

知 使用Oratompp+dispatch进行单条记录中存在回车的数据表的ETL工作

谢超凡 2017-05-21 发表

Q：数据表中每一条数据成为一条记录，由于业务系统设计或者使用者操作不当的原因，有的记录中会存在回车符，使用kettle进行ETL时可以对这种数据进行清洗，但是kettle面对数据量巨大（上亿条记录）速度过慢，只能使用Oratompp+dispatch进行ETL工作。

A：在执行oratompp的脚本命令中使用行分隔符关键字"line_separator"，例如：./OratoMPP --user='u02495/i39@10.1.17.8:1521/HNYY' --query="select * FROM 病案主页" --file='/opt/data/BAZY.txt' --field="|++|" --line_separator=="|@@" --format=3 --parallel=4

注意，行分隔符需要和列分隔符（field关键字）设置的不同，且尽量复杂。

同时，在dispatch的配置文件中，增加oratompp执行命令中设置的行分隔符"line_separator=="|@@"