数据迁移之PIG篇—如何处理HBASE数据含换行符问题

目录

[一、背景 2](#_Toc420157768)

[二、PIG代码修改 2](#_Toc420157769)

[三、示例 4](#_Toc420157770)

# 一、背景

从HBase向关系库迁移数据的时候，方便的工具不多，PIG是可选工具之一。当HBase表字段数据中含有换行符的时候，所有工具可能都会失效，PIG也不例外。通过本文给大家介绍如何通过PIG处理这个问题。

先看下，通过PIG导出HBase数据脚本是怎样的。



导出数据时，可以指定列分割符（PigStorage第一个参数），但是却没有可以指定换行符的地方，如果导出时，我们可以自定义换行符，那么前面我们提到的问题就可以迎刃而解了。

# 二、PIG代码修改

翻看下PIG的代码，发现PIG的输出主要由PigTextOutputFormat内部类PigLineRecordWriter处理，其中对行结束符的处理代码如下：



通过源码，我们可以看到在PIG的输出中，行结束符是程序写死的‘\n’，看来我们就不得不改造下PIG代码了。

我的修改方式是增加PigStorage带行结束符的构造函数，然后把这个行结束符带到输出类中，当然也可以采用其它方式处理，这里就不一一介绍了。

详细修改就不一一列举，只提几处关键修改点。



增加PigTextOutputFormat带行结束符的构造函数。



把传入PigTextOutputFormat的行结束符传递给PigLineRecordWriter。



增加PigStorage带行结束符的构造函数。



把传入的行结束符传递给PigTextOutputFormat。

把编译后的文件放到PIG的pig-0.12.0-cdh5.0.4-withouthadoop.jar、pig-0.12.0-cdh5.0.4.jar里就可以使用了。

# 三、示例

假设HBASE表test的f1:info字段里含有换行符\n,以此为例，看如何把该表数据正确导入到ORACLE库中。

第一步，PIG脚本，我们指定行结束符’$$’，脚本如下：



第二步，对输出文件进行处理（替换数据字段里的换行符，替换$$为\n），处理脚本（shell脚本）如下（test.sh）：



第三步，处理PIG输出文件



第一个参数是PIG的输出文件，第二个参数PIG指定的行结束符，第三个是处理后的文件。

第四步，把处理后的数据文件导入到ORACLE中，这个可选方式就比较多了，不作说明。

第五步，把导入后的数据内容替换回换行符\n。

以ORACLE为例： update mytest set sinfo = replace(sinfo,'##',chr(10))

最终结果：



如需要修改后的PIG JAR包，可联系kongml@yonyou.com。