51CTO WOT

从数据到洞察：利用DataOps加速AI模型开发

代立冬
白鲸开源科技联合创始人 & CTO

关于我

白鲸开源科技联合创始人

Apache 孵化器导师

Apache DolphinScheduler PMC Chair

Apache SeaTunnel PMC

ApacheCon 亚洲大数据湖仓论坛出品人

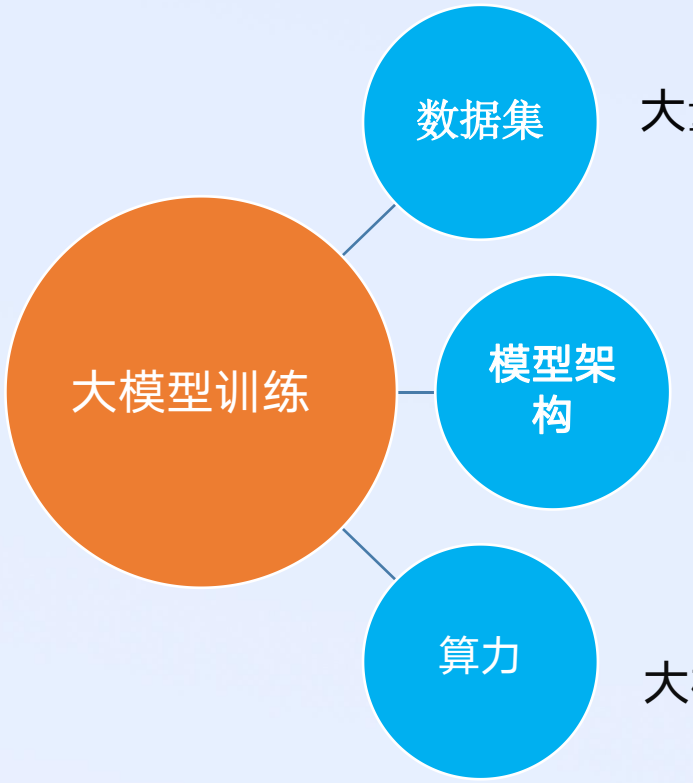
中国科协 “2023开源创新榜” 优秀人物

目录 CONTENTS

PART · 01

DataOps 核心理念

DataOps 在大模型开发中的作用



大量、多样化、高质量的数据是训练和微调大模型的基础

包括网络层数、隐藏层的大小、参数的类型等

大模型训练需要极其强大计算资源，包括高性能的GPU，如 A100 等

DataOps与AI模型开发的融合，将加速AI模型的开发周期，提升模型的准确性和效率。

大模型训练流程：数据工程 + 算法工程 + Ops

数据工程阶段

算法工程阶段

数据源 → 数据准备&数据流管理 → 数据质量管理 → 大模型训练 → 模型评估与测试 → AI应用

非结构化数据

向量数据

互联网/物联网

API/SaaS

传统数据库

信创数据库

跨数据源聚合

多层次数据验证

自动化数据加工 Co-Pilot

实时+批量整合

流处理 Stream Processing

数据版本管理

数据质量管控

数据影响分析

训练编排 Pipeline

预训练 (Pretraining)

指令微调 (Supervised Finetuning)

增强学习 (Reinforcement Learning)

性能评估

模型验证

模型优化

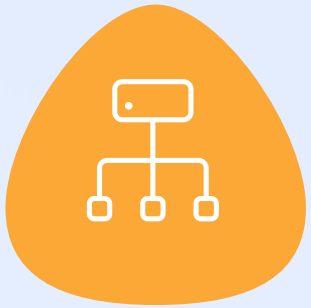
AIGC AI-Generated Content

BI Business Intelligence

应用 Apps

企业面临的数据挑战

- 企业内拥有多组 “数据平台”，数据资源和流程分散在各部门，难以掌控



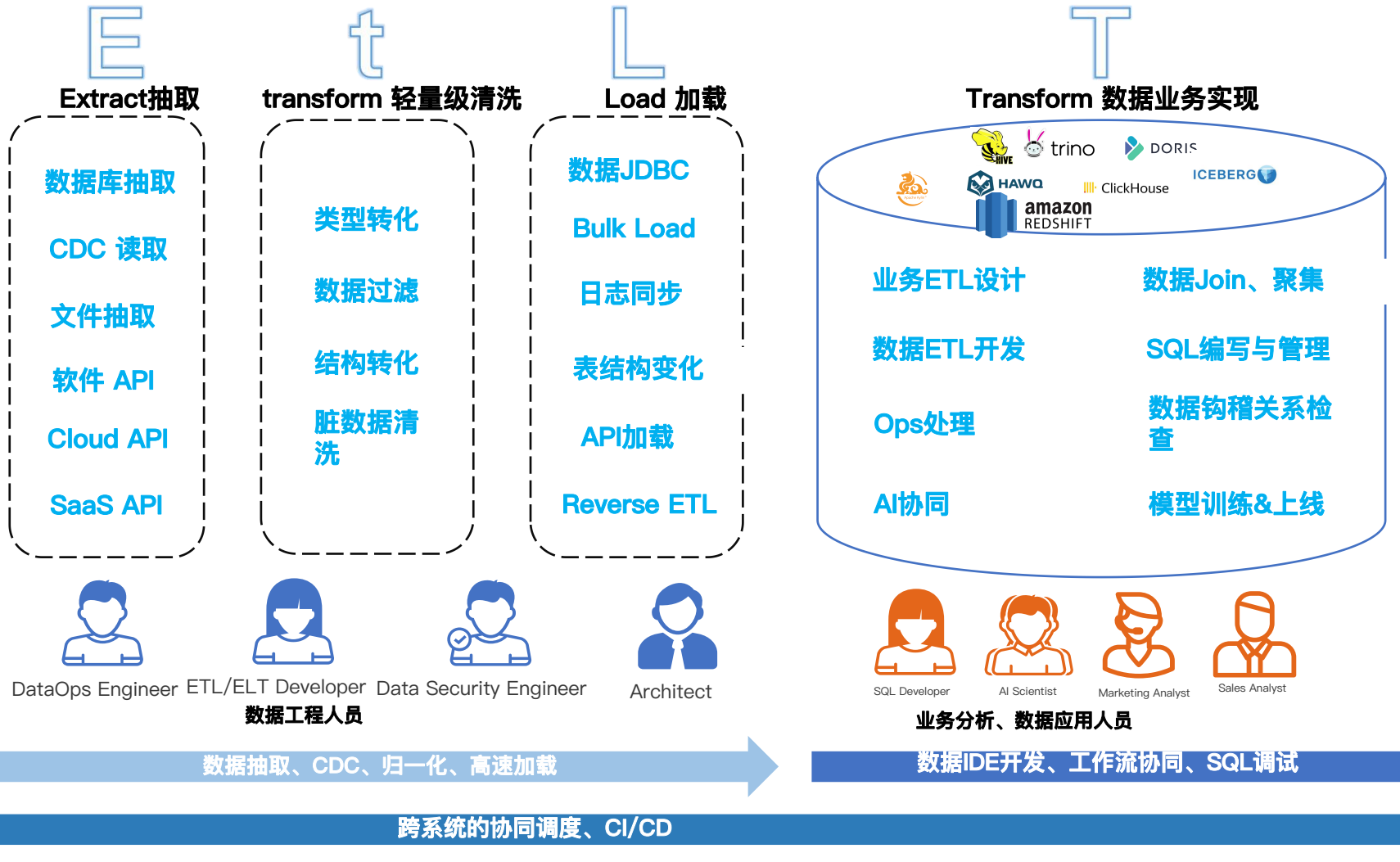
- 大数据、流数据、AI数据加工缺乏工具管控形成了企业新的“蜘蛛网”

- 企业大数据开发处于“野蛮生长状态”，整体研发管理距离应用开发DevOps流程相差甚远

- 多种新兴数据引擎、云原生、新数据架构的变化缺乏管控，数据血缘、同步、调度与数据发展严重落后

新技术环境下 EtLT 架构出现

云、SaaS、本地混合数据源让传统的数据处理流程从ETL、ELT变为能更加快速满足业务需求的 EtLT 架构，EtLT 能更加敏捷应对离线/实时数据湖、数据仓库、AI 模型训练当中的复杂多变的数据需求场景。



白鲸开源是 DataOps 领域领先的开源原生公司

白鲸开源是一家开源原生的 DataOps 商业公司，由多个 Apache Foundation Member成立，80%员工都是 Apache Committer，主导 2 个 Apache 顶级开源项目(Apache DolphinScheduler, Apache SeaTunnel）。

同时根据全球最佳实践发布商业版版本 -- WhaleStudio ，帮助企业解决内部多数据源、多数据系统复杂的**数据集成，数据开发、工作流编排运维及部署、数据质量管控、团队敏捷协作**等一些列问题，并在 6000 多家企业中得到实践和使用。

DataOps 的优势

增强数据质量

DataOps 强调数据质量的重要性，通过自主数据校验，提高数据的准确性，确保模型微调所需的优质数据

提升开发效率

DataOps 通过自动化和优化数据处理流程，减少人工干预，提高数据处理 和AI 模型的开发效率。

促进团队协作

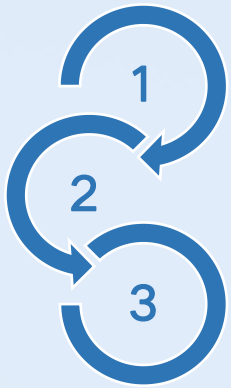
DataOps 提倡团队间协作和沟通，通过共享工具和平台，有助于构建高效、敏捷的数据 & AI 模型开发团队 。

DataOps 关键实践之任务调度平台

51CTO WOT

Apache DolphinScheduler：云原生调度引擎

大数据工作流痛点



多个任务单元

存在时间顺序

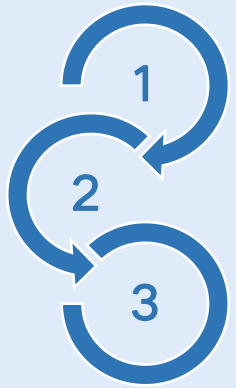
存在依赖关系

+

执行频率高

数据量、任务量大

云原生



企业级场景痛点

脚本 + 大数据组件

</> + cron

- 代码复用性低
- 集群化部署与拓展过于复杂
- 开源数据组件更新升级频率高

上一代调度引擎

Azkaban　Apache Airflow

- 非可视化设计 智能化程度低
- 扩展能力弱 受限于单点瓶颈
- 多云异构数据管理能力弱

新一代调度引擎

DolphinScheduler　Star: 12.1k

- 可视化任务调度 支持多任务类型
- 去中心分布式设计 高稳定可用
- 百万数据量级 任务稳定运行

5 + 年社区运营

520 + 开发者贡献

6000 + 用户使用

全球国内外企业在使用 DolphinScheduler

China unicom中国联通

联通数科早期使用原商业调度系统支撑着全域数据平台加工与调度，以接口机配合Shell（HiveSQL）为主的开发编排运维模式，处理日均数万的流程实例和日均**上百万的 Job** 作业，对比闭源调度工具、Airflow、Azkaban后，最终选择DS。

- 满足业务需求和调度功能要求
- 满足大数据量要求
- 用户使用成本低

阿里云　中国平安 PING AN 金融·科技　奇安信

高性能、大批量数据调度

SHEIN

早期使用Airflow调度全球任务，但因为分布式支持问题、无可视化问题导致系统开发效率和稳定性堪忧，同时也无法支持K8S和全球的云原生部署。选择从Airflow迁移至DS。

- 全球云部署、K8S支持
- 分布式去中心化以保证稳定
- 全量替换Airflow
- 解决全球大数据调度易用问题，赋能分析团队快速开发调度任务

Shopee　McDonald's　有赞

全球云部署、易用数据开发

荔枝FM

过去大数据调度使用SQL/Shell/Python脚本和其他大数据组件完成整个AI流程，面临不易用且难复用问题。 使用基于DS的AI开发平台后，荔枝FM将获取数据、数据预处理、模型训练、模型预测、模型评估和模型发布过程抽象成组件，用 DAG 串联，使用拖拽和配置的方式实现低代码开发。

- 实现对海量数据存算
- 可以复用ML流程
- DAG 执行引擎

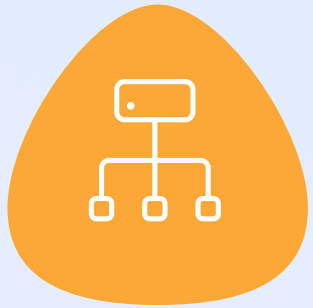
360

AI/ML Ops调度融合

DataOps关键实践
之数据集成工具

企业数据集成面临的挑战

- 数据源多达几百种，版本间不兼容，而且不断有新的出现



- 数据丢失与重复，无法一致性
- 出现问题无法回滚或者断点继续执行
- 同步过程不透明，缺少监控

- 频繁读取 binlog 对数据源端影响大
- 大事务、Schema 变更影响下游
- 低吞吐高时延导致数据无法及时到达

- 离线同步和实时同步常被分开管理，维护困难
- 数据割接人工进行

51CTO WOT

SeaTunnel：新一代实时多源数据同步工具

Star: 7.4k

上百种源数据库/地点

MySQL
PostgreSQL
Kafka
MongoDB
Elastic
Hive
Druid
Redis
AWS Aurora
Hudi
Kudu
HBase
InfluxDB
Neo4j
Feishu

数据同步与集成

原有解决方案

Spark
Flink
.....

Sea Tunnel

SeaTunnel Universal API

Table API
Source API
Sink API
Engine API

SeaTunnel Engine

- 批量数据全量、增量集成
- 实时数据集成
- 批量无主键增量集成等

160 +
连接器数量

3x
版本迭代

500+%
接口数增长

其他解决方案

Airbyte
比 Airbyte性能快30倍

DataX
单机比DataX性能快2.6倍

目标数据库/地点

MySQL
PostgreSQL
Kafka
MongoDB
Elastic
TiDB
Druid
Redis
Hive
Iceberg
Kudu
HBase
AWS Redshift
SelectDB
StarRocks
AWS S3
ClickHouse
Snowflake

SeaTunnel 特点

- 简单易用，开箱即用，不依赖HDFS，Flink，Spark
- 全可视化操作

- 丰富的数据源支持，目前已经支持 100+ 种数据源

- 批流 、实时、CDC一体化配置
- 无主键增量数据集成
- 整库同步、表结构自动变更

- 存算分离架构设计
- 高性能数据同步
- 支持节点动态伸缩

- 全量到增量无锁化自动切换
- 读缓冲(一个源到多个目标数据源，只用一次读取)
- 动态速率控制，对源端和目标端压力可控

- 支持 Schema Evolution
- 断点续传
- 实现 Exactly-Once 一次语义，保证数据一致性

- 云组件支持
  - K8s支持
  - AWS Redshift、S3, RDS, DynamoDB
  - 阿里 OSS File，Max, TableStore等

- 无中心化设计确保系统的高可用，支持多云
- 支持每日千亿级数据量同步

为什么重新造轮子 – 自研同步引擎？

其它引擎

01 容错粒度大

当进行多表同步时，任何一张表出现问题，整个作业都会失败停止，导致所有表的同步都会延迟

02 资源浪费严重

每个作业只能同步一张表，当我们有多张表需要同步时，需要启动多个Job来运行，这种方式非常的浪费资源

03 JDBC连接数过多

每个task只能处理一张表，所有每张表至少需要一个JDBC连接来读取或写入数据。当进行多表同步和整库同步时，需要大量的JDBC连接

04 不支持数据缓存

CDC场景下，有些企业设置的数据库日志留存时间比较短，作业失败时，数据库日志处理也会停止，会出现源端数据库日志被清除的情况

05 不支持表结构变更

在CDC场景下，DDL变更的检测和下游应用非常重要，目前Datax/Spark/Flink/FlinkCDC都无法支持

SeaTunnel 同步引擎- Zeta

01 易用省资源

WhaleTunnel 利用Zeta引擎Dynamic Thread Sharing技术，提高CPU利用率，不依赖HDFS，Spark等复杂组件，更好的单机处理性能

02 可视化开发与运维

WhaleTunnel 使用智能辅助的可视化UI技术，帮助用户快速建立单表多字段、多表多字段、SaaS和非结构化到数据库等复杂的数据集成任务并监控

03 多方式确保一致性

支持无中心HA和更细粒度的作业回滚机制，结合多阶段提交与CheckPoint机制，确保数据一致的同时避免大量回滚导致性能下降

04 极致CDC&批量性能

WhaleTunnel 实现 zero-copy 技术，不需要序列化开销，列式内存格式增加大吞吐，可以最大化利用网络带宽，支持Oracle/SQL Server/Mysql 等

05 多复用/数据库日志多表解析

WhaleTunnel支持多表或整库同步，解决JDBC连接过多的问题；支持多表或整库数据库日志读取解析，解决CDC多表同步场景下需要重复解析日志的问题。

51CTO WOT

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

bilibili

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，数据集群规模几十台，日均记录数上千亿，日均数据量在100T以上。

滴滴 腾讯云 去哪儿旅行 oppo vip.com唯品会 多点DMALL bilibili

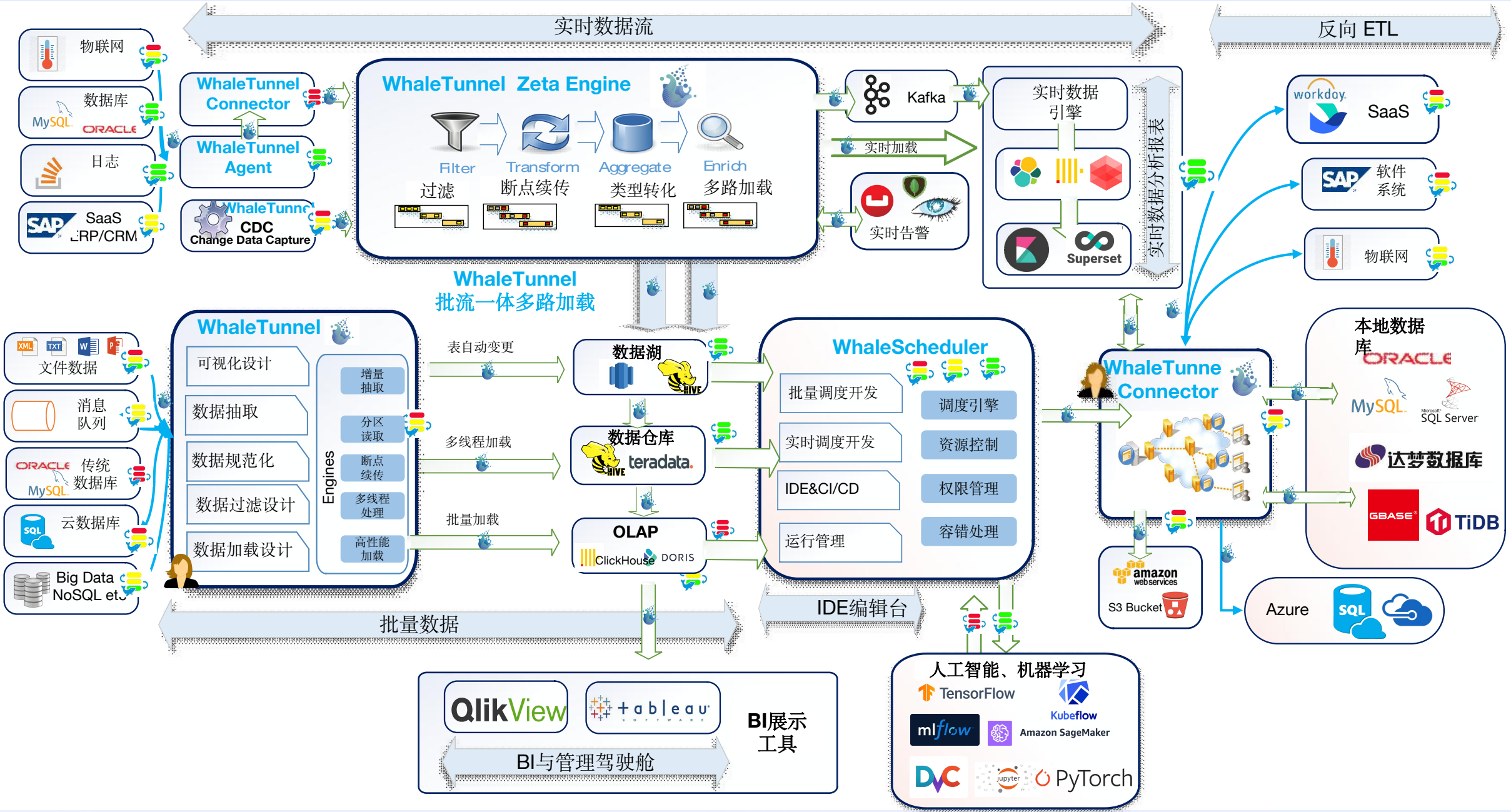
Shopee YOOZOO 游族网络 FOXCONN 富士康科技集团 新浪微博 weibo.com coupang ByteDance 字节跳动 ZUI YOU

松果出行 BOTON 宝通 中国移动 China Mobile 虎牙直播 国家电网 STATE GRID 震旦集團 AURORA GROUP GRIDSUM 国双

PART · 02

白鲸开源 DataOps 实践

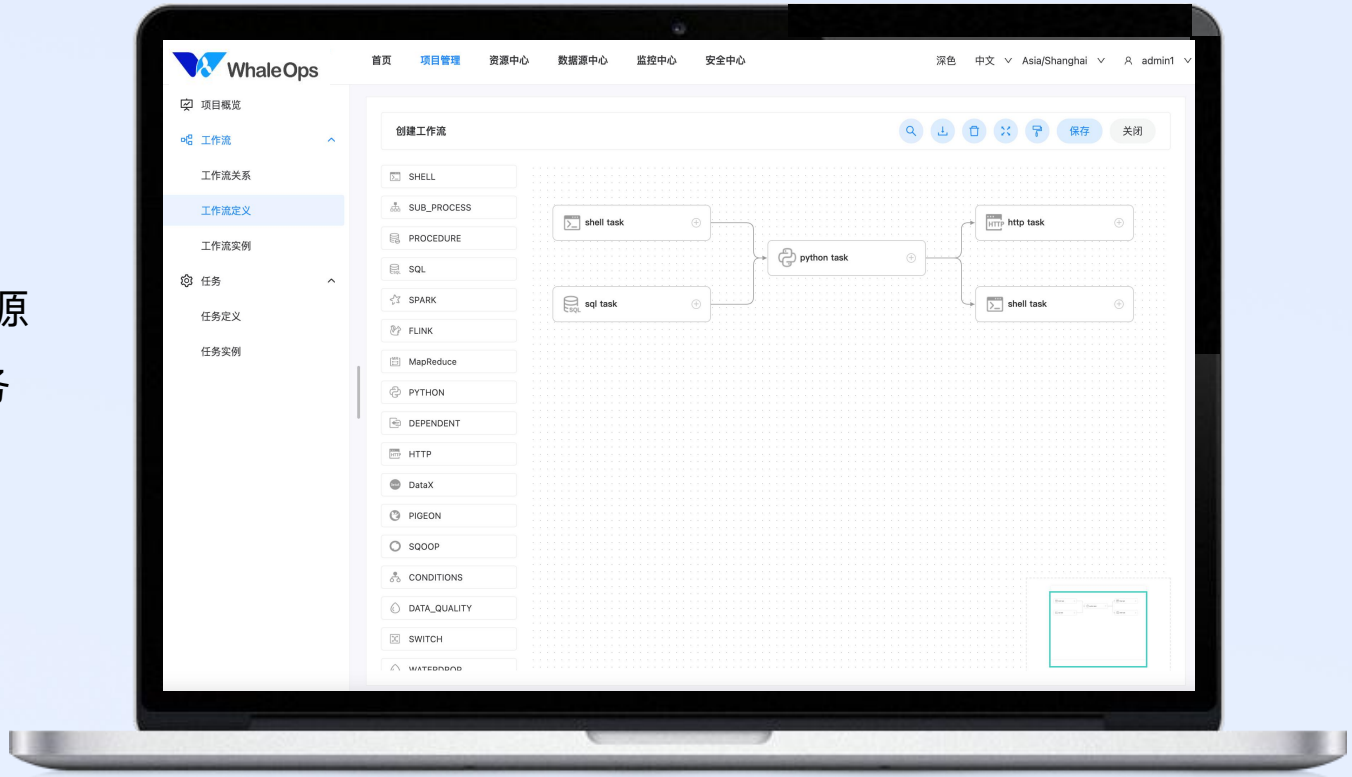
白鲸开源企业数据打造全流程 DataOps



稳定高效的企业级 DataOps 平台

基于 Apache DolphinScheduler 和 Apache SeaTunnel 的产品 WhaleStudio ，是分布式、云原生并带有强大可视化界面的 DataOps 系统，增加了商业客户所需的企业级特性：

- 完全自主研发，上下游生态圈广阔，支持 160+ 种数据源
- 全面支持云原生—云、仓、湖 实时/离线批流一体化任务管控
- 低代码实现企业大数据的操作系统和高速公路
- 完善的 DataOps 流程可无缝集成代码工具
- 丰富的数据源对接和传统 ETL 数据组件支持
- 一站式完成从开发-》测试-》上线-》运维闭环



全部项目 10
项目概览
工作流
调度任务
离线任务实例
实时任务实例
加冕列表
隔离列表
离线同步
离线任务定义
离线任务实例
实时同步
实时任务定义
实时任务实例

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： –

自动刷新 5s

已读取数据量（行） Row MB
3,560,000

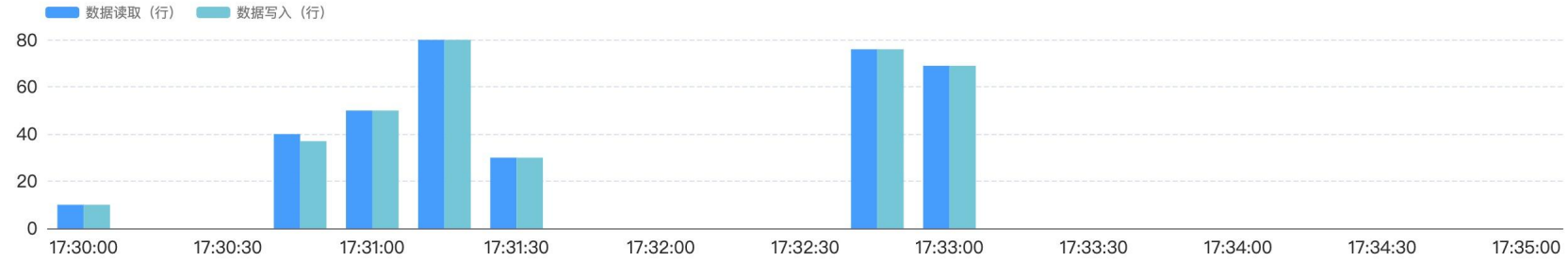
已写入数据量（行） Row MB
3,560,000

Insert: n
Update: n
Delete: n
DDL: n

当前数据延迟（秒）
1.03 10%↘
Maximum：20.33
Average：2

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 最近 5min 1hour 1day 1week



实时增量 搜索表名称

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

1 10 / 页 跳至

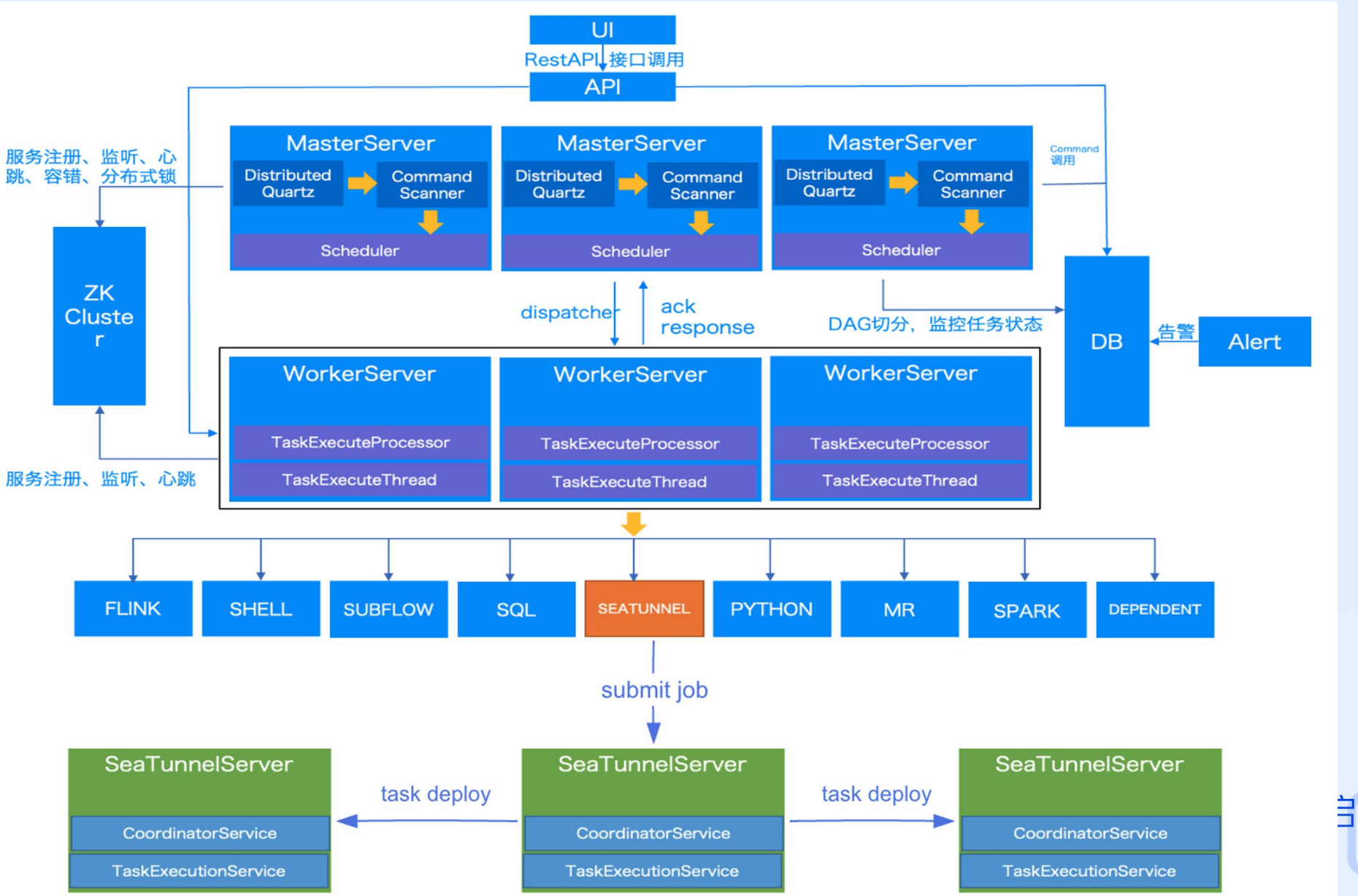
任务基本信息
同步任务定义 同步任务定义名称
业务模型 实时同步
状态 暂停
执行用户 admin
开始时间 2024-1-1 15:22:22
结束时间 –
最近操作用户 admin
最近操作时间 –

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

技术架构


UI
RestAPI 接口调用
API
服务注册、监听、心跳、容错、分布式锁
MasterServer
Distributed Quartz
Command Scanner
Scheduler
MasterServer
Distributed Quartz
Command Scanner
Scheduler
MasterServer
Distributed Quartz
Command Scanner
Scheduler
Command 调用
ZK Cluster
dispatche
ack response
DAG切分，监控任务状态
DB
告警
Alert
WorkerServer
TaskExecuteProcessor
TaskExecuteThread
WorkerServer
TaskExecuteProcessor
TaskExecuteThread
WorkerServer
TaskExecuteProcessor
TaskExecuteThread
服务注册、监听、心跳
FLINK
SHELL
SUBFLOW
SQL
SEATUNNEL
PYTHON
MR
SPARK
DEPENDENT
submit job
SeaTunnelServer
CoordinatorService
TaskExecutionService
task deploy
SeaTunnelServer
CoordinatorService
TaskExecutionService
task deploy
SeaTunnelServer
CoordinatorService
TaskExecutionService

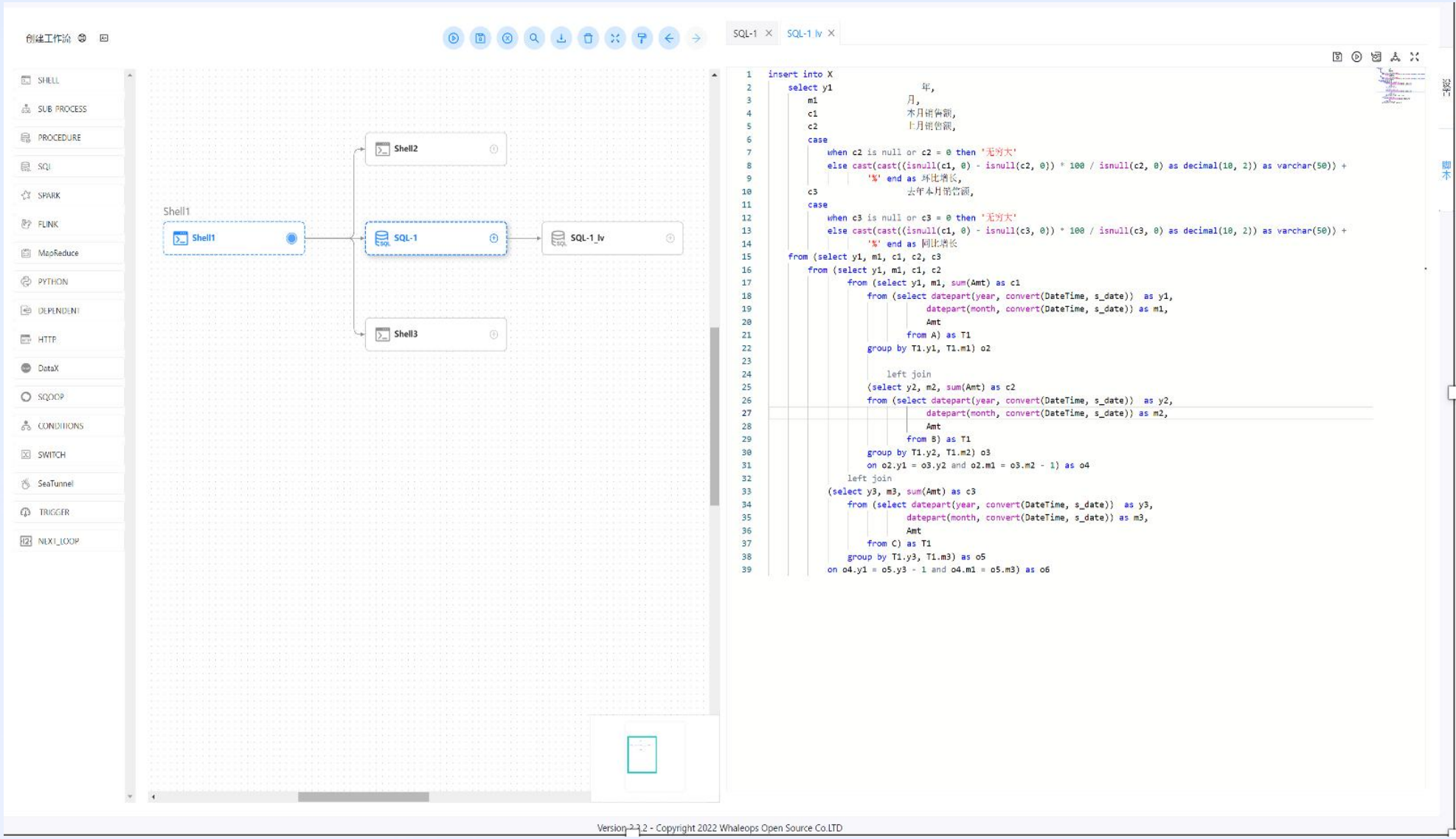
WhaleStudio - 工作流编排能力

1. 支持各类计算任务组件：Amazon DMS、Azure Datafactory，Amazon Datasync、Apache Linkis，DataX，Sqoop，SeaTunnel等

2. 支持各类云数据库和计算架构，支持 K8S、MLDB。

3. 平台采用插件式设计，支持自由扩展数据源支持。

4. 可视化的数据源管理，数据源统一集中管理，一次配置，到处使用，大大减少配置修改带来的工作量。



支持160种数据源接口，多种数据集成方式

WhaleTunnel支持 160种数据源，例如MySQL，SAP Hana，Oracle，DB2，SQLServer，Gbase，Kafka，ClickHouse，RedShift、达梦等
平台采用插件式设计，支持自由扩展数据源支持

支持多种

- 批量数据全量、增量集成
- 实时数据集成
- 批量无主键增量集成等

商业版支持商业数据库实时CDC

- Mysql
- PostGreSQL
- SQLServer
- Oracle
- DB2
- AWS Aurora
- 翰高
- StarRocks
- 达梦
- 人大金仓
- PolarDB

*目前支持160种数据源，详情参见Excel

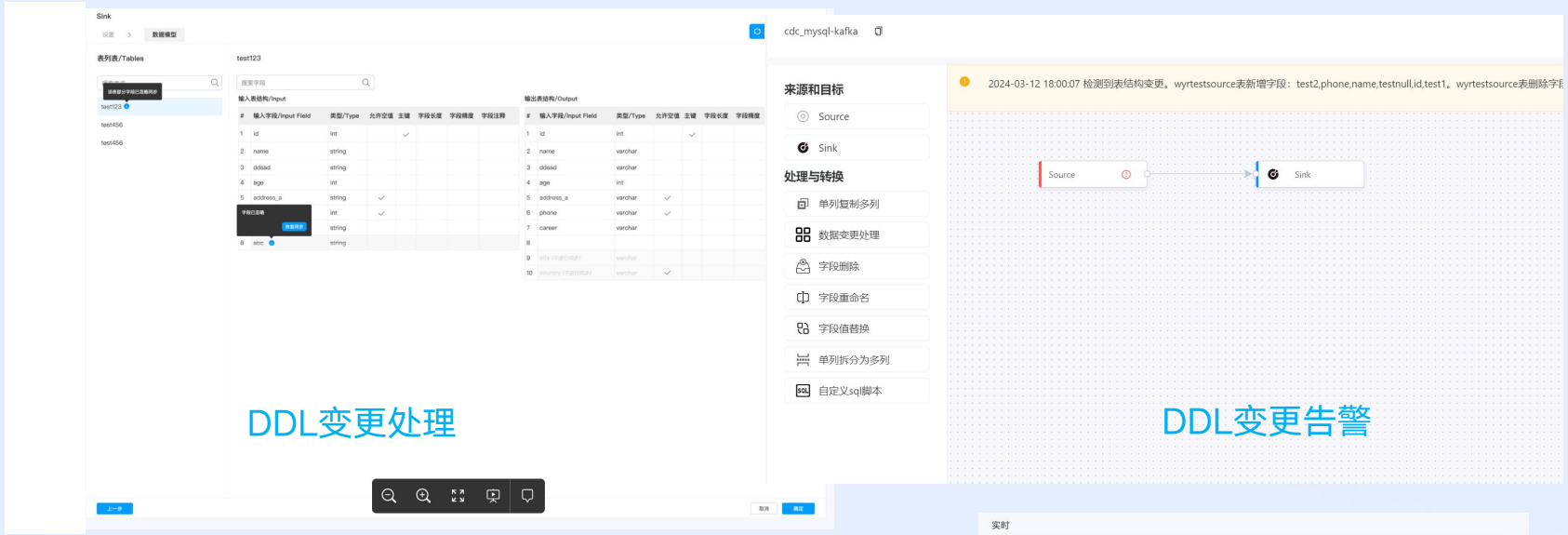
实时数据同步支持DDL变更触发暂停、报警以及延迟告警

实时数据处理支持多种实时数据监测处理：

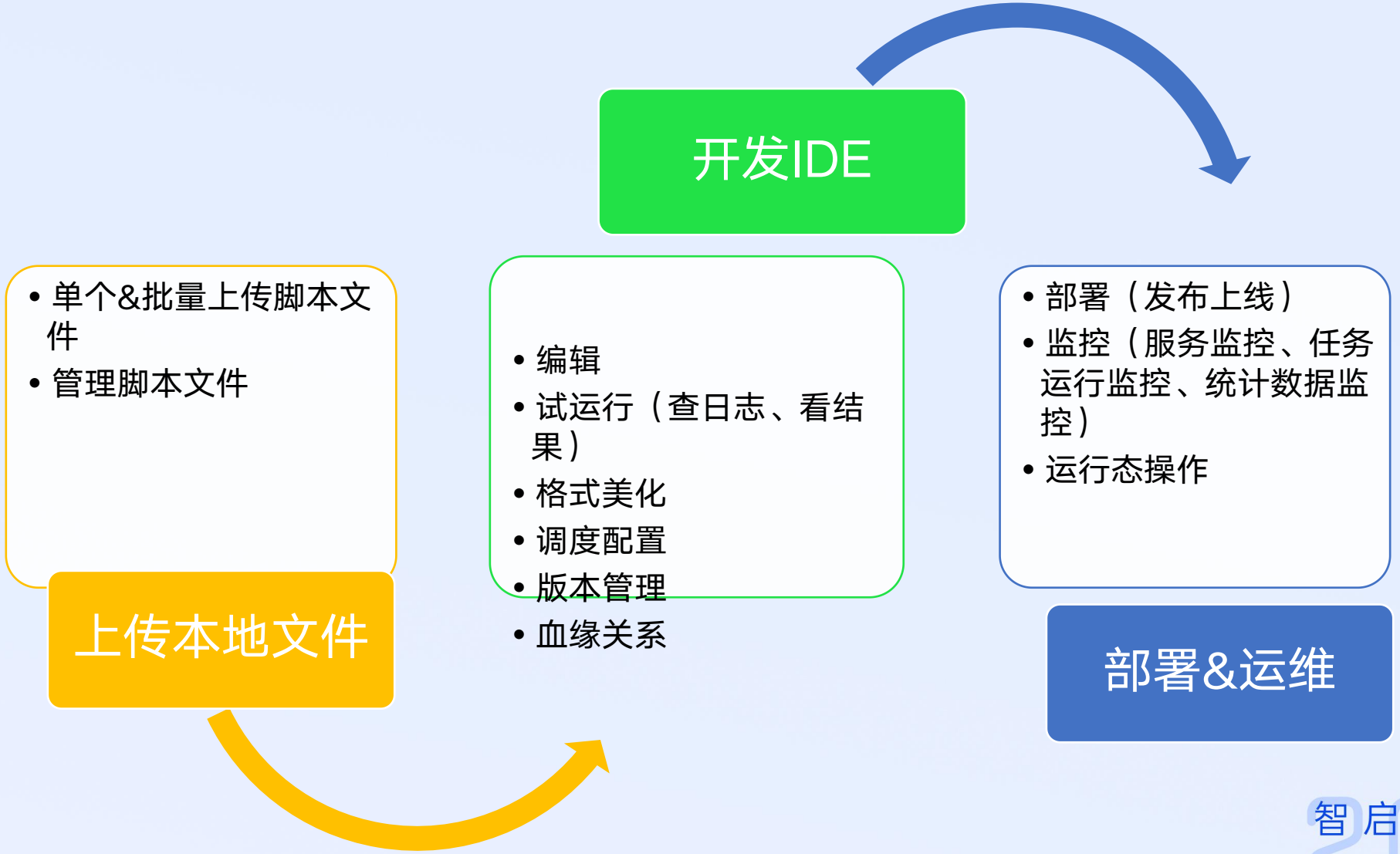
- DDL变更暂停
- DDL变更告警
- DDL暂停加表
- DDL手工处理

支持多种方式控制&监测速率：

- 数据采集速率控制
- 并发控制
- 数据延迟告警
- 数据全量完成告警
- 数据CDC增量启动告警

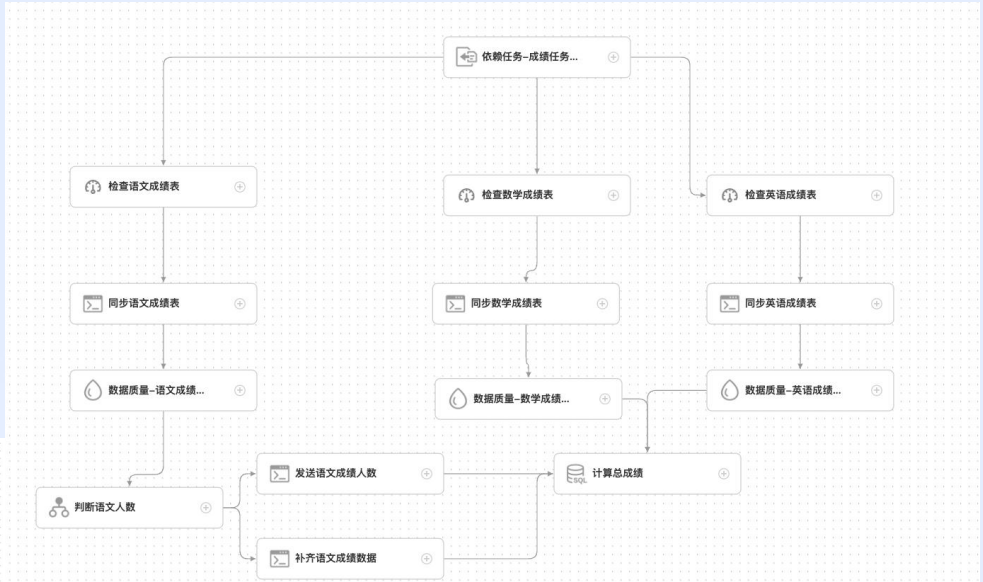


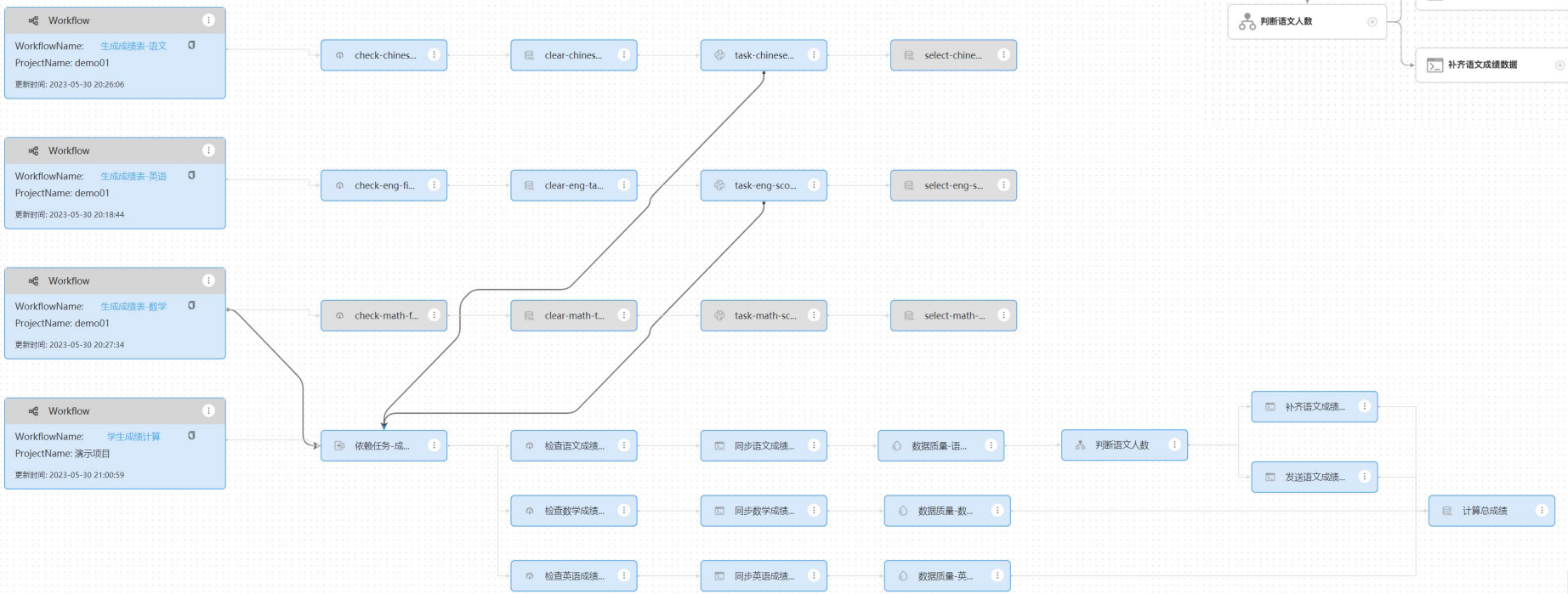
数据开发能力- 在线 IDE & 集成 ， 实现 DataOps

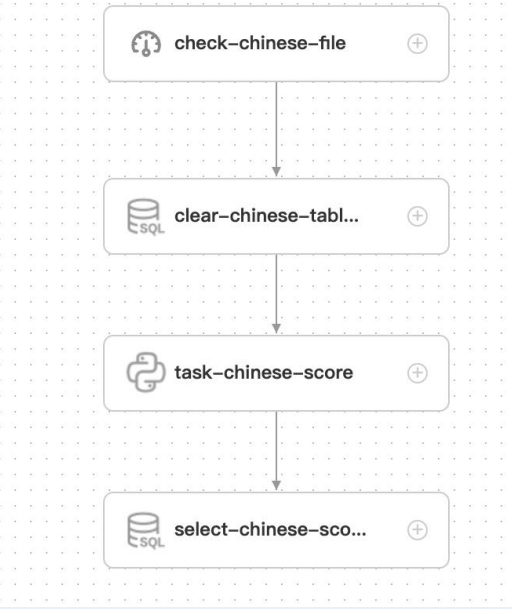


数据质量把控 – 数据工作流血缘关系

- 全局跨工作流的任务和及实例间的依赖关系
- 结合任务与表定义，实现表及血缘分析以及任务操作
- 支持实例级别的依赖链路展示
- 支持全局视图进行停止、暂停、重跑、依赖链重跑等操作



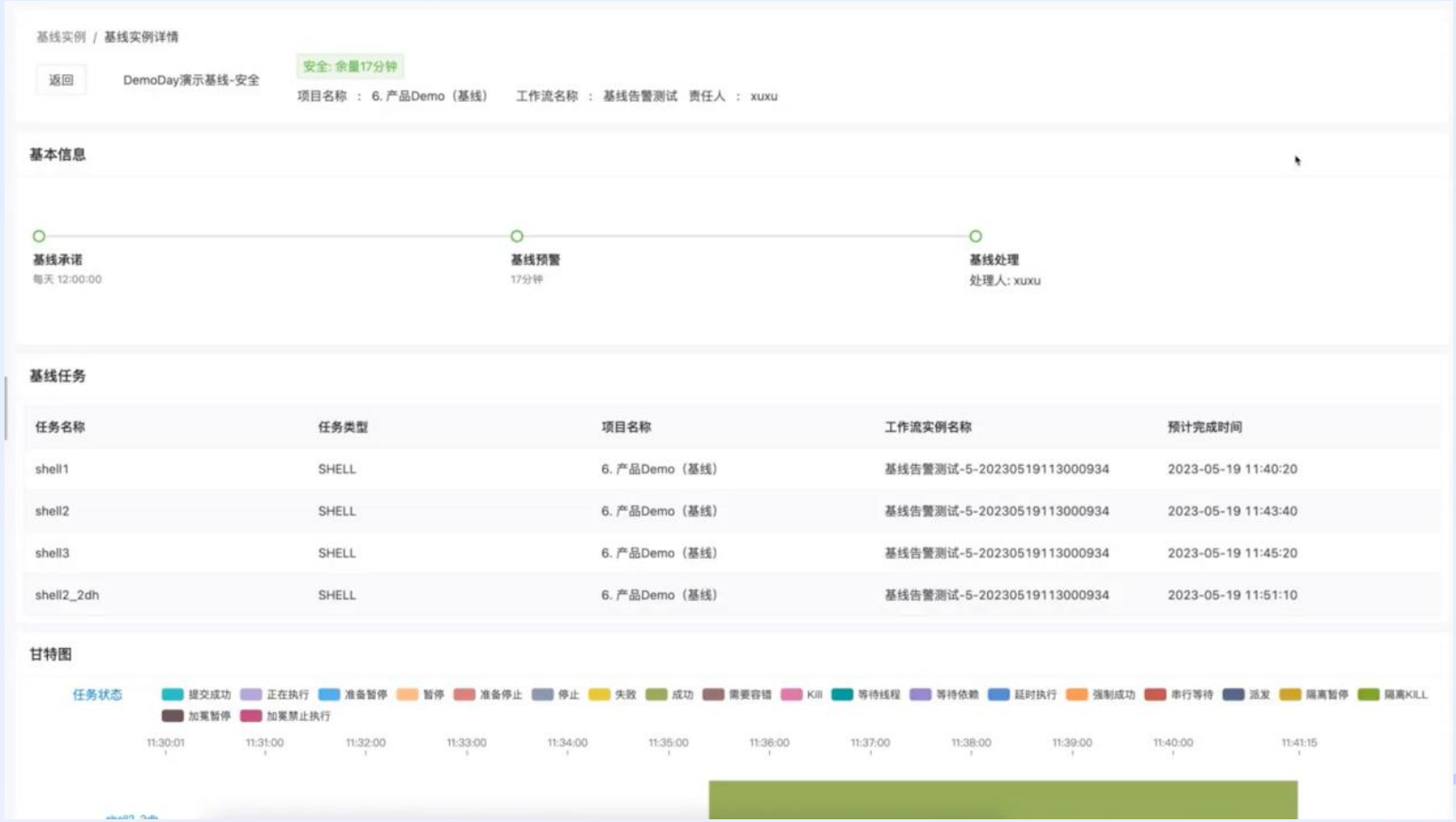




智能基线–提前预知任务延迟从而提前采取行动

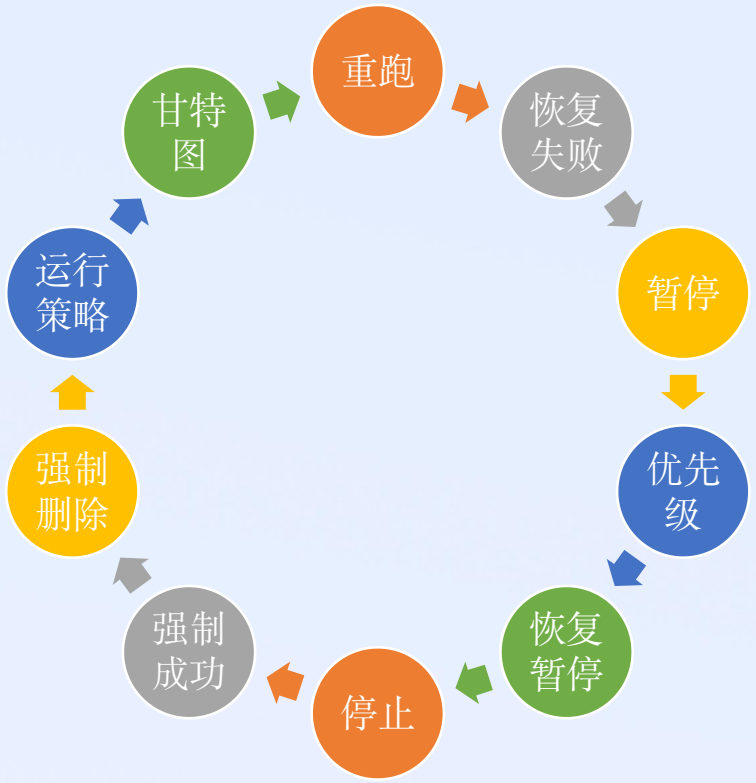
配置智能基线帮助“智能”告警:

- ✓ 定义核心任务基线，多一双“智能”的眼镜
- ✓ 根据任务的执行历史只能推算时长
- ✓ 设置安全预警时间，智能告警

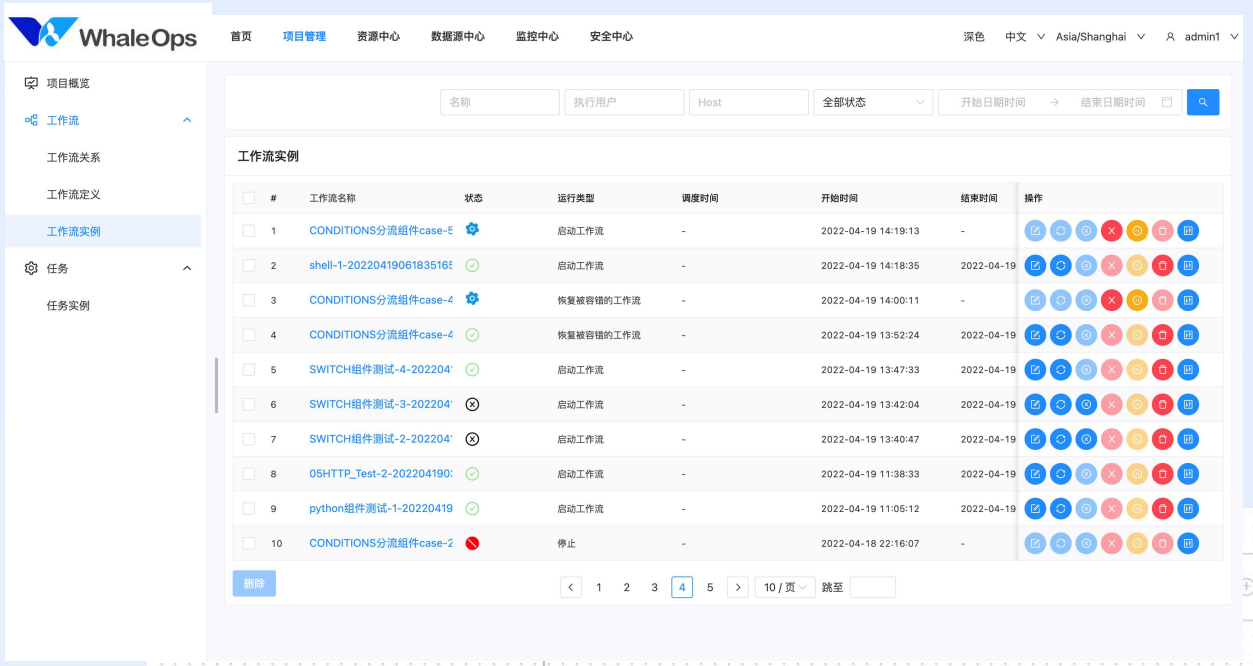


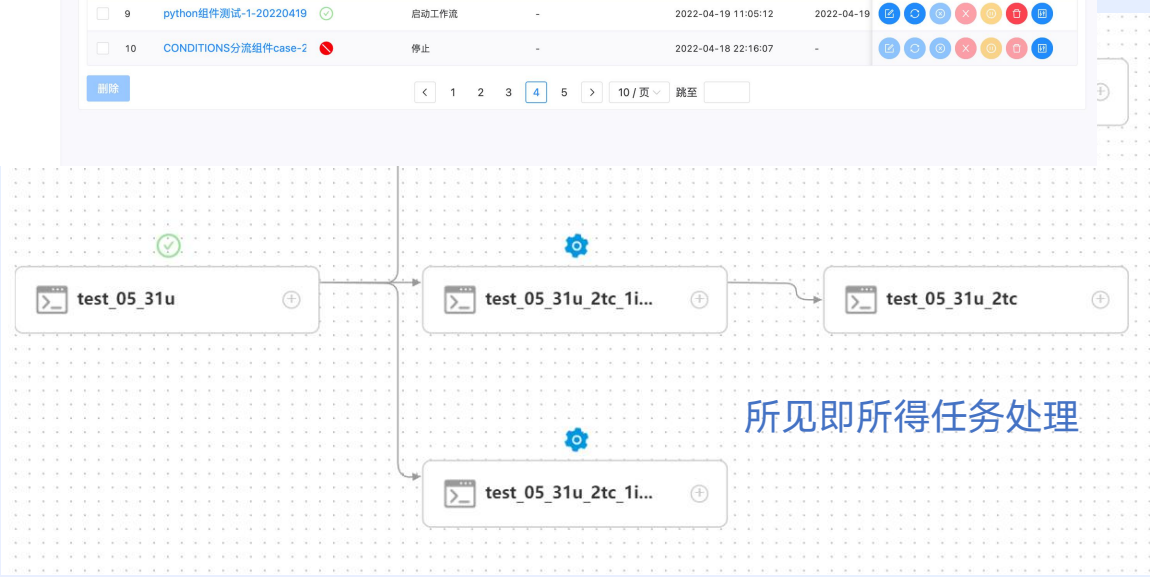
多种运维管理手段 —— 帮助运维人员快速处理故障

任务上线之后，面对各种突发情况，有多种手段来确保在任务发生异常时可以协助运维人员快速处理异常。



对工作流实例和任务实例进行丰富的人工干预功能





PART · 03

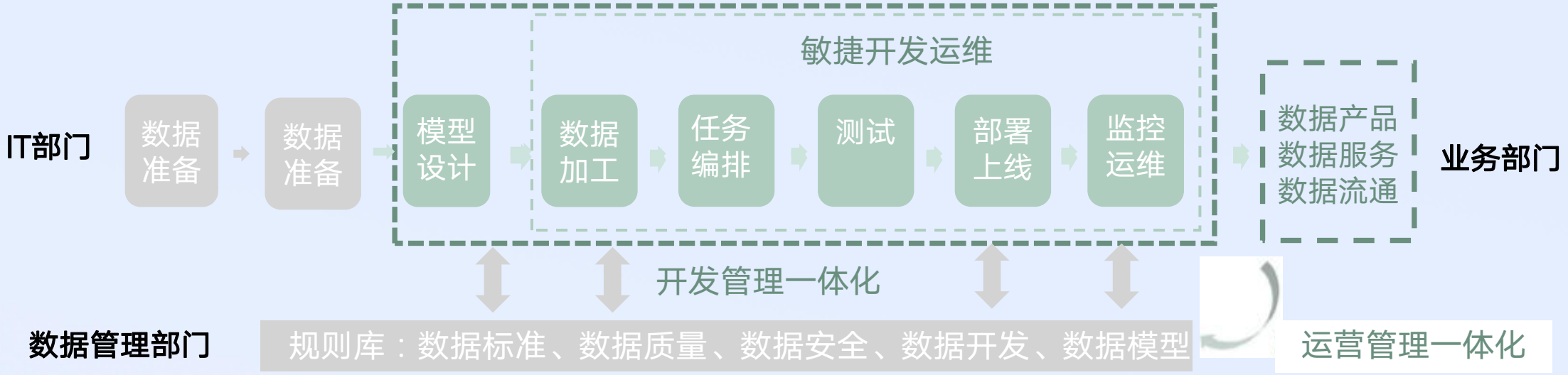
实践案例

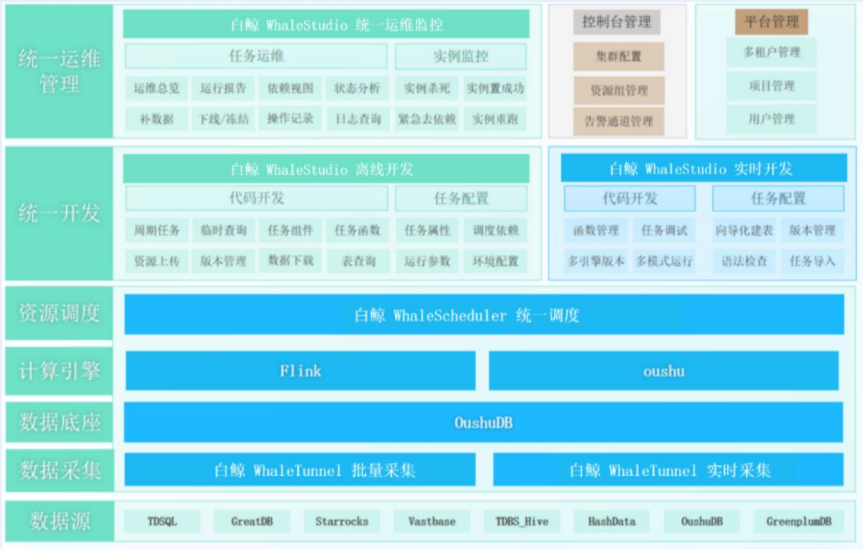
中信建投——当前数据开发中存在的挑战

反洗钱监管加强	需要实时、全面地整合业务数据	传统数据流程和时效性无法满足当前需求
监管报送复杂性	多样化的报送频次和优先级	自动化挑战：传统调度系统难以适应
新兴数据源与国产信创生态	湖仓一体架构对数据开发提出更高要求	
实时性需求提升	实时计算盈亏、资产账户、持仓股票信息	批流一体架构对大数据研发提出更高要求
系统稳定性	数据处理结果对市场交易和公司经营至关重要	对系统稳定性和可靠性有极高要求

中信建投——中信建投DataOps布局

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，累计编排定义工作流超过3000个，上线任务数量接近16000个,交易日平均运行工作流实例数量超过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。





实际日历			20220321 周一		20220322 周二		20220323 周三		20220317 20220324 周四		20220318 20220325 周五		20220319 20220326 周六		20220320 20220327 周日		20220328 周一	
			是否交易日	是否跑批日	是否交易日	是否跑批日	是否交易日	是否跑批日	是否交易日	是否跑批日	是否交易日	是否跑批日	是否交易日	是否跑批日	是否交易日	是否跑批日	是否交易日	是否跑批日
			是	是	是	是	是	是	是	是	是	否	否	是	否	否	是	是
切日时间	日历类型	跑数类型	是否跑批	跑数日期	是否跑批	跑数日期	是否跑批	跑数日期	是否跑批	跑数日期	是否跑批	跑数日期	是否跑批	跑数日期	是否跑批	跑数日期	是否跑批	跑数日期
16点前	交易日历	T-1	跑	20220318	跑	20220321	跑	20220322	跑	20220323	跑	20220324	不跑		不跑		跑	20220328
		T-2	跑	20220317	跑	20220318	跑	20220321	跑	20220322	跑	20220323	不跑		不跑		跑	20220324
	跑批日历	T-1	不跑		跑	20220321	跑	20220322	跑	20220323	跑	20220324	不跑		跑	20220325	不跑	
		T-2	不跑		跑	20220318	跑	20220321	跑	20220322	跑	20220323	不跑		跑	20220324	不跑	
	交易跑批日历	T-1	不跑		跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	跑	20220325	不跑	
	自然日历	T-1	跑	20220320	跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	跑	20220326	跑	20220327
		T-2	跑	20220319	跑	20220320	跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	跑	20220326
16点	16点后翻牌为当日日期，16点前为昨日日期																	
16点后	交易日历	T-1	跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	不跑		不跑		跑	20220328
		T-2	跑	20220318	跑	20220321	跑	20220322	跑	20220323	跑	20220324	不跑		不跑		跑	20220325
	跑批日历	T-1	跑	20220321	跑	20220322	跑	20220323	跑	20220324	不跑		跑	20220325	不跑		跑	20220328
		T-2	跑	20220318	跑	20220321	跑	20220322	跑	20220323	不跑		跑	20220324	不跑		跑	20220325
	交易跑批日历	T-1	跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	跑	20220325	不跑		跑	20220328
	自然日历	T-1	跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	跑	20220326	跑	20220327	跑	20220328
		T-2	跑	20220320	跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	跑	20220326	跑	20220327
24点切日	自然日历	T-1	跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	跑	20220326	跑	20220327	跑	20220328
		T-2	跑	20220320	跑	20220321	跑	20220322	跑	20220323	跑	20220324	跑	20220325	跑	20220326	跑	20220327

PART · 04

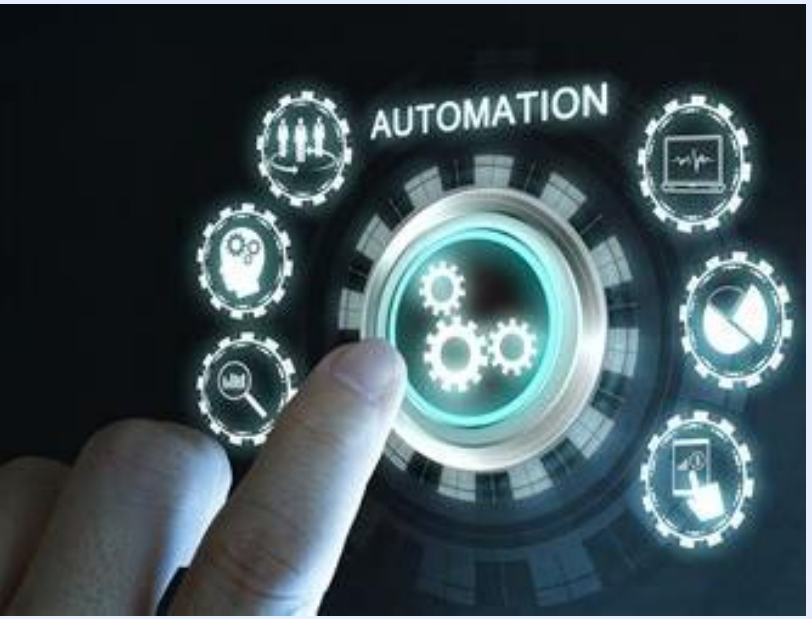
DataOps 未来

AI 加持下的 DataOps 未来

大模型在数据处理流程中可以扮演多种角色，提高整个数据处理流程的效率和智能化水平。大模型将应用于以下方面：

智能调度策略	数据处理涉及复杂的任务调度，大模型可以分析历史作业执行情况、资源使用状况，从而预测未来的工作流需求，智能地调度任务和分配资源。减少延迟，提高整体处理速度，并优化资源利用率。
数据质量检测与清洗	在数据同步过程中，大模型可以辅助自动检测数据质量问题，比如识别异常值、缺失数据或不一致性。通过机器学习算法，模型可以学习数据特征，自动清洗和修正数据，确保数据同步后的质量。
智能数据分类与标签	对于需要分类或标签化的数据，大模型可以自动分析数据内容，对其进行分类或附加有意义的标签，特别是在多模态数据处理场景下，这对于后续的数据分析和应用至关重要。
自适应数据同步策略	根据网络状况、数据变化频率和业务需求，大模型可以帮助动态调整数据同步策略，比如选择最合适的同步频率、确定优先级高的数据流，以优化同步效率和减少带宽消耗。
自动化异常处理	在数据传输或处理过程中遇到异常时，大模型可以基于历史数据和模式识别，自动识别异常原因并触发相应的处理机制，减少人工干预，提高处理效率。
Copilot	

AI 加持下的 DataOps 未来



智能化数据处理

随着AI技术的发展，DataOps将越来越智能化，实现更高效、自动化的数据处理。

数据安全保护

数据安全保护将成为DataOps未来发展的重要方向，如何在确保数据安全的前提下实现数据共享和利用将成为关键。

跨平台/云数据治理

随着企业业务的多平台化和全球化，跨平台的数据治理将成为DataOps的重要发展趋势。

相关资源

https://seatunnel.apache.org

https://dolphinscheduler.apache.org

白鲸开源官网： https://www.whaleops.com

51CTO WOT

谢 谢 观 看

THANKS

智启新纪
慧创万物
2024