

1.Taier简介

2.对比目前数据中台架构的优缺点

2.1 多租户多集群管理;目前中台也有租户管理,但Taier租户配置集群管理更优

2.2 数据源配置,依赖于DatasourceX的插件包;

2.2.1 目前中台没有数据源管理

2.3 数据同步

2.3.1 目前中台的数据同步工具是datax,是离线数据同步

2.3.2 Taier的数据同步底层是基于Flink分布式架构

2.4 实时/离线任务开发

2.4.1 目前中台主要是sparksql开发离线任务,然后通过接口启动任务,airflow调度

2.4.2 Taier 支持SparkSql 目前支持的版本spark-2.1.3; flinkSql后续会开源

2.5 任务调度

2.5.1 目前中台是用airflow调度,需要写调度脚本

2.5.1 Taier同样支持百万级并发调度能力,可视化任务调度配置,多种依赖与调度方式,与airflow区别不大; 但优点是Taier的任务调度是已经集成的功能,不需要额外写调度脚本,额外的调度工具,可以在线开发sparkSql任务,直接提交到调度

2.6 任务运维

2.6.1 目前中台和Taier的任务运维相差不大,都具有下面功能

2.7 监控告警

2.7.1 目前中台通过airflow 有邮件监控告警机制

2.7.1 Taier 目前没有邮件监控告警机制

2.8 kerberos

2.8.1 目前中台没有支持kerberos

2.8.2 Taier支持kerberos

2.9 Taier官方文档介绍的优点

1.Taier简介

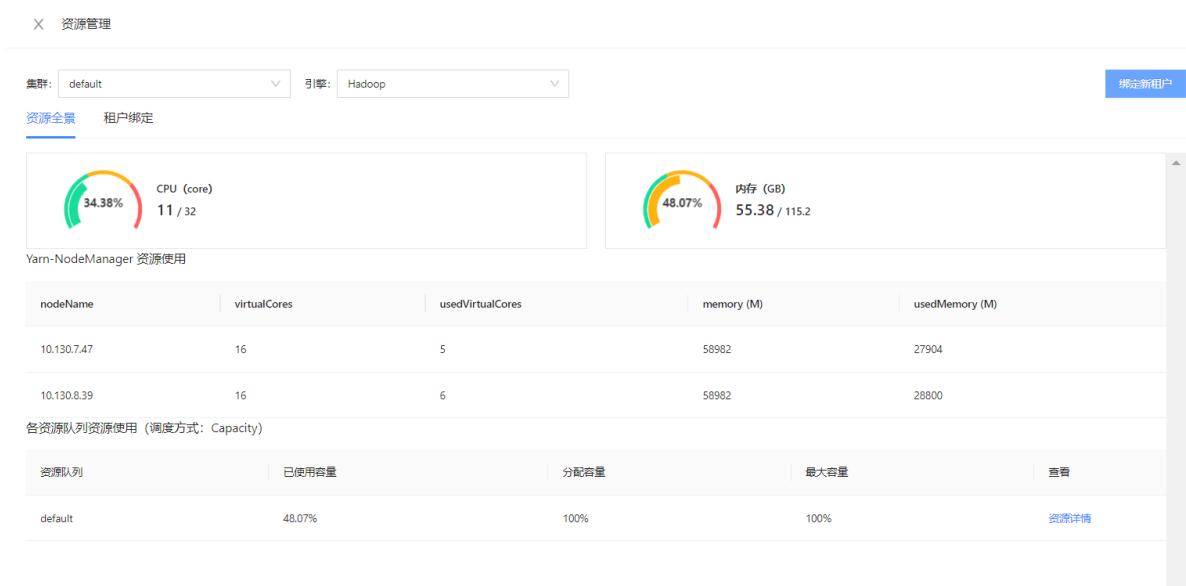
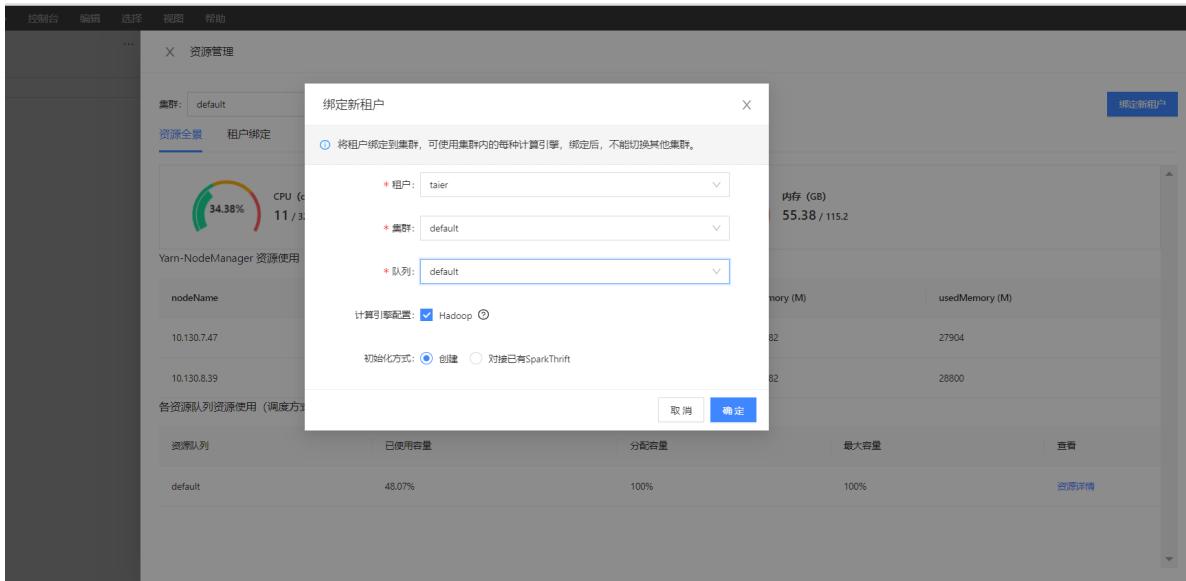
官文文档:

<https://dtstack.github.io/Taier/docs/guides/introduction>

2.对比目前数据中台架构的优缺点

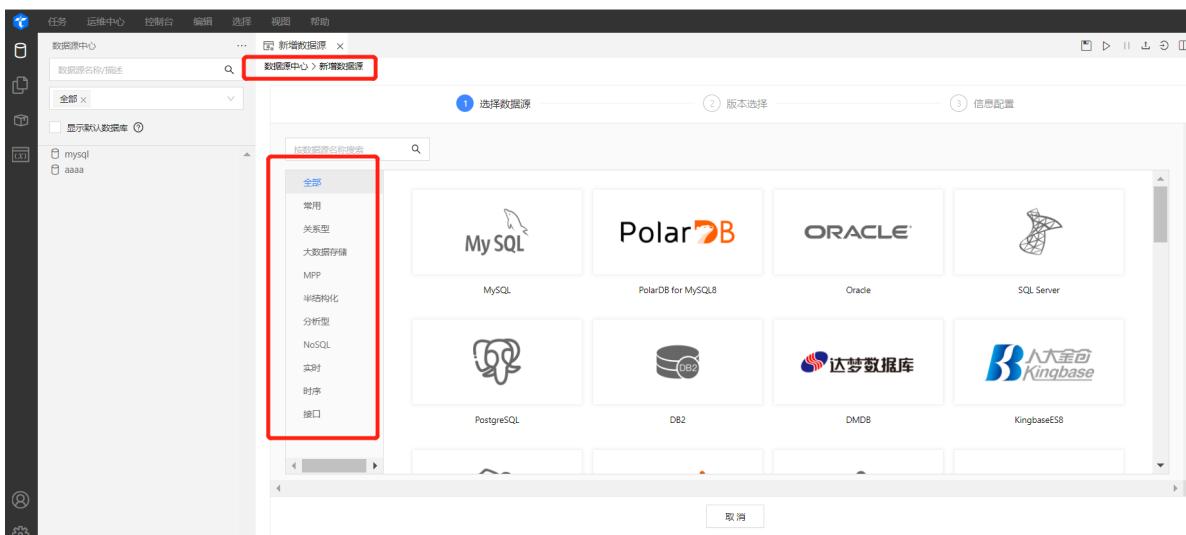
2.1 多租户多集群管理;目前中台也有租户管理,但Taier租户配置集群管理更优

不同的集群使用不同的资源,然后让不同环境下的租户绑定到不同的集群,比如测试环境租户绑定到测试集群、预发环境租户绑定预发集群,从而实现对资源的隔离使用



2.2 数据源配置, 依赖于DatasourceX的插件包;

2.2.1 目前中台没有数据源管理



2.3 数据同步

2.3.1 目前中台的数据同步工具是datax,是离线数据同步

2.3.2 Taier的数据同步底层是基于Flink分布式架构

1. 支持对10+种存储系统进行数据读/写,支持离线同步和实时同步,对比datax的优点是支持实时同步

	Database Type	Source	Sink	Lookup
Batch Synchronization	MySQL	doc	doc	doc
	TiDB		reference mysql	reference mysql
	Oracle	doc	doc	doc
	Doris		doc	
	SqlServer	doc	doc	doc
	PostgreSQL	doc	doc	doc
	DB2	doc	doc	doc
	ClickHouse	doc	doc	doc
	Greenplum	doc	doc	
	KingBase	doc	doc	
Stream Synchronization	MongoDB	doc	doc	doc
	SAP HANA	doc	doc	
	ElasticSearch7	doc	doc	doc
	FTP	doc	doc	
	HDFS	doc	doc	
	Stream	doc	doc	
	Redis		doc	doc
	Hive		doc	
	Solr	doc	doc	
	File	doc		

2. 支持可视化配置,字段可以自由映射,不需要写json脚本

测试数据同步

任务开发 > 测试数据同步

1 数据来源 2 选择目标 3 字段映射 4 通道控制 5 预览保存

* 数据源: mysql (MySQL)

* 表名(批量): ods_bugsdaily

数据过滤①: 请参考相关SQL语法填写where过滤语句(不要填写where关键字)。该过滤语句通常用作增量同步

切分键②:

高级配置③: 以JSON格式添加高级参数, 例如对关系型数据库可配置fetchSize

[数据预览](#)

[下一步](#)

测试数据同步

任务开发 > 测试数据同步

1 数据来源 2 选择目标 3 字段映射 4 通道控制 5 预览保存

* 数据同步目标: mysql (MySQL)

* 表名: ods_bugsdaily_copy1

一键生成目标表

导入前准备语句: 请输入导入数据前执行的SQL脚本

导入后准备语句: 请输入导入数据后执行的SQL脚本

* 主键冲突: insert into (当主键/约束冲突, 报脏数据)

高级配置④: 以JSON格式添加高级参数, 例如对关系型数据库可配置fetchSize

测试数据同步

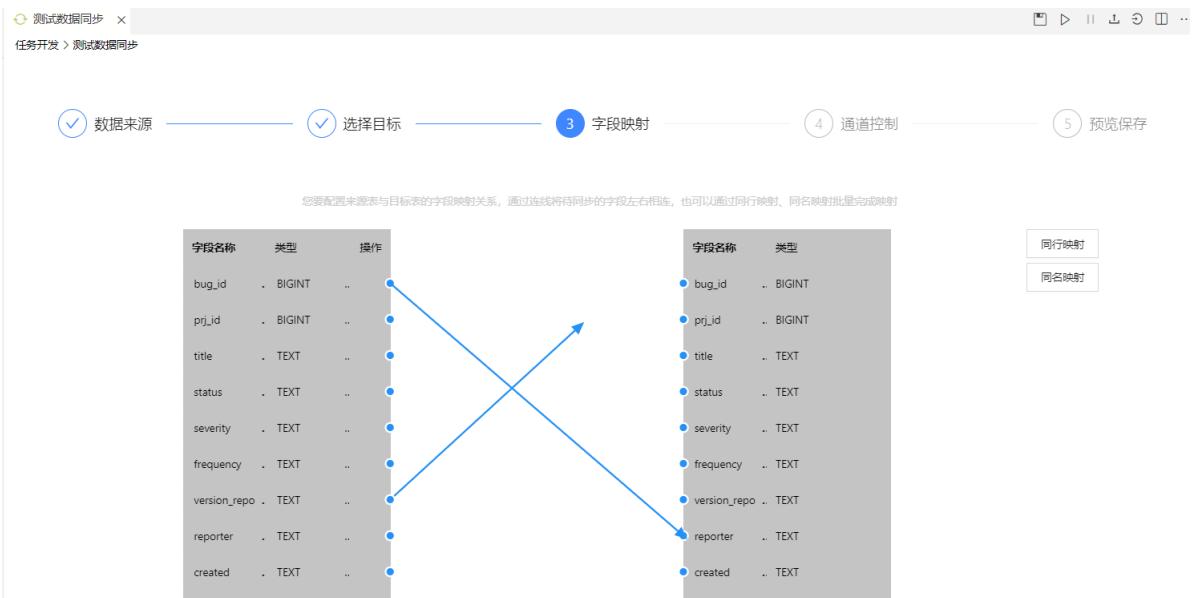
任务开发 > 测试数据同步

1 数据来源 2 选择目标 3 字段映射 4 通道控制 5 预览保存

您要配置来源表与目标表的字段映射关系, 通过连线将待同步的字段左右相连, 也可以通过同行映射、同名映射批量完成映射

字段名称	类型	操作	字段名称	类型
bug_id	BIGINT	..	bug_id	BIGINT
prj_id	BIGINT	..	prj_id	BIGINT
title	TEXT	..	title	TEXT
status	TEXT	..	status	TEXT
severity	TEXT	..	severity	TEXT
frequency	TEXT	..	frequency	TEXT
version_repo	TEXT	..	version_repo	TEXT
reporter	TEXT	..	reporter	TEXT
created	TEXT	..	created	TEXT
in_progress_i	TEXT	..	in_progress_i	TEXT

同行映射
 同名映射



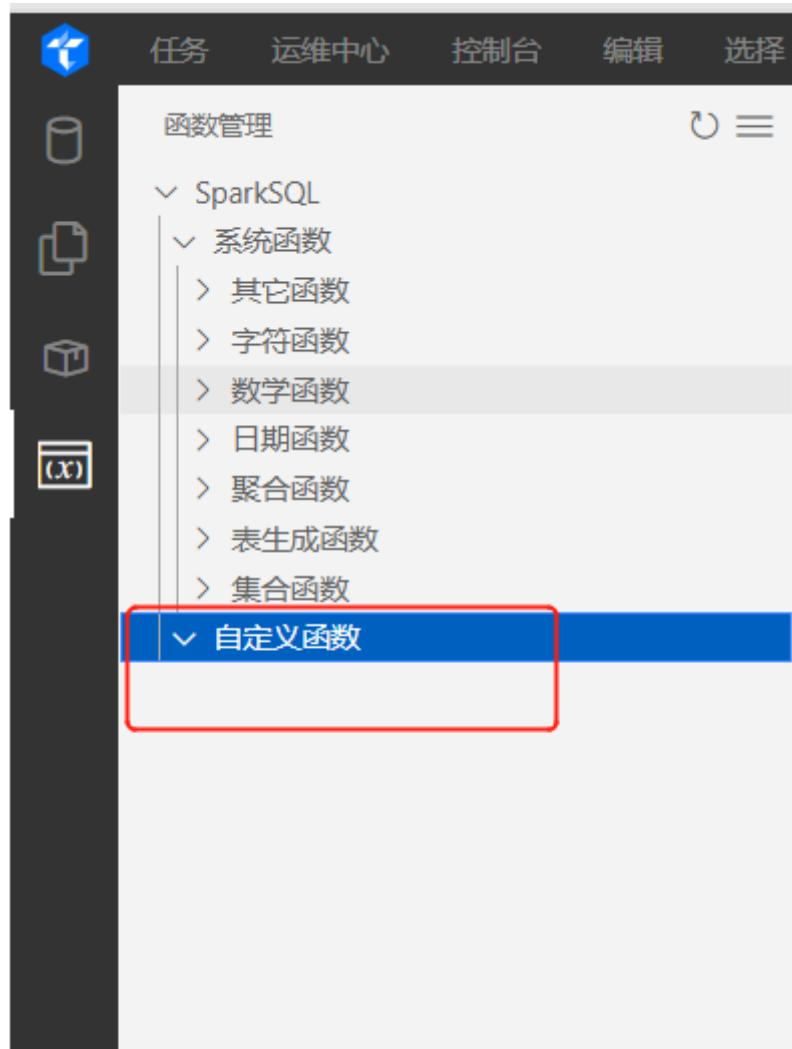
3. 支持传输速率设置, 脏数据管理, 支持断点续传, 是基于flink的快照

2.4 实时/离线任务开发

2.4.1 目前中台主要是sparksql开发离线任务, 然后通过接口启动任务, airflow调度

2.4.2 Taier 支持SparkSql 目前支持的版本spark-2.1.3; flinkSql后续会开源

1. 自定义函数



3. 可以直接前端页面写sparkSql任务,页面操作运行,提交调度

```

1 create table if not exists ods_order_header (
2     ,order_header_id string comment '订单头id'
3     ,order_date bigint comment '订单日期'
4     ,shop_id string comment '店铺id'
5     ,customer_id string comment '客户id'
6     ,order_status bigint comment '订单状态'
7     ,pay_date bigint comment '支付日期'
8
9 )comment '销售订单明细表'
10 PARTITIONED BY (ds string);
11
12 create table if not exists ods_order_detail (
13     ,order_header_id string comment '订单头id'
14     ,order_detail_id string comment '订单明细分id'
15     ,item_id string comment '商品id'
16     ,count int double comment '商品数量'
17     ,unit_price double comment '商品单价'
18     ,dist_amount double comment '折扣金额'
19 )comment '销售订单明细分表';

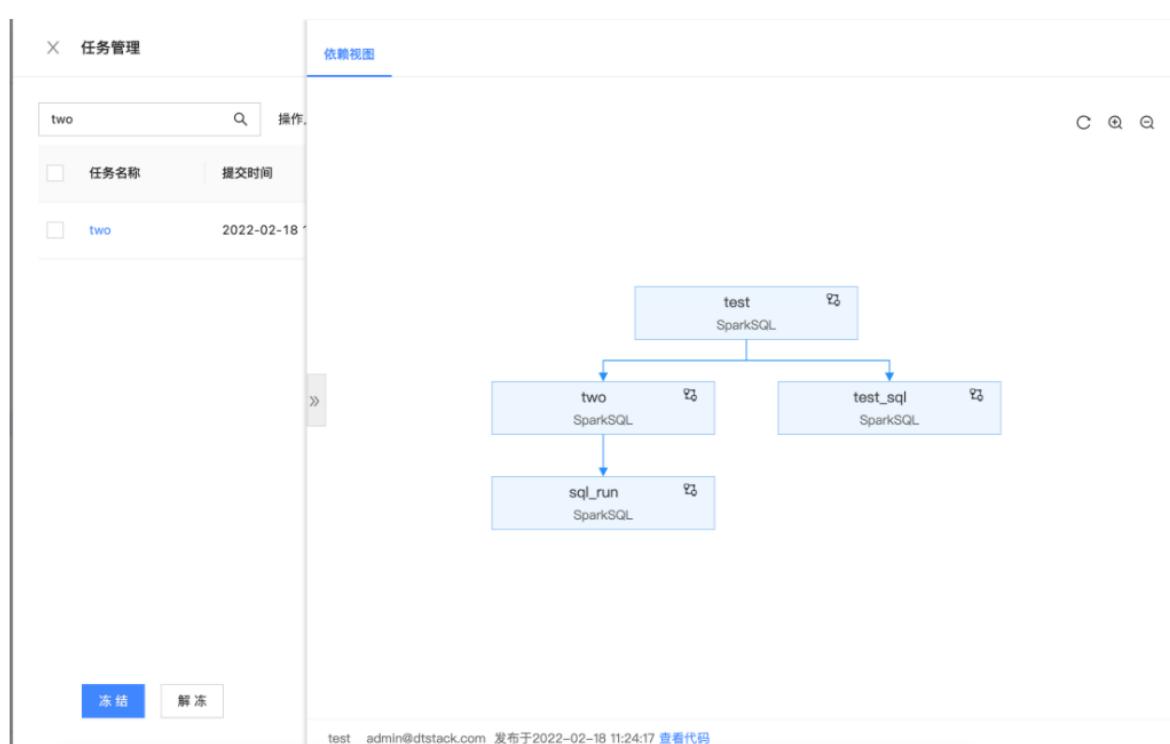
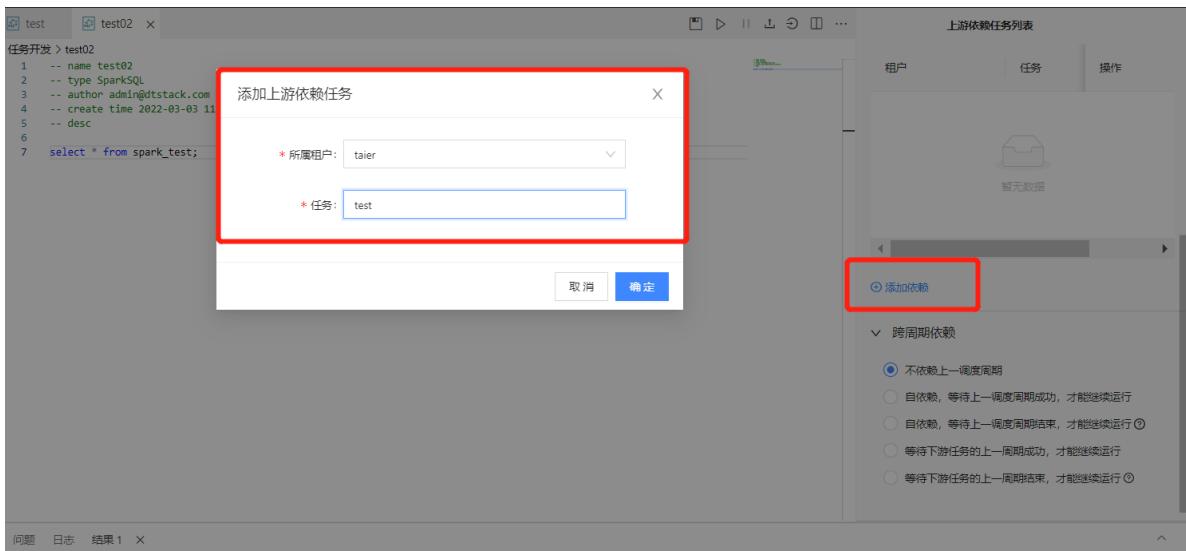
```

2.5 任务调度

2.5.1 目前中台是用airflow调度,需要写调度脚本

2.5.1 Taier同样支持百万级并发调度能力,可视化任务调度配置,多种依赖与调度方式,与airflow区别不大; 但优点是Taier的任务调度是已经集成的功能,不需要额外写调度脚本,额外的调度工具,可以在线开发sparkSql任务,直接提交到调度

```
任务 运维中心 控制台 编辑 选择 视图 帮助  
浏览 > 打开的编辑器  
任务开发  
> 原始数据层(ODS)  
> 敏感明层(DWD)  
> 敏仓汇总层(DWS)  
> 应用数据层(ADS)  
> 数据维度(DIM)  
test  
○ 测试数据同步  
... (code editor window)  
提交至调度
```



2.6 任务运维

2.6.1 目前中台和Taier的任务运维相差不大,都具有下面功能

1. 任务上下游追溯
2. 支持修复重跑, 暂停, kill等多种操作方式
3. 明细日志追踪定位

任务管理						
按任务名称搜索		操作人:	请选择操作人	<input type="checkbox"/> 我的任务	<input type="checkbox"/> 今日修改的任务	<input type="checkbox"/> 冻结的任务
任务名称	提交时间	任务类型	调度周期	操作人	操作	
test	2022-03-03 11:47:04	SparkSQL	天任务	admin@dtstack.com	补数据	修改
test02	2022-03-03 11:48:41	SparkSQL	天任务	admin@dtstack.com	补数据	修改

2.7 监控告警

2.7.1 目前中台通过airflow 有邮件监控告警机制

2.7.1 Taier 目前没有邮件监控告警机制

2.8 kerberos

2.8.1 目前中台没有支持kerberos

2.8.2 Taier支持kerberos

2.9 Taier官方文档介绍的优点

稳定性

- 单点故障: 去中心化的分布式模式
- 高可用方式: `Zookeeper`
- 过载处理: `分布式节点` + `两级存储策略` + `队列机制`。每个节点都可以处理任务调度与提交; 任务多时会优先缓存在内存队列, 超出可配置的队列最大数量值后会全部落数据库; 任务处理以队列方式消费, 队列异步从数据库获取可执行实例
- 实战检验: 得到数百家企业客户生产环境实战检验

易用性

- 支持大数据作业 `Spark`、`Flink` 的调度，
- 支持众多的任务类型，目前支持 `Spark SQL`、`Flinkx`



TIP
后续将开源： `SparkMR`、`PySpark`、`FlinkMR`、`Python`、`Shell`、`Jupyter`、`Tensorflow`、`Pytorch`、`HadoopMR`、`Kylin`、`Odps`、SQL类任务(`MySQL`、`PostgreSQL`、`Hive`、`Impala`、`Oracle`、`SQLServer`、`TiDB`、带格式的:突出显示 `greenplum`、`inceptor`、`kingbase`、`presto`)

- 可视化工作流配置：支持封装工作流、支持单任务运行，不必封装工作流、支持拖拽模式绘制DAG
- DAG监控界面：运维中心、支持集群资源查看，了解当前集群资源的剩余情况、支持对调度队列中的任务批量停止、任务状态、任务类型、重试次数、任务运行机器、可视化变量等关键信息一目了然
- 调度时间配置：可视化配置
- 多集群连接：支持一套调度系统连接多套 `Hadoop` 集群

多版本引擎

- 支持 `Spark`、`Flink` 等引擎的多个版本共存，例如可同时支持 `Flink1.10`、`Flink1.12`（后续开源）

Kerberos支持

- `Spark`、`Flink`

系统参数

- 丰富，支持3种时间基准，且可以灵活设置输出格式

扩展性

- 设计之初就考虑分布式模式，目前支持整体 `Taier` 水平扩容方式；
- 调度能力随集群线性增长；