# BigtopProposal

## Bigtop - Apache Hadoop Ecosystem Packaging and Test

### Abstract

Bigtop - a project for the development of packaging and tests of the Hadoop ecosystem.

### Proposal

The primary goal of Bigtop is to build a community around the packaging and interoperability testing of Hadoop-related projects. This includes testing at various levels (packaging, platform, runtime, upgrade, etc...) developed by a community with a focus on the system as a whole, rather than individual projects.

Build, packaging and integration test code that depends upon official releases of the Apache Hadoop-related projects (HDFS, MapReduce, HBase, Hive, Pig, ZooKeeper, etc...) will be developed and released by this project. As bugs and other issues are found we expect these to be fixed upstream.

### Background

The initial packaging and test code for Bigtop was developed by Cloudera to package projects from the Apache Hadoop ecosystem and provide a consistent, inter-operable framework.

### Rationale

Hadoop defines itself as:

```
The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. Hadoop
includes these subprojects:

* Hadoop Common: The common utilities that support the other Hadoop subprojects.
* HDFS: A distributed file system that provides high throughput access to application data.
* MapReduce: A software framework for distributed processing of large data sets on compute clusters.
```

There are also several other Hadoop-related projects at Apache. Some TLP examples include HBase, Hive, Mahout, ZooKeeper, and Pig. There are also several new projects in the Incubator such as HCatalog, Hama and Sqoop.

From a packaging and deployment perspective, the current loosely-coupled nature of the project has limitations:

1. Insufficient building against trunk versions of dependent projects (in the style of Apache Gump).
2. Insufficient testing against the trunk versions of dependent projects.
3. No consistent packaging for the Linux servers which provide the main Hadoop datacenter platform.
4. No functional testing against multi-machine clusters as part of the regular automated build process. This is due to a lack of a physical or virtual Hadoop cluster for testing, and not enough test suites designed to run against a live cluster with known datasets.

The intent of this project is to build a community where the projects are brought together, packaged, and tested for interoperability.

Projects such as Apache Whirr (incubating), which deploy and use a collection of Hadoop-related projects, would benefit from the interoperability testing done by Bigtop, rather than picking and testing project combinations themselves.

### Initial Goals

Much of the code for Bigtop has been released by Cloudera under the Apache 2.0 license for over two years.

Some current goals include:

- create a set of packages for the Hadoop ecosystem, over a wide range of platforms
- interoperability test these projects
- document project sets that are known to work well together

Bigtop's release artifact would consist of a single tarball of packaging and test code that, when built, would produce source and binary Linux packages for the upstream projects.

## Current Status

### Meritocracy

Bigtop was originally developed and released as an open source packaging infrastructure, CDH, by Cloudera.

## Community

The community is primarily the original developers at Cloudera, however a number of contributions to the packaging specifications have been accepted from outside contributors. Growing a diverse community is the main reason to bring Bigtop to the Apache Incubator.

## Core Developers

The core developers for Bigtop project are:

- Andrew Bayer has extensive expertise with build tools, specifically Jenkins continuous integration and Maven.
- Peter Linnell has contributed to the RPM packaging.
- Bruno Mahé has overseen much of the development of the RPM and Debian packaging system.
- Roman Shaposhnik and Konstantin Boudnik designed and implemented the system testing framework.

Many of the committers to the Bigtop project have contributed towards Hadoop or related Apache projects (Alejandro Abdelnur, Konstantin Boudnik, Eli Collins, Alan Gates, Patrick Hunt, Steve Loughran, Owen O'Malley, John Sichi, Michael Stack, Tom White) and are familiar with Apache principals and philosophy for community driven software development.

## Alignment

We expect projects in Bigtop to be drawn from Hadoop and related projects at Apache. Bigtop will complement these projects (Hadoop, Pig, Hive, HBase, etc...) by providing an environment for contributors interested in building more complex data processing pipelines to work together integrating more than a single project into a well-tested whole.

# Known Risks

## Orphaned Products

The contributors are leading vendors of Hadoop-based technologies and have a long standing in the Hadoop community. There is minimal risk of this work becoming non-strategic and the contributors are confident that a larger community will form within the project in a relatively short space of time.

## Inexperience with Open Source

All code developed for Bigtop has been open sourced under the Apache 2.0 license. Most committers of Bigtop project are intimately familiar with the Apache model for open-source development and are experienced with working with new contributors.

## Homogeneous Developers

The initial set of committers is from a small set of organizations and numerous existing Apache projects. We expect that once approved for incubation, the project will attract new contributors from more organizations and will thus grow organically.

## Reliance on Salaried Developers

It is expected that Bigtop will be developed on salaried and volunteer time, although all of the initial developers will work on it mainly on salaried time.

## Relationships with Other Apache Products

Bigtop depends upon other Apache Projects including Apache Hadoop, Apache HBase, Apache Hive, Apache Pig, Apache Zookeeper, Apache Thrift, Apache Avro, Apache Whirr. The build system uses Apache Ant and Apache Maven.

## An Excessive Fascination with the Apache Brand

We would like Bigtop to become an Apache project to further foster a healthy community of contributors and consumers around interoperability, testing and packaging of Hadoop projects. Since Bigtop directly interacts with many Apache Hadoop-related projects and solves important problems of many Hadoop users, residing in the the Apache Software Foundation will increase interaction with the larger community.

# Documentation

- Bigtop will develop its own documentation detailing how to build, test, install, configure and debug.

# Initial Source

- https://github.com/cloudera/bigtop

## Source and Intellectual Property Submission Plan

- The initial source is already licensed under the Apache License, Version 2.0.

https://github.com/cloudera/bigtop

## External Dependencies

The required external dependencies are all Apache License or compatible licenses.

## Cryptography

Bigtop doesn't use cryptography itself, however Hadoop projects use standard APIs and tools for SSH and SSL communication where necessary.

# Required Resources

## Mailing lists

- bigtop-private (with moderated subscriptions)
- bigtop-dev
- bigtop-commits
- bigtop-user

## Subversion Directory

https://svn.apache.org/repos/asf/incubator/bigtop

## Issue Tracking

JIRA BIGTOP (Bigtop)

## Other Resources

The existing code already has unit and integration tests so we would like a Jenkins instance to run them whenever a new patch is submitted. This can be added after project creation.

To test RPM & deb install/uninstall and upgrade, it is useful to have a set of Virtual Machine images in known states, and servers that can bring them up. It should be possible to use Apache Whirr to choreograph the VM setup/teardown, so these tests could be performed against VMs on developer desktops or large scale VM-hosting platforms. For the latter, VM hosting time would be appreciated.

# Initial Committers

- Alejandro Abdelnur (tucu at cloudera dot com)
- Andre Arcilla (arcilla at yahoo-inc dot com)
- Andrew Bayer (abayer at cloudera dot com)
- Konstantin Boudnik (cos at apache dot org)
- Eli Collins (eli at apache dot org)
- Travis Crawford (travis at twitter dot com)
- Bruno Mahé (bruno at cloudera dot com)
- Alan Gates (gates at apache dot org)
- Patrick Hunt (phunt at apache dot org)
- Peter Linnell (plinnell at cloudera dot com)
- Steve Loughran (stevel at apache dot org)
- Owen O'Malley (omalley at apache dot org)
- James Page (James.page at canonical dot com)
- Roman Shaposhnik (rvs at cloudera dot com)
- John Sichi (jvs at apache dot org)
- Michael Stack (stack at apache dot org)
- Tom White (tomwhite at apache dot org)
- Andrei Savu (asavu at apache dot org)
- Edward J. Yoon (edwardyoon at apache dot org)

# Affiliations

- Alejandro Abdelnur, Cloudera
- Andre Arcilla, Yahoo! Inc.
- Andrew Bayer, Cloudera
- Konstantin Boudnik, free lancer
- Eli Collins, Cloudera
- Travis Crawford, Twitter
- Bruno Mahé, Cloudera
- Alan Gates, Yahoo!
- Patrick Hunt, Cloudera
- Peter Linnell, Cloudera
- Steve Loughran, HP Laboratories
- Owen O'Malley, Yahoo!
- James Page, Canonical
- Roman Shaposhnik, Cloudera
- John Sichi, Facebook
- Michael Stack, StumbleUpon
- Tom White, Cloudera
- Andrei Savu, Adobe
- Edward J. Yoon, Korea Telecom

# Sponsors

## Champion

- Patrick Hunt

## Nominated Mentors

- Patrick Hunt
- Tom White
- Owen O'Malley
- Alan Gates
- Steve Loughran

## Sponsoring Entity

- Apache Incubator PMC