

Social Computing and Big Data Analytics

社群運算與大數據分析

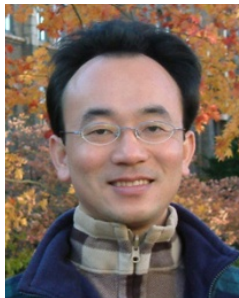
Fundamental Big Data:

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
(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
(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
(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,

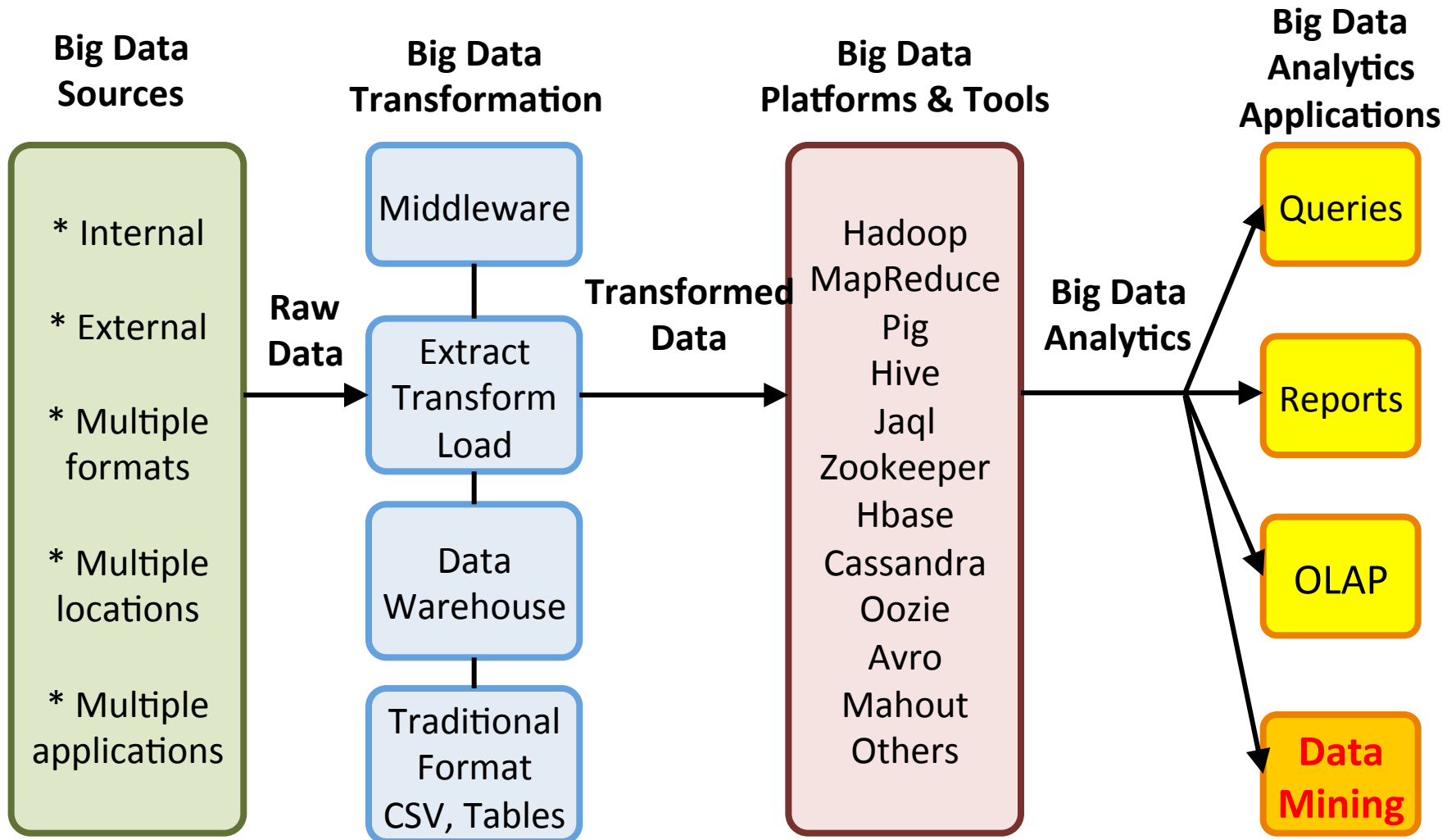
Hadoop and Spark Ecosystem

(大數據基礎：

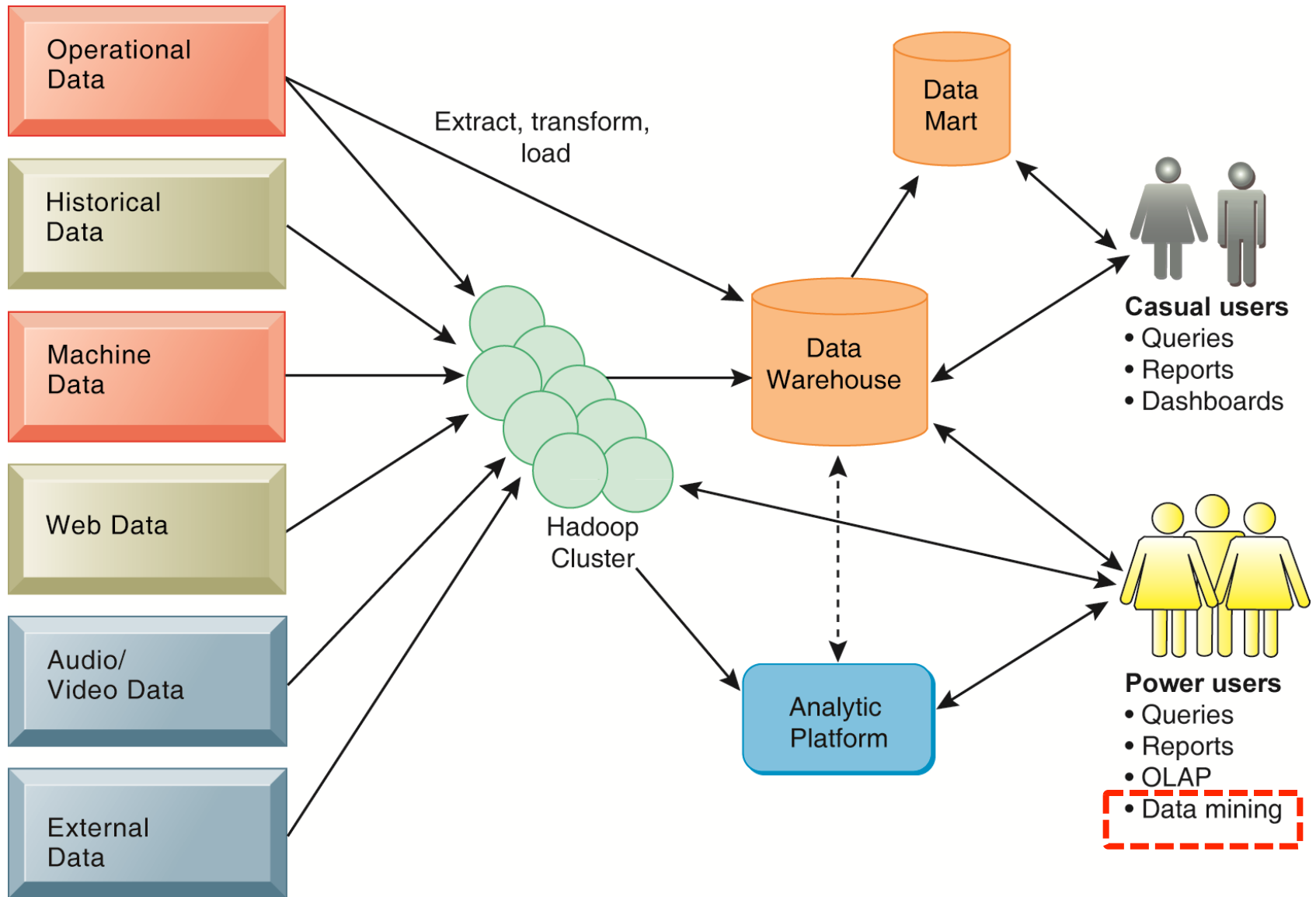
MapReduce典範、

Hadoop與Spark生態系統)

Architecture of Big Data Analytics



Business Intelligence (BI) Infrastructure



Fundamental Big Data:
MapReduce Paradigm,
Hadoop and Spark
Ecosystem



VISUAL ANALYTICS

DYNAMIC & INTERACTIVE

Dashboard Graph
Map

ENHANCE

Understanding Investigation
User Experience



BIG ANALYTICS

QUERY & FILTER

Complex queries
 R^2I^2

DETECT

Anomalies
Communities
Typologies

PREDICT

Tending
Real-time
Prediction

DECIDE

Simulation
Optimization



BIG DATA – Batch



BIG DATA – Real Time



Complex by nature



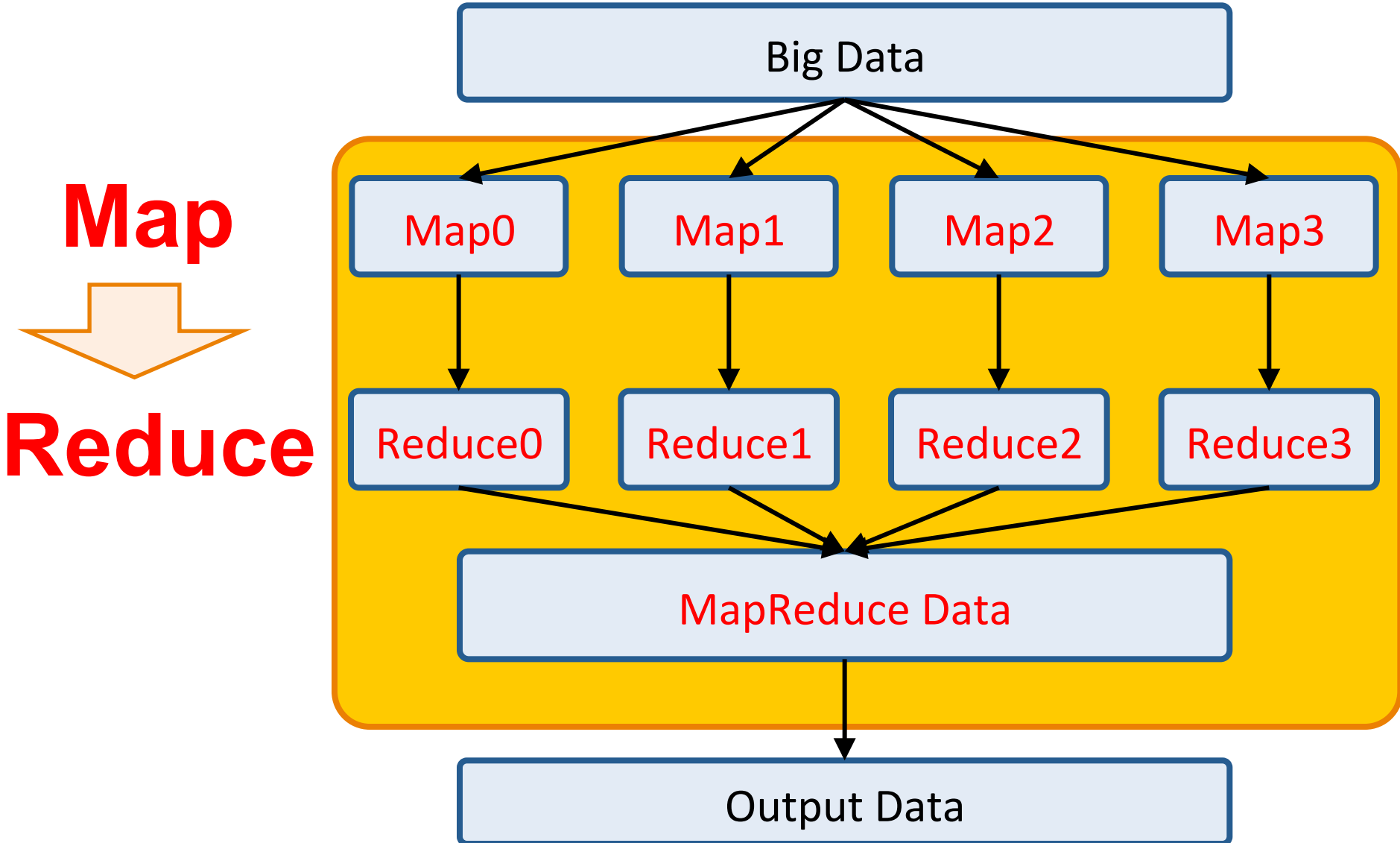
DATA

Complex by structure



MapReduce Paradigm

MapReduce Paradigm



Hadoop Ecosystem



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



MapReduce

Processing

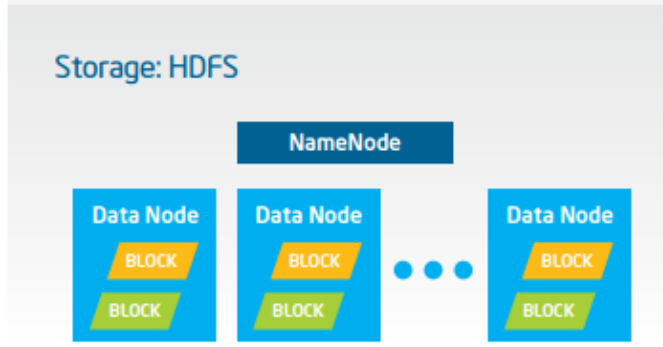
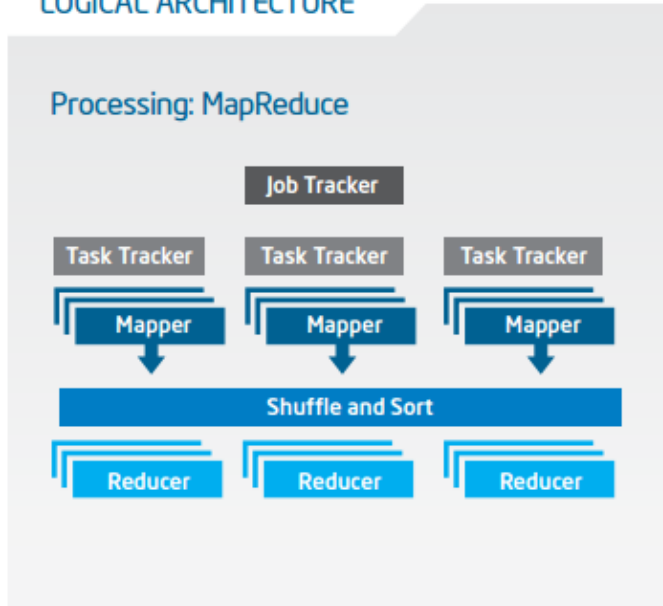


HDFS

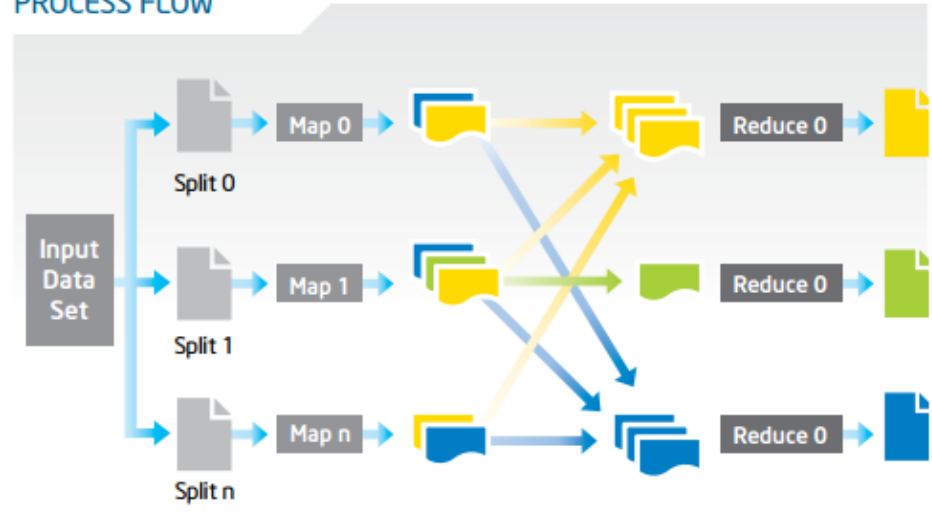
Storage

Big Data with Hadoop Architecture

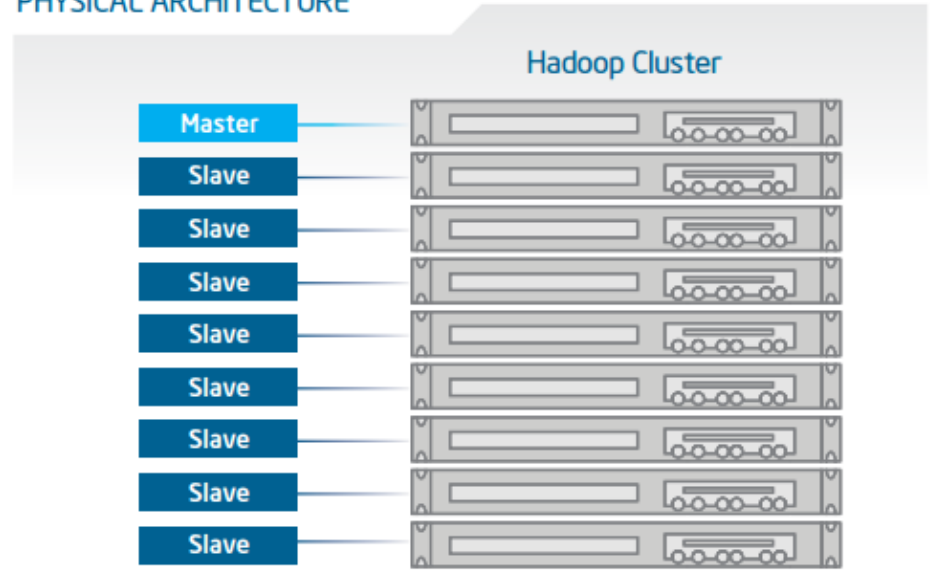
LOGICAL ARCHITECTURE



PROCESS FLOW



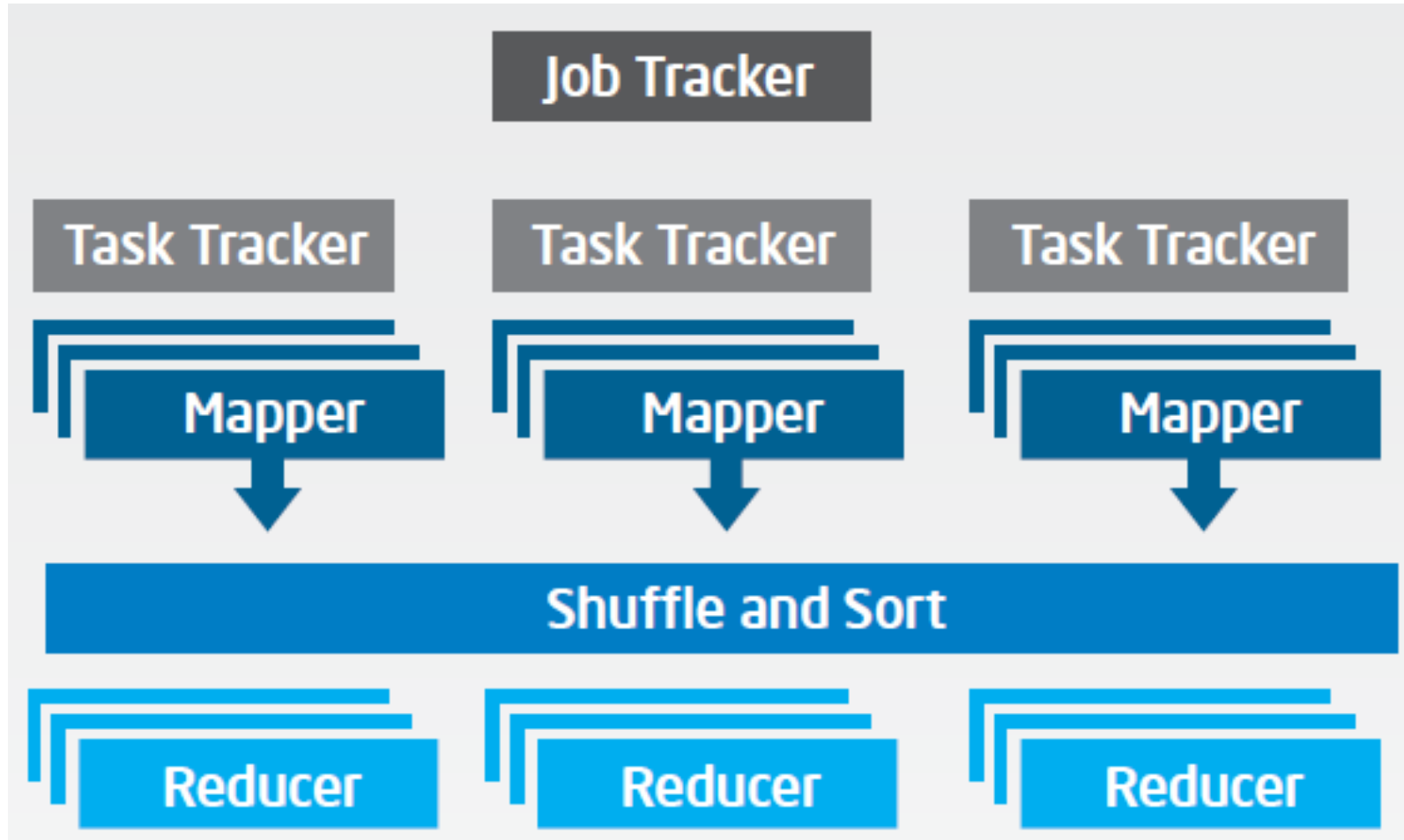
PHYSICAL ARCHITECTURE



Big Data with Hadoop Architecture

Logical Architecture

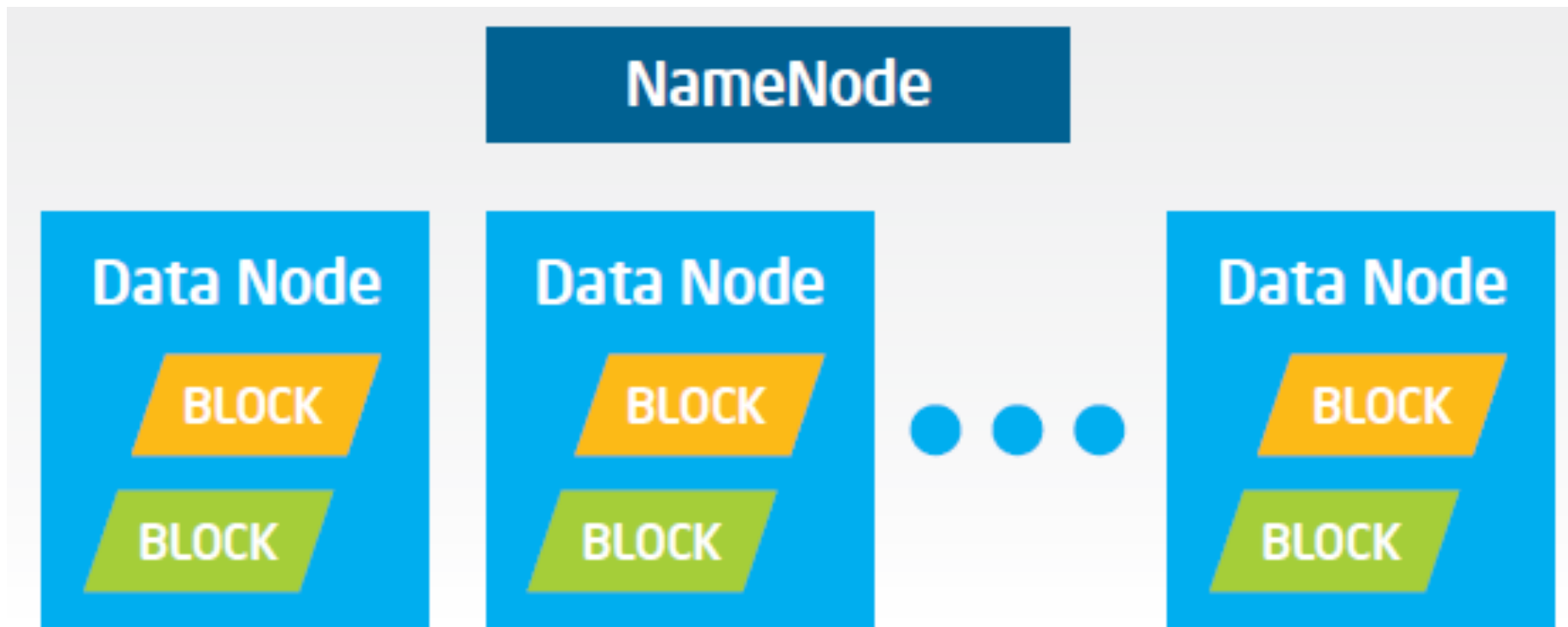
Processing: MapReduce



Big Data with Hadoop Architecture

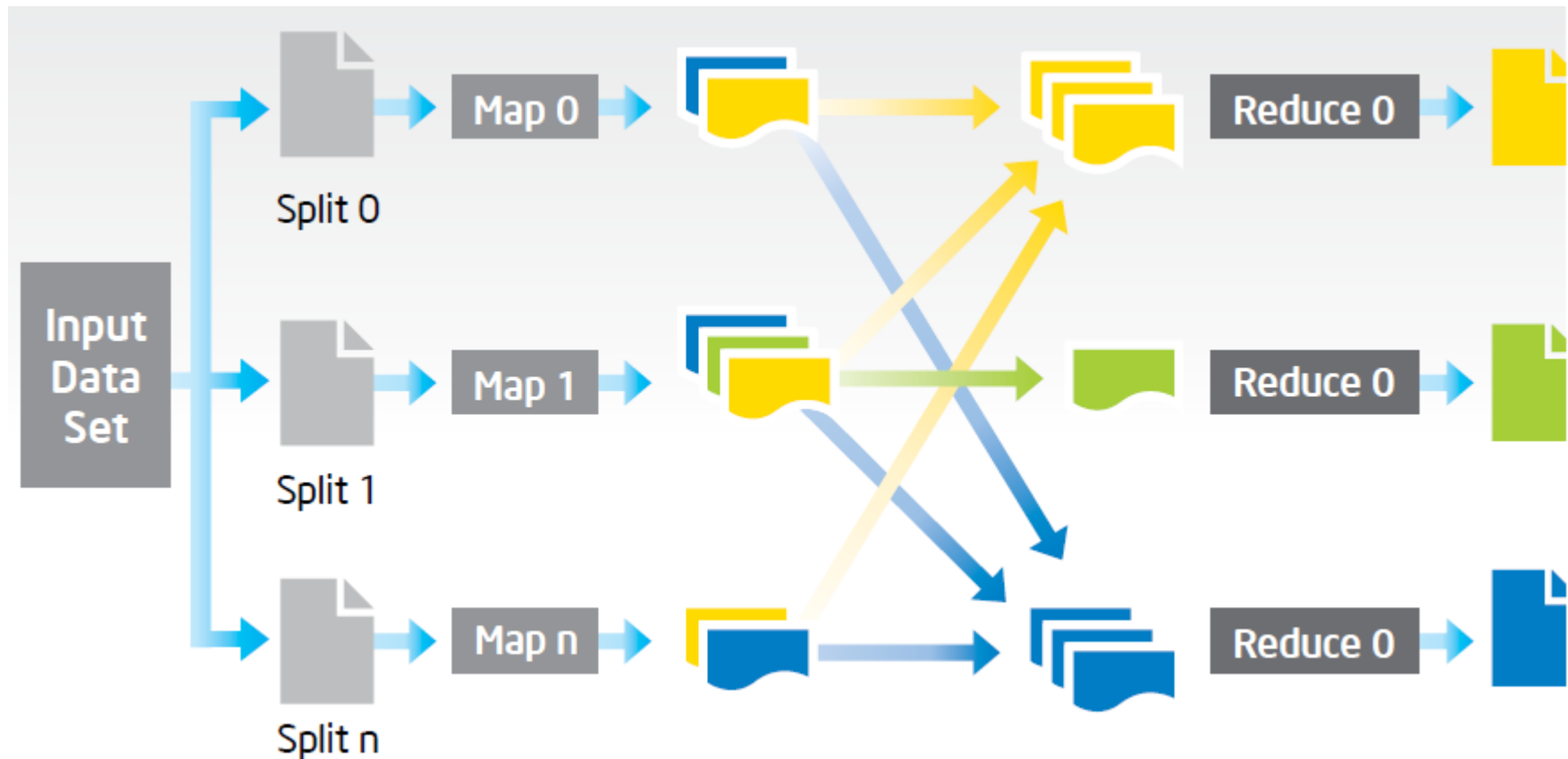
Logical Architecture

Storage: HDFS



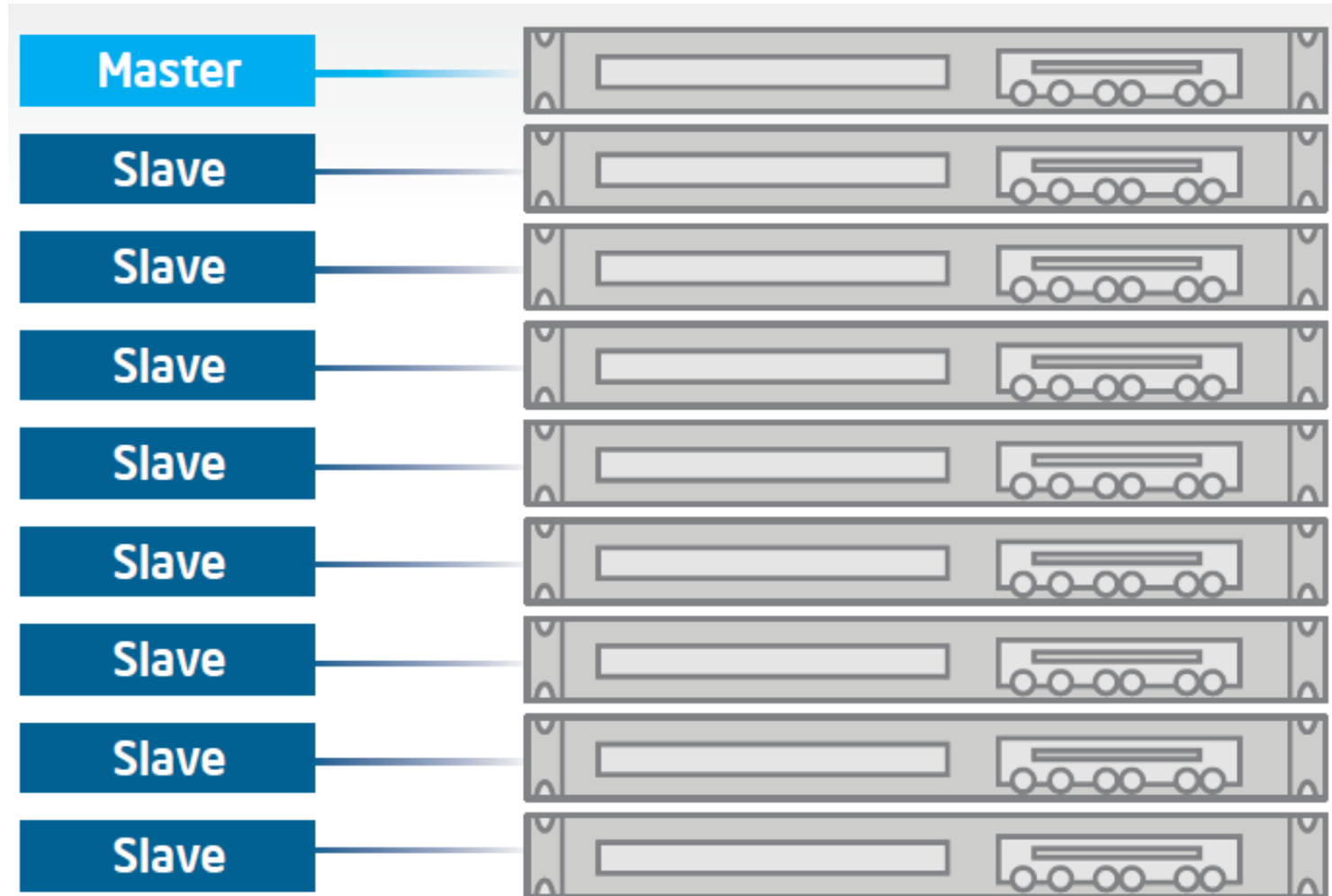
Big Data with Hadoop Architecture

Process Flow

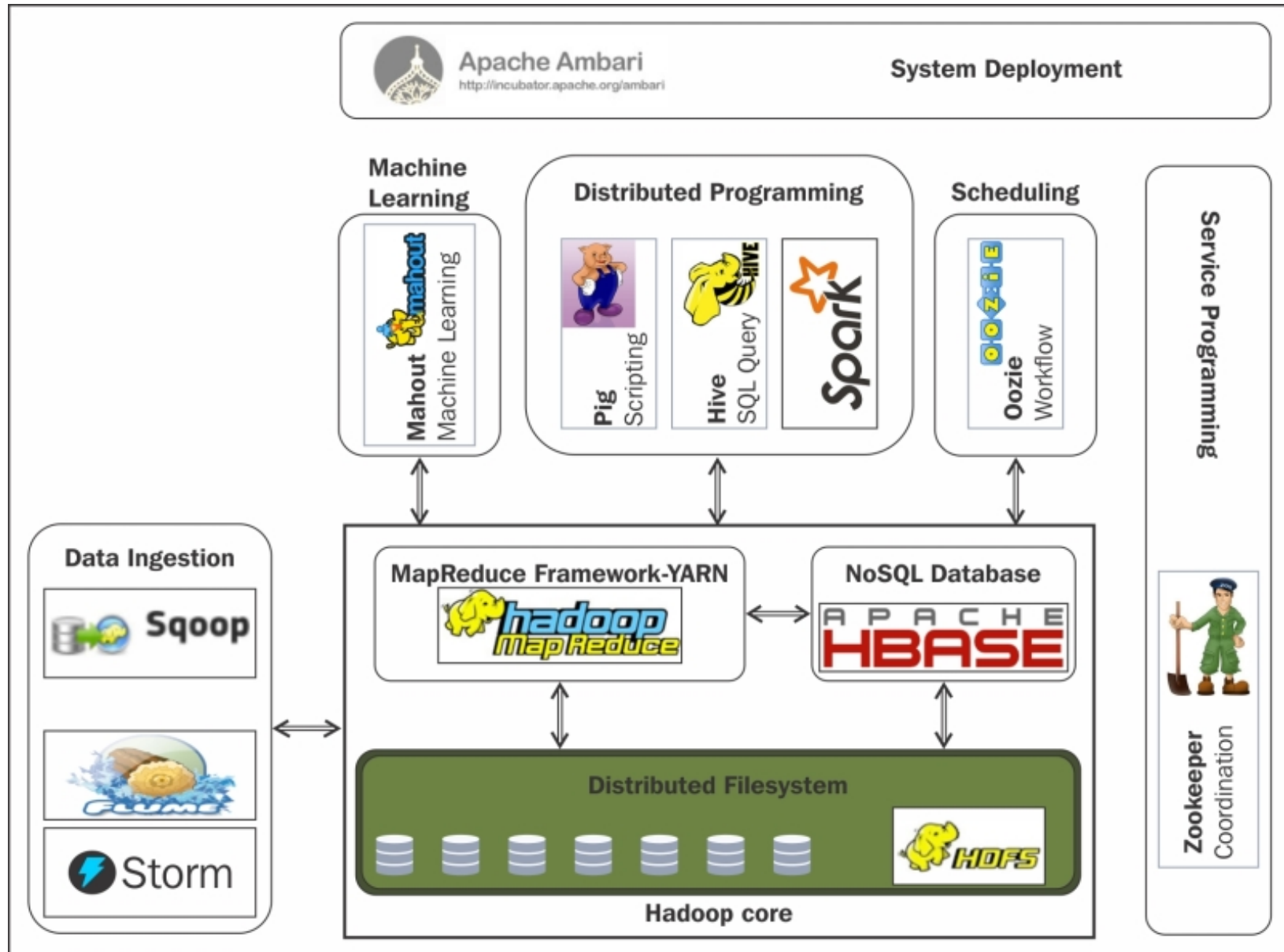


Big Data with Hadoop Architecture

Hadoop Cluster

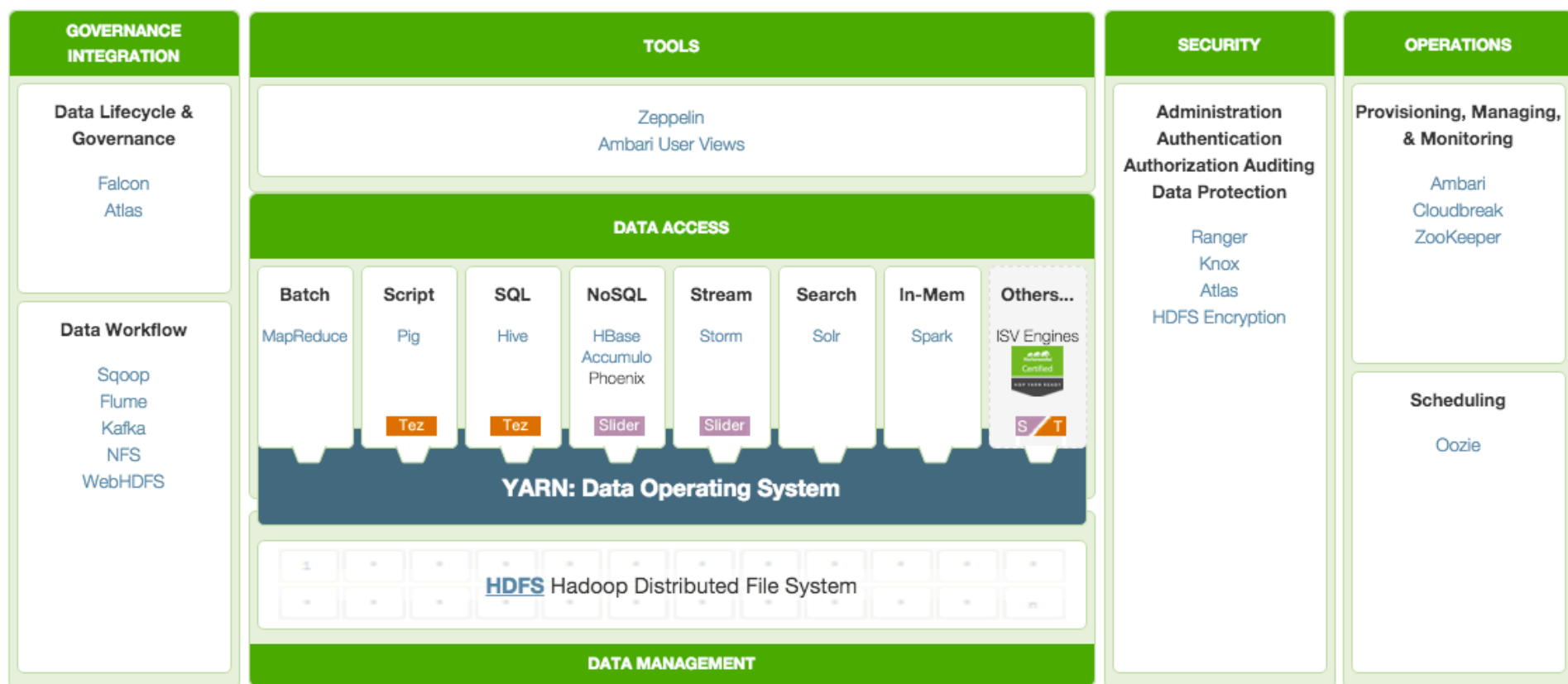


Hadoop Ecosystem



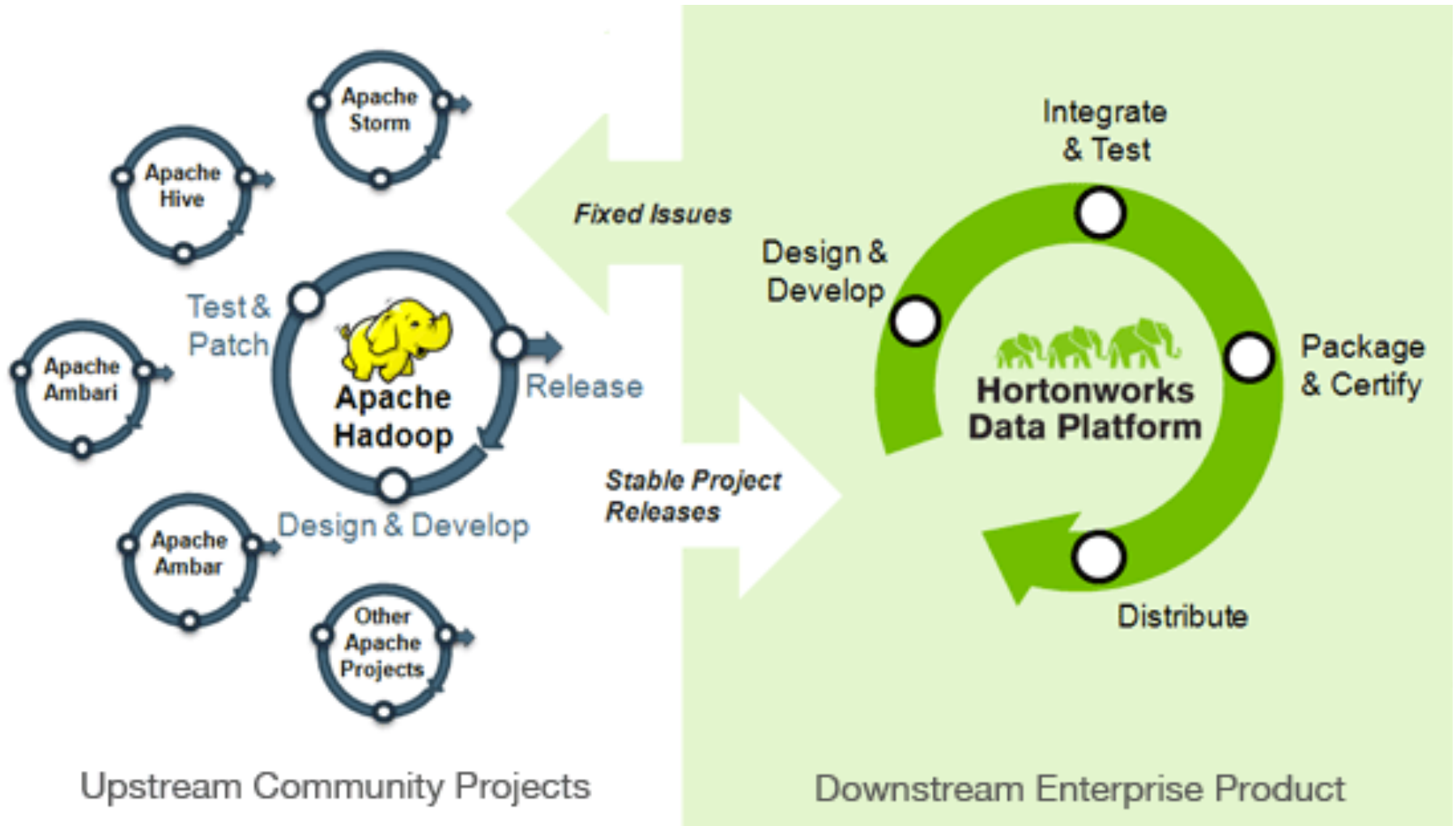
HDP (Hortonworks Data Platform)

A Complete Enterprise Hadoop Data Platform

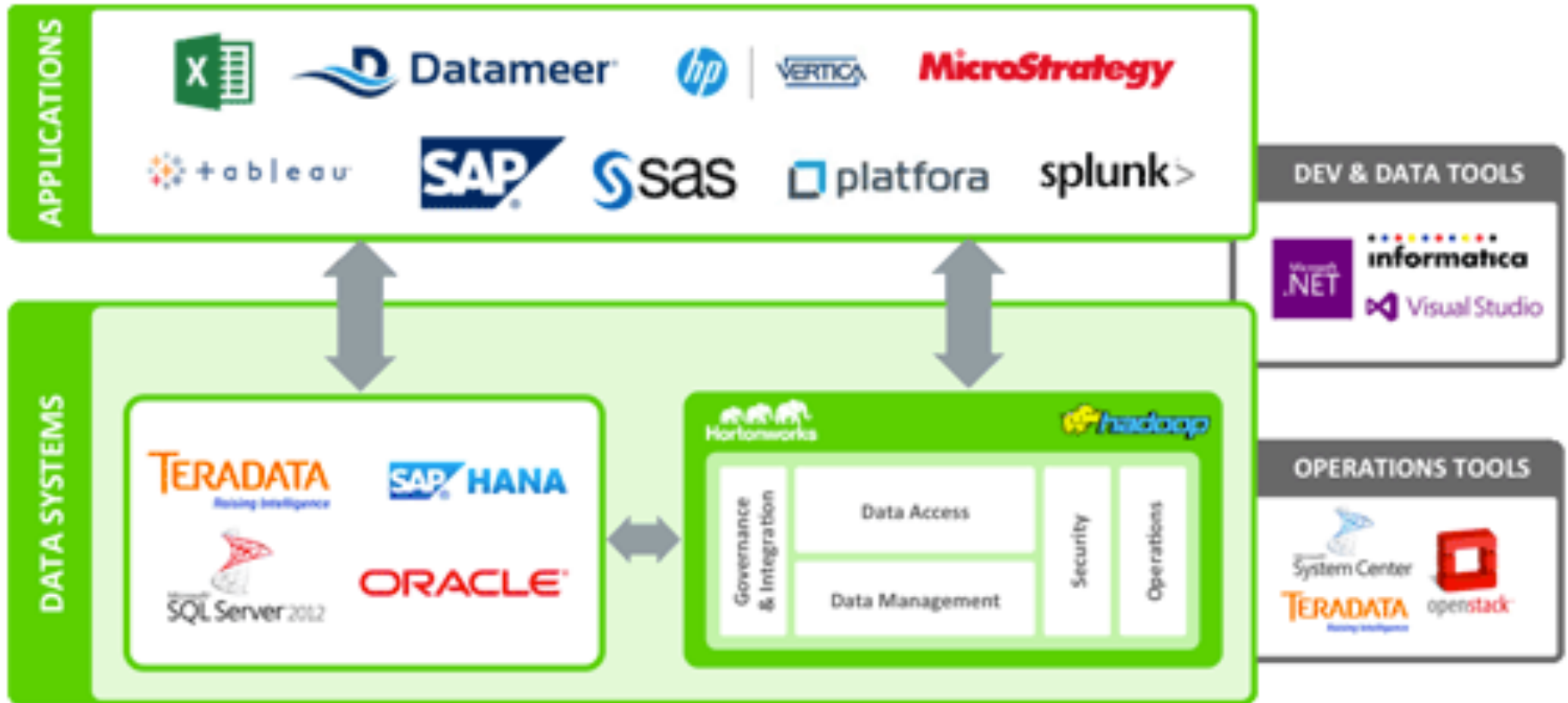


Apache Hadoop

Hortonworks Data Platform



Hadoop and Data Analytics Tools



Hadoop 1 → Hadoop 2

Hadoop 1

- Silos & Largely batch
- Single Processing engine



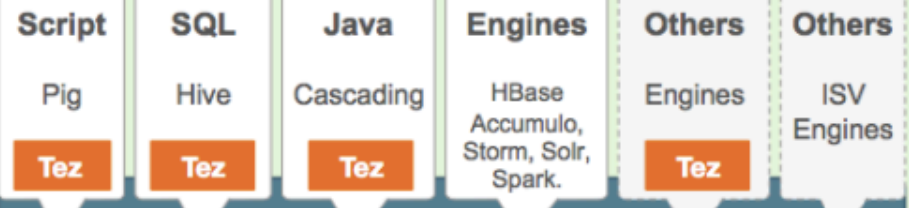
MapReduce

(Cluster Resource Management & Data Processing)



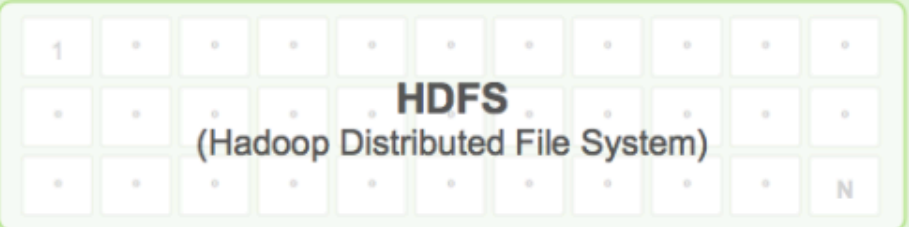
Hadoop 2 w/ Tez

- Multiple Engines, Single Data Set
- Batch, Interactive & Real-Time

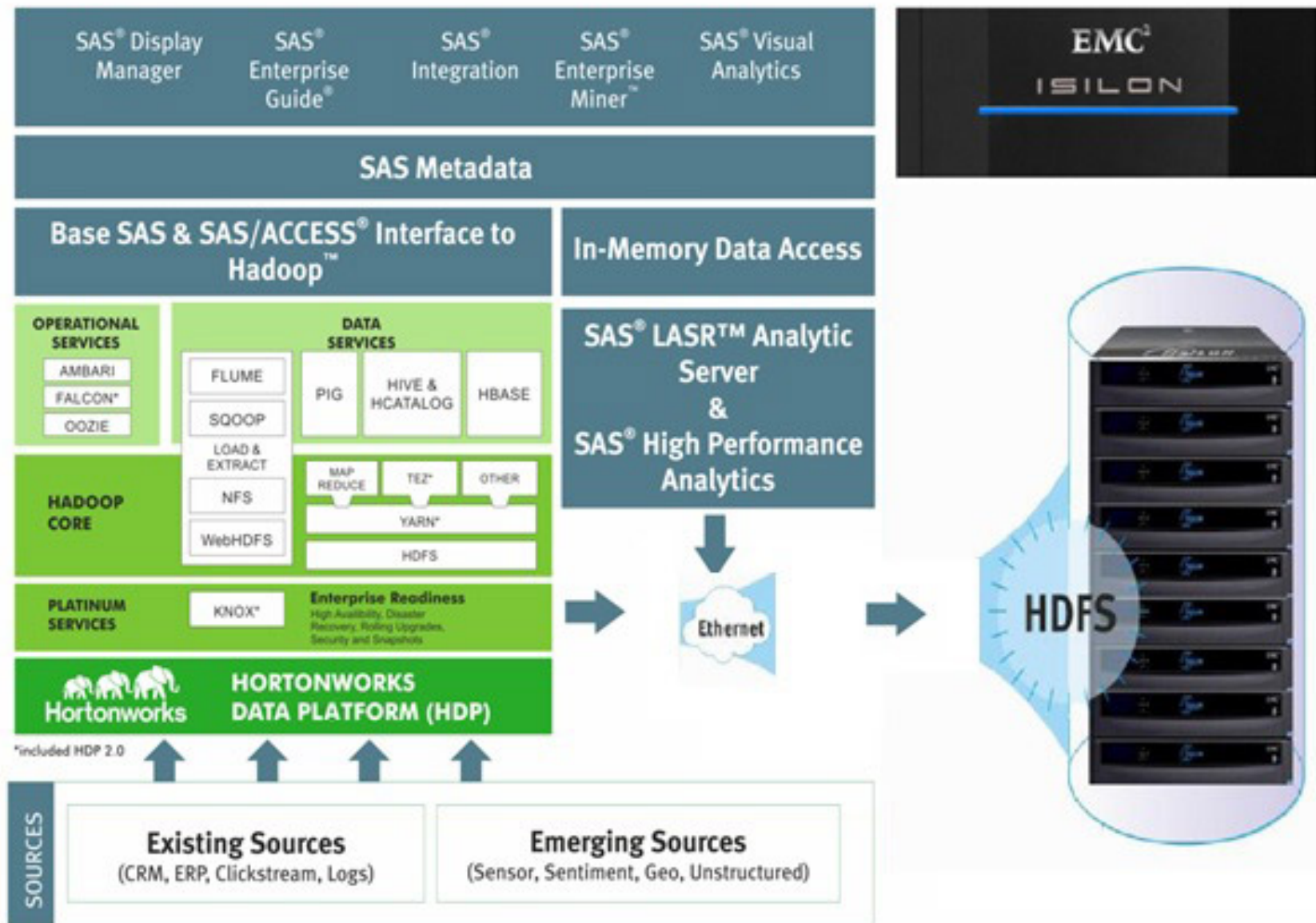


YARN: Data Operating System

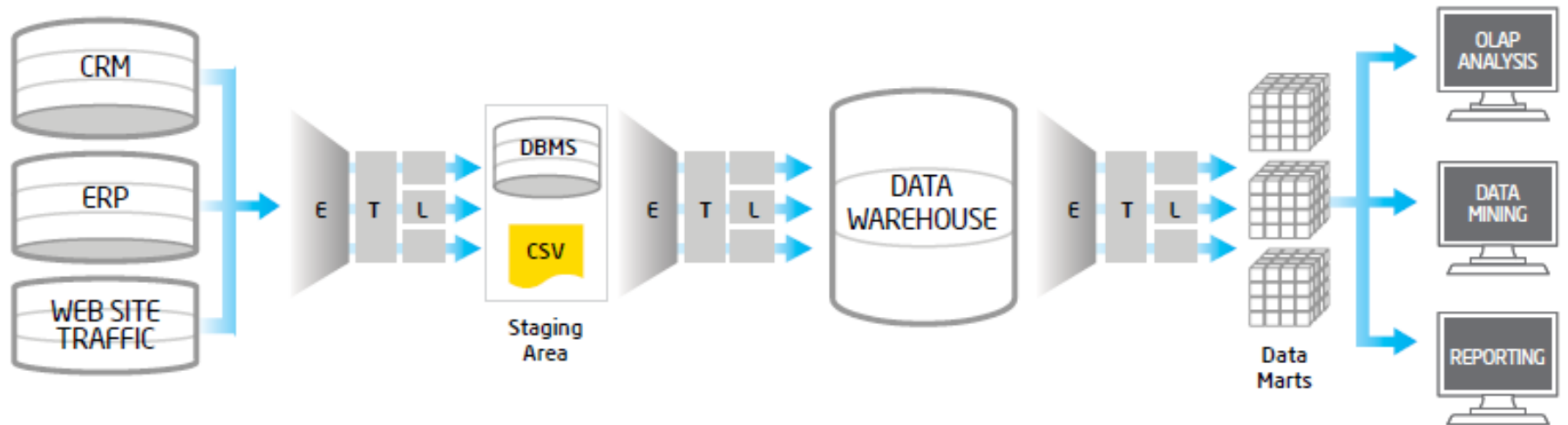
(Cluster Resource Management)



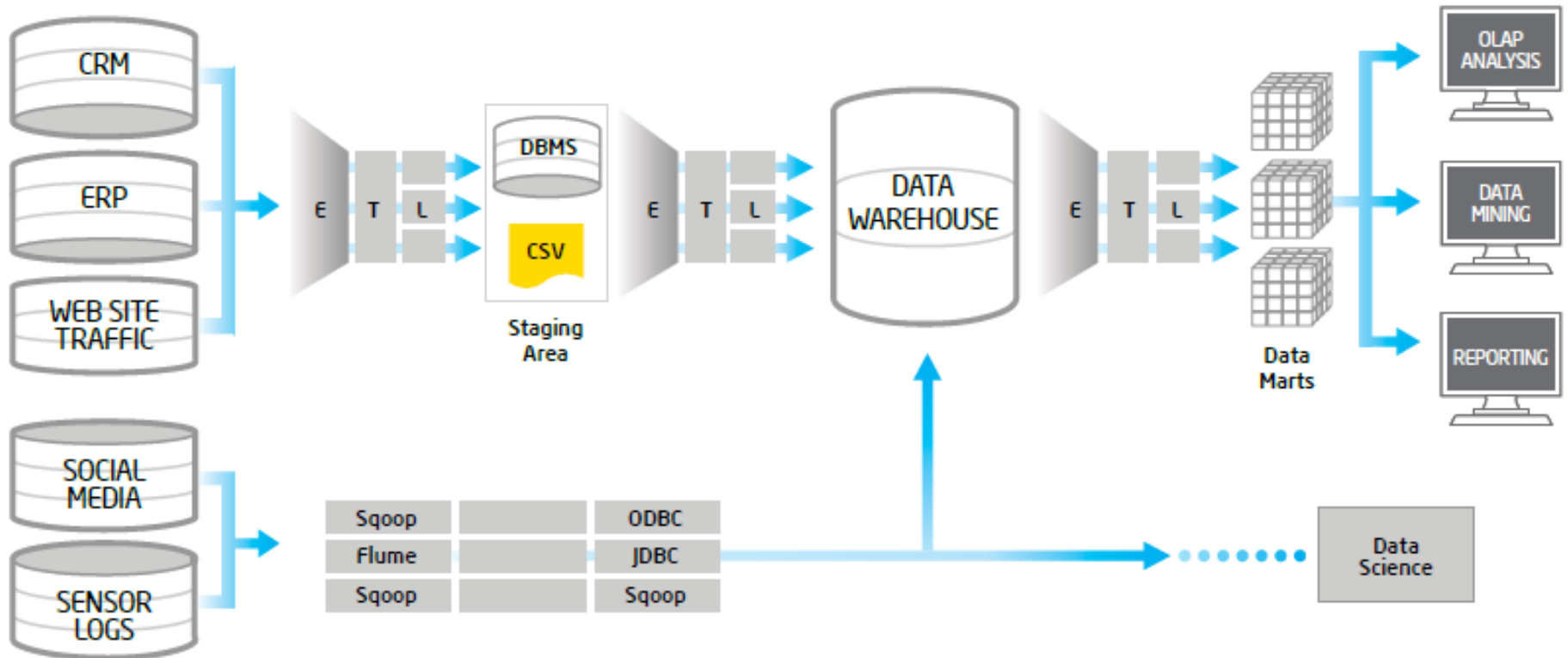
Big Data Solution



Traditional ETL Architecture



Offload ETL with Hadoop (Big Data Architecture)



Spark Ecosystem



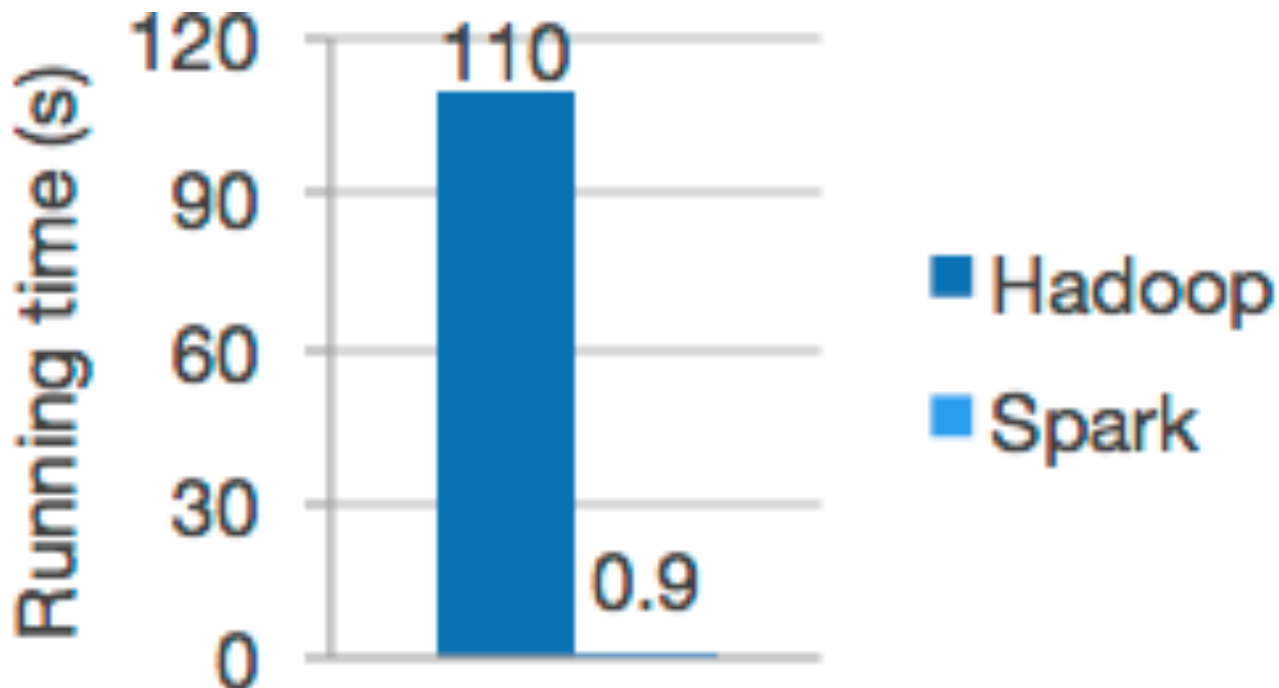
Lightning-fast cluster computing

Apache Spark

**is a fast and general engine
for**

large-scale data processing.

Logistic regression in Hadoop and Spark



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

Ease of Use

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



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)
```

Spark and Hadoop





Spark Ecosystem

Spark
SQL

Spark
Streaming

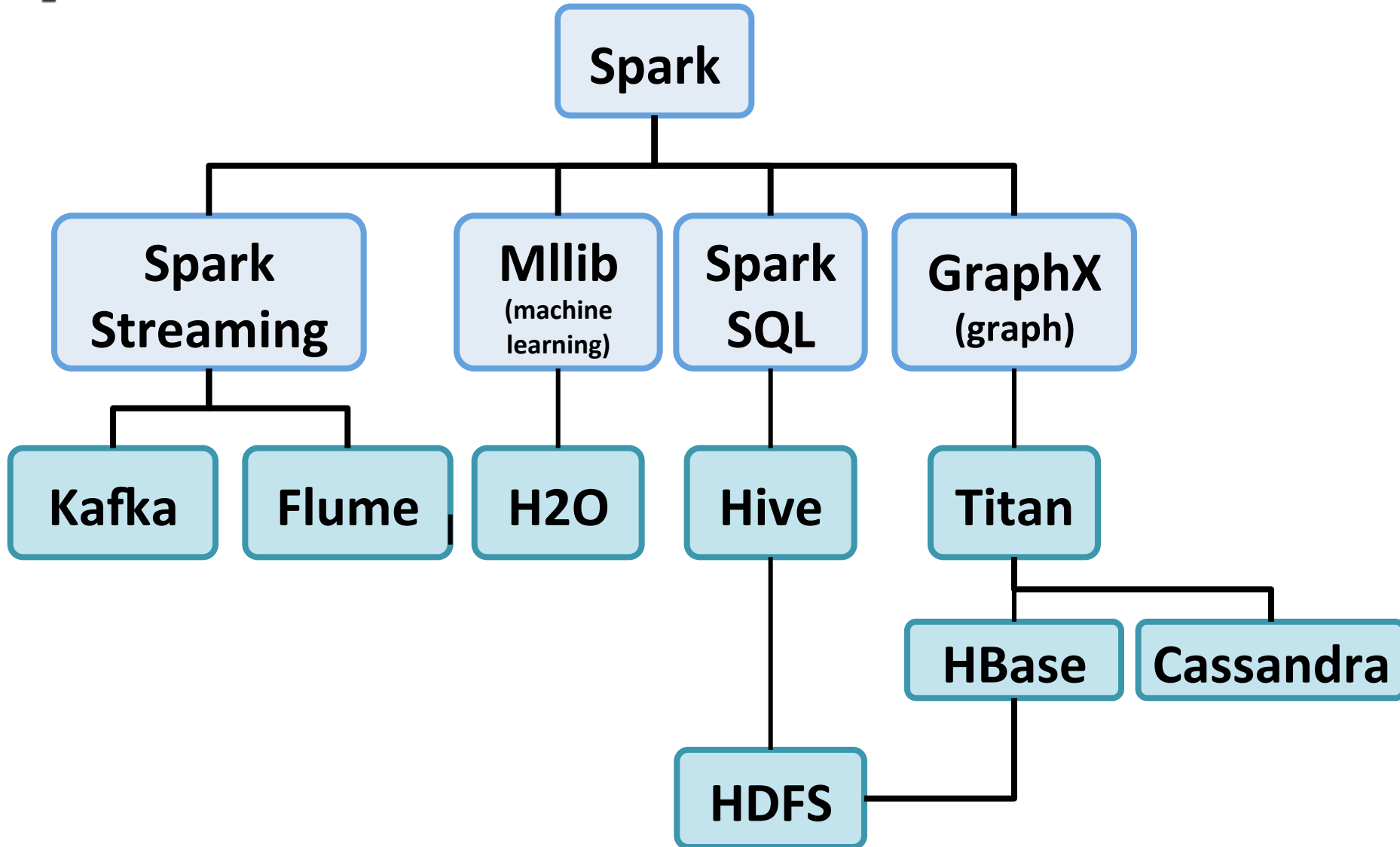
MLlib
(machine
learning)

GraphX
(graph)

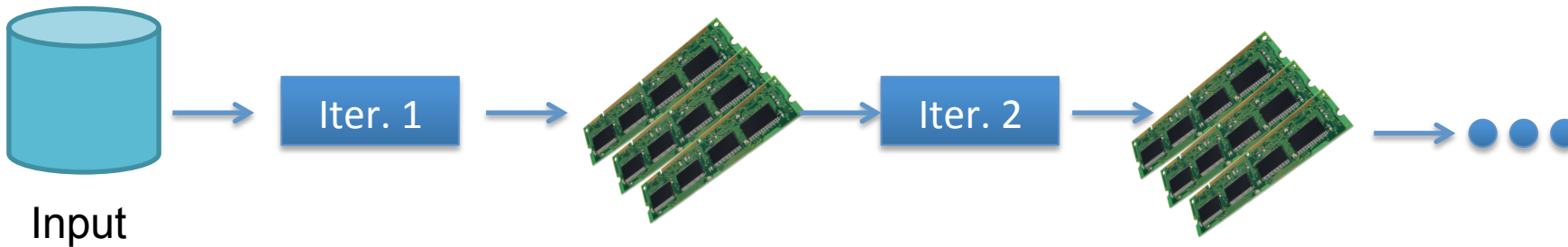
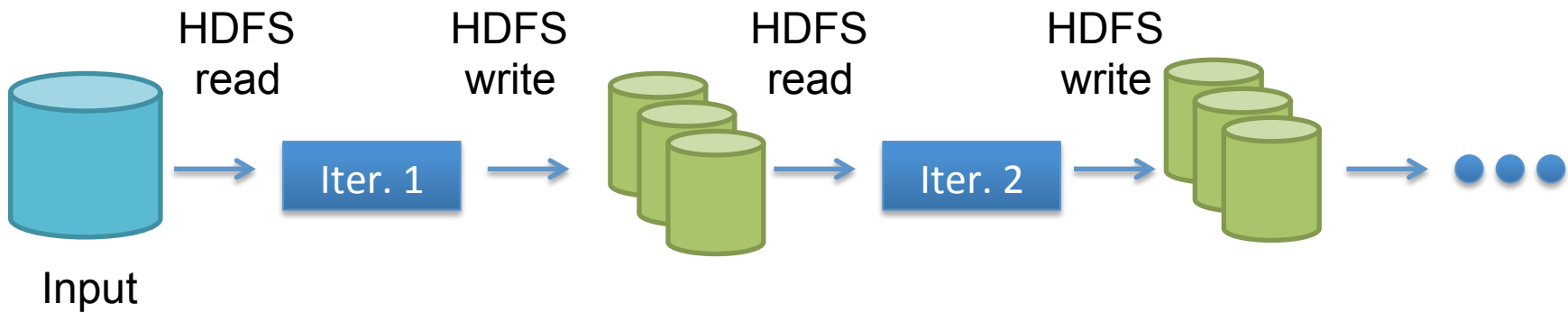
Apache Spark



Spark Ecosystem

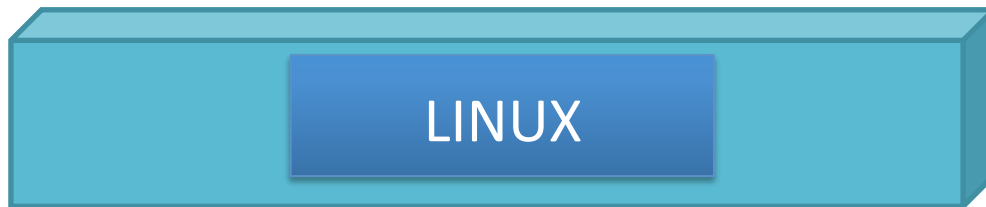
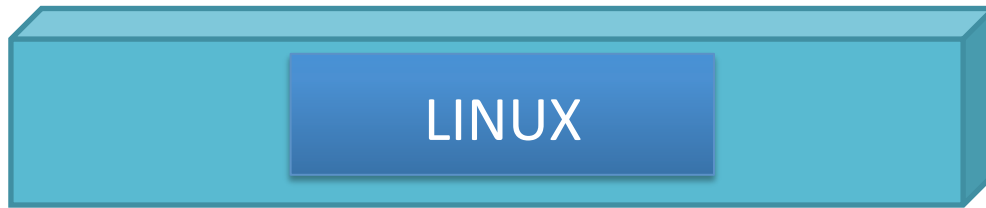


Hadoop vs. Spark



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

Hadoop: Linux Based Software



Appliance

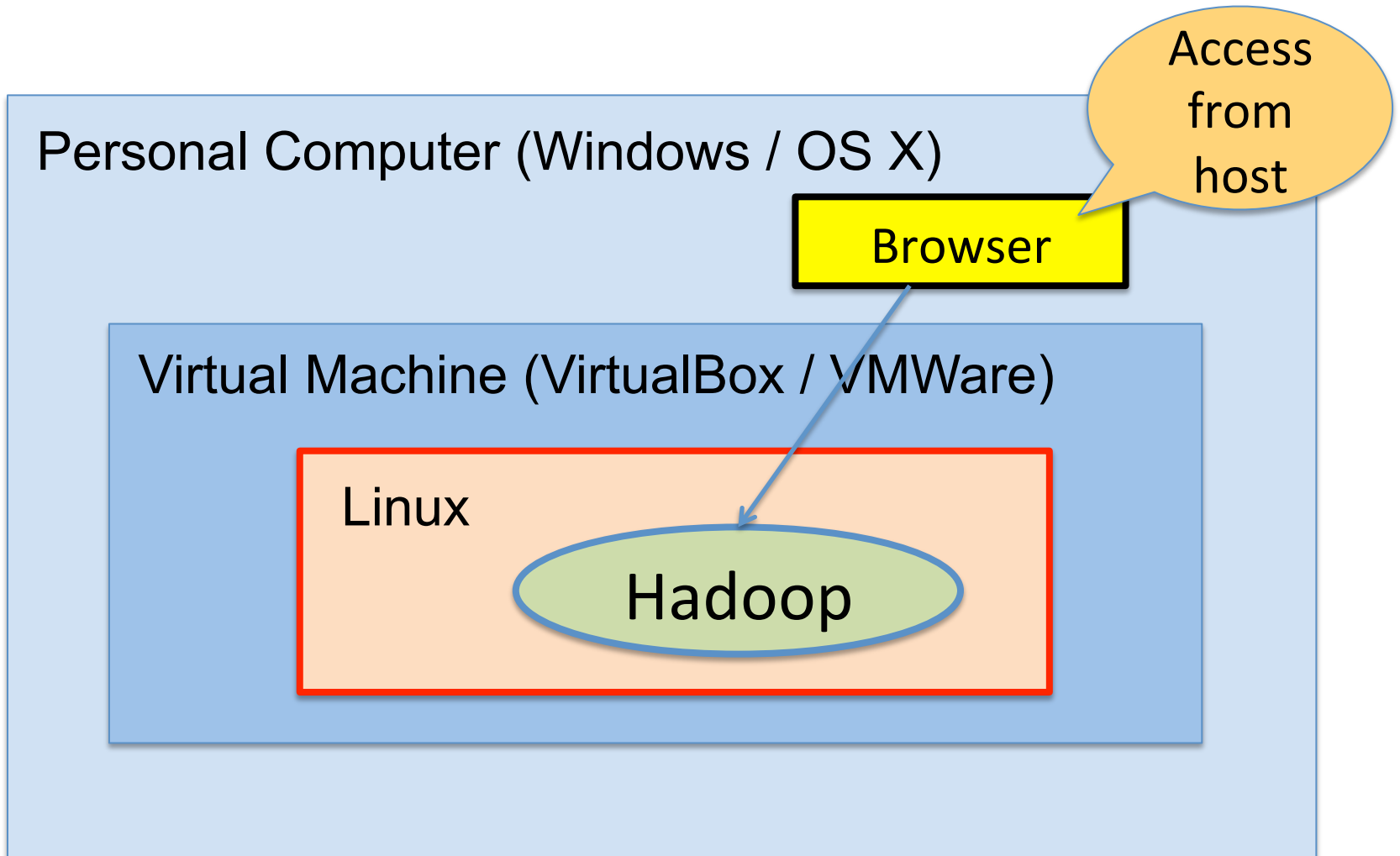
Personal Computer (Windows / OS X)

Virtual Machine (VirtualBox / VMWare)

Linux

Hadoop

Connection to Hadoop



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

Step 1. Download and Install VirtualBox



Step 2. Download Appliance



Step 3. Import Appliance



Step 4. Configure Virtual Machine (VM)



Step 5. Start Virtual Machine (VM)



Step 6. Test Connection From Host

Virtual Box

← → ↻ <https://www.virtualbox.org> ☆



VirtualBox

search...
[Login](#) [Preferences](#)

Welcome to VirtualBox.org!

VirtualBox is a powerful x86 and AMD64/Intel64 [virtualization](#) product for enterprise as well as home use. Not only is VirtualBox an extremely feature rich, high performance product for enterprise customers, it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2. See "[About VirtualBox](#)" for an introduction.

Presently, VirtualBox runs on Windows, Linux, Macintosh, and Solaris hosts and supports a large number of [guest operating systems](#) including but not limited to Windows (NT 4.0, 2000, XP, Server 2003, Vista, Windows 7, Windows 8, Windows 10), DOS/Windows 3.x, Linux (2.4, 2.6, 3.x and 4.x), Solaris and OpenSolaris, OS/2, and OpenBSD.

VirtualBox is being actively developed with frequent releases and has an ever growing list of features, supported guest operating systems and platforms it runs on. VirtualBox is a community effort backed by a dedicated company: everyone is encouraged to contribute while Oracle ensures the product always meets professional quality criteria.

Download
VirtualBox **5.0**

Hot picks:

- Pre-built virtual machines for developers at [Oracle Tech Network](#)
- **Hyperbox** Open-source Virtual Infrastructure Manager [project site](#)
- **phpVirtualBox** AJAX web interface [project site](#)
- **IQEmu** automated Windows VM creation, application integration <http://mirage335-site.member.hacdc.org:6380/wiki/Category:IQEmu>

News Flash

- **New January 19th, 2016 VirtualBox 5.0.14 released!**
Oracle today released a 5.0 maintenance release which improves stability and fixes regressions. See the [Changelog](#) for details.
- **New January 19th, 2016 VirtualBox 4.3.36 released!**
Oracle today released maintenance releases which improve stability and fixes regressions. See the [Changelog](#) for details.
- **New July 9th, 2015 VirtualBox 5.0 released!**
Read the official [press release](#) for details.
- **Important February, 2015 We're hiring!**
Looking for a new challenge? We're looking for [generic product developers \(Russia\)](#).

[More information...](#)

ORACLE

<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)

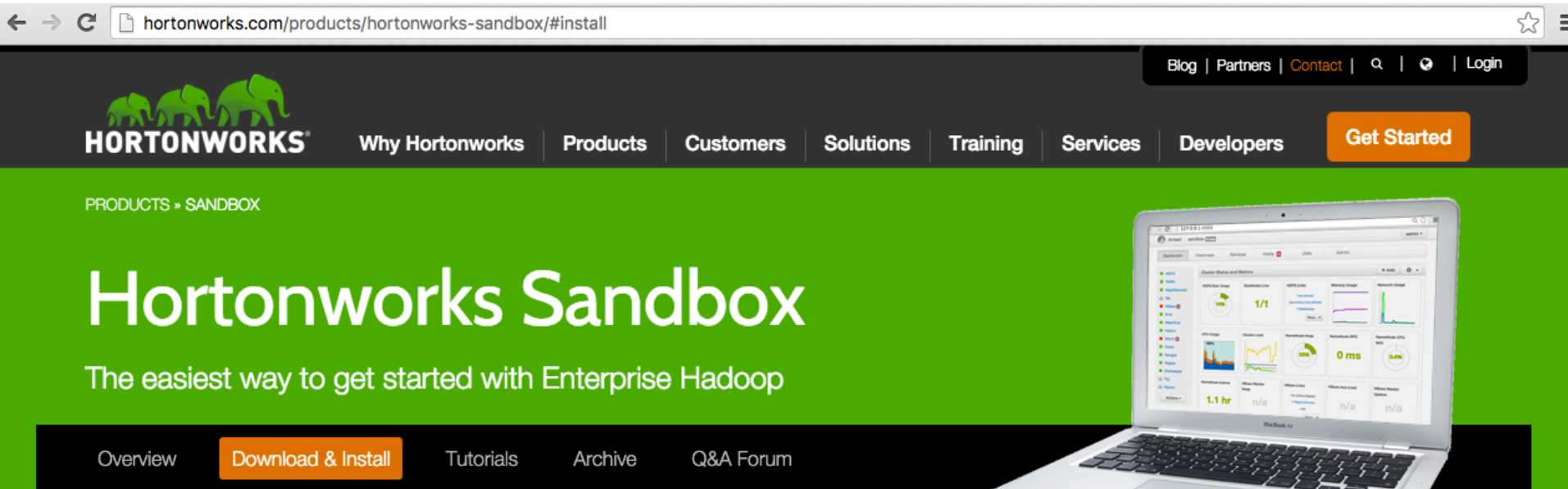
Step 5. Start Virtual Machine (VM)

Step 6. Test Connection From Host



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

No data center, no cloud service and no internet connection needed! Full control of the environment. Easily extend with additional components or try the various Hortonworks technical previews. Always updated with latest edition.

HDP 2.4 on Hortonworks Sandbox

Runs on VirtualBox or VMware



for VirtualBox
[Mac & Windows](#)

for VirtualBox
(HDP 2.4 - 9.91 GB)

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

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



Tutorials

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

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:

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 Lap Around Apache Spark](#)
If you have any errors in completing this tutorial. Please ask questions or notify us on [Hortonworks Community Connection!](#) Introduction This...

Apache Hadoop

Apache > Hadoop >



Search with Apache Solr Search

Last Published: 02/13/2016 15:31:55

Top Wiki

About

Welcome

- What Is Apache Hado...
- Getting Started ...
- Download Hadoop
- Who Uses Hadoop?...
- News
- Releases
- Mailing Lists
- Issue Tracking
- Who We Are?
- Who Uses Hadoop?
- Buy Stuff
- Sponsorship
- Thanks
- Privacy Policy
- Bylaws
- License
- Documentation
- Related Projects



Welcome to Apache™ Hadoop®!

What Is Apache Hadoop?

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

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

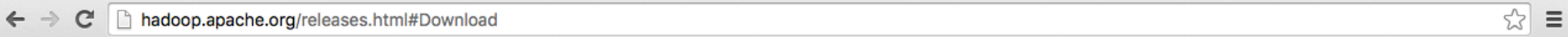
Other Hadoop-related projects at Apache include:

- **Ambari™:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™:** A data serialization system.
- **Cassandra™:** A scalable multi-master database with no single points of failure.
- **Chukwa™:** A data collection system for managing large distributed systems.
- **HBase™:** A scalable, distributed database that supports structured data storage for large tables.
- **Hive™:** A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™:** A Scalable machine learning and data mining library.
- **Pig™:** A high-level data-flow language and execution framework for parallel computation.

<http://hadoop.apache.org/>

Apache Hadoop

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



Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-256.

Version	Release Date	Tarball	GPG	SHA-256
2.6.4	11 February, 2016	source	signature	F755D961 18316335..
		binary	signature	C58F08D2 E0B13035..
2.7.2	25 January, 2016	source	signature	7D48E61B 5464A765..
		binary	signature	49AD740F 85D27FA3..
2.6.3	17 Dec, 2015	source	signature	FA0C71B5 CB33A7FD..
		binary	signature	ADA83D8C 2FF72D46..
2.6.2	28 Oct, 2015	source	signature	6996A4A8 0FCE9109..
		binary	signature	56F630D7 0D4C7850..
2.7.1	06 July, 2015	source	signature	53F3001C 457AD8FD..
		binary	signature	991DC34E A42A80B2..
2.5.2	19 Nov, 2014	source	signature	139EF872 09C5637E..
		binary	signature	0BDB4850 A3825208..

To verify Hadoop releases using GPG:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the signature file `hadoop-X.Y.Z-src.tar.gz.asc` from [Apache](#).
3. Download the [Hadoop KEYS](#) file.
4. `gpg --import KEYS`
5. `gpg --verify hadoop-X.Y.Z-src.tar.gz.asc`

To perform a quick check using SHA-256:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the checksum `hadoop-X.Y.Z-src.tar.gz.mds` from [Apache](#).
3. `shasum -a 256 hadoop-X.Y.Z-src.tar.gz`

All previous releases of Hadoop are available from the [Apache release archive](#) site.

Many third parties distribute products that include Apache Hadoop and related tools. Some of these are listed on the [Distributions wiki page](#).

Release Notes

11 February, 2016: Release 2.6.4 available

Apache Hadoop 2.6.4 is a point release in the 2.6.x release line, and fixes a few critical issues in 2.6.3.

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

Apache Hadoop 2.7.2

hadoop.apache.org/docs/r2.7.2/



Apache > Hadoop

Wiki | git | Last Published: 2016-01-26 | Version: 2.7.2

General

Overview

- Single Node Setup
- Cluster Setup
- Hadoop Commands
- Reference
- FileSystem Shell
- Hadoop Compatibility
- Interface Classification
- FileSystem Specification

Common

- CLI Mini Cluster
- Native Libraries
- Proxy User
- Rack Awareness
- Secure Mode
- Service Level
- Authorization
- HTTP Authentication
- Hadoop KMS
- Tracing

HDFS

- HDFS User Guide
- HDFS Commands
- Reference
- High Availability With QJM
- High Availability With NFS Federation
- ViewFs Guide
- HDFS Snapshots
- HDFS Architecture
- Edits Viewer
- Image Viewer
- Permissions and HDFS Quotas and HDFS HFTP
- C API libhdfs
- WebHDFS REST API

Apache Hadoop 2.7.2

Apache Hadoop 2.7.2 is a minor release in the 2.x.y release line, building upon the previous stable release 2.7.1.

Here is a short overview of the major features and improvements.

- Common
 - Authentication improvements when using an HTTP proxy server. This is useful when accessing WebHDFS via a proxy server.
 - A new Hadoop metrics sink that allows writing directly to Graphite.
 - [Specification work](#) related to the Hadoop Compatible Filesystem (HCFS) effort.
- HDFS
 - Support for POSIX-style filesystem extended attributes. See the [user documentation](#) for more details.
 - Using the [OfflineImageViewer](#), clients can now browse an fsimage via the WebHDFS API.
 - The NFS gateway received a number of supportability improvements and bug fixes. The Hadoop portmapper is no longer required to run the gateway, and the gateway is now able to reject connections from unprivileged ports.
 - The SecondaryNameNode, JournalNode, and DataNode web UIs have been modernized with HTML5 and Javascript.
- YARN
 - YARN's REST APIs now support write/modify operations. Users can submit and kill applications through REST APIs.
 - The timeline store in YARN, used for storing generic and application-specific information for applications, supports authentication through Kerberos.
 - The Fair Scheduler supports dynamic hierarchical user queues, user queues are created dynamically at runtime under any specified parent-queue.

Getting Started

The Hadoop documentation includes the information you need to get started using Hadoop. Begin with the [Single Node Setup](#) which shows you how to set up a single-node Hadoop installation. Then move on to the [Cluster Setup](#) to learn how to set up a multi-node Hadoop installation.

Hadoop: Setting up a Single Node Cluster



General

- Overview
- Single Node Setup
- Cluster Setup
- Hadoop Commands
- Reference
- FileSystem Shell
- Hadoop Compatibility
- Interface Classification
- FileSystem Specification

Common

- CLI Mini Cluster
- Native Libraries
- Proxy User
- Rack Awareness
- Secure Mode
- Service Level
- Authorization
- HTTP Authentication
- Hadoop KMS
- Tracing

HDFS

- HDFS User Guide
- HDFS Commands
- Reference
- High Availability With QJM
- High Availability With NFS Federation
- ViewFs Guide
- HDFS Snapshots
- HDFS Architecture
- Edits Viewer
- Image Viewer
- Permissions and HDFS Quotas and HDFS HFTP
- C API libhdfs
- WebHDFS REST API
- HttpFS Gateway

Hadoop: Setting up a Single Node Cluster.

- Hadoop: Setting up a Single Node Cluster.
 - Purpose
 - Prerequisites
 - Supported Platforms
 - Required Software
 - Installing Software
 - Download
 - Prepare to Start the Hadoop Cluster
 - Standalone Operation
 - Pseudo-Distributed Operation
 - Configuration
 - Setup passphraseless ssh
 - Execution
 - YARN on a Single Node
 - Fully-Distributed Operation

Purpose

This document describes how to set up and configure a single-node Hadoop installation so that you can quickly perform simple operations using Hadoop MapReduce and the Hadoop Distributed File System (HDFS).

Prerequisites

Supported Platforms

- GNU/Linux is supported as a development and production platform. Hadoop has been demonstrated on GNU/Linux clusters with 2000 nodes.
- Windows is also a supported platform but the followings steps are for Linux only. To set up Hadoop on Windows, see [wiki page](#).

Hadoop Cluster Setup

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



Apache > Hadoop > Apache Hadoop Project Dist POM > Apache Hadoop 2.7.2

Wiki | git | Apache Hadoop | Last Published: 2016-01-26 | Version: 2.7.2

General

- Overview
- Single Node Setup
- Cluster Setup
- Hadoop Commands
- Reference
- FileSystem Shell
- Hadoop Compatibility
- Interface Classification
- FileSystem Specification

Common

- CLI Mini Cluster
- Native Libraries
- Proxy User
- Rack Awareness
- Secure Mode
- Service Level
- Authorization
- HTTP Authentication
- Hadoop KMS
- Tracing

HDFS

- HDFS User Guide
- HDFS Commands
- Reference
- High Availability With QJM
- High Availability With NFS Federation
- ViewFs Guide
- HDFS Snapshots
- HDFS Architecture
- Edits Viewer
- Image Viewer
- Permissions and HDFS
- Quotas and HDFS
- HFTP
- C API libhdfs
- WebHDFS REST API
- HttpFS Gateway
- Short Circuit Local Reads

- **Hadoop Cluster Setup**
 - Purpose
 - Prerequisites
 - Installation
 - Configuring Hadoop in Non-Secure Mode
 - Configuring Environment of Hadoop Daemons
 - Configuring the Hadoop Daemons
 - Monitoring Health of NodeManagers
 - Slaves File
 - Hadoop Rack Awareness
 - Logging
 - Operating the Hadoop Cluster
 - Hadoop Startup
 - Hadoop Shutdown
 - Web Interfaces

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.

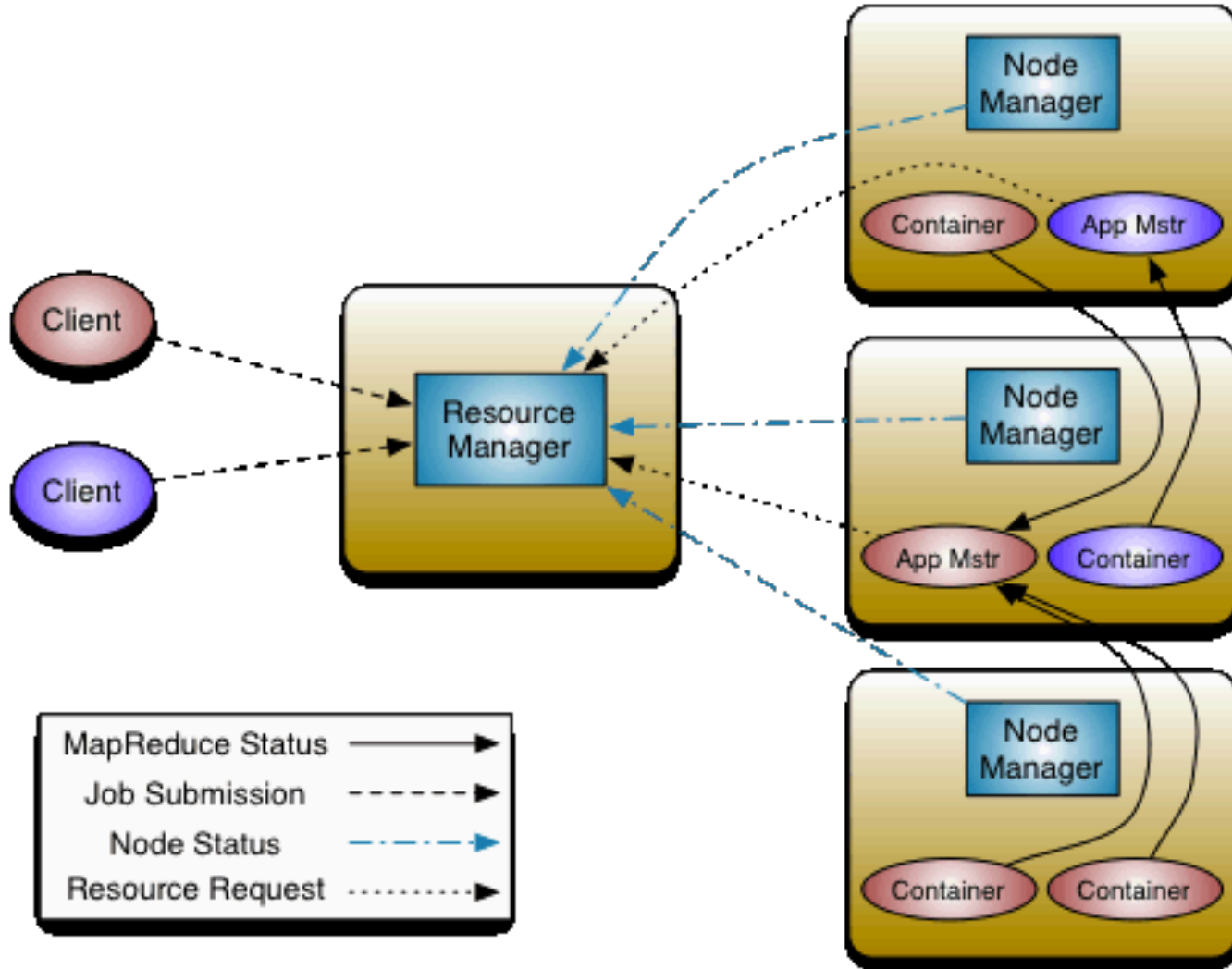
Prerequisites

- 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

Apache Hadoop YARN



Apache Spark

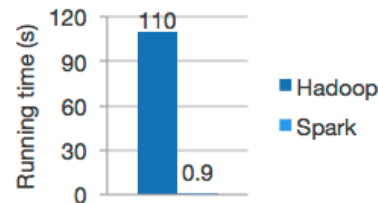


Apache Spark™ is a fast and general engine for large-scale data processing.

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](#)

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](http://www.slideshare.net/deepakramanathan/sas-modernization-architectures-big-data-analytics)