The Project Split
What does it mean to you?

Aaron Kimball
Cloudera Inc.
July 15, 2009
The Project Split

Images © Encyclopedia of New Zealand (http://www.teara.govt.nz/), the Astigan Society (www.astigan.com)
The Project Split

- Hadoop Core
  - Hadoop Common
  - HDFS
  - MapReduce
Why Split the Project?

- Developer email traffic getting out of hand
  - Several dozen JIRA messages/day
- Project code base getting large
  - 300,000+ lines of Java
- Users interested in deploying HDFS without MapReduce
Some history…

- Decoupling process has been on-going for a while
  - hadoop-site.xml split into mapred-site, hdfs-site, etc.
  - All-purpose “bin/hadoop” now split into “bin/hadoop,” “bin/mapred,” and “bin/hdfs” in trunk
What Happened?

- svn repository for Hadoop-core split into three sub-repositories:
  - hadoop-common
  - hadoop-hdfs
  - hadoop-mapreduce

- HADOOP JIRA project split into three:
  - HADOOP (a.k.a Hadoop “common”)
  - HDFS
  - MAPREDUCE

- Mailing lists split
User-facing mailing lists

- core-user@hadoop.apache.org renamed to common-user
  - If you subscribed to core-user, your subscription was ported
  - Please don’t CC messages to both core-user@ and common-user@ :)

- mapreduce-user, hdfs-user also created
  - These are seeing very little traffic in practice
Developer mailing lists

- For each of {common, hdfs, mapreduce}:
  - dev – general developer commentary. Also issue creation, resolution
  - commits – svn commit hook, wiki updates
  - issues – all JIRA updates (comments, etc.)
... And plenty of subproject lists...

- And \{commits, dev, user\} for each of:
  - avro
  - chukwa
  - hbase
  - hive
  - pig
  - zookeeper

- ... and general@hadoop.apache.org!
Which matter to *you*

- common-user is the de facto community-wide list
  - See also mapreduce-user, hdfs-user
- Subscribe to –dev to keep an eye on development
- Subscribe to –issues and –commits for the firehose
End-user impact

- Nothing for now
  - Stable versions (18, 19, 20) not affected at all
  - Users of the 20 branch will see future releases (e.g. 0.20.1) as a single project

- Longer term…
  - Separate subprojects will release separate tarballs
  - Users of Cloudera’s distribution will install multiple RPMs at the same time
Configuration Changes

- hadoop-site.xml is deprecated in 0.20; you should already be moving to mapred-site, hdfs-site for configuration
  - These same files will work for 0.21+
  - hadoop-site.xml not supported in 0.21
Job Launch Changes

- Starting in 0.21, using “bin/hadoop” for everything is deprecated

- Support for generic bin/hadoop ends by 1.0
API Changes

- New MapReduce API introduced in 0.20
- Version in 0.20 doesn’t fully work
  - See MAPREDUCE-179, MAPREDUCE-565
  - Will be fixed by 0.20.1

- But *after* 0.20.1 is released, it’s time to upgrade your code…
Running Hadoop from Trunk

- Now that Hadoop’s got three different projects, how do they actually connect?

- You’re crazy; don’t do this yet.
- No, really.
- Ok, fine.
- Here’s one way to do this. It’s probably not the 100%-optimal solution.
Running Hadoop: Checkout

- Check out the various repositories

```
svn checkout http://svn.apache.org/repos/asf/hadoop/common/trunk hadoop-common

svn checkout http://svn.apache.org/repos/asf/hadoop/mapreduce/trunk mapred

svn checkout http://svn.apache.org/repos/asf/hadoop/hdfs/trunk hdfs
```
Running Hadoop: Fix a bug

- The `bin/hadoop`, etc scripts don't quite work ;)

- Download the patch at: http://issues.apache.org/jira/browse/HADOOP-6152

- Apply the patch with:
  
  ```bash
  cd path/to/hadoop-common
  patch -p0 < /path/to/HADOOP-6152.patch
  ```
Running Hadoop: Build Hadoop

- In each project directory, run:
  ```
  ant jar
  ```
  ... to build the jars

- In hadoop-common, build the full package:
  ```
  ant package –Djava5.home=/path/to/jdk5 \ 
  –Dforrest.home=/path/to/apache-forrest-0.8
  ```
Running Hadoop: Copy Jars

- From now on, let \texttt{HADOOP\_HOME} be:
  \begin{verbatim}
  hadoop-common/build/hadoop-core-0.21.0-dev
  \end{verbatim}

- Copy the .jar files from \texttt{hdfs/build/} and \texttt{mapred/build/} into \texttt{$HADOOP\_HOME$}

- Copy \texttt{hadoop-core-0.21.0-dev.jar} into \texttt{$HADOOP\_HOME/lib$}
Running Hadoop: Configure

- Edit $HADOOP_HOME/conf/hadoop-env.sh
  - Set JAVA_HOME to your Java installation
- Edit $HADOOP_HOME/conf/mapred-site.xml, hdfs-site.xml, core-site.xml
  - e.g., to set up pseudo-distributed mode.
- Set HADOOP_HDFS_HOME to /path/to/svn/hdfs
  - export HADOOP_HDFS_HOME=/home/aaron/src/hdfs
  - Not the build/hadoop-hdfs-0.21/ subdirectory
  - (I think needing to set this explicitly is a bug...)
Running Hadoop: Run it!

```bash
cd $HADOOP_HOME
bin/hdfs namenode -format
bin/start-dfs.sh
bin/start-mapred.sh
```
Conclusions

- Some turbulence affecting Hadoop developers, not likely end users for now
- Remember to change your mail filter from core-user to common-user
- Future releases will have more fine-grained components