BIGDATA INTEGRATION AND STORAGE – PROG8450
LAB 06 – Data Compression using Hive

Name: Geedhu KV
Student No: 8899510

Compression Codec	File Size(MB)	Compression Ration (%)
Gzip	45.5	29.45
Snappy	76.6	49.58
LZ4	80.1	51.84
Bzip2	37.1	24.01
Original Data	154.5	

Gzip

```
hive> CREATE TABLE stocks_gzip AS
    > SELECT * FROM stocks;
Query ID = cloudera_20231014073232_85b4dc9c-fb5c-427b-a2dd-d7facf3f8d45
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1697290899419_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1697290899419_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1697290899419_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-10-14 07:32:34,845 Stage-1 map = 0%,  reduce = 0%
2023-10-14 07:33:15,274 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 27.34 sec
MapReduce Total cumulative CPU time: 27 seconds 340 msec
Ended Job = job_1697290899419_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/.hive-staging_hive_2023-10-14_07-32-07_332_5905082902776074893-1/-ext-10001
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/stocks_gzip
Table stocks_db.stocks_gzip stats: [numFiles=1, numRows=3006503, totalSize=47691102, rawDataSize=153021182]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 27.34 sec   HDFS Read: 162049514 HDFS Write: 47691191 SUCCESS
Total MapReduce CPU Time Spent: 27 seconds 340 msec
OK
Time taken: 72.871 seconds
hive> SHOW TABLES;
OK
stocks
stocks_gzip
Time taken: 0.024 seconds, Fetched: 2 row(s)
```

```
hive> DESCRIBE FORMATTED stocks_gzip;
OK
# col_name              data_type               comment

ymd                     date
symbol                  string
price_open              float
price_high              float
price_low               float
price_close             float
price_adj_close         float
volume                  int

# Detailed Table Information
Database:               stocks_db
Owner:                  cloudera
CreateTime:             Sat Oct 14 07:33:19 PDT 2023
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/stocks_gzip
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   true
        numFiles                1
        numRows                 3006503
        rawDataSize             153021182
        totalSize               47691102
        transient_lastDdlTime   1697293999

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.311 seconds, Fetched: 38 row(s)
```

```
hive> !hadoop fs -ls /user/hive/warehouse/stocks_db.db/stocks_gzip;
Found 1 items
-rwxrwxrwx   1 cloudera supergroup   47691102 2023-10-14 07:33 /user/hive/warehouse/stocks_db.db/stocks_gzip/000000_0.gz
hive> select * from stocks_gzip
    > LIMIT 10;
OK
2005-09-27      AAL     21.05   21.4    19.1    19.3    18.19   961200
2005-09-28      AAL     19.3    20.53   19.2    20.5    19.33   5747900
2005-09-29      AAL     20.4    20.58   20.1    20.21   19.05   1078200
2005-09-30      AAL     20.26   21.05   20.18   21.01   19.81   3123300
2005-10-03      AAL     20.9    21.75   20.9    21.5    20.27   1057900
2005-10-04      AAL     21.44   22.5    21.44   22.16   20.89   1768800
2005-10-05      AAL     22.1    22.31   21.75   22.2    20.93   904300
2005-10-06      AAL     22.6    23.0    22.4    22.58   21.29   1614200
2005-10-07      AAL     22.25   22.6    21.8    22.15   20.88   753400
2005-10-10      AAL     22.28   22.29   22.1    22.21   20.94   538700
Time taken: 0.137 seconds, Fetched: 10 row(s)
```

```
hive> !hadoop fs -ls -h /user/hive/warehouse/stocks_db.db/stocks_gzip;
Found 1 items
-rwxrwxrwx   1 cloudera supergroup     45.5 M 2023-10-14 07:33 /user/hive/warehouse/stocks_db.db/stocks_gzip/000000_0.gz
hive>
```

Snappy

```
hive> SET mapred.output.compression.codec= org.apache.hadoop.io.compress.SnappyCodec;
```

```
hive> CREATE TABLE stocks_snappy_crct AS select * from stocks;
Query ID = cloudera_20231014074848_de560e7b-d4f3-46c4-8472-d107e7e2829b
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1697290899419_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1697290899419_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1697290899419_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-10-14 07:48:49,578 Stage-1 map = 0%,  reduce = 0%
2023-10-14 07:49:12,555 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 16.31 sec
MapReduce Total cumulative CPU time: 16 seconds 310 msec
Ended Job = job_1697290899419_0003
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/.hive-staging_hive_2023-10-14_07-48-40_952_7020487823066724486-1/-ext-10001
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/stocks_snappy_crct
Table stocks_db.stocks_snappy_crct stats: [numFiles=1, numRows=3006503, totalSize=80348599, rawDataSize=153021182]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 16.31 sec   HDFS Read: 162049575 HDFS Write: 80348695 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 310 msec
OK
Time taken: 33.615 seconds
hive> !hadoop fs -ls /user/hive/warehouse/stocks_db.db/stocks_snappy_crct;
Found 1 items
-rwxrwxrwx   1 cloudera supergroup   80348599 2023-10-14 07:49 /user/hive/warehouse/stocks_db.db/stocks_snappy_crct/000000_0.snappy
hive> select * from stocks_snappy_crct limit 10;
OK
2005-09-27      AAL     21.05   21.4    19.1    19.3    18.19   961200
2005-09-28      AAL     19.3    20.53   19.2    20.5    19.33   5747900
2005-09-29      AAL     20.4    20.58   20.1    20.21   19.05   1078200
2005-09-30      AAL     20.26   21.05   20.18   21.01   19.81   3123300
2005-10-03      AAL     20.9    21.75   20.9    21.5    20.27   1057900
2005-10-04      AAL     21.44   22.5    21.44   22.16   20.89   1768800
2005-10-05      AAL     22.1    22.31   21.75   22.2    20.93   904300
2005-10-06      AAL     22.6    23.0    22.4    22.58   21.29   1614200
2005-10-07      AAL     22.25   22.6    21.8    22.15   20.88   753400
2005-10-10      AAL     22.28   22.29   22.1    22.21   20.94   538700
Time taken: 0.099 seconds, Fetched: 10 row(s)
```

```
hive> !hadoop fs -ls -h /user/hive/warehouse/stocks_db.db/stocks_snappy_crct;
Found 1 items
-rwxrwxrwx   1 cloudera supergroup     76.6 M 2023-10-14 07:49 /user/hive/warehouse/stocks_db.db/stocks_snappy_crct/000000_0.snappy
```

LZ4

```
hive> SET mapred.output.compression.codec= org.apache.hadoop.io.compress.Lz4Codec;
```

```
hive> CREATE TABLE stocks_lz4 AS select * from stocks;
Query ID = cloudera_20231014084545_43e3f7ea-7226-4955-b0c7-43e0942ad4d1
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1697290899419_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1697290899419_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1697290899419_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-10-14 08:45:14,961 Stage-1 map = 0%,  reduce = 0%
2023-10-14 08:45:43,622 Stage-1 map = 83%,  reduce = 0%, Cumulative CPU 15.55 sec
2023-10-14 08:45:44,678 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 16.53 sec
MapReduce Total cumulative CPU time: 16 seconds 530 msec
Ended Job = job_1697290899419_0005
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/.hive-staging_hive_2023-10-14_08-45-04_086_7303518760333487370-1/-ext-10001
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/stocks_lz4
Table stocks_db.stocks_lz4 stats: [numFiles=1, numRows=3006503, totalSize=83965822, rawDataSize=153021182]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 16.53 sec   HDFS Read: 162049567 HDFS Write: 83965910 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 530 msec
OK
Time taken: 43.332 seconds
```

```
hive> !hadoop fs -ls /user/hive/warehouse/stocks_db.db/stocks_lz4;
Found 1 items
-rwxrwxrwx   1 cloudera supergroup   83965822 2023-10-14 08:45 /user/hive/warehouse/stocks_db.db/stocks_lz4/000000_0.lz4
hive>
```

```
hive> !hadoop fs -ls -h /user/hive/warehouse/stocks_db.db/stocks_lz4;
Found 1 items
-rwxrwxrwx   1 cloudera supergroup     80.1 M 2023-10-14 08:45 /user/hive/warehouse/stocks_db.db/stocks_lz4/000000_0.lz4
```

BZIP2

```
hive> SET mapred.output.compression.codec= org.apache.hadoop.io.compress.BZip2Codec;
hive> CREATE TABLE stocks_bzip2 AS select * from stocks;
Query ID = cloudera_20231014084848_52137053-a180-42a7-8b1b-1552c01d570f
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1697290899419_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1697290899419_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1697290899419_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-10-14 08:49:01,957 Stage-1 map = 0%,  reduce = 0%
2023-10-14 08:49:25,436 Stage-1 map = 83%,  reduce = 0%, Cumulative CPU 21.11 sec
2023-10-14 08:49:28,403 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.06 sec
MapReduce Total cumulative CPU time: 24 seconds 60 msec
Ended Job = job_1697290899419_0006
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/.hive-staging_hive_2023-10-14_08-48-55_445_3413062807276177971-1/-ext-10001
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/stocks_db.db/stocks_bzip2
Table stocks_db.stocks_bzip2 stats: [numFiles=1, numRows=3006503, totalSize=38884471, rawDataSize=153021182]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 24.06 sec   HDFS Read: 162049569 HDFS Write: 38884561 SUCCESS
Total MapReduce CPU Time Spent: 24 seconds 60 msec
OK
Time taken: 37.152 seconds
hive> !hadoop fs -ls /user/hive/warehouse/stocks_db.db/stocks_bzip2;
Found 1 items
-rwxrwxrwx   1 cloudera supergroup   38884471 2023-10-14 08:49 /user/hive/warehouse/stocks_db.db/stocks_bzip2/000000_0.bz2
```

```
hive> !hadoop fs -ls -h /user/hive/warehouse/stocks_db.db/stocks_bzip2;
Found 1 items
-rwxrwxrwx   1 cloudera supergroup     37.1 M 2023-10-14 08:49 /user/hive/warehouse/stocks_db.db/stocks_bzip2/000000_0.bz2
```

Original Data

```
hive> !hadoop fs -ls -h /BigData/stocks.csv;
-rw-r--r--   1 cloudera supergroup    154.5 M 2023-09-16 08:17 /BigData/stocks.csv
hive>
```