

DORIS
SUMMIT



利用 Apache SeaTunnel 批流加载至 Apache Doris

郭炜 白鲸开源 CEO

目录

1. Apache SeaTunnel 介绍
2. SeaTunnel+Doris 快速搭建批流一体数据仓库
3. Apache SeaTunnel 未来 Roadmap 介绍

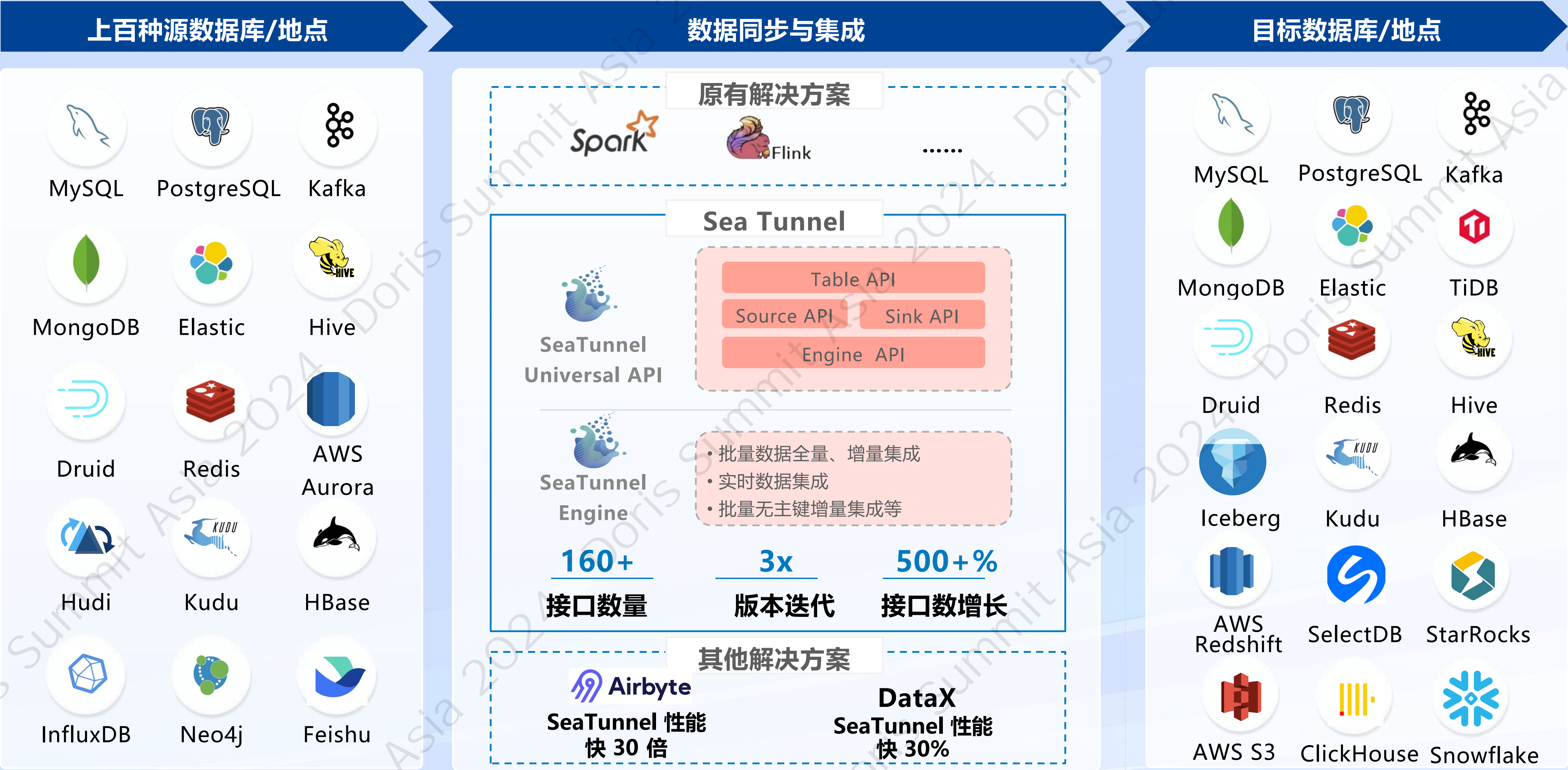
01

Apache SeaTunnel 介绍

160+ 数据源批流一体的数据集成工具

Apache SeaTunnel：新一代实时多源数据同步工具——大数据高速公路

Github Star: 8.1k



生态工具 Connectors: 目前支持超过 160+ Connectors

1	Source Type	DataSource	Type	1	Source Type	DataSource	Type	1	Source Type	DataSource	Type	1	Source Type	DataSource	Type
2	Cloud	AmazonDynamoDB	Sink	40	Database	Neo4j	Sink	78	FileSystem	Gitlab	Source	116	Database	Phoenix	Source
3	Cloud	AWS Aurora	Source	41	Database	Oracle	Sink	79	FileSystem	GoogleSheets	Source	117	Database	PostgreSql	Source
4	System	Assert	Sink	42	Cloud	OssFile	Sink	80	Database	DB2/IBM	Source	118	MQ	Pulsar	Source
5	Cloud	AWS Aurora	Sink	43	Database	Phoenix	Sink	81	Database	Doris	Source	119	MQ	RabbitMQ	Source
6	Cloud	AWS RDS	Sink	44	FileSystem	OssJindoFile	Sink	82	System	Fake	Source	120	Catch	Redis	Source
7	Cloud	AWS RDS	Source	45	Database	Postgresql	Sink	83	Database	Gbase 8a	Source	121	Database	Redshift	Source
8	Database	Cassandra	Sink	46	MQ	RabbitMQ	Sink	84	Database	Greenplum	Source	122	Cloud	S3File	Source
9	Database	ClickHouse	Sink	47	Catch	Redis	Sink	85	FileSystem	HdfsFile	Source	123	FileSystem	SftpFile	Source
10	Database	ClickHouseFile	Sink	48	Database	Redshift	Sink	86	Database	Hive	Source	124	SaaS	Slack	Source
11	System	Console	Sink	49	Cloud	S3File	Sink	87	System	Http	Source	125	System	Socket	Source
12	Cloud	DataHub	Sink	50	Cloud	S3Redshift	Sink	88	Database	Hudi	Source	126	Database	Sqlite	Source
13	Database	DB2	Sink	51	FileSystem	S3File	Sink	89	Database	Apache Iceberg	Source	127	Database	StarRocks	Source
14	SaaS	DingTalk	Sink	52	Cloud	SelectDB Cloud	Sink	90	Database	InfluxDB	Source	128	Cloud	Tablestore	Source
15	Database	Doris	Sink	53	SaaS	Sentry	Sink	91	Database	IoTDB	Source	129	Database	Teradata	Source
16	Database	Elasticsearch	Sink	54	FileSystem	SftpFile	Sink	92	Database	JDBC	Source	130	Database	TDengine	Source
17	System	Email	Sink	55	System	Socket	Sink	93	FileSystem	jira	Source	131	MQ	Apache Pulsar	Source
18	SaaS	Enterprise WeChat	Sink	56	Database	Sqlite	Sink	94	FileSystem	Klaviyo	Source	132	Database	PolarDB	Source
19	SaaS	FeiShu	Sink	57	Database	SqlServer	Sink	95	MQ	Kafka	Source	133	Database	SQL Server	Source
20	FileSystem	FtpFile	Sink	58	Database	StarRocks	Sink	96	Database	Kudu	Source	134	Database	TiDB	Source
21	Database	Gbase 8a	Sink	59	Database	TDengine	Sink	97	SaaS	Lemlist	Source	135	System	Source Common Options	Source
22	Database	Greenplum	Sink	60	Cloud	Tablestore	Sink	98	System	LocalFile	Source	136	Database	Informix	Source
23	FileSystem	HdfsFile	Sink	61	Database	Teradata	Sink	99	Cloud	Maxcompute	Source	137	Database	MongoDB	Source
24	Database	Hive	Sink	62	System	Sink Common Options	Sink	100	Database	MongoDB	Source	138	Database	TD-SQL-MySQL	Source
25	System	Http	Sink	63	Database	PolarDB	Sink	101	SaaS	MyHours	Source	139	Database	TD-SQL-MySQL	Source
26	Database	InfluxDB	Sink	64	Database	SQL Server	Sink	102	Database	MySQL	Source				
27	Database	IoTDB	Sink	65	Database	TiDB	Sink	103	Database	MySQLCDC	Source				
28	Cloud	Snowflake	Sink	66	Database	MongoDB	Sink	104	Database	Neo4j	Source				
29	Database	JDBC	Sink	67	Database	Iceberg	Sink	105	Database	Paimon	Sink				
30	MQ	Kafka	Sink	68	Database	Informix	Sink	106	Cloud	GoogleFirestore	Sink				
31	Database	Kudu	Sink	69	Database	TD-SQL-MySQL	Sink	107	Cloud	Snowflake	Source				
32	System	LocalFile	Sink	70	Database	TD-SQL-MySQL	Sink	108	Database CDC	MySQL CDC	Source				
33	Database	SAPHana	Sink	71	Cloud	AmazonDynamoDB	Source	109	SaaS	Notion	Source				
34	Database	SAPHana	Source	72	Database	Cassandra	Source	110	SaaS	OneSignal	Source				
35	Database	Vertica	Sink	73	Database	ClickHouse	Source	111	Database	OpenMldb	Source				
36	Database	Vertica	Source	74	Database	Elasticsearch	Source	112	Database	Oracle	Source				
37	Cloud	Maxcompute	Sink	75	Database	FakeSource	Source	113	Cloud	OssFile	Source				
38	Database	MongoDB	Sink	76	FileSystem	FtpFile	Source	114	Cloud	OssJindoFile	Source				
39	Database	MySQL	Sink	77	FileSystem	Github	Source	115	Database	PersistIQ	Source				

Search **"SeaTunnel connector"** on Google or seatunnel.apache.org

Apache SeaTunnel 典型案例

跨云数据准备

JPMorgan & Chase

美国最大商业银行

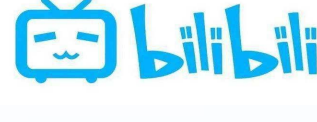
解决多云异构环境下，需要异构数据打通，将 AWS Aruora，DynamoDB，SFTP 数据实时同步到 ES，S3，Snowflake 下

异构数据实时数据同步



超大型客户

解决多数据源数据每日出入数据库以及每日出入仓同步数据问题，数据集群规模**30+台**，日均记录数量级上**千亿**，日均数据量在 **100TB** 以上。



相关资源



官网: <https://seatunnel.apache.org>



GitHub : <https://github.com/apache/incubator-seatunnel>



Slack: <https://apacheseatunnel.slack.com>



X.com : <https://x.com/ASFSeaTunnel>



B 站 : <https://space.bilibili.com/1542095008>

02

SeaTunnel+Doris 快速搭建批流一体数据仓库

全面取代 Lambda 架构的批流一体

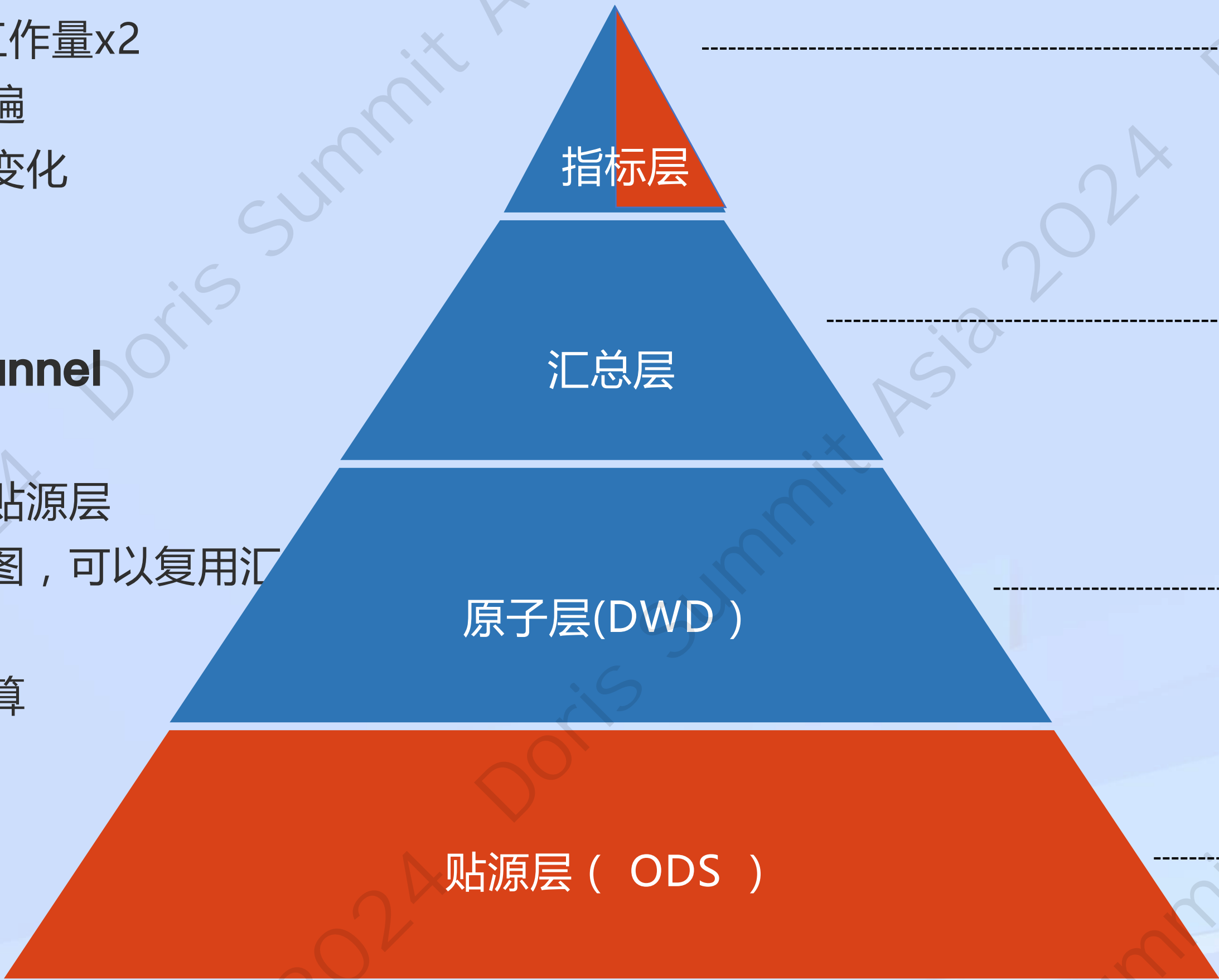
Apache Doris+Apache SeaTunnel 全面淘汰Lambda架构

传统Lambda痛点

- 批和流开发2套代码，工作量x2
- 流跑完，批还要覆盖一遍
- 当日实时数据第二天会变化
- 流+批 两套集群

Apache Doris+SeaTunnel

- 贴源层采用全实时架构
- 原子层数据也来自实时贴源层
- 指标层采用实时物化视图，可以复用汇总层数据
- 实时指标不需要二次计算



指标层，半流半批：

- 流用于实时数据，由于数据物化 SQL 和数据源一样，数据不会存在差异
- 批数据和以前一样，只对需要实时数据进行物化即可

汇总层，批量运行：

- 批量数据复杂处理运行效率更高
- 数据量比较大，大宽表支持上层业务
- 调度支持批量运行与跨层次依赖

原子层，批量运行：

- 批量数据复杂处理运行效率更高
- 模型设计更加规范
- 调度支持“流停批跑，批停流跑”

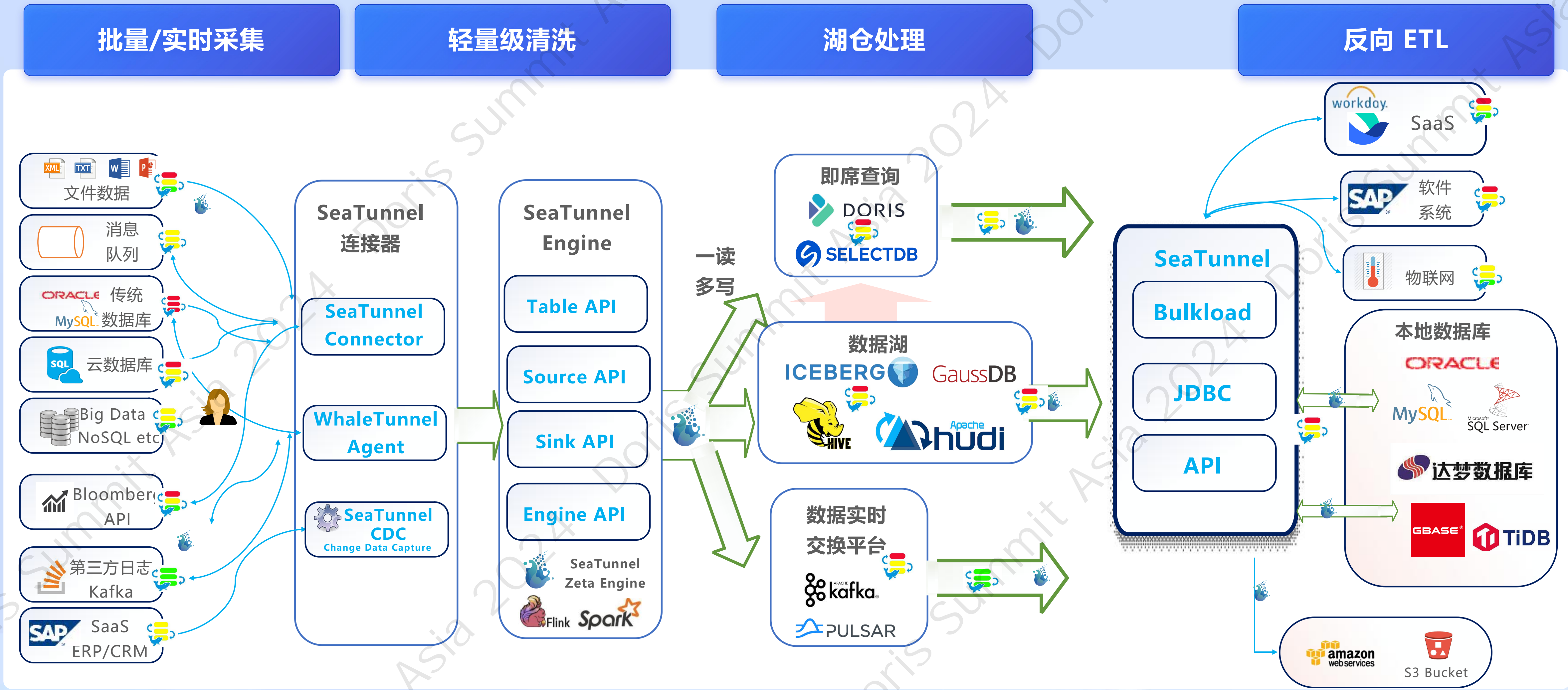
贴源层，实时接入：

- 历史数据+实时数据一个任务自动切换处理
- 单任务支持多表/加表/多连接
- 自动DDL变更（WhaleTunnel）

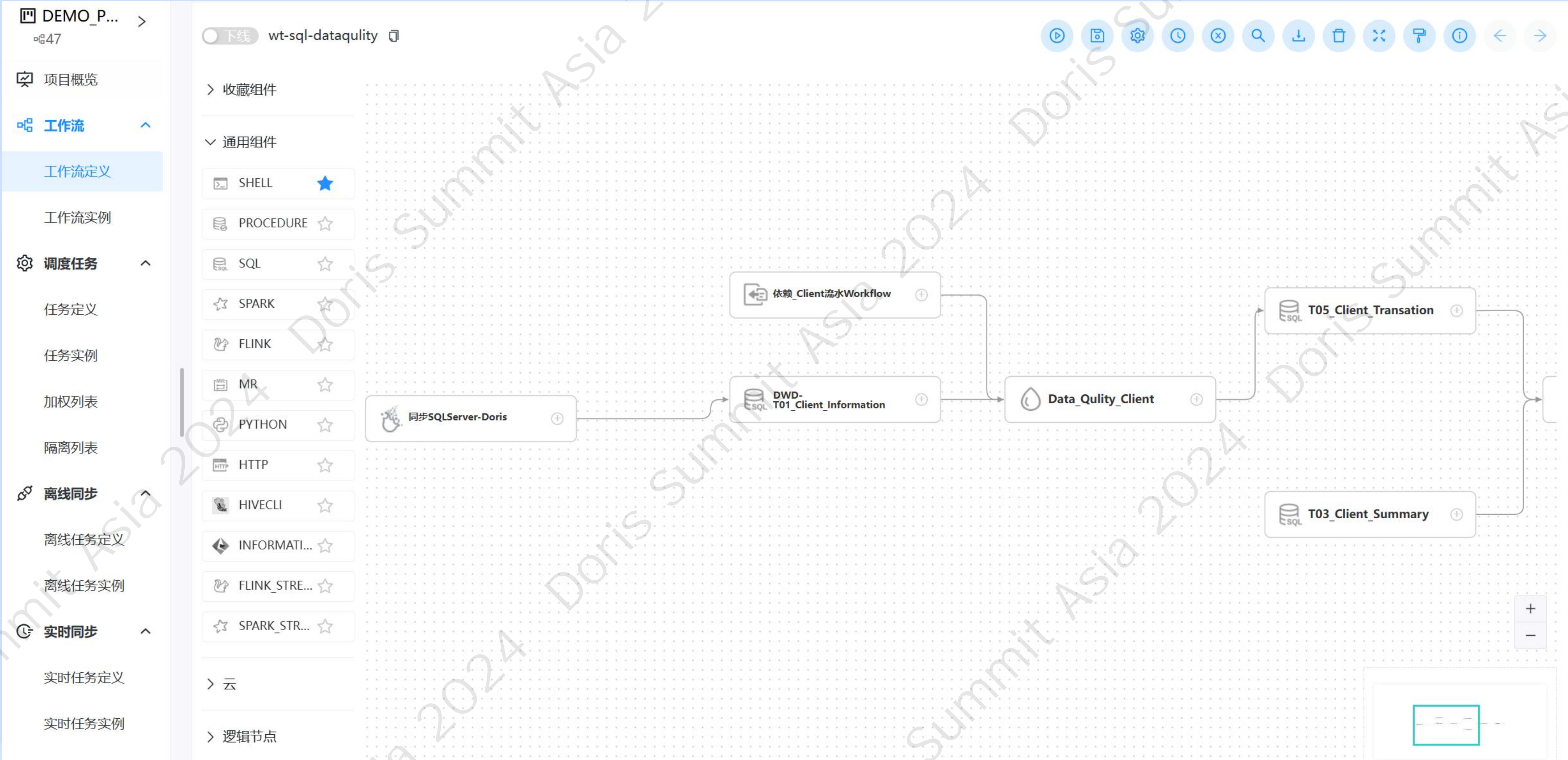
Apache Doris+SeaTunnel 常见的批流一体的数据仓库模型设计架构

实时数仓的最佳组合：Doris(SelectDB) x SeaTunnel(WhaleTunnel)

使用 Doris+SeaTunnel 可以实现多种数据源的实时获取，一读多写的模式实时写入仓库，也可以进入数据湖，通过 Doris 外表模式加载实现实时指标层场景。



Doris(SelectDB)xSeaTunnel(WhaleTunnel)xDolphinScheduler(WhaleScheduler)



一站式解决批 + 流的同步和数据开发问题

03

Apache SeaTunnel 未来 Roadmap 介绍

大模型+易用性+可监测性加强

Apache SeaTunnel 未来 Roadmap 介绍

更快、更好用

作为一个数据集成平台，SeaTunnel 将不断专注于解决数据集成领域的需求和问题。持续从数据源的数量、数据同步的性能和易用性上满足用户的需求。

连接器丰富

- 支持更多向量数据库
- ...



支持多表 Source + Sink SeaTunnel Web 的开发




CDC 支持 DDL 变更 流速控制



K8S+Yarn 支持



Apache SeaTunnel 对向量和大模型的支持

 Apache SeaTunnel

Quick Start - V2

Concepts

Connector-V2

Transform-V2

Transform Common Options

Copy

DynamicCompile

Embedding

FieldMapper

FilterRowKind

Filter

JsonPath

LLM

Replace

Split

SQL Functions

SQL UDF

SQL

HomeDocumentDownloadCommunityBlogUserCasesTeamUsersASFGitHubSecurityEnglish

Home>Transform-V2>Embedding

Version: 2.3.8

Embedding

Embedding Transform Plugin

Description

The Embedding transform plugin leverages embedding models to convert text data into vectorized representations. This transformation can be applied to various fields. The plugin supports multiple model providers and can be integrated with different API endpoints.

Options

Name	Type	Required	Default Value	Description
model_provider	enum	yes	-	The model provider for embedding. Options may include QIANFAN, OPENAI, etc.

基于 Apache SeaTunnel 的商业版：WhaleTunnel

- 简单易用，开箱即用，不依赖 HDFS，Flink，Spark 集群
- 全可视化操作，支持可视化运维与监控配置
- 支持信创，目前支持 198 种数据源
- 整库同步、表结构自动变更
- 与WhaleScheduler全面集成，完成传参和编排工作
- 根据调度日历、数据日期（牌）等参数进行传递和上下游触发

来源和目标

Source

Sink

处理与转换

单列复制多列

数据变更处理

字段删除

字段重命名

字段值替换

单列拆分为多列

Oracle CDC A...

数据变更处理

SelectDB-Ali

Source

设置

数据模型

节点名称 *

Oralce Agent

场景模式 *

多表同步

源名称 *

st02-oracle-agent

数据库 *

033_guangzhou

待同步表 *

待同步表 2/18

请输入

ddltest_sink

map

已选中表 0/3

请输入

alerttestsource_sink

alerttestsource_sinksink

重新加载表

WhaleOps

首页

项目管理

影响分析

资源中心

数据质量

源中心

监控中心

配置中心

全部项目

Test-2430854580

暂停

恢复

同步概览

运行日志

运行记录

告警列表

运行中断时间: -

已读取数据量 (行)

3,560,000

已写入数据量 (行)

3,560,000

Insert: n

Update: n

Delete: n

DDL: n

当前数据延迟 (秒)

1.03

100%

全量阶段同步速率

读取速率(行/秒): 33214

处理速率(行/秒): 33145

读取速率(MB/秒): 10

处理速率(MB/秒): 10

同步数据

数据读取 (行)

数据写入 (行)

实时增量

源表

目标表

Insert

Update

Delete

DML

DDL

上次同步时间

testwyr1

TESTWYR1

100,000

100,000

0

200,000

0

2024-1-1 17:33:04

testwyr2

TESTWYR2

100,000

100,000

18

200,018

5

2024-1-1 17:31:15

testwyr3

TESTWYR3

5,000

0

0

5,000

0

2024-1-1 17:30:44

任务基本信息

同步任务定义

同步任务定义名称

业务模型

实时同步

状态

暂停

执行用户

admin

开始时间

2024-1-1 15:22:22

结束时间

-

最近操作用户

admin

最近操作时间

-

任务进度

结构迁移

表3/3

全量初始化

表3/3

实时增量

已运行 2h

启动参数

运行类型

直接启动

优先级

medium

Worker分组

default

告警通知

失败告警

运维1组

实时增量延迟

运维1组

DDL事件通知

运维1组

全量同步完成通知

运维1组

白鲸开源-DolphinScheduler&SeaTunnel 核心开发者打造的 WhaleStudio

商业案例



中信建投证券 CHINA SECURITIES

- 中信建投是中国前三大券商之一。
- WhaleStudio满足用户数据一体化编辑、上线、数据管控、复杂时间管理等需求，充分提高中信建投数据研发效率。
- 平台应用于公司反洗钱、实时盈亏计算、监管报送、数据精算等多个核心应用，累计编排定义 workflows 超过3000个，上线任务数量接近16000个，交易日平均运行 workflows 实例数量超过5000个，日均任务执行任务数量超过20000个。
- 目前公司各业务线数据处理任务还在持续上线DataOps平台，整个平台规模还在持续增长中

其他案例



- 中国银行是中国前三大银行之一
- 在 WhaleStudio 上开发了超过 10,000 个大数据任务
- 整合了 10 个系统和数据库（包括 Oracle、Informix、MySQL 等）



- 中国人寿是中国前三大保险公司之一。
- WhaleStudio 在 8台服务器上运行了超过 100万个 SQL 任务。
- 超过 20 个部门和 36 家子公司正在使用 WhaleStudio 开发大数据作业，提高整体开发效率。

AWS客户



- WhaleStudio 帮助用户快速从 AWS 数据库、Oracle 数据库和 ERP 系统收集数据，并将其加载到 AWS Redshift 的 ODS 层。然后，使用 WhaleStudio 进行 SQL 开发和调试，他们可以完成汇总层和整个数据仓库的开发。
- WhaleOps 公司是 AWS 的重要技术合作伙伴。用户可以在 AWS Marketplace 上一键购买和部署与 WhaleStudio 相关的服务。
- 用户也可以在自己的数据中心部署 WhaleStudio，实现混合云和多云部署。

其他客户案例



谢谢观看
THANK YOU!



扫码加入 Apache SeaTunnel 社群