Leveraging Docker for Hadoop Build Automation and Big Data Stack Provisioning

Apache Big Data North America 2017

PRESENTED BY Evans **Ye** | May 16, 2017

Who am I

- Software Engineer @ Y! APAC Data Team
- Building data products for...




- Apache Bigtop PMC chair

Outline

- Quick Intro to Apache Bigtop
- Docker for Bigtop Packaging
- Docker for Bigtop Provisioner
- Docker for Bigtop Sandbox
- Release

Quick Intro to Apache Bigtop

Linux Distributions

debian

YAHOO!

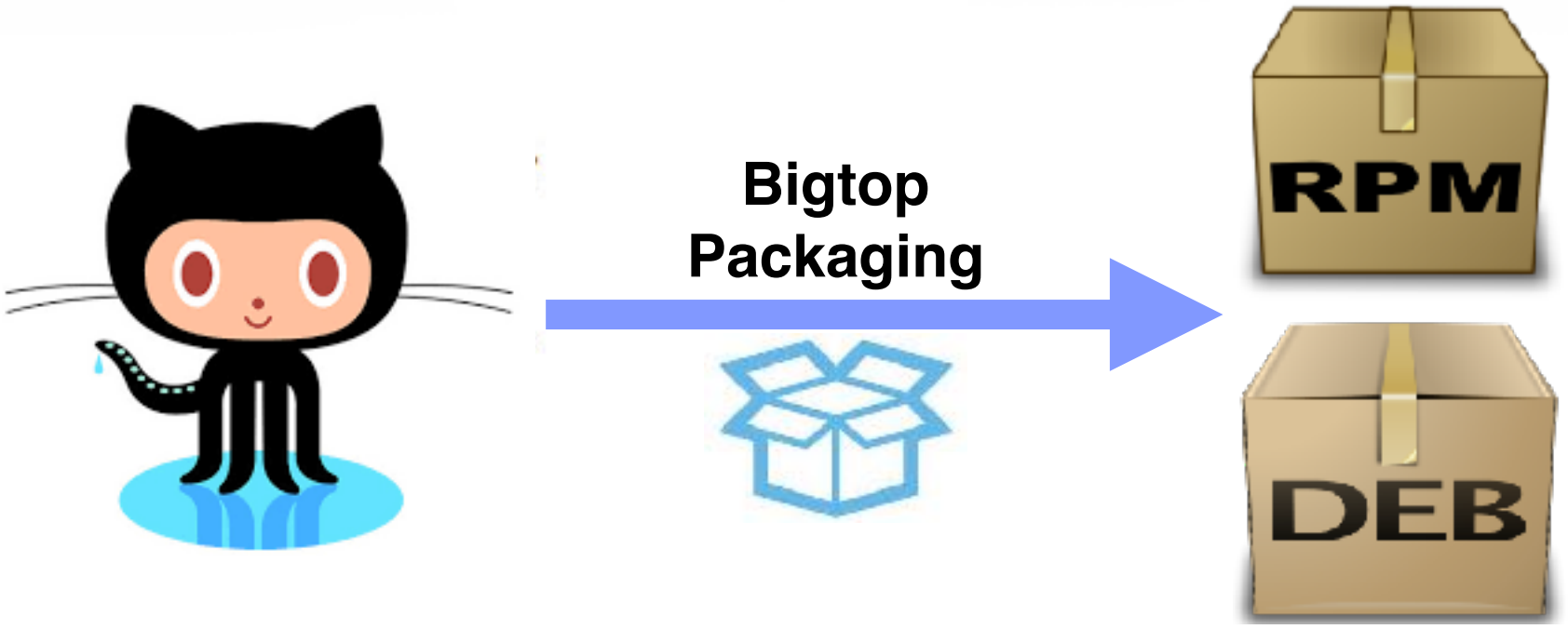
Hadoop Distributions

But there're some other great Hadoop ecosystem components..

How do I add patches?

YAHOO!

From source code to packages



YAHOO!

Supported components

YAHOO!

Bigtop feature set

Packaging **Testing** **Deployment** **Virtualization**

for you to easily build your own Big Data Stack

Docker for Bigtop Packaging

Preparing build environment

Tool requirements for building Bigtop

On all systems	Also on RPM-based systems	Also on DEB-based systems
• Java JDK 1.6	• lzo-devel	• libxslt1-dev
• Apache Ant	• zlib-devel	• libkrb5-dev
• Apache Maven	• fuse-devel	• libldap2-dev
• wget	• openssl-devel	• libmysqlclient-dev
• tar	• python-devel	• libsasl2-dev
• git	• libxml2-devel	• libsqlite3-dev
• subversion	• libxslt-devel	• libxml2-dev
• gcc	• cyrus-sasl-devel	• python-dev
• gcc-c++	• sqlite-devel	• python-setuptools
• make	• mysql-devel	• liblzo2-dev
• fuse	• openldap-devel	• libzip-dev
• protobuf-compiler	• rpm-build	• libfuse-dev
• autoconf	• createrepo	• libssl-dev
• automake	• redhat-rpm-config (RedHat/CentOS only)	• build-essential
• libtool		• dh-make
• sharutils		• debhelper
• xmlto		• devscripts
		• reprepro

Preparing build environment

Tool requirements for building Bigtop

On all systems	Also on RPM-based systems	Also on DEB-based systems
• Java JDK 1.6	• lzo-devel	• libxslt1-dev
• Apache Ant	• zlib-devel	• libkrb5-dev
• Apache Maven	• fuse-devel	• libldap2-dev
• wget	• openssl-devel	• libmysqlclient-dev
• tar	• python-devel	• libsasl2-dev
• git	• [illegible]2-devel	• libsqlite3-dev
• subversion	• [illegible]	• libxml2-dev
• gcc	• [illegible]	• python-dev
• gcc-c++	• sqlite-devel	• python-setuptools
• make	• mysql-devel	• liblzo2-dev
• fuse	• openldap-devel	• libzip-dev
• protobuf-compiler	• rpm-build	• libfuse-dev
• autoconf	• createrepo	• libssl-dev
• automake	• redhat-rpm-config (RedHat/CentOS only)	• build-essential
• libtool		• dh-make
• sharutils		• debhelper
• xmlto		• devscripts
		• reprepro

...

Seriously ?

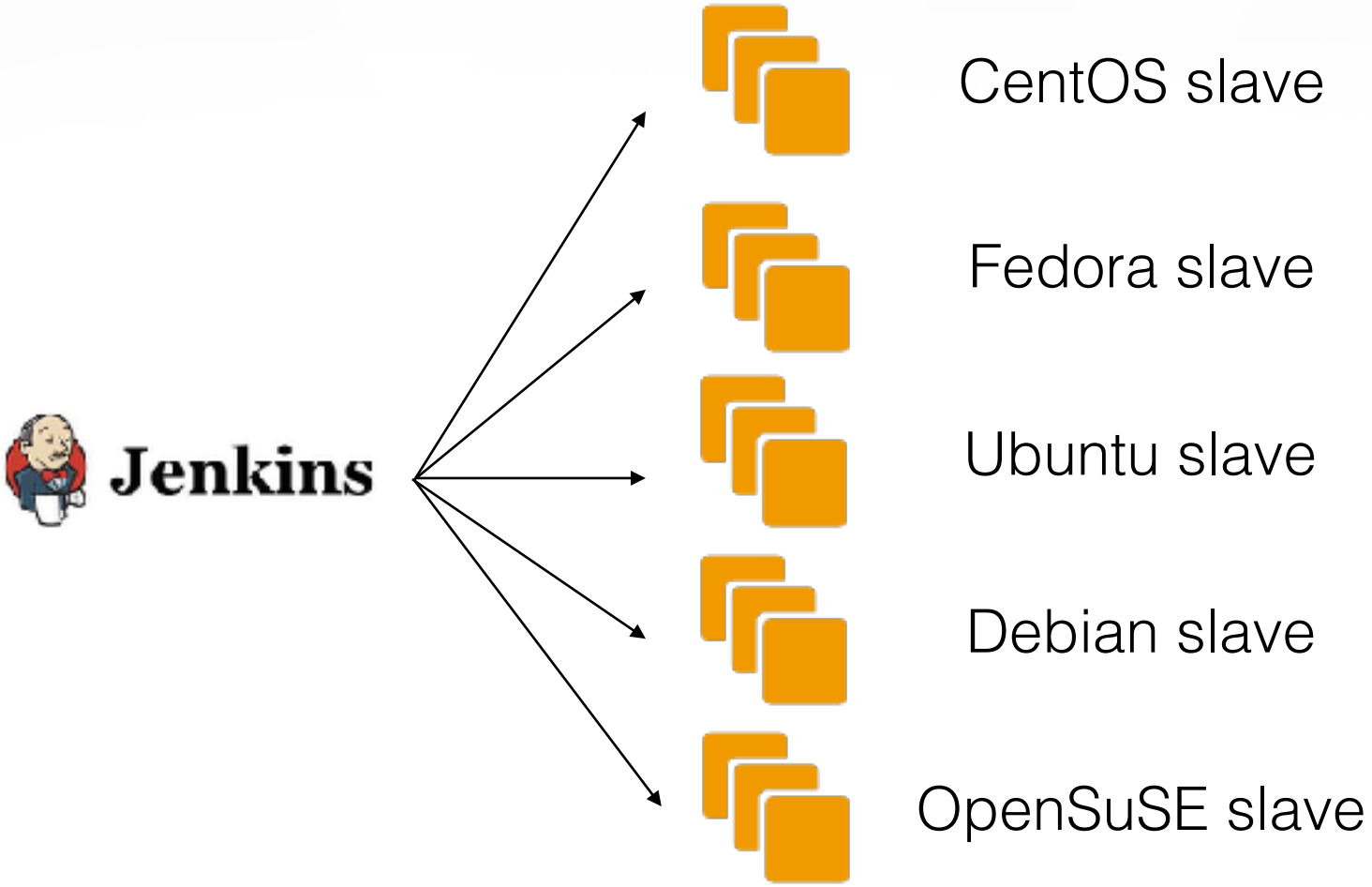
Bigtop Toolchain

- Puppet recipes to install required libraries, build tools
- To prepare a build environment:

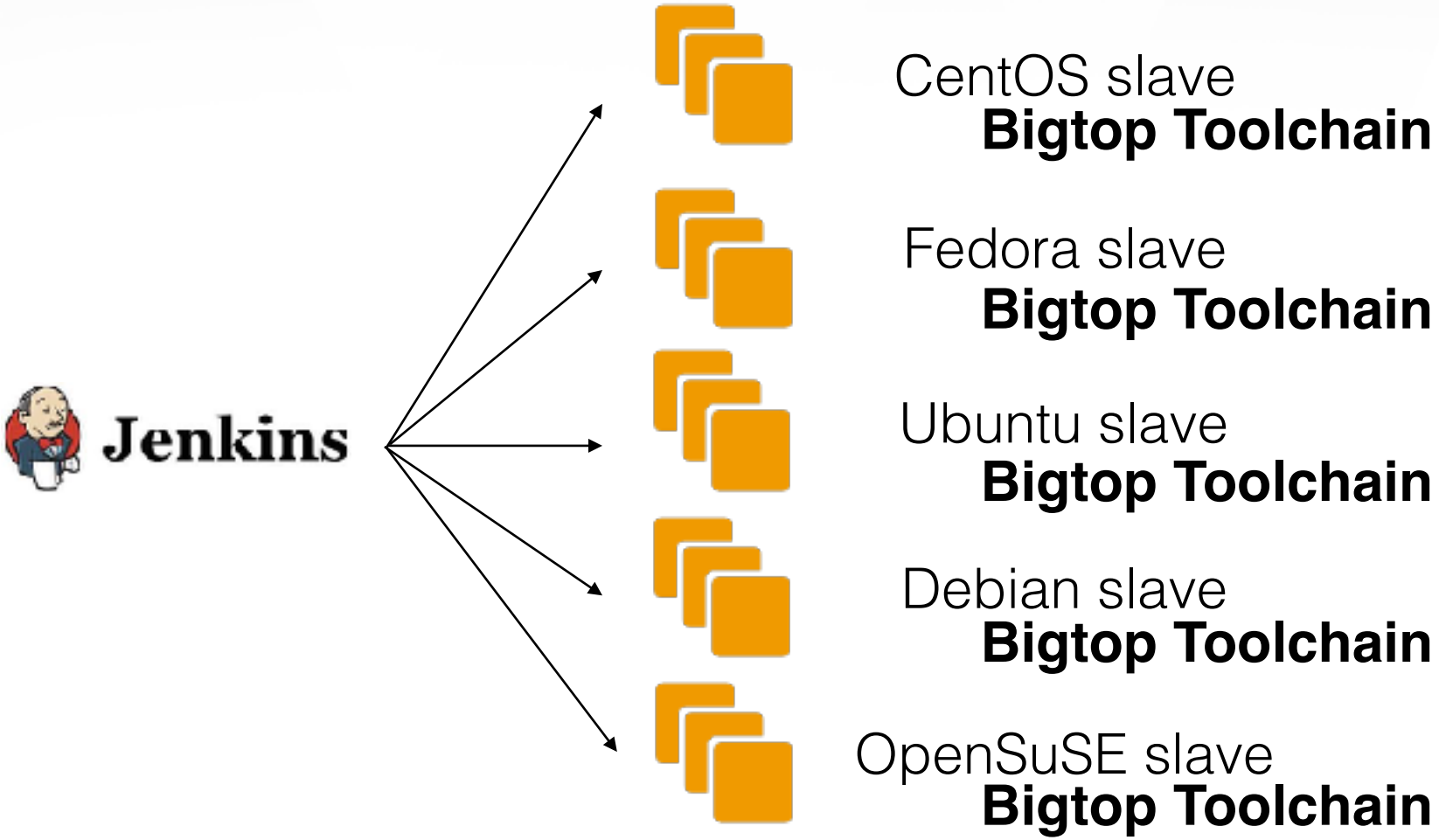
```
git clone https://github.com/apache/bigtop.git
cd bigtop
./bigtop_toolchain/bin/puppetize.sh
./gradlew toolchain
```

- Prerequisite :
  - Java

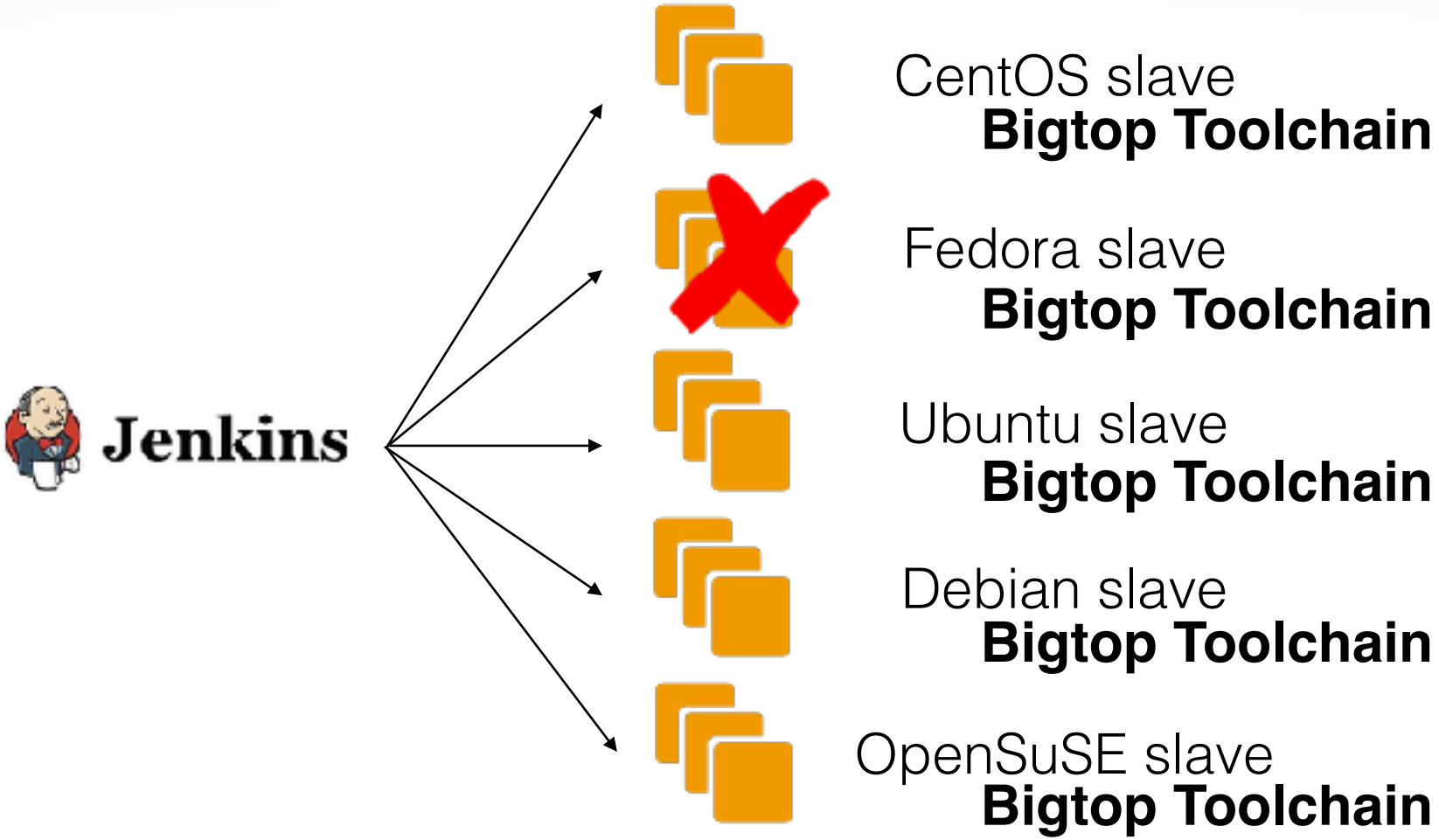
CI Infrastructure



CI Infrastructure

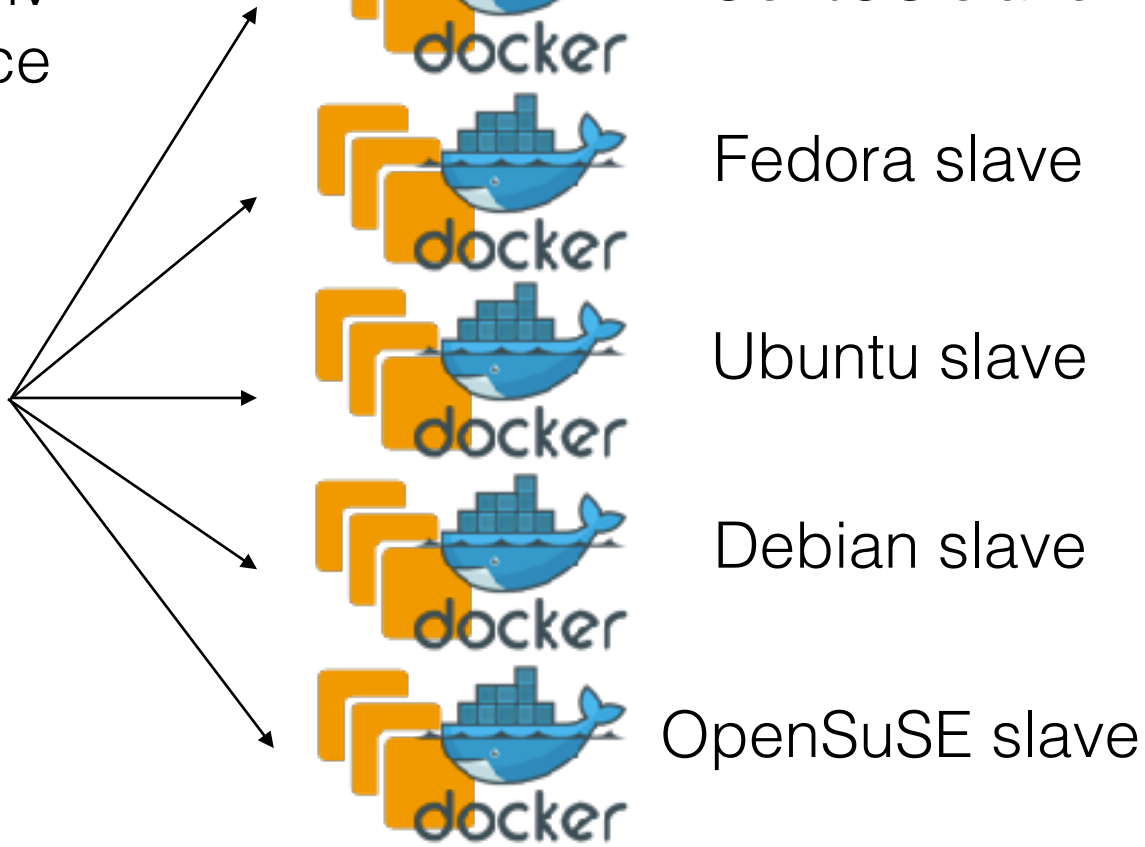


CI Infrastructure



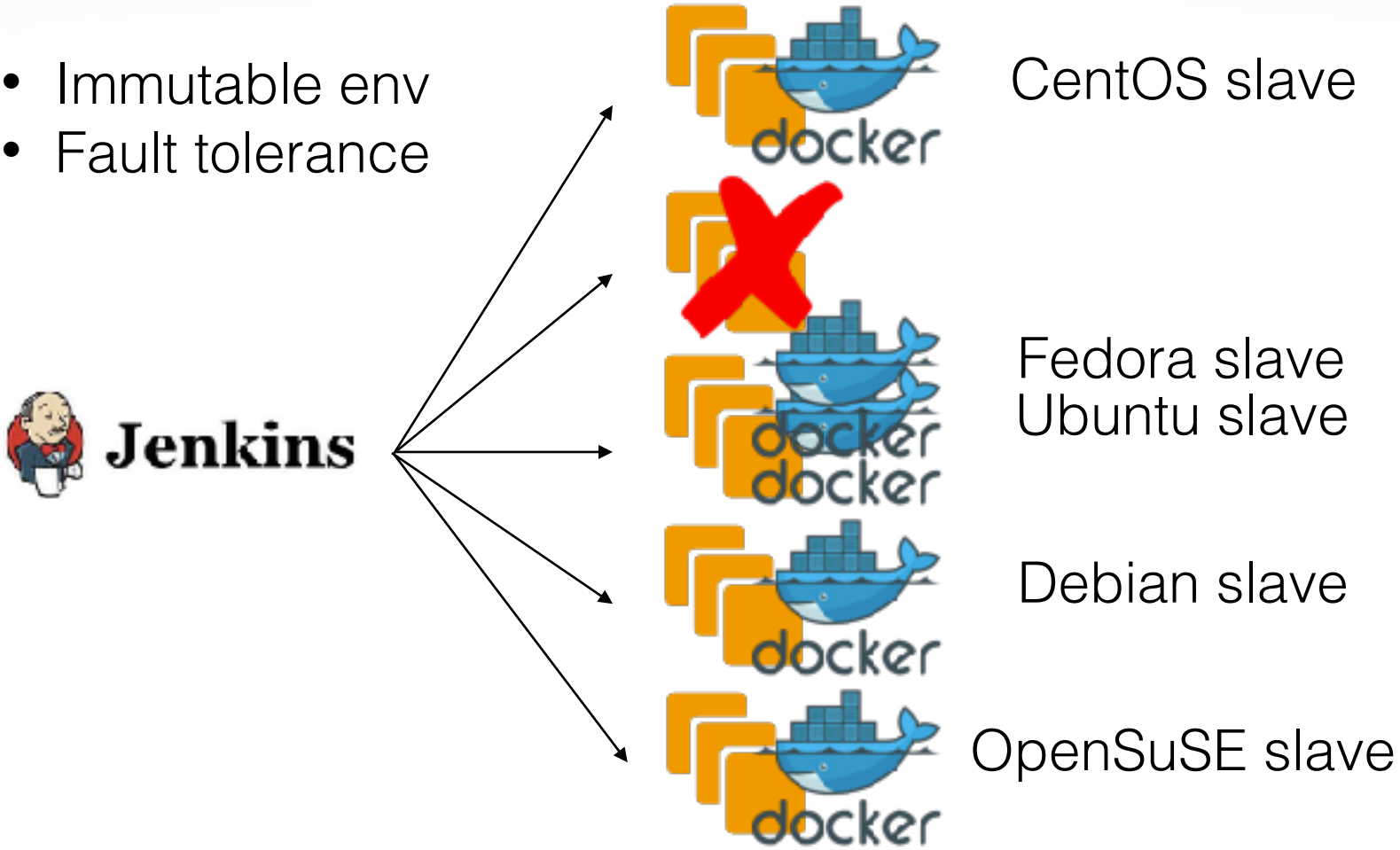
Dockerlized CI Infrastructure

- Immutable env
- Fault tolerance



Dockerlized CI Infrastructure

- Immutable env
- Fault tolerance



YAHOO!

How to build packages

- Execute shell

```
# OS=debian-8
# COMPONENT=hadoop

docker run -u jenkins --rm \
-v `pwd`:/bigtop --workdir /bigtop \
bigtop/slaves:trunk-$OS \
bash -l -c "./gradlew allclean $COMPONENT-pkg"
```

- Bigtop CI Setup Guide

Bigtop master

https://ci.bigtop.apache.org/view/Packages/job/Bigtop-trunk-packages/

YAHOO!

Bigtop early mission accomplished

MAPR

cloudera

amazon web services™

All major Hadoop distros leverage Bigtop to build its foundation

Leveraged by app providers…

wanDISCO

YAHOO!

Get out from the Apache dome

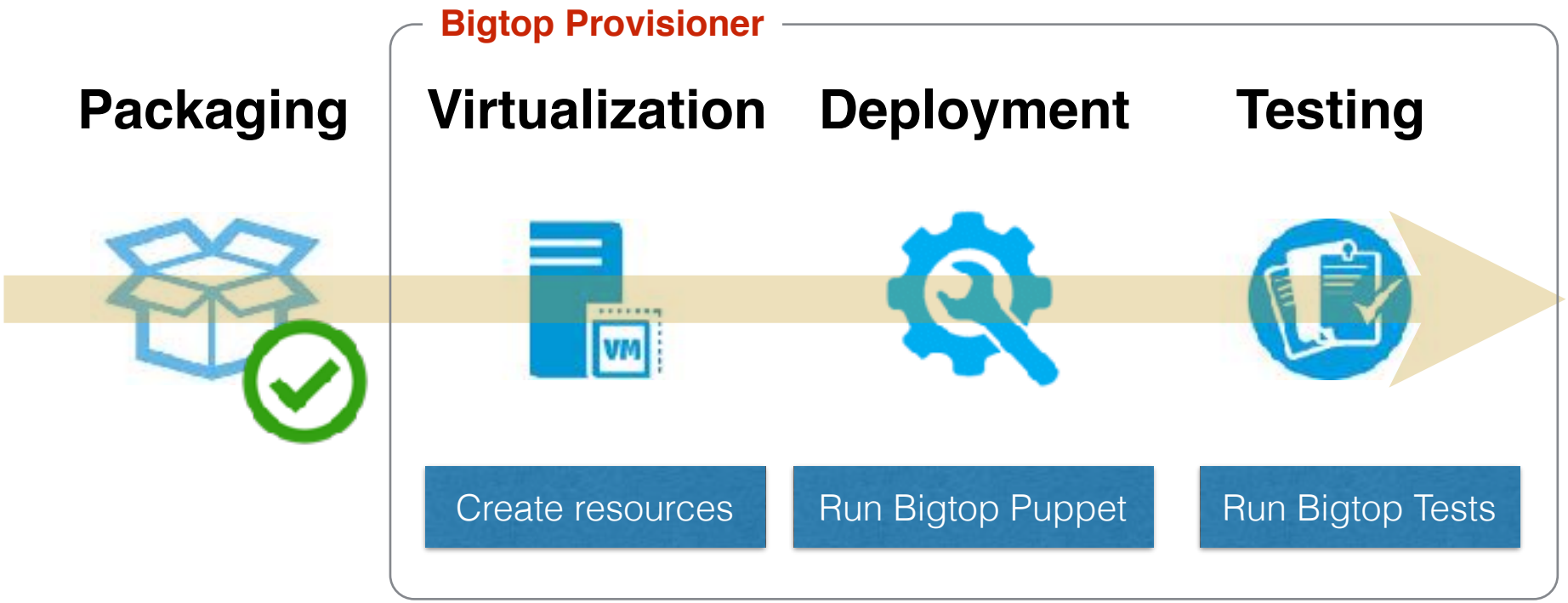
New focus and target user

- Data engineers vs Distro. builders
- Solution diversity:
  - Streaming: Flink, Apex
  - In-memory cache: Alluxio, Ignite
  - Non apache: QFS, GPDB
- User/developer tools:
  - Bigtop Provisioner
  - Bigtop Sandbox
- Big data stack references

Docker for Bigtop Provisioner

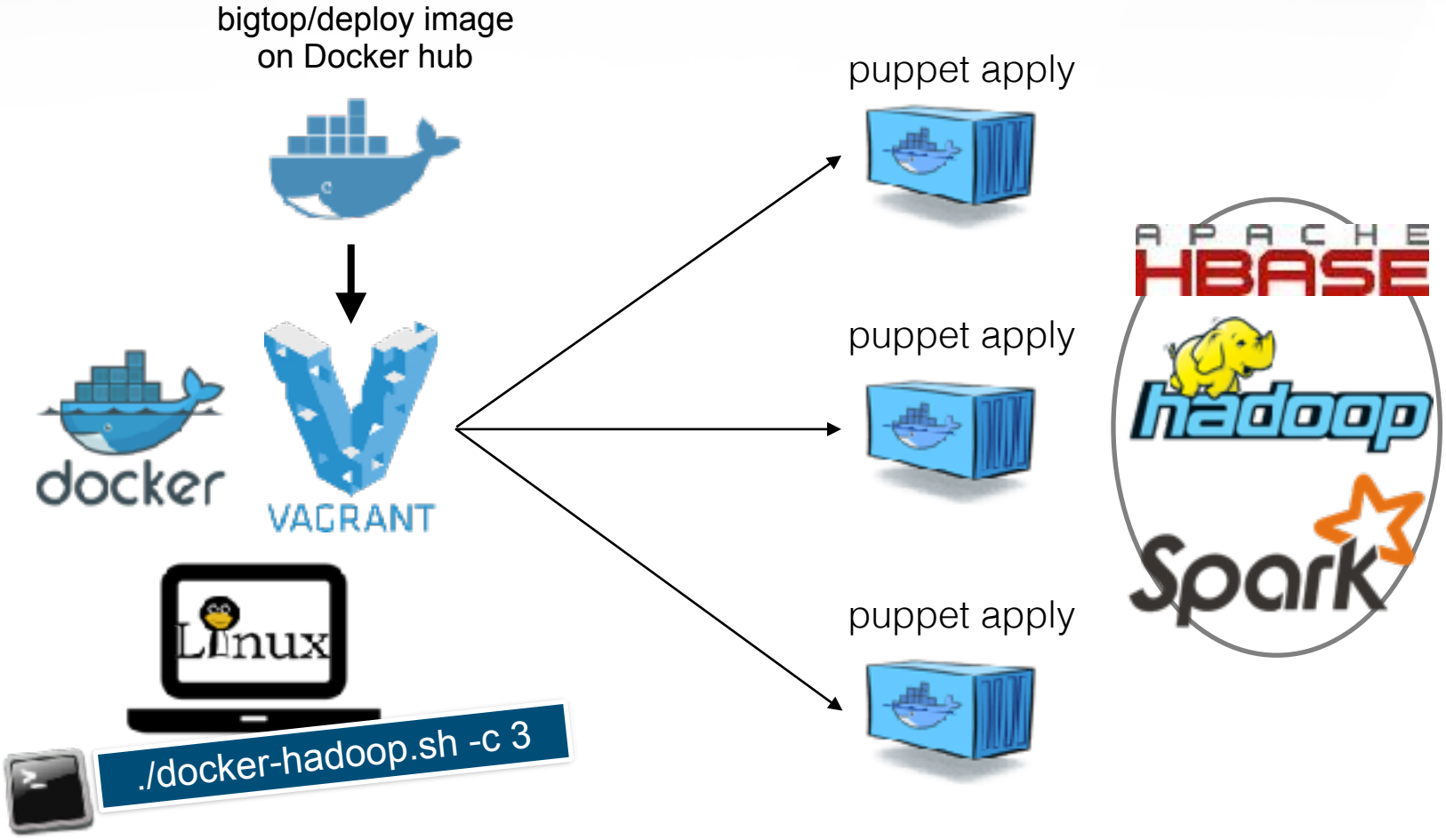
Bigtop Provisioner

- A tool to demonstrate full life cycle of Bigtop



YAHOO!

One click Hadoop provisioning (Bigtop 1.0.0)

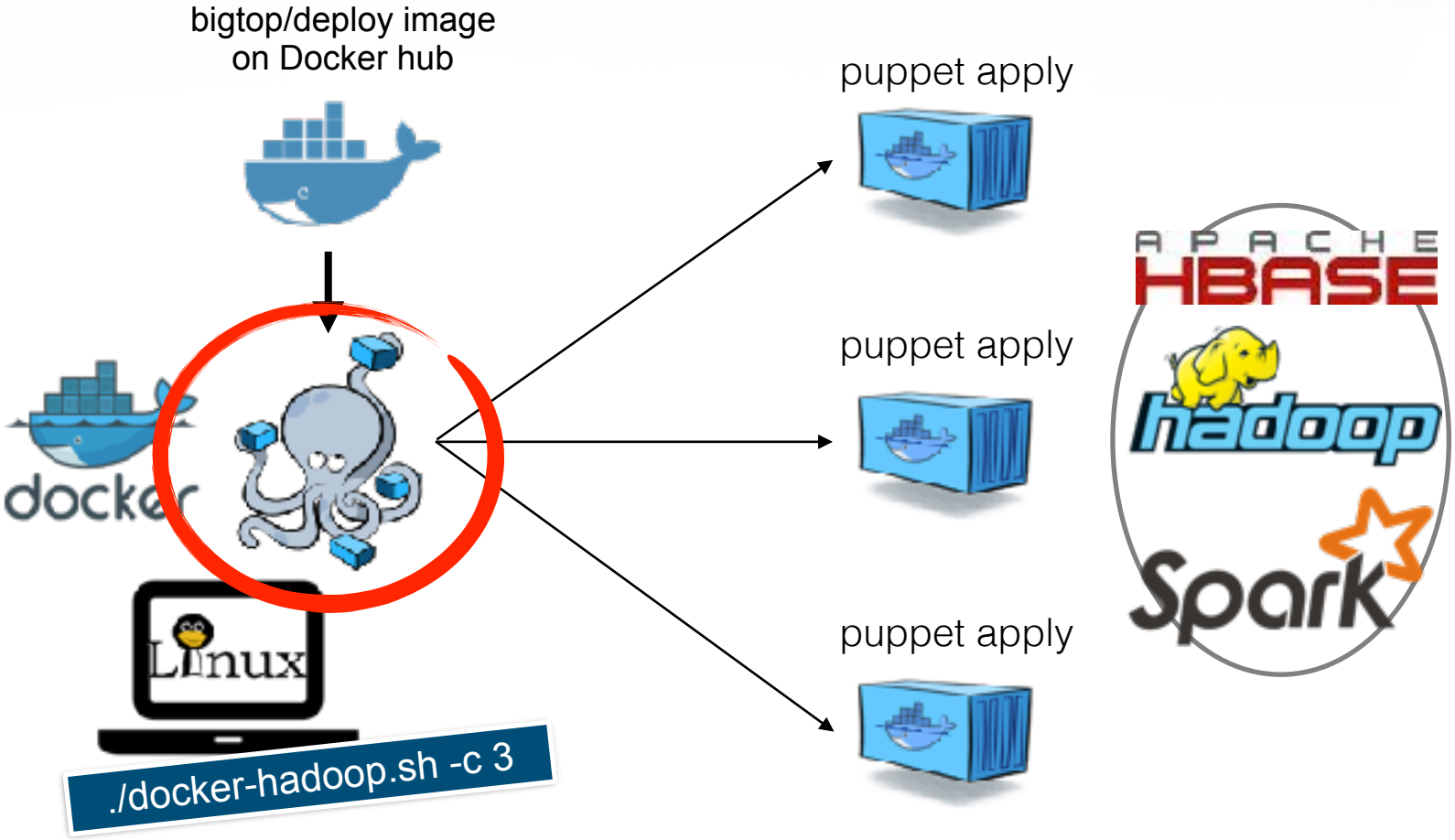


YAHOO!

What's the problem with Vagrant's Docker Provider?

- Need to add vagrant public key into docker images
- Too many issues with auto-created boot2docker VM
- A bug for docker provider keep opening for 2ys
  - Waiting for machine to boot' hangs infinitely
- Can not share same code for different providers anyway
- Not all the docker options supported in Vagrantfile
- ^#?& slow

Replaced by docker-compose (Bigtop 1.2.0)



YAHOO!

Advantages

- No need to create customized image beforehand
- Better compatibility with Docker's native solutions
- Clear, simple yaml file for orchestration settings
- Supports new features such as overlay network
- Leverage Swarm for multi-node cluster deployment
- Fast —> better user experience

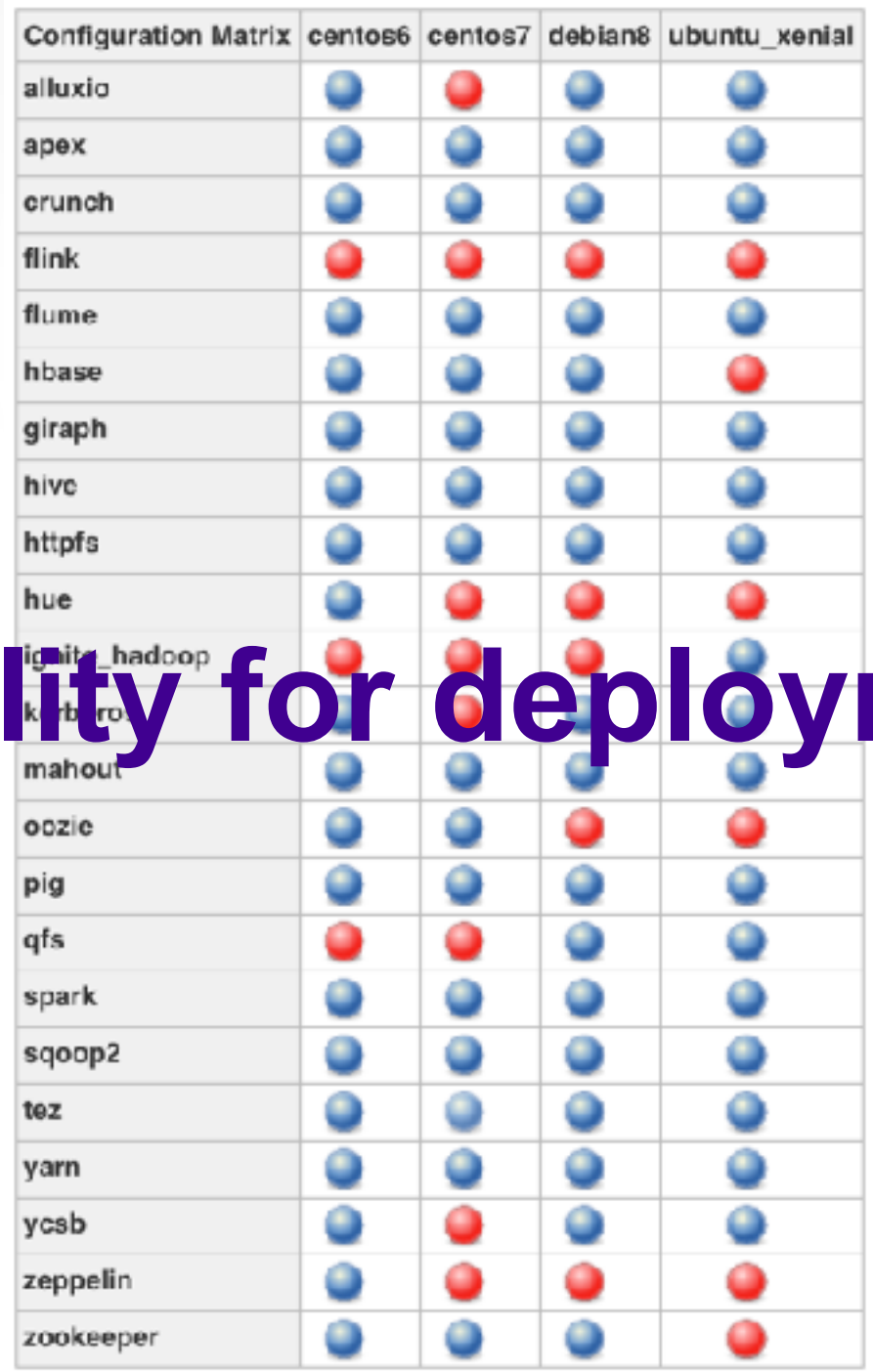
How to run Docker Provisioner

- Execute shell

```
# See bigtop/provisioner/docker/*.yaml
CONFIG=YOUR_CUSTOM_CONF.yaml

# provision
./gradlew -Pconfig=${CONFIG} -Pnum_instances=1 \
docker-provisioner

# destroy provisioned cluster
./gradlew docker-provisioner-destroy
```

- Bigtop CI Setup Guide

Visibility for deployments



Configuration Matrix	centos6	centos7	debian8	ubuntu_xenial
alluxio				
apex				
crunch				
flink				
flume				
hbase				
giraph				
hive				
httpfs				
hue				
ignite_hadoop				
mahout				
oozie				
pig				
qfs				
spark				
sqoop2				
tez				
yarn				
ycsb				
zeppelin				
zookeeper				

YAHOO!

Use Cases

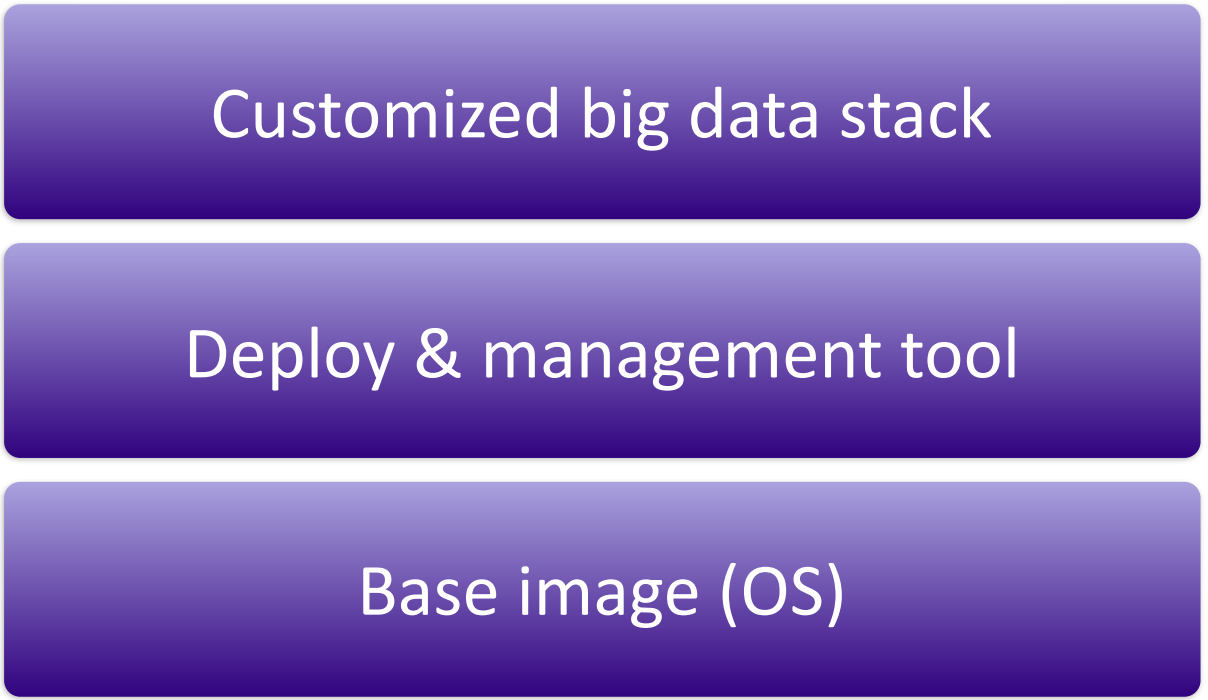
- **For application developers, cluster admins, users**
  - Run a Hadoop cluster to test your code on
  - Try & test configurations before applying to Production
  - Play around with Bigtop Big Data Stacks
- **For contributors**
  - Easy to test your packaging, deployment, testing code
- **For Distro. builders**
  - CI matrix —> patch upstream code made easier

Docker for Bigtop Sandbox

Introducing Bigtop Sandbox

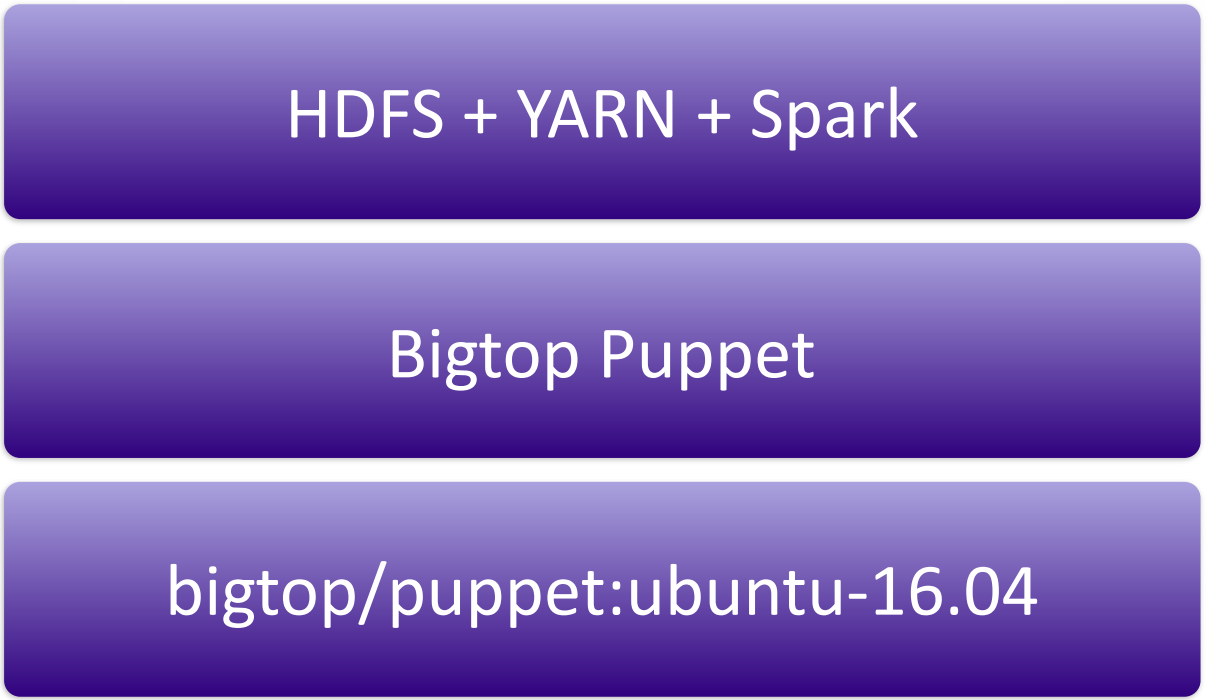
- Easiest way to get started
- Docker images that has Bigtop stacks installed and configured
- Pseudo cluster up & running w/ zero installation
- Command-line tool for you to build your own stack

Docker Image layer Interface

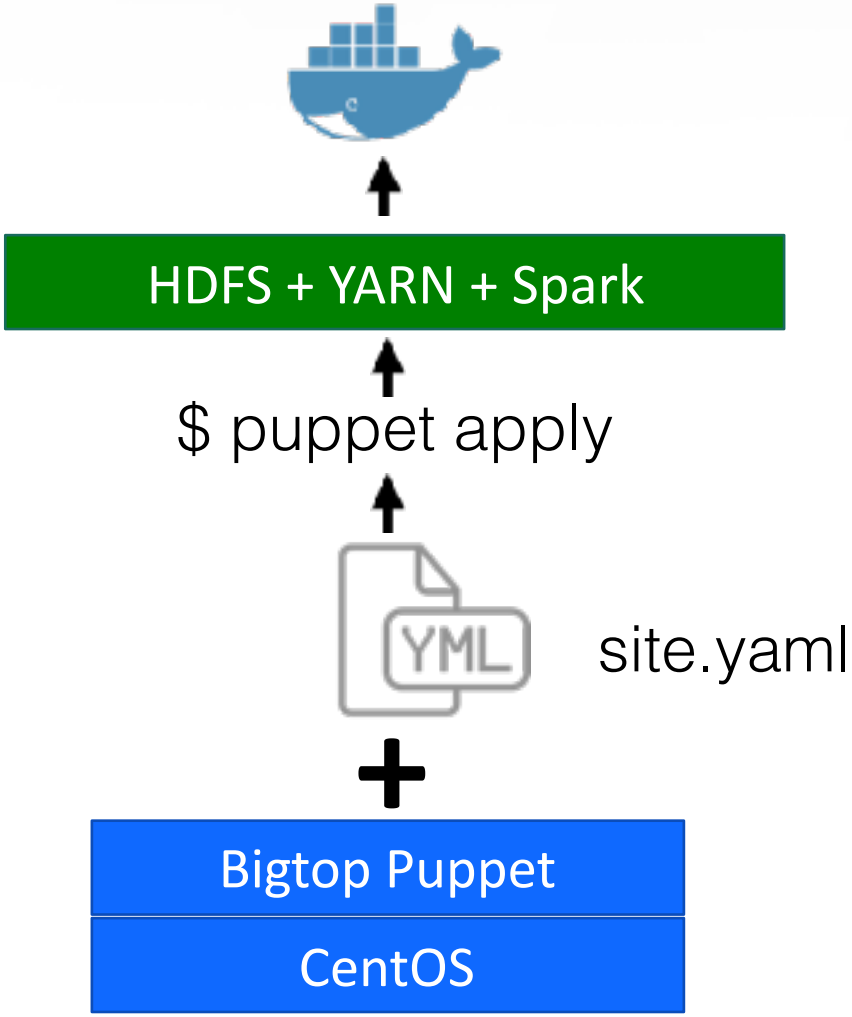


YAHOO!

Docker Image layer
Concrete implementation



Building images



YAHOO!

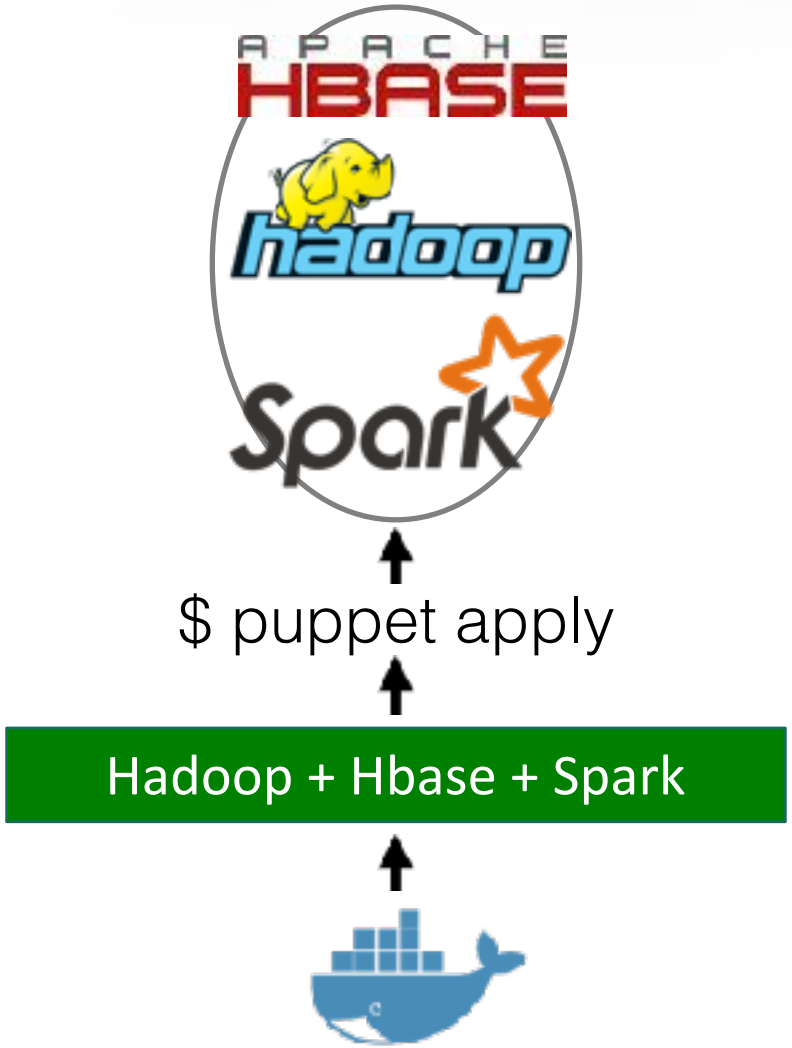
How to build

```
git clone https://github.com/apache/bigtop.git
cd bigtop/docker/sandbox

./build.sh -a evansye -o ubuntu-16.04 \
-c hdfs,yarn,spark
```

- Specify custom conf:

```
./build.sh-a evansye -o ubuntu-16.04 \
-f site.yaml -t apache_big_data_2017_miami
```

Running images



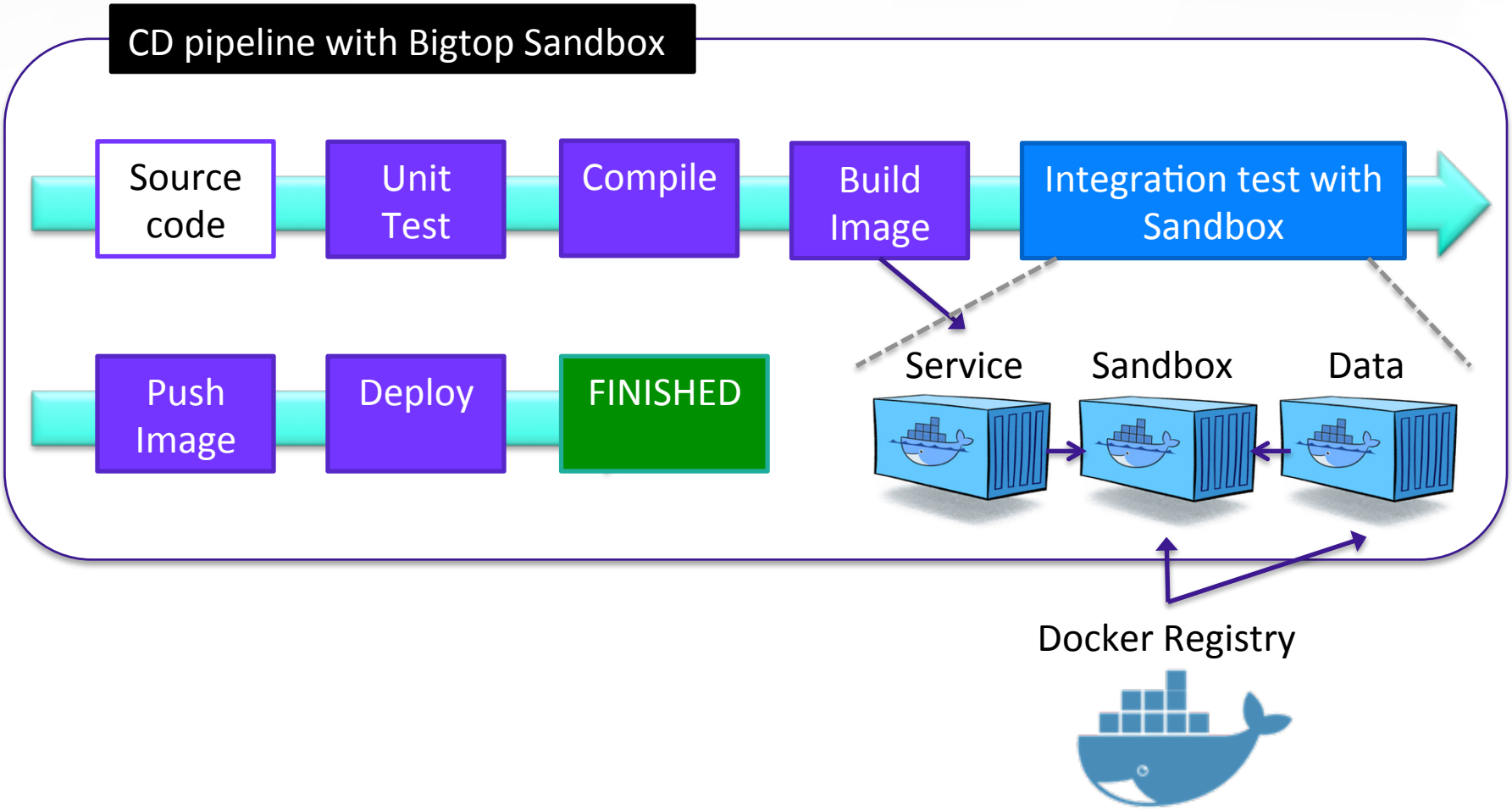
How to run

```
docker run --name sandbox -d \
-p 50070:50070 -p 8088:8088 \
bigtop/sandbox:apache_big_data_2017_miami

docker logs -f sandbox

docker exec sandbox spark-example SparkPi
```

	Bigtop Provisioner	Bigtop Sandbox
Scalable	V	X
Portable	X	V
Flexibility	High	Medium
Speed	> 2 mins	> 15 secs
Requires Network	V	X

	Bigtop Provisioner	Bigtop Sandbox
Data engineers	Multi-node cluster testing	Build/use sandboxes for dev & test
Ops	Multi-node cluster testing	Single node testing
Contributors	Test packages, puppet recipes, test cases	Test packages, puppet recipes, test cases
Distro. Builders	Test packages, puppet recipes, test cases	Provide Sandboxes

Integration test in CI/CD pipeline



YAHOO!

Future

- Production deployment using Sandbox image
  - --net host or SDN
  - External volumes for fsimage, data, logs, etc
  - Cluster orchestration
    - Kubernetes?

Release

Bigtop 1.2.0 Released Apr., 2017

- New components:
  - Ambari 2.5.0
  - GPDB 5.0.0-alpha.0 (Greenplum)
- Featured upgrade:
  - Hadoop 2.7.3
  - Spark 2.1.0
  - Kafka 0.10.1.1
  - HBase 1.1.3
  - and more

What's new in Bigtop 1.2.0?

- New features:
  - Juju bigtop charms
  - Bigtop Sandbox (alpha)
- Improvement:
  - Bigtop Docker Provisioner made faster

Juju Cloud Weather Report

Bundle	Date	Latest Test Results
cwr_bundle_bigdata_dev_hadoop_processing	Apr 13, 2017 at 03:01	AWS ✔ Azure ○ GCE ✔ lxd ▲
cwr_bundle_bigdata_dev_spark_processing	Apr 13, 2017 at 06:41	AWS ✔ Azure ○ GCE ▲ lxd ✘
cwr_bundle_hadoop_hbase	May 12, 2017 at 20:17	AWS ✔ Azure ✔ GCE ▲ lxd ○
cwr_bundle_hadoop_kafka	May 12, 2017 at 22:57	AWS ✔ Azure ✔ GCE ▲ lxd ○
cwr_bundle_hadoop_processing	May 12, 2017 at 20:37	AWS ✔ Azure ✔ GCE ▲ lxd ○
cwr_bundle_hadoop_spark	May 12, 2017 at 17:04	AWS ✔ Azure ✔ GCE ✔ lxd ○
cwr_bundle_spark_processing	May 12, 2017 at 23:26	AWS ✔ Azure ✔ GCE ▲ lxd ○
cwr_charm_release_ubuntu_devenv_in_cs_kwmonroe_java_devenv	May 10, 2017 at 20:05	AWS ✔ Azure ○ GCE ○ lxd ○

✔ Test Passed ✘ Test Failed ▲ Infrastructure Failure ○ No Test Result

http://bigtop.charm.qa/

YAHOO!

Road ahead

- AARCH 64 support
- Enhance support set in Bigtop Puppet
- Extend the CI matrix to Bigtop Tests
- Ambari Bigtop integration
- Big data stack references

We want you!

- Join mailing list, ask questions, suggest features, etc
- Contribute (components, tutorials, docs)
- Report bugs

- Reference
  - Home page: http://bigtop.apache.org/
  - mailing list: http://bigtop.apache.org/mail-lists.html
  - Document: https://cwiki.apache.org/confluence/display/BIGTOP/Index
  - Source code: https://github.com/apache/bigtop
  - Packages: https://www.apache.org/dist/bigtop/bigtop-1.2.0/repos/
  - JIRA: https://issues.apache.org/jira/browse/BIGTOP

Thank you !

Questions?