CSE545
빅데이터처리 및 실습
(Big Data Processing and Practice)
Practice 02: Hadoop

담당교수: 전강욱(컴퓨터공학부)
kw.chon@koreatech.ac.kr

개요

- **실습명**
  - 하둡 설치 및 개발 환경 구축

- **목표**
  - 하둡을 설치하고, 기본적인 명령어 학습

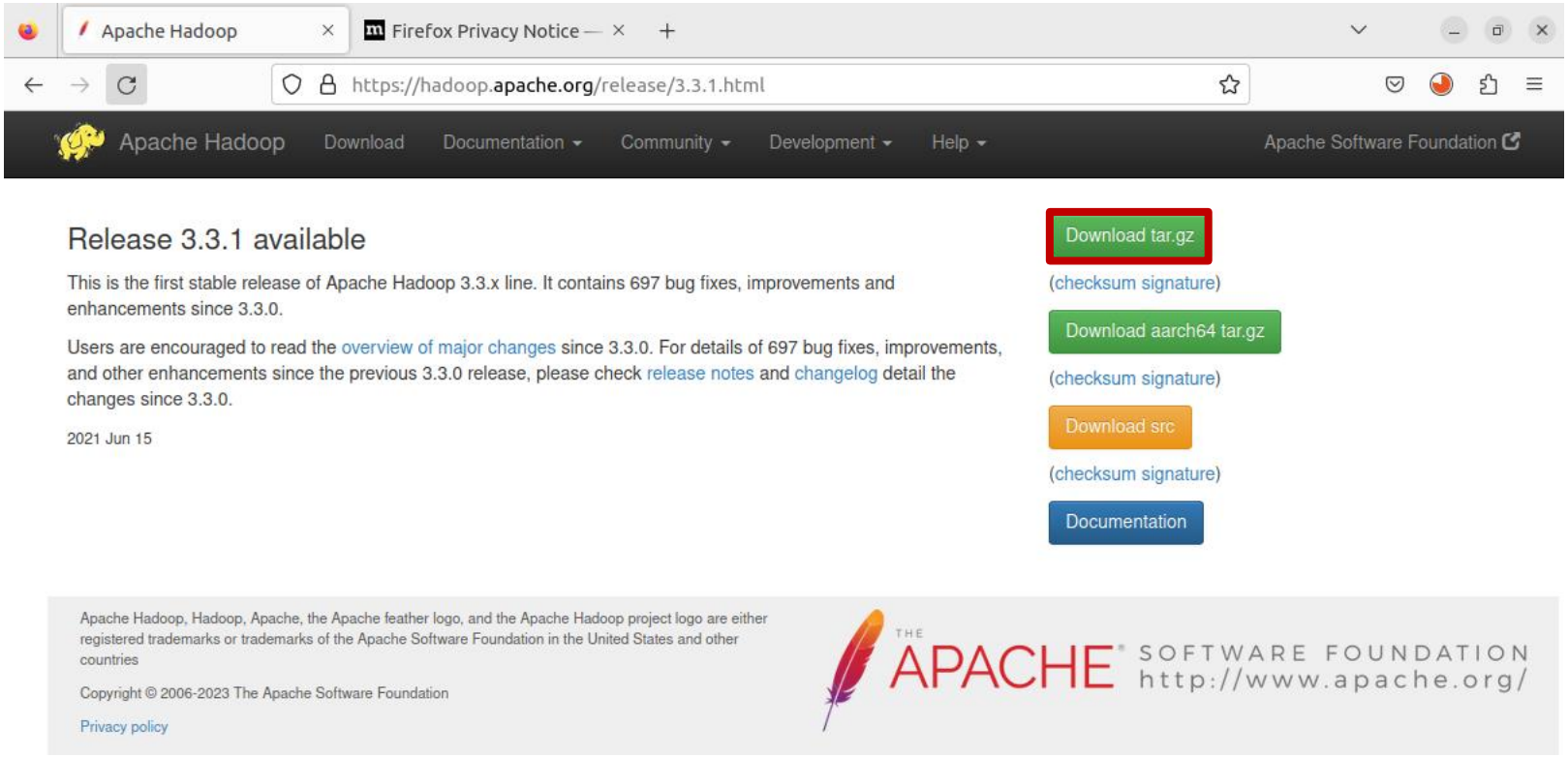
Java 설치

- **Java (jdk 1.8이상) 설치**
  - $ sudo apt-get update
  - $ sudo apt-get install openjdk-8-jdk
  - $ java –version

- **JAVA_HOME 설정**
  - Java 경로 확인
    - $readlink –f $(which java)
    - 예: /usr/lib/jvm/java-8-openjdk-amd64/bin/java
  - Java 경로를 등록
    - $sudo vim /etc/profile
    - 아래 내용(예시)을 /etc/profile 에 붙여 넣기
      - Export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
      - Export PATH=$PATH:$JAVA_HOME/bin
    - $source /etc/profile

Hadoop 설치

- **Hadoop 다운로드**
  - https://Hadoop.apache.org/release/3.3.1.html



Hadoop 설치 (계속)

- 하둡 설치 디렉토리는 "/home/Hadoop/Install”로 가정

- 1. 압축 풀기
  - $ tar –xvfz /home/Hadoop/Install/Hadoop-3.3.1.tar.gz

- 2. Hadoop 설정 디렉토리로 이동
  - $ cd /home/Hadoop/Install/Hadoop-3.3.1/etc/Hadoop

- 3. Hadoop 설정 파일(hadoop-env.sh) 수정 (예제)
  - $ export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/bin/java
  - $ export HADOOP_HOME= /home/Hadoop/Install/Hadoop-3.3.1
  - $ export HADOOP_CONF= /home/Hadoop/Install/Hadoop-3.3.1/etc/hadoop

Hadoop 설치 (계속)

- **4. 설정 파일 수정 (core-site.xml)**

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

- **5. 설정 파일 수정 (hdfs-site.xml)**

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

Hadoop 설치 (계속)

- **6. 설정파일 수정 (mapred-site.xml)**

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
    <property>
        <name>mapreduce.application.classpath</name>
        <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$
        HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*
        </value>
    </property>
</configuration>
```

Hadoop 설치 (계속)

- **7. 설정파일 수정 (yarn-site.xml)**

```
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
    <property>
        <name>yarn.nodemanager.env-whitelist</name>
        <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOO
        P_HDFS_HOME,
        HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCAC
        HE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME
        </value>
    </property>
</configuration>
```

Hadoop 설치 (계속)

- **8. SSH 설정**
  - $ sudo apt-get install openssh-server
  - $ ssh-keygen –t rsa –P '' –f ~/.ssh/id_rsa
  - $ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
  - $ chmod 0600 ~/.ssh/authorized_keys
  - $ ssh localhost
    - 접속 확인

Hadoop 설치 (계속)

- **9. NameNode (HDFS 메타 데이터 관리) 포맷**
  - $ hdfs namenode –format
  - NameNode를 포맷하면 dfs.namenode.name.dir 경로의 fsimage와 edits 파일을 초기화 시킴
  - 설치 후 최초 1회만 수행 함

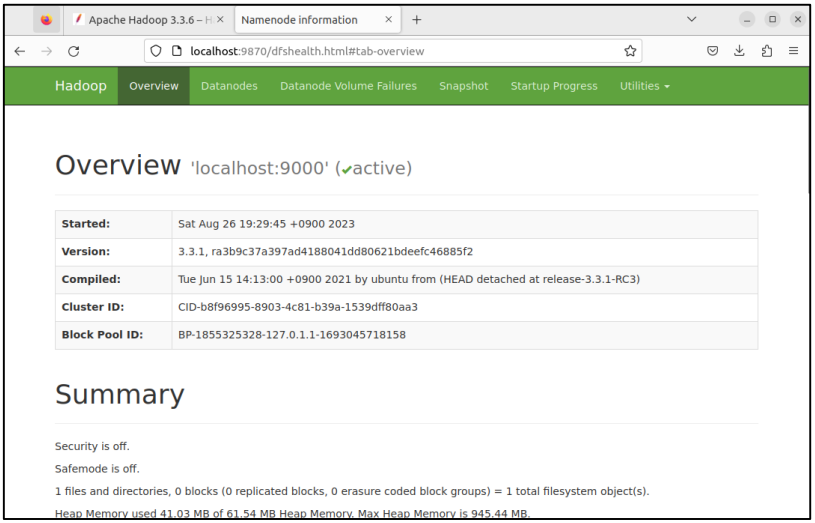
Hadoop 설치 (계속)

- **10. HDFS 데몬 실행**
  - $ sbin/start-dfs.sh

```
hadoop@hadoop-VirtualBox:~/Install/hadoop-3.3.1$ ./sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hadoop-VirtualBox]
hadoop-VirtualBox: Warning: Permanently added 'hadoop-virtualbox' (ED25519) to the list of known hosts.
hadoop@hadoop-VirtualBox:~/Install/hadoop-3.3.1$ jps
29697 SecondaryNameNode
29382 NameNode
29803 Jps
29500 DataNode
```

HDFS 데몬 실행

jps 명령어를 통해서 Java Process 확인



NameNode의 Web Interface 가 접속이 잘되면 성공
- http://localhost:9870/

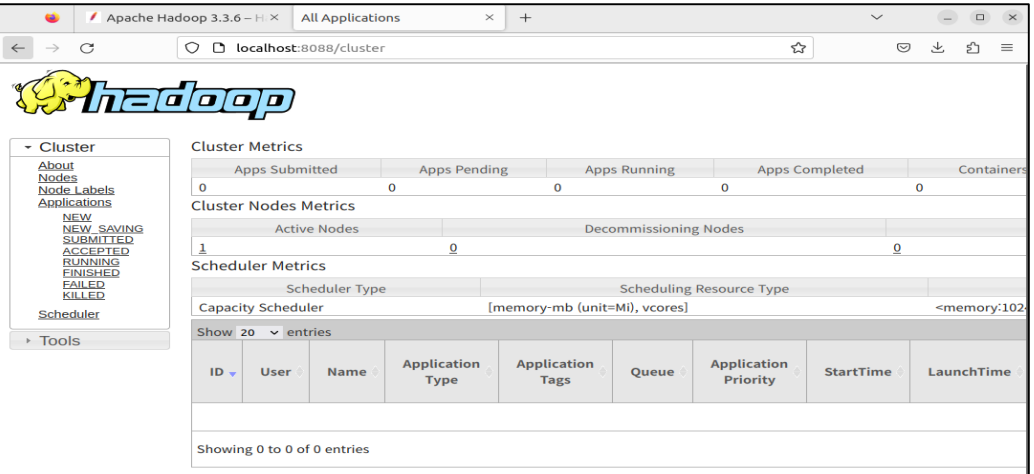
Hadoop 설치 (계속)

- 11. YARN 데몬 실행
  - $ sbin/start-yarn.sh

```
hadoop@hadoop-VirtualBox:~/Install/hadoop-3.3.1$ sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@hadoop-VirtualBox:~/Install/hadoop-3.3.1$ jps
30369 Jps
29697 SecondaryNameNode
29382 NameNode
29500 DataNode
29933 ResourceManager
30046 NodeManager
```

YARN 데몬 실행

jps 명령어를 통해서 Java Process 확인



Resource Manager의 Web Interface 가 접속이 잘되면 성공
- http://localhost:8088

Hadoop 설치 (계속)

- **하둡 설치 디렉토리는 "/home/Hadoop/Install”로 가정**

- **12. 예제 프로그램 (word count) 실행**
  - $ **hdfs dfs –mkdir –p** /user/hadoop/test
    - HDFS에 경로 (/user/hadoop/test) 생성
  - $ **hdfs dfs –put** /home/Hadoop/Install/Hadoop-3.3.1/etc/Hadoop/core-site.xml /user/Hadoop/test
    - Local에 있는 core-site.xml파일을 HDFS에 적재
  - $ **yarn jar** /home/Hadoop/Install/Hadoop-3.3.1 /share/hadoop/mapreduce/Hadoop-mapreduce-example-3.3.1.jar wordcount test output
    - 테스트용으로 이미 구현된 wordcount예제를 수행

제출물

- **Hadoop 설치를 위한 단계별로 캡쳐화면 생성하여 제출**
  - 본인이 한 것임을 증명할 수 있도록 캡쳐화면 (파일 이름 등)에 본인 학번이 포함

개발환경 구축

담당교수: 전강욱(컴퓨터공학부)
kw.chon@koreatech.ac.kr

Eclipse & Maven 설치

- **Eclipse**
  - 이클립스 다운로드 페이지: http://eclipse.org/downloads
    - 최신 버전 설치
  - $ tar xvfz ./eclipse-inst-jre-linux64.tar.gz
  - $ cd eclipse-installer
  - $ ./eclipse-inst
  - 이후 eclipse for Java Developer 설치

- **Maven**
  - $ sudo apt update
  - $ sudo apt install maven
  - $ mvn -version

개발환경 구축

- **Eclipse 실행 후 Maven 프로젝트 생성**
  - File → New → Maven Project
- **pom.xml 수정 (MapReduce Dependency 추가)**

```
<dependencies>
	<dependency>
		<groupId>org.apache.hadoop</groupId>
		<artifactId>hadoop-mapreduce-client-core</artifactId>
		<version>3.0.0</version>
	</dependency>
	<dependency>
		<groupId>org.apache.hadoop</groupId>
		<artifactId>hadoop-common</artifactId>
		<version>3.0.0</version>
	</dependency>
	<dependency>
		<groupId>org.apache.hadoop</groupId>
		<artifactId>hadoop-mapreduce-client-jobclient</artifactId>
		<version>3.0.0</version>
	</dependency>
</dependencies>
```

출처:
https://jangunmp1.github.io/tips/2019/04/17/hadoop.html

개발환경 구축 (계속)

- **MapReduce 코드 작성 (Package 목록)**

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;
```

개발환경 구축 (계속)

- **MapReduce 코드 작성 (Main 함수 및 Map 함수)**

```
public class WordCount extends Configured implements Tool{
    public static void main(String[] args) throws Exception {
        ToolRunner.run(new WordCount(), args);
    }
    public static class WCMapper extends Mapper<Object, Text, Text, IntWritable>{
        Text word = new Text();
        IntWritable one = new IntWritable(1);
        @Override protected void map(Object key, Text value,
                        Mapper<Object, Text, Text, IntWritable>.Context context)
                        throws IOException, InterruptedException
    {
        StringTokenizer st = new StringTokenizer(value.toString());
        while(st.hasMoreTokens()) {
            word.set(st.nextToken());
            context.write(word, one);
        }
    }
}
```

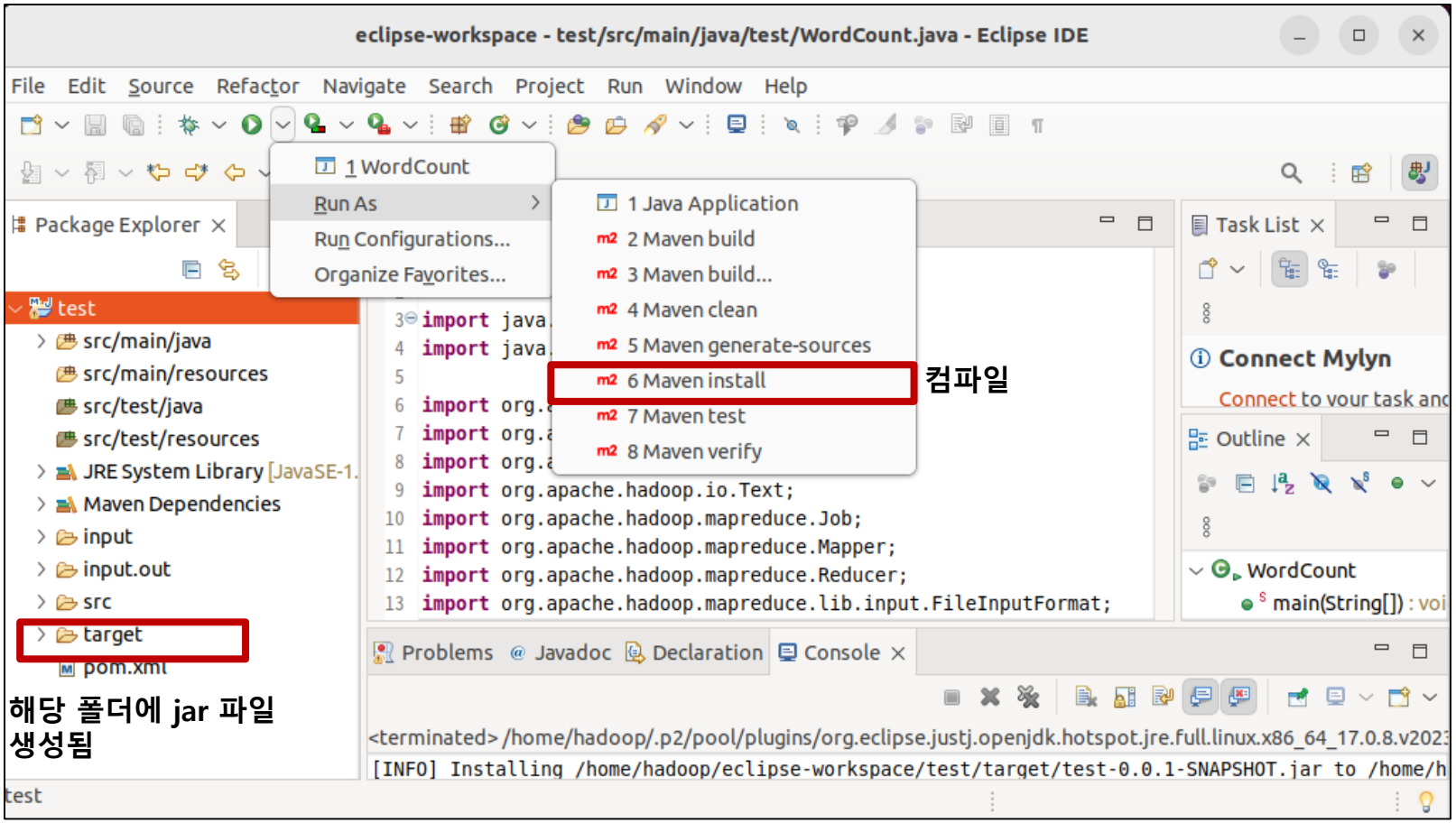
개발환경 구축 (계속)

- **MapReduce 코드 작성 (Reduce 함수)**

```
public static class WCReducer extends
            Reducer<Text, IntWritable, Text, IntWritable>
{
    IntWritable oval = new IntWritable();
    @Override protected void reduce(Text key, Iterable<IntWritable>
                                                          values,
        Reducer<Text, IntWritable, Text, IntWritable>.Context context)
                throws IOException, InterruptedException
    {
        int sum = 0;
        for(IntWritable value : values)
        {
            sum += value.get();
        } oval.set(sum);
        context.write(key, oval);
    }
}
```

개발환경 구축 (계속)

- **Maven 컴파일 (jar파일 생성)**
  - Eclipse에서 [Run As] → [Maven Install]



개발환경 구축 (계속)

- **jar 파일 실행 명령어**
  - yarn jar ./test-0.0.1-SNAPSHOT.jar test.WordCount /hadoop/test

Jar File | Main Class | Input Path on HDFS

```
hadoop@hadoop-VirtualBox:~/eclipse-workspace/test/target$ yarn jar ./test-0.0.1-SNAPSHOT.jar test.WordCount /hadoop/test/
2023-08-26 23:56:37,763 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2023-08-26 23:56:38,445 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.stag
ing/job_1693045913947_0002
2023-08-26 23:56:38,747 INFO input.FileInputFormat: Total input files to process : 1
2023-08-26 23:56:38,831 INFO mapreduce.JobSubmitter: number of splits:1
2023-08-26 23:56:39,068 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1693045913947_0002
2023-08-26 23:56:39,068 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-08-26 23:56:39,276 INFO conf.Configuration: resource-types.xml not found
2023-08-26 23:56:39,281 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-08-26 23:56:40,084 INFO impl.YarnClientImpl: Submitted application application_1693045913947_0002
2023-08-26 23:56:40,160 INFO mapreduce.Job: The url to track the job: http://hadoop-VirtualBox:8088/proxy/application_1693045913947_
0002/
2023-08-26 23:56:40,161 INFO mapreduce.Job: Running job: job_1693045913947_0002
2023-08-26 23:56:48,381 INFO mapreduce.Job: Job job_1693045913947_0002 running in uber mode : false
2023-08-26 23:56:48,382 INFO mapreduce.Job:  map 0% reduce 0%
2023-08-26 23:56:53,441 INFO mapreduce.Job:  map 100% reduce 0%
2023-08-26 23:56:58,481 INFO mapreduce.Job:  map 100% reduce 100%
2023-08-26 23:56:58,485 INFO mapreduce.Job: Job job_1693045913947_0002 completed successfully
2023-08-26 23:56:58,603 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=5450
```

제출물

- **단계별로 캡쳐화면 생성하여 제출**
  - 본인이 한 것임을 증명할 수 있도록 캡쳐화면(파일 이름 등)에 본인 학번이 포함

감사합니다.

Contact: kw.chon@koreatech.ac.kr