Apache DolphinScheduler 3.2.0 features preview and OceanBase Integration

Kyle Zhike Chen

Data Engineer Manager at GoTo Financial

Linkedin: kk17

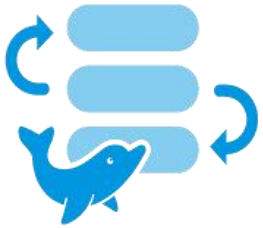
GitHub: kk17

Aug 21, 2023

DolphinScheduler

Agenda

- Introduce to DolpinScheduler
- 3.1.x & 3.2.0 feature preview
- Workflow Demo - OceanBase integration



Introduce to DolpinScheduler

What is DolphinScheduler

Workflow Platform

- Complex Dependence
- Concurrency&Limitations
- Retry & Alert & Backfill
- Monitoring & Metric

Drag and Drop First

- Drag&Drop First
- Python API & Open API
- Yaml Definition

A distributed and extensible workflow scheduler platform with powerful DAG visual interfaces

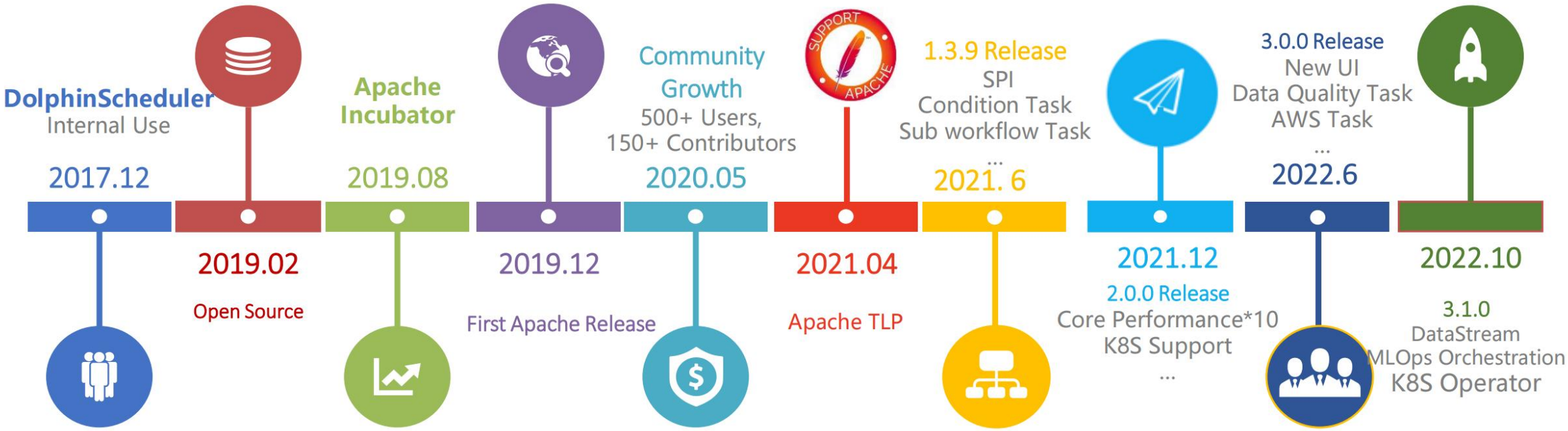
High Availability & Performance

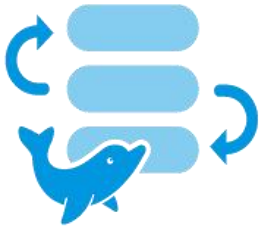
- Decentralization
- Native HA Queue
- Fault Tolerance
- 1m+ tasks in prod env

Plug-in Based Design

- Tasks: **40+**
- DataSources: **11+**
- Alerts:**10**
- Registration: **3**

DolphinScheduler History





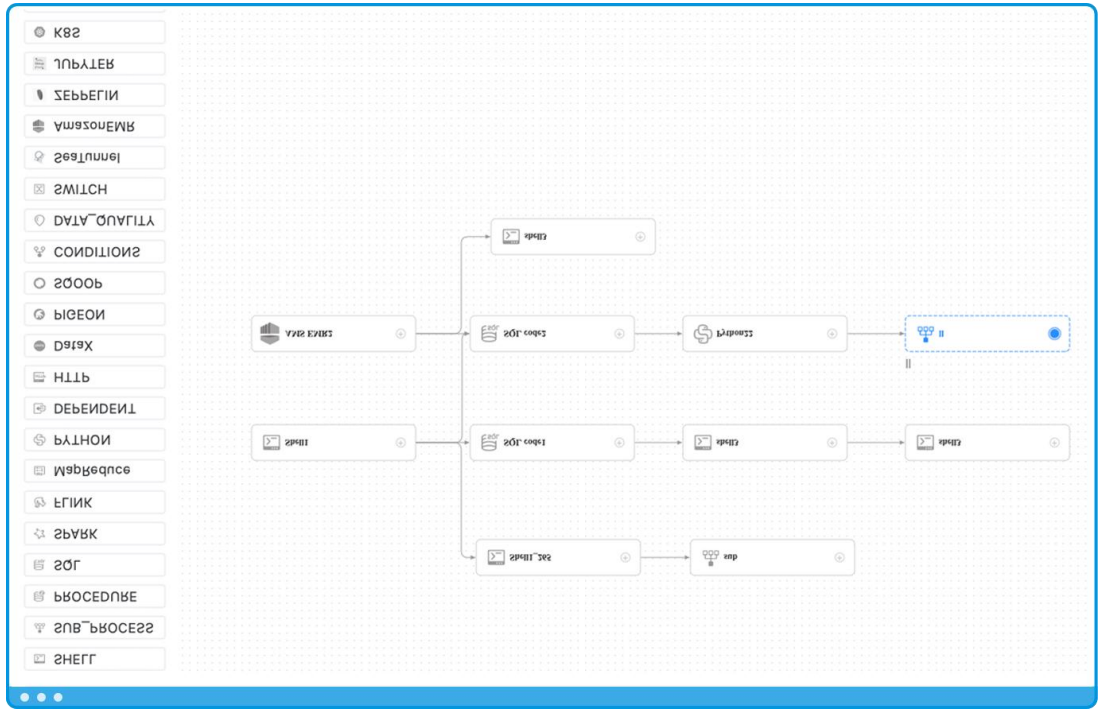
DolphinScheduler 3.1.x Features

Currently the laster version is 3.1.8

Agile and flexible workflow management

- Drag & Drop(WYSIWYG) workflow creation
- Python / Yaml generate workflow
- API workflow management

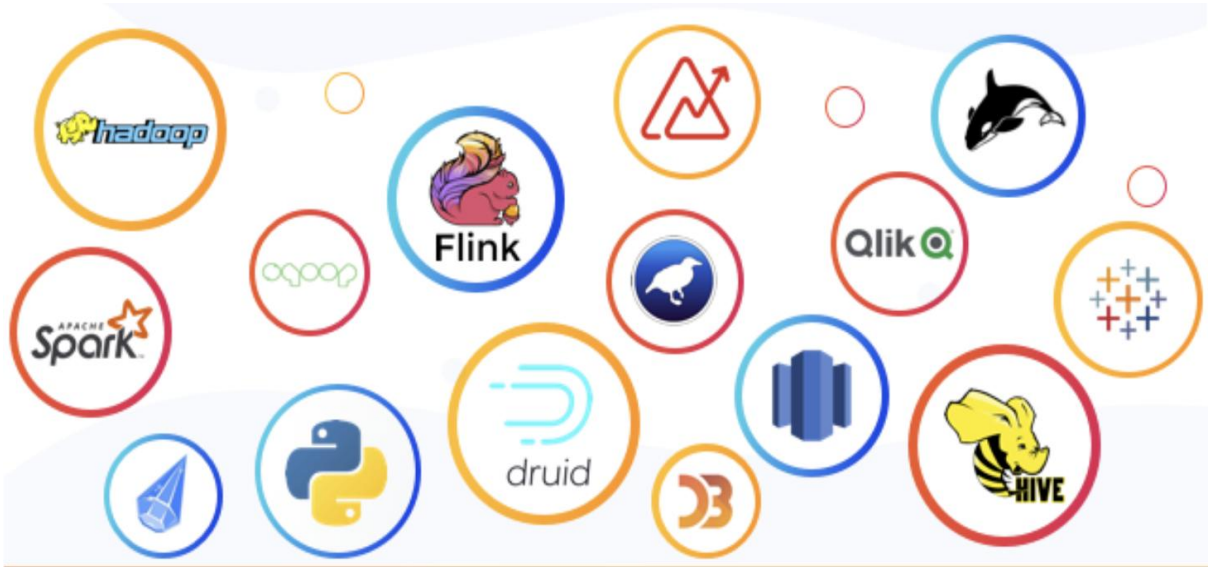
Suitable for both developers and non-developers



Batch & Streaming Job Support

- Spark, Hive, MR, DataX, etc
- SQL Data sources: MySQL, Presto, Trino, etc
- Flink, Sparking streaming Support

One place big data workflow platform



Data Quality Support

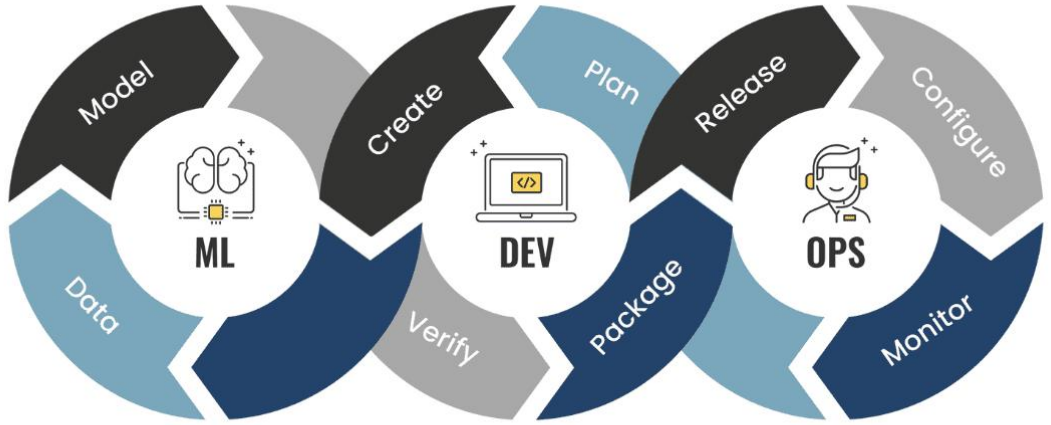
- Build in Data Quality rules
- Custom SQL Check
- Seamlessly integrated into the workflow

Improve data quality with minimal operation costs

Rule Name	Rule Type
NullCheck	Single Table
Custom Sql	Custom Sql
Multi Table Accuracy	Multi Table Accuracy
Multi Table Compare	Multi Table Compare
FieldLengthCheck	Single Table
UniquenessCheck	Single Table
RegexpCheck	Single Table
TimelinessCheck	Single Table
EnumerationCheck	Single Table
TableCountCheck	Single Table

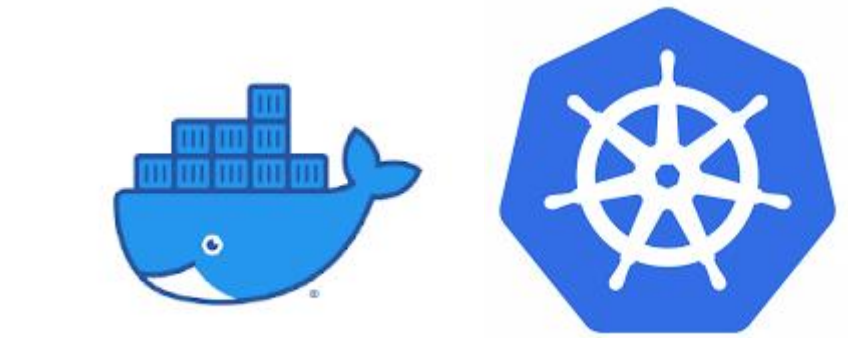
MLOps Support

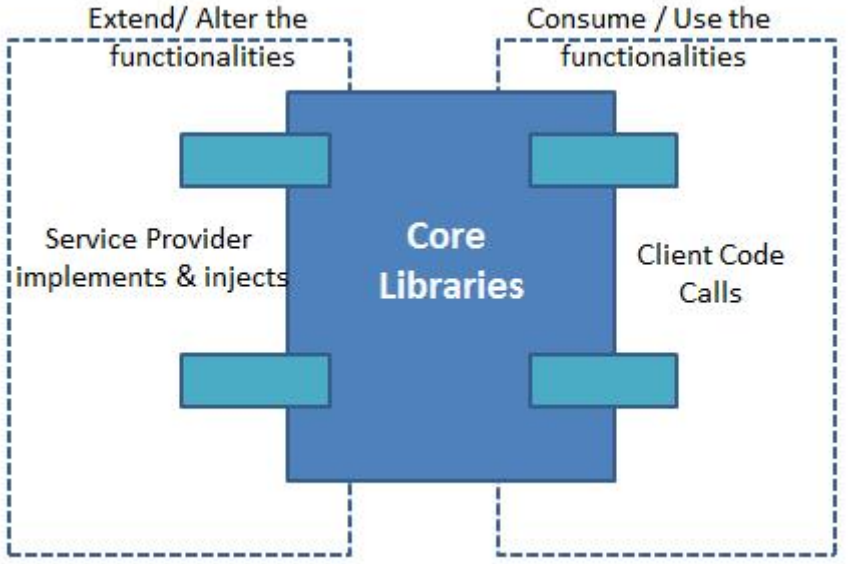
- KubeFlow, MLflow, Sagemaker,DVC
- Jupyter, PyTorch
- OpenMLDB



Cloud-native & Extensibility

- Docker/Kubernetes deployment
- Kubernetes Task
- User-defined Task via SPI





DolphinScheduler 3.2.0 Features preview

- Add default tenant
- Add support for multiple data sources: Snowflake, Kyuubi, OceanBase, etc
- Add new task types
- Add caching support for tasks
- Enhance existing task types (e.g. Sqoop, SQL, etc.)
- Enhance architecture (e.g. Alert supports HA, SSO support, etc.)
- Specify workflow execution forward and backward when re-running tasks
- Add support for remote logs from OSS, GCS, S3
- Get real-time logs from Kubernetes pods
- Enhance task parameters
- Add support for Alibaba Cloud OSS in the resource center
- Enhance the Restful API
- Add support for ETCD and JDBC registry centers

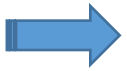
DolphinScheduler vs Airflow

	DolphinScheduler	Airflow
Program language	written in Java	written in Python
DAG definition	Drag and Drop First	Python DAG files
DAG versioning	build in version control support	integrate with git-sync
Component integration	Rich integration of big data and ML components	general component
Backfill	UI support	command line
Multi-tenancy	yes	partially support
Streaming job	streaming task support	no streaming task support
Data Quality	Yes	no

DolphinScheduler

Workflow creation demo

Oceanbase integration as an example

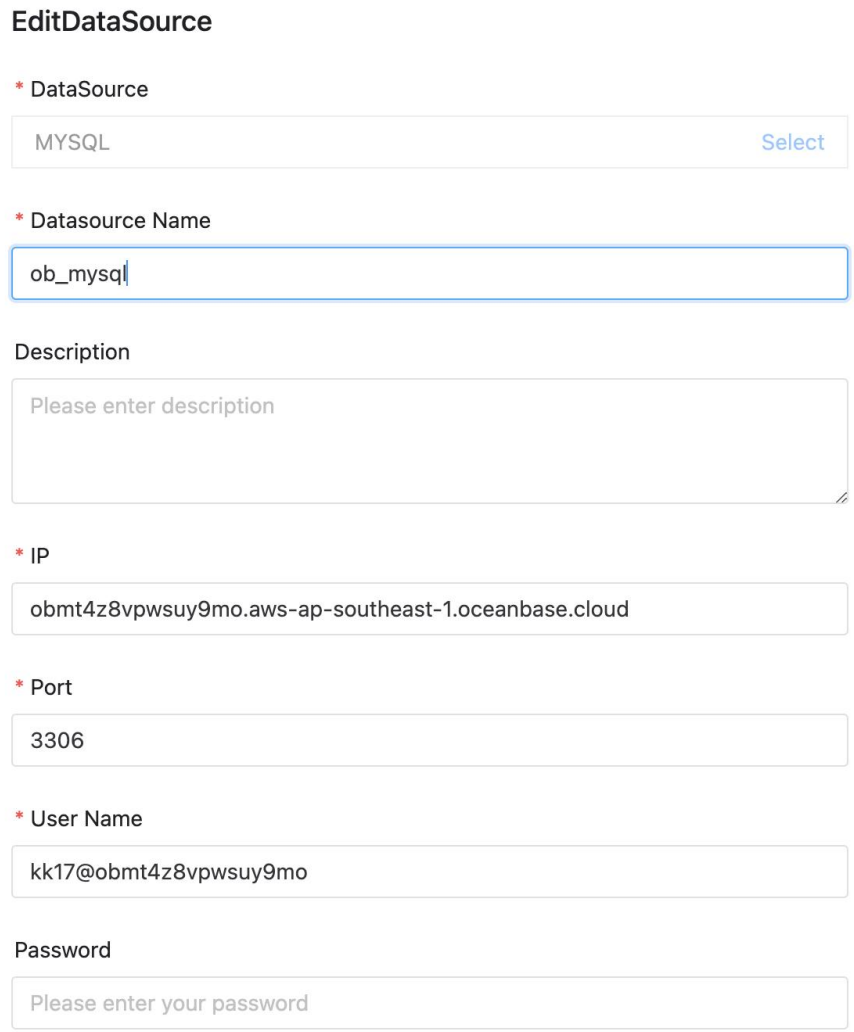


Define OceanBase Datasource

Before DolpinScheduler 3.2.0, use MySQL type
After 3.2.0, can use MySQL or OceanBase type

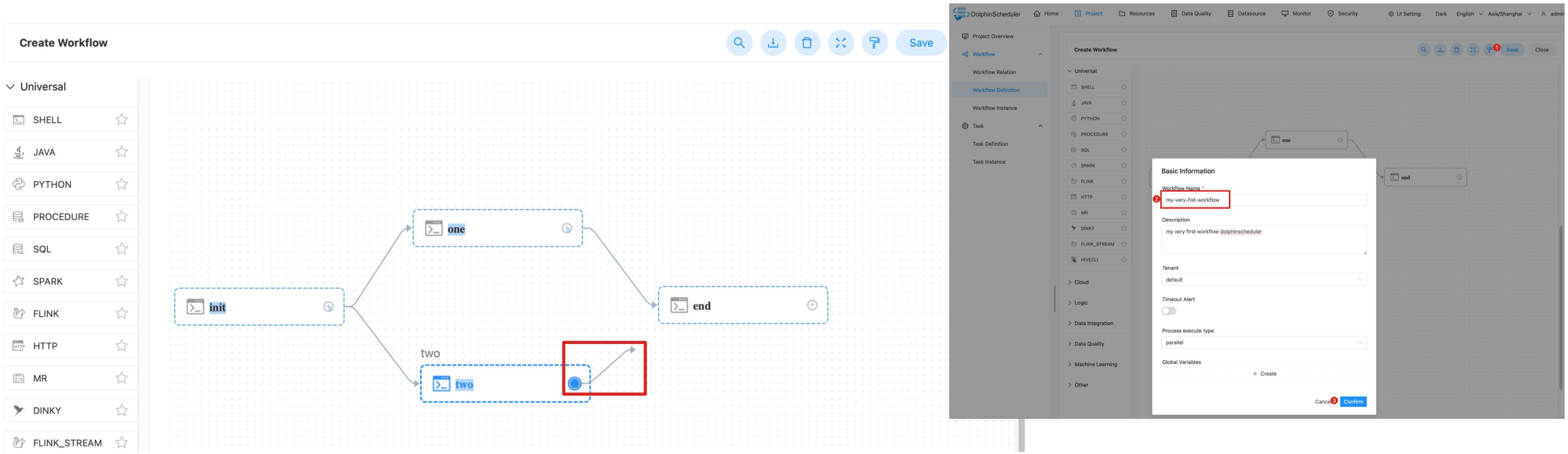
UI: Datasource -> Create Sources

need both test and prod datasource



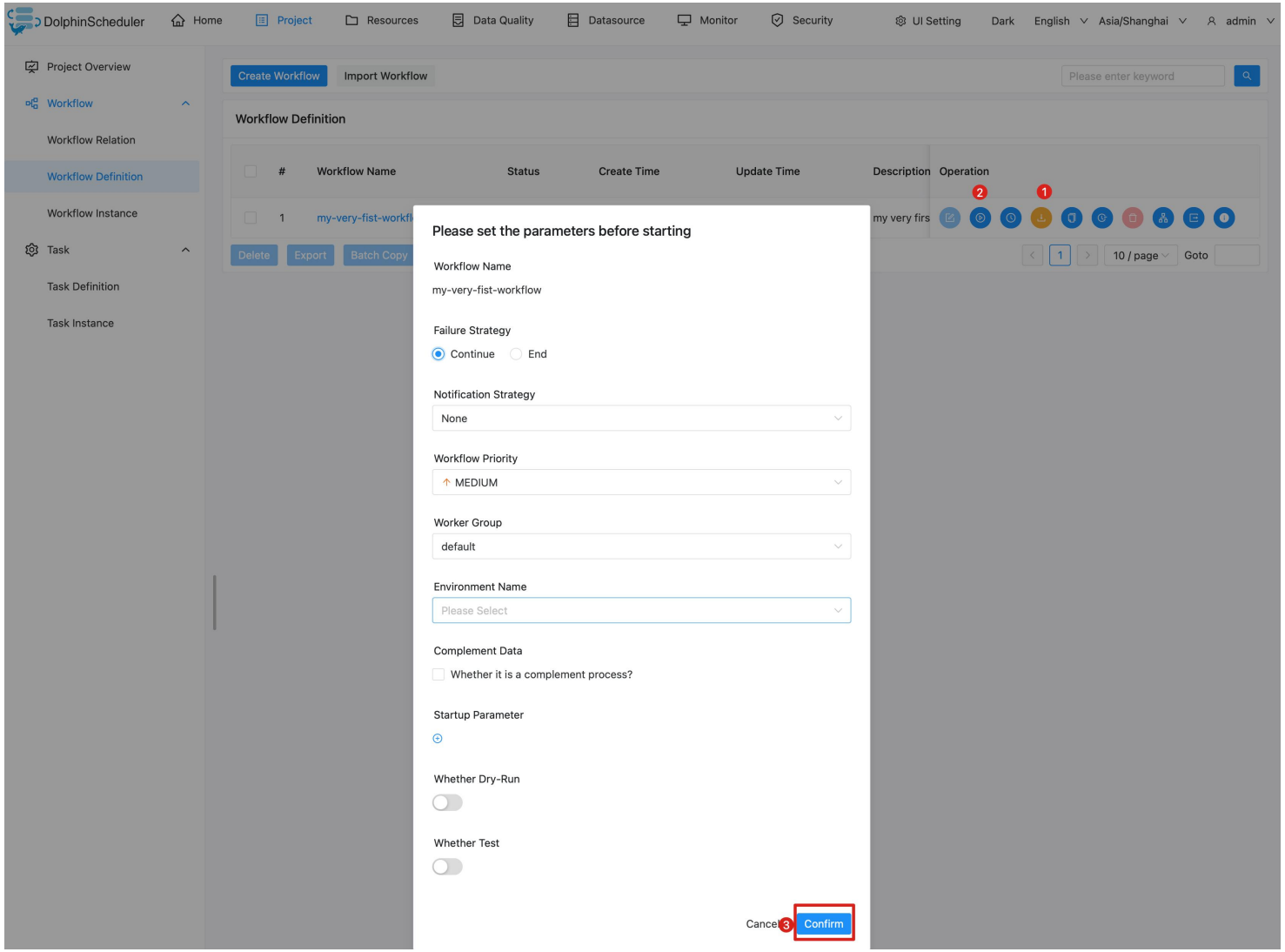
Create workflow definition

UI: Project->Workflow definition
Drag & Drop to craete SQL type task, and set dependence



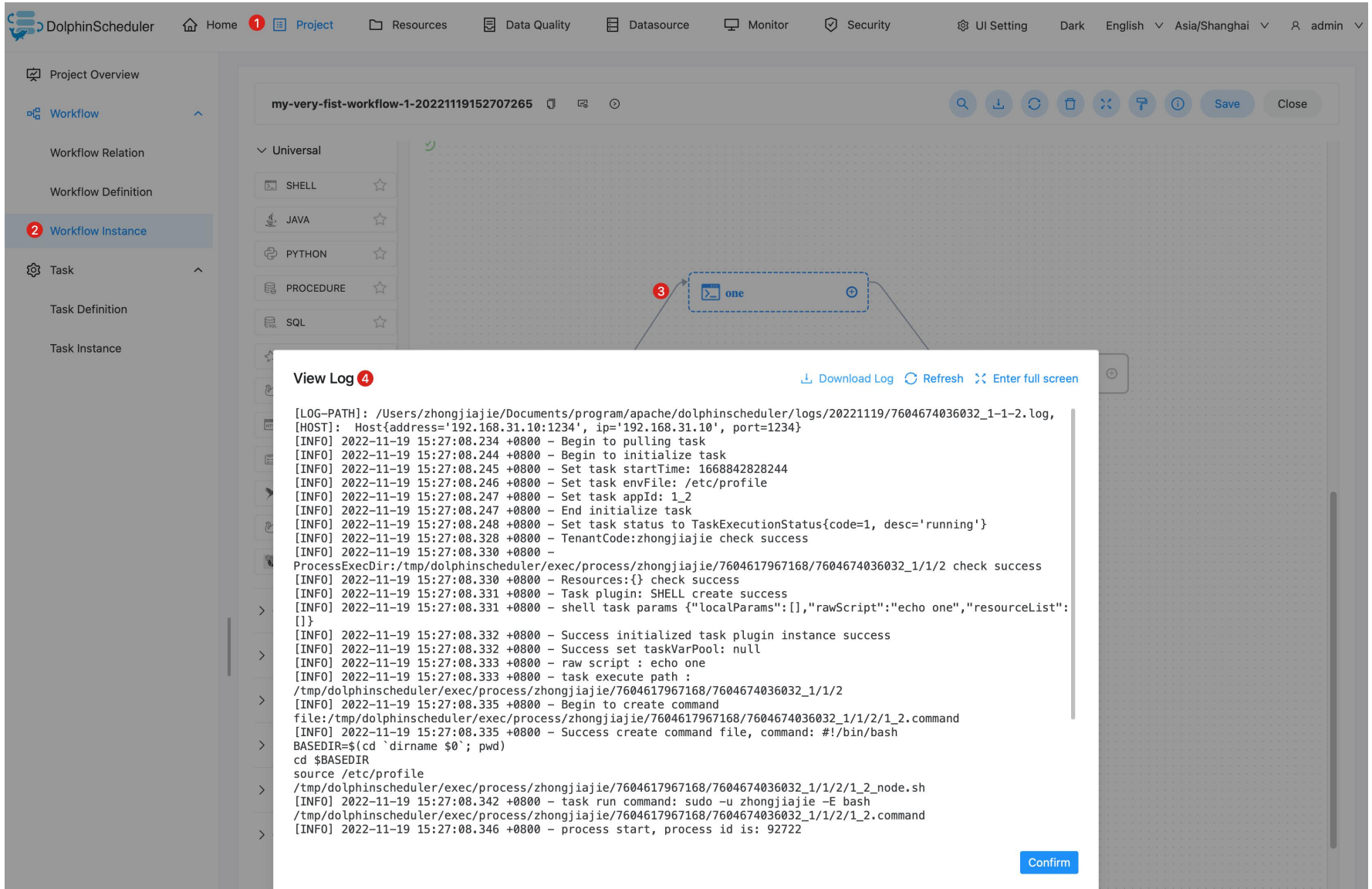
Publish and Run workflow

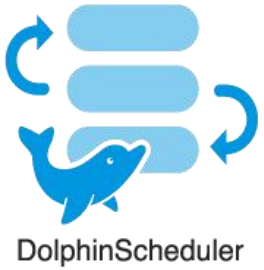
Save, publish and trigger the workflow



Check status

View workflow instance status and view log





Community

Users & Integrations

Some of Our Users

IBM, vmware, cisco, DELL, Lenovo, NOKIA, accenture, Tencent, WPS, Walmart, SHEIN, McDonald's, Fordeal

Some of Our Integrations

Apache Spark, Flink, hadoop, Hive, kubernetes, PyTorch, mlflow, DVC, jupyter, aws

Contact Us

- **Troubleshooting**
  - User Mail List: users@dolphinscheduler.apache.org
  - Slack: https://asf-dolphinscheduler.slack.com/channels/troubleshooting
- **Bug & Features Request & Features Discussion**
  - Developer Mail List: dev@dolphinscheduler.apache.org
  - GitHub Issue: https://github.com/apache/dolphinscheduler/issues
  - GitHub Pull Requests: https://github.com/apache/dolphinscheduler/pulls
- **Announcements**
  - Mail List: Both dev and users metions above
  - Twitter: https://twitter.com/dolphinschedule
  - Slack: https://asf-dolphinscheduler.slack.com/channels/announcements