# AutoLoad使用文档

|  |  |
| --- | --- |
| 版本 | 2.0.1 |
| 日期 | 2016-01-25 |
| 修订人 | 何云洋 |

* 前言：

AutoLoad是一个把csv文件导入到mysql的程序，主要针对于数据仓库所涉及的表，一般使用最多有维度（Dimension）表和事实（Fact）表，针对这类型的表，可实现增量导入，如每天将T+1的数据导入到mysql，本程序可支持任意导入的频度，这要看你怎么使用。

* 流程：

1. 配置目标Mysql数据库；（配置表：AL\_DataSource）
2. 从源库中查询出维度表与事实表生成csv，csv文件名需要按照一定规则定义，以便于正则表达式能很好的匹配；（配置表：[AL\_InputGroup]、AL\_InputFile、[AL\_InputFileField]）
3. 根据生成文件名，匹配需要截取的维度；（配置表：[AL\_InputFileConstant]）
4. 根据维度删除该批次的数据（删除可能出现的重复记录数，配置表：[AL\_InputFileBefore]）
5. 执行命令开始导入数据

* 配置表结构与详解





* 程序调用说明

如：java -jar xxx.jar org.blazer.dataload.handle.ALHandle \\* \\*

第一个参数是AL\_InputGroup表的RecordID

第二个参数是AL\_InputFile表的RecordID

两者是and的关系,匹配所有则填写\\*,RecordID以,分割

支持NotIn和In语法,默认是In

如：java -jar AutoLoad.jar org.blazer.dataload.handle.ALHandle NotIn3 In100,101,102

IN和逗号以及ID中间不能有空格

如：java –jar AutoLoad.jar AL \\* \\*

* 需要注意的问题

1. 为了提高事实表的插入效率，程序会将维度表读入内存中，因此维度表不应该太大，否则就需要调整内存，一般情况下，默认即可。
2. 为了提高事实表的插入效率，程序会将csv文件读入内存中，因此csv文件也不应该太大，如果是流水日志似的数据，则可以实行其他策略进行插入，这里需要扩展程序。

* 测试案例：

维度表：

周期维度表（Dim\_PeriodKey）通用维度表

日期维度表（Dim\_DayKey）通用维度表

产品等级维度表（Dim\_Product\_Level）

产品维度表（Dim\_Product）

销售员维度表（Dim\_Salesman）

事实表

销售事实表（Fact\_Sales）

表结构的SQL文件：



假设你现在有一个csv文件fact\_sales\_1\_20160128.csv：

一级 MacPro 张三 120000 160000

一级 MacPro 李四 90000 120000

一级 MacPro 王五 45000 60000

一级 MacPro 赵六 15000 20000

一级 MacBook 张三 15000 20000

注意：字段处理、空值处理：

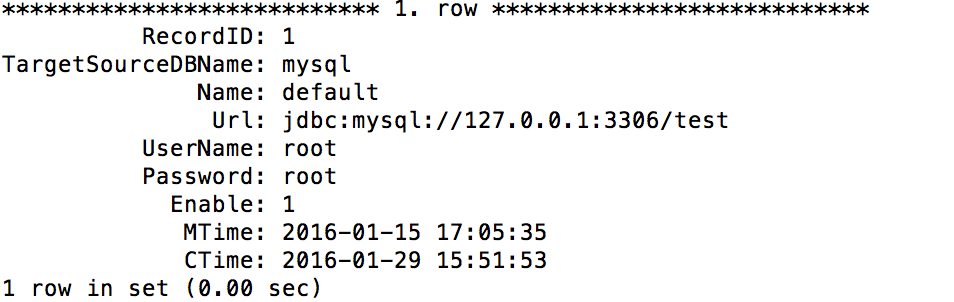
* 1. 当没有读到该列，或者该列的值为null（不区分大小）、#NULL#、”N，的时候，程序会转换成数据库中的NULL值。
  2. 当该列值为#NULL\_STR#时，程序会转换成数据库中的null字符串（小写）。
  3. 当以上情况都没匹配到，则用该列真实字符串值

要说明一下：前三列是维度，分别是维度表Dim\_Product\_Level、Dim\_Product、Dim\_Salesman，后两列是聚合值

现在，我需要将这样一个csv文件，导入到mysql中。

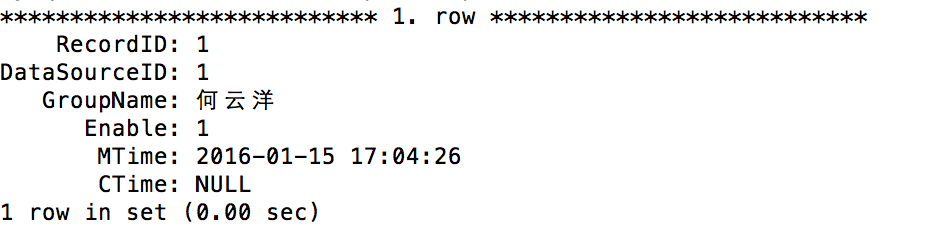
配置过程图文详解：

1. 配置AL\_DataSource



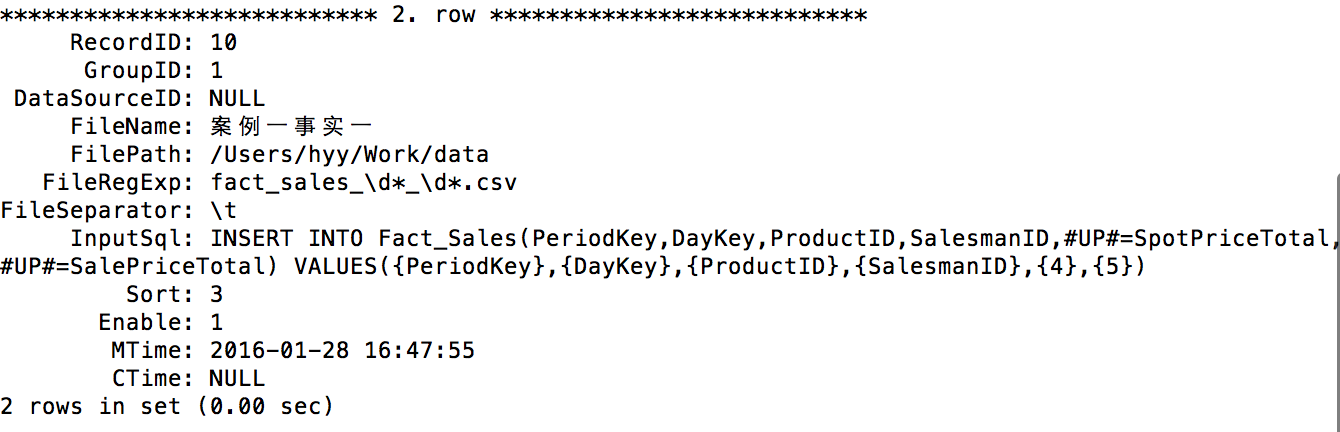
配置数据源，我这里是本地测试环境

1. 配置AL\_InputGroup



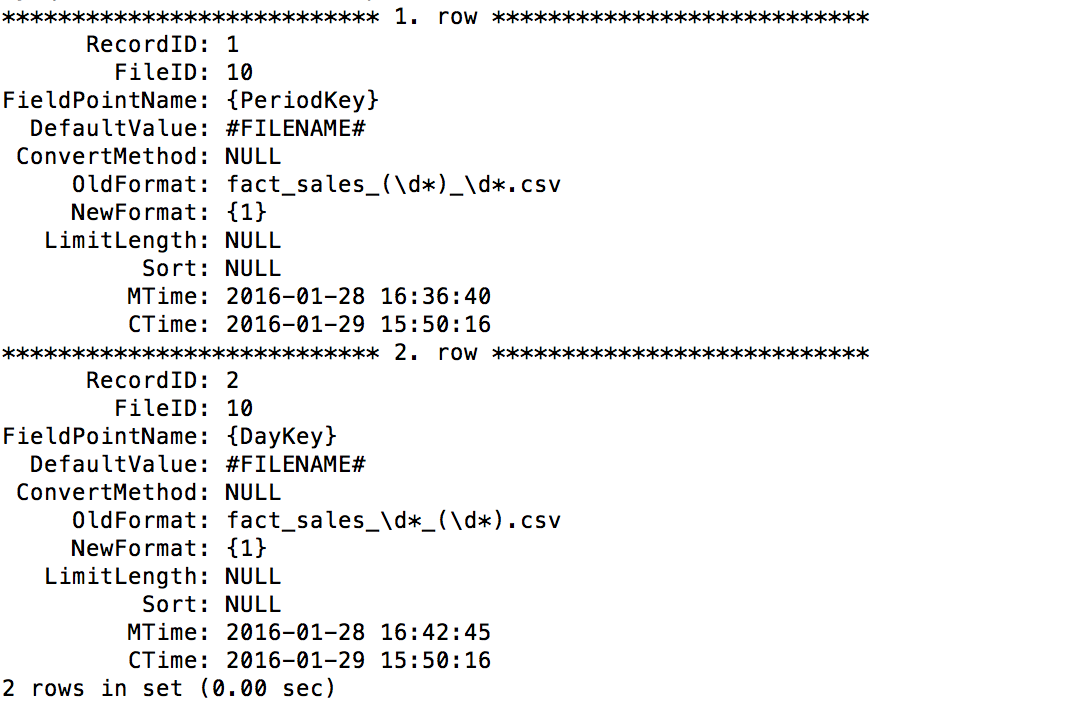
这里配置分组表，主要是将一组业务的csv分到一个组。并且使用同一个数据源

1. 配置AL\_InputFile



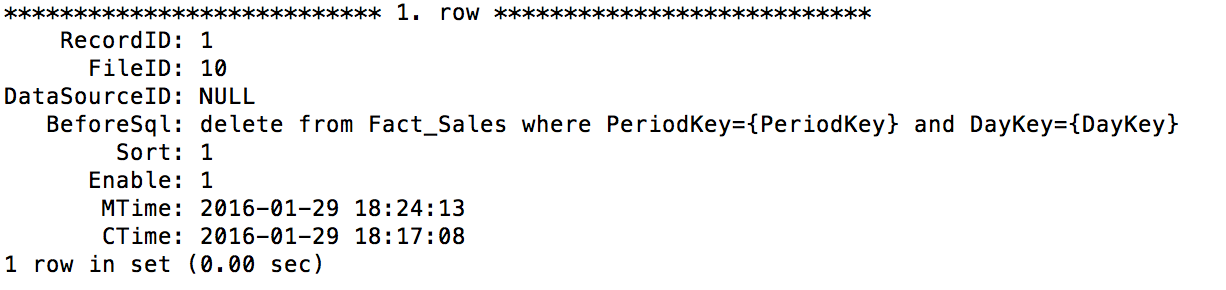
这里DataSourceID可以不填写，会自动引用GroupID中的DataSourceID

1. 配置AL\_InputFileConstant

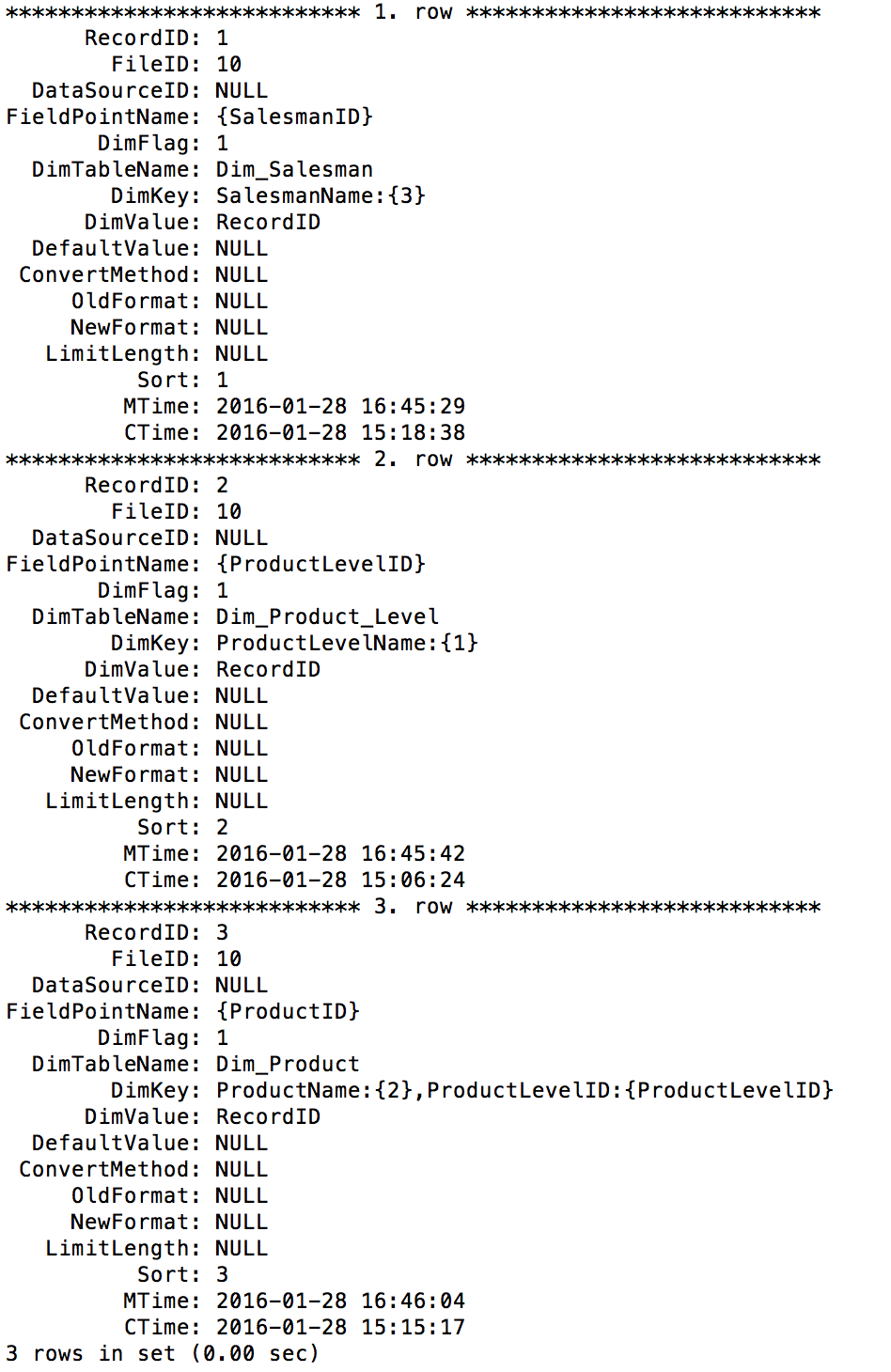


FieldPointName:{PeriodKey}是我自定义的变量，可以在之后配置中使用

1. 配置AL\_InputFileBefore



1. 配置AL\_InputFileField



这里要详细说明一下，这里配置的多重维度（维度里面包含维度），我们看到第三条记录的DimKey里面包含着第二条记录的ProductLevelID，如果需要配置的不是一个维度，是csv中的一个字段，那么可以参照AL\_InputFileConstant配置，只做转换使用。

1. 系统日志表，无需配置，可查询每次执行的信息。

