Env: Hive on MapR

Goal: Sometimes users want to change the number of mapper for Hive jobs for the better performance.

Solution:
mapred.map.tasks doesn't work for hive jobs.
You need to change mapreduce.input.fileinputformat.split.maxsize(256000000 by default).

```
hive> set mapreduce.input.fileinputformat.split.maxsize;
mapreduce.input.fileinputformat.split.maxsize=256000000
hive> select count(*) from books;
Query ID = mapr_20161111115327_cdf2e675-3dd6-4972-812c-10f189bc9764
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1478774222583_0002, Tracking URL = http://beta1:8088/proxy/application_1478774222583_0002/ (http://beta1:8088/proxy/application_1478774222583_0002/)
Kill Command = /opt/mapr/hadoop/hadoop-2.7.0/bin/hadoop job  -kill job_1478774222583_0002
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 1
2016-11-11 11:53:34,473 Stage-1 map = 0%,  reduce = 0%
...
```

If you increase it, the number of mapper will decrease.

```
hive> set mapreduce.input.fileinputformat.split.maxsize=512000000;
hive> select count(*) from books;
Query ID = mapr_20161111115422_57dce440-f361-463e-a40d-482c216c0010
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1478774222583_0003, Tracking URL = http://beta1:8088/proxy/application_1478774222583_0003/ (http://beta1:8088/proxy/application_1478774222583_0003/)
Kill Command = /opt/mapr/hadoop/hadoop-2.7.0/bin/hadoop job  -kill job_1478774222583_0003
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 1
2016-11-11 11:54:28,936 Stage-1 map = 0%,  reduce = 0%
...
```

If you decrease it, the number of mapper will increase.

```
hive> set mapreduce.input.fileinputformat.split.maxsize=128000000;
hive> select count(*) from books;
Query ID = mapr_20161111115500_2145df17-d60a-4f2a-bb83-f674e5ef4abc
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1478774222583_0004, Tracking URL = http://beta1:8088/proxy/application_1478774222583_0004/ (http://beta1:8088/proxy/application_1478774222583_0004/)
Kill Command = /opt/mapr/hadoop/hadoop-2.7.0/bin/hadoop job  -kill job_1478774222583_0004
Hadoop job information for Stage-1: number of mappers: 9; number of reducers: 1
2016-11-11 11:55:07,533 Stage-1 map = 0%,  reduce = 0%
...
```

If you increase the max size, it's good also to change the mfs chunksize(268435456 by default) of the warehouse directory to the bigger size.
Because it affects the data locality.

hadoop mfs -setchunksize <chunk size> /user/hive/warehouse/<db name>.db/<table name>/

Example:
hadoop mfs -setchunksize 536870912 /user/hive/warehouse/db1.db/table1/

```
hive> load data inpath '/user/mapr/input' into table db1.table1;
hive> set mapreduce.input.fileinputformat.split.maxsize=512000000;
```

Note:
You have to change the chunksize before you load the data into the table.
If you didn't, the data will be split by the default chunk size.
If you change the chunksize of the parent directory('/user/hive/warehouse/db1.db' on the above example), new subdirectories inherits it.