BIGDATA HADOOP AND SPARK DEVELOPER

STOCK EXCHANGE DATA ANALYSIS PROJECT - SNAPSHOTS

RAVEENA PRABAKARAN

TIGER ANALYTICS

1) Create a data pipeline using sqoop to pull the data from the table below from MYSQL server into Hive.

   -> stockcompanies

-> stockprices

2) Create a new hive table with the following fields by joining the above two hive tables. Please use appropriate Hive built-in functions for columns (a,b,e and h to l).

- Trading_year: Should contain YYYY for each record
- Trading_month: Should contain MM or MMM for each record
- Symbol: Ticker code
- CompanyName: Legal name of the listed company
- State: State to be extracted from headquarters value.
- Sector: Business vertical of the listed company
- Sub_Industry: Business domain of the listed company within a sector
- Open: Average of intra-day opening price by month and year for each listed company
- Close: Average of intra-day closing price by month and year for each listed company
- Low: Average of intra-day lowest price by month and year for each listed company
- High: Average of intra-day highest price by month and year for each listed company
- Volume: Average of number of shares traded by month and year for each listed company

-> stock_data5

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 20.81 sec   HDFS Read: 52169279 HDFS Write: 2790018 HDFS EC Read: 0 SUCCESS
Stage-Stage-5: Map: 1   Cumulative CPU: 4.67 sec   HDFS Read: 2797801 HDFS Write: 4489513 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 25 seconds 480 msec
OK
Time taken: 45.884 seconds
hive> select * from stock_data5 limit 5;
OK
2010    01      A       Agilent Technologies Inc      California      Health Care      Health Care Equipment   21.72   21.61   21.4    21.86   4208442.11
2011    01      A       Agilent Technologies Inc      California      Health Care      Health Care Equipment   30.29   30.29   29.96   30.65   4496845.0
2012    01      A       Agilent Technologies Inc      California      Health Care      Health Care Equipment   28.54   28.78   28.22   29.08   5069975.0
2013    01      A       Agilent Technologies Inc      California      Health Care      Health Care Equipment   31.2    31.26   30.97   31.45   4567819.05
2014    01      A       Agilent Technologies Inc      California      Health Care      Health Care Equipment   42.01   42.04   41.66   42.36   3494200.0
Time taken: 0.064 seconds, Fetched: 5 row(s)
hive>
```

DATA ANALYSIS USING HIVE

3) Find the top five companies that are good for investment

```
2022-03-23 10:14:35,415 Stage-3 map = 0%,  reduce = 0%
2022-03-23 10:14:41,518 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.82 sec
2022-03-23 10:14:47,613 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.4 sec
MapReduce Total cumulative CPU time: 3 seconds 400 msec
Ended Job = job_1640258093152_29384
Launching Job 6 out of 7
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>
22/03/23 10:14:48 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm81
Starting Job = job_1640258093152_29385, Tracking URL = http://ip-10-0-21-22.ec2.internal:8088/proxy/application_1640258093152_29385/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1640258093152_29385
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2022-03-23 10:14:54,686 Stage-4 map = 0%,  reduce = 0%
2022-03-23 10:15:00,805 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 1.77 sec
2022-03-23 10:15:05,903 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 3.6 sec
MapReduce Total cumulative CPU time: 3 seconds 600 msec
Ended Job = job_1640258093152_29385
MapReduce Jobs Launched:
Stage-Stage-11: Map: 1   Cumulative CPU: 4.18 sec   HDFS Read: 4497123 HDFS Write: 20808 HDFS EC Read: 0 SUCCESS
Stage-Stage-12: Map: 1   Cumulative CPU: 3.71 sec   HDFS Read: 4497118 HDFS Write: 21370 HDFS EC Read: 0 SUCCESS
Stage-Stage-9: Map: 1   Cumulative CPU: 3.1 sec   HDFS Read: 26730 HDFS Write: 20808 HDFS EC Read: 0 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 3.4 sec   HDFS Read: 25661 HDFS Write: 323 HDFS EC Read: 0 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 3.6 sec   HDFS Read: 6073 HDFS Write: 342 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 990 msec
OK
Netflix Inc.     1536.0158311345647
Regeneron        1382.2714681440443
Ulta Salon Cosmetics & Fragrance Inc    1174.9378418697165
United Rentals; Inc.    1064.340239912759
Alaska Air Group Inc     878.5555555555554
Time taken: 106.643 seconds, Fetched: 5 row(s)
```

4) Show the best-growing industry by each state, having at least two or more industries mapped.

-> companyanalysis1

```
number of reduce tasks is set to 0 since there's no reduce operator
22/03/23 10:56:01 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm81
Starting Job = job_1640258093152_29471, Tracking URL = http://ip-10-0-21-22.ec2.internal:8088/proxy/application_1640258093152_29471/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1640258093152_29471
Hadoop job information for Stage-10: number of mappers: 1; number of reducers: 0
2022-03-23 10:56:07,143 Stage-10 map = 0%,  reduce = 0%
2022-03-23 10:56:13,252 Stage-10 map = 100%,  reduce = 0%, Cumulative CPU 3.16 sec
MapReduce Total cumulative CPU time: 3 seconds 160 msec
Ended Job = job_1640258093152_29471
Moving data to directory hdfs://nameservice1/user/hive/warehouse/bdhsproject.db/companyanalysis1
MapReduce Jobs Launched:
Stage-Stage-7: Map: 1  Reduce: 1   Cumulative CPU: 6.36 sec   HDFS Read: 4498484 HDFS Write: 33238 HDFS EC Read: 0 SUCCESS
Stage-Stage-12: Map: 1   Cumulative CPU: 3.32 sec   HDFS Read: 4497242 HDFS Write: 20808 HDFS EC Read: 0 SUCCESS
Stage-Stage-13: Map: 1   Cumulative CPU: 3.47 sec   HDFS Read: 4497235 HDFS Write: 21370 HDFS EC Read: 0 SUCCESS
Stage-Stage-10: Map: 1   Cumulative CPU: 3.16 sec   HDFS Read: 48444 HDFS Write: 30609 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 310 msec
OK
Time taken: 83.792 seconds
hive> select * from companyanalysis1 limit 5;
OK
Minnesota       Industrial Conglomerates        3M Company      112.60817307692305
Georgia Life & Health Insurance AFLAC Inc       38.92215568862274
Pennsylvania    Electrical Components & Equipment       AMETEK Inc      192.15081886792448
Texas   Integrated Telecommunications Services  AT&T Inc        55.250282273248494
Illinois        Pharmaceuticals AbbVie  73.49295774647888
```

-> industry_growth

```
hive> select * from industry_growth limit 50;
OK
California      Apparel Retail  261.8019767671436
California      Application Software     120.31201428397625
California      Biotechnology   189.855597859821
California      Health Care Equipment   147.29321666303753
California      Home Entertainment Software      298.3868203665769
California      Internet Software & Services     336.56041800779366
California      MultiUtilities  62.895455776620366
California      Networking Equipment    16.60504085306744
California      REITs   131.08399426163498
California      Retail REITs    107.50814027970536
California      Semiconductor Equipment 147.9653267908011
California      Semiconductors  282.4593566657631
Connecticut     Industrial Conglomerates         75.02785859563353
Florida Hotels; Resorts & Cruise Lines  137.37933900437363
Florida Industrial Conglomerates        177.39424996713996
Florida Internet Software & Services     163.10267039273256
Illinois        Construction & Farm Machinery & Heavy Trucks    73.11111617775501
Illinois        Health Care Equipment   41.94627917684228
Illinois        Industrial Machinery    136.03903698890892
Illinois        Packaged Foods & Meats  55.38509384373717
Illinois        REITs   65.02753772568181
Ireland Industrial Conglomerates        136.6124307695158
Ireland Pharmaceuticals 245.14488964557233
Maryland        Cable & Satellite       85.87024737549106
Massachusetts   Biotechnology   260.2045747013016
Massachusetts   Health Care Equipment   169.22809835926253
Massachusetts   REITs   68.24052887161051
Massachusetts   Semiconductors  285.17414566777273
Michigan        MultiUtilities  143.76404421430016
Minnesota       Packaged Foods & Meats  166.69719818630523
Missouri        Industrial Conglomerates        85.58764014977949
New Jersey      Health Care Equipment   120.88893239332754
New Jersey      Research & Consulting Services  115.46492537344713
```

-> best-growing industry by each state

```
22/03/23 11:30:05 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm81
Starting Job = job_1640258093152_29554, Tracking URL = http://ip-10-0-21-22.ec2.internal:8088/proxy/application_1640258093152_29554/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1640258093152_29554
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2022-03-23 11:30:12,035 Stage-4 map = 0%,  reduce = 0%
2022-03-23 11:30:19,213 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 3.65 sec
MapReduce Total cumulative CPU time: 3 seconds 650 msec
Ended Job = job_1640258093152_29554
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.43 sec   HDFS Read: 10966 HDFS Write: 833 HDFS EC Read: 0 SUCCESS
Stage-Stage-4: Map: 1   Cumulative CPU: 3.65 sec   HDFS Read: 7158 HDFS Write: 1409 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 80 msec
OK
California      Internet Software & Services     336.56041800779366
Connecticut     Industrial Conglomerates         75.02785859563353
Florida Industrial Conglomerates        177.39424996713996
Illinois        Industrial Machinery    136.03903698890892
Ireland Pharmaceuticals 245.14488964557233
Maryland        Cable & Satellite       85.87024737549106
Massachusetts   Semiconductors  285.17414566777273
Michigan        MultiUtilities  143.76404421430016
Minnesota       Packaged Foods & Meats  166.69719818630523
Missouri        Industrial Conglomerates        85.58764014977949
New Jersey      Health Care Equipment   120.88893239332754
New York        Diversified Financial Services  244.27936670090116
North Carolina  Apparel; Accessories & Luxury Goods     234.5365478318533
Ohio    Banks   171.58635235361513
Oklahoma        Oil & Gas Exploration & Production      129.87772667777514
Pennsylvania    Diversified Chemicals   163.20832113323664
Texas   Airlines        571.0256197386555
United Kingdom  Insurance Brokers       101.0504280362911
Virginia        Aerospace & Defense     255.37107674669656
Washington      Internet & Direct Marketing Retail      323.86488316718413
Wisconsin       Electric Utilities      131.94734372014784
Time taken: 43.652 seconds, Fetched: 21 row(s)
hive>
```

5) For each sector find the following.

-> sectorwisegrowth

```
2022-03-23 11:45:20,195 Stage-5 map = 0%,  reduce = 0%
2022-03-23 11:45:26,312 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 3.15 sec
MapReduce Total cumulative CPU time: 3 seconds 150 msec
Ended Job = job_1640258093152_29595
Moving data to directory hdfs://nameservice1/user/hive/warehouse/bdhsproject.db/sectorwisegrowth
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 7.62 sec   HDFS Read: 4500562 HDFS Write: 3561 HDFS EC Read: 0 SUCCESS
Stage-Stage-4: Map: 1 Reduce: 1   Cumulative CPU: 6.84 sec   HDFS Read: 4500570 HDFS Write: 3561 HDFS EC Read: 0 SUCCESS
Stage-Stage-5: Map: 1   Cumulative CPU: 3.15 sec   HDFS Read: 9536 HDFS Write: 3047 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 610 msec
OK
Time taken: 61.978 seconds
hive> select * from sectorwisegrowth limit 20;
OK
Consumer Discretionary 2010    14.404047387211953
Consumer Discretionary 2011    4.860126582278504
Consumer Discretionary 2012    7.730506329113922
Consumer Discretionary 2013    24.392572590013653
Consumer Discretionary 2014    9.603809523809531
Consumer Discretionary 2015    6.202380952380949
Consumer Discretionary 2016    14.681904761904747
Consumer Staples       2010    4.182352941176468
Consumer Staples       2011    4.158529411764697
Consumer Staples       2012    4.577941176470589
Consumer Staples       2013    8.819092436974778
Consumer Staples       2014    9.322857142857139
Consumer Staples       2015    4.892626984126977
Consumer Staples       2016    3.183055555555569
Energy 2010    7.710909090909105
Energy 2011    -1.5948484848484891
Energy 2012    -1.9490909090909128
Energy 2013    12.5844444444444436
Energy 2014    -6.264444444444443
Energy 2015    -10.099444444444443
Time taken: 0.055 seconds, Fetched: 20 row(s)
hive>
```

- Worst year

```
    at org.apache.logging.log4j.core.appender.AbstractManager.getManager(AbstractManager.java:114)
    at org.apache.logging.log4j.core.appender.OutputStreamManager.getManager(OutputStreamManager.java:114)
    at org.apache.logging.log4j.core.appender.RandomAccessFileManager.getFileManager(RandomAccessFileManager.java:74)
    at org.apache.hadoop.hive.ql.log.HushableRandomAccessFileAppender.createAppender(HushableRandomAccessFileAppender.java:178)
    ... 35 more

Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
22/03/23 11:54:50 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm81
Starting Job = job_1640258093152_29629, Tracking URL = http://ip-10-0-21-22.ec2.internal:8088/proxy/application_1640258093152_29629/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1640258093152_29629
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2022-03-23 11:54:56,172 Stage-4 map = 0%,  reduce = 0%
2022-03-23 11:55:02,272 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 2.88 sec
MapReduce Total cumulative CPU time: 2 seconds 880 msec
Ended Job = job_1640258093152_29629
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.97 sec   HDFS Read: 10935 HDFS Write: 536 HDFS EC Read: 0 SUCCESS
Stage-Stage-4: Map: 1   Cumulative CPU: 2.88 sec   HDFS Read: 6839 HDFS Write: 646 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 850 msec
OK
Consumer Discretionary  2011    4.8601265822278504
Consumer Staples        2016    3.1830555555555569
Energy  2015    -10.89944444444443
Financials      2011    -6.860689655517235
Health Care     2016    2.0805084745762485
Industrials     2015    -2.640000000000029
Information Technology  2011    -2.9025396825396896
Materials       2011    -3.967083333333335
Real Estate     2013    -4.463448275862078
Telecommunications Services     2015    -2.29399999999999
Utilities       2015    -6.473928571428566
Time taken: 42.649 seconds, Fetched: 11 row(s)
hive>
```

- b. Best year

```
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:313)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:227)

Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
22/03/23 12:00:26 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm81
Starting Job = job_1640258093152_29645, Tracking URL = http://ip-10-0-21-22.ec2.internal:8088/proxy/application_1640258093152_29645/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1640258093152_29645
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2022-03-23 12:00:33,731 Stage-4 map = 0%,  reduce = 0%
2022-03-23 12:00:39,820 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 3.12 sec
MapReduce Total cumulative CPU time: 3 seconds 120 msec
Ended Job = job_1640258093152_29645
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.33 sec   HDFS Read: 11059 HDFS Write: 536 HDFS EC Read: 0 SUCCESS
Stage-Stage-4: Map: 1   Cumulative CPU: 3.12 sec   HDFS Read: 6943 HDFS Write: 646 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 450 msec
OK
Consumer Discretionary  2013    24.392572590011653
Consumer Staples        2014    9.328571428571439
Energy  2016    18.642222222222223
Financials      2016    16.326612903225815
Health Care     2014    24.325254237288092
Industrials     2016    19.804925404944584
Information Technology  2013    15.538281249999997
Materials       2016    19.036400000000015
Real Estate     2014    18.892413793103472
Telecommunications Services     2014    5.05999999999995
Utilities       2014    10.268571428571427
Time taken: 55.214 seconds, Fetched: 11 row(s)
hive>
```

- c. Stable year

```
    at org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask.executeInProcess(MapredLocalTask.java:393)
    at org.apache.hadoop.hive.ql.exec.mr.ExecDriver.main(ExecDriver.java:779)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:313)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:227)
Caused by: java.lang.IllegalStateException: ManagerFactory [org.apache.logging.log4j.core.appender.RandomAccessFileManager$RandomAccessFileManagerFactory@611f8128]
  unable to create manager for [/var/log/hive/operation_logs/d9f68a7d-7b28-43a6-a953-ef42f03a630a/raveenaprabakartigeranaly_20220323120539_ec6ecdd5-8d66-46e6-8e41-c
5cdc94a402d] with data [org.apache.logging.log4j.core.appender.RandomAccessFileManager$FactoryData@79c01d37]
    at org.apache.logging.log4j.core.appender.AbstractManager.getManager(AbstractManager.java:114)
    at org.apache.logging.log4j.core.appender.OutputStreamManager.getManager(OutputStreamManager.java:114)
    at org.apache.logging.log4j.core.appender.RandomAccessFileManager.getFileManager(RandomAccessFileManager.java:74)
    at org.apache.hadoop.hive.ql.log.HushableRandomAccessFileAppender.createAppender(HushableRandomAccessFileAppender.java:178)
    ... 42 more

Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
22/03/23 12:06:08 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm81
Starting Job = job_1640258093152_29649, Tracking URL = http://ip-10-0-21-22.ec2.internal:8088/proxy/application_1640258093152_29649/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1640258093152_29649
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2022-03-23 12:06:15,321 Stage-4 map = 0%,  reduce = 0%
2022-03-23 12:06:22,442 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 3.1 sec
MapReduce Total cumulative CPU time: 3 seconds 100 msec
Ended Job = job_1640258093152_29649
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.24 sec   HDFS Read: 11806 HDFS Write: 536 HDFS EC Read: 0 SUCCESS
Stage-Stage-4: Map: 1   Cumulative CPU: 3.1 sec   HDFS Read: 6868 HDFS Write: 87 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 340 msec
OK
Time taken: 43.742 seconds
hive>
```