



O'REILLY®

oscon

open source convention

oscon.com

#oscon

Getting Hadoop, Hive and HBase up and running in less than 15 mins

OSCON 2013

Mark Grover

@mark_grover, Cloudera Inc.

www.github.com/markgrover/oscon-bigtop

About me

- Committer on Apache Bigtop
- Contributor to Apache Hadoop, Hive, Sqoop, Flume
- Software Engineer at Cloudera

Bart

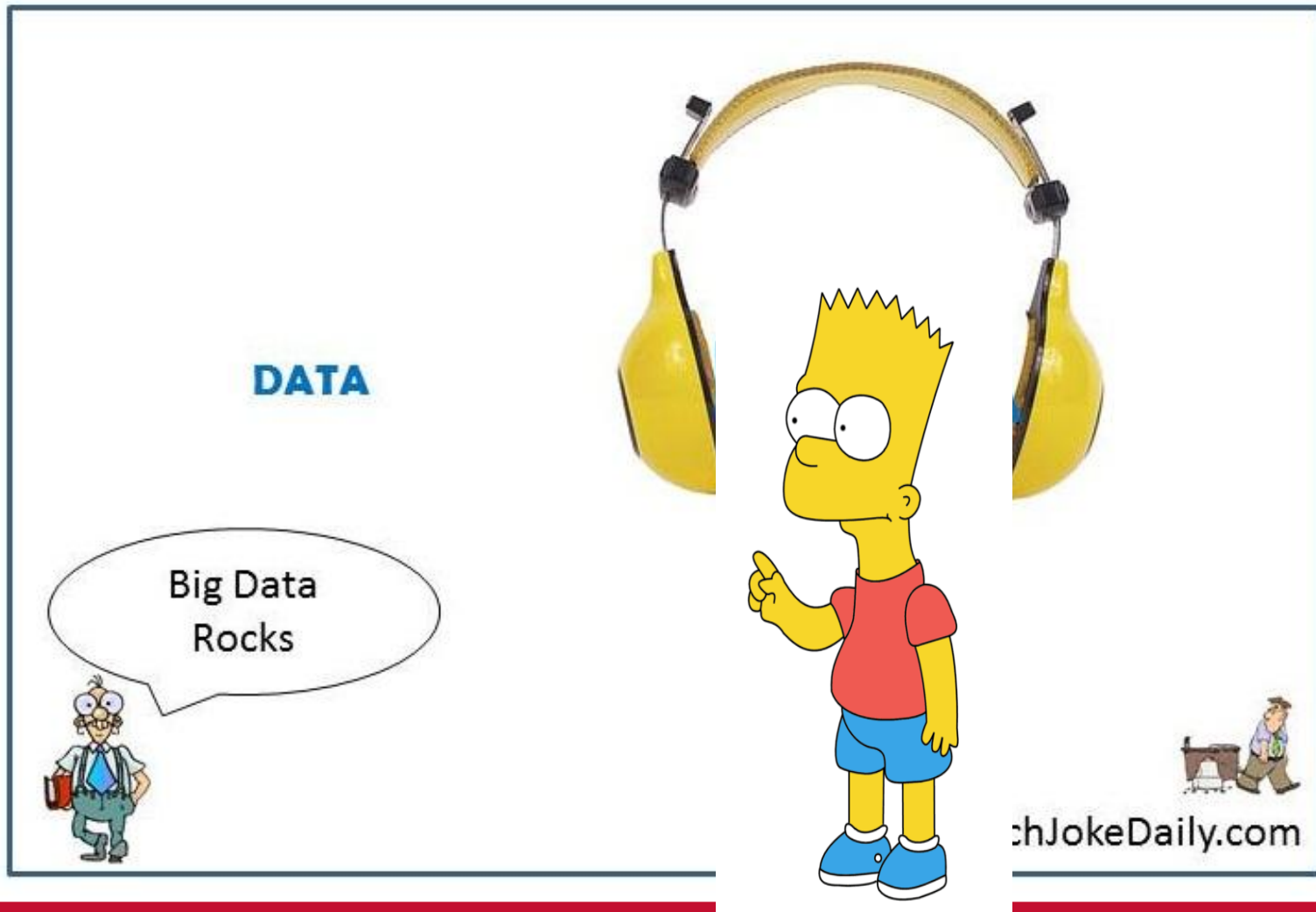


fromemptyhands.blogspot.com

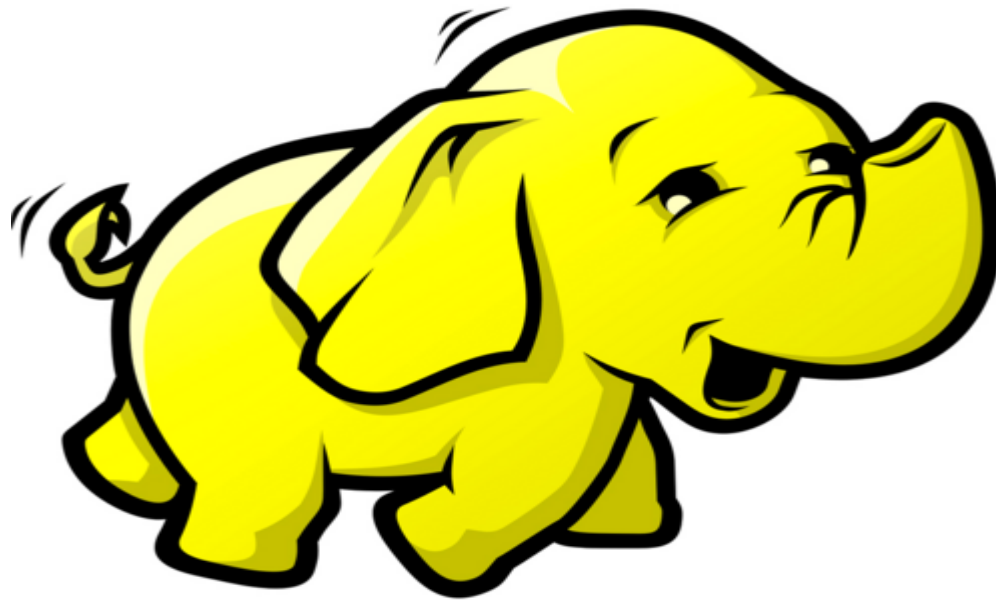
Big Data Rocks



Big Data Rocks



Bart meets the elephant

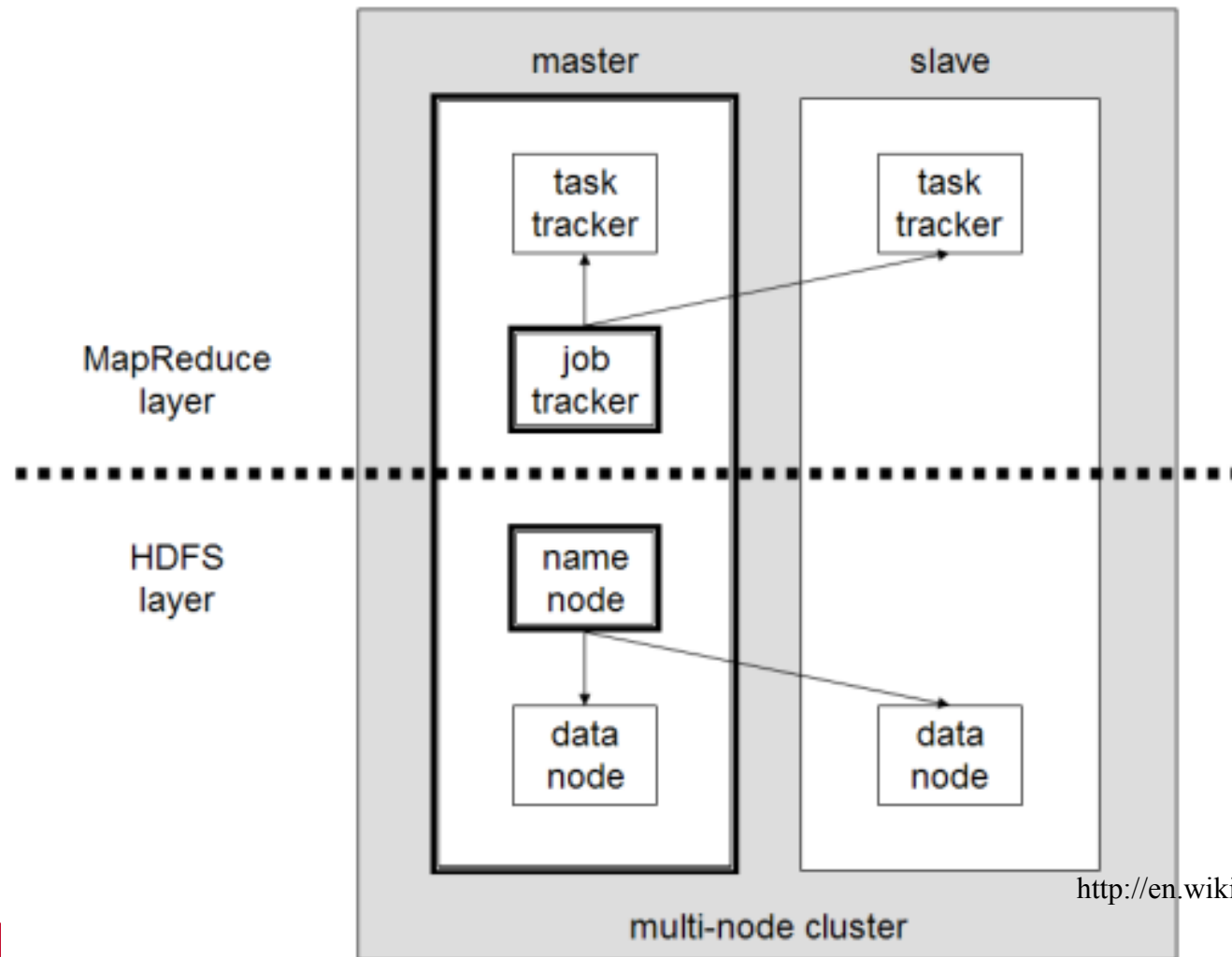


Apache Hadoop!!!

What is Hadoop?

- Distributed batch processing system
- Runs on commodity hardware

What is Hadoop?



http://en.wikipedia.org/wiki/Apache_Hadoop

Installing Hadoop on 1 node

- Download Hadoop tarball
- Create working directories
- Populate configs: core-site.xml, hdfs-site.xml...
- Format namenode
- Start hadoop daemons
- Run MR job!

Grrrr....

Error: JAVA_HOME is not set and
could not be found.

Oops...Environment variables

- Set up environment variables

```
$ export JAVA_HOME=/usr/lib/jvm/default-  
java
```

```
$ export HADOOP_MAPRED_HOME=/opt/hadoop
```

```
$ export HADOOP_COMMON_HOME=/opt/hadoop
```

```
$ export HADOOP_HDFS_HOME=/opt/hadoop
```

```
$ export YARN_HOME=/opt/hadoop
```

```
$ export HADOOP_CONF_DIR=/opt/hadoop/conf
```

```
$ export YARN_CONF_DIR=/opt/hadoop/conf
```

Wait.....What?

```
org.apache.hadoop.security.AccessControlExc  
ption: Permission denied: user=vagrant,  
access=WRITE,  
inode="/":hdfs:supergroup:drwxr-xr-x  
at  
org.apache.hadoop.hdfs.server.namenode.FSP  
ermissionChecker.check(FSPermissionChecker  
.java:205)  
at  
org.apache.hadoop.hdfs.server.namenode.FSP  
ermissionChecker.check(FSPermissionChecker  
.java:186)
```

Oops...HDFS directories for YARN

```
■ sudo -u hdfs hadoop fs -mkdir -p /user/  
$USER
```

```
sudo -u hdfs hadoop fs -chown $USER:$USER  
user/$USER
```

```
sudo -u hdfs hadoop fs -chmod 770 /user/  
$USER
```

```
sudo -u hdfs hadoop fs -mkdir /tmp
```

```
sudo -u hdfs hadoop fs -chmod -R 1777 /tmp
```

```
sudo -u hdfs hadoop fs -mkdir -p /var/log/  
hadoop-yarn
```

```
sudo -u hdfs hadoop fs -chown  
yarn:mapred /var/log/hadoop-yarn
```

•
•

Running a MR job

- Tada!

Frustrating!



hwalls.com

Wouldn't it be nice...

to have an easier process to **install** and **configure**
hadoop

Hive mailing list

On Thu, Jan 31, 2013 at 11:42 AM, Bart Simpson <bart@thesimpsons.com> wrote:

Howdy Hivers!

Can you tell me if the latest version of Hadoop (X) is supported with the latest version of Hive (Y)?

Hive

*On Thu, Jan 31, 2013 at 12:01 PM, The Hive Dude
<thehivedude@gmail.com> wrote:*

We only tested latest Hive version (Y) with an older Hadoop version (X') but it **should** work with the latest version of Hadoop (X).

Yours truly,

The Hive Dude

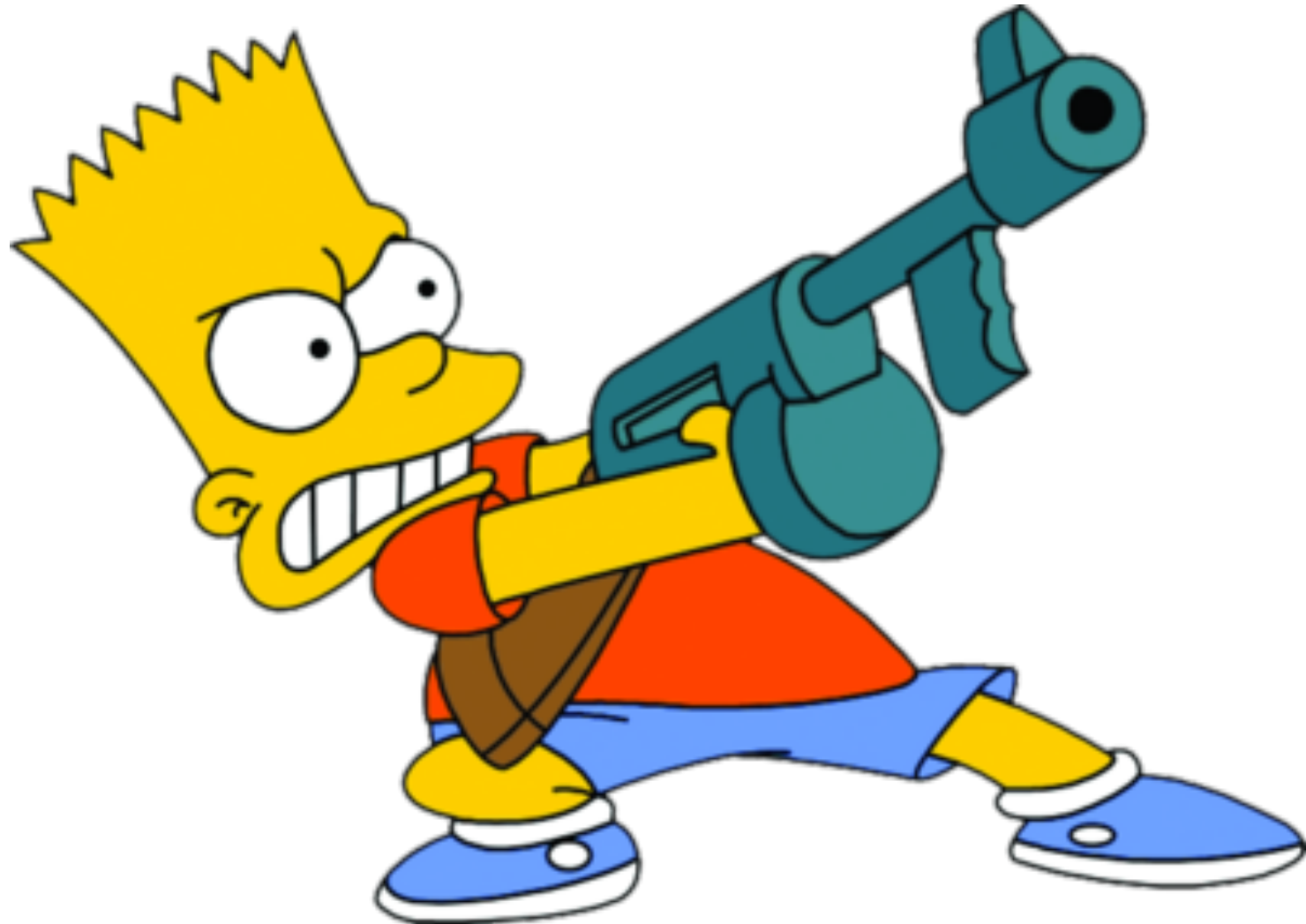
Latest Hive with Latest Hadoop

```
Job running in-process (local Hadoop)
Hadoop job information for null: number of
mappers: 1; number of
reducers: 0
2012-06-27 09:08:24,810 null map = 0%,
reduce = 0%
Ended Job = job_1340800364224_0002 with
errors
Error during job, obtaining debugging
information...
```

.

.

Grr....



Wouldn't it be nice...

If someone **integration tested** these projects

So what do we see?

Installing and configuring hadoop ecosystem is hard

There is lack of integration testing

So what do we see?

Installing and configuring a bigtop ecosystem is hard

Lack of integration testing

Apache Bigtop

Apache Bigtop

Makes installing and configuring hadoop projects
easier

Integration testing among various projects

Apache Bigtop




- Apache Top Level Project
- Generates packages of various Hadoop ecosystem components for various distros
- Provides deployment code for various projects
- Convenience artifacts available e.g. `hadoop-conf-pseudo`
- Integration testing of latest project releases

Installing Hadoop (without Bigtop)

- Download Hadoop tarball
- Create working directories
- Populate configs: core-site.xml, hdfs-site.xml...
- Format namenode
- Start hadoop daemons
- Set environment variables
- Create directories in HDFS
- Run MR job!

Installing Hadoop (without Bigtop)

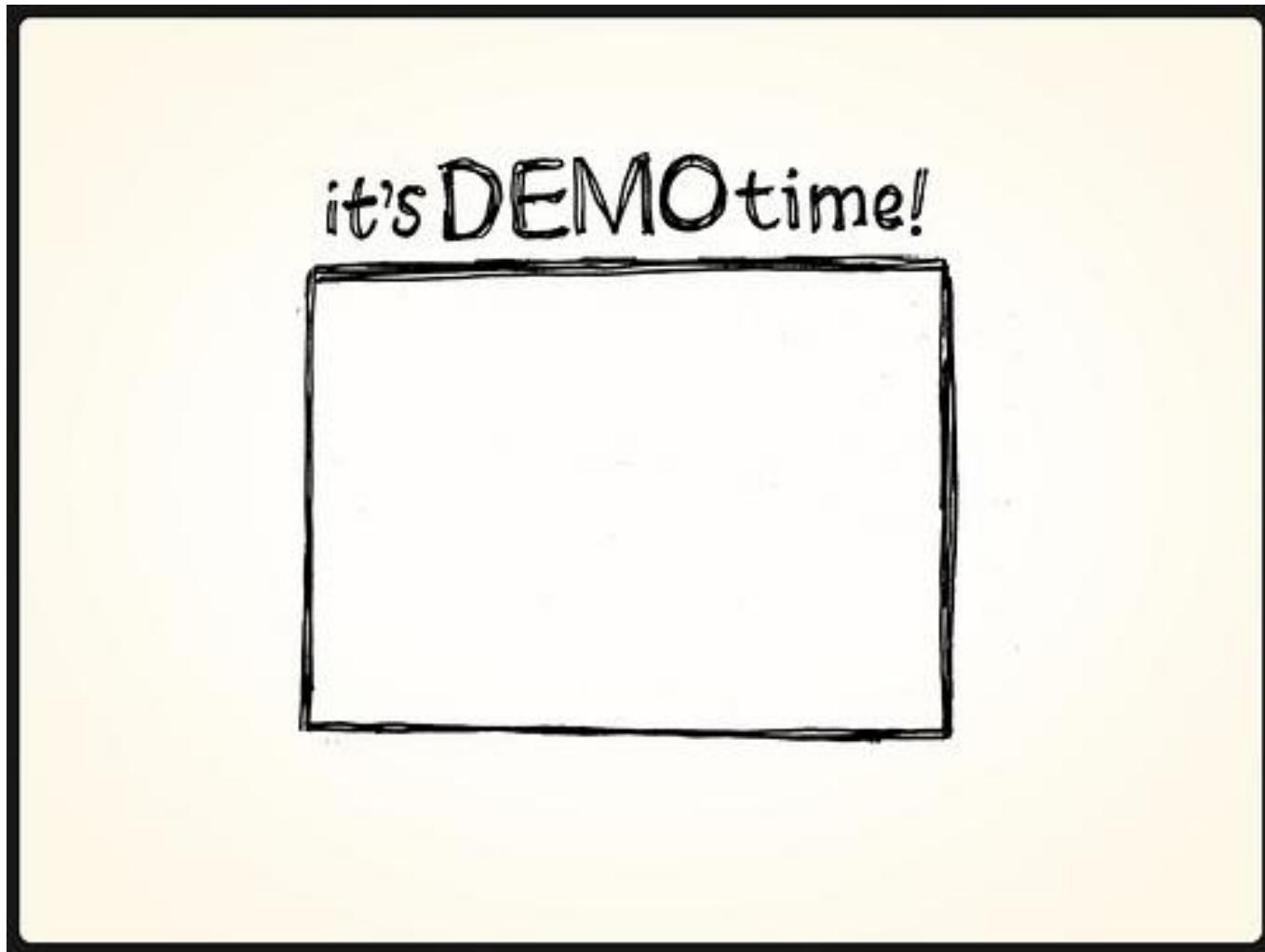
- Download Hadoop tarball
 - Create directories
 - Populate `hadoop-site.xml`
 - Format namenode
 - Start hadoop
 - Run MR
- 

Installing Hadoop (with Bigtop)

```
sudo apt-get install hadoop-conf-pseudo  
sudo service hadoop-hdfs-namenode init  
sudo service hadoop-hdfs-namenode start  
sudo service hadoop-hdfs-datanode start  
. /usr/lib/hadoop/libexec/init-hdfs.sh
```

Run your MR job!

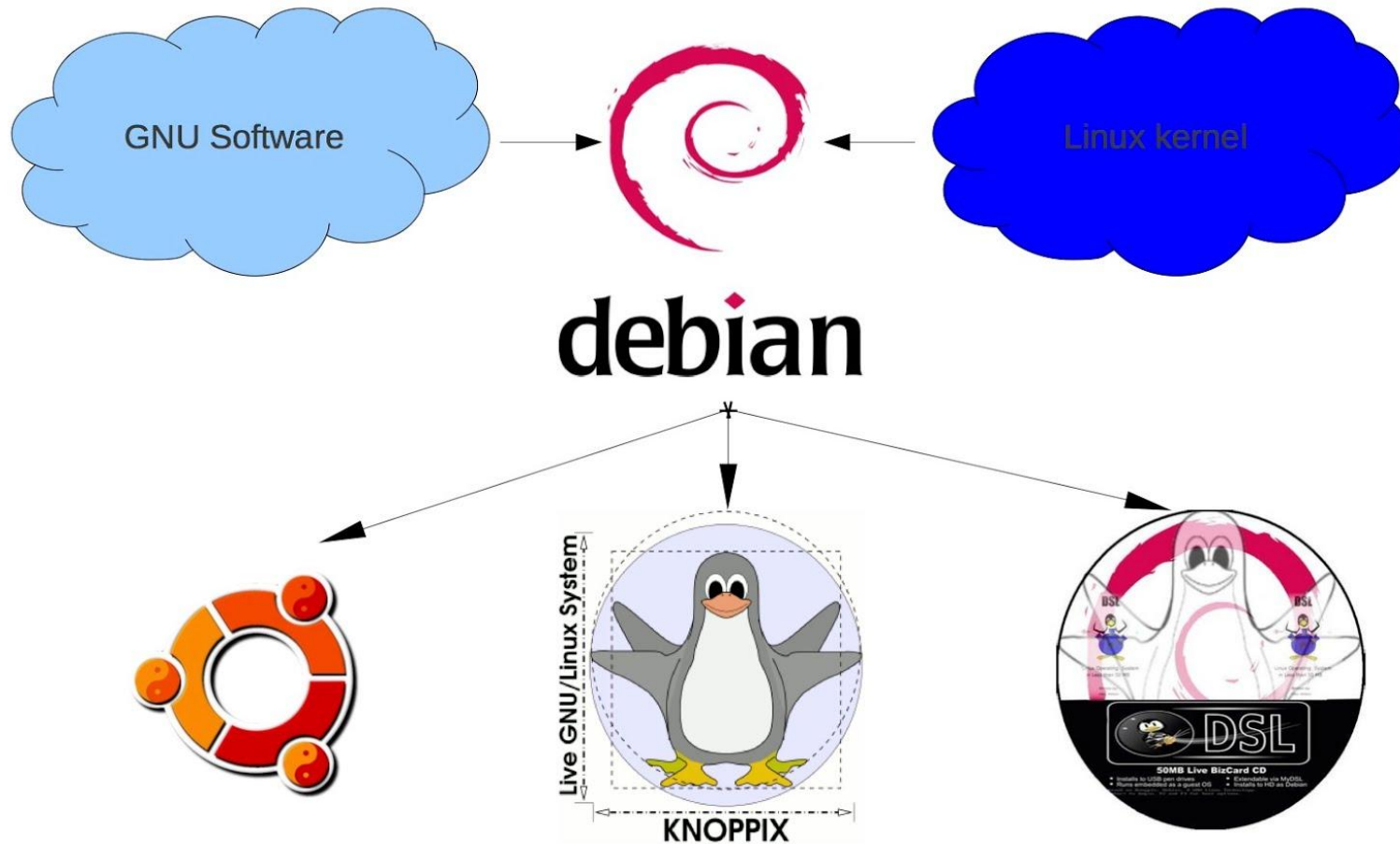
Demo



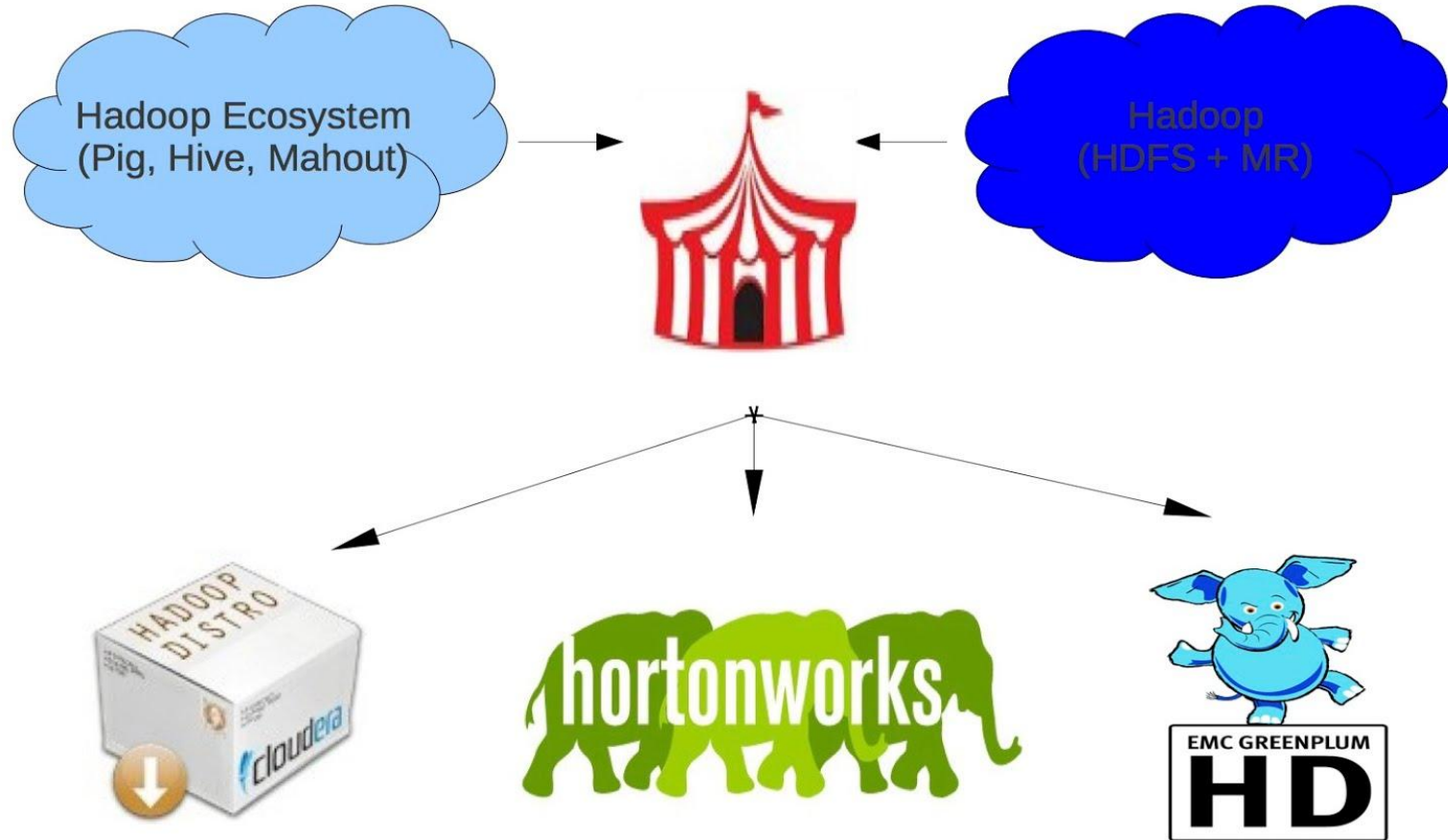
Integration testing

- Most individual projects don't perform integration testing
 - No HBase tarball that runs out of box with Hadoop2
- Complex combinatorical problem
 - How can we test that all versions of project X work with all versions of project Y?
 - We can't!
- Testing based on
 - Packaging
 - Platform
 - Runtime
 - Upgrade

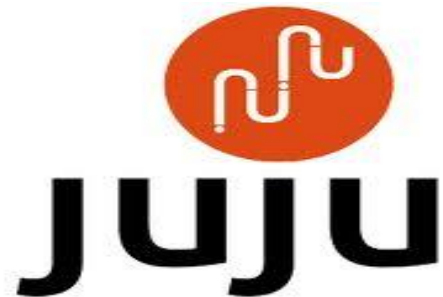
What Debian did to Linux



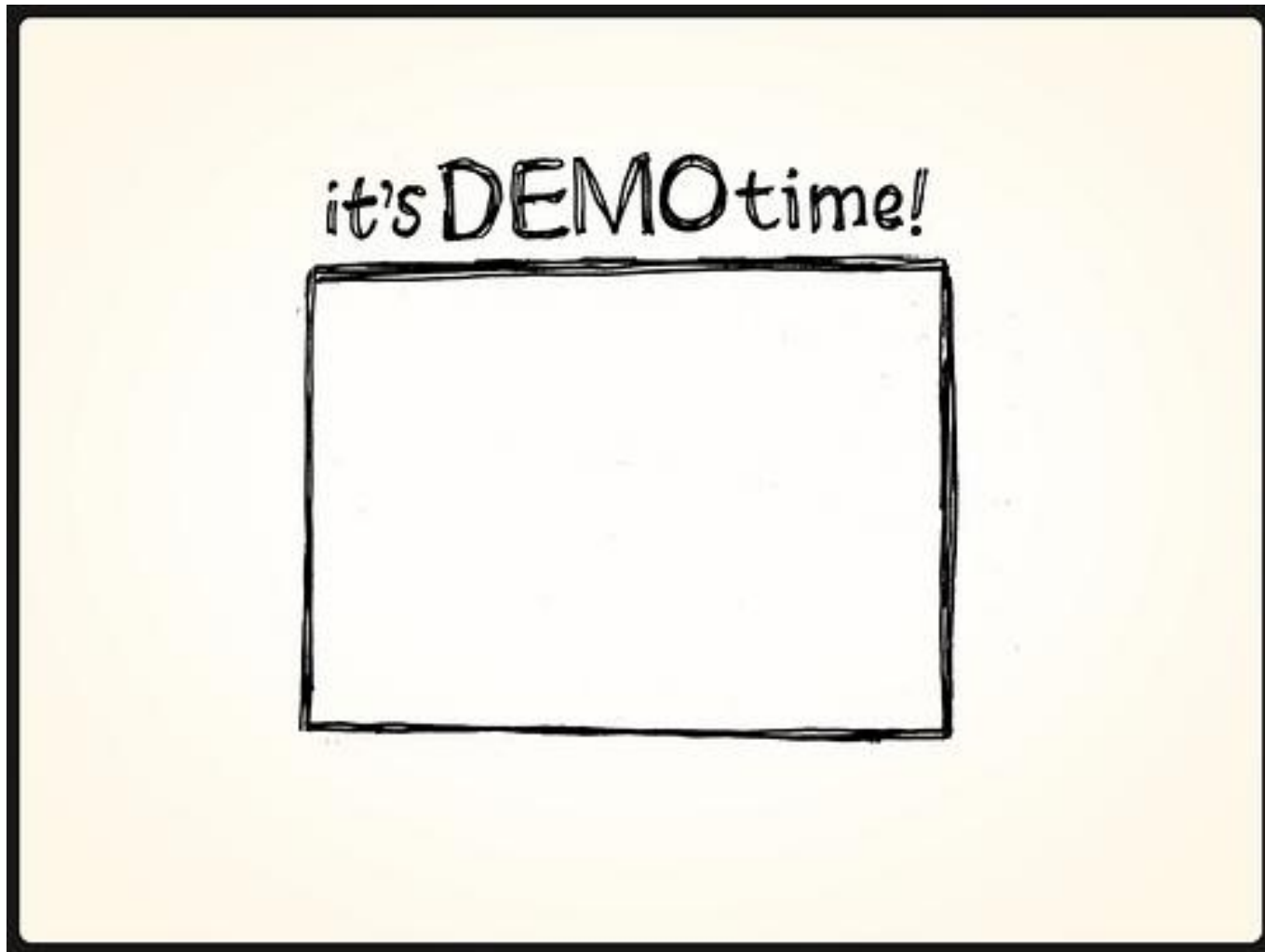
What Bigtop is doing to Hadoop



Who uses Bigtop?



Demo



But MongoDB is web scale, are you?



Xtranormal.com

Deploying larger clusters with Bigtop

- Puppet recipes for various components (Hadoop, Hive, HBase, etc.)
- Integration with Apache Whirr for easier testing

Why use Bigtop?

- Easier deployment of tested upstream artifacts
- Artifacts are integration tested!
- A distribution of the community, by the community, for the community

Apache Bigtop

Makes installing and configuring hadoop projects
easier

Integration testing among various projects

Questions?

- Twitter:
mark_grover
- Code for the demo

<http://github.com/markgrover/oscon-bigtop>

