gpload

按照一个YAML格式的控制文件的定义运行一个装载作业。

概要

```
gpload -f control_file [-l log_file] [-h hostname] [-p port]
   [-U username] [-d database] [-W] [--gpfdist_timeout seconds]
   [--no_auto_trans] [[-v | -V] [-q]] [-D]

gpload -?

gpload --version
```

先决条件

要执行gpload命令的客户机必须具有下列要求：

- Python 2.6.2或更新版本，装有pygresql (Python的PostgreSQL接口包)，和pyyaml。注意Python及所需的Python库被包含在Greenplum数据库安装包中， 因此如果在gpload运行的机器上安装有Greenplum数据库，用户就不需要单独安装Python。
  Note: Greenplum数据库的Windows装载客户端仅支持Python 2.5(您可以从https://www.python.org 获取)
  。
- gpfdist并行文件分发程序被安装在$PATH中。这个程序位于Greenplum 数据库的$GPHOME/bin目录下。
- gpload客户机可以访问（被访问）Greenplum数据库集群（Master和Segments）中所有主机。
- gpload客户及可以访问（被访问）所有可能用来装载数据的主机（ETL服务器）。

描述

gpload是一个数据装载工具，它扮演着Greenplum数据库外部表并行装载 特性的接口的角色。通过一个用YAML格式控制文件定义的装载说明，gpload 调用Greenplum数据库的并行文件服务器 (gpfdist)执行 文件装载，基于源数据的定义创建一个外部表定义，并且指定INSERT、UPDATE或MERGE操作把源数据装载到数据库中的目标表中。
Note: gpfdist和gpload是在Greenplum的主版本级别有效的。 例如，Greenplum 4.x版本的gpfdist不能用于Greenplum 5.x或6.x版本。
Note: 如果目标表的列名为保留关键字、有大写字母或包含任何双引号， 那么MERGE和UPDATE操作不被支持。

在目标表上指定多个同时的装载操作时，操作包括在YAML控制文件(控制文件格式见控制文件格式)的SQL集合中指定的任何SQL命令会在单个事务中 执行以防止数据不一致。

选项

-f *control_file*
: **必选项。**包含装载说明详情的YAML文件。请见 控制文件格式。

--gpfdist_timeout *seconds*
: 为gpfdist并行文件分发程序发送响应设置超时时间。 输入一个从0到30秒的值（输入"0"会禁用超时）。 注意在高流量网络上可能需要增加这个值。

-l *log_file*
: 指定在哪里写日志文件。默认是~/gpAdminLogs/gpload_YYYYMMDD。 有关日志文件的更多信息请见日志文件格式。

--no_auto_trans
: 如果用户在目标表上执行单个装载操作，可指定--no_auto_trans 禁用把装载操作作为单个事务处理的特性。
  默认情况下，在一个目标表上执行多个同时的操作时，gpload把每个装载操作处理为 单个事务以防止不一致的数据。

-q (无屏幕输出)
: 运行在静默模式中。命令输出不会被显示在屏幕上，但仍将被写入到日志文件。

-D (调试模式)
: 检查错误情况，但是不执行装载。

-v (详细模式)
: 在装载步骤被执行时，显示它们的详细输出。

-V (非常详细模式)
: 显示非常详细的输出。

-? (显示帮助)

显示帮助，然后退出。

--version

显示这个工具的版本，然后退出。

连接选项

-d *database*

要装载到的数据库。如果没有指定，则从装载控制文件、环境变量$PGDATABASE读取或者默认为当前系统用户名。

-h *hostname*

指定Greenplum的Master数据库服务器在其上运行的机器的主机名。如果没有指定， 会从装载控制文件、环境变量$PGHOST读取或者默认为 localhost。

-p *port*

指定Greenplum的Master数据库服务器在其上监听连接的TCP端口。如果没有指定， 会从装载控制文件、环境变量$PGPORT读取或者默认为5432。

-U *username*

要用其进行连接的数据库角色名。如果没有指定，会从装载控制文件、环境变量$PGUSER读取或者默认为当前系统用户名。

-W (强制口令提示)

强制口令提示。如果没有指定，会从环境变量$PGPASSWORD、 $PGPASSFILE指定的口令文件或~/.pgpass中的口令文件中读取口令。如果这些都没有设置，即使没有提供-W， gpload也将提示要求一个口令。

控制文件格式

gpload控制文件使用 YAML 1.1 文档格式，然后为定义Greenplum数据库装载操作的多个步骤实现了其自身的模式。 该控制文件必须是一个有效的YAML文档。

gpload程序按顺序处理控制文件文档并且使用缩进（空格）来判断文档层次以及 各个部分之间的关系。空格的使用是有意义的。不能把空格简单地用于格式化目的，并且不能使用制表符。

一个装载控制文件的基础结构是：

```
---
VERSION: 1.0.0.1
DATABASE: db_name
USER: db_username
HOST: master_hostname
PORT: master_port
GPLOAD:
   INPUT:
    - SOURCE:
                 LOCAL_HOSTNAME:
                     - hostname_or_ip
                 PORT: http_port
             | PORT_RANGE: [start_port_range, end_port_range]
                 FILE:
                     - /path/to/input_file
                 SSL: true | false
                 CERTIFICATES_PATH: /path/to/certificates
    - FULLY_QUALIFIED_DOMAIN_NAME: true | false
    - COLUMNS:
                     - field_name: data_type
    - TRANSFORM: 'transformation'
      - TRANSFORM_CONFIG: 'configuration-file-path'
      - MAX_LINE_LENGTH: integer
      - FORMAT: text | csv
      - DELIMITER: 'delimiter_character'
      - ESCAPE: 'escape_character' | 'OFF'
      - NULL_AS: 'null_string'
      - FORCE_NOT_NULL: true | false
      - QUOTE: 'csv_quote_character'
      - HEADER: true | false
      - ENCODING: database_encoding
    - ERROR_LIMIT: integer
    - LOG_ERRORS: true | false
   EXTERNAL:
      - SCHEMA: schema | '%'
   OUTPUT:
    - TABLE: schema.table_name
    - MODE: insert | update | merge
    - MATCH_COLUMNS:
                     - target_column_name
    - UPDATE_COLUMNS:
                     - target_column_name
    - UPDATE_CONDITION: 'boolean_condition'
    - MAPPING:
```

```
                target_column_name: source_column_name | 'expression'
    PRELOAD:
     - TRUNCATE: true | false
     - REUSE_TABLES: true | false
     - STAGING_TABLE: external_table_name
     - FAST_MATCH: true | false
    SQL:
     - BEFORE: "sql_command"
     - AFTER: "sql_command"
```

VERSION

可选。gpload控制文件模式的版本。当前版本是1.0.0.1。

DATABASE

可选。指定Greenplum数据库系统要连接到哪个数据库。如果没有指定则默认为$PGDATABASE。 如果$PGDATABASE也没有设置，则默认为当前系统用户名。用户还可以在命令行上用 -d选项指定数据库。

USER

可选。指定用于连接的数据库角色。如果没有指定，默认为当前用户或者$PGUSER （如果设置）。用户还可以在命令行上用-U选项指定数据库角色。

如果运行gpload的用户不是Greenplum数据库的超级用户， 那么必须为该用户授予适当的权限。更多信息请见*Greenplum数据库参考指南*。

HOST

可选。指定Greenplum数据库的Master主机名。如果没有指定，默认为localhost或者 $PGHOST （如果设置）。用户还可以在命令行上用-h 选项指定Master主机名。

PORT

可选。指定Greenplum数据库的Master端口。如果没有指定，默认为5432或者 $PGPORT （如果设置）。用户还可以在命令行上用-p 选项指定Master端口。

GPLOAD

必需。开始装载说明部分。GPLOAD说明必须定义有一个INPUT 小节和一个OUTPUT小节。

INPUT

必需。定义要装载的输入数据的位置和格式。gpload 将在当前主机上启动gpfdist文件分布程序的一个或者更多 实例并且在Greenplum数据库中创建指向源数据的外部表定义。注意在其上运行 gpload的主机必须对所有的Greenplum数据库主机（Master和Segment） 通过网络可访问。

SOURCE

必需。 INPUT说明的SOURCE块定义源文件的位置。 一个INPUT小节可以定义多个SOURCE块。每个定义 的SOURCE块对应于将在本地机器上启动的一个gpfdist文件分布程序的实例。 每个定义的SOURCE块必须有一个FILE说明。

更多关于使用gpfdist并行文件服务器和单个以及多个 gpfdist实例的信息，请见*Greenplum数据库管理员指南* 中的“装载和卸载数据”部分。

LOCAL_HOSTNAME

可选。指定gpload运行其上的本地机器 的主机名或者IP地址。如果这个机器被配置有多个网络接口卡（NICs）， 用户可以指定每块NIC的主机名或者IP，以便允许网络流量同时使用所有的NIC。 默认是仅使用本地机器的主要主机名或者IP。

PORT

可选。指定gpfdist 文件分布程序应该使用的特定端口号。用户还可以提供一个PORT_RANGE 来从指定的范围中选择可用的端口。如果PORT和 PORT_RANGE同时被定义，那么PORT优先。 如果PORT和PORT_RANGE都没有定义， 默认为在8000和9000之间选择一个可用端口。

如果在LOCAL_HOSTNAME中声明多个主机名， 这个端口号被用于所有主机。如果用户想要使用所有的NICs装载一个给定目录位置的同一个 文件或者文件集合，这种配置就是用户想要的。

PORT_RANGE

可选。可被用来代替PORT提供一个端口号范围， gpload可以从其中为这个gpfdist 文件分布程序实例选择一个可用的端口。

FILE

必需。指定本地文件系统上的一个文件位置、命名管道或者目录位置， 其中包含要被装载的数据。用户可以声明多于一个文件， 只要所有指定文件中数据的格式相同。

如果这些文件被使用gzip或者bzip2 （有.gz或者.bz2文件扩展名）压缩， 这些文件将被自动解压缩（在用户路径中有gunzip或者 bunzip2）。

在指定要装载哪些源文件时，用户可以使用通配符 (*) 或其他C风格的模式匹配来指示多个文件。被指定的文件假定在相对于gpload 被执行的当前目录的位置（或者用户可以声明绝对路径）。

SSL

可选。指定SSL加密的使用。 如果SSL被设置为true， gpload 用--ssl启动gpfdist 服务器并且使用gpfdists://协议。

CERTIFICATES_PATH

当SSL为true时必需； 当SSL为false或者没有指定时 不能指定这个参数。CERTIFICATES_PATH 中指定的位置必须包含下列文件：

- 服务器证书文件 server.crt
- 服务器私钥文件 server.key
- 可信证书授权 root.crt

根目录(/)不能被指定为 CERTIFICATES_PATH。

FULLY_QUALIFIED_DOMAIN_NAME

可选。指定gpload是否把主机名解析成完全限定的域名 （FQDN）或者本地主机名。如果值被设置为true， 名称会被解析到FQDN。如果该值被设置为false， 则解析到本地主机名。默认是false。

在某些情况下可能要求一个完全限定的域名。例如，如果Greenplum 数据库系统在与ETL应用不同的域且该域能够被gpload访问。

COLUMNS

可选。以field_name:data_type 这样的格式指定源数据文件的模式。源文件中的DELIMITER 字符是分隔两个数据值域（列）的东西。一行由一个换行字符（0x0a决定）。

如果输入COLUMNS没有指定，则使用输出TABLE的模式， 意味着源数据必须与目标表具有相同的列序、列数以及数据格式。

默认的source-to-target映射基于这一节定义的列名与目标TABLE 中列名之间的匹配。默认映射可以使用MAPPING小节覆盖。

TRANSFORM

可选。指定传递给gpload的输入转换的名字。 有关XML转换的信息，请见*Greenplum数据库管理员指南* 中的“装载和卸载数据”。

TRANSFORM_CONFIG

当TRANSFORM被指定时，这个元素是必需的。 指定在上面TRANSFORM参数中指定的转换的 配置文件位置。

MAX_LINE_LENGTH

可选。一个整数，指定传递给gpload 的XML转换数据中一行的最大长度。

FORMAT

可选。指定源数据文件的格式：纯文本（TEXT）格式， 逗号分隔值（CSV）格式。如果没有指定，这个默认为 TEXT。更多有关源数据格式的信息，请见 *Greenplum数据库管理员指南*中的“装载和卸载数据”。

DELIMITER

可选。指定在每行数据内分隔列的单个ASCII字符。在TEXT模式中默认是一个制表符， 在CSV模式中默认是一个逗号。用户还可以指定一个非可打印ASCII字符或者非可打印 Unicode字符，例如：″\x1B″或者 ″\u001B″。 对于非可打印字符也支持转义字符串语法E’character-code’。 ASCII或Unicode字符必须被封闭在单引号中。例如：E’\x1B’或者 E’\u001B’。

ESCAPE

指定用于C转义序列（例如\n、\t、 \100等等）以及转义可能被当作行列定界符的数据字符的 单个字符。确保选择一个在实际列数据中任何地方都没有使用的转义字符。文本 格式文件的默认转义字符是一个\（反斜线）， csv格式文件的默认转义字符是 一个″（双引号）。不过可以指定另一个字符来表示转义。 还可以在文本格式文件中通过指定’OFF’值作为转义值来 禁用转义。这对于其中嵌有很多不准备作为转义字符的反斜线的文本格式的Web 日志数据非常有用。

NULL_AS

可选。指定表示空值的字符串。TEXT模式中默认是\N （反斜线-N）， CSV模式中默认是没有引用的空的值。 即便在TEXT模式中，对于想要把空值与空字符串区分 开来的情况，用户也可以使用空字符串。任何匹配这个字符串的源数据项将 被认为是一个空值。

FORCE_NOT_NULL

可选。在CSV模式中，处理每个被指定的列，仿佛它被引用并且因此 不是一个NULL值。对于CSV模式中的默认空值字符串（两个定界符之 间什么都没有），这导致缺失的值被计算为长度为零的字符串。

QUOTE

当FORMAT是CSV时， 这个元素是必需的。为CSV模式指定引用 字符。默认是双引号（″）。

HEADER

可选。指定数据文件中的第一行是一个头部行（包含列名）并且不 应被包括在要被装载的数据中。如果使用多个数据源文件，所有的 文件必须有一个头部行。默认是假定输入文件没有头部行。

ENCODING

可选。源数据的字符集编码。可指定一个字符串常量（例如’SQL_ASCII’）、 一个整数编码编号，或者指定’DEFAULT’以使用默认客户端编码。 如果没有指定，默认的客户端编码会被使用。有关支持的字符集的信息， 请见*Greenplum数据库参考指南*。

ERROR_LIMIT

可选。为这个装载操作启用单行错误隔离模式。当被启用时， 在输入被处理期间只要没有达到错误限制计数，任何Greenplum数据库 Segment会抛弃有格式错误的输入行。如果错误限制没有达到，所有 好的行将会被装载并且任何错误行都将被抛弃或者被捕捉在错误日志 信息中。默认是在遇到第一个错误时中止装载操作。注意单行错误隔离 只适用于有格式错误的数据行，例如有额外或者缺失的属性、有错误 数据类型的属性或者有无效的客户端编码序列。如果遇到约束错误（例如主键约束）仍将导致装载操作中止。有关处理装载错误的信息， 请见*Greenplum数据库管理员指南*中的“装载和卸载数据”。

LOG_ERRORS

当ERROR_LIMIT被声明时，这个元素是可选的。 值可以是true或者false。 默认值是false。如果值是true， 当运行在单行错误隔离模式中时，格式错误的行会被内部记录下来。 用户可以用Greenplum数据库的内建SQL函数gp_read_error_log(’table_name’)检查格式错误。如果在装载数据时检测到格式错误，gpload 会用包含错误信息的表的名字生成一个警告消息，看起来类似于这个消息。

```
timestamp|WARN|1 bad row, please use GPDB built-in function gp_read_error_log('table-name')
   to access the detailed error row
```

如果LOG_ERRORS: true被指定，必须指定REUSE_TABLES: true 以便在Greenplum数据库的错误日志中保留格式错误。如果没有指定REUSE_TABLES: true， 错误信息会在gpload操作后被删除。只有关于格式错误的总结信息会被返回。 用户可以用Greenplum数据库的函数gp_truncate_error_log()从错误日志 中删除格式错误。
更多有关处理装载错误的信息，请见*Greenplum数据库管理员指南*中的"装载和卸载数据"。 有关gp_read_error_log()函数的信息，请见*Greenplum数据库参考指南* 中的CREATE EXTERNAL TABLE命令。

EXTERNAL
可选。定义gpload创建的外部表数据库对象所属的方案。
默认是使用Greenplum数据库的search_path。

SCHEMA
当EXTERNAL被声明时，这个元素是必需的。 外部表所在的方案的名称。如果该方案不存在，会返回一个错误。
如果%（百分号字符）被指定， 会使用OUTPUT小节中TABLE 指定的表名的方案。如果这个表名没有指定一个方案，则会使用默认方案。

OUTPUT
必需。定义要被装载到数据库中的目标表和最终数据列值。

TABLE
必需。要装载到其中的目标表名。

MODE
可选。如果没有指定，则默认为INSERT。 有三种可用的装载模式:
INSERT - 使用下列方法装载数据到目标表中:

```
INSERT INTO target_table SELECT * FROM input_data;
```

UPDATE - 更新目标表中MATCH_COLUMNS 属性值等于输入数据并且UPDATE_CONDITION为true（可选条件） 的行的UPDATE_COLUMNS。如果目标表的列名为保留关键字、 有大写字母或包含双引号（""），那么不支持UPDATE。
MERGE - 插入新行并且更新FOOBAR 属性值等于相应输入数据而且MATCH_COLUMNS为true （可选条件）的已有行的UPDATE_COLUMNS。当源数据 中的MATCH_COLUMNS值在目标表数据中没有相应值时会 被标识成新行。在那种情况下，源文件中的**整个行**会被插入，而不仅仅 是MATCH和UPDATE列。如果有多个 相等的新MATCH_COLUMNS值，只有其中一个新行将被插入。 使用UPDATE_CONDITION可过滤掉要抛弃的行。 如果目标表的列名为保留关键字、 有大写字母或包含双引号（""），那么不支持MERGE。

MATCH_COLUMNS
如果MODE为UPDATE或者 MERGE，则这个元素是必需的。指定被用作更新 的连接条件的列。对于要在目标表中更新的行，指定目标列中的属性值 必须等于相应的源数据列值。

UPDATE_COLUMNS
如果MODE为UPDATE 或者MERGE，则这个元素是必需的。指定对 符合MATCH_COLUMNS条件和可选 UPDATE_CONDITION的行要更新的列。

UPDATE_CONDITION
可选。指定目标表中要被更新的行（在MERGE 情况下是要被插入的行）必须满足的一个布尔条件（类似于在 WHERE子句中声明的那样）。

MAPPING
可选。如果指定一个映射，它会覆盖默认的source-to-target列映射。 默认的source-to-target映射基于源COLUMNS 小节定义的列名与目标TABLE中列名之间的匹配。 映射可以被指定为:
target_column_name: source_column_name
或者
target_column_name: 'expression'
其中*expression*是在查询的SELECT 列表中指定的任意表达式，例如常量值、列引用、操作符调用、函数调用等等。

PRELOAD
可选。指定在装载操作之前运行的操作。目前唯一的预装载操作是TRUNCATE。

TRUNCATE
可选。如果设置为true，gpload将在装载目标表之前 移除其中所有的行。

REUSE_TABLES
可选。如果设置为true，gpload将不会删除它创建的外部表 对象和阶段性对象。这些对象将被重用于未来使用同一装载说明的装载操作。这会 提高小型装载的性能（正在进行的到同一目标表的小型装载）。
如果LOG_ERRORS: true被指定，REUSE_TABLES: true 必须被指定以保留Greenplum数据库错误日志中的格式错误。如果REUSE_TABLES: true 没有被指定，格式错误信息会在gpload操作之后被删除。
如果*external_table_name*存在，工具会使用存在的表。 如果OUTPUT表结构与数据库中表结构不吻合，工具会返回错误。

STAGING_TABLE
可选。指定gpload操作过程中创建的临时外部表的名称。 该外部表会被gpfdist使用。 REUSE_TABLES: true必须被指定。如果REUSE_TABLES为false或没有指定， STAGING_TABLE会被忽略。默认情况下， gpload会采用随机名称创建一个临时外部表。
如果*external_table_name*包含点号(.)， gpload会返回错误。如果表已经存在，工具会使用该表。 如果表的结构与OUTPUT表定义的结构不匹配，工具会返回错误。

工具会使用EXTERNAL部分定义的SCHEMA值作为 *external_table_name*。如果 SCHEMA的值为%， *external_table_name*的名字会和目标表相同， TABLE的表结构与OUTPUT 部分定义的相同。
如果没有设置SCHEMA，工具会在数据库中搜索表 (在search_path中定义的模式下)，如果找不到该表， *external_table_name*表会在PUBLIC 模式下被创建。

FAST_MATCH
可选项。如果设置为true，在重用外部表时，gpload仅 搜索匹配外部表对象的数据库。工具不会从pg_attribute 检查外部表的列名和列类型。设置该值可以在重用外部表的前提下提高， gpload性能。但是如果实际的列不匹配，该工具会在实际执行 时返回错误
默认值为false，工具会去预先检查外部表定义的列名和列类型。
REUSE_TABLES: true也必须被定义。如果 REUSE_TABLES为false或者没定义，并且 FAST_MATCH: true被指定， gpload会返回告警信息。

SQL
可选。定义在装载操作之前或者之后要运行的SQL命令。用户可以指定多个BEFORE 或者AFTER命令。按照想要的执行顺序列出命令。

BEFORE
可选。在装载操作开始之前要运行的一个SQL命令。将命令封闭在引号中。

AFTER
可选。在装载操作完成之后要运行的一个SQL命令。将命令封闭在引号中。

日志文件格式

gpload输出的日志文件具有下面的格式：

```
timestamp|level|message
```

其中*timestamp*的形式是：YYYY-MM-DD HH:MM:SS，*level*是DEBUG、 LOG、 INFO、 ERROR中间的一个， 而*message*是普通文本消息。

日志文件中可能让人感兴趣的一些INFO消息是（其中# 对应于实际的秒数、数据的单位或者失败的行）：

```
INFO|running time: #.## seconds
INFO|transferred #.# kB of #.# kB.
INFO|gpload succeeded
INFO|gpload succeeded with warnings
INFO|gpload failed
INFO|1 bad row
INFO|# bad rows
```

注解

如果用户的数据库对象名使用双引号标识符（定界的标识符）创建，用户必须在gpload 控制文件中用单引号指定定界的名称。例如，如果用户这样创建一个表：

```
CREATE TABLE "MyTable" ("MyColumn" text);
```

用户的YAML格式的gpload控制文件应该按如下方式引用上述表和列名：

```
- COLUMNS:
   - '"MyColumn"': text
OUTPUT:
   - TABLE: public.'"MyTable"'
```

如果YAML控制文件包含Greenplum数据库4.3.x的ERROR_TABLE元素， gpload会提示一个告警，显示表明ERROR_TABLE 是不支持的，如果LOG_ERRORS和REUSE_TABLE 被设置为true，加载错误信息会被正常处理。当运行在单行事务处理模式时， 格式错误的行会被记录到内部数据库日志。

示例

按my_load.yml中的定义运行一个装载作业：

```
gpload -f my_load.yml
```

装载控制文件的例子:

```
---
VERSION: 1.0.0.1
DATABASE: ops
USER: gpadmin
HOST: mdw-1
PORT: 5432
GPLOAD:
   INPUT:
    - SOURCE:
         LOCAL_HOSTNAME:
           - etl1-1
           - etl1-2
                    - etl1-3
           - etl1-4
         PORT: 8081
         FILE:
                     - /var/load/data/*
    - COLUMNS:
           - name: text
                     - amount: float4
                     - category: text
                     - descr: text
           - date: date
    - FORMAT: text
    - DELIMITER: '|'
       - ERROR_LIMIT: 25
    - LOG_ERRORS: true
   OUTPUT:
    - TABLE: payables.expenses
    - MODE: INSERT
   PRELOAD:
    - REUSE_TABLES: true
      SQL:
      - BEFORE: "INSERT INTO audit VALUES('start', current_timestamp)"
      - AFTER: "INSERT INTO audit VALUES('end', current_timestamp)"
```

另见

gpfdist, *Greenplum数据库参考指南*中的CREATE EXTERNAL TABLE。