

CONTENTS

Preliminary Topics
事前準備

01

Hadoop Eco System

02

Hadoop Architecture

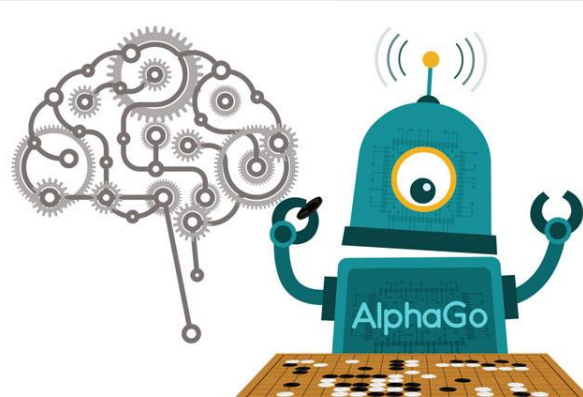
03

Hadoop Environment
実験環境

04

05 Hadoop Command

06 Reference Books



Forgetting Curve



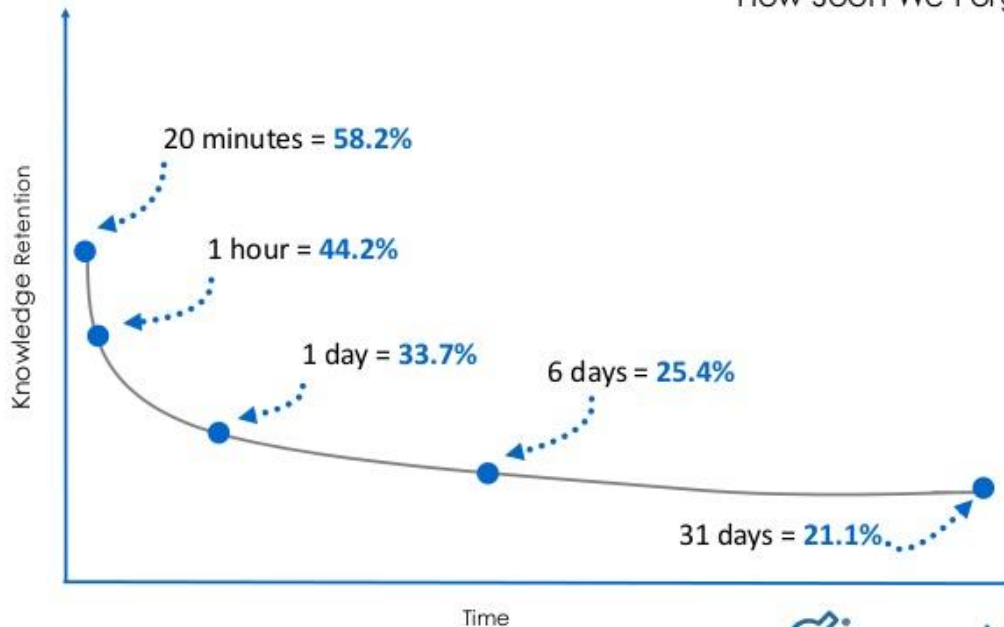
Forgetting Curve

Hermann Ebbinghaus

4

Ebbinghaus Curve

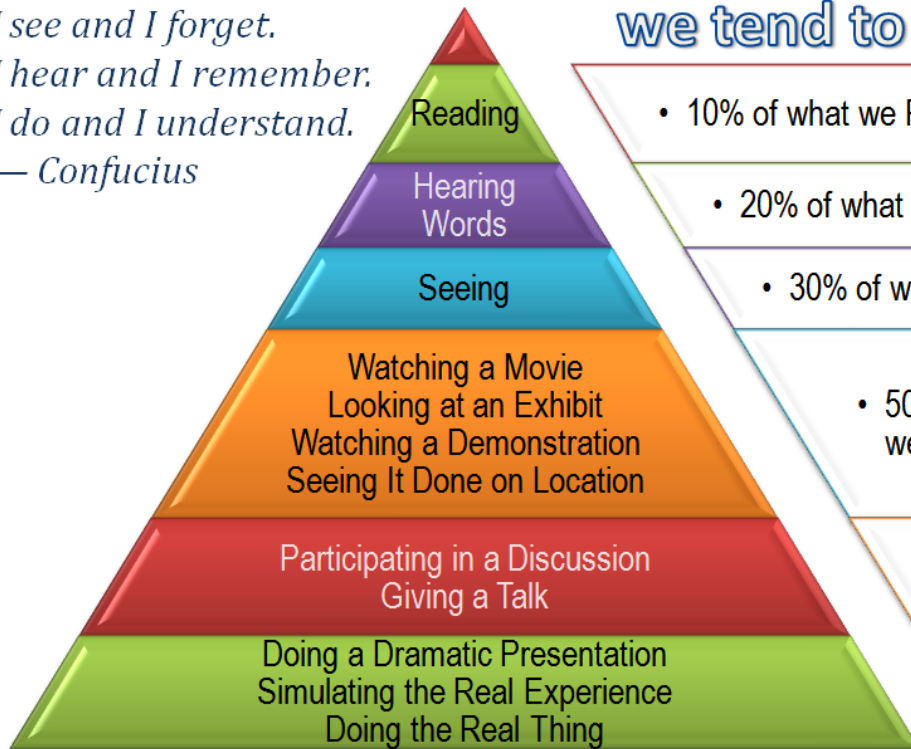
How Soon We Forget



Source: Purdue University Graph

The Cone Of Learning

*I see and I forget.
I hear and I remember.
I do and I understand.
— Confucius*



After 2 weeks,
we tend to remember ...

- 10% of what we READ
- 20% of what we HEAR
- 30% of what we SEE
- 50% of what we SEE & HEAR
- 70% of what we SAY
- 90% of what we SAY & DO

P
a
s
s
i
v
e

A
c
t
i
v
e

Source: Edgar Dale (1969)

PART 1



Preliminary Topics
事前準備



Java SE Public Updates			
Major Release	GA Date	End of Public Updates Notification	End of Public Updates
5.0	May 2004	Apr 2008	Oct 2009
6	Dec 2006	Feb 2011	Feb 2013
7	Jul 2011	Mar 2014	Apr 2015
8	Mar 2014	TBD	Sep 2017*

* or later, depending on factors described above.

<http://www.oracle.com/technetwork/java/eol-135779.html>





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

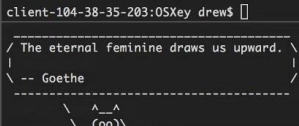
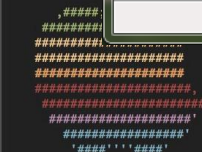


Official Website :
<https://www.centos.org/>
<https://www.centos.org/download/>
<https://wiki.centos.org/Download>



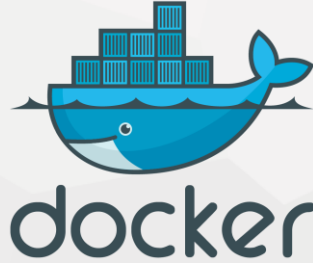
Official Website :
<https://www.redhat.com/en>

10





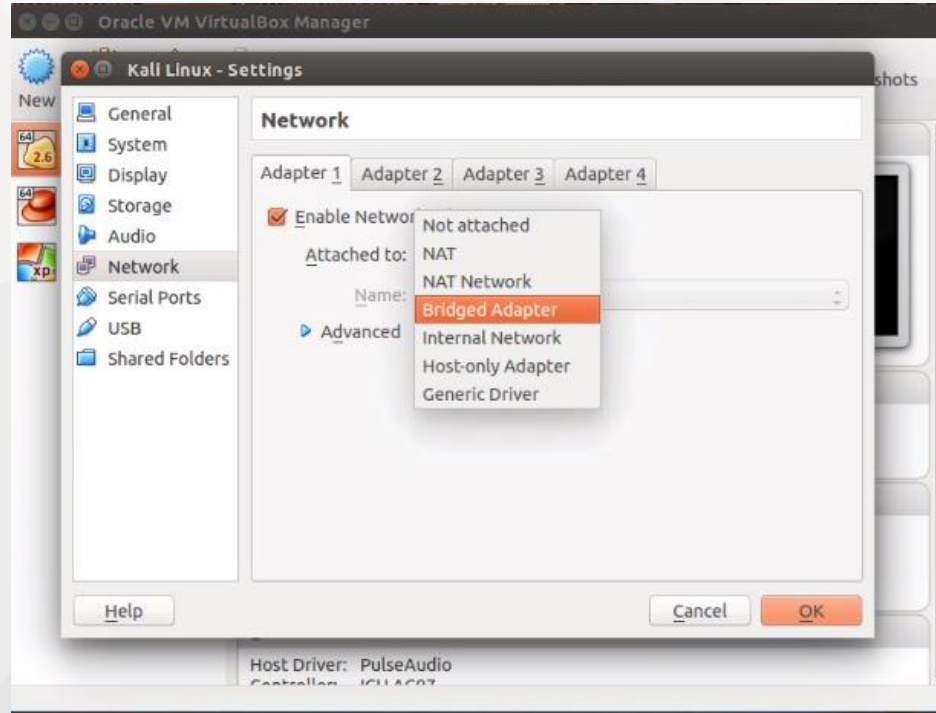
VirtualBox



CITRIX[®]

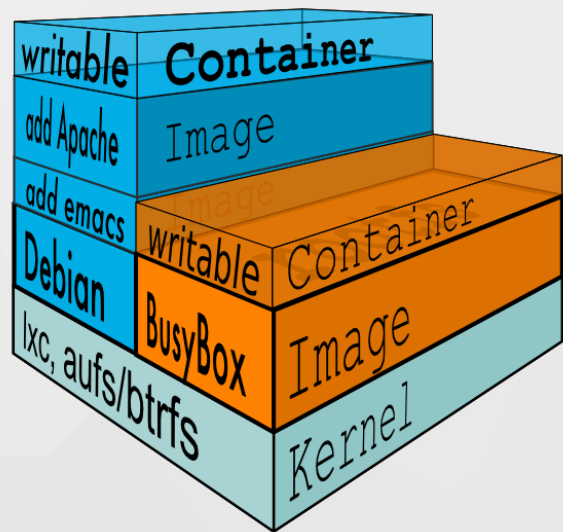
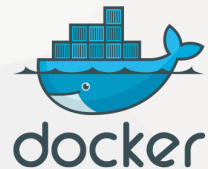


Virtualization with
KVM on **Linux[™]**






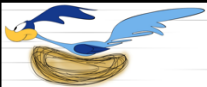


Official Website :
<https://www.virtualbox.org/>
<https://www.virtualbox.org/wiki/Downloads>

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


 Watch ▾
13


Mirror of Apache Hadoop common

 8,109 commits


 104 branches

 174 releases


 13 contributors







branch: trunk ▾

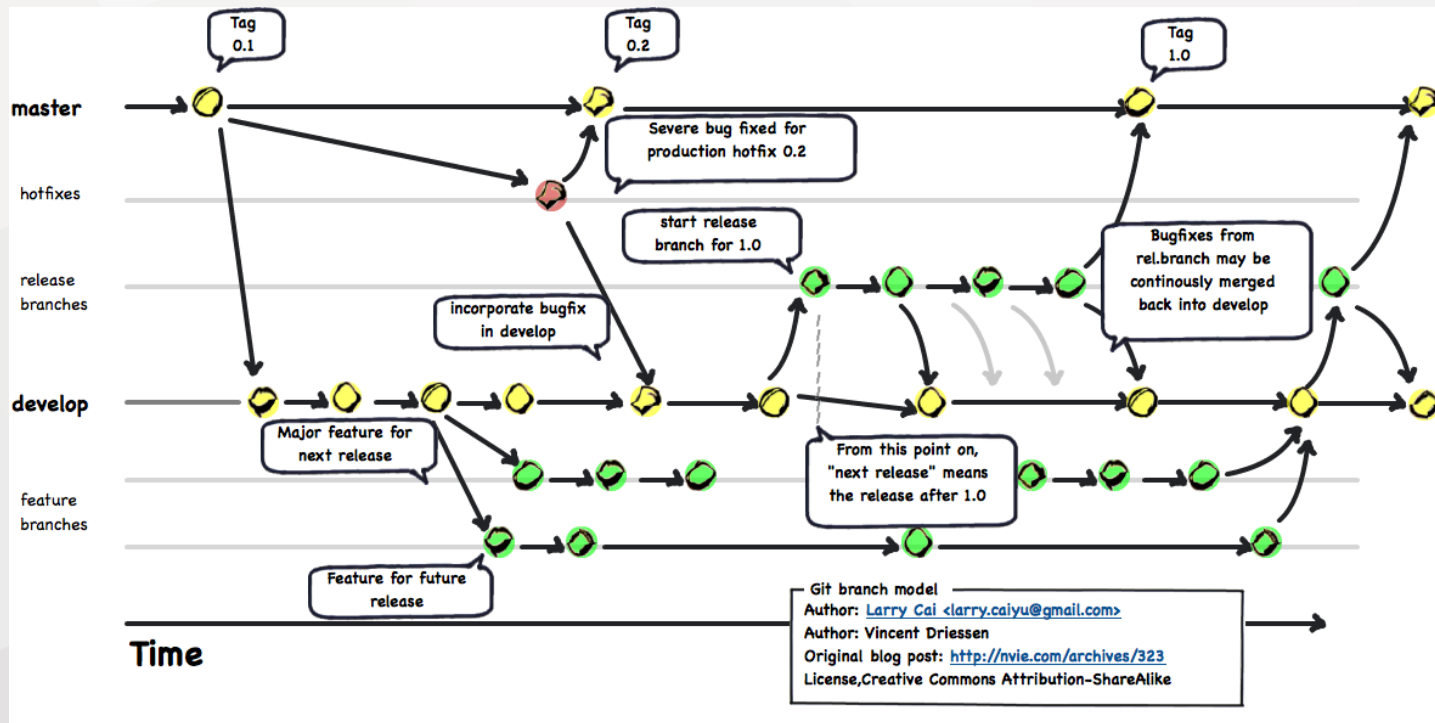
hadoop-common / 

Move HDFS-5276 to 2.3.0 in CHANGES.txt ...

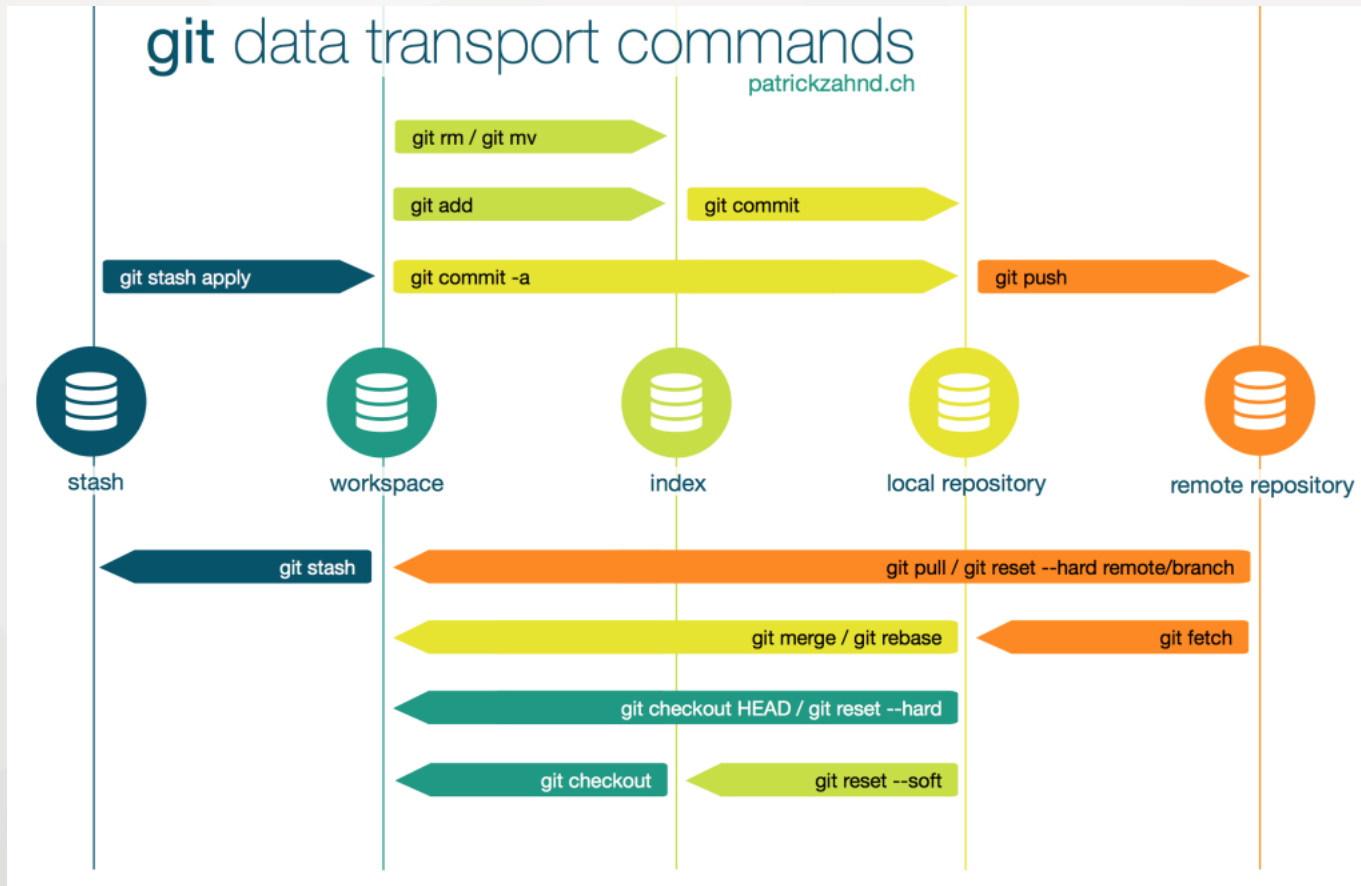
 Luke Lu authored 4 hours ago

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



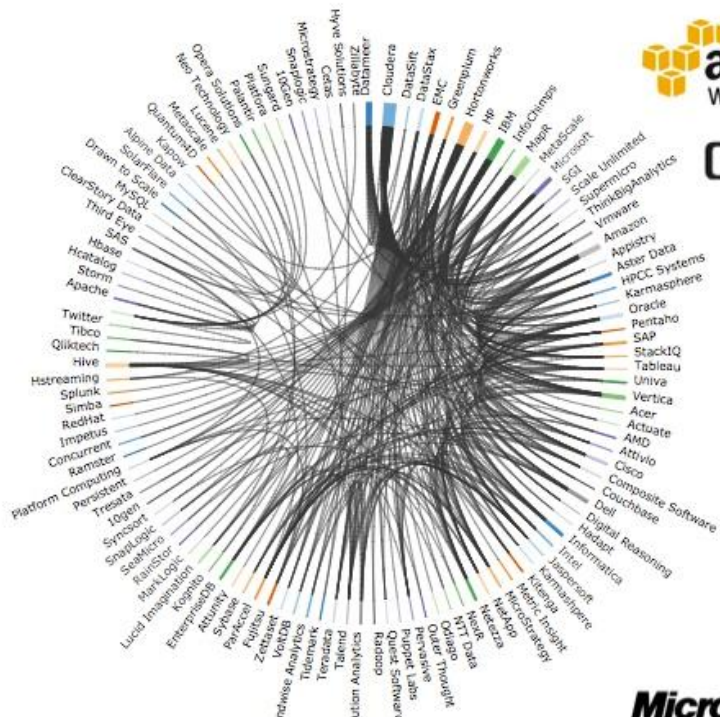
Official Website :
<https://git-scm.com/>
 The latest stable Release : v2.11.1





PART2

Hadoop Eco System



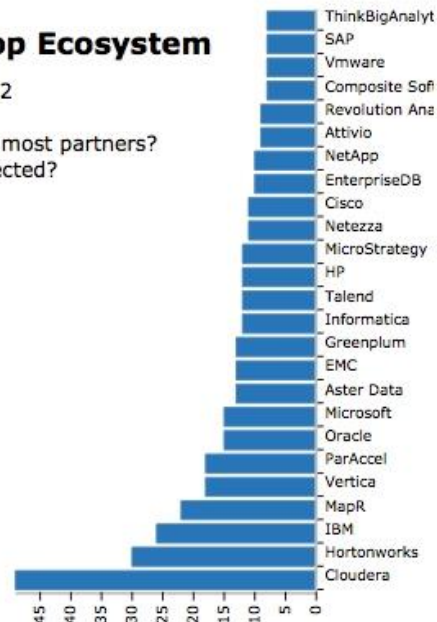
cloudera



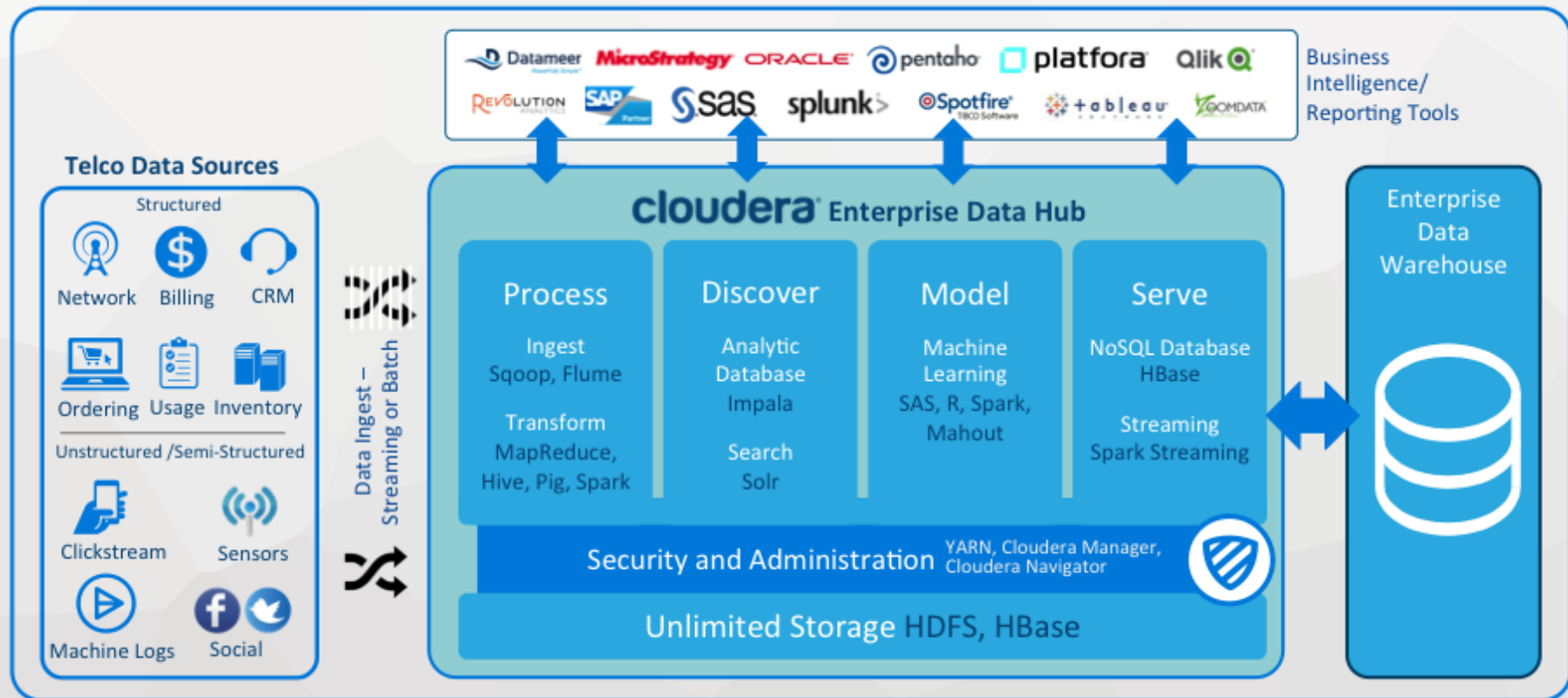
The Hadoop Ecosystem

June 21, 2012

























Who has the most partners?
Who is connected?



brought to you by Datameer
Powerfully Simple™

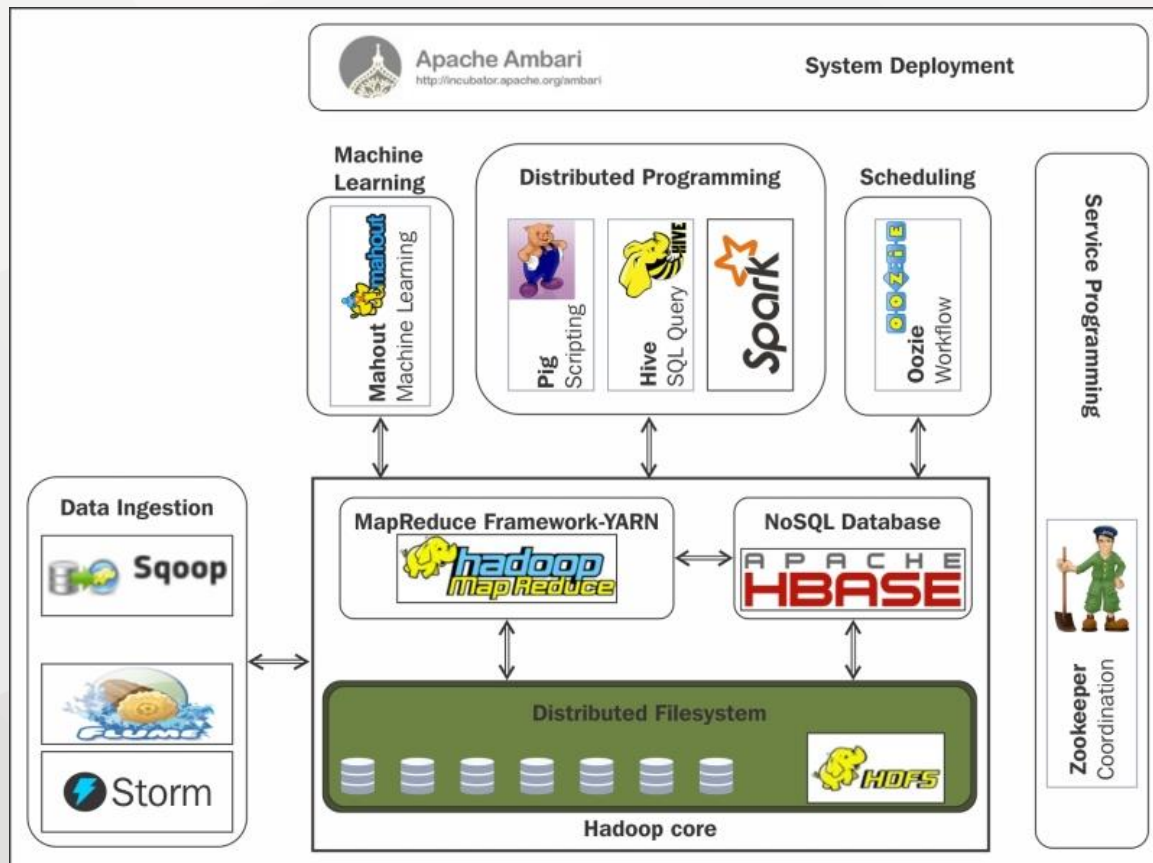


Hadoop Eco System

VISUALIZATION LAYER AND API'S				RESTful API
ANALYTICAL TOOLS				
ANALYSIS LAYER			 Lightning-Fast Cluster Computing	 HIVE 
PROCESSING LAYER		 Motor de Eventos	 Motor de Workflow	
STORAGE LAYER				
REAL TIME EVENT ENGINE		 Motor de Eventos		
INTEGRATION LAYER (ETL)				
SOURCES				

CAPA DE GESTIÓN	
HERRAMIENTAS ADMINISTRACIÓN	
HERRAMIENTAS MONITORIZACIÓN	 
HERRAMIENTAS DIAGNÓSTICO	 

Hadoop Eco System

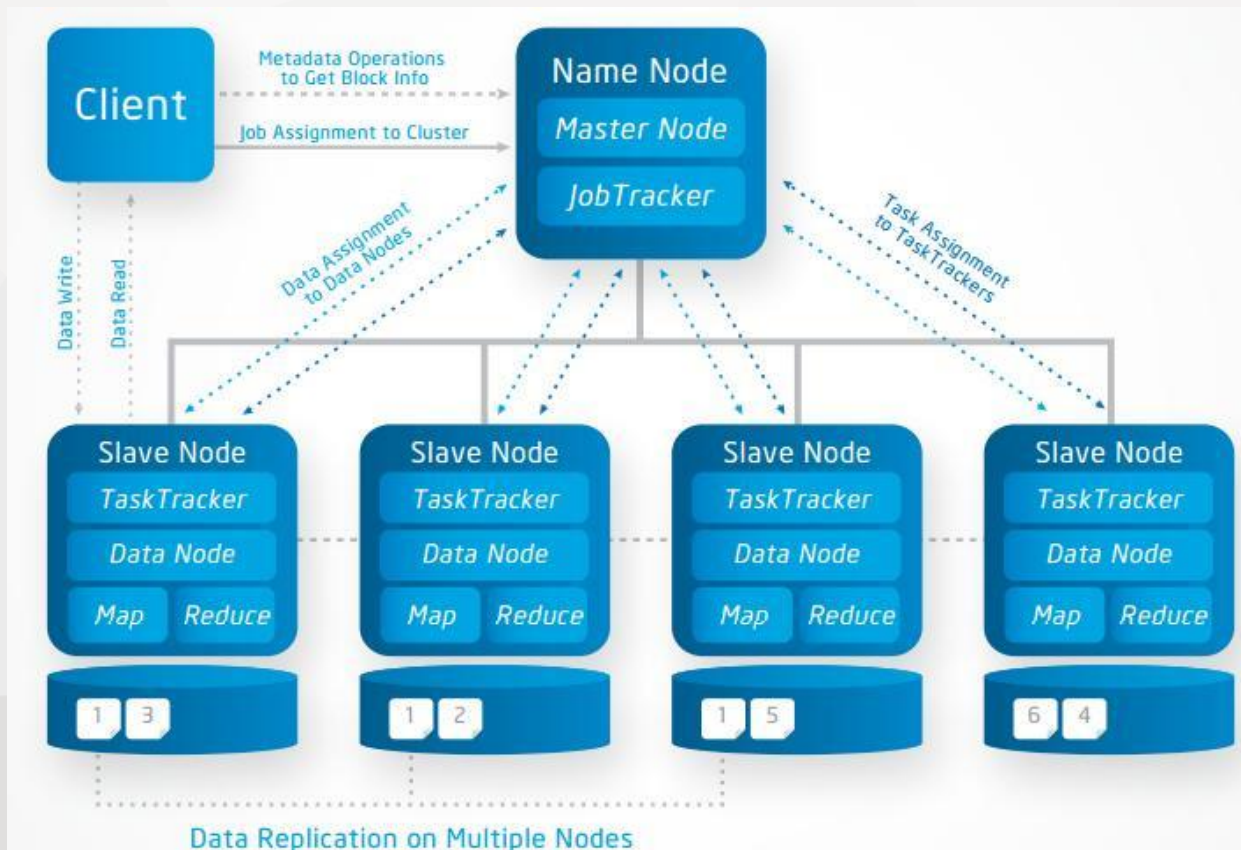




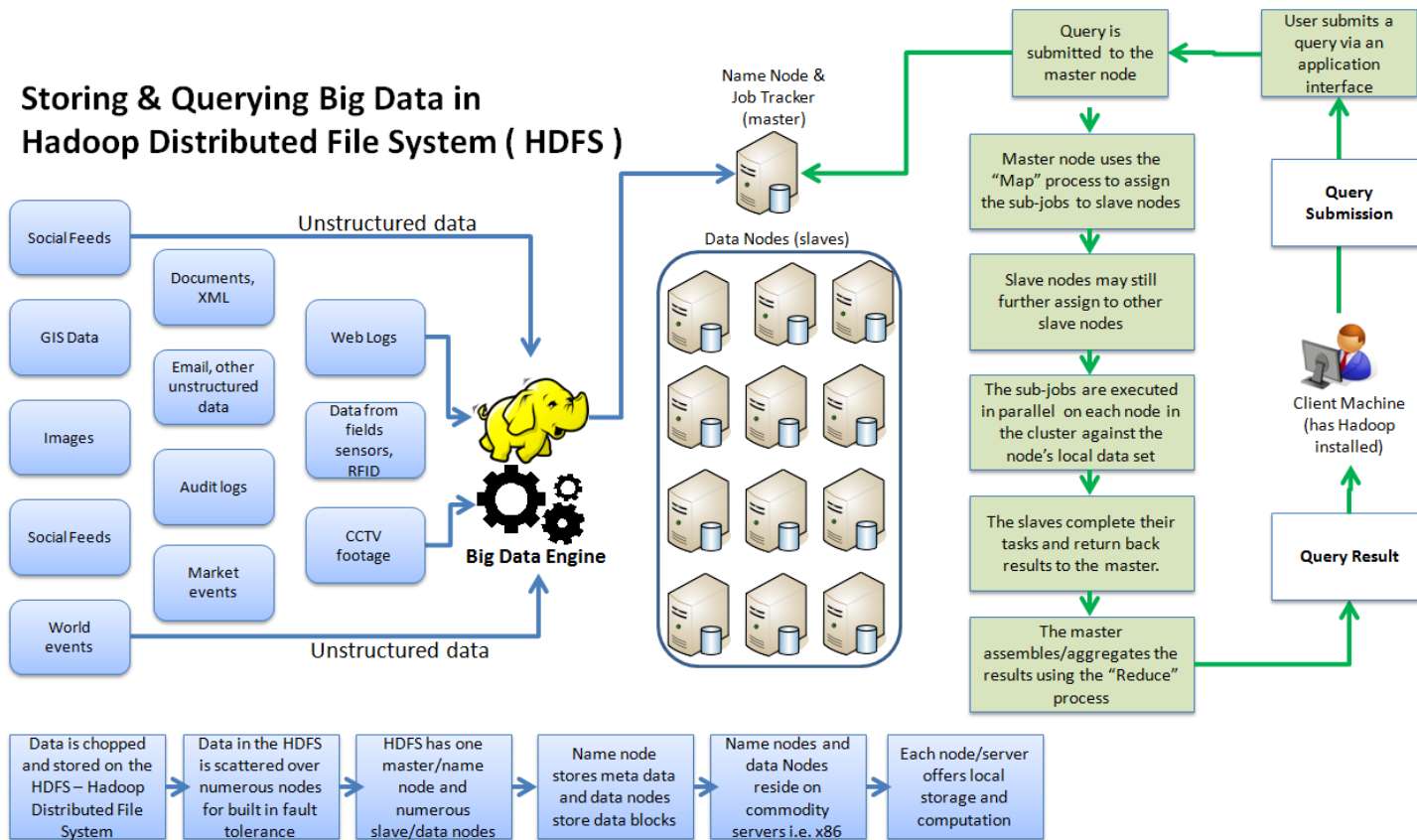
PART3

Hadoop Architecture

Architecture

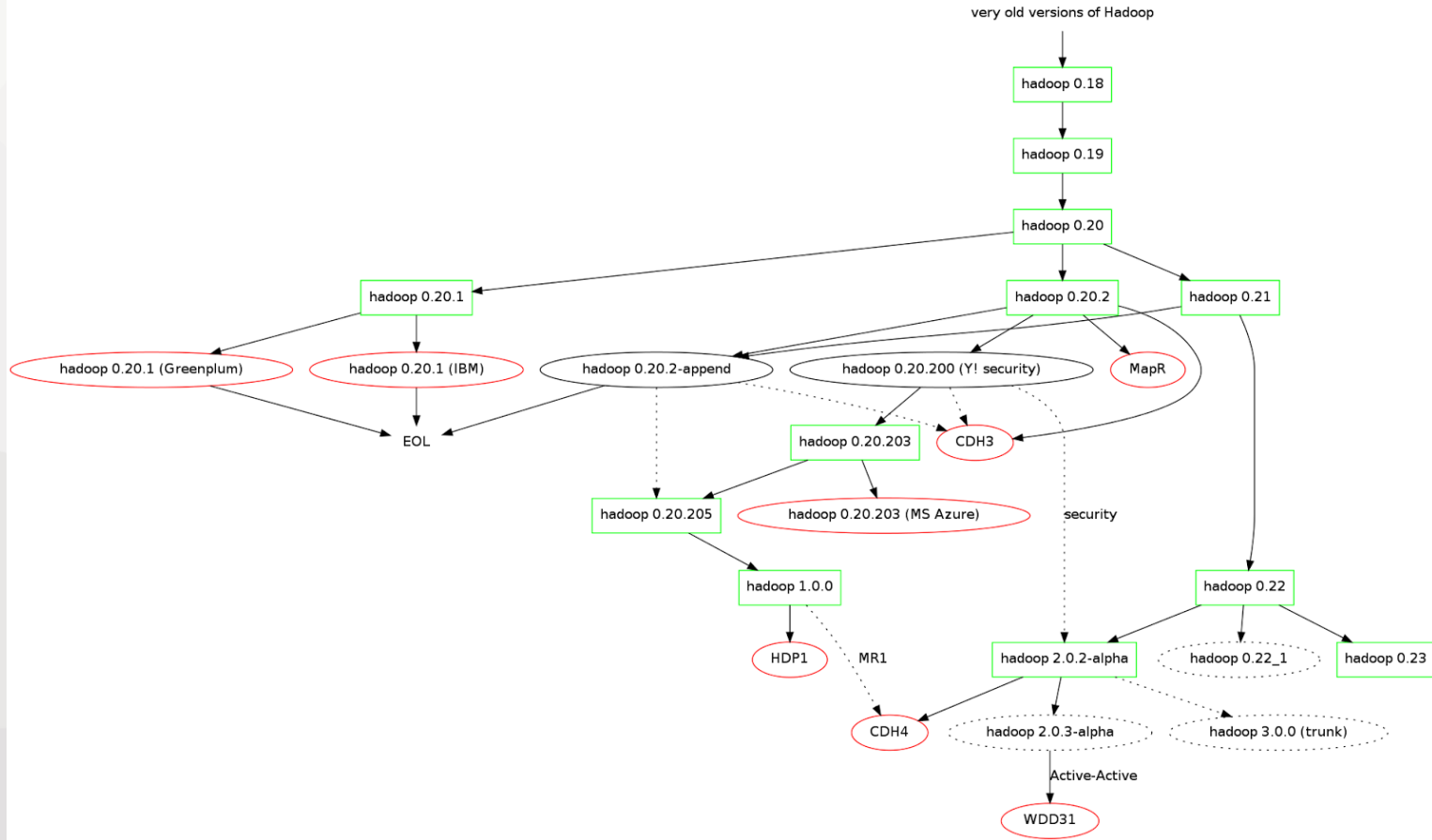


Storing & Querying Big Data in Hadoop Distributed File System (HDFS)



Apache Hadoop Version

25



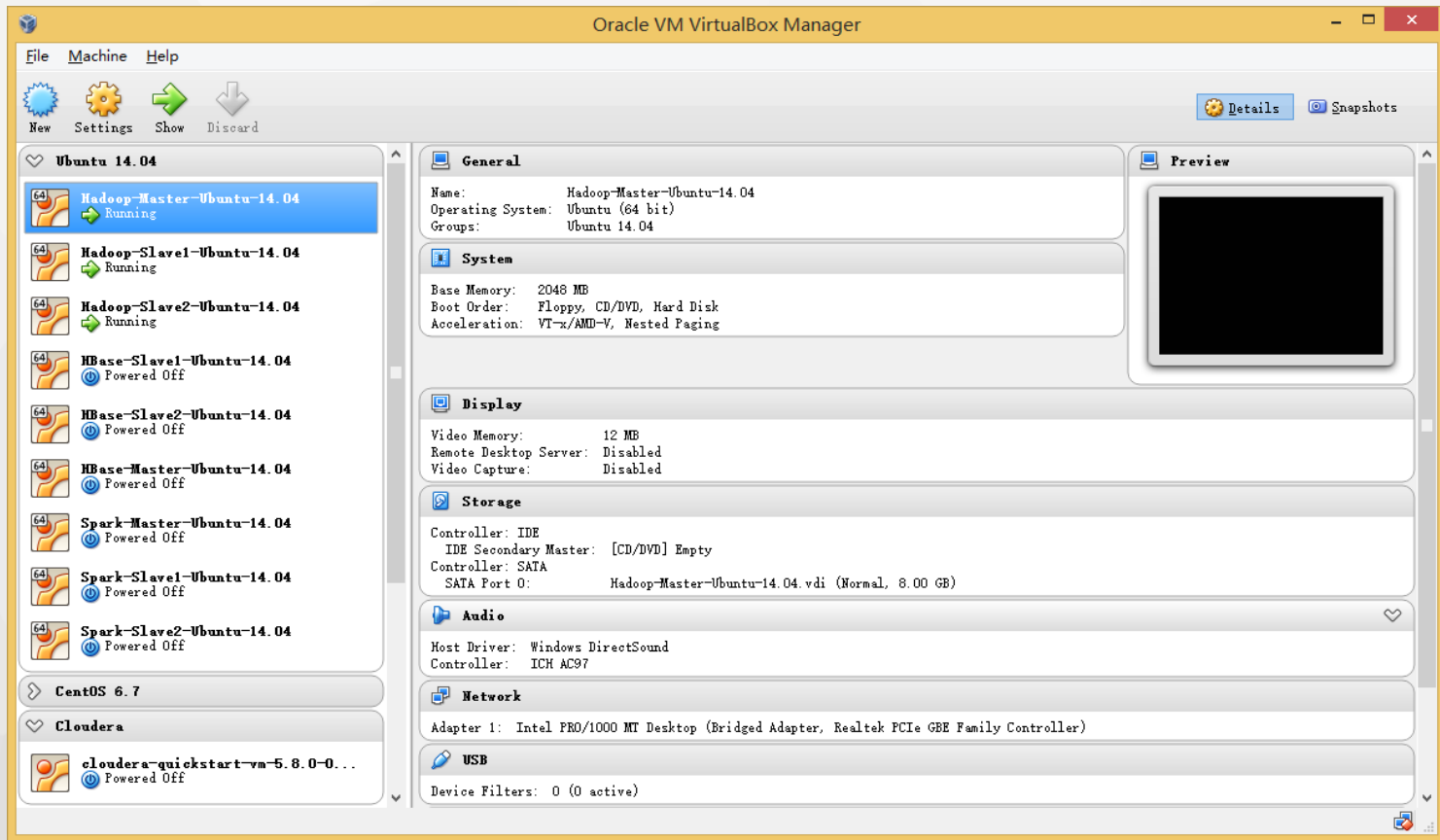


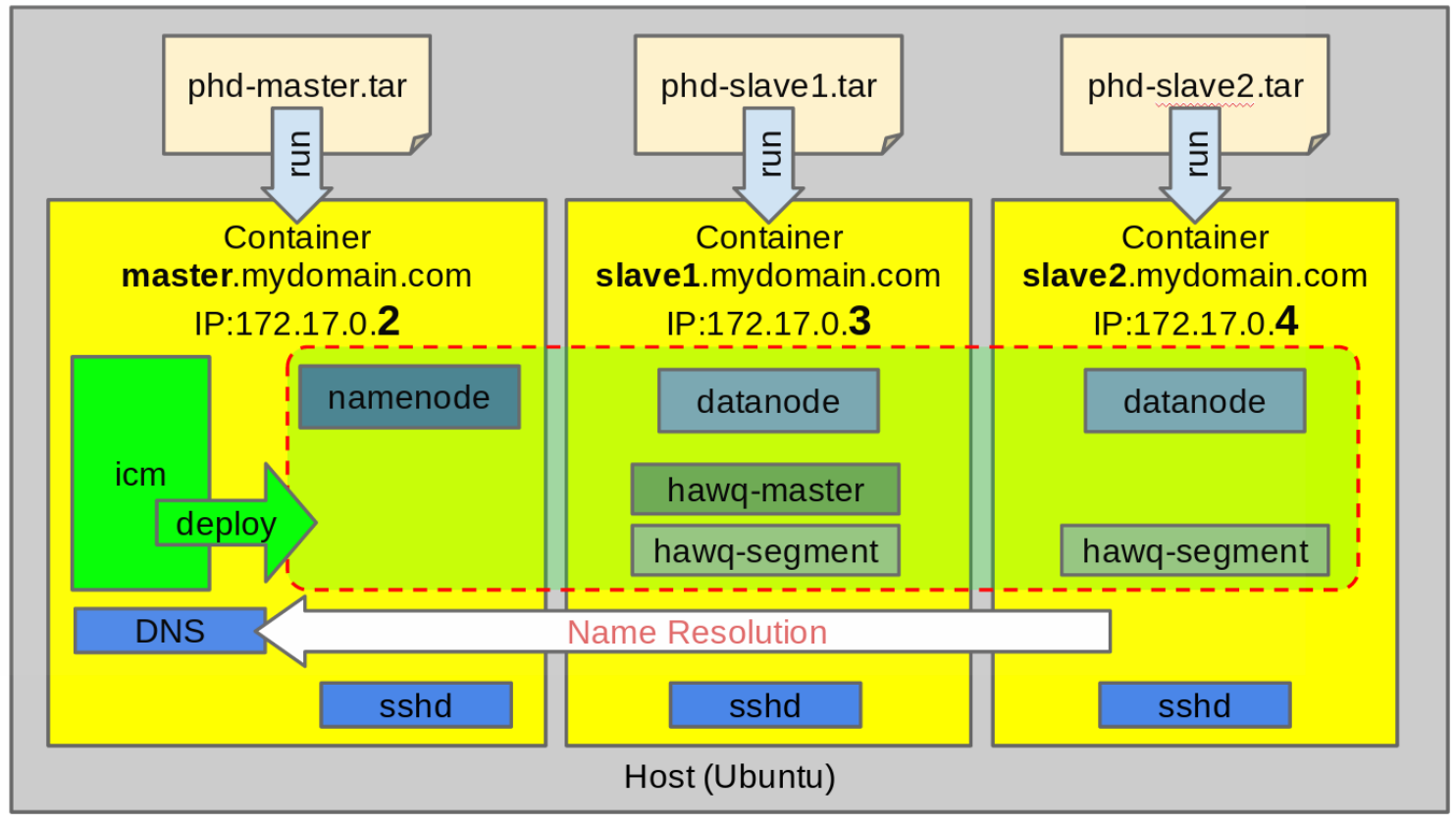
PART4

Hadoop Environment

Oracle Virtual Box

27







Hadoop configuration files

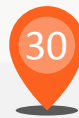


Table 9-1. Hadoop configuration files

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

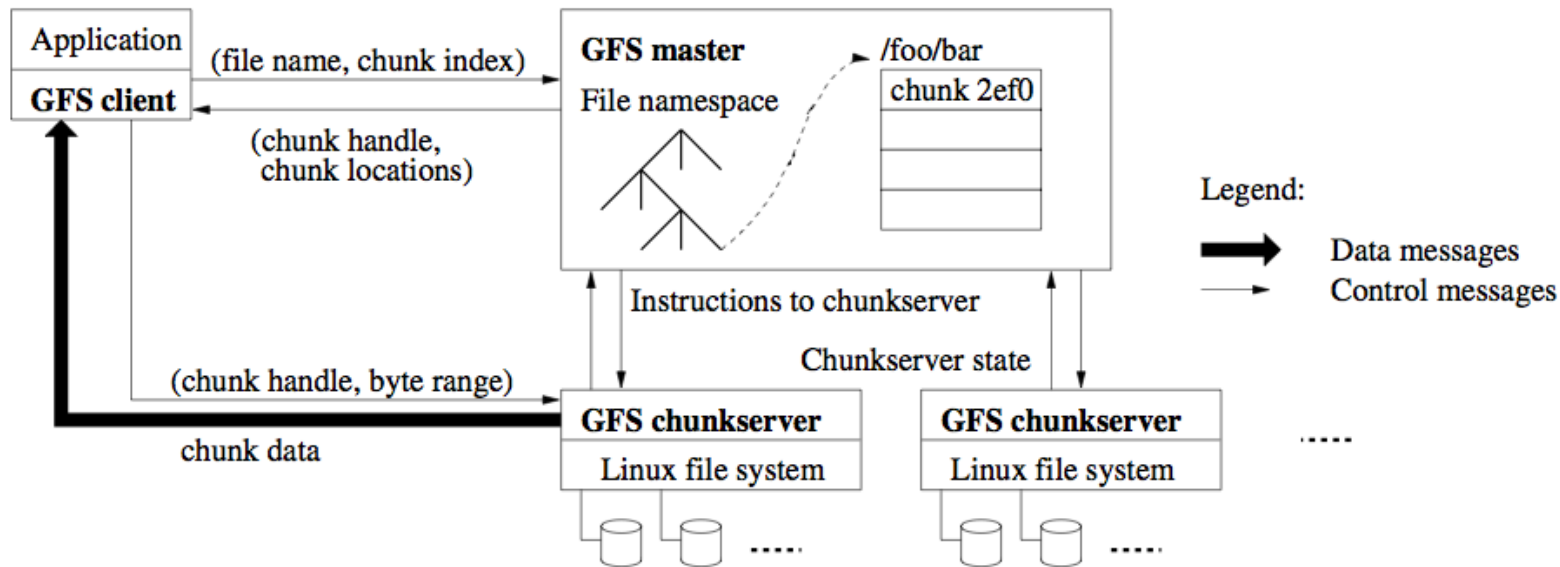


Figure 1: GFS Architecture

The Google File System

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

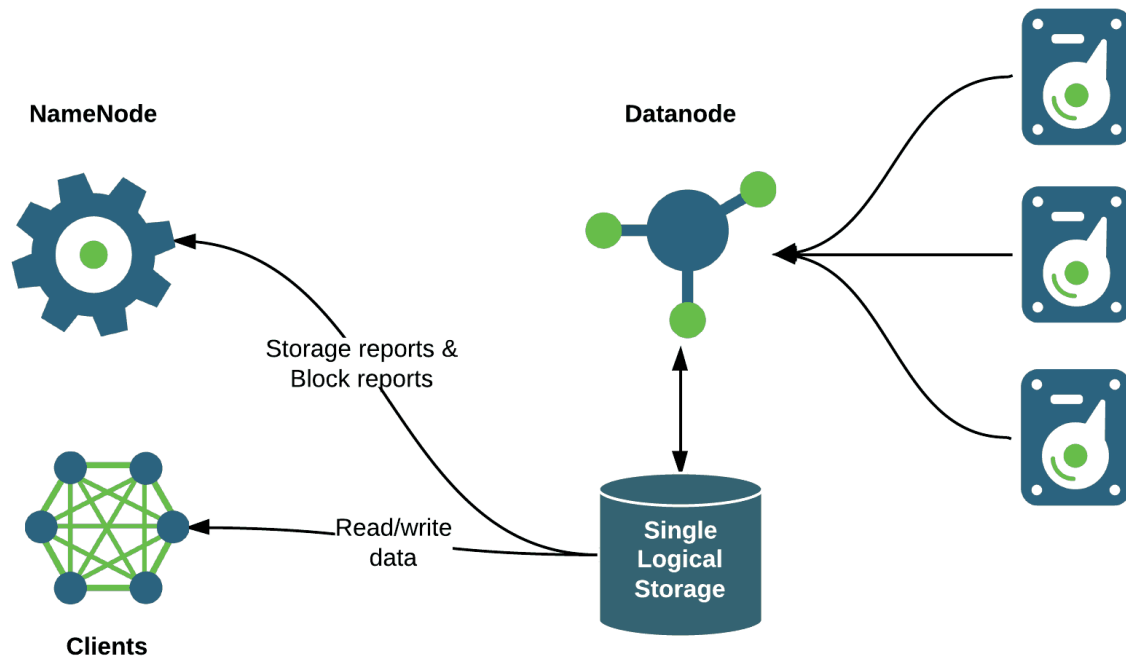


Figure 1: A DataNode presented itself as a single logical storage

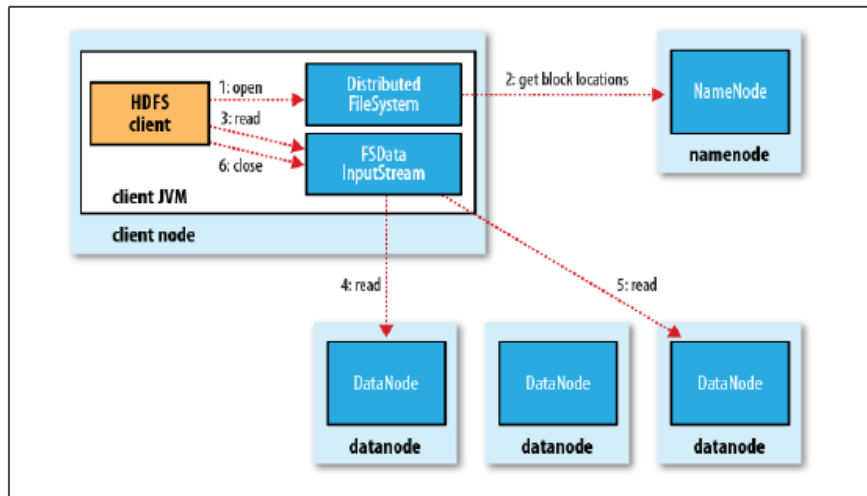


Figure 3-2. A client reading data from HDFS

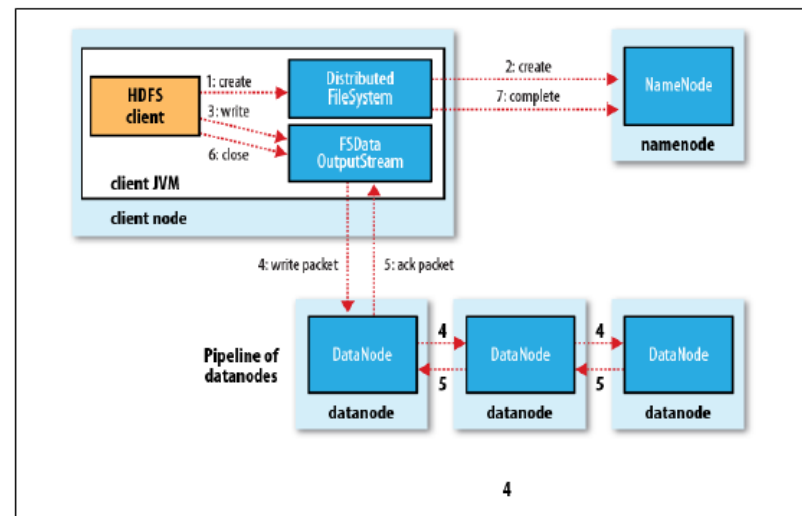


Figure 3-4. A client writing data to HDFS

```
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

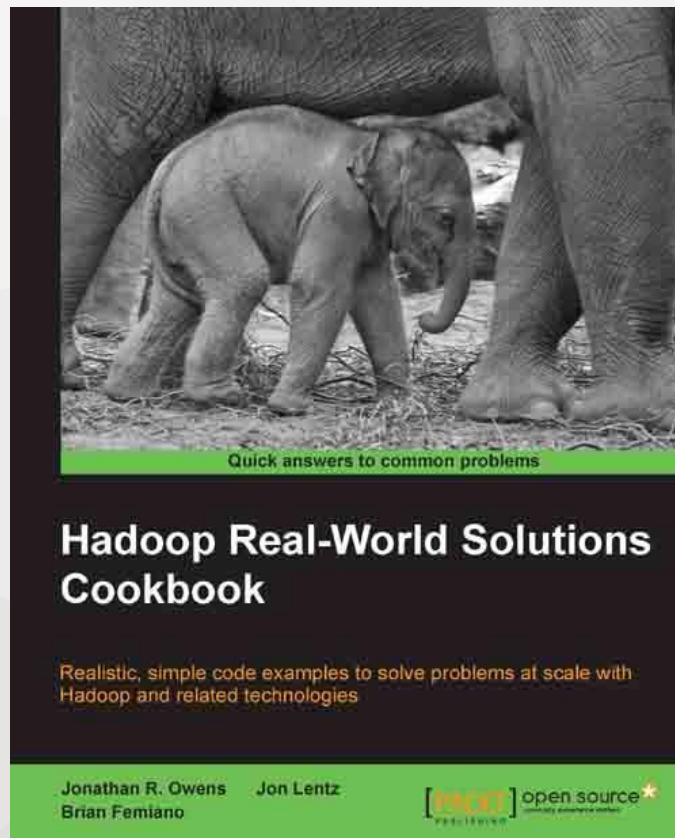
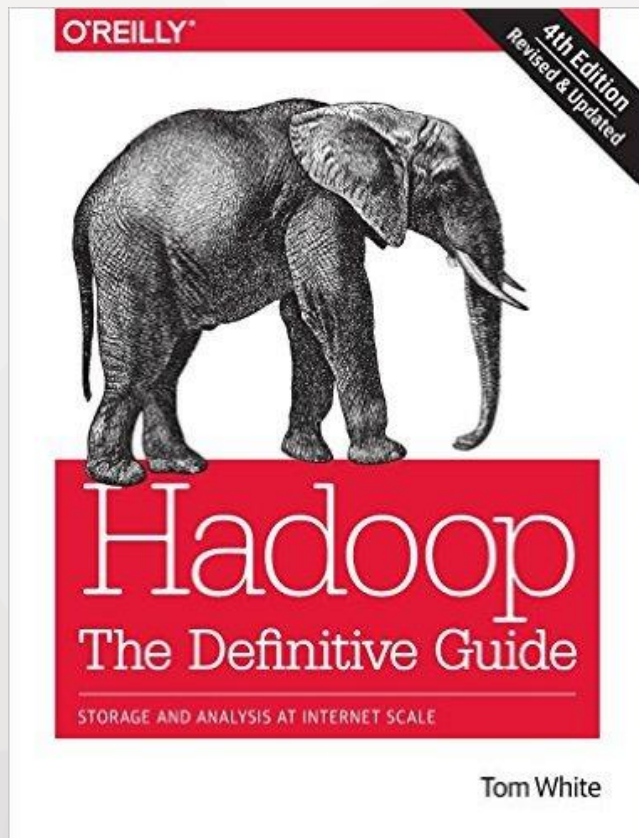
Hadoop, 2013.



PART6



Reference Books





HomeWork



The End