Defining Hadoop

Defining Apache Hadoop

This is a draft, please provide feedback to the hadoop-general mailing list

Excluding quoted text, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

This document is to help clarify ways that other organizations may choose to incorporate Apache Hadoop software into their products or services, and in particular, provide guidance as to appropriate naming styles for third party software related to Apache Hadoop software.

Apache Product Naming

The source code of the Apache™ Hadoop® software is released under the Apache License, as is the source code for the many other Hadoop-related Apache products.

The trademark policy for all Apache Software Foundation (ASF) projects including Hadoop is defined by the Apache Trademark Policy. In particular, much like any other organization's trademark for software products, it is important to understand:

The following uses of ASF trademarks are probably infringing:

- Confusingly similar software product names.
- Software service offerings that are for anything other than official ASF-distributed software.
- Company names that may be associated in customer's minds with ASF or its trademarked project software."

The key point is that the only products that may be called Apache Hadoop or Hadoop are the official releases by the Apache Hadoop project as managed by that Project Management Committee (PMC). It's also important to remember that HADOOP is a registered trademark of the Apache Software Foundation.

Derivative Works

All products which include the official Apache Hadoop artifacts, or included artifacts that are somehow on the source code used to generate these artifacts are *derivative works*. The Apache License enumerates the licensing conditions you must comply with for such derivative works.

Products that are derivative works of Apache Hadoop are not Apache Hadoop, and may not call themselves versions of Apache Hadoop, nor Distributions of Apache Hadoop.

Derivative works may choose to declare that they are *Powered by Apache™ Hadoop®*. Please see our FAQ entry on Powered By naming styles.

Domain Names

The use of the name *Hadoop* in domain names is covered by the Apache Third Party Domain Name Branding Policy.

Compatibility

Some products have been released that have been described as "compatible" with Hadoop, even though parts of the Hadoop codebase have either been changed or replaced. The Apache™ Hadoop® developer team are not a standards body: they do not qualify such (derivative) works as compatible. Nor do they feel constrained by the requirements of external entities when changing the behavior of Apache Hadoop software or related Apache software.

- The definition of the signatures of the Hadoop interfaces and classes is the Apache Source tree, under revision control.
- The definition of semantics of the Hadoop interfaces and classes is the Apache Source tree, including its test classes.
- The verification that the actual semantics of an Apache Hadoop release is compatible with the expected semantics is that the test suites in the Apache codebase pass, and that Hadoop users within the open source community have tested the release running at production scale in their datacenters.
- Bug reports can highlight incompatibility with expectations of community users, and once incorporated into tests form part of the compatibility testing.
- Beta testing of forthcoming releases of Apache Hadoop are of great value in finding unexpected problems, and so not only benefit the product, they benefit the beta testers, who can more confident that their code will work in the final release.
- The Hadoop source tree has annotations to mark any interface as Public or Private, and Stable vs Unstable, independently of the Java public /private annotations.
- Private interfaces can and will change between releases; anyone who uses them must test against future releases, and is strongly urged to join the Hadoop developer mailing lists to track forthcoming changes.
- Interfaces marked as unstable/evolving are just that; they should not be relied on.
- Even interfaces marked as Stable may change -if not in the actual binary signature, then in the semantics.

The key point is this: the Apache Hadoop codebase defines what Apache Hadoop is, so only that codebase can not only assert that it is 100% compatible with Apache Hadoop, but back it up implicitly: Apache Hadoop is compatible with Apache Hadoop, because it is Apache Hadoop.

Other entities may claim that other products (including derivative works) are compatible with Apache Hadoop. The Apache Hadoop development team is not a standards body, and cannot confirm or deny such assertions. All that we can say is "there is no official certification that a product is compatible with Hadoop, other than when a release of the Apache source tree is declared a new release of Apache Hadoop itself".

For background on this, please consult the email thread Defining Hadoop Compatibility in the hadoop-general list, bearing in mind that the participants in the discussion are engineers, not lawyers or trademark experts, so their opinions cannot be considered normative.

Examples

Here are some example naming/branding options and their issues. "Automotive Joe" is an entirely fictional character and bears no resemblance to any individual or company.

INAPPROPRIATE: Automotive Joe's Hadoop

"AJH is the first version of Hadoop for the Automotive Industry!"

Bad. Hadoop is used in a product name, which infringes on Apache's Hadoop mark. Additionally, there are no "versions of Hadoop," except those that Apache releases. What Joe is selling is something "powered by Hadoop", or a derivative work, but it isn't Hadoop. Finally, the acronym "AJH" is derived from Hadoop. Even if the product title is changed, that acronym is likely to be infringing.

INAPPROPRIATE: Automotive Joe's Hadoop Distribution

"Which Hadoop distribution should you use? Automotive Joe's Distribution!"

There is only one distribution of Hadoop: Apache Hadoop. Everything else is a derivative work.

Yes, we know about CDH, but that acronym has been grandfathered in -its product description no longer describes itself as a distribution of Hadoop. It is a distribution "that includes Apache Hadoop".

INAPPROPRIATE: "Automotive Hadoop(TM)" by Joe's Automotive

"Automotive Hadoop" is a trademark of Joe's Automotive."

This is a clear infringement on Apache's Hadoop registered mark, a mark held in many countries.

INAPPROPRIATE: Camshaft: it's a Hadoop for the Automotive industry

It's good that Joe has created his own product name and brand, but saying "a Hadoop" is trouble. If it does contain Apache Hadoop-related artifacts, then it breaks the trademark rules. If it doesn't contain ASF code, then it falls foul of the Generic Trademark problem: the ASF don't want their products to be generified, and will send a note reminding Joe of their rights and obligations.

APPROPRIATE: Camshaft: Joe's datamining solution for the Automotive industry

"Do you know where your trucks are? With Camshaft, the power of Apache Hadoop can be used to examine your trucks logs and show up unusual actions and identify improvements in scheduling, improving time and fuel economy. "

Good: it defines a new product "Camshaft", and opts to use the Apache Hadoop brand to emphasize its heritage. The marketing text sells the product.

APPROPRIATE: Automotive Joe's "Apache Hadoop for Automotive Engineers"

"Continuing Automotive Joe's best selling series, including the popular titles "Spark Gap tuning" and "Datacenter fabric: architecture and implementation", the book "Hadoop for Automotive Engineers" explains Apache Hadoop in an easy and practical way. As with the rest of the series, the cover is designed to be easy to wipe oil off. "

Good: provided it credits Apache properly, this appears to be a good book title. Furthermore, because it's the "Automotive Joe" book series, and not "Automotive Joe's Hadoop" series, the series doesn't infringe anything. Please see our FAQ entry on using Apache marks in book titles.

INAPPROPRIATE: Automotive Joe: Hadooping the motor industry

Bad: This is into the word of Generic trademarks again. Hadoop is a noun, not a verb or an adjective.

INAPPROPRIATE: Crankshaft: Automotive Joe's complete rewrite of Hadoop"

Bad because it isn't Apache Hadoop. It's hard to say "rewrite", and better to discuss the features.

Better to say "Crankshaft: a Big Data engine and filesystem that resembles Apache Hadoop"

INAPPROPRIATE: Automotive Joe's Hadoop filesystem

"Automotive Joe's Hadoop filesystem makes Hadoop installations faster and more reliable!"

Hadoop does support multiple filesystems, and has a reasonably stable interface for them. Example filesystems include Amazon S3, POSIX-compatible native filesystems, and others. For this reason, the only "Hadoop" filesystem that should use Hadoop in its brand name is "HDFS": Hadoop Filesystem, but any other filesystem is welcome to declare their support for Hadoop.

The Apache Hadoop project has attempted to define the behavior of HDFS and so inform other filesytem implementors what that they need to do to integrate with Hadoop, with tests for this in the Hadoop source tree and Apache Bigtop. Following the HCFS work will help you integrate with Hadoop —but it does not grant any rights to product naming, or the right to state categorically that your filesystem is totally compatible with the behavior of HDFS expected by program.

APPROPRIATE: Automotive Joe's JoeFS filesystem -with support for Apache Hadoop

"JoeFS now supports Apache Hadoop! Simply by adding the JoeFS binding JAR files to your Hadoop installation, your MapReduce jobs can run against JoeFS. This offers a natively mountable filesystem, better support for small files which helps Apache Pig jobs. With stable append operation, HBase is also supported."

This is good because it creates its own filesystem brand (which is wider than just Hadoop), shows how easy it is to switch to it, and states some clear benefits of using JoeFS rather than HDFS.

Appropriate: Crankshaft: Joe's filesystem and MapReduce Engine

"Gearbox is Automotive Joe's distributed filesystem, on top of which you can run Apache Hadoop's MapReduce engine version 0.21".

Provided that it really is the Apache 0.21 distribution's JARs that run against the Gearbox filesystem, this seems good. If Automotive Joe's development team has had to make changes ~~rather than just add a new filesystem support JAR~~ then the derivative work naming rules will kick in.

INAPPROPRIATE: Automotive Joe's Crankshaft: 100% compatible with Apache Hadoop

Bad, because "100% compatible" is a meaningless statement. Even Apache releases have regressions; cases were versions are incompatible *even when the Java interfaces don't change*. A statement about compatibility ought to be qualilified "Certified by Joe's brother Bob as 100% compatible with Apache Hadoop(TM)". In the US, the marketing team may be able to get way with the "100% compatible" claim, but in some EU countries, sticking that statement up your web site is a claim that residents can demand the vendor justifies, or take it down.

OK: Automotive Joe's Crankshaft: like Apache Hadoop only faster

This could be a defensible statement, though saying "3.4X faster" isn't. Yes, your app may have terasorted on your 20 node cluster faster than the published benchmarks for Hadoop 0.20.1, but remember that many of the big Hadoop users don't publicise the fact ~~let alone their benchmarks~~ and terasorting isn't the primary role of their cluster. Furthermore Hadoop is evolving, and your statements may soon be invalid.

Finally, criticising Hadoop is not polite. If you feel you ever have need to work with the Hadoop developers, this isn't a good way to start building a relationship.

INAPPROPRIATE: Automotive Joe's Hadoop to Go!

"Automotive Joe offers you a Hadoop as a Service offering, where you can pay by the hour for our custom Hadoop-enabled truck to deliver you on-demand Hadoop."

No Hadoop-as-a-Service offering can use Hadoop in it's product name. Notice how Amazon call theirs "Elastic MR"? That lets them create their own brand, and stops them trying to trade off the Apache Hadoop brand.

INAPPROPRIATE: Automotive Joe's Apache Hadoop Storage Infrastructure

"Lots of datacenters are moving to container-hosted racks, but only Automotive Joe delivers a Hadoop-compatible storage system built into a truck"

There are a number of filesystems that work with Apache Hadoop as well as the location-aware HDFS: anyone is free to implement the Hadoop filesystem interfaces, and so provide a binding to their new or existing filesystem. However, this doesn't mean that the vendor can use Hadoop in the product name.

Ask First: Automotive Joe's Hadoop Party Fest

"In a three day event, attendees will learn about leading edge Hadoop trends and Spark Gap tuning, while consuming large quantities of fine beverages."

Review Apache 3rd party events page and talk to the Apache Conference Planning team (concom@apache.org) before using Hadoop in conference titles.

Ask First: Automotive Joe's Certification Program for Apache Hadoop

Talk to VP, Brand Management for the ASF on trademarks@hadoop.apache.org first.

TALK TO THE ASF FIRST : Press Release Automotive Joe: donating their crankshaft technologies to Apache Hadoop

"Today, Automotive Joe released the interfaces to enable the Apache Hadoop team to run Hadoop against Automotive Joe's Crankshaft filesystem"

It's not appropriate to make press releases about Apache software unless you have the permission of the ASF Press Team. They can advise you on content, and will help you co-ordinate a press release from the ASF too. Announcing some contribution or milestone in an Apache project without co-ordinating it with them isn't a good way to collaborate with the ASF.

Additionally, it's only when code is checked in to the source tree that something becomes part of Hadoop. Until then, it's a JIRA issue that runs a serious risk of being ignored unless it is compelling enough to the entire Hadoop community. To get your patches in:

- Work against trunk, not with a past release.
- Get on the -user lists and be helpful, to show that you care about the product.
- Get on the -dev mailing lists, and participate in the discussions, to show that you understand the technology. That can include reviewing other patches, and testing them. Finding problems in other people's patches helps you get a feel for the process as well as the code.
- Make sure your patch builds on the supported Hadoop OS/Java releases, meets the style guidelines, and adds new tests.
- Make sure your patch scales. This is a big fear of some of the major Hadoop users: that a change to the scalability bottlenecks of the system -the Job Tracker and Name Node - may work on Joe's ten machine cluster, but hurt scale at the high end. Be careful when going near this code, especially if you are unknown and untrusted.
- Don't think that submitting some interface as a JIRA issue is a success worthy of publicity. Wait until it has been committed -and work with the entire -dev team to get that to happen, bearing in mind they are busy with lots of other issues, and not all full time developers.

PLEASE TALK TO THE ASF FIRST : Press Release Automotive Joe's crankshaft technologies now support Apache Hadoop

It is, of course, perfectly reasonable to make a press release about your own software, as long as you don't imply any endorsement from the Apache Hadoop project.

Their Apache Press Team's details are on the ASF Press Team love to be involved in such announcements, as they can provide some advice and co-ordinate your announcement with matching announcements from the ASF itself.