

原创

Hive文件存储格式（TEXTFILE、ORC、PARQUET三者的存储格式的压缩对比和查询速度对比

2018-12-06 11:51:06

isea_you

阅读数 7052

☆ 收藏

更多

版权声明：本文为博主原创文章，遵循 [CC 4.0 BY-SA](#) 版权协议，转载请附上原文出处链接和本声明。
本文链接：https://blog.csdn.net/qq_31807385/article/details/84796880

综述：

Hlve的文件存储格式有四种：TEXTFILE、SEQUENCEFILE、ORC、PARQUET，前面两种是行式存储，后面两种是列式存储；所谓的存储格式就是在建表的时候指定的将表中的数据按照什么样子的存储方式，如果指定了A方式，那么在向表中插入数据的时候，将会使用该方式向HDFS中添加相应的数据类型。

如果为textfile的文件格式，直接load就OK，不需要走MapReduce；如果是其他的类型就需要走MapReduce了，因为其他的类型都涉及到了文件的MapReduce的压缩方式来实现。

总结：

比对三种主流的文件存储格式TEXTFILE、ORC、PARQUET

压缩比：ORC > Parquet > textFile（textfile没有进行压缩）

查询速度：三者几乎一致

案例证明：

```
1 | 1, textfile, 创建表，存储数据格式为TEXTFILE
2 | create table log_text (
3 | track_time string,
4 | url string,
5 | session_id string,
6 | referer string,
7 | ip string,
8 | end_user_id string,
9 | city_id string
10 | )
11 | row format delimited fields terminated by '\t'
12 | stored as textfile ;
```



6



2



举报

```
13
14 向表中加载数据
15
16 load data local inpath '/opt/module/datas/log.data' into table log_text ;
17 查看表中数据大小
18
19 这个过程不会走MapReduce，而是直接将文件上传到了HDFS，在HDFS上文件的名称还叫log.data
20 dfs -du -h /user/hive/warehouse/db_hive.db/log_text;
21 18.1 M  /user/hive/warehouse/db_hive.db/log_text/log.data
22
23 2,ORC,创建表，存储数据格式为ORC
24 create table log_orc(
25   track_time string,
26   url string,
27   session_id string,
28   referer string,
29   ip string,
30   end_user_id string,
31   city_id string
32 )
33 row format delimited fields terminated by '\t'
34 stored as orc ;
35
36 向表中加载数据
37 insert into table log_orc select * from log_text ;
38
39 查看表中数据大小
40 这个过程要走MapReduce，而且文件是按照列式存储的，还会对文件进行压缩，Orc默认使用的压缩方式是
41 zlib因此会更加节省空间，hadoop上是新的文件名，
42
43 hive (db_hive)> dfs -du -h /user/hive/warehouse/db_hive.db/log_orc;
44 2.8 M  /user/hive/warehouse/db_hive.db/log_orc/000000_0
45
46
47 3,Parquet,创建表，存储数据格式为parquet
48 create table log_parquet(
49   track_time string,
50   url string,
51   session_id string,
52   referer string,
53   ip string,
54   end_user_id string,
55   city_id string
```



6



2



举报

```
56 | )
57 | row format delimited fields terminated by '\t'
58 | stored as parquet ;
59 |
60 | 向表中加载数据
61 | insert into table log_parquet select * from log_text ;
62 |
63 | 查看表中数据大小这个过程要走MapReduce，而且文件是按照列式存储的，因此会更加节省空间，
64 | hadfs上是新的文件名，
65 | hive (db_hive)> dfs -du -h /user/hive/warehouse/db_hive.db/log_parquet;
66 | 13.1 M /user/hive/warehouse/db_hive.db/log_parquet/000000_0
67 |
68 | 存储文件的压缩比总结：
69 | ORC > Parquet > textFile
70 |
71 |
72 | select count(*) from log_text;
73 |
74 | select count(*) from log_orc;
75 |
76 | select count(*) from log_parquet;
77 |
78 | 存储文件的查询速度总结：查询速度相近。
```



6



2



压缩和存储的结合：

在建表的时候，如果我们指定了列式存储的方式，他会默认使用对于的压缩方式将我们的数据进行压缩，与此同时我们能够自己定制在文件存储的时候使用什么样子的压缩方式，例子如下：

```
1 | 1. 创建一个非压缩的的ORC存储方式
2 | create table log_orc_none(
3 | track_time string,
4 | url string,
5 | session_id string,
6 | referer string,
7 | ip string,
8 | end_user_id string,
9 | city_id string
10 | )
11 | row format delimited fields terminated by '\t'
12 | stored as orc tblproperties ("orc.compress"="NONE");
```



举报

```
13 插入数据 14 | hive (default)> insert into table log_orc_none select * from log_text ;
15
16 2. 创建一个SNAPPY压缩的ORC存储方式
17 create table log_orc_snappy(
18   track_time string,
19   url string,
20   session_id string,
21   referer string,
22   ip string,
23   end_user_id string,
24   city_id string
25 )
26 row format delimited fields terminated by '\t'
27 stored as orc tblproperties ("orc.compress"="SNAPPY");
28 插入数据
29 hive (default)> insert into table log_orc_snappy select * from log_text ;
30
31 3. 创建一个默认压缩的ORC存储方式
32 create table log_orc(
33   track_time string,
34   url string,
35   session_id string,
36   referer string,
37   ip string,
38   end_user_id string,
39   city_id string
40 )
41 row format delimited fields terminated by '\t'
42 stored as orc ;
43
44 向表中加载数据
45 insert into table log_orc select * from log_text ;
46
47 对比三者的压缩比:
48
49 hive (db_hive)> dfs -du -h /user/hive/warehouse/db_hive.db/log_orc_none;
50 18.1 M  /user/hive/warehouse/db_hive.db/log_orc_none/log.data
51
52 hive (db_hive)> dfs -du -h /user/hive/warehouse/db_hive.db/log_orc_snappy;
53 3.8 M  /user/hive/warehouse/db_hive.db/log_orc_snappy/000000_0
54
55 hive (db_hive)> dfs -du -h /user/hive/warehouse/db_hive.db/log_orc;
```



6



2





举报


```
56 | 2.8 M /user/hive/warehouse/db_hive.db/log_orc/000000_057 |
58 | 总结:
59 | 没有压缩的orc格式相当于textfile, 默认的压缩格式压缩比最大, snappy对数据进行了压缩
60 | orc存储文件默认采用ZLIB压缩, ZLIB采用的是deflate压缩算法。因此比snappy压缩的小。
61 | 文件没有压缩的话, HDFS上显示的是原来的文件名, 如果压缩的话, 使用类似于000000_0的文件名
```


总结:


- 比对三种主流的文件存储格式TEXTFILE、ORC、PARQUET
- 压缩比: ORC > Parquet > textFile (textfile没有进行压缩)
- 查询速度: 三者几乎一致
- HDFS上显示的是原来的文件名, 如果压缩的话, 使用类似于000000_0的文件名


6





2













赏



文章最后发布于: 2018-11-51:06

Hive的几种常见压缩格式 (ORC, Parquet, Sequencefile, RCfile, Avro) 的读写查询性能测试

阅读数 8302

一.测试背景工作中想把历史的APP日志结构化到Hive中进行查询, 由于数据较大, 需要进行压缩, 根据Hive官方提... 博文 来自: 人唯优的博客




想对作者说点什么

yyy雨 3小时前 #2楼

数据量太少, ORC查询速度要明显优于textFile



hellobeach 8个月前 #1楼

18.1 M 数据的话90%的时间都是JVM的启动时间。

3

hive常见的几种文件存储格式与压缩方式的结合-----Parquet格式+snappy压缩 以及ORC格式+snappy压...

阅读数 8001

一.使用Parquet存储数据数据使用列存储之前是普通的行存储, 下面是行存储的的文件大小, 这个HDFS上的数据使... 博文 来自: wyz0516071128的...



举报