











zixi0825 [Refactor][Engine] Refactor engine module (#581) ✓

1360764 · 3 weeks ago

.github	[Fix][Workflows] Fix mvnw p...	2 years ago
.mvn/wrapper	[Feature][CI] Add ci test (#208)	3 years ago
bin	[Feature][JMX] Add JMX su...	2 years ago
datavines-cli	[Feature][Sserver] add exec...	4 years ago
datavines-client	[Feature][Server] Support D...	2 years ago
datavines-common	[Feature][Connector] add m...	last month
datavines-connector	[Feature][Connector] Suppor...	last month
datavines-core	[Feature][Server] Add metad...	last month
datavines-dist	[Refactor][Engine] Refactor ...	3 weeks ago
datavines-engine	[Refactor][Engine] Refactor ...	3 weeks ago
datavines-metric	[Refactor][SPI] Refactor spi ...	last month
datavines-notification	[Refactor][SPI] Refactor spi ...	last month
datavines-registry	[Refactor][SPI] Refactor spi ...	last month
datavines-runner	[Refactor][Engine] Refactor ...	3 weeks ago
datavines-server	[Refactor][Engine] Refactor ...	3 weeks ago
datavines-spi	[Refactor][SPI] Refactor spi ...	last month
datavines-ui	[Fix][UI] Fix login required w...	last month
	[Feature][Engine] Add flink e...	last year

deploy		
 docs/img	[Fix][Engine] Fix spark engin...	11 months ago
 scripts/sql	[Feature][Server] Add metad...	last month
 tools/checkstyle	[Feature][Build] add license ...	4 years ago
 .gitignore	[Refactor][Engine] Refactor ...	3 weeks ago
 HEADER	[Feature][Build] add license ...	4 years ago
 LICENSE	init	4 years ago
 README.md	[Fix][Doc] fix readme doc wo...	2 years ago
 README.zh-CN.md	[Fix][Doc] fix readme doc wo...	2 years ago
 mvnw	[Feature][Server] Add data q...	2 years ago
 pom.xml	[Refactor][Engine] Refactor ...	3 weeks ago

 **README**  Apache-2.0 license

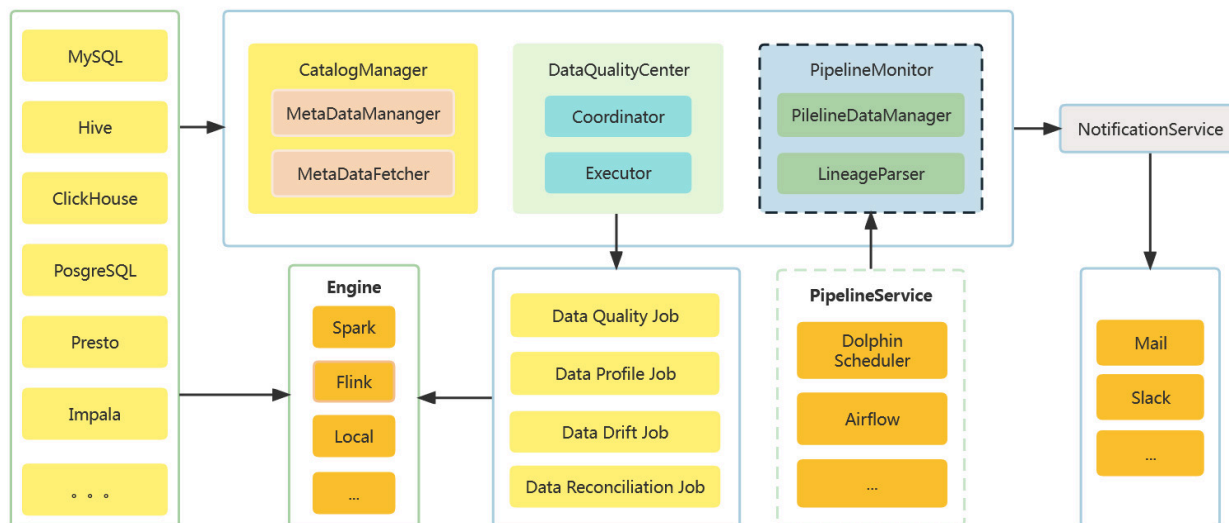


Datavines

document **English** 文档 中文版

Data quality is used to ensure the accuracy of data in the process of integration and processing. It is also the core component of DataOps. DataVines is an easy-to-use data quality service platform that supports multiple metric.

Architecture Design



Install

Need: Maven 3.6.1 and later

```
$ mvn clean package -Prelease -DskipTests
```



Features

Data Catalog

- Obtain **data source metadata** regularly to construct data directory
- Regular monitoring of **metadata changes**
- **Tag management** with support for metadata

The screenshot shows the MySQL interface for the 'datavines' database. The sidebar on the left lists various databases and schemas. The main content area displays a summary card with the following metrics:

Last Scan Time	Tables	Labels	Rules	Use heat
2023-07-16 17:44:03	43	0	0	0

Below the summary card, there is a table showing schema changes:

Table	Last Refresh Time	Column	Metrics
dv_actual_values	2023-07-16 17:44:12	8	0
dv_catalog_entity_definition	2023-07-16 17:44:12	10	0
dv_catalog_entity_instance	2023-07-16 17:44:12	14	0
dv_catalog_entity_metric_job_rel	2023-07-16 17:44:12	8	0
dv_catalog_entity_profile	2023-07-16 17:44:12	7	0
dv_catalog_entity_rel	2023-07-16 17:44:12	6	0
dv_catalog_entity_tag_rel	2023-07-16 17:44:12	7	0

Data Quality

- Built-in **27** data quality check rules
- Support **4** data quality check rule types
 - Single Table-Column Check
 - Single Table Custom **SQL** check
 - Cross Table Accuracy Check
 - Two Table Value Comparison Check
- Support schedule tasks for check
- Support **SLA** for **check result alert**

Data Quality Metric Job X

Job Configuration | Schedule Configuration | SLA Configuration | Configuration File

* Metric: Filter condition:

* Database:

* table:

* Column:

Expected value configuration

* Expected value type:

Verify configuration

* Formula: * Compare: * Threshold:

Execution engine configuration

* Execution engine:

Other config

Number of retries: Retry interval:

Data Profile

- Support timing execution of data detection, output **data profile report**
- Support **automatically identify** column types to automatically match appropriate data profile indicators
- Support **table row number trend** monitoring
- Support **data distribution** view

MySQL

course

datavines

- dv_actual_values
- dv_catalog_entity_definition
- dv_catalog_entity_instance
 - id
 - uuid
 - type
 - datasource_id
 - fully_qualified_name
 - display_name
 - description
 - properties
 - owner
 - version
 - status
 - create_time
 - update_time
 - update_by
 - dv_catalog_entity_metric_job_rel
 - dv_catalog_entity_profile
 - dv_catalog_entity_rel

MySQL > datavines > dv_catalog_entity_instance

Table dv_catalog_entity_instance

Last Scan Time	Columns	Labels	Rules	Use heat
2023-07-16 17:44:03	14	0	1	0

Profile | Column | Metrics | Schema Changes | Issues

Column	Type	Null	NotNull	Unique	Distinct
datasource_id	BIGINT	0 [0.00%]	36001 [100.00%]	0 [0.00%]	> [0.01%]
fully_qualified_name	VARCHAR	0 [0.00%]	36001 [100.00%]	11431 [31.75%]	21119 [58.66%]

Top 10

Maximum: zebra.zebra_pr...
 Minimum: cbs
 Max Length: 85.00
 Min Length: 2.00
 Avg Length: 40.61

Plug-in Design

The platform is based on plug-in design, and the following modules support user-defined plug-ins to expand

- **Data Source:** MySQL , Impala , StarRocks , Doris , Presto , Trino , ClickHouse , PostgreSQL are already supported
- **Check Rules:** 27 check rules such as built-in null value check, non-null check, enumeration check, etc.
- **Job Execution Engine:** Two execution engines Spark and Local have been supported. The Spark engine currently only supports the Spark2.4 version, and the Local engine is a local execution engine developed based on JDBC , without relying on other execution engines.
- **Alert Channel:** Supported Email
- **Error Data Storage:** MySQL and local files are already supported (only Local execution engine is supported)
- **Registry:** Already supports MySQL , PostgreSQL and ZooKeeper

Multiple Execute Modes

- Provide **Web page** to configure check jobs, run jobs, view job execution logs, view error data and check results
- Support **online generation** job running scripts, submit jobs through `datavines-submit.sh` , can be used in conjunction with the scheduling system

Data Quality Metric Job ×

Job Configuration Schedule Configuration SLA Configuration Configuration File

Copy

Download

```
{
  "name": "COLUMN_BLANK(#)_task_1689603554248",
  "executePlatformType": "client",
  "executePlatformParameter": {},
  "engineType": "local",
  "engineParameter": {},
  "parameter": {
    "connectorParameter": {
      "type": "mysql",
      "parameters": {
        "database": "cbs",
        "password": "123456",
        "port": "3306",
        "host": "localhost",
        "user": "root",
        "properties": "useUnicode=true&characterEncoding=UTF-8&useSSL=false&serverTimezone=Asia/Shanghai"
      }
    },
    "metricParameterList": [
      {
        "metricType": "column_blank",
        "metricParameter": {
          "table": "cbs_ratio",
          "metricName": "column_blank"
        }
      }
    ]
  }
}
```

Easy Deployment & High Availability

- Less platform dependency, easy to deploy
- Minimal only rely on `MySQL` to start the project and complete the check of data quality operations
- Support horizontal expansion, automatic fault tolerance
- **Decentralized design**, `server` node supports horizontal expansion to improve performance
- Job **Automatic Fault Tolerance**, to ensure that jobs are not lost or repeated

Environmental Dependency

1. java runtime environment: jdk8
2. If the data volume is small, or the goal is merely for functional verification, you can use JDBC engine
3. If you want to run DataVines based on Spark, you need to ensure that your server has spark installed

Quick Start

Click [Document](#) for more information

Development

Click [Document](#) for more information

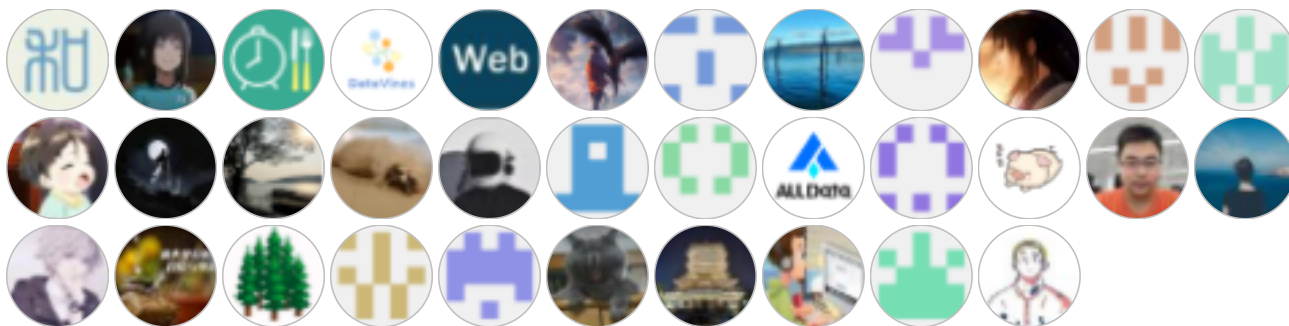
Contribution

PRs **welcome**

You can submit any ideas as [pull requests](#) or as [GitHub issues](#).

If you're new to posting issues, we ask that you read [How To Ask Questions The Smart Way](#) (This guide does not provide actual support services for this project!), [How to Report Bugs Effectively](#) prior to posting. Well written bug reports help us help you!

Thank you to all the people who already contributed to Datavines!



License

Datavines is licensed under the [Apache License 2.0](#). Datavines relies on some third-party components, and their open source protocols are also Apache License 2.0 or compatible with Apache License 2.0. In addition, Datavines also directly references or modifies some codes in Apache DolphinScheduler, SeaTunnel and Dubbo, all of which are Apache License 2.0. Thanks for contributions to these projects.

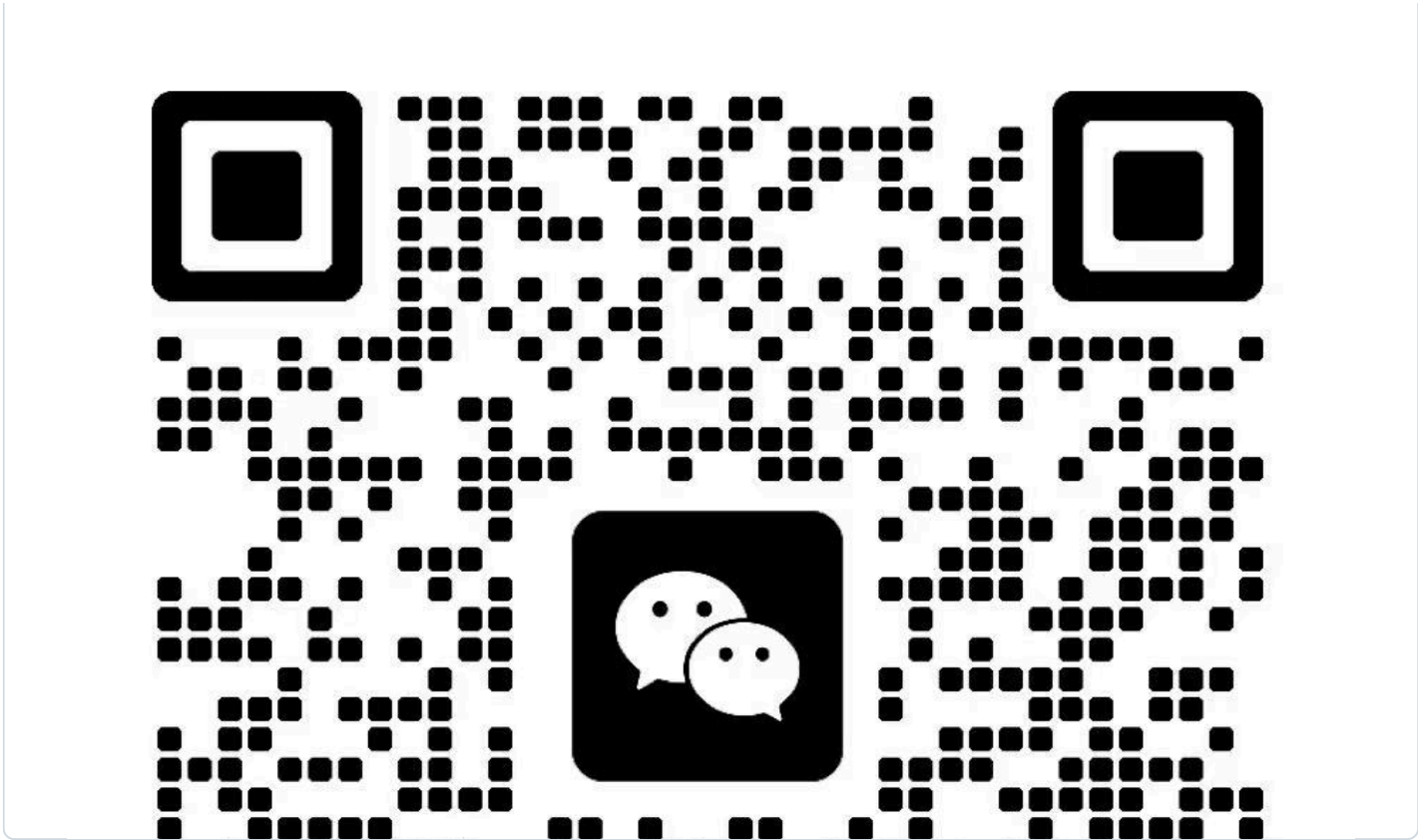
Social Media

- WeChat Official Account (in Chinese, scan the QR code to follow)

A promotional banner with an orange background. On the left, the text 'Datavines' is written in a large white font, followed by 'Data Observability Platform' in a smaller white font, and 'Follow us !' at the bottom left. On the right side, there is a large QR code with the Datavines logo in the center.

Contact Author

- Notes "Datavines" When Adding Me On WeChat



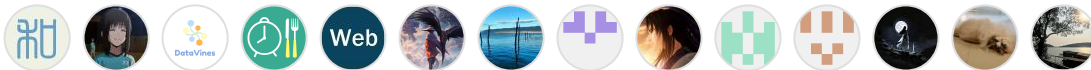
Releases

No releases published

Packages

No packages published

Contributors 34



[+ 20 contributors](#)

Languages

