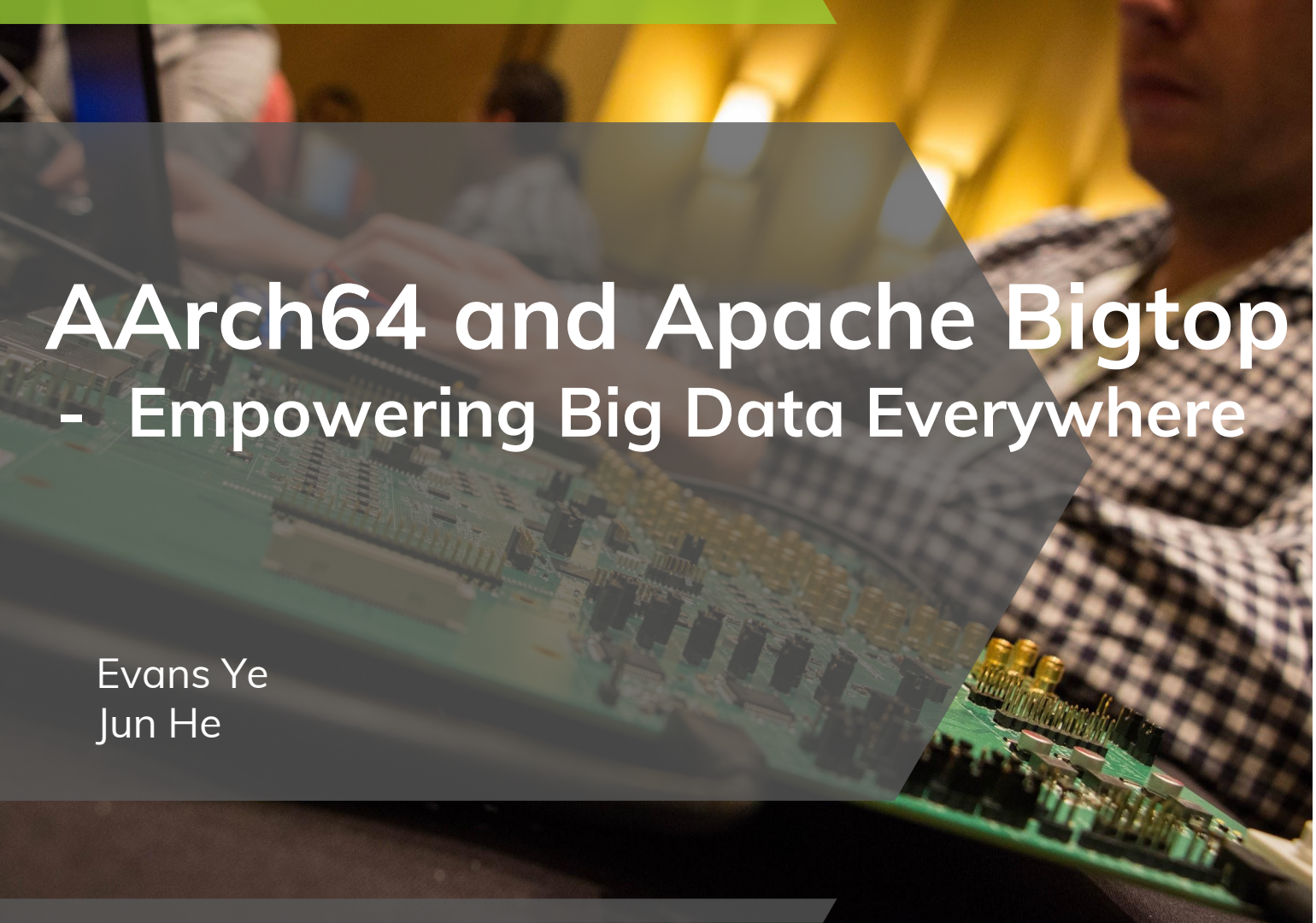


AArch64 and Apache Bigtop
- Empowering Big Data Everywhere

Evans Ye
Jun He

Linaro connect
Bangkok 2019

Evans Ye - Intro

Member of the Apache Software Foundation

- Spread the Apache Way
- Mentorship

Apache Bigtop PMC member, Committer, former VP

- About to introduce

Director of Taiwan Data Engineering Association (TDEA)

- Promote OSS, big data related technology
- Hold conference, workshop, meetup

What is Apache Bigtop?

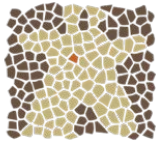
Package **Hadoop ecosystem** to RPM/DEB artifacts

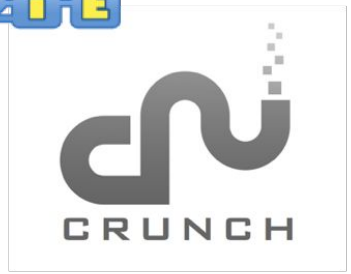
Purely open source Hadoop Distribution

Support 25 Hadoop Ecosystem Components









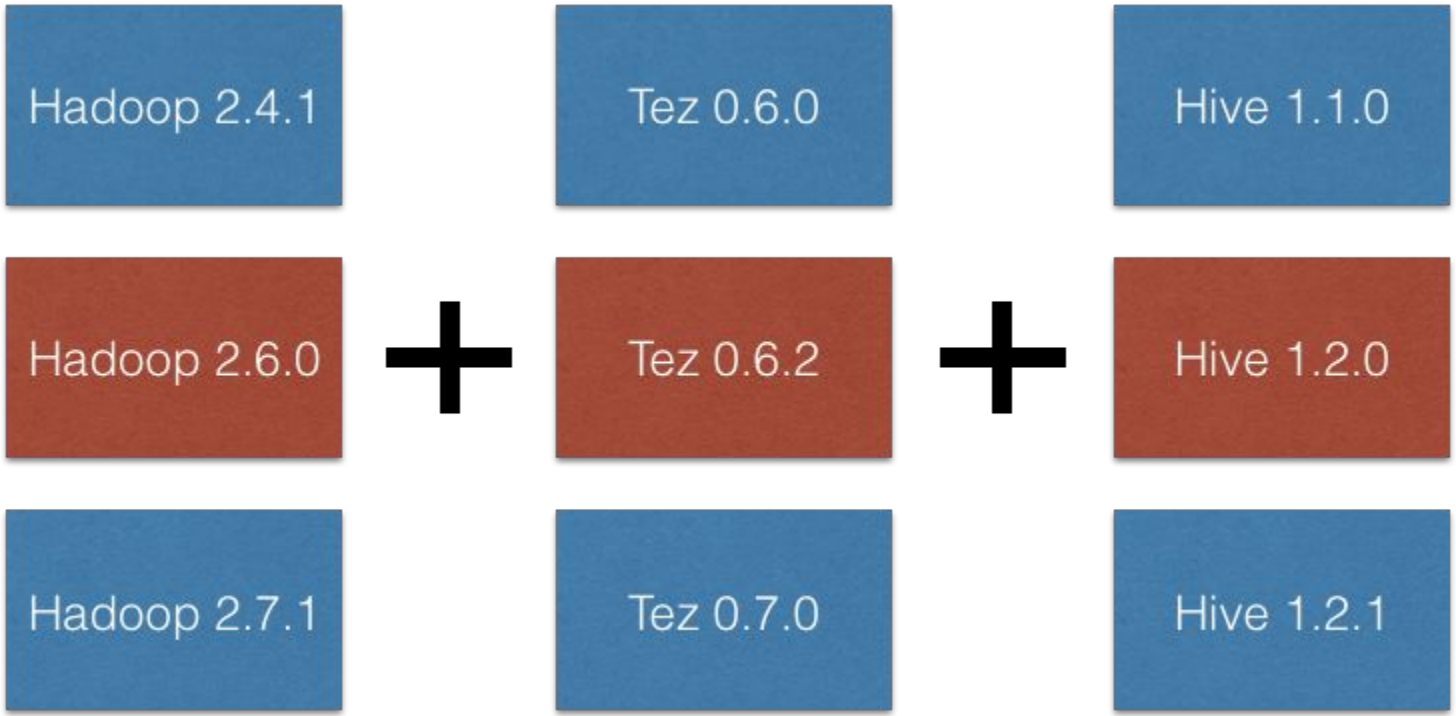
X

Foundation for commercial Hadoop Distros/services

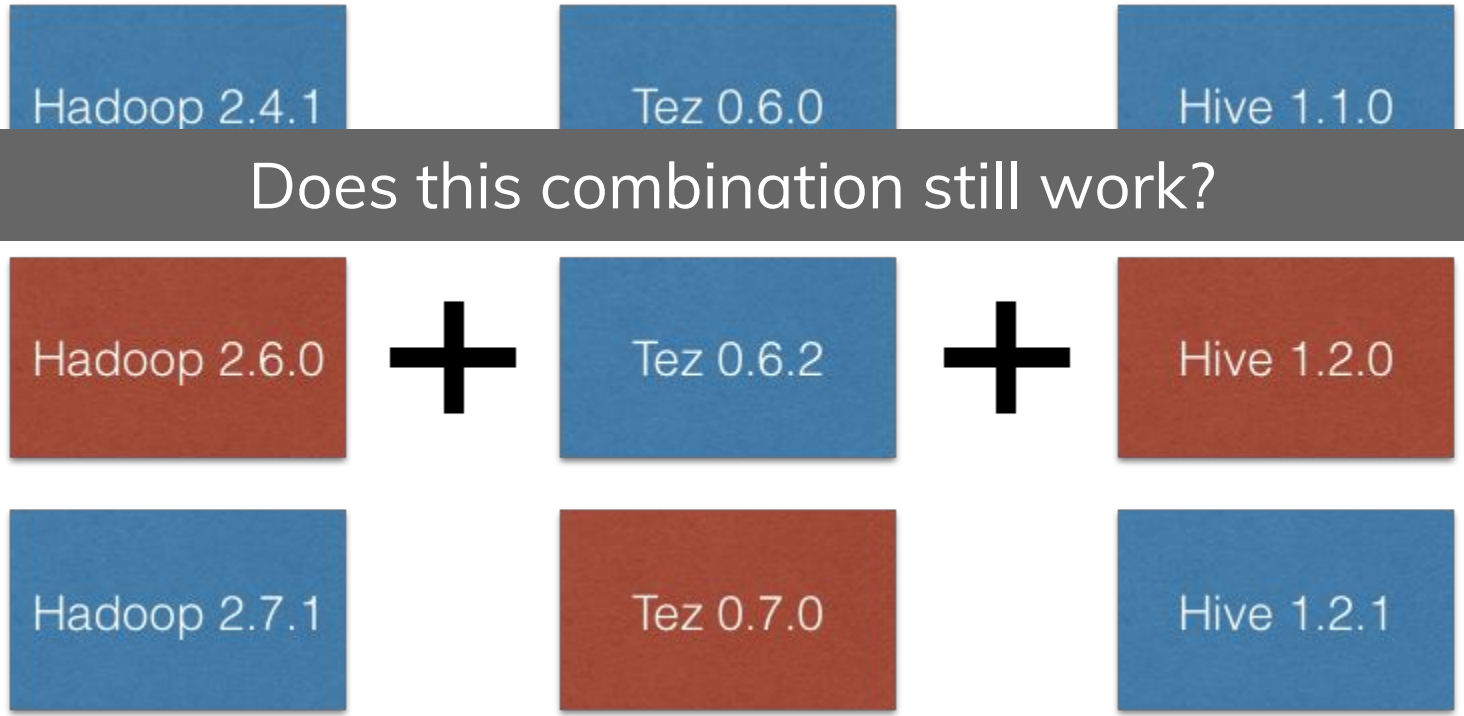
All major Hadoop distros leverage Bigtop to build its foundation

Leveraged by app providers…

Why Apache Bigtop ?



Why Apache Bigtop ?



Bigtop Feature Set

Packaging **Containerlization** **Deployment** **Testing**

Total solution to build your own Big Data Stack

Bigtop Toolchain

A set of Puppet recipes to install required **libraries**, **build tools**

To prepare a bigtop build environment:

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

Prerequisite:

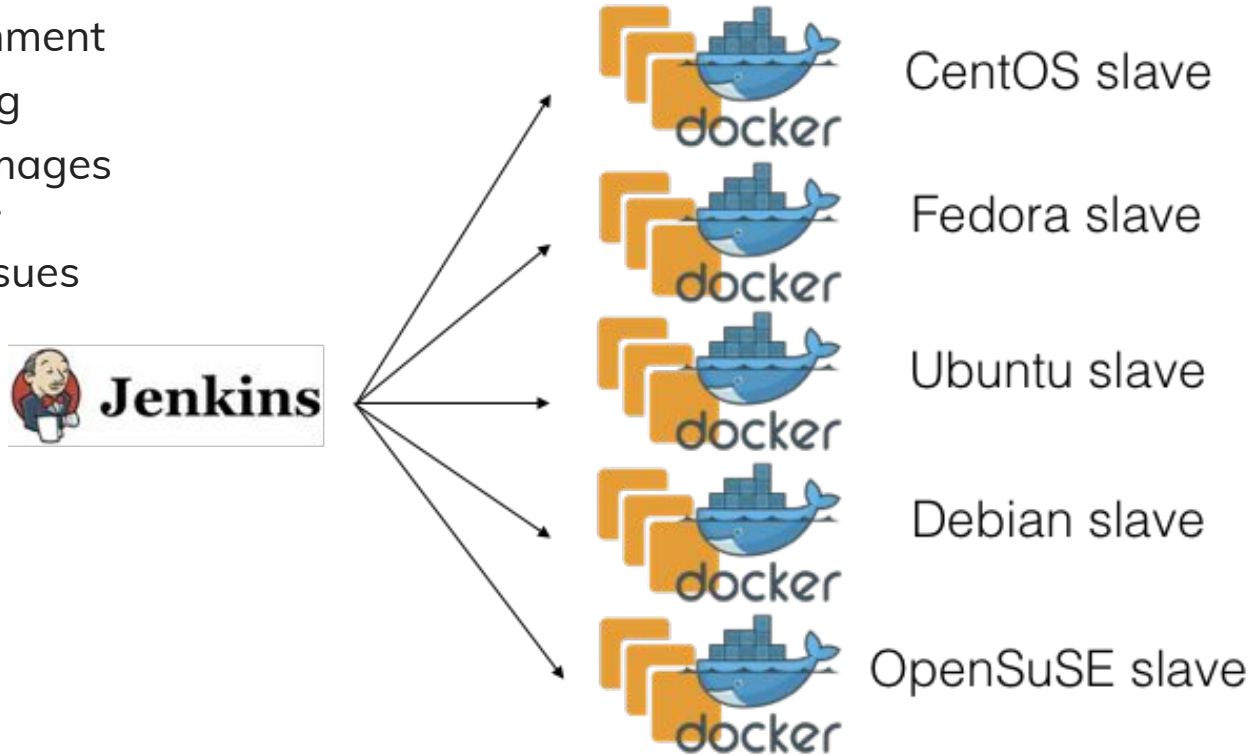
- Java

Containerized build infra

Immutable build environment

Super friendly for porting

- Prepare aarch64 images
- Try build on docker
- Fix compatibility issues



Bigtop Package

Framework to build Hadoop ecosystem components into RPM/DEB packages

Two ways:

- Release tarball -> build -> (patch) -> package
- Git branch/commit -> build -> (patch) -> package

How to:

- $ ./gradlew hadoop-pkg-ind

Why patch?

- Lots of compatibility issue
- Say Spark works well with hive and oozie, but got no luck with Zeppelin...
- We focus on the entire distribution

Bigtop Puppet & Test

Bigtop Puppet:

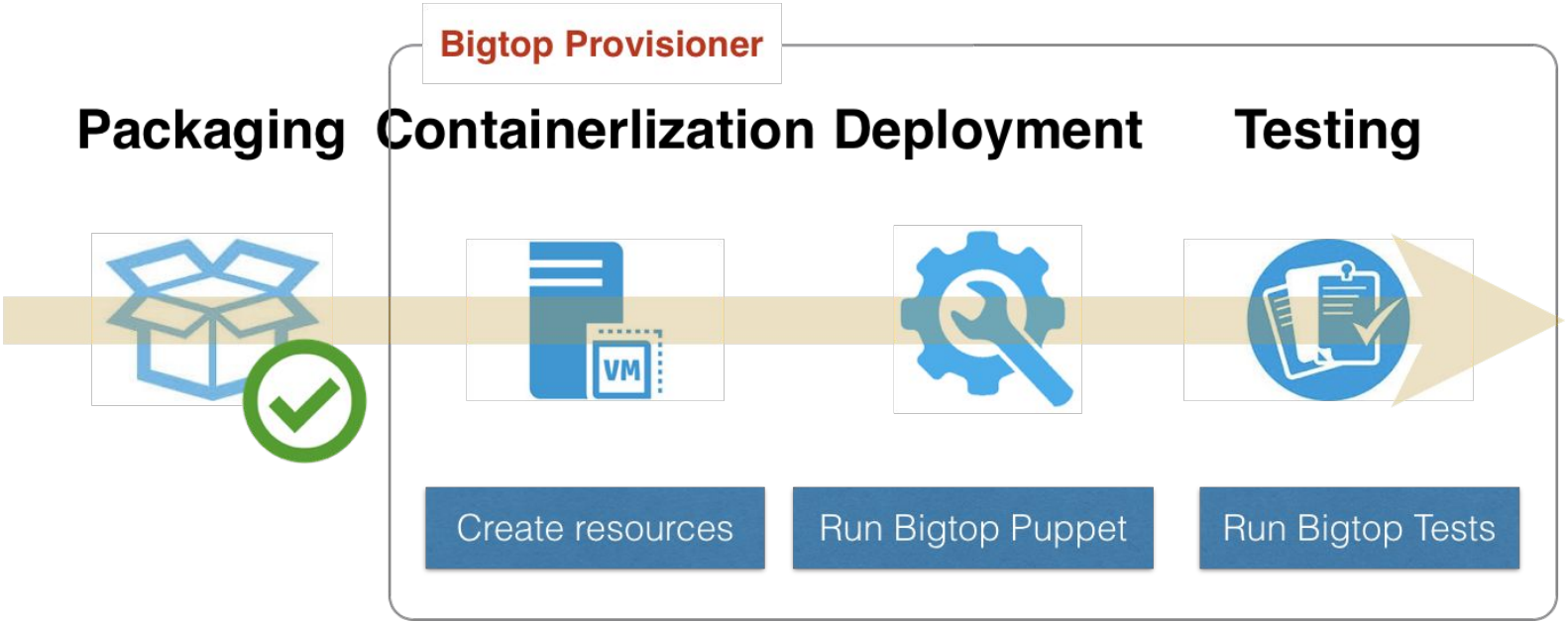
- A set of Puppet recipes to **deploy** Hadoop ecosystem components

Bigtop Test

- Bigtop Test Framework
  - Test utilities for writing tests in Java/Groovy
- Bigtop Smoke Test
  - Bunch of built-in smoke tests (quick diagnosis)
- Bigtop Integration Test
  - Bunch of built-in integration tests (coverage)
- Bigtop Package Test
  - Designed to find bugs in the packages before deployed

Bigtop Provisioner

Integrated provisioning solution to deploy and test Bigtop stack on Docker



Bigtop Sandbox

Bigtop stack built as image to be easily consumed

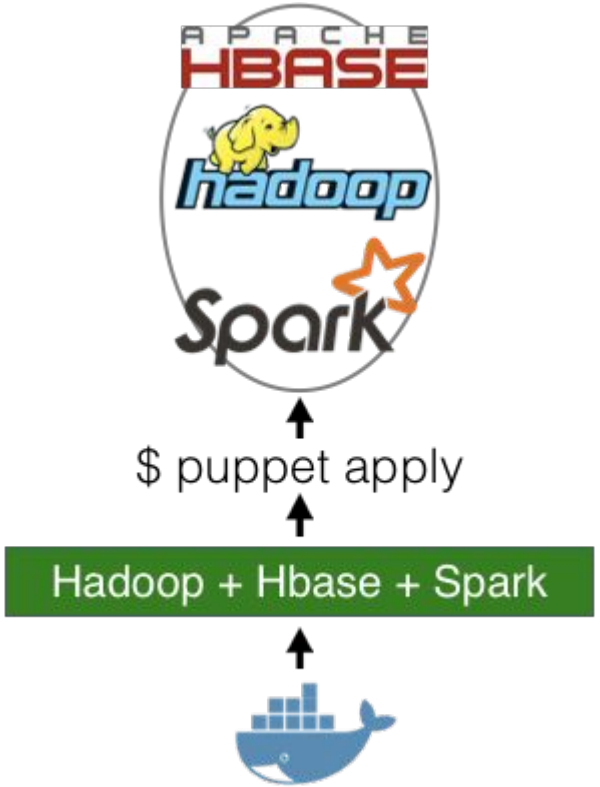
- How to:

```
docker run --name sandbox -d \
-p 50070:50070 -p 8088:8088 \
bigtop/sandbox:linaroconnect2019_spark

docker logs -f sandbox

docker exec sandbox spark-example SparkPi
```

- Quick start environment
- Handy image for applications to do integration test



Bigtop Integration Test Framework 2.0

Full support to build and test inside docker with one-stop seamlessly integrated entry at ./gradlew

- Package
  - $ ./gradlew spark-pkg-ind repo-ind
- Deploy & Test
  - $ ./gradlew docker-provisioner \
    -Penable_local_repo \
    -Pstack="hdfs,yarn,spark" \
    -Psmoke_tests=spark;
- Build -> Deploy -> Test lifecycle in one stop
  - $ ./gradlew spark-pkg-ind repo-ind docker-provisioner \
    -Penable_local_repo \
    -Pstack="hdfs,yarn,spark" \
    -Psmoke_tests="spark";

Bigtop Integration Test Framework 2.0

- Build directly from branch or commit hash:
  - $ ./gradlew allclean kafka-pkg-ind \
    -Pgit_repo=https://github.com/apache/kafka.git \
    -Pgit_ref=trunk \
    -Pgit_commit_hash=dc0601a1c604bea3f426ed25b6c20176ff444079 \
    -Pbase_version=2.2.0;

  - Advantages:
    - For developer to quickly evaluate the result
    - Code that brokes Integration can be discovered earlier in dev

Apache Bigtop: v1.4

Timeline: Upcoming Early April, 2019!

Features:

- Integration Test Framework 2.0
  - one-stop seamlessly integrated entry at **./gradlew** to build and test inside docker
- Smoke Test CI Matrix go online
  - https://ci.bigtop.apache.org/view/Test/job/Bigtop-trunk-smoke-tests
- Version bumps
  - Hadoop 2.8.5, Spark 2.2.3, Kafka 2.1.1, Flume 1.9.0, Alluxio 1.8.1
- More built-in test coverage
  - Hive, Flink, Giraph
- A Lot of improvements and bug fixes!
  - **100** JIRAs resolved

Jun He - Intro

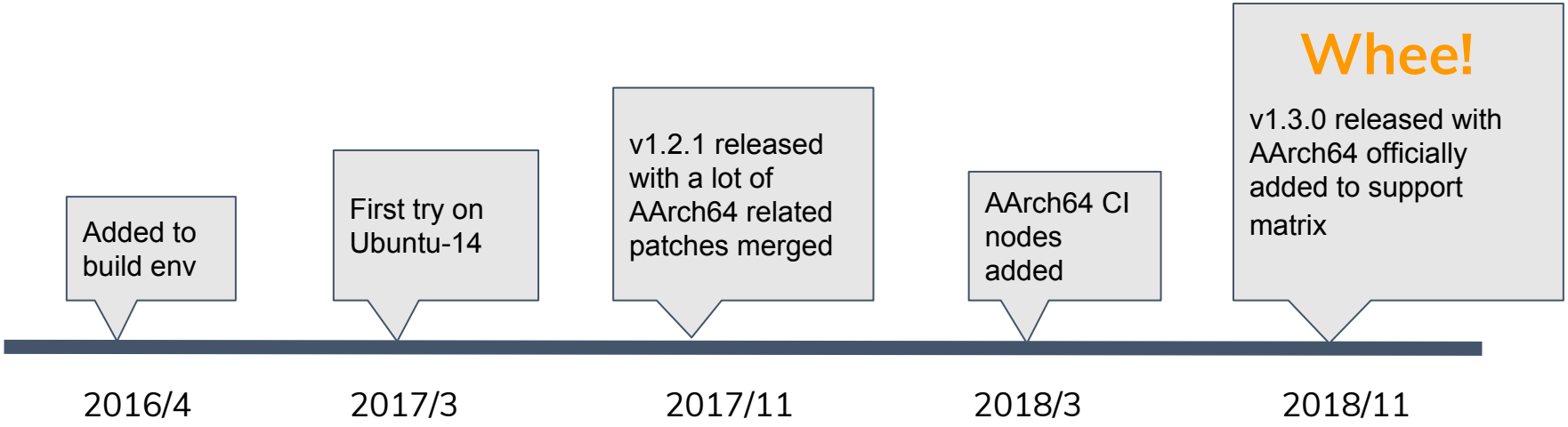
Apache Bigtop PMC member, Committer

- Now you get it ...

Lead of Enterprise Workloads Team in Arm OSS Group

- Enable and optimize Data Science/Storage stacks on Arm64
- Contribute to build a diverse software ecosystem

Apache Bigtop on AArch64



What we learned so far

- Dependency issues
  - Native binaries: protobuf, phantomjs, ...
  - Jars with native binaries embedded: levedb-jni, ignite-shmem, jffi, snappy-java ...
  - Version mismatch: slf4j, log4j, log4j2, ...
- Cyclic references take a lot of effort to fix
- Tests are important

Where is Big Data heading ?

There will be more and more big data tools and integrations on the cloud

- Lots of money goes into cloud vendor's pocket

K8S is taking up the whole industry, including big data

- HDFS on K8S, Spark on K8S, Flink on K8S, etc
- One single platform for OLTP, OLAP, ML/AI

More focus on user experience (can do -> perform well -> **easy to use**)

- NewSQL
- More user friendly APIs

Apache Bigtop: Future Roadmap

Focus on components that maximize the core value of big data

- Processing: Spark, Flink, Hive
- Storage: Hadoop, Kafka
- NoSQL: HBase, Cassandra

Cloud / K8S native support (operators) for build, deploy, and test

Embrace cloud(AWS/GCP/Azure) and introduce more integrations

Demo

Questions ?

Dev Mailing lists

Wiki page

CI page

Jira link

Linaro Collaborate page

Contact details :

Evans Ye: evansye@apache.org

Jun He: jun.he@arm.com