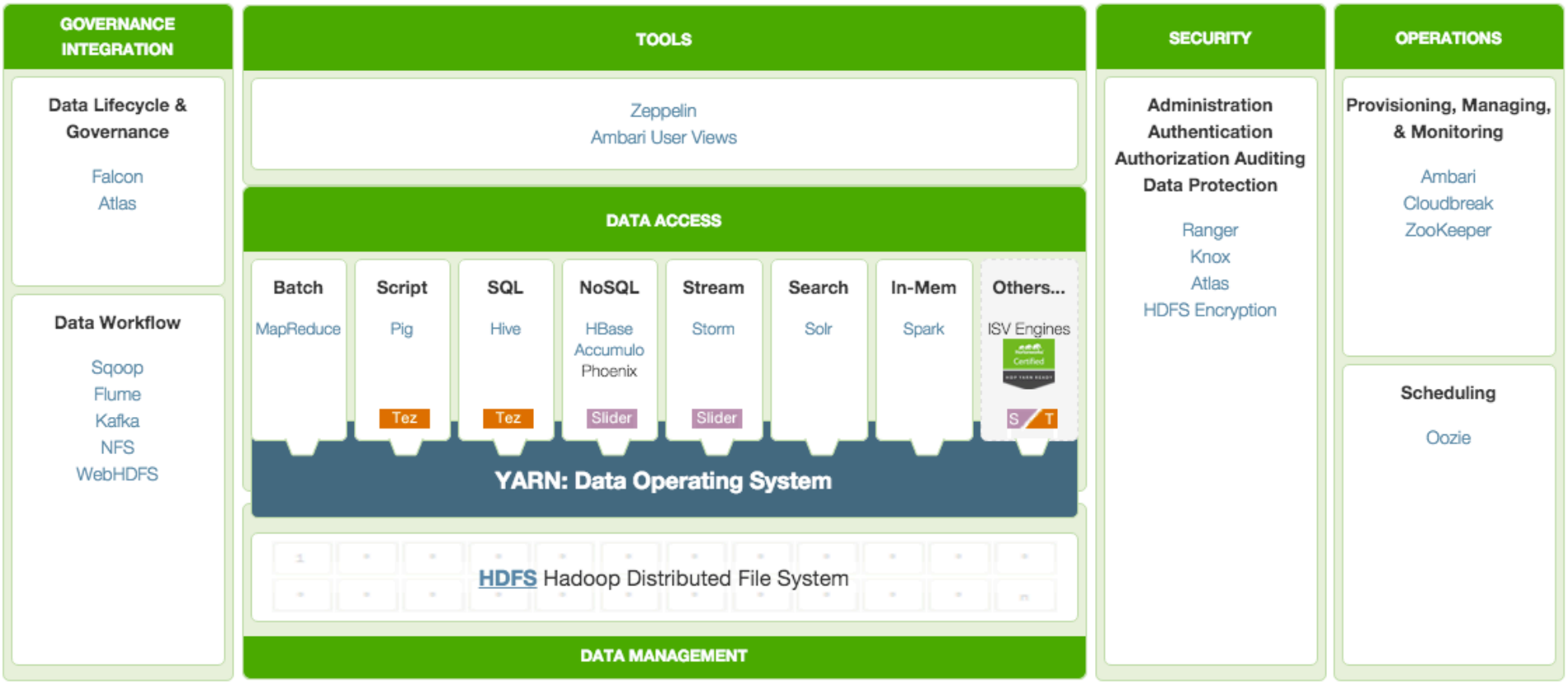
# Hadoop Ecosystem




Source: Shiva Achari (2015), Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing

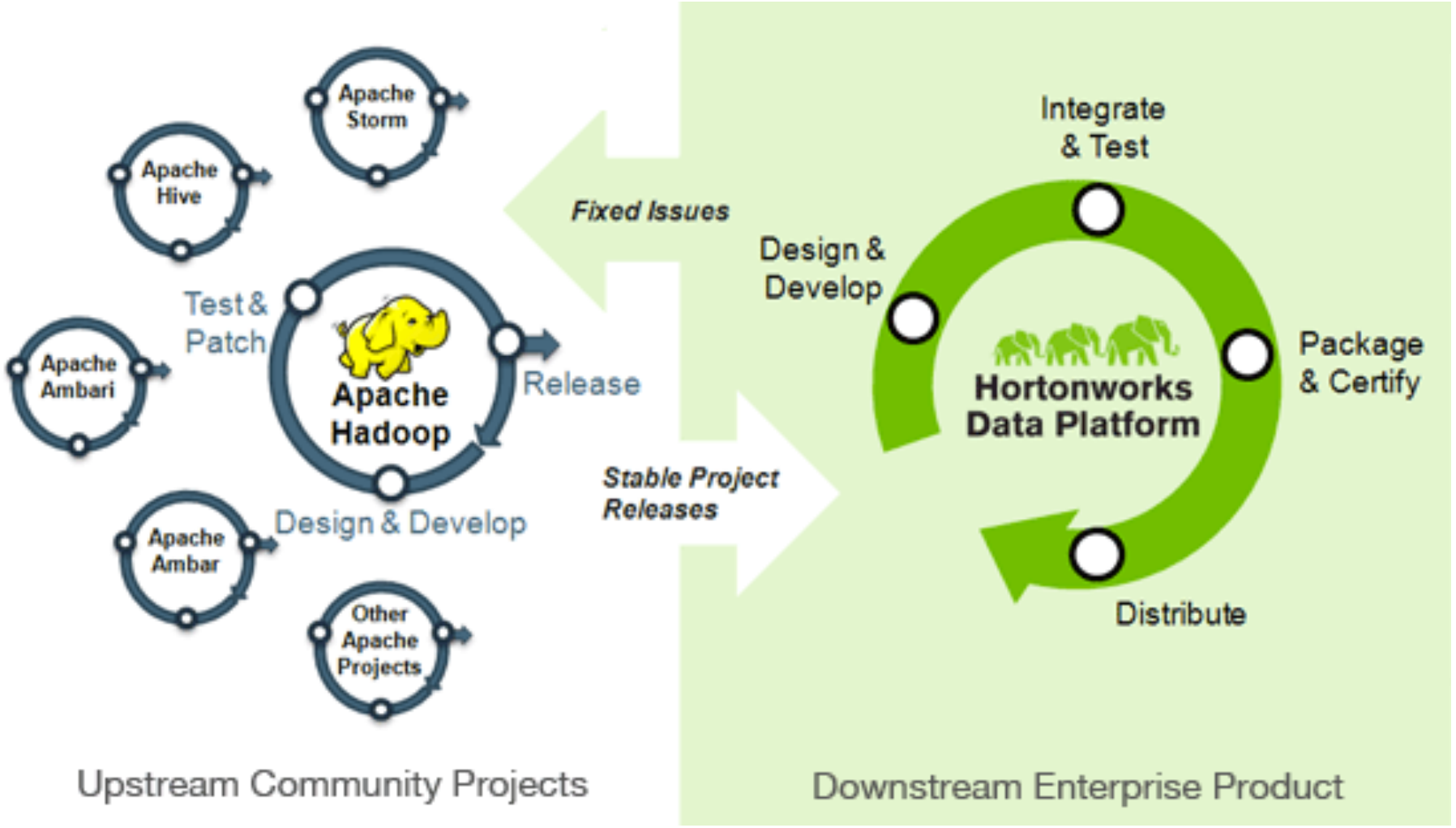
# HDP (Hortonworks Data Platform)

## A Complete Enterprise Hadoop Data Platform



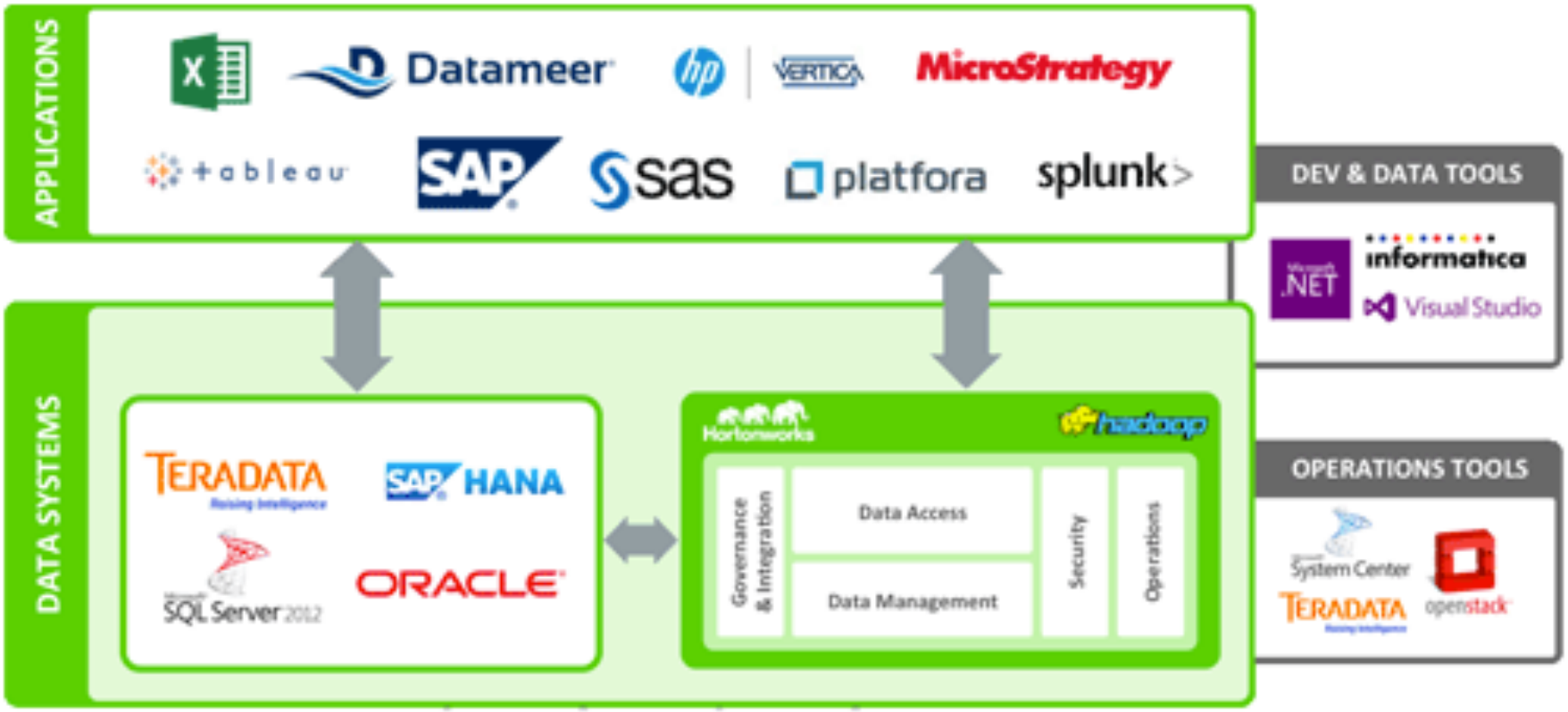

Source: http://hortonworks.com/hdp/

# Apache Hadoop
# Hortonworks Data Platform



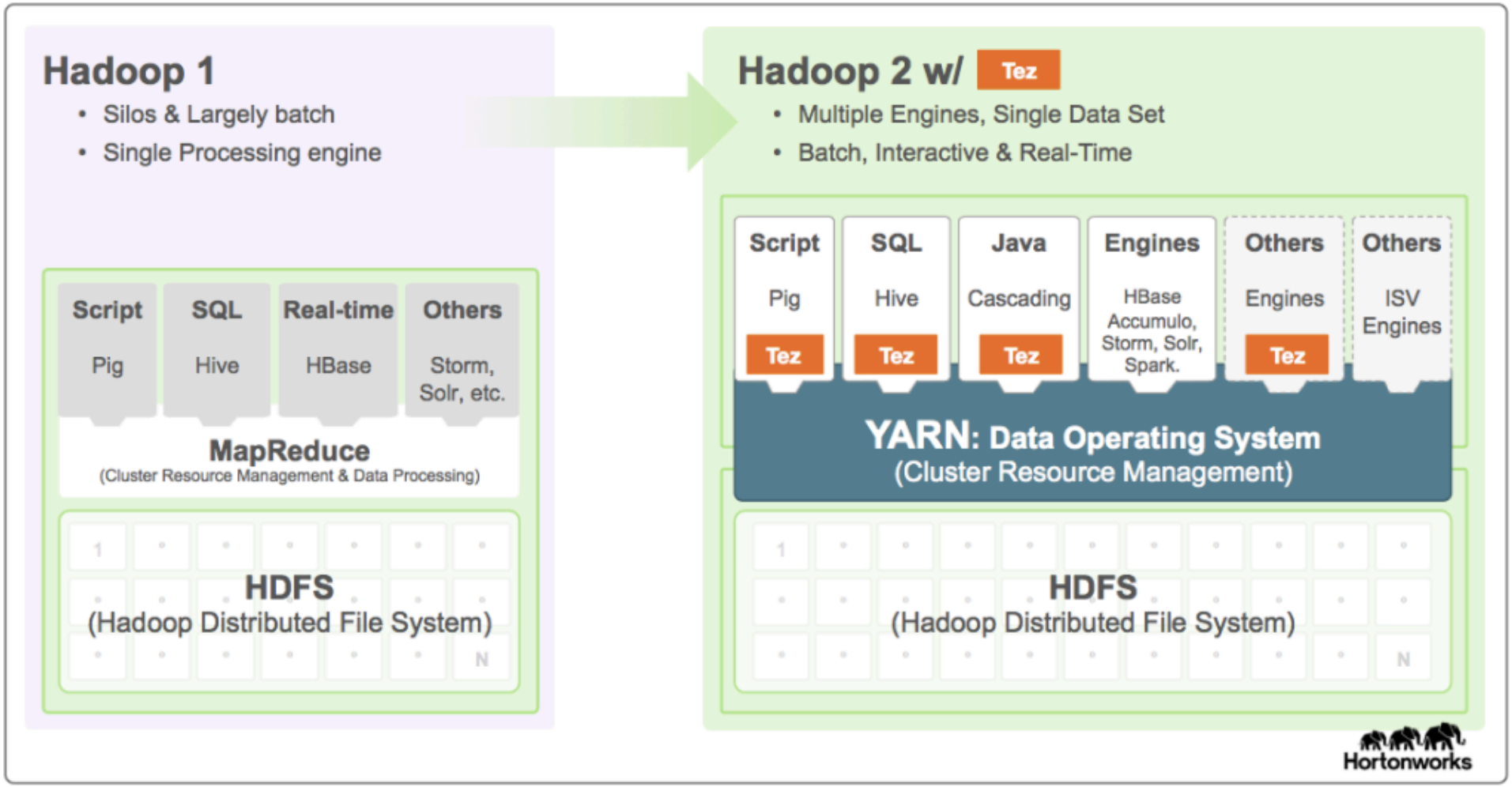

Source: http://hortonworks.com/hdp/

# Hadoop and Data Analytics Tools



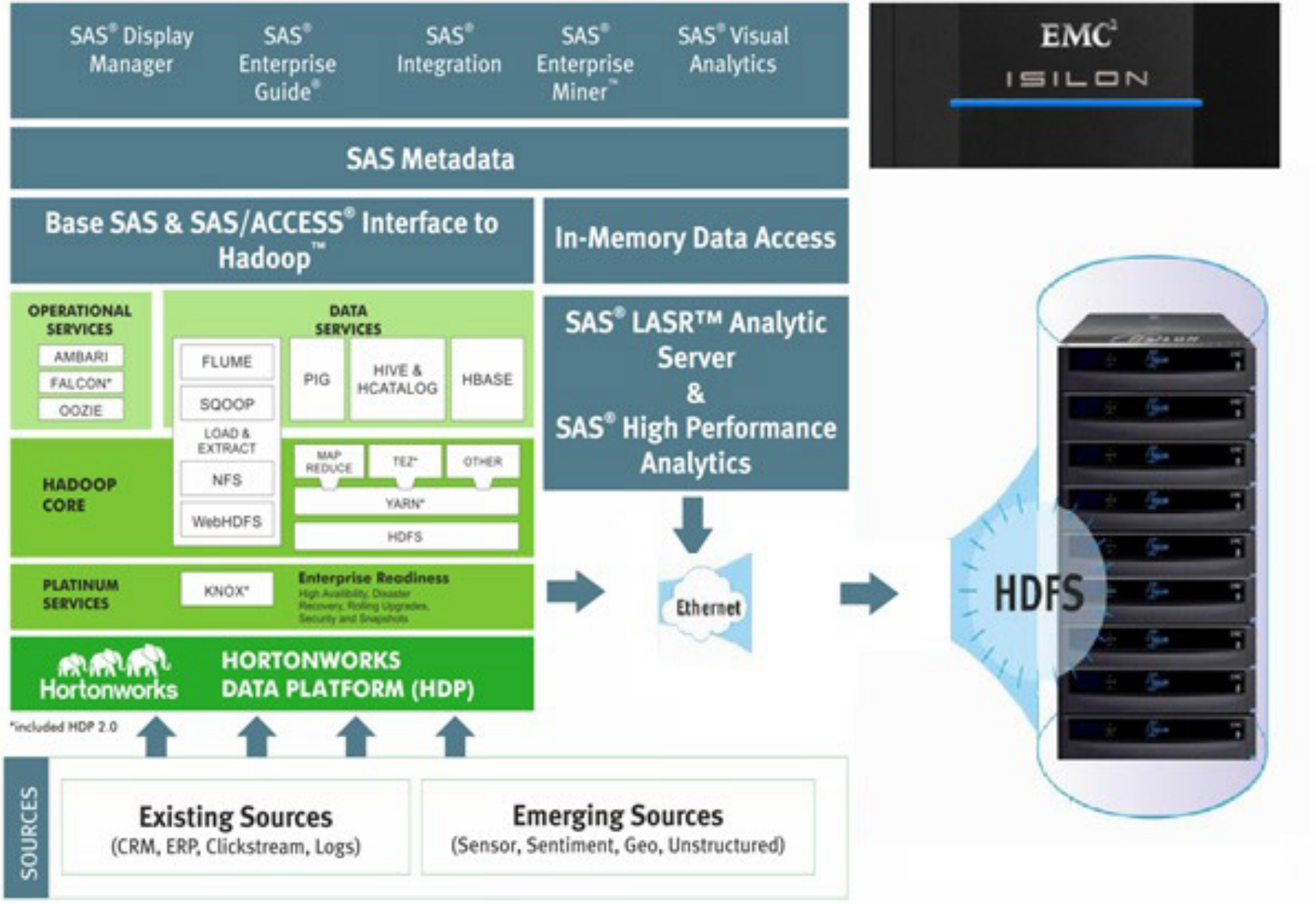

Source: http://hortonworks.com/hdp/

# Hadoop 1 → Hadoop 2



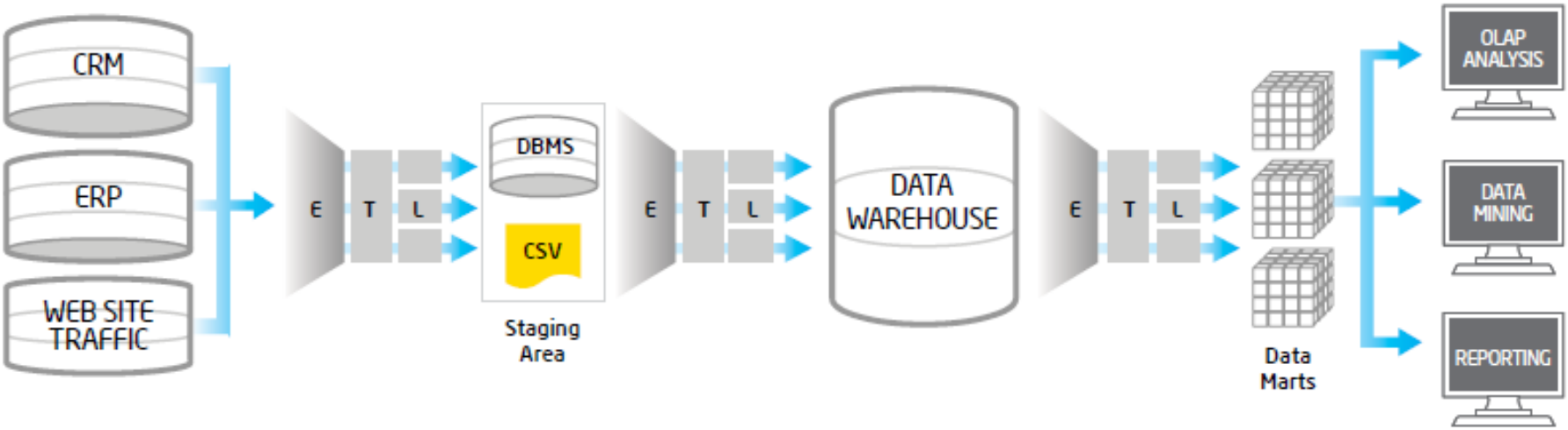

Source: http://hortonworks.com/hadoop/tez/

# Big Data Solution



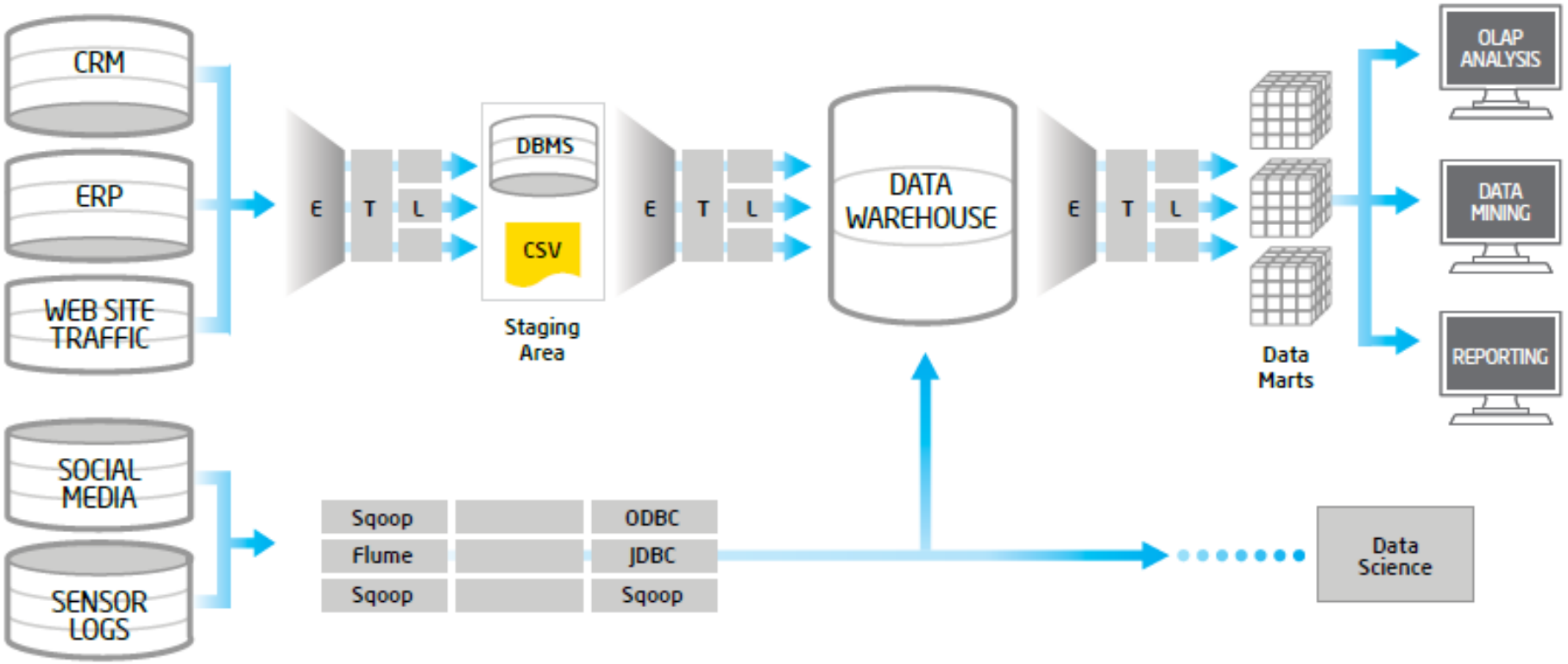

Source: http://www.newera-technologies.com/big-data-solution.html

# Traditional ETL Architecture




Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

# Offload ETL with Hadoop (Big Data Architecture)




Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

# Spark Ecosystem

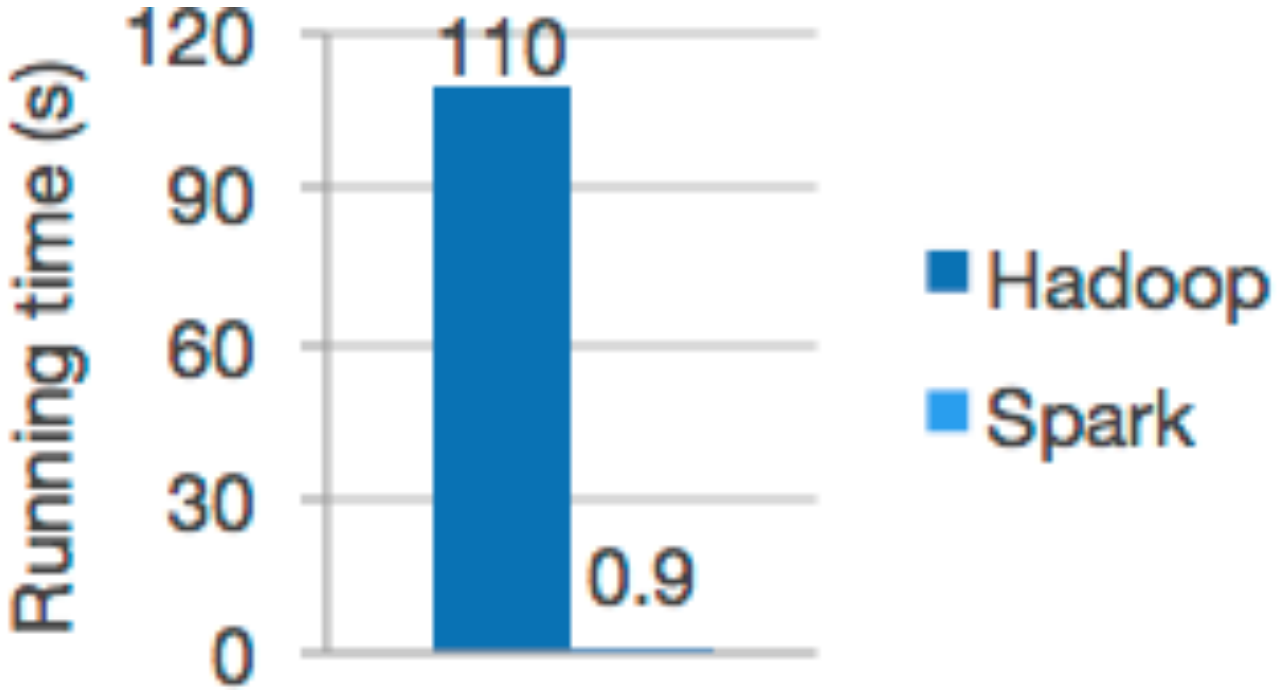

*Lightning-fast cluster computing*

# Apache Spark

**is a fast and general engine for large-scale data processing.**

Source: http://spark.apache.org/

# Logistic regression in Hadoop and Spark




Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Source: http://spark.apache.org/

# Ease of Use

- Write applications quickly in Java, Scala, Python, R.

Source: http://spark.apache.org/

# Word count in Spark's Python API

```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Source: http://spark.apache.org/

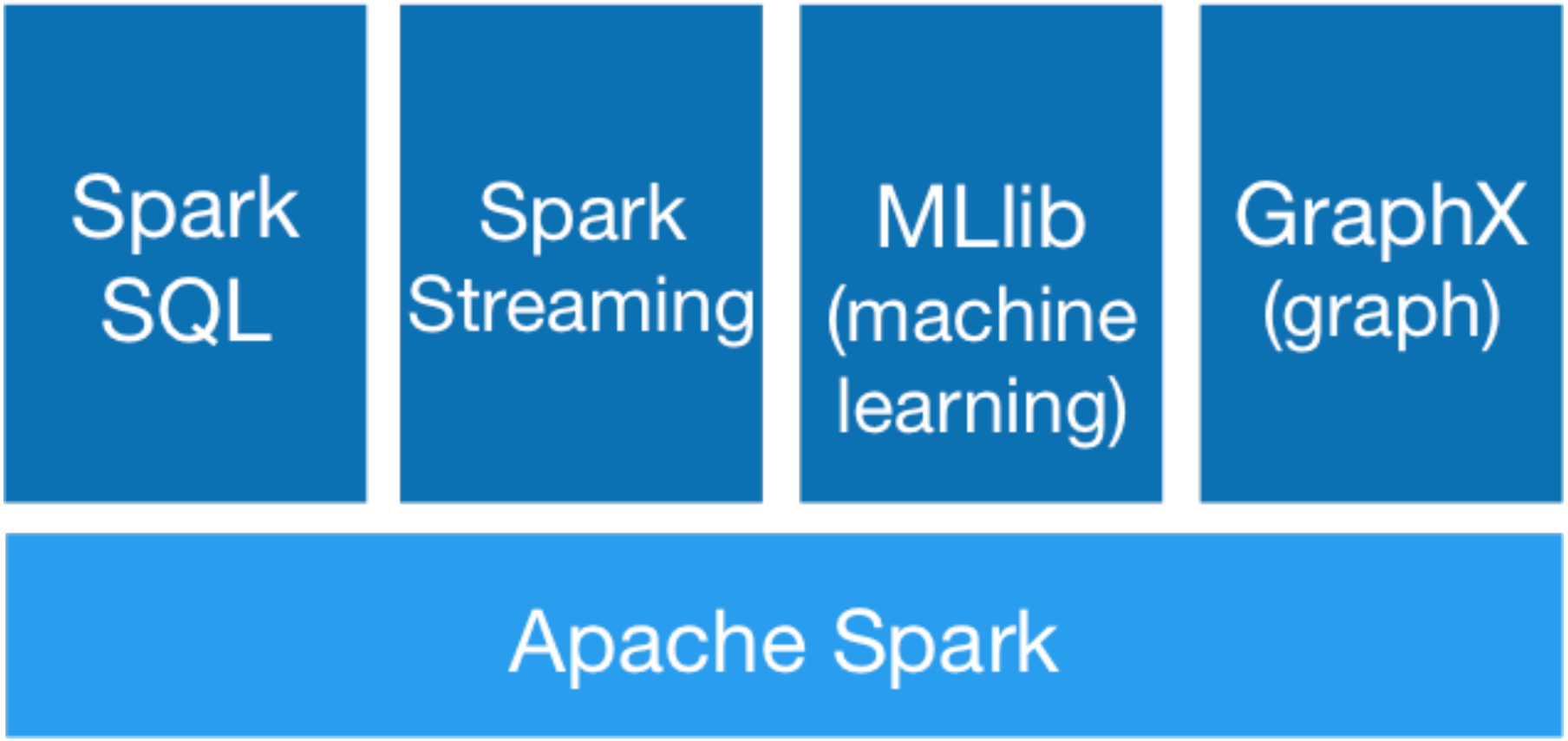
# Spark and Hadoop





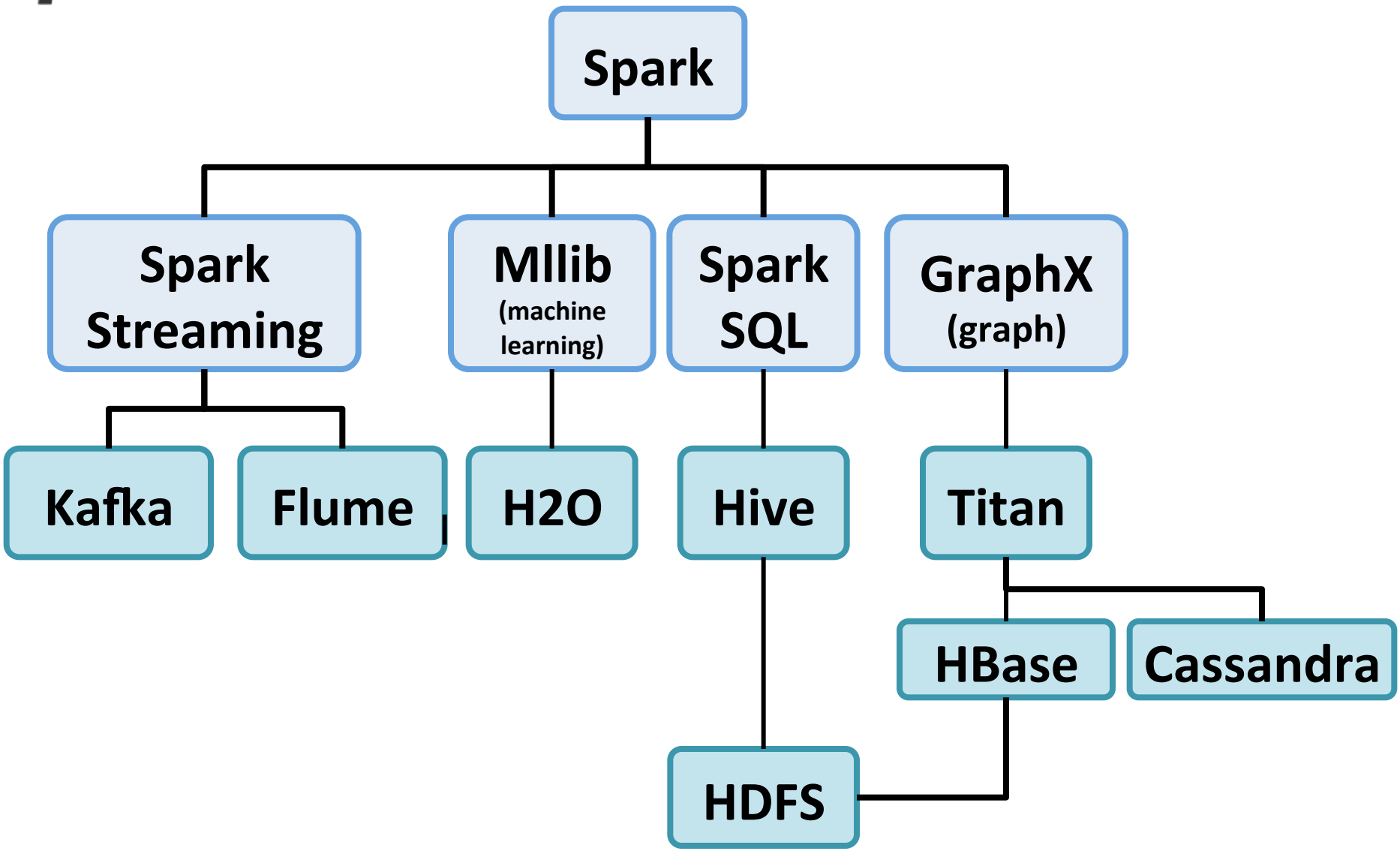

Source: http://spark.apache.org/

# Spark Ecosystem



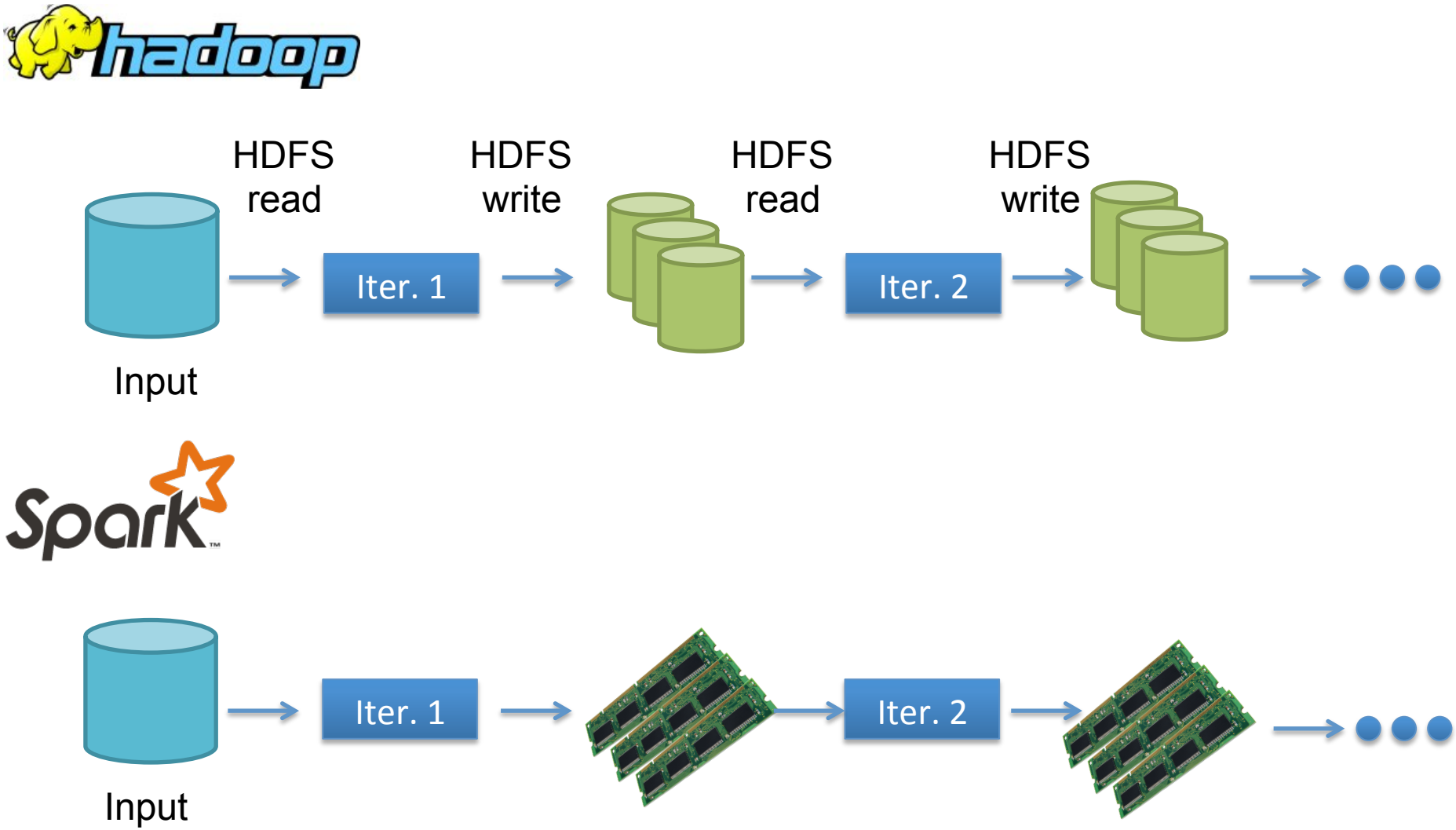

Source: http://spark.apache.org/

# Spark Ecosystem




Source: Mike Frampton (2015), Mastering Apache Spark, Packt Publishing

# Hadoop vs. Spark




Source: Shiva Achari (2015), Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing

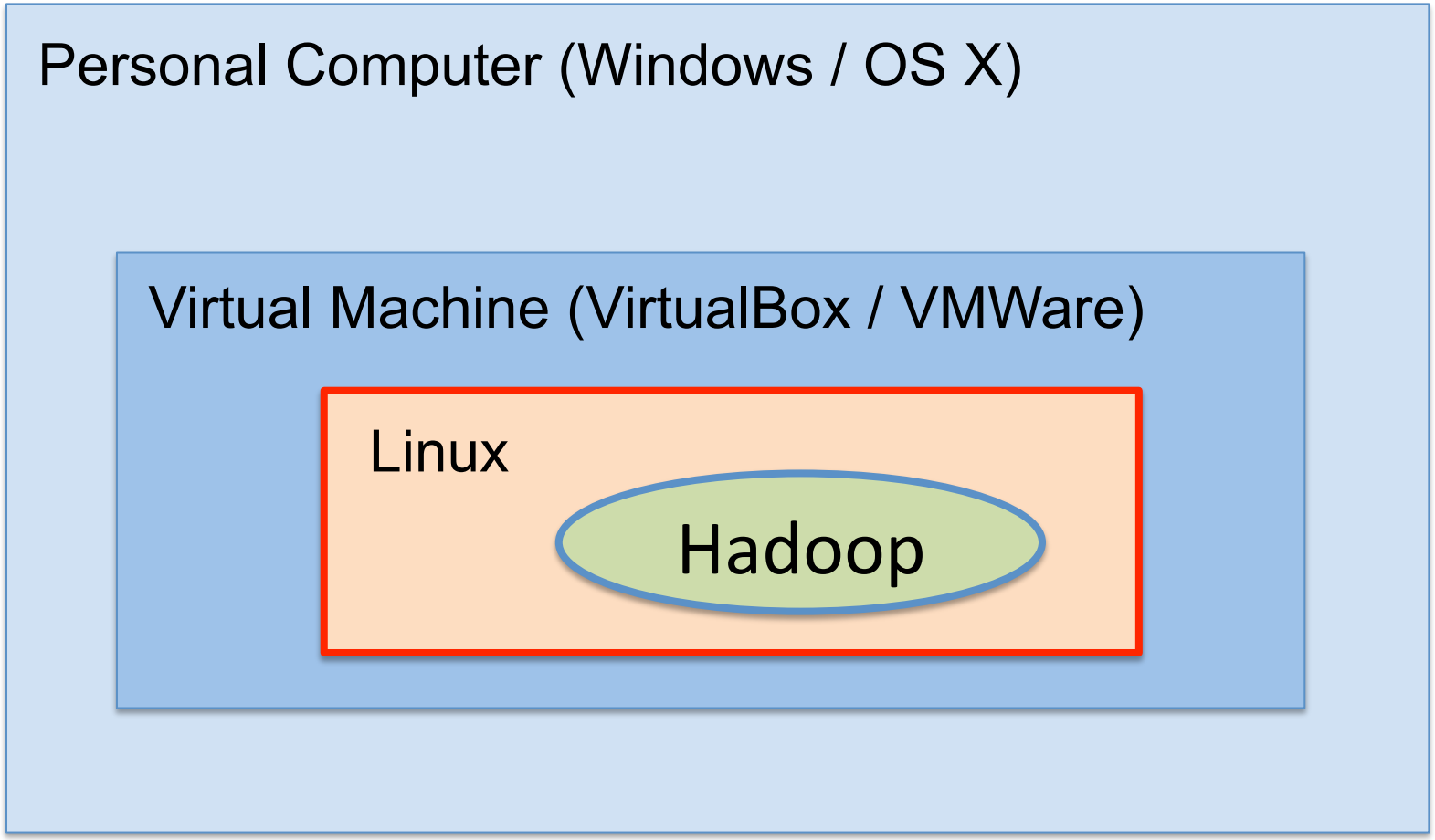
# Steps to Install Hadoop on a Personal Computer (Windows/OS X)

Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

# Hodoop: Linux Based Software

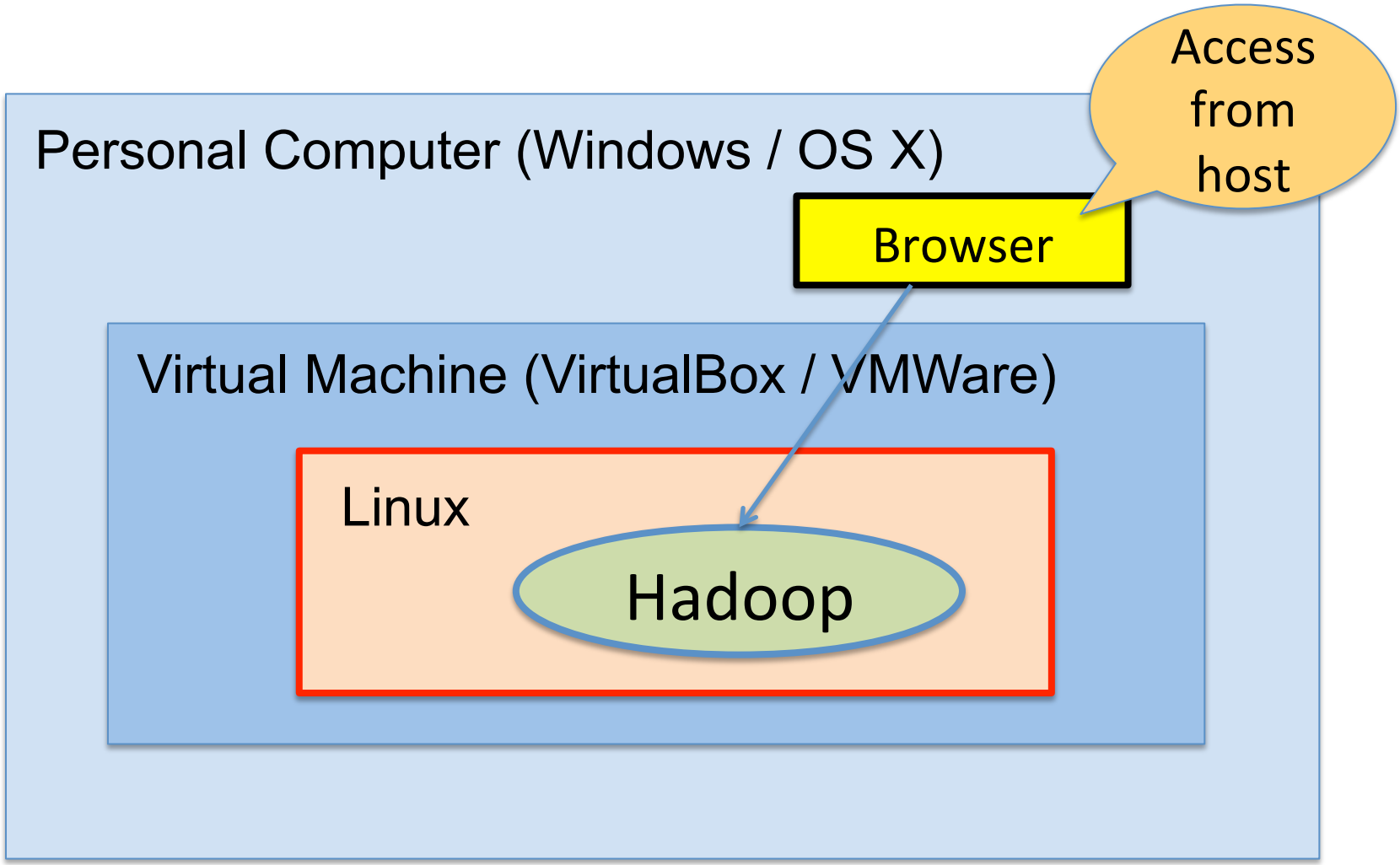

Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

# Appliance



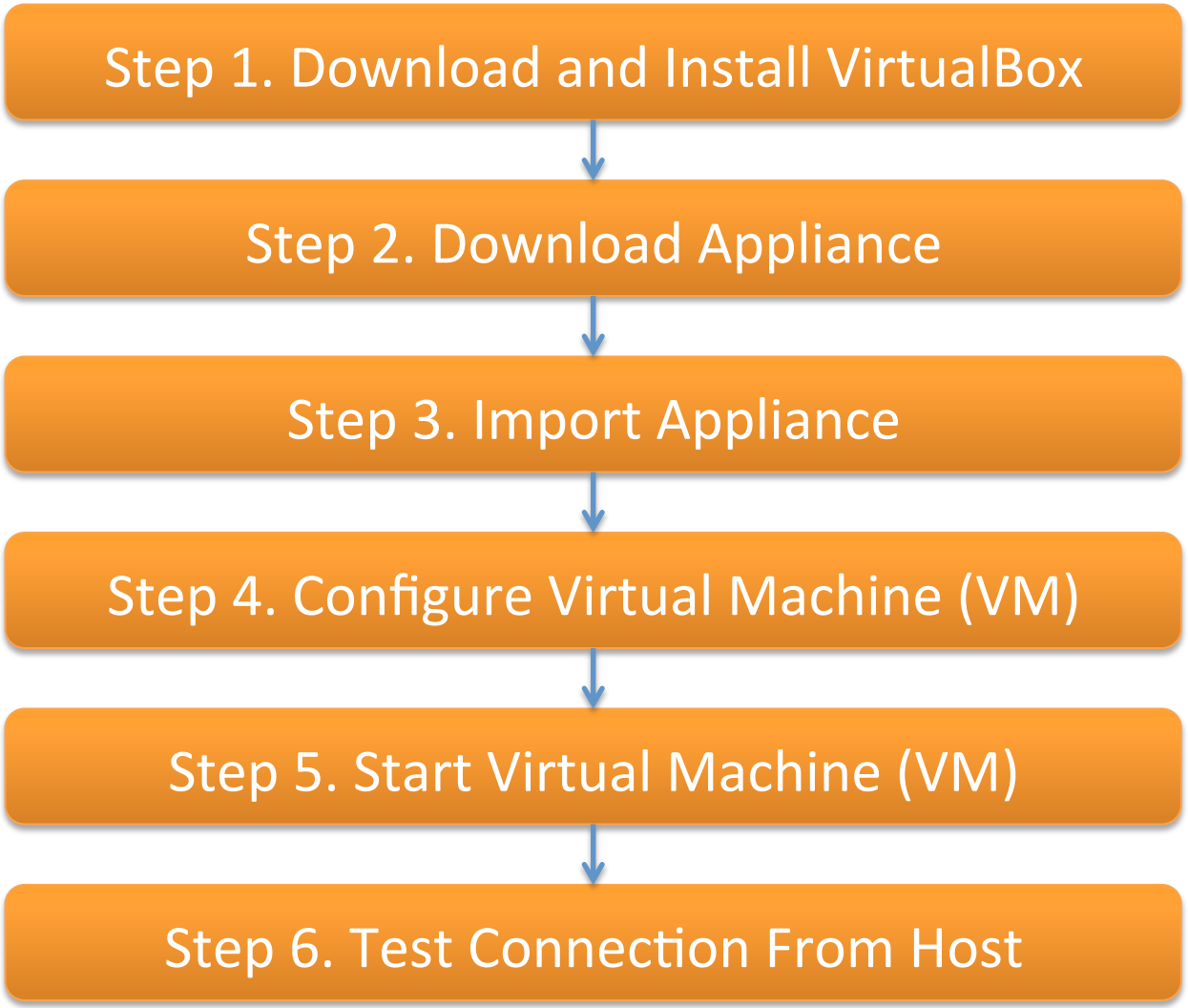

Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

# Connection to Hadoop




Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

# Steps to Install Hadoop on a Personal Computer (Windows/OS X)




Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

# Virtual Box


https://www.virtualbox.org/

# Steps to Install Hadoop on a Personal Computer (Windows/OS X)


Step 1. Download and Install VirtualBox

Step 2. Download Appliance

Hortonworks Sandbox

Step 3. Import Appliance

Step 4. Configure Virtual Machine (VM)

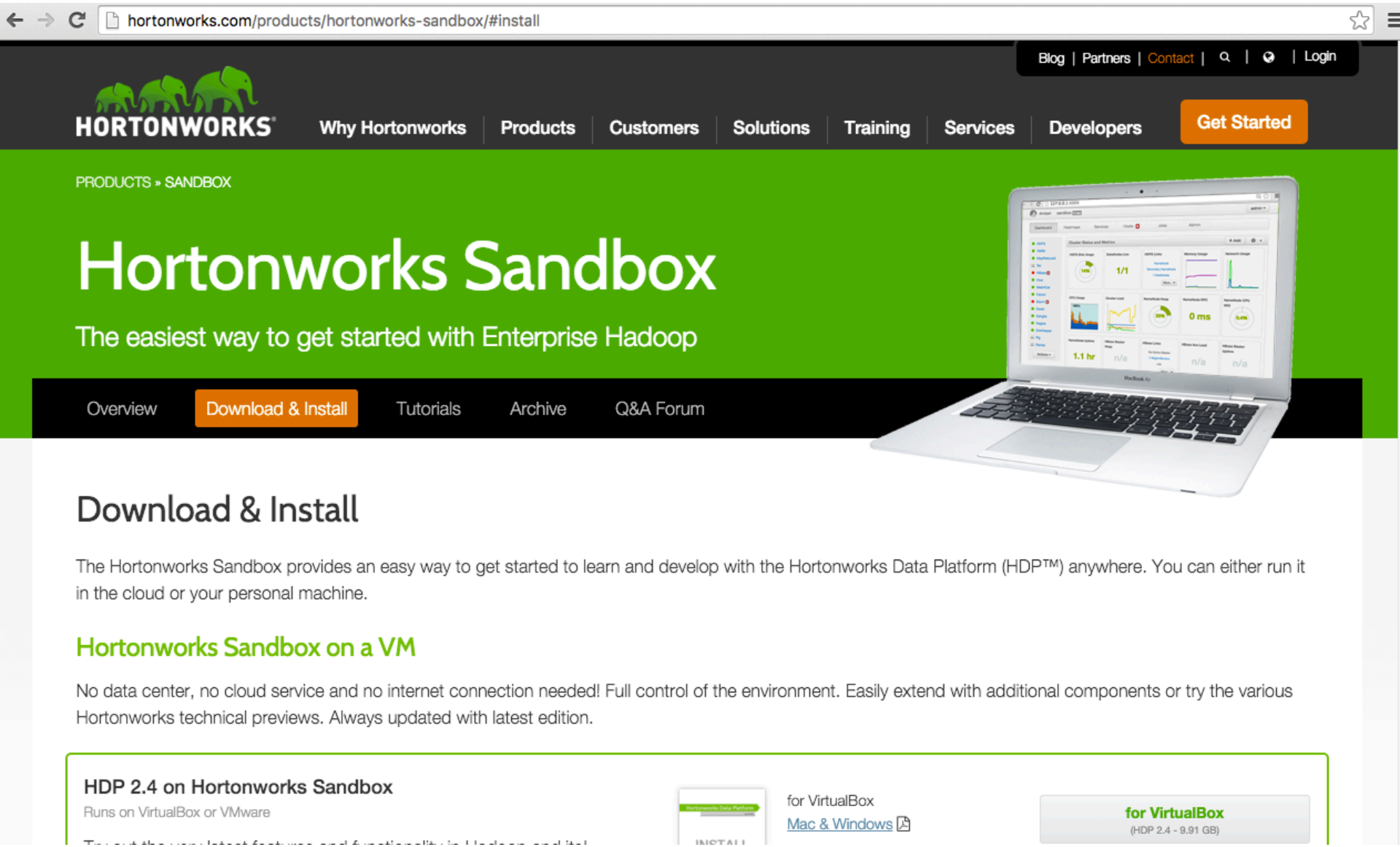
Step 5. Start Virtual Machine (VM)

Step 6. Test Connection From Host

Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

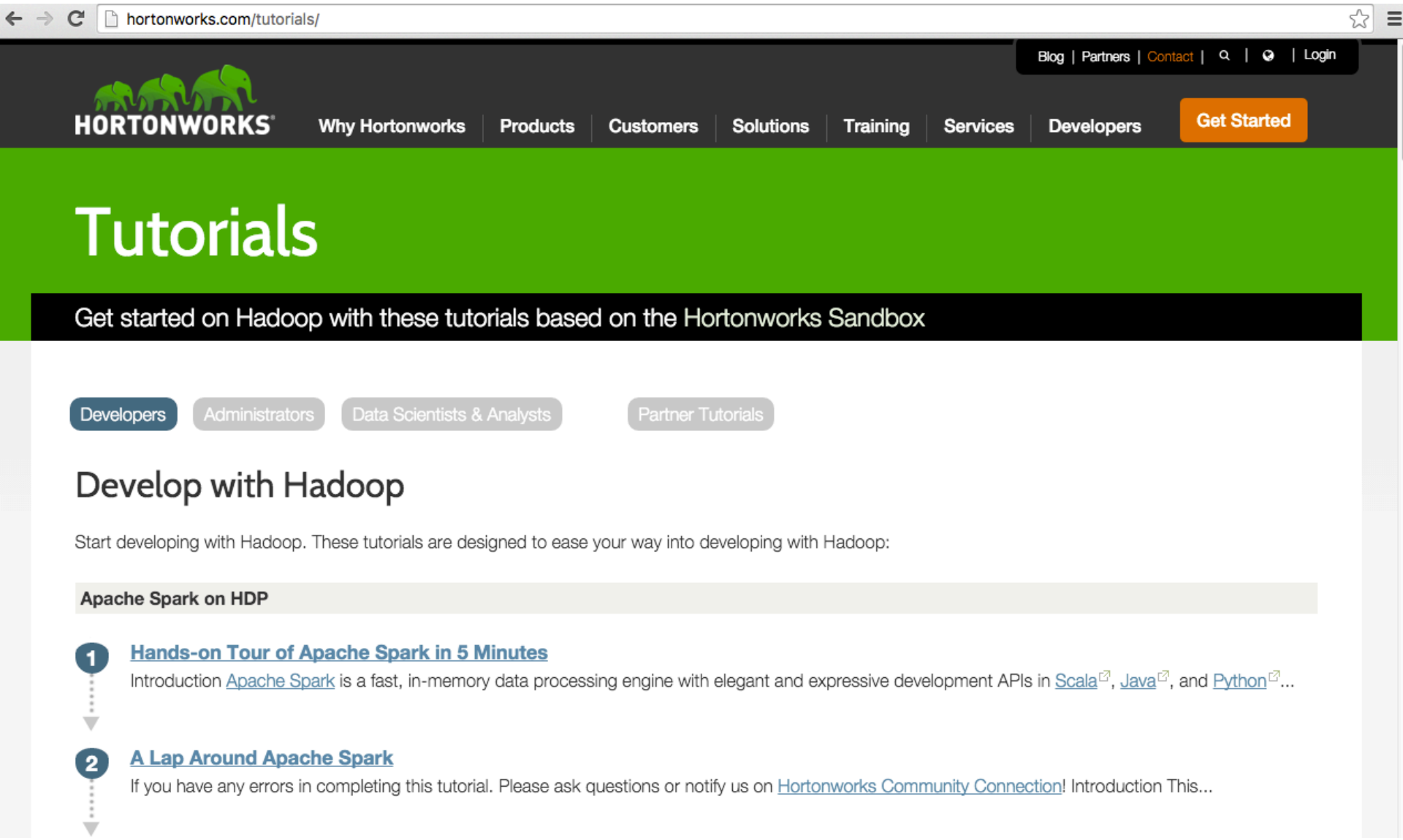
# Hortonworks Sandbox

## The easiest way to get started with Enterprise Hadoop




http://hortonworks.com/products/hortonworks-sandbox/#install

# Get started on Hadoop with these tutorials based on the Hortonworks Sandbox



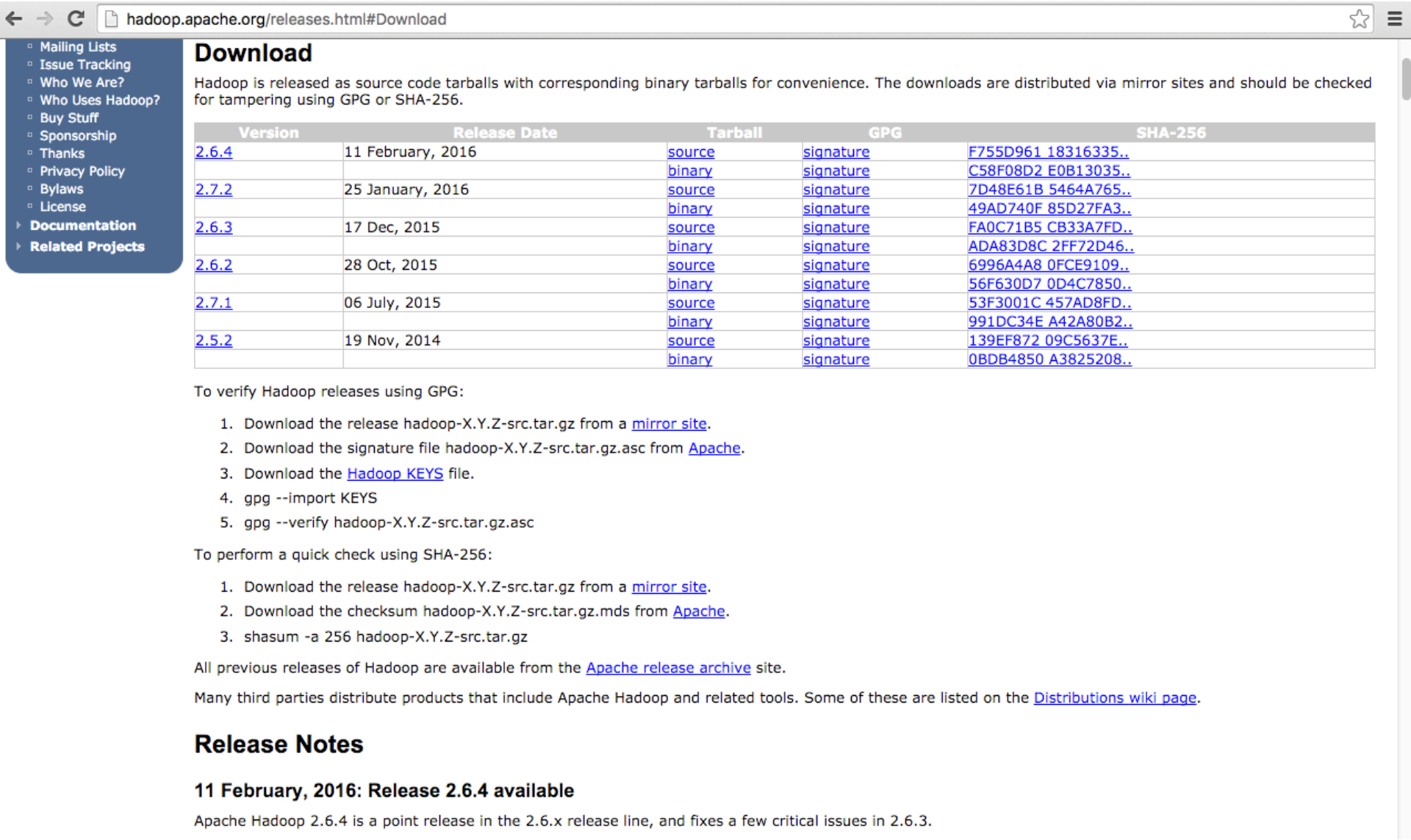

http://hortonworks.com/tutorials/

# Apache Hadoop


http://hadoop.apache.org/

# Apache Hadoop

http://hadoop.apache.org/releases.html#Download



# Apache Hadoop

hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/releasenotes.html

## Hadoop 2.7.2 Release Notes

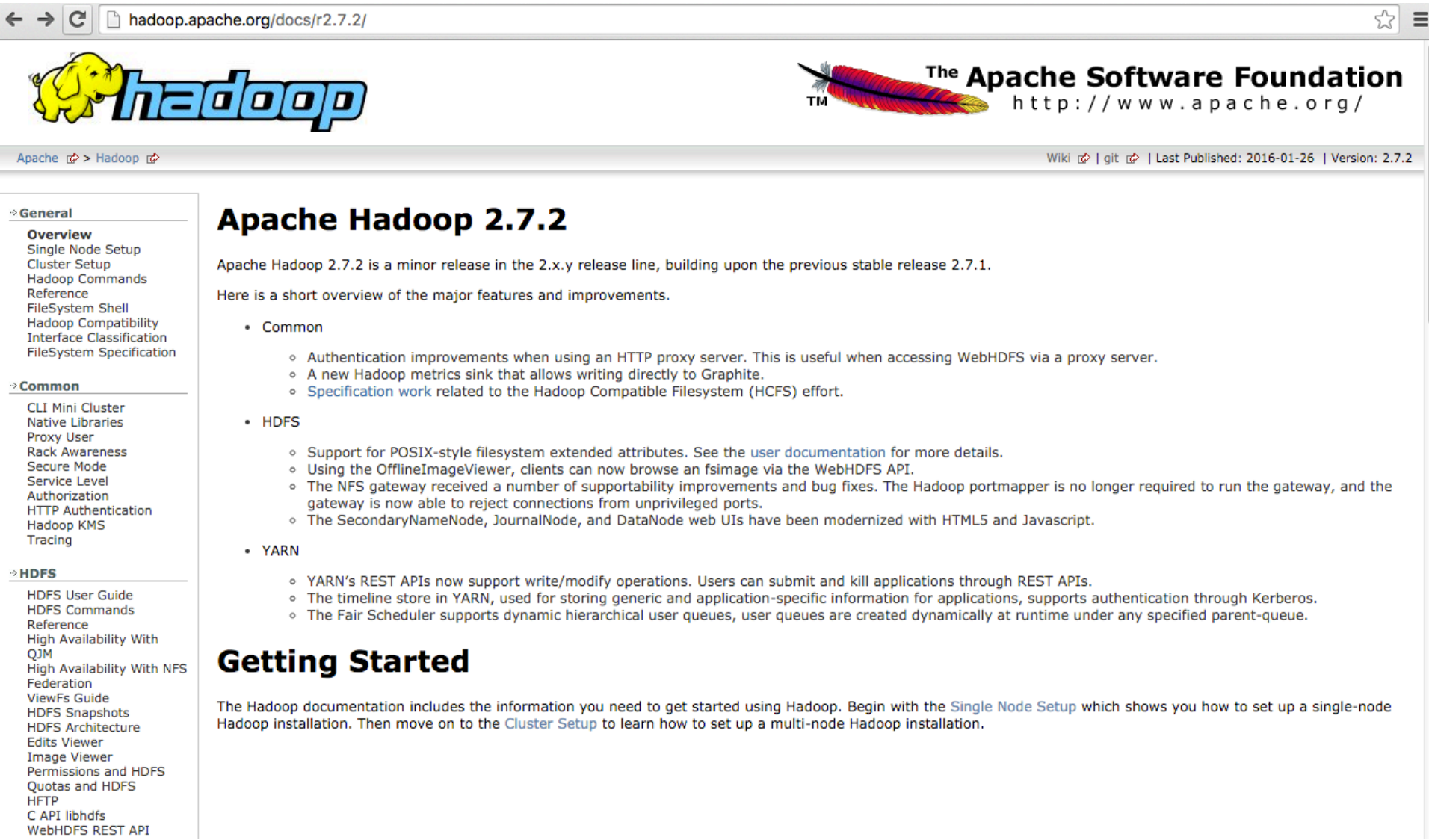
These release notes include new developer and user-facing incompatibilities, features, and major improvements.

### Changes since Hadoop 2.7.1

- YARN-4434. Minor bug reported by Takashi Ohnishi and fixed by Weiwei Yang (documentation , nodemanager)
  **NodeManager Disk Checker parameter documentation is not correct**
- YARN-4424. Blocker bug reported by Yesha Vora and fixed by Jian He
  **Fix deadlock in RMAppImpl**
- YARN-4365. Major bug reported by Jason Lowe and fixed by Kuhu Shukla (resourcemanager)
  **FileSystemNodeLabelStore should check for root dir existence on startup**
- YARN-4354. Blocker bug reported by Jason Lowe and fixed by Jason Lowe (nodemanager)
  **Public resource localization fails with NPE**
- YARN-4348. Blocker bug reported by Tsuyoshi Ozawa and fixed by Tsuyoshi Ozawa
  **ZKRMStateStore.syncInternal shouldn't wait for sync completion for avoiding blocking ZK's event thread**
- YARN-4344. Critical bug reported by Varun Vasudev and fixed by Varun Vasudev (resourcemanager)
  **NMs reconnecting with changed capabilities can lead to wrong cluster resource calculations**
- YARN-4326. Major bug reported by MENG DING and fixed by MENG DING
  **Fix TestDistributedShell timeout as AHS in MiniYarnCluster no longer binds to default port 8188**
- YARN-4321. Major bug reported by Varun Saxena and fixed by Varun Saxena (resourcemanager)
  **Incessant retries if NoAuthException is thrown by Zookeeper in non HA mode**
- YARN-4320. Major bug reported by Varun Saxena and fixed by Varun Saxena
  **TestJobHistoryEventHandler fails as AHS in MiniYarnCluster no longer binds to default port 8188**
- YARN-4313. Major bug reported by Jian He and fixed by Jian He
  **Race condition in MiniMRYarnCluster when getting history server address**
- YARN-4312. Major bug reported by Varun Saxena and fixed by Varun Saxena
  **TestSubmitApplicationWithRMHA fails on branch-2.7 and branch-2.6 as some of the test cases time out**

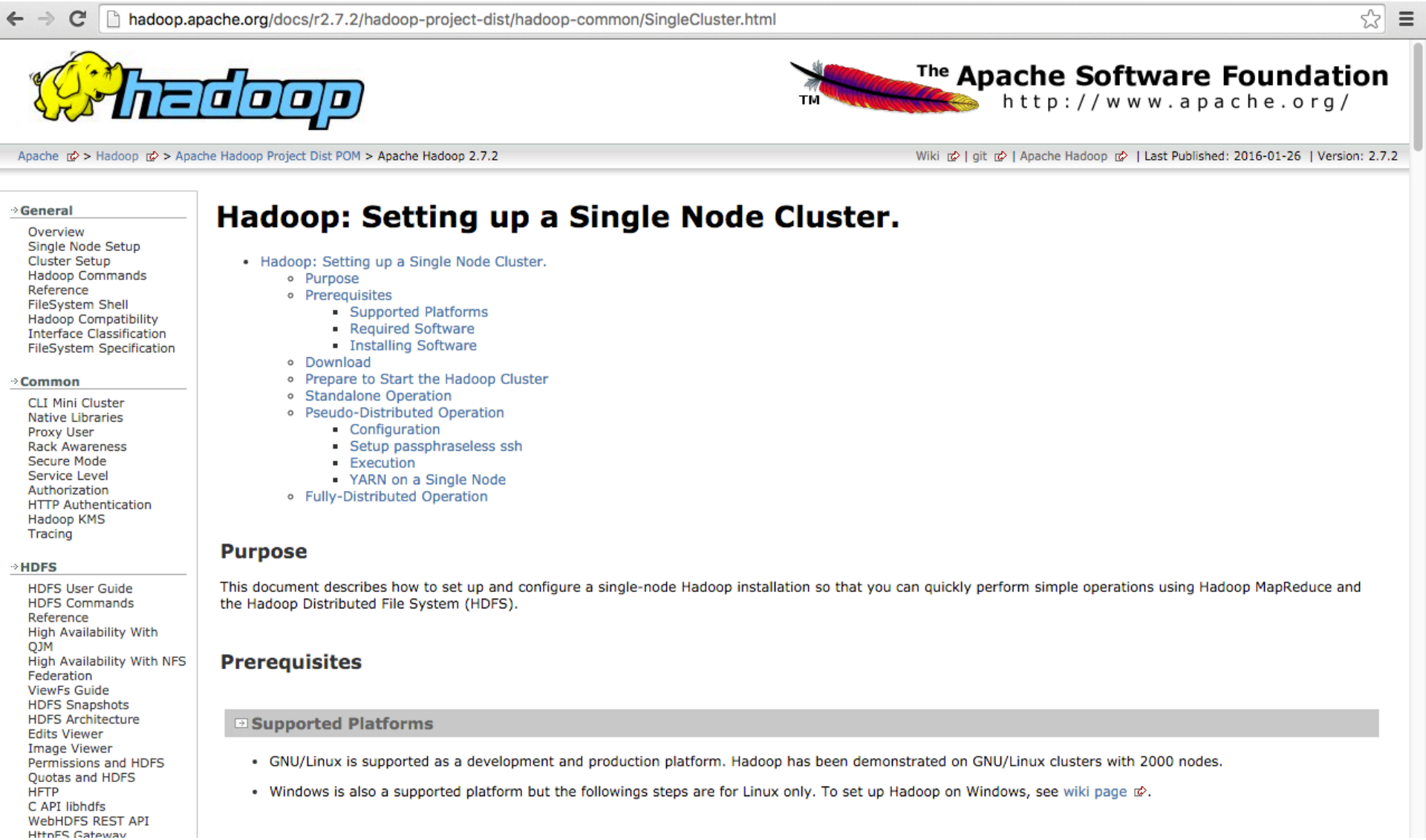
Source: http://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/releasenotes.html

# Apache Hadoop 2.7.2




Source: http://hadoop.apache.org/docs/r2.7.2/

# Hadoop: Setting up a Single Node Cluster




Source: http://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/SingleCluster.html

# Hadoop Cluster Setup

hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/ClusterSetup.html






## Hadoop Cluster Setup

### Purpose

This document describes how to install and configure Hadoop clusters ranging from a few nodes to extremely large clusters with thousands of nodes. To play with Hadoop, you may first want to install it on a single machine (see Single Node Setup).

This document does not cover advanced topics such as Security or High Availability.

### Prerequisites

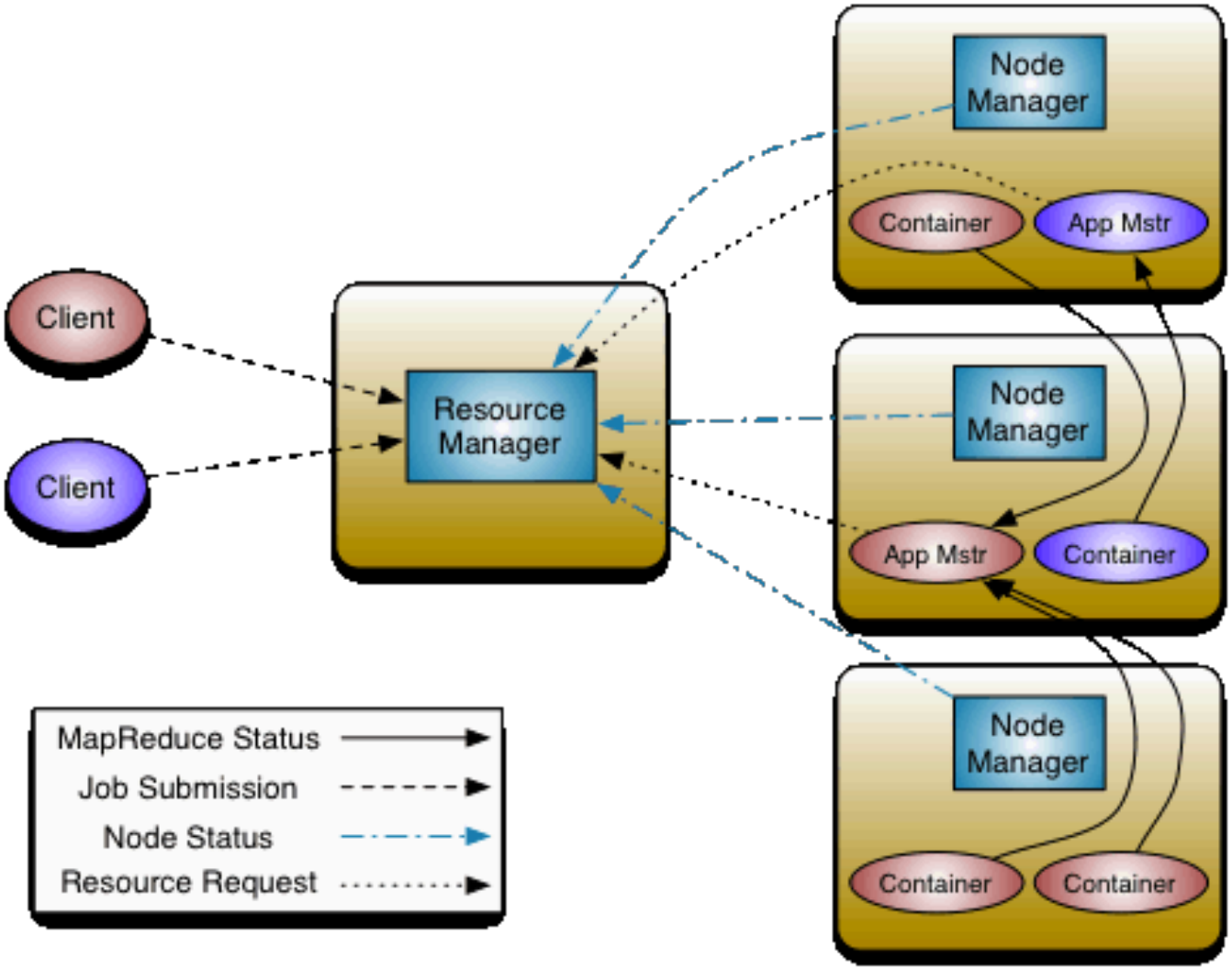
- Install Java. See the Hadoop Wiki for known good versions.
- Download a stable version of Hadoop from Apache mirrors.

### Installation

Installing a Hadoop cluster typically involves unpacking the software on all the machines in the cluster or installing it via a packaging system as appropriate for your

Source: http://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/ClusterSetup.html

# Apache Hadoop YARN




Source: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

# Apache Spark

spark.apache.org

Spark *Lightning-fast cluster computing*

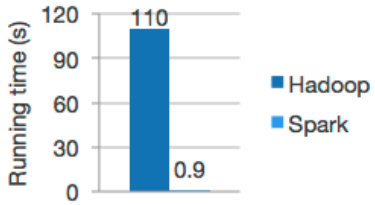
Download | Libraries | Documentation | Examples | Community | FAQ | Apache Software Foundation

**Apache Spark™** is a fast and general engine for large-scale data processing.

## Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.




Logistic regression in Hadoop and Spark

### Latest News

Submission is open for Spark Summit San Francisco (Feb 11, 2016)

Spark Summit East (Feb 16, 2016, New York) agenda posted (Jan 14, 2016)

Spark 1.6.0 released (Jan 04, 2016)

CFP for Spark Summit East 2016 is closing soon! (Nov 19, 2015)

Archive

Download Spark

## Ease of Use

Write applications quickly in Java, Scala, Python, R.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python and R shells.

```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

### Built-in Libraries:

SQL and DataFrames
Spark Streaming
MLlib (machine learning)
GraphX (graph)

Third-Party Packages

http://spark.apache.org/

# References


- EMC Education Services (2015),
  Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley
- Shiva Achari (2015),
  Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing
- Mike Frampton (2015),
  Mastering Apache Spark, Packt Publishing
- Deepak Ramanathan (2014),
  SAS Modernization architectures - Big Data Analytics, http://www.slideshare.net/deepakramanathan/sas-modernization-architectures-big-data-analytics