The Pilot Of Hadoop

2017.2   XenRon

📊 CONTENTS 📊

PART1

Preliminary Topics
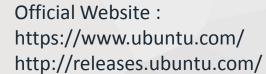**事前準備**

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

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

SSH Terminal

Virtual Box

Official Website :
https://www.virtualbox.org/
https://www.virtualbox.org/wiki/Downloads
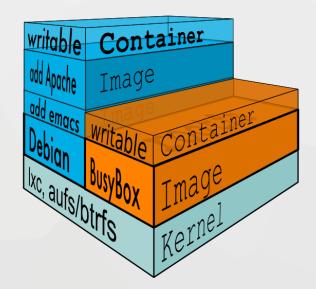
Virtual networking :
https://www.virtualbox.org/manual/ch06.html
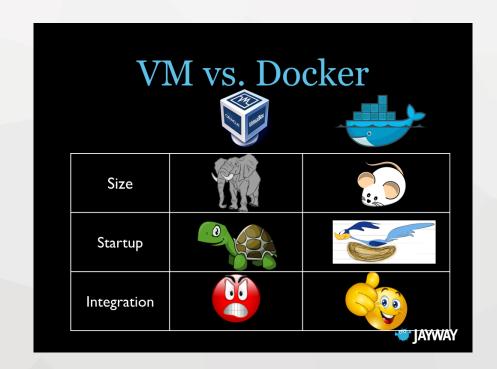
VM vs. Docker

Size		
Startup		
Integration		

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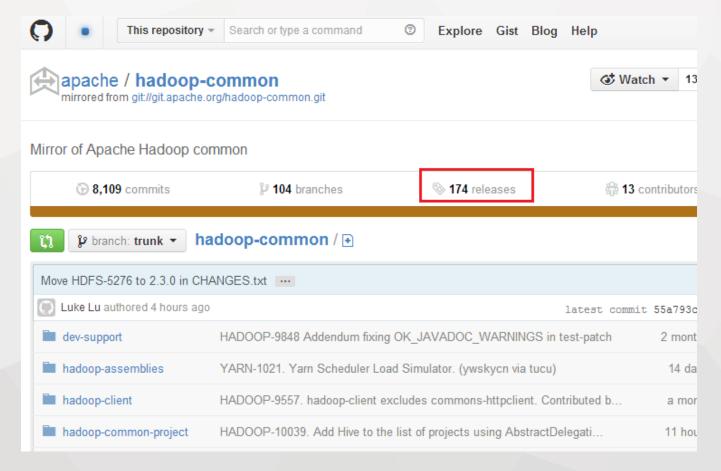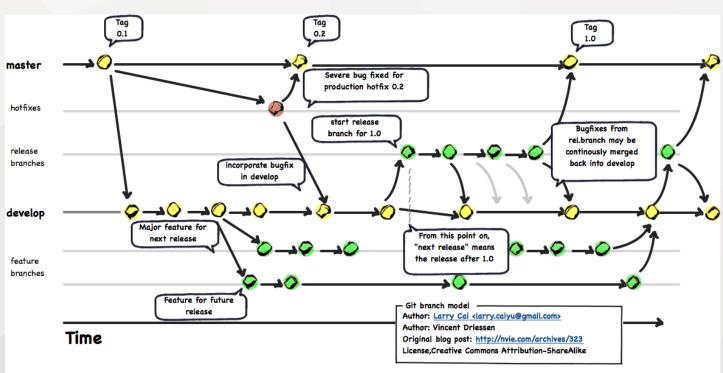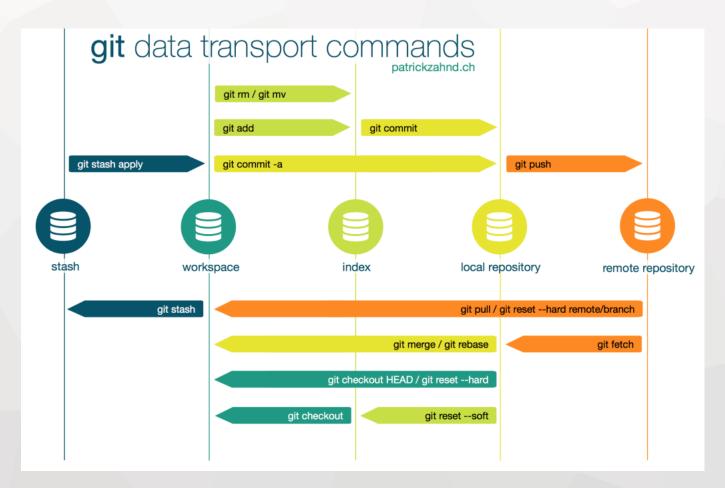
Official Website :
https://git-scm.com/
The latest stable Release : v2.11.1

git data transport commands
patrickzahnd.ch

PART2

Hadoop Eco System

VISUALIZATION LAYER AND API'S	Java	HTML5	D3	RESTful API
ANALYTICAL TOOLS	pentaho	mahout	R	
ANALYSIS LAYER	Java		Spark (Lightning-Fast Cluster Computing)	HIVE	cloudera IMPALA
PROCESSING LAYER	hadoop MapReduce	Bidoop Motor de Eventos	Bidoop Motor de Workflow	
STORAGE LAYER	hadoop HDFS	HIVE	APACHE HBASE	mongoDB
REAL TIME EVENT ENGINE	Bidoop Motor de Eventos			
INTEGRATION LAYER (ETL)	FLUME	chukwa	AVRO	sqoop
SOURCES			WWW		Logs XML		

CAPA DE GESTIÓN

Chef

HERRAMIENTAS ADMINISTRACIÓN

OOZIE

HERRAMIENTAS MONITORIZACIÓN

Ganglia

Nagios

HERRAMIENTAS DIAGNÓSTICO

Pacemaker

Zoo Keeper

PART3

Hadoop Architecture

Architecture

Data Replication on Multiple Nodes

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

Designed by Sri Prakash, November 2012

PART4
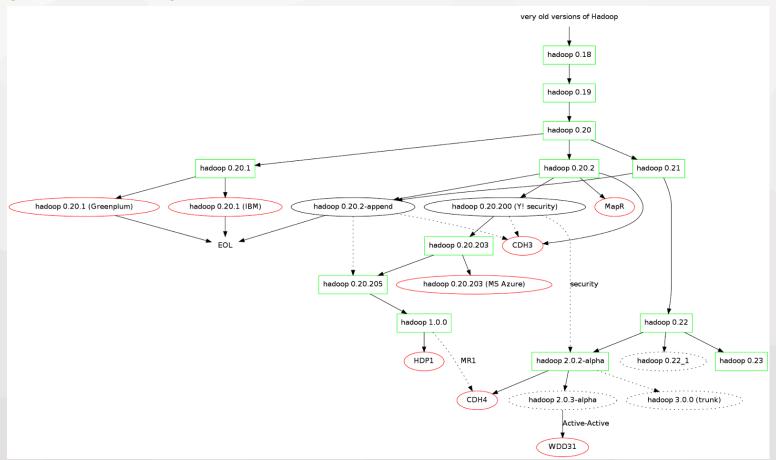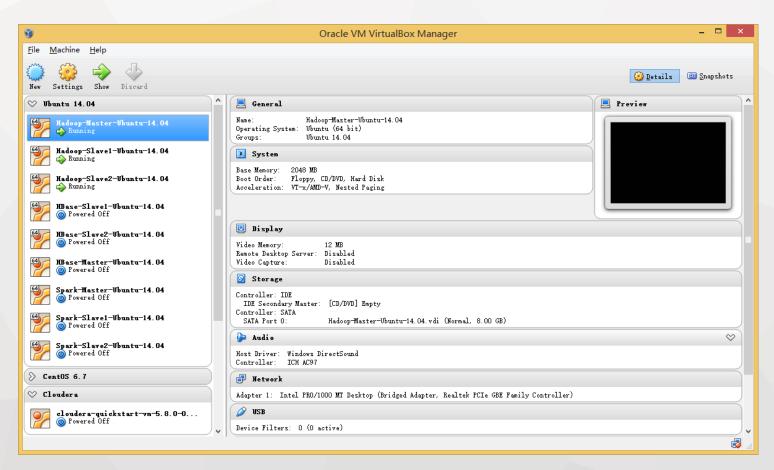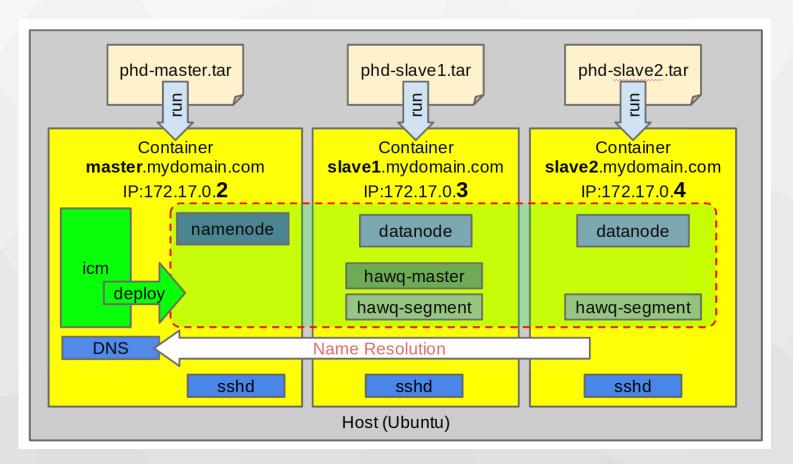
Hadoop Environment

Oracle Virtual Box

phd-master.tar

phd-slave1.tar

phd-slave2.tar

run

run

run

Container
master.mydomain.com
IP:172.17.0.**2**

Container
slave1.mydomain.com
IP:172.17.0.**3**

Container
slave2.mydomain.com
IP:172.17.0.**4**

icm

deploy

namenode

datanode

datanode

hawq-master

hawq-segment

hawq-segment

DNS

Name Resolution

sshd

sshd

sshd

Host (Ubuntu)

		Master	Slave	Master	Slave
Master	Slave	HBase-Base		Spark-Base	
Hadoop-Base					
DNS					
OS					

Hadoop configuration files
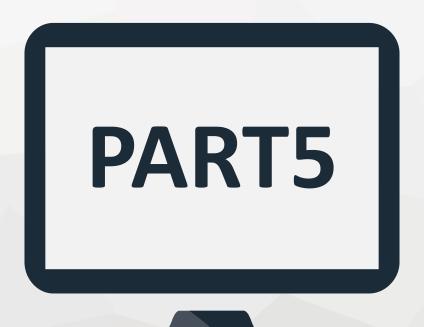
Table 9-1. Hadoop configuration files

Filename	Format	Description
hadoop-env.sh	Bash script	Environment variables that are used in the scripts to run Hadoop
core-site.xml	Hadoop configuration XML	Configuration settings for Hadoop Core, such as I/O settings that are common to HDFS and MapReduce
hdfs-site.xml	Hadoop configuration XML	Configuration settings for HDFS daemons: the namenode, the secondary namenode, and the datanodes
mapred-site.xml	Hadoop configuration XML	Configuration settings for MapReduce daemons: the jobtracker, and the tasktrackers
masters	Plain text	A list of machines (one per line) that each run a secondary namenode
slaves	Plain text	A list of machines (one per line) that each run a datanode and a task-tracker
hadoop-metrics .properties	Java Properties	Properties for controlling how metrics are published in Hadoop (see "Metrics" on page 352)
log4j.properties	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
hadoop-env.sh	Bash script	Environment variables that are used in the scripts to run Hadoop
mapred-env.sh	Bash script	Environment variables that are used in the scripts to run MapReduce (overrides variables set in hadoop-env.sh)
yarn-env.sh	Bash script	Environment variables that are used in the scripts to run YARN (overrides variables set in hadoop-env.sh)
core-site.xml	Hadoop configuration XML	Configuration settings for Hadoop Core, such as I/O settings that are common to HDFS, MapReduce, and YARN
hdfs-site.xml	Hadoop configuration XML	Configuration settings for HDFS daemons: the namenode, the secondary namenode, and the datanodes
mapred-site.xml	Hadoop configuration XML	Configuration settings for MapReduce daemons: the job history server
yarn-site.xml	Hadoop configuration XML	Configuration settings for YARN daemons: the resource manager, the web app proxy server, and the node managers
slaves	Plain text	A list of machines (one per line) that each run a datanode and a node manager
hadoop-metrics2 .properties	Java properties	Properties for controlling how metrics are published in Hadoop (see Metrics and JMX)
log4j.properties	Java properties	Properties for system logfiles, the namenode audit log, and the task log for the task JVM process (Hadoop Logs)
hadoop-policy.xml	Hadoop configuration XML	Configuration settings for access control lists when running Hadoop in secure mode
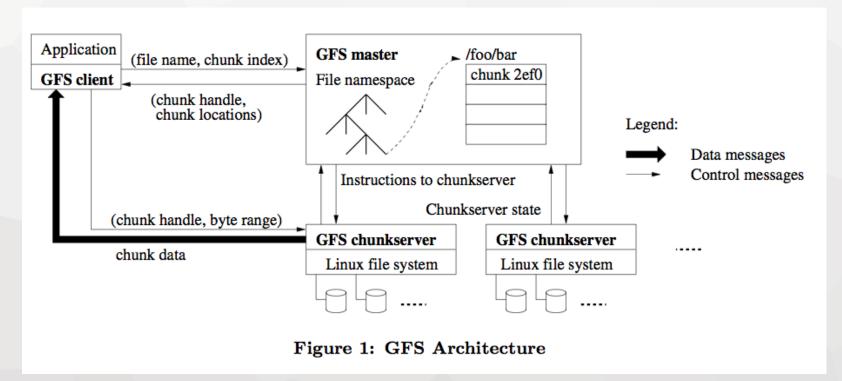
Hadoop 2.x
Hadoop.The.Definitive.Guide.4th.Edition

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

NameNode

Datanode

Storage reports &
Block reports

Read/write
data

**Single
Logical
Storage**

Clients

Figure 1: A DataNode presented itself as a single logical storage
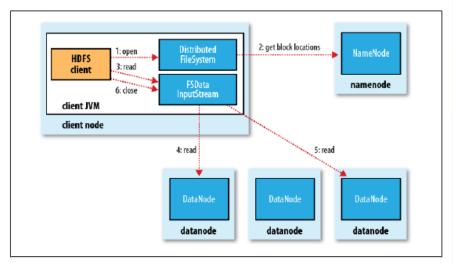
Figure 3-2. A client reading data from HDFS
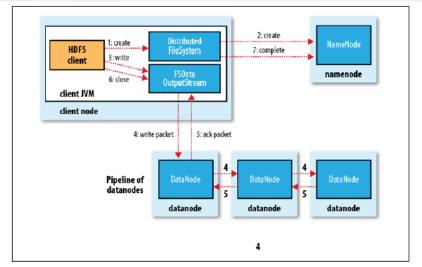
Figure 3-4. A client writing data to HDFS

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$
```
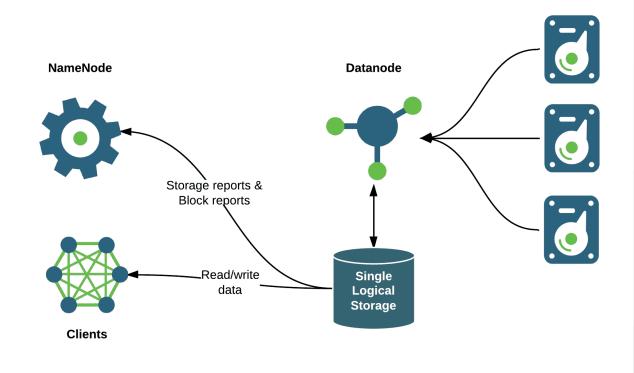
Official Website :

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
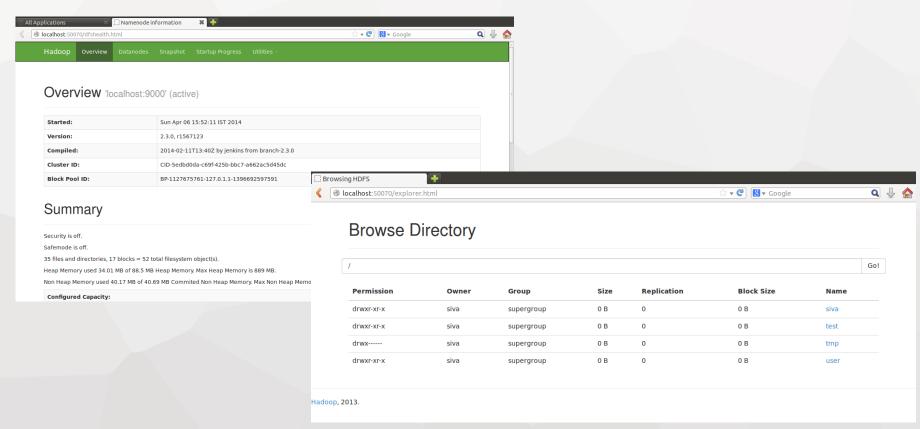
export JAVA_HOME=/usr/java/jdk1.6.0_26/ for service in /etc/init.d/hadoop-0.20-* do sudo $service start done	STARTING THE PROCESSES
Hadoop NameNode (http://<hostname>:50070) Hadoop JobTracker (http://<hostname>:50030)	MONITORING PAGES
for service in /etc/init.d/hadoop-0.20-* do sudo $service stop done	STOPPING THE PROCESSES

HADOOP HDFS

hadoop fs -ls <path>	LIST FILES
hadoop fs -mkdir <path>	MAKE DIRECTORY
hadoop fs -rmr <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

HDFS Web UI

PART6

Reference Books

O'REILLY

4th Edition
Revised & Updated

Hadoop
The Definitive Guide

STORAGE AND ANALYSIS AT INTERNET SCALE

Tom White

Quick answers to common problems

Hadoop Real-World Solutions Cookbook

Realistic, simple code examples to solve problems at scale with Hadoop and related technologies

Jonathan R. Owens   Jon Lentz
Brian Femiano

[PACKT] open source

HomeWork

The End