How Bigtop Leveraged Docker for Build Automation and One-Click Hadoop Provisioning

Evans Ye

Apache Big Data 2015
Budapest

APACHE: BIG_DATA EUROPE

TREND MICRO™

Securing Your Journey to the Cloud

Who am I

- Apache Bigtop PMC member
- Software Engineer at Trend Micro
- Develop Big Data platform
- Develop cyber security solutions using Big Data

TREND MICRO™

Outline

- What is Apache Bigtop?
- Achieving Build Automation
- One-Click Hadoop Provisioning
- Bigtop at Trend Micro

TREND MICRO™

What is Apache Bigtop ?

TREND MICRO™

Bigtop is a project for

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

for you to easily build your own Big Data Stack

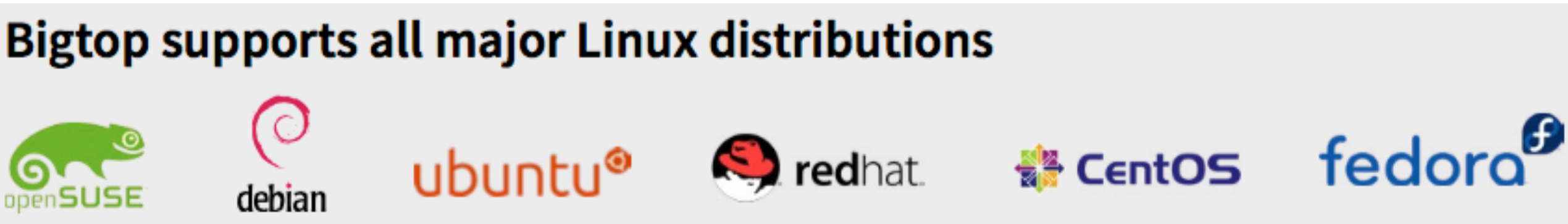
TREND MICRO™

Packaging

- **Build RPM, DEB packages for Hadoop ecosystem**
  - $ yum -y install hadoop
  - $ apt-get upgrade hbase
- **Why not just untar ?**
  - Version control, upgrade/downgrade, dependency management
  - Unified view of configuration, log, binary directories
  - Daemons for better process management

TREND MICRO™

Supported Linux distro.



- You can find binary repos here:
  - http://www.apache.org/dist/bigtop/bigtop-1.0.0/repos/

TREND MICRO™

Supported components

Testing

- A fundamental testing problem may happen in your Big Data Stack, too

Hadoop 2.4.1	Tez 0.6.0	Hive 1.1.0
Hadoop 2.6.0	+ Tez 0.6.2	+ Hive 1.2.0
Hadoop 2.7.1	Tez 0.7.0	Hive 1.2.1

TREND MICRO™

Testing

- A fundamental testing problem may happen in your Big Data Stack, too

Hadoop 2.4.1		Tez 0.6.0		Hive 1.1.0
Hadoop 2.6.0	+	Tez 0.6.2	+	Hive 1.2.0
Hadoop 2.7.1		Tez 0.7.0		Hive 1.2.1

TREND MICRO™

Testing

- A fundamental testing problem may happen in your Big Data Stack, too

Does this combination still work?

Hadoop 2.6.0	+	Tez 0.6.2	+	Hive 1.2.0
Hadoop 2.7.1		Tez 0.7.0		Hive 1.2.1

TREND MICRO™

There’s a need for integration tests !

Bigtop Tests

- Bigtop has a testing framework that provides APIs
- Built-in tests that Bigtop provides:
  - **smoke-tests**
    - Basic Hadoop ecosystem interoperability
  - **integration tests**
    - Advanced tests runs from jar files

TREND MICRO™

Okay,
then how to run those tests?

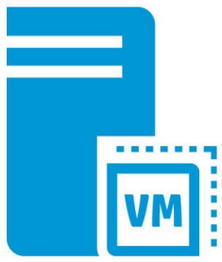
TREND MICRO™

Deployment

- Use Bigtop Puppet to deploy a fully functional, distributed Hadoop cluster
- Integration tests running on a fully distributed cluster
- **Bigtop Puppet**
  - Masterless Puppet
  - Numbers of components are supported
  - Kerberos enabled cluster deployment

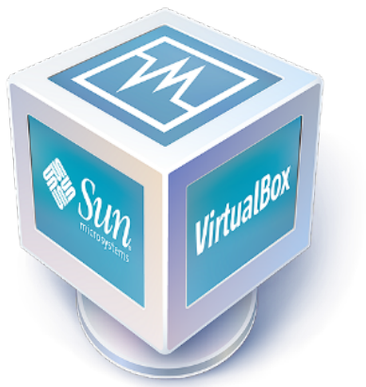
Where to deploy ?

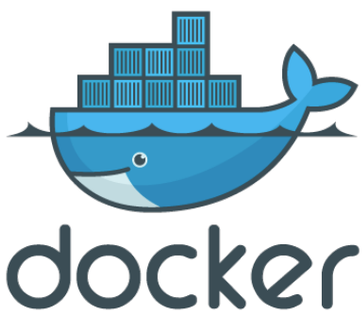
TREND MICRO™



Virtualization

- Automate the infrastructure creation & setup
- **Bigtop Provisioner**
  - Virtualbox VMs auto-provisioning
  - Docker containers auto-provisioning





TREND MICRO™

From user point of view

- **Hadoop App developers**
  - Provision Hadoop cluster to test your code on
- **Cluster administrators**
  - Use Bigtop Puppet to deploy and manage your cluster
  - Run Bigtop tests to ensure your cluster is working
- **Vendors**
  - Build your own Hadoop distribution, customized from Bigtop bits

TREND MICRO™

What is Apache Bigtop?

Achieving Build Automation

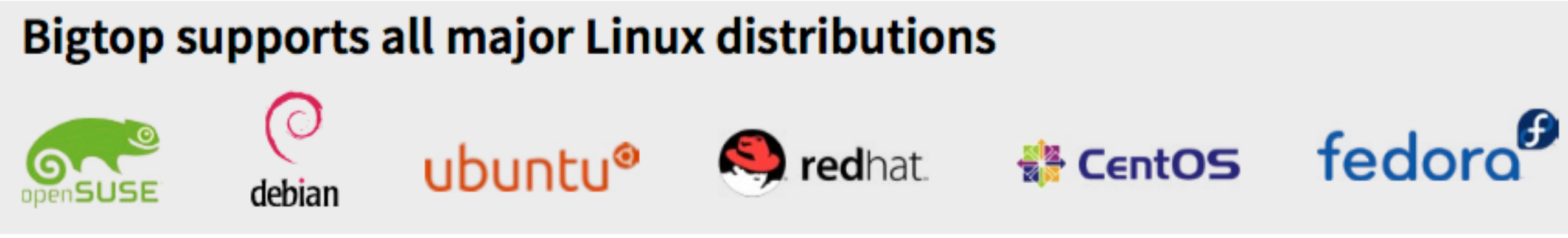
One-Click Hadoop Provisioning

Bigtop at Trend Micro

TREND MICRO™

What's the challenge ?

- As you can see,
  we support a number of Linux distributions **(m)**



TREND MICRO

What's the challenge ?

- Numbers of components **(n)**

We need to build them all

—> **m * n**

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 • Apache Ant • Apache Maven • wget • tar • git • subversion • gcc • gcc-c++ • make • fuse • protobuf-compiler • autoconf • automake • libtool • sharutils • xmlto	• lzo-devel • zlib-devel • fuse-devel • openssl-devel • python-devel • libxml2-devel • libxslt-devel • cyrus-sasl-devel • sqlite-devel • mysql-devel • openldap-devel • rpm-build • createrepo • redhat-rpm-config (RedHat/CentOS only)	• libxslt1-dev • libkrb5-dev • libldap2-dev • libmysqlclient-dev • libsasl2-dev • libsqlite3-dev • libxml2-dev • python-dev • python-setuptools • liblzo2-dev • libzip-dev • libfuse-dev • libssl-dev • build-essential • dh-make • debhelper • 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 • Apache Ant • Apache Maven • wget • tar • git • subversion • gcc • gcc-c++ • make • fuse • protobuf-compiler • autoconf • automake • libtool • sharutils • xmlto	• lzo-devel • zlib-devel • fuse-devel • openssl-devel • python-devel • [illegible]2-devel • [illegible]t-devel • [illegible]sal-devel • sqlite-devel • mysql-devel • openldap-devel • rpm-build • createrepo • redhat-rpm-config (RedHat/CentOS only)	• libxslt1-dev • libkrb5-dev • libldap2-dev • libmysqlclient-dev • libsasl2-dev • libsqlite3-dev • libxml2-dev • python-dev • python-setuptools • liblzo2-dev • libzip-dev • libfuse-dev • libssl-dev • build-essential • dh-make • debhelper • devscripts • reprepro

■ ■ ■

Seriously ?

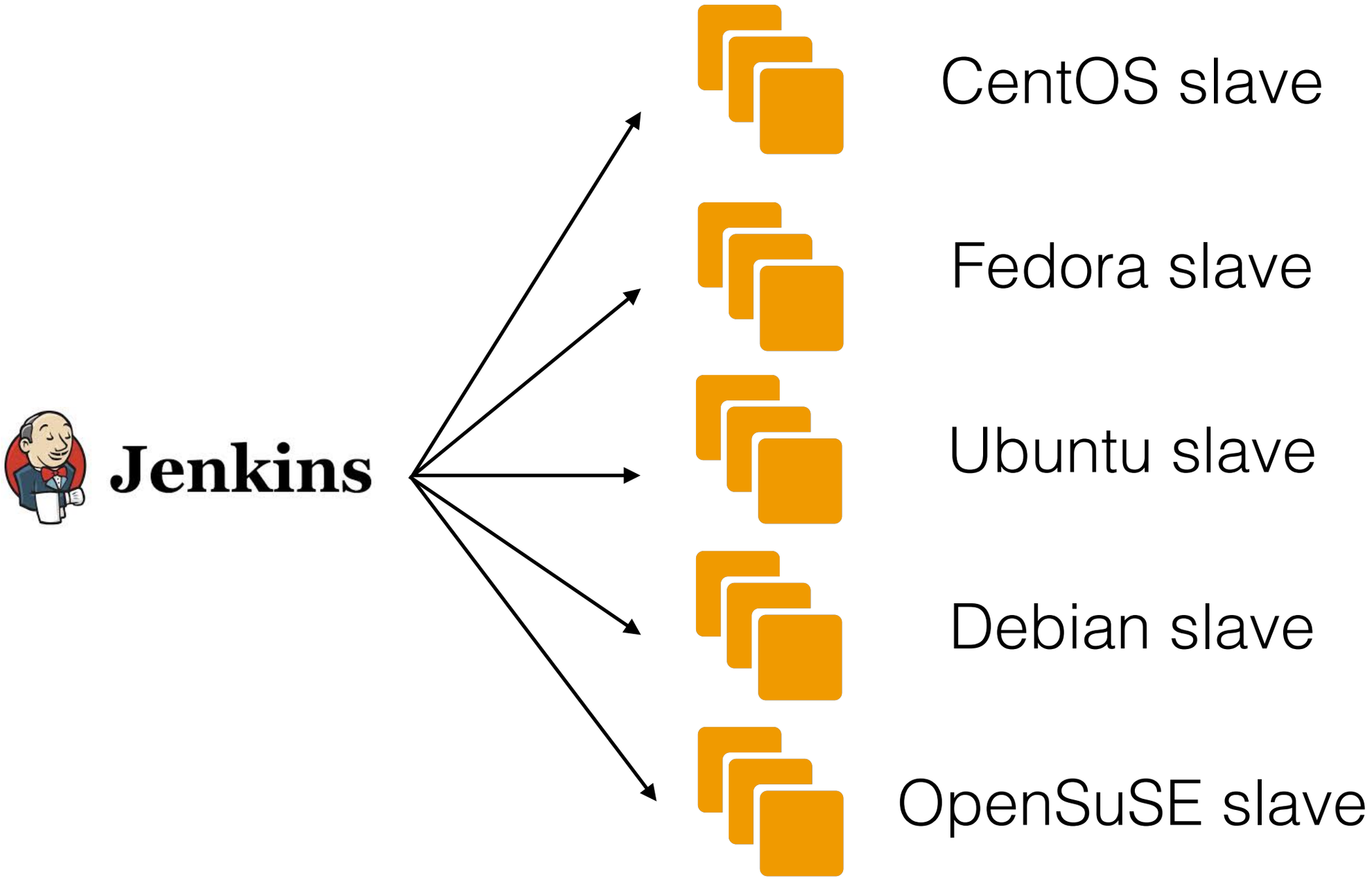
TREND MICRO™

Bigtop Toolchain

- Puppet recipes automatically install required libraries, build tools
- To transform a machine into a build environment, simply do
  - $ gradle toolchain
- **Prerequisites :**
  - Puppet 3.x
  - Java 7

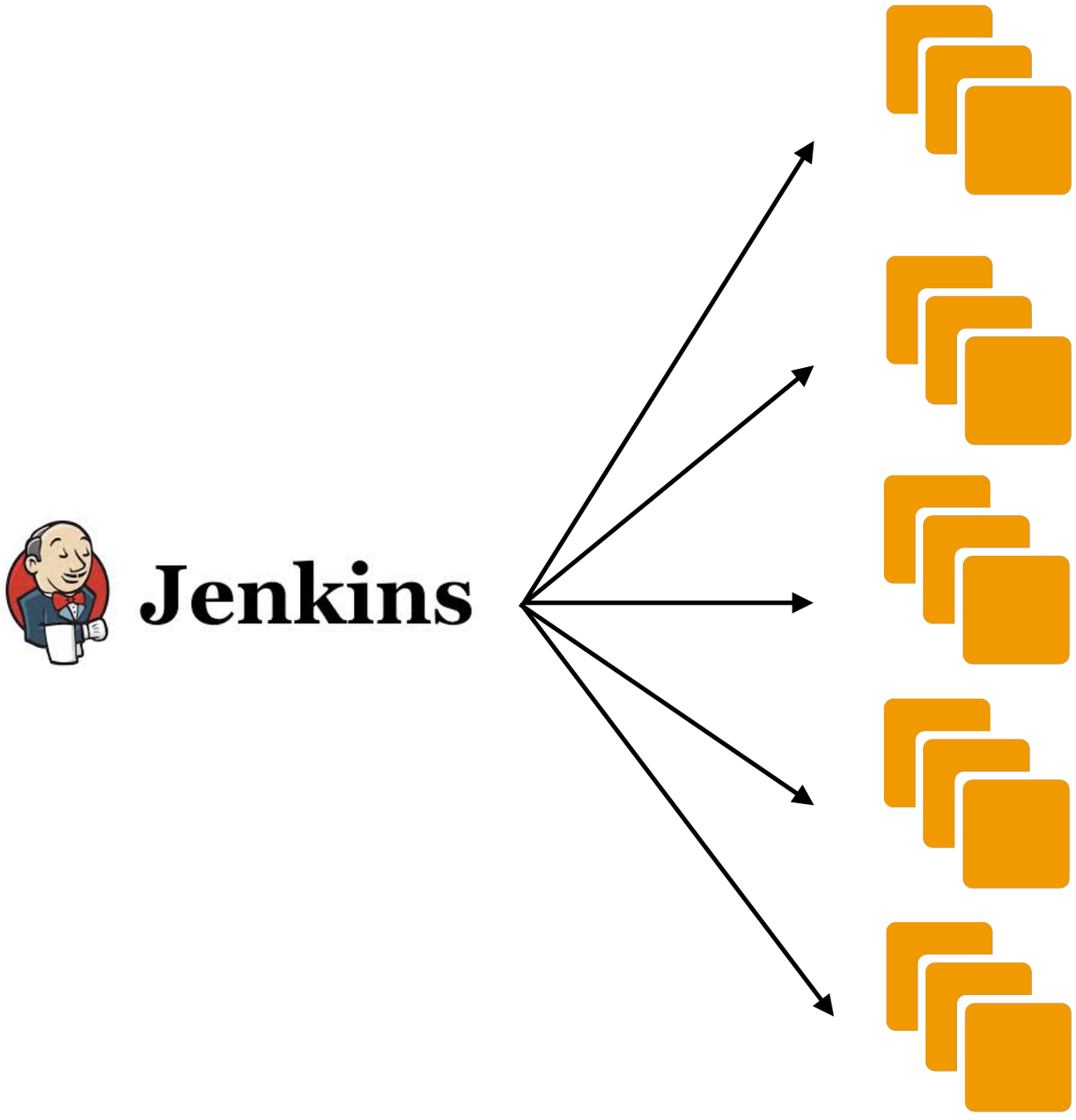
TREND MICRO™

CI infra



TREND MICRO™

CI infra



CentOS slave
Bigtop Toolchain

Fedora slave
Bigtop Toolchain

Ubuntu slave
Bigtop Toolchain

Debian slave
Bigtop Toolchain

OpenSuSE slave
Bigtop Toolchain

TREND MICRO™

https://www.flickr.com/photos/kit4na/6385016345

TREND MICRO™

Hi Bigtop, How do I test my packaging code?

We got Bigtop Toolchain for you.

A contributor

A committer

TREND MICRO™

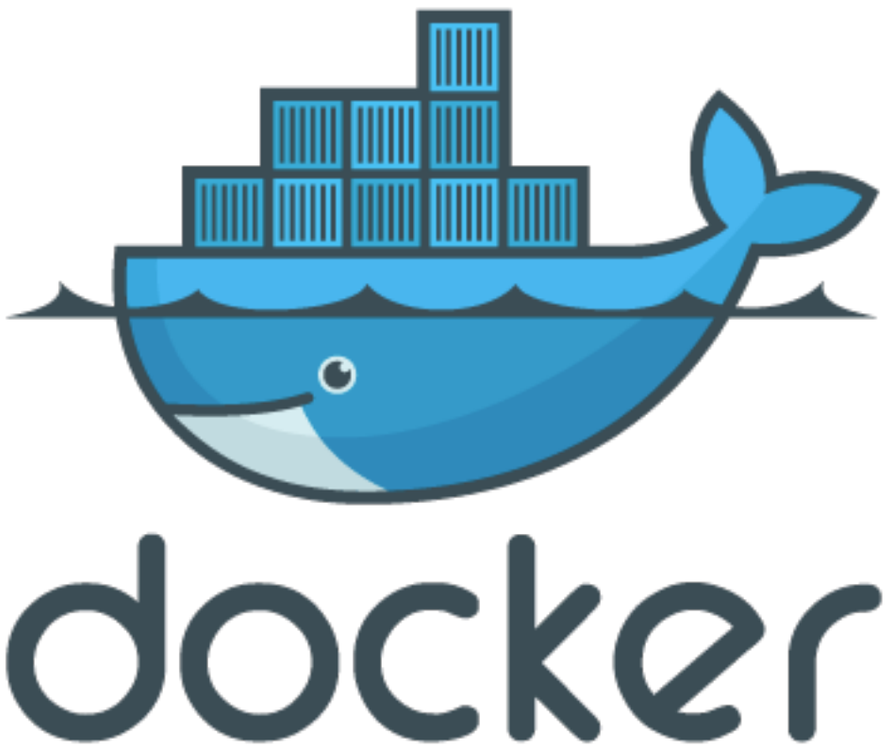
I need to setup all the Linux environments on my laptop?

(Don’t want to answer…)

A contributor

A committer

TREND MICRO™



TREND MICRO™

Docker - How it works

- Three key techniques in Docker
  - Use Linux Kernel's **Namespaces** to create isolated resources
  - Use Linux Kernel's **cgroups** to constrain resource usage
  - **Union file system** that supports git-like image creation

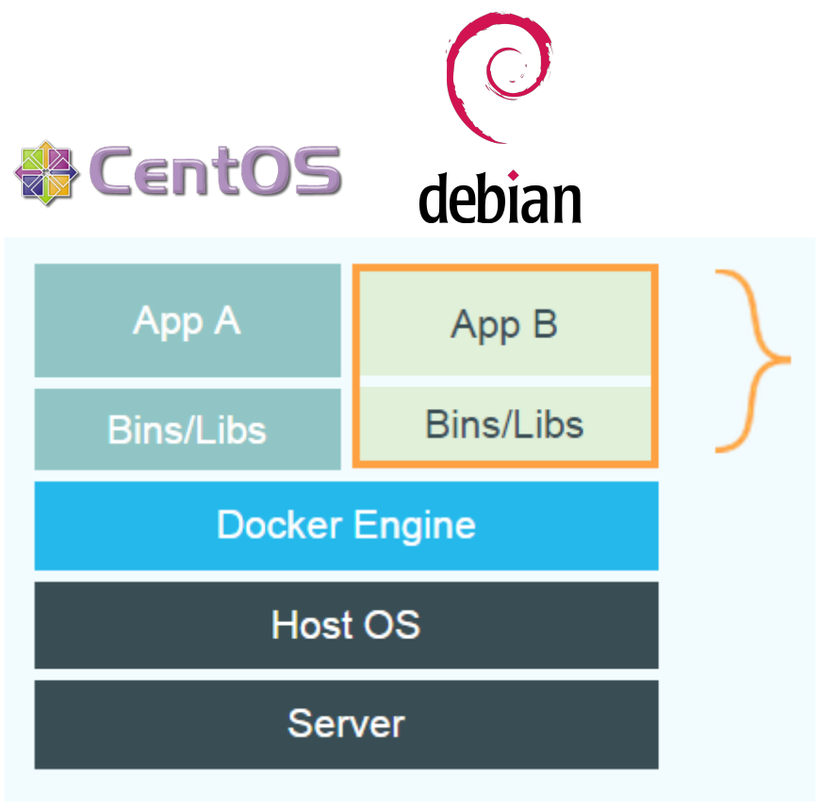
TREND MICRO™

The nice things about it

- Lightweight
- Fast creation
- Repeatable
- Portability
  - runs on *any* Linux environment

TREND MICRO™

Best fit in our use case



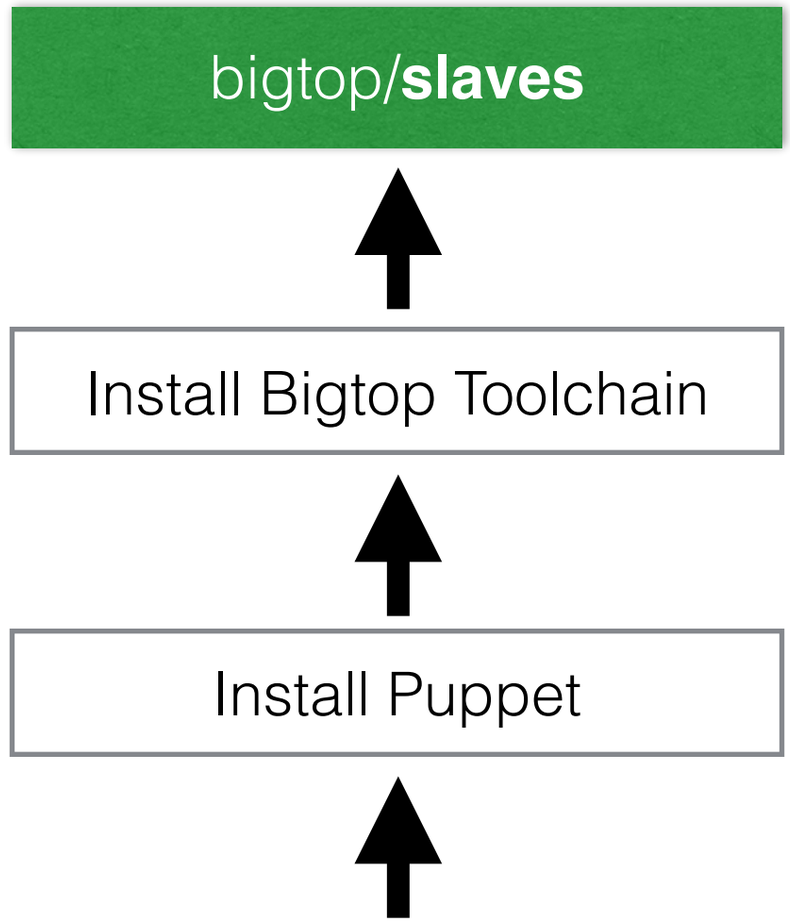
OS specific Bins/Libs

- Run on any kind of Linux distribution in one single machine

TREND MICRO

Ship build environment in Docker images

Bigtop/slaves images



Dockerhub official images

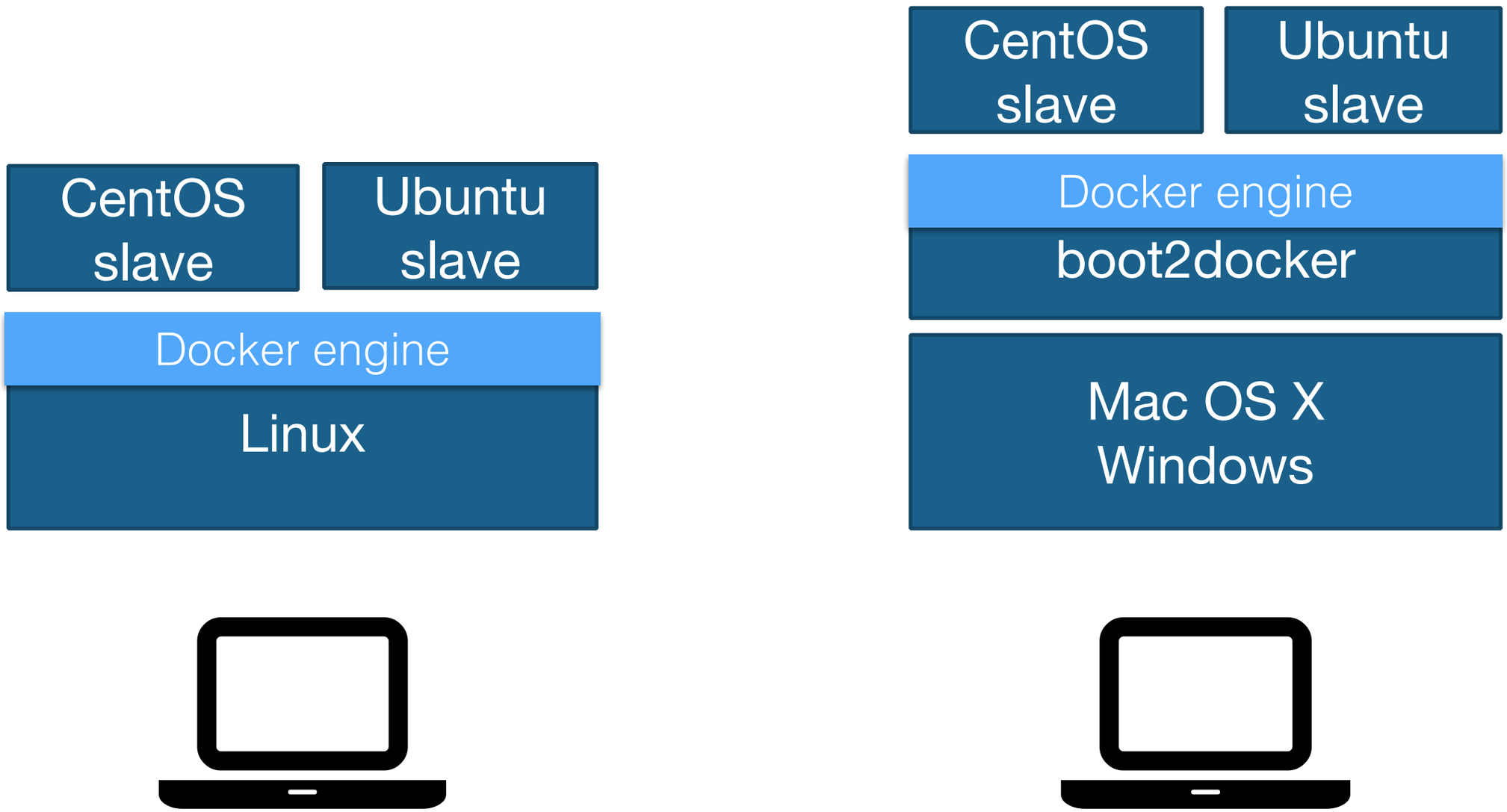
TREND MICRO™

One click to build packages

- $ git clone https://github.com/apache/bigtop.git
- $ docker run \
  --rm \
  --volume `pwd`/bigtop:/bigtop \
  --workdir /bigtop \
  bigtop/slaves:centos-7 \
  bash -l -c ‘./gradlew rpm’

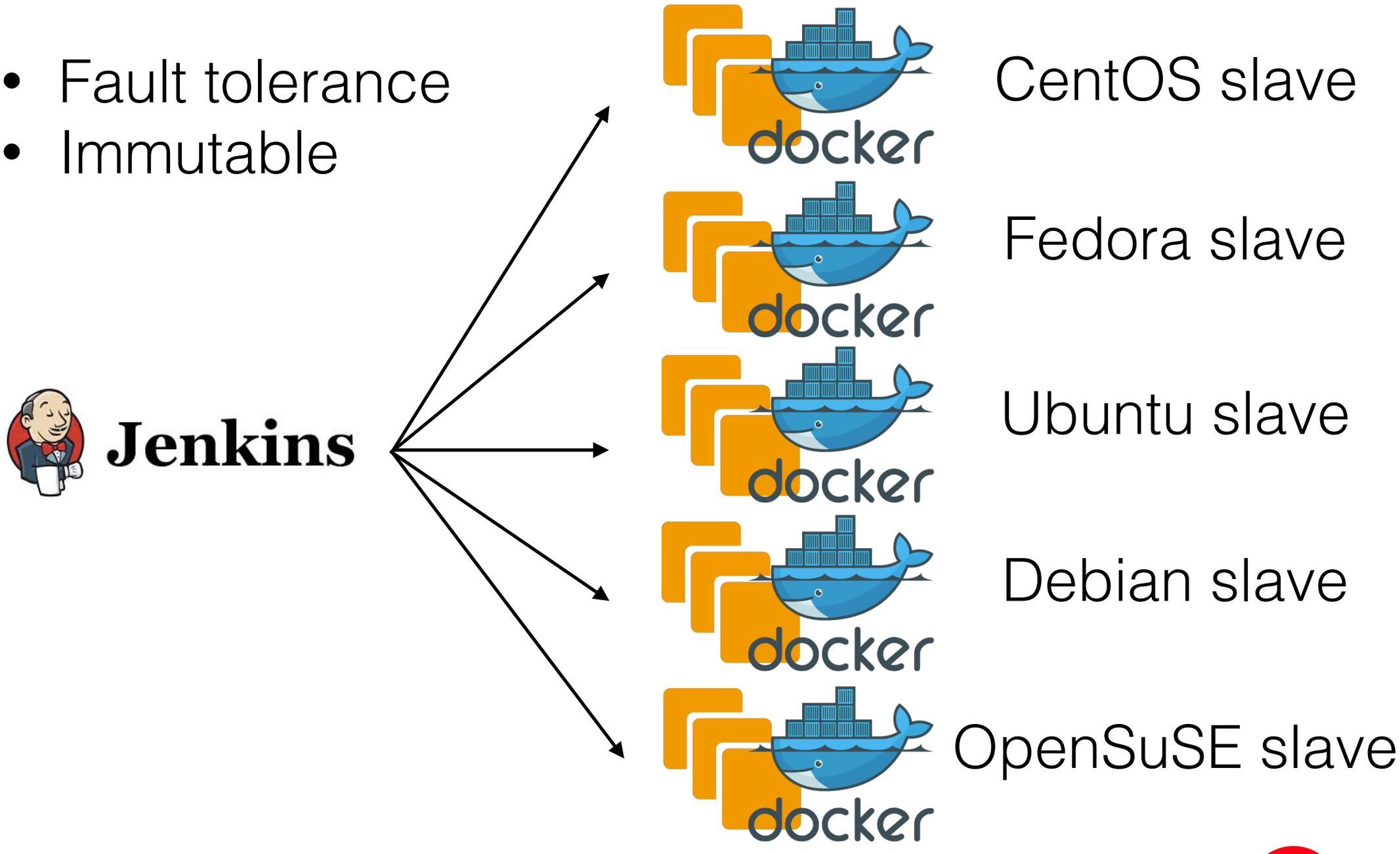
TREND MICRO™

Now, easy to build & test on your laptop



Flexible CI infra

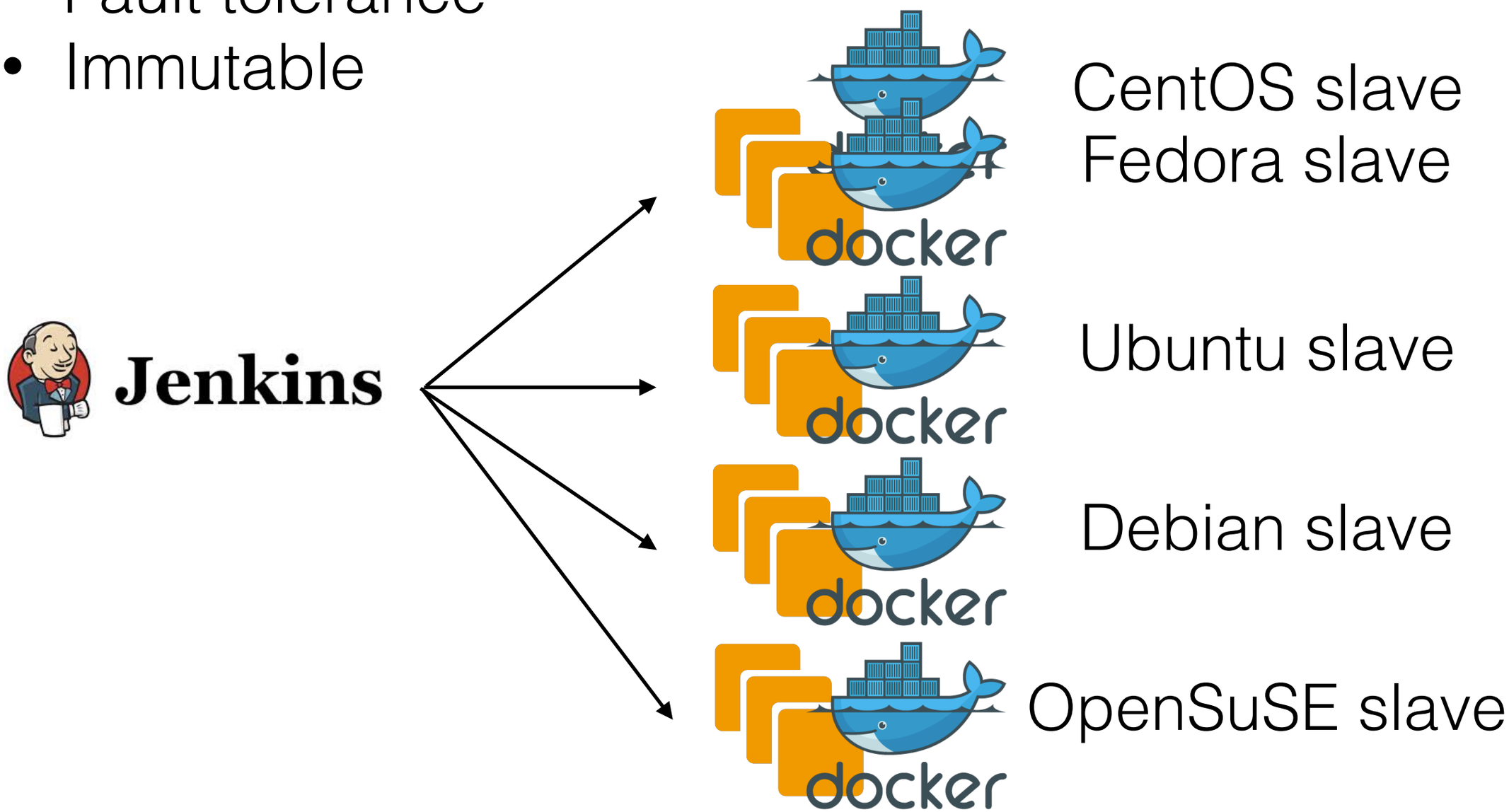
- Fault tolerance
- Immutable



TREND MICRO

Flexible CI infra

- Fault tolerance
- Immutable



TREND MICRO™

What is Apache Bigtop?

Achieving Build Automation

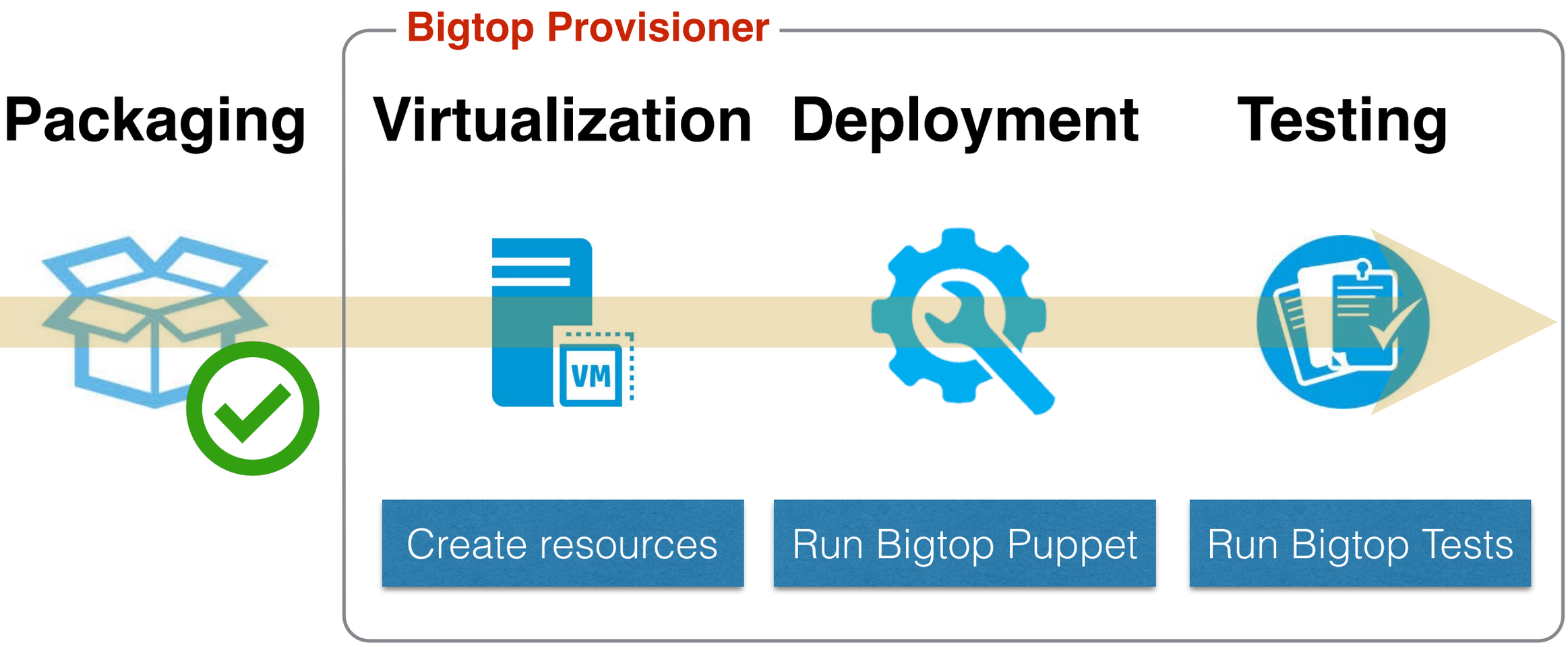
One-Click Hadoop Provisioning

Bigtop at Trend Micro

TREND MICRO™

Bigtop Provisioner

- A tool to demonstrate the full life cycle of Bigtop



TREND MICRO™

The goal

- **Fast iterative development**
  - Test your code in the cluster, on your laptop, w/o human intervention
- **Flexibility**
  - Choose any combination of components as you want
- **Responsive CI**
  - Integration tests that get you the result in mins
- **A Big Data Stack playground**
  - Spark + Tachyon, Spark + Ignite, etc

TREND MICRO™

Bigtop Provisioner

Vagrant w/ provider

+

Automation Code

+

Bigtop Puppet

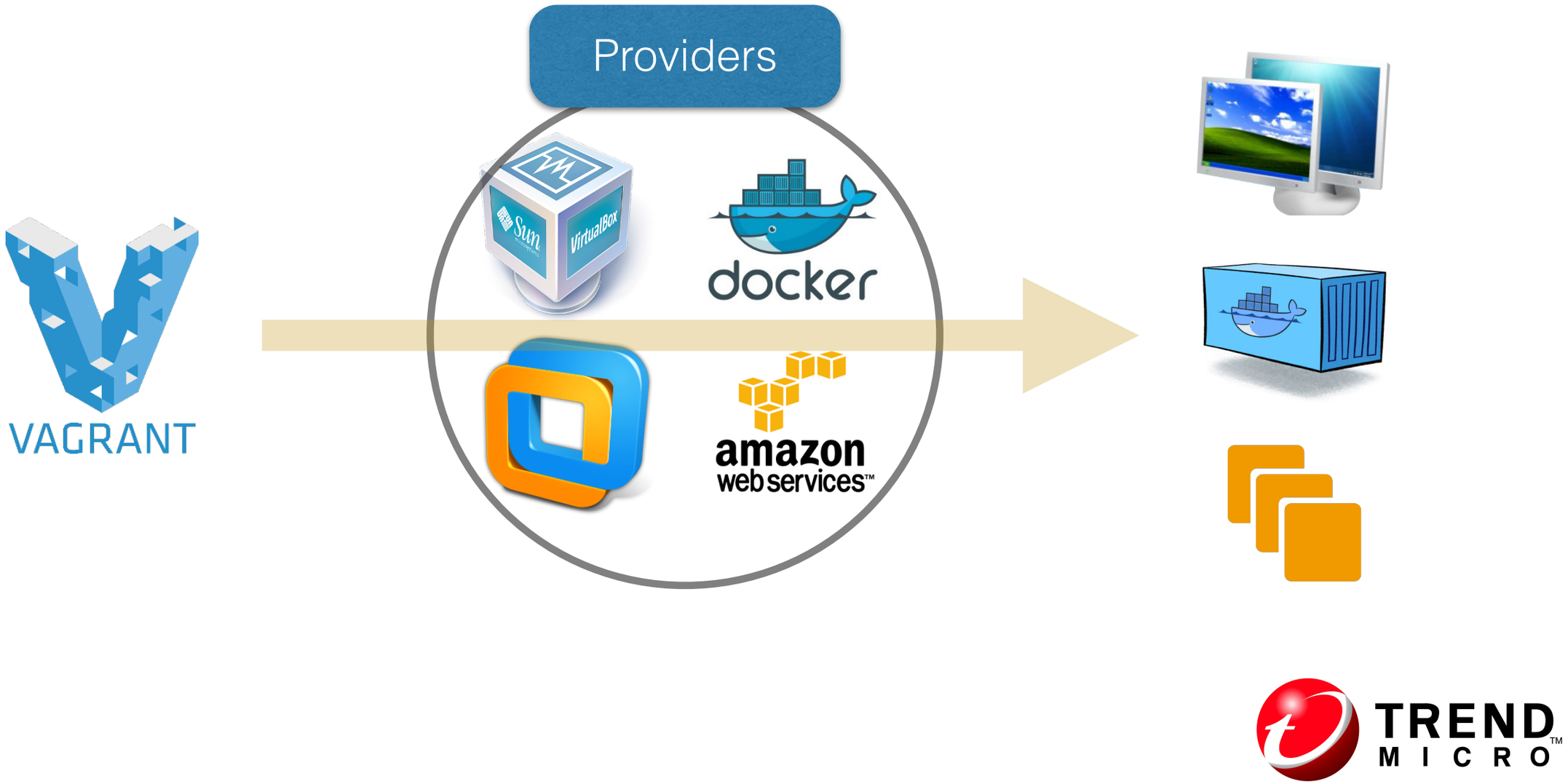
=

One-click Hadoop Provisioning

TREND MICRO™

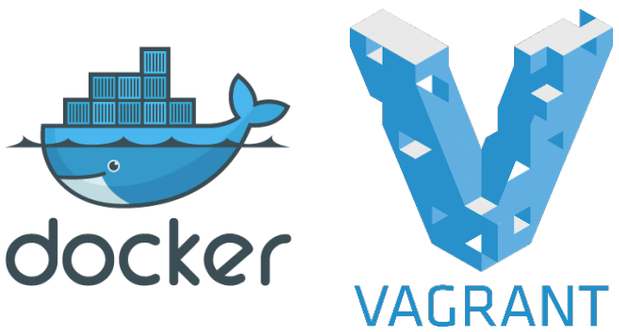
Vagrant

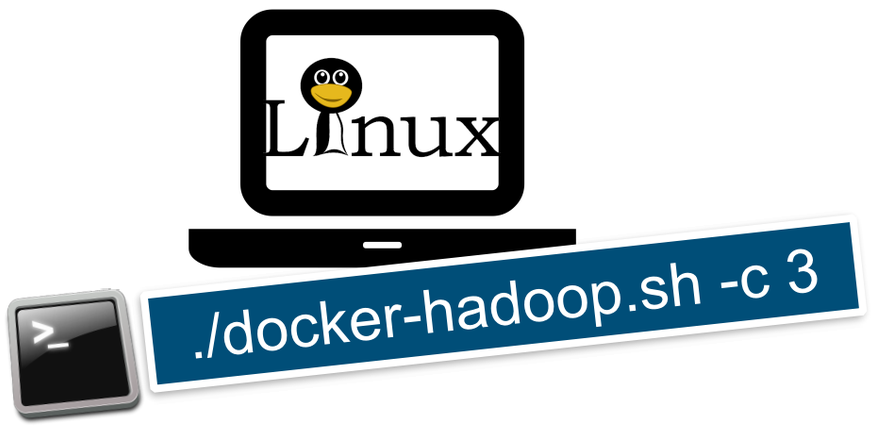
- We use Vagrant as an abstraction layer to support different kind of resource providers



TREND MICRO™

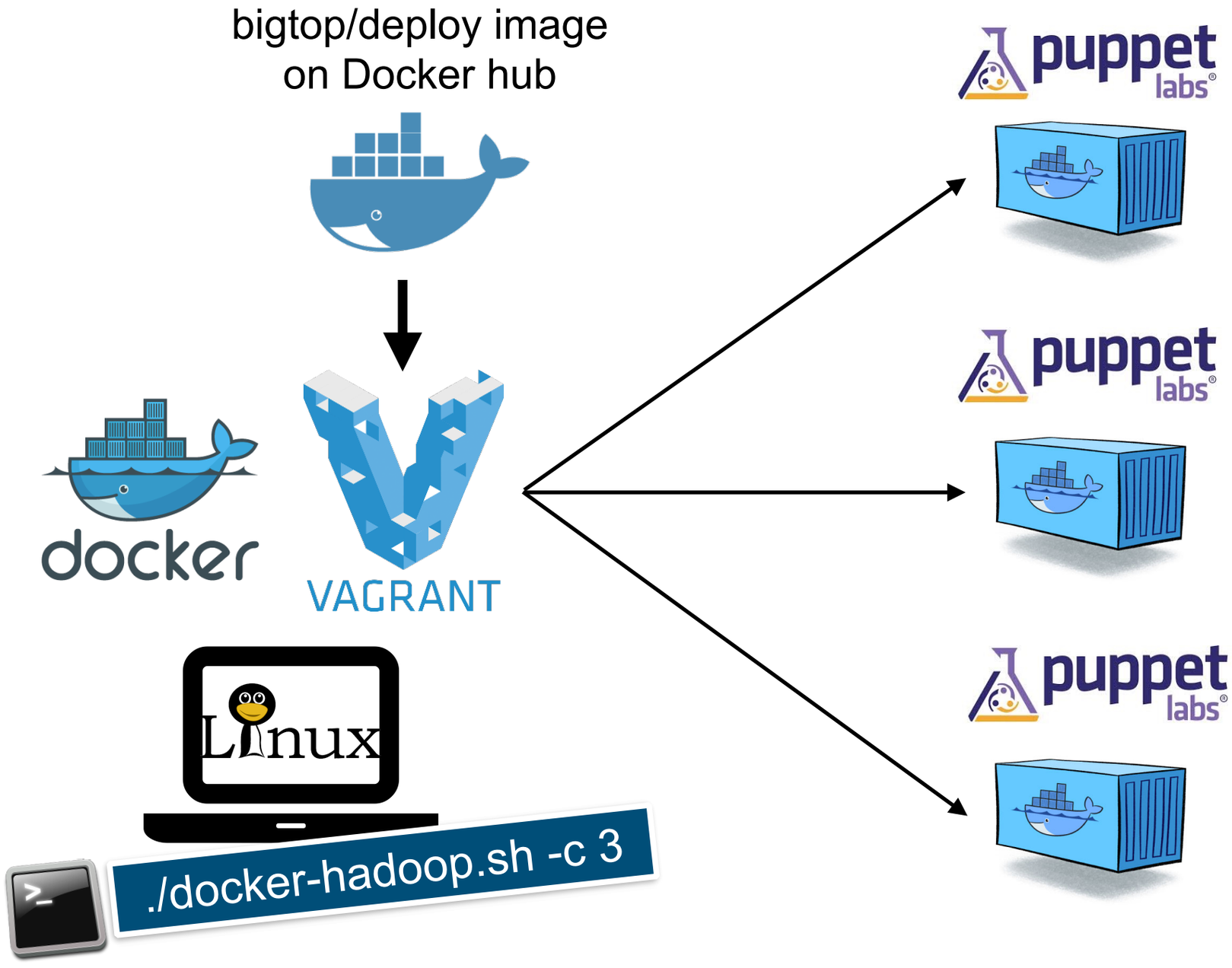
One click Hadoop provisioning





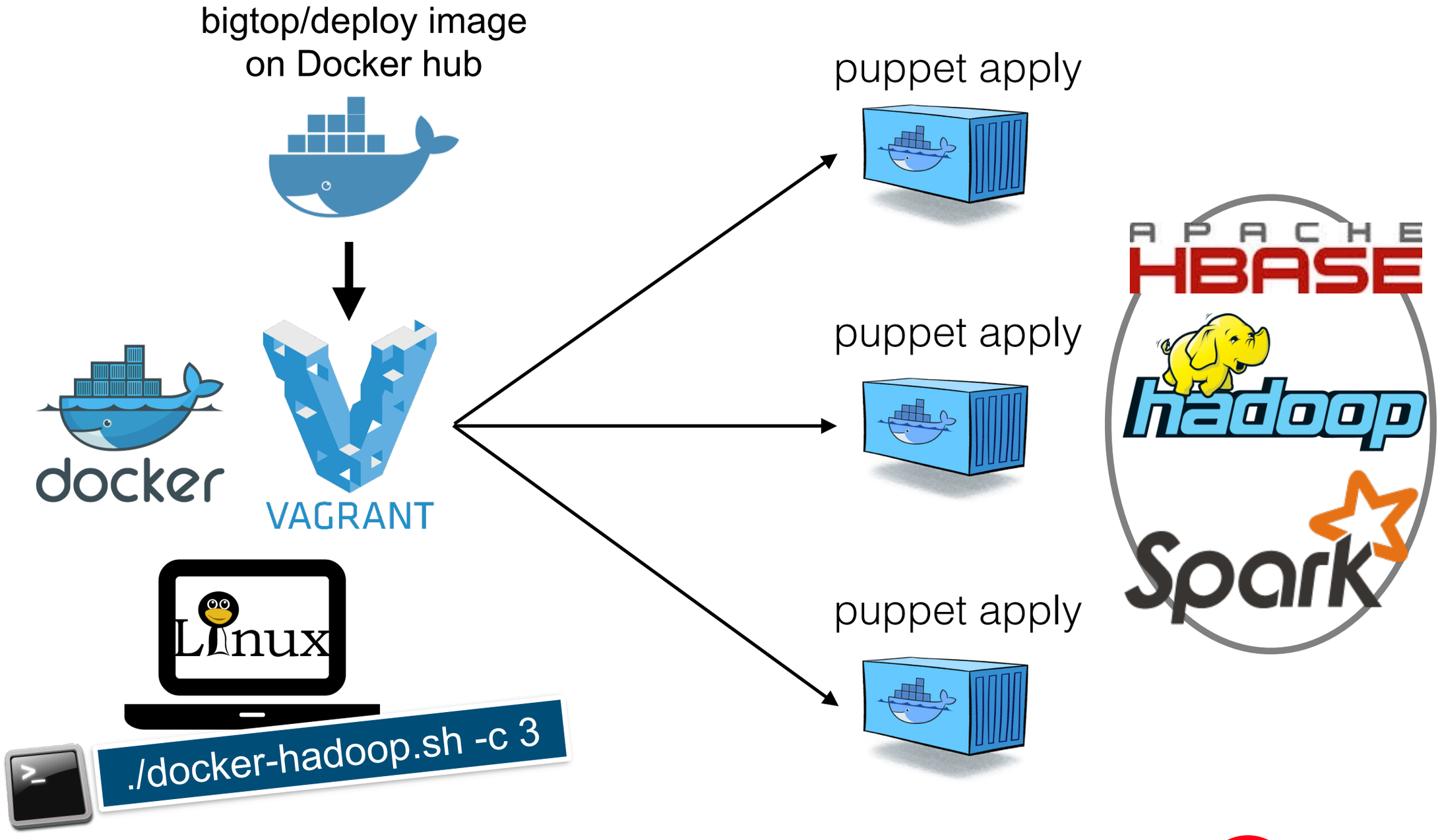
TREND MICRO™

One click Hadoop provisioning



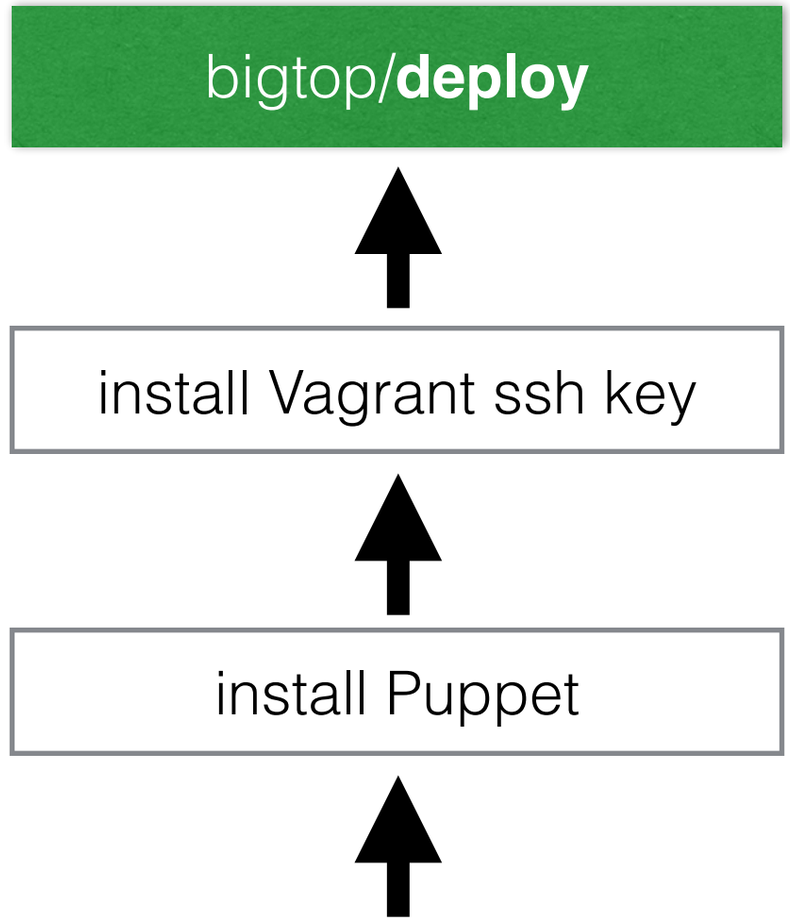
TREND MICRO

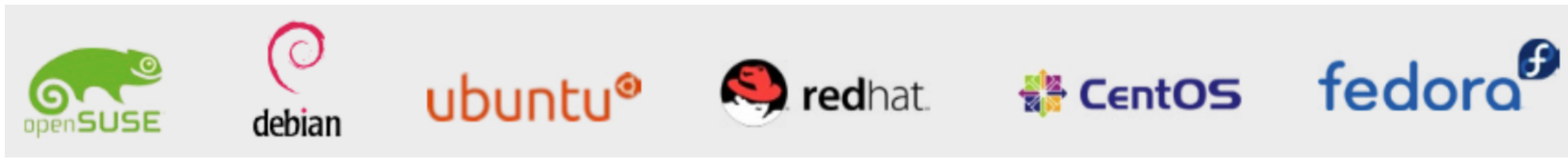
One click Hadoop provisioning



TREND MICRO™

Bigtop/deploy images





Dockerhub official images

TREND MICRO

Bigtop image hierarchy

bigtop/**slaves**

diff

bigtop/**deploy**

bigtop/**puppet**

=

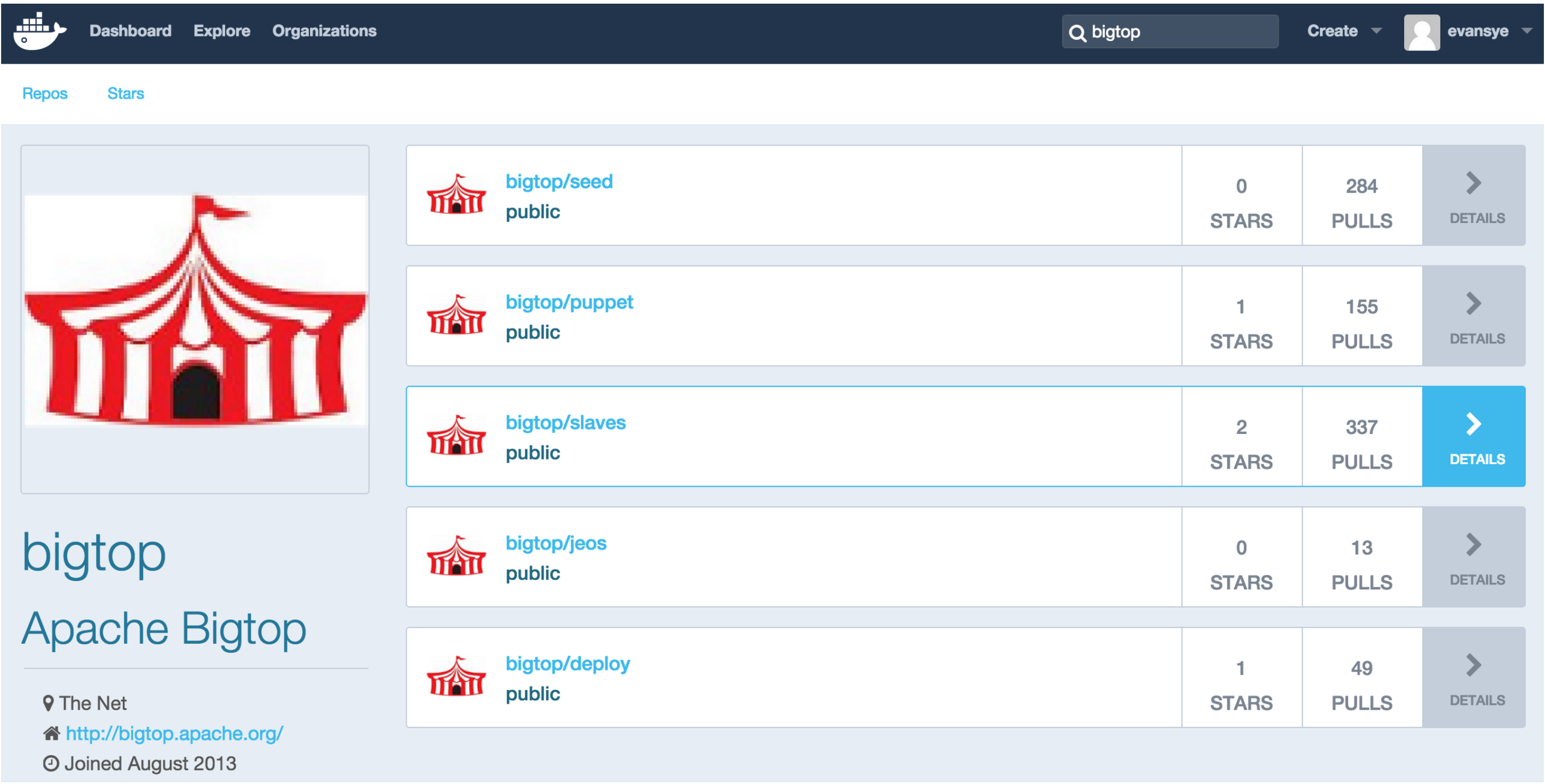
bigtop/**puppet**

Dockerhub official images

openSUSE debian ubuntu redhat CentOS fedora

TREND MICRO

All on the Docker hub



https://hub.docker.com/u/bigtop/

TREND MICRO

A demo is worth 1k words

Oops! demo fail…

https://asciinema.org/a/55aw3zl2g3dzsfe69198homrl

TREND MICRO™

Bigtop Provisioner

- Supported providers in Bigtop 1.0.0 release
  - Virtaulbox VM
  - Docker container
- OpenStack support is in master branch now

TREND MICRO™

Use cases

- **For Hadoop app 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 Stack
- **For contributors**
  - Easy to test your packaging, deployment, testing code
- **For vendors**
  - CI out of the box —> patch upstream code made easier

TREND MICRO™

What is Apache Bigtop?

Achieving Build Automation

One Click Hadoop Provisioning

Bigtop at Trend Micro

TREND MICRO™

Trend Micro Hadoop (TMH)

- Use Bigtop as the basis for our internal custom distribution of Hadoop

- Apply **community**, **internal patches** to upstream projects for business and operational need

- Newest TMH7 is based on Bigtop 1.0 SNAPSHOT

TREND MICRO™

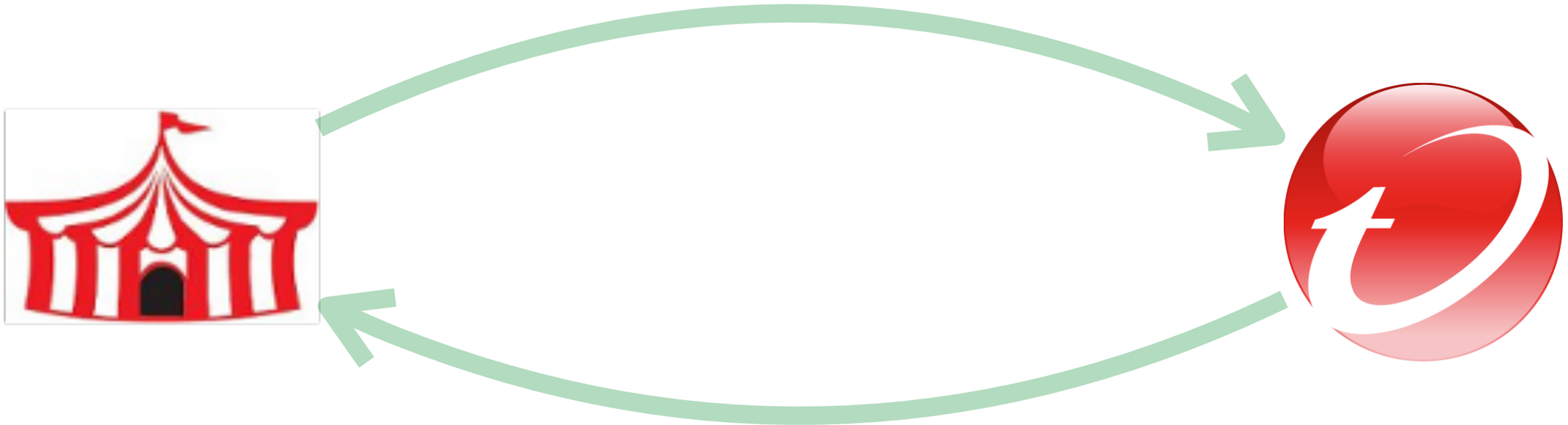
Working with community made our life easier

- Knowing community status made TMH7 release based on Bigtop 1.0 SNAPSHOT possible

TREND MICRO™

Working with community made our life easier

- Knowing community status made TMH7 release based on Bigtop 1.0 SNAPSHOT possible

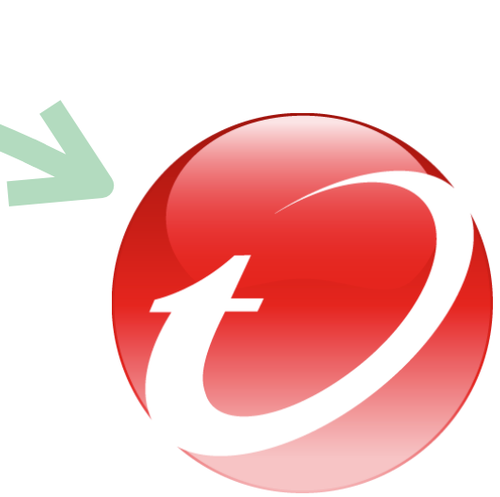


- Contribute Bigtop Provisioner, packaging code, puppet recipes, bugfixes, CI infra, anything!

TREND MICRO™

Working with community made our life easier

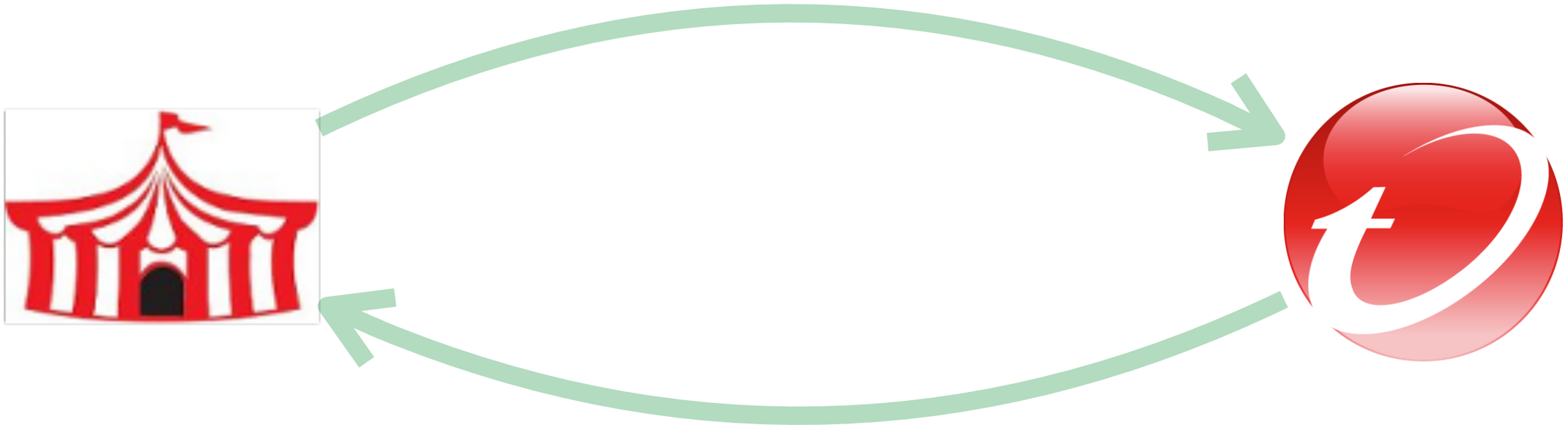
- Leverage Bigtop smoke tests and integration tests with Bigtop Provisioner to evaluate TMH7



TREND MICRO™

Working with community made our life easier

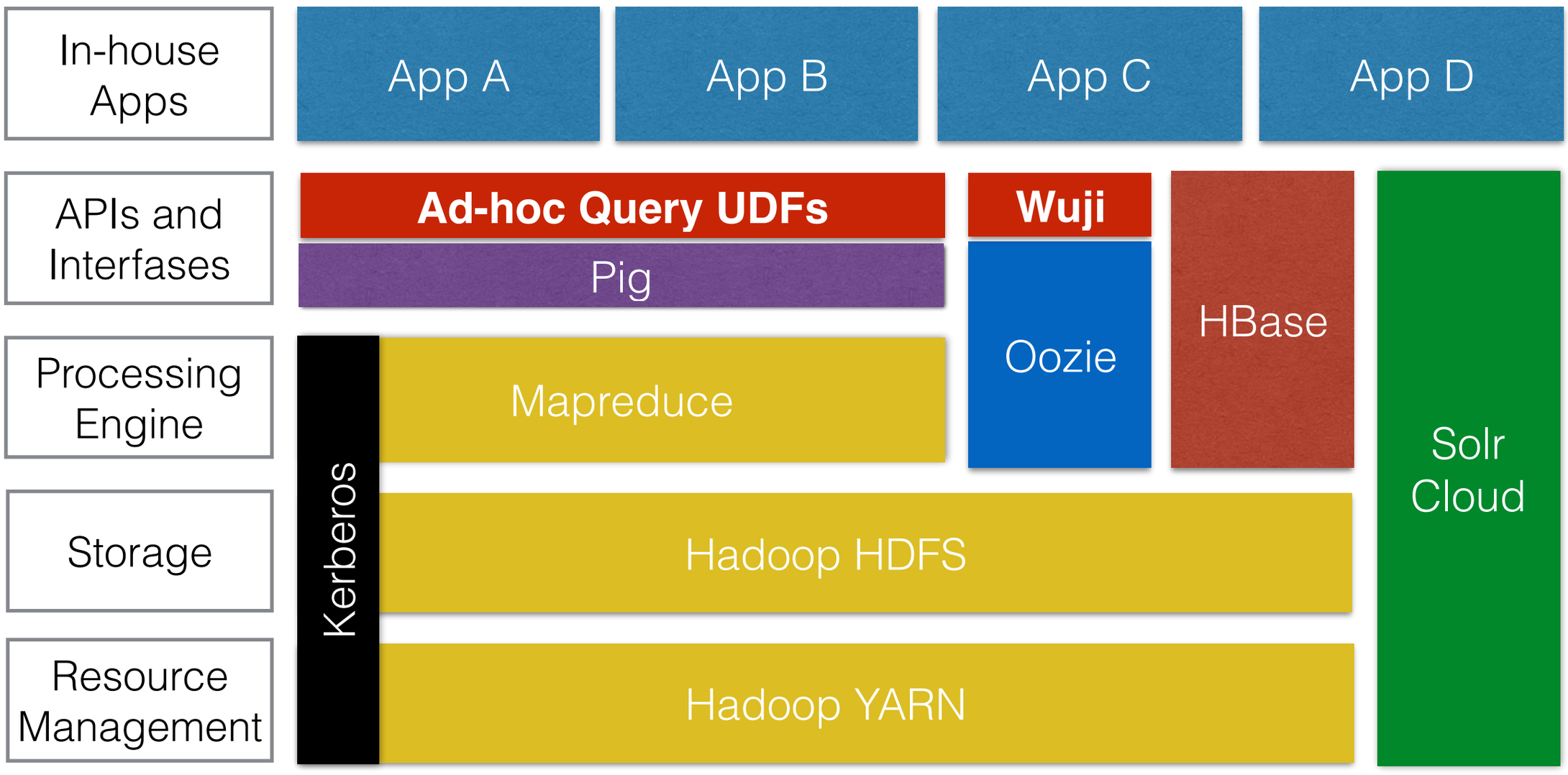
- Leverage Bigtop smoke tests and integration tests with Bigtop Provisioner to evaluate TMH7



- Contribute feedback, evaluation, use case through Production level adoption

TREND MICRO™

Trend Micro Big Data Stack Powered by Bigtop



TREND MICRO™

Challenges we’re facing !

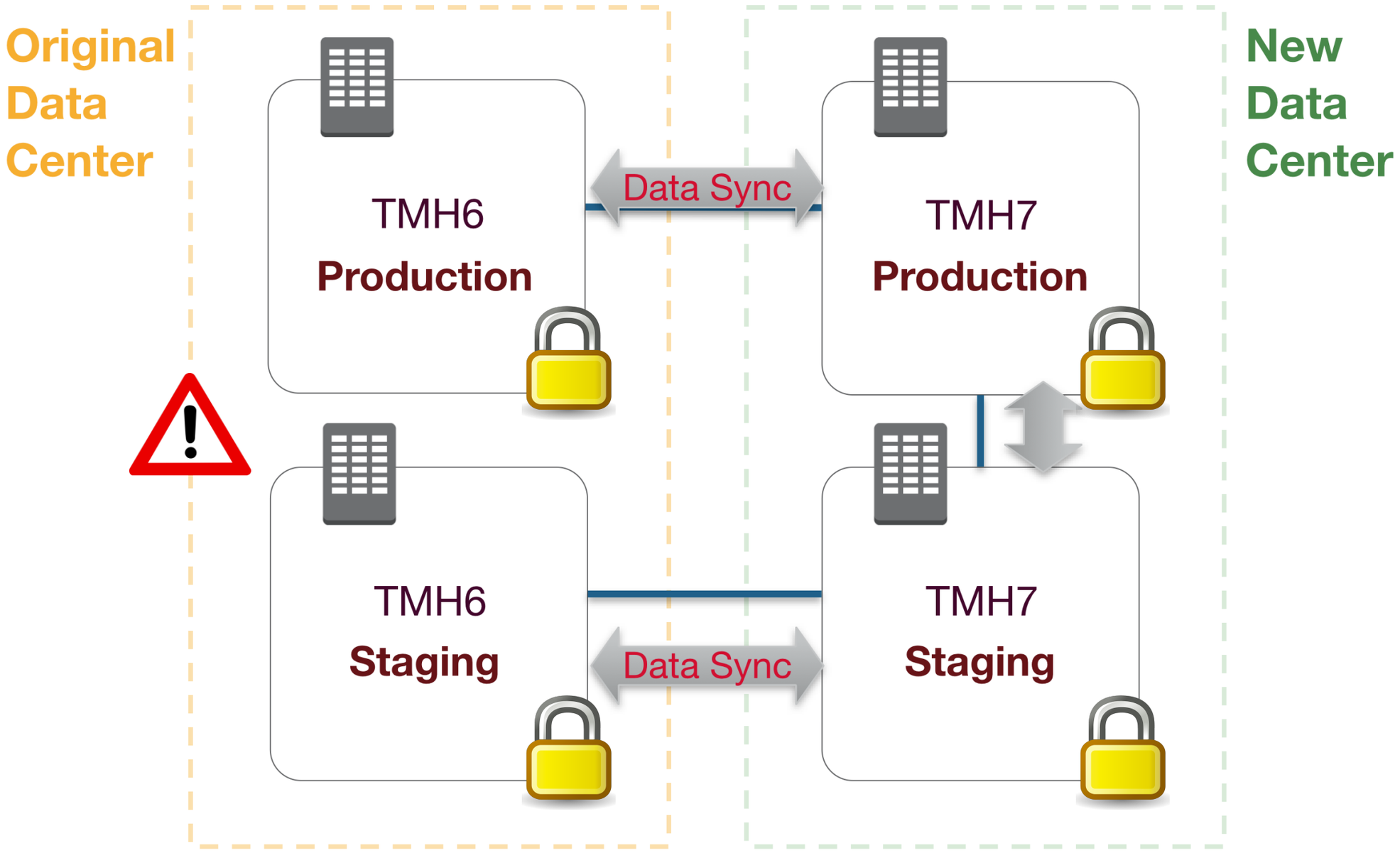
TREND MICRO™

Migration x Upgrade

- We’re moving our cluster to another Data Center
- New cluster will be running **TMH7**
- Need Distcp to sync data from old **TMH6** cluster
- Both are **Kerberos** enabled secure clusters

TREND MICRO™

Sync data between clusters



Distcp between TMH6 and TMH7 ?

Hadoop 2.0.0

Hadoop 2.6.0

Factors that matter

- Hadoop version
- Protocols (hdfs, hftp, webhdfs)
- Where to issue Distcp (Where to run MR job)
- Secure or not

TREND MICRO™

Things always get complicated with Kerberos…

TREND MICRO™

Kerberos cross realm authentication

- We need to do Distcp across 4 secure clusters
  —> Kerberos cross realm auth across 4 clusters

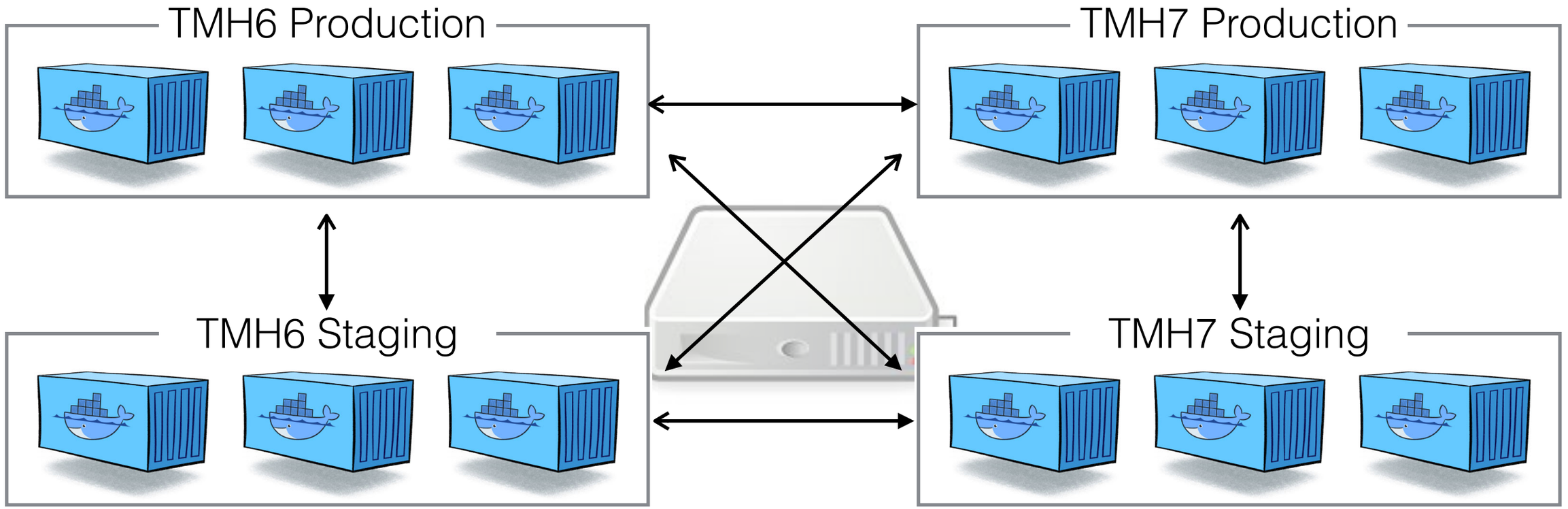
- Our Hadoop management tool needs to support this through auto-configuring

- Developing the management tool is challenging !

TREND MICRO™

Docker comes to the rescue

- Run multiple secure HA clusters on Docker
- Dev & test the management tool iteratively



TREND MICRO™

Hadoop apps dev & test on Docker

Hadoocker

- A Devops toolkit for Hadoop app developer
  to develop and test its code on

- Provides fixed Big Data Stack preload images
  —> dev & test env up and running w/o deployment
  —> support end-to-end CI test for apps

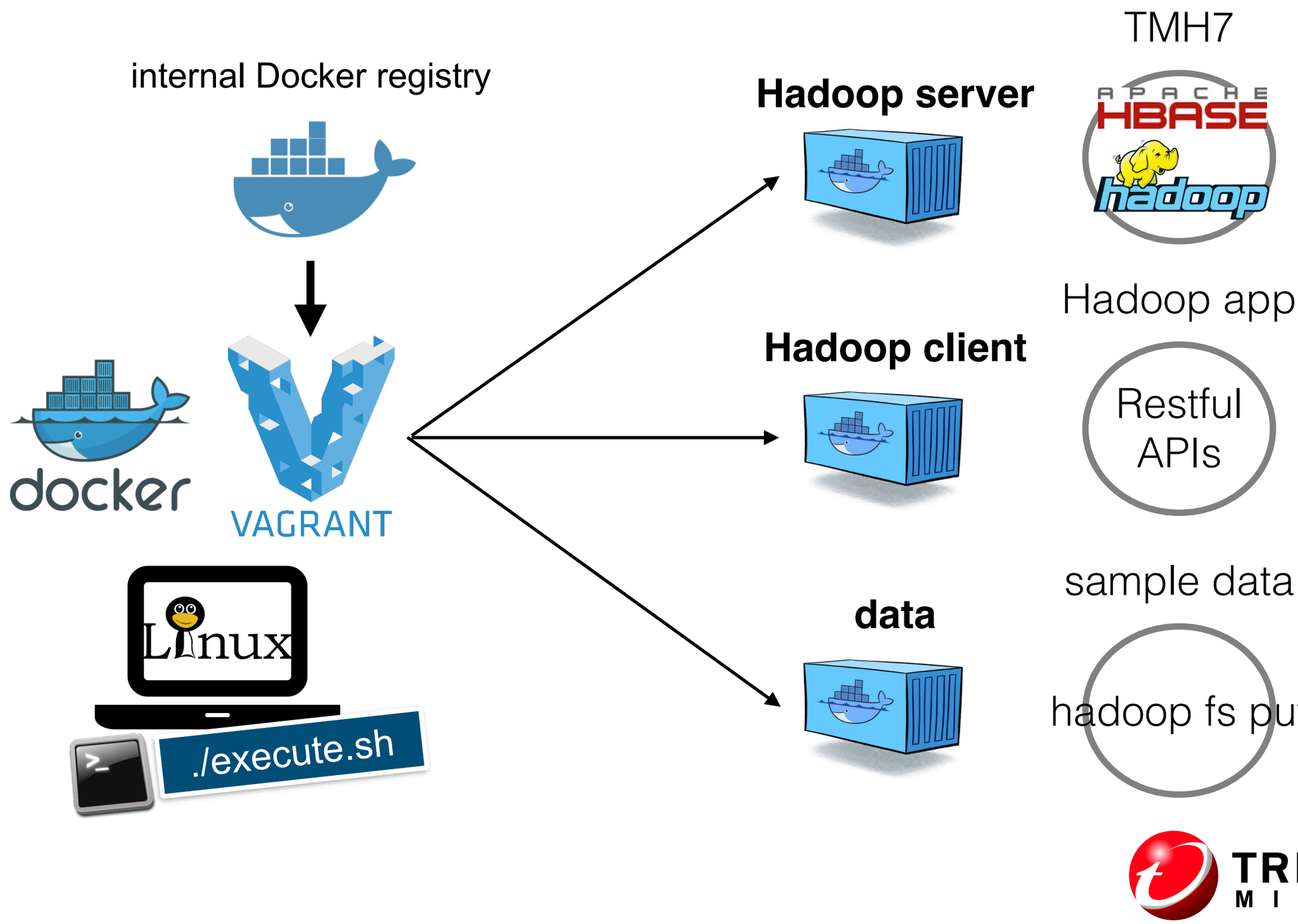
- A Hadoop env for apps to test against our new
  Hadoop distribution

- Same features are also delivered in Vagrant boxes

- https://github.com/evans-ye/hadoocker

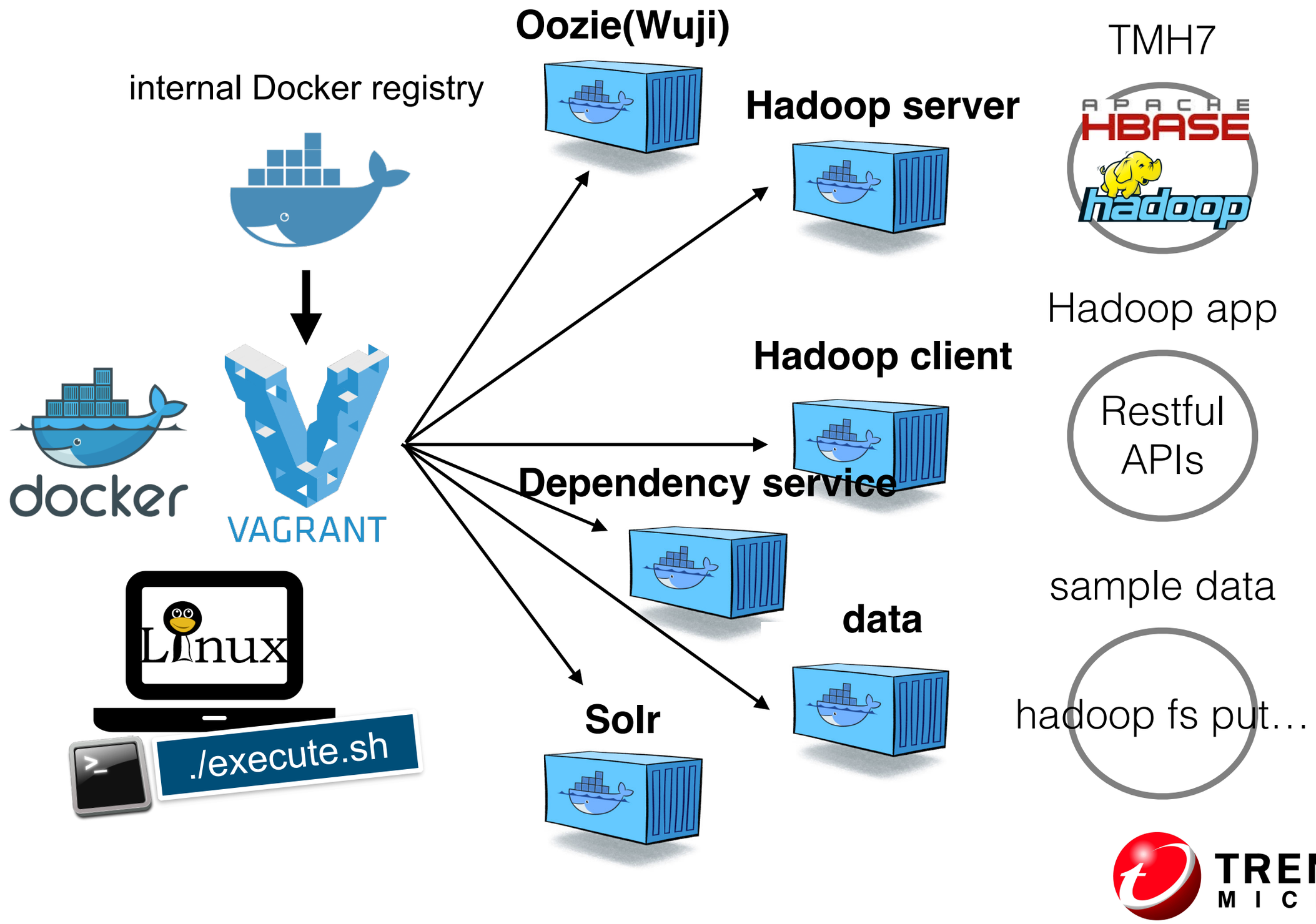
TREND MICRO™

Docker based dev & test env



TREND MICRO

Docker based dev & test env



Summary

TREND MICRO™

Summary

- Bigtop helps you to build your own Big Data Stack
- Building packages made easier by offering immutable build environment through Docker images
- Bigtop Provisioner creates you a Hadoop cluster by one click with flexibility
- Use Docker to simulate complex environment to ease development and testing efforts
- Ship apps in Docker images to dev & test anywhere

TREND MICRO™

Wait,

how’s the demo ?

TREND MICRO™

Thank you !

Questions ?