LESSON 1

사전 준비사항

사전 준비사항

전제조건

- 하이퍼바이저나 도커 설치(리눅스 네이티브 인스톨시 불필요)
- 우분투 서버 이미지 다운로드 설치
  - vagrant나 도커 사용할 경우 이미지 검색 및 설치
- 자바 설치
  - 오픈 자바나 오라클 자바 설치
- SSH 설치
  - 원격접속용, putty 설치(윈도우용 SSH 클라이언트)
- NANO(기본에디터) 설치
  - 도커의 경우 따로 인스톨 해줘야 함

실습동영상
하이퍼바이저/리눅스/vagrant/자바/ssh설치

LESSON 2

하둡의 수행모드

하둡의 수행모드

- 스탠드얼론(Standalone)
  - 기본수행모드
  - HDFS 사용하지 않음
  - 서버 없이 간단한 프로그램 테스트용
- 의사분산모드(Pseudo Distributed Mode)
  - HDFS 사용(입출력 모두)
  - 한 대의 서버에 모든 서버 수행
    - 네임노드/데이터노드/잡트래커/데이터노드
- 완전분산모드(Full Distributed Mode)
  - HDFS사용
  - 각 서버를 여러 대의 시스템에 나눠 수행
  - 하둡을 제대로 사용하는 모드

LESSON 3

하둡 1.2.1 환경설정

하둡 1.2.1 환경설정

- 하둡 다운로드

  wget http://archive.apache.org/dist/hadoop/core/hadoop-1.2.1/hadoop-1.2.1.tar.gz

- 압축해제

  tar xvfz hadoop-1.2.1.tar.gz

- 에디터 수행

  nano ~/.profile 또는 nano ~/.bashrc

- 환경변수 설정 및 저장(마지막에 추가)

  export HADOOP_HOME=/home/vagrant/hadoop-1.2.1

  export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

하둡 1.2.1 환경설정

- 환경설정 반영

  source ~/.profile 또는 source ~/.bashrc

- 환경변수 설정 확인

  echo $HADOOP_HOME / echo $PATH

하둡 1.2.1 환경설정

하둡 1.2.1 - 스탠드얼론

하둡 수행

hadoop jar $HADOOP_HOME/hadoop-examples-1.2.1.jar wordcount $HADOOP_HOME/README.txt ~/output

결과 확인

nano /output/part-r-00000 또는 cat /output/part-r-00000

하둡 1.2.1 환경설정

워드카운트 예제 실행

$ hadoop jar hadoop-examples-1.2.1.jar wordcount README.txt ~/wordcount-output

```
Warning: $HADOOP_HOME is deprecated.

13/11/20 11:48:42 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/11/20 11:48:42 INFO input.FileInputFormat: Total input paths to process : 1
13/11/20 11:48:42 WARN snappy.LoadSnappy: Snappy native library not loaded
13/11/20 11:48:42 INFO mapred.JobClient: Running job: job_local_0001
13/11/20 11:48:42 INFO util.ProcessTree: setsid exited with exit code 0
13/11/20 11:48:42 INFO mapred.Task:  Using ResourceCalculatorPlugin : org.apache.hadoop.util.LinuxResourceCalculatorPlugin@44e03e45
13/11/20 11:48:42 INFO mapred.MapTask: io.sort.mb = 100
13/11/20 11:48:42 INFO mapred.MapTask: data buffer = 79691776/99614720
13/11/20 11:48:42 INFO mapred.MapTask: record buffer = 262144/327680
13/11/20 11:48:42 INFO mapred.MapTask: Starting flush of map output
13/11/20 11:48:42 INFO mapred.MapTask: Finished spill 0
13/11/20 11:48:42 INFO mapred.Task: Task:attempt_local_0001_m_000000_0 is done. And is in the process of commiting
13/11/20 11:48:43 INFO mapred.JobClient:  map 0% reduce 0%
13/11/20 11:48:45 INFO mapred.LocalJobRunner:
13/11/20 11:48:45 INFO mapred.Task: Task 'attempt_local_0001_m_000000_0' done.
13/11/20 11:48:45 INFO mapred.Task:  Using ResourceCalculatorPlugin : org.apache.hadoop.util.LinuxResourceCalculatorPlugin@4b7aa8c8
13/11/20 11:48:45 INFO mapred.LocalJobRunner:
13/11/20 11:48:45 INFO mapred.Merger: Merging 1 sorted segments
13/11/20 11:48:45 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1832 bytes
13/11/20 11:48:45 INFO mapred.LocalJobRunner:
13/11/20 11:48:45 INFO mapred.Task: Task:attempt_local_0001_r_000000_0 is done. And is in the process of commiting
13/11/20 11:48:45 INFO mapred.LocalJobRunner:
13/11/20 11:48:45 INFO mapred.Task: Task attempt_local_0001_r_000000_0 is allowed to commit now
```

하둡 1.2.1 환경설정

워드카운트 결과

$ cat ~/wordcount-output/part-r-00000

(BIS), 1
(ECCN) 1
(TSU) 1
(see 1
5D002.C.1, 1
740.13) 1
<http://www.wassenaar.org/> 1
Administration 1
Apache 1
BEFORE 1
BIS 1
Bureau 1
Commerce, 1
Commodity 1
Control 1
Core 1
Department 1
ENC 1
Exception 1
Export 2
For 1
Foundation 1
Government 1
Hadoop 1
Hadoop, 1
Industry 1
Jetty 1
License 1
Number 1
Regulations, 1
SSL 1
Section 1
Security 1
See 1
Software 2
Technology 1
The 4
This 1
U.S. 1
Unrestricted 1
about 1
algorithms. 1
and 6
and/or 1
another 1
any 1
as 1
asymmetric 1
at: 2
both 1
by 1
check 1
classified 1
code 1
code. 1
concerning 1
country 1
country's 1
country, 1
cryptographic 3
currently 1
details 1
distribution 2
eligible 1
encryption 3
exception 1
export 1
following 1
for 3
form 1
from 1
functions 1
has 1
have 1
http://hadoop.apache.org/core/ 1
http://wiki.apache.org/hadoop/ 1
if 1
import, 2
in 1
included 1
includes 2
information 2
information. 1
is 1
it 1
latest 1
laws, 1
libraries 1
makes 1
manner 1
may 1
more 2
mortbay.org. 1
object 1
of 5
or 2
our 2
performing 1
permitted. 1
please 2
policies 1
possession, 2
project 1
provides 1
re-export 2
regulations 1
reside 1
restrictions 1
security 1
see 1
software 2
software, 2
software. 2
software: 1
source 1
the 8
this 3
to 2
under 1
use, 2
uses 1
using 2
visit 1
website 1
which 2
wiki, 1
with 1
written 1
you 1
your 1

실습동영상
hadoop-1.2.1 standalone

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

- 실행해야 할 서버(Server-Daemon)
  - HDFS: Name Node/Secondary NameNode/DataNode
  - MapReduce: JobTracker/TaskTracker
- ssh 공개키기반 자동로그인 설정
- 환경설정 파일 수정

  $HADOOP_HOME/conf/hadoop-env.sh 수정(JAVA_HOME 설정)
  $HADOOP_HOME/conf/mapred-site.xml
  $HADOOP_HOME/conf/hdfs-site.xml
  $HADOOP_HOME/conf/core-site.xml
- HDFS 포맷

  $ hadoop namenode -format
- 데몬 수행 및 수행확인

  $ start-all.sh(deprecated) / start-dfs.sh + start-mapred.sh
  $ jps

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

- conf/hadoop-env.sh 수정

  export JAVA_HOME =

- 자바 홈 디렉토리 설정
  - 우분투의 경우 /usr/lib/jvm 밑에 있음
    실제 디렉토리 확인 필요
  - 오픈자바(JDK)의 경우
    /usr/lib/jvm/java-7-openjdk-amd64
  - 오라클 자바의 경우
    /usr/lib/jvm/java-8-oracle

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

ssh 설치

$ sudo apt-get install openssh-server

ssh 자동로그인 설정

$ ssh-keygen -t dsa -P "" -f ~/.ssh/id_dsa

```
Generating public/private dsa key pair.
Created directory '/home/sjha/.ssh'.
Your identification has been saved in /home/sjha/.ssh/id_dsa.
Your public key has been saved in /home/sjha/.ssh/id_dsa.pub.
The key fingerprint is:
5f:36:aa:9c:9e:c8:0c:04:93:92:2e:53:7a:ae:cf:8d sjha@ubuntu-server
The key's randomart image is:
+--[ DSA 1024]----+
|           |
| . .       |
|o =        |
|.+ o       |
|+.. .  S  +  |
|.+ .   . + .  |
| . .   o   |
| o o + o +    |
|..E . +.*     |
+-----------------+
```

$ cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

$ cat authorized_keys

ssh-dss AAAAB3NzaC1kc3MAAACBALT/4wKb0sYolzrhH66oatJU5x/533GO/soiB58s1DfUUF6Hp7cKFbZvCVimFSmepMx1XTlLjDA/BwBtWGWwz9XF2fEWzov5
LFR9yKGomfWU9yoj+nbzlFRuLdy0lolu4oSgXwNtWtDicPsjBxysUidvplvI7k2OZFnXXnFUoKUbAAAAFQCvlkXdAu/B+vGZYGPqvg7COnFBjwAAAIByKVuvrDt9/m
NUQ+eBZcrvEBWrbiRsUVFAXw4NQDm8MmLu9/ZqJEdozySwNiEenQC0WKtltcJRc2iPvV17XTiXCsWrcyvUNwxmGLQrteyDeidPa0V5A7YG5r5dTwlPLfrTUqRyfm
1DyUlkdig7G9XQL+3ydXn7AND/nVpQ2PcDbwAAAIAX2Y/mXvcjGTqaE2Fl+M6kCEZguWPTUmLTsOXOvEQOtZLoXGeNTOuXCNtJrdEQD4sn+7jRMMWrWCw
CnbFRZ4dcfPBVSIXmCrBljvfnua5lgEfOgUSoobmaPbJEG3D6vstvmr65isb6yfSPg3yvfio4Q85QqG8MnYQGAG7Q/bW+oQ== sjha@ubuntu-server
ssh-dss AAAAB3NzaC1kc3MAAACBALT/4wKb0sYolzrhH66oatJU5x/533GO/soiB58s1DfUUF6Hp7cKFbZvCVimFSmepMx1XTlLjDA/BwBtWGWwz9XF2fEWzov5
LFR9yKGomfWU9yoj+nbzlFRuLdy0lolu4oSgXwNtWtDicPsjBxysUidvplvI7k2OZFnXXnFUoKUbAAAAFQCvlkXdAu/B+vGZYGPqvg7COnFBjwAAAIByKVuvrDt9/m
NUQ+eBZcrvEBWrbiRsUVFAXw4NQDm8MmLu9/ZqJEdozySwNiEenQC0WKtltcJRc2iPvV17XTiXCsWrcyvUNwxmGLQrteyDeidPa0V5A7YG5r5dTwlPLfrTUqRyfm
1DyUlkdig7G9XQL+3ydXn7AND/nVpQ2PcDbwAAAIAX2Y/mXvcjGTqaE2Fl+M6kCEZguWPTUmLTsOXOvEQOtZLoXGeNTOuXCNtJrdEQD4sn+7jRMMWrWCw
CnbFRZ4dcfPBVSIXmCrBljvfnua5lgEfOgUSoobmaPbJEG3D6vstvmr65isb6yfSPg3yvfio4Q85QqG8MnYQGAG7Q/bW+oQ== sjha@ubuntu-server

- 테스트(비밀번호없이 자동로그인 되는지 확인)

  $ ssh localhost

  $ exit

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

conf/mapred-site.xml

```
<configuration>
  <property>
    <name>mapred.job.tracker</name>
    <value>localhost:9001</value>
  </property>
</configuration>
```

conf/hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

conf/core-site.xml

```
<configuration>
  <property>
   <name>fs.default.name</name>
   <value>hdfs://localhost:9000</value>
  </property>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/home/vagrant/temp</value>
  </property>
</configuration>
```

/home/vagrant아래에 temp폴더 생성(자신의 홈디렉토리에)

mkdir /home/vagrant/temp

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

HDFS 포맷

```
$ hadoop namenode -format
Warning: $HADOOP_HOME is deprecated.

13/11/20 13:19:29 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = ubuntu-server/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 1.2.1
STARTUP_MSG:   build = https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.0 -r 1335192; compiled by 'hortonfo' on Tue May  8 20:31:25 UTC 2012
************************************************************/
13/11/20 13:19:29 INFO util.GSet: VM type       = 64-bit
13/11/20 13:19:29 INFO util.GSet: 2% max memory = 19.33375 MB
13/11/20 13:19:29 INFO util.GSet: capacity      = 2^21 = 2097152 entries
13/11/20 13:19:29 INFO util.GSet: recommended=2097152, actual=2097152
13/11/20 13:19:30 INFO namenode.FSNamesystem: fsOwner=sjha
13/11/20 13:19:30 INFO namenode.FSNamesystem: supergroup=supergroup
13/11/20 13:19:30 INFO namenode.FSNamesystem: isPermissionEnabled=true
13/11/20 13:19:30 INFO namenode.FSNamesystem: dfs.block.invalidate.limit=100
13/11/20 13:19:30 INFO namenode.FSNamesystem: isAccessTokenEnabled=false accessKeyUpdateInterval=0 min(s), accessTokenLifetime=0 min(s)
13/11/20 13:19:30 INFO namenode.NameNode: Caching file names occuring more than 10 times
13/11/20 13:19:31 INFO common.Storage: Image file of size 110 saved in 0 seconds.
13/11/20 13:19:31 INFO common.Storage: Storage directory /tmp/hadoop-sjha/dfs/name has been successfully formatted.
13/11/20 13:19:31 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at ubuntu-server/127.0.1.1
************************************************************/
```

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

데몬 실행

$ start-all.sh 또는 start-dfs.sh + start-mapred.sh

```
Warning: $HADOOP_HOME is deprecated.
starting namenode, logging to /home/sjha/hadoop-1.2.1/libexec/../logs/hadoop-sjha-namenode-ubuntu-server.out
localhost: starting datanode, logging to /home/sjha/hadoop-1.2.1/libexec/../logs/hadoop-sjha-datanode-ubuntu-server.out
localhost: starting secondarynamenode, logging to /home/sjha/hadoop-1.0.3/libexec/../logs/hadoop-sjha-secondarynamenode-ubuntu-server.out
starting jobtracker, logging to /home/sjha/hadoop-1.0.3/libexec/../logs/hadoop-sjha-jobtracker-ubuntu-server.out
localhost: starting tasktracker, logging to /home/sjha/hadoop-1.2.1/libexec/../logs/hadoop-sjha-tasktracker-ubuntu-server.out
```

$ jps

```
5225 JobTracker
5332 TaskTracker
5134 SecondaryNameNode
4636 NameNode
4878 DataNode
5535 Jps
```

$ stop-all.sh

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

워드카운트 예제 실행

- **HDFS상에 입력파일 업로드**
  $ hadoop fs -mkdir /input
  Warning: $HADOOP_HOME is deprecated.
- **HDFS상에 폴더생성(입력파일용)**
  $ hadoop fs -copyFromLocal README.txt /input
  Warning: $HADOOP_HOME is deprecated.
- **업로드 확인**
  $ hadoop fs -ls /input
  Warning: $HADOOP_HOME is deprecated.
  Found 1 items
  -rw-r--r-- 1 sjha supergroup 1366 2013-11-20 13:36 /input/README.txt

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

예제 실행

$ hadoop jar hadoop-examples-1.2.1.jar wordcount /input/README.txt /output

```
Warning: $HADOOP_HOME is deprecated.

13/11/20 13:38:53 INFO input.FileInputFormat: Total input paths to process : 1
13/11/20 13:38:53 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/11/20 13:38:53 WARN snappy.LoadSnappy: Snappy native library not loaded
13/11/20 13:38:54 INFO mapred.JobClient: Running job: job_201311201329_0001
13/11/20 13:38:55 INFO mapred.JobClient:  map 0% reduce 0%
13/11/20 13:39:09 INFO mapred.JobClient:  map 100% reduce 0%
13/11/20 13:39:21 INFO mapred.JobClient:  map 100% reduce 100%
13/11/20 13:39:26 INFO mapred.JobClient: Job complete: job_201311201329_0001
13/11/20 13:39:26 INFO mapred.JobClient: Counters: 29
13/11/20 13:39:26 INFO mapred.JobClient:   Job Counters
13/11/20 13:39:26 INFO mapred.JobClient:     Launched reduce tasks=1
13/11/20 13:39:26 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=14693
13/11/20 13:39:26 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
13/11/20 13:39:26 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
13/11/20 13:39:26 INFO mapred.JobClient:     Launched map tasks=1
13/11/20 13:39:26 INFO mapred.JobClient:     Data-local map tasks=1
13/11/20 13:39:26 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=10944
13/11/20 13:39:26 INFO mapred.JobClient:   File Output Format Counters
13/11/20 13:39:26 INFO mapred.JobClient:     Bytes Written=1306
13/11/20 13:39:26 INFO mapred.JobClient:   FileSystemCounters
13/11/20 13:39:26 INFO mapred.JobClient:     FILE_BYTES_READ=1836
13/11/20 13:39:26 INFO mapred.JobClient:     HDFS_BYTES_READ=1469
13/11/20 13:39:26 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=46779
13/11/20 13:39:26 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=1306
13/11/20 13:39:26 INFO mapred.JobClient:   File Input Format Counters
13/11/20 13:39:26 INFO mapred.JobClient:     Bytes Read=1366
```

하둡 1.2.1 환경설정

하둡 1.2.1-의사분산모드

예제 실행

```
13/11/20 13:39:26 INFO mapred.JobClient:   Map-Reduce Framework
13/11/20 13:39:26 INFO mapred.JobClient:     Map output materialized bytes=1836
13/11/20 13:39:26 INFO mapred.JobClient:     Map input records=31
13/11/20 13:39:26 INFO mapred.JobClient:     Reduce shuffle bytes=1836
13/11/20 13:39:26 INFO mapred.JobClient:     Spilled Records=262
13/11/20 13:39:26 INFO mapred.JobClient:     Map output bytes=2055
13/11/20 13:39:26 INFO mapred.JobClient:     Total committed heap usage (bytes)=222101504
13/11/20 13:39:26 INFO mapred.JobClient:     CPU time spent (ms)=1560
13/11/20 13:39:26 INFO mapred.JobClient:     Combine input records=179
13/11/20 13:39:26 INFO mapred.JobClient:     SPLIT_RAW_BYTES=103
13/11/20 13:39:26 INFO mapred.JobClient:     Reduce input records=131
13/11/20 13:39:26 INFO mapred.JobClient:     Reduce input groups=131
13/11/20 13:39:26 INFO mapred.JobClient:     Combine output records=131
13/11/20 13:39:26 INFO mapred.JobClient:     Physical memory (bytes) snapshot=258703360
13/11/20 13:39:26 INFO mapred.JobClient:     Reduce output records=131
13/11/20 13:39:26 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=1949204480
13/11/20 13:39:26 INFO mapred.JobClient:     Map output records=179
```

결과보기

```
$ hadoop fs -cat /output/part-r-00000 | more
```

실습동영상
hadoop-1.2.1-pseudo-distributed

실습동영상
hadoop-1.2.1-full-distributed