HADOOP

이영환

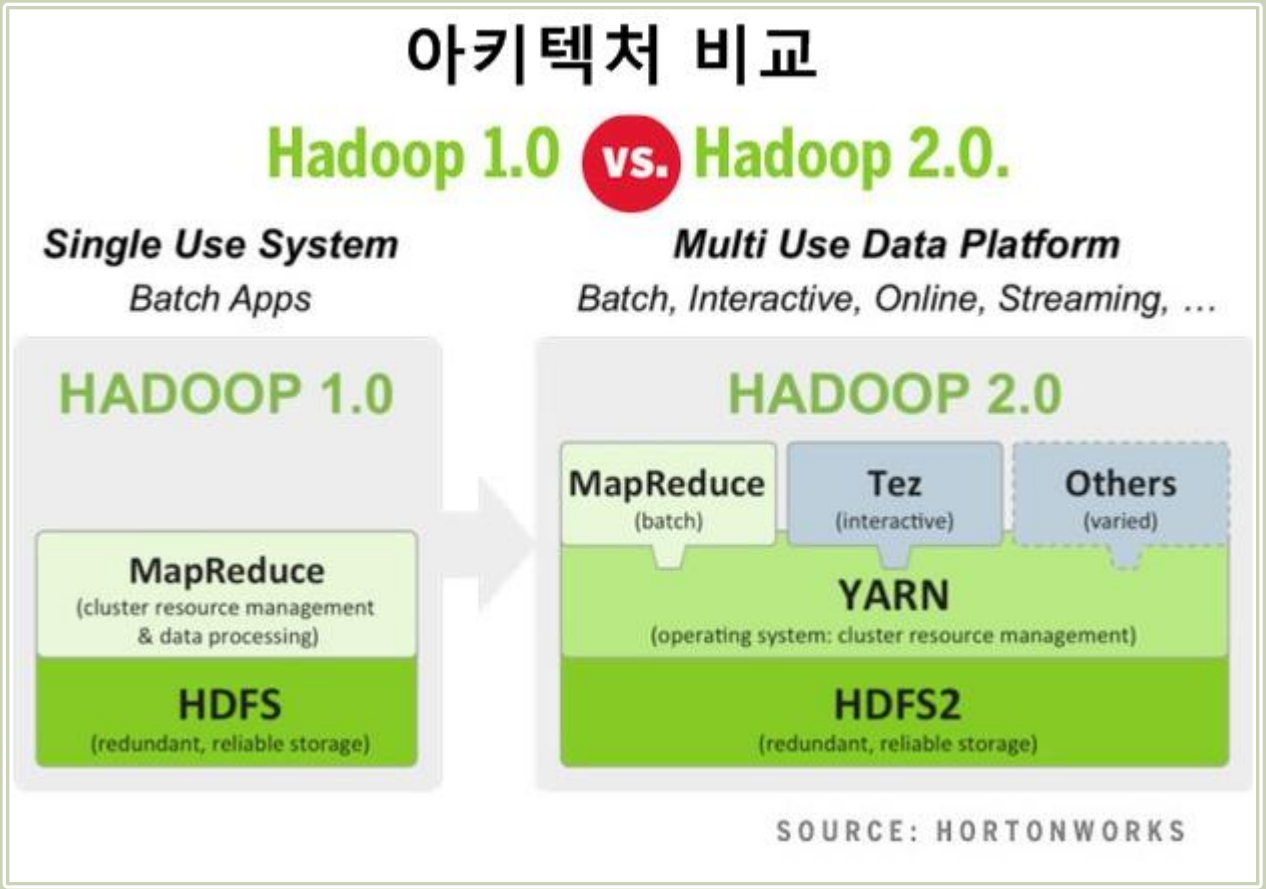
HADOOP 수행 3가지 모드

- 1. Standalone 모드
  - standalone 모드는 하둡의 기본 모드이다. 하둡 소스를 받아 압축을 푼 상태에서, 하둡은 사용자의 하드웨어에 대한 정보를 가지고 있지 않다. 따라서, 하둡은 로컬 머신에서만 실행된다. 다른 노드와 통신할 필요가 없기 때문에 standalone 모드에서는 HDFS를 사용하지 않고 다른 데몬들도 실행시키지 않는다. 이 모드의 목적은 독립적으로 MapReduce 프로그램의 로직을 개발하고 디버깅하는데 있다. 따라서 다른 데몬들과 서로 주고받는 부가 작업이 필요없다.

- 2. Pseudo-distributed 모드
  - pseudo-distributed 모드는 클러스터가 한 대로 구성되어 있고, 모든 데몬 역시 이 한 대의 컴퓨터에서 실행된다. 이 모드는 코드 디버깅 시 standalone 모드에서의 기능을 보완할 수 있는데, 메모리 사용정도, HDFS 입출력 관련 문제, 다른 데몬과의 상호작용에서 발생하는 일을 검사할 수 있다. standalone과 pseudo-distributed 모드는 모두 개발이나 디버깅 목적으로 사용된다. 실제 하둡 클러스터는 fully distributed 모드에서 실행된다.

- 3. Fully distributed 모드
  - 모든 기능이 갖추어진 클러스터 구성이며, 분산 저장과 분산 연산의 장점을 누릴 수 있다. 클러스터를 설명할 때는 아래와 같은 서버명을 사용한다.
  - master – 클러스터의 master 노드로서, NameNode와 JobTracker 데몬을 제공한다.
  - backup – SNN(Secondary NameNode 데몬)을 제공하는 서버
  - hadoop1, hadoop2, hadoop3, … – DataNode와 TaskTracker 데몬을 실행하는 slave들

HADOOP 1.X와 HADOOP 2.X 비교



HADOOP 1.2.1 설치

- 1. Hadoop Download
  - ~$ wget http://archive.apache.org/dist/hadoop/core/hadoop-1.2.1/hadoop-1.2.1.tar.gz
  - 다운로드 후 오류가 나면 다시 받아야 한다.(여러 번 할 수도 있다.)
  - 파일 지우기 명령어 : rm filename
  - 디렉토리 지우기 명령어 : rm -r directoryname
    - -r 옵션은 비어 있지 않아도 지운다.(root 권한으로 지우면 깔끔하게 지워진다.)
- 2. 압축해제
  - tar xvfz haddop-1.2.1.tar.gz
  - /home/ubuntu/hadoop-1.2.1 이라는 디렉토리에 압축이 풀린다.

- 3. 에디터로 설정 파일 열기
  - nano ~/.profile
- 4. 환경변수 설정 및 저장(아래 내용을 마지막에 추가해 준다.)
  - export HADOOP_HOME=/home/ubuntu/hadoop-1.2.1
  - export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
- 5. 환경변수 설정 반영
  - source ~/.profile
- 6. 환경변수 설정 확인
  - echo $HADOOP_HOME

STANDALONE으로 하둡실행

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

  - hadoop : 하둡을 실행해라
  - jar : 실행할 형태는 jar이다.
  - $HADOOP_HOME/hadoop-examples-1.2.1.jar : 예제 프로그램이 들어 있는 jar 파일
  - wordcount : wordcont를 실행하겠다.
  - $HADOOP_HOME/README.txt : 입력되는 데이터 파일
  - ~/output : 출력되는 위치
  - Standalone으로 실행했으므로 내컴퓨터에서 실행하게 된다.

- 실행 결과 확인
  - cat ~/output/part-r-00000 | more
  - cat : 파일 텍스트 내용을 확인하는 명령어
  - ~/output/part-r-00000 : 확인하려는 파일명
  - | more : 한페이지씩 출력하고 멈춰있다.
    - Enter를 치면 한줄씩, spacebar 키를 치면 한페이지씩 나온다.

PSEUDO-DISTRIBUTED 모드로 하둡실행

- 실행해야 할 서버(Server-Daemon)
  - HDFS : Name Node, Secondary Name Node, Data Node
  - MapReduce : Jon Tracker, Task Tracker
- SSH 공개키 기반 자동 로그인 설정
- 환경설정 파일 수정
  - $HADOOP_HOME/conf/hadoop_env.sh (JAVA_HOME 설정)
  - $HADOOP_HOME/conf/mapred-site.xml
  - $HADOOP_HOME/conf/hdfs-site.xml
  - $HADOOP_HOME/conf/core-site.xml

하둡 1.2.1 의사분산모드(1대로 실행)

- 환경설정
  - $HADOOP_HOME/conf/hadoop-env.sh 수정
    - nano $HADOOP_HOME/conf/hadoop-env.sh
    - export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
      - #을 풀고 자바의 위치를 잡아준다.

SSH 자동 로그인 설정

- ssh 설치
  - $ sudo apt-get install openssh-server
  - $ sudo apt-get install openssh-client
- ssh 폴더 만들기
  - mkdir ~/.ssh
- ssh 자동 로그인 설정
  - ssh 키 생성
  - $ ssh-keygen -t rsa
- 사용자 권한 파일에 생성된 키를 추가한다.
  - cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

하둡 1.2.1 환경 설정

- $HADOOP_HOME/conf/mapred-site.xml 의 configration tag를 완성한다.

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

- $HADOOP_HOME/conf/hdfs-site.xml 의 configration tag를 완성한다.

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

- $HADOOP_HOME/conf/core-site.xml 의 configration tag를 완성한다.

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

- HDFS 포맷
  - $hadoop namenode -format
- 데몬 수행 및 확인
  - start-all.sh
    - 모든 데몬 실행하는 명령어
  - stop-all.sh
    - 모든 데몬 중지시키는 명령어
  - jps
    - 실행되고 있는 데몬을 확인하는 명령어