

# User guide to create an output file of normalized attribute extraction for new items

## Introduction

This user guide provides instructions to guide users to run the code [ATTRIBUTE\\_EXTRACTION\\_NEW\\_012324 - Databricks \(azuredatabricks.net\)](#) to create an output file of normalized attribute extraction for new items.

- [Introduction](#)
- [Normalized Attribute Extraction Steps](#)
  - [Step 1: Create a CSV file which contains new items since last update](#)
  - [Step 2: Upload the CSV file to Azure Databricks](#)
  - [Step 3: Create /download the output file](#)
  - [Reference](#)

---

## Normalized Attribute Extraction Steps

### Step 1: Create a CSV file which contains new items since last update

- Connect SQL server via Azure Data Studio to access PIM table.
- Follow screenshot1-1, select 'pimmart\_prod', and the server address '[11-eastus2-3534-idq-sqlmi.cc4c669dc8d1.database.windows.net](#)' will show automatically, click 'Connect' to connect SQL server to production environment.

**Screenshot1-1**

**Connection**

Recent Browse

Clear List

pimmart\_prod

PIMMART-DEV

**Connection Details**

Connection type: Microsoft SQL Server

Input type: ☒ Parameters ☐ Connection String

Server \*: 11-eastus2-3534-idq-sqlmi.cc4c669dc8d1.database.windows.net

Authentication type: Azure Active Directory - Universal with MFA support

Account: Yue, Sophia (NonEmp) - sophia.yue@kroger.com

Database: <Default>

Encrypt: Mandatory (True)

Trust server certificate: False

Server group: <Do not save>

Name (optional): pimmart\_prod

Advanced...

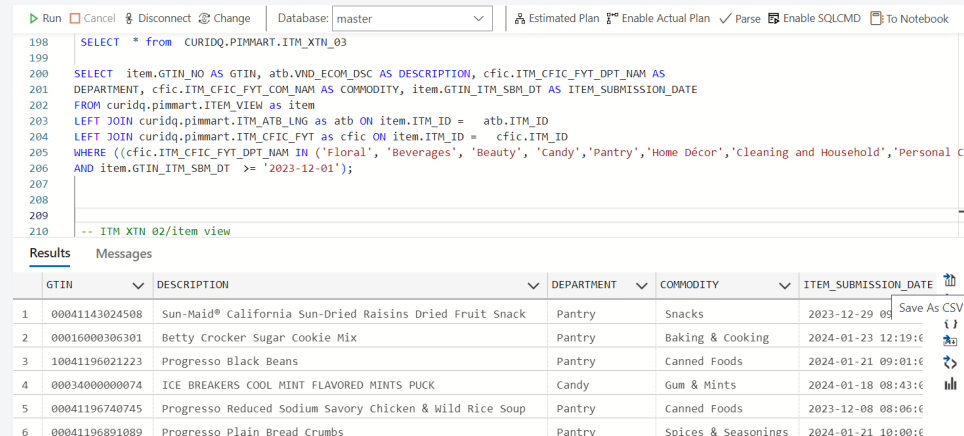
Connect Cancel

- For the query below, replace yyyy-mm-dd with last run date and run the query. If the last run date is 2023-12-01, the yyyy-mm-dd should be 2023-12-01

```
SELECT item.GTIN_NO AS GTIN, atb.VND_ECOM_DSC AS DESCRIPTION, cfic.ITM_CFIC_FYT_DPT_NAM AS  
DEPARTMENT, cfic.ITM_CFIC_FYT_COM_NAM AS COMMODITY, item.GTIN_ITM_SBM_DT AS ITEM_SUBMISSION_DATE  
FROM curidq.pimmart.ITEM_VIEW as item  
LEFT JOIN curidq.pimmart.ATB_LNG as atb ON item.ITEM_ID = atb.ITEM_ID  
LEFT JOIN curidq.pimmart.ITM_CFIC_FYT as cfic ON item.ITEM_ID = cfic.ITEM_ID  
WHERE ((cfic.ITM_CFIC_FYT_DPT_NAM IN ('Floral', 'Beverages', 'Beauty', 'Candy', 'Pantry', 'Home Décor', 'Cleaning and  
Household', 'Personal Care', 'Kitchen & Dining') OR cfic.ITM_CFIC_FYT_COM_NAM = 'Ice Cream')  
AND item.GTIN_ITM_SBM_DT >= 'yyyy-mm-dd');
```

- Save the result to a CSV file with naming convention as SampleInputWeekly\_mmddyy.csv. If run date is 012424, the mmddyy should be 012424.
  - Run the SQL, once you have the query result displayed, mouse over the top icon of right sidebar, 'save as CSV' will pop up on the screen from the 'Results' window. Please refer to screenshot1-2.

### Screenshot1-2

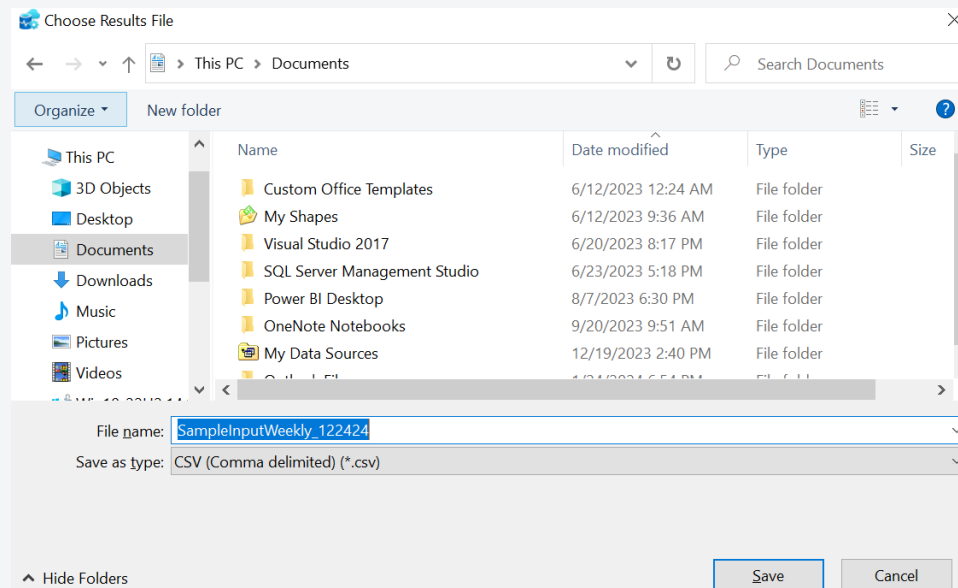


```
198 SELECT * FROM CURIDQ.PIMMART.ITEM_XTH_03
199
200 SELECT item.GTIN_NO AS GTIN, atb.VND_ECOM_DSC AS DESCRIPTION, cfc.ITH_CFIC_FYT_DPT_NAM AS
201 DEPARTMENT, cfc.ITH_CFIC_FYT_COM_NAM AS COMMODITY, item.GTIN_ITH_SBM_DT AS ITEM_SUBMISSION_DATE
202 FROM curidq.pimmart.ITEM_VIEW as item
203 LEFT JOIN curidq.pimmart.ITH_ATB_LNG as atb ON item.ITH_ID = atb.ITH_ID
204 LEFT JOIN curidq.pimmart.ITH_CFIC_FYT as cfc ON item.ITH_ID = cfc.ITH_ID
205 WHERE ((cfc.ITH_CFIC_FYT_DPT_NAM IN ('Floral', 'Beverages', 'Beauty', 'Candy', 'Pantry', 'Home Décor', 'Cleaning and Household', 'Personal Ca
206 AND item.GTIN_ITH_SBM_DT >= '2023-12-01');
207
208
209
210 -- ITH XTH 02/item view
```

	GTIN	DESCRIPTION	DEPARTMENT	COMMODITY	ITEM_SUBMISSION_DATE
1	00041143024508	Sun-Maid® California Sun-Dried Raisins Dried Fruit Snack	Pantry	Snacks	2023-12-29 08:00:00
2	00016000306301	Betty Crocker Sugar Cookie Mix	Pantry	Baking & Cooking	2024-01-23 12:19:00
3	10041196021223	Progresso Black Beans	Pantry	Canned Foods	2024-01-21 09:01:00
4	00034000000074	ICE BREAKERS COOL MINT FLAVORED MINTS PUCK	Candy	Gum & Mints	2024-01-18 08:43:00
5	00041196740745	Progresso Reduced Sodium Savory Chicken & Wild Rice Soup	Pantry	Canned Foods	2023-12-08 08:06:00
6	00041196891089	Proerroso Plain Bread Crumbs	Pantrv	Spices & Seasonines	2024-01-21 10:00:00

- Click the icon on the top, the screen of 'Choose Results file' will pop up. Follow the naming convention to enter the file name, choose the location to save the csv file, and click 'save' to save the csv file. In screenshot1-3, the folder/location to save the CSV file is 'Documents' and the filename is 'SampleInputWeekly\_122424'.

### Screenshot1-3



Choose Results File

← → ↑ ↓ This PC > Documents Search Documents

Organize New folder

Name	Date modified	Type	Size
Custom Office Templates	6/12/2023 12:24 AM	File folder	
My Shapes	6/12/2023 9:36 AM	File folder	
Visual Studio 2017	6/20/2023 8:17 PM	File folder	
SQL Server Management Studio	6/23/2023 5:18 PM	File folder	
Power BI Desktop	8/7/2023 6:30 PM	File folder	
OneNote Notebooks	9/20/2023 9:51 AM	File folder	
My Data Sources	12/19/2023 2:40 PM	File folder	

File name: SampleInputWeekly\_122424

Save as type: CSV (Comma delimited) (\*.csv)

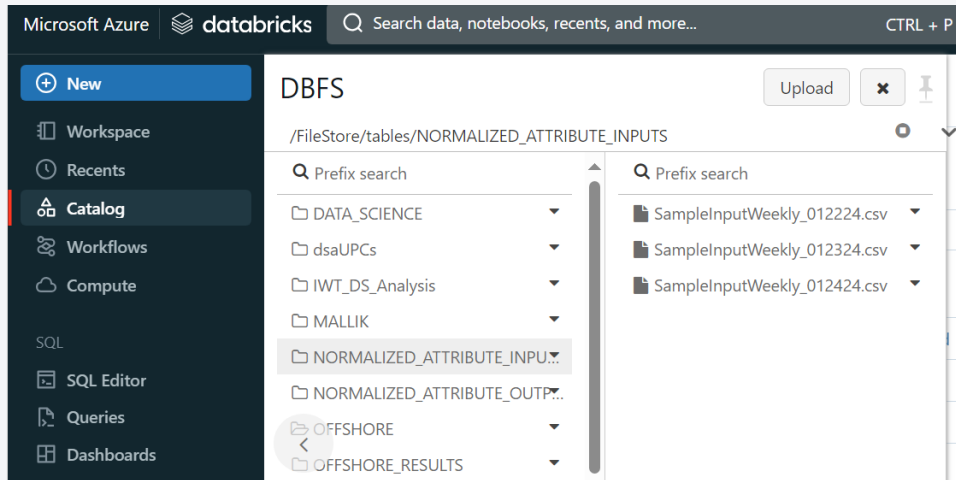
Save Cancel

## Step 2: Upload the CSV file to Azure Databricks

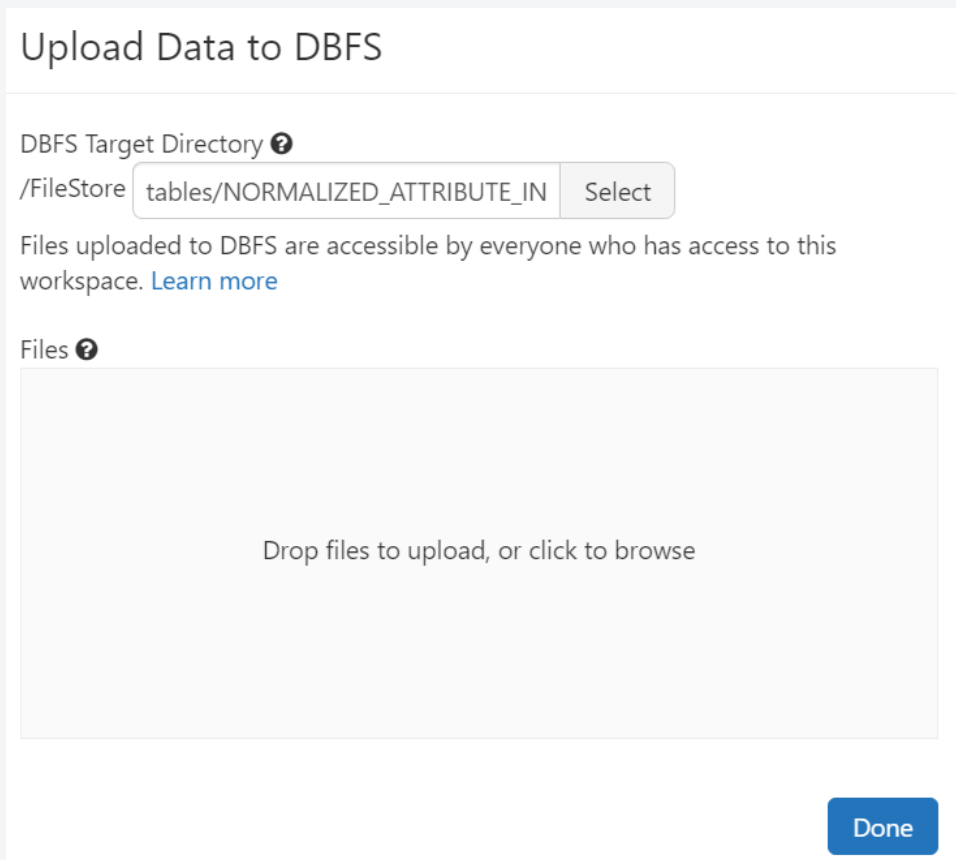
- The input file is required to upload to Azure Databricks and save it in: /dbfs/FileStore/tables/NORMALIZED\_ATTRIBUTE\_INPUTS
- Upload the input file
  - It requires to have Azure Databricks access to upload the input file. Please raise the APT for the access: [jcs000-mx-insights-dev-8666-domain name](#). The profile is domain-specific, and you can get the [domain name](#) from manager.
  - Steps to upload the input file
    - Sing in Azure Databricks.
    - Click 'Catalog' from the left pane.
    - Click 'Browse DBFS' on the top screen to list DBFS file as shown on the second left pane from the screenshot2-1.
    - Click the folder 'NORMALIZED\_ATTRIBUTE\_INPUTS' to show all the files in folder on the third left pane from the screenshot2-1.
    - Click 'Upload' on top of the screen, the screen 'Upload Data to DBFS' will pop up. Please refer to screenshot2-2
    - Click 'click to browse' from screenshot2-2 to pop up screenshot2-3 to open file.
    - Navigate the file location and file name you want to upload

- Click 'open' from screenshot2-3, the screenshot2-4 will pop-up to show the file to be uploaded.
- Click 'done' to upload the file.

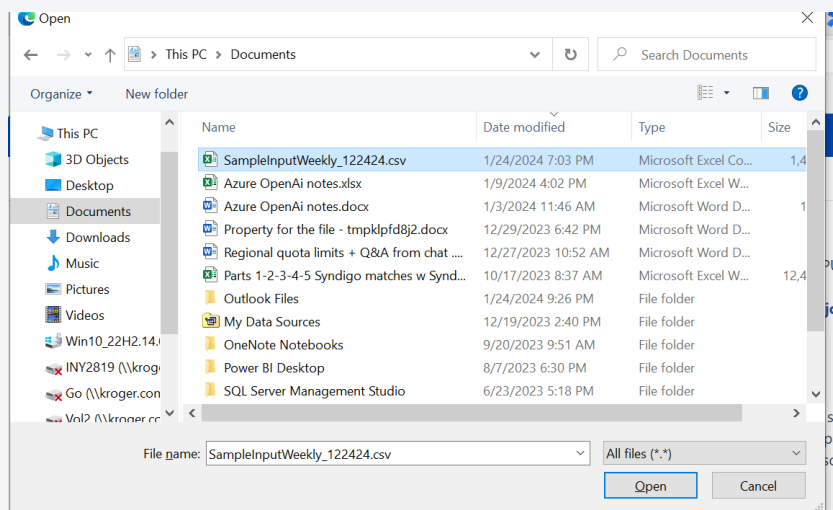
#### Screenshot2-1



#### Screenshot2-2



#### Screenshot2-3



**Screenshot2-4**



### Step 3: Create/download the output file

- After the input file had been uploaded to Databricks, run the script [ATTRIBUTE\\_EXTRACTION\\_NEW\\_012324 - Databricks \(azuredatabricks.net\)](#) to create an output file of normalized attribute extraction for new items.
- The script is a Databricks notebook written in Python with two types of cell:
  - Markdown cells are for document purpose.
  - Code cells with/without cell title are codes. After navigating the script, run the code cells in sequence.
- Script to create/download the output file
  - The script has 4 cells.
  - The type of cell 'cmd 1' and 'cmd2' are 'Markdown'. There is no need to run the cell.
  - The type of cell 'cmd 3' is 'Python' code to set up the environment and load up the code. To run the cell, please click the "Run" button in the notebook toolbar or click the triangle pointed to right in the cell or use the keyboard shortcut "Shift+Enter". The details of code and output are hidden. Double click "show code"/"Show result" to view the details.
  - The type of cell 'cmd 4' is 'Python' code to run the latest input file to generate the output (This might take a few minutes)
    - The result will show the process for all attributes and departments.
    - Click the button 'Download File' or the URL to download the output file.

ATTRIBUTE\_EXTRACTION\_NEW\_012324Python☆

FileEditViewRunHelpLast edit was 21 hours agoProvide feedback

Run allDMP Access - StageScheduleShare

Cmd 1

Markdown

▢ ▾ - ✕

### Versions of libraries used

numpy==1.24.3 pandas==1.3.4 rapidfuzz==3.6.1 ipywidgets==8.1.1

Cmd 2

Markdown

▢ ▾ - ✕

### Output File path:

/dbfs/FileStore/tables/NORMALIZED\_ATTRIBUTE\_OUTPUTS

Cmd 3

Python

▶ ▾ - ✕

Run to set up the environment and load up the code

Show code

Show result

🔔 1

Cmd 4

Python

▶ ▾ - ✕

Run the latest input file to generate the output (This might take a few minutes)

Show code

```
Loading the latest input file: /dbfs/FileStore/tables/NORMALIZED_ATTRIBUTE_INPUTS/SampleInputWeekly_122424-1.csv
Completed loading up configs.
Color Kitchen & Dining
Total input items: [ 476 x 5 ]
Completed loading up configs.
Color Home_Decor_Rugs
Total input items: [ 28 x 5 ]
Completed loading up configs.
Color Home_Decor_NoRugs
Total input items: [ 596 x 5 ]
Completed loading up configs.
Color Floral
Total input items: [ 193 x 5 ]
Completed loading up configs.
Flavor Pantry
Total input items: [ 4807 x 5 ]
Completed loading up configs.
Flavor Beverages
Total input items: [ 1483 x 5 ]
Completed loading up configs.
Flavor Ice Cream
```

Download File

If the above link doesn't work, try pasting this URL into your browser.  
[http://adb-4812933386228410.10.azuredatabricks.net/files/tables/NORMALIZED\\_ATTRIBUTE\\_OUTPUTS/Normalized\\_Attributes\\_01252024\\_csv](http://adb-4812933386228410.10.azuredatabricks.net/files/tables/NORMALIZED_ATTRIBUTE_OUTPUTS/Normalized_Attributes_01252024_csv)

## Reference

[Automation of normalized attribute extraction for new items.](#)