# Hadoop Versions and Dependency Shading

Dependency Shading has been extended in this pull request: https://github.com/apache/flink/pull/454

Before the change, Flink had a "flink-shaded" module which contained a relocated guava dependency. All guava dependencies inside the individual modules were set to the 'provided' scope.

There are three areas where we shade our dependencies:

## Internal Shading: Hadoop Dependencies

Internal shading is about hiding some of Hadoop's dependencies from Flink.
Flink supports a broad range of Hadoop versions, which depend on different versions of certain dependencies.
Right now, we shade

- everything in the *com.google* namespace (guava and protobuf)
- *org.objectweb.asm*
- *org.jboss.netty*

They all end up in *org.apache.flink.hadoop.shaded.**
The shading of hadoop dependencies is done in "**flink-shaded-hadoop**" and its sub-modules.
Originally, we wanted to to handle the different hadoop versions inside this module, but the maven-shade-plugin doesn't properly support maven build profiles when creating the dependency reduced pom (a pom file which doesn't contain the shaded dependencies. In our case they won't contain guava, asm or netty anymore).
To resolve that, the "**flink-shaded-hadoop**" is the parent of some version specific flink-shaded-hadoop modules. The parent contains the configuration of the shading (the sub modules will inherit the configuration).
The submodules have the following purpose:

- **flink-shaded-hadoop1** Is for all *hadoop 0.2X and 1.XX versions*. It contains only hadoop-core + some dependency exclusions
- **flink-shaded-hadoop2** is for all hadoop versions starting from 2.x. It contains dependencies for hadoop-common, hadoop-hdfs, hadoop-mapreduce-client-core (for the hadoop compatibility stuff in flink-java). This module is only used for the hadoop 2.0.0-alpha build. It contains only some MapReduce interfaces and the HDFS client. It does not contain YARN.
- **flink-shaded-include-yarn** *This module is for Hadoop 2.2.0+.* This module contains the same dependencies as the hadoop2 module above + some dependencies for YARN.
- **flink-shaded-include-yarn-tests** This module is like flink-shaded-include-yarn but it contains all classes required for flink-yarn-tests. The dependencies are NOT in the 'tests' scope, because the maven-shade-plugin can not relocate them from there. We need this module to consistently shade all required classes when running the tests.

All other modules in flink which need something from hadoop need to depend on the right flink-shaded-hadoop submodule. Therefore the flink-parent pom exposes the *shading-artifact.name* variable. It is set depending on the requested hadoop dependency set.
To depend on Hadoop, add the following entry to your pom:

```
<dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>${shading-artifact.name}</artifactId>
        <version>${project.version}</version>
</dependency>
```

**NOTE**: It is very important not to put guava, protobuf, netty and asm (== all shaded hadoop dependencies) into the dependencyManagement section of the parent pom. If they are there, they'll influence the version we'll pack into the shaded hadoop jar!

## External Shading: Guava, ASM

External shading hides some of Flink's dependencies to Flink users. This allows users to use whatever version of Guava or ASM they want.
The external shading is controlled in the flink-parent pom, in the maven-shade-plugin.

As per status of the pull request mentioned above, a project depending on: flink-java, flink-streaming-core, flink-streaming-connectors, flink-clients, flink-scala, flink-streaming-scala will see the following Flink dependencies:

```
[INFO] --- maven-dependency-plugin:2.8:list (default-cli) @ debug-quickstart ---
[INFO]
[INFO] The following files have been resolved:
[INFO]    org.objenesis:objenesis:jar:2.1:compile
[INFO]    com.google.protobuf:protobuf-java:jar:2.5.0:compile
[INFO]    com.twitter:bijection-core_2.10:jar:0.7.2:compile
[INFO]    com.rabbitmq:amqp-client:jar:3.3.1:compile
```

```
[INFO]    org.apache.flink:flink-compiler:jar:0.9-SNAPSHOT:compile
[INFO]    com.typesafe.akka:akka-remote_2.10:jar:2.3.7:compile
[INFO]    commons-fileupload:commons-fileupload:jar:1.3.1:compile
[INFO]    org.uncommons.maths:uncommons-maths:jar:1.2.2a:compile
[INFO]    com.google.code.simple-spring-memcached:spymemcached:jar:2.8.4:compile
[INFO]    org.scalamacros:quasiquotes_2.10:jar:2.0.1:compile
[INFO]    stax:stax-api:jar:1.0.1:compile
[INFO]    org.apache.flink:flink-streaming-connectors:jar:0.9-SNAPSHOT:compile
[INFO]    org.apache.flink:flink-streaming-core:jar:0.9-SNAPSHOT:compile
[INFO]    org.tukaani:xz:jar:1.0:compile
[INFO]    org.apache.thrift:libthrift:jar:0.6.1:compile
[INFO]    commons-logging:commons-logging:jar:1.1.1:compile
[INFO]    com.twitter:chill-java:jar:0.5.2:compile
[INFO]    redis.clients:jedis:jar:2.4.2:compile
[INFO]    org.apache.avro:avro-ipc:jar:1.7.6:compile
[INFO]    com.esotericsoftware.kryo:kryo:jar:2.24.0:compile
[INFO]    org.codehaus.jackson:jackson-mapper-asl:jar:1.9.13:compile
[INFO]    de.javakaffee:kryo-serializers:jar:0.27:compile
[INFO]    org.slf4j:slf4j-api:jar:1.7.7:compile
[INFO]    org.apache.commons:commons-pool2:jar:2.0:compile
[INFO]    org.fusesource.leveldbjni:leveldbjni-all:jar:1.8:compile
[INFO]    jline:jline:jar:0.9.94:compile
[INFO]    com.twitter:chill-thrift:jar:0.5.2:compile
[INFO]    org.apache.httpcomponents:httpcore:jar:4.2:compile
[INFO]    log4j:log4j:jar:1.2.17:compile
[INFO]    com.amazonaws:aws-java-sdk:jar:1.8.1:compile
[INFO]    com.twitter:chill-protobuf:jar:0.5.2:compile
[INFO]    com.twitter:bijection-avro_2.10:jar:0.7.2:compile
[INFO]    com.twitter:chill_2.10:jar:0.5.2:compile
[INFO]    com.thoughtworks.paranamer:paranamer:jar:2.3:compile
[INFO]    org.apache.httpcomponents:httpclient:jar:4.2.5:compile
[INFO]    commons-lang:commons-lang:jar:2.5:compile
[INFO]    com.fasterxml.jackson.core:jackson-databind:jar:2.1.1:compile
[INFO]    joda-time:joda-time:jar:2.5:compile
[INFO]    com.twitter:chill-bijection_2.10:jar:0.5.2:compile
[INFO]    com.fasterxml.jackson.core:jackson-core:jar:2.1.1:compile
[INFO]    org.apache.flink:flink-shaded-include-yarn:jar:0.9-SNAPSHOT:compile
[INFO]    org.apache.kafka:kafka_2.10:jar:0.8.2.0:compile
[INFO]    org.eclipse.jetty:jetty-continuation:jar:8.0.0.M1:compile
[INFO]    org.eclipse.jetty:jetty-util:jar:8.0.0.M1:compile
[INFO]    io.netty:netty-all:jar:4.0.24.Final:compile
[INFO]    org.scala-lang:scala-reflect:jar:2.10.4:compile
[INFO]    com.yammer.metrics:metrics-core:jar:2.2.0:compile
[INFO]    commons-cli:commons-cli:jar:1.2:compile
[INFO]    com.google.code.findbugs:jsr305:jar:1.3.9:compile
[INFO]    org.apache.zookeeper:zookeeper:jar:3.4.6:compile
[INFO]    com.fasterxml.jackson.core:jackson-annotations:jar:2.1.1:compile
[INFO]    org.slf4j:slf4j-log4j12:jar:1.7.7:compile
[INFO]    org.eclipse.jetty:jetty-servlet:jar:8.0.0.M1:compile
[INFO]    org.codehaus.jettison:jettison:jar:1.1:compile
[INFO]    com.typesafe.akka:akka-slf4j_2.10:jar:2.3.7:compile
[INFO]    org.xerial.snappy:snappy-java:jar:1.0.5:compile
[INFO]    org.apache.flink:flink-java:jar:0.9-SNAPSHOT:compile
[INFO]    org.apache.flink:flink-streaming-scala:jar:0.9-SNAPSHOT:compile
[INFO]    org.apache.commons:commons-compress:jar:1.4.1:compile
[INFO]    org.apache.commons:commons-lang3:jar:3.3.2:compile
[INFO]    com.esotericsoftware.minlog:minlog:jar:1.2:compile
[INFO]    org.apache.flink:flink-runtime:jar:0.9-SNAPSHOT:compile
[INFO]    org.eclipse.jetty:jetty-server:jar:8.0.0.M1:compile
[INFO]    org.mortbay.jetty:servlet-api:jar:3.0.20100224:compile
[INFO]    commons-collections:commons-collections:jar:3.2.1:compile
[INFO]    io.netty:netty:jar:3.5.12.Final:compile
[INFO]    org.apache.flink:flink-core:jar:0.9-SNAPSHOT:compile
[INFO]    org.apache.kafka:kafka-clients:jar:0.8.2.0:compile
[INFO]    com.101tec:zkclient:jar:0.3:compile
[INFO]    com.github.scopt:scopt_2.10:jar:3.2.0:compile
[INFO]    com.typesafe:config:jar:1.2.1:compile
[INFO]    org.apache.mina:mina-core:jar:2.0.4:compile
[INFO]    org.apache.commons:commons-math:jar:2.2:compile
[INFO]    org.scala-lang:scala-compiler:jar:2.10.4:compile
[INFO]    junit:junit:jar:4.4:compile
```

```
[INFO]    org.apache.flink:flink-scala:jar:0.9-SNAPSHOT:compile
[INFO]    org.apache.sling:org.apache.sling.commons.json:jar:2.0.6:compile
[INFO]    org.codehaus.jackson:jackson-core-asl:jar:1.9.13:compile
[INFO]    commons-io:commons-io:jar:2.4:compile
[INFO]    net.jpountz.lz4:lz4:jar:1.2.0:compile
[INFO]    org.apache.flink:flink-clients:jar:0.9-SNAPSHOT:compile
[INFO]    org.eclipse.jetty:jetty-io:jar:8.0.0.M1:compile
[INFO]    org.apache.avro:avro:jar:1.7.6:compile
[INFO]    com.typesafe.akka:akka-actor_2.10:jar:2.3.7:compile
[INFO]    org.eclipse.jetty:jetty-security:jar:8.0.0.M1:compile
[INFO]    commons-codec:commons-codec:jar:1.3:compile
[INFO]    org.eclipse.jetty:jetty-http:jar:8.0.0.M1:compile
[INFO]    com.twitter:chill-avro_2.10:jar:0.5.2:compile
[INFO]    org.scala-lang:scala-library:jar:2.10.4:compile
```

To add more dependencies to this external shading, add them to the artifactSet in the flink-parent/pom.xml

## Final binary fat jar

Before the pull request mentioned above, Flink was not building a single fat-jar for the binary distribution. This was only done for YARN to reduce the number of files necessary to distribute to all containers.
With this change, we always create one fat-jar containing all flink jars and their dependencies.
If we would stick with the old approach (a jar file for each Flink module), we would end up with a very big binary distribution in the end, because all jars would contain all shaded dependencies (asm, guava etc.).