

Social Computing and Big Data Analytics

社群運算與大數據分析

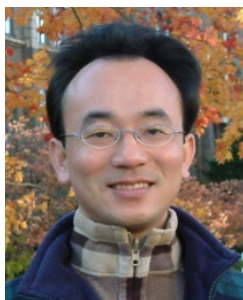
巨量資料基礎：

MapReduce典範、Hadoop與Spark生態系統 (Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem)

1052SCBDA03

MIS MBA (M2226) (8606)

Wed, 8,9, (15:10-17:00) (B505)



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/>

2017-03-01



課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2017/02/15	Course Orientation for Social Computing and Big Data Analytics (社群運算與大數據分析課程介紹)
2	2017/02/22	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	2017/03/01	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	2017/03/08	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	2017/03/15	Big Data Analytics with Numpy in Python (Python Numpy 大數據分析)
6	2017/03/22	Finance Big Data Analytics with Pandas in Python (Python Pandas 財務大數據分析)
7	2017/03/29	Text Mining Techniques and Natural Language Processing (文字探勘分析技術與自然語言處理)
8	2017/04/05	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	2017/04/12	Social Media Marketing Analytics (社群媒體行銷分析)
10	2017/04/19	期中報告 (Midterm Project Report)
11	2017/04/26	Deep Learning with Theano and Keras in Python (Python Theano 和 Keras 深度學習)
12	2017/05/03	Deep Learning with Google TensorFlow (Google TensorFlow 深度學習)
13	2017/05/10	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)

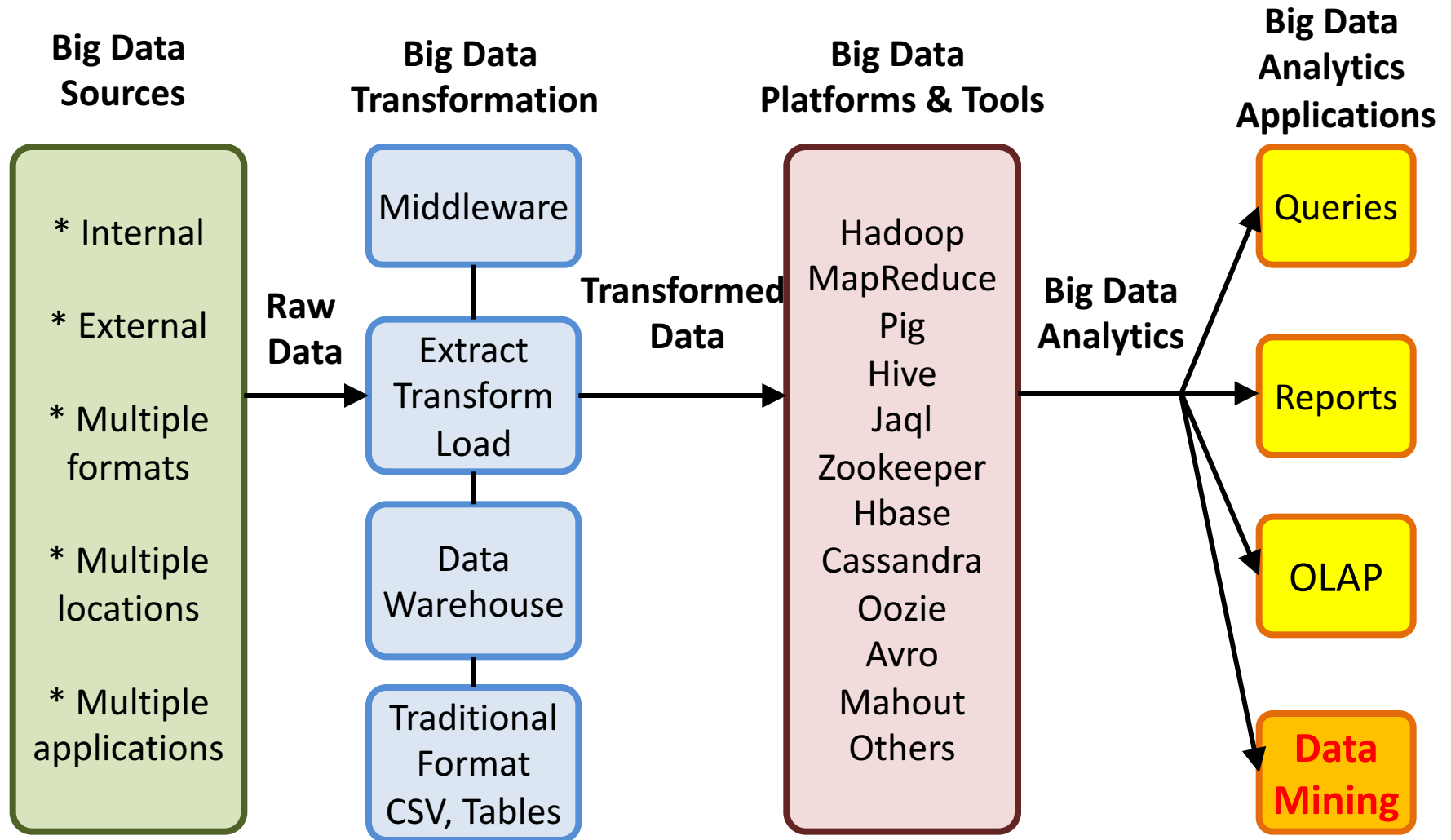
週次 (Week)	日期 (Date)	內容 (Subject/Topics)
14	2017/05/17	Social Network Analysis (社會網絡分析)
15	2017/05/24	Measurements of Social Network (社會網絡量測)
16	2017/05/31	Tools of Social Network Analysis (社會網絡分析工具)
17	2017/06/07	Final Project Presentation I (期末報告 I)
18	2017/06/14	Final Project Presentation II (期末報告 II)

2017/03/01

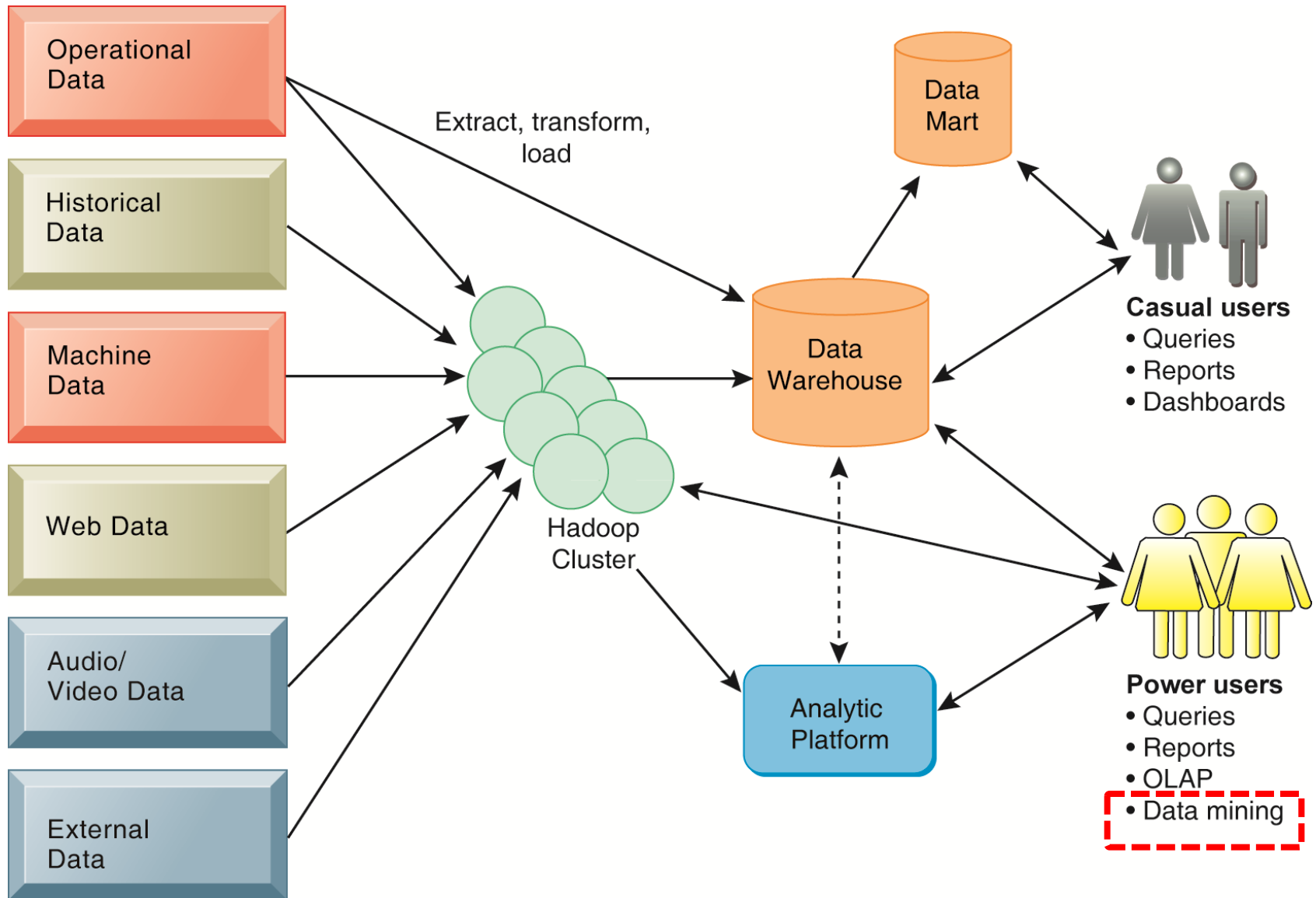
巨量資料基礎：

**MapReduce典範、
Hadoop與Spark生態系統
(Fundamental Big Data:
MapReduce Paradigm,
Hadoop and Spark Ecosystem)**

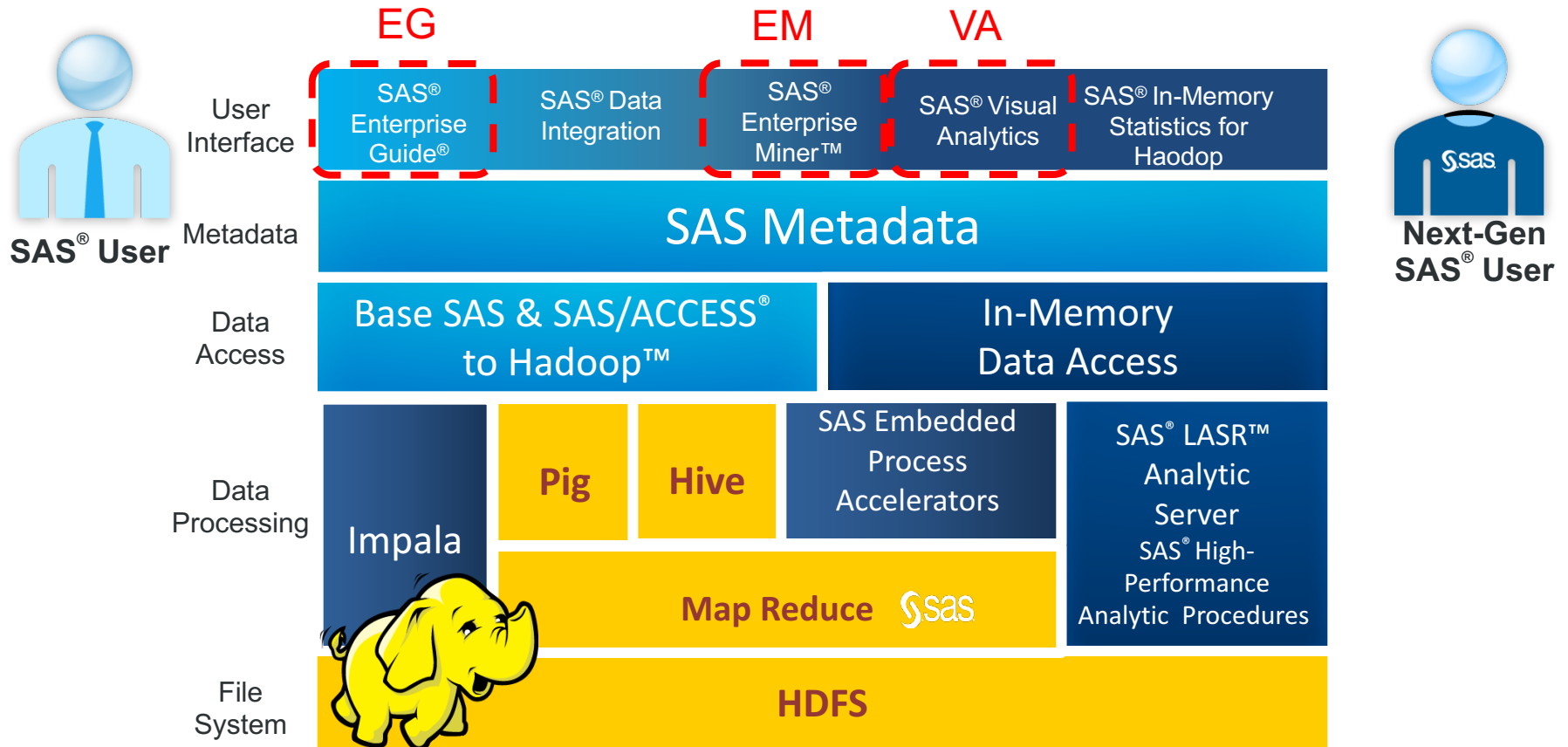
Architecture of Big Data Analytics



Business Intelligence (BI) Infrastructure



SAS® Within the HADOOP ECOSYSTEM



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
 R^2I^2

DETECT

Anomalies
Communities
Typologies

PREDICT

Trending
Real-time
Prediction

DECIDE

Simulation
Optimization



BIG DATA – Batch



BIG DATA – Real Time

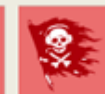


Complex by nature



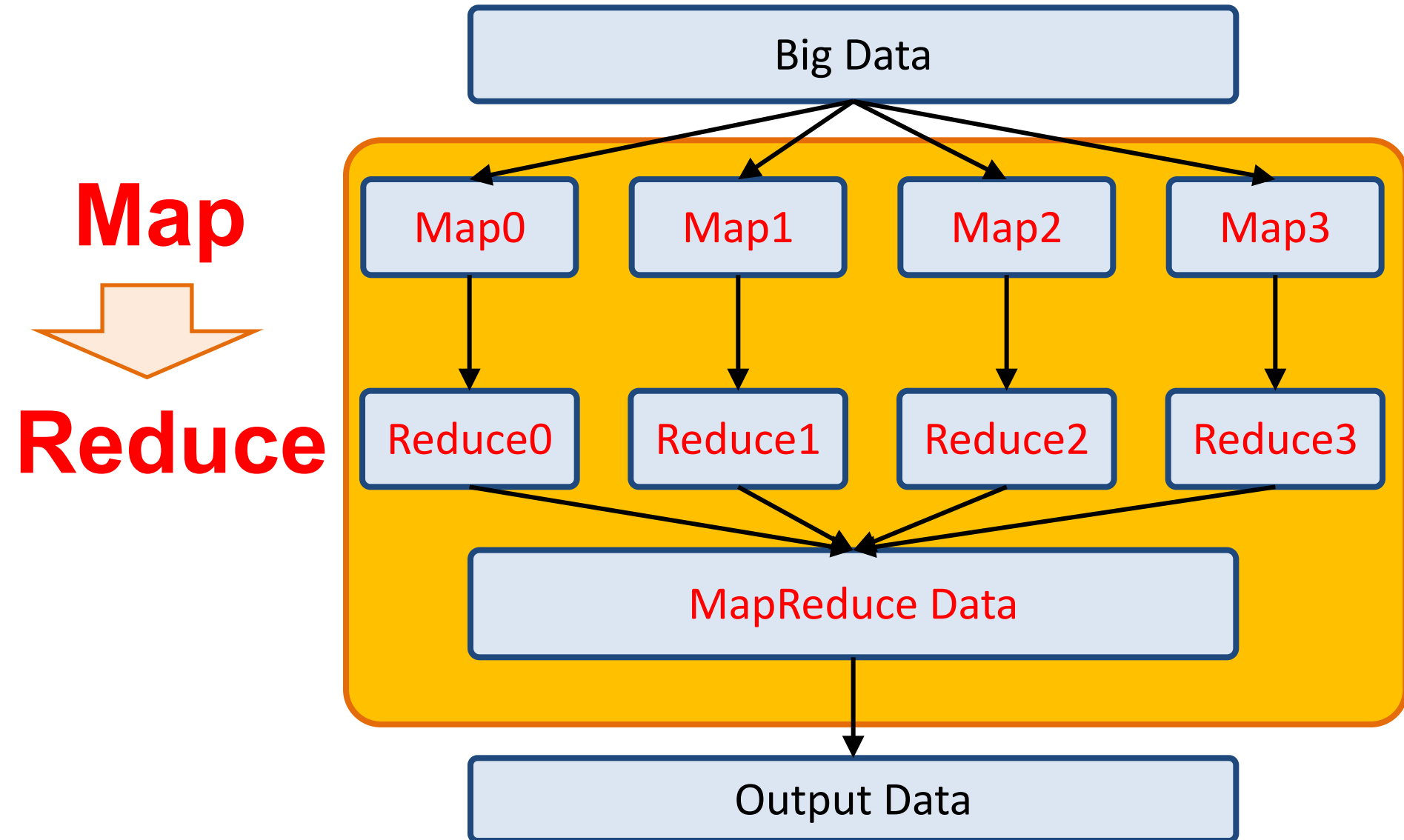
DATA

Complex by structure



MapReduce Paradigm

MapReduce Paradigm



MapReduce Word Count

Input

Dog Love Cat
Bird Love Bird
Dog Bird Cat

MapReduce Word Count

Input

Output



Dog Love Cat
Bird Love Bird
Dog Bird Cat

Bird, 3
Cat, 2
Dog, 2
Love, 2

MapReduce Word Count

Input

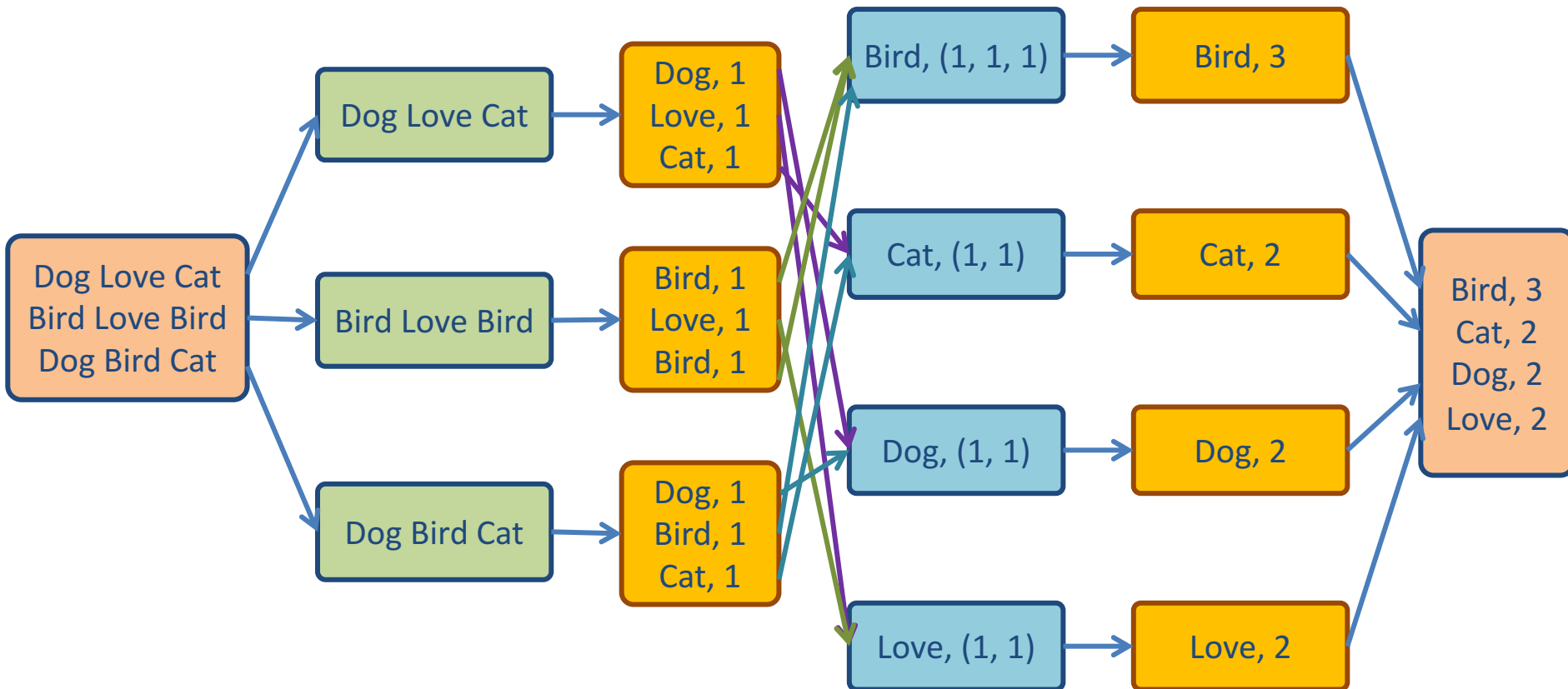
Split

Map

Shuffle

Reduce

Output



Hadoop Ecosystem



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



MapReduce

Processing



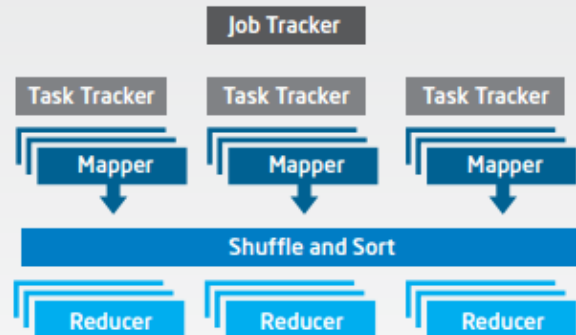
HDFS

Storage

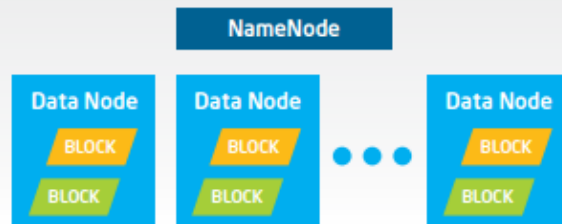
Big Data with Hadoop Architecture

LOGICAL ARCHITECTURE

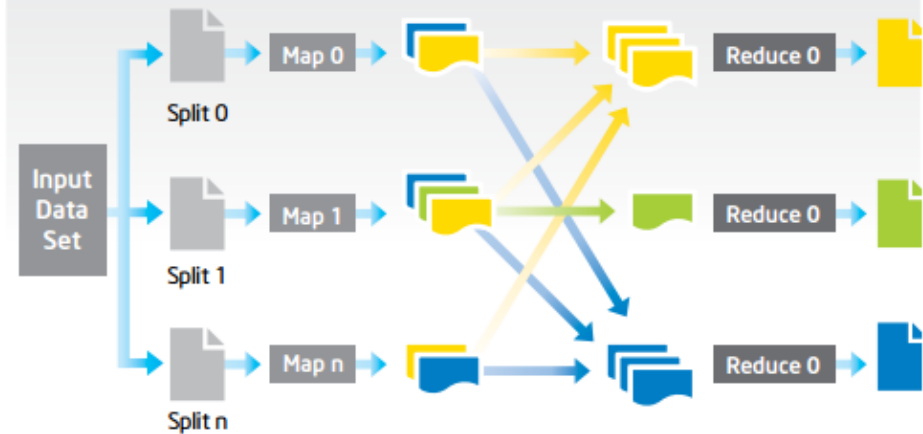
Processing: MapReduce



Storage: HDFS

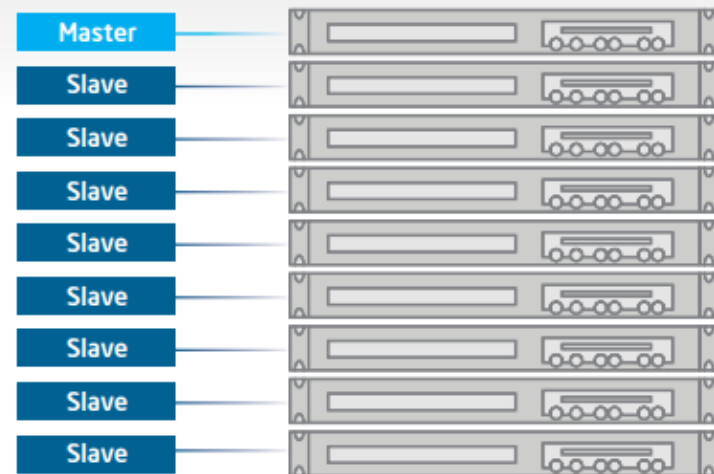


PROCESS FLOW



PHYSICAL ARCHITECTURE

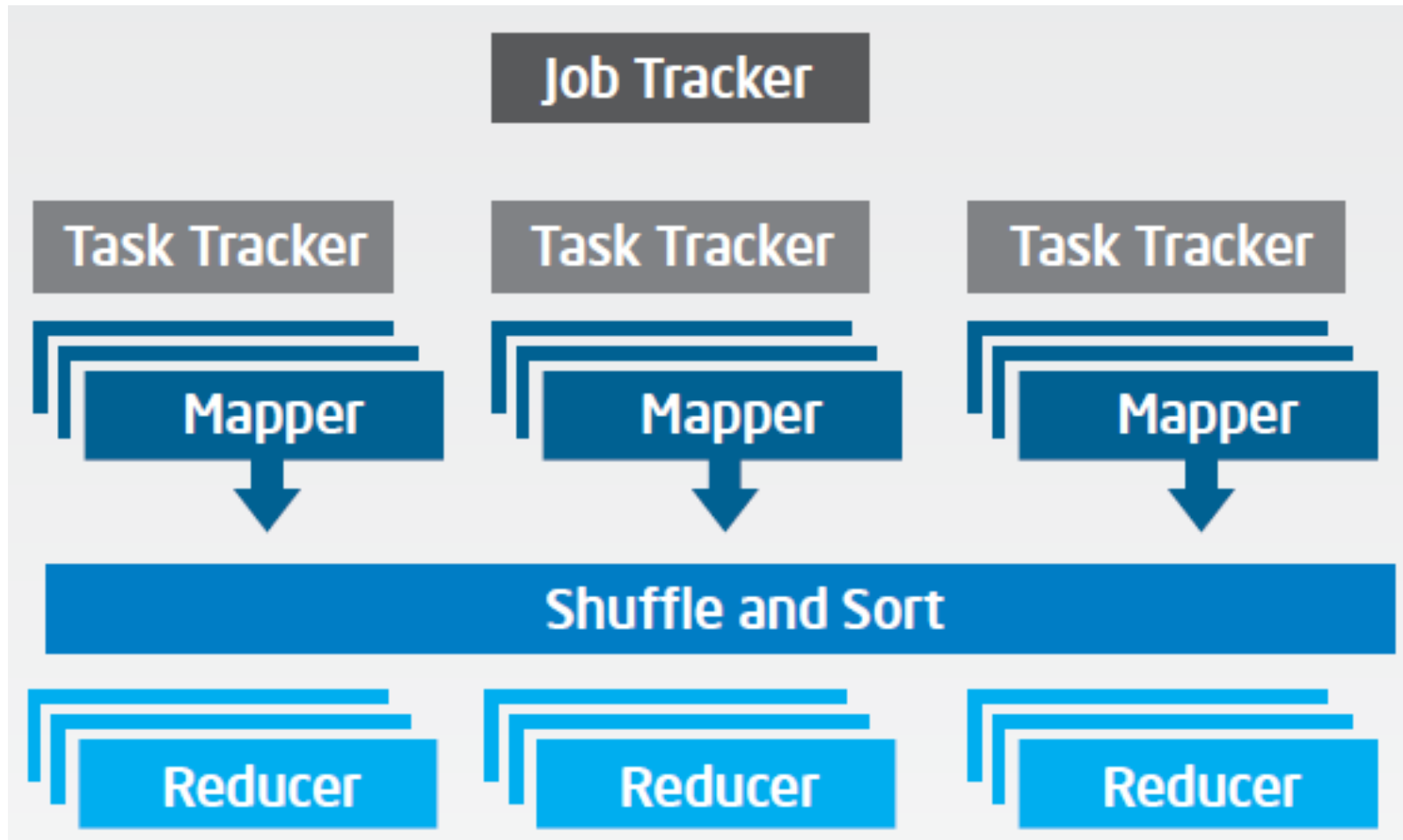
Hadoop Cluster



Big Data with Hadoop Architecture

Logical Architecture

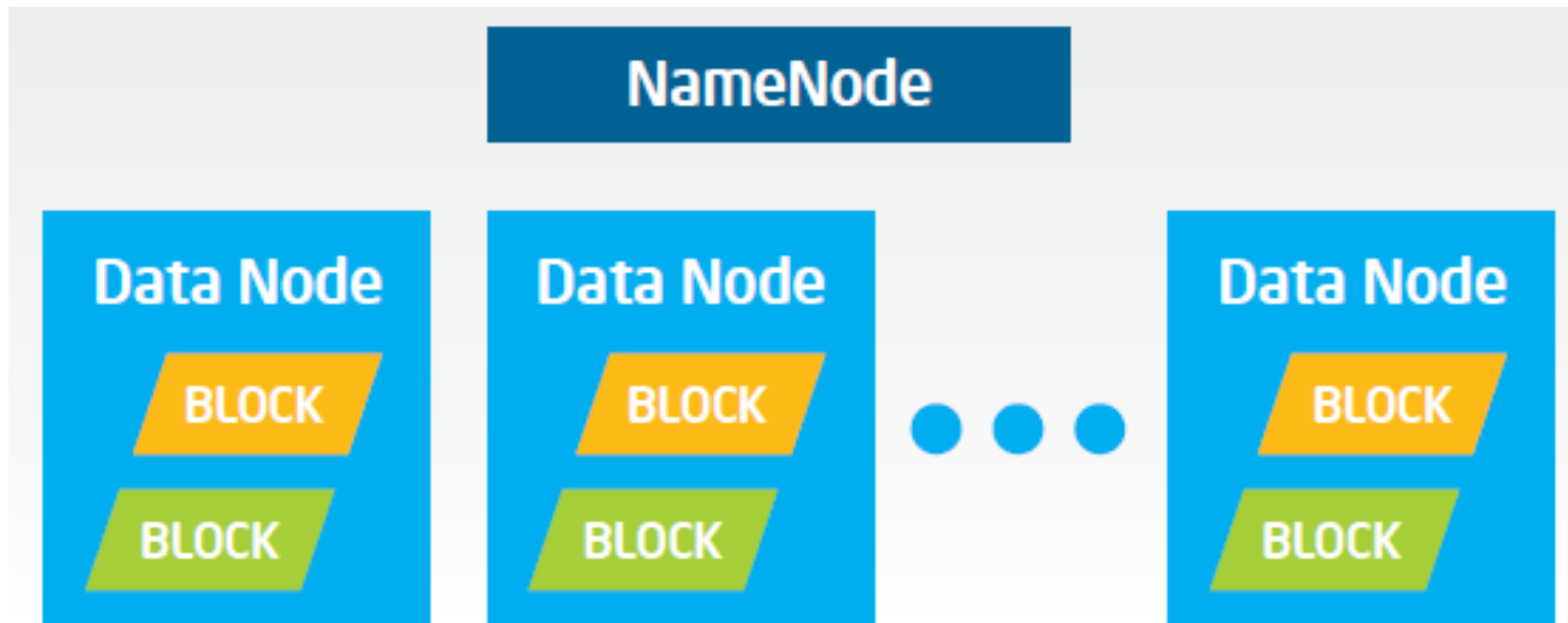
Processing: MapReduce



Big Data with Hadoop Architecture

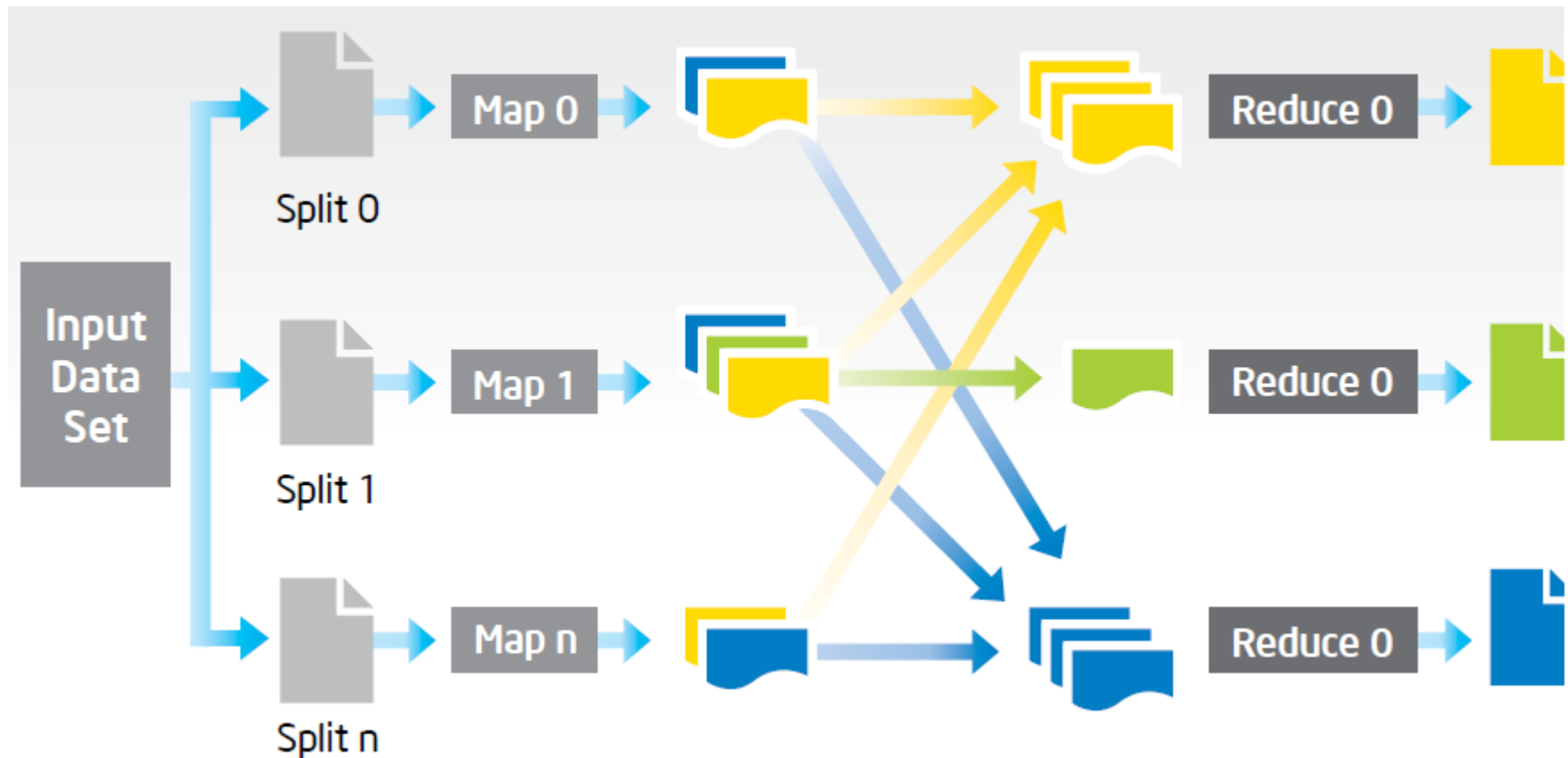
Logical Architecture

Storage: HDFS



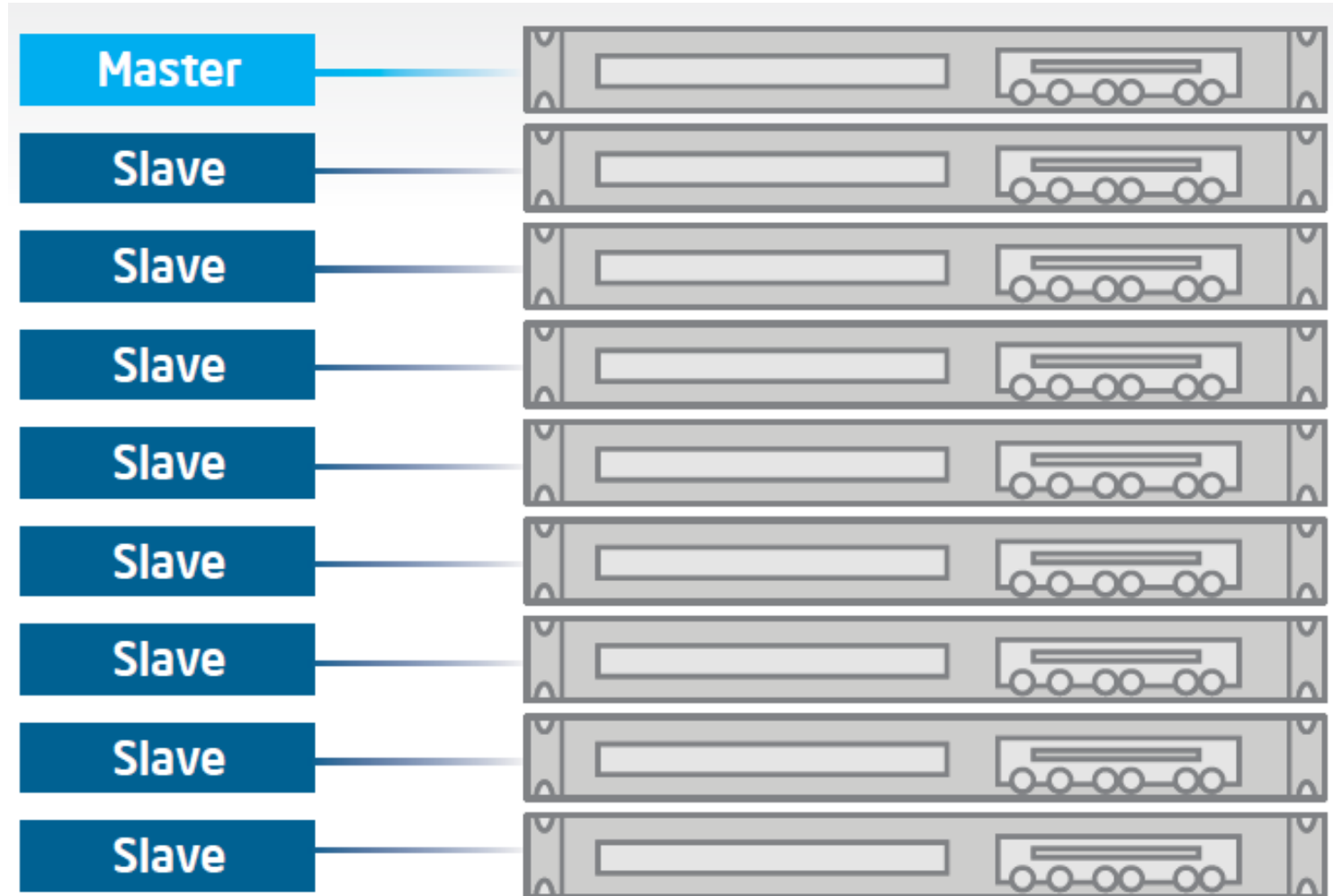
Big Data with Hadoop Architecture

Process Flow

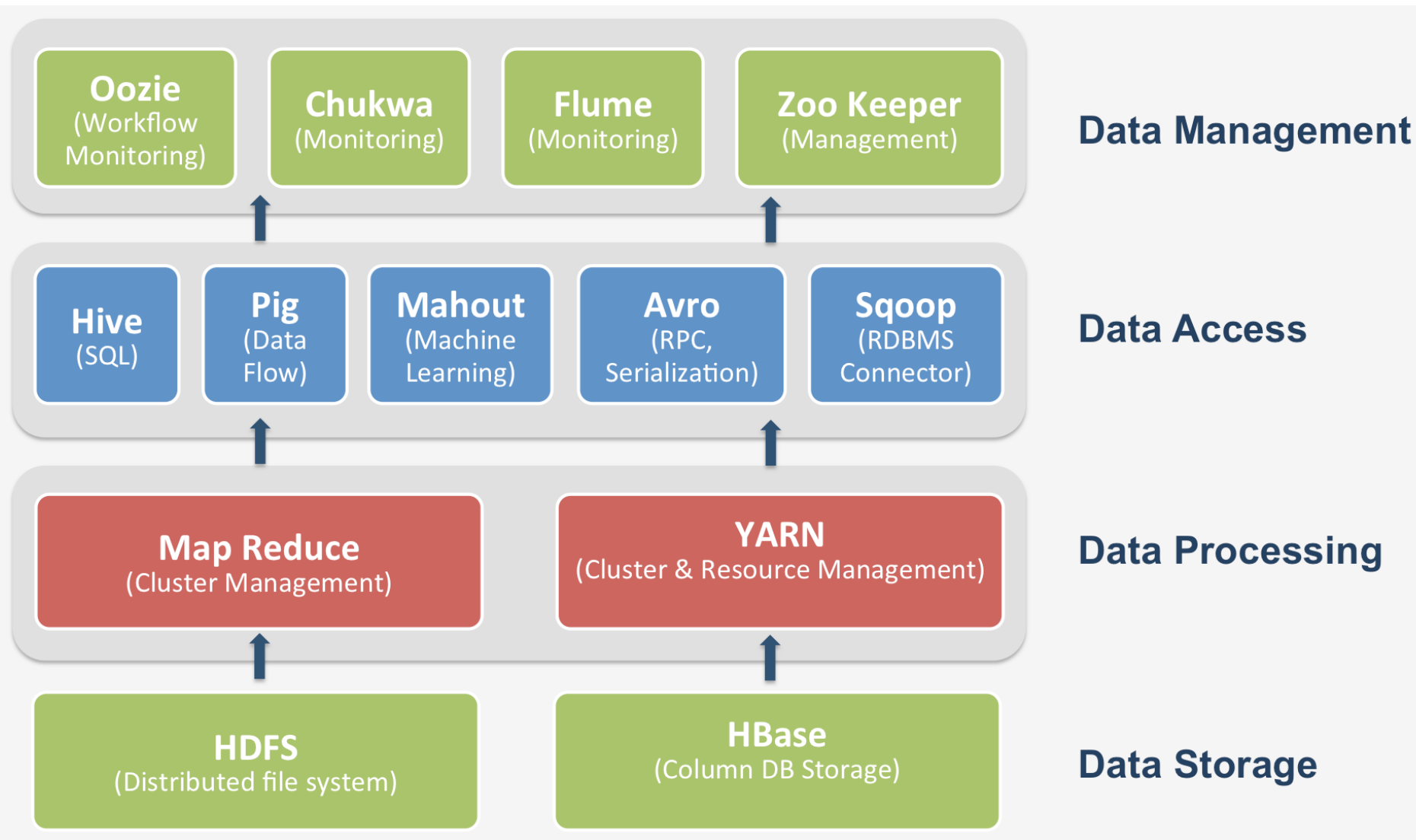


Big Data with Hadoop Architecture

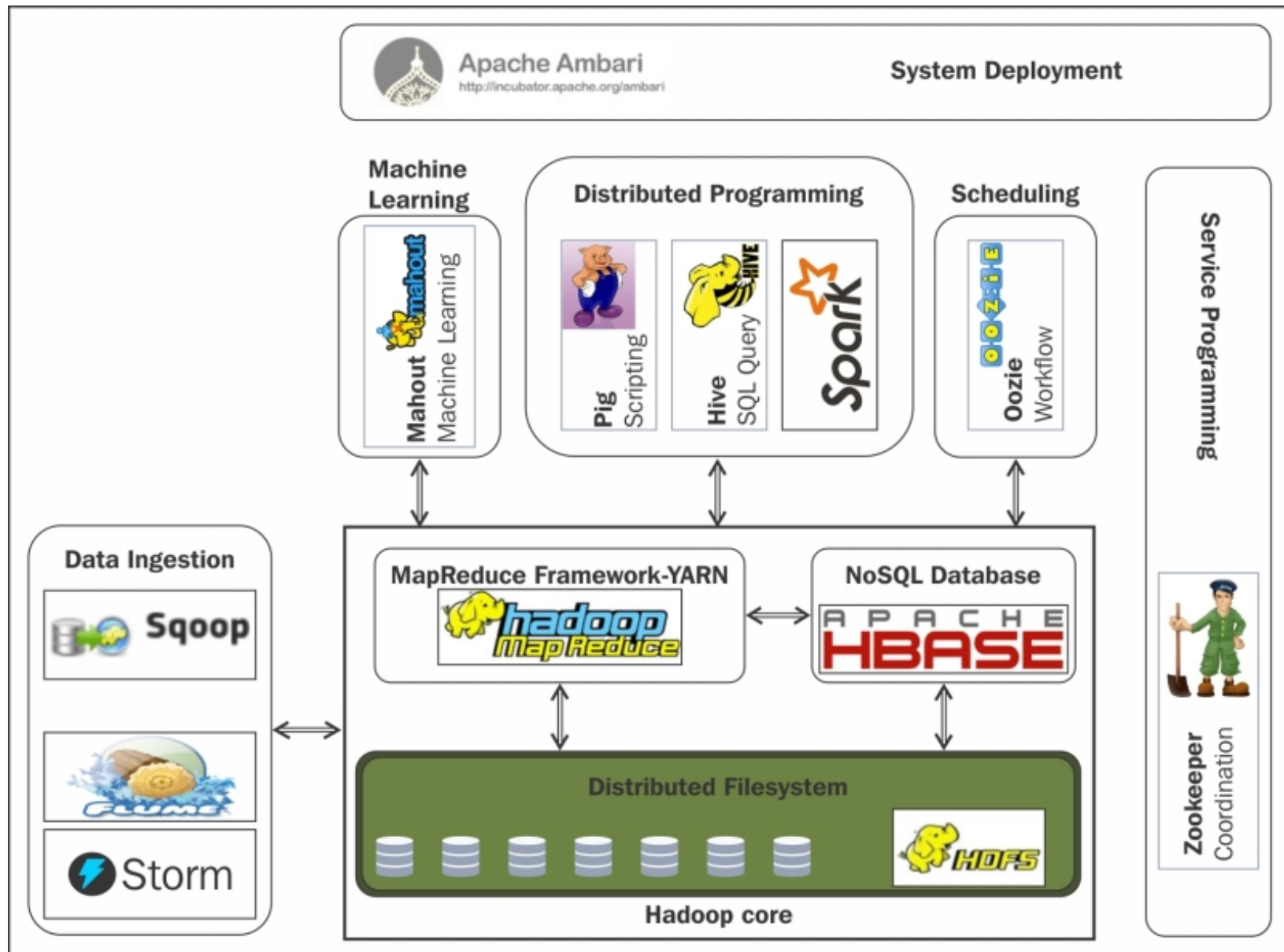
Hadoop Cluster



Hadoop Ecosystem

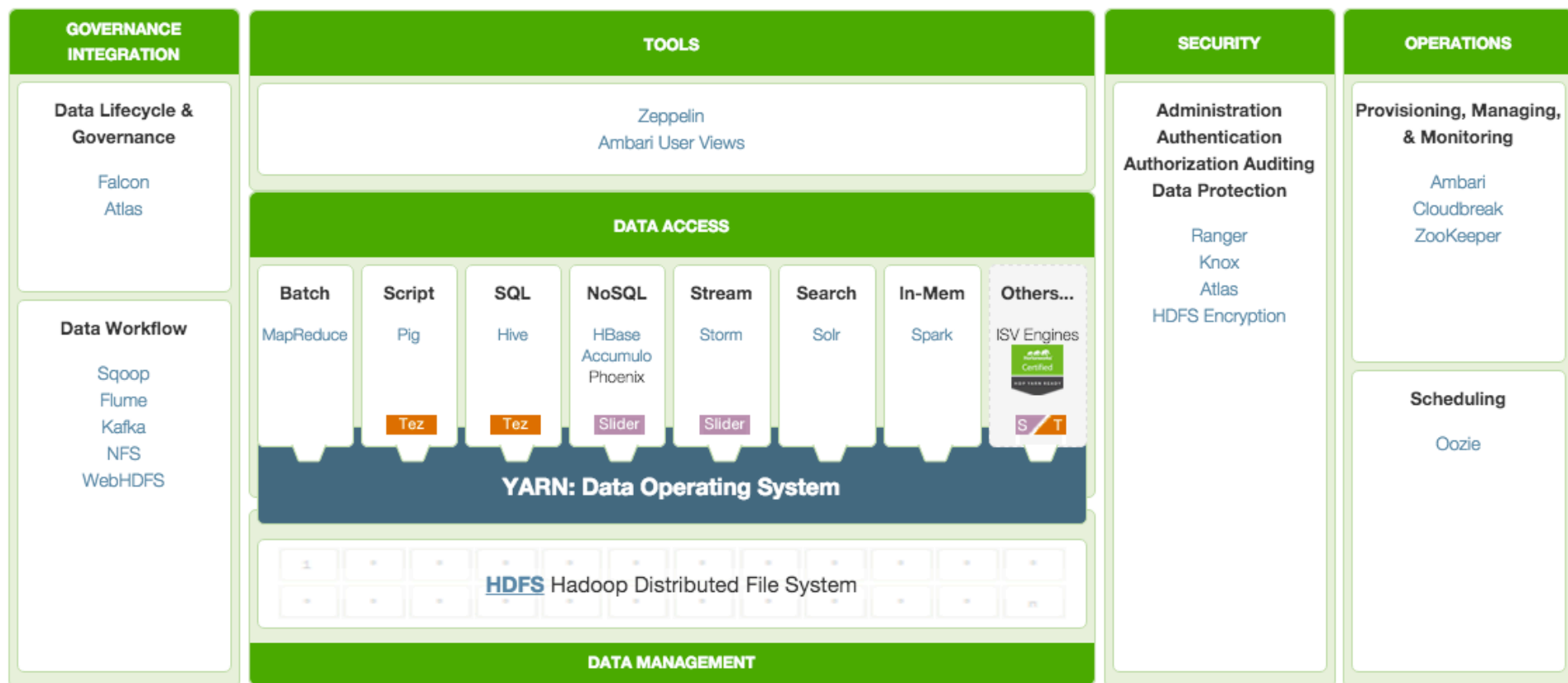


Hadoop Ecosystem



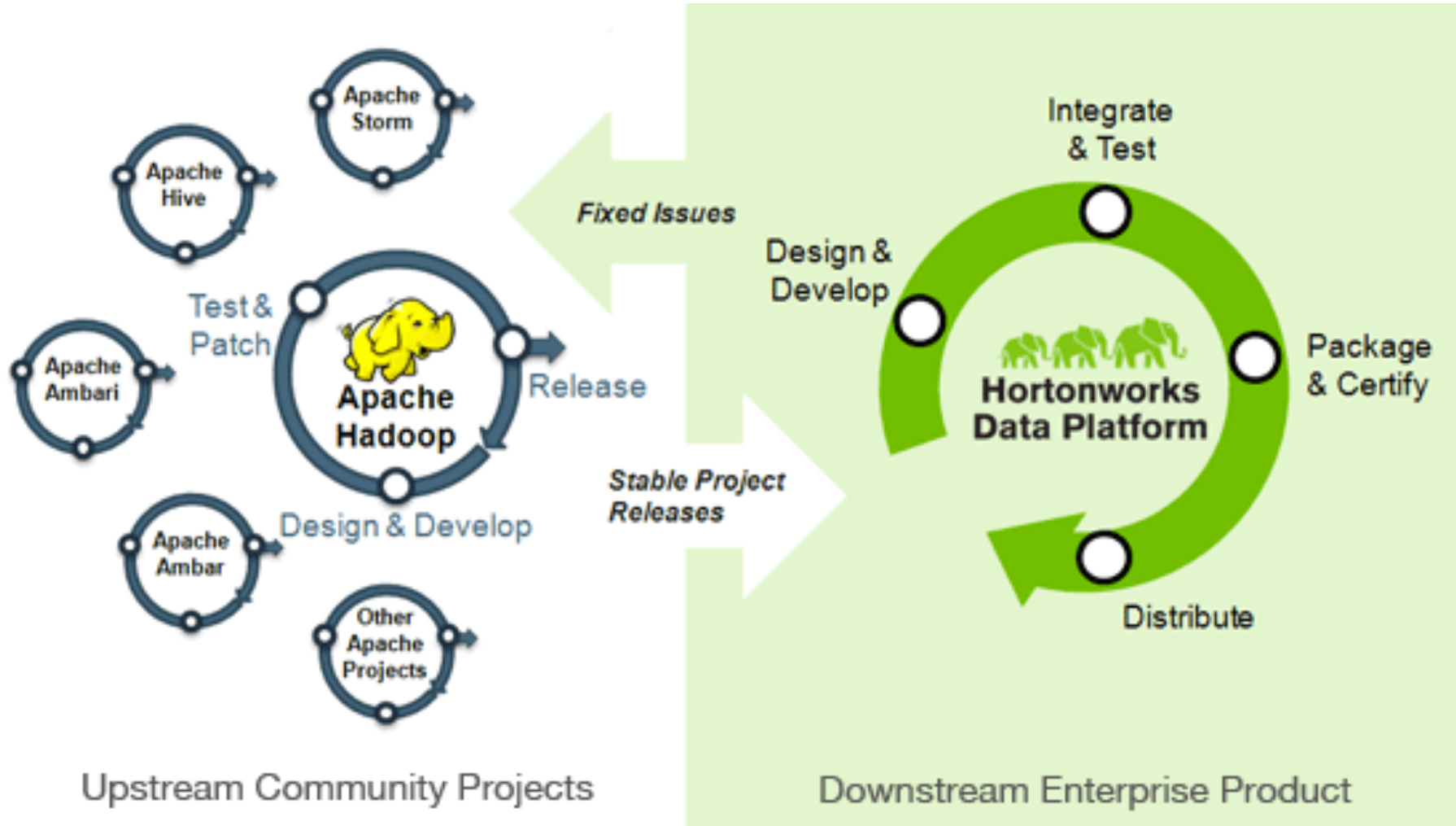
HDP (Hortonworks Data Platform)

A Complete Enterprise Hadoop Data Platform

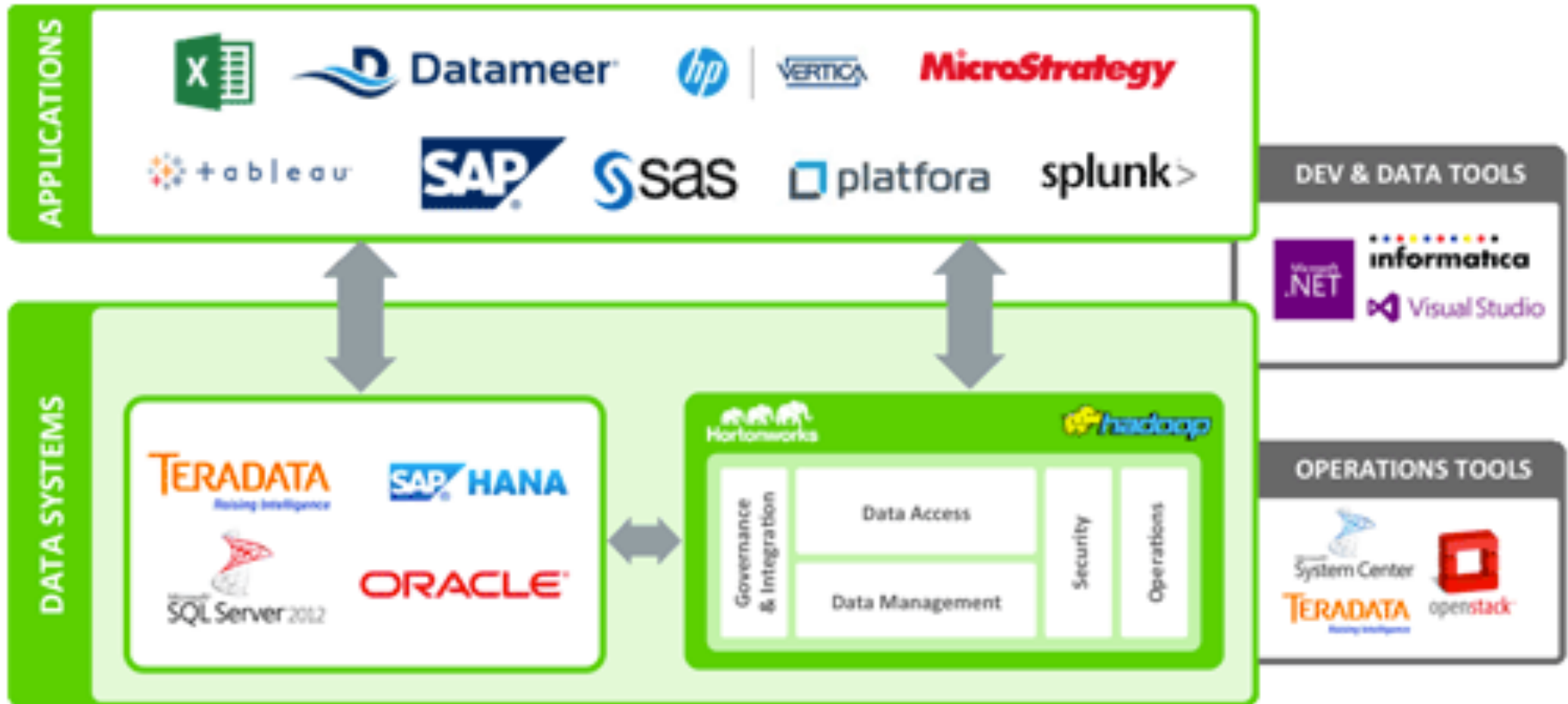


Apache Hadoop

Hortonworks Data Platform



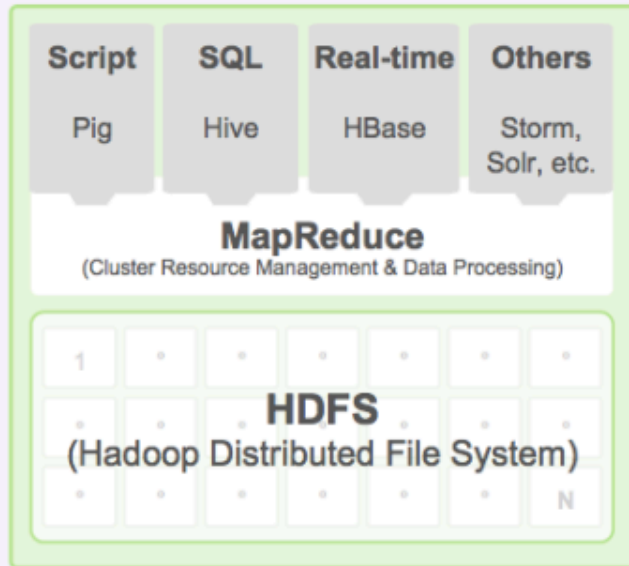
Hadoop and Data Analytics Tools



Hadoop 1 → Hadoop 2

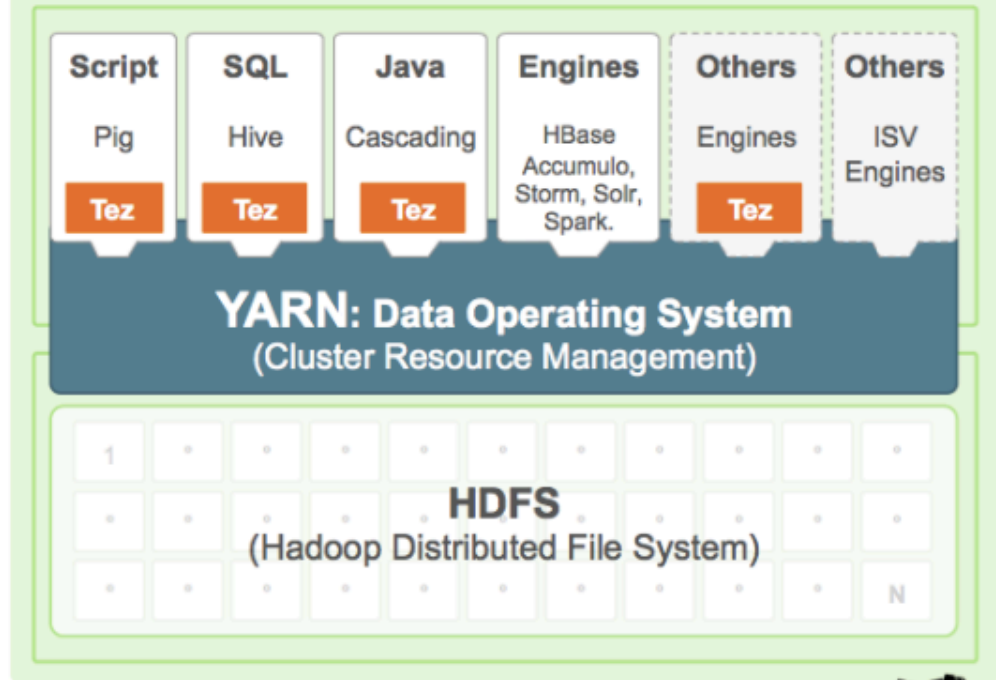
Hadoop 1

- Silos & Largely batch
- Single Processing engine

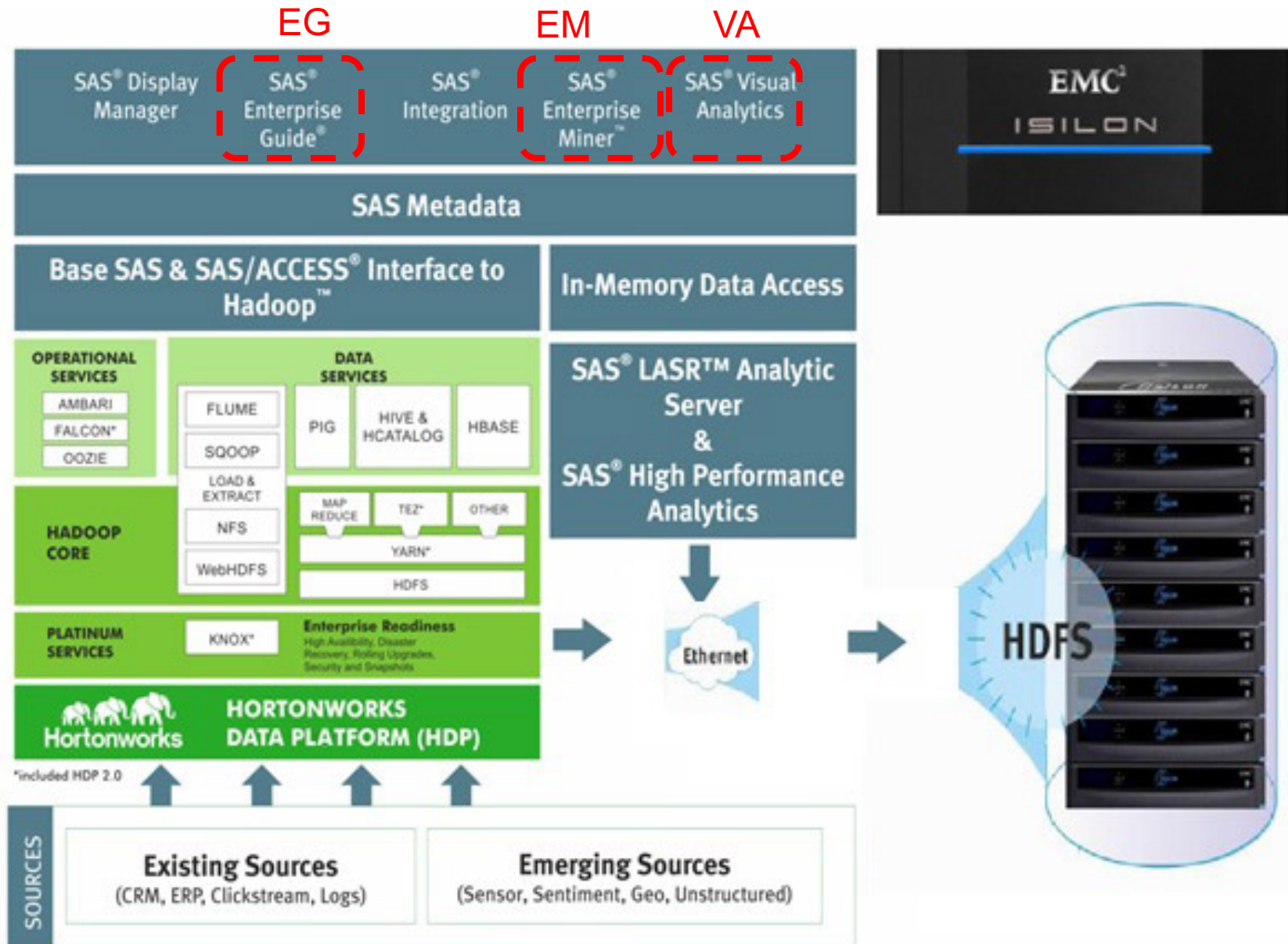


Hadoop 2 w/ Tez

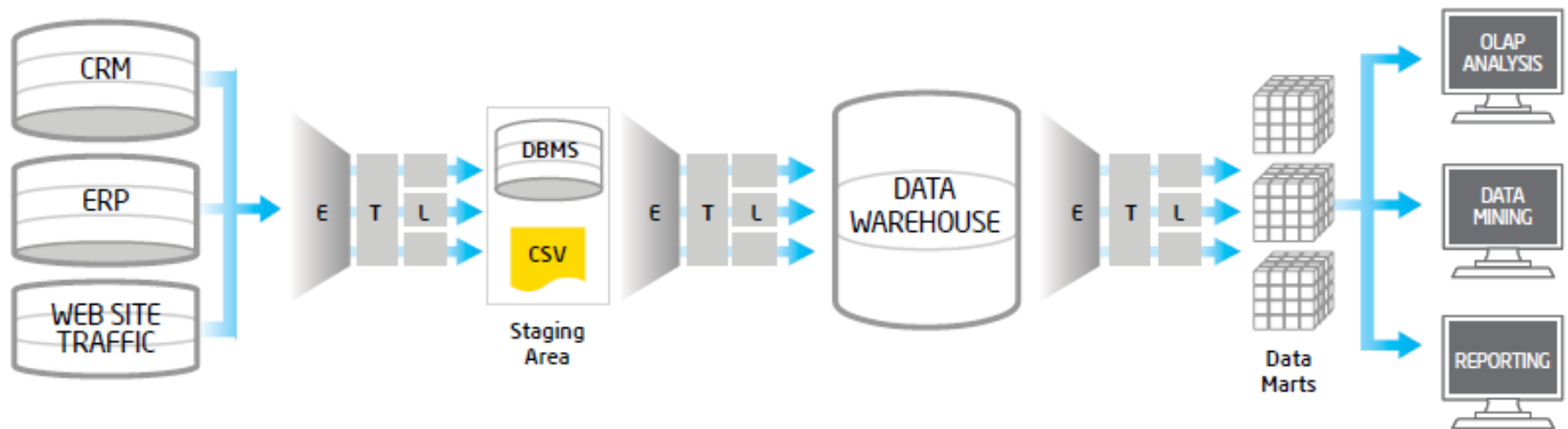
- Multiple Engines, Single Data Set
- Batch, Interactive & Real-Time



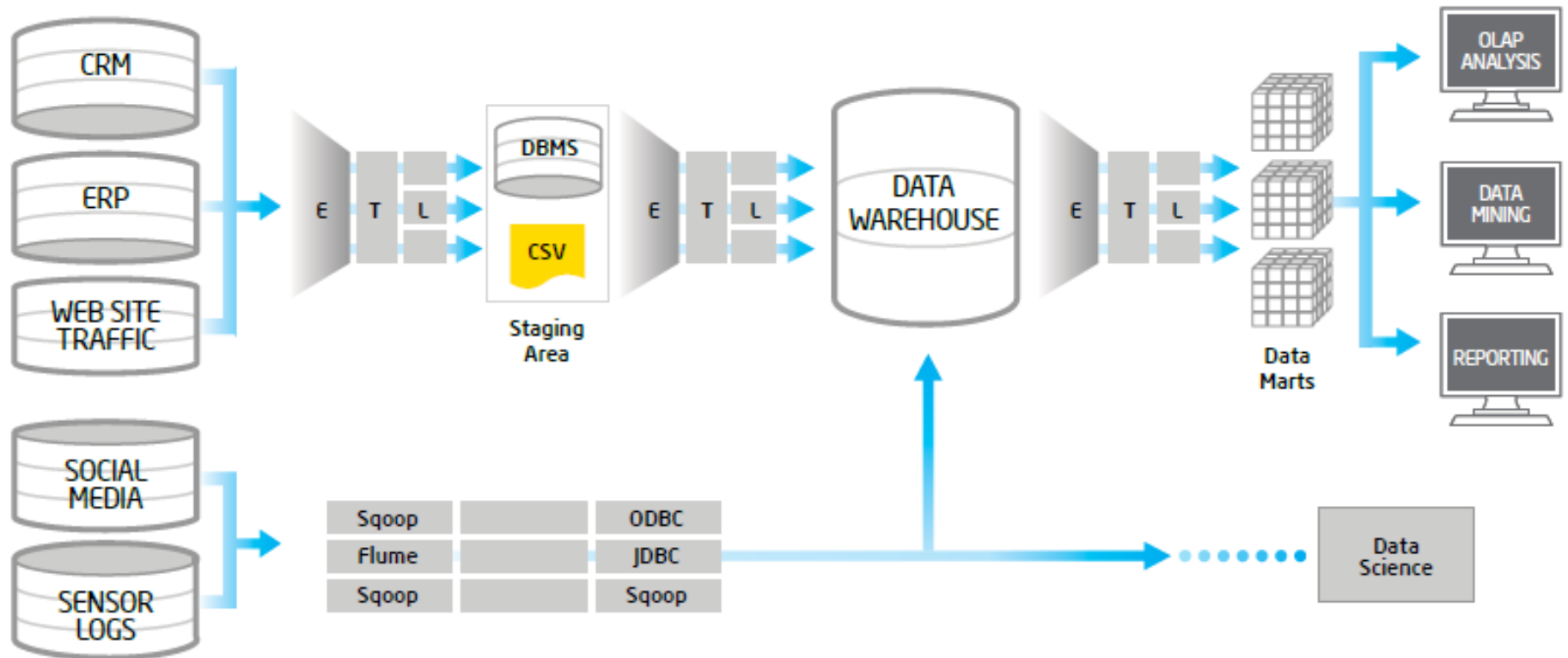
Big Data Solution



Traditional ETL Architecture



Offload ETL with Hadoop (Big Data Architecture)



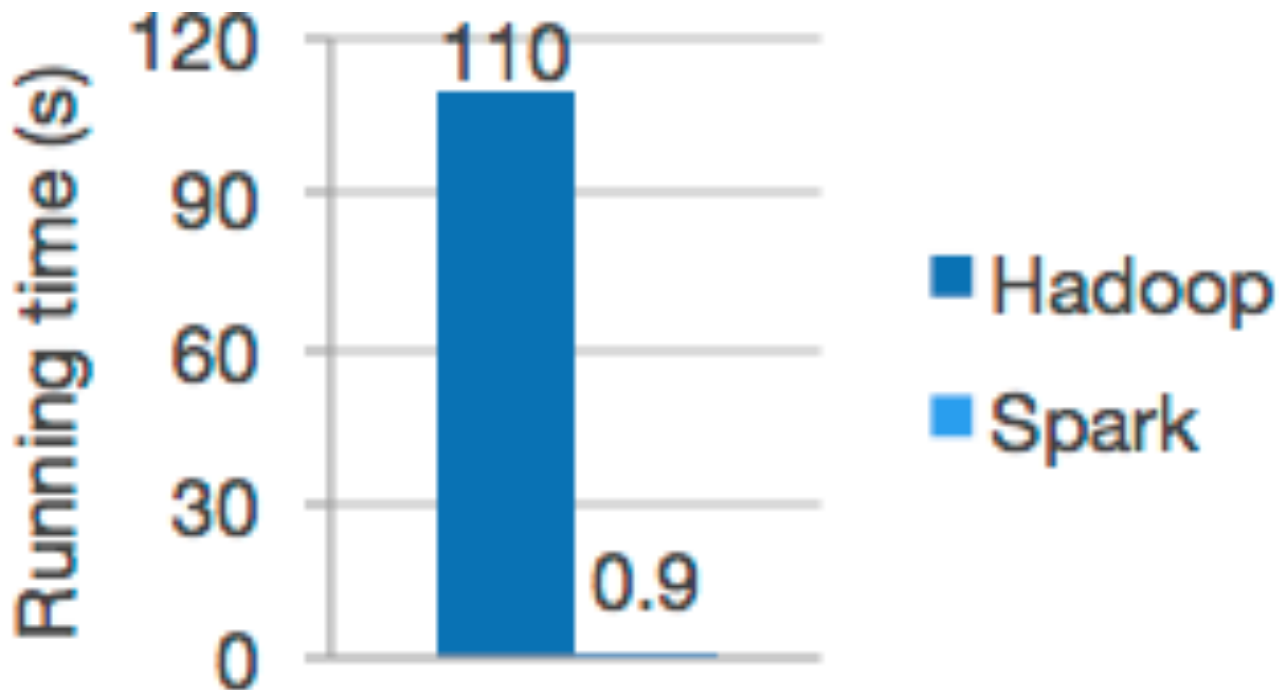
Spark Ecosystem



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

Spark SQL +
DataFrames

Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark Core API

R

SQL

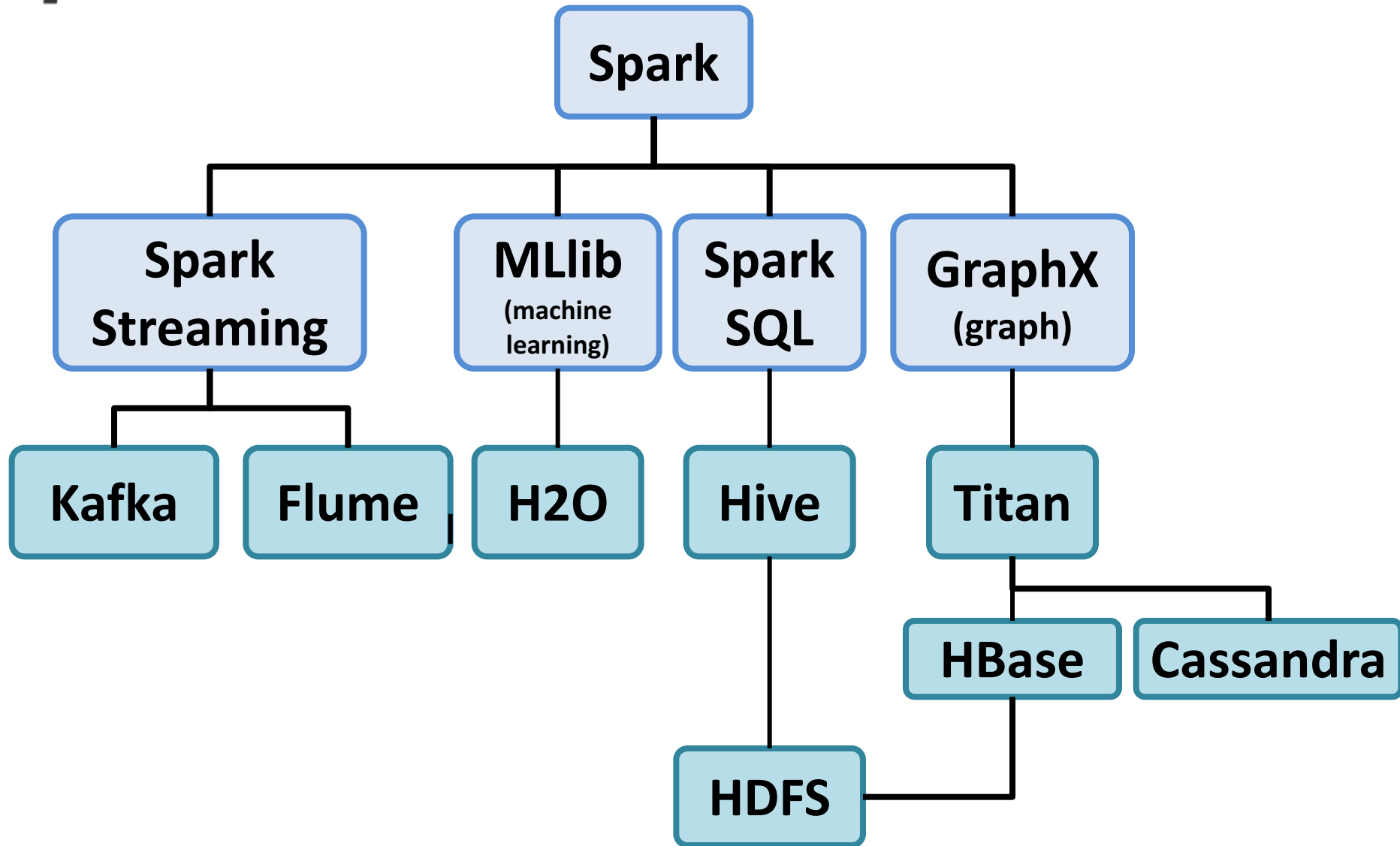
Python

Scala

Java



Spark Ecosystem



SMACK Stack

- **Spark**



- fast and general engine for distributed, large-scale data processing

- **Mesos**



- cluster resource management system that provides efficient resource isolation and sharing across distributed applications

- **Akka**



- a toolkit and runtime for building highly concurrent, distributed, and resilient message-driven applications on the JVM

- **Cassandra**



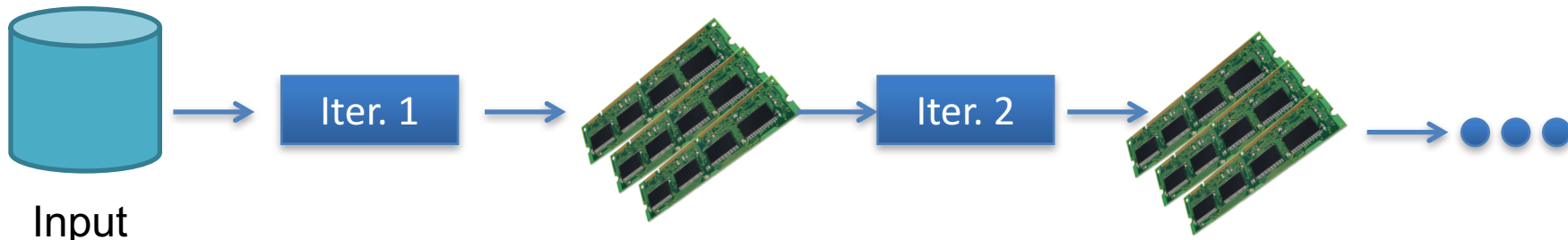
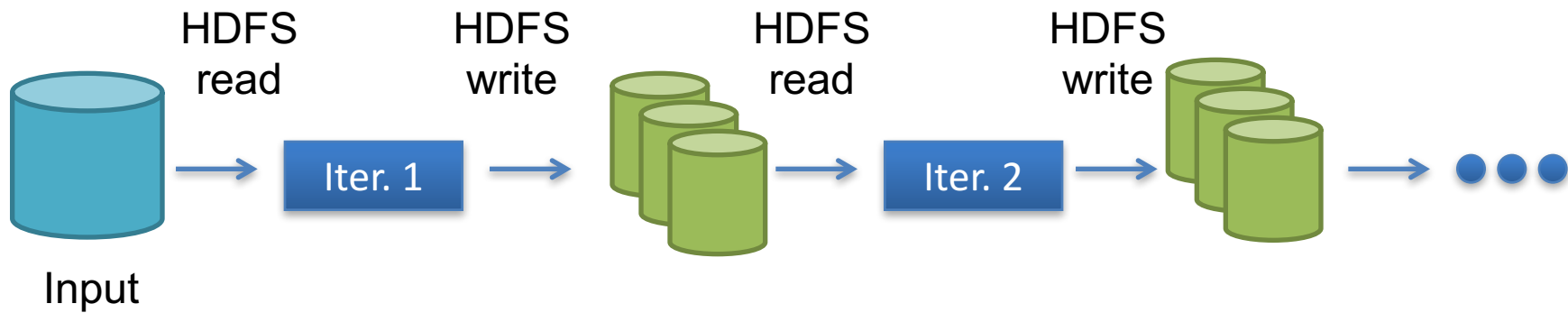
- distributed, highly available database designed to handle large amounts of data across multiple datacenters

- **Kafka**



- a high-throughput, low-latency distributed messaging system designed for handling real-time data feeds

Hadoop vs. Spark

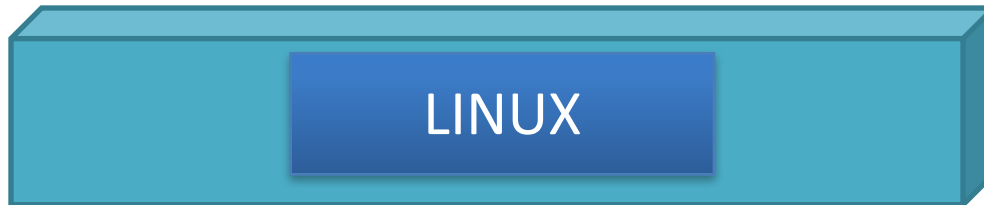
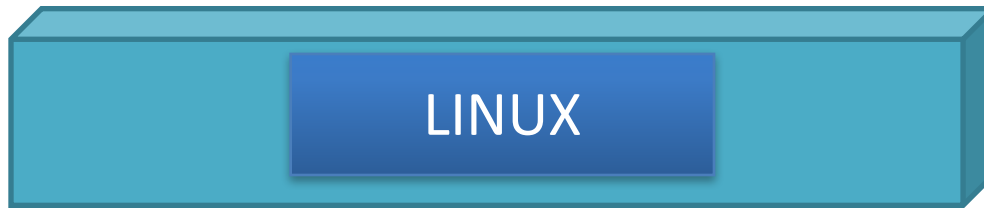


Hadoop Distribution

- Apache Hadoop
 - <http://hadoop.apache.org/>
- Amazon Elastic Map Reduce (EMR)
 - <https://aws.amazon.com/emr/>
- Cloudera CDH
 - <https://www.cloudera.com/downloads.html>
- Hortonworks Sandbox
 - <https://hortonworks.com/products/sandbox/>

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

Hadoop: Linux Based Software



Appliance

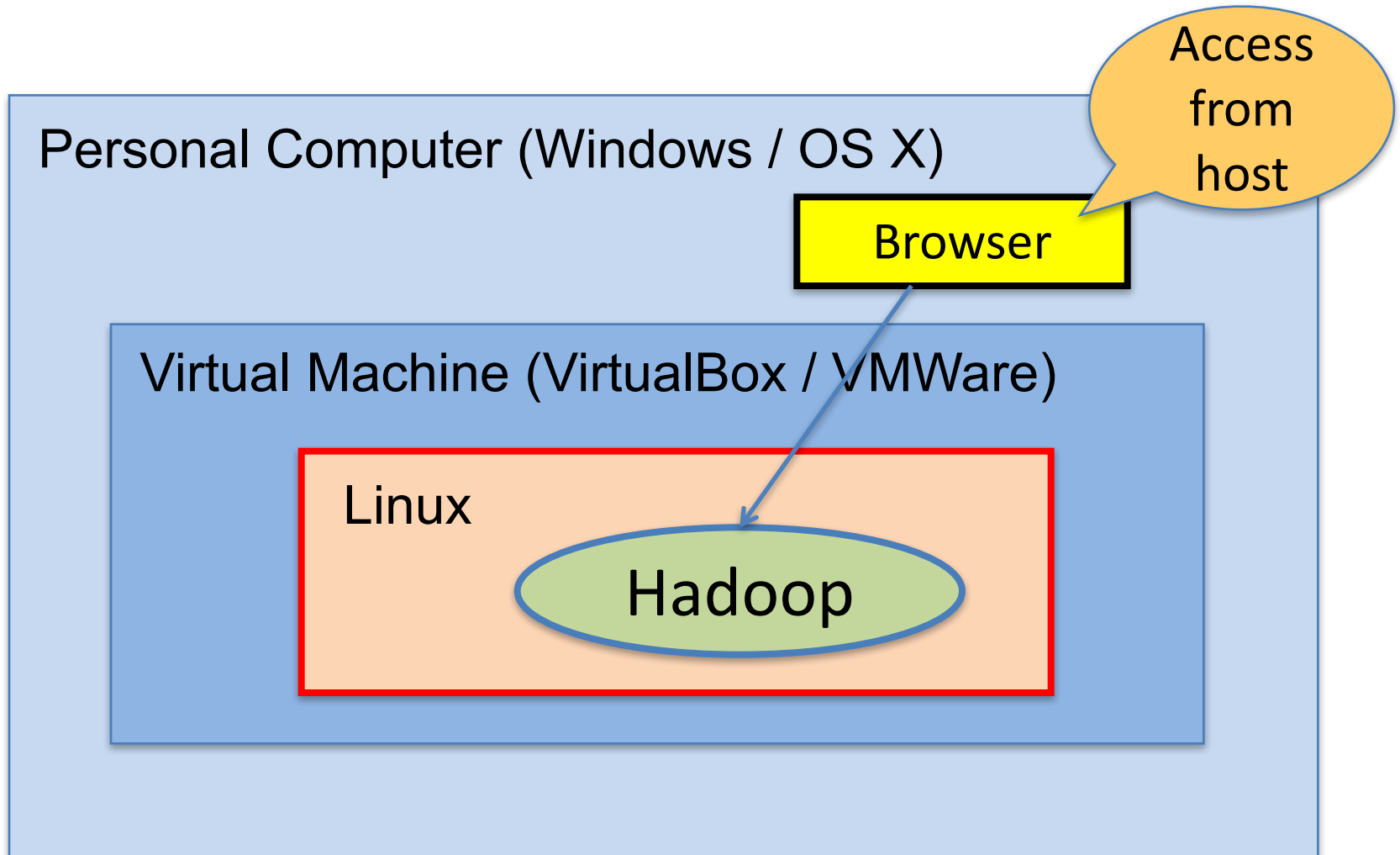
Personal Computer (Windows / OS X)

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



VirtualBox

[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 5.1

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 17th, 2017 VirtualBox 5.1.14 released!**
Oracle today released a 5.1 maintenance release which improves stability and fixes regressions. See the [Changelog](#) for details.
- **Important December 2nd, 2016 We're hiring!**
Looking for a new challenge? We're looking for a [GUI developer \(Germany/European Union\)](#).
- **New July 12th, 2016 VirtualBox 5.1 released!**
Many enhancements and improvements. Read more in the [announcement](#).

[More information...](#)

ORACLE

<https://www.virtualbox.org/>

- [About](#)
- [Screenshots](#)
- [Downloads](#)
- [Documentation](#)
 - [End-user docs](#)
 - [Technical docs](#)
- [Contribute](#)
- [Community](#)

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

COMMUNITY

BLOGS

PARTNERS

CONTACT US



SUPPORT LOGIN



ENGLISH ▼



Products

Solutions

Customers

Services & Support

About Us

GET STARTED

GET STARTED TODAY WITH HORTONWORKS SANDBOX



Ready to Get Started?

DOWNLOAD SANDBOX

OVERVIEW

WHAT'S NEW

DOWNLOADS

RESOURCES

≡ MENU

[products](#)

SHARE



NEWSLETTER



Contact Sales?



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

[COMMUNITY](#)[BLOGS](#)[PARTNERS](#)[CONTACT US](#)[SUPPORT LOGIN](#)[ENGLISH](#) ▼[Products](#)[Solutions](#)[Customers](#)[Services & Support](#)[About Us](#)[GET STARTED](#)

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

Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs in Scala, Java, Python, and R that allow data workers to efficiently execute machine learning algorithms that require fast iterative access to datasets (see [Spark API Documentation](#) for more info).

Spark on Apache Hadoop YARN enables deep integration with Hadoop [...]

[MENU](#)[SHARE](#)[NEWSLETTER](#)[Contact Sales?](#)

Apache Hadoop

Apache > Hadoop >



Top

Wiki

Search with Apache Solr

Search

Last Published: 01/27/2017 02:33:46

About

Welcome

- What Is Apache Hado...
- Getting Started ...
- Download Hadoop
- Who Uses Hadoop?...
- News
- Releases
- Release Versioning
- Mailing Lists
- Issue Tracking
- Who We Are?
- Who Uses Hadoop?
- Buy Stuff
- Sponsorship
- Thanks
- Privacy Policy
- Bylaws
- Committer criteria
- 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/>



PDF

Apache Hadoop

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

Apache > Hadoop >



Search with Apache Solr

Search

Last Published: 01/27/2017 02:33:46

Top

Wiki

About

- Welcome
- Releases
- Download
- Release Notes
- Release Versioning
- Mailing Lists
- Issue Tracking
- Who We Are?
- Who Uses Hadoop?
- Buy Stuff
- Sponsorship
- Thanks
- Privacy Policy
- Bylaws
- Committer criteria
- License

Documentation

Related Projects

Apache Hadoop Releases

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
3.0.0-alpha2	25 January, 2017	source	signature	checksum file
		binary	signature	checksum file
3.0.0-alpha1	03 September, 2016	source	signature	checksum file
		binary	signature	checksum file
2.7.3	25 August, 2016	source	signature	227785DC 6E3E6EF8..
		binary	signature	D489DF38 08244B90..
2.6.5	08 October, 2016	source	signature	3A843F18 73D9951A..
		binary	signature	001AD18D 4B6D0FE5..
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`

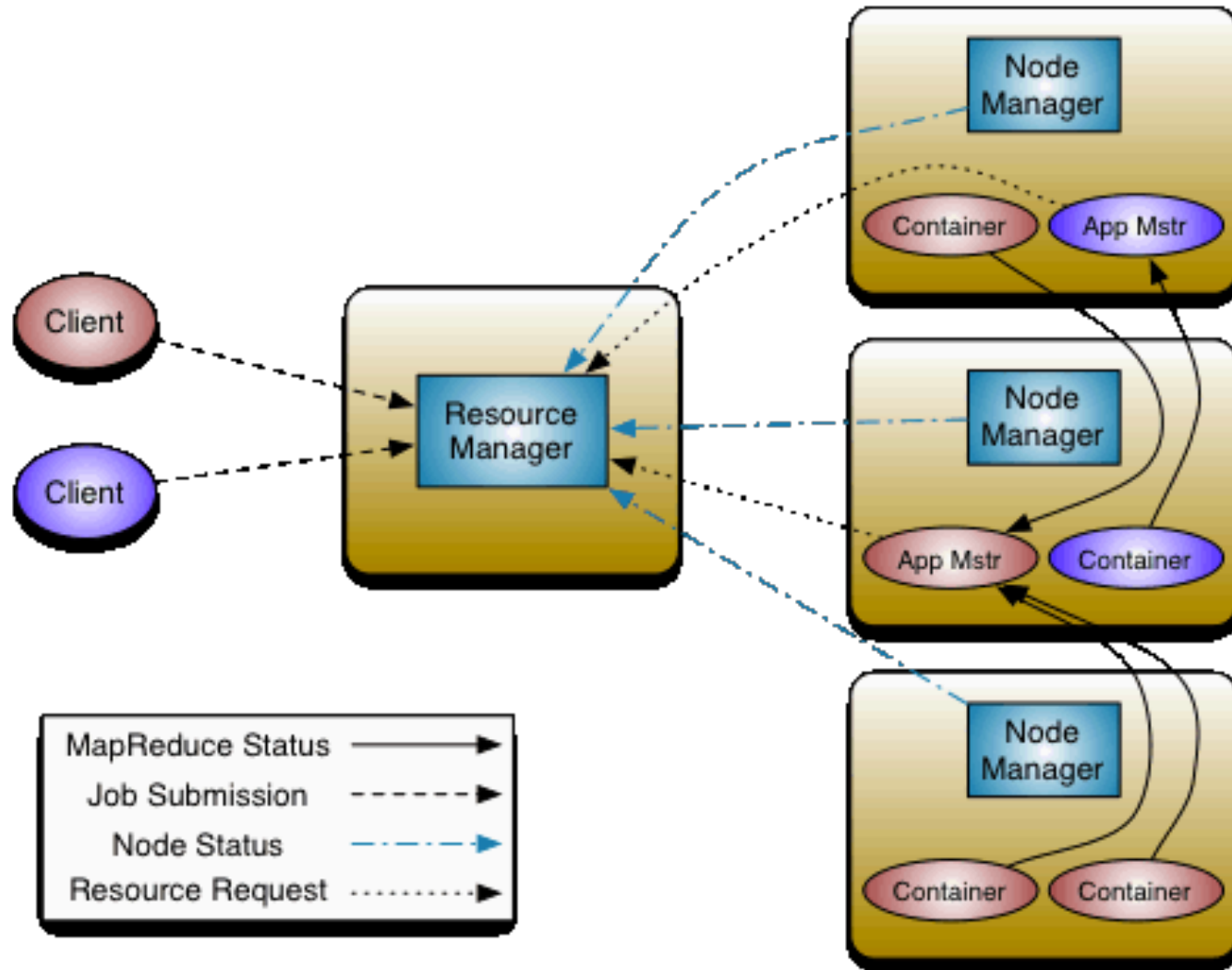
To perform a quick check using SHA-256:

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](#).



PDF

Apache Hadoop YARN



Apache Spark

[Download](#)[Libraries ▾](#)[Documentation ▾](#)[Examples](#)[Community ▾](#)[Developers ▾](#)[Apache Software Foundation ▾](#)

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

Latest News

Spark Summit East (Feb 7-9th, 2017, Boston) agenda posted (Jan 04, 2017)

Spark 2.1.0 released (Dec 28, 2016)

Spark wins CloudSort Benchmark as the most efficient engine (Nov 15, 2016)

Spark 2.0.2 released (Nov 14, 2016)

[Archive](#)[Download Spark](#)

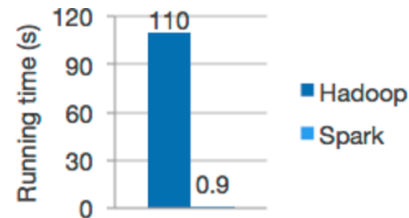
Built-in Libraries:

[SQL and DataFrames](#)[Spark Streaming](#)[MLlib \(machine learning\)](#)[GraphX \(graph\)](#)[Third-Party Projects](#)

Speed

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

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



Logistic regression in Hadoop and 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

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

Big Data Landscape 2016 (Version 3.0)

Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

Hadoop in the Cloud
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Qubole

Spark
databricks, GridGain, TACHYON, NEXUS

Cluster Services
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Qubole

Analytics

NoSQL Databases
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Qubole

NewSQL Databases
SAP, Clustrix, Pivotal, paradigm4, nuodb, splicemachine, memsql, VOLTDB, MariaDB, citusdata, deepdb, Trafilon, Cockroach LABS

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITALINSIGHT

Analytics Platforms
Microsoft, GUAVUS, Datameer, Bottlenose, interana

Data Science Platforms
context relevant, DataRobot, ARIMO, MODE, plotly, dataiku, tonian, DOMINO, sense, yhat, ALGORITHMIA

Visualization
tableau, Qlik, looker, Roambi, SISENSE, GCHDATA, datorama, CHARTIO

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, blueyonder, Livefyre, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, sense, QUANTIFIND, ACTIONIQ, fusemachines, ENGAGIO

Customer Service
MEDALLIA, ATTENTIVITY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appurri, Wiseio

Human Capital
gild, Connectifier, textio, entelo, hiQ, RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Graph Databases
neo4j, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, NETEZZA, Action, cognitio, XSOLO, dremio

Cloud EDW
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Qubole

Data Transformation
alteryx, talend, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica, MuleSoft, snapLogic, BedrockData, xplenty

BI Platforms
Power BI, amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Qubole

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
Hootsuite, NETBASE, DATASIFT, track, bitly, synthetio, simplereach

Ad Optimization
AppNexus, MediaMath, critico, OpenX, rocketfuel, Integral, theTradeDesk, Adgorithms, distillery, Livelihood, TAPAD, DataXu, Appier, MOAT

Security
CYLANCE, CounterTrack, cyberreason, ThreatMetrix, AREA1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
facebook, Clara, KASIST, lumiata

Management / Monitoring
New Relic, APPDYNAMICS, amazon, actifio, Numerify, splunk, DATADOG, DRIVEN, Anodot

Security
TANUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Qubole

App Dev
apigee, CASK, Typesafe, DRIVEN

Crowd-sourcing
amazon, Mechanical Turk, CrowdPower, WorkFusion

Real-Time
amazon, METAMARKETS, strim, confluent, DATAFORENT, dataArtisans

Machine Learning
Azure Machine Learning, H2O, Dato, SKYTREE, rapidminer, DATAFPM, deepstream, VISEN, PredictionIO, glowfish

Speech & NLP
NarrativeScience, NUNANCE, WolframAlpha, semantic, ARRIA, nora, Numenta, HyperScience, clarifai, DEXTER, MetaMind

Horizontal AI
IBM Watson, Cortana, Viv, nervana, nora, Numenta, HyperScience, clarifai, DEXTER, MetaMind

Publisher Tools
outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
Socrata, OPENGOV, EN, FiscalNote, enigma, mark43, OpenDataSoft

Finance
affirm, LendingClub, OnDeck, Kreditech, zest finance, LendUp, Kabbage, tidemark, uora, Dataminr, Lenddo, KENSHC, AIDYA, iSENTIUM, Quantopian, sentient

Cross-Infrastructure/Analytics
amazon, Google, Microsoft, IBM, SAP, sas, hp, Oracle, NetApp, VMware, TIBCO, TERADATA, ORACLE, NetApp

Education / Learning
KNEWTON, Clever, Declara, PANORAMA, knowte

Life Sciences
23andMe, Counsyl, Recombine, KYRUUS, zymogen, HealthTap, METABIOTA, ZEPHYR, HEALTH, ovia, Gingerio, transcriptic, Glow, enlitic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, STITCH FIX, WorkFusion, BLUE RIVER, TACHYON, Seeq, FarmLogs, SwiftKey, HowGood, select, NIGHT MACHINE, statmuse, BOEYER

Framework
hadoop, YARN, Spark, MESOS, TEZ, Flink, CDAP

Query / Data Flow
SLAMDATA, DRILL, Google Cloud Dataflow

Data Access
cassandra, HBASE, mongoDB, couchDB, riak, kafka, nifi, OPENTSDB

Coordination
talend, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
ScalaLab, Numpy, SciPy

Open Source

Machine Learning
mlilb, Aerosolve, Caffe, FeatureFu, DIMSUM, WEKA, jupyter, DL4J

Search
elasticsearch, Solr, Lucene, Zepplin

Security
Apache Ranger, Visualizaton

Data Sources & APIs

Health
Apple, JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara, AUGURY, estimate

Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, S&P CAPITAL IQ, YODLEE, PREMIER, quandle, xignite, CB INSIGHTS, mattermark, StockTwits, Estimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location / People / Entities
axiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTO, factual, PlaceIQ, CIRCULATE, placemeter, BASIS, Sense

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, PLURALSIGHT, INSIGHT, DataCamp, DataElite, The Data Incubator, METIS

Last Updated 3/23/2016

© Matt Turck (@mattturck), Jim Hao (@jimrho), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

Source: <http://mattturck.com/2016/02/01/big-data-landscape/>

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>