BI文件入库工具部署指南

(仅供内部使用）

For internal use only

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 拟制:  Prepared by | 张政勇 |  | 日期：  Date | 2014-07-21 |
| 审核:  Reviewed by |  |  | 日期：  Date | yyyy-mm-dd |
| 批准:  Granted by |  |  | 日期：  Date | yyyy-mm-dd |



华为技术有限公司

Huawei Technologies Co., Ltd.

版权所有 侵权必究

All rights reserved

修订记录Revision record·

|  |  |  |  |
| --- | --- | --- | --- |
| 日期  Date | 修订版本Revision version | 修改描述  change Description | 作者  Author |
| 2014-07-21 | V0.1 | 初稿 | 张政勇 |
| 2015-03-21 | V0.2 | 1. 分区名按下划线分割，通过配置选择某段为分区 2. 增加配置属性和命令行参数fileskip和fileskiptime，为true时库中已经存在的文件不重复入库，默认关闭 | 张政勇 |
|  |  |  |  |
|  |  |  |  |

[1 概述 4](#_Toc393995269)

[2 部署 4](#_Toc393995270)

[2.1 前置条件 4](#_Toc393995271)

[2.2 部署步骤 4](#_Toc393995272)

[3 配置 4](#_Toc393995273)

[3.1 环境配置 4](#_Toc393995274)

[3.2 业务配置 6](#_Toc393995275)

[3.2.1 文件入库到HDFS 6](#_Toc393995276)

[3.2.2 HDFS数据导入MYSQL 8](#_Toc393995277)

[3.2.3 文件入库到HDFS 10](#_Toc393995278)

[3.3 日志配置 11](#_Toc393995279)

[4 运行 12](#_Toc393995280)

[4.1 tcc自动调度 12](#_Toc393995281)

[4.2 手动执行 12](#_Toc393995282)

[4.3 日常维护 13](#_Toc393995283)

[5 附录 13](#_Toc393995284)

[5.1 FAQ 13](#_Toc393995285)

# 概述

BI文件入库工具用于业务数据入库，主要包含以下功能：

* 文件入库到HDFS
* HDFS导出到MYSQL
* HDFS数据导入文件

**说明：老业务不进行迁移，新业务逐步采用**

# 部署

## 前置条件

1. Java

推送工具采用java语言写就，故要确保java运行环境正常。安装合适的JDK，版本要求在1.6及以上。

1. 服务器

因涉及到读取MFS文件系统数据，需部署到数据通道。

## 部署步骤

1. 创建用户

为入库工具在数据通道服务器创建1个单独的操作系统用户，比如Dataload；

1. 软件包上传

将软件包DataLoad.zip上传至*/MFS/Share/Dataload/Share/bin/Datalod/*目录下，并解压。

1. 修改权限

赋予logs、work、data、appconf目录所有用户读写权限

赋予DataLoad.sh脚本运行权限

1. 开启数据通道服务器的hiveserver服务

nohup hive -–service hiveserver &

# 配置

## 环境配置

环境配置文件EnvConf.xml用于配置数据源、入库文件配置信息等运行环境信息。包括：

* TCC监听地址

上报告警事件用，预留，功能未实现

* 各种数据源

Mysql、Hive数据源

* 文件配置信息

每个业务入库到的HDFS文件系统的信息

* 业务配置文件路径

每个业务的业务配置的配置文件路径。

EnvConf.xml配置项：

|  |  |  |
| --- | --- | --- |
| 配置项名称 | 含义 | 可能取值 |
| TCC | TCC监听 | IP + 端口 |
| DBSource | 数据库数据源 | 数据库类型、IP、端口、数据库名称、用户名、密码、追加参数 |
| Fileconf | 文件配置信息 | HDFS文件路径 |
| TransportService | 传输服务 | IP、端口、用户名、登陆方式、目标路径等 |

典型配置示例如下（红色字体部分通常是需要配置的）：

<?xml version="1.0" encoding="UTF-8"?>

<Conf>

<!--

数据目标数据库，支持myql

可多个，以id唯一标示，业务根据需要自行配置

-->

<!-- 数据源数据库 Hive -->

<DBSource id="Hive">

<Type>Hive</Type>

<!-- 如果为空会取去本地IP -->

<Host>192.168.22.57</Host>

<Port>10000</Port>

<Database>default</Database>

<Username></Username>

<!-- AES加密 -->

<Password></Password>

<AdditionalInfo/>

</DBSource>

<!-- 数据目标数据库 Myql -->

<DBSource id="Mysql\_FR">

<Type>mysql</Type>

<Host>192.168.22.52</Host>

<Port>3306</Port>

<Database>z00215119</Database>

<Username>nj\_developer</Username>

<!-- AES加密 -->

<Password>907BAA918553463611A915128A40BBAE</Password>

<AdditionalInfo/>

</DBSource>

<!--

文件配置信息

可多个，以id唯一标示，业务根据需要自行配置

-->

<Fileconf id="VMallProd">

<!-- hive/hdfs存储目录 -->

<Hadoopdir>/user/hive/</Hadoopdir>

</Fileconf>

<!--

TCC监听上报告警事件用，预留，功能未实现

-->

<TCC>

<Host>10.120.5.45</Host>

<Port>8888</Port>

</TCC>

<AppConf>

<!-- 业务应用配置信息路径 -->

<include file="appconf/vmall/VMallProd\_Conf.xml"/>

</AppConf>

</Conf>

## 业务配置

### 文件入库到HDFS

将DATAPUSH推送到MFS文件系统的文件入库到HDFS文件系统。

FileToHDFS配置项说明：

|  |  |  |
| --- | --- | --- |
| 配置项名称 | 含义 | 可能取值 |
| action | 任务唯一标识 | filetohdfs\_开头的任务标示，tcc脚本中-a与之对应 |
| InputFileList | 输入文件列表 | 输入文件路径匹配规则 |
| InputFileMinCount | 输入文件个数 | 输入文件个数 超过个数任务结束，如果配置成-1，该参数无效，只考虑等待时间 |
| WaitInputMinutes | 等待输入时间（分钟） | 超过此等待时间任务结束，注意此时间一般小于tcc中配置的超时时间  超过等待时间任务没有达到输入个数(为-1时认为是1个)就报错文件未到达处理 |
| TimeOutCheckFile | 超时后检测个数 | 超过后是否检测文件个数 |
| DumpFileLoad | 空文件是否入库 | 空文件是否入库 默认true |
|  | 空文件是否算入库个数 |  |
| FileColumns | 文件列数 | 文件列数校验 <=0表示不校验 |
| Separator | 文件字段的分割符 | 在文件列数校验配置时才有效，默认使用\\0x01 |
| FileSkip | 是否跳过重名入库文件 | 是否跳过重名入库文件，为true时跳过重名文件，需要注意：命令行属性配置最优先 |
| FileSkipTime | 跳过重名文件时段 | 跳过重名文件时段，如果没有配置，默认0-24都跳过，FileSkip未false时，此属性不生效 |
| Tablename | 表名 | 入库HDFS的HIVE的表名 |
| Tablelocation | 表存储位置 | 表存储位置 |
| Partition | 分区 | none 设置成无分区每批入库都是全量覆盖  pt\_y 年  pt\_m 月  pt\_d 天  pt\_h 小时  pt\_min 分钟  也可以是时间的组合，之间用,隔开，例如pt\_m,pt\_h会先按月再按天分区 |
| PartitionPosition | 分区字段位置 | 除时间分区外的其他分区字段位置，没有设置时，按时间前的字段依次为分区位置，下标从0开始，举例：p1\_p2\_20150301120000.txt，分区pt,p1,p2对应 |

文件个数、超时时间、超时检测个数的关系

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| m>1 | n>1 |  |  |  |  |
| 文件个数 | 等待时间 | 超时后检测个数 | 结束条件 | 出错 | 成功 |
| -1 | n | - | 等待n分钟 | 文件入库0个出错 | 文件入库>=0个成功 |
| m | -1 | - | 只执行一次 | 文件入库<m个报错 | 文件入库>=m个成功 |
| m | n | TRUE | 文件到达m个 或者等待n分钟 | 文件入库<m个出错 | 文件入库>=m个成功 |
| m | n | FALSE | 文件到达m个 或者等待n分钟 | 文件入库0个出错 | 文件入库>=m个成功 |

典型配置实例如下：

<FileToHDFS action="filetohdfs\_tmp\_zzy\_test1\_9">

<!-- 文件配置信息 -->

<Fileconf>VMallProd</Fileconf>

<!-- 输入文件配置-->

<MultiInputFileConf>

<!-- 输入文件列表-->

<InputFileList>/home/z00215119/data/tmp\_zzy\_test1\_13/data\*.txt\*</InputFileList>

<!-- 输入文件个数 如果配置成-1，该参数无效，只考虑等待时间 -->

<InputFileMinCount>2</InputFileMinCount>

<!-- 空文件是否入库 默认true -->

<DumpFileLoad>fasle</DumpFileLoad>

<!-- 等待输入时间（分钟） -->

<WaitInputMinutes>1</WaitInputMinutes>

<!-- 超过等待时间是否检查文件个数，默认true -->

<TimeOutCheckFile>true</TimeOutCheckFile>

<!-- 文件列数 <=0表示不校验 -->

<FileColumns>5</FileColumns>

<!-- 文件字段的分割符，默认使用\\0x01 -->

<Separator>,</Separator>

<!--是否跳过重名入库文件，为true时跳过重名文件，命令行属性配置最优先-->

<FileSkip>true</FileSkip>

<!--跳过重名文件时段，如果没有配置，默认0-24都跳过-->

<FileSkipTime>0-8</FileSkipTime>

</MultiInputFileConf>

<!-- 表名 -->

<Tablename>tmp\_zzy\_test1\_9</Tablename>

<!-- 表存储位置 -->

<Tablelocation>tmp\_zzy\_test1\_9</Tablelocation>

<!-- 分区 none pt\_y pt\_m pt\_d pt\_h pt\_min -->

<Partition>pt\_d,pt\_h,version\_code,devip</Partition>

<!--分区字段位置，时间分区除外，下标从0开始-->

<PartitionPosition>1,2</PartitionPosition>

</FileToHDFS>

### HDFS数据导入MYSQL

将HIVE中的表数据（全表数据、分区数据、按指定条件的数据）入库到MYSQL。

可以配置的项包括：

* 任务唯一标识
* 源数据库
* 目标数据库
* 表名
* 分区信息
* 抽取SQL语句
* 表别名
* 最大允许出错记录数
* 最大出错记录百分比

HDFSToDB配置项说明：

|  |  |  |
| --- | --- | --- |
| 配置项名称 | 含义 | 可能取值 |
| action | 任务唯一标识 | filetohdfs\_开头的任务标示，tcc脚本中-a与之对应 |
| SourceDB | 源数据库标识 | 与环境配置文件数据源对应 |
| DestDB | 目标数据标识 | 与环境配置文件数据源对应 |
| Tablename | 抽取表名 | 要抽取HIVE的的表名 |
| Partition | 抽取分区 | 转换关系对照  pt\_d：pt\_d >= ${startDate} and pt\_d <= ${endDate}  pt\_w：pt\_w = ${start-endDate}  pt\_m：pt\_m = ${startMonth} |
| ExtractSql | 提取SQL语句 | 提取表的HIVE 数据提取SQL语句，此配置项目和表名、分区互斥，如果均配置以表名、分区优先  内置7个时间变量，入库工具会将其替换为实际值  ${startDate}:周期开始年月日，格式yyyyMMdd  ${endDate} ：周期结束年月日，格式yyyyMMdd  ${startWeek} ：周期开始年月日的第一个星期，格式yyyyMMdd-yyyyMMdd  ${startMonth} ：周期开始年月，格式yyyyMM  ${startHour}: 周期开始小时，格式hh  ${startMinute}:周期开始分钟，格式mm  ${start-endDate} ：周期开始结束年月日，格式yyyyMMdd-yyyyMMdd |
| AliasTablename | mysql中对应的表名 | 如果hive和mysql对应的表名相同就不用配置  ExtractSql不为空时，AliasTablename也不能为空 |
| MaxErrorNumber | 最大允许出错记录数 | 当出错记录数到达该门槛时，终止加载并告警，当为-1时表示不限制出错记录数 |
| MaxPercentNumber | 最大出错记录百分比 | 当出错记录百分比到达该门槛时，终止加载并告警，当为-1时表示不限制出错百分比 |
|  |  |  |

典型配置实例如下：

<!-- HDFS导出到MYSQL: action唯一标示一个Job ，全局唯一-->

<HDFSToDB action="hdfstodb\_tmp\_zzy\_test1\_7">

<!-- 源数据库标识，环境配置文件中获取 -->

<SourceDB>Hive</SourceDB>

<!-- 目标数据标识，环境配置文件中获取 -->

<DestDB>Mysql\_FR</DestDB>

<!-- 抽取表名 表名和ExtractSql 必须有一个不是空 -->

<Tablename>tmp\_zzy\_test1\_7</Tablename>

<!-- 抽取分区，默认规则如下

pt\_d：pt\_d >= ${startDate} and pt\_d <= ${endDate}

pt\_w：pt\_w = ${start-endDate}

pt\_m：pt\_m = ${startMonth}

-->

<Partition>pt\_d</Partition>

<!-- 提取表的HIVE 数据提取SQL语句，此配置项目和表名、分区互斥，如果均配置以表名、分区优先

内置7个时间变量，入库工具会将其替换为实际值

${startDate}: 周期开始年月日，格式yyyyMMdd

${endDate} ： 周期结束年月日，格式yyyyMMdd

${startWeek} ： 周期开始年月日的第一个星期，格式yyyyMMdd-yyyyMMdd

${startMonth} ： 周期开始年月，格式yyyyMM

${startHour}: 周期开始小时，格式hh

${startMinute}: 周期开始分钟，格式mm

${start-endDate} ：周期开始结束年月日，格式yyyyMMdd-yyyyMMdd

-->

<ExtractSql>

<![CDATA[

SELECT dev\_id FROM tmp\_zzy\_test1\_7 WHERE pt\_d=${startDate} or pt\_d=${endDate}

]]>

</ExtractSql>

<!-- mysql中对应的表名，如果hive和mysql对应的表名相同就不用配置

ExtractSql不为空时，AliasTablename也不能为空 -->

<AliasTablename>tmp\_zzy\_test1\_7</AliasTablename>

<!-- 当出错记录数到达该门槛时，终止加载并告警，当为-1时表示不限制出错记录数 -->

<MaxErrorNumber>4</MaxErrorNumber>

<!-- 当出错记录百分比到达该门槛时，终止加载并告警，当为-1时表示不限制出错百分比-->

<MaxPercentNumber>33</MaxPercentNumber>

</HDFSToDB>

### HDFS导出到文件

将HIVE中的表数据（全表数据、分区数据、按指定条件的数据）入库到MYSQL。

可以配置的项包括：

* 任务唯一标识
* 源数据库
* 目标文件目录
* 导出文件字段的分割符
* 表名
* 分区信息
* 抽取SQL语句
* 导出的文件名

HDFSToDB配置项说明：

|  |  |  |
| --- | --- | --- |
| 配置项名称 | 含义 | 可能取值 |
| action | 任务唯一标识 | filetohdfs\_开头的任务标示，tcc脚本中-a与之对应 |
| DestFilePath | 目标文件目录 | 目标文件目录，需要具有写权限 |
| Separator | 导出文件字段的分割符 | 默认使用\001 |
| Tablename | 抽取表名 | 要抽取HIVE的的表名 |
| Partition | 抽取分区 | 转换关系对照  pt\_d：pt\_d >= ${startDate} and pt\_d <= ${endDate}  pt\_w：pt\_w = ${start-endDate}  pt\_m：pt\_m = ${startMonth} |
| ExtractSql | 提取SQL语句 | 提取表的HIVE 数据提取SQL语句，此配置项目和表名、分区互斥，如果均配置以表名、分区优先  内置7个时间变量，入库工具会将其替换为实际值  ${startDate}:周期开始年月日，格式yyyyMMdd  ${endDate} ：周期结束年月日，格式yyyyMMdd  ${startWeek} ：周期开始年月日的第一个星期，格式yyyyMMdd-yyyyMMdd  ${startMonth} ：周期开始年月，格式yyyyMM  ${startHour}: 周期开始小时，格式hh  ${startMinute}:周期开始分钟，格式mm  ${start-endDate} ：周期开始结束年月日，格式yyyyMMdd-yyyyMMdd |
| AliasFilename | 导出的文件名 | 如果导出文件名与和hive对应的表名相同就不用配置 |

典型配置实例如下：

<!-- HDFS导出到文件: action唯一标示一个Job ，全局唯一-->

<HDFSToFile action="hdfstofile\_tmp\_zzy\_test1\_7">

<!-- 源数据库标识，环境配置文件中获取 -->

<SourceDB>Hive</SourceDB>

<!-- 目标文件目录 -->

<DestFilePath>/data1/z00215119/temp/testfiletohdfs/exportdata</DestFilePath>

<!-- 导出文件字段的分割符，默认使用\001 -->

<Separator>,</Separator>

<!-- 抽取表名 -->

<Tablename>tmp\_zzy\_test1\_7</Tablename>

<!-- 抽取分区，默认规则如下

pt\_d：pt\_d >= ${startDate} and pt\_d <= ${endDate}

pt\_w：pt\_w = ${start-endDate}

pt\_m：pt\_m = ${startMonth}

-->

<Partition>pt\_d</Partition>

<!-- 提取表的HIVE 数据提取SQL语句，此配置项目和表名、分区互斥，如果均配置以表名、分区优先

内置7个时间变量，入库工具会将其替换为实际值

${startDate}: 周期开始时间，格式yyyyMMdd

${endDate} ： 周期结束时间，格式yyyyMMdd

${startWeek} ： 周期开始年月日的第一个星期，格式yyyyMMdd-yyyyMMdd

${startMonth} ： 周期结束时间，格式yyyyMM

${startHour}: 周期开始时间，格式hh

${startMinute}: 周期开始时间，格式mm

${start-endDate} ：周期结束时间，格式yyyyMMdd-yyyyMMdd

-->

<ExtractSql>

<![CDATA[

SELECT dev\_id , pt\_d FROM tmp\_zzy\_test1\_7 WHERE pt\_d=${startWeek}

]]>

</ExtractSql>

<!-- 导出的文件名，如果导出文件名与和hive对应的表名相同就不用配置

入库工具也会将上述7个变量替换为实际值-->

<AliasFilename>tmp\_zzy\_test1\_7\_2</AliasFilename>

</HDFSToFile>

## 日志配置

入库工具采用log4j作为日志组件，可以设置日志级别、输出目的，对日志输出进行控制。

通常，调试阶段，日志级别可设置为DEBUG级别输出。

在线运行时需将日志级别设置为INFO输出！

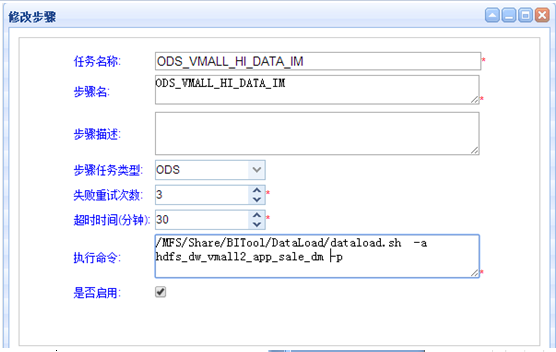
#日志级别

log4j.rootLogger=DEBUG, console, file

# 运行

## tcc自动调度

* + - 1. 在tcc的任务步骤中指定脚本全路径和action
      2. 不要忘记周期-p



## 手动执行

通用的命令行命令Dataload.sh

通用参数：

-a 任务标识 对应下章节降到的应用标示action字段

-p 任务周期信息 周期开始日期+周期结束日期+周期类型+周期长度

-D 属性配置文件，使用;分割，目前只有支撑一个属性fileskip=false|true

例子：Dataload.sh -a hdfs\_dw\_vmall2\_app\_sale\_dm -p ‘20140628-0000 20140628-0000 D 1’ -D fileskip=false

注意：**-p 后面一定要带单引号，否则参数会被截断**

## 日常维护

无

# 附录

## FAQ

暂无