在 HDFS 上创建 UDW 外部表

前言

UDW 支持 Greenplum 和 udpg 两种类型的数据仓库，两种数据类型的数据仓库都可以通过创建 hdfs 外部表来实现对 hdfs 的读写操作。

Greenplum 和 udpg 采用了不同的方法访问 hdfs。如果您使用的是 UDW Greenplum 请参考第一章 gphdfs:在 hdfs 上创建 UDW Greenplum 的外部表，如果您使用的是 UDW udpg 请参考第二章 udwhdfs：在 HDFS 上创建 UDW udpg 外部表。

第一章 gphdfs:在 hdfs 上创建 UDW Greenplum 的外部表

1.1gphdfs 的使用简介

Greenplum 通过 gpwhdfs 打通了 Greenplum 和 hdfs 之间的联系，可以通过在 Greenplum 上创建连接 hdfs 的外部表，我们就可以通过 Greenplum 直接对 hdfs 的数据访问，也可以通过创建可写的外部表，把 Greenplum 的数据或者分析结果写入 hdfs。

1.1.1 创建 gphdfs 可读外部表

```
CREATE [READABLE] EXTERNAL TABLE table_name
    ( column_name data_type [, ...] | LIKE other_table )
     LOCATION ('gphdfs://hdfs_host[:port]/path/file')
     FORMAT 'TEXT'
          [( [HEADER]
             [DELIMITER [AS] 'delimiter' | 'OFF']
             [NULL [AS] 'null string']
             [ESCAPE [AS] 'escape' | 'OFF']
             [NEWLINE [ AS ] 'LF' | 'CR' | 'CRLF']
             [FILL MISSING FIELDS] )]
         | 'CSV'
          [( [HEADER]
             [QUOTE [AS] 'quote']
             [DELIMITER [AS] 'delimiter']
```

```
            [NULL [AS] 'null string']
            [FORCE NOT NULL column [, ...]]
            [ESCAPE [AS] 'escape']
            [NEWLINE [ AS ] 'LF' | 'CR' | 'CRLF']
            [FILL MISSING FIELDS] )]
        | 'CUSTOM' (Formatter=<formatter specifications>)
    [ ENCODING 'encoding' ]
    [ [LOG ERRORS INTO error_table] SEGMENT REJECT LIMIT count
      [ROWS | PERCENT] ]
```

1.1.2 创建 gphdfs 可写外部表

```
CREATE WRITABLE EXTERNAL TABLE table_name
 ( column_name data_type [, ...] | LIKE other_table )
  LOCATION('gphdfs://hdfs_host[:port]/path')
   FORMAT 'TEXT'
           [( [DELIMITER [AS] 'delimiter']
           [NULL [AS] 'null string']
           [ESCAPE [AS] 'escape' | 'OFF'] )]
       | 'CSV'
           [([QUOTE [AS] 'quote']
           [DELIMITER [AS] 'delimiter']
           [NULL [AS] 'null string']
           [FORCE QUOTE column [, ...]] ]
           [ESCAPE [AS] 'escape'] )]
        | 'CUSTOM' (Formatter=<formatter specifications>)
  [ ENCODING 'write_encoding' ]
  [ DISTRIBUTED BY (column, [ ... ] ) | DISTRIBUTED RANDOMLY ]
```

1.2gphdfs 的使用举例

1.2.1 gphdfs 可读外部表

1.创建一个可读外部表（请把下面的 ip 替换成自己的 namenode ip）

```
create EXTERNAL table catalog_page
(
    cp_catalog_page_sk       integer                 ,
    cp_catalog_page_id       char(16)                ,
    cp_start_date_sk         integer                 ,
    cp_end_date_sk           integer                 ,
```

```
    cp_department             varchar(50)                ,
    cp_catalog_number          integer                   ,
    cp_catalog_page_number     integer                    ,
    cp_description            varchar(100)               ,
    cp_type                   varchar(100)
) LOCATION ('gphdfs://10.19.171.28:8020/udw_hdfs')
FORMAT 'csv' (DELIMITER '|');
```

2.hdfs 中/udw_hdfs 目录下的数据如下所示（支持 lzo 压缩）

```
[hadoop@uhadoop-valnw3-master1 data]$ hdfs dfs -ls /udw_hdfs
Found 2 items
-rw-r--r--   3 hadoop   postgres    4182385 2016-06-27 15:28 /udw_hdfs/catalog_page.dat
-rw-r--r--   3 postgres postgres    1781887 2016-06-23 14:57 /udw_hdfs/catalog_page.dat.lzo
```

3.通过 greenplum 的 sql 就可以访问 hdfs 上述目录的数据

```
udwhdfs=# select count(*) from catalog_page ;
 count
-------
 60000
(1 row)

udwhdfs=# select * from catalog_page limit 5;
 cp_catalog_page_sk | cp_catalog_page_id | cp_start_date_sk | cp_end_date_sk | cp_department | cp_catalog_number | cp_catalog_page_number |
                                cp_description                                 |  cp_type
-------------------+--------------------+------------------+----------------+---------------+-------------------+------------------------+----
----------------------------------------------------------------------------------------+-----------
                 1 | AAAAAAAABAAAAAAA   |          2450815 |        2450996 | DEPARTMENT    |                 1 |                      1 | In
general basic characters welcome. Clearly lively friends conv                          | bi-annual
                 2 | AAAAAAAACAAAAAAA   |          2450815 |        2450996 | DEPARTMENT    |                 1 |                      2 | Eng
lish areas will leave prisoners. Too public countries ought to become beneath the years.  | bi-annual
                 3 | AAAAAAAADAAAAAAA   |          2450815 |        2450996 | DEPARTMENT    |                 1 |                      3 | Tim
es could not address disabled indians. Effectively public ports c                      | bi-annual
                 4 | AAAAAAAAEAAAAAAA   |          2450815 |                |               |                 1 |                        |
                                                                                                                  | bi-annual
                 5 | AAAAAAAAFAAAAAAA   |          2450815 |        2450996 | DEPARTMENT    |                 1 |                      5 | Cla
ssic buildings ensure in a tests. Real years may not receive open systems. Now broad m     | bi-annual
(5 rows)
```

1.2.2 gphdfs 可写外部表

1.创建 hdfs 可写外部表（请把下面的 ip 替换成自己的 namenode ip）

```
create WRITABLE EXTERNAL table t_catalog_page
(
    cp_catalog_page_sk        integer                   ,
    cp_catalog_page_id        char(16)                   ,
    cp_start_date_sk          integer                  ,
    cp_end_date_sk            integer                  ,
    cp_department             varchar(50)                ,
    cp_catalog_number          integer                   ,
    cp_catalog_page_number     integer                    ,
    cp_description            varchar(100)               ,
    cp_type                   varchar(100)
```

```
) LOCATION ('gphdfs://10.19.171.28:8020/udw_write_hdfs')
FORMAT 'csv' (DELIMITER '|');
```

2.创建 hdfs 对应目录，并修改权限

```
hdfs dfs -mkdir -p /udw_write_hdfs
hdfs dfs -chown postgres:postgres /udw_write_hdfs
```

3.把 catalog_page 表格中的内容通过 greenplum 写入 hdfs

```
udwhdfs=# INSERT INTO t_catalog_page SELECT * FROM catalog_page;
INSERT 0 60000
```

4.在 hdfs 中查看写入结果

```
hdfs dfs -ls /udw_write_hdfs

091769 2016-06-27 16:06 /udw_write_hdfs/0_1466754774-0000000029
180187 2016-06-27 16:12 /udw_write_hdfs/0_1466754774-0000000053
090616 2016-06-27 16:06 /udw_write_hdfs/1_1466754774-0000000029
184583 2016-06-27 16:12 /udw_write_hdfs/1_1466754774-0000000053
```

备注：1. UDW 的 greenplum 目前支持 CDH5 相关的 Hadoop 版本

2.更多 greenplum 的 gphdfs 的使用请参考：

http://gpdb.docs.pivotal.io/4350/admin_guide/load/topics/g-gphdfs-protocol.html

第二章 udwhdfs：在 HDFS 上创建 UDW udpg 外部表

2.1 udwhdfs 的使用简介

udwhdfs 打通了 udw(udpg)和 hdfs 之间的联系，可以通过在 udw(udpg)上创建连接 hdfs 的外部表，我们就可以通过 udw(udpg)直接对 hdfs 的数据访问，也可以通过创建可写的外部表，把 udw(udpg)的数据或者分析结果写入 hdfs。

2.1.1 创建 udwhdfs 可读外部表

```
CREATE [READABLE] EXTERNAL TABLE table_name
    ( column_name data_type [, ...] | LIKE other_table )
     LOCATION ('udwhdfs://hdfs_host[:port]/path/file')
     FORMAT 'TEXT'
```

```
        [( [HEADER]
           [DELIMITER [AS] 'delimiter' | 'OFF']
           [NULL [AS] 'null string']
           [ESCAPE [AS] 'escape' | 'OFF']
           [NEWLINE [ AS ] 'LF' | 'CR' | 'CRLF']
           [FILL MISSING FIELDS] )]
      | 'CSV'
        [( [HEADER]
           [QUOTE [AS] 'quote']
           [DELIMITER [AS] 'delimiter']
           [NULL [AS] 'null string']
           [FORCE NOT NULL column [, ...]]
           [ESCAPE [AS] 'escape']
           [NEWLINE [ AS ] 'LF' | 'CR' | 'CRLF']
           [FILL MISSING FIELDS] )]
      | 'CUSTOM' (Formatter=<formatter specifications>)
  [ ENCODING 'encoding' ]
  [ [LOG ERRORS INTO error_table] SEGMENT REJECT LIMIT count
    [ROWS | PERCENT] ]
```

2.1.2 创建 udwhdfs 可写外部表

```
CREATE WRITABLE EXTERNAL TABLE table_name
 ( column_name data_type [, ...] | LIKE other_table )
 LOCATION('udwhdfs://hdfs_host[:port]/path')
  FORMAT 'TEXT'
         [( [DELIMITER [AS] 'delimiter']
         [NULL [AS] 'null string']
         [ESCAPE [AS] 'escape' | 'OFF'] )]
     | 'CSV'
         [([QUOTE [AS] 'quote']
         [DELIMITER [AS] 'delimiter']
         [NULL [AS] 'null string']
         [FORCE QUOTE column [, ...]] ]
         [ESCAPE [AS] 'escape'] )]
      | 'CUSTOM' (Formatter=<formatter specifications>)
[ ENCODING 'write_encoding' ]
```

```
    [ DISTRIBUTED BY (column, [ ... ] ) | DISTRIBUTED RANDOMLY ]
```

2.2 udwhdfs 的使用举例

2.2.1 udwhdfs 可读外部表

1.创建一个可读外部表（请把下面的 ip 替换成自己的 namenode ip）

```
create EXTERNAL table catalog_page
(
    cp_catalog_page_sk       integer                ,
    cp_catalog_page_id       char(16)               ,
    cp_start_date_sk         integer                ,
    cp_end_date_sk           integer                ,
    cp_department            varchar(50)            ,
    cp_catalog_number        integer                ,
    cp_catalog_page_number    integer               ,
    cp_description           varchar(100)           ,
    cp_type                  varchar(100)
) LOCATION ('udwhdfs://10.19.171.28:8020/udw_hdfs')
FORMAT 'csv' (DELIMITER '|');
```

2.hdfs 中/udw_hdfs 目录下的数据如下所示（支持 lzo 压缩）

```
[hadoop@uhadoop-valnw3-master1 data]$ hdfs dfs -ls /udw_hdfs
Found 2 items
-rw-r--r--   3 hadoop   postgres    4182385 2016-06-27 15:28 /udw_hdfs/catalog_page.dat
-rw-r--r--   3 postgres postgres    1781887 2016-06-23 14:57 /udw_hdfs/catalog_page.dat.lzo
```

3.通过 udw(udpg)的 sql 就可以访问 hdfs 上述目录的数据

```
udwhdfs=# select count(*) from catalog_page ;
 count
-------
 60000
(1 row)

udwhdfs=# select * from catalog_page limit 5;
 cp_catalog_page_sk | cp_catalog_page_id | cp_start_date_sk | cp_end_date_sk | cp_department | cp_catalog_number | cp_catalog_page_number |
                    cp_description                    |   cp_type
--------------------+--------------------+------------------+----------------+---------------+-------------------+------------------------+----
----------------------------------------------------------------------------------------------+-----------
                  1 | AAAAAAAABAAAAAAA   |          2450815 |        2450996 | DEPARTMENT    |                 1 |                      1 | In
general basic characters welcome. Clearly lively friends conv                                | bi-annual
                  2 | AAAAAAAACAAAAAAA   |          2450815 |        2450996 | DEPARTMENT    |                 1 |                      2 | Eng
lish areas will leave prisoners. Too public countries ought to become beneath the years.     | bi-annual
                  3 | AAAAAAAADAAAAAAA   |          2450815 |        2450996 | DEPARTMENT    |                 1 |                      3 | Tim
es could not address disabled indians. Effectively public ports c                            | bi-annual
                  4 | AAAAAAAAEAAAAAAA   |          2450815 |                |               |                 1 |                        |
                                                                                              | bi-annual
                  5 | AAAAAAAAFAAAAAAA   |          2450815 |        2450996 | DEPARTMENT    |                 1 |                      5 | Cla
ssic buildings ensure in a tests. Real years may not receive open systems. Now broad m       | bi-annual
(5 rows)
```

2.2.1 udwhdfs 可写外部表

1.创建 hdfs 可写外部表（请把下面的 ip 替换成自己的 namenode ip）

```
create WRITABLE EXTERNAL table t_catalog_page
(
    cp_catalog_page_sk       integer                    ,
    cp_catalog_page_id       char(16)                   ,
    cp_start_date_sk         integer                   ,
    cp_end_date_sk           integer                   ,
    cp_department           varchar(50)                ,
    cp_catalog_number        integer                   ,
    cp_catalog_page_number    integer                   ,
    cp_description           varchar(100)              ,
    cp_type                 varchar(100)
) LOCATION ('udwhdfs://10.19.171.28:8020/udw_write_hdfs')
FORMAT 'csv' (DELIMITER '|');
```

2.创建 hdfs 对应目录，并修改权限

```
hdfs dfs -mkdir -p /udw_write_hdfs
hdfs dfs -chown postgres:postgres /udw_write_hdfs
```

3.把 catalog_page 表格中的内容通过 udw(udpg)写入 hdfs

```
udwhdfs=# INSERT INTO t_catalog_page SELECT * FROM catalog_page;
INSERT 0 60000
```

4.在 hdfs 中查看写入结果

```
hdfs dfs -ls /udw_write_hdfs

091769 2016-06-27 16:06 /udw_write_hdfs/0_1466754774-0000000029
180187 2016-06-27 16:12 /udw_write_hdfs/0_1466754774-0000000053
090616 2016-06-27 16:06 /udw_write_hdfs/1_1466754774-0000000029
184583 2016-06-27 16:12 /udw_write_hdfs/1_1466754774-0000000053
```