

百川入海 数据迁移至 Apache Doris 最佳实践

王华杰

飞轮科技 资深架构师

Apache StreamPark PPMC 成员

目录

1. 需求与痛点
2. 数据迁移连接器
3. X2Doris 百川入海
4. X2Doris 未来规划

1 需求与痛点

需求：单表/多表/整库数据高效迁移，增量迁移

准备工作	迁移过程	迁移结果
<ul style="list-style-type: none">• 创建Doris表：单表/多表/整库 创建 Doris 表• 选择工具：有工具？自己开发？	<ul style="list-style-type: none">• 可用：生产使用必备• 性能：高效迁移，支持分布式？	<ul style="list-style-type: none">• 准确性：保证准确性。

痛点

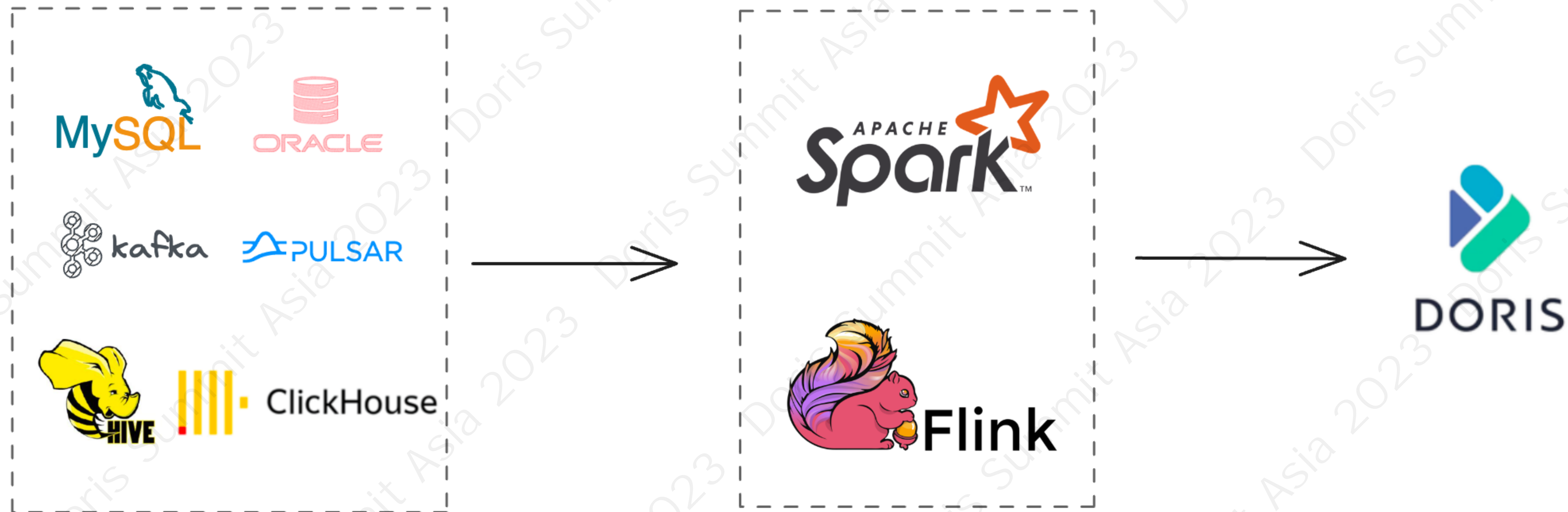
- 无自动建表
- 无工具可选
- 工具上手门槛高

- 不稳定
- 单机部署
- 性能低

- 无法保证准确性
- 无 Failover 机制

2 数据迁移连接器

数据迁移（离线 | 实时）

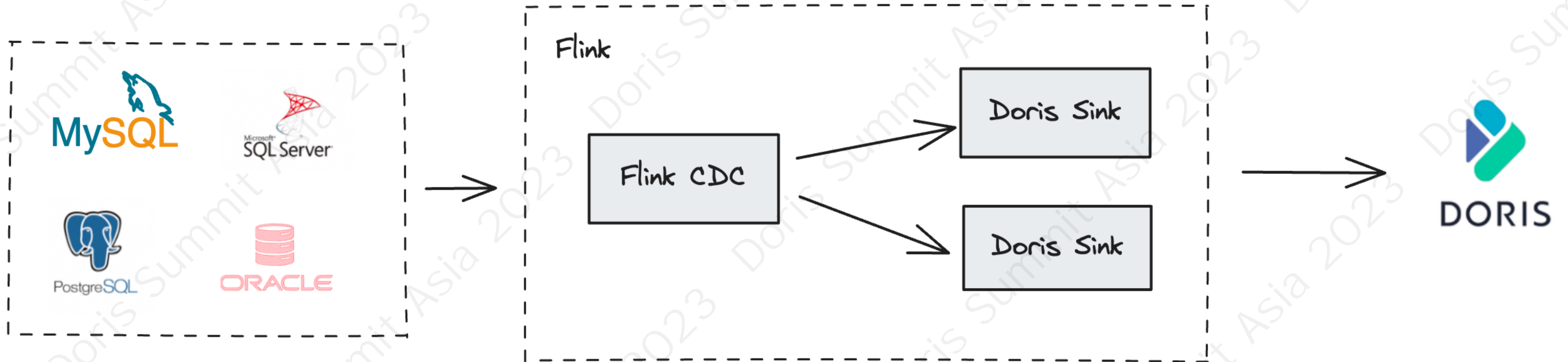


TP 数据库实时同步的痛点

1. 手动映射上下游字段，维护大量 FlinkSQL 脚本
2. 多表同步场景下，不共用 CDC Source，源端压力大
3. FlinkSQL 无法同步 Schema Change 操作，DataStream 开发成本大
4. 大量数据表的预先创建

集成 FlinkCDC，让数据实时迁移更简单

Flink Doris Connector 通过内部集成 FlinkCDC，可以预先在 Doris 中创建表，并且无需配置上下游表映射，快速同步上有数据至 Doris 中。



MySQL | Oracle 万表入 Doris

Doris-Flink-Connector 1.4.0

一键实现万表 MySQL 整库同步

MySQL 整库同步示例

使用 Flink Doris Connector 一键同步 MySQL 表到 Doris 中

```
<FLINK_HOME>/bin/flink run \  
-Dexecution.checkpointing.interval=10s \  
-Dparallelism.default=1 \  
-c org.apache.doris.flink.tools.cdc.CdcTools \  
lib/flink-doris-connector-1.16-1.4.0.jar \  
mysql-sync-database \  
--database test_db \  
--mysql-conf hostname=127.0.0.1 \  
--mysql-conf username=root \  
--mysql-conf password=123456 \  
--mysql-conf database-name=mysql_db \  
--including-tables "tbl|test.*" \  
--sink-conf fenodes=127.0.0.1:8030 \  
--sink-conf username=root \  
--sink-conf password=123456 \  
--sink-conf jdbc-url=jdbc:mysql://127.0.0.1:9030 \  
--sink-conf sink.label-prefix=label1 \  
--table-conf replication_num=1
```

核心配置

--database	同步到Doris的数据库名
--including-tables	需要同步的MySQL表, 支持正则表达式
--mysql-conf	MySQL CDCSource的配置项
--sink-conf	DorisSink的配置项
--table-conf	Doris表的配置项, 即Properties的配置

ORACLE 整库同步示例

使用 Flink Doris Connector 一键同步 ORACLE 表到 Doris 中

```
<FLINK_HOME>/bin/flink run \  
-Dexecution.checkpointing.interval=10s \  
-Dparallelism.default=1 \  
-c org.apache.doris.flink.tools.cdc.CdcTools \  
./lib/flink-doris-connector-1.16-1.5.0.jar \  
oracle-sync-database \  
--database test_db \  
--oracle-conf hostname=127.0.0.1 \  
--oracle-conf port=1521 \  
--oracle-conf username=admin \  
--oracle-conf password=admin123 \  
--oracle-conf database-name=HELOWIN \  
--oracle-conf schema-name=ADMIN \  
--including-tables "PERSONS.*" \  
--sink-conf fenodes=127.0.0.1:8030 \  
--sink-conf username=root \  
--sink-conf password=\  
--sink-conf jdbc-url=jdbc:mysql://127.0.0.1:9030 \  
--sink-conf sink.label-prefix=label \  
--table-conf replication_num=1
```

核心配置

- database 同步到Doris的数据库名
- including-tables 需要同步的MySQL表,支持正则表达式
- oracle-conf Oracle CDCSource的配置项
- sink-conf DorisSink的配置项
- table-conf Doris表的配置项,即Properties的配置

部署 doris-flink-connector 最佳实践

APACHE
StreamPark

系统管理

StreamPark

项目管理

作业管理

变量管理

设置中心

版本: 2.1.1

stars 3.4k forks 871

团队: default

admin

上传jar文件:

单击或拖动 jar 到此区域以上传
支持单次上传。您可以在此处上传本地 jar 以支持当前作业

flink-doris-connector-1.16-1.4.0.jar

* 程序入口类: org.apache.doris.flink.tools.cdc.CdcTools

* 作业名称: Flink整库同步作业

程序参数:

1 mysql-sync-database \
2 --database test_db \
3 --mysql-conf hostname=127.0.0.1 \
4 --mysql-conf username=root \
5 --mysql-conf password=123456 \
6 --mysql-conf database-name=mysql_db \
7 --including-tables "tbl|test.*" \
8 --sink-conf fenodes=127.0.0.1:8030 \
9 --sink-conf username=root \
10 --sink-conf password=123456 \
11 --sink-conf jdbc-url=jdbc:mysql://127.0.0.1:9030 \
12 --sink-conf sink.label-prefix=label1 \
13 --table-conf replication_num=1

描述: MySQL整库同步作业

取消 提交

部署 doris-flink-connector 最佳实践

APACHE StreamPark

System

StreamPark

Project

Application

Variable

Settings

Apache StreamPark, Make stream processing easier!

Version: 2.1.1

stars 3.4k forks 871

Team

Available Task Slots

0

Task Slots 3 | Task Managers 3

Running Jobs

3

Total Task 3 | Running Task 3

JobManager Memory

16,384

Total JobManager Mem 16384 MB

TaskManager Memory

16,384

Total TaskManager Mem 16384 MB

Tags

Owner

Job Type

Name

+ Add New

Application Name	Flink Version	Tags	Run Status	Release Status	Duration	Modified Time	Owner	Operation
JAR aliyun-pgslave-hdworkdochd-2-d...	1.16.0		RUNNING	DONE SUCCESS	1m 37s	2023-10-17 09:54:11	admin	
JAR aliyun-pgslave-slavebillnohd-2-do...	1.16.0		RUNNING	DONE SUCCESS	19h 46m 57s	2023-10-17 09:54:11	admin	
JAR aliyun-pgslave-2-doris	1.16.0		RUNNING	DONE SUCCESS	18h 23m 32s	2023-10-17 09:54:16	admin	

show 1 ~ 3 records, total 3 1 10 / page

Website

Document

Copyright ©2023 The Apache Software Foundation. Apache StreamPark, StreamPark, and its feather logo are trademarks of The Apache Software Foundation

用户好评

在使用 Doris-Flink-Connector 之前，我们一般是通过 DataX 定时从业务系统中抽取数据，当进行全量同步时，抽取数据会对业务系统造成一定的压力，且该方式只能做到小时级的同步。Doris-Flink-Connector 可以实现一键化脚本操作，为我们减少了繁杂的手工配置流程，高效稳定的实现了整库数据快速同步。

—— 郑煤机数耘科技 资深大数据工程师 杨开元

Doris-Flink-Connector 一键操作即可快速实现 Oracle 数据整库同步到 Doris，节省了手动配置以及编写复杂同步代码的步骤，避免了手动同步中可能出现数据不一致的问题。不仅能提高数据的准确性和可靠性，也极大提升了工作的效率。

—— 海程邦达 资深大数据工程师 王新

原先将 Oracle 数据同步到 Doris 中时，需要手动创建 Source 和 Sink 表，而使用 Doris-Flink-Connector 后可以实现多表、整库数据一键同步，极大简化了开发流程，该工具还能实现字段类型自动转换，数据同步更加简单便捷。

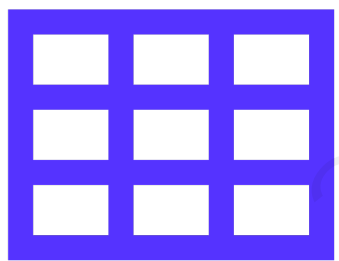
—— Apache Doris 某企业用户

3 X2Doris 百川入海

X2Doris 是 SelectDB 开发的，专门用于将各种类型的数据迁移到 Apache Doris (Selectdb Cloud) 中的核心工具，该工具集 **自动建doris表** 和 **数据迁移** 为一体，支持 **分布式** 部署，**超高性能** 的将各类数据 往 Doris(SelectDB Cloud) 迁移上迁移，整个过程 **可视化的平台** 操作，非常 **简单易用**，减轻数据同步到 Doris (SelectDB Cloud) 中的门槛。

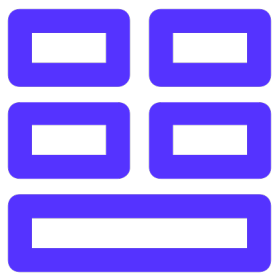


核心特性



自动建表

自动识别源表中的表结构，部分数据库支持**自动创建**目标 Doris 表，针对 Hive,支持**引导式**设置信息建表



多源支持

支持 **Hive, Doris, Doris系** 等数据源迁移至 Apache Doris, 更多数据源正在支持中。



简单易用

安装简单，解包启动即可。全程**可视化**，**平台化**操作，**引导式**作业创业，非常简单易用。



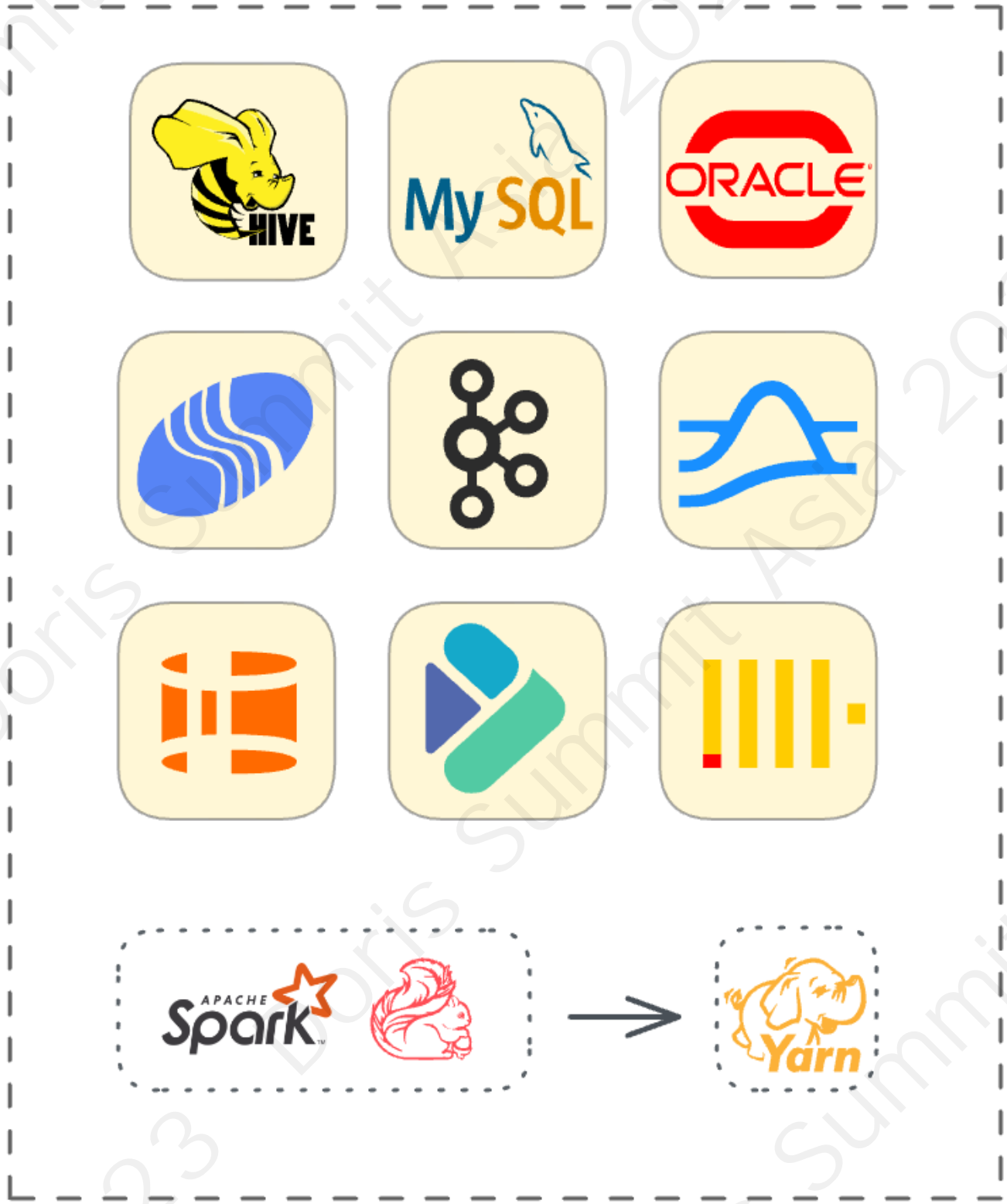
极速稳定

支持**单机**和**分布式**部署，经过诸多用户生产环境**海量数据**的检验，不论是速度还是**稳定性**，**准确性**，都做到行业前列

X2Doris 架构图



作业管理中心



同步作业核心层



写入目标

作业首页

作业中心

系统设置

作业名称

Please enter

创建时间

Start date

→

End date

Reset

Search

作业中心

作业名称

YARN Id

Doris Table

作业状态

操作

Doris:dwd.dwd_proj_conf_inco	application_1686904829848_0398	dwd.dwd_proj_conf_inco	FINISHED	
Doris:dwd.dwd_proj_budg	application_1686904829848_0397	dwd.dwd_proj_budg	FINISHED	
Doris:dwd.dwd_proj_borr_cost_adju	application_1686904829848_0396	dwd.dwd_proj_borr_cost_adju	FINISHED	
Doris:dwd.dwd_proj_bonu_deta	application_1686904829848_0395	dwd.dwd_proj_bonu_deta	FINISHED	
Doris:dwd.dwd_proj_bill	application_1686904829848_0394	dwd.dwd_proj_bill	FINISHED	
Doris:dwd.dwd_proj_bid_serv_fee	application_1686904829848_0393	dwd.dwd_proj_bid_serv_fee	FINISHED	
Doris:dwd.dwd_proj_atte_addr	application_1686904829848_0392	dwd.dwd_proj_atte_addr	FINISHED	
Doris:dwd.dwd_proj_appl_proj_stag	application_1686904829848_0391	dwd.dwd_proj_appl_proj_stag	FINISHED	
Doris:dwd.dwd_prod_proj_actu_cost	application_1686904829848_0390	dwd.dwd_prod_proj_actu_cost	FINISHED	
Doris:dwd.dwd_prec_pro_purc_cost	application_1686904829848_0388	dwd.dwd_prec_pro_purc_cost	FINISHED	
Doris:dwd.dwd_plti_reco	application_1686904829848_0387	dwd.dwd_plti_reco	FINISHED	
Doris:dwd.dwd_pers_sign	application_1686904829848_0386	dwd.dwd_pers_sign	FINISHED	
Doris:dwd.dwd_pers_pers_mont_snap	application_1686904829848_0389	dwd.dwd_pers_pers_mont_snap	FINISHED	
Doris:dwd.dwd_oms_rent_pub	application_1686904829848_0384	dwd.dwd_oms_rent_pub	FINISHED	
Doris:dwd.dwd_budg_item_deta	application_1686904829848_0383	dwd.dwd_budg_item_deta	FINISHED	
Doris:dwd.dwd_cost_erm_details_official	application_1686904829848_0379	dwd.dwd_cost_erm_details_official	FINISHED	

Version: 1.0.0

admin

Live Demo

超高性能

1. 单机 1G 内存情况下实测 5KM DataX 耗时 1.5 分钟, X2Doris 50 秒, **快 50%**
2. 国内某头部行业公司POC实测: 日志场景, 45个大字段, 速度可到 **2 分钟 / 1 亿条记录 / 45 G** 数据迁移
3. 比其他数据同类型的迁移工具, 速度快 **2 到 10** 倍
4. 对底层使用的 Spark 连接器进行增强, **进一步提升速度**。相关开源版本支持 **Array, Map, Bitmap** 类型的数据库迁移

内测用户反馈



阿庆

- 1、x2doris这个工具很好，可以作为doris生态组件的一部分，比如像GP的生态组件，用户用了GP后，对GP的生态组件依赖性很大，doris也可以开发一批成熟的组件来增加用户粘性
- 2、x2doris可以继续增加支持的源数据库列表，比如pg系、mysql系，甚至时序数据库也可以尝试一下度

—— 社区用户的反馈



X2Doris在杭银消金落地，打通了我司Hive到Doris的通路，为更多的上游数据同步进Doris提供了快捷便利的工具，希望X2Doris的功能越来越丰富，可以一站式的解决其他数据源到Doris的同步，真正实现Doris一站式数据平台的使用体验！

——杭银消金大数据架构师 周其进

正式开放使用

通过 [SelectDB 官网](#)和[公众号](#)了解更多信息

4 X2Doris 未来规划

1. 支持更多的数据源



计划支持 ClickHouse, Greenplum, Oracle, ADB... 等更多的数据源迁移至 Doris

2. 支持增量同步



Flink

支持 **Apache Flink**，借助 Flink CDC 及其生态连接器，完成大量数据源的**实时入 Doris** 的能力，进一步无缝对接 doris-flink-connector，让用户**可视化，平台化**使用增量**实时同步**的能力，实现**批流一体**，一个平台搞定所有数据的同步

3. 开放操作API



开发 **操作接口 API**，把作业的启动，停止，状态查询等常见操作**开放操作接口**，方便用户更好的与第三方**系统集成**，比如：集成定时调度等能力，按天启动作业迁移数据



获取更多社区动态与最佳实践

Apache Doris 官方平台:

- Apache Doris 官网: doris.apache.org
- Apache Doris GitHub: github.com/apache/doris/

获取更多峰会资料:

- Doris Summit 峰会官网: doris-summit.org.cn
- Doris Summit 峰会回放: <https://space.bilibili.com/1196172099/channel/collectiondetail?sid=1824324>