Assessment 4: Managing data with Hive and Pig in Hadoop

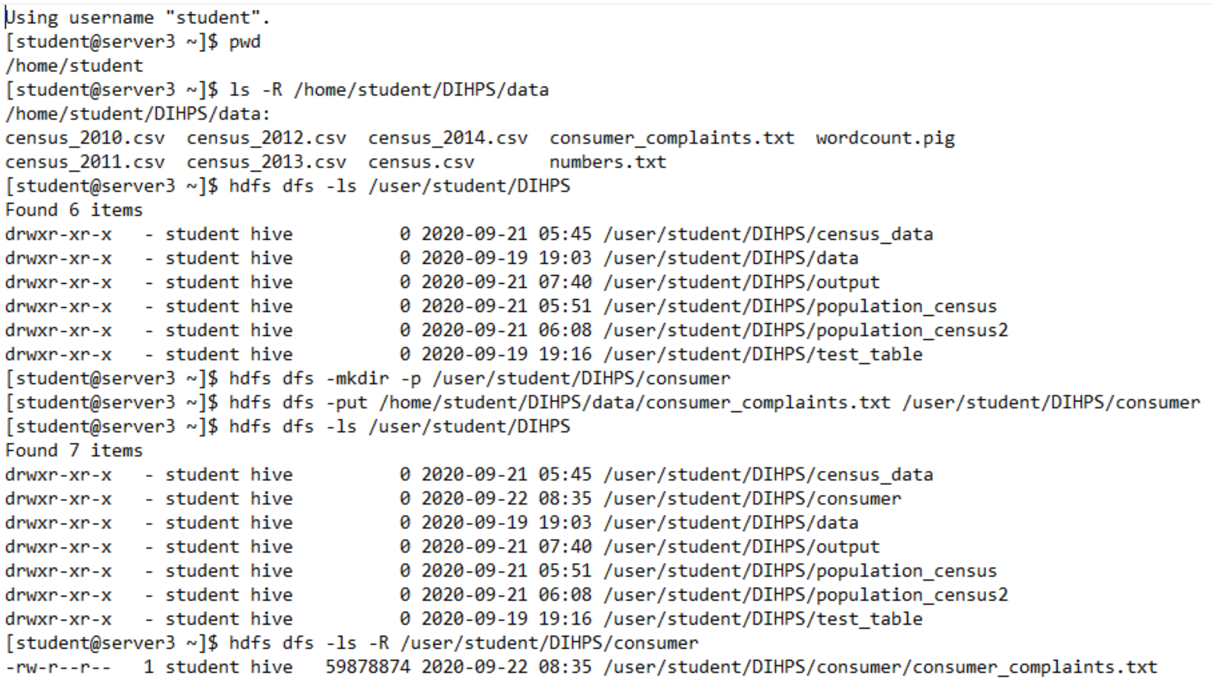
MA5831 – Advanced Data Processing and Analysis using SAS

13848336 Nikki Fitzherbert

Technical Preparation and Setup

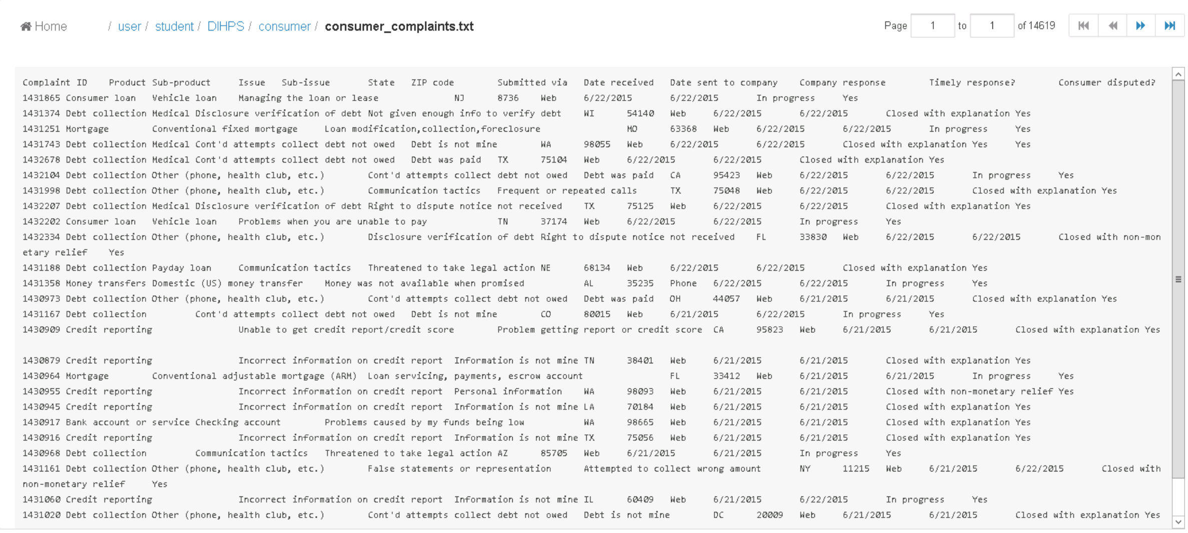
A couple of preparatory steps were required before any querying of the consumer complaints data could be performed. These are outlined briefly in this and the next section.

The following code extract from the MRemoteNG session shows how it was verified that the required data file (*consumer_complaints.txt*) did in fact exist, how a new folder called *consumer* was created in the *DIHPS* database in the Hadoop Distributed File System (HDFS) and how a copy of the data file was moved to the new folder.



```
Using username "student".
[student@server3 ~]$ pwd
/home/student
[student@server3 ~]$ ls -R /home/student/DIHPS/data
/home/student/DIHPS/data:
census_2010.csv  census_2012.csv  census_2014.csv  consumer_complaints.txt  wordcount.pig
census_2011.csv  census_2013.csv  census.csv       numbers.txt
[student@server3 ~]$ hdfs dfs -ls /user/student/DIHPS
Found 6 items
drwxr-xr-x   - student hive          0 2020-09-21 05:45 /user/student/DIHPS/census_data
drwxr-xr-x   - student hive          0 2020-09-19 19:03 /user/student/DIHPS/data
drwxr-xr-x   - student hive          0 2020-09-21 07:40 /user/student/DIHPS/output
drwxr-xr-x   - student hive          0 2020-09-21 05:51 /user/student/DIHPS/population_census
drwxr-xr-x   - student hive          0 2020-09-21 06:08 /user/student/DIHPS/population_census2
drwxr-xr-x   - student hive          0 2020-09-19 19:16 /user/student/DIHPS/test_table
[student@server3 ~]$ hdfs dfs -mkdir -p /user/student/DIHPS/consumer
[student@server3 ~]$ hdfs dfs -put /home/student/DIHPS/data/consumer_complaints.txt /user/student/DIHPS/consumer
[student@server3 ~]$ hdfs dfs -ls /user/student/DIHPS
Found 7 items
drwxr-xr-x   - student hive          0 2020-09-21 05:45 /user/student/DIHPS/census_data
drwxr-xr-x   - student hive          0 2020-09-22 08:35 /user/student/DIHPS/consumer
drwxr-xr-x   - student hive          0 2020-09-19 19:03 /user/student/DIHPS/data
drwxr-xr-x   - student hive          0 2020-09-21 07:40 /user/student/DIHPS/output
drwxr-xr-x   - student hive          0 2020-09-21 05:51 /user/student/DIHPS/population_census
drwxr-xr-x   - student hive          0 2020-09-21 06:08 /user/student/DIHPS/population_census2
drwxr-xr-x   - student hive          0 2020-09-19 19:16 /user/student/DIHPS/test_table
[student@server3 ~]$ hdfs dfs -ls -R /user/student/DIHPS/consumer
-rw-r--r--   1 student hive   59878874 2020-09-22 08:35 /user/student/DIHPS/consumer/consumer_complaints.txt
```

This extract from Hue shows that a copy of the *consumer complaints* data was successfully transferred into the HDFS.

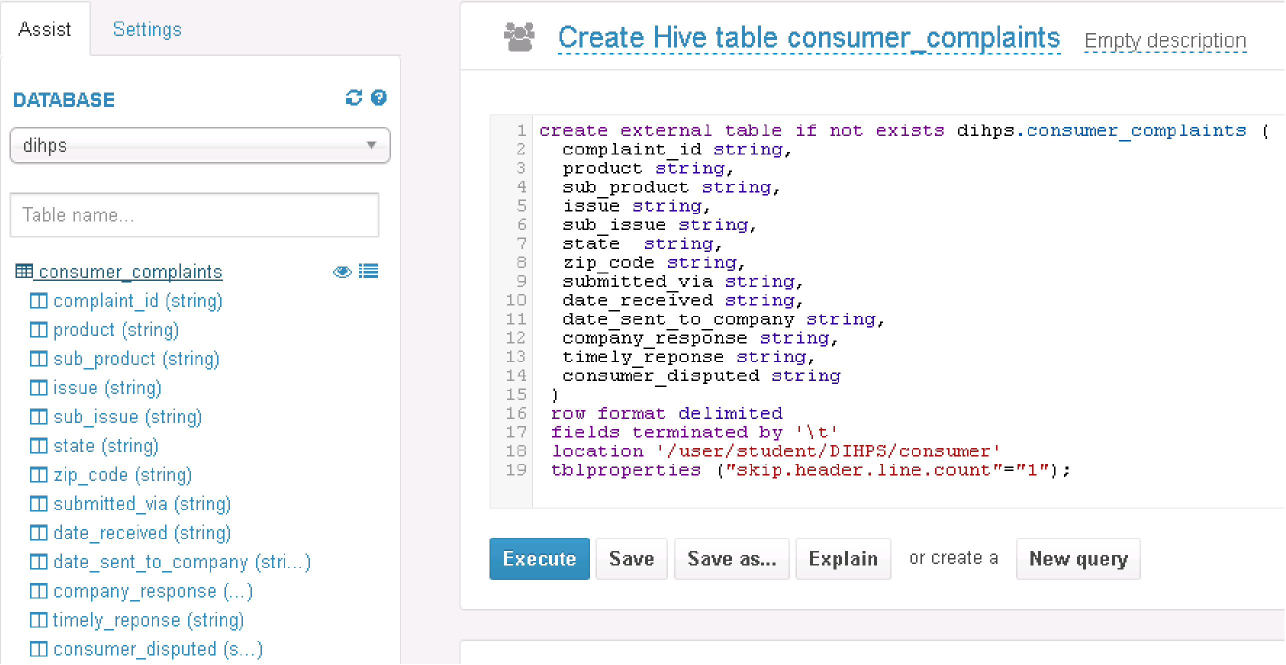

Home / user / student / DIHPS / consumer / consumer_complaints.txt
Page 1 to 1 of 14619

```
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?
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
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
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
```

Data Processing Tasks

The final data preparation step was to create a Hive table schema appropriate to support any subsequent query operations using HiveQL. Therefore, an external table called *consumer_complaints* was created using the ‘create table’ command in Hue, with the location of the underlying data indicated by the ‘location’ property, and the format of the stored data indicated by the ‘row format’ and ‘fields terminated by’ properties. The two advantages of creating an external table rather than a managed table were it was more flexible in that other tools could also query the data, and the source data remained intact and untouched.

Note that for simplicity, all fields were read in as strings and the ‘tblproperties’ option was used to skip the first row of the underlying data file as it contained the field names.

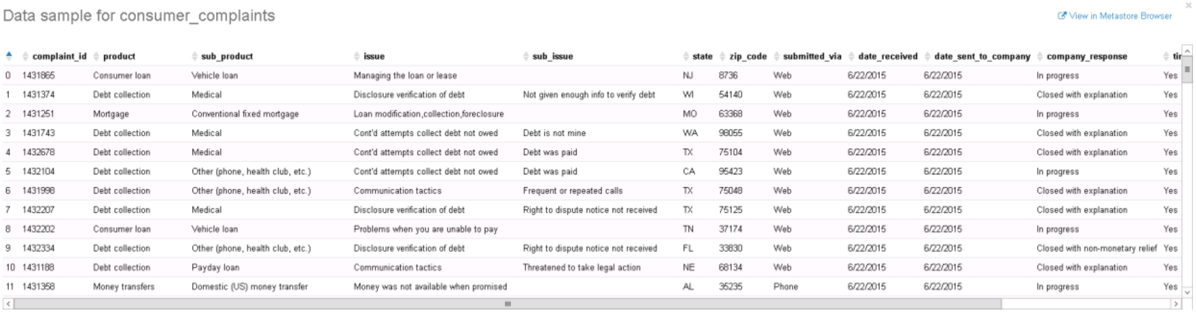


Create Hive table consumer_complaints Empty description

consumer_complaints
- complaint_id (string)
- product (string)
- sub_product (string)
- issue (string)
- sub_issue (string)
- state (string)
- zip_code (string)
- submitted_via (string)
- date_received (string)
- date_sent_to_company (stri...)
- company_response (...)
- timely_reponse (string)
- consumer_disputed (s...)

```
create external table if not exists dihps.consumer_complaints (
  complaint_id string,
  product string,
  sub_product string,
  issue string,
  sub_issue string,
  state  string,
  zip_code string,
  submitted_via string,
  date_received string,
  date_sent_to_company string,
  company_response string,
  timely_reponse string,
  consumer_disputed string
 )
 row format delimited
 fields terminated by '\t'
 location '/user/student/DIHPS/consumer'
 tblproperties ("skip.header.line.count"="1");
```

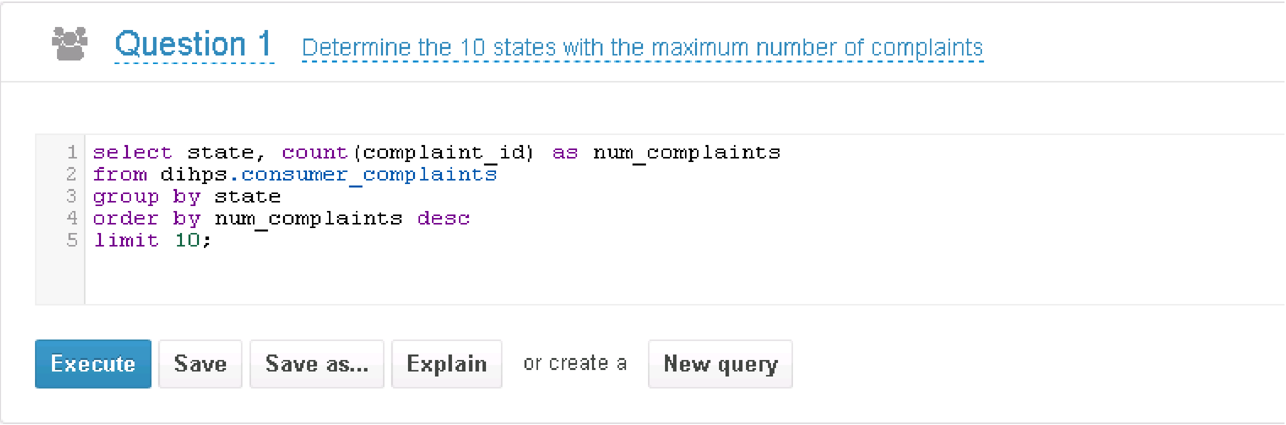
The following extract shows that the Hive table was successfully created and populated with the expected consumer complaints data.



Data sample for consumer_complaints

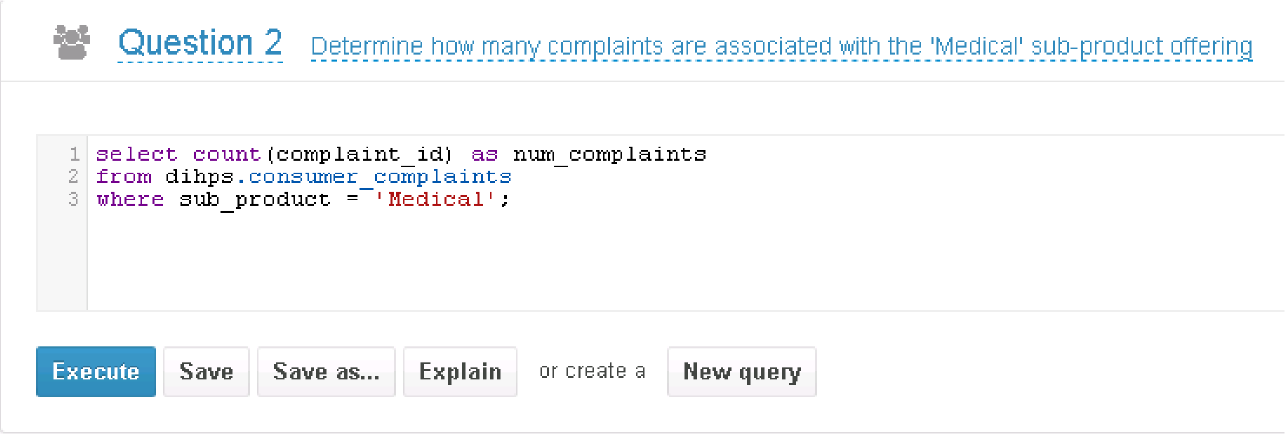
	complaint_id	product	sub_product	issue	sub_issue	state	zip_code	submitted_via	date_received	date_sent_to_company	company_response	ti
0	1431865	Consumer loan	Vehicle loan	Managing the loan or lease		NJ	8736	Web	6/22/2015	6/22/2015	In progress	Yes
1	1431374	Debt collection	Medical	Disclosure verification of debt	Not given enough info to verify debt	WI	54140	Web	6/22/2015	6/22/2015	Closed with explanation	Yes
2	1431251	Mortgage	Conventional fixed mortgage	Loan modification,collection,foreclosure		MO	63368	Web	6/22/2015	6/22/2015	In progress	Yes
3	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
4	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
5	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
6	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
7	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
8	1432202	Consumer loan	Vehicle loan	Problems when you are unable to pay		TN	37174	Web	6/22/2015	6/22/2015	In progress	Yes
9	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-monetary relief	Yes
10	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
11	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

Reporting and Coding Tasks – Hive Queries



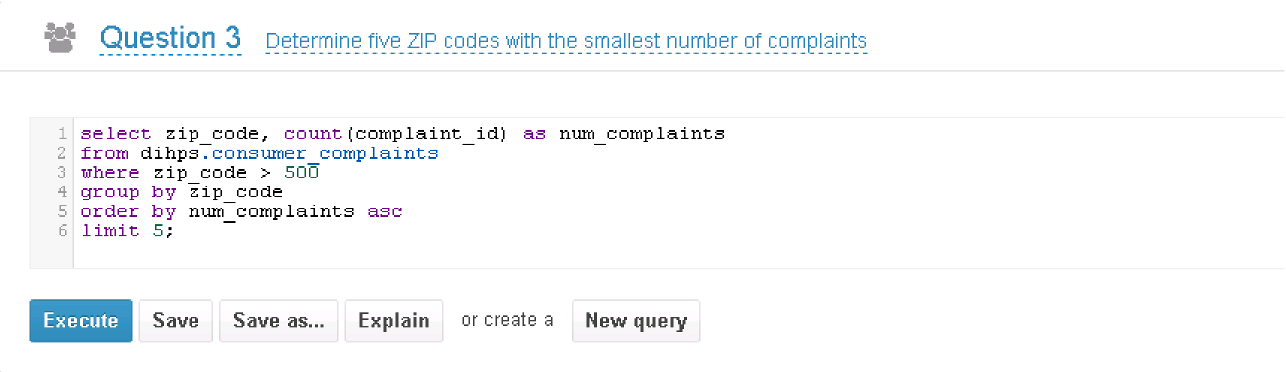
Question 1 Determine the 10 states with the maximum number of complaints

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



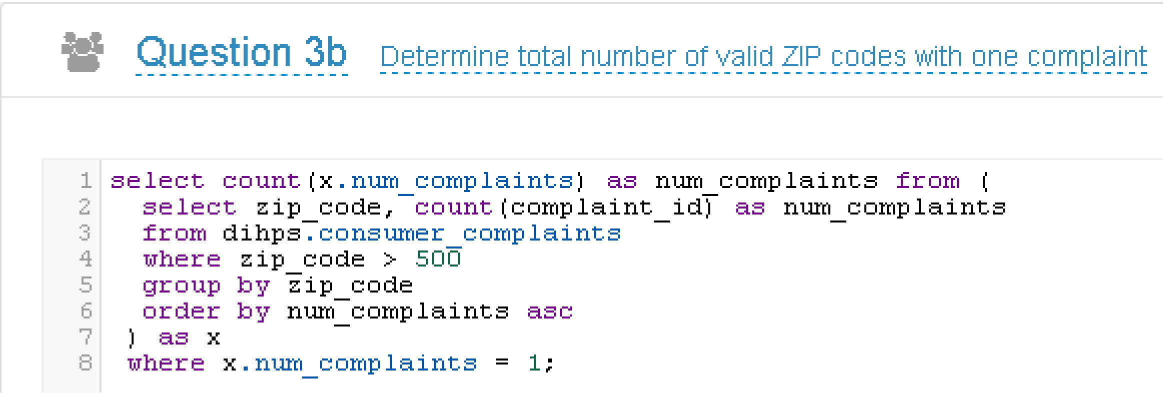
Question 2 Determine how many complaints are associated with the 'Medical' sub-product offering

```sql
select count(complaint_id) as num_complaints
from dihps.consumer_complaints
where sub_product = 'Medical';
```



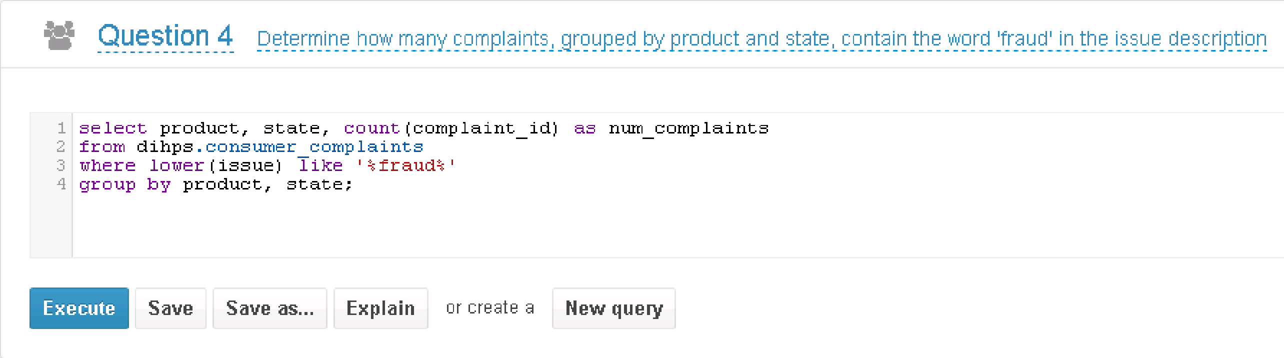
Question 3 Determine five ZIP codes with the smallest number of complaints

```sql
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;
```



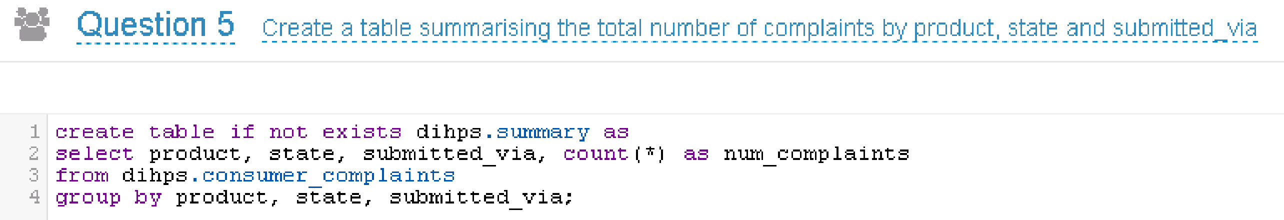
Question 3b Determine total number of valid ZIP codes with one complaint

```sql
select count(x.num_complaints) as num_complaints from (
   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
 ) as x
 where x.num_complaints = 1;
```



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

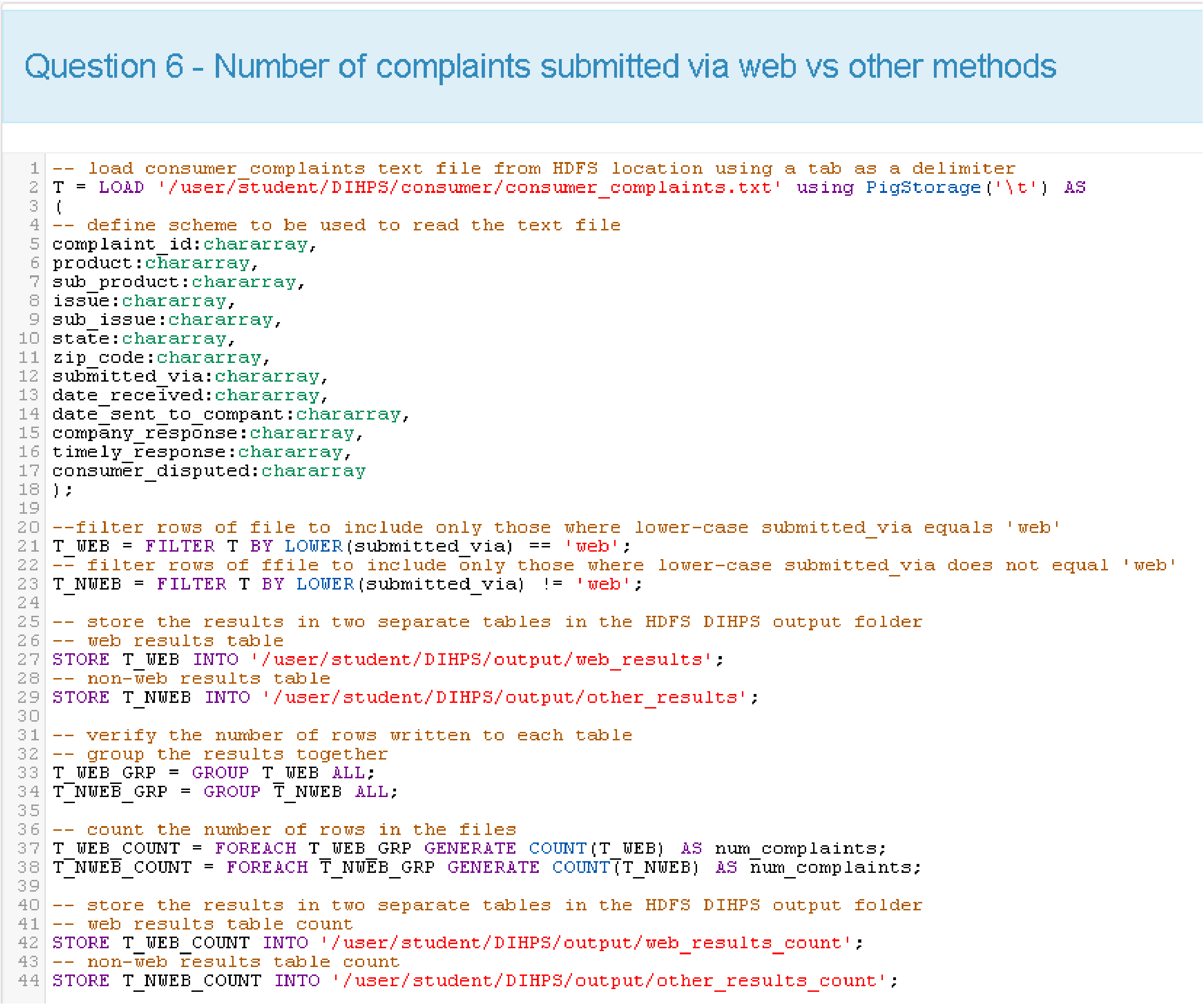
```sql
select product, state, count(complaint_id) as num_complaints
from dihps.consumer_complaints
where lower(issue) like '%fraud%'
group by product, state;
```



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

```sql
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;
```

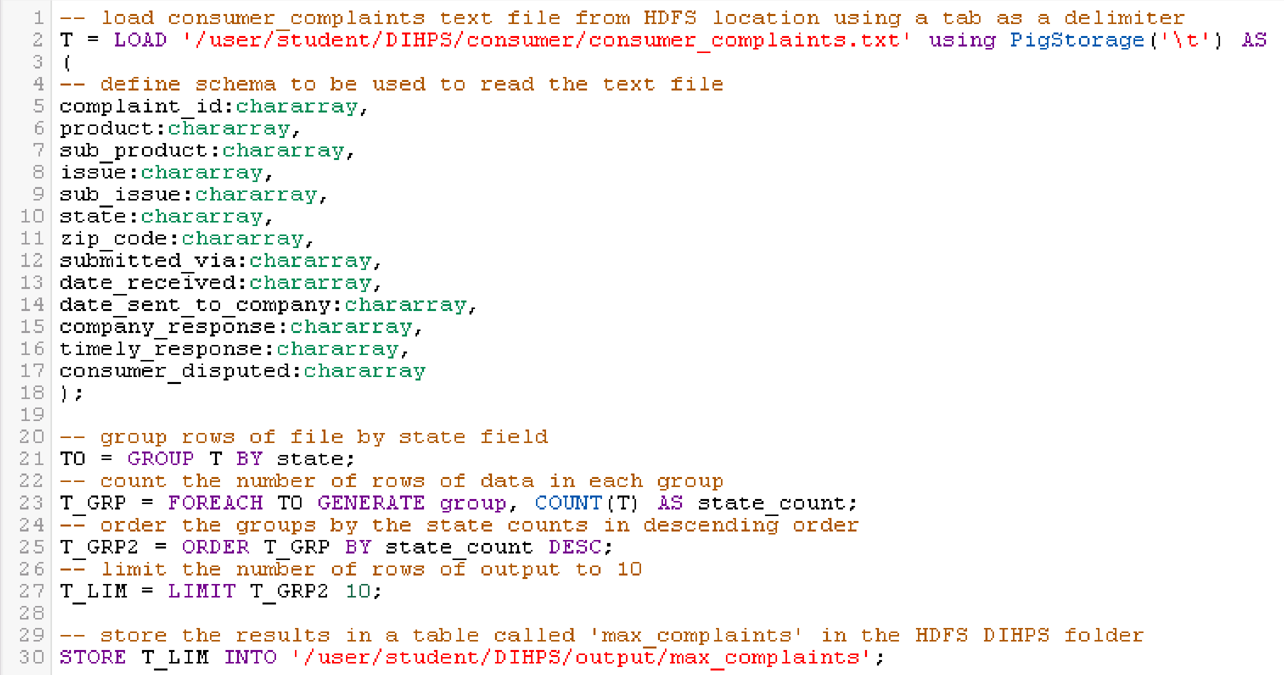
Reporting and Coding Tasks – Pig Latin scripts



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

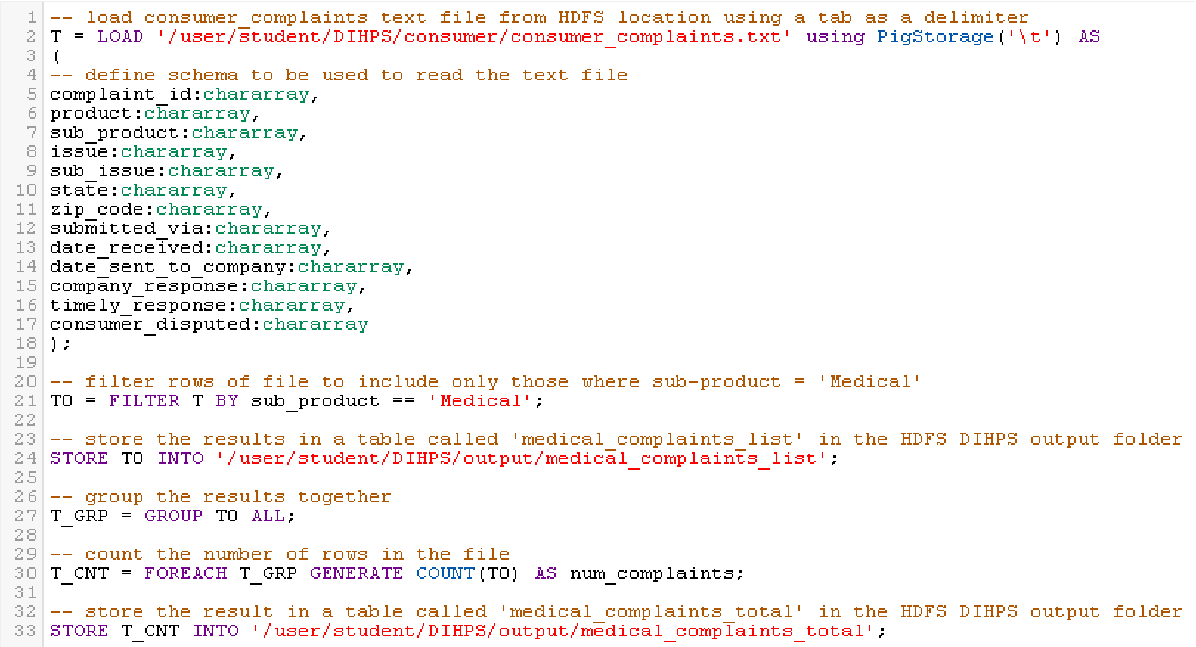
```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define scheme to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,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

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



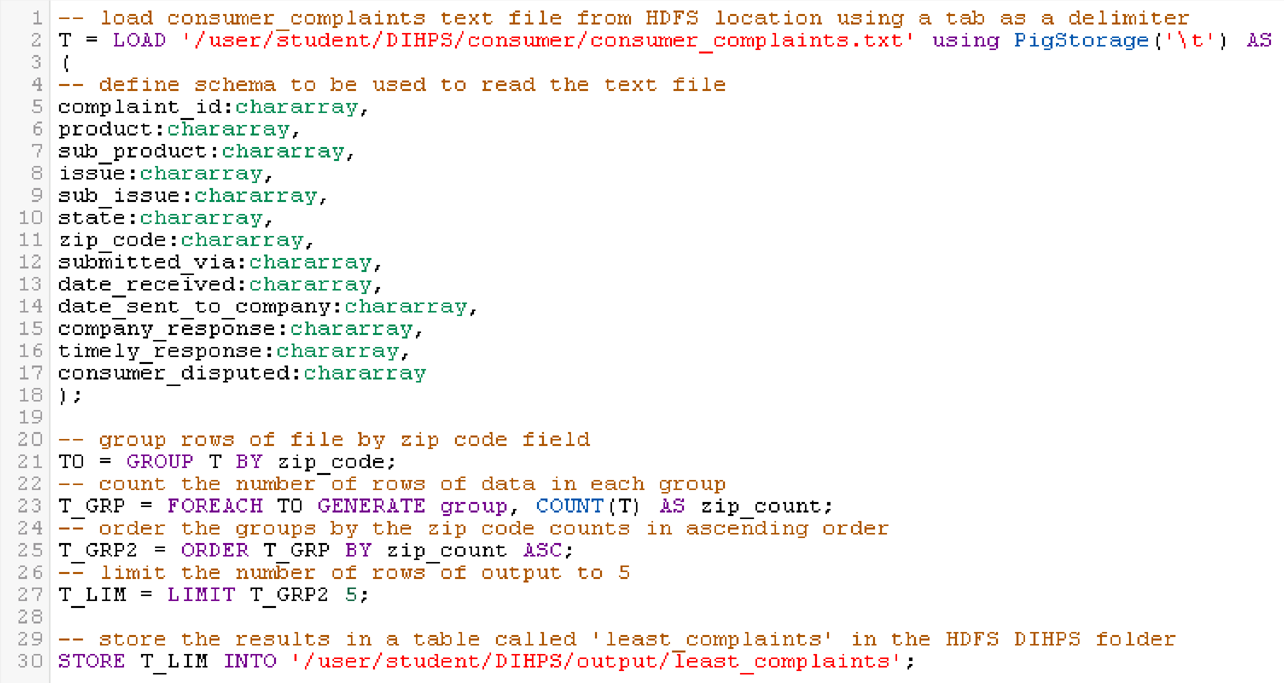
```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define schema to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,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

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



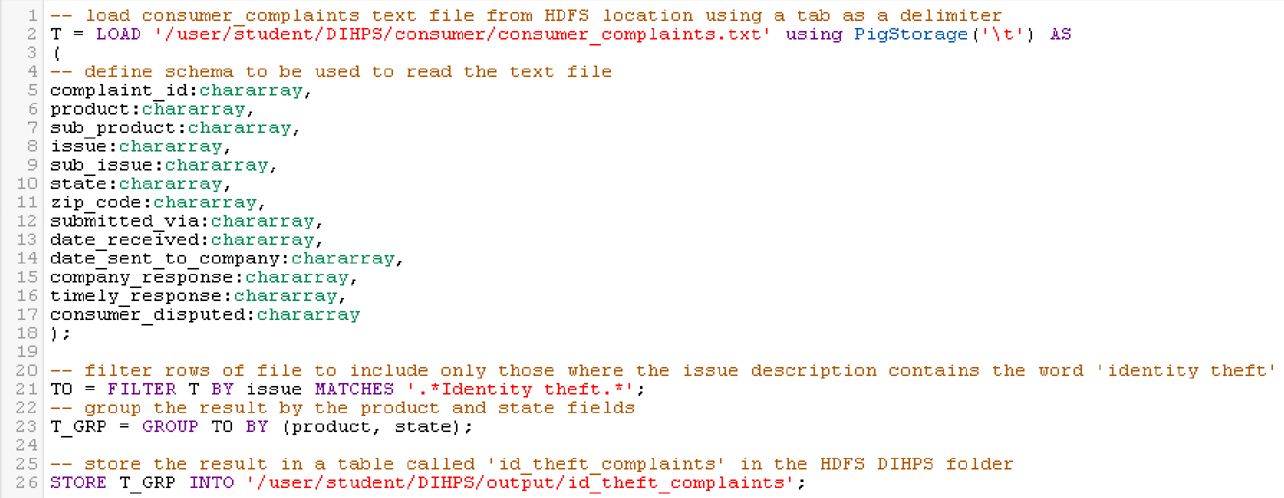
```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define schema to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,
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'
T0 = FILTER T BY sub_product == 'Medical';

-- store the results in a table called 'medical_complaints_list' in the HDFS DIHPS output folder
STORE T0 INTO '/user/student/DIHPS/output/medical_complaints_list';

-- group the results together
T_GRP = GROUP T0 ALL;

-- count the number of rows in the file
T_CNT = FOREACH T_GRP GENERATE COUNT(T0) 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';
```

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



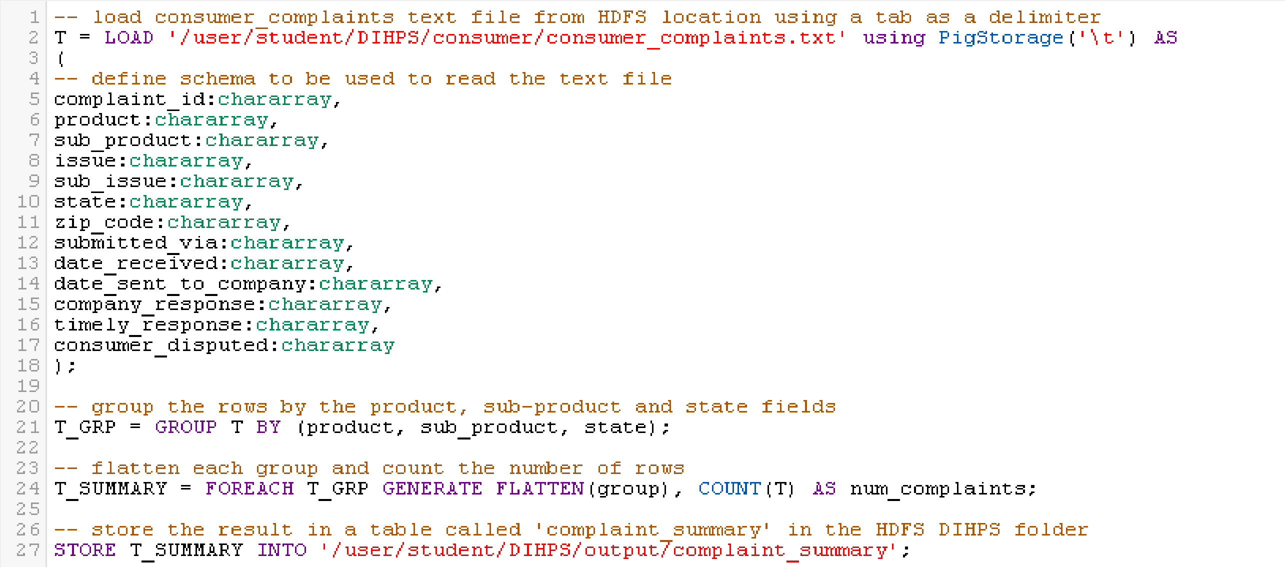
```
-- load consumer_complaints text file from HDFS location using a tab as a delimiter
T = LOAD '/user/student/DIHPS/consumer/consumer_complaints.txt' using PigStorage('\t') AS
(
-- define schema to be used to read the text file
complaint_id:chararray,
product:chararray,
sub_product:chararray,
issue:chararray,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

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'
TO = FILTER T BY issue MATCHES '.*Identity theft.*';
-- group the result by the product and state fields
T_GRP = GROUP TO 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';
```

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';
```