Hive:CSV Serde介绍以使用

CSV格式的文件也称为逗号分隔值（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号。在本文中的CSV格式的数据就不是简单的逗号分割的），其文件以纯文本形式存储表格数据（数字和文本）。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符。通常，所有记录都有完全相同的字段序列。

在Hive的0.14版(可以看这里https://issues.apache.org/jira/browse/HIVE-7777)中提供了一个CSVSerde解析器，我们可以很方便的使用Hive解析CSV格式的数据，这个SerDe支持很多种格式的CSV数据，但是如果你数据中包含了换行符，这就不行了。如果你需要在Hive中使用CSVSerde，只需要在建表语句中指定Serde为org.apache.hadoop.hive.serde2.OpenCSVSerde。建表语句如下：

```
CREATE TABLE iteblog_table(a string, b string, ...)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = "₩t",
  "quoteChar"     = "'",
  "escapeChar"    = "₩₩"
)
STORED AS TEXTFILE;
```

默认的分隔符是

```
DEFAULT_ESCAPE_CHARACTER ₩
DEFAULT_QUOTE_CHARACTER  "
DEFAULT_SEPARATOR        ,
```

可以通过separatorChar、quoteChar以及escapeChar指定数据的分割。其实org.apache.hadoop.hive.serde2.OpenCSVSerde的实现依赖了Open-CSV 2.3库，如果你使用的Hive低于Hive 0.14版，你可以自己写一个类似功能的Serde；或者使用https://github.com/ogrodnek/csv-serde。其实Hive 0.14中的OpenCSVSerde就是基于这个实现的。

本博客文章除特别声明，全部都是原创！
原创文章版权归过往记忆大数据（过往记忆）所有，未经许可不得转载。
本文链接: 【】（）