Analysys 易观
你要的数据能力

Apache DolphinScheduler

- 分布式易扩展的可视化 ETL 调度系统

代立冬
PPMC & 易观大数据平台总监

DolphinScheduler 部分用户案例(排名不分先后)

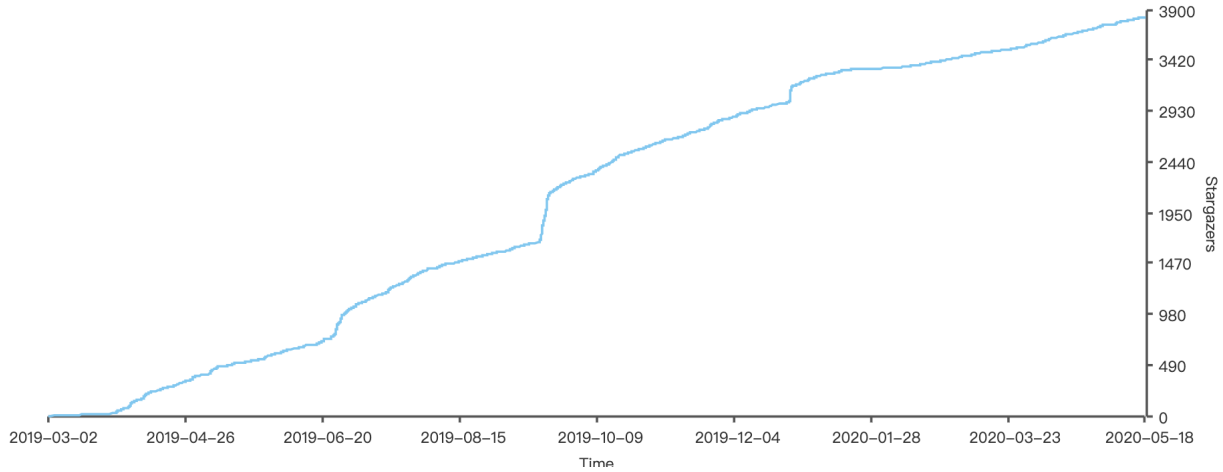
Analysys 易观 你要的数据能力

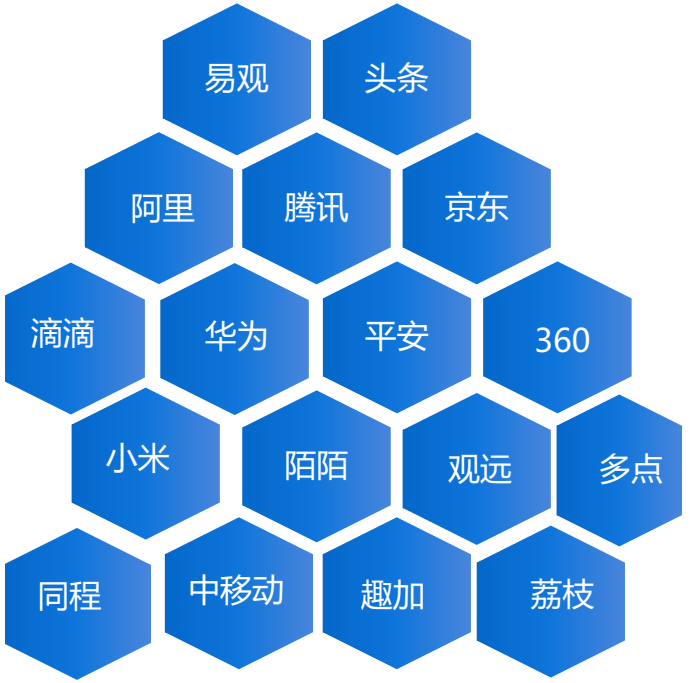
DolphinScheduler 社区情况

Analysys 易观
你要的数据能力

- Apache DolphinScheduler Group 3 (481)
- Apache DolphinScheduler Group 4 (471)
- Apache DolphinScheduler Group 2 (484)
- Apache DolphinScheduler Group 1 (496)
- DolphinScheduler Developer Group (163)
- Apache DolphinScheduler Group 5 (480)
- Apache DolphinScheduler Group 6 (93)
- DolphinScheduler 直播讲师群 (10)
- DolphinScheduler PPMC Group (18)
- DolphinScheduler 前沿用户研发讨论 (39)

Dolphin Scheduler for Big Data



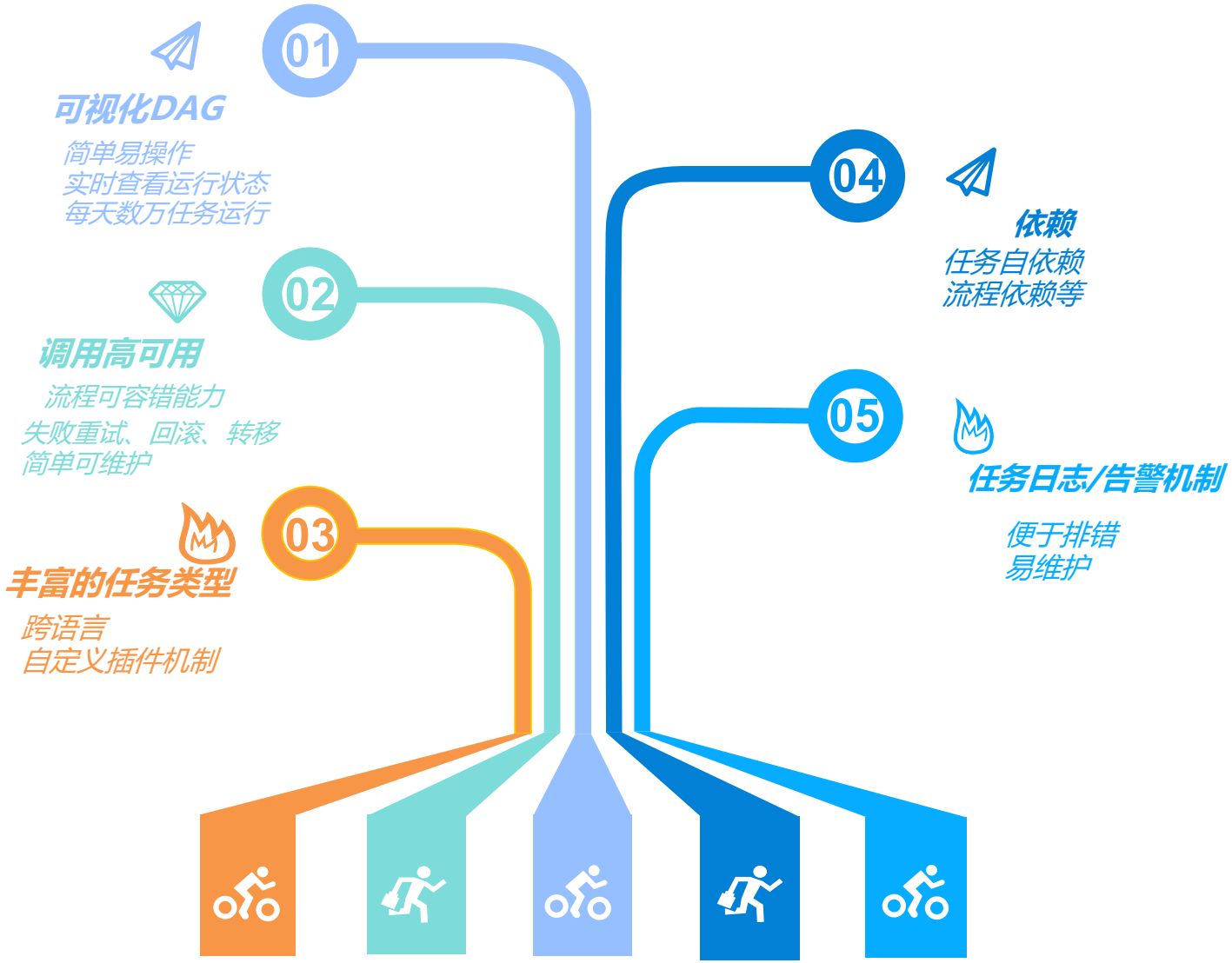


贡献者分布

缘何研发DolphinScheduler？

Analysys 易观
你要的数据能力

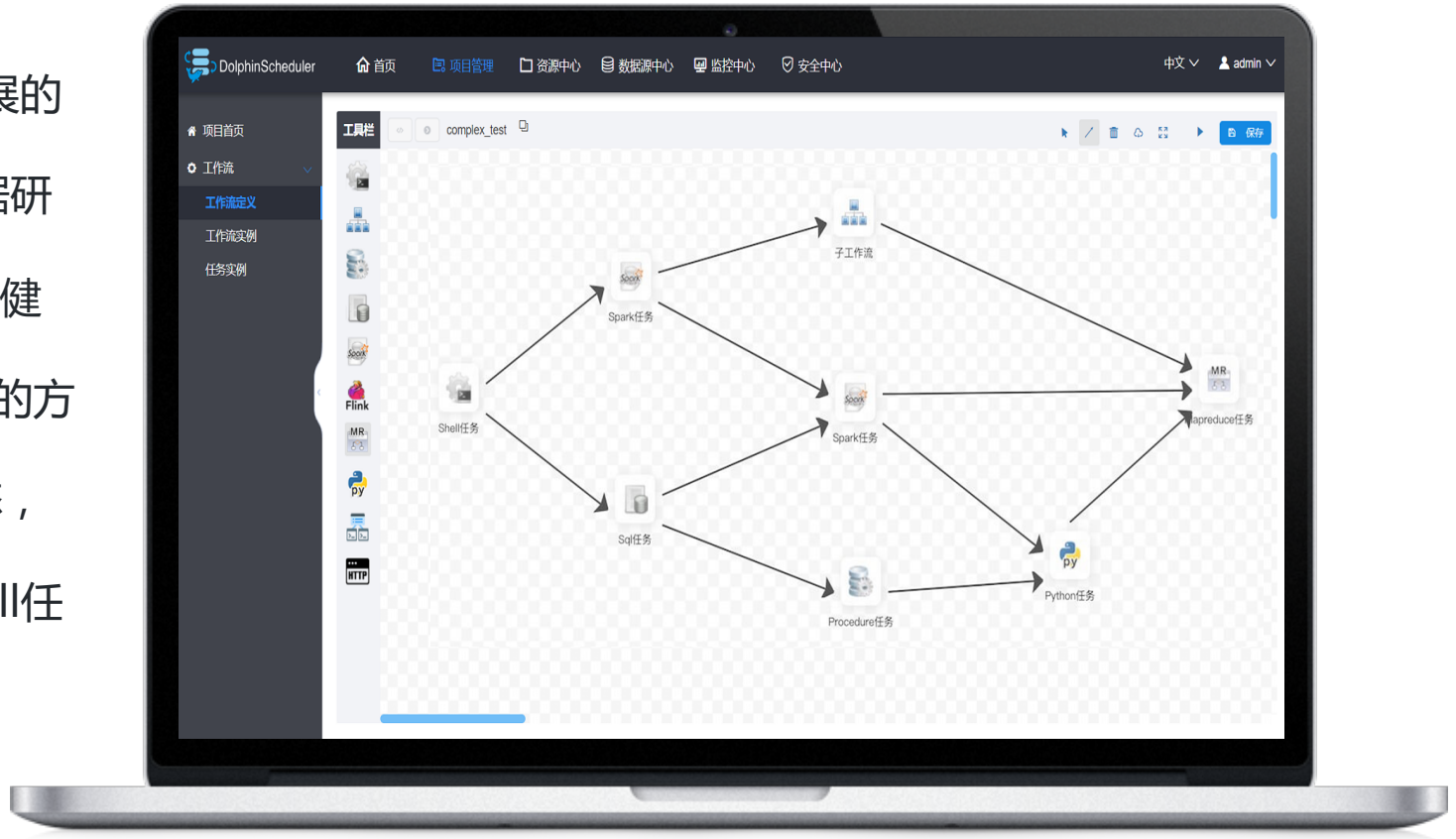
调度需求



Apache DolphinScheduler 简介

Analysys 易观
你要的数据能力

Apache DolphinScheduler是一个分布式易扩展的可视化DAG工作流任务调度开源系统。解决数据研发ETL 错综复杂的依赖关系，不能直观监控任务健康状态等问题。DolphinScheduler以DAG流式的方式将Task组装起来，可实时监控任务的运行状态，同时支持重试、从指定节点恢复失败、暂停及Kill任务等操作



Apache DolphinScheduler 特性

Analysys 易观
你要的数据能力

高可靠性

去中心化的多Master和多Worker, 自身支持HA功能, 实现超大规模任务调度，采用任务队列和自身保护机制来避免过载，不会造成机器卡死

简单易用

DAG监控界面，所有流程定义都是可视化，通过拖拽任务定制DAG，通过API方式与第三方系统对接, 一键部署

丰富的使用场景

支持暂停恢复操作. 支持多租户，更好的应对大数据的使用场景. 支持更多的任务类型，如 spark, hive, mr, python, sub_process, shell

高扩展性

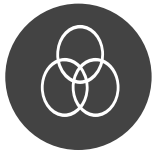
支持自定义任务类型，调度器使用分布式调度，调度能力随集群线性增长，Master和Worker支持动态上下线

Apache DolphinScheduler 能力

Analysys 易观
你要的数据能力

- Task以DAG形式关联，实时监控任务的状态。

- 支持Shell、MR、Spark、SQL、依赖等10多种任务类型。



- 去中心化设计确保系统的稳定、高可用。

- 工作流优先级、任务优先级，全局参数及局部自定义参数

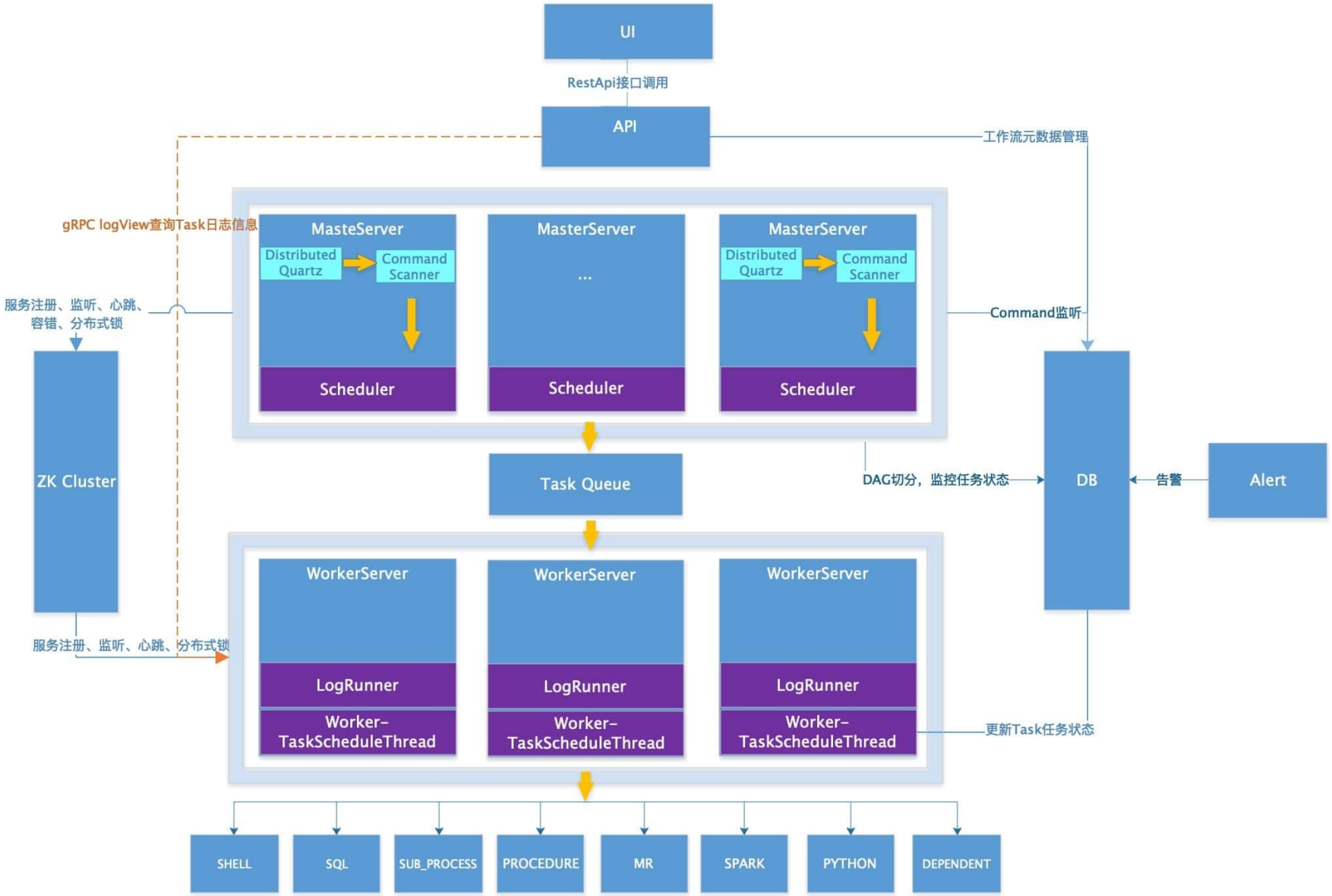
- 工作流可定时、依赖、手动、暂停/停止/恢复

- 完善的系统服务监控，任务超时告警/失败。

- 支持补数、多租户、日志在线查看及资源在线管理

- 支持每日十万数据量级任务稳定运行

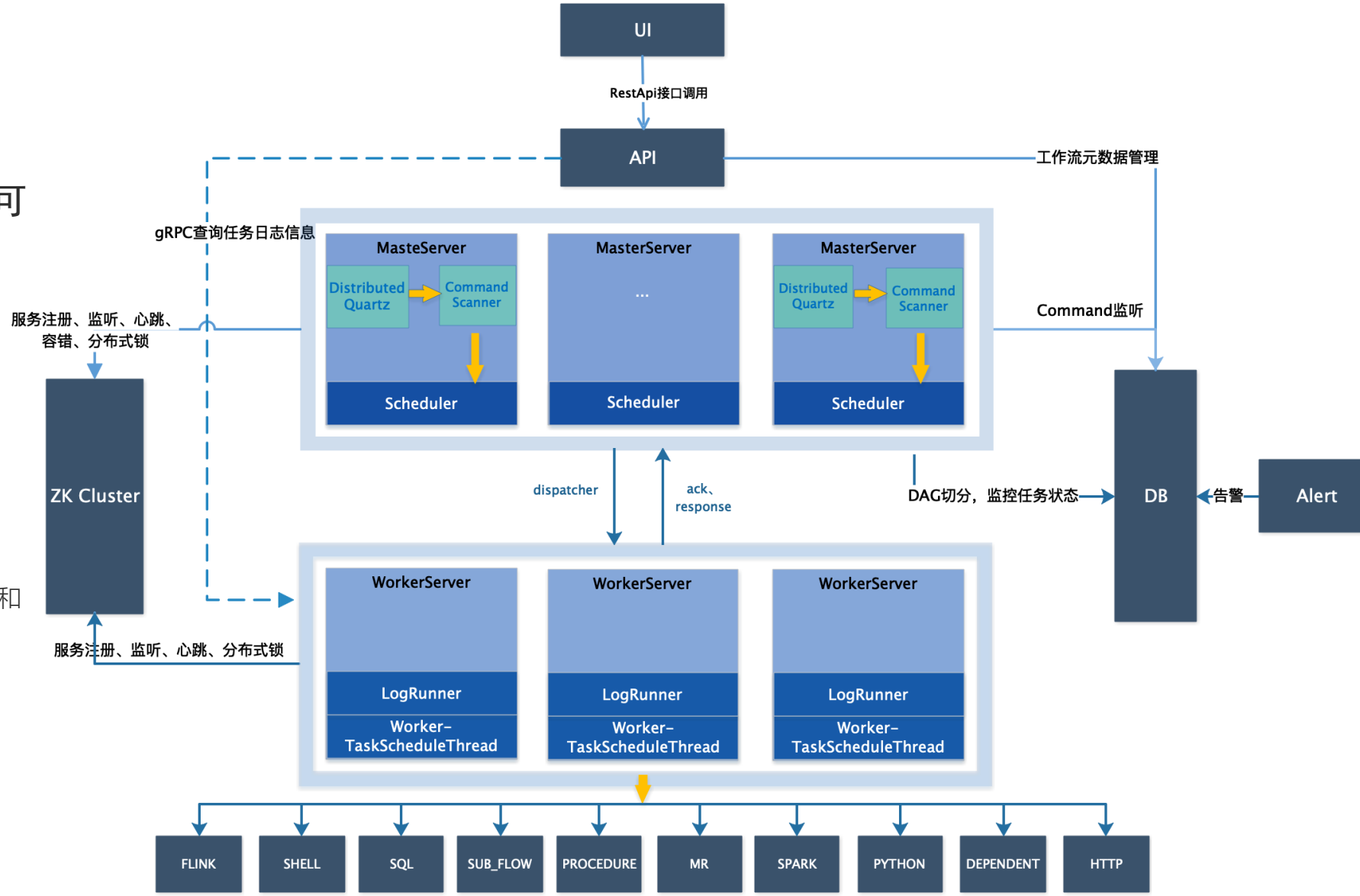
DolphinScheduler 前世



DolphinScheduler 今生 - 1.3.0

Analysys 易观
你要的数据能力

- ✓ 数据库减压，减少极端情况下的可能造成的调度延时
- ✓ Worker去DB、职责更单一
- ✓ Master和Worker直接通信，降低延时
- ✓ Master多种策略分发任务
  - Worker节点的三种选择：随机、循环和CPU和内存的线性加权负载平衡



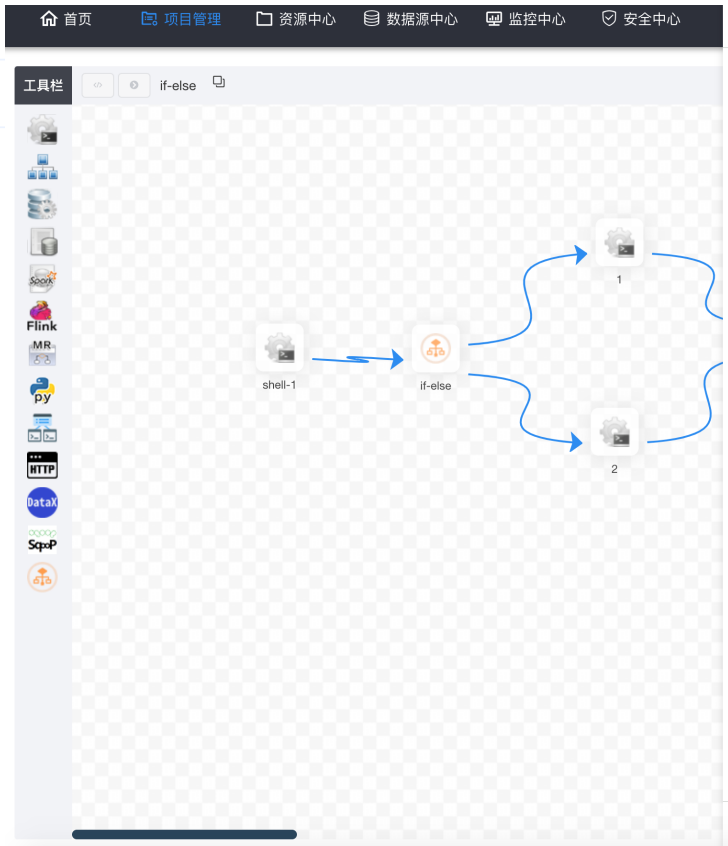
DolphinScheduler 1.3.0 新特性 – 资源中心多目录

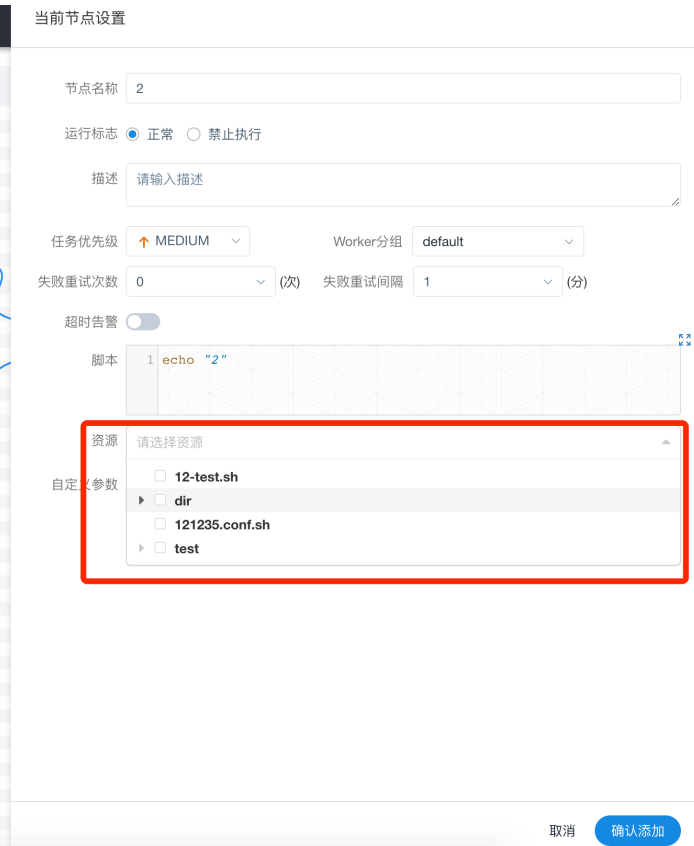
Analysys 易观
你要的数据能力

文件管理

创建文件夹 | 创建文件 | 上传文件

编号	名称	是否文件夹	文件名称
1	test	是	test
2	121235.conf.sh	否	12345.sh
3	dir	是	dir
4	12-test.sh	否	12-test.sh





DolphinScheduler 1.3.0 新特性 – Datax

Analysys 易观
你要的数据能力

当前节点设置

节点名称 请输入名称(必填)

运行标志 正常 禁止执行

描述 请输入描述

任务优先级 MEDIUM Worker分组 default

失败重试次数 0 (次) 失败重试间隔 1 (分)

超时告警

自定义模版

数据源 MYSQL 请选择数据源

sql语句 1

目标库 MYSQL 请选择数据源

目标表 请输入目标表名

目标库前置sql

目标库后置sql

限流(字节数) 0 (KB，0代表不限制)

限流(记录数) 1000 (0代表不限制)

当前节点设置

节点名称 请输入名称(必填)

运行标志 正常 禁止执行

描述 请输入描述

任务优先级 MEDIUM Worker分组 default

失败重试次数 0 (次) 失败重试间隔 1 (分)

超时告警

自定义模版

json 1

自定义参数 prop(必填) value(选填)

prop(必填) value(选填)

prop(必填) value(选填)

自定义模板

DolphinScheduler 1.3.0 新特性 – Sqoop

当前节点设置

流向 import

数据来源

类型 MYSQL

数据源 MYSQL 请选择数据源

模式 表单 SQL

sql语句 1

Hive类型映射

Java类型映射

数据目的

类型 HDFS

目标路径 请输入目标路径(必填)

是否删除目录

压缩类型 snappy lzo gzip no

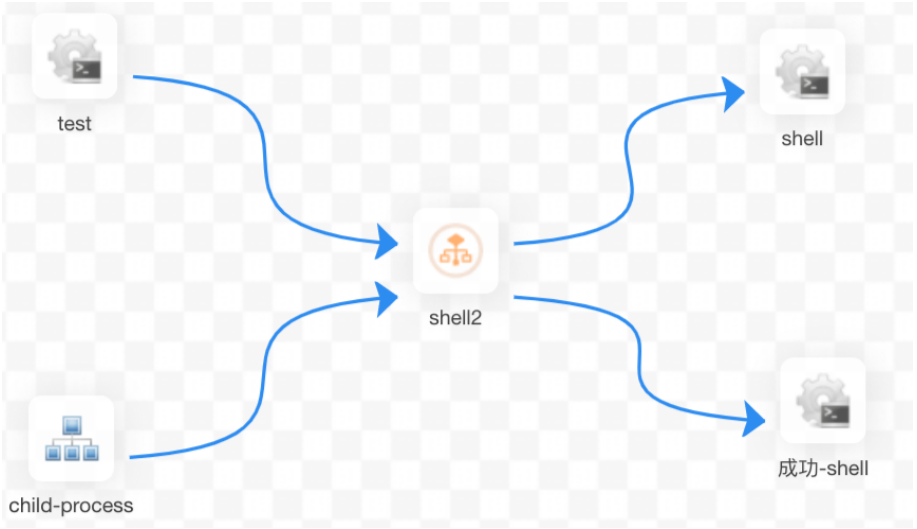
保存格式 avro sequence text parquet

列分隔符 请输入列分隔符

行分隔符 请输入行分隔符

并发度 1

DolphinScheduler 1.3.0 新特性 – 条件分支



当前节点设置

节点名称 shell2

运行标志 正常 禁止执行

描述 请输入描述

任务优先级 MEDIUM

Worker分组 default

失败重试次数 0 (次)

失败重试间隔 1 (分)

状态 成功

分支流转 成功-shell

状态 失败

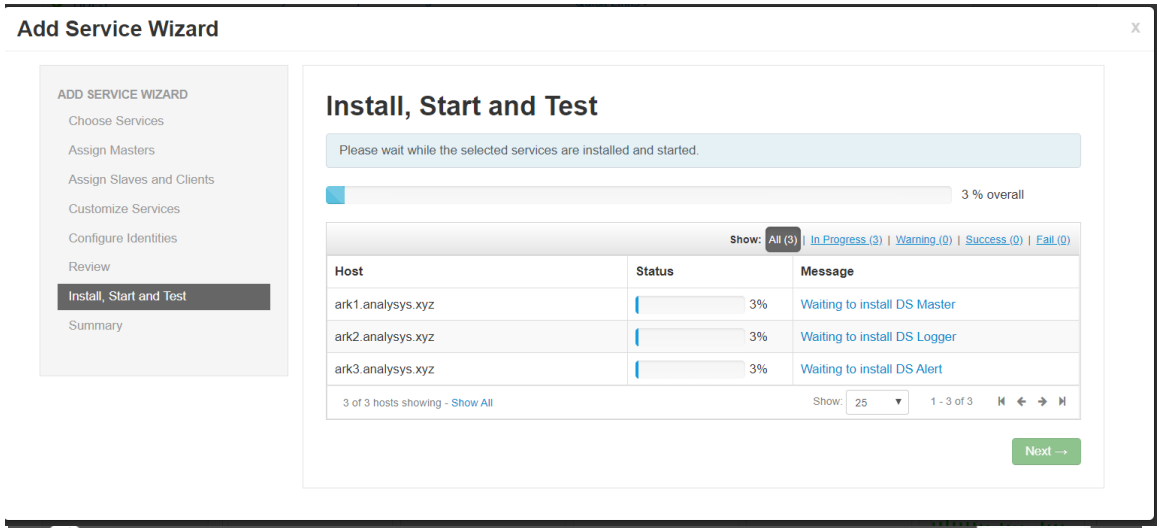
分支流转 shell

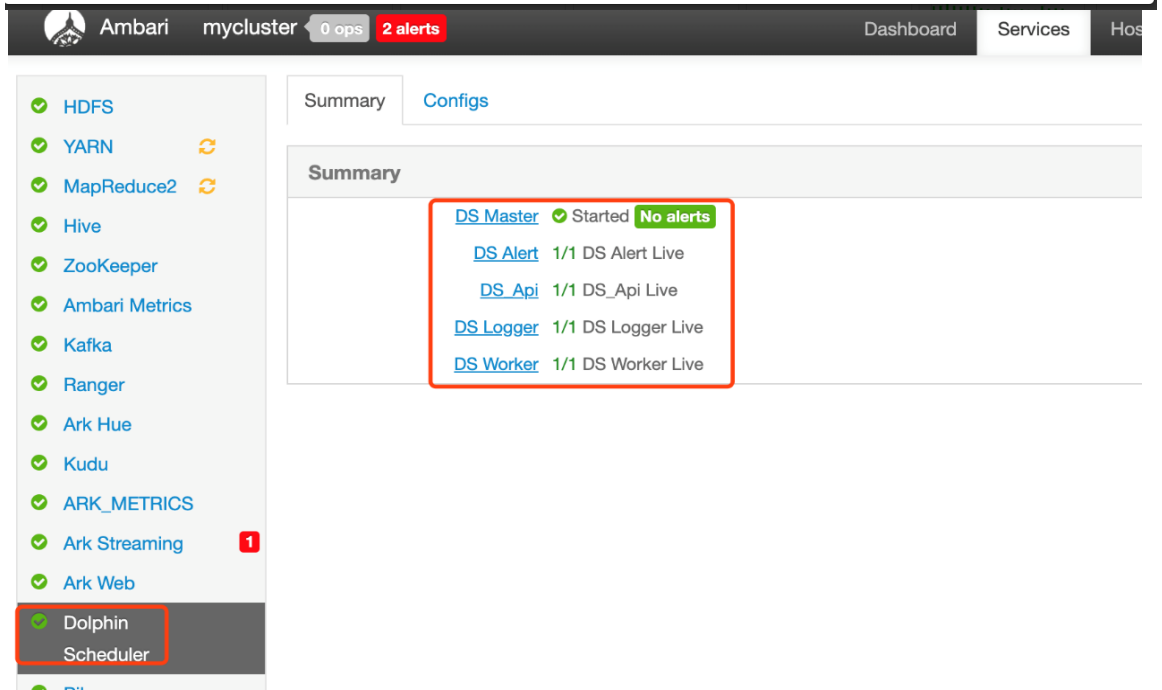
超时告警

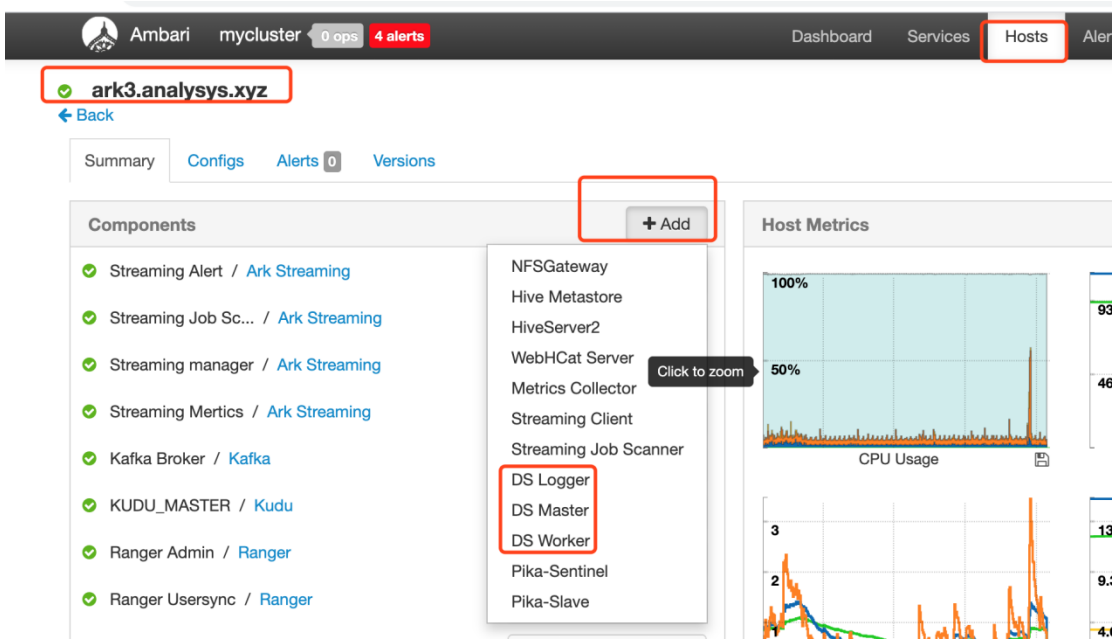
自定义参数

DolphinScheduler 1.3.0 新特性 – Ambari插件

Analysys 易观
你要的数据能力





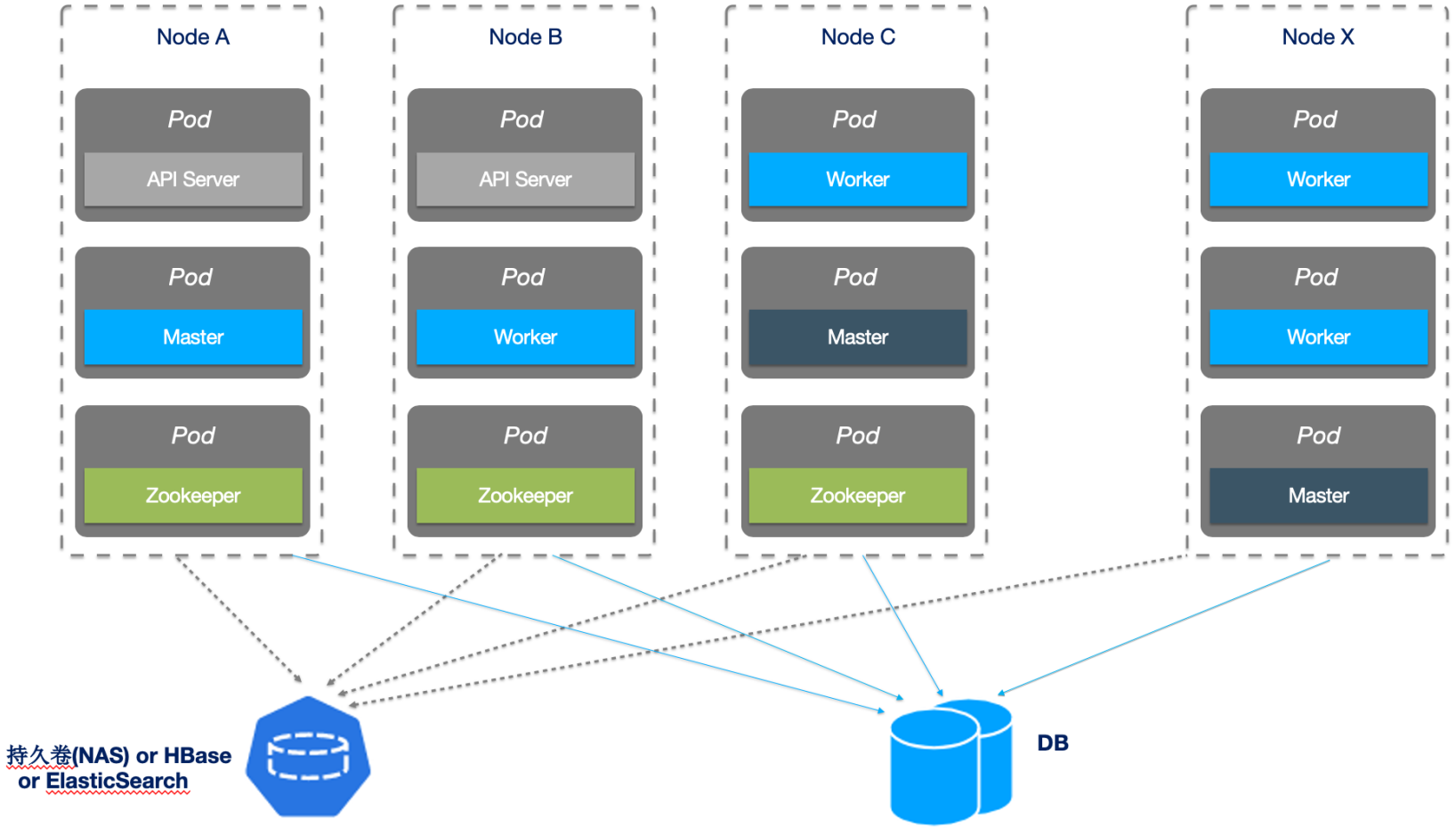


DolphinScheduler 1.3.0 新特性 – K8S 支持

优点：
- 动态扩展
- Graceful shutdown
- 本身的维护成本低

缺点:
- K8S运维经验

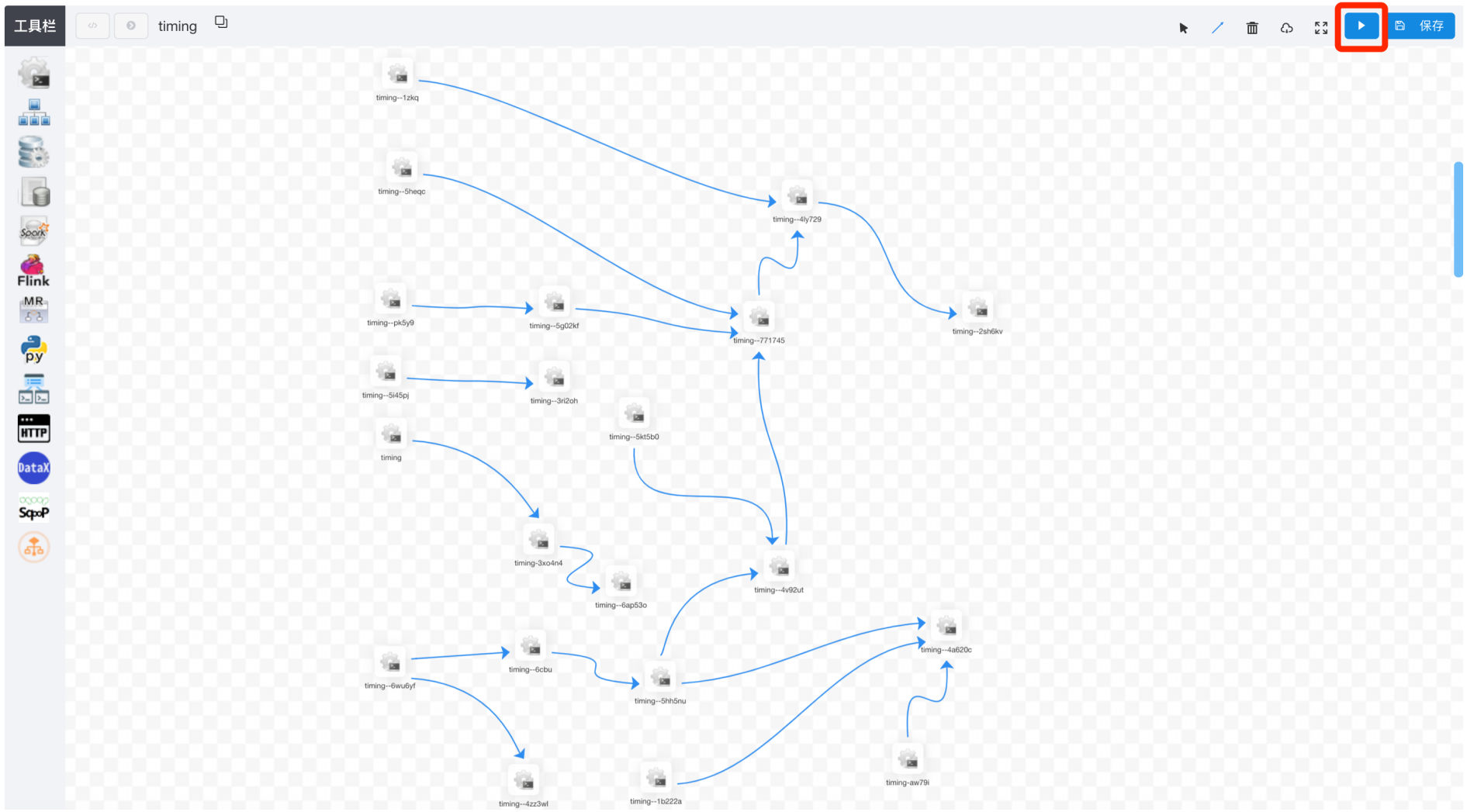
待实现：
- 任务Log需要持久化共享存储



DolphinScheduler 1.3.0 新特性 – DAG一键格式化

Analysys 易观
你要的数据能力

适合open api调用场景

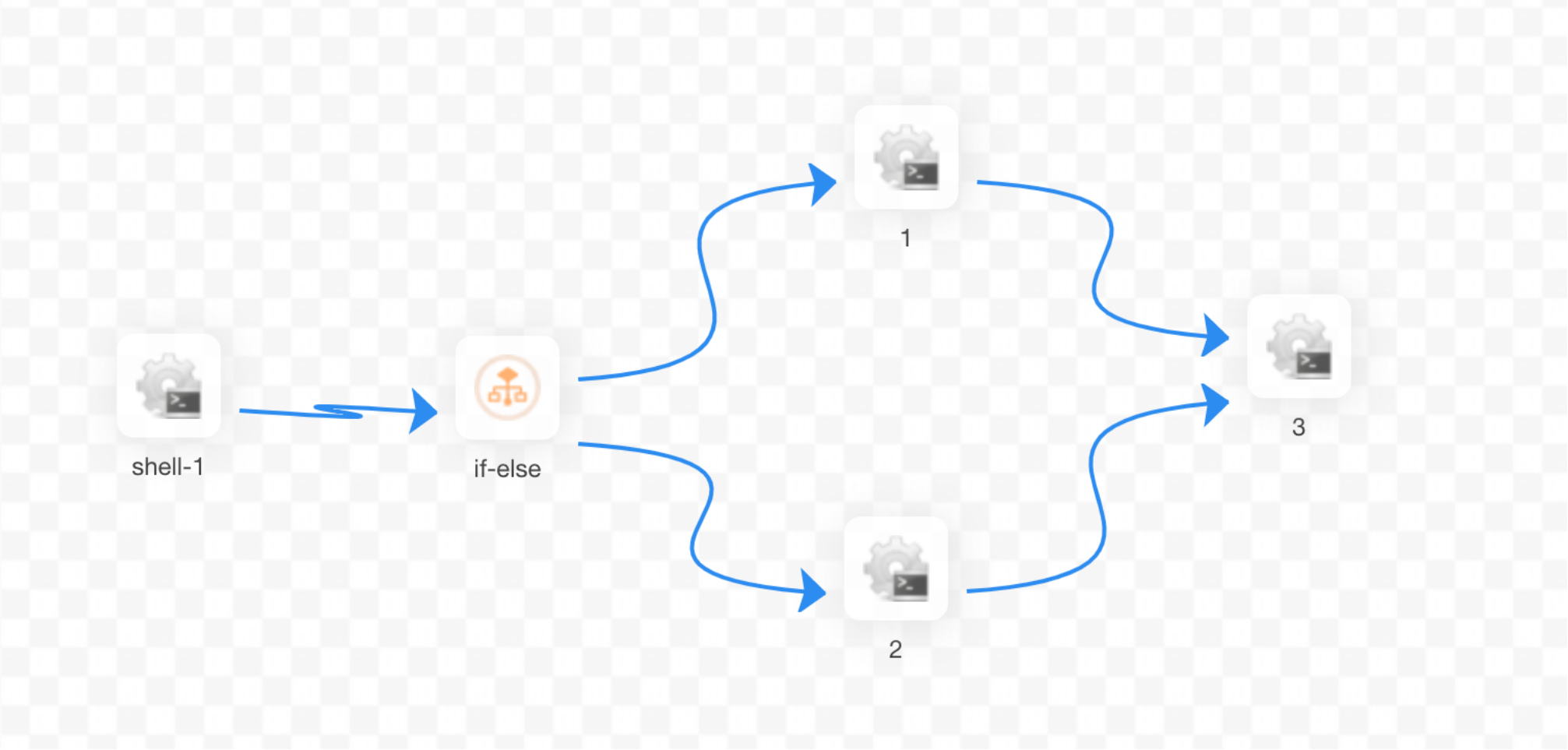


DolphinScheduler 1.3.0 新特性

Analysys 易观
你要的数据能力

- ✓ 支持Windows系统运行任务
- ✓ 批量导出和导入工作流
- ✓ 工作流复制
- ✓ 删流程实例级联删除任务日志
- ✓ 简化配置，优化部署体验
- ✓ 完善自动化CI、CD
- ✓ 添加钉钉告警

DolphinScheduler 1.3.0 新特性 – 流程图美化



DolphinScheduler Roadmap

Analysys 易观
你要的数据能力

总体依照社区需求和关注度来安排功能优先级

- master重构：建立 api 和 master 直接通信等
- 任务参数传递
- 任务类型插件化Plugin
- 工作流触发
- 数据质量
- 工作流血缘关系
- 列表依赖(上游依赖)
- 告警服务化，提供API
- 支持多集群上线发布
- 工作流版本管理
- 权限改造
- Easy to use

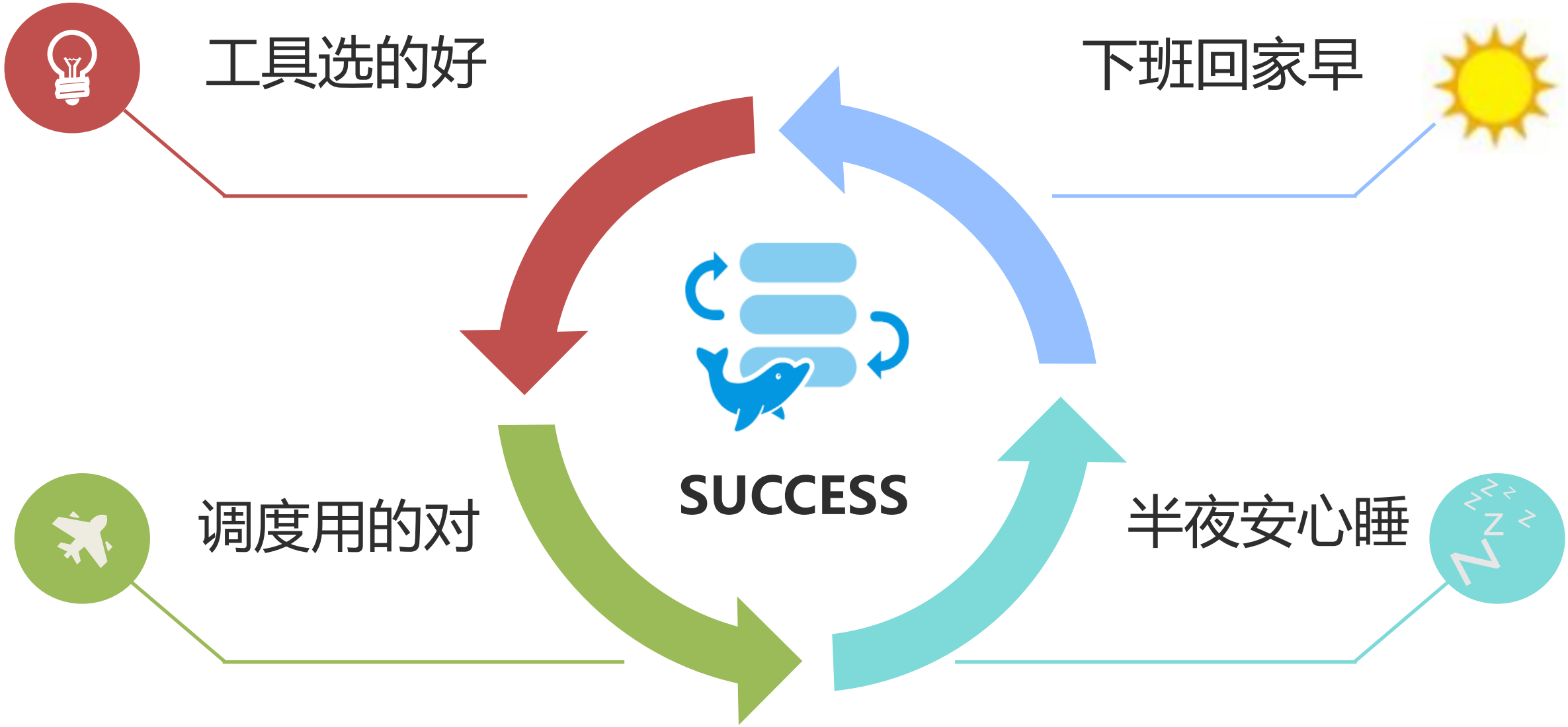
如果有好建议或有兴趣，欢迎邮件讨论

更多参考 work plan: https://github.com/apache/incubator-dolphinscheduler/projects/1

DolphinScheduler 项目发展历程

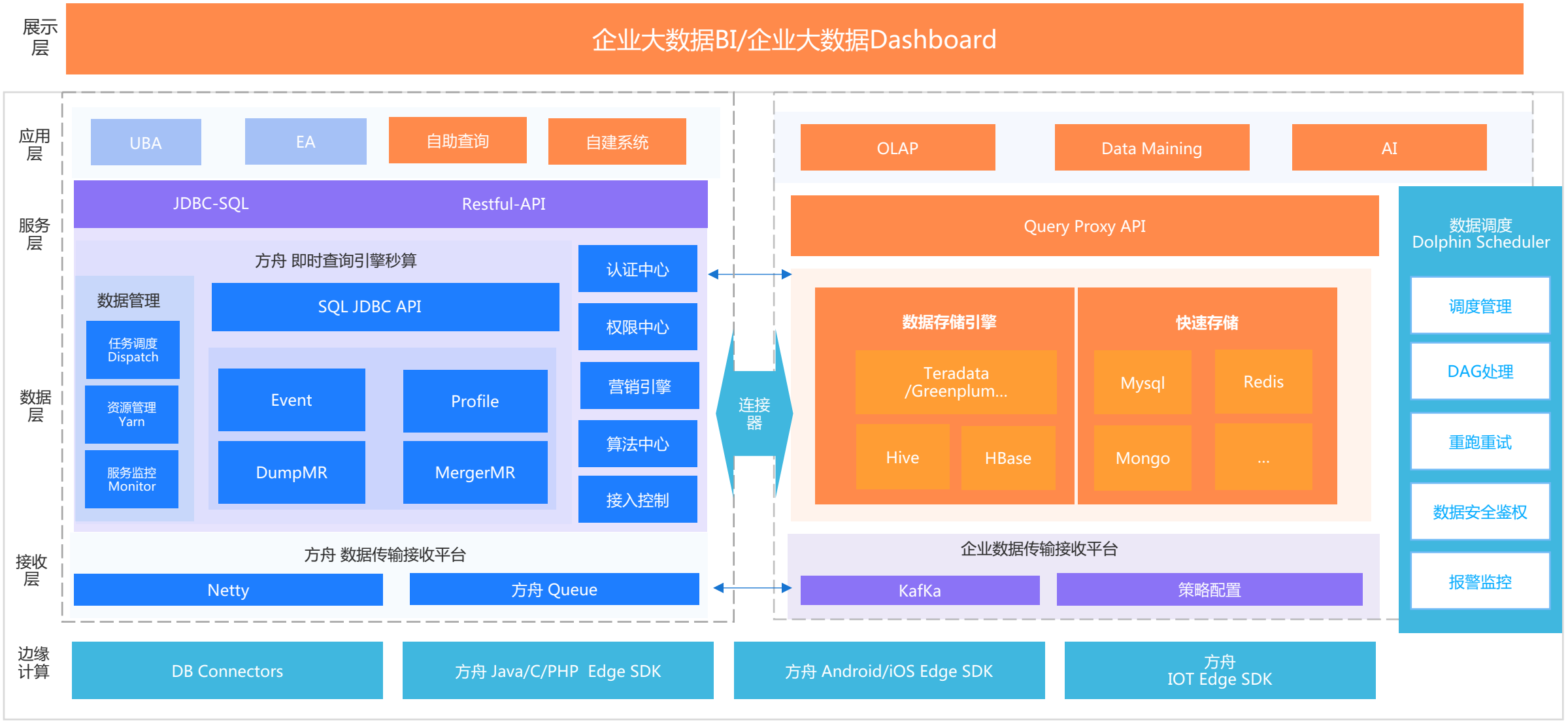
Slogan

Analysys 易观
你要的数据能力



DS调度让易观方舟智能数据平台与企业数仓融合，打造统一数据平台

Analysys 易观
你要的数据能力



DolphinScheduler 资源

Analysys 易观
你要的数据能力

- 在线DEMO: http://106.75.43.194:8888/
- 官网：https://dolphinscheduler.apache.org
- 开源地址： https://github.com/apache/incubator-dolphinscheduler

欢迎加入贡献队伍：
https://dolphinscheduler.apache.org/zh-cn/docs/development/contribute.html

获得帮助：

- Submit an issue
- Mail to dev-subscribe@dolphinscheduler.apache.org, follow the reply to subscribe the mail list.

Analysys 易观
你要的数据能力

数据驱动 精益成长

- 易观方舟
- 易观千帆
- 易观万像

网址： www.analysys.cn

客户热线：4006-010-230 / 4006-010-231

微博：@Analysys易观

加入社区