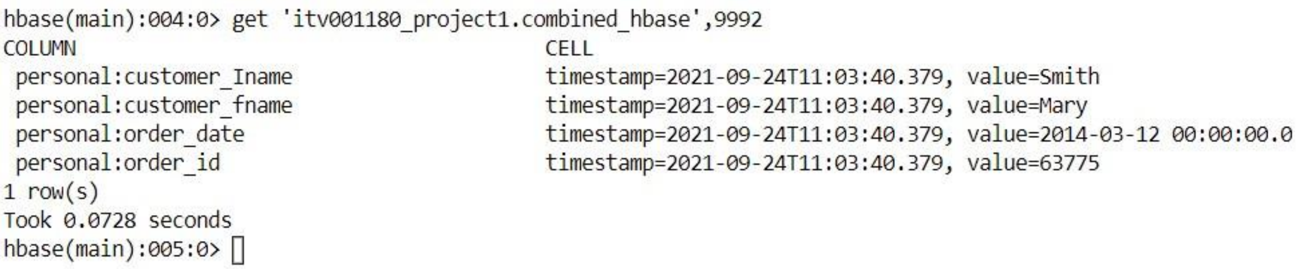
Bigdata Training – Project

Kaustubh Gharat

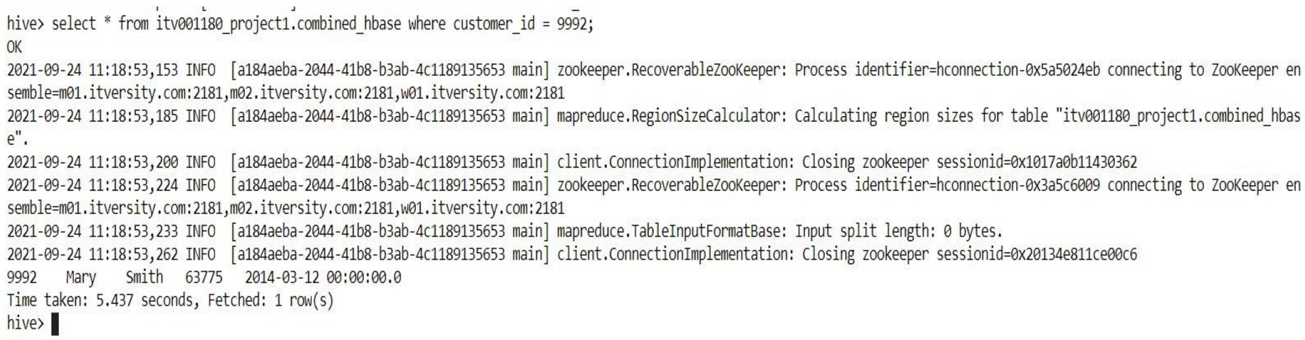
Project 1: Customer 360

Combined HBase Table:



```
hbase(main):004:0> get 'itv001180_project1.combined_hbase',9992
COLUMN                                      CELL
 personal:customer_Iname                    timestamp=2021-09-24T11:03:40.379, value=Smith
 personal:customer_fname                    timestamp=2021-09-24T11:03:40.379, value=Mary
 personal:order_date                        timestamp=2021-09-24T11:03:40.379, value=2014-03-12 00:00:00.0
 personal:order_id                          timestamp=2021-09-24T11:03:40.379, value=63775
1 row(s)
Took 0.0728 seconds
hbase(main):005:0> 
```

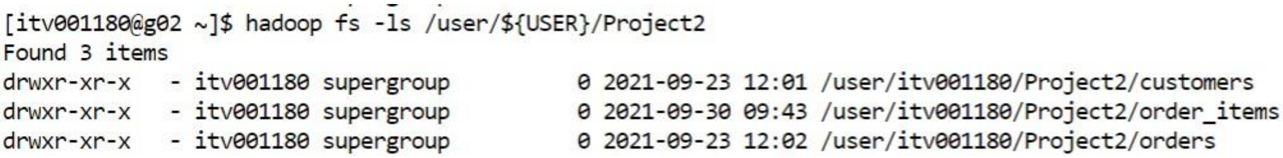
Combined Hive Table:



```
hive> select * from itv001180_project1.combined_hbase where customer_id = 9992;
OK
2021-09-24 11:18:53,153 INFO  [a184aeba-2044-41b8-b3ab-4c1189135653 main] zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x5a5024eb connecting to ZooKeeper en
semble=m01.itversity.com:2181,m02.itversity.com:2181,w01.itversity.com:2181
2021-09-24 11:18:53,185 INFO  [a184aeba-2044-41b8-b3ab-4c1189135653 main] mapreduce.RegionSizeCalculator: Calculating region sizes for table "itv001180_project1.combined_hbas
e".
2021-09-24 11:18:53,200 INFO  [a184aeba-2044-41b8-b3ab-4c1189135653 main] client.ConnectionImplementation: Closing zookeeper sessionid=0x1017a0b11430362
2021-09-24 11:18:53,224 INFO  [a184aeba-2044-41b8-b3ab-4c1189135653 main] zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x3a5c6009 connecting to ZooKeeper en
semble=m01.itversity.com:2181,m02.itversity.com:2181,w01.itversity.com:2181
2021-09-24 11:18:53,233 INFO  [a184aeba-2044-41b8-b3ab-4c1189135653 main] mapreduce.TableInputFormatBase: Input split length: 0 bytes.
2021-09-24 11:18:53,262 INFO  [a184aeba-2044-41b8-b3ab-4c1189135653 main] client.ConnectionImplementation: Closing zookeeper sessionid=0x20134e811ce00c6
9992    Mary    Smith   63775   2014-03-12 00:00:00.0
Time taken: 5.437 seconds, Fetched: 1 row(s)
hive> 
```

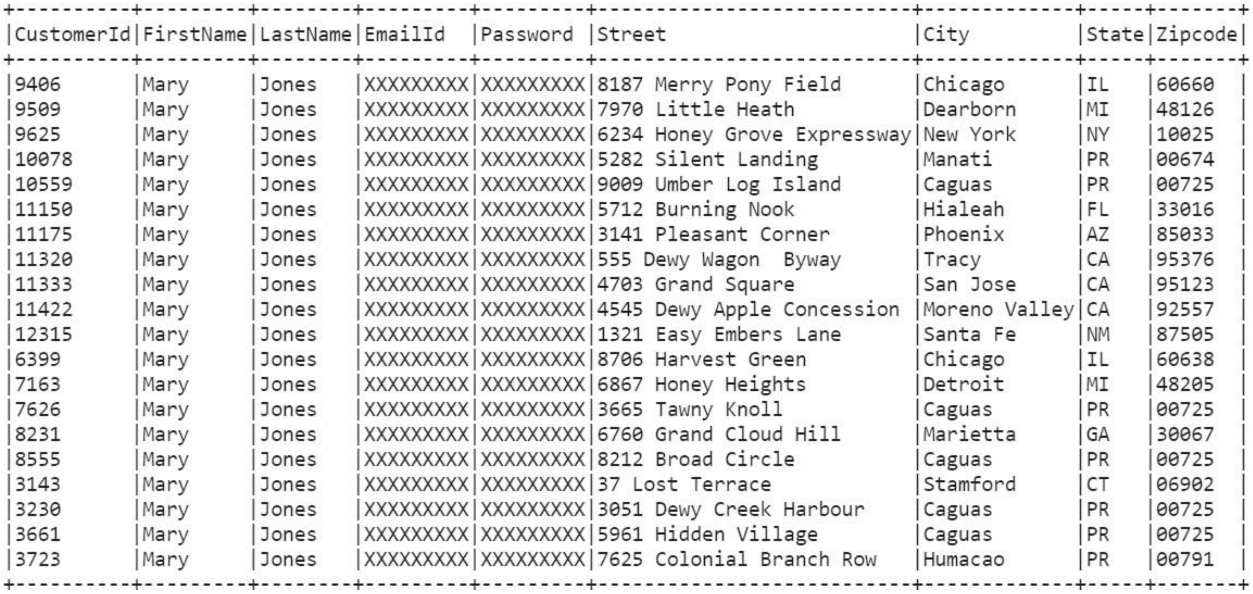
Project 2: A1Mart Online Retail Analysis

- Import data from MySQL to HDFS



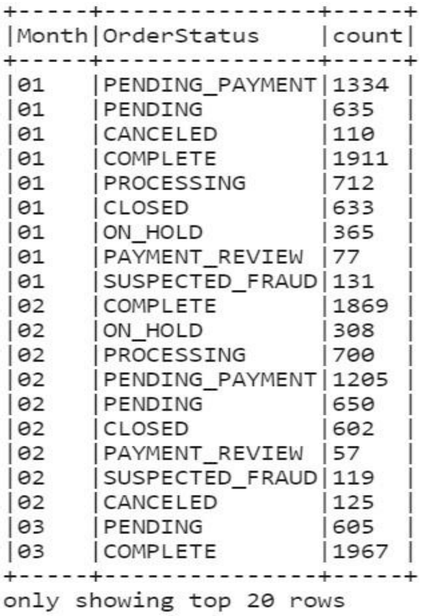
```
[itv001180@g02 ~]$ hadoop fs -ls /user/${USER}/Project2
Found 3 items
drwxr-xr-x   - itv001180 supergroup          0 2021-09-23 12:01 /user/itv001180/Project2/customers
drwxr-xr-x   - itv001180 supergroup          0 2021-09-30 09:43 /user/itv001180/Project2/order_items
drwxr-xr-x   - itv001180 supergroup          0 2021-09-23 12:02 /user/itv001180/Project2/orders
```

- Retrieve all records for particular customer



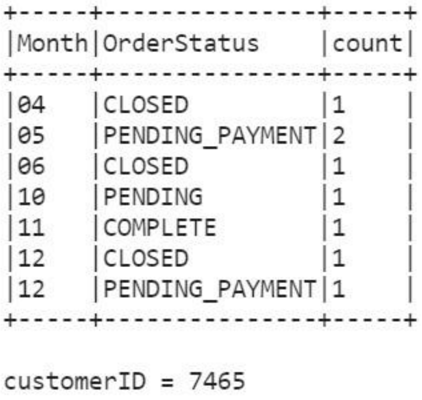
```
+----------+---------+--------+---------+---------+---------------------------+-------------+-----+-------+
|CustomerId|FirstName|LastName|EmailId  |Password |Street                     |City         |State|Zipcode|
+----------+---------+--------+---------+---------+---------------------------+-------------+-----+-------+
|9406      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|8187 Merry Pony Field      |Chicago      |IL   |60660  |
|9509      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|7970 Little Heath          |Dearborn     |MI   |48126  |
|9625      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|6234 Honey Grove Expressway|New York     |NY   |10025  |
|10078     |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|5282 Silent Landing        |Manati       |PR   |00674  |
|10559     |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|9009 Umber Log Island      |Caguas       |PR   |00725  |
|11150     |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|5712 Burning Nook          |Hialeah      |FL   |33016  |
|11175     |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|3141 Pleasant Corner       |Phoenix      |AZ   |85033  |
|11320     |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|555 Dewy Wagon  Byway      |Tracy        |CA   |95376  |
|11333     |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|4703 Grand Square          |San Jose     |CA   |95123  |
|11422     |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|4545 Dewy Apple Concession |Moreno Valley|CA   |92557  |
|12315     |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|1321 Easy Embers Lane      |Santa Fe     |NM   |87505  |
|6399      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|8706 Harvest Green         |Chicago      |IL   |60638  |
|7163      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|6867 Honey Heights         |Detroit      |MI   |48205  |
|7626      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|3665 Tawny Knoll           |Caguas       |PR   |00725  |
|8231      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|6760 Grand Cloud Hill      |Marietta     |GA   |30067  |
|8555      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|8212 Broad Circle          |Caguas       |PR   |00725  |
|3143      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|37 Lost Terrace            |Stamford     |CT   |06902  |
|3230      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|3051 Dewy Creek Harbour    |Caguas       |PR   |00725  |
|3661      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|5961 Hidden Village        |Caguas       |PR   |00725  |
|3723      |Mary     |Jones   |XXXXXXXXX|XXXXXXXXX|7625 Colonial Branch Row   |Humacao      |PR   |00791  |
+----------+---------+--------+---------+---------+---------------------------+-------------+-----+-------+
```

- List count of orders based on status and month



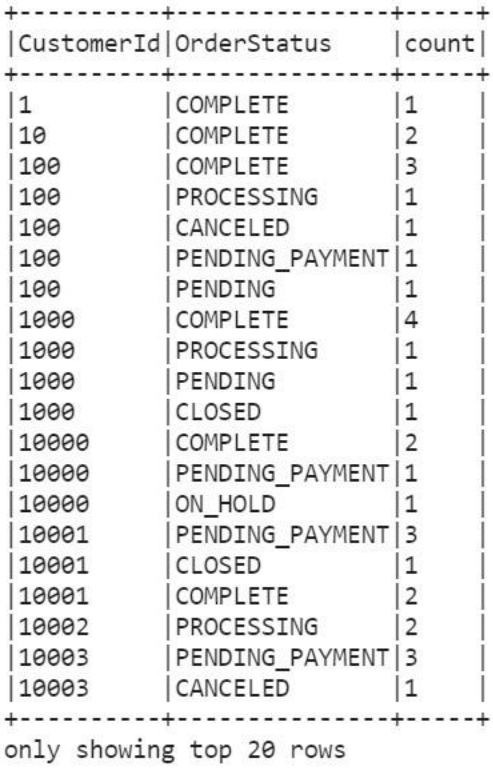
```
+-----+---------------+-----+
|Month|OrderStatus    |count|
+-----+---------------+-----+
|01   |PENDING_PAYMENT|1334 |
|01   |PENDING        |635  |
|01   |CANCELED       |110  |
|01   |COMPLETE       |1911 |
|01   |PROCESSING     |712  |
|01   |CLOSED         |633  |
|01   |ON_HOLD        |365  |
|01   |PAYMENT_REVIEW |77   |
|01   |SUSPECTED_FRAUD|131  |
|02   |COMPLETE       |1869 |
|02   |ON_HOLD        |308  |
|02   |PROCESSING     |700  |
|02   |PENDING_PAYMENT|1205 |
|02   |PENDING        |650  |
|02   |CLOSED         |602  |
|02   |PAYMENT_REVIEW |57   |
|02   |SUSPECTED_FRAUD|119  |
|02   |CANCELED       |125  |
|03   |PENDING        |605  |
|03   |COMPLETE       |1967 |
+-----+---------------+-----+
only showing top 20 rows
```

- List count of orders based on status and month for particular customer



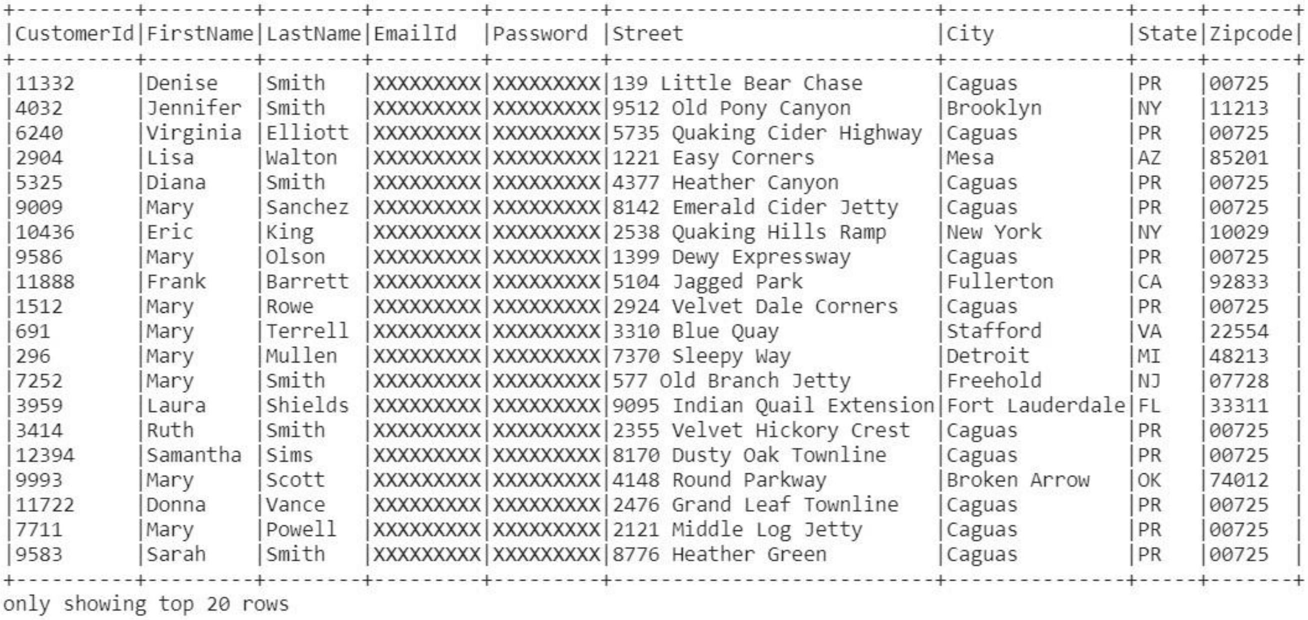
```
+-----+--------------+-----+
|Month|OrderStatus   |count|
+-----+--------------+-----+
|04   |CLOSED        |1    |
|05   |PENDING_PAYMENT|2    |
|06   |CLOSED        |1    |
|10   |PENDING       |1    |
|11   |COMPLETE      |1    |
|12   |CLOSED        |1    |
|12   |PENDING_PAYMENT|1    |
+-----+--------------+-----+

customerID = 7465
```

- List count of orders based on customer and status



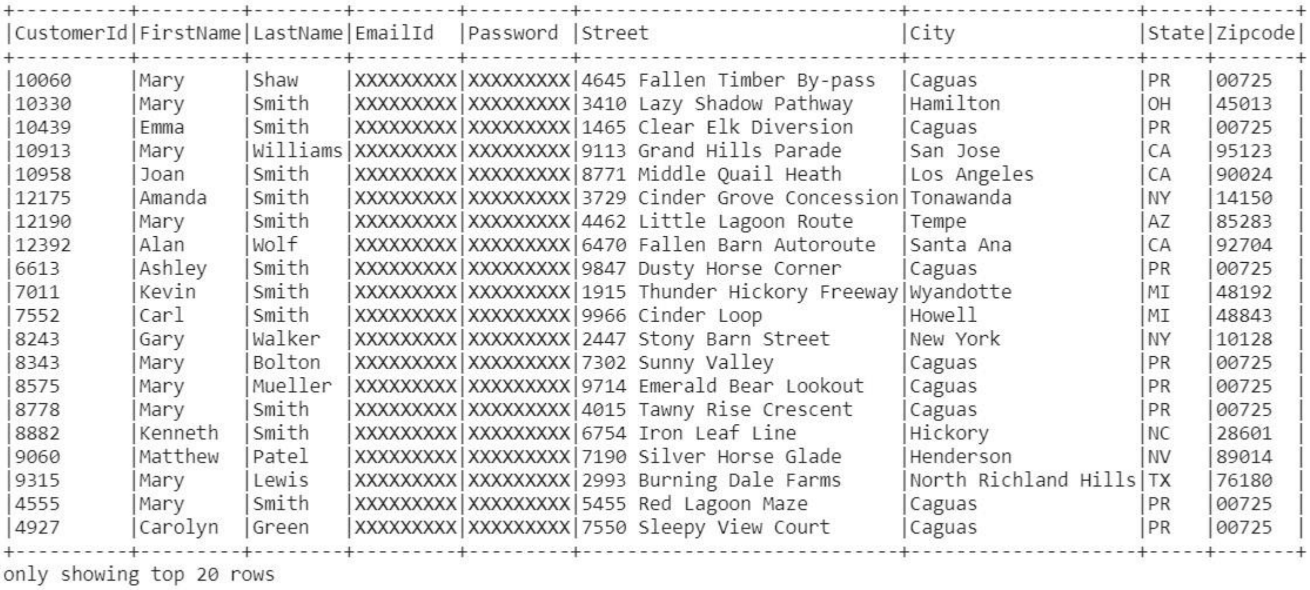
```
+----------+--------------+-----+
|CustomerId|OrderStatus   |count|
+----------+--------------+-----+
|1         |COMPLETE      |1    |
|10        |COMPLETE      |2    |
|100       |COMPLETE      |3    |
|100       |PROCESSING    |1    |
|100       |CANCELED      |1    |
|100       |PENDING_PAYMENT|1    |
|100       |PENDING       |1    |
|1000      |COMPLETE      |4    |
|1000      |PROCESSING    |1    |
|1000      |PENDING       |1    |
|1000      |CLOSED        |1    |
|10000     |COMPLETE      |2    |
|10000     |PENDING_PAYMENT|1    |
|10000     |ON_HOLD       |1    |
|10001     |PENDING_PAYMENT|3    |
|10001     |CLOSED        |1    |
|10001     |COMPLETE      |2    |
|10002     |PROCESSING    |2    |
|10003     |PENDING_PAYMENT|3    |
|10003     |CANCELED      |1    |
+----------+--------------+-----+
only showing top 20 rows
```

- Find the customers who have placed orders



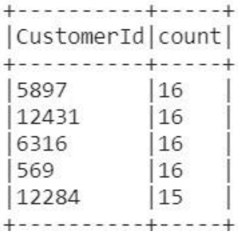
```
+----------+---------+--------+---------+---------+---------------------------+---------------+-----+-------+
|CustomerId|FirstName|LastName|EmailId  |Password |Street                     |City           |State|Zipcode|
+----------+---------+--------+---------+---------+---------------------------+---------------+-----+-------+
|11332     |Denise   |Smith   |XXXXXXXXX|XXXXXXXXX|139 Little Bear Chase      |Caguas         |PR   |00725  |
|4032      |Jennifer |Smith   |XXXXXXXXX|XXXXXXXXX|9512 Old Pony Canyon       |Brooklyn       |NY   |11213  |
|6240      |Virginia |Elliott |XXXXXXXXX|XXXXXXXXX|5735 Quaking Cider Highway |Caguas         |PR   |00725  |
|2904      |Lisa     |Walton  |XXXXXXXXX|XXXXXXXXX|1221 Easy Corners          |Mesa           |AZ   |85201  |
|5325      |Diana    |Smith   |XXXXXXXXX|XXXXXXXXX|4377 Heather Canyon        |Caguas         |PR   |00725  |
|9009      |Mary     |Sanchez |XXXXXXXXX|XXXXXXXXX|8142 Emerald Cider Jetty   |Caguas         |PR   |00725  |
|10436     |Eric     |King    |XXXXXXXXX|XXXXXXXXX|2538 Quaking Hills Ramp    |New York       |NY   |10029  |
|9586      |Mary     |Olson   |XXXXXXXXX|XXXXXXXXX|1399 Dewy Expressway       |Caguas         |PR   |00725  |
|11888     |Frank    |Barrett |XXXXXXXXX|XXXXXXXXX|5104 Jagged Park           |Fullerton      |CA   |92833  |
|1512      |Mary     |Rowe    |XXXXXXXXX|XXXXXXXXX|2924 Velvet Dale Corners   |Caguas         |PR   |00725  |
|691       |Mary     |Terrell |XXXXXXXXX|XXXXXXXXX|3310 Blue Quay             |Stafford       |VA   |22554  |
|296       |Mary     |Mullen  |XXXXXXXXX|XXXXXXXXX|7370 Sleepy Way            |Detroit        |MI   |48213  |
|7252      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|577 Old Branch Jetty       |Freehold       |NJ   |07728  |
|3959      |Laura    |Shields |XXXXXXXXX|XXXXXXXXX|9095 Indian Quail Extension|Fort Lauderdale|FL   |33311  |
|3414      |Ruth     |Smith   |XXXXXXXXX|XXXXXXXXX|2355 Velvet Hickory Crest  |Caguas         |PR   |00725  |
|12394     |Samantha |Sims    |XXXXXXXXX|XXXXXXXXX|8170 Dusty Oak Townline    |Caguas         |PR   |00725  |
|9993      |Mary     |Scott   |XXXXXXXXX|XXXXXXXXX|4148 Round Parkway         |Broken Arrow   |OK   |74012  |
|11722     |Donna    |Vance   |XXXXXXXXX|XXXXXXXXX|2476 Grand Leaf Townline   |Caguas         |PR   |00725  |
|7711      |Mary     |Powell  |XXXXXXXXX|XXXXXXXXX|2121 Middle Log Jetty      |Caguas         |PR   |00725  |
|9583      |Sarah    |Smith   |XXXXXXXXX|XXXXXXXXX|8776 Heather Green         |Caguas         |PR   |00725  |
+----------+---------+--------+---------+---------+---------------------------+---------------+-----+-------+
only showing top 20 rows
```

- Find the customers who have not placed orders yet



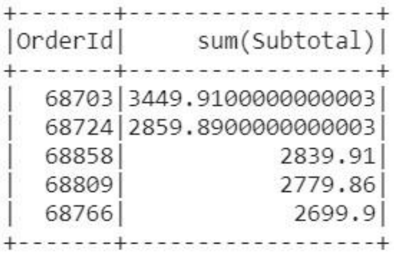
```
+----------+---------+--------+---------+---------+----------------------------+--------------------+-----+-------+
|CustomerId|FirstName|LastName|EmailId  |Password |Street                      |City                |State|Zipcode|
+----------+---------+--------+---------+---------+----------------------------+--------------------+-----+-------+
|10060     |Mary     |Shaw    |XXXXXXXXX|XXXXXXXXX|4645 Fallen Timber By-pass  |Caguas              |PR   |00725  |
|10330     |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|3410 Lazy Shadow Pathway    |Hamilton            |OH   |45013  |
|10439     |Emma     |Smith   |XXXXXXXXX|XXXXXXXXX|1465 Clear Elk Diversion    |Caguas              |PR   |00725  |
|10913     |Mary     |Williams|XXXXXXXXX|XXXXXXXXX|9113 Grand Hills Parade     |San Jose            |CA   |95123  |
|10958     |Joan     |Smith   |XXXXXXXXX|XXXXXXXXX|8771 Middle Quail Heath     |Los Angeles         |CA   |90024  |
|12175     |Amanda   |Smith   |XXXXXXXXX|XXXXXXXXX|3729 Cinder Grove Concession|Tonawanda           |NY   |14150  |
|12190     |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|4462 Little Lagoon Route    |Tempe               |AZ   |85283  |
|12392     |Alan     |Wolf    |XXXXXXXXX|XXXXXXXXX|6470 Fallen Barn Autoroute  |Santa Ana           |CA   |92704  |
|6613      |Ashley   |Smith   |XXXXXXXXX|XXXXXXXXX|9847 Dusty Horse Corner     |Caguas              |PR   |00725  |
|7011      |Kevin    |Smith   |XXXXXXXXX|XXXXXXXXX|1915 Thunder Hickory Freeway|Wyandotte           |MI   |48192  |
|7552      |Carl     |Smith   |XXXXXXXXX|XXXXXXXXX|9966 Cinder Loop            |Howell              |MI   |48843  |
|8243      |Gary     |Walker  |XXXXXXXXX|XXXXXXXXX|2447 Stony Barn Street      |New York            |NY   |10128  |
|8343      |Mary     |Bolton  |XXXXXXXXX|XXXXXXXXX|7302 Sunny Valley           |Caguas              |PR   |00725  |
|8575      |Mary     |Mueller |XXXXXXXXX|XXXXXXXXX|9714 Emerald Bear Lookout   |Caguas              |PR   |00725  |
|8778      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|4015 Tawny Rise Crescent    |Caguas              |PR   |00725  |
|8882      |Kenneth  |Smith   |XXXXXXXXX|XXXXXXXXX|6754 Iron Leaf Line         |Hickory             |NC   |28601  |
|9060      |Matthew  |Patel   |XXXXXXXXX|XXXXXXXXX|7190 Silver Horse Glade     |Henderson           |NV   |89014  |
|9315      |Mary     |Lewis   |XXXXXXXXX|XXXXXXXXX|2993 Burning Dale Farms     |North Richland Hills|TX   |76180  |
|4555      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|5455 Red Lagoon Maze        |Caguas              |PR   |00725  |
|4927      |Carolyn  |Green   |XXXXXXXXX|XXXXXXXXX|7550 Sleepy View Court      |Caguas              |PR   |00725  |
+----------+---------+--------+---------+---------+----------------------------+--------------------+-----+-------+
only showing top 20 rows
```

- Find top 5 customers
    - Highest number of orders



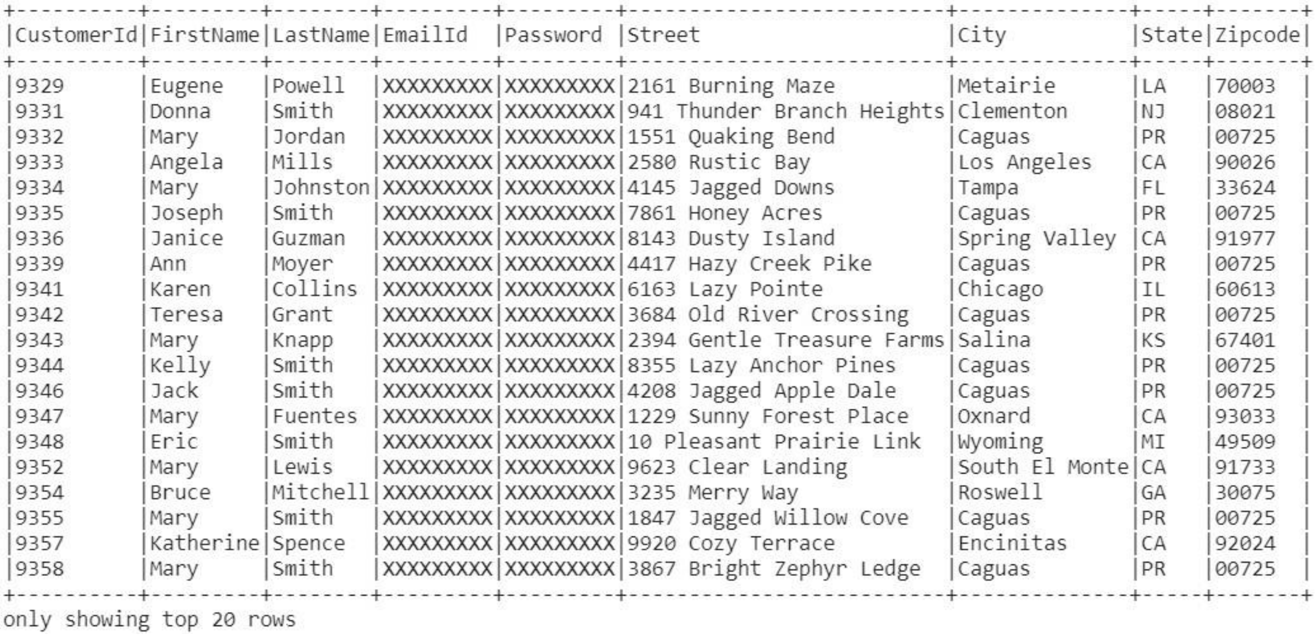
```
+----------+-----+
|CustomerId|count|
+----------+-----+
|5897      |16   |
|12431     |16   |
|6316      |16   |
|569       |16   |
|12284     |15   |
+----------+-----+
```

- Highest sum of total orders



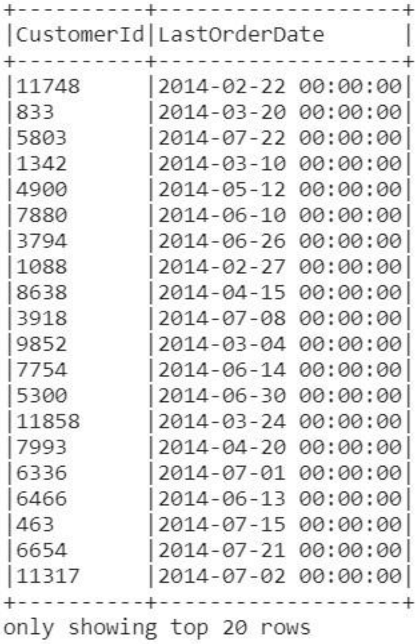
```
+-------+------------------+
|OrderId|     sum(Subtotal)|
+-------+------------------+
|  68703|3449.9100000000003|
|  68724|2859.8900000000003|
|  68858|           2839.91|
|  68809|           2779.86|
|  68766|            2699.9|
+-------+------------------+
```

- Find the customer who did not order in last 1 month or for long time



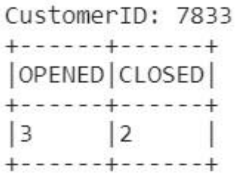
```
+----------+---------+--------+---------+---------+--------------------------+--------------+-----+-------+
|CustomerId|FirstName|LastName|EmailId  |Password |Street                    |City          |State|Zipcode|
+----------+---------+--------+---------+---------+--------------------------+--------------+-----+-------+
|9329      |Eugene   |Powell  |XXXXXXXXX|XXXXXXXXX|2161 Burning Maze         |Metairie      |LA   |70003  |
|9331      |Donna    |Smith   |XXXXXXXXX|XXXXXXXXX|941 Thunder Branch Heights|Clementon     |NJ   |08021  |
|9332      |Mary     |Jordan  |XXXXXXXXX|XXXXXXXXX|1551 Quaking Bend         |Caguas        |PR   |00725  |
|9333      |Angela   |Mills   |XXXXXXXXX|XXXXXXXXX|2580 Rustic Bay           |Los Angeles   |CA   |90026  |
|9334      |Mary     |Johnston|XXXXXXXXX|XXXXXXXXX|4145 Jagged Downs         |Tampa         |FL   |33624  |
|9335      |Joseph   |Smith   |XXXXXXXXX|XXXXXXXXX|7861 Honey Acres          |Caguas        |PR   |00725  |
|9336      |Janice   |Guzman  |XXXXXXXXX|XXXXXXXXX|8143 Dusty Island         |Spring Valley |CA   |91977  |
|9339      |Ann      |Moyer   |XXXXXXXXX|XXXXXXXXX|4417 Hazy Creek Pike      |Caguas        |PR   |00725  |
|9341      |Karen    |Collins |XXXXXXXXX|XXXXXXXXX|6163 Lazy Pointe          |Chicago       |IL   |60613  |
|9342      |Teresa   |Grant   |XXXXXXXXX|XXXXXXXXX|3684 Old River Crossing   |Caguas        |PR   |00725  |
|9343      |Mary     |Knapp   |XXXXXXXXX|XXXXXXXXX|2394 Gentle Treasure Farms|Salina        |KS   |67401  |
|9344      |Kelly    |Smith   |XXXXXXXXX|XXXXXXXXX|8355 Lazy Anchor Pines    |Caguas        |PR   |00725  |
|9346      |Jack     |Smith   |XXXXXXXXX|XXXXXXXXX|4208 Jagged Apple Dale    |Caguas        |PR   |00725  |
|9347      |Mary     |Fuentes |XXXXXXXXX|XXXXXXXXX|1229 Sunny Forest Place   |Oxnard        |CA   |93033  |
|9348      |Eric     |Smith   |XXXXXXXXX|XXXXXXXXX|10 Pleasant Prairie Link  |Wyoming       |MI   |49509  |
|9352      |Mary     |Lewis   |XXXXXXXXX|XXXXXXXXX|9623 Clear Landing        |South El Monte|CA   |91733  |
|9354      |Bruce    |Mitchell|XXXXXXXXX|XXXXXXXXX|3235 Merry Way            |Roswell       |GA   |30075  |
|9355      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|1847 Jagged Willow Cove   |Caguas        |PR   |00725  |
|9357      |Katherine|Spence  |XXXXXXXXX|XXXXXXXXX|9920 Cozy Terrace         |Encinitas     |CA   |92024  |
|9358      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|3867 Bright Zephyr Ledge  |Caguas        |PR   |00725  |
+----------+---------+--------+---------+---------+--------------------------+--------------+-----+-------+
only showing top 20 rows
```

- Find the last order date for all customers



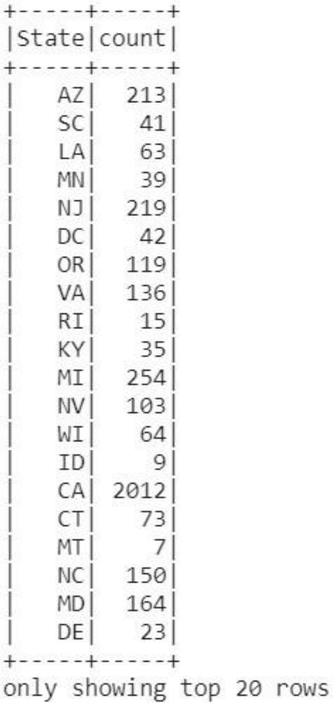
```
+----------+-------------------+
|CustomerId|LastOrderDate      |
+----------+-------------------+
|11748     |2014-02-22 00:00:00|
|833       |2014-03-20 00:00:00|
|5803      |2014-07-22 00:00:00|
|1342      |2014-03-10 00:00:00|
|4900      |2014-05-12 00:00:00|
|7880      |2014-06-10 00:00:00|
|3794      |2014-06-26 00:00:00|
|1088      |2014-02-27 00:00:00|
|8638      |2014-04-15 00:00:00|
|3918      |2014-07-08 00:00:00|
|9852      |2014-03-04 00:00:00|
|7754      |2014-06-14 00:00:00|
|5300      |2014-06-30 00:00:00|
|11858     |2014-03-24 00:00:00|
|7993      |2014-04-20 00:00:00|
|6336      |2014-07-01 00:00:00|
|6466      |2014-06-13 00:00:00|
|463       |2014-07-15 00:00:00|
|6654      |2014-07-21 00:00:00|
|11317     |2014-07-02 00:00:00|
+----------+-------------------+
only showing top 20 rows
```

- Find open and close number of orders for a customer



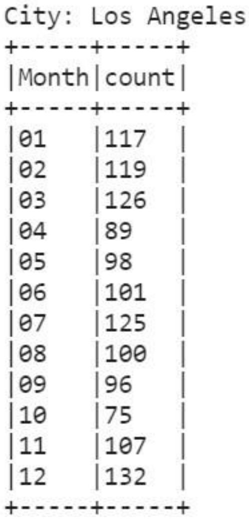
```
CustomerID: 7833
+------+------+
|OPENED|CLOSED|
+------+------+
|3     |2     |
+------+------+
```

- Find number of customers in every state



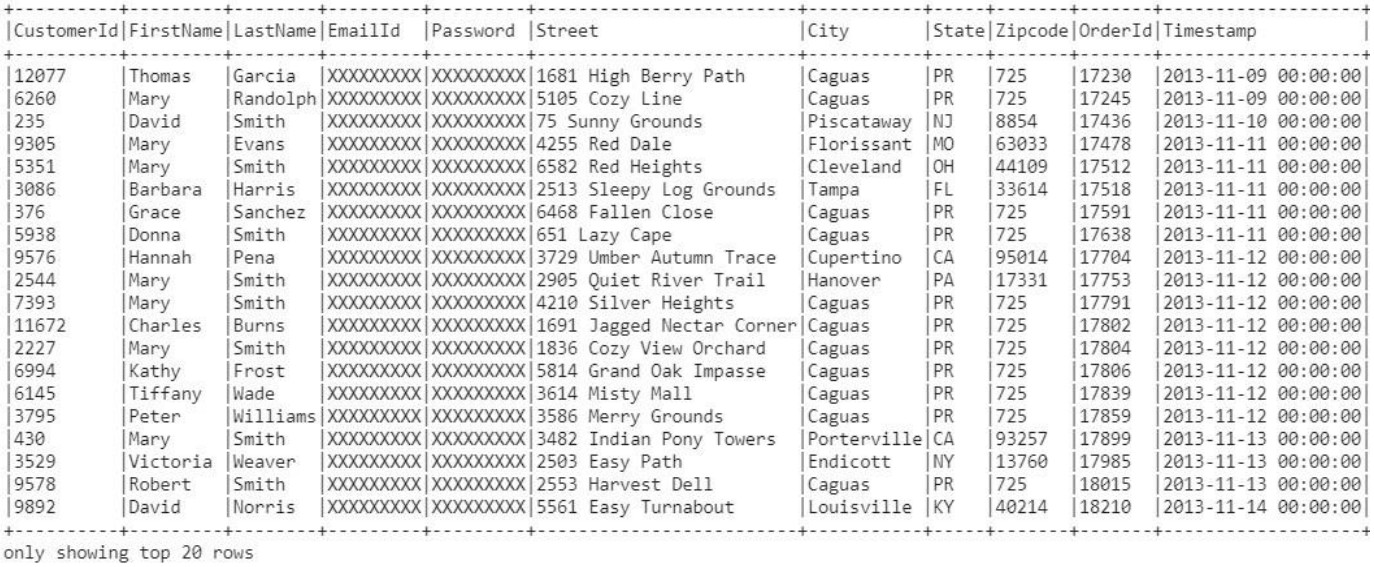
```
+-----+-----+
|State|count|
+-----+-----+
|   AZ|  213|
|   SC|   41|
|   LA|   63|
|   MN|   39|
|   NJ|  219|
|   DC|   42|
|   OR|  119|
|   VA|  136|
|   RI|   15|
|   KY|   35|
|   MI|  254|
|   NV|  103|
|   WI|   64|
|   ID|    9|
|   CA| 2012|
|   CT|   73|
|   MT|    7|
|   NC|  150|
|   MD|  164|
|   DE|   23|
+-----+-----+
only showing top 20 rows
```

- 3 additional requirements
    - Number of orders based on month for a particular city



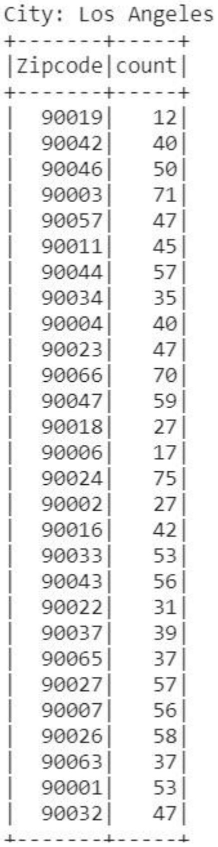
```
City: Los Angeles
+-----+-----+
|Month|count|
+-----+-----+
|01   |117  |
|02   |119  |
|03   |126  |
|04   |89   |
|05   |98   |
|06   |101  |
|07   |125  |
|08   |100  |
|09   |96   |
|10   |75   |
|11   |107  |
|12   |132  |
+-----+-----+
```

- Customer details whose orders are suspected to be fraud



```
+----------+---------+--------+---------+---------+-------------------------+-----------+-----+-------+-------+-------------------+
|CustomerId|FirstName|LastName|EmailId  |Password |Street                   |City       |State|Zipcode|OrderId|Timestamp          |
+----------+---------+--------+---------+---------+-------------------------+-----------+-----+-------+-------+-------------------+
|12077     |Thomas   |Garcia  |XXXXXXXXX|XXXXXXXXX|1681 High Berry Path     |Caguas     |PR   |725    |17230  |2013-11-09 00:00:00|
|6260      |Mary     |Randolph|XXXXXXXXX|XXXXXXXXX|5105 Cozy Line           |Caguas     |PR   |725    |17245  |2013-11-09 00:00:00|
|235       |David    |Smith   |XXXXXXXXX|XXXXXXXXX|75 Sunny Grounds         |Piscataway |NJ   |8854   |17436  |2013-11-10 00:00:00|
|9305      |Mary     |Evans   |XXXXXXXXX|XXXXXXXXX|4255 Red Dale            |Florissant |MO   |63033  |17478  |2013-11-11 00:00:00|
|5351      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|6582 Red Heights         |Cleveland  |OH   |44109  |17512  |2013-11-11 00:00:00|
|3086      |Barbara  |Harris  |XXXXXXXXX|XXXXXXXXX|2513 Sleepy Log Grounds  |Tampa      |FL   |33614  |17518  |2013-11-11 00:00:00|
|376       |Grace    |Sanchez |XXXXXXXXX|XXXXXXXXX|6468 Fallen Close        |Caguas     |PR   |725    |17591  |2013-11-11 00:00:00|
|5938      |Donna    |Smith   |XXXXXXXXX|XXXXXXXXX|651 Lazy Cape            |Caguas     |PR   |725    |17638  |2013-11-11 00:00:00|
|9576      |Hannah   |Pena    |XXXXXXXXX|XXXXXXXXX|3729 Umber Autumn Trace  |Cupertino  |CA   |95014  |17704  |2013-11-12 00:00:00|
|2544      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|2905 Quiet River Trail   |Hanover    |PA   |17331  |17753  |2013-11-12 00:00:00|
|7393      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|4210 Silver Heights      |Caguas     |PR   |725    |17791  |2013-11-12 00:00:00|
|11672     |Charles  |Burns   |XXXXXXXXX|XXXXXXXXX|1691 Jagged Nectar Corner|Caguas     |PR   |725    |17802  |2013-11-12 00:00:00|
|2227      |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|1836 Cozy View Orchard   |Caguas     |PR   |725    |17804  |2013-11-12 00:00:00|
|6994      |Kathy    |Frost   |XXXXXXXXX|XXXXXXXXX|5814 Grand Oak Impasse   |Caguas     |PR   |725    |17806  |2013-11-12 00:00:00|
|6145      |Tiffany  |Wade    |XXXXXXXXX|XXXXXXXXX|3614 Misty Mall          |Caguas     |PR   |725    |17839  |2013-11-12 00:00:00|
|3795      |Peter    |Williams|XXXXXXXXX|XXXXXXXXX|3586 Merry Grounds       |Caguas     |PR   |725    |17859  |2013-11-12 00:00:00|
|430       |Mary     |Smith   |XXXXXXXXX|XXXXXXXXX|3482 Indian Pony Towers  |Porterville|CA   |93257  |17899  |2013-11-13 00:00:00|
|3529      |Victoria |Weaver  |XXXXXXXXX|XXXXXXXXX|2503 Easy Path           |Endicott   |NY   |13760  |17985  |2013-11-13 00:00:00|
|9578      |Robert   |Smith   |XXXXXXXXX|XXXXXXXXX|2553 Harvest Dell        |Caguas     |PR   |725    |18015  |2013-11-13 00:00:00|
|9892      |David    |Norris  |XXXXXXXXX|XXXXXXXXX|5561 Easy Turnabout      |Louisville |KY   |40214  |18210  |2013-11-14 00:00:00|
+----------+---------+--------+---------+---------+-------------------------+-----------+-----+-------+-------+-------------------+
only showing top 20 rows
```

- Number of orders based on zip code for a particular city



```
City: Los Angeles
+-------+-----+
|Zipcode|count|
+-------+-----+
|  90019|   12|
|  90042|   40|
|  90046|   50|
|  90003|   71|
|  90057|   47|
|  90011|   45|
|  90044|   57|
|  90034|   35|
|  90004|   40|
|  90023|   47|
|  90066|   70|
|  90047|   59|
|  90018|   27|
|  90006|   17|
|  90024|   75|
|  90002|   27|
|  90016|   42|
|  90033|   53|
|  90043|   56|
|  90022|   31|
|  90037|   39|
|  90065|   37|
|  90027|   57|
|  90007|   56|
|  90026|   58|
|  90063|   37|
|  90001|   53|
|  90032|   47|
+-------+-----+
```

Optimizations:

- The given data is in text file format (csv format). As we have learnt, text file format consumes a lot of storage and processing becomes slow. So, we need to convert this file to Hadoop-specific file format.
- Since we need to process only some columns for most of the issues, I prefer column-based file format rather than row-based file format.
- Since we are working on Apache Spark and we won't be using Hive, I prefer Parquet file format over ORC.
- For data reading purpose, we can use Hbase over Hive, since it involves less execution time.
- Although sql-query is more user-friendly, dataframes are more preferred over sql as sql functions might have to be imported explicitly and might affect the performance.
- Data could be partitioned into different files based on "state". It will help in running State-specific queries or even city-specific queries and can save a lot of processing.
- If the job involves multiple operations, it should be sequenced such that the amount of data that needs to be processed for later operations is less. It could be done if one of the operations is filter.