上海 云数智聚 砥柱笃行

# CHINA APACHE HADOOP MEETUP 2022

🕒 2022年09月24日 9:00

📍 上海古井假日酒店4楼

# Apache Linkis Data Processing Practice

李孟

2022.10.22

# Background

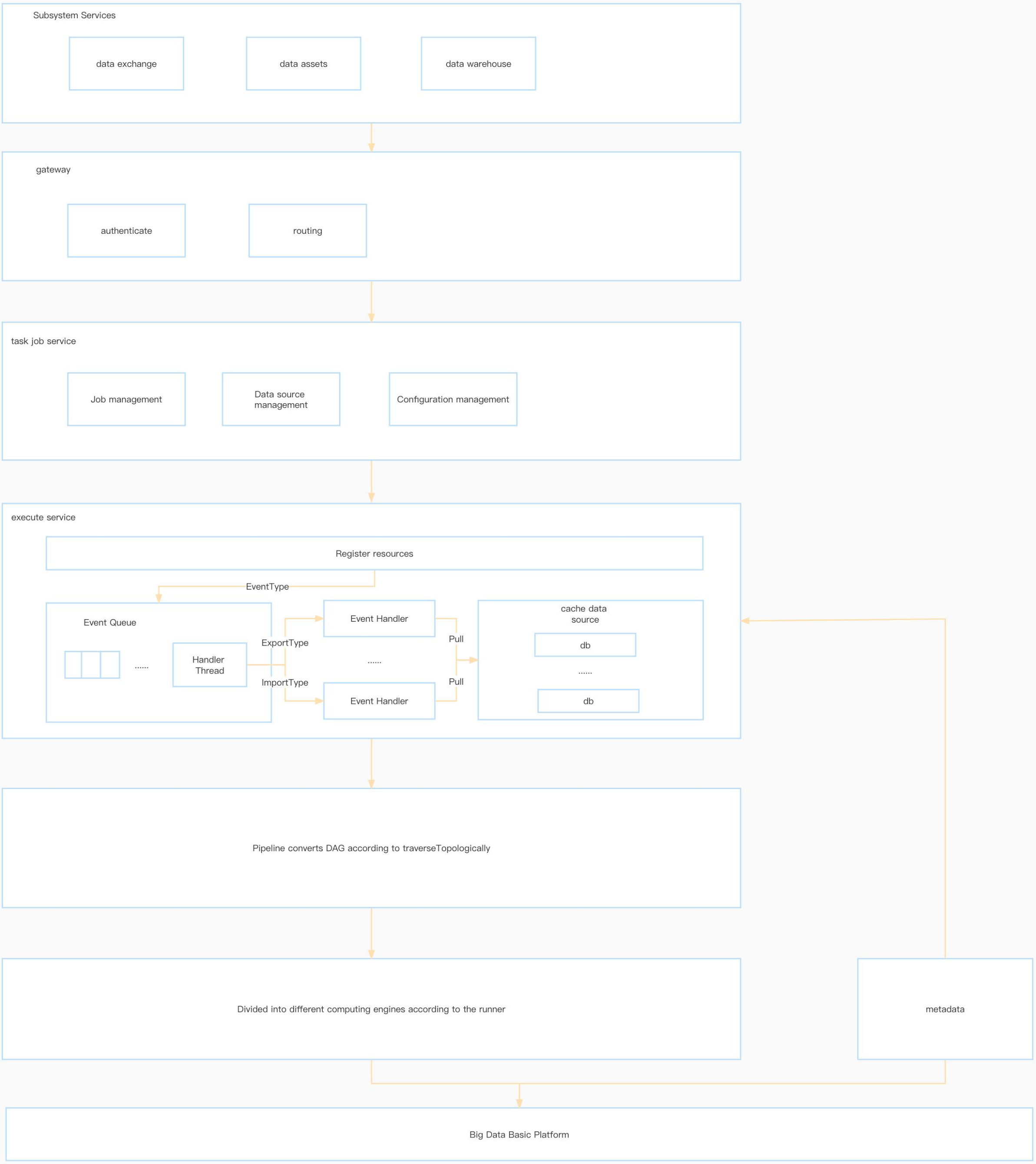
The basic composition of the original version of the big data platform framework:

Beam data processing and Hive Hook as metadata

# Background

shortcoming

- external access single
- high development cost
- metadata access is complex



Subsystem Services

data exchange

data assets

data warehouse

gateway

authenticate

routing

task job service

Job management

Data source management

Configuration management

execute service

Register resources

EventType

Event Queue

......

Handler Thread

ExportType

ImportType

Event Handler

......

Event Handler

Pull

Pull

cache data source

db

......

db

Pipeline converts DAG according to traverseTopologically

Divided into different computing engines according to the runner

metadata

Big Data Basic Platform

# Selection

Apache Livy

Apache Zeppelin

Netflix Geine

openLooKeng

Apache Linkis

# Selection

**Linkis builds a decoupling computing middleware layer with the ability to connect, expand, manage, orchestrate and reuse**

- Computing Governance Services
- Public Enhancement Services
- Microservice Governance Service
- Complete component base
- Community

# Scenario and Value

DataSphereStudio(DSS)

Linkis

Scriptis

DolphinScheduler

Exchangis

Streamis

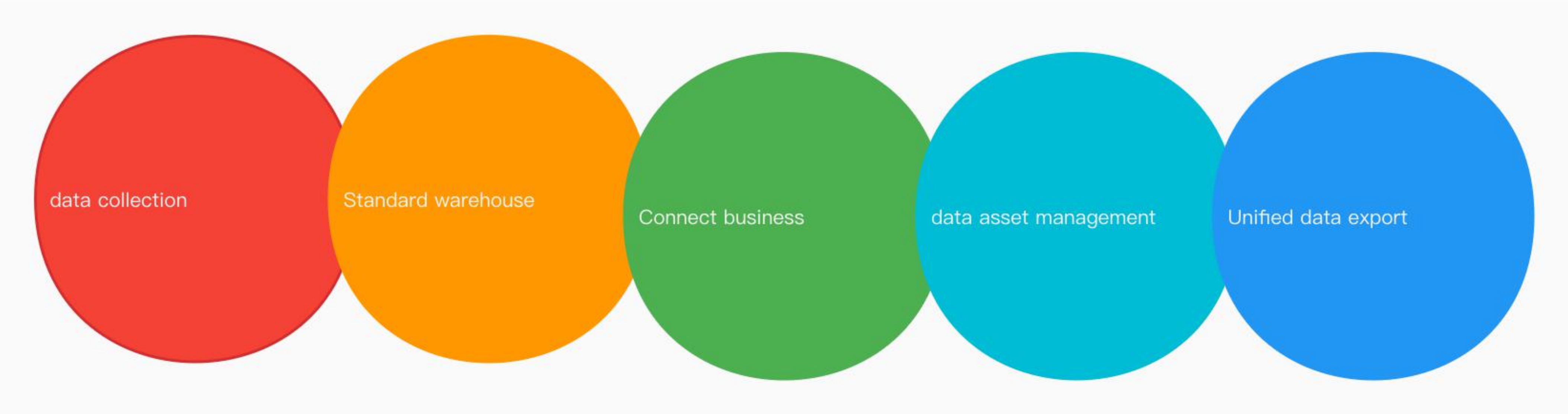
……

# Scenario and Value

## Provide the basis for data processing

- data assets
- data model
- data development
- data scheduling

# Scenario and Value



data collection

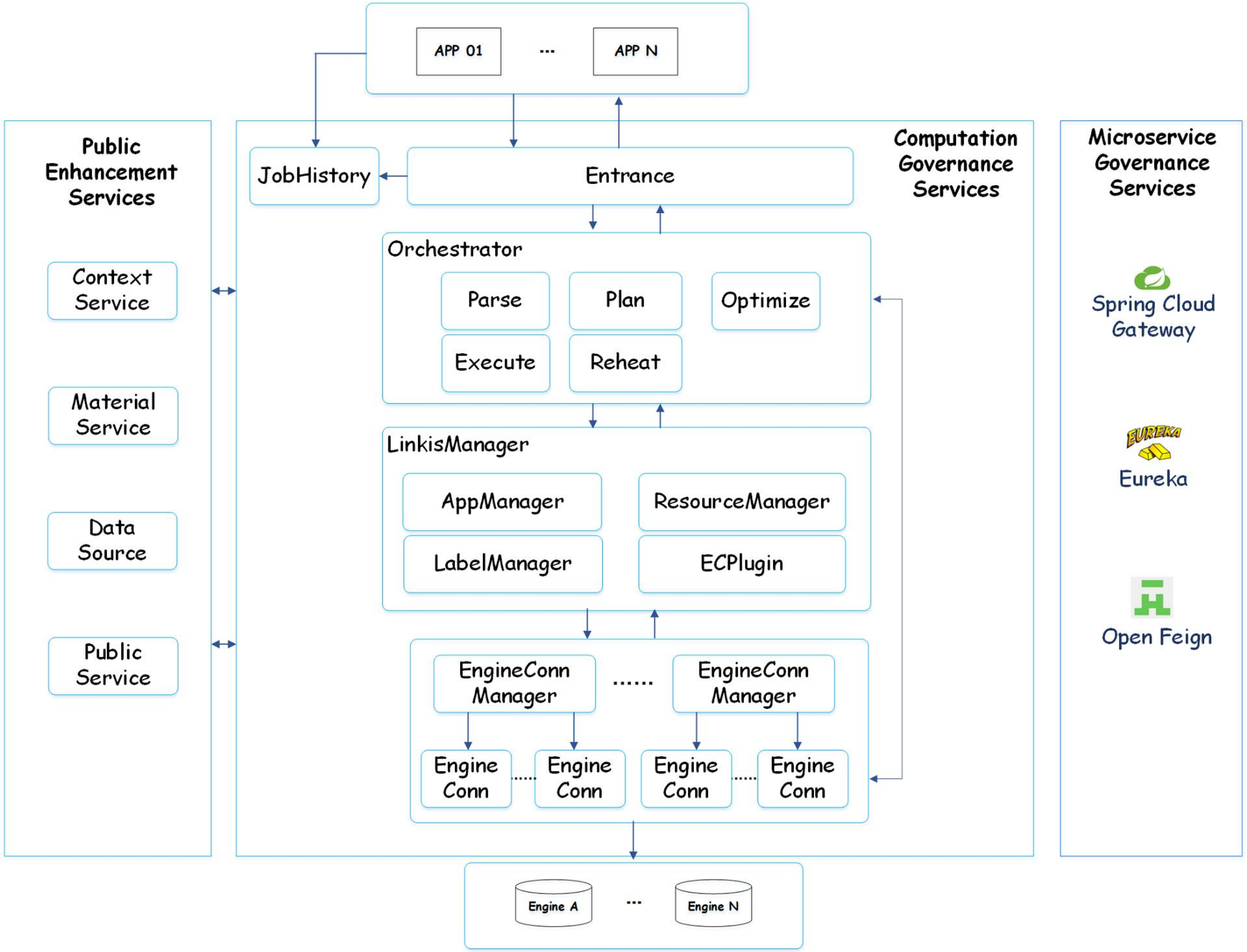
Standard warehouse

Connect business

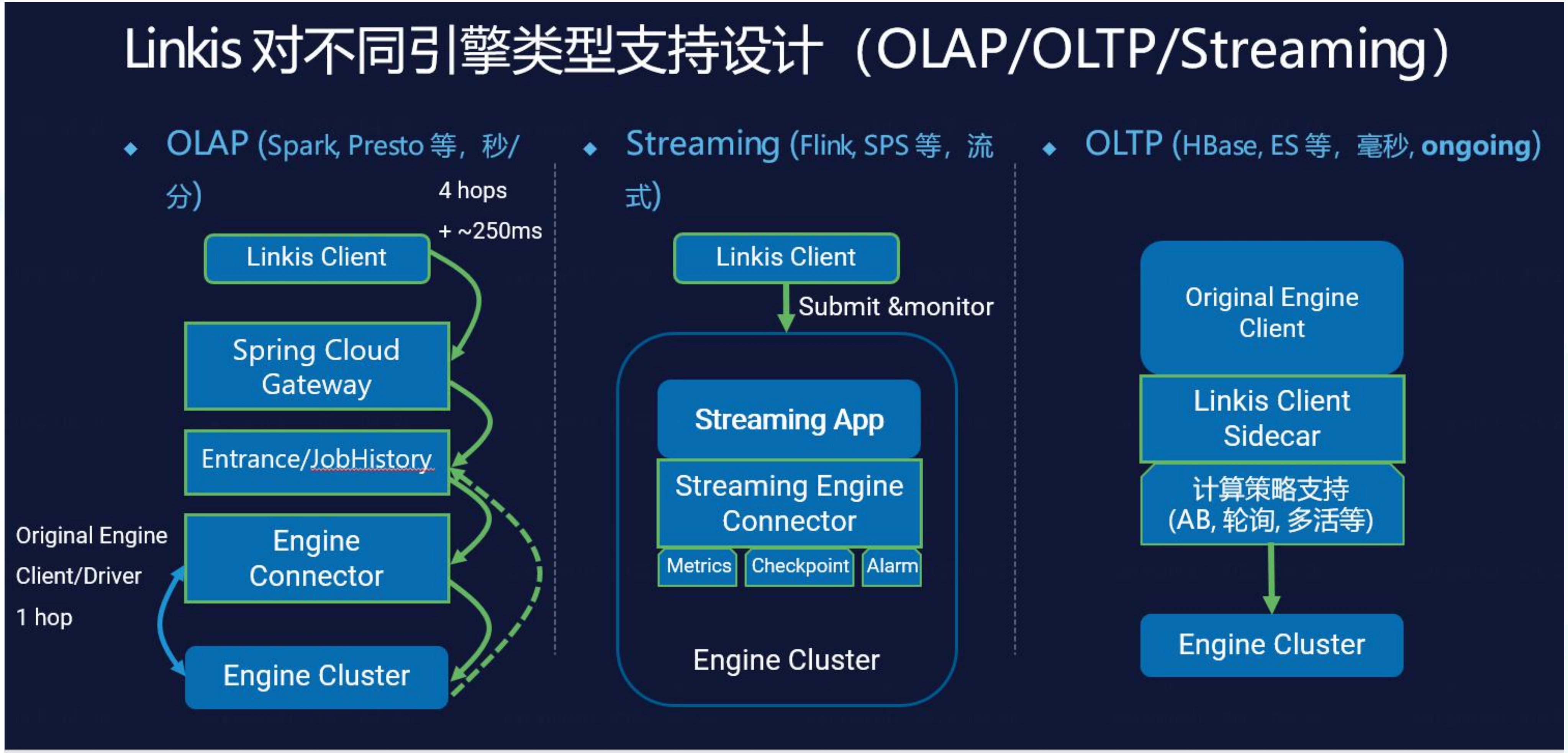
data asset management

Unified data export

# Architecture-Linkis



# Architecture-Linkis



## Linkis 对不同引擎类型支持设计（OLAP/OLTP/Streaming）

- OLAP (Spark, Presto 等，秒/分)
- Streaming (Flink, SPS 等，流式)
- OLTP (HBase, ES 等，毫秒, **ongoing**)

4 hops

+ ~250ms

Linkis Client

Spring Cloud Gateway

Entrance/JobHistory

Original Engine Client/Driver

1 hop

Engine Connector

Engine Cluster

Linkis Client

Submit &monitor

Streaming App

Streaming Engine Connector

Metrics

Checkpoint

Alarm

Engine Cluster

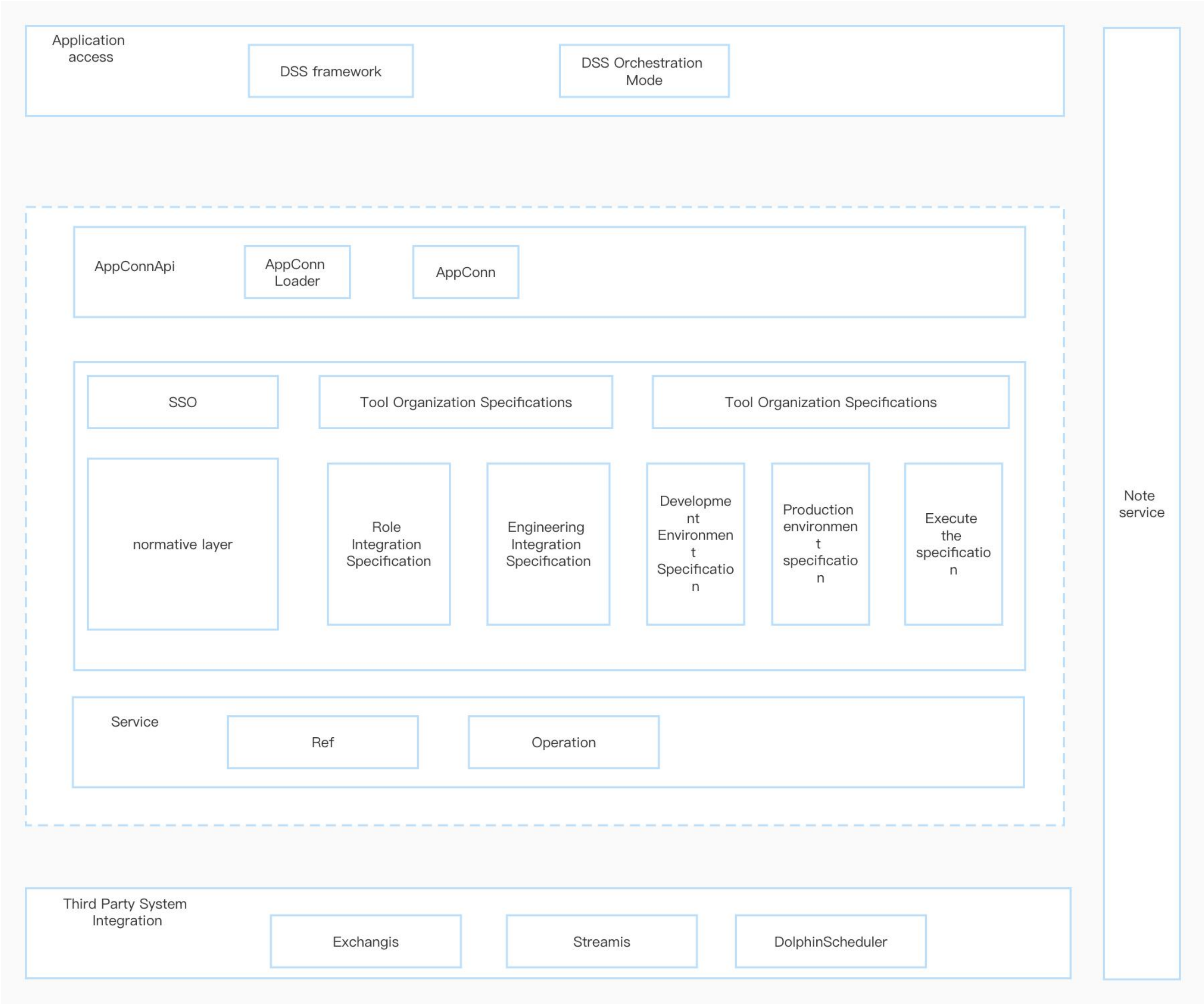
Original Engine Client

Linkis Client Sidecar

计算策略支持 (AB, 轮询, 多活等)

Engine Cluster

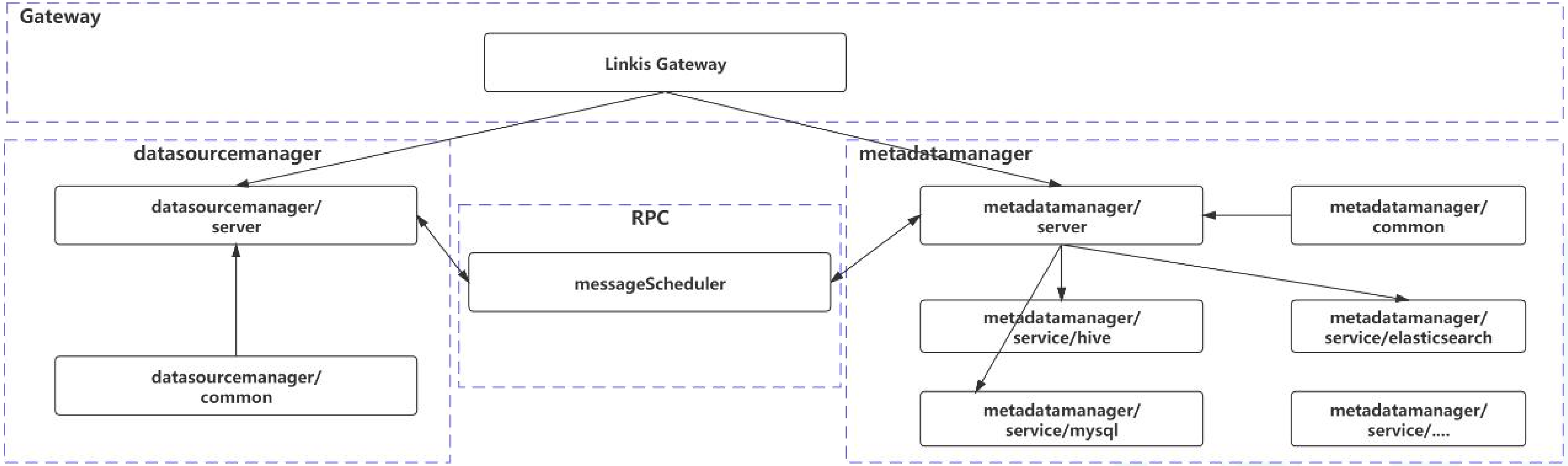
# Architecture-DSS


Application access
DSS framework
DSS Orchestration Mode
AppConnApi
AppConn Loader
AppConn
SSO
Tool Organization Specifications
Tool Organization Specifications
normative layer
Role Integration Specification
Engineering Integration Specification
Development Environment Specification
Production environment specification
Execute the specification
Service
Ref
Operation
Third Party System Integration
Exchangis
Streamis
DolphinScheduler
Note service

# Data Governance-DataSource

- Basic functions of a data source service provider
- The data source service provides some basic information about the data source environment
- Supports parameter dynamic lookup library table generation for multiple data source types
- Supports data sources to be divided according to user, creation source, and creation system
- Support for different data source connection tests
- A metadata query service that supports a certain data source

# Data Governance-DataSource


Gateway
Linkis Gateway
datasourcemanager
datasourcemanager/ server
datasourcemanager/ common
RPC
messageScheduler
metadatamanager
metadatamanager/ server
metadatamanager/ common
metadatamanager/ service/hive
metadatamanager/ service/elasticsearch
metadatamanager/ service/mysql
metadatamanager/ service/....

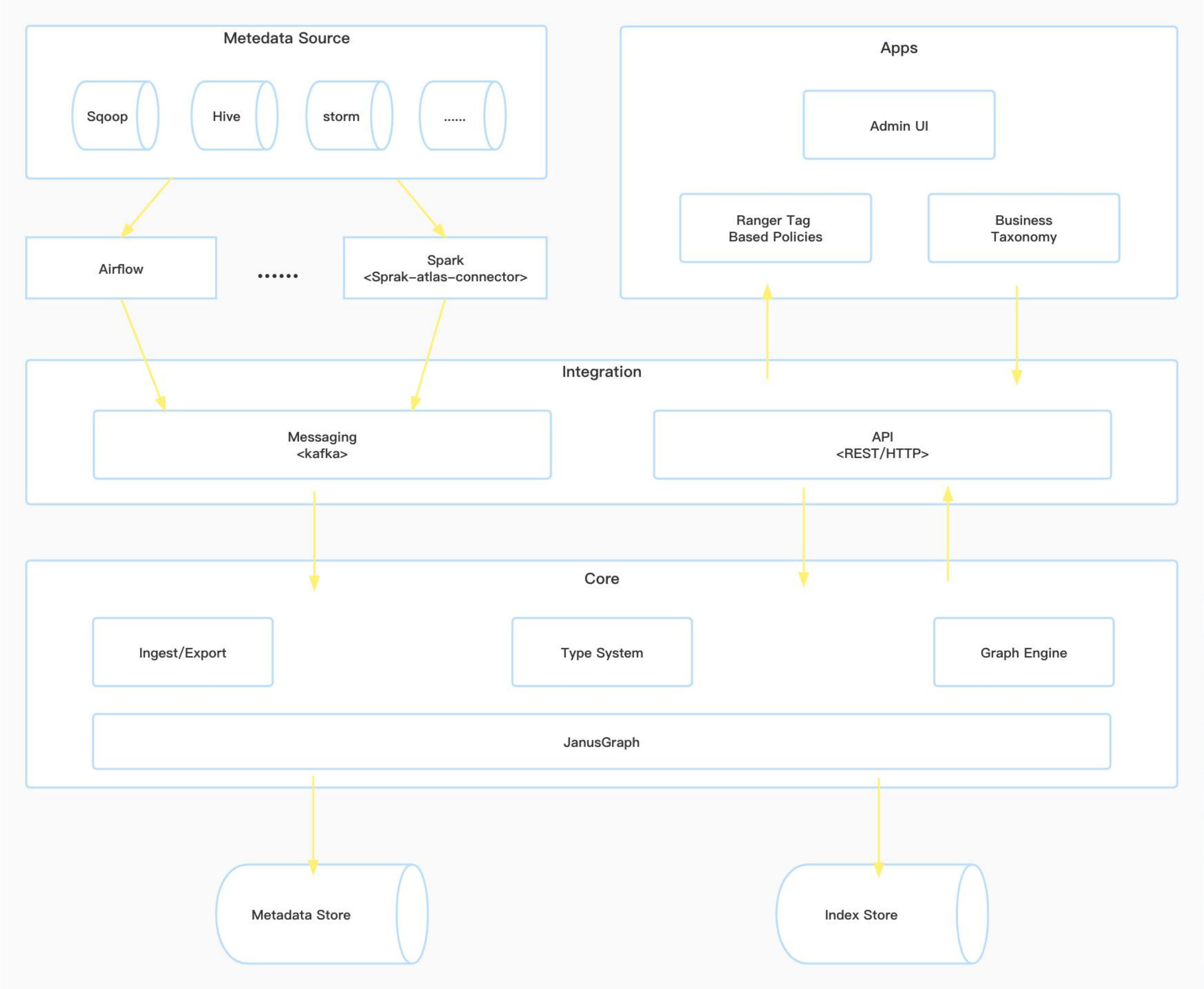
# Data Governance-Apache Atlas

Atlas is a data governance and metadata framework tightly coupled to the Hadoop ecosystem.

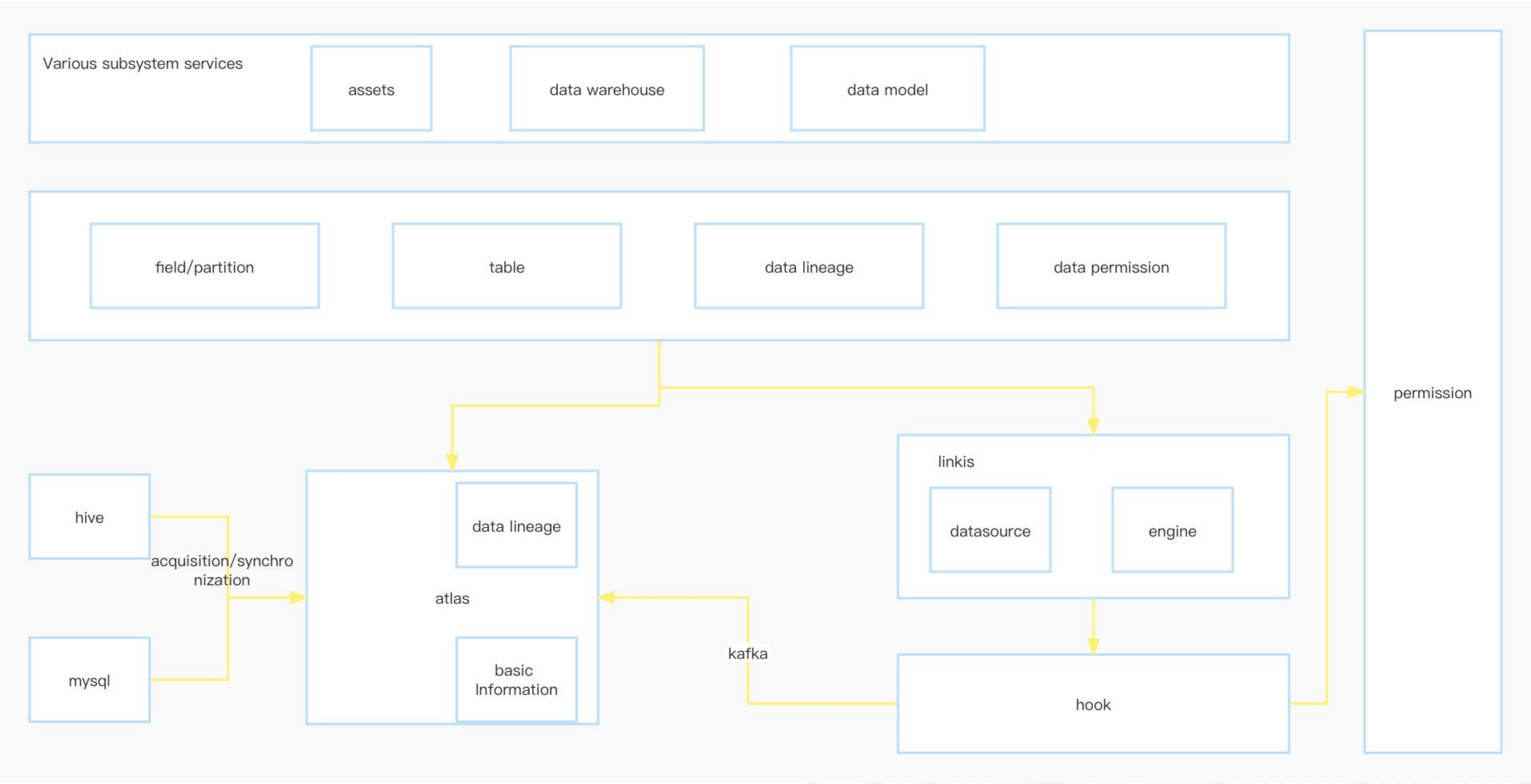
# Data Governance-Apache Atlas

- type system
- graphics engine
- capture/export
- api
- messaging
- metadata sources
- atlas admin ui
- tag based policies

# Data Governance-Apache Atlas



# Data Governance-Apache Atlas

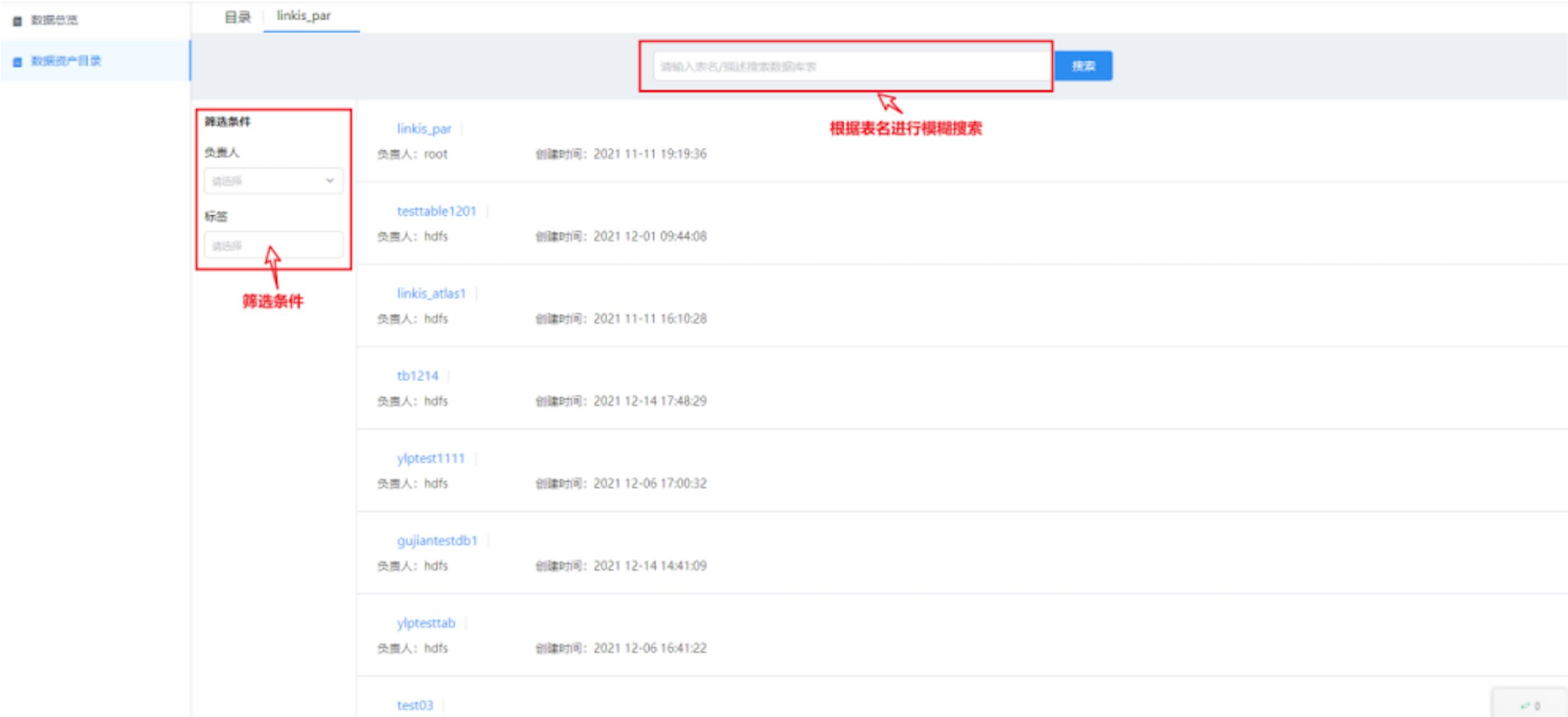


# Data Governance-Assets

Data asset management to make data traceable, usable, and trusted

- asset overview
- asset catalog

# Data Governance-Assets



根据表名进行模糊搜索

筛选条件

linkis_par
负责人：root 创建时间：2021 11-11 19:19:36

testtable1201
负责人：hdfs 创建时间：2021 12-01 09:44:08

linkis_atlas1
负责人：hdfs 创建时间：2021 11-11 16:10:28

tb1214
负责人：hdfs 创建时间：2021 12-14 17:48:29

ylptest1111
负责人：hdfs 创建时间：2021 12-06 17:00:32

gujiantestdb1
负责人：hdfs 创建时间：2021 12-14 14:41:09

ylptesttab
负责人：hdfs 创建时间：2021 12-06 16:41:22

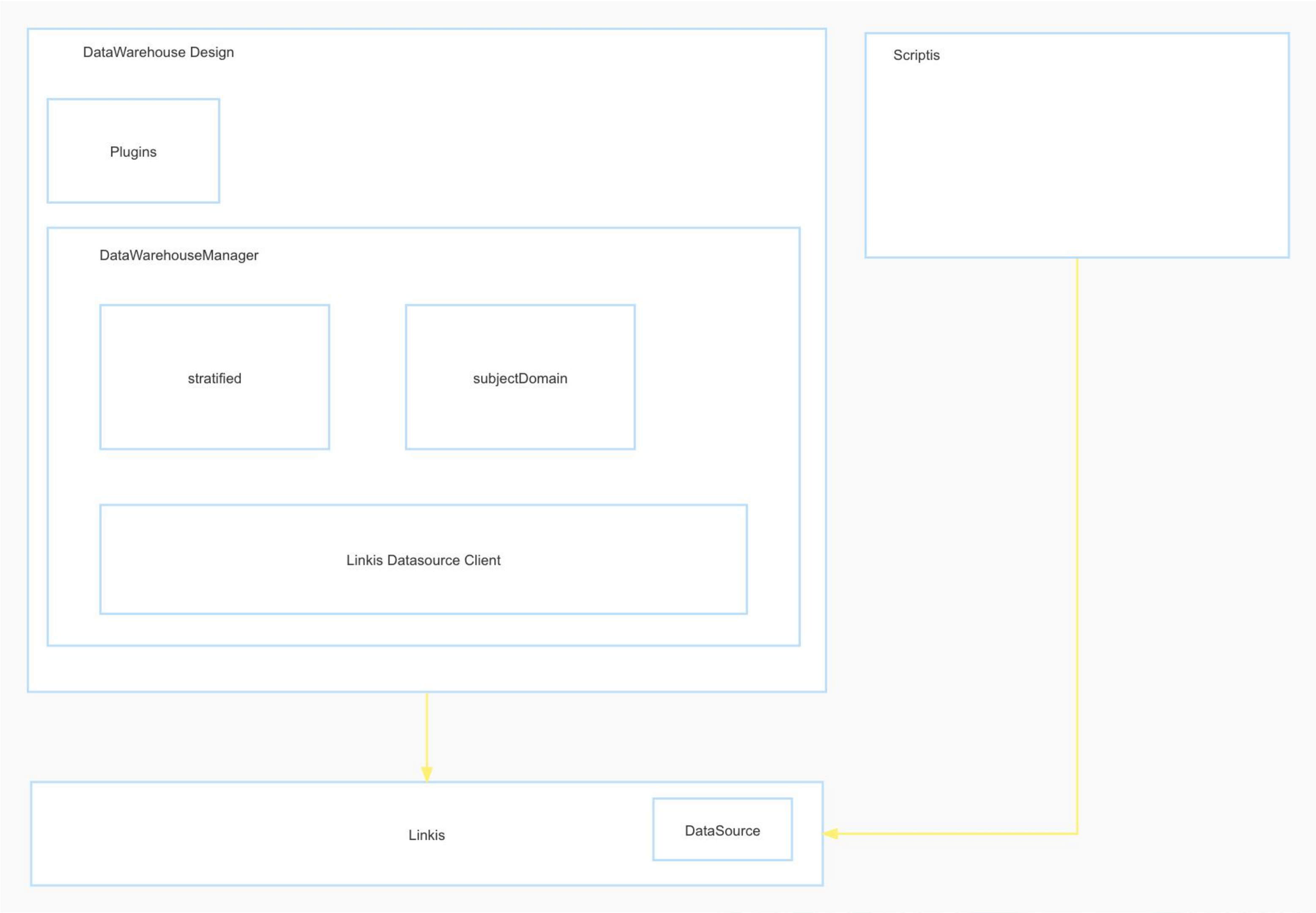
test03

# Data Governance-DataWarehouse

**Data warehouse standard specification management.**

- Subject Domain Management
- Data warehouse management
- Modifier management
- Statistical cycle management

# Data Governance-DataWarehouse



# Data Governance-DataModel

The content described by the data model has three parts, namely data operations, data constraints and data structures.

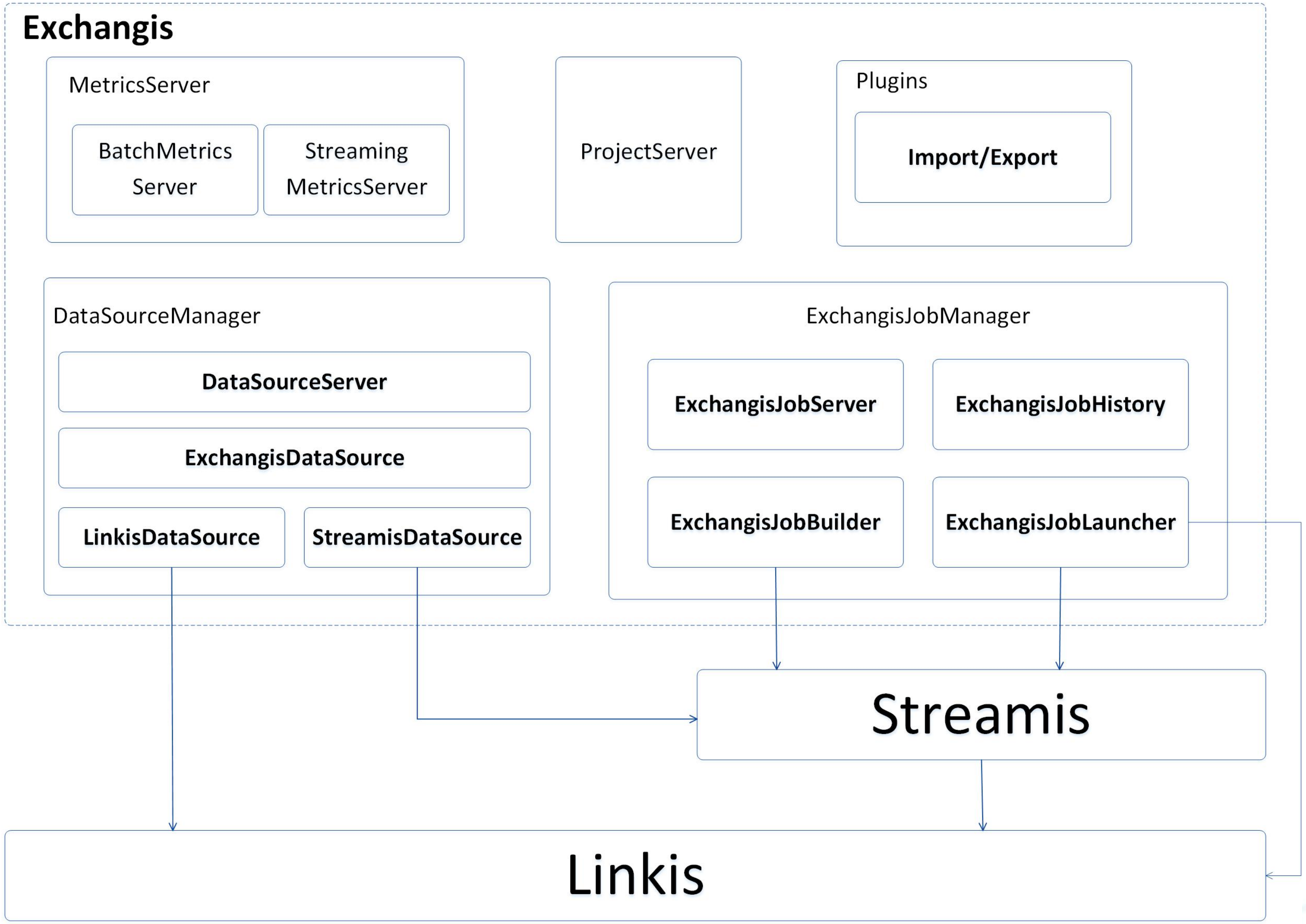
- table management
- dimension management
- metric management
- indicator management
- tag management

# Data Processing-Exchangis

A data exchange tool that supports the synchronization of structured and unstructured data transfers between heterogeneous data sources.

- Lightweight data source management
- High stability, fast response data synchronization task execution
- Open up with DSS workflow, one-stop big data development portal

# Data Processing-Exchangis



# Data Processing-Dolphinscheduler

Apache DolphinScheduler is a distributed and easily extensible visual DAG workflow task scheduling open source system.

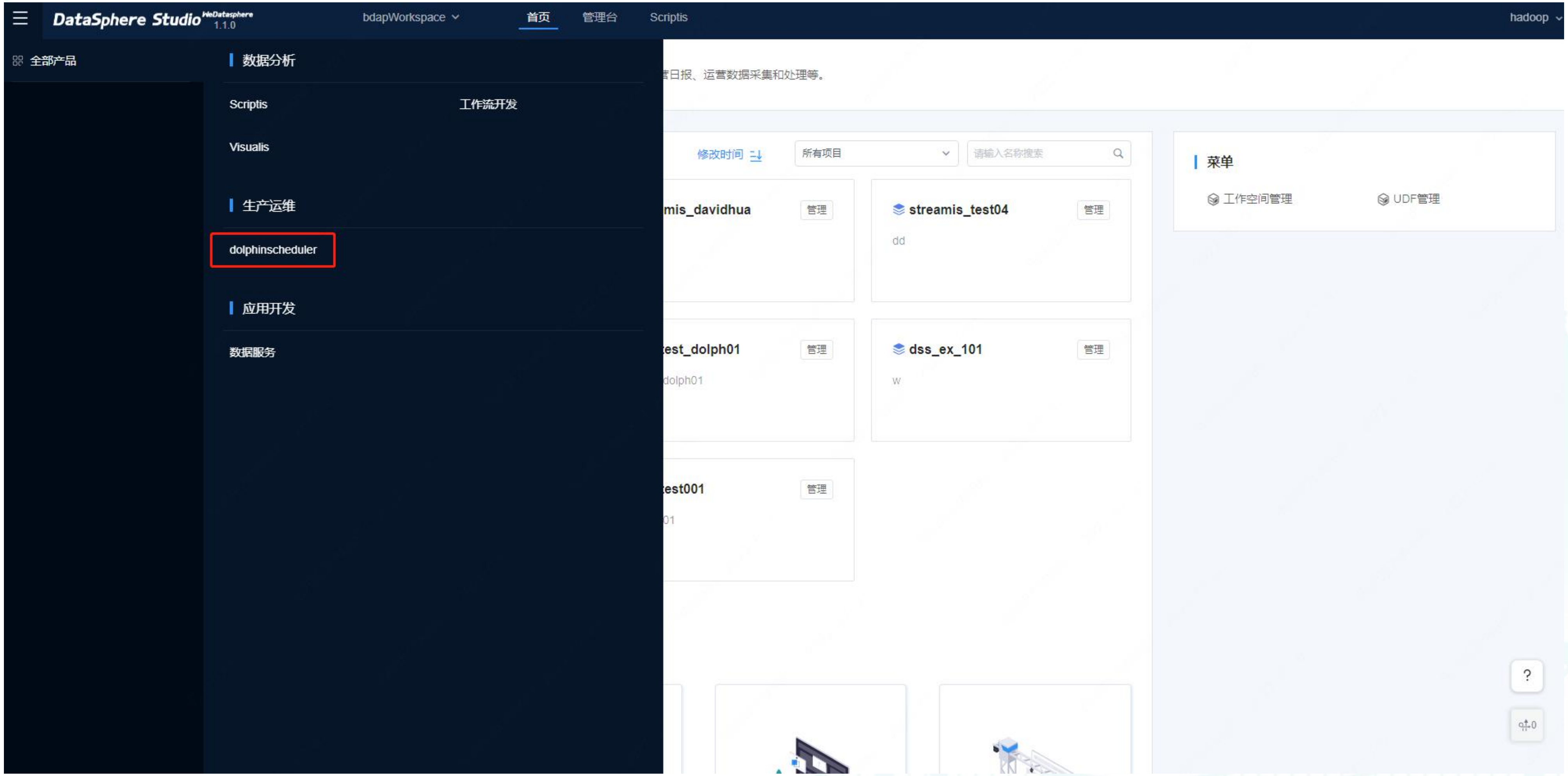
# Data Processing-Dolphinscheduler

- AppConn is the core concept of DSS that can easily and quickly integrate various upper-layer web systems.
- Dolphinscheduler can publish integration components to the scheduler with DSS publishing function

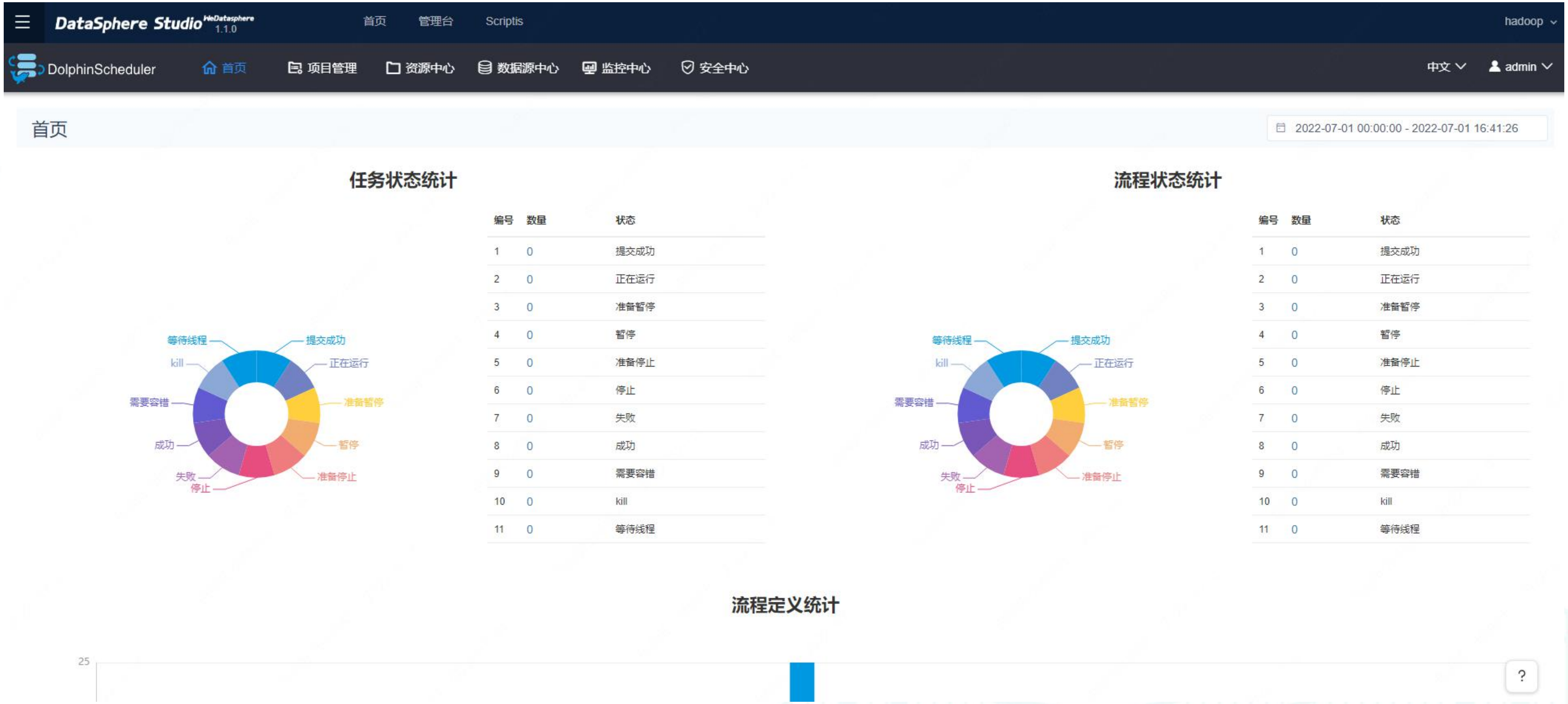
# Data Processing-Dolphinscheduler

DolphinScheduler can normally schedule the workflow node jobs of DataSphere Studio. You also need to install the dss-dolphinscheduler-client plugin, which is used to execute DSS workflow node jobs.

# Data Processing-Dolphinscheduler



# Data Processing-Dolphinscheduler



首页

2022-07-01 00:00:00 - 2022-07-01 16:41:26

任务状态统计

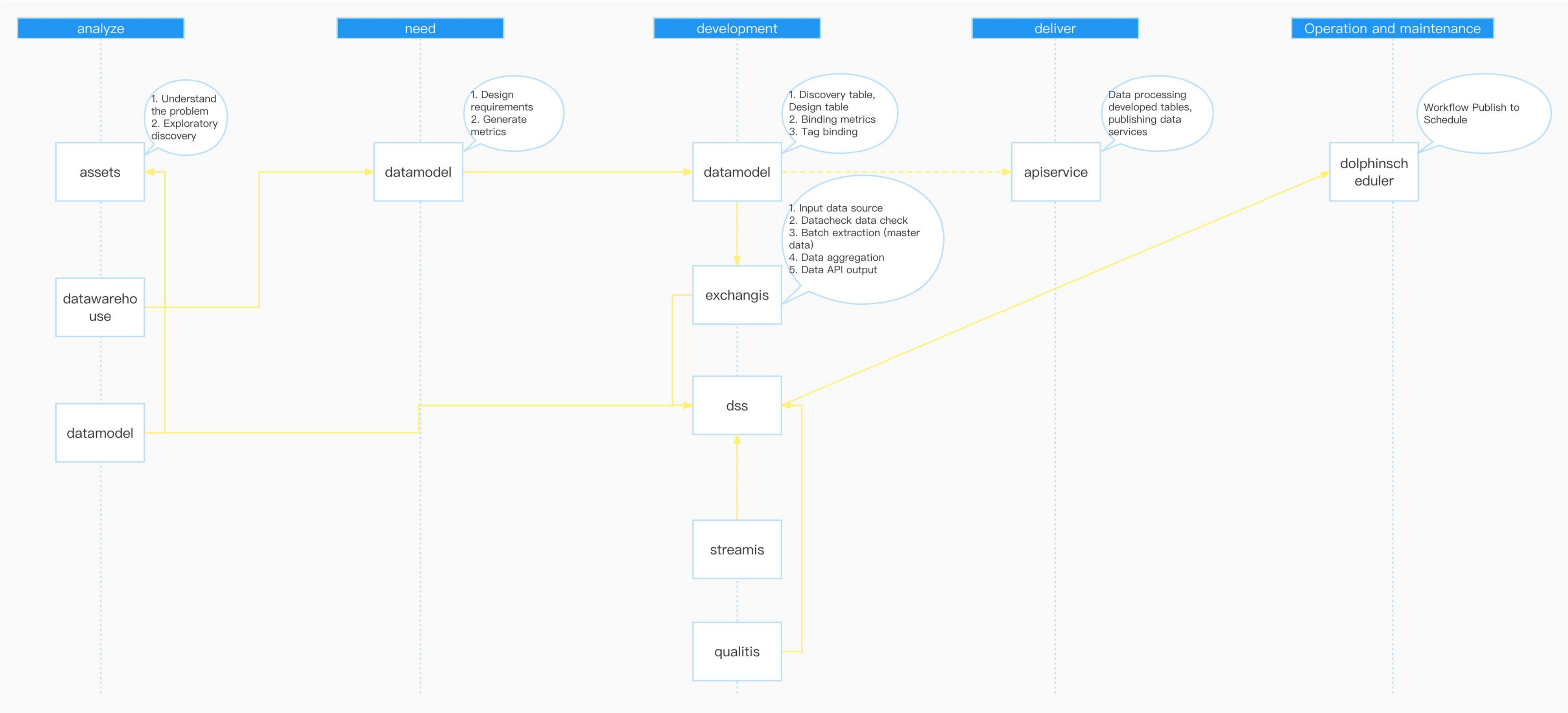
| 编号 | 数量 | 状态 |
|---|---|---|
| 1 | 0 | 提交成功 |
| 2 | 0 | 正在运行 |
| 3 | 0 | 准备暂停 |
| 4 | 0 | 暂停 |
| 5 | 0 | 准备停止 |
| 6 | 0 | 停止 |
| 7 | 0 | 失败 |
| 8 | 0 | 成功 |
| 9 | 0 | 需要容错 |
| 10 | 0 | kill |
| 11 | 0 | 等待线程 |

流程状态统计

| 编号 | 数量 | 状态 |
|---|---|---|
| 1 | 0 | 提交成功 |
| 2 | 0 | 正在运行 |
| 3 | 0 | 准备暂停 |
| 4 | 0 | 暂停 |
| 5 | 0 | 准备停止 |
| 6 | 0 | 停止 |
| 7 | 0 | 失败 |
| 8 | 0 | 成功 |
| 9 | 0 | 需要容错 |
| 10 | 0 | kill |
| 11 | 0 | 等待线程 |

流程定义统计

# Data Link


analyze
need
development
deliver
Operation and maintenance
1. Understand the problem
2. Exploratory discovery
assets
datawarehouse
datamodel
1. Design requirements
2. Generate metrics
datamodel
1. Discovery table, Design table
2. Binding metrics
3. Tag binding
datamodel
1. Input data source
2. Datacheck data check
3. Batch extraction (master data)
4. Data aggregation
5. Data API output
exchangis
dss
streamis
qualitis
Data processing developed tables, publishing data services
apiservice
Workflow Publish to Schedule
dolphinscheduler

# Future Outlook

- Improvement of data integration capabilities
- Data governance related service release
- WeDataSphere community-related components continue to improve and strive for Apache

# Open Source

All the components shared this time have been open sourced in the community. I look forward to more friends to join and communicate with each other so that better products can be iterated. I always believe that the open source community is the soul of the product, and it is connected with its pulse to make Products have better development.

Over the years, I have met many friends in the WeDataSphere community, learned a lot, and gave back the results to the community. I will continue to work hard and continue to walk with you.

# THANK YOU

云数智聚　砥柱笃行