

Lessons in Apache Software integration

Roman Shaposhnik
rvs@apache.org
Cloudera Inc.

\$ whoami

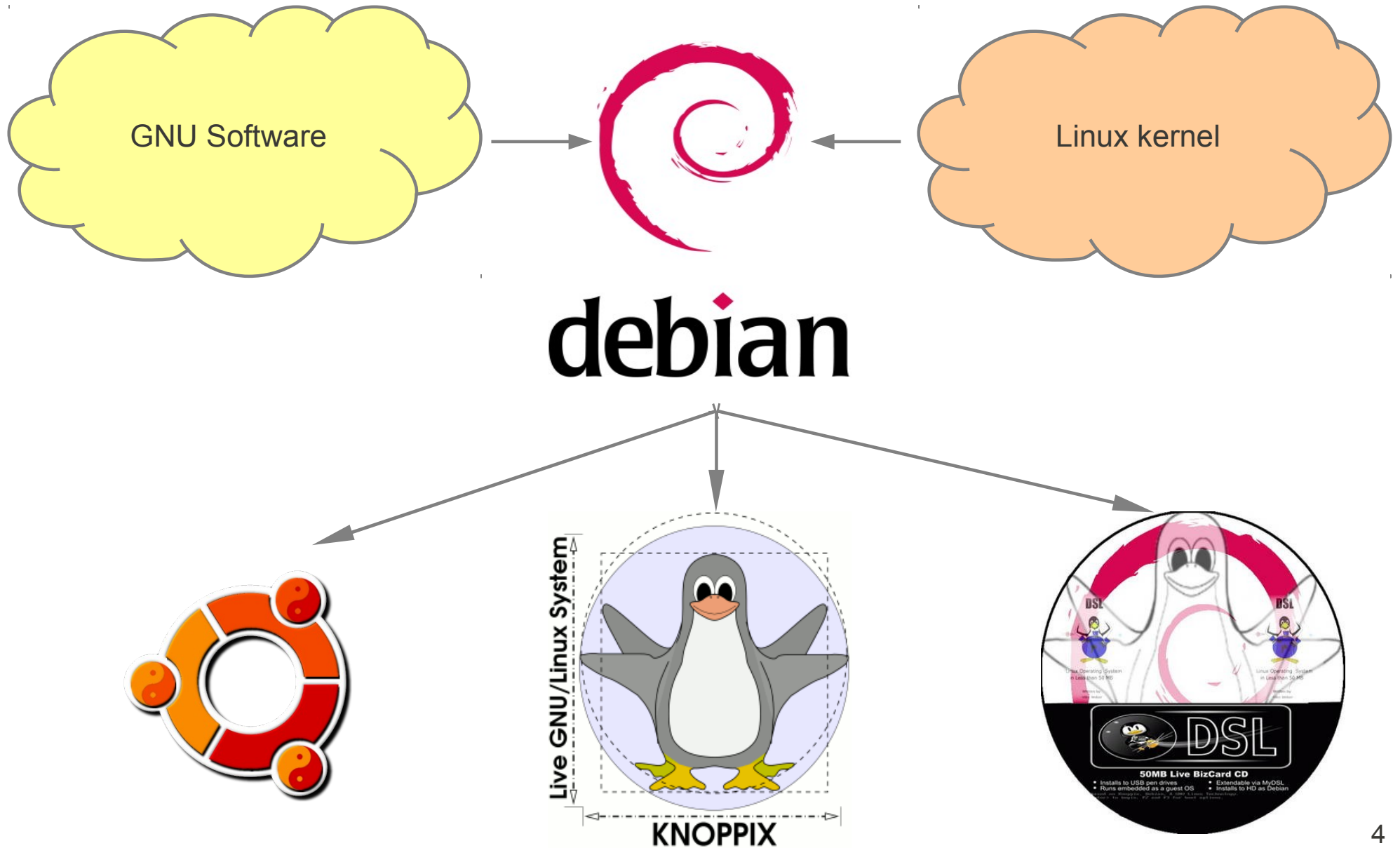
- An open source (UNIX) software developer
 - Linux kernel, C/C++ compilers, FFmpeg, Plan9
- A Hadoop guy
- Apache Software Foundation Incubator PMC
 - [Bigtop], Hadoop Development Tools, Celix, Helix
- VP of Apache Bigtop

Apache Bigtop

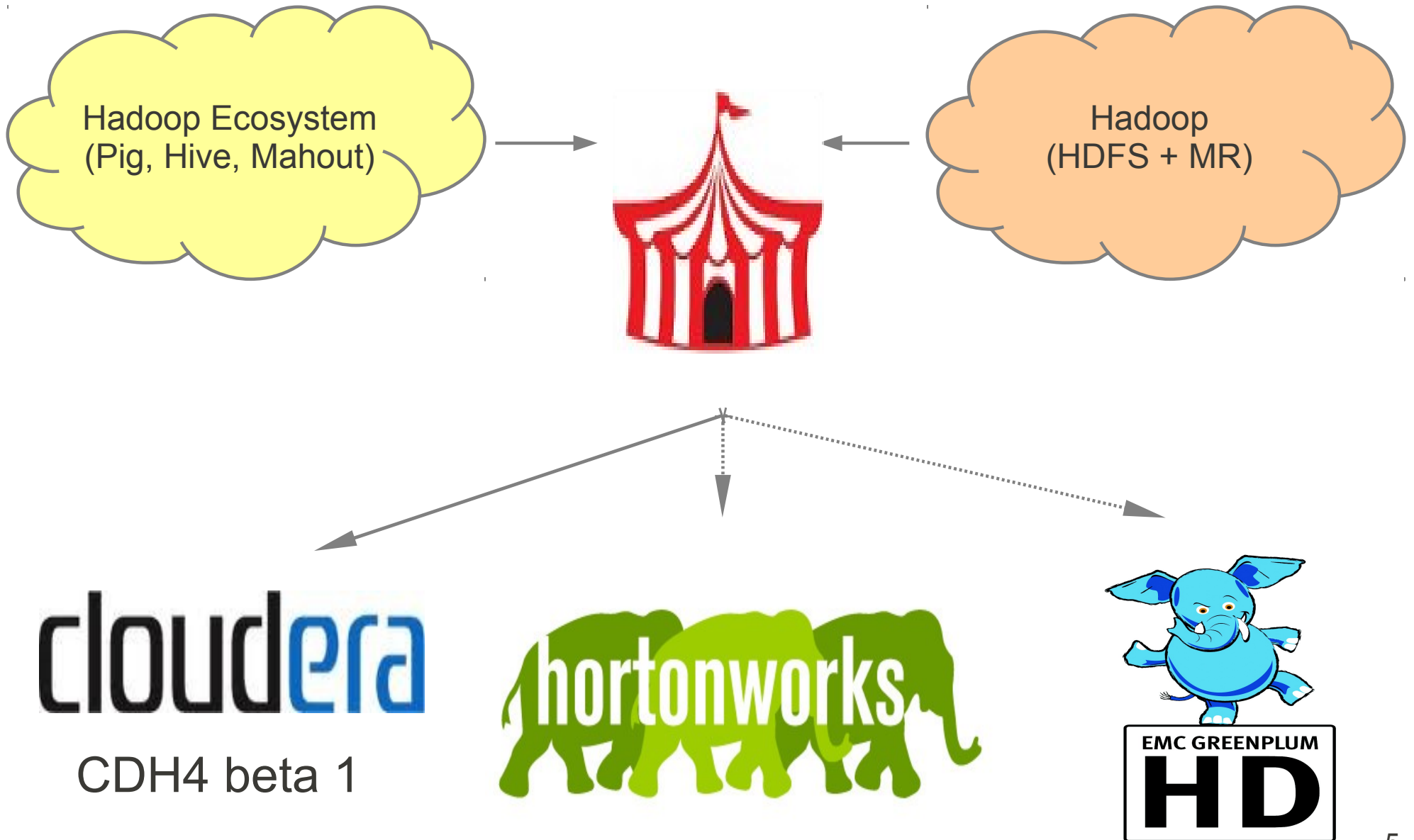
“open-source software related to a system for integration, packaging, deployment and validation of a big data management software distribution based on Apache Hadoop”



Remember what Debian did to Linux?



Bigtop is trying to do it with Hadoop



What is missing, really?

- Bigdata management platform view
- An generalist 'yin' for specialist 'yang'
- Shared, community driven:
 - Use cases
 - Best practices
 - Upcoming standards
 - Integration with









Portland Farmers Market



WELCOME

Wednesdays 10am-2pm NW Park & Salmon	Saturdays 8:30-2pm NW Park at PSU
Thursdays 4pm-8pm NW 10th & Johnson	

WHOLEFOODSMARKET

WHOLEFOODSMARKET



One way of using ASF software:

```
$ wget http://apache.org/httpd.tar.gz
```

```
$ tar xzvf httpd.tar.gz
```

```
$ cd httpd
```

```
$ ./configure ; make
```

```
$ make install
```

ERROR: can't write to /usr/local/bin

```
$ sudo make install
```

A different way:

```
$ sudo apt-get install httpd
```

Would you like to also upgrade your conf?



An “ultimate” way:

```
$ bigtop launch-cluster -config ./hbase.ini
```

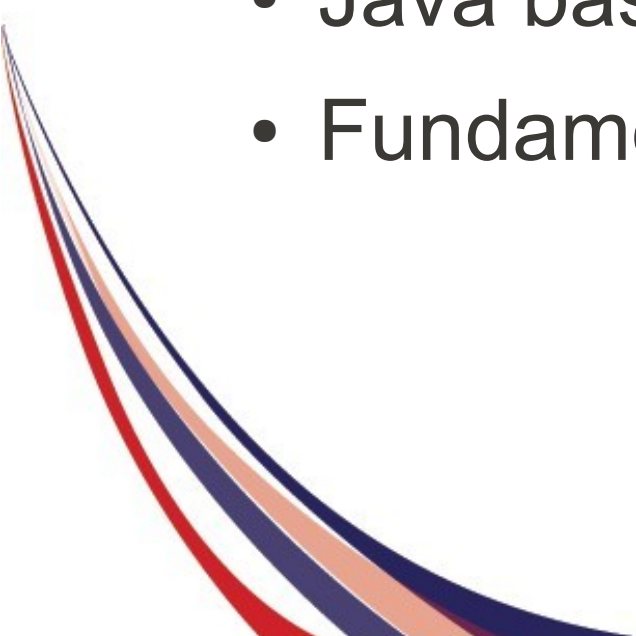


Aren't we already there?

```
$ whirr launch-cluster --config ./hbase.ini
```



Key challenges

- A really diverse set of components
 - High churn APIs
 - Asynchronous development cycles
 - Combinatoric explosion of dependencies
 - Java based
 - Fundamentally distributed applications
- 

ZooKeeper (coordination)

HUE (web based UI)

Pig (DQL)

Hive (SQL)

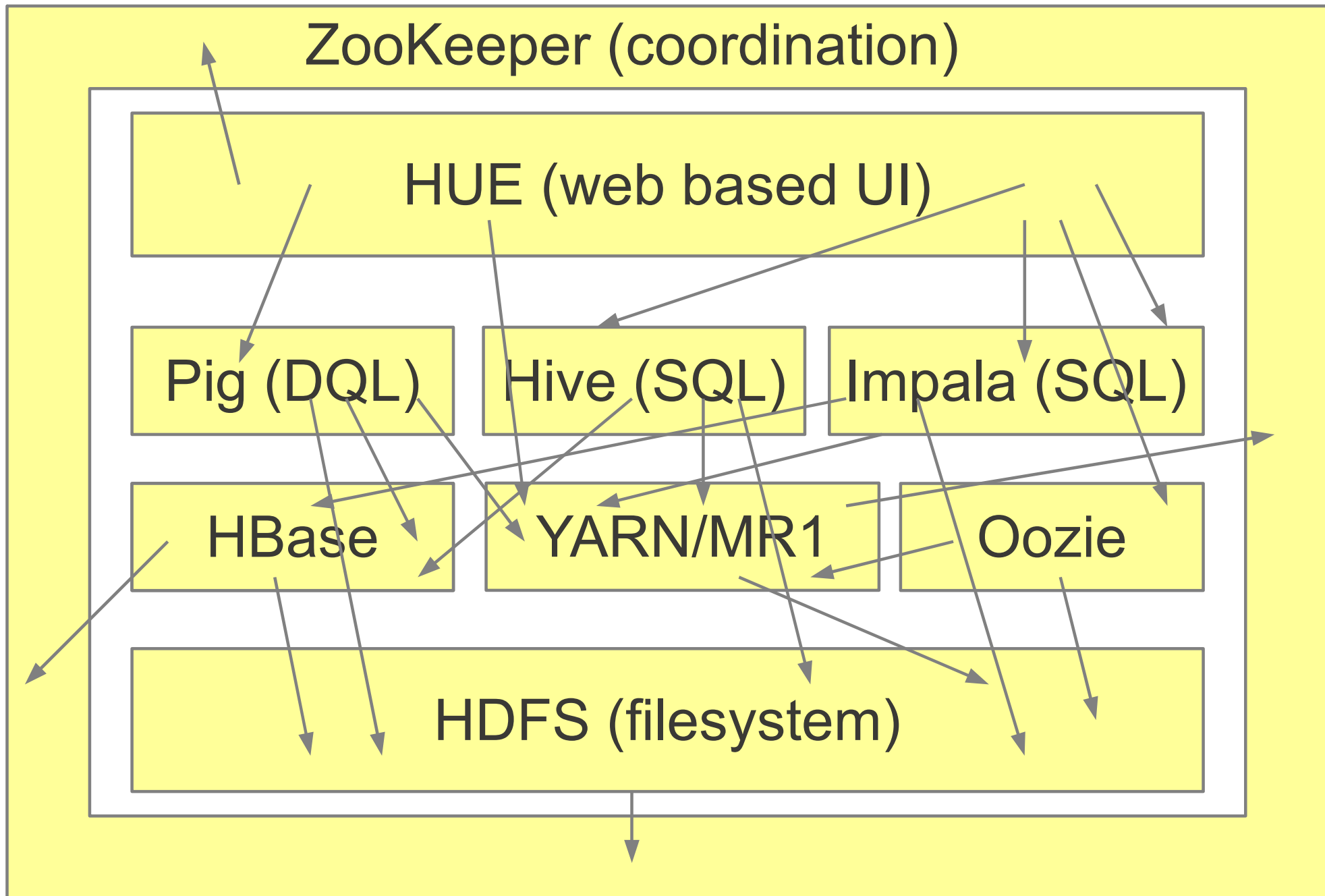
Impala (SQL)

HBase

YARN/MR1

Oozie

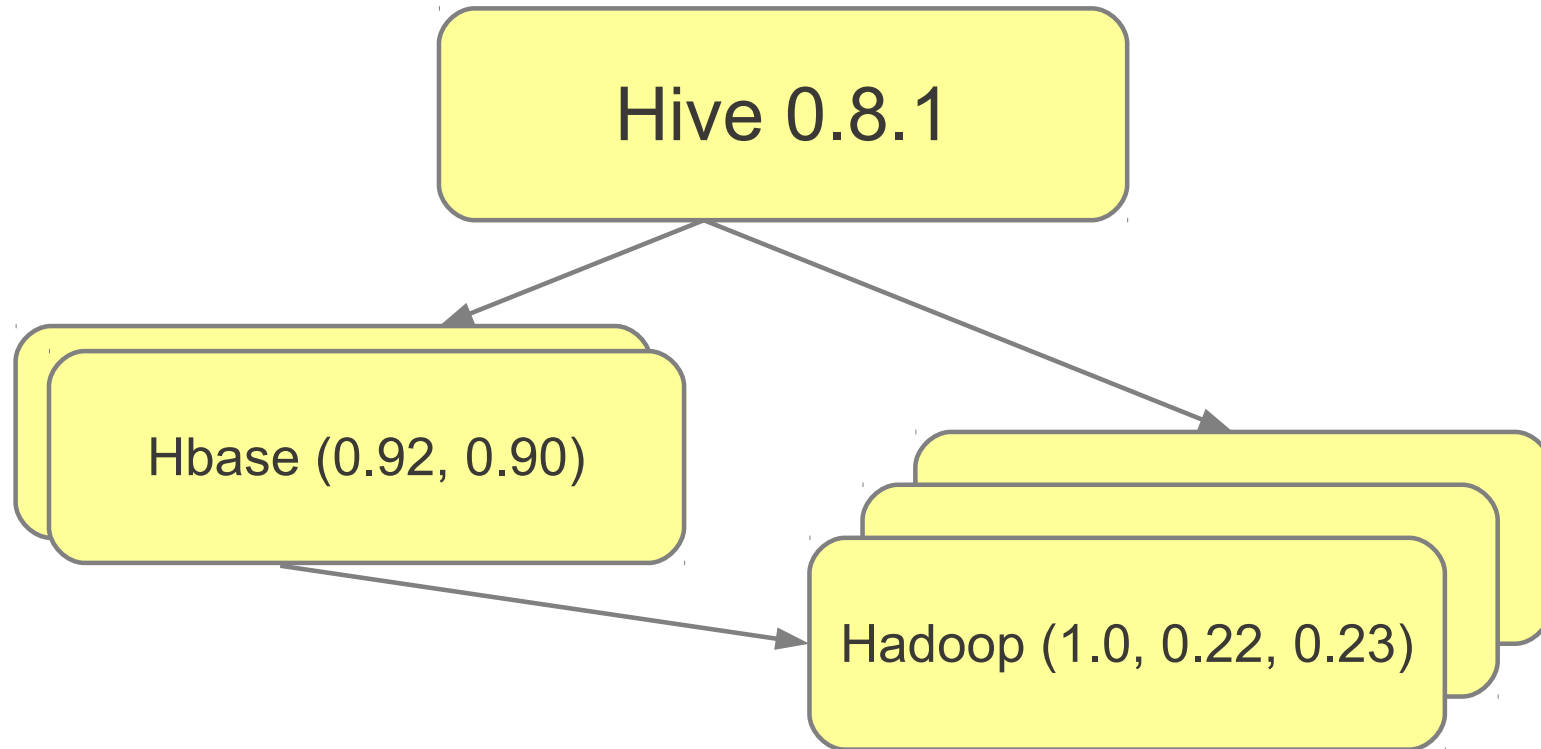
HDFS (filesystem)



It is a jungle out there

- Zookeeper
- Hadoop
 - HDFS
 - YARN
 - MR1
 - HTTPFS
- HBase
- Pig
- Hive
- Impala
- Sqoop
- Oozie
- Whirr
- Mahout
- Flume
- Giraph
- Hama
- Hue
- Solr
- Crunch
- JDK/JRE
- Kerberos
- Ganglia
- Nagios
- JSVC
- Tomcat
- Utils
- Postgress
- HTTPD

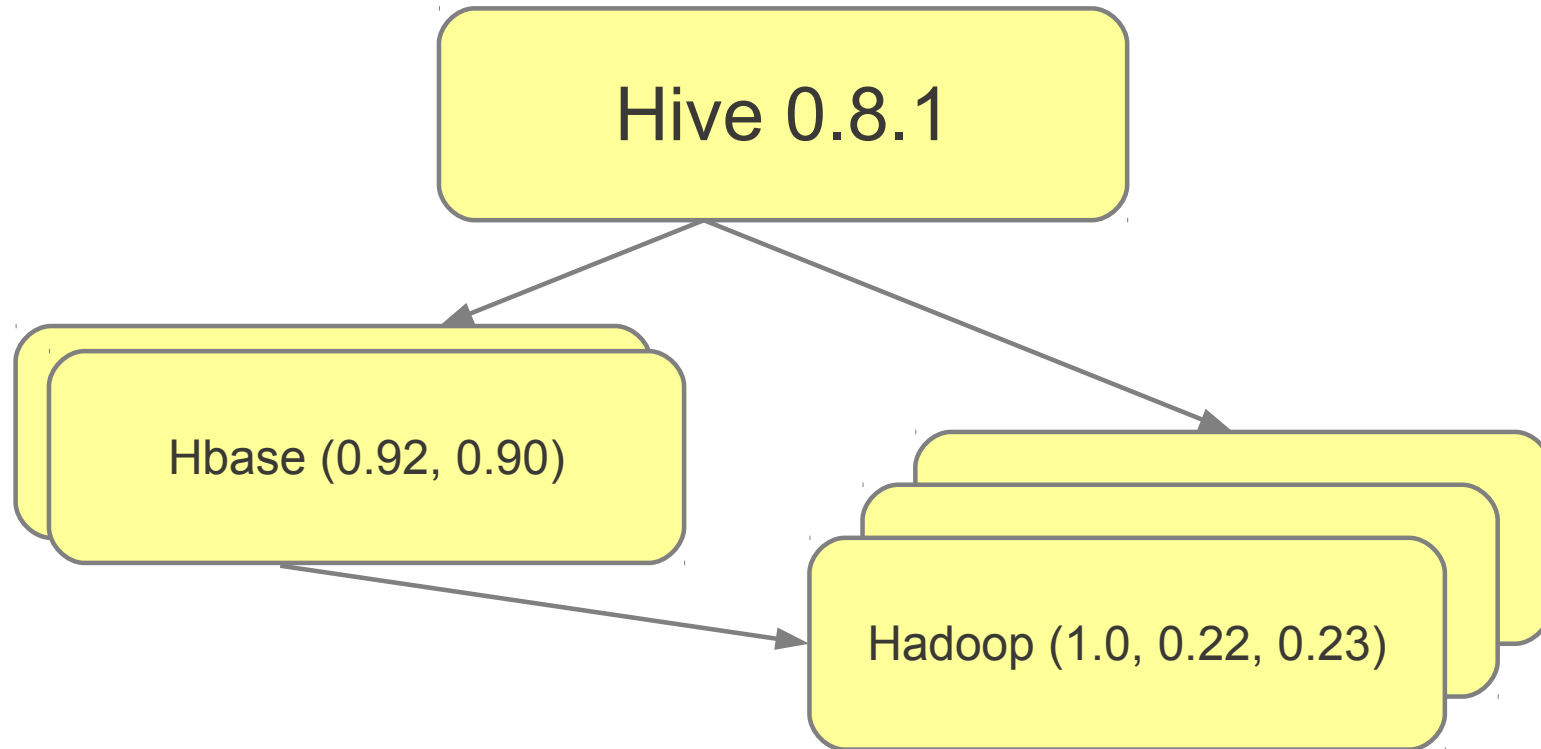
Dependencies Inferno:



A million dollar question:

```
$ tar xzvf hive-0.8.1.tar.gz  
$ ls hive-0.8.1/lib
```

Dependencies Inferno:



A million dollar question:

```
$ tar xzvf hive-0.8.1.tar.gz
```

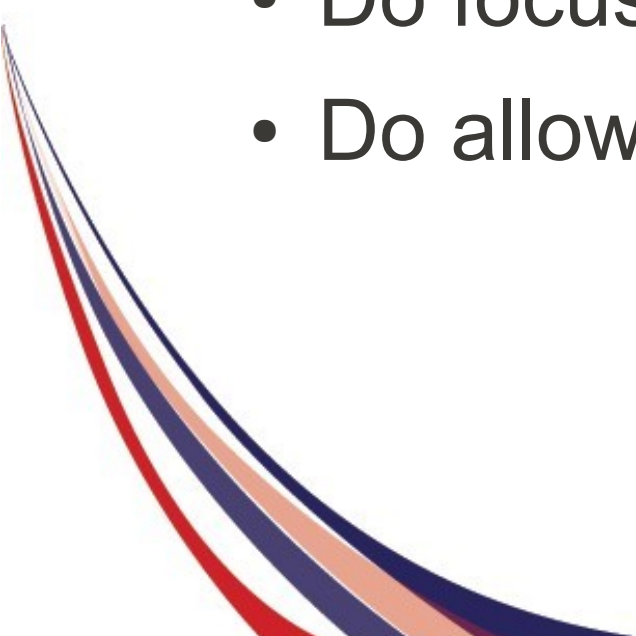
```
$ ls hive-0.8.1/lib
```

```
hbase-0.89.jar log4j-1.2.15.jar log4j-1.2.16.jar
```

Lessons in cat herding

- ~~Admitting the problem~~
- On origins of suffering
- You can't make “them” do “it”
- The real world is highly asynchronous
- The art of making friends
- YLH

The origin of suffering is attachment

- Don't get attached to your code
 - Don't waste your time on ill-maintained code
 - Don't second guess your users
 - Do provide capabilities, not polices
 - Do focus on specialization
 - Do allow customization
- 

You can't make “them” do “it”

- Don't expect common dependencies
- Don't expect agreement on use cases
- Don't ask – offer:

```
<groupId>org.apache.hadoop</groupId>
```

```
<artifactId>hadoop-core</artifactId>
```

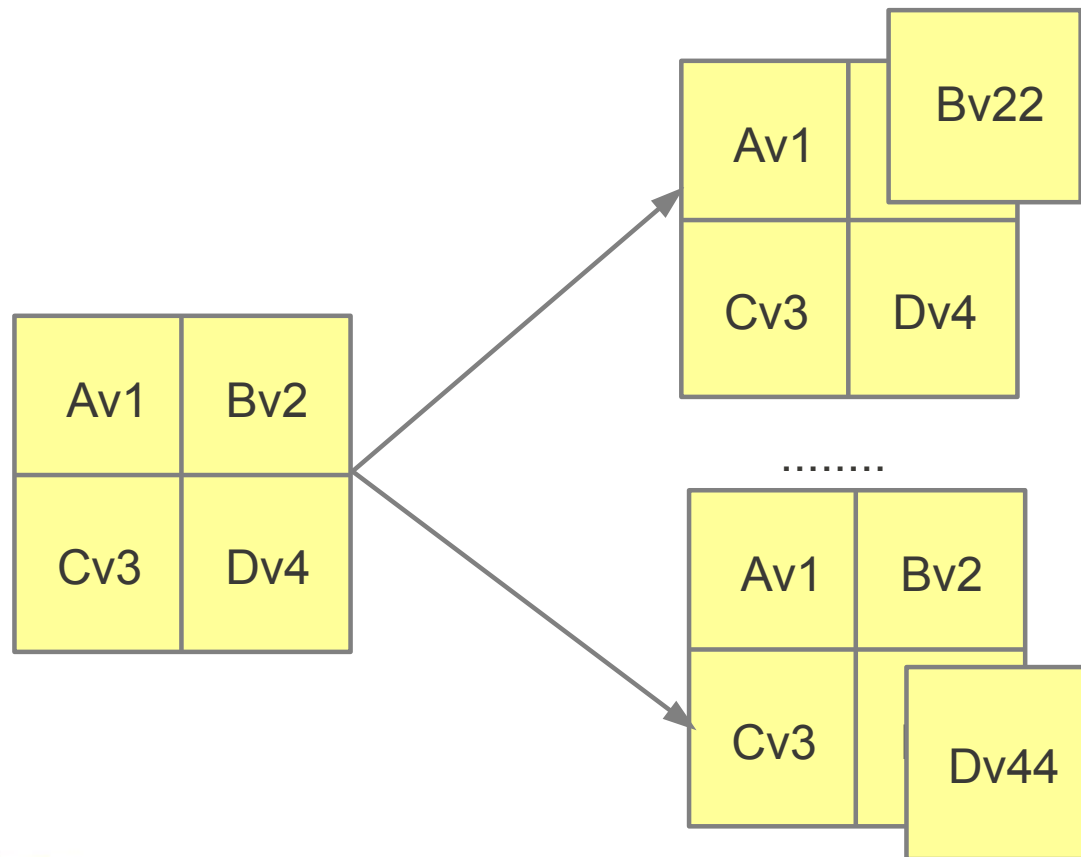
```
<version>${hadoop.version}</version>
```

```
<optional>>true</optional>
```

```
<classifier>hadoop-2.0.2</classifier>
```

Embrace asynchronous nature

- Don't expect flag days
- Don't expect agreement on releases
- Do practice Last Known Good Builds



Make yourself indispensable

- Be nice
- Do provide glue code
- Do provide tons of automation
- Do provide missing testing
- Do participate in upstream communities:
 - RC votes
 - Release Planning

What does Bigtop offer:

- Community focused on all of the above
- Software for:
 - Integration
 - Build (make, Maven)
 - Packaging (RPM, DEB)
 - Deployment (Puppet)
 - Testing (iTest)
- A continuous integration Jenkins server

Who's on-board?

- Cloudera
 - CDH4 is 100% based on Bigtop (hadoop v2)
- WANdisco
- TrendMicro
- Hortonworks, EMC, EBay, Intel (partially)
- Canonical
 - Ubuntu Server: Hadoop and Bigdata blueprint
- Illumos (early stages of interest)

What's happening

- A special release: Bigtop 0.3.0-incubating
 - Hadoop 1.0.1
- Last stable release: Bigtop 0.5.0
 - Hadoop 2.0.2-alpha
- Next stable release: Bigtop 0.6.0
 - End of Mar 2013 release
 - Hadoop 2.0.3-beta (DANGER! DANGER!)
 - Major focus on developers

What does Bigtop need?

- More of you!
 - “Silicon Valley Hands-on Programming”
<http://www.meetup.com/HandsOnProgrammingEvents/>
- More infrastructure for build/test
 - EC2, Supercell, EMC magic cluster, CloudStack
- More integration tests
 - Convince your bosses to commit to Bigtop
- Validate upstream release using Bigtop

How to get in touch

- Bigtop home @Apache:
 - <http://bigtop.apache.org/>
- Hangout places:
 - `{dev,user}@bigtop.apache.org`
 - `#bigtop` on Freenode
- Roman Shaposhnik
 - rvs@apache.org, rvs@cloudera.com