Assessment 4: Managing data with Hive and Pig in Hadoop

MA5831 – Advanced Data Processing and Analysis using SAS

13848336 Nikki Fitzherbert

Reporting and Coding Tasks – Hive Queries

Question 1

The following image extract from the Hive query editor in Hue presents the code used to determine the 10 states with the maximum number of complaints (that is, the top 10 states by total number of complaints).

The *state* field was selected from the *consumer_complaints* table along with a new field called *num_complaints* that counted all the complaints using the *complaint_id* field. The complaints were then grouped into their respective states and counted up before being ordered by *num_complaints* in descending order. Finally, the number of rows of output was limited to 10 rows.

The output indicates that the 10 states with the most complaints were (in descending order) California, Florida, Texas, New York, Georgia, New Jersey, Pennsylvania, Illinois, Maryland and Virginia.

Question 1 Determine the 10 states with the maximum number of complaints

```
select state, count(complaint_id) as num_complaints
from dihps.consumer_complaints
group by state
order by num_complaints desc
limit 10;
```

Execute | Save | Save as... | Explain | or create a | New query

Recent queries | Query | Log | Columns | Results | Chart

	state	num_complaints
0	CA	60528
1	FL	39381
2	TX	29817
3	NY	27974
4	GA	17705
5	NJ	16588
6	PA	14637
7	IL	14332
8	MD	13136
9	VA	13136

Question 2

The following image extract from the Hive query editor in Hue presents the code used to determine how many complaints were associated with the “Medical” sub-product offering.

A new field called *num_complaints* was created that counted all the complaints using the *complaint_id* field. No other fields apart from this one were selected from the *consumer_complaints* table. Initial exploration of the data had indicated that there was only one sub-product offering containing the word “medical”, so it was appropriate to limit the results using a simple WHERE statement and the desired sub-product name.

The output indicates that there were 8,271 complaints associated with the “Medical” sub-product offering.

Question 3

The following image extract from the Hive query editor in Hue presents the code used to identify five ZIP codes with the smallest number of complaints.

The *zip_code* field was selected from the *consumer_complaints* table along with a new field called *num_complaints* that counted all the complaints using the *complaint_id* field. Initial exploration of the data had revealed that there were data quality issues, so the results were explicitly limited to those where the value of *zip_code* was greater than 500 using a WHERE statement. Further detail about the identified data quality issues are discussed further in the following paragraph. The results were grouped by *zip_code*, ordered by *num_complaints* in ascending order to ensure zip codes with the lowest number of complaints were listed first in the output, and finally limited to five rows of output.

Initial querying of the *consumer_complaints* data had revealed that there were unexpected values in the zip code field. In particular, there were zip code values in the range of 200 to 499, which are not currently in use by the United States Postal Service. All US states, international territories and military bases have a zip code of 500 or higher, which is why that value was chosen for the WHERE clause condition. The exploration also identified that there were instances where a record's state value did not match its zip code value and vice versa. For example, zip codes in the state of California start with the number 9, but there was at least one record with a different value. This particular issue had implications for the results of all queries involving the state and/or zip code fields, but was ignored for now as rectifying data quality issues was outside the scope of this exercise.

The final notable observation was that there were more than five unique zip codes for which there was only a single complaint (6,190 to be exact), so the output presented was only one of many possible valid solutions.

Question 3 Determine five ZIP codes with the smallest number of complaints

```
select zip_code, count(complaint_id) as num_complaints
from dihps.consumer_complaints
where zip_code > 500
group by zip_code
order by num_complaints asc
limit 5;
```

Execute Save Save as... Explain or create a New query

Recent queries Query Log Columns Results Chart

	zip_code	num_complaints
0	10066	1
1	10043	1
2	10050	1
3	99926	1
4	1000	1

Question 4

The following image extract from the Hive query editor in Hue presents the code used to determine how many complaints, grouped by product and state, were associated with the word “fraud” in the issue description.

The *product* and *state* fields were selected from the *consumer_complaints* table along with a new field called *num_complaints* that counted all the complaints using the *complaint_id* field. The results were limited to those where *issue* contained the word “fraud” in any position within the row. Also, the string function lower() was used in the WHERE statement to captured any variation in capitalisation of the word “fraud” within the field. The WHERE statement contained these additions because the description of the *issue* field – ‘primary reason that the customer filed the complaint, such as checking (cheque) account charges’ – suggested that it might have contained free text and therefore “fraud” might have appeared in a number of different forms within the field values. Finally, the results were grouped by *product* and *state*.

There were too many unique combinations of *product* and *state* for the entire output to be displayed here. Instead, the image extract shows the first 11 rows, which are all from the “credit card” product offering. It indicates for example that there were 590 complaints recorded against the state of California, only a single complaint each originated from military bases located in Europe, the Middle East or Canada (AE) and military bases located in the United States (AA), and there were 70 credit card complaints for which no state code had been recorded.

Question 4 Determine how many complaints, grouped by product and state, contain the word 'fraud' in the issue description

```
select product, state, count(complaint_id) as num_complaints
from dihps.consumer_complaints
where lower(issue) like '%fraud%'
group by product, state;
```

Execute | Save | Save as... | Explain | or create a | New query

Recent queries | Query | Log | Columns | Results | Chart

	product	state	num_complaints
0	Credit card		70
1	Credit card	AA	1
2	Credit card	AE	1
3	Credit card	AK	7
4	Credit card	AL	37
5	Credit card	AR	19
6	Credit card	AZ	81
7	Credit card	CA	590
8	Credit card	CO	93
9	Credit card	CT	45
10	Credit card	DC	21

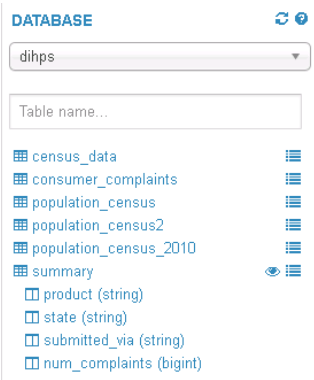
Question 5

The following image extract from the Hive query editor in Hue presents the code used to create a new table that summarised the total number of complaints by product, state and submitted_via.

The first line of code created a new Hive table in the *DIHPS* database called *summary*. The 'if not exists' option was not strictly necessary in this case as a quick visual perusal of the tables already in that database had verified that no other table with the same name already existed. The option is usually included in a CREATE TABLE statement to ensure that a pre-existing table with the same name is not overwritten, which is particularly useful if it contains the output of a different query. The table was then populated with *consumer_complaints* data as follows: First, the *product*, *state* and *submitted_via* fields were selected along with a new field called *num_complaints* that counted up all the complaints. Second, the complaints were grouped according to product offering type, state and how each was submitted.

Question 5 Create a table summarising the total number of complaints by product, state and submitted_via

```
create table if not exists dihps.summary as
select product, state, submitted_via, count(*) as num_complaints
from dihps.consumer_complaints
group by product, state, submitted_via;
```



As with the previous query, there were too many unique combinations of *product*, *state* and *submitted_via* for the entire output to be displayed here. Instead, the image extract shows the first 13 rows, which are all from the "bank account or service" product offering, but different states and the full list of possible values for *submitted_via*. It indicates for example that there were three complaints associated with a bank account or service product that were submitted over the phone and originated from a military base located in the United States (AA).

Data sample for summary

	product	state	submitted_via	num_complaints
0	Bank account or service		Email	8
1	Bank account or service		Fax	32
2	Bank account or service		Phone	232
3	Bank account or service		Postal mail	40
4	Bank account or service		Referral	413
5	Bank account or service		Web	211
6	Bank account or service	AA	Phone	3
7	Bank account or service	AA	Web	1
8	Bank account or service	AE	Fax	1
9	Bank account or service	AE	Phone	1
10	Bank account or service	AE	Postal mail	1
11	Bank account or service	AE	Referral	7
12	Bank account or service	AE	Web	9

Reporting and Coding Tasks – Pig Latin scripts

Question 6

The following image extract from the Pig query editor in Hue presents the code used to create two separate tables called *web_results* and *other_results*, which separated the consumer complaints according to whether they were submitted via the web or any other method.

The same script was also used to verify that 254,550 complaint records had been submitted via a web submission form (the contents of *web_results*) and 154,850 complaint records had been submitted by phone, postal mail, referral, fax or email (the contents of *other_results*).

Question 6 - Number of complaints submitted via web vs other methods

```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define scheme to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,
sub_issue:chararray,
state:chararray,
zip_code:chararray,
submitted_via:chararray,
date_received:chararray,
date_sent_to_compant:chararray,
company_response:chararray,
timely_response:chararray,
consumer_disputed:chararray
);

--filter rows of file to include only those where lower-case submitted_via equals 'web'
T_WEB = FILTER T BY LOWER(submitted_via) == 'web';
-- filter rows of ffile to include only those where lower-case submitted_via does not equal 'web'
T_NWEB = FILTER T BY LOWER(submitted_via) != 'web';

-- store the results in two separate tables in the HDFS DIHPS output folder
-- web results table
STORE T_WEB INTO '/user/student/DIHPS/output/web_results';
-- non-web results table
STORE T_NWEB INTO '/user/student/DIHPS/output/other_results';

-- verify the number of rows written to each table
-- group the results together
T_WEB_GRP = GROUP T_WEB ALL;
T_NWEB_GRP = GROUP T_NWEB ALL;

-- count the number of rows in the files
T_WEB_COUNT = FOREACH T_WEB_GRP GENERATE COUNT(T_WEB) AS num_complaints;
T_NWEB_COUNT = FOREACH T_NWEB_GRP GENERATE COUNT(T_NWEB) AS num_complaints;

-- store the results in two separate tables in the HDFS DIHPS output folder
-- web results table count
STORE T_WEB_COUNT INTO '/user/student/DIHPS/output/web_results_count';
-- non-web results table count
STORE T_NWEB_COUNT INTO '/user/student/DIHPS/output/other_results_count';
```

This image extract displays part of the contents of the new *web_results* table, and confirms that only complaints submitted via the web were copied across.

Home / user / student / DIHPS / output / web_results / **part-m-00000**

Page 1 to 1 of 9065

```
1431865 Consumer loan   Vehicle loan    Managing the loan or lease              NJ      8736    Web     6/22/2015       6/22/2015       In progress     Yes
1431374 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    WI      54140   Web     6/22/2015       6/22/2015       Closed with explanation Yes
1431251 Mortgage        Conventional fixed mortgage     Loan modification,collection,foreclosure                MO      63368   Web     6/22/2015       6/22/2015       In progress     Yes
1431743 Debt collection Medical Cont'd attempts collect debt not owed   Debt is not mine        WA      98055   Web     6/22/2015       6/22/2015       Closed with explanation Yes     Yes
1432678 Debt collection Medical Cont'd attempts collect debt not owed   Debt was paid   TX      75104   Web     6/22/2015       6/22/2015       Closed with explanation Yes
1432104 Debt collection Other (phone, health club, etc.)        Cont'd attempts collect debt not owed   Debt was paid   CA      95423   Web     6/22/2015       6/22/2015       In progress     Yes
1431998 Debt collection Other (phone, health club, etc.)        Communication tactics   Frequent or repeated calls      TX      75048   Web     6/22/2015       6/22/2015       Closed with explanation Yes
1432207 Debt collection Medical Disclosure verification of debt Right to dispute notice not received    TX      75125   Web     6/22/2015       6/22/2015       Closed with explanation Yes
1432202 Consumer loan   Vehicle loan    Problems when you are unable to pay             TN      37174   Web     6/22/2015       6/22/2015       In progress     Yes
1432334 Debt collection Other (phone, health club, etc.)        Disclosure verification of debt Right to dispute notice not received    FL      33830   Web     6/22/2015       6/22/2015       Closed with non-mon
etary relief    Yes
1431188 Debt collection Payday loan     Communication tactics   Threatened to take legal action NE      68134   Web     6/22/2015       6/22/2015       Closed with explanation Yes
1430973 Debt collection Other (phone, health club, etc.)        Cont'd attempts collect debt not owed   Debt was paid   OH      44057   Web     6/21/2015       6/21/2015       Closed with explanation Yes
1431167 Debt collection         Cont'd attempts collect debt not owed   Debt is not mine        CO      80015   Web     6/21/2015       6/22/2015       In progress     Yes
1430909 Credit reporting                Unable to get credit report/credit score        Problem getting report or credit score  CA      95823   Web     6/21/2015       6/21/2015       Closed with explanation Yes

1430879 Credit reporting                Incorrect information on credit report  Information is not mine TN      38401   Web     6/21/2015       6/21/2015       Closed with explanation Yes
1430964 Mortgage        Conventional adjustable mortgage (ARM)  Loan servicing, payments, escrow account                FL      33412   Web     6/21/2015       6/21/2015       In progress     Yes
1430955 Credit reporting                Incorrect information on credit report  Personal information     WA      98093   Web     6/21/2015       6/21/2015       Closed with non-monetary relief Yes
1430945 Credit reporting                Incorrect information on credit report  Information is not mine LA      70184   Web     6/21/2015       6/21/2015       Closed with explanation Yes
1430917 Bank account or service Checking account        Problems caused by my funds being low           WA      98665   Web     6/21/2015       6/21/2015       Closed with explanation Yes
1430916 Credit reporting                Incorrect information on credit report  Information is not mine TX      75056   Web     6/21/2015       6/21/2015       Closed with explanation Yes
1430968 Debt collection         Communication tactics   Threatened to take legal action AZ      85705   Web     6/21/2015       6/21/2015       In progress     Yes
1431161 Debt collection Other (phone, health club, etc.)        False statements or representation      Attempted to collect wrong amount       NY      11215   Web     6/21/2015       6/22/2015       Closed with
non-monetary relief     Yes
1431060 Credit reporting                Incorrect information on credit report  Information is not mine IL      60409   Web     6/21/2015       6/22/2015       In progress     Yes
1431020 Debt collection Other (phone, health club, etc.)        Cont'd attempts collect debt not owed   Debt is not mine        DC      20009   Web     6/21/2015       6/21/2015       Closed with explanation Yes

1430991 Mortgage        FHA mortgage    Loan modification,collection,foreclosure                CA      92404   Web     6/21/2015       6/21/2015       In progress     Yes
```

Similarly, this image extract displays part of the contents of the new *other_results* table, and confirms that no web-submitted complaints were copied across.

Home / user / student / DIHPS / output / other_results / **part-m-00000**

Page 1 to 1 of 5455

```
Complaint ID    Product Sub-product     Issue   Sub-issue       State   ZIP code        Submitted via   Date received   Date sent to company    Company response        Timely response?        Consumer disputed?
1431358 Money transfers Domestic (US) money transfer    Money was not available when promised           AL      35235   Phone   6/22/2015       6/22/2015       In progress     Yes
1429289 Debt collection Other (phone, health club, etc.)        Disclosure verification of debt Right to dispute notice not received    CA      91607   Postal mail     6/19/2015       6/23/2015       Closed with
explanation     Yes
1429322 Debt collection         Communication tactics   Threatened to take legal action WI      53149   Phone   6/19/2015       6/23/2015       Closed with explanation Yes
1429534 Consumer loan   Installment loan        Managing the loan or lease              CA      92027   Referral        6/19/2015       6/22/2015       In progress     Yes
1429488 Consumer loan   Vehicle loan    Managing the loan or lease              MD      20706   Referral        6/19/2015       6/23/2015       In progress     Yes
1429570 Consumer loan   Vehicle loan    Managing the loan or lease              IL      60441   Referral        6/19/2015       6/22/2015       In progress     Yes
1430137 Bank account or service Other bank product/service      Deposits and withdrawals                CA      30087   Referral        6/19/2015       6/22/2015       Closed with explanation Yes
1430509 Bank account or service Checking account        Account opening, closing, or management         PA      19609   Referral        6/19/2015       6/22/2015       In progress     Yes
1430494 Bank account or service Checking account        Deposits and withdrawals                FL      33139   Referral        6/19/2015       6/22/2015       Closed with monetary relief     Yes
1430546 Bank account or service Checking account        Deposits and withdrawals                PR      936     Referral        6/19/2015       6/22/2015       Closed with explanation Yes     Yes
1429047 Bank account or service Other bank product/service      Using a debit or ATM card               AZ      85024   Referral        6/19/2015       6/22/2015       In progress     Yes
1426726 Debt collection Credit card     Cont'd attempts collect debt not owed   Debt was paid   VA      23224   Fax     6/18/2015       6/22/2015       In progress     Yes
1427219 Debt collection         Disclosure verification of debt Not given enough info to verify debt    FL      32207   Phone   6/18/2015       6/22/2015       In progress     Yes
1427377 Money transfers International money transfer    Money was not available when promised                   16      Phone   6/18/2015       6/22/2015       In progress     Yes
1427685 Debt collection Other (phone, health club, etc.)        Cont'd attempts collect debt not owed   Debt was paid   WA      98501   Referral        6/18/2015       6/23/2015       Closed with non-monetary re
lief    Yes
1428423 Mortgage        Other mortgage  Loan modification,collection,foreclosure                FL      33025   Referral        6/18/2015       6/23/2015       Closed with explanation Yes
1426655 Consumer loan   Vehicle loan    Managing the loan or lease              SC      29063   Phone   6/18/2015       6/22/2015       Closed with explanation Yes
1424841 Consumer loan   Installment loan        Managing the loan or lease              AR      72112   Phone   6/17/2015       6/19/2015       In progress     Yes
1425040 Consumer loan   Installment loan        Taking out the loan or lease            OH      44112   Postal mail     6/17/2015       6/19/2015       Closed with explanation Yes
1424991 Debt collection Other (phone, health club, etc.)        Communication tactics   Frequent or repeated calls      AZ      85712   Postal mail     6/17/2015       6/19/2015       Closed with explanation Yes

1425268 Debt collection         Cont'd attempts collect debt not owed   Debt is not mine        OK      74115   Phone   6/17/2015       6/22/2015       Closed with explanation Yes
1425132 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    TX      77521   Phone   6/17/2015       6/19/2015       In progress     Yes
1425315 Debt collection         Cont'd attempts collect debt not owed   Debt is not mine        NM      87114   Phone   6/17/2015       6/23/2015       Closed with explanation Yes
1425515 Debt collection Payday loan     Communication tactics   Threatened to take legal action TX      77586   Phone   6/17/2015       6/19/2015       Closed with explanation Yes
1425477 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    AL      35061   Phone   6/17/2015       6/19/2015       Closed with explanation Yes
```

The following image extracts display the number of rows written to the *web_results* and -*other_results* tables respectively. Note that the header row containing field names was also copied across into the latter table, which meant that the count was out by one.

Home / user / student / DIHPS / output / web_results_count / **part-r-00000**

```
254550
```

Home / user / student / DIHPS / output / other_results_count / **part-r-00000**

```
154851
```

Question 7

The following image extract from the Pig query editor in Hue presents the code used to create a table called *max_complaints* that listed the 10 states with the maximum number of complaints.

Question 7 - Top 10 states with the maximum number of complaints

```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define schema to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,
sub_issue:chararray,
state:chararray,
zip_code:chararray,
submitted_via:chararray,
date_received:chararray,
date_sent_to_company:chararray,
company_response:chararray,
timely_response:chararray,
consumer_disputed:chararray
);

-- group rows of file by state field
T0 = GROUP T BY state;
-- count the number of rows of data in each group
T_GRP = FOREACH T0 GENERATE group, COUNT(T) AS state_count;
-- order the groups by the state counts in descending order
T_GRP2 = ORDER T_GRP BY state_count DESC;
-- limit the number of rows of output to 10
T_LIM = LIMIT T_GRP2 10;

-- store the results in a table called 'max_complaints' in the HDFS DIHPS folder
STORE T_LIM INTO '/user/student/DIHPS/output/max_complaints';
```

The output indicates that the 10 states with the most complaints were (in descending order) California, Florida, Texas, New York, Georgia, New Jersey, Pennsylvania, Illinois, Virginia and Maryland.

Home / user / student / DIHPS / output / max_complaints / **part-r-00000**

CA	60528
FL	39381
TX	29817
NY	27974
GA	17705
NJ	16588
PA	14637
IL	14332
VA	13136
MD	13136

Question 8

The following image extract from the Pig query editor in Hue presents the code used to create a table called *medical_complaints_list* that contained all complaints associated with the “Medical” sub-product offering. The same script was used to create a table called *medical_complaints_total* that calculated and output the total number of consumer complaints associated with the “Medical” sub-product offering.

Question 8 - Complaints associated with the 'Medical' sub-product offering

```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define schema to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,
sub_issue:chararray,
state:chararray,
zip_code:chararray,
submitted_via:chararray,
date_received:chararray,
date_sent_to_company:chararray,
company_response:chararray,
timely_response:chararray,
consumer_disputed:chararray
);

-- filter rows of file to include only those where sub-product = 'Medical'
TO = FILTER T BY sub_product == 'Medical';

-- store the results in a table called 'medical_complaints_list' in the HDFS DIHPS output folder
STORE TO INTO '/user/student/DIHPS/output/medical_complaints_list';

-- group the results together
T_GRP = GROUP TO ALL;

-- count the number of rows in the file
T_CNT = FOREACH T_GRP GENERATE COUNT(TO) AS num_complaints;

-- store the result in a table called 'medical_complaints_total' in the HDFS DIHPS output folder
STORE T_CNT INTO '/user/student/DIHPS/output/medical_complaints_total';
```

This image displays an extract of the output from the new *medical_complaints_list* table.

Home / user / student / DIHPS / output / medical_complaints_list / **part-m-00000**

Page 1 to 1 of 315

```
1431374 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    WI      54140   Web     6/22/2015       6/22/2015       Closed with explanation Yes
1431743 Debt collection Medical Cont'd attempts collect debt not owed   Debt is not mine        WA      98055   Web     6/22/2015       6/22/2015       Closed with explanation Yes     Yes
1432678 Debt collection Medical Cont'd attempts collect debt not owed   Debt was paid   TX      75104   Web     6/22/2015       6/22/2015       Closed with explanation Yes
1432207 Debt collection Medical Disclosure verification of debt Right to dispute notice not received    TX      75125   Web     6/22/2015       6/22/2015       Closed with explanation Yes
1428833 Debt collection Medical Improper contact or sharing of info     Talked to a third party about my debt   GA      30144   Web     6/20/2015       6/20/2015       Closed with non-monetary relief Yes
1428990 Debt collection Medical Cont'd attempts collect debt not owed   Debt was paid   AZ      85035   Web     6/19/2015       6/19/2015       Closed with explanation Yes
1429538 Debt collection Medical Cont'd attempts collect debt not owed   Debt is not mine        IL      61103   Web     6/19/2015       6/19/2015       Closed with explanation Yes
1429346 Debt collection Medical Cont'd attempts collect debt not owed   Debt was paid   NJ      7042    Web     6/19/2015       6/19/2015       Closed with explanation Yes
1429466 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    NJ      8045    Web     6/19/2015       6/19/2015       Closed with explanation Yes
1429723 Debt collection Medical Disclosure verification of debt Not disclosed as an attempt to collect  MD      21222   Web     6/19/2015       6/19/2015       In progress     Yes
1430103 Debt collection Medical Communication tactics   Frequent or repeated calls      NC      28056   Web     6/19/2015       6/19/2015       Closed with non-monetary relief Yes
1427578 Debt collection Medical False statements or representation      Indicated committed crime not paying    NJ      8882    Web     6/18/2015       6/18/2015       In progress     Yes
1427153 Debt collection Medical Communication tactics   Frequent or repeated calls      MA      2601    Web     6/18/2015       6/22/2015       In progress     Yes
1427131 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    TX      75248   Web     6/18/2015       6/18/2015       Closed with explanation Yes
1427121 Debt collection Medical Communication tactics   Used obscene/profane/abusive language   AZ      86505   Web     6/18/2015       6/18/2015       Closed with explanation Yes     Yes
1427407 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    FL      34984   Web     6/18/2015       6/18/2015       Closed with explanation Yes
1427151 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    KY      40069   Web     6/18/2015       6/18/2015       Closed with explanation Yes
1429162 Debt collection Medical Cont'd attempts collect debt not owed   Debt was paid   NY      10601   Web     6/18/2015       6/19/2015       In progress     Yes
1429020 Debt collection Medical Cont'd attempts collect debt not owed   Debt was paid   OH      44142   Web     6/18/2015       6/18/2015       Closed with explanation Yes
1424860 Debt collection Medical Improper contact or sharing of info     Contacted employer after asked not to   FL      33435   Web     6/17/2015       6/17/2015       Closed with explanation Yes
1424862 Debt collection Medical Cont'd attempts collect debt not owed   Debt is not mine        FL      33435   Web     6/17/2015       6/17/2015       Closed with explanation Yes
1425132 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    TX      77521   Phone   6/17/2015       6/19/2015       In progress     Yes
1425325 Debt collection Medical Cont'd attempts collect debt not owed   Debt was paid   PA      19382   Web     6/17/2015       6/17/2015       Closed with explanation Yes
1425741 Debt collection Medical Cont'd attempts collect debt not owed   Debt is not mine        WA      98506   Web     6/17/2015       6/17/2015       Closed with explanation Yes
1425734 Debt collection Medical Cont'd attempts collect debt not owed   Debt is not mine        AZ      85138   Web     6/17/2015       6/17/2015       Closed with explanation Yes     Yes
1425477 Debt collection Medical Disclosure verification of debt Not given enough info to verify debt    AL      35061   Phone   6/17/2015       6/19/2015       Closed with explanation Yes
1425680 Debt collection Medical Disclosure verification of debt Not given enough info to verify deb
```

The output from the second part of the script indicated that there were 8,271 complaints associated with the “Medical” sub-product offering.

Home / user / student / DIHPS / output / medical_complaints_total / **part-r-00000**

```
8271
```

Question 9

The following image extract from the Pig query editor in Hue presents the code used to create a table called *least_complaints* that listed five zip codes with the smallest number of complaints.

Question 9 - Five zip codes with the least number of complaints

```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define schema to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,
sub_issue:chararray,
state:chararray,
zip_code:chararray,
submitted_via:chararray,
date_received:chararray,
date_sent_to_company:chararray,
company_response:chararray,
timely_response:chararray,
consumer_disputed:chararray
);

-- group rows of file by zip code field
TO = GROUP T BY zip_code;
-- count the number of rows of data in each group
T_GRP = FOREACH TO GENERATE group, COUNT(T) AS zip_count;
-- order the groups by the zip code counts in ascending order
T_GRP2 = ORDER T_GRP BY zip_count ASC;
-- limit the number of rows of output to 5
T_LIM = LIMIT T_GRP2 5;

-- store the results in a table called 'least_complaints' in the HDFS DIHPS folder
STORE T_LIM INTO '/user/student/DIHPS/output/least_complaints';
```

This image displays the output from the new *least_complaints* table. Recall from question three that since there are 6,190 valid and unique zip codes with a single complaint, the output represents only one of many possible valid solutions.

Home / user / student / DIHPS / output / least_complaints / **part-r-00000**

```
61235	1
74818	1
26337	1
74821	1
74832	1
```

Question 10

The following image extract from the Pig query editor in Hue presents the code used to create a table called *id_theft_complaints* that listed all complaints associated with identity theft issues grouped by the product and state fields.

Question 10 - Complaints associated with identity theft

```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define schema to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,
sub_issue:chararray,
state:chararray,
zip_code:chararray,
submitted_via:chararray,
date_received:chararray,
date_sent_to_company:chararray,
company_response:chararray,
timely_response:chararray,
consumer_disputed:chararray
);

-- filter rows of file to include only those where the issue description contains the word 'identity theft'
T0 = FILTER T BY issue MATCHES '.*Identity theft.*';
-- group the result by the product and state fields
T_GRP = GROUP T0 BY (product, state);

-- store the result in a table called 'id_theft_complaints' in the HDFS DIHPS folder
STORE T_GRP INTO '/user/student/DIHPS/output/id_theft_complaints';
```

This image displays an extract of the output from the new *id_theft_complaints* table.

Home / user / student / DIHPS / output / id_theft_complaints / **part-r-00000**

Page 1 of 133

```
(Credit card,AA)        {(418844,Credit card,,Identity theft / Fraud / Embezzlement,,AA,34041,Web,5/29/2013,5/30/2013,Closed with explanation,Yes,No)}
(Credit card,AE)        {(649225,Credit card,,Identity theft / Fraud / Embezzlement,,AE,9630,Web,12/30/2013,1/7/2014,Closed with explanation,Yes,No)}
(Credit card,AK)        {(83733,Credit card,,Identity theft / Fraud / Embezzlement,,AK,99615,Referral,5/18/2012,6/6/2012,Closed with explanation,Yes,Yes),(1016168,Credit card,,Identity theft / Fraud / Embezzleme
nt,,AK,72335,Phone,9/5/2014,9/10/2014,Closed with explanation,Yes,No),(579248,Credit card,,Identity theft / Fraud / Embezzlement,,AK,99611,Web,10/31/2013,11/4/2013,Closed with explanation,Yes,No),(1338172,Credit
card,,Identity theft / Fraud / Embezzlement,,AK,99835,Referral,4/20/2015,4/23/2015,Closed,Yes,),(485121,Credit card,,Identity theft / Fraud / Embezzlement,,AK,99654,Phone,8/8/2013,8/12/2013,Closed with explanati
on,Yes,No),(558552,Credit card,,Identity theft / Fraud / Embezzlement,,AK,99708,Web,10/14/2013,10/14/2013,Closed with explanation,Yes,No),(363801,Credit card,,Identity theft / Fraud / Embezzlement,,AK,99801,Phon
e,3/22/2013,3/26/2013,Closed with non-monetary relief,Yes,No)}
(Credit card,AL)        {(633961,Credit card,,Identity theft / Fraud / Embezzlement,,AL,35901,Web,12/13/2013,12/13/2013,Closed with explanation,Yes,No),(1403890,Credit card,,Identity theft / Fraud / Embezzlemen
t,,AL,36109,Web,6/2/2015,6/3/2015,Closed with explanation,Yes,),(365580,Credit card,,Identity theft / Fraud / Embezzlement,,AL,36117,Postal mail,3/26/2013,3/27/2013,Closed with explanation,Yes,No),(83465,Credit
card,,Identity theft / Fraud / Embezzlement,,AL,36503,Referral,2/2/2012,2/2/2012,Closed without relief,Yes,No),(393188,Credit card,,Identity theft / Fraud / Embezzlement,,AL,36265,Referral,4/26/2013,4/29/2013,Cl
osed with monetary relief,Yes,No),(685721,Credit card,,Identity theft / Fraud / Embezzlement,,AL,35748,Web,1/25/2014,2/9/2014,Closed with non-monetary relief,Yes,No),(74966,Credit card,,Identity theft / Fraud /
Embezzlement,,AL,35645,Web,5/10/2012,5/11/2012,Closed with relief,Yes,No),(105089,Credit card,,Identity theft / Fraud / Embezzlement,,AL,35603,Web,6/20/2012,6/21/2012,Closed with monetary relief,Yes,No),(391745,
Credit card,,Identity theft / Fraud / Embezzlement,,AL,36801,Referral,4/25/2013,4/26/2013,Closed with explanation,Yes,No),(1041451,Credit card,,Identity theft / Fraud / Embezzlement,,AL,36344,Web,9/23/2014,9/23/
2014,Closed with monetary relief,Yes,No),(861508,Credit card,,Identity theft / Fraud / Embezzlement,,AL,36605,Web,5/20/2014,5/20/2014,Closed with non-monetary relief,Yes,No),(987915,Credit card,,Identity theft /
Fraud / Embezzlement,,AL,35504,Web,8/17/2014,8/20/2014,Closed with explanation,Yes,No),(861509,Credit card,,Identity theft / Fraud / Embezzlement,,AL,36605,Web,5/20/2014,5/20/2014,Closed with non-monetary relie
f,Yes,No),(1058821,Credit card,,Identity theft / Fraud / Embezzlement,,AL,35226,Web,10/6/2014,10/6/2014,Closed with explanation,Yes,Yes),(1027172,Credit card,,Identity theft / Fraud / Embezzlement,,AL,36093,Refe
rral,9/12/2014,9/17/2014,Closed with explanation,Yes,Yes),(1395134,Credit card,,Identity theft / Fraud / Embezzlement,,AL,35173,Web,5/28/2015,6/1/2015,Closed with explanation,Yes,Yes),(83461,Credit card,,Identit
y theft / Fraud / Embezzlement,,AL,36503,Postal mail,1/10/2012,1/13/2012,Closed with relief,Yes,No),(926921,Credit card,,Identity theft / Fraud / Embezzlement,,AL,36867,Phone,7/8/2014,7/11/2014,Closed with expla
nation,Yes,No),(1258438,Credit card,,Identity theft / Fraud / Embezzlement,,AL,35071,Web,2/26/2015,2/26/2015,Closed with monetary relief,Yes,No),(1301200,Credit card,,Identity theft / Fraud / Embezzlement,,AL,35
988,Web,3/25/2015,3/25/2015,Closed with explanation,Yes,No),(9082,Credit card,,Identity theft / Fraud / Embezzlement,,AL,35640,Web,1/2/2012,1/3/2012,Closed with relief,Yes,No),(834914,Credit card,,Identity theft
/ Fraud / Embezzlement,,AL,35244,Postal mail,4/30/2014,5/6/2014,Closed with non-monetary relief,Yes,No),(611755,Credit card,,Identity theft / Fraud / Embezzlement,,AL,36775
```

Question 11

The following image extract from the Pig query editor in Hue presents the code used to create a table called *complaint_summary* that summarised the complaints by product, sub-product and state.

Question 11 - Number of complaints grouped by product, sub-product and state

```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define schema to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,
sub_issue:chararray,
state:chararray,
zip_code:chararray,
submitted_via:chararray,
date_received:chararray,
date_sent_to_company:chararray,
company_response:chararray,
timely_response:chararray,
consumer_disputed:chararray
);

-- group the rows by the product, sub-product and state fields
T_GRP = GROUP T BY (product, sub_product, state);

-- flatten each group and count the number of rows
T_SUMMARY = FOREACH T_GRP GENERATE FLATTEN(group), COUNT(T) AS num_complaints;

-- store the result in a table called 'complaint_summary' in the HDFS DIHPS folder
STORE T_SUMMARY INTO '/user/student/DIHPS/output/complaint_summary';
```

This image displays an extract of the output from the new *complaint_summary* table as there were too many possible combinations of the three fields to be able to display the entire table. It indicates, for example, that there were 51 complaints from Illanois associated with the VA mortgage sub-product offering and only two from Guam.

Home / user / student / DIHPS / output / complaint_summary / part-r-00000 Page 1 of 21

Product	Sub-product	State	1
Mortgage	VA mortgage	AA	1
Mortgage	VA mortgage	AE	13
Mortgage	VA mortgage	AK	10
Mortgage	VA mortgage	AL	28
Mortgage	VA mortgage	AP	4
Mortgage	VA mortgage	AR	20
Mortgage	VA mortgage	AZ	80
Mortgage	VA mortgage	CA	225
Mortgage	VA mortgage	CO	59
Mortgage	VA mortgage	CT	13
Mortgage	VA mortgage	DC	4
Mortgage	VA mortgage	DE	16
Mortgage	VA mortgage	FL	263
Mortgage	VA mortgage	FM	1
Mortgage	VA mortgage	GA	172
Mortgage	VA mortgage	GU	2
Mortgage	VA mortgage	HI	15
Mortgage	VA mortgage	IA	20
Mortgage	VA mortgage	ID	10
Mortgage	VA mortgage	IL	51
Mortgage	VA mortgage	IN	31
Mortgage	VA mortgage	KS	14
Mortgage	VA mortgage	KY	39
Mortgage	VA mortgage	LA	28
Mortgage	VA mortgage	MA	24
Mortgage	VA mortgage	MD	107
Mortgage	VA mortgage	ME	14
Mortgage	VA mortgage	MI	45

Total word count: 1,640 words