March / April 2016

Prepare, Secure, and Publish

by Mark Rittman

Automate ingesting, profiling, and transforming data so you can use it in big data cloud and other environments.

If you are working on an initiative involving big data, analytics, or data warehousing in your organization, you know that adding new datasources and making data suitable for use by analysts can be complicated, largely manual tasks. In the past, these tasks required technical skills and would have been carried out by the IT department. Now, with Oracle Big Data Preparation Cloud Service, available in Oracle Cloud, business users and data domain experts can perform these tasks by using just a web browser.

Oracle Big Data Preparation Cloud Service automates many of the tasks involved in ingesting, profiling, and transforming data, so you can use it in big data and other environments. Oracle Big Data Preparation Cloud Service uses machine learning and natural language processing to detect common data patterns and to alert you to sensitive data items (such as credit card numbers) so that you can easily obfuscate them. You can apply transformations and enrichments by using a point-and-click graphical interface. All you need to do then is apply your knowledge of the data domain you're working in to confirm the recommendations the tool makes and to add your own enhancements before publishing the data for use downstream in other applications.

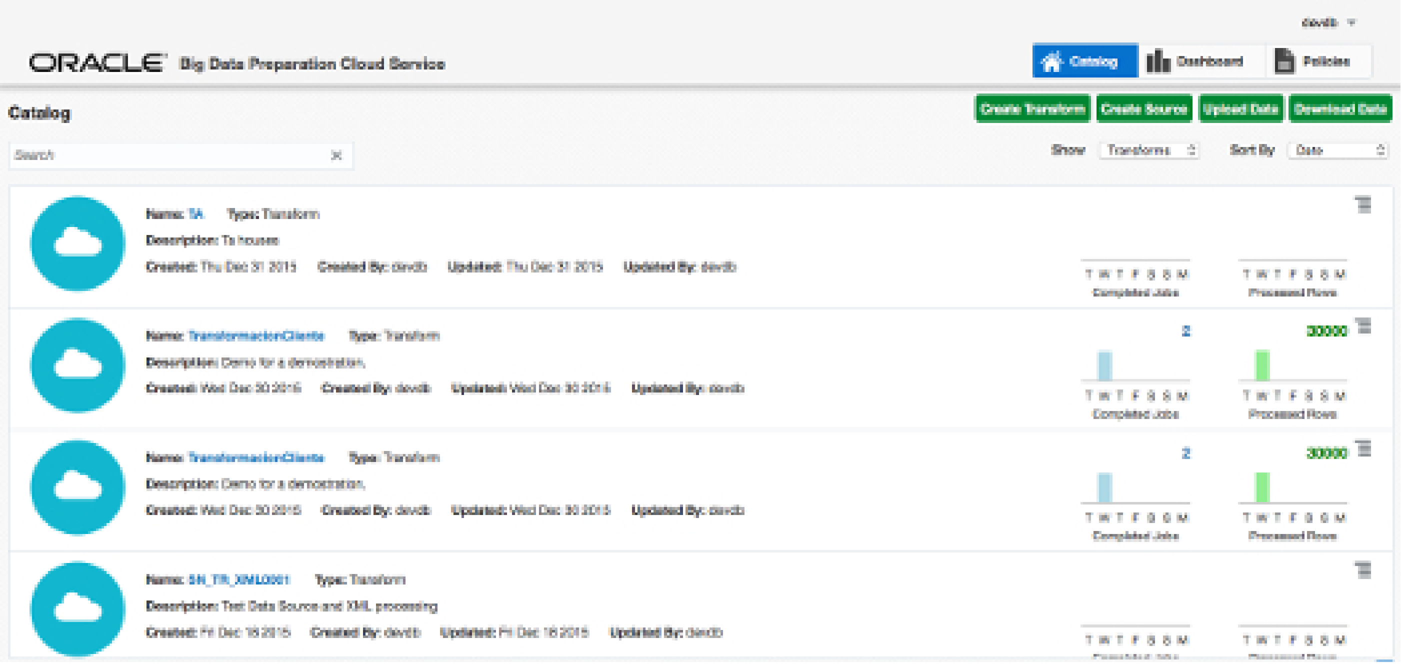
For the example in this article, I'll take a spreadsheet file of customer data that needs to be prepared for use before being made available for other users in an organization. As is often the case, the file comes with no information about the data it contains, and my job is to properly name the columns in the file, make sure all sensitive data is obfuscated, and add whatever enrichments I can to make the contents more useful downstream.

To follow along with the example, request a trial subscription for Oracle Big Data Preparation Cloud Service on Oracle Cloud and download the example file.

Uploading the File and Profiling Its Contents

Let's start by logging in to Oracle Big Data Preparation Cloud Service, uploading my data file, and then profiling the data so it is ready for transformation.

1. Using your web browser, navigate to the URL given to you for your instance of Oracle Big Data Preparation Cloud Service and log in with the username, password, and identity domain provided to you. After successful authentication, you will see the Catalog page, shown in **Figure 1**, which lists the transforms and sources created by you and others in this cloud environment.



Name: TA Type: Transform
Description: Ta houses
Created: Thu Dec 31 2015 Created By: devdb Updated: Thu Dec 31 2015 Updated By: devdb

Name: TransformacionCliente Type: Transform
Description: Demo for a demostration.
Created: Wed Dec 30 2015 Created By: devdb Updated: Wed Dec 30 2015 Updated By: devdb

Name: TransformacionCliente Type: Transform
Description: Demo for a demostration.
Created: Wed Dec 30 2015 Created By: devdb Updated: Wed Dec 30 2015 Updated By: devdb

Name: SN_TR_XML0001 Type: Transform
Description: Test Data Source and XML processing
Created: Fri Dec 18 2015 Created By: devdb Updated: Fri Dec 18 2015 Updated By: devdb

Figure 1: Catalog page

2. Unzip the o62ba-2867973.zip file. Click the **Upload Data** button to begin the process of uploading the unzipped spreadsheet file to Oracle Storage Cloud Service. When prompted, select **External Client** as the source type.

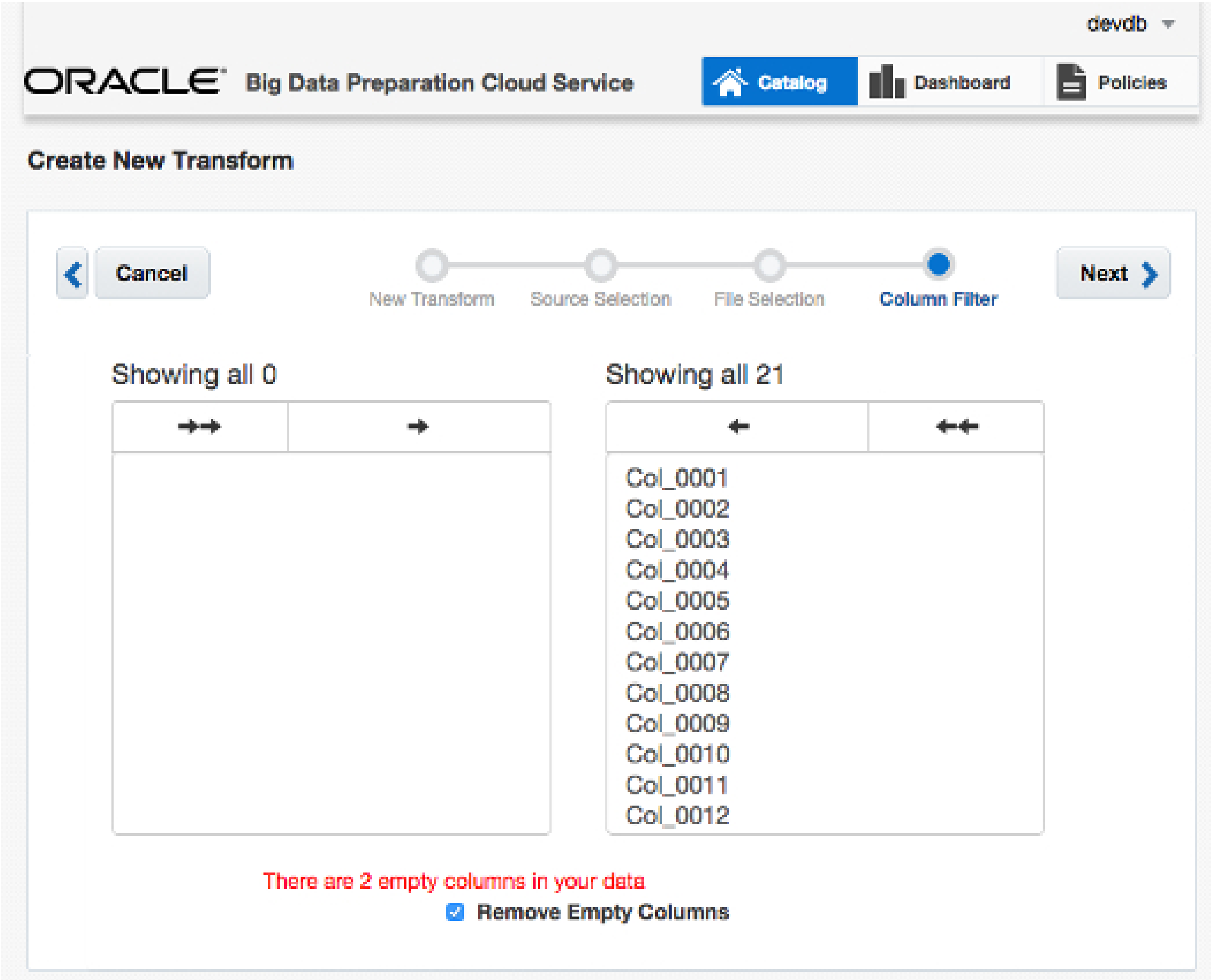
3. Use the **Browse** button to select the file to upload, and then in the Oracle Storage Cloud Service directory browser, select a directory, such as **/Data**, in which to save the file. When you are done, click the **Upload File** button to upload your spreadsheet file to the cloud environment.

4. Navigate back to the Catalog page in Oracle Big Data Preparation Cloud Service, and click the **Create Transform** button to start the process that first profiles the data and then enables you to transform the data. When prompted, enter the following details for the name and description of the transform, and then click **Next** to proceed.

New Transform Name : `Customer_Info_Transform`
Description : `Transformation to prepare customer file and remove sensitive details`

5. Now choose the Oracle Storage Cloud Service source you created earlier to store your data file, and then use the directory browser and file picker to locate and select the file you uploaded. Leave the **Contains Header** checkbox deselected and the **Smart Sample** checkbox selected, so that the profiling process will know that the first row in your file isn't column names and so it will use only a sample of the file's data to determine column contents rather than reading all the contents.
Click **Next** to proceed.

6. You will then be presented with the list of columns shown in **Figure 2**. Note that at this point, the column names are generic, default values (because in Step 5, you indicated that the file didn't contain a header row that could be used to provide column names). However, you can set the column names later (in most cases, automatically, based on recommendations the tool will provide).



Showing all 0

Showing all 21

Col_0001
Col_0002
Col_0003
Col_0004
Col_0005
Col_0006
Col_0007
Col_0008
Col_0009
Col_0010
Col_0011
Col_0012

There are 2 empty columns in your data

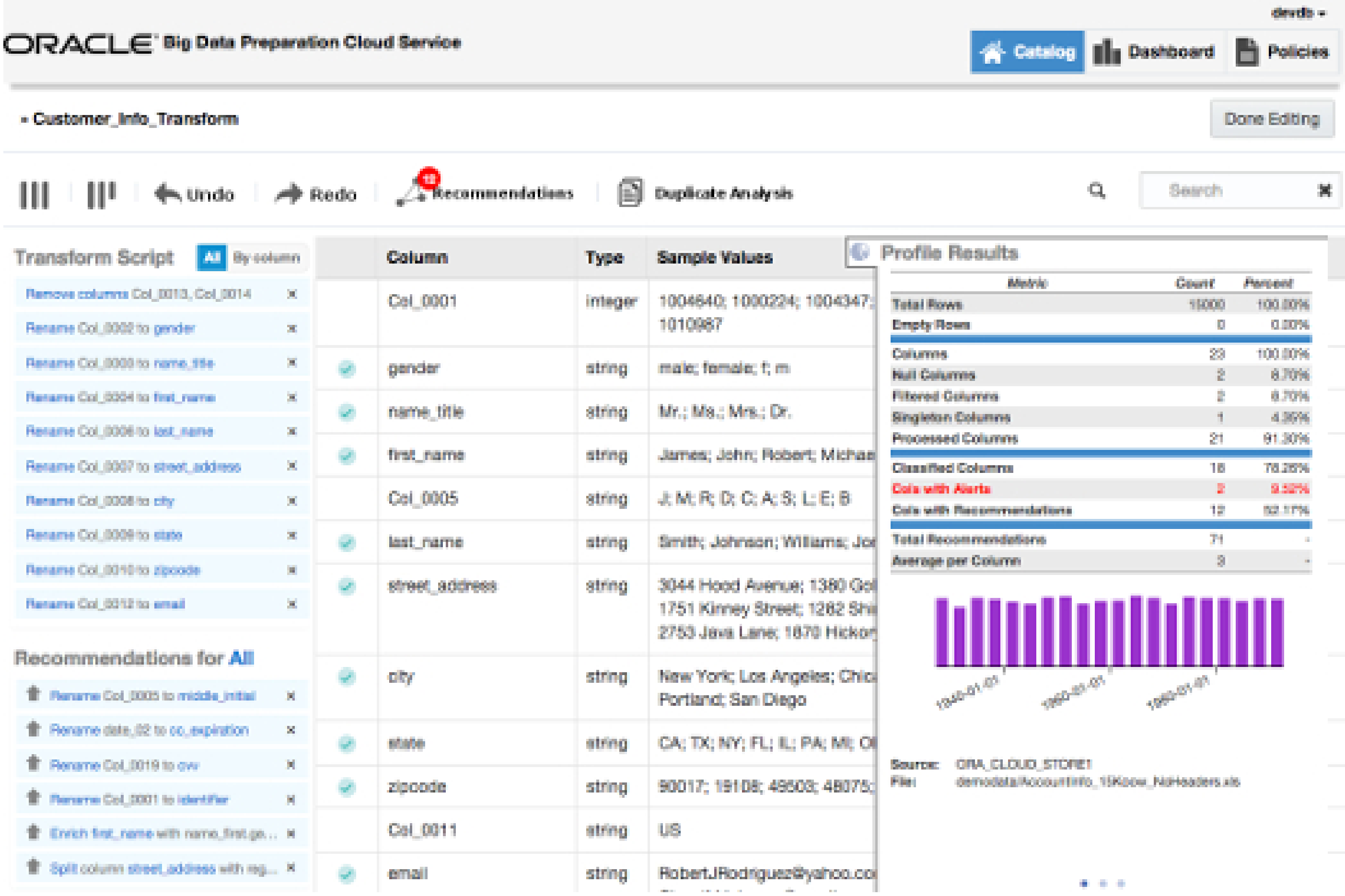
☑ Remove Empty Columns

Figure 2: Create New Transform Page

7. For now, leave the **Remove Empty Columns** checkbox selected so the next stage of the process will consider only columns that contain data. Then click Next to proceed to the main transform authoring page.

Reviewing, Transforming, and Obfuscating Data in Your File

Figure 3 shows the main transform authoring page you'll now use to review the data and metadata that was determined during the initial profiling process. Then you'll specify transformations and other "data wrangling" actions to apply to the data set before you make it available to users in your big data environment.



Transform Script — All | By column

- Remove columns Col_0013, Col_0014 ×
- Rename Col_0002 to gender ×
- Rename Col_0003 to name_title ×
- Rename Col_0004 to first_name ×
- Rename Col_0006 to last_name ×
- Rename Col_0007 to street_address ×
- Rename Col_0008 to city ×
- Rename Col_0009 to state ×
- Rename Col_0010 to zipcode ×
- Rename Col_0012 to email ×

Recommendations for All

- Rename Col_0005 to middle_initial ×
- Rename date_02 to cc_expiration ×
- Rename Col_0019 to cvv ×
- Rename Col_0001 to identifier ×
- Enrich first_name with name_first_ge... ×
- Split column street_address with reg... ×

	Column	Type	Sample Values
	Col_0001	integer	1004640; 1000224; 1004347; 1010987
✓	gender	string	male; female; f; m
✓	name_title	string	Mr.; Ms.; Mrs.; Dr.
✓	first_name	string	James; John; Robert; Michae
	Col_0005	string	J; M; R; D; C; A; S; L; E; B
✓	last_name	string	Smith; Johnson; Williams; Jo
✓	street_address	string	3044 Hood Avenue; 1380 Go 1751 Kinney Street; 1262 Shi 2753 Java Lane; 1870 Hickor
✓	city	string	New York; Los Angeles; Chic Portland; San Diego
✓	state	string	CA; TX; NY; FL; IL; PA; MI; O
✓	zipcode	string	90017; 19108; 49503; 48075;
	Col_0011	string	US
✓	email	string	RobertJRodriguez@yahoo.co

Profile Results

Metric	Count	Percent
Total Rows	15000	100.00%
Empty Rows	0	0.00%
Columns	23	100.00%
Null Columns	2	8.70%
Filtered Columns	2	8.70%
Singleton Columns	1	4.35%
Processed Columns	21	91.30%
Classified Columns	18	78.26%
Cols with Alerts	2	9.52%
Cols with Recommendations	12	52.17%
Total Recommendations	71	-
Average per Column	3	-

Source: ORA_CLOUD_STORE1
File: demodata/AccountInfo_15Krow_NoHeaders.xls

Figure 3: Transform Authoring Page

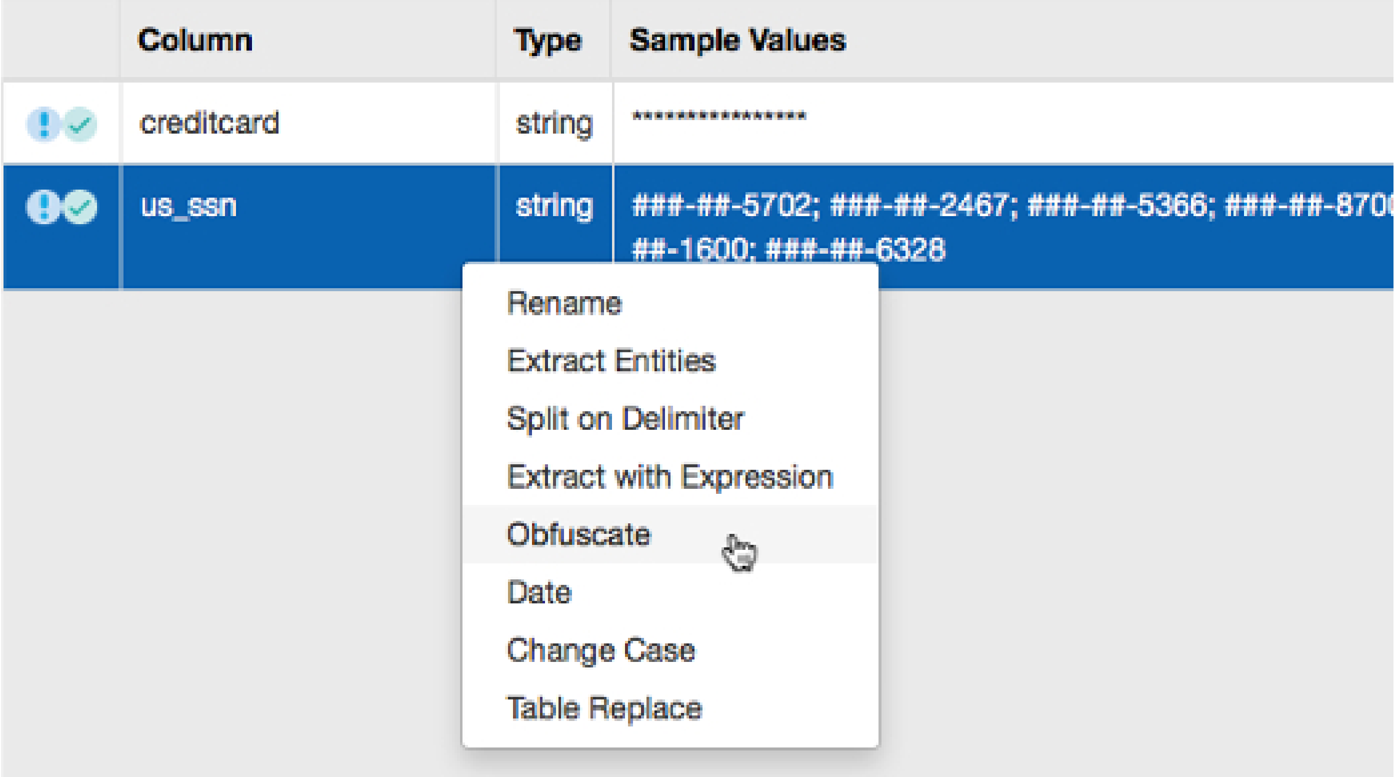
Transformations that will be applied automatically to columns in the data set are listed at the top of the **Transform Script** panel at the left side of the page. Typically the transformations include renaming columns for which there is a clear match with known data patterns.

Recommendations—which are transformations you might want to apply but that are not run automatically, because the pattern match is weak or the transformation alters data values—are shown in the **Recommendations for All** panel at the bottom left of the page.

Now let's start working through the transformations you might want to apply. First you'll look at the highest-priority items on the page, which are the columns associated with the red Cols with Alerts metric in the **Profile Results** panel at the right. Then you'll apply one of the recommended transformations listed in the **Recommendations for All** panel.

1. In the **Profile Results** panel, click the **2** next to the Cols with Alerts metric. Clicking the count next to a metric filters the list of columns displayed in the center of the page. So, in this case, only the **creditcard** and **us_ssn** columns (automatically renamed for you) will be shown.

2. To obfuscate the values in the **us_ssn** column so that no sensitive customer data makes its way into downstream applications or analysis environments, start by hovering your cursor next to the us_ssn column's name, clicking the menu icon next to it, and selecting **Obfuscate** from the menu, as shown in **Figure 4**.



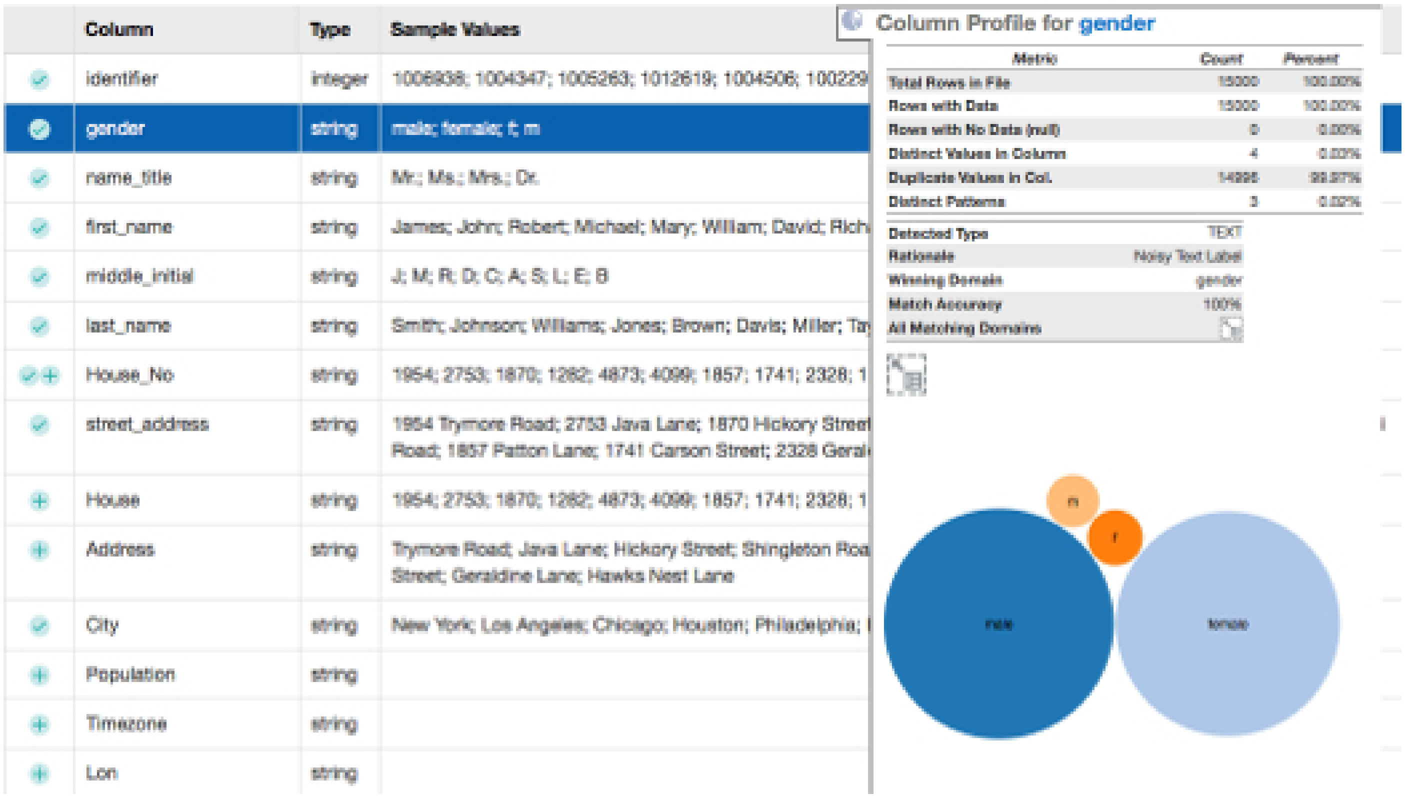
	Column	Type	Sample Values
	creditcard	string	****************
	us_ssn	string	###-##-5702; ###-##-2467; ###-##-5366; ###-##-870 ##-1600; ###-##-6328

Figure 4: Obfuscate Menu Item

3. Now, repeat Step 2 to obfuscate the values in the **creditcard** column.

4. If you now check the list of transformations listed in the **Transform Script** panel, you should see the two obfuscation transformations added to the list of transformations that were automatically created by the profiling process. When you're done with the obfuscations, click the **Metadata View** button in the top left corner of the page to redisplay the full list of columns and their accompanying metadata items. Now you are ready to set up another transformation—this time, to standardize values in the gender column.

5. Locate the gender column in the list, and click to select it, as shown in **Figure 5**. Note that "male" and "female" are domain values but that there are some incidences of "m" and "f" values that were probably entered into the data file in error. These values are likely to cause problems for anyone looking to analyze the values in that column.



	Column	Type	Sample Values
✓	identifier	integer	1006938; 1004347; 1005263; 1012619; 1004506; 100229
✓	gender	string	male; female; f; m
✓	name_title	string	Mr.; Ms.; Mrs.; Dr.
✓	first_name	string	James; John; Robert; Michael; Mary; William; David; Rich
✓	middle_initial	string	J; M; R; D; C; A; S; L; E; B
✓	last_name	string	Smith; Johnson; Williams; Jones; Brown; Davis; Miller; Ta
✓ +	House_No	string	1954; 2753; 1870; 1282; 4873; 4099; 1857; 1741; 2328; 1
✓	street_address	string	1954 Trymore Road; 2753 Java Lane; 1870 Hickory Street Road; 1857 Patton Lane; 1741 Carson Street; 2328 Geral
+	House	string	1954; 2753; 1870; 1282; 4873; 4099; 1857; 1741; 2328; 1
+	Address	string	Trymore Road; Java Lane; Hickory Street; Shingleton Roa Street; Geraldine Lane; Hawks Nest Lane
✓	City	string	New York; Los Angeles; Chicago; Houston; Philadelphia;
+	Population	string	
+	Timezone	string	
+	Lon	string	

Column Profile for gender

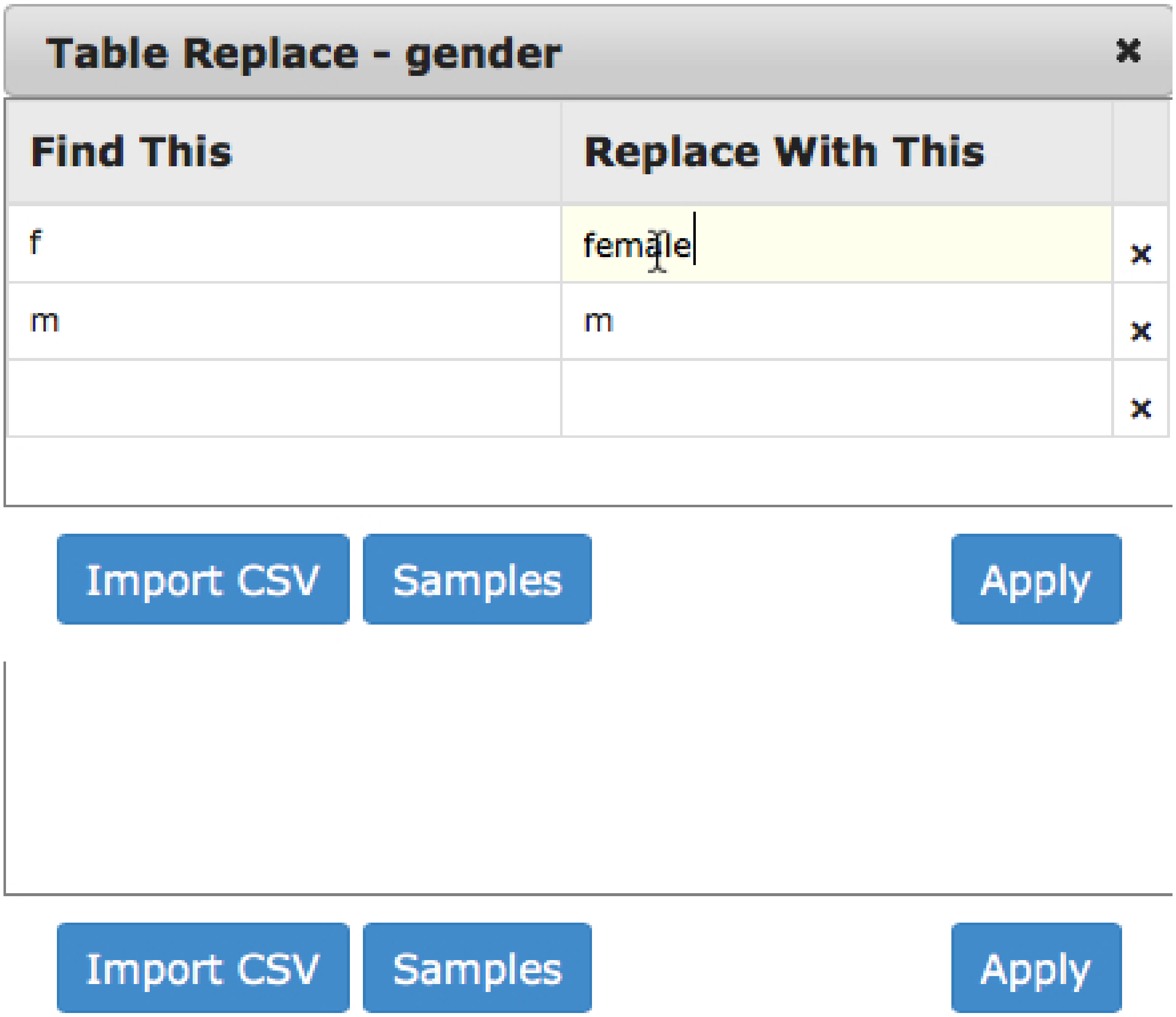
Metric	Count	Percent
Total Rows in File	15000	100.00%
Rows with Data	15000	100.00%
Rows with No Data (null)	0	0.00%
Distinct Values in Column	4	0.03%
Duplicate Values in Col.	14996	99.97%
Distinct Patterns	3	0.02%

Detected Type	TEXT
Rationale	Noisy Text Label
Winning Domain	gender
Match Accuracy	100%
All Matching Domains	

Figure 5: Values in the Gender Column

6. To begin standardizing the values in the gender column so that only “male” and “female” are used, click the menu icon next to the column’s name and choose Table Replace from the menu that is displayed.

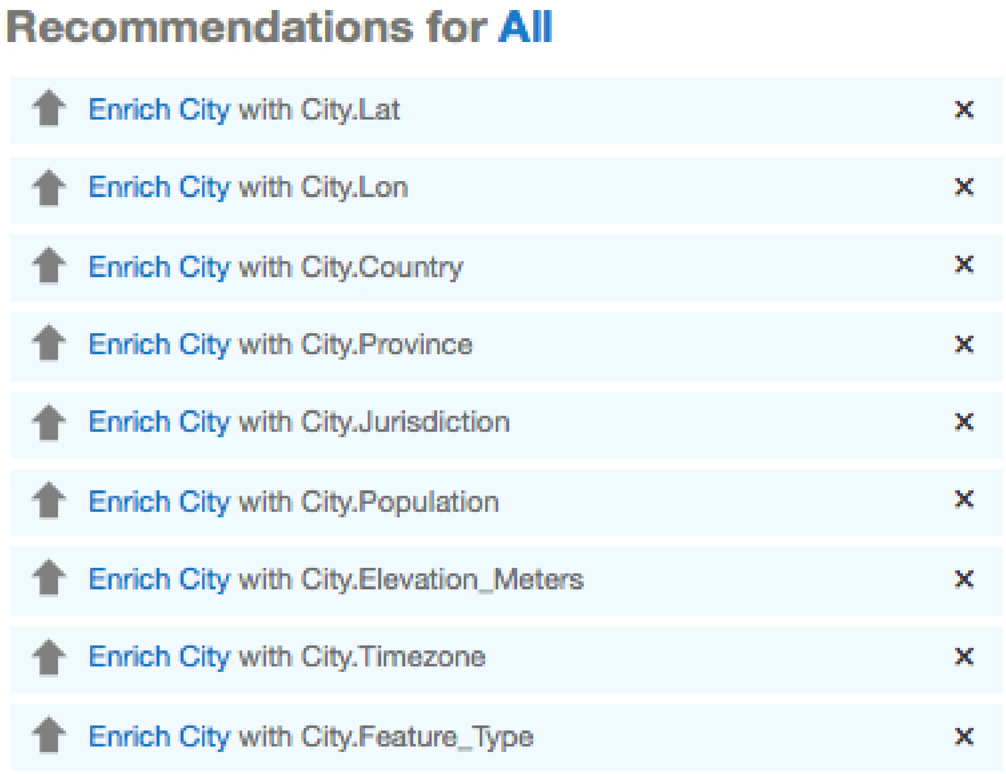
7. Then, in the Table Replace dialog box, first click the **Samples** button to show the set of column values. Then remove the “male” and “female” values by clicking the **x** next to them. To replace the “f” and “m” values that remain, type female and male, respectively, into the **Replace With This** column, as shown in **Figure 6**. Then click **Apply** to close the dialog box and return to the transformation editor screen.



Find This	Replace With This
f	female
m	m

Figure 6: Table Replace Dialog Box

8. As the final transformation, you'll use some of the enrichment recommendations to add population, latitude, longitude, and other data discovered earlier. To add the enrichment recommendations shown in **Figure 7** to the file you publish, locate the recommendations in the **Recommendations for All** panel and then click the up arrow next to each recommendation you want to add to the main set of transformations for this file.



Recommendations for All

- Enrich City with City.Lat ✕
- Enrich City with City.Lon ✕
- Enrich City with City.Country ✕
- Enrich City with City.Province ✕
- Enrich City with City.Jurisdiction ✕
- Enrich City with City.Population ✕
- Enrich City with City.Elevation_Meters ✕
- Enrich City with City.Timezone ✕
- Enrich City with City.Feature_Type ✕

Figure 7: Enrichment Recommendations

Once you're finished, click the **Done Editing** button in the top-right corner of the screen to close this page and return to the Catalog page.

Publishing and Accessing the Prepared Data File

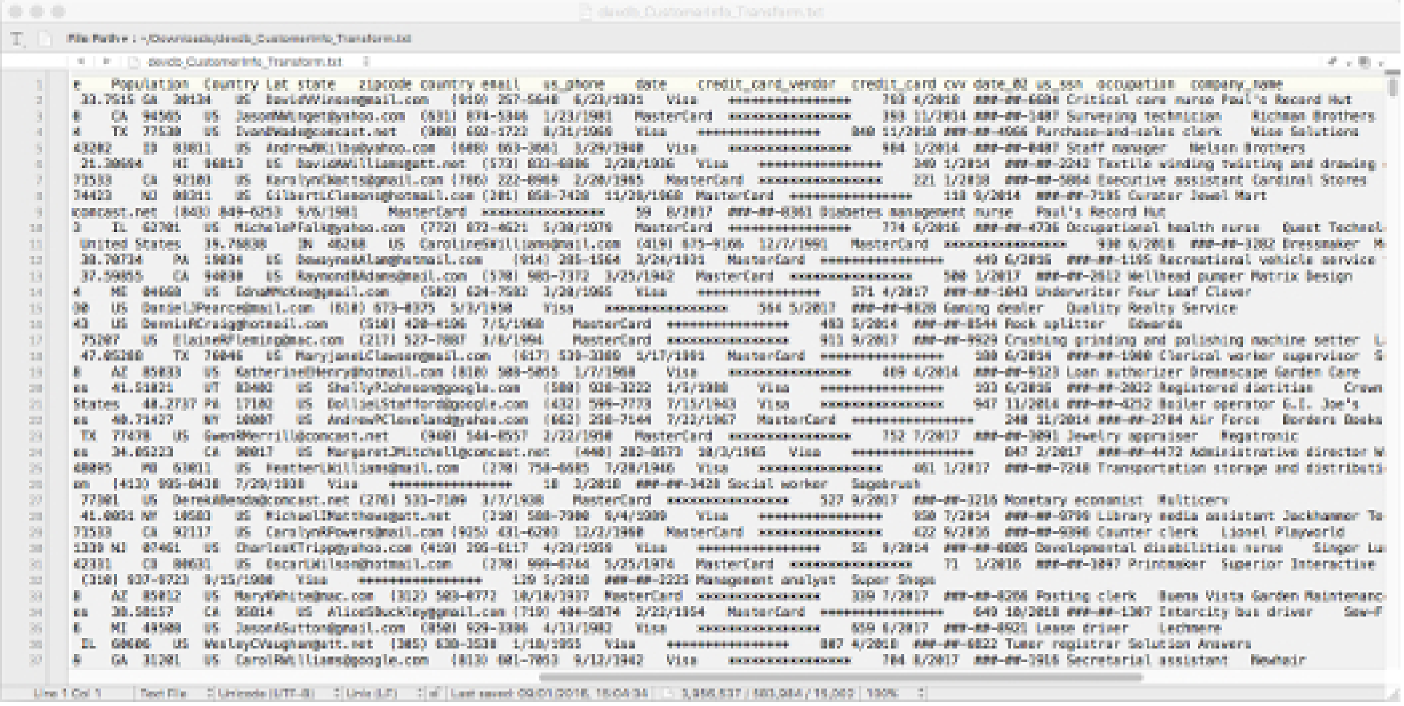
Now that you have defined the transformations, enrichments, and column obfuscations for this data set, you can publish the prepared data file back to Oracle Storage Cloud Service for later download to your workstation or you can transfer the prepared file to another environment. To publish your prepared file, take the following steps:

1. On the Catalog page in Oracle Big Data Preparation Cloud Service, navigate to the transformation you created a moment ago and click the context menu icon to the right of it. When the menu is displayed, select **Publish** from the listed options.

2. On the Publish page, select **Default_Cloud_Storage** for **Target** and then click the **Publish** button to proceed. You should then see the output from the publish process displayed in your web browser and, when the process is complete, you should see a message indicating that the publish process succeeded.

3. To download the published file and view its contents, return to the Catalog page and click the **Download Data** button. On the Select Source Type page, select the Oracle Storage Cloud Service source type and then **Default_Cloud_Storage** as the source. Navigate to the /Publish directory, and locate the published file. (Your username is appended to the name of the transformation file.) Once you have located the file, select it and click the **Download** button to download it from Oracle Cloud to your workstation.

4. Finally, open the prepared file with a text editor and check that the prepared file contains not only the data you selected from the original data file but also the columns added by the enrichment process and the standardized gender values. Also check that all the sensitive customer data is obfuscated, as shown in **Figure 8**.


devdb_CustomerInfo_Transform.txt
File Path ▸ : ~/Downloads/devdb_CustomerInfo_Transform.txt
devdb_CustomerInfo_Transform.txt
Line 1 Col 1 | Text File | Unicode (UTF-8) | Unix (LF) | Last saved: 03/01/2018, 15:04:34 | 3,856,537 / 583,064 / 15,002 | 100%

```
e    Population  Country Lat state   zipcode country email   us_phone    date    credit_card_vendor  credit_card cvv date_02 us_ssn  occupation  company_name
 33.7515 GA  30134   US  DavidVWinean@mail.com   (910) 257-5648  6/21/1931   Visa    ****************    703 4/2018  ###-##-6604 Critical care nurse Paul's Record Hut
0    CA  94565   US  JasonMWinget@yahoo.com (631) 874-5346  1/23/1981   MasterCard  ****************    393 11/2014 ###-##-1487 Surveying technician    Richman Brothers
4    TX  77530   US  IvanMWade@comcast.net  (908) 602-1722  8/31/1960   Visa    ****************    840 11/2018 ###-##-4066 Purchase-and-sales clerk    Wise Solutions
43202    ID  83811   US  AndrewRKilby@yahoo.com (608) 863-3661  3/29/1940   Visa    ****************    984 1/2014  ###-##-0407 Staff manager   Nelson Brothers
 21.30604    HI  96813   US  DavidAWilliams@att.net  (573) 833-6806  3/28/1936   Visa    ****************    340 1/2014  ###-##-2243 Textile winding twisting and drawing
73533    CA  92101   US  KerstynCWatts@gmail.com (786) 222-8909  2/20/1965   MasterCard  ****************    221 1/2018  ###-##-5804 Executive assistant Cardinal Stores
34423    NJ  08311   US  GilbertLClemons@hotmail.com (201) 658-7428  11/26/1960  MasterCard  ****************    118 9/2014  ###-##-7185 Curator Jewel Mart
comcast.net  (843) 849-6253  9/6/1981    MasterCard  ****************    59  8/2017  ###-##-6361 Diabetes management nurse   Paul's Record Hut
3    IL  62701   US  MichelePFalk@yahoo.com (772) 872-4621  5/30/1976   MasterCard  ****************    774 6/2016  ###-##-4736 Occupational health nurse   Quest Technol
 United States   39.76838    IN  46208   US  CarolineSWilliams@mail.com  (419) 675-9166  12/7/1991   MasterCard  ****************    930 6/2016  ###-##-3382 Dressmaker  M
 38.78734    PA  19034   US  DewaynaAAlan@hotmail.com    (914) 285-1564  3/24/1931   MasterCard  ****************    440 6/2016  ###-##-1195 Recreational vehicle service
 37.59855    CA  94030   US  RaymondBAdams@mail.com  (578) 985-7372  3/25/1942   MasterCard  ****************    508 1/2017  ###-##-2812 Wellhead pumper Matrix Design
4    ME  04660   US  EdnaMMcKee@gmail.com    (582) 624-7582  3/20/1965   Visa    ****************    571 4/2017  ###-##-1043 Underwriter Four Leaf Clover
90   US  DanielJPearce@mail.com  (610) 673-8375  5/3/1950    Visa    ****************    564 5/2017  ###-##-8828 Gaming dealer   Quality Realty Service
43   US  DennisRCraig@hotmail.com    (510) 428-6106  7/5/1960    MasterCard  ****************    483 5/2014  ###-##-8544 Rock splitter   Edwards
 75207   US  ElaineRFleming@mac.com  (217) 527-7887  3/8/1994    MasterCard  ****************    913 9/2017  ###-##-9929 Crushing grinding and polishing machine setter  L
 47.05280    TX  76046   US  MaryjaneLClawson@mail.com   (617) 539-3389  1/17/1983   MasterCard  ****************    300 6/2014  ###-##-1900 Clerical worker supervisor  S
0    AZ  85033   US  KatherineDHenry@hotmail.com (810) 589-5055  1/7/1968    Visa    ****************    409 4/2014  ###-##-9123 Loan authorizer Dreamscape Garden Care
es   41.51821    UT  83402   US  ShellyFJohnson@google.com   (508) 928-3222  1/5/1988    Visa    ****************    393 6/2016  ###-##-2022 Registered dietitian    Crown
States   40.2737 PA  17102   US  DollieLStafford@google.com  (432) 596-7773  7/15/1943   Visa    ****************    947 11/2014 ###-##-4252 Boiler operator G.I. Joe's
es   40.71427    NY  10007   US  AndrewPCleveland@yahoo.com  (662) 258-7144  7/22/1967   MasterCard  ****************    248 11/2014 ###-##-2784 Air Force   Borders Books
 TX  77478   US  GwenRMerrill@comcast.net    (940) 544-8557  2/22/1950   MasterCard  ****************    752 7/2017  ###-##-3091 Jewelry appraiser   Megatronic
es   34.05223    CA  90017   US  MargaretJMitchell@comcast.net   (440) 282-8173  10/3/1965   Visa    ****************    047 2/2017  ###-##-4472 Administrative director W
40005    MO  63011   US  HeatherLWilliams@mail.com   (270) 750-6685  7/28/1946   Visa    ****************    461 1/2017  ###-##-7240 Transportation storage and distributi
om   (413) 895-0438  7/29/1938   Visa    ****************    18  3/2018  ###-##-3428 Social worker   Sagebrush
 77301   US  DerekWBenda@comcast.net (276) 531-7189  3/7/1938    MasterCard  ****************    527 9/2017  ###-##-3216 Monetary economist  Multicorp
 41.0851 NY  10583   US  MichaelIMatthews@att.net    (239) 588-7900  9/4/1989    Visa    ****************    850 7/2014  ###-##-9798 Library media assistant Jackhammer Te
73533    CA  92117   US  CarolynRPowers@mail.com (925) 431-6283  12/2/1960   MasterCard  ****************    422 9/2016  ###-##-9396 Counter clerk   Lionel Playworld
1339 NJ  07461   US  CharlesKTripp@yahoo.com (410) 295-6117  4/29/1936   Visa    ****************    35  9/2014  ###-##-6085 Developmental disabilities nurse    Singer Lu
42331    CO  80631   US  OscarLWilson@hotmail.com    (270) 999-6744  5/25/1976   MasterCard  ****************    71  1/2016  ###-##-3897 Printmaker  Superior Interactive
 (310) 937-8723  9/15/1980   Visa    ****************    129 5/2018  ###-##-2223 Management analyst  Super Shops
0    AZ  85012   US  MaryKWhite@mac.com  (312) 503-0772  10/10/1937  MasterCard  ****************    339 7/2017  ###-##-8266 Posting clerk   Buena Vista Garden Maintenanc
es   38.58157    CA  95814   US  AliceSBuckley@gmail.com (718) 404-5874  2/22/1954   MasterCard  ****************    640 10/2018 ###-##-1307 Intercity bus driver    Sew-F
6    MI  49509   US  JasonASutton@gmail.com  (850) 929-3306  4/13/1982   Visa    ****************    659 6/2017  ###-##-8921 Lease driver    Lechmere
 IL  60606   US  WesleyCVaughan@att.net  (385) 630-3538  1/18/1955   Visa    ****************    887 4/2018  ###-##-6822 Tuner registrar Solution Answers
0    GA  31201   US  CarolRWilliams@google.com   (813) 881-7053  9/12/1942   Visa    ****************    704 8/2017  ###-##-1916 Secretarial assistant   Newhair
```

Figure 7: Enrichment Recommendations

Conclusion

Much of the work involved in big data and other analytic projects involves preparing, obfuscating, and enriching datasources so they are useful for analysis and do not contain customer data that is sensitive or private. Oracle Big Data Preparation Cloud Service puts this capability into the hands of data domain experts and makes the process quick and easy to perform, by automating much of the work and running as a service in Oracle Cloud.