Hive：简单查询不启用Mapreduce job而启用Fetch task

写在前面的话
，学Hive这么久了，发现目前国内还没有一本完整的介绍Hive的书籍，而且互联网上面的资料很乱，于是我决定写一些关于《Hive的那些事》
序列文章，分享给大家。我会在接下来的时间整理有关Hive的资料，如果对Hive的东西感兴趣，请关注本博客。https://www.iteblog.com/archives/tag/hive-technology/

如果你想查询某个表的某一列，Hive默认是会启用MapReduce Job来完成这个任务，如下：

```
hive> SELECT id, money FROM m limit 10;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Cannot run job locally: Input Size (= 235105473) is larger than
hive.exec.mode.local.auto.inputbytes.max (= 134217728)
Starting Job = job_1384246387966_0229, Tracking URL =
http://l-datalogm1.data.cn1:9981/proxy/application_1384246387966_0229/
Kill Command = /home/q/hadoop-2.2.0/bin/hadoop job
-kill job_1384246387966_0229
hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 0
2013-11-13 11:35:16,167 Stage-1 map = 0%,  reduce = 0%
2013-11-13 11:35:21,327 Stage-1 map = 100%,  reduce = 0%,
 Cumulative CPU 1.26 sec
2013-11-13 11:35:22,377 Stage-1 map = 100%,  reduce = 0%,
 Cumulative CPU 1.26 sec
MapReduce Total cumulative CPU time: 1 seconds 260 msec
Ended Job = job_1384246387966_0229
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 1.26 sec
HDFS Read: 8388865 HDFS Write: 60 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 260 msec
OK
1	122
1	185
1	231
1	292
1	316
1	329
1	355
1	356
1	362
```

```
1	364
Time taken: 16.802 seconds, Fetched: 10 row(s)
```

我们都知道，启用MapReduce Job是会消耗系统开销的。对于这个问题，从Hive0.10.0版本开始，对于简单的不需要聚合的类似SELECT <col> from <table> LIMIT n语句，不需要起MapReduce job，直接通过Fetch task获取数据，可以通过下面几种方法实现：

方法一：

```
hive> set hive.fetch.task.conversion=more;
hive> SELECT id, money FROM m limit 10;
OK
1	122
1	185
1	231
1	292
1	316
1	329
1	355
1	356
1	362
1	364
Time taken: 0.138 seconds, Fetched: 10 row(s)
```

上面 set hive.fetch.task.conversion=more;开启了Fetch任务，所以对于上述简单的列查询不在启用MapReduce job！

方法二：

```
bin/hive --hiveconf hive.fetch.task.conversion=more
```

方法三：

上面的两种方法都可以开启了Fetch任务，但是都是临时起作用的；如果你想一直启用这个功能，可以在${HIVE_HOME}/conf/hive-site.xml里面加入以下配置：

```
<property>
  <name>hive.fetch.task.conversion</name>
  <value>more</value>
  <description>
```

```
  Some select queries can be converted to single FETCH task
  minimizing latency.Currently the query should be single
  sourced not having any subquery and should not have
  any aggregations or distincts (which incurrs RS),
  lateral views and joins.
  1. minimal : SELECT STAR, FILTER on partition columns, LIMIT only
  2. more    : SELECT, FILTER, LIMIT only (+TABLESAMPLE, virtual columns)
 </description>
</property>
```

这样就可以长期启用Fetch任务了，很不错吧，也赶紧去试试吧！