Airflow & Bigtop: Modernize and integrate time-proven OSS stack with Apache Airflow

Kengo Seki, Masatake Iwasaki
NTT DATA

AIRFLOW SUMMIT

3.0

Who we are

- Open source software developers / data engineers @NTT DATA

- Kengo Seki
  - Committer & PMC member of Apache Airflow and Apache Bigtop
  - ASF member

- Masatake Iwasaki
  - Committer & PMC member of Apache Hadoop and Apache Bigtop
  - ASF member

Agenda

- What is Apache Bigtop

- Introducing Bigtop features using the Airflow component

- Why we chose Airflow as our workflow orchestrator

- BigPetStore example DAG

What is Apache Bigtop

What is Apache Bigtop

AIRFLOW SUMMIT

- "Bigtop is an ASF project for Infrastructure Engineers and Data Scientists looking for comprehensive packaging, testing, and configuration of the leading open source big data components." (from https://bigtop.apache.org/)

- Bigtop is a Hadoop distribution developed by an open-source community and makes it easier for users to build and test their data platform.

Bigtop Features

1. Building components and packaging on widely-used Linux distributions

2. Deployment automation

3. Smoke tests

4. Provisioning components onto Docker containers

Each detail will be explained later.

History of Bigtop

	1.0.0	1.1.0	1.2.1	1.3.0	1.4.0	1.5.0
Ambari			2.5.0	2.6.1	2.6.1	2.6.1
Flink			1.1.3	1.4.2	1.6.4	1.6.4
Hadoop	2.6.0	2.7.1	2.7.3	2.8.4	2.8.5	2.10.1
HBase	0.98.12	0.98.12	1.1.9	1.3.2	1.3.2	1.5.0
Hive	1.0.0	1.2.1	1.2.1	2.3.3	2.3.3	2.3.6
Kafka	0.8.1.1	0.8.1.1	0.10.1.1	0.10.2.2	0.10.2.2	2.4.0
Oozie	4.0.1	4.2.0	4.3.0	4.3.0	4.3.0	4.3.0
Spark	1.3.1	1.5.1	2.1.1	2.2.1	2.2.3	2.4.5
Zeppelin		0.5.6	0.7.2	0.7.3	0.7.3	0.8.2
Zookeeper	3.4.6	3.4.6	3.4.6	3.4.6	3.4.6	3.4.13

Timeline: 2011 — 2012 — 2015 — 2016 — 2017 — 2018 — 2019 — 2020

AIRFLOW SUMMIT

History of Bigtop

AIRFLOW SUMMIT

	3.0.0	3.0.1	3.1.0	3.1.1	3.2.0	3.2.1
Ambari	2.7.5	2.7.5	2.7.5	2.7.5	2.7.5	2.8.0
Flink	1.11.3	1.11.6	1.11.6	1.11.6	1.15.0	1.15.3
Hadoop	3.2.2	3.2.2	3.2.3	3.2.4	3.3.4	3.3.5
HBase	2.2.6	2.2.6	2.4.8	2.4.8	2.4.13	2.4.13
Hive	3.1.2	3.1.2	3.1.2	3.1.2	3.1.3	3.1.3
Kafka	2.4.1	2.4.1	2.8.1	2.8.1	2.8.1	2.8.1
Oozie	5.2.1	5.2.1	5.2.1	5.2.1	5.2.1	5.2.1
Spark	3.0.1	3.0.1	3.1.2	3.1.2	3.2.1	3.2.1
Zeppelin	0.9.0	0.10.0	0.10.0	0.10.0	0.10.1	0.10.1
Zookeeper	3.4.14	3.4.14	3.5.9	3.5.9	3.5.9	3.5.9

2021   2022   2023   2024

History of Bigtop

AIRFLOW SUMMIT

	2024 3.3.0	2025 3.4.0	3.5.0
Airflow			2.10.5
Ambari			
Flink	1.16.2	1.20.0	1.20.1
Hadoop	3.3.6	3.3.6	3.3.6
HBase	2.4.17	2.6.1	2.6.2
Hive	3.1.3	4.0.1	4.0.1
Kafka	2.8.2	3.4.1	3.4.1
Oozie			
Ranger	2.4.0	2.6.0	2.7.0
Spark	3.3.4	3.5.3	3.5.6
Zeppelin	0.11.0	0.11.2	0.11.2
Zookeeper	3.7.2	3.8.4	3.8.4

Introducing Bigtop features using the Airflow component

Bigtop Features

1. Build components as native packages of widely-used Linux distributions

2. Deployment automation

3. Smoke tests

4. Provisioning components onto Docker containers

Building packages locally

1. Download and extract Bigtop tarball

```
$ curl -sLO https://downloads.apache.org/bigtop/bigtop-3.5.0/bigtop-3.5.0-project.tar.gz
$ tar zxf bigtop-3.5.0-project.tar.gz
$ cd bigtop-3.5.0
```

2. Install Puppet then tools including compilers and libraries

```
$ sudo bigtop_toolchain/bin/puppetize.sh
$ sudo puppet apply ¥
 --modulepath=".:/etc/puppet/modules:/usr/share/puppet/modules:/etc/puppet/code/modules" ¥
 -e "include bigtop_toolchain::installer"
(or `./gradlew toolchain` if supported JDK is already available)
```

```
$ ./gradlew bigtop-utils-pkg airflow-pkg repo
$ ls output
```

3. build RPM/DEB packages then create YUM/APT repository

Building packages in containers

```
(installing Docker)                    1. Download and extract Bigtop tarball

$ curl -sLO https://downloads.apache.org/bigtop/bigtop-3.5.0/bigtop-3.5.0-project.tar.gz
$ tar zxf bigtop-3.5.0-project.tar.gz
$ cd bigtop-3.5.0                       2. build RPM/DEB packages then create
                                        YUM/APT repository in Docker container


$ ./gradlew bigtop-utils-pkg-ind airflow-pkg-ind repo-ind ¥
    -POS=rockylinux-9 ¥
    -Pprefix=3.5.0 ¥
    -Pdocker-run-option="--privileged" ¥
    -Pmvn-cache-volume=true


$ ls output
```

Installing Bigtop pre-built RPM packages (Rocky Linux 9)

1. Download and install repository definition

```
$ sudo curl -L -o /etc/yum.repos.d/bigtop.repo ¥
    https://downloads.apache.org/bigtop/bigtop-3.5.0/repos/rockylinux-9/bigtop.repo
```

2. Import GPG key

```
$ sudo rpm --import ¥
    https://downloads.apache.org/bigtop/bigtop-3.5.0/repos/GPG-KEY-bigtop
```

```
$ sudo dnf install airflow bigtop-utils
$ /usr/lib/airflow/bin/airflow version
2.10.5
```

3. Update repository information and Install components via package manager

Installing Bigtop pre-built DEB packages (Ubuntu 24.04)

1. Download and install repository definition

```
$ sudo curl -o /etc/apt/sources.list.d/bigtop.list ¥
    https://downloads.apache.org/bigtop/bigtop-3.5.0/repos/ubuntu-24.04/bigtop.list
```

2. Import GPG key

```
$ curl -s https://downloads.apache.org/bigtop/bigtop-3.5.0/repos/GPG-KEY-bigtop ¥
    | sudo apt-key add -
```

3. Update repository information and Install components via package manager

```
$ sudo apt-get update
$ sudo apt-get install airflow bigtop-utils
$ /usr/lib/airflow/bin/airflow version
2.10.5
```

Provisioning by Puppet manifest on Rocky Linux 9

```
$ curl -sLO https://downloads.apache.org/bigtop/bigtop-3.5.0/bigtop-3.5.0-project.tar.gz
$ tar zxf bigtop-3.5.0-project.tar.gz
$ cd bigtop-3.5.0
$ sudo bigtop_toolchain/bin/puppetize.sh
$ cat > bigtop-deploy/puppet/hieradata/site.yaml <<'EOF'
bigtop::bigtop_repo_uri: http://repos.bigtop.apache.org/releases/3.5.0/rockylinux/9/$basearch
bigtop::hadoop_head_node: "host1.example.com"
hadoop_cluster_node::cluster_nodes: ["host1.example.com ", "host2.example.com ", host3.example.com "]
hadoop_cluster_node::cluster_components: ["bigtop-utils", "airflow"]
EOF

$ cp -r bigtop-deploy/puppet/hiera* /etc/puppet/
$ puppet apply ¥
      --hiera_config=/etc/puppet/hiera.yaml ¥
      --modulepath=./bigtop-deploy/puppet/modules:/usr/share/puppet/modules:etc/puppet/code/modules ¥
      ./bigtop-deploy/puppet/manifests
```

Provisioning by Puppet manifest on Ubuntu 24.04

```
$ curl -sLO https://downloads.apache.org/bigtop/bigtop-3.5.0/bigtop-3.5.0-project.tar.gz
$ tar zxf bigtop-3.5.0-project.tar.gz
$ cd bigtop-3.5.0
$ sudo bigtop_toolchain/bin/puppetize.sh
$ cat > bigtop-deploy/puppet/hieradata/site.yaml <<'EOF'
bigtop::bigtop_repo_uri: http://repos.bigtop.apache.org/releases/3.5.0/ubuntu/24.04/$(ARCH)
bigtop::hadoop_head_node: "host1.example.com"
hadoop_cluster_node::cluster_nodes: ["host1.example.com ", "host2.example.com ", host3.example.com "]
hadoop_cluster_node::cluster_components: ["bigtop-utils", "airflow"]
EOF

$ cp -r bigtop-deploy/puppet/hiera* /etc/puppet/
$ puppet apply ¥
    --hiera_config=/etc/puppet/hiera.yaml ¥
    --modulepath=./bigtop-deploy/puppet/modules:/usr/share/puppet/modules ¥
    ./bigtop-deploy/puppet/manifests
```

Running smoke tests

```
$ curl -sLO https://downloads.apache.org/bigtop/bigtop-3.5.0/bigtop-3.5.0-project.tar.gz
$ tar zxf bigtop-3.5.0-project.tar.gz
$ cd bigtop-3.5.0

$ . /usr/lib/bigtop-utils/bigtop-detect-javahome
$ export AIRFLOW_HOME=/usr/lib/airflow
$ ./gradlew bigtop-tests:smoke-tests:airflow:test -Psmoke.tests --info
```

provisioning and smoke-tests on containers

```
(installing Docker and Docker Compose; building packages and creating local repo )
$ cd provisioner/docker
$ ./docker-hadoop.sh --create 1 --memory 16g ¥
                     --image bigtop/puppet:3.5.0-rockylinux-9 ¥
                     --repo file:///bigtop-home/output ¥
                     --enable-local-repo --disable-gpg-check ¥
                     --stack bigtop-utils,airflow ¥
                     --smoke-tests airflow

$ ./docker-hadoop.sh --list
NAME                             IMAGE                              COMMAND        SERVICE    CREATED
STATUS         PORTS
20251002_144838_r3229_bigtop_1   bigtop/puppet:3.5.0-rockylinux-9   "/sbin/init"   bigtop     3 minutes
ago    Up 3 minutes

$ ./docker-hadoop.sh --exec 1 /bin/bash
[root@17a680a4f14a /]#
```

Why we chose Airflow as our new workflow orchestrator

Why we chose Airflow as our new workflow orchestrator

- De-facto standard of workflow orchestrator

- Widespread userbase and vibrant community

- Affinity with the existing Bigtop components

Affinity with the existing Bigtop software stack

- Bigtop is composed of various ASF products, e.g., Hadoop, Spark, Flink, Kafka

- Airflow's diverse providers enable integration between them

- Enterprise-ready features are also supported, e.g., Kerberos support, proven scalability, fine-grained security

Airflow integration example: BigPetStore DAG

BigPetStore: Simple ETL, data analytics and ML application

1. Generate synthetic
sales transaction data

generate_task
■ success
SparkSubmitOperator

2. Read transaction data, normalize
and load them into separated tables

transform_task
SparkSubmitOperator

3. Analyze transactions by some axes

analyze_task
SparkSubmitOperator

3. Recommend products for customers
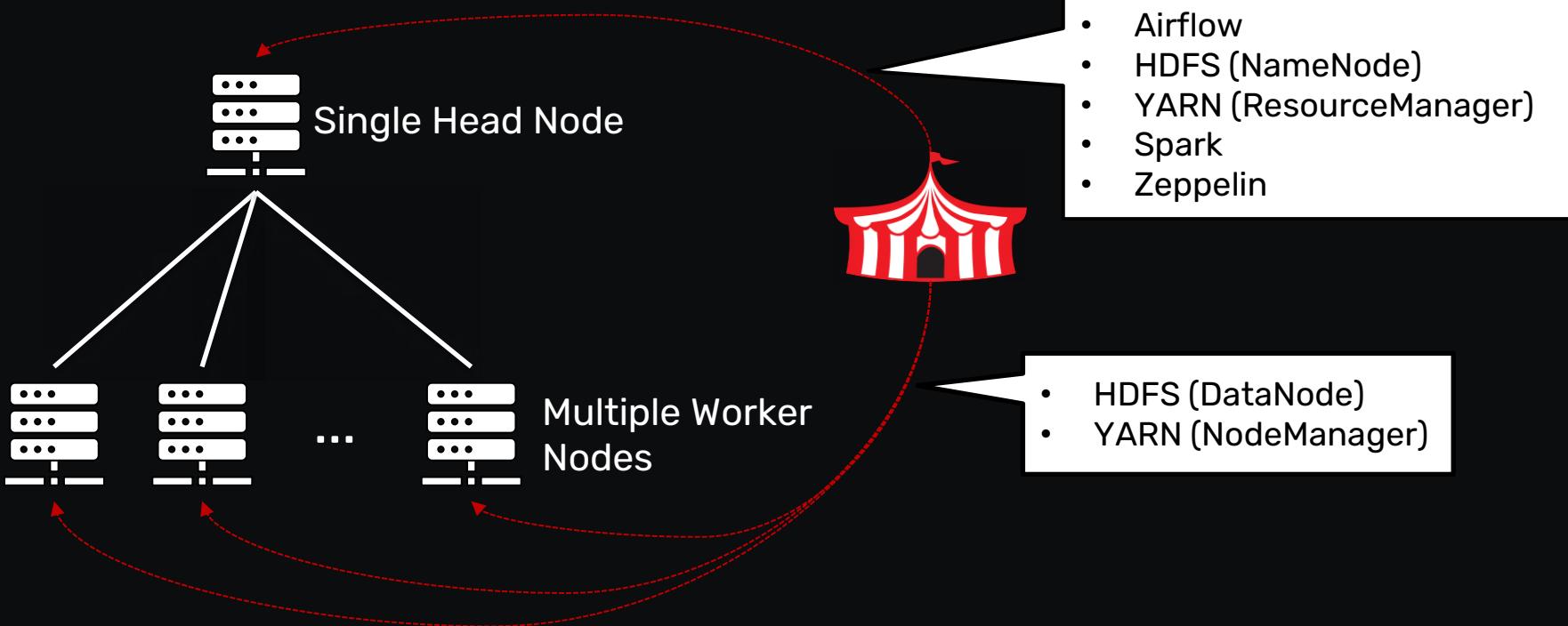
recommend_task
SparkSubmitOperator

For details, see
https://github.com/apache/bigtop/blob/rel/3.5.0/bigtop-bigpetstore/bigpetstore-spark/README.md

System Composition

Single Head Node

Multiple Worker Nodes

- Airflow
- HDFS (NameNode)
- YARN (ResourceManager)
- Spark
- Zeppelin

- HDFS (DataNode)
- YARN (NodeManager)

Deployment process

- Download Bigtop 3.5.0 and install Puppet on each node

```
$ curl -sIO https://dlcdn.apache.org/bigtop/bigtop-3.5.0/bigtop-3.5.0-project.tar.gz
$ tar xf bigtop-3.5.0-project.tar.gz
$ sudo bigtop-3.5.0/bigtop_toolchain/bin/puppetize.sh
```

- Enable "install BigPetStore DAG" option (and some required settings) by modifying the configuration file

```
$ tail -5 bigtop-3.5.0/bigtop-deploy/puppet/hieradata/bigtop/cluster.yaml
# Airflow
airflow::server::executor: "LocalExecutor"          Default parameters regarding Airflow
airflow::server::load_examples: false
airflow::server::sql_alchemy_conn: "postgresql+psycopg2://..."
airflow::server::install_bigpetstore_example: false
$ cat <<EOT > bigtop-3.5.0/bigtop-deploy/puppet/hieradata/site.yaml
bigtop::hadoop_head_node: fqdn.of.head.node
hadoop::hadoop_storage_dirs: [/data]          Override parameters in accordance with our requirements
hadoop_cluster_node::cluster_components: [airflow, hdfs, yarn, spark, zeppelin]
airflow::server::install_bigpetstore_example: true
EOT
```

Deployment process (continued)

- Distribute configuration file from head node to all of worker nodes

```
$ for node in $NODES; do scp bigtop-3.5.0/bigtop-deploy/puppet/hieradata/site.yaml ¥
> ${node}:~/bigtop-3.5.0/bigtop-deploy/puppet/hieradata/site.yaml; done
```

- Copy configuration files to Puppet's default location and apply them on each node

```
$ sudo cp -r bigtop-3.5.0/bigtop-deploy/puppet/hiera* /etc/puppet/
$ sudo puppet apply --modulepath bigtop-3.5.0/bigtop-deploy/puppet/modules:/usr/share/puppet/modules ¥
> bigtop-3.5.0/bigtop-deploy/puppet/manifests
(snip)
Notice: /Stage[main]/Airflow::Server/Package[airflow]/ensure: created
Notice: /Stage[main]/Airflow::Server/File[/var/lib/airflow/airflow.cfg]/ensure: defined content as
'{sha256}acb4d3f8a5e9b1abe56b4b6d0aba5ae611e8fbbe10a17bbaeadfbd71be345d7d'
Notice: /Stage[main]/Airflow::Server/Service[airflow-webserver]/ensure: ensure changed 'stopped' to 'running'
Notice: /Stage[main]/Airflow::Server/Exec[install-spark-provider]/returns: executed successfully
Notice: /Stage[main]/Airflow::Server/File[/var/lib/airflow/dags]/ensure: created
Notice: /Stage[main]/Airflow::Server/File[/var/lib/airflow/dags/example_bigpetstore.py]/ensure: defined
content as '{sha256}ad9a3c7ef4b76fa1cfbd047207ff48143405b6a687d901690a3d693c7fc5b711'
(snip)
Notice: Applied catalog in 570.91 seconds
```

Airflow

⚙ 12:20 UTC ▾ →] Log In

Sign In

Enter your login and password below:

Username:

👤 | admin

Password:

🔑 | •••••|

Sign In

Hadoop

Overview   Datanodes   Datanode Volume Failures   Snapshot   Startup Progress

Utilities ▾

Overview

'node0.fmkb5agmaolunimw4xslea2kkh.lx.internal.cloudapp.net:8020'
(✔active)

Started:	Thu Sep 25 20:58:13 +0900 2025
Version:	3.3.6, r79fb2857c7c52d680cbaba0e5cb1e9a0f0d33132
Compiled:	Thu Aug 14 22:06:00 +0900 2025 by jenkins from (HEAD detached at 79fb2857)
Cluster ID:	CID-4449d4e0-e2da-4713-bf61-936c8e079071
Block Pool ID:	BP-914695455-10.0.1.7-1758801490963

Zeppelin

Notebook ▾   Job

Q Search

● anonymous ▾

Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!
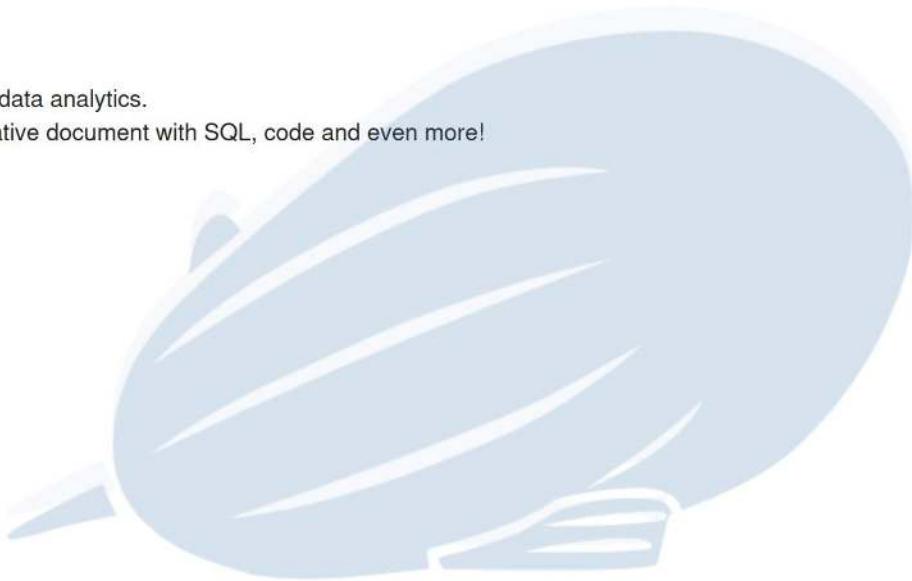
Notebook ⟳

⬆ Import note

▯ Create new note

Q Filter

- ◼ Flink Tutorial
- ◼ Miscellaneous Tutorial
- ◼ Python Tutorial
- ◼ R Tutorial
- ◼ Spark Tutorial

Help

Get started with Zeppelin documentation

Community

Please feel free to help us to improve Zeppelin,
Any contribution are welcome!

Some additional steps

- Build BigPetStore's jar file ourselves on head node

```
$ cd bigtop-3.5.0/bigtop-data-generators
$ ../gradlew clean publishToMavenLocal
$ cd ../bigtop-bigpetstore/bigpetstore-spark
$ ../../gradlew shadowJar
$ mv build/libs/bigpetstore-spark-3.5.0-all.jar /tmp/
```

- Create Airflow user's home directory on HDFS

```
$ sudo -u hdfs hdfs dfs -mkdir /user/airflow
$ sudo -u hdfs hdfs dfs -chown airflow:hadoop /user/airflow
```

- These steps will be automatically done in the future release

DAGs

All **1**	Active **0**	Paused **1**		Running **0**	Failed **0**		Filter DAGs by tag

ⓘ	DAG ⇕	Owner ⇕	Runs ⓘ	Schedule
⬤	**bigpetstore_dag**	airflow	◯ ◯ ◯ ◯	None ⓘ

« ‹ **1** › »

Trigger DAG: bigpetstore_dag

DAG conf Parameters

bigpetstore_jar_path: /tmp/bigpetstore-spark-3.5.0-all.jar

Generated Configuration JSON and Dagrun Options ▾

◉ Unpause DAG when triggered

[Trigger]   [Cancel]
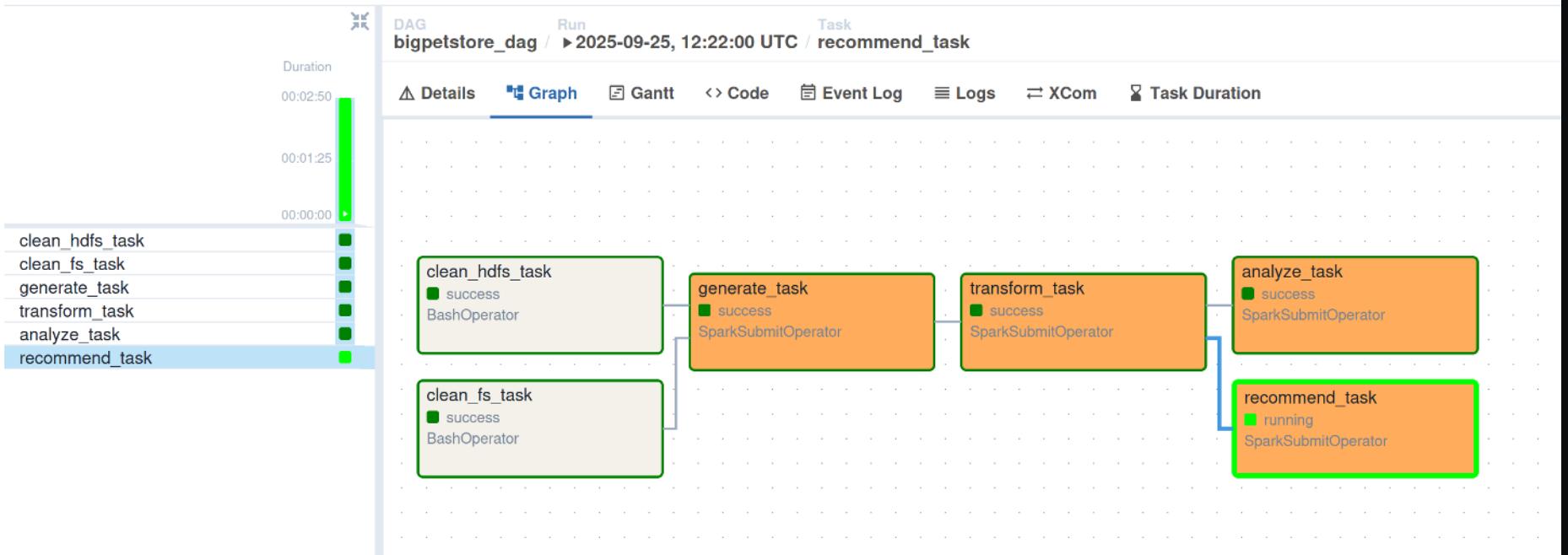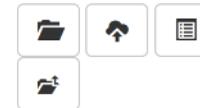
Browse Directory

/user/airflow	Go!

Show 25 entries   Search: [   ]

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
☐	drwxr-xr-x	airflow	hadoop	0 B	Sep 25 21:25	0	0 B	.sparkStaging	🗑
☐	drwxr-xr-x	airflow	hadoop	0 B	Sep 25 21:23	0	0 B	generated_data	🗑
☐	drwxr-xr-x	airflow	hadoop	0 B	Sep 25 21:24	0	0 B	transformed_data	🗑

Showing 1 to 3 of 3 entries   Previous 1 Next

Untitled Note 1

default ▾
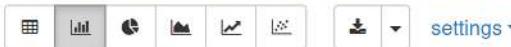
```
202,
193,
289,
```
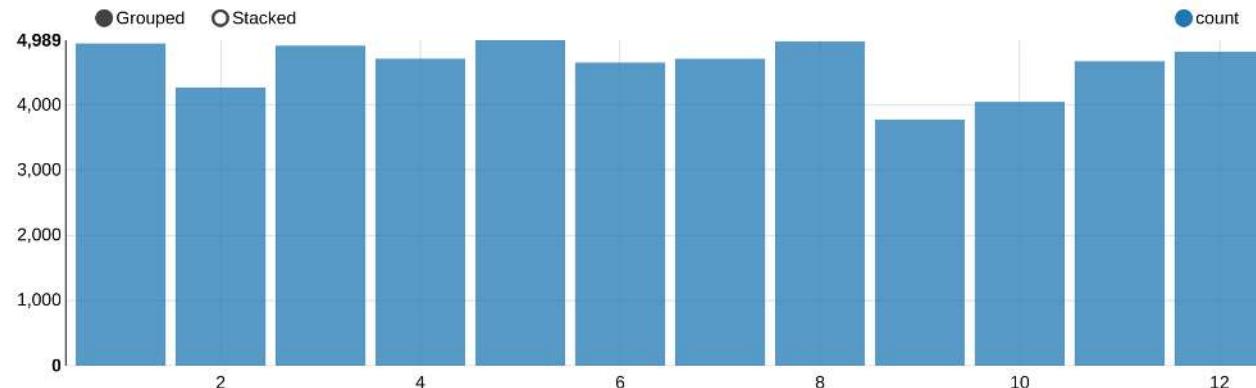
Took 3 sec. Last updated by anonymous at September 25 2025, 9:37:18 PM.

FINISHED

```
%sh
echo %table
echo -e month\\tcount
jq -r '.transactionsByMonth[] | [.month, .count] | @tsv' /tmp/PetStoreStats.json
```

● Grouped   ○ Stacked

● count
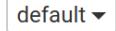
Took 0 sec. Last updated by anonymous at September 25 2025, 9:37:39 PM. (outdated)

Notebook | **Job**

anonymous

default

```
%sh
cat /tmp/recommendations.json | jq .recommendations[:10]
```

FINISHED

```
[
  {
    "customerId": 0,
    "productIds": [
      202,
      193,
      141,
      289,
      40
    ]
  },
  {
    "customerId": 1,
    "productIds": [
      202,
      193,
      289,
      141
```

Took 3 sec. Last updated by anonymous at September 25 2025, 9:37:18 PM.

Conclusion

AIRFLOW SUMMIT

- Bigtop is an ASF project that offers software distribution for big data processing

- Bigtop adopted Airflow as its new workflow orchestrator since v3.5.0

- By using Bigtop, users can easily build their own data platform including Airflow

- Airflow is a keystone of the integration between diverse Bigtop components

- BigPetStore DAG is an example of that integration and we'll add more diverse usecases and workloads in the future, e.g., streaming, GenAI, etc.

Questions?

Kengo.Seki sekikn@apache.org
Masatake Iwasaki iwasakims@apache.org