HKG18-220: State of Big Data on AArch64 - Apache BigTop

Ganesh Raju, Jun He & Naresh Bhat
Big Data Team, LEG
Agenda

- Why and What is Bigtop
- Bigtop patches for AArch64
- Challenges of porting
- Walk through of setup and installation process
- Demo on Provisioning and Smoke Tests
Why Apache Bigtop?

- Hadoop is a collection of many components
  - Numerous versions (Dependency hell)
  - Lots of patches
  - No stable development environment with certified binaries
  - No proper integrated tests
- With Bigtop - Build, Deploy in cluster with puppet, Configure, Install and Test
- Juju orchestration
- Blueprints
- Seamless integration into CI
What is in Apache Bigtop?

- **Output**
  - A set of binaries (deb and rpm) just like HDP, ODPi, CDH, etc
  - Docker images
  - Docker sandbox images
- **Integration code, Packaging code, Deployment code, Orchestration code**
- **Validation code**
  - Integration tests
    - Clean slate provisioning
    - Dependency integration artifacts
    - Versioned test artifacts
    - Plug and play artifacts
    - JVM-based artifacts
  - Packaging tests
  - Smoke tests
- **Continuous Integration**
Consumers of Bigtop

Some of consumers of Bigtop

- ODPi
- Hortonworks
- Amazon
- Canonical
- EMC
- Pivotal
- Infosys
- Capgemini
- Ebay
- Intel
- TrendMicro
- WANdisco
Bigtop v1.2

A few highlights of the v1.2 series release include:

● 6 Distros, 2 archs (x86, and ppc64le) supported
  ○ ARM support with v1.3
● A newly introduced Bigtop Sandbox feature
● A faster Docker Provisioner which is rewritten to fully embrace Docker ecosystem
● OpenJDK 8 support
● Hadoop 2.7.3, Spark 2.1.1, HBase 1.1.9, and Zeppelin 0.72 are used
● And many upgrades of the ecosystem projects
  (Apex, Crunch, Flume, Ignite, Mahout, Oozie, Phoenix, and many others)
## Components of Bigtop

<table>
<thead>
<tr>
<th>Component</th>
<th>Version</th>
<th>Version</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>alluxio</td>
<td>v1.0.1</td>
<td>Greenplum gpdb v5.0.0-alpha.0</td>
<td>Apache pig v0.15.0</td>
</tr>
<tr>
<td>Apache ambari</td>
<td>v2.5.0</td>
<td>Apache hadoop v2.7.3</td>
<td>Quantcast qfs v1.1.4</td>
</tr>
<tr>
<td>Apache apex</td>
<td>v3.5.0</td>
<td>Apache hama v0.7.0</td>
<td>Apache solr v4.10.4</td>
</tr>
<tr>
<td>groovy</td>
<td>v2.4.10</td>
<td>Apache hbase v1.1.3</td>
<td>Apache spark 1.1 v1.6.2</td>
</tr>
<tr>
<td>Apache commons - jsvc</td>
<td>v1.0.15</td>
<td>Apache hive v1.2.1</td>
<td>Apache spark 2.0 v2.1.1</td>
</tr>
<tr>
<td>Apache tomcat</td>
<td>v6.0.45</td>
<td>Apache hue v3.11.0</td>
<td>Apache sqoop v1 v1.4.6</td>
</tr>
<tr>
<td>bigtop_utils</td>
<td>v1.2.0</td>
<td>Apache ignite v1.9.0</td>
<td>Apache sqoop v2 v1.99.4</td>
</tr>
<tr>
<td>Apache crunch</td>
<td>v0.14.0</td>
<td>Apache kafka v0.10.1.1</td>
<td>Apache tajo v0.11.1</td>
</tr>
<tr>
<td>Pig UDF datafu</td>
<td>v1.3.0</td>
<td>kite v1.1.0</td>
<td>Apache tez v0.6.2</td>
</tr>
<tr>
<td>Apache flink</td>
<td>v1.1.3</td>
<td>Apache mahout v0.12.2</td>
<td>ycsb v0.4.0</td>
</tr>
<tr>
<td>Apache flume</td>
<td>v1.7.0</td>
<td>Apache oozie v4.3.0</td>
<td>Apache zeppelin v0.7.0</td>
</tr>
<tr>
<td>Apache giraph</td>
<td>v1.1.0</td>
<td>Apache phoenix v4.9.0-HBase-1.1</td>
<td>Apache zookeeper v3.4.6</td>
</tr>
</tbody>
</table>
Contributions from ARM Ecosystem

- AArch64 CI nodes are running on Linaro DevCloud
  - 3 distros are supported: Debian-9, Fedora-26, Ubuntu-16.04
    https://ci.bigtop.apache.org/job/Bigtop-trunk-packages/

<table>
<thead>
<tr>
<th>Configuration Matrix</th>
<th>centos-7</th>
<th>debian-9</th>
<th>debian-9-aarch64</th>
<th>fedora-26</th>
<th>fedora-26-aarch64</th>
<th>fedora-26-ppc64le</th>
<th>opensuse-42.3</th>
<th>ubuntu-16.04</th>
<th>ubuntu-16.04-aarch64</th>
<th>ubuntu-16.04-ppc64le</th>
</tr>
</thead>
<tbody>
<tr>
<td>alluxio</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>ambari</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>apex</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>bigtop-groovy</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>bigtop-jsvc</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>bigtop-tomcat</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>bigtop-utils</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>crunch</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Contributions from ARM Ecosystem

● Patches to enable components on AArch64
  ○ Build
    ■ Hadoop, Solr, Hbase, Ignite, …
  ○ Package
    ■ Hama, Solr, Oozie, Hue, …
  ○ Deploy
    ■ Service scripts, Automation scripts, Dockerfiles,…
  ○ Test
    ■ SmokeTests and Provisioner settings
Contributions from ARM Ecosystem

● Lessons learned
  ○ Dependency issues
    ■ Native binaries: protobuf, phantomjs, …
    ■ Jars: levedb-jni, ignite-shmem, jffi, …
    ■ Version mismatch: slf4j, log4j, log4j2, …
  ○ Repos
    ■ Official release did not support aarch64
      ● Had to create private/local repo
Challenges

- Cyclic references take a lot of effort to fix
- Though most big data companies all use Bigtop, there has not been contributions coming in from them
- With founders moving out end of last year, and lead committers changing focus, the project has lost momentum
Roadmap

- Make Bigtop a 1st class citizen on Kubernetes
- Test out Puppet deployment code in variety of different scenarios including developers spinning up test clusters via Docker deployer
- Improve Docker deployer to be more developer friendly and hook it back into Gradle
- Provide predefined sample stacks for specific use cases. For example:
  - Machine Learning: Hadoop+Spark+Zeppelin
  - Streaming: Hadoop+Kafka+Flink
- Create more tests
- Enable Ambari to install Bigtop stack. Utilize the work done for ODPi
Sandbox

- What is Sandbox?
  A tool to build and run big data pseudo cluster using Docker

- Command to generate sandbox image
  
  $ ./build.sh -a bigtop -o centos-7 -c "hdfs, yarn, spark, ignite"

- You can do a dry run using option “--dryrun” command

- How to run the sandbox image
  
  $ docker run -d -p 50070:50070 bigtop/sandbox:centos-7_hdfs
Smoke Tests

- Uses yaml file to configure located under <BIGTOP_SRC_TOP>/provisioner/docker/
- Components to configure
  - docker image
  - distro type
  - components to install
  - components to test
  - JDK
- Environment check
  $ ./docker-hadoop.sh -E
- Execution
  $ cd provisioner/docker
  $ ./docker-hadoop.sh -C <smoke_test_cfg_yaml> -c <node_count> -s -d
Glossary

- Linaro Collaborate page
- Bigtop wiki page
- Smoke Test Collaborate page
- Smoke Test Results
Thank You

#HKG18
HKG18 keynotes and videos on: connect.linaro.org
For further information: www.linaro.org