GIAC | 全球互联网架构大会
GLOBAL INTERNET ARCHITECTURE CONFERENCE

蘑菇街作业调度系统Jarvis的架构与实现

刘洋（炎寻）@蘑菇街
高级技术专家

2017.thegiac.com

Agenda

1. 任务调度系统背景知识
2. Jarvis的架构与实现
3. Jarvis在提升系统易用性方面的工作
4. Jarvis在提升系统可维护性方面的工作
5. Jarvis的现状与未来计划

任务调度系统背景知识

- 调度系统分类
- DAG工作流类调度系统在数据平台的角色
- Jarvis的基本概念

调度系统分类

资源调度系统

- Yarn
- Mesos
- Omega
- Borg
- Fuxi
- Gaia
- Normandy

任务调度系统

- Oozie
- Azkaban
- Chronos
- Zeus
- Lhotse
- SchedulerX
- Elastic-job
- Saturn
- Jarvis

任务调度系统-定时分片类系统

场景定位

- 类似任务，批量执行
- 大任务拆分成多个小任务，分布式执行
- TBSchedule，SchedulerX，Elastic-job, Saturn

关注目标

- 不漏不重，负载均衡，弹性扩容，失效转移
- 精确定时触发，强实时性和高可靠性

业务影响

- 对所调度的任务往往有代码侵入性要求
- 有些系统还要求常驻Daemon进程，用于协调本地作业的管理和通讯

任务调度系统-DAG工作流类系统

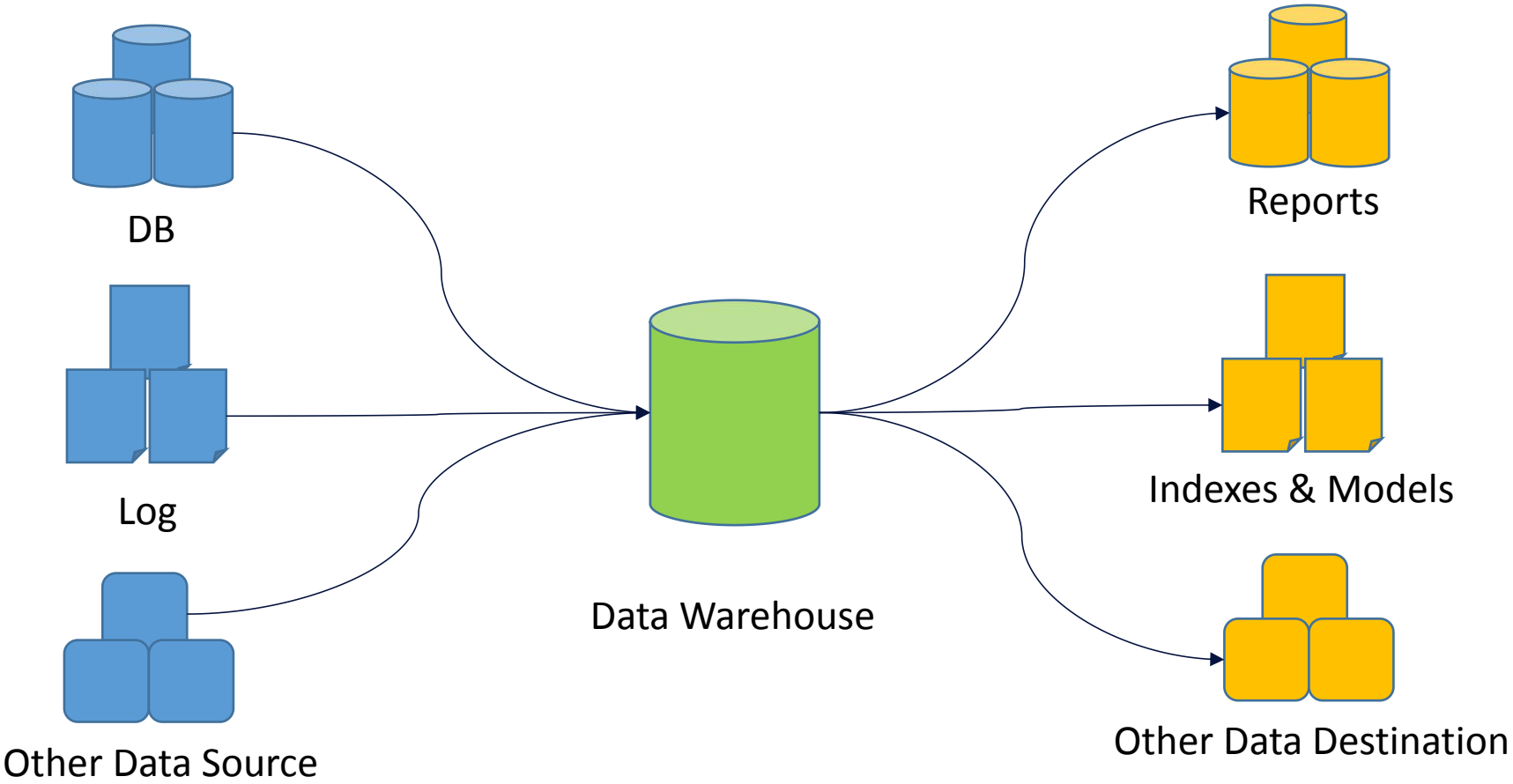
场景定位

- 所服务的往往是流程依赖比较复杂的场景
- 可能涉及到成百上千个相互交叉依赖关联的作业
- Oozie，Azkaban，Chronos，Lhotse，Jarvis

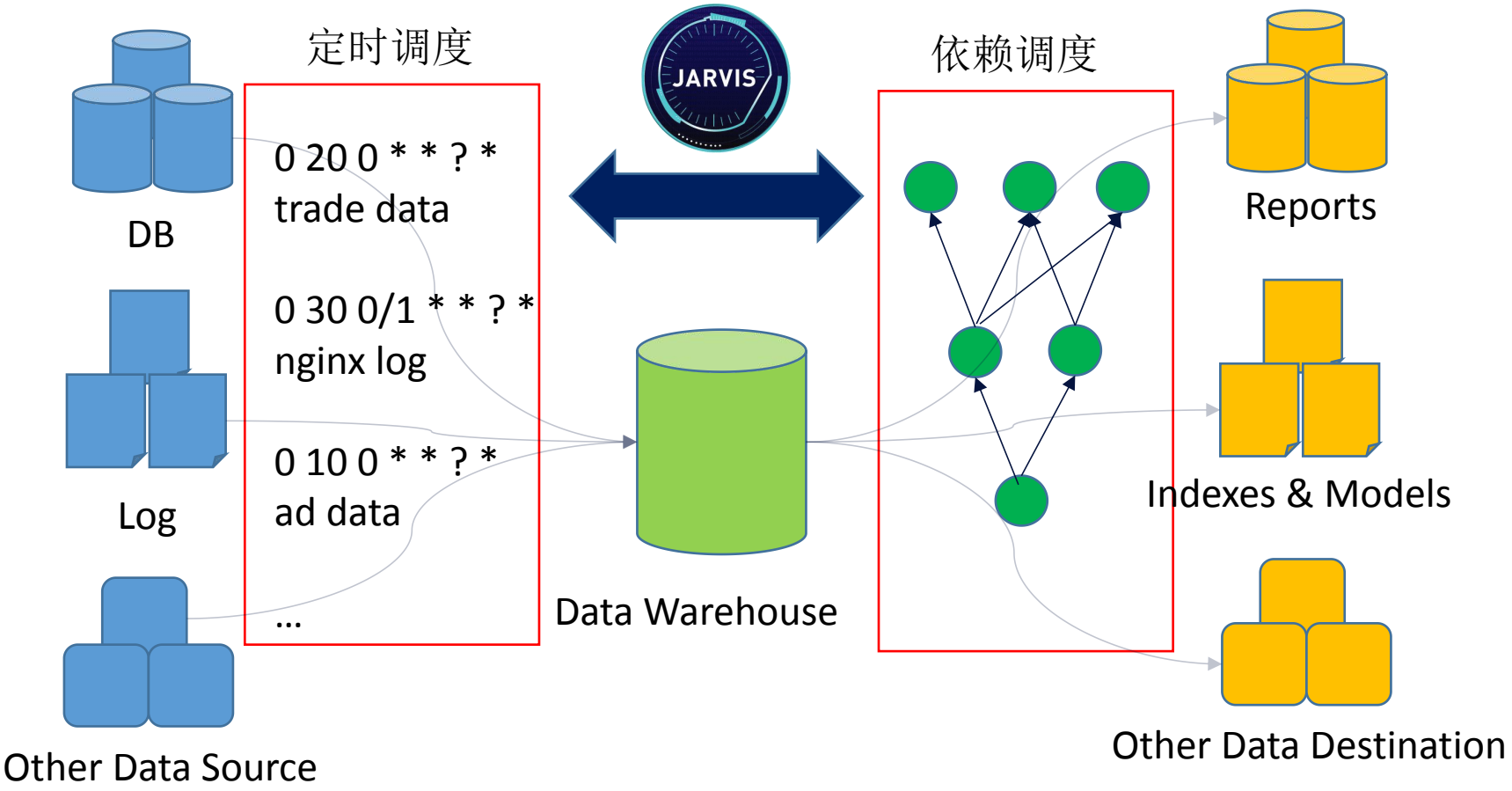
关注目标

- 丰富灵活的触发机制：时间，依赖，混合
- 灵活的作业变更管理，流程管控
- 优先级管理，业务隔离，错误跟踪，异常报警
- 系统和业务健康度监控，性能优化和问题诊断

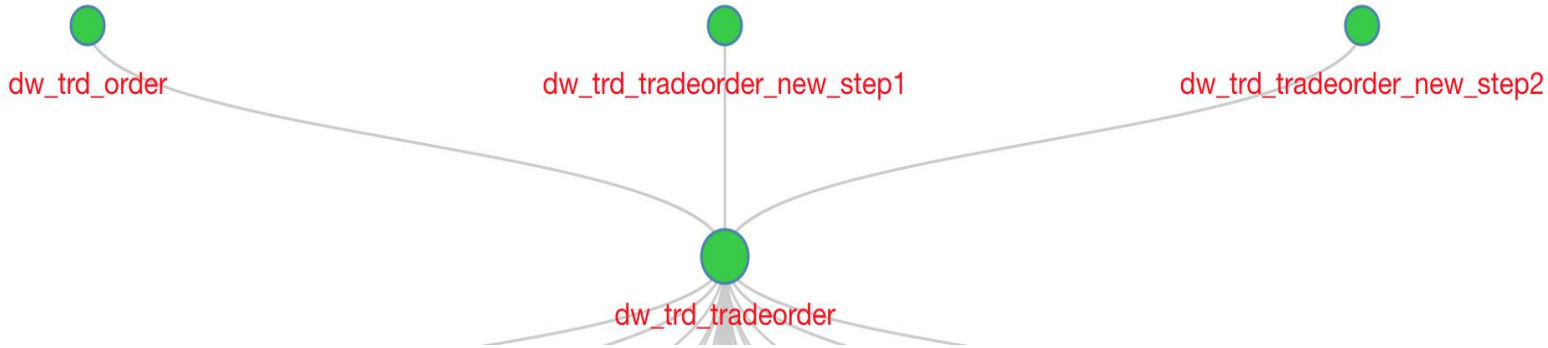
DAG工作流类调度系统在数据平台的角色



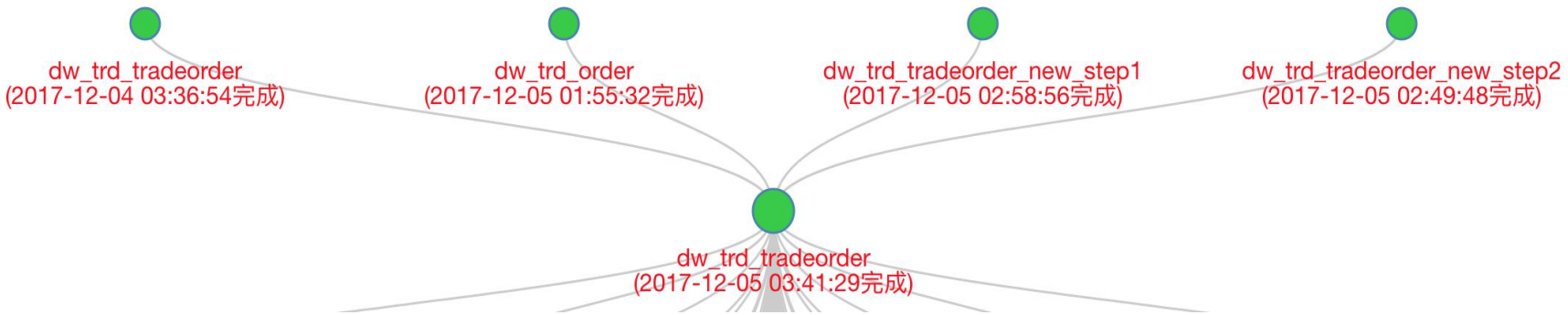
DAG工作流类调度系统在数据平台的角色



Jarvis的基本概念-Job与Task



Job



Task

Jarvis的基本概念-执行计划与执行实例

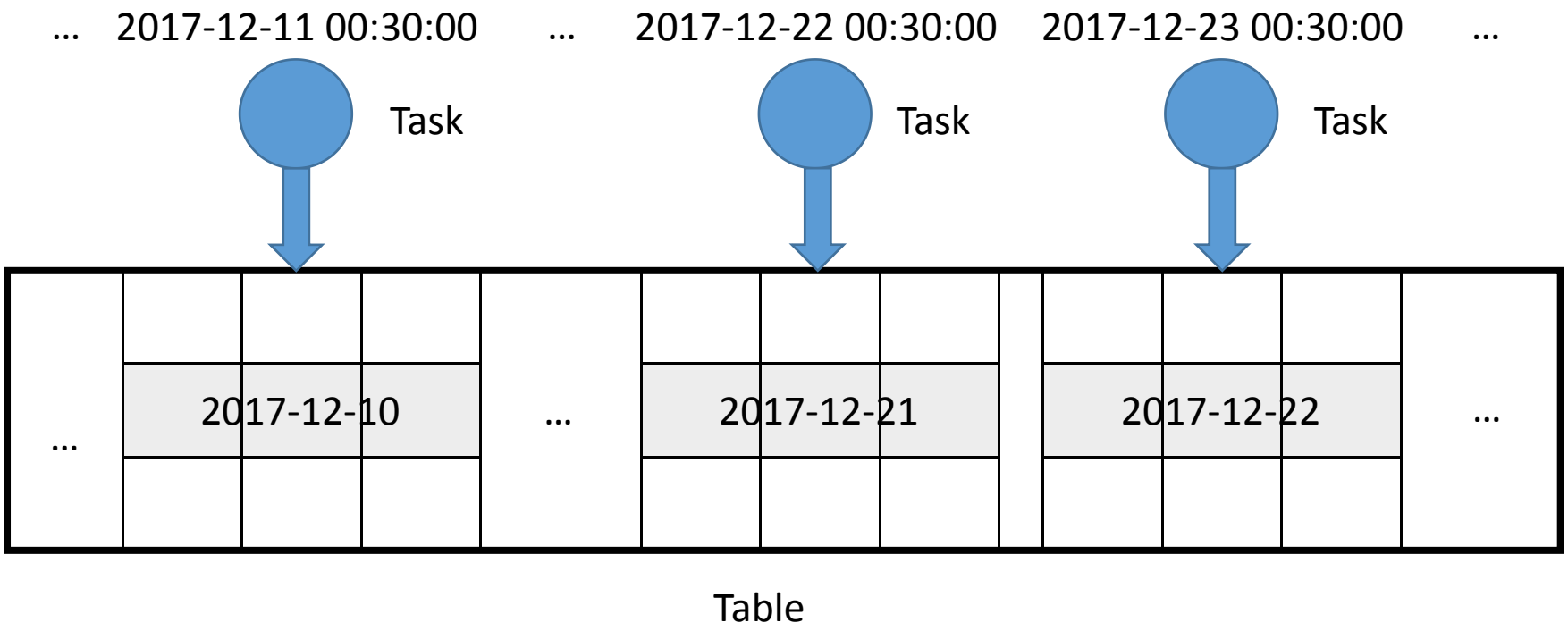
静态执行列表

- 根据作业计划
  提前生成并持久化任务执行列表
- 遍历检查列表
  满足条件触发执行
- 代表：Oozie，Azkaban
  众多公有云上的workflow服务

动态执行列表

- 不提前固化任务执行列表
  根据触发条件动态生成
- 通过时间或上游任务触发
  根据当前依赖关系生成任务实例
- 代表：Zeus, chronos
  我司的Jarvis调度系统

Jarvis的基本概念-调度时间与数据时间



Agenda

1. 任务调度系统背景知识
2. Jarvis的架构与实现
3. Jarvis在提升系统易用性方面的工作
4. Jarvis在提升系统可维护性方面的工作
5. Jarvis的现状与未来计划

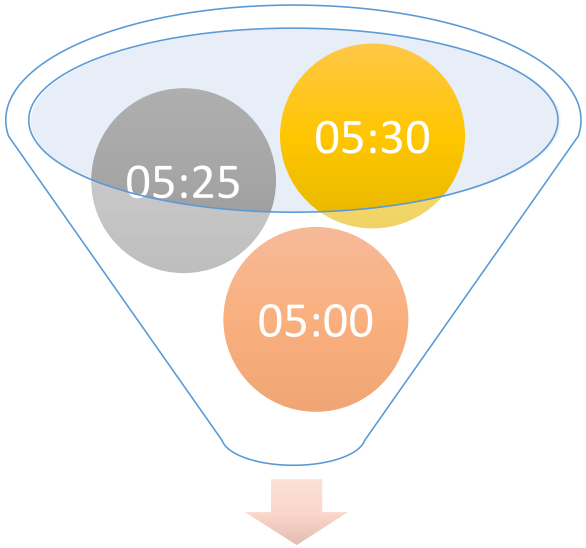
Jarvis的架构与实现

- 核心问题
- 整体架构图
- 组件详细介绍

核心问题

核心问题-时间调度

- 纯时间任务是依赖树的根
- 时间调度 = 优先队列 + 轮询检测



核心问题-依赖调度

- 依赖区间

  1.表达式：（”yyyy-MM-dd HH:00:00”，x(n), x(m)）

  （1）x是偏移单位：年y，月M，周w，天d，时h，分m，秒s

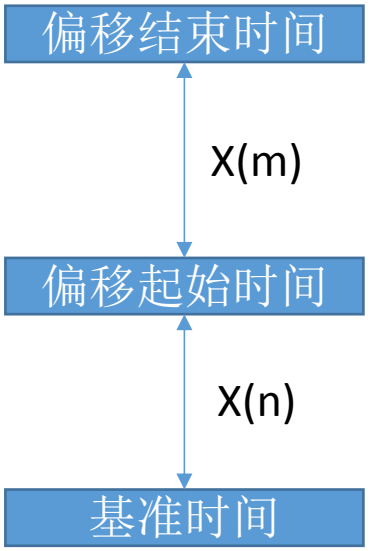
  （2）n, m是偏移值：正数向前，负数向后

  （3）起始偏移基于基准时间，结束偏移基于起始偏移

  2.简化版配置

  （1）cd-当天，（yyyy-MM-dd 00:00:00, d(-1), d(1)）

  （2）d(n)-前n天，（yyyy-MM-dd 00:00:00, d(0), d(n)）

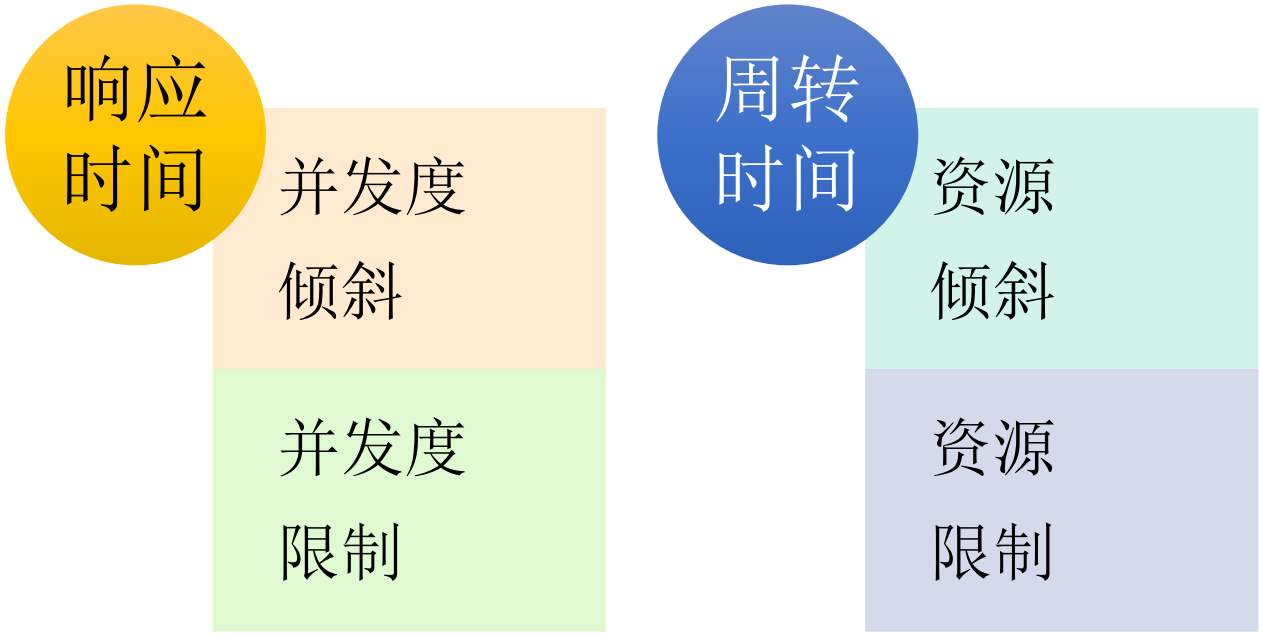


核心问题-依赖调度

- 依赖策略
  当上游Job在依赖区间里面有多个Task时，需要依赖策略：
  （1）All(*)：该区间的所有Tasks均需要成功
  （2）Any(+)：该区间的所有Tasks至少成功一个
  （3）First(n)：该区间前n个Tasks需要成功
  （4）Last(n)：该区间后n个Tasks需要成功
  （5）Continuous(n)：该区间连续n个Tasks成功

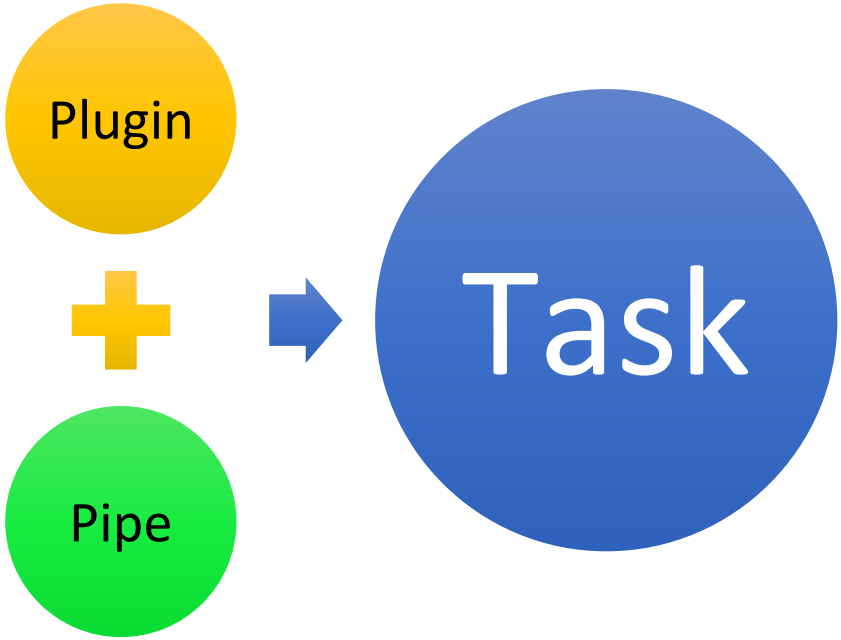
核心问题-多租户调度

- 资源有限
- 不同业务（租户）优先级不同

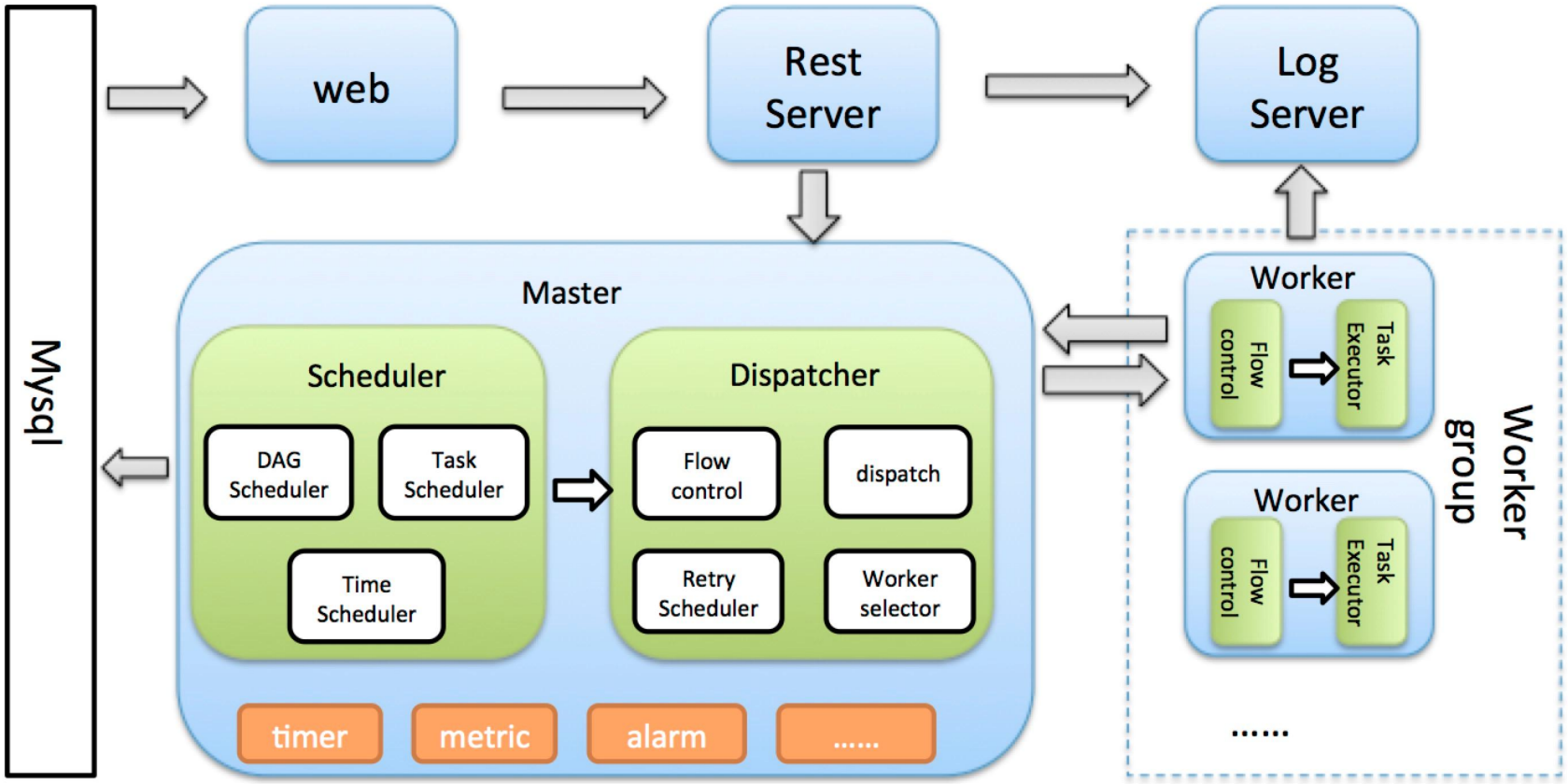


核心问题-任务执行

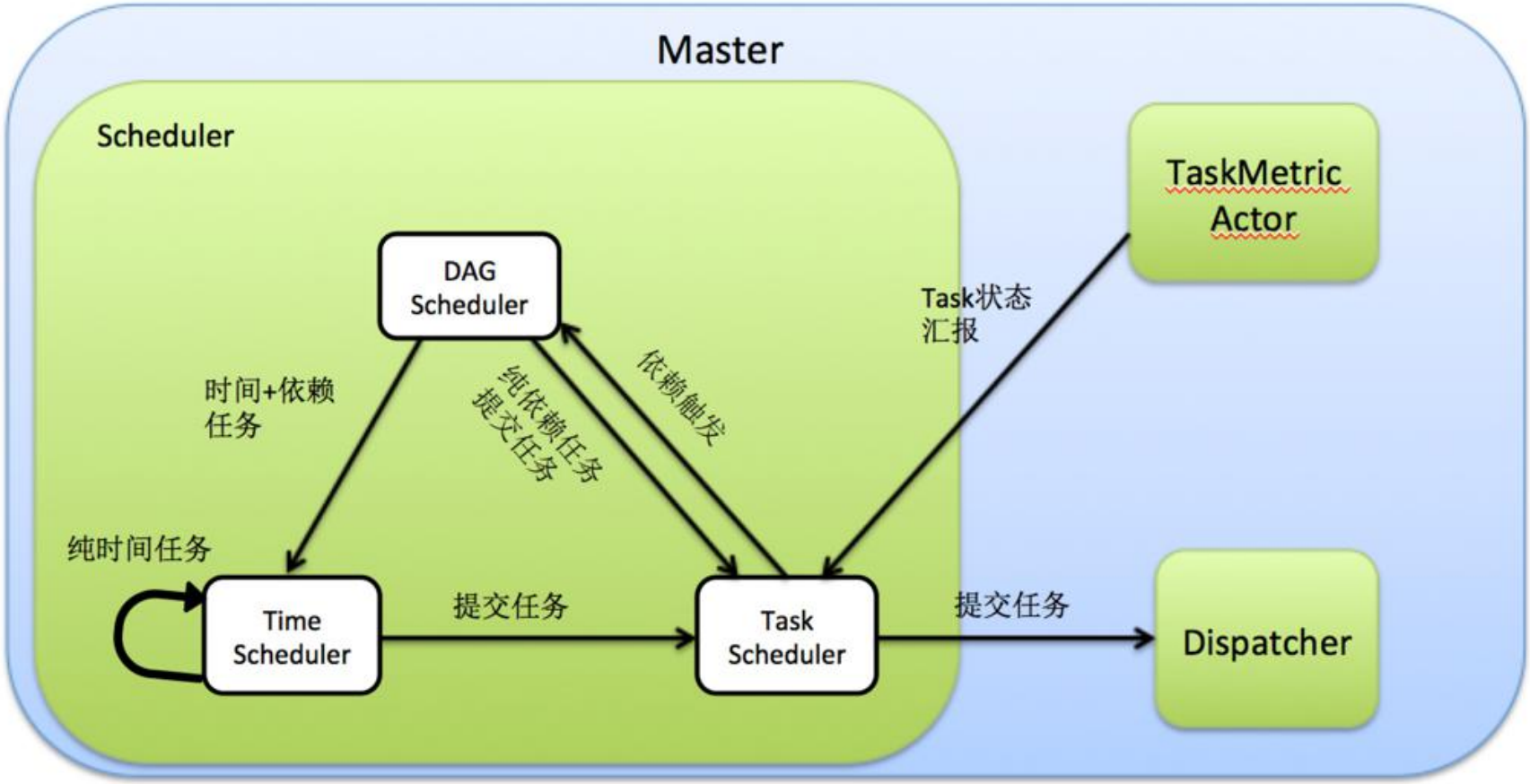
- Hive
- Java（MR）
- Spark/Spark SQL
- Presto
- Shell
- TEZ
- Hadoop Streaming
- Funnel
- DQC
- Kylin
- TheBFG
- …



整体架构图


Mysql
web
Rest Server
Log Server
Master
Scheduler
DAG Scheduler
Task Scheduler
Time Scheduler
Dispatcher
Flow control
dispatch
Retry Scheduler
Worker selector
timer
metric
alarm
.....
Worker
Flow control
Task Executor
Worker
Flow control
Task Executor
.....
Worker group

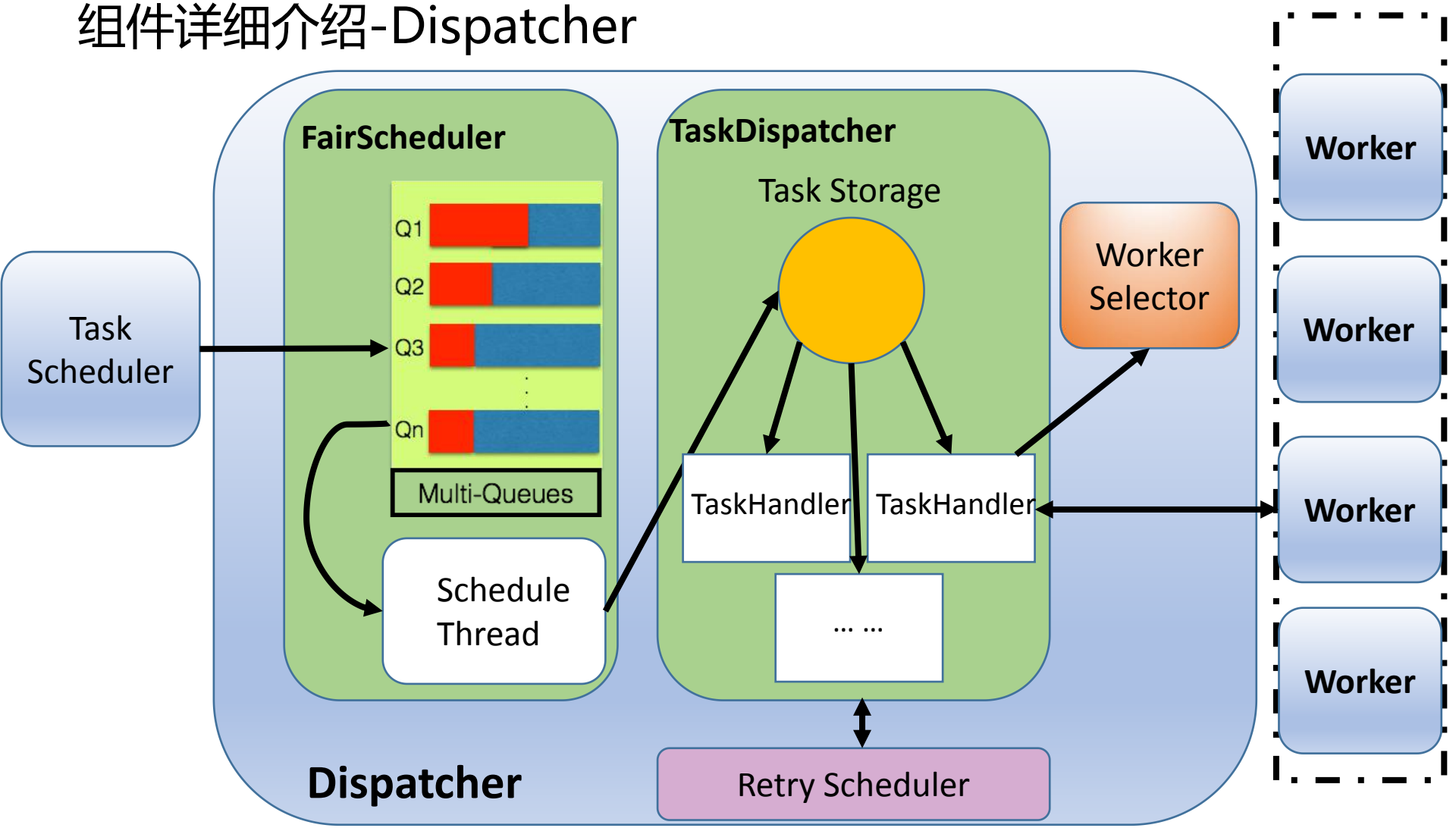
组件详细介绍-Scheduler



组件详细介绍-Scheduler

- Time Scheduler
  对纯时间限制任务进行调度，内部循环扫描，找到调度时间到达的任务提交给Task Scheduler，如果是纯时间任务则计算下一周期调度实例放入Time Scheduler无限循环
- DAG Scheduler
  内部通过JobGraph维护了所有的Job依赖关系，任何Task成功事件都会发送给DAG Scheduler去触发该Task对应Job的下游Job检查依赖，依赖通过的放入TimeScheduler
- Task Scheduler
  所有除了时间/依赖调度之外与Task相关的处理都在这里，自身维护了TaskGraph来反映每天调度的Task依赖图，提供提交任务，失败重跑，自依赖检查，暂停，恢复等功能

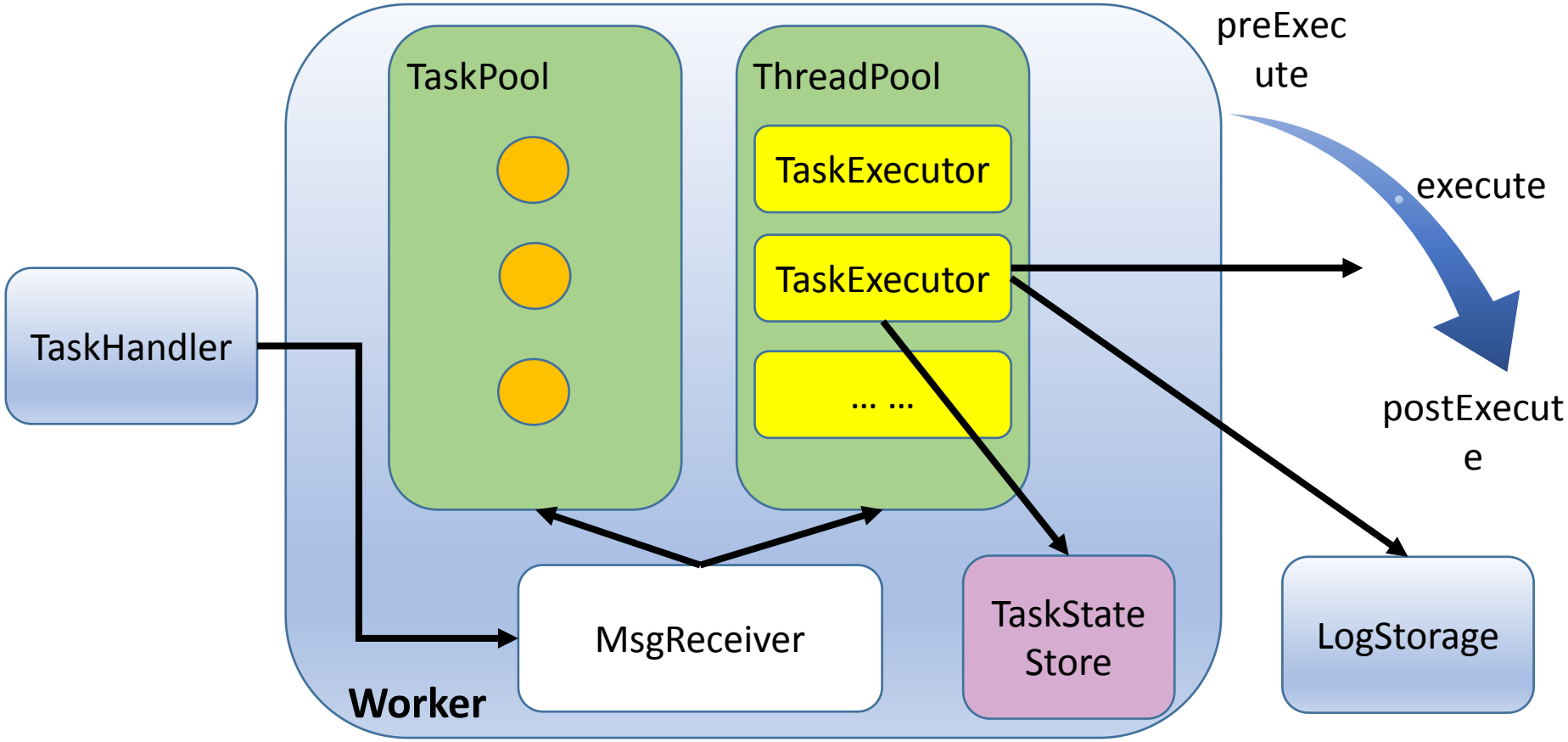
组件详细介绍-Dispatcher



组件详细介绍-Dispatcher

- Fair Scheduler
  使用多队列min-max weighted公平算法进行多租户任务调度
- Task Dispatcher
  使用生产者消费者模型来存储发送Task到Worker开始实际运行
- Worker Selector
  每个Task发送前会根据负载均衡或者灰度策略选定Worker
- Retry Scheduler
  在Worker负载过高，无法发送Task到Worker或者被Worker拒绝时加入到重试调度器，等待一段时间后重新发送

组件详细介绍-Worker



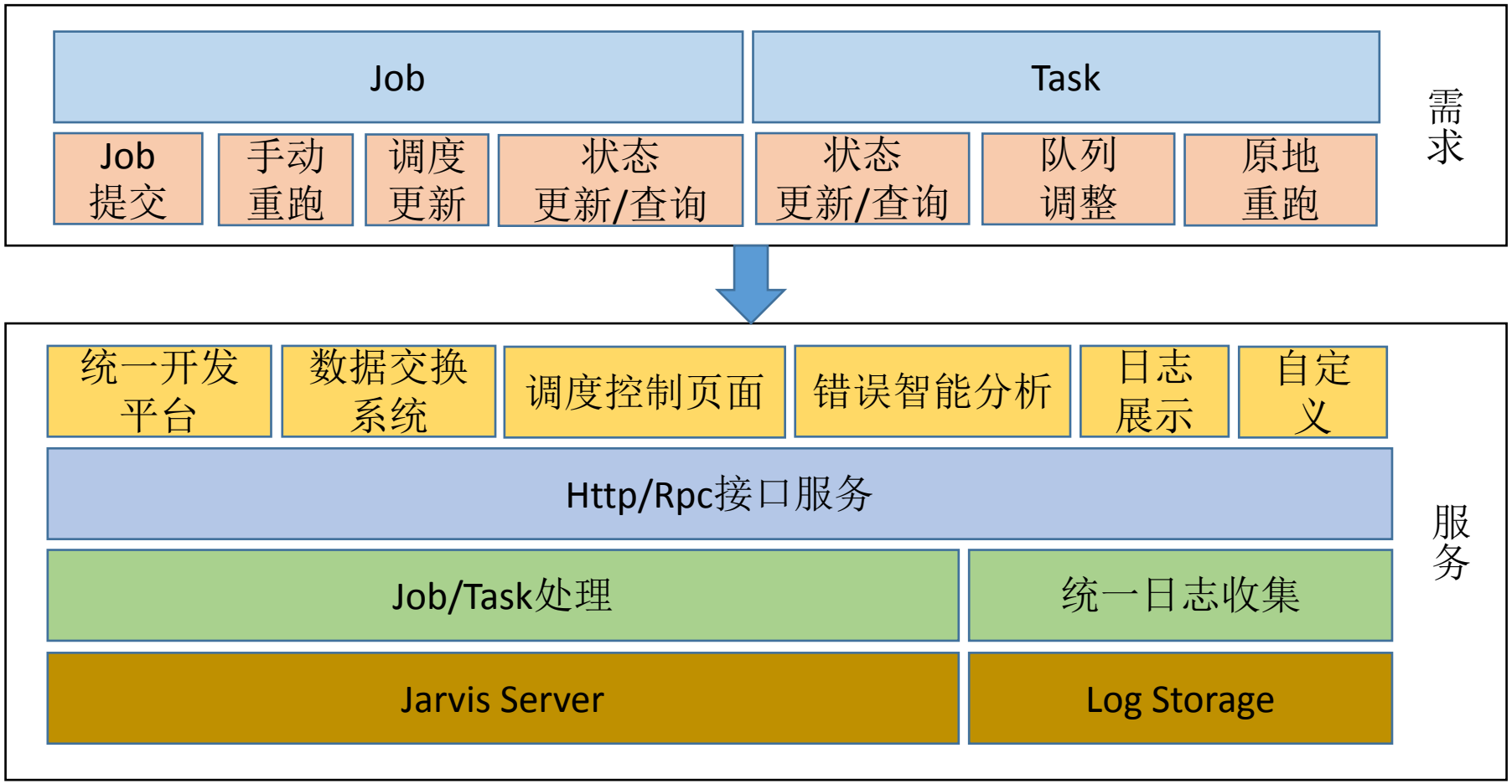
组件详细介绍-Worker

- MsgReceiver
  接收Server提交过来的Task，进行任务Context解析，去重与提交
- TaskPool
  Task状态的缓存，用于去重
- TaskExecutor
  Task执行线程，运行任务且实时发送运行日志到统一日志服务器
- TaskState Store
  Task状态持久化，Worker意外崩溃时可以恢复Task状态

Agenda

1. 任务调度系统背景知识
2. Jarvis的架构与实现
3. Jarvis在提升系统易用性方面的工作
4. Jarvis在提升系统可维护性方面的工作
5. Jarvis的现状与未来计划

Jarvis在提升系统易用性方面的工作



执行Id 即taskId | 任务Id 即jobId | 任务名称 支持正则或mysql模糊匹配,不支持*_out这种 | 应用 全部

数据日期 今天任务的数据日期为昨天 | 发布人 任务发布人 | 任务类型 全部 | 关键 全部 是 否

开始调度 2017-12-06 | 执行人 执行人 | 部门 全部 | 产品线 全部

结束调度 2017-12-06 | 执行类型 全部 周期任务 重刷数据 单次任务 | 状态 全部 等待 池子 运行 成功 失败 killed 暂停

1 查询 重置

关键路径进度(只统计所选日期的周期任务) 完成度: 99.55% 预计剩余执行时间: 5分36秒

等待 2073 | 池子 1380 | 运行 71 | 成功 47180 | 失败 2767 | Killed 126 | 暂停 0

批量操作: 暂停 | 删除此任务 | 标为成功 | 加速 | Kill | 原地重跑 | 停用报警 | 恢复 2

显示列 下载

		执行Id	任务名	状态	类型	调度时间	数据时间	开始时间	用时	进度	操作
+		12105953	funnel_mobile_track_network_v2_log		funnel	12-06 20:44:00	2017-12-05	12-06 20:44:02	1分26秒	20.43%	操作
+		12156153	analyseData		hive 单次任务	12-06 20:43:34	2017-12-05	12-06 20:43:36	47秒	100%	
+		12156146	sparkJar_example_1-鲤伴-20171206204308	灰度	sparkJar 单次任务	12-06 20:43:08	2017-12-05	12-06 20:43:09	2分19秒	0%	操作
+		12156123	双12会场_fcid点击率-艾米-20171206204055		hive 单次任务	12-06 20:40:57	2017-12-05	12-06 20:40:58	3分34秒	100%	
+		12155979	稼轩_运费险黑名单数据订单特征		hive 重刷	12-06 20:27:59	2016-08-30	12-06 20:40:01	5分27秒	40.05%	操作
+		12155977	稼轩_运费险黑名单数据订单特征		hive 重刷	12-06 20:27:59	2016-08-28	12-06 20:28:03	17分25秒	48.64%	操作
+		12155975	稼轩_运费险黑名单数据订单特征		hive 重刷	12-06 20:27:59	2016-08-26	12-06 20:28:03	17分25秒	57.35%	操作

执行详细信息

任务ID	3492672	任务名称	st_trd_real_category_shop_rank_top_d	worker IP	[redacted]
执行ID	12465617 hive	执行者	[redacted]	状态	运行
调度时间	2017-12-13 17:00:00	数据时间	2017-12-12 17:00:00	执行进度	3.4091%
开始时间	2017-12-13 17:46:24	结束时间	-	本次耗时	9分31秒(最近平均:17分18秒)
参数	hiveSqlAppNumber:{"1":32}				

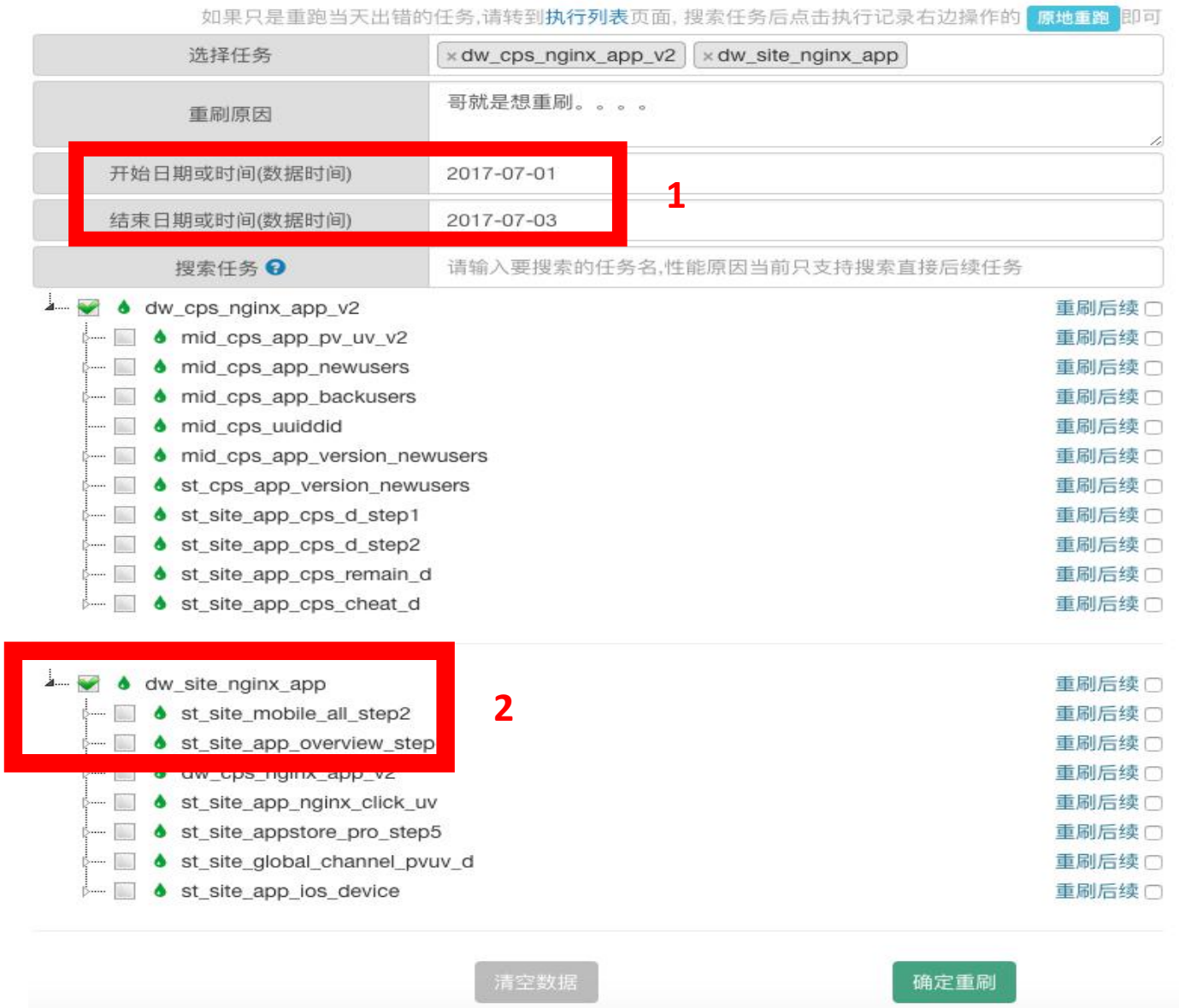
执行内容 | 日志 | 输出结果

```
[redacted]
Transaction isolation: TRANSACTION_REPEATABLE_READ
No rows affected (0.054 seconds)
No rows affected (0.002 seconds)
No rows affected (0.001 seconds)
No rows affected (0.001 seconds)
No rows affected (0.092 seconds)
Total jobs = 3
INFO  : Stage-1 is selected by condition resolver.
INFO  : Number of reduce tasks not specified. Estimated from input data size: 61
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:87
INFO  : Submitting tokens for job: job_1510656640015_2675968
```

禁用 删除 暂停

任务基本信息

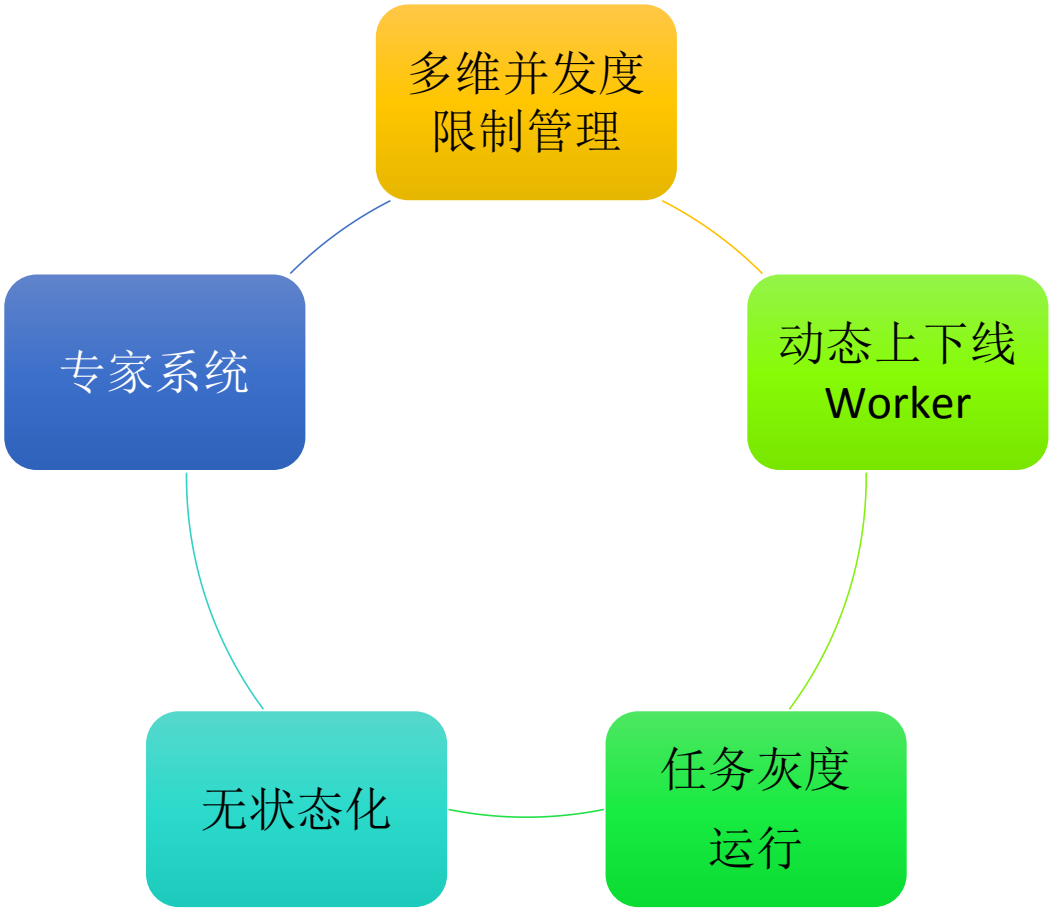
id	93574	名称	dw_trd_tradeorder	状态	启用
调度时间	每天 0点15分	类型	hive	串行/并行	串行
优先级	非常高	是否临时	否	部门	平台技术 （数据仓库&产品）
发布人	[redacted]	代理人	无	报警人	[redacted]
重试次数	3次	重试间隔	30秒	最大并发	10
应用	jarvis-web	worker组	hive集群	产品线	交易-基础
开始日期	2000-01-01	结束日期	2999-01-01	超时时长	不失效
业务产出时间	05:00:00				
参数	hiveSqlAppNumber:{"1":2}				
内容 显示					


如果只是重跑当天出错的任务,请转到执行列表页面, 搜索任务后点击执行记录右边操作的 原地重跑 即可
选择任务
dw_cps_nginx_app_v2
dw_site_nginx_app
重刷原因
哥就是想重刷。。。。
开始日期或时间(数据时间)
2017-07-01
结束日期或时间(数据时间)
2017-07-03
1
搜索任务
请输入要搜索的任务名,性能原因当前只支持搜索直接后续任务
dw_cps_nginx_app_v2
mid_cps_app_pv_uv_v2
mid_cps_app_newusers
mid_cps_app_backusers
mid_cps_uuiddid
mid_cps_app_version_newusers
st_cps_app_version_newusers
st_site_app_cps_d_step1
st_site_app_cps_d_step2
st_site_app_cps_remain_d
st_site_app_cps_cheat_d
dw_site_nginx_app
st_site_mobile_all_step2
st_site_app_overview_step
st_site_app_nginx_click_uv
st_site_appstore_pro_step5
st_site_global_channel_pvuv_d
st_site_app_ios_device
2
重刷后续
清空数据
确定重刷

Agenda

1. 任务调度系统背景知识
2. Jarvis的架构与实现
3. Jarvis在提升系统易用性方面的工作
4. Jarvis在提升系统可维护性方面的工作
5. Jarvis的现状与未来计划

Jarvis在提升系统可维护性方面的工作



新增并发度限制策略

并发度限制策略列表

启用 暂停

2 3

显示列

策略id	直接子策略ids	限制条件(*代表所有)	匹配方向	并发度上限	开始时间	结束时间	操作
1	无	调度类型:正常调度 业务类型:非KPI任务	正向	200	00:00:00	06:30:00	编辑 暂停 删除
2	无	任务类型:hive	正向	280	-	-	编辑 暂停 删除
3	无	调度类型:手动重刷	正向	5	00:00:00	09:00:00	编辑 暂停 删除
4	无	调度类型:一次性任务	正向	80	-	-	编辑 暂停 删除
5	无	应用名称:jarvis-web,xmen,ironman,report,dqc,mgs_lrm,stark	反向	20	-	-	编辑 暂停 删除
6	无	任务类型:shell	正向	120	00:00:00	02:00:00	编辑 暂停 删除
7	8	执行人:* 调度类型:手动重刷	正向	50	-	-	编辑 暂停 删除
8	无	执行人:etlprd 调度类型:手动重刷	正向	200	-	-	编辑 暂停 删除
10	无	任务ids:* 调度类型:手动重刷	正向	30	-	-	编辑 暂停 删除
11	无	应用名称:dqc	正向	50	-	-	编辑 暂停 删除

1

首页 Worker管理

Worker Worker Group

Group名称 全部 IP 全部 port 全部 是否停用 全部 查询 重置

显示方式 显示列 下载

IP	端口	所属Group	状态	执行任务数	是否启动	是否停用	操作
[redacted]203	10001	hive2es集群	✔上线中	0	已启动	-	状态
[redacted].149	10001	标准权限模型灰度验证集群	⊘未上线	-	-	-	状态
[redacted].165	10001	hive集群	✔上线中	25	已启动	-	状态
[redacted].68	10001	hive集群	✔上线中	29	已启动	-	状态
[redacted].167	10001	hive集群	✔上线中	16	已启动	-	状态
[redacted].166	10001	hive集群	✔上线中	24	已启动	-	停用
[redacted].162	10001	dqc	✔上线中	0	已启动	-	状态
[redacted].164	10001	数据交换测试集群	✔上线中	0	已启动	-	状态
[redacted].153	10011	标准权限模型灰度验证集群	✔上线中	1	已启动	-	状态
[redacted].154	10011	标准权限模型灰度验证集群	✔上线中	0	已启动	-	状态

1 2 3

首页 > 系统管理 > 灰度发布配置

灰度功能开关(开启中): 开启

新增灰度策略

灰度策略列表

启用 暂停

1 2

显示列

策略id	灰度功能名	筛选条件	灰度策略	开始时间	结束时间	附加信息	负责人	添加时间	操作
42	hbase升版本测试	发布人:	灰度比例:100% 灰度机器:	00:00:00	23:55:00	{}		2017-10-23 16:25:28	编辑 暂停 删除
41	SparkSession灰度	任务类型:sparkSQL,sparkSession,spark,sparkJar 是否关键:false	灰度比例:100% 灰度机器:	00:00:00	23:55:00	{}		2017-10-17 18:26:22	编辑 暂停 删除
39	EasyHive灰度	任务类型:easyHive 是否关键:false	灰度比例:100% 灰度机器	00:00:00	23:55:00	{}		2017-08-01 14:39:05	编辑 暂停 删除
37	数据交换灰度功能	任务名称:hive_clean_shell 是否关键:false	灰度比例:100% 灰度机器:	00:00:00	23:55:00	{}		2017-07-28 15:16:56	编辑 暂停 删除
31	TEZ灰度-所有hive	任务类型:tez	灰度比例:100% 灰度机器:	00:00:00	23:55:00	{}		2017-07-12 17:50:47	编辑 暂停 删除

Agenda

1. 任务调度系统背景知识
2. Jarvis的架构与实现
3. Jarvis在提升系统易用性方面的工作
4. Jarvis在提升系统可维护性方面的工作
5. Jarvis的现状与未来计划

Jarvis的现状与问题

现状

- 经过快两年的开发和持续改进，前文所描述的系统功能，易用性，可维护性目标，都已经实现
- 日常稳定承载约2万个固定周期调度作业，以及同样数量级的一次性任务作业和重刷任务作业

问题

- 部分业务逻辑实现过于定制化，不利于系统功能的后续拓展和调整
- 在突发峰值或者极端高负载情况下的系统稳定性还需要经历更多的复杂场景来加以磨练

Jarvis的未来改进计划

系统整体业务健康度检测和评估手段改进

- 三个层面的监控并不够：硬件指标, 系统和进程, 组件和链路
- 加强系统监控综合评估能力, 建设业务专家系统

自动测试体系的完善

- 单元测试不足以发现大流量负载，复杂并发场景下隐蔽Bug
- 需要构建随机生成测试用例和模拟组件失效模式的测试体系

功能扩展

- 计算引擎
  - 即席计算
  - 机器学习
- 分片调度，水平扩展

开源

- 开源不是一个目标，而是用来提高产品质量的手段
- 重要的是开放思想，目的是让大家一起参与，共同努力，共同受益，而不是光晒代码

GIAC | 全球互联网架构大会
GLOBAL INTERNET ARCHITECTURE CONFERENCE

GIAC

全球互联网架构大会

GLOBAL INTERNET ARCHITECTURE CONFERENCE

扫码关注GIAC公众号

2017.thegiac.com