**基于akka技术栈的任务调度引擎功能与实现技术介绍**

目 录

[1 简介 3](#_Toc47871954)

[2 类似产品特性对比 3](#_Toc47871955)

[3 架构介绍 4](#_Toc47871956)

[3.1 系统架构 4](#_Toc47871957)

[3.1.1 Rest Service 4](#_Toc47871958)

[3.1.2 Java Api 4](#_Toc47871959)

[3.1.3 Workflow Builder（工作流构建器） 4](#_Toc47871960)

[3.1.4 Workflow Runner（工作流执行器） 5](#_Toc47871961)

[3.1.4.1 工作流运行 5](#_Toc47871962)

[3.1.4.2 工作流的重试 6](#_Toc47871963)

[3.1.5 Task Client（任务客户端） 7](#_Toc47871964)

[3.1.6 Task Executor（任务执行端） 7](#_Toc47871965)

[3.1.6.1 任务执行 7](#_Toc47871966)

[3.1.6.2 优先级管理： 8](#_Toc47871967)

[3.1.7 Pub/Sub System（发布订阅系统） 8](#_Toc47871968)

[3.1.8 Metrics Listener（集群指标监听器） 8](#_Toc47871969)

[3.1.9 Fallback Listener（故障恢复监听器） 8](#_Toc47871970)

[3.1.10 二次开发模组 9](#_Toc47871971)

[3.1.10.1 Trigger Module（触发器模组） 9](#_Toc47871972)

[3.1.10.2 Repository Module（持久化仓库模组） 10](#_Toc47871973)

[3.1.10.3 FileSystem Module（文件系统模组） 10](#_Toc47871974)

[3.1.10.4 StdHandler Module（任务控制台输出管理模组） 10](#_Toc47871975)

[3.1.10.5 Expression Evaluator Module（表达式求值模组） 10](#_Toc47871976)

[3.1.10.6 Pluggable Executor Module（自定义任务执行模组） 11](#_Toc47871977)

# 简介

基于akka的任务调度引擎是一个响应式，完全去中心化，可动态缩扩容，支持跨集群任务调度，指定节点执行任务等功能的调度引擎，目前正在开发中。已可以实现工作流的正常执行、停止、重试；工作流定时调度，暂停，继续，停止；状态查看等功能。

# 类似产品特性对比

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Oozie** | **Dolphin** | **此调度引擎** |
| 多租户 | 不支持 | 支持 | 支持 |
| 算子 | 不支持python，flink；对shell，java支持度不高 | 好 | 好 |
| 表达式 | 支持el表达式，难以二次开发 | 不支持 | 支持开源组件Aviator，可二次开发。 |
| 工作流失败策略 | 可自定义失败策略 | 不支持，若任务失败则工作流直接停止。 | 可自定义失败策略 |
| 工作流重试策略 | 可仅重新运行失败节点，也可跳过指定的节点。并且重试时可修改运行参数。 | 不支持，直接重新运行所有节点，并且没有运行参数的设置。 | 提供不同的重试策略，并且重试时可修改运行参数 |
| 高可用性 | 无 | 通过zookeeper分布式锁动态选举master，保证高可用。 | 基于Akka Cluster完全去中心化，保证高可用。 |
| 插件开发 | 可自定义算子 | 可自定义算子 | 可自定义多个模块  详见3.10节 |

# 架构介绍

## 系统架构

调度引擎的架构如图所示：



下面依次分析图中的各个模块：

### Rest Service

调度引擎提供整一套Rest Api来管理工作流和调度任务的执行，以及集群，工作流或调度任务的状态查看。类似Elasticsearch，在集群的所有节点都可以使用Rest Api进行操作。

### Java Api

Java Api基于Rest Api，用于java应用程序的开发。

### Workflow Builder（工作流构建器）

该模块在收到Rest Service提交的工作流DAG图后，会通过Akka Stream Graph构建一个可执行的工作流，工作流由算子组成，算子种类有：

|  |  |  |  |
| --- | --- | --- | --- |
| **算子名称** | **输入端口数** | **输出端口数** | **功能说明** |
| start | 0 | 1 | 启动一个动作流 |
| end | 1 | 0 | 结束一个工作流 |
| task | 1 | 1 | 执行一个任务，比如java，shell等 |
| fork | 1 | n | 并行的通知下游算子执行 |
| decision | 1 | n | 根据算子的条件表达式选择一个下游算子执行 |
| join | n | 1 | 收集上游的算子执行情况，当所有上游算子都执行完后，通知下游算子执行。用在fork算子的下游 |
| converge | n | 1 | 当上游的其中一个算子执行完后，就会通知下游的算子执行。  用在decision算子的下游。 |

### Workflow Runner（工作流执行器）

Workflow Builder中构建好的工作流会进入到该模块，负责工作流的运行。

#### 工作流运行

工作流的运行包括两种失败策略：

* CONTINUE：当一个任务算子失败时继续执行下游算子，直到工作流结束；
* SHUTDOWN：当一个任务算子失败时立即停止。

运行流程如图所示：



当工作流开始执行时，只需要在工作流的输入端提交一个GraphMeta对象，GraphMeta对象内包含了工作流的运行参数和运行时的状态信息。当Task算子获取GraphMeta对象后会生成一个任务，提交给下方模块Task Client来执行，并监听任务的执行状态。当任务完成后会返回一个TaskMeta对象。GraphMeta和TaskMeta结合生成一个新的GraphMeta并通知到下游，然后下游的Task算子会继续这个过程。当整个工作流的流程结束后，GraphMeta对象中会存有所有运行的算子的状态信息。通过这种方式，可以实现工作流的运行。

#### 工作流的重试

工作流的重试包括三种重试策略：

* RUN\_ALL：当一个任务算子失败时继续执行下游算子，直到工作流结束；
* RUN\_ONLY\_FAILED：只运行失败的算子；
* RUN\_ALL\_AFTER\_FAILED：运行失败算子和所有在其下游的算子。

第1种重试策略与直接运行一次工作流类似，第2、3种则需要根据工作流前一次运行的状态来确定运行策略，其流程如图所示：



由于工作流第一次运行时输出的Graph Meta已经有了所有的Task Meta信息，所以只需要将这个Graph Meta添加上重试策略重新输入到工作流即可，task算子会根据重试策略和Graph Meta中上一次运行的结果判断是否要再运行一遍，由此来实现不同策略的工作流重试。

### Task Client（任务客户端）

Workflow Runner的task算子生成的任务会提交到该模块，该模块基于Akka Cluster Client，负责将任务提交到目标节点进行任务的执行或关闭，并实时监控节点的运行状态。

任务的详细执行流程参考2.1.6节。

### Task Executor（任务执行端）

#### 任务执行

流程如下图所示：



任务客户端首先会询问执行端是否可用。执行端回复可用的同时会创建一个ExecutionInstance。任务客户端收到确认后会创建一个ClientInstance。之后遍由两个instance来进行该任务的交互。其中：

* ClientInstance负责监听状态信息和提交kill命令；
* ExecutionInstance负责向ClientInstance推送任务的状态信息并向**RunnerManager(执行管理器)**提交任务，RunnerManager会返回一个hook用来kill掉任务。
* RunnerManager(执行管理器)实际上是一个拥有固定线程的线程池，线程会根据任务配置将其交给一个TaskRunner(执行器)对象来执行任务，TaskRunner每次执行一个任务都会通过java的Process类创建一个子进程。这样能够保证资源隔离，并且可以使用不同的系统用户或kerberos租户来执行任务，java任务也能使用jvm参数。

#### 优先级管理：

每个任务都带有一个优先级值，当任务提交进RunnerManager时首先会进入一个优先级队列会根据优先级排序，线程会循环从队列中拉取任务并执行，从而实现按优先级运行。

### Pub/Sub System（发布订阅系统）

任务调度引擎支持Akka的发布订阅系统，每一个节点都会订阅一个或多个topic。当TaskCient请求任务时，请求会根据任务配置的topic来随机找到一个拥有相同topic的节点。用户可以根据需求来配置需要的topic，比如：

* 按业务划分节点的topic；
* 对于一个跨集群的调度引擎，各个集群的节点都订阅一个各自的topic，这样就可以实现跨集群的任务调度；
* 每个节点都订阅一个独立的topic，这样就可以已节点为粒度来执行任务，不过在这种情况下若某节点故障可能会导致工作流执行失败。

### Metrics Listener（集群指标监听器）

用于监控集群节点的运行指标，比如堆内存，cpu占用等。

### Fallback Listener（故障恢复监听器）

当集群的某个节点故障或失效时该模块负责将失效节点管理的定时调度任务转移至其他节点运行。

调度引擎的定时调度任务在集群中的执行方式如图所示：



图中的Coordinator就是定时调度任务，由Trigger来进行管理，每个集群的节点都会管理一部分的Coordinator，且每个coordinator都有一个版本号，初始为0。Trigger根据coordinator定时触发工作流运行，工作流的任务算子会由Task Client分发给对应节点的Task Executor来执行。

当集群的其中一个节点失效时，Fallback Listener会将该节点管理的Coordinator重新提交到其他节点运行，并将版本号加1，失效节点如果还在运行，则会发现自己管理的Coordinator版本号要低于当前提交的版本号，于是停止调度。如图所示：



### 二次开发模组

#### Trigger Module（触发器模组）

工作流需要周期性的执行，周期性执行的策略可能是每隔一段时间执行一次，也可能是监听某个文件目录出现变化就执行一次等等。

调度引擎提供com.oceanum.trigger.Trigger接口来供开发者实现不同的触发策略，接口包含启动、停止、暂停、继续四个方法。

#### Repository Module（持久化仓库模组）

负责持久化和查询工作流的dag配置、工作流运行状态、任务运行状态等等。

调度引擎提供com.oceanum.persistence.RepositoryFactory接口来供开发者实现不同数据库下的持久化方案，推荐使用json格式存储的数据库。

#### FileSystem Module（文件系统模组）

任务算子的执行文件需要存放在一个方便管理、上传、下载的地方，由该模组管理。

调度引擎提供com.oceanum.file.FileSystem接口来供开发者实现不同的文件系统。

#### StdHandler Module（任务控制台输出管理模组）

Runner Manager中执行的任务的控制台输出会由该模组来管理，也可以通过该模组查看输出内容。

调度引擎提供com.oceanum.exec.StdHandlerFactory接口来供开发者实现不同的任务控制台输出管理方案，接口方法包括对stdout和stderr输出的存储和查看。

#### Expression Evaluator Module（表达式求值模组）

调度引擎支持的表达式求值组件基于开源组件Aviator开发，可以动态生成任务的配置。目前已支持多种函数及表达式，以下为几个例子：

* 输出一小时后的时间的格式化字符串：

|  |
| --- |
| date.format('yyyy-MM-dd HH:mm:ss', date.now() + duration.hour(1)) |

* 判断文件是否存在并且文件大小大于1KB：

|  |
| --- |
| fs.exists('/tmp/test/file') && (fs.size('/tmp/test/file') > 1024) |

#### Pluggable Executor Module（自定义任务执行模组）

用于自定义任务的开发。