

Hadoop 없이 MapReduce 테스트 하기

김병곤

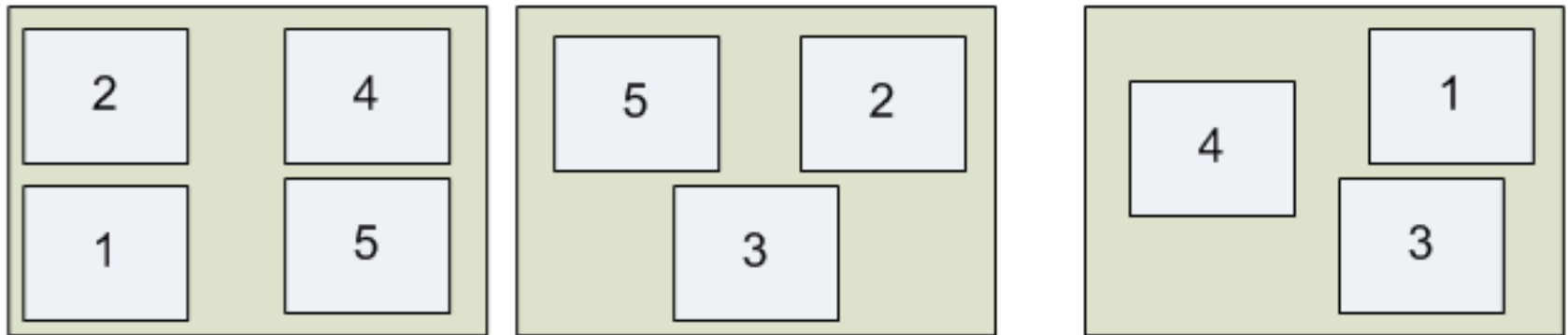
fharenheit@gmail.com

파일 시스템 : HDFS

NameNode:
Stores metadata only

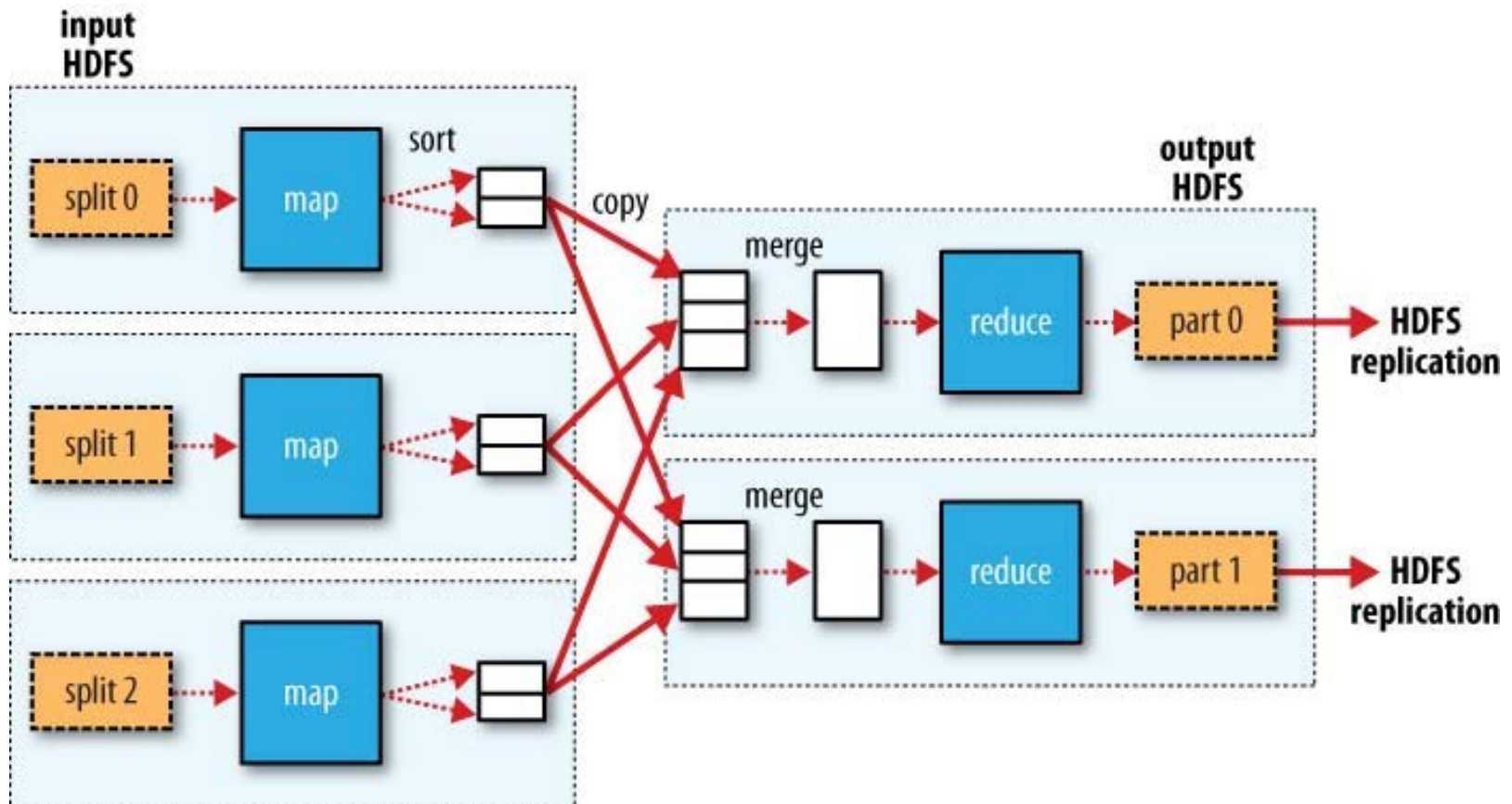
METADATA:
/user/aaron/foo → 1, 2, 4
/user/aaron/bar → 3, 5

DataNodes: Store blocks from files



프로그래밍 모델 : MapReduce

- HDFS의 파일을 처리하기 위한 프로그래밍 모델



WordCount

- Hadoop의 MapReduce Framework 동작을 이해하는 핵심 예제
- 각각의 ROW에 하나의 Word가 있을 때 Word의 개수를 알아내는 예제

입력 파일(Mapper의 Input)	출력 파일(Reduce Output)
hadoop	apache 1
apache	cloud 1
page	cluster 1
hive	copywrite 1
hbase	hadoop 2
cluster	hbase 1
hadoop	hive 1
page	page 2
cloud	
copywrite	

WordCount

입력 파일

```
hadoop
apache
page
hive
hbase
-----
cluster
hadoop
page
cloud
copywrite
```

Mapper



```
hadoop,1
apache,1
page,1
hive,1
hbase,1
```

Mapper



```
cluster,1
hadoop,1
page,1
cloud,1
copywrite,1
```

```
apache,<1>
cloud,<1>
cluster,<1>
copywrite,<1>
hadoop,<1,1>
hbase,<1>
hive,<1>
page,<1,1>
```

Reducer



출력 파일

```
apache 1
cloud 1
cluster 1
copywrite 1
hadoop 2
hbase 1
hive 1
page 2
```

MapReduce가 가지는 특징

- Map과 Reduce가 네트워크를 경계로 동작한다.
- Map의 Output Key를 중심으로 Reduce에서 데이터를 취합한다.
- Map, Reduce, Combiner, Partitioner, Input Format, Output Format, Multiple Output, Comparator 등등 다양한 구성 요소가 동작에 영향을 준다.
- 파일을 직접 다룬다.
- 분산 환경에서 동작한다.
- 대용량 파일을 다루므로 처리하는데 시간이 오래 걸린다.

MapReduce의 개발시 주의할 점

- 로그 파일이 크므로 처리하는데 오랜 시간이 소요되므로 시간을 단축 시키는 것은 매우 큰 비용이 절감됨
- 현장에서 발생하는 로그는 훨씬 더 다양한 케이스가 존재하므로 사전에 충분한 검증이 이루어지지 않으면 추후 급격한 비용이 발생(일반 개발은 저리 가라!!)
- 개발 기간보다 테스트 기간이 더 길 수 있다.
- 데이터를 이해하는 눈썰미가 꽤 장점으로 작용한다.

MRUnit이란 뭐냐?

- Hadoop의 내장 Object를 Mock Object로 구현한 단위 테스트 프레임워크
- Cloudera가 개발해서 Apache에 기증
- 최근 Top Level Project로 승격
- 문서 없음. 기대하지 마시길...
- 직접 빌드해서 사용하세요. 매우 친절하지 않습니다.

MRUnit이 없다면 그대는?

- Hadoop Cluster에 MR Job 실행하면서 고생하게 됩니다.
- Pseudo Mode에서 뭐 좀 해보려고 하겠죠
 - 생산성이 도저히 나오지 않을 거고, 메모리도 부족할 겁니다.
- 결과 파일과 입력 파일을 검증하는데 고생 좀 할 겁니다.

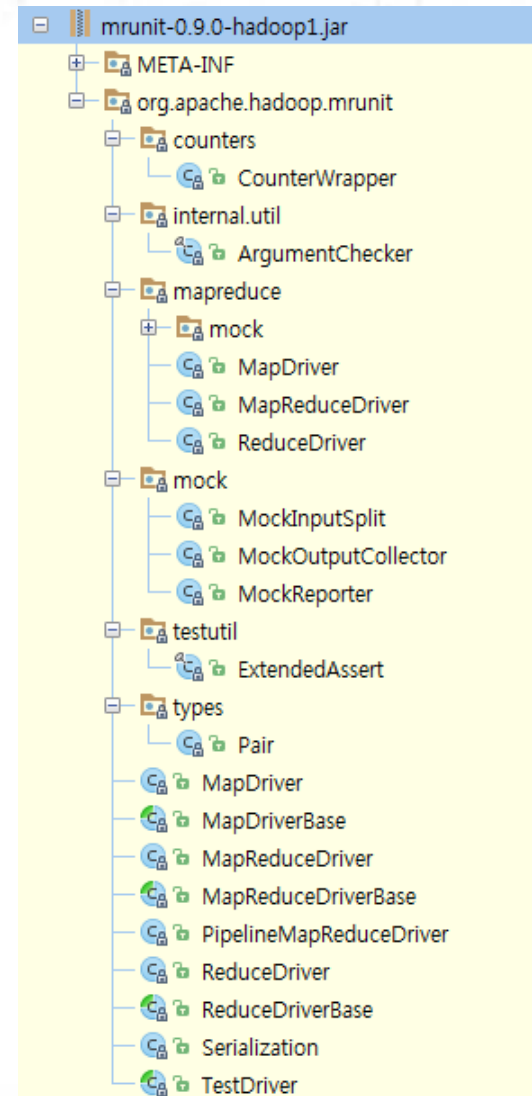
MRUnit은 어디서 구하나요?

The screenshot shows the GitHub repository page for `apache/mrunit`. The page is mirrored from `git://git.apache.org/mrunit.git`. It features a navigation bar with options like 'Code', 'Network', 'Pull Requests', and 'Graphs'. Below this, there are buttons for cloning the repository in various ways (Windows, ZIP, HTTP, SSH, Git Read-Only) and a 'Read-Only access' button. The repository is currently on the `trunk` branch. The main content area shows a list of commits, with the most recent one being `MRUNIT-161: some distributed cache apis not working - JobContext getC...` by Brock Noland, committed 18 days ago. Other commits include updates to `.gitignore`, `BIN-NOTICE.txt`, `BUILD.txt`, `CHANGES.txt`, `LICENSE.txt`, and `NOTICE.txt`, as well as a commit adding the `mrunit` framework to HADOOP-5518.

File	Commit Message	Author	Time
src	MRUNIT-161: some distributed cache apis not working - JobContext getC...	Brock Noland	18 days ago
.gitignore	MRUNIT-168: .gitignore does not exist	Brock Noland	a month ago
BIN-NOTICE.txt	MRUNIT-103: JUnit license not in NOTICE.txt in a binary tarball	Brock Noland	9 months ago
BUILD.txt	MRUNIT-96: Cleanup release: remove unnecessary artifacts from nexus ...	Jim Donofrio	10 months ago
CHANGES.txt	Update CHANGES.txt to include latest changeset from 0.9	Brock Noland	9 months ago
LICENSE.txt	MRUNIT-150: remove unnecessary log4j dependency	Jim Donofrio	5 months ago
NOTICE.txt	MRUNIT-96: Cleanup release: remove unnecessary artifacts from nexus ...	Jim Donofrio	10 months ago
README.txt	HADOOP-5518. Add contrib/mrunit, a MapReduce unit test framework. Con...	[cutting]	4 years ago

MRUnit은 어디서 구하나요?

- 지금까지 봤던 Apache Top Level Project에서 가장 소스코드가 없습니다.
- Map, Reduce, MapReduce를 별도로 테스트할 수 있는 Driver가 제공됩니다.
- MRUnit은 Map, Reduce 테스트 그 자체만 집중합니다.
 - Partitioner 같은 것 테스트 못합니다.



MRUnit 초기화

// Map, Reduce, MapReduce 테스트 범위에 따라서 Driver를 다르게 생성합니다.

```
public class GroupByMapReduceTest {  
  
    private Mapper mapper;  
    private Reducer reducer;  
    private MapReduceDriver driver;  
  
    @Before  
    public void setUp() {  
        mapper = new GroupByMapper();  
        reducer = new GroupByReducer();  
        driver = new MapReduceDriver(mapper, reducer);  
    }  
    ...  
}
```

MRUnit 테스트 케이스 작성

```
public class GroupByMapReduceTest {
    ...
    @Test
    public void groupBy() {
        Configuration conf = new Configuration();
        conf.set("inputDelimiter", ",");
        conf.set("keyValueDelimiter", ",");
        conf.set("valueDelimiter", ",");
        conf.set("allowDuplicate", "false");
        conf.set("allowSort", "false");
        conf.set("groupByKey", "0");

        driver.setConfiguration(conf);

        driver.withInput(new LongWritable(1), new Text("홍길동,a,b"));
        driver.withInput(new LongWritable(2), new Text("홍길동,b"));
        driver.withOutput(NullWritable.get(), new Text("홍길동,a,b"));
        driver.runTest();
    }
}
```

MRUnit의 테스트 케이스 위치

The screenshot shows the Eclipse IDE interface with the following components:

- Package Explorer:** Displays the project structure. The package `org.openflamingo.mapreduce.etl.groupby` is expanded, showing `GroupByMapReduceTest.java` in the `src/test/java` directory. Red boxes highlight `src/test/java` and `GroupByMapReduceTest.java`. Red arrows point from `src/main/java` to `GroupByMapReduceTest.java` and from `GroupByDriver.java`, `GroupByMapper.java`, and `GroupByReducer.java` to `GroupByMapReduceTest.java`.
- Editor:** Shows the source code of `GroupByMapReduceTest.java`. The code includes a package declaration, imports, a Javadoc comment, and the start of a `public class GroupByMapReduceTest` with private fields for `Mapper`, `Reducer`, and `MapReduceDriver`.
- Outline:** Lists the project's Javadoc artifacts, including `javacoc [default]`, `javacoc-1.6-html`, `javacoc-1.7`, `javacoc-1.7-html`, `javacoc-rtf`, and `javacoc-zip`.
- Console:** Shows a message: `<terminated> mapreduce-template build.xml [Ant Build] C:#Java#jdk1.6.0_32#bin#javaw.exe (20`

org.openflamingo.mapreduce.etl.groupby.GroupByMapReduceTest.java - mapreduce-template/src/test/java

MRUnit의 테스트 케이스 실행

The screenshot displays the Eclipse IDE interface with three main components:

- Package Explorer (Left):** Shows the project structure for 'mapreduce-template'. The package 'org.openflamingo.mapreduce.etl.groupby' is expanded, and the file 'GroupByMapReduceTest.java' is highlighted with a red box.
- Context Menu (Center):** A right-click context menu is open over the selected file. The 'Run As' option is highlighted with a red box.
- Run Dialog (Right):** The 'Run As' dialog is open, showing two options: '1 Run on Server' and '2 JUnit Test'. The 'JUnit Test' option is highlighted with a red box.

The Eclipse IDE title bar indicates the project is 'Java - mapreduce-template/src/test/java/org...'. The Package Explorer shows the following structure:

- mapreduce-template
 - src/test/java
 - org.openflamingo.mapreduce.etl.groupby
 - GroupByMapReduceTest.java
 - src/test/resources
 - src/main/java
 - com.yourcompany.hadoop.mapreduce
 - org.openflamingo.mapreduce
 - org.openflamingo.mapreduce.etl.groupby
 - org.openflamingo.mapreduce.etl.groupby.driver
 - src/main/resources
 - JRE System Library [jdk1.6.0_32]
 - Referenced Libraries
 - etc
 - lib
 - src
 - target
 - build.xml
 - oom.xml

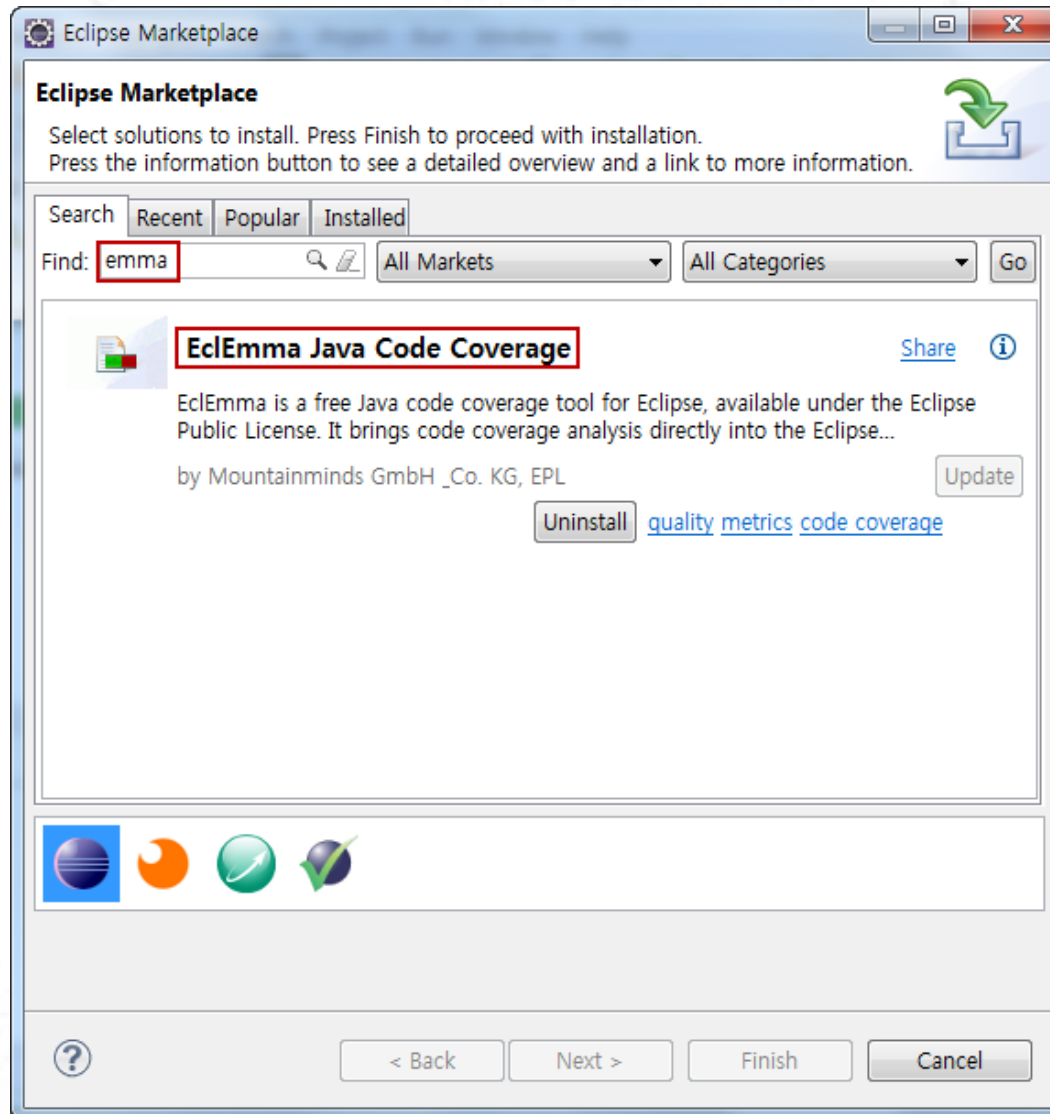
The Run Dialog shows the following options:

- 1 Run on Server (Alt+Shift+X, R)
- 2 JUnit Test (Alt+Shift+X, T)

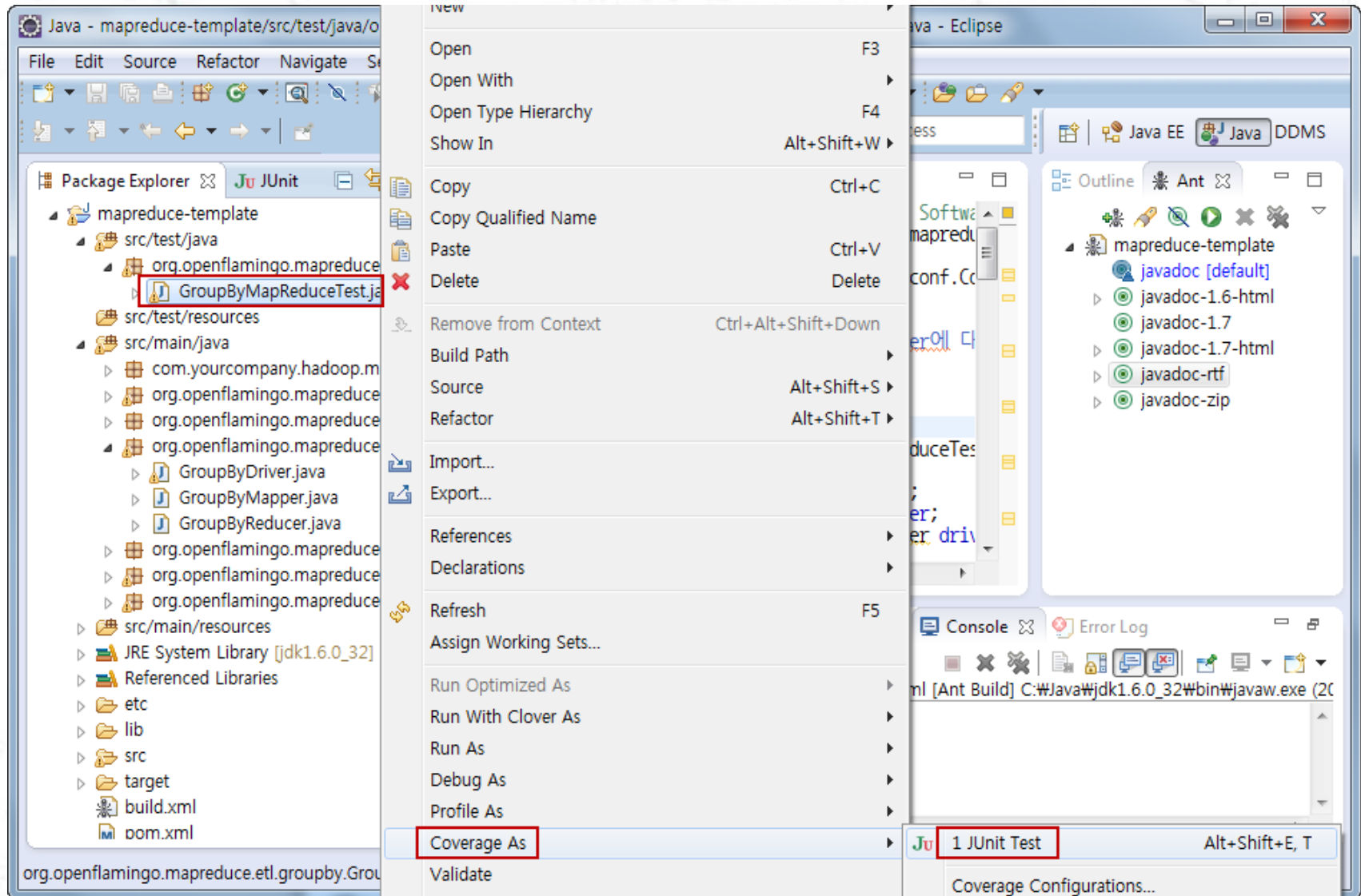
The Console window at the bottom shows the output of the Ant build process:

```
[Ant Build] C:\Java\jdk1.6.0_32\bin\javaw.exe (20...
```

Emma Code Coverage 설치



Code Coverage (1)



Code Coverage (2)

The screenshot shows the Eclipse IDE interface. The main editor displays the source code for `GroupByReducer.java`. The code includes a comment in Korean, a private boolean field `allowSort`, and two overridden methods: `setup` and `reduce`. The `setup` method is highlighted in green, indicating it is covered. The `reduce` method is partially visible.

The Coverage view at the bottom of the IDE shows the following data:

Element	Coverage	Covered Instructio...	Missed Instruct
org.openflamingo.mapreduce.etl.groupby	56.9 %	189	
GroupByDriver.java	0.0 %	0	
GroupByMapper.java	91.9 %	57	
GroupByReducer.java	100.0 %	132	

JBoss Community

JBoss Community (<http://www.jboss.org>)
Korea JBoss User Group (<http://cafe.naver.com/jbossug>)