

CONTENTS

Preliminary Topics
事前準備

01

Hadoop Eco System

02

Hadoop Architecture

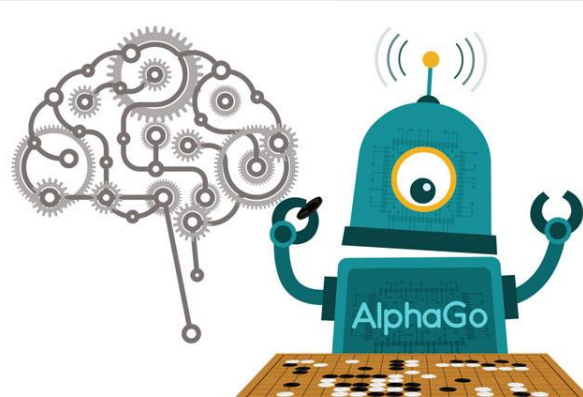
03

Hadoop Environment
実験環境

04

05 Hadoop Command

06 Reference Books



PART1



Preliminary Topics
事前準備



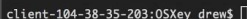
Official Website :
<https://www.ubuntu.com/>
<http://releases.ubuntu.com/>



Official Website :
<https://www.centos.org/>
<https://www.centos.org/download/>
<https://wiki.centos.org/Download>



5



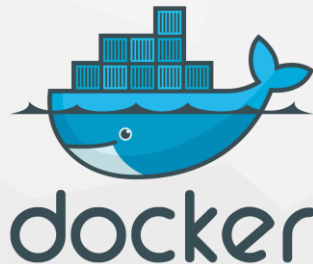
```

/ The eternal feminine draws us upward. \
|                                         |
\ -- Goethe                             /
-----
      \   ^__^
        (oo)\_____)
           (_____)

```



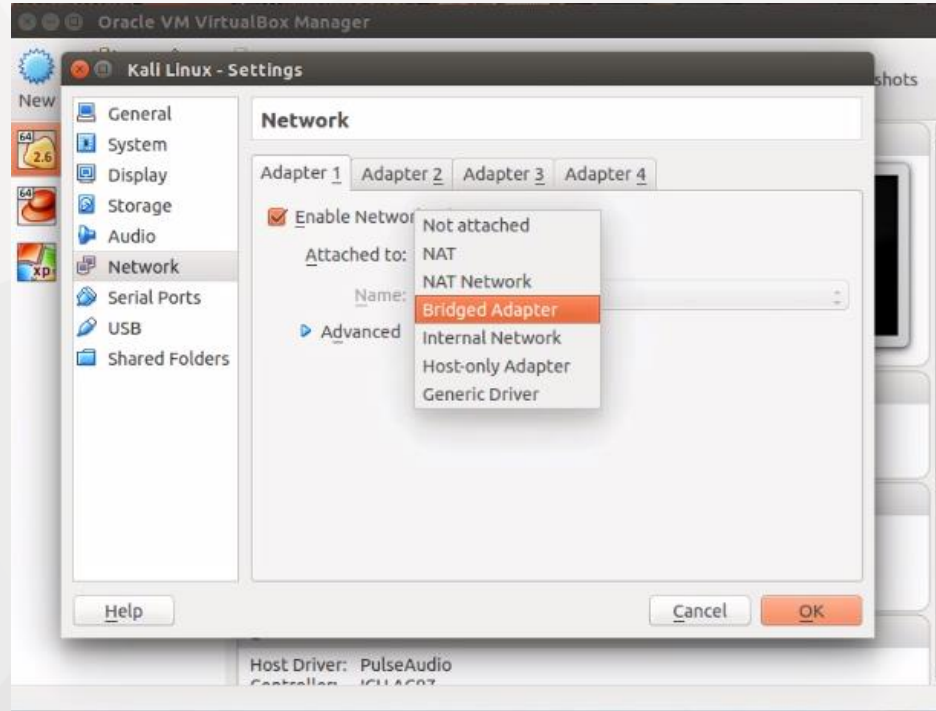
VirtualBox



CITRIX[®]

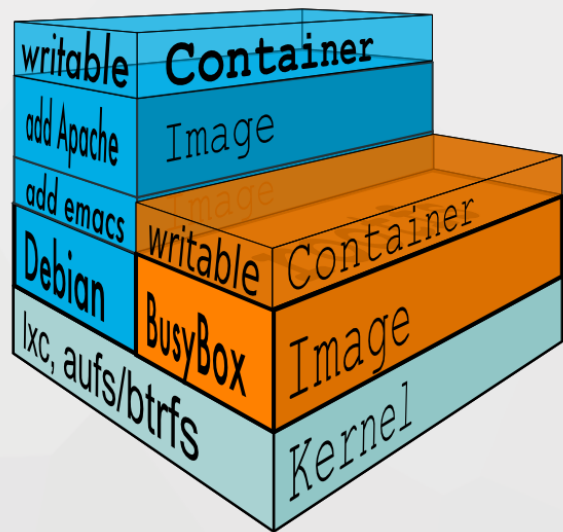
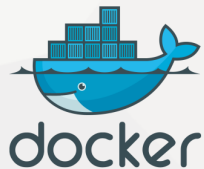


Virtualization with
KVM on **Linux[™]**






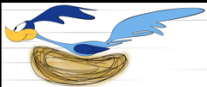


Official Website :
<https://www.virtualbox.org/>
<https://www.virtualbox.org/wiki/Downloads>

Virtual networking :
<https://www.virtualbox.org/manual/ch06.html>





VM vs. Docker




Size		
Startup		
Integration		





This repository ▾

Search or type a command 

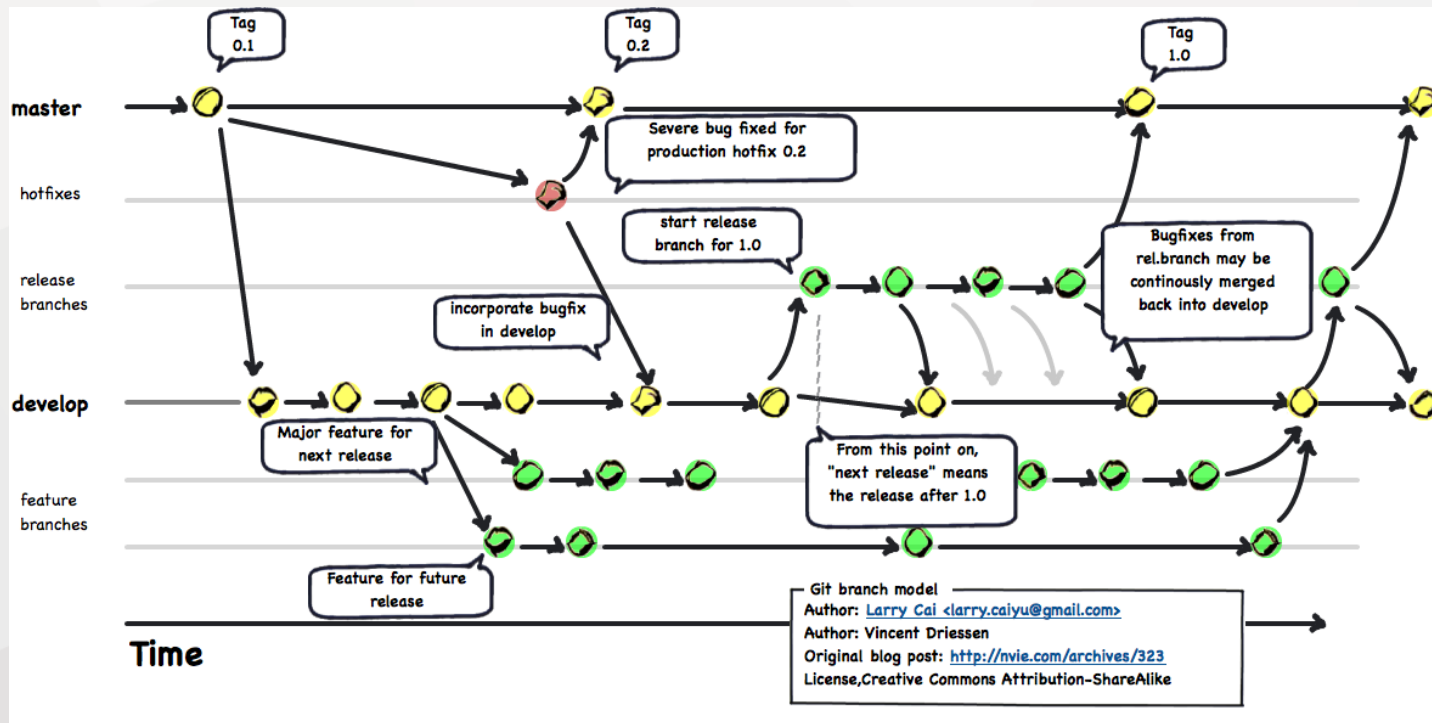
Explore Gist Blog Help

 **apache / hadoop-common** Watch ▾ 13mirrored from [git://git.apache.org/hadoop-common.git](https://git.apache.org/hadoop-common.git)

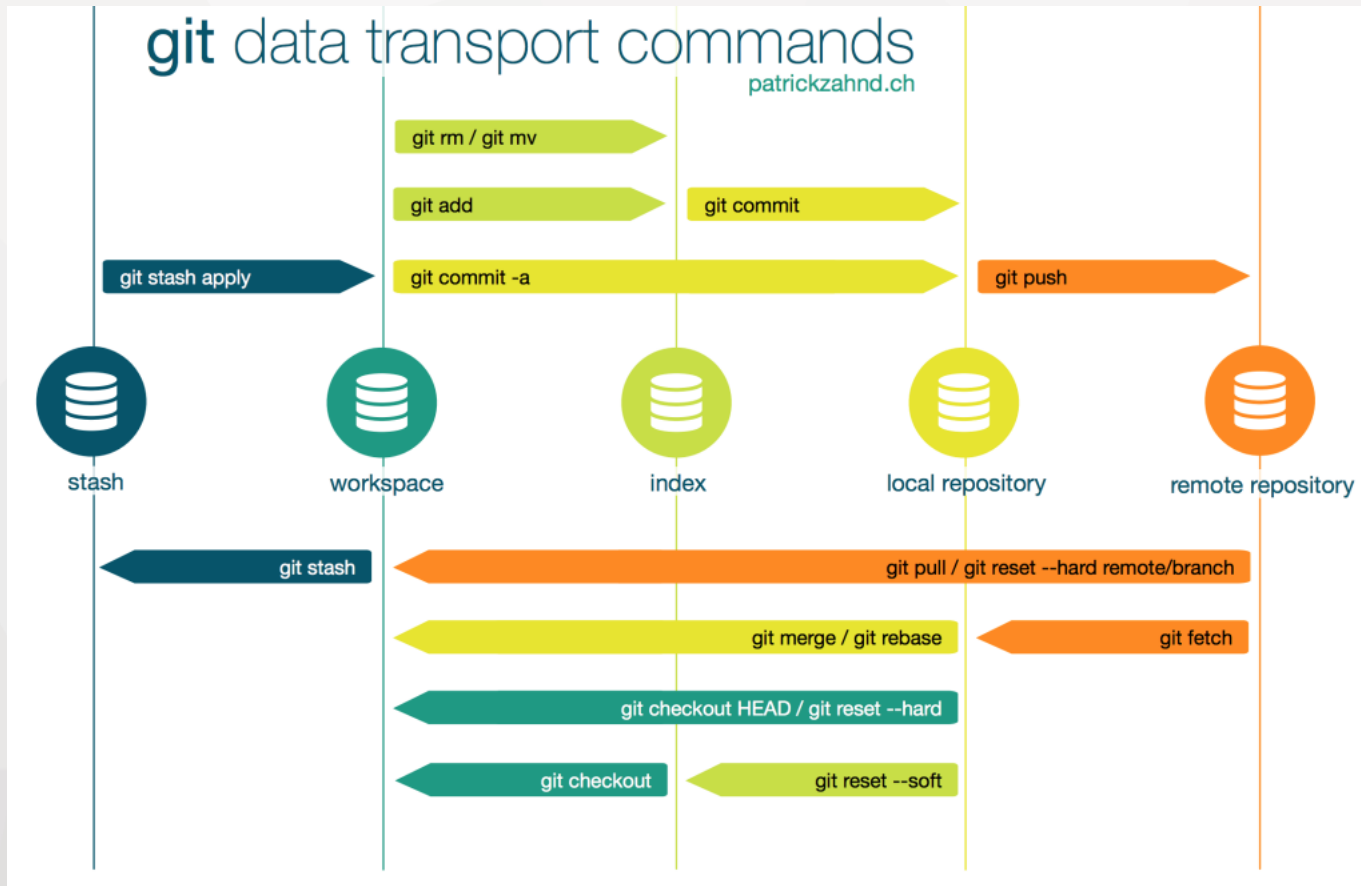
Mirror of Apache Hadoop common

 **8,109** commits  **104** branches  **174** releases  **13** contributors branch: **trunk** ▾ **hadoop-common** / Move HDFS-5276 to 2.3.0 in CHANGES.txt  Luke Lu authored 4 hours ago latest commit 55a793c

 dev-support	HADOOP-9848 Addendum fixing OK_JAVADOC_WARNINGS in test-patch	2 mont
 hadoop-assemblies	YARN-1021. Yarn Scheduler Load Simulator. (ywsykcn via tucu)	14 da
 hadoop-client	HADOOP-9557. hadoop-client excludes commons-httpclient. Contributed b...	a mor
 hadoop-common-project	HADOOP-10039. Add Hive to the list of projects using AbstractDelegati...	11 hou



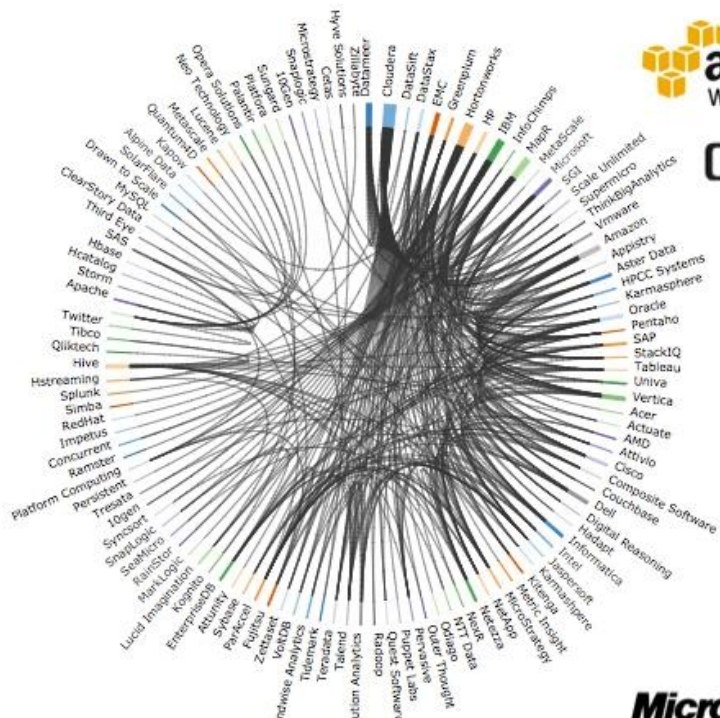
Official Website :
<https://git-scm.com/>
 The latest stable Release : v2.11.1





PART2

Hadoop Eco System



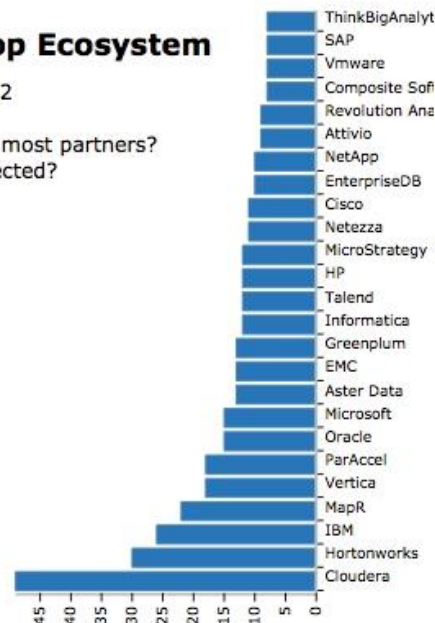
cloudera



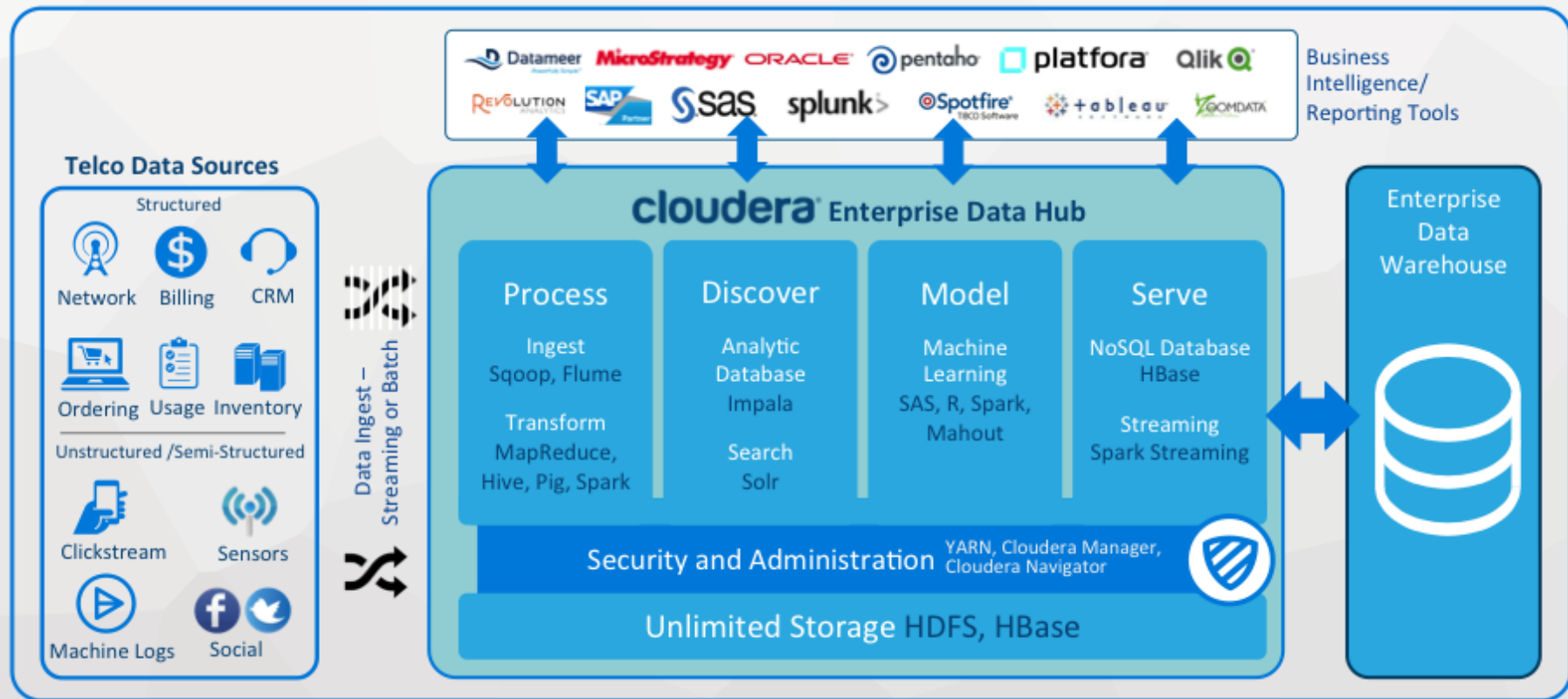
The Hadoop Ecosystem

June 21, 2012

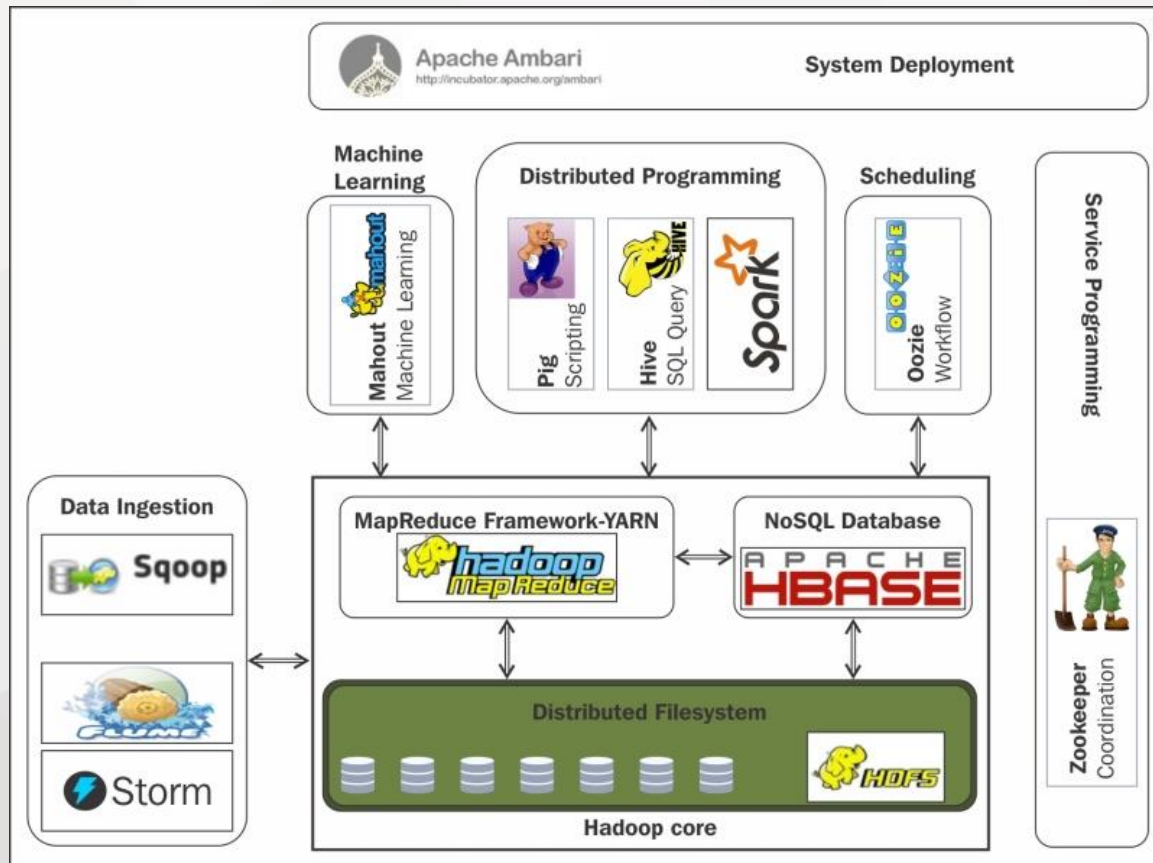
Who has the most partners?
Who is connected?



brought to you by Datameer
Powerfully Simple™

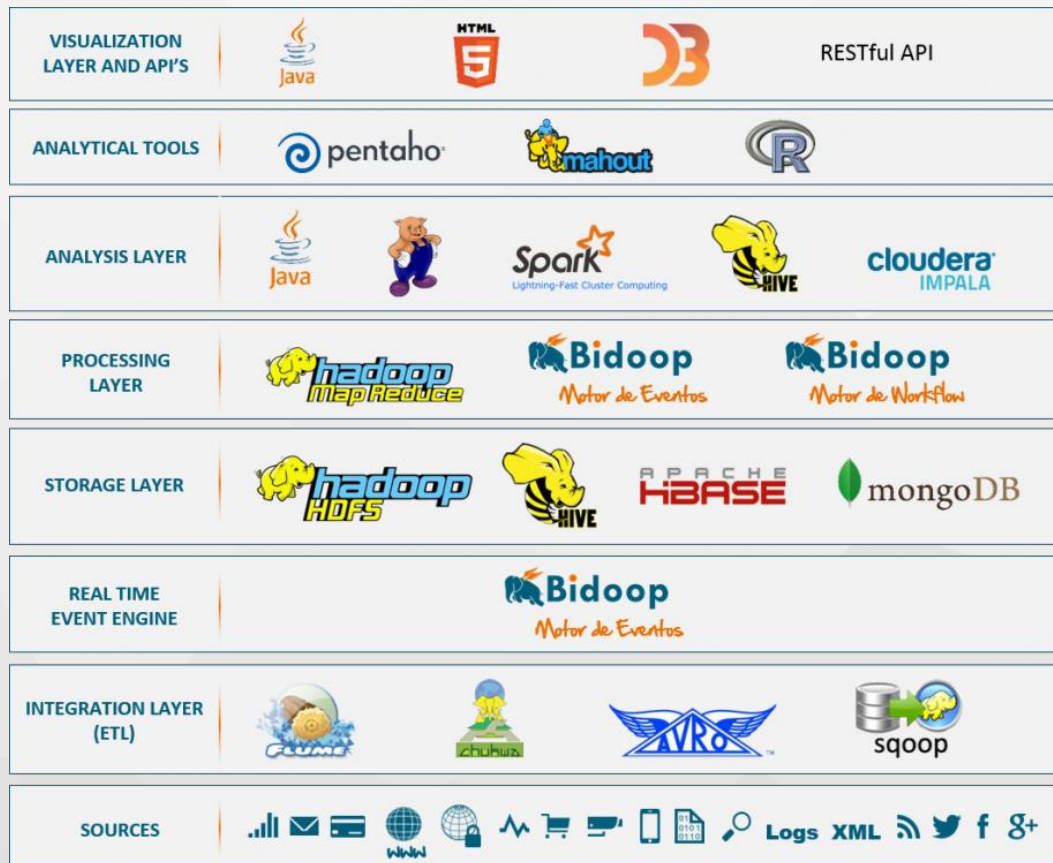


Hadoop Eco System



Hadoop Eco System

16

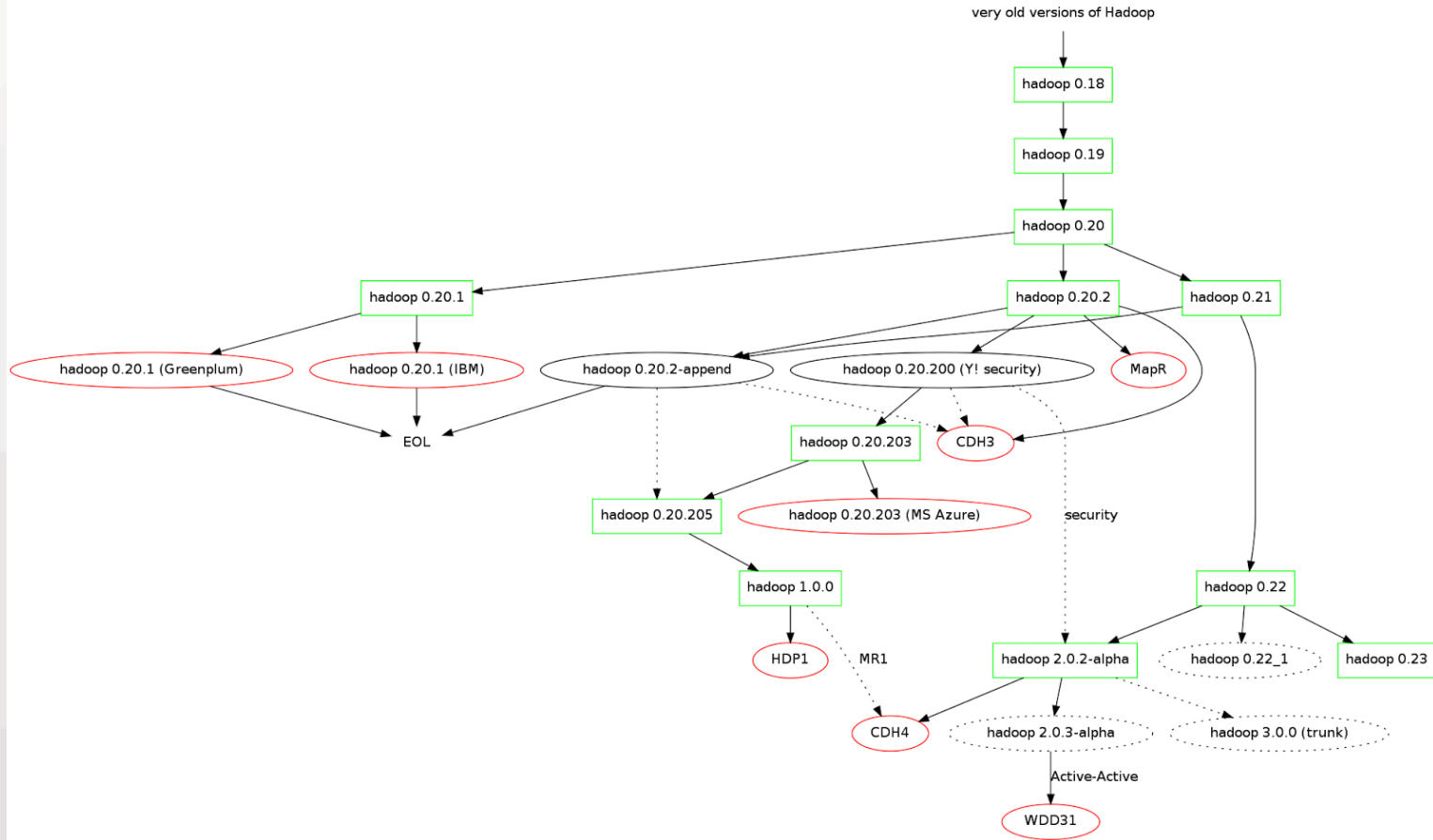




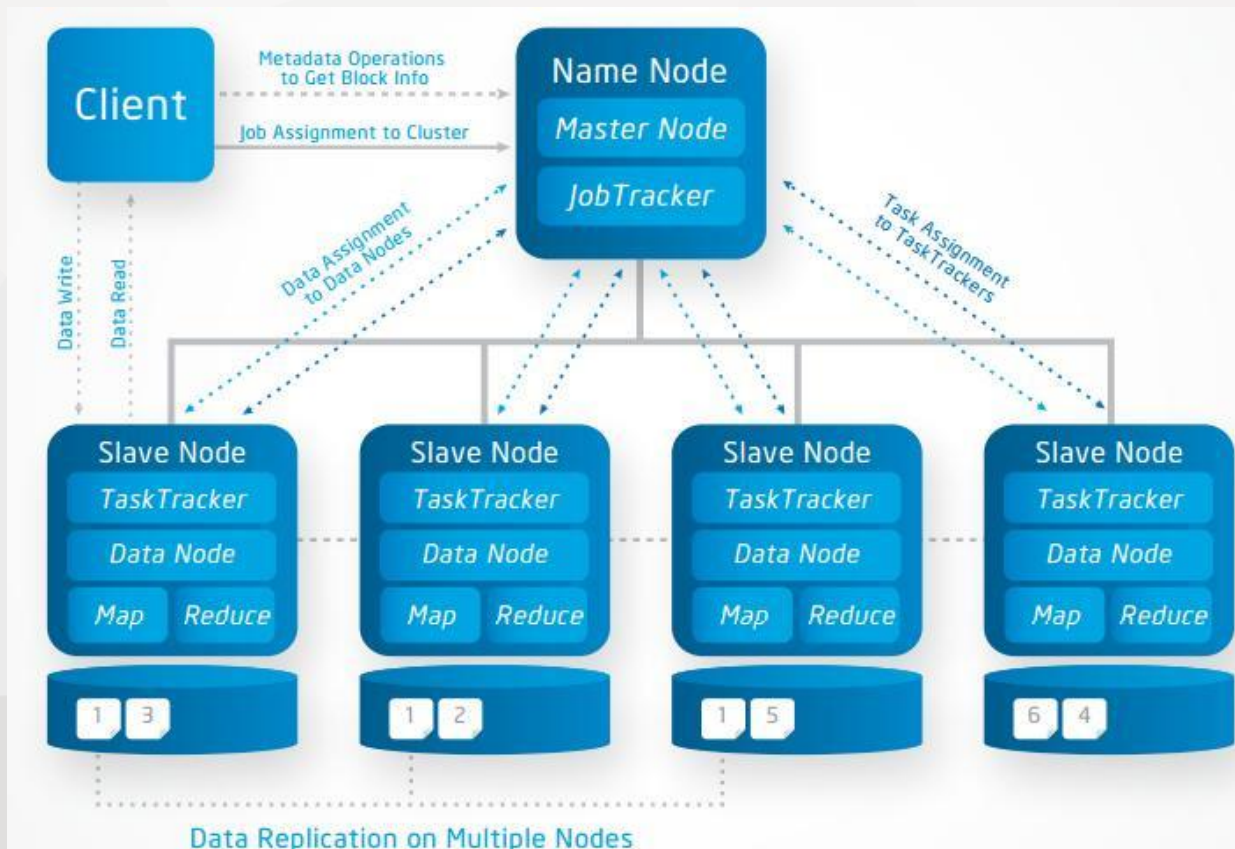
PART3

Hadoop Architecture

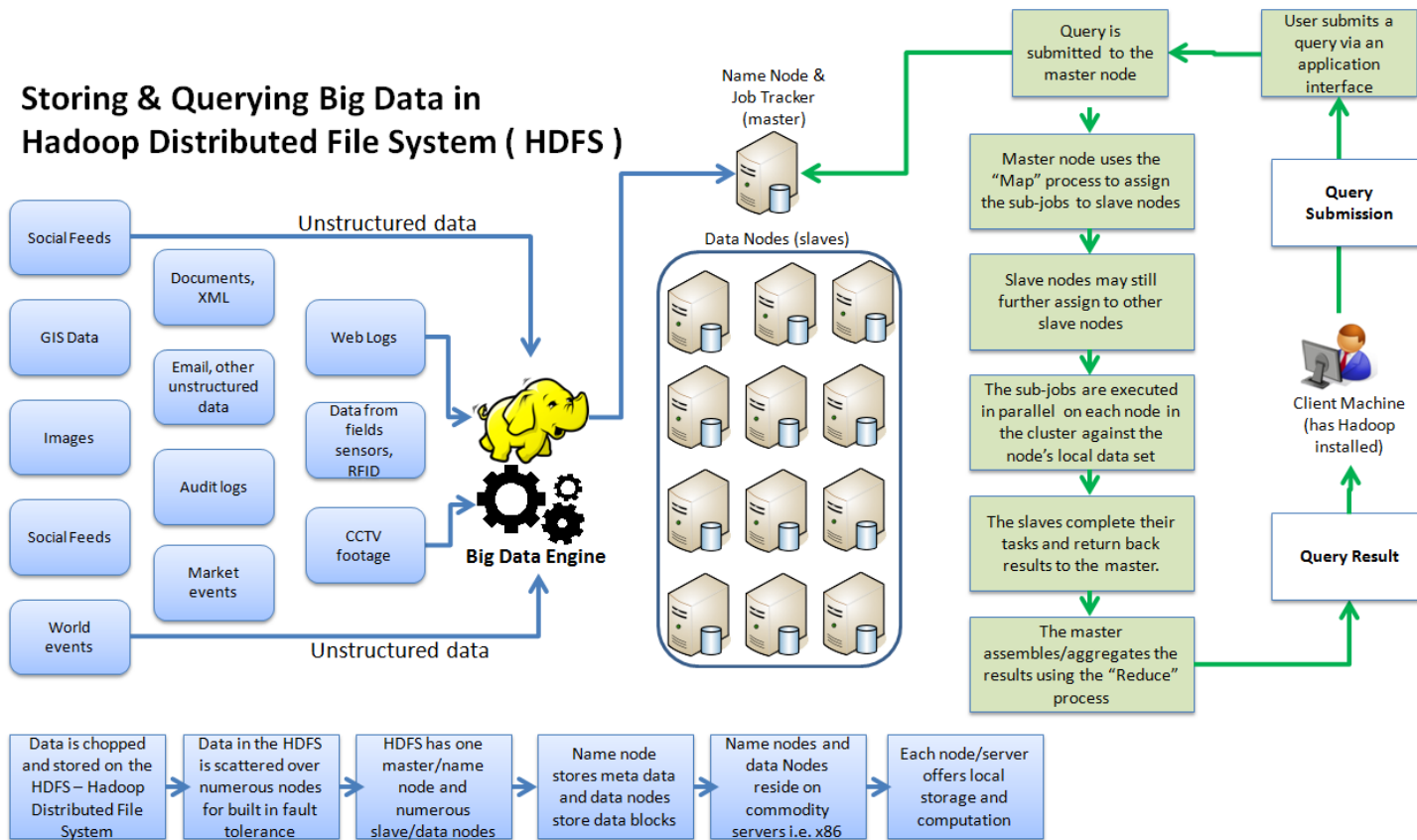
18



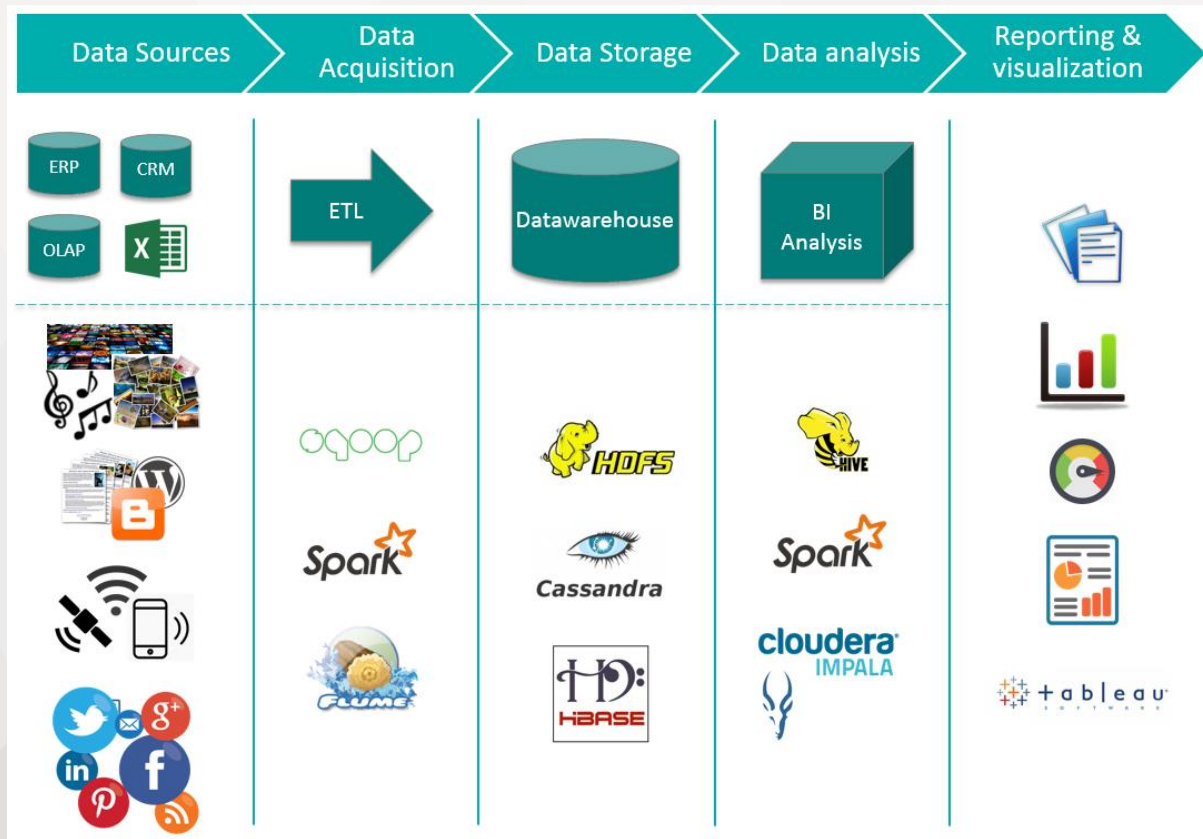
Architecture



Storing & Querying Big Data in Hadoop Distributed File System (HDFS)



Architecture





PART4

Hadoop Environment

Oracle Virtual Box

The screenshot displays the Oracle VM VirtualBox Manager interface. On the left, a list of virtual machines is shown, including 'Hadoop-Master-Ubuntu-14.04' (Running), 'Hadoop-Slave1-Ubuntu-14.04' (Running), 'Hadoop-Slave2-Ubuntu-14.04' (Running), 'HBase-Slave1-Ubuntu-14.04' (Powered Off), 'HBase-Slave2-Ubuntu-14.04' (Powered Off), 'HBase-Master-Ubuntu-14.04' (Powered Off), 'Spark-Master-Ubuntu-14.04' (Powered Off), 'Spark-Slave1-Ubuntu-14.04' (Powered Off), 'Spark-Slave2-Ubuntu-14.04' (Powered Off), 'CentOS 6.7', and 'cloudera-quickstart-vm-5.8.0-0...'. The 'Hadoop-Master-Ubuntu-14.04' VM is selected, and its details are shown on the right. The details pane includes sections for General, System, Display, Storage, Audio, Network, and USB. The General section shows the VM name, operating system, and groups. The System section shows base memory, boot order, and acceleration. The Display section shows video memory, remote desktop server, and video capture. The Storage section shows controllers and ports. The Audio section shows the host driver and controller. The Network section shows the adapter and controller. The USB section shows device filters.

Oracle VM VirtualBox Manager

File Machine Help

New Settings Show Discard

Details Snapshots

Ubuntu 14.04

- Hadoop-Master-Ubuntu-14.04** Running
- Hadoop-Slave1-Ubuntu-14.04** Running
- Hadoop-Slave2-Ubuntu-14.04** Running
- HBase-Slave1-Ubuntu-14.04** Powered Off
- HBase-Slave2-Ubuntu-14.04** Powered Off
- HBase-Master-Ubuntu-14.04** Powered Off
- Spark-Master-Ubuntu-14.04** Powered Off
- Spark-Slave1-Ubuntu-14.04** Powered Off
- Spark-Slave2-Ubuntu-14.04** Powered Off

CentOS 6.7

Cloudera

- cloudera-quickstart-vm-5.8.0-0...** Powered Off

General

Name: Hadoop-Master-Ubuntu-14.04
Operating System: Ubuntu (64 bit)
Groups: Ubuntu 14.04

System

Base Memory: 2048 MB
Boot Order: Floppy, CD/DVD, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging

Display

Video Memory: 12 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage

Controller: IDE
IDE Secondary Master: [CD/DVD] Empty
Controller: SATA
SATA Port 0: Hadoop-Master-Ubuntu-14.04.vdi (Normal, 8.00 GB)

Audio

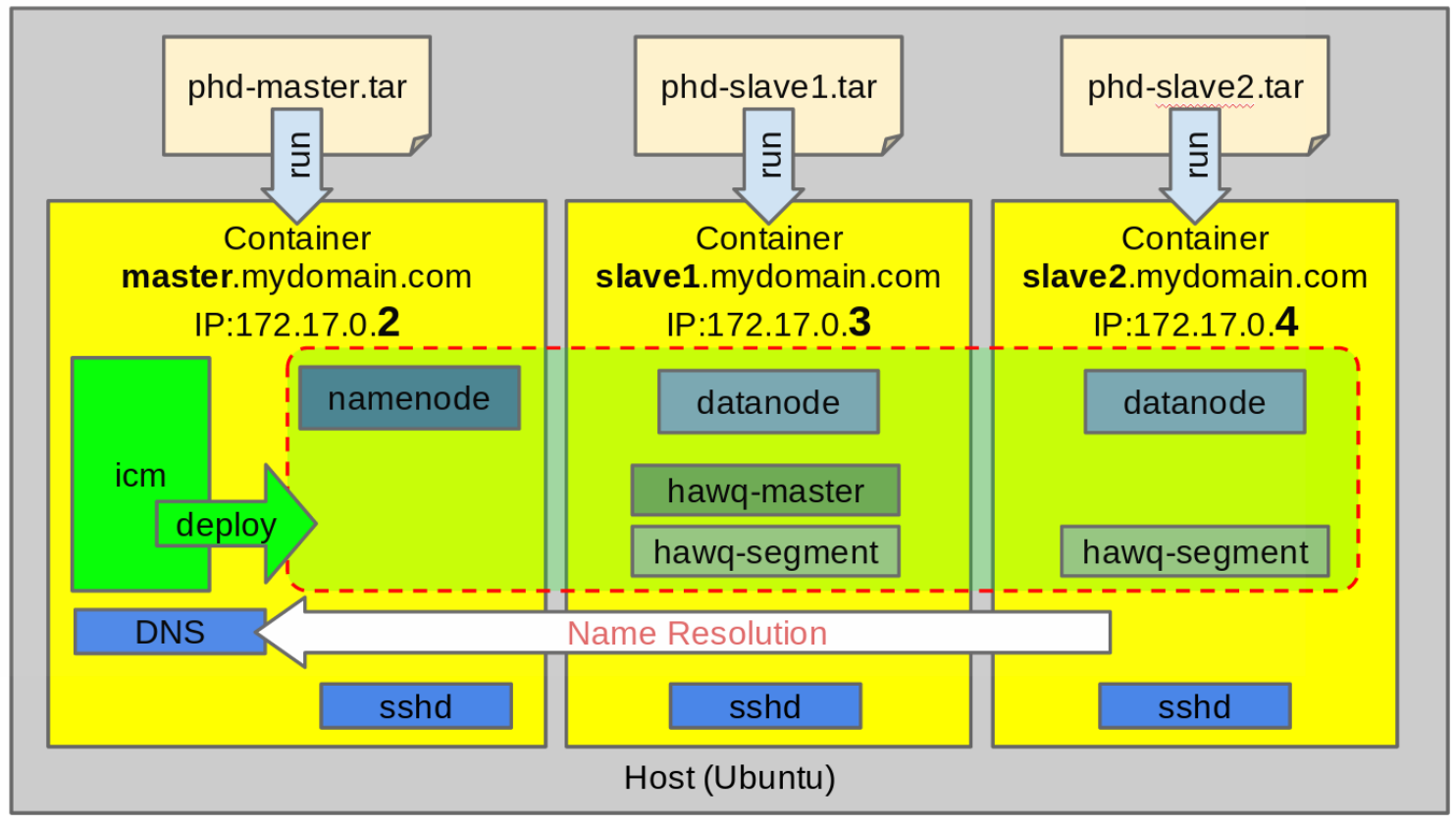
Host Driver: Windows DirectSound
Controller: ICH AC97

Network

Adapter 1: Intel PRO/1000 MT Desktop (Bridged Adapter, Realtek PCIe GBE Family Controller)

USB

Device Filters: 0 (0 active)





Hadoop configuration files

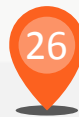


Table 9-1. Hadoop configuration files

Filename	Format	Description
<i>hadoop-env.sh</i>	Bash script	Environment variables that are used in the scripts to run Hadoop
<i>core-site.xml</i>	Hadoop configuration XML	Configuration settings for Hadoop Core, such as I/O settings that are common to HDFS and MapReduce
<i>hdfs-site.xml</i>	Hadoop configuration XML	Configuration settings for HDFS daemons: the namenode, the secondary namenode, and the datanodes
<i>mapred-site.xml</i>	Hadoop configuration XML	Configuration settings for MapReduce daemons: the jobtracker, and the tasktrackers
<i>masters</i>	Plain text	A list of machines (one per line) that each run a secondary namenode
<i>slaves</i>	Plain text	A list of machines (one per line) that each run a datanode and a task-tracker
<i>hadoop-metrics .properties</i>	Java Properties	Properties for controlling how metrics are published in Hadoop (see "Metrics" on page 352)
<i>log4j.properties</i>	Java Properties	Properties for system logfiles, the namenode audit log, and the task log for the tasktracker child process ("Hadoop Logs" on page 175)

Hadoop 1.x
Hadoop.The.Definitive.Guide.3rd.Edition

Table 10-1. Hadoop configuration files

Filename	Format	Description
<i>hadoop-env.sh</i>	Bash script	Environment variables that are used in the scripts to run Hadoop
<i>mapred-env.sh</i>	Bash script	Environment variables that are used in the scripts to run MapReduce (overrides variables set in <i>hadoop-env.sh</i>)
<i>yarn-env.sh</i>	Bash script	Environment variables that are used in the scripts to run YARN (overrides variables set in <i>hadoop-env.sh</i>)
<i>core-site.xml</i>	Hadoop configuration XML	Configuration settings for Hadoop Core, such as I/O settings that are common to HDFS, MapReduce, and YARN
<i>hdfs-site.xml</i>	Hadoop configuration XML	Configuration settings for HDFS daemons: the namenode, the secondary namenode, and the datanodes
<i>mapred-site.xml</i>	Hadoop configuration XML	Configuration settings for MapReduce daemons: the job history server
<i>yarn-site.xml</i>	Hadoop configuration XML	Configuration settings for YARN daemons: the resource manager, the web app proxy server, and the node managers
<i>slaves</i>	Plain text	A list of machines (one per line) that each run a datanode and a node manager
<i>hadoop-metrics2 .properties</i>	Java properties	Properties for controlling how metrics are published in Hadoop (see Metrics and JMX)
<i>log4j.properties</i>	Java properties	Properties for system logfiles, the namenode audit log, and the task log for the task JVM process (Hadoop Logs)
<i>hadoop-policy.xml</i>	Hadoop configuration XML	Configuration settings for access control lists when running Hadoop in secure mode

Hadoop 2.x
Hadoop.The.Definitive.Guide.4th.Edition



PART5

HDFS & Hadoop Command

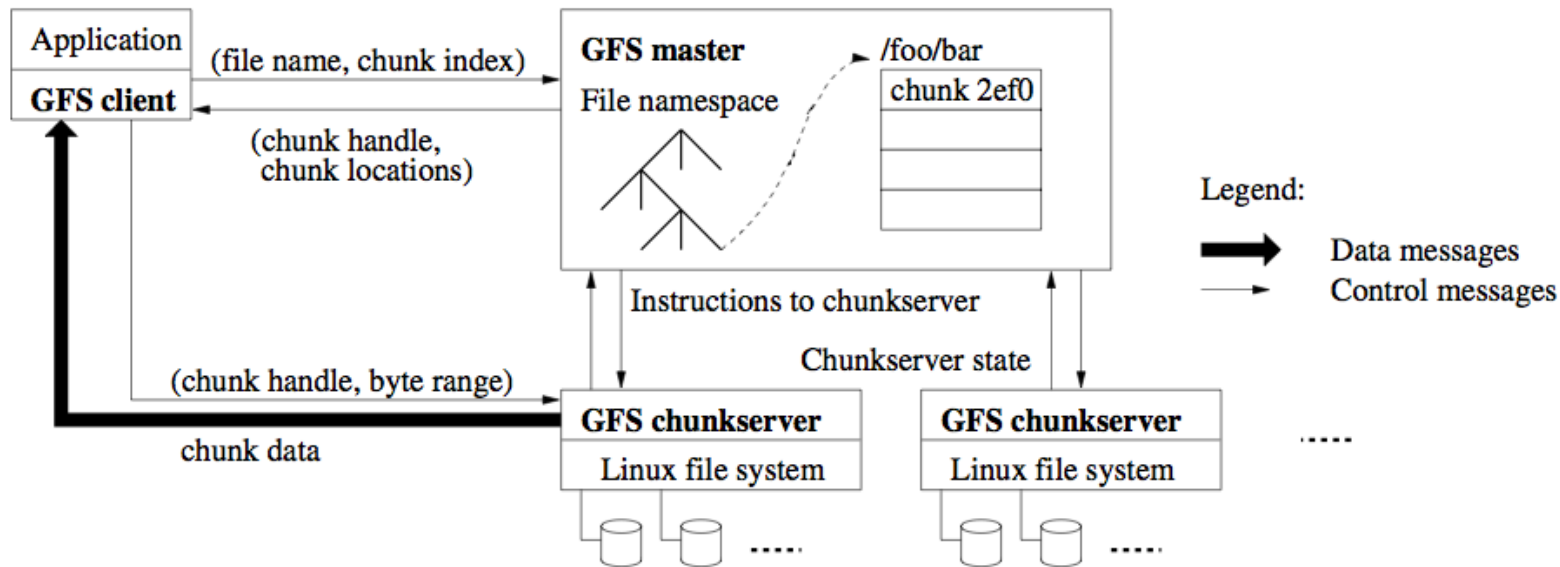


Figure 1: GFS Architecture

The Google File System

<https://static.googleusercontent.com/media/research.google.com/ja//archive/gfs-sosp2003.pdf>

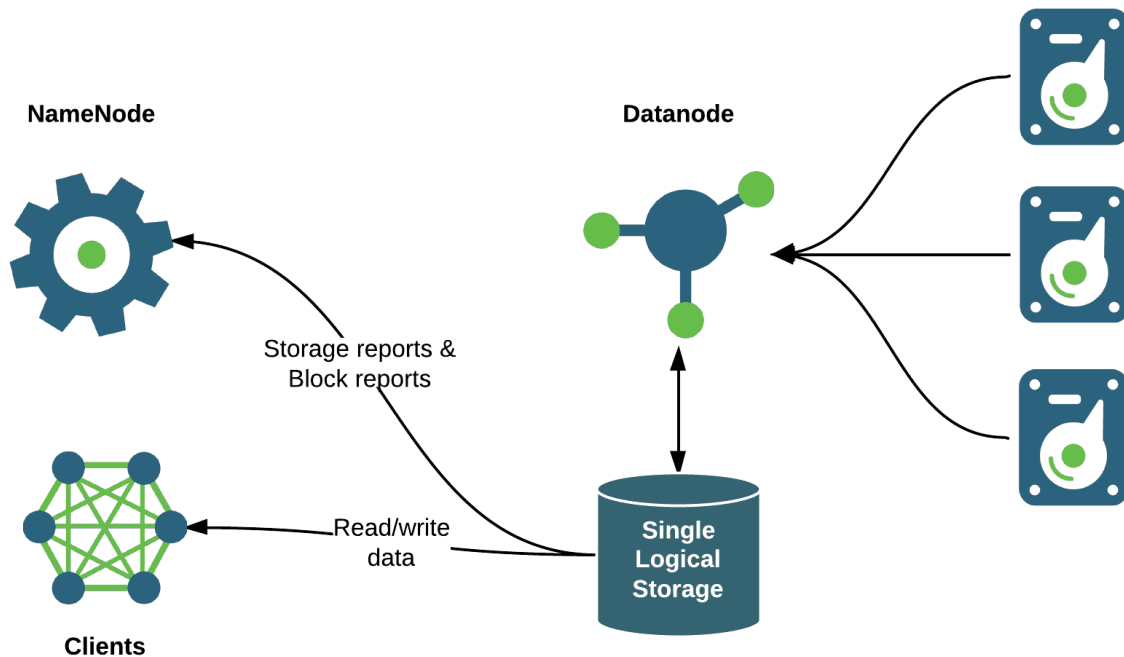


Figure 1: A DataNode presented itself as a single logical storage

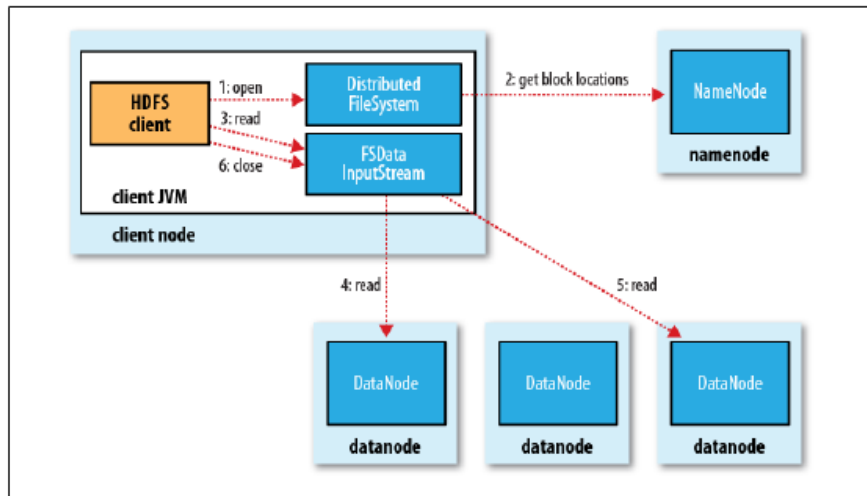


Figure 3-2. A client reading data from HDFS

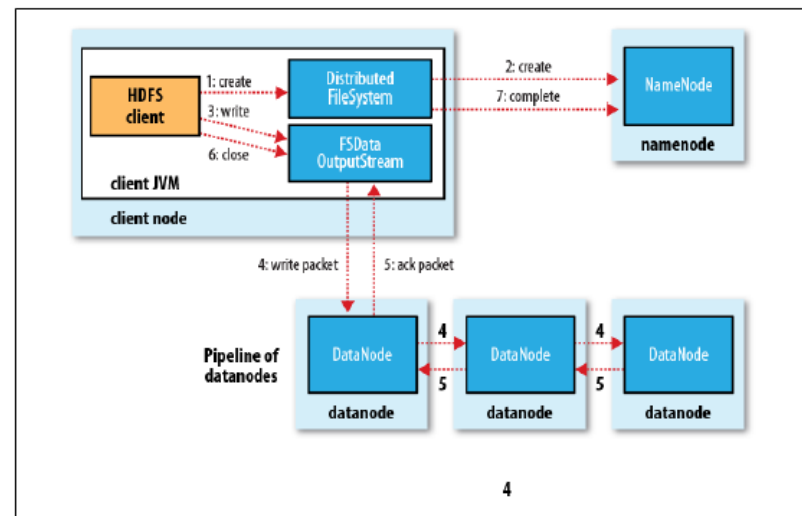


Figure 3-4. A client writing data to HDFS

```
hduser@kannandreams: /usr/local/hadoop/bin

Usage: hadoop [--config confdir] COMMAND
where COMMAND is one of:
  namenode -format      format the DFS filesystem
  secondarynamenode    run the DFS secondary namenode
  namenode              run the DFS namenode
  datanode              run a DFS datanode
  dfsadmin              run a DFS admin client
  mradmin               run a Map-Reduce admin client
  fsck                  run a DFS filesystem checking utility
  fs                    run a generic filesystem user client
  balancer              run a cluster balancing utility
  oiv                   apply the offline fsimage viewer to an fsimage
  fetchdt               fetch a delegation token from the NameNode
  jobtracker            run the MapReduce job Tracker node
  pipes                 run a Pipes job
  tasktracker           run a MapReduce task Tracker node
  historyserver         run job history servers as a standalone daemon
  job                   manipulate MapReduce jobs
  queue                 get information regarding JobQueues
  version               print the version
  jar <jar>             run a jar file
  distcp <srcurl> <desturl> copy file or directories recursively
  distcp2 <srcurl> <desturl> DistCp version 2
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
  classpath             prints the class path needed to get the
                        Hadoop jar and the required libraries
  daemonlog             get/set the log level for each daemon
or
  CLASSNAME             run the class named CLASSNAME
Most commands print help when invoked w/o parameters.
hduser@kannandreams: /usr/local/hadoop/bin$
```

Official Website :

<https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/FileSystemShell.html>

https://hadoop.apache.org/docs/r1.0.4/cn/hdfs_shell.html

HADOOP



```
export JAVA_HOME=/usr/java/jdk1.6.0_26/
for service in /etc/init.d/hadoop-0.20-*
do
    sudo $service start
done
```

STARTING THE PROCESSES

Hadoop NameNode
(http://<hostname>:50070)

Hadoop JobTracker
(http://<hostname>:50030)

MONITORING PAGES

```
for service in /etc/init.d/hadoop-0.20-*
do
    sudo $service stop
done
```

STOPPING THE PROCESSES

HADOOP HDFS



```
hadoop fs -ls <path>
```

LIST FILES

```
hadoop fs -mkdir <path>
```

MAKE DIRECTORY

```
hadoop fs -rmdir <path>
```

REMOVE DIRECTORY

```
hadoop fs -put <local_file>
<hdfs_path>
```

LOAD FILE

```
hadoop fs -rm <file>
```

REMOVE FILE

```
hadoop fs -cat <file>
hadoop fs -tail <file>
```

VIEW FILE

```
hadoop fs -getmerge <hdfs_
directory> <local_output_file>
```

MERGE MULTIPLE PART FILES

All Applications

Namenode Information

localhost:50070/dfshealth.html

Hadoop

Overview

Datanodes

Snapshot

Startup Progress

Utilities -

Overview

'localhost:9000' (active)

Started:	Sun Apr 06 15:52:11 IST 2014
Version:	2.3.0, r1567123
Compiled:	2014-02-11T13:40Z by jenkins from branch-2.3.0
Cluster ID:	CID-5edbd0da-c69f-425b-bbc7-a662ac5d45dc
Block Pool ID:	BP-1127675761-127.0.1.1-1396692597591

Summary

Security is off.

Safemode is off.

35 files and directories, 17 blocks = 52 total filesystem object(s).

Heap Memory used 34.01 MB of 88.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 40.17 MB of 40.69 MB Committed Non Heap Memory. Max Non Heap Memo

Configured Capacity:

Browsing HDFS

localhost:50070/explorer.html

Browse Directory

/

Go!

Permission	Owner	Group	Size	Replication	Block Size	Name
drwxr-xr-x	siva	supergroup	0 B	0	0 B	siva
drwxr-xr-x	siva	supergroup	0 B	0	0 B	test
drwx-----	siva	supergroup	0 B	0	0 B	tmp
drwxr-xr-x	siva	supergroup	0 B	0	0 B	user

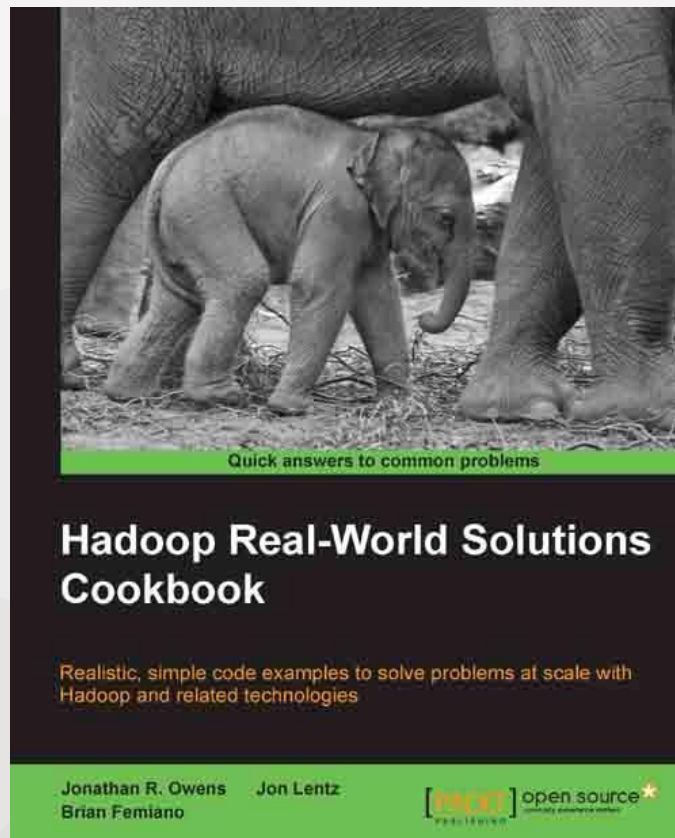
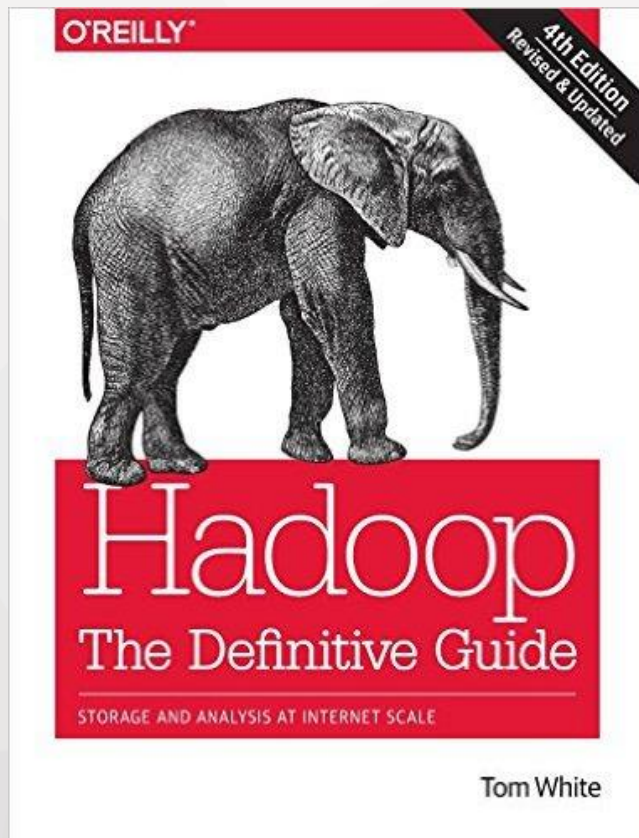
Hadoop, 2013.



PART6



Reference Books



A stylized illustration of a computer monitor. The monitor has a dark blue frame and a white screen. On the screen, the words "The End" are written in a bold, dark blue, sans-serif font. The monitor is supported by a dark blue stand. The background is a light gray with a pattern of overlapping, irregular polygons.

The End