



Big Data & Data Science Project Update

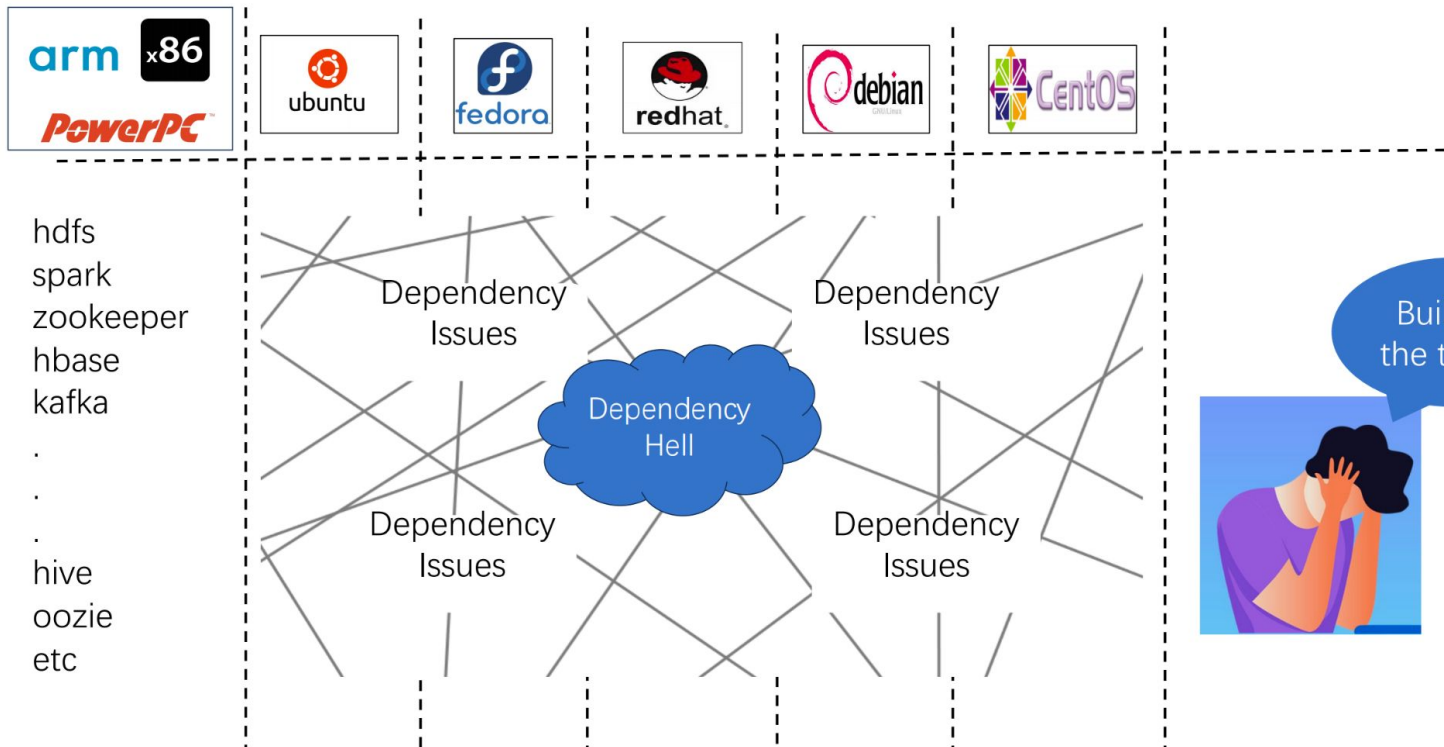
Zhiguo Wu
Kevin Zhao

Agenda

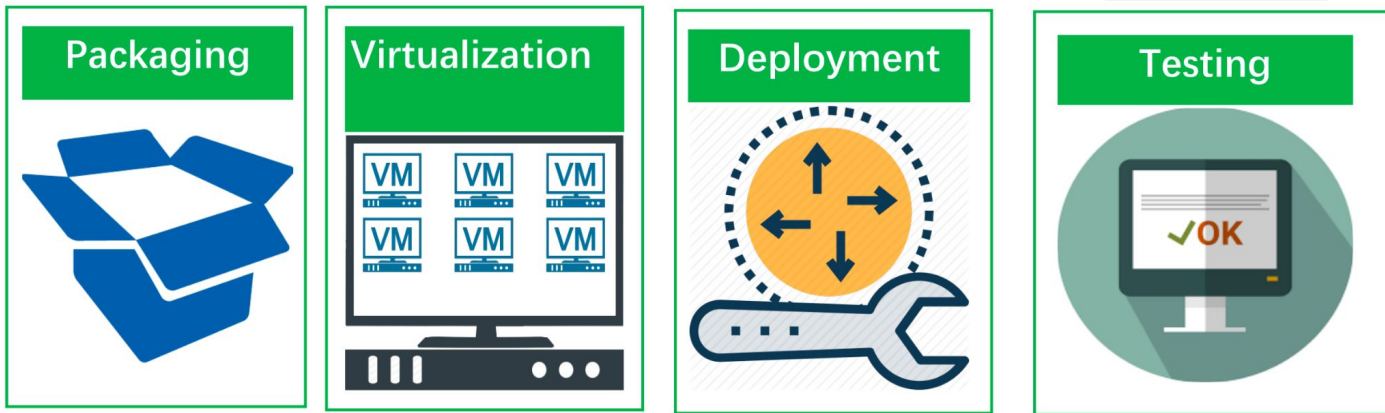
- Bigtop & Bigtop Manager Update
- Gluten & Velox Update

Bigtop & Bigtop Manager Update

Challenges in ecosystem



What is Bigtop - Features Snapshot



Total solution to provision Big Data Stack



Multi Arch

arm

x86

PowerPC™

Bigtop 3.4.0 - In Progress

- Distro upgrade
 - Rocky Linux 9
 - Fedora 40
 - openEuler 22.03
 - Debian 12
 - Ubuntu 24.04

➤ Services upgrade

Components	in v3.3	in v3.4
alluxio	2.9.3	=> 2.9.6
bigtop-groovy	2.5.4	=> 2.5.4
bigtop-jsvc	1.2.4	=> 1.4.0
bigtop-utils	3.3.0	=> 3.4.0
bigtop-select	3.3.0	=> 3.4.0
flink	1.16.2	=> 1.20.0
hadoop	3.3.6	=> 3.3.x or 3.4.x
hbase	2.4.17	=> 2.6.1
hive	3.1.3	=> 4.0.1
kafka	2.8.2	=> 3.4.1 (for the compatibility with Spark 3.5.3)
livy	0.8.0	=> 0.8.0
phoenix	5.1.3	=> 5.2.1
ranger	2.4.0	=> 2.5.0
solr	8.11.2	=> 8.11.4
spark	3.3.4	=> 3.5.3 (4.x is not officially released as of writing)
tez	0.10.2	=> 0.10.4 (required for Hive 4)
zeppelin	0.11.0	=> 0.11.2
zookeeper	3.7.2	=> 3.8.4

Kengo Seki - Friday, February 21, 2025 3:51:21 PM GMT+8

Hi everyone,

As I've submitted BIGTOP-4360 [1], I'd like to propose incorporating Apache Airflow [2] into our stack, because we don't have any job scheduler/workflow orchestrator since we've dropped Oozie from 3.3.0.

Rationales in my mind are as follows:

- * As Astronomer and major cloud vendors provide the managed service of Airflow [3][4][5][6], it is the de-facto standard of open-source workflow orchestrator.
- * Airflow community is very active [7], so we don't need to be worried about its stagnation for several years at least.
- * I'm also an Airflow committer, so I can be responsible for maintaining it as a Bigtop component.

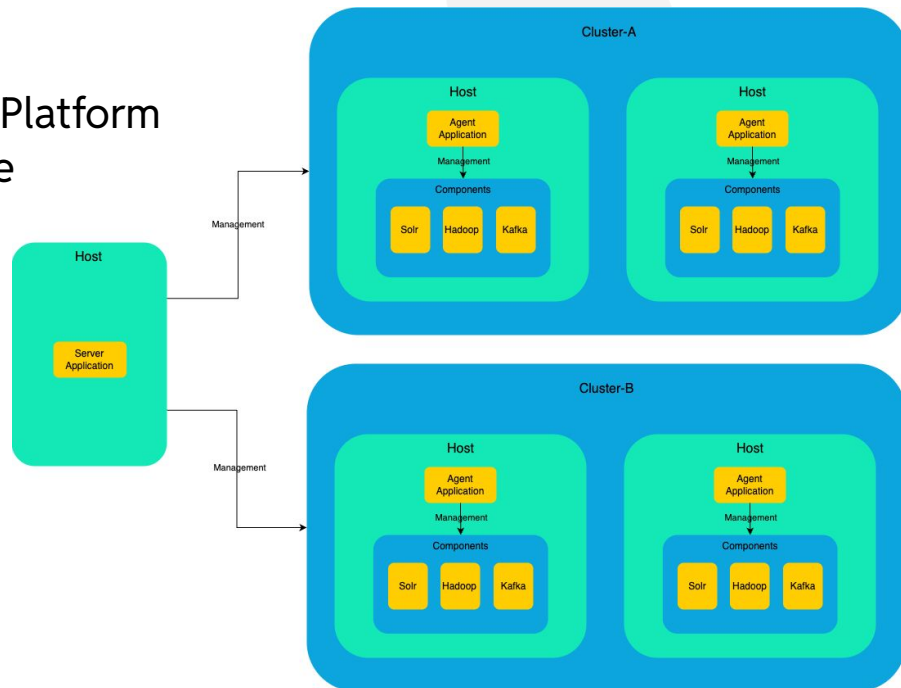
Any thoughts or comments?

[1]: <https://issues.apache.org/jira/browse/BIGTOP-4360>
 [2]: <https://airflow.apache.org/>
 [3]: <https://www.astronomer.io/>
 [4]: <https://aws.amazon.com/managed-workflows-for-apache-airflow/>
 [5]: <https://learn.microsoft.com/en-us/fabric/data-factory/create-apache-airflow-jobs>
 [6]: <https://cloud.google.com/composer>
 [7]: <https://github.com/apache/airflow/graphs/contributors>

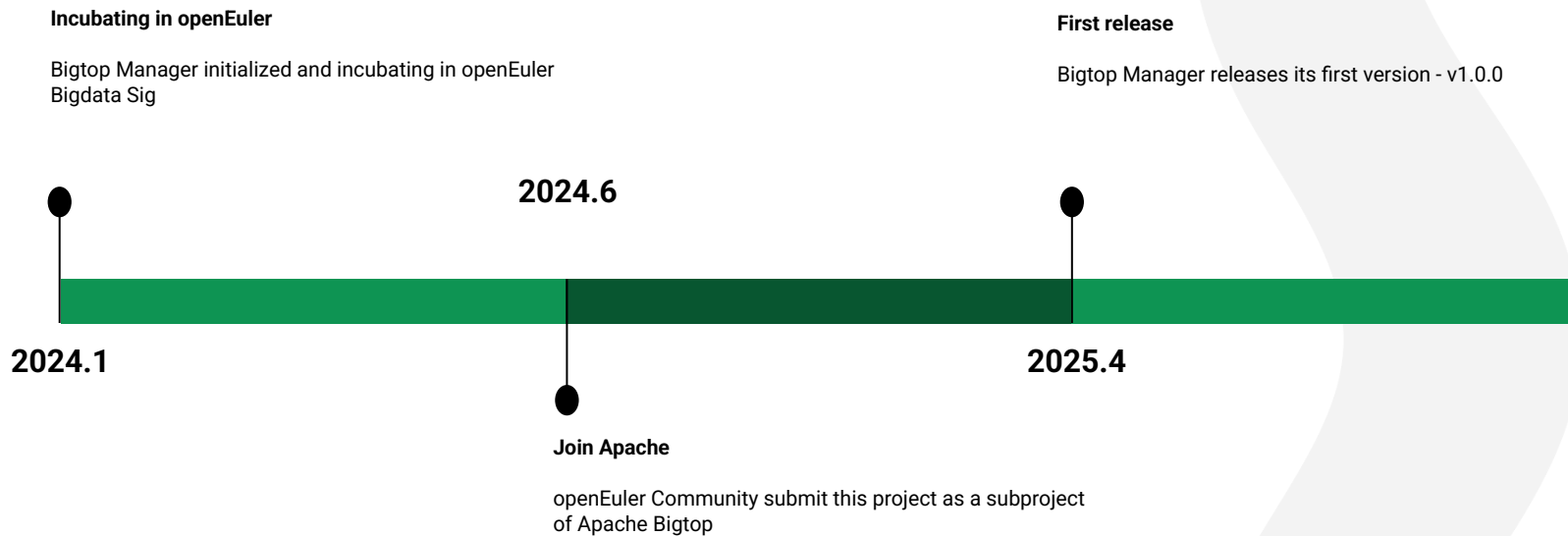
Kengo Seki <sekikn@apache.org>

What is Bigtop Manager


- ❑ Easy deployment solution for Bigtop
- ❑ Modern Bigdata Cluster Management Platform
- ❑ AI-Driven Operations and Maintenance
- ❑ Flexible Multi-Cluster Management



Timeline







Preview

BIGTOP MANAGER
STREAMLINE DATA INFRASTRUCTURE

ClusterSystem

Administrator



Cluster



hadoop

Infrastructure

Stacks

Host




Create Cluster




hadoop 
hadoop




Add ServiceMore Operations





OverviewServiceHostUserJob





Service NameRestartStatus





ZooKeeper
3.7.2-1 
Restart 

Solr
8.11.2-2 
Restart 

Kafka
2.8.2-1 
Restart 







Copyright ©2011-2025 The Apache Software Foundation . All rights reserved.

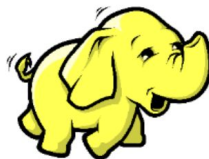
Accomplishments

- Released Apache Bigtop Manager 1.0.0
- Refactor API Interfaces for backend deployment/management function
- Use tarballs to replace deb/rpm based packaged to make the cluster management distro-independent.
- Multi-stack support.
 - Bigtop stack: Spark, HBase, Hive
 - Infra stack: Prometheus, Grafana, Mysql
 - Extra stack: SeaTunnel

Stack and Service

❑ Bigtop Stack(3.3.0)

- ❑ Hadoop
- ❑ HBase
- ❑ Hive
- ❑ ZooKeeper
- ❑ Kafka
- ❑ Spark
- ❑ Flink
- ❑ ...



❑ Infra Stack(1.0.0)

- ❑ Grafana
- ❑ Prometheus
- ❑ MySQL



❑ Extra Stack(1.0.0)

- ❑ SeaTunnel



.....

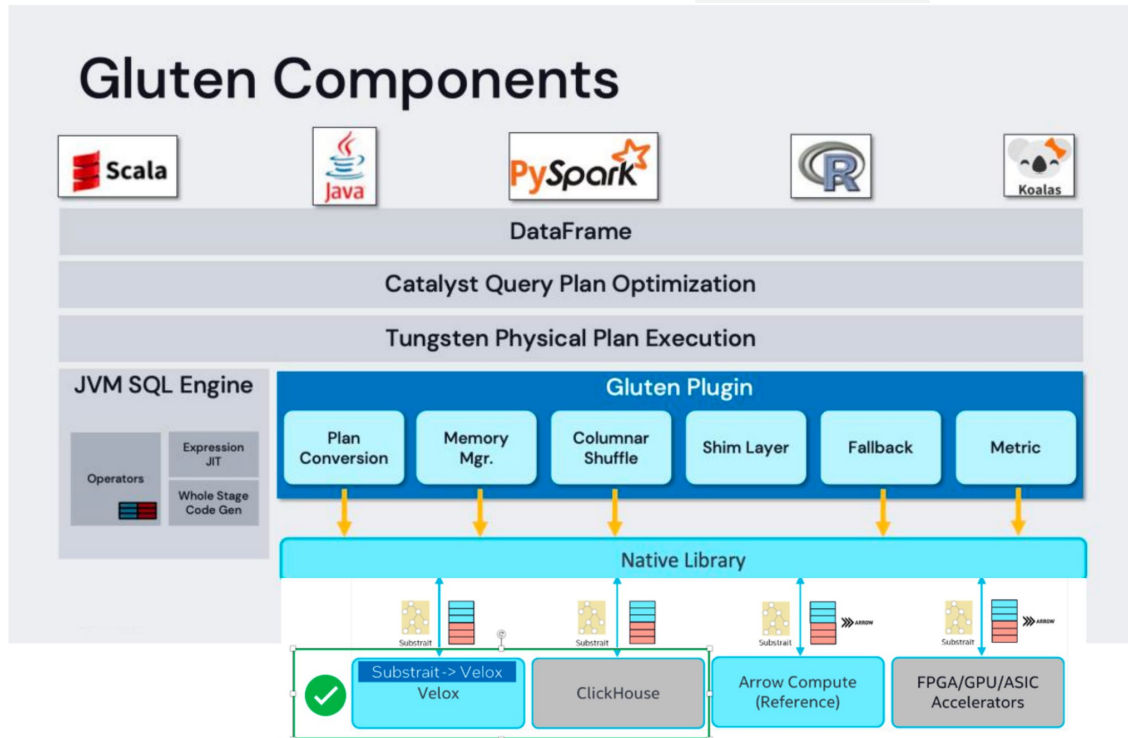
Future plans

- Bigtop
 - Bigtop 3.4.0 Release
 - Upgrade openEuler to 24.03
- Bigtop Manager
 - Next Release - V1.1.0
 - Metrics/Alerts system
 - MCP Server(A2A?)
 - User experience optimization
 - Service stability enhancement
 - ZooKeeper/Hadoop/Hive and etc.
 - New services
 - StarRocks/Doris and etc.
 - Kerberos(Best effort)
 - openEuler
 - Include to openEuler yum repository

Gluten & Velox Update

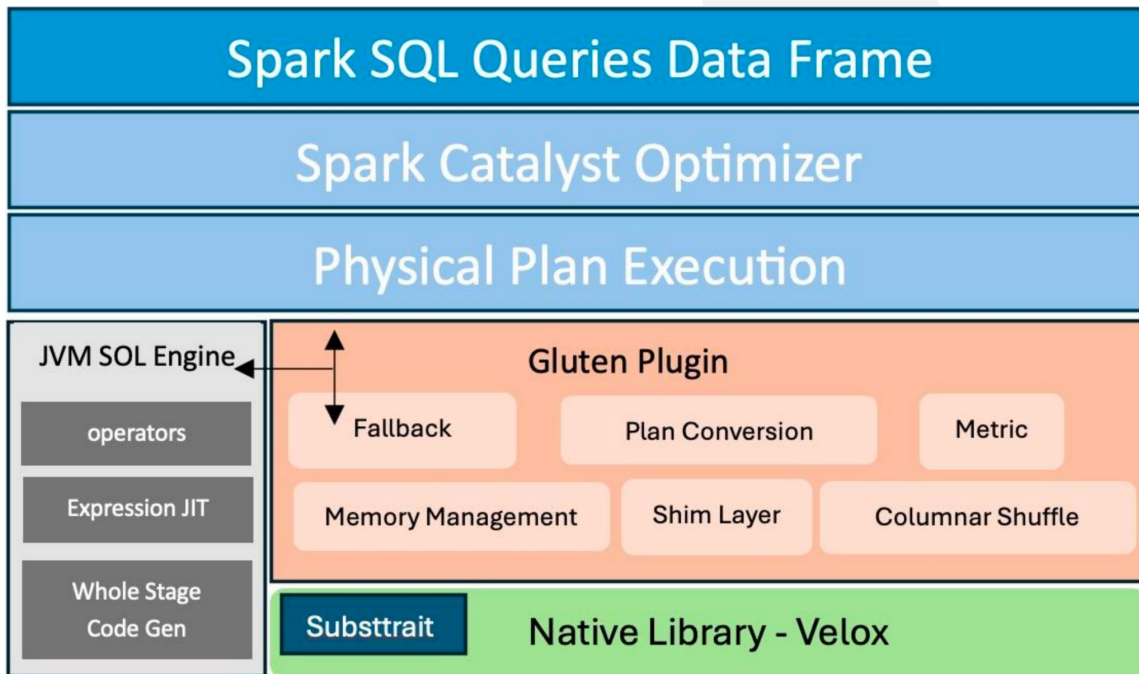
What is Gluten

- A Middle Layer to Offload Spark SQL Queries to Native Engines
- Offload the compute-intensive data processing part to native code
- “Glue” native libraries with Spark SQL
- Enabling the use of hardware accelerators



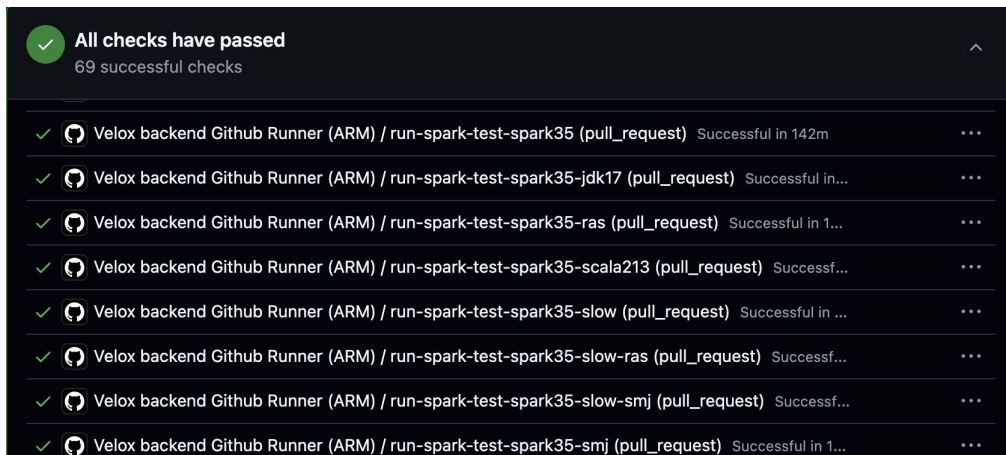
What is Velox

- Open-Source unified execution Native engine
- Serve as a native backend for Gluten
- Data-intensive operations: expression evaluation, aggregation, sorting, joining ...



Gluten Status - CI

- Velox
 - [fix\(parquet\): Group index should not reset during aggregation pushdown](#)
- Gluten
 - [\[GLUTEN-8802\]\[VL\] Support build static/dynamic docker images for arm](#)
 - [\[GLUTEN-8802\]\[VL\] Add specific jdk version for CentOS 8 docker image](#)
 - [\[GLUTEN-8894\]\[VL\] Fix buffer overflow in jStringToCString on arm](#)



Gluten Status - openEuler

- Velox
 - [misc: Throw invalid time zone instead of runtime error when timezone not found](#)
- Gluten
 - [\[GLUTEN-8709\]\[VL\] Support build on openEuler 24.03 LTS with Velox backend](#)

Future plans

- Gluten upstream CI on Arm64
 - Distros: CentOS8/openEuler24.03
 - Spark versions: 3.4/3.5
- Gluten and Velox support on openEuler
 - Distros: openEuler24.03



Thank You!