# 概述

实现db到hdfs数据导入功能，提供高效的分布式并行处理能力，可以采用数据库分区、按字段分区、分页方式并行批处理抽取db数据到hdfs文件系统中。有效解决按字段分区数据不均匀问题。

灵活的作业任务处理模式：可以增量方式执行作业任务，作业可以重新执行，重新执行时只需执行未完成的作业任务即可；可以在原有作业切分的基础进一步切分出子任务，提升系统处理数据效率。

## 技术体系

**Bboss ioc：**轻量级ioc容器，ioc扩展属性配置语法

**Bboss持久层：**查询数据行处理器，动态数据源管理，连接池数据源和非连接池数据源，表分区信息查询等

**Bboss分布式事件框架**：基于JGroups，动态管理作业节点（服务节点和数据处理节点），包括作业节点的动态加入、动态离开等；在管理节点、数据处理节点之间分发和接收各种作业处理指令事件

**Bboss mvc：**实现监控管理应用的控制器,在监控服务节点中，通过mvc 容器启动监听器方法中启动作业服务

**Bboss 标签库，jquery等**：实现监控管理应用的视图层

**Hadoop Hdfs客户端**：用来连接hadoop hdfs文件系统

**Sqllite：**在监控节点中保存作业配置，后续考虑保存作业状态

**Jetty：**运行监控管理应用模块的web应用容器

**Bboss应用执行容器：**用来启动作业管理监控应用、作业数据处理应用、启动jetty容器

## 基本概念

**指令：**可以在管理节点上或者数据处理节点执行的一系列操作，比如数据抽取操作、停止数据源操作、删除hdfs文件操作等等

**作业：**一个完整的数据抽取指令，比如表数据抽取到hdfs作业等，停止数据源作业指令等，一个作业可以被分解为做个作业任务，然后交个数据处理节点执行。数据处理节点会给每个作业分配独立的作业任务处理队列和作业工作线程，作业之间互不干扰；处理队列长度和工作线程数可以在作业定义中进行配置（可以根据作业对应的数据量级别、服务器资源配置以及数据库并发处理能力进行灵活配置）。

**作业任务：**由作业分解而成，将由数据处理节点中的分配给对应作业的工作线程处理执行，如果对应的作业没有空闲的作业线程，则将作业任务加入作业对应的工作队列，如果作业工作队列也满了，则阻塞等待直到有空闲的作业队列或者作业工作线程。

**管理节点：**管理、监控、调度作业的服务器

**数据处理节点：**处理和执行作业任务的服务器

## 整体架构

服务器类型

管理节点

数量：只有一台，

作用：

* 执行所有作业指令的入口，监控所有数据节点作业状态（正常、异常、运行、等待、终止、完成等）和进展情况（数据任务处理状态，成功数据行数，异常数据行数，总的记录行数等等），管理节点也可以作为数据节点。
* 分解作业任务，根据作业定义，将作业分解成多个任务，然后将作业任务均匀地分配给数据处理节点执行
* 调度任务，根据数据处理节点作业任务执行情况，将任务繁忙的作业任务调配到相对较空闲的作业数据处理节点
* 停止作业任务，在作业执行过程中可以停止正在执行的作业
* 重新执行已经完成的作业或者重新执行修改了配置的作业
* 管理hdfs文件：删除hdfs文件和目录
* 管理作业使用的数据源：停止不需要的数据源，或者作业数据源参数修改调整后，为了加载新的配置而停止数据源
* 采集作业状态数据源：通过异步事件机制，采集分布在每个数据处理节点上的作业任务执行状态数据，包括：作业执行状态、作业任务执行状态、数据抽取进度、异常情况、已经完成作业任务清单、未完成作业任务清单、失败作业任务清单

数据节点

数量：1…n台

作用：

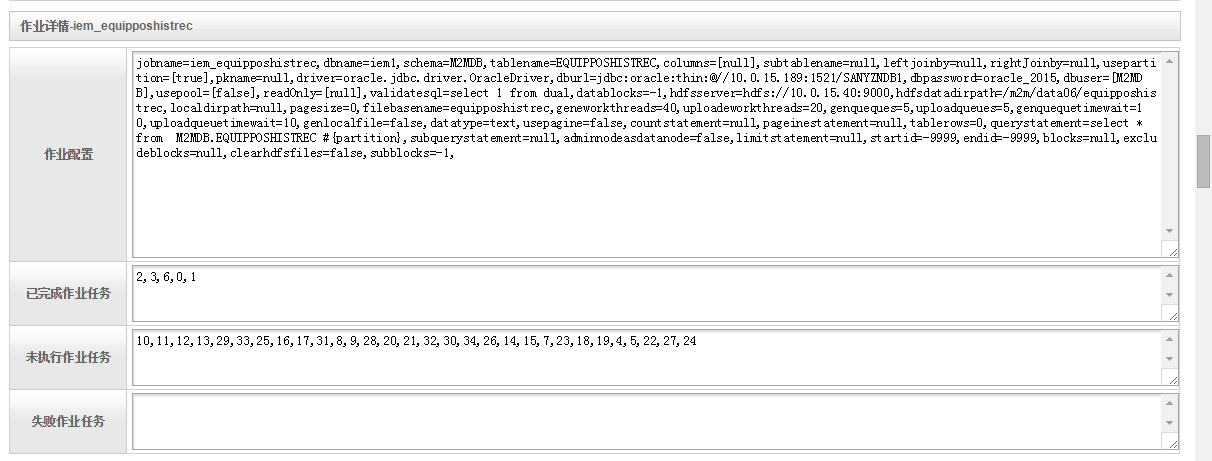
* 接收并处理管理节点分派的作业任务，为每个作业分配独立的工作处理线程池和工作队列
* 处理管理节点提交的作业监控指令，将各个作业的实施处理情况以异步事件方式返回给管理节点



# 作业监控









# 作业指令

指令：

执行作业指令

停止作业指令

同步作业状态指令

提交新作业指令

执行作业指令类型：

db数据抽取到hdfs

删除hdfs文件

停止数据源指令

任务重新分派指令

下面分别介绍各种指令配置及使用方法

# 指令详解

作业指令配置语法采用bboss ioc配置语法，大量采用扩展属性配置语法，配置灵活。

## 删除hdfs目录/文件指令

|  |  |  |  |
| --- | --- | --- | --- |
| 指令说明 | 一次性删除hdfs中的多个或者一个文件或者文件夹，重新执行作业时可以使用这个指令删除hdfs文件或者目录 | | |
| 指令参数 | 参数描述 | 默认值 | 是否必填 |
| name | 作业唯一名称 | 无 | 必填 |
| hdfsserver | 指定hdfs服务器 | 无 | 必填 |
| deletefiles | 指定要删除的文件或者文件夹，多个以逗号分隔。 | 无 | 必填 |
| 使用实例 | <property name="testdelete"  hdfsserver="hdfs://10.0.15.40:9000"  deletefiles="/m2m/sss"  /> | | |

## 停止数据源指令

|  |  |  |  |
| --- | --- | --- | --- |
| 指令说明 | 停止数据源，或者调整数据库参数后使用这个指令 | | |
| 指令参数 | 参数描述 | 默认值 | 是否必填 |
| name | 作业唯一名称 | 无 | 必填 |
| stopdbnames | 数据源名称，多个数据源用逗号分隔 | 无 | 必填 |
| 使用实例 | <property name="stopds" stopdbnames="hx125seldom" /> | | |

## 作业任务重新调度指令

## 执行作业指令

### 单任务作业指令

|  |  |  |  |
| --- | --- | --- | --- |
| 指令说明 | 一个作业只对应一个任务，适用于数据量比较小的表数据抽取到hdfs，可以通过target属性指定将作业提交给对应的数据处理节点来，不指定target则交个管理作业节点来执行（前提是管理节点也作为数据处理节点）或者数据处理节点 | | |
| 指令参数 | 参数描述 | 默认值 | 是否必填 |
| name | 作业唯一名称 | 无 | 必填 |
| dbname | 数据源名称，多个数据源用逗号分隔 | 无 | 必填 |
| dbname | 数据源名称，多个数据源用逗号分隔 | 无 | 必填 |
| driver | 数据库驱动程序，部署时必须要将对于数据库的驱动程序jar文件放到应用的lib目录中 | 无 | 必填 |
| dburl | 对应数据库的jdbc连接地址 |  |  |
| dbuser | 数据库账号名称 |  |  |
| dbpassword | 数据库账号访问口令 |  |  |
| validatesql | 连接池校验sql，非连接池模式不需要指定 |  |  |
| usepool | 标识数据源是否采用连接池，默认为true ，true采用，false不采用，在作业任务数数量比较少，则可以设置为false，作业任务比较多的情况下，可以采用连接池，重用连接 | true | 可选 |
| hdfsserver | Hdfs服务器地址 |  |  |
| hdfsdatadir | 作业数据保存到hdfs服务器的文件夹目录 |  |  |
| tablename | 指定数据库表名称 |  | 可选 |
| querystatement | 指定查询语句，如果指定了查询语句，则无需指定表名称 |  | 可选 |
| filebasename | 作业数据文件的基本文件名前缀 |  |  |
| schema | 数据库schema名称，oracle表空间查询时，必须指定 |  | 可选 |
| pkname | 表字段分区查询字段名称 |  | 可选 |
| columns | 指定要查询的字段名称列表，多个用逗号分隔 |  |  |
| single | 单任务作业标识，true 单任务，false 多任务 | false | 可选 |
| target | 指定单任务作业的数据节点，可以指定也可以不指定，如果指定的话，则在指定的数据作业节点上执行，如果不指定如果管理节点作为数据节点则直接在管理节点上执行，否则任意挑选一个作业节点执行 |  | 可选 |
| datatype | Hdfs文件存储记录数据格式，text：#分隔字段及字段值，json：json格式存储作业数据 |  |  |
| 使用实例 | <property name="qzj\_Sany\_vehicle" dbname="qzj"  driver="oracle.jdbc.driver.OracleDriver" dburl="jdbc:oracle:thin:@//10.16.1.12:1521/iem"  dbuser="SANY" dbpassword="qzjiem#EDC2011" validatesql="select 1 from dual"  usepool="false" hdfsserver="hdfs://10.0.15.40:9000" hdfsdatadir="/m2m/data05/common"  filebasename="Sany\_vehicle"  tablename="Sany\_vehicle"  single="true"  target="BPIT-10006673-2810"  datatype="text" /> | | |

### 按字段分区作业指令

|  |  |  |  |
| --- | --- | --- | --- |
| 指令说明 | 停止数据源，或者调整数据库参数后使用这个指令 | | |
| 指令参数 | 参数描述 | 默认值 | 是否必填 |
| name | 作业唯一名称 | 无 | 必填 |
| dbname | 数据源名称，多个数据源用逗号分隔 | 无 | 必填 |
| driver |  |  |  |
| dburl |  |  |  |
| dbuser |  |  |  |
| dbpassword |  |  |  |
| validatesql |  |  |  |
| usepool |  |  |  |
| hdfsserver |  |  |  |
| hdfsdatadir |  |  |  |
| tablename |  |  |  |
| filebasename |  |  |  |
| schema |  |  |  |
| pkname |  |  |  |
| columns |  |  |  |
| datablocks |  |  |  |
| geneworkthreads |  |  |  |
| genqueques |  |  |  |
| genquequetimewait |  |  |  |
| datatype |  |  |  |
| clearhdfsfiles |  |  |  |
| usepagine |  |  |  |
| startid |  |  |  |
| endid |  |  |  |
| blocks |  |  |  |
| excludeblocks |  |  |  |
| 使用实例 | 见后续实例清单 | | |

#### 实例一

按主键分区作业 自动查询分区范围（startid和endid），比较慢的数据库耗时比较长

<property name="hx20111212\_20120331"

dbname="hx243"

driver="oracle.jdbc.driver.OracleDriver"

dburl="jdbc:oracle:thin:@//10.11.16.243:1521/ISS"

dbuser="HX"

dbpassword="edcr9637\_"

validatesql="select 1 from dual"

usepool="true"

hdfsserver="hdfs://10.0.15.40:9000"

hdfsdatadir="/m2m/data04/hx20111212\_20120331"

tablename="is\_static\_20111212\_20120331"

filebasename="is\_static"

schema="ISS"

pkname="ST\_PID"

columns=""

datablocks="10000"

geneworkthreads="10"

genqueques="5"

genquequetimewait="10"

datatype="text"

clearhdfsfiles="true"

usepagine="false"

/>

#### 实例二

按主键分区作业 指定分区范围（startid和endid）

<property name="hx20111212\_20120331"

dbname="hx243"

driver="oracle.jdbc.driver.OracleDriver"

dburl="jdbc:oracle:thin:@//10.11.16.243:1521/ISS"

dbuser="HX"

dbpassword="edcr9637\_"

validatesql="select 1 from dual"

usepool="true"

hdfsserver="hdfs://10.0.15.40:9000"

hdfsdatadir="/m2m/data04/hx20111212\_20120331"

tablename="is\_static\_20111212\_20120331"

filebasename="is\_static"

schema="ISS"

pkname="ST\_PID"

columns=""

datablocks="10000"

geneworkthreads="10"

genqueques="5"

genquequetimewait="10"

datatype="text"

genlocalfile="false"

clearhdfsfiles="true"

usepagine="false"

startid="323404849"

endid="30039753017" />

### 按表分区作业指令

|  |  |  |  |
| --- | --- | --- | --- |
| 指令说明 | 停止数据源，或者调整数据库参数后使用这个指令 | | |
| 指令参数 | 参数描述 | 默认值 | 是否必填 |
| name | 作业唯一名称 | 无 | 必填 |
| dbname | 数据源名称，多个数据源用逗号分隔 | 无 | 必填 |
| driver |  |  |  |
| dburl |  |  |  |
| dbuser |  |  |  |
| dbpassword |  |  |  |
| validatesql |  |  |  |
| usepool |  |  |  |
| hdfsserver |  |  |  |
| hdfsdatadir |  |  |  |
| tablename |  |  |  |
| filebasename |  |  |  |
| schema |  |  |  |
| pkname |  |  |  |
| columns |  |  |  |
| datablocks |  |  |  |
| geneworkthreads |  |  |  |
| genqueques |  |  |  |
| genquequetimewait |  |  |  |
| datatype |  |  |  |
| clearhdfsfiles |  |  |  |
| usepartition |  |  |  |
| blocks |  |  |  |
| excludeblocks |  |  |  |
| partitions |  |  |  |
| excludepartitions |  |  |  |
| 使用实例 | 见后续实例清单 | | |

示例以

<property name="iem\_equipposhistrec" dbname="iem1"

driver="oracle.jdbc.driver.OracleDriver" dburl="jdbc:oracle:thin:@//10.0.15.189:1521/SANYZNDB1"

dbuser="M2MDB" dbpassword="oracle\_2015" validatesql="select 1 from dual"

usepool="false" hdfsserver="hdfs://10.0.15.40:9000" hdfsdatadir="/m2m/data06/equipposhistrec"

filebasename="equipposhistrec"

schema="M2MDB"

geneworkthreads="40" genqueques="5"

genquequetimewait="10"

tablename="EQUIPPOSHISTREC"

usepartition="true"

datatype="text" />

# 版本发布及部署