# 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.



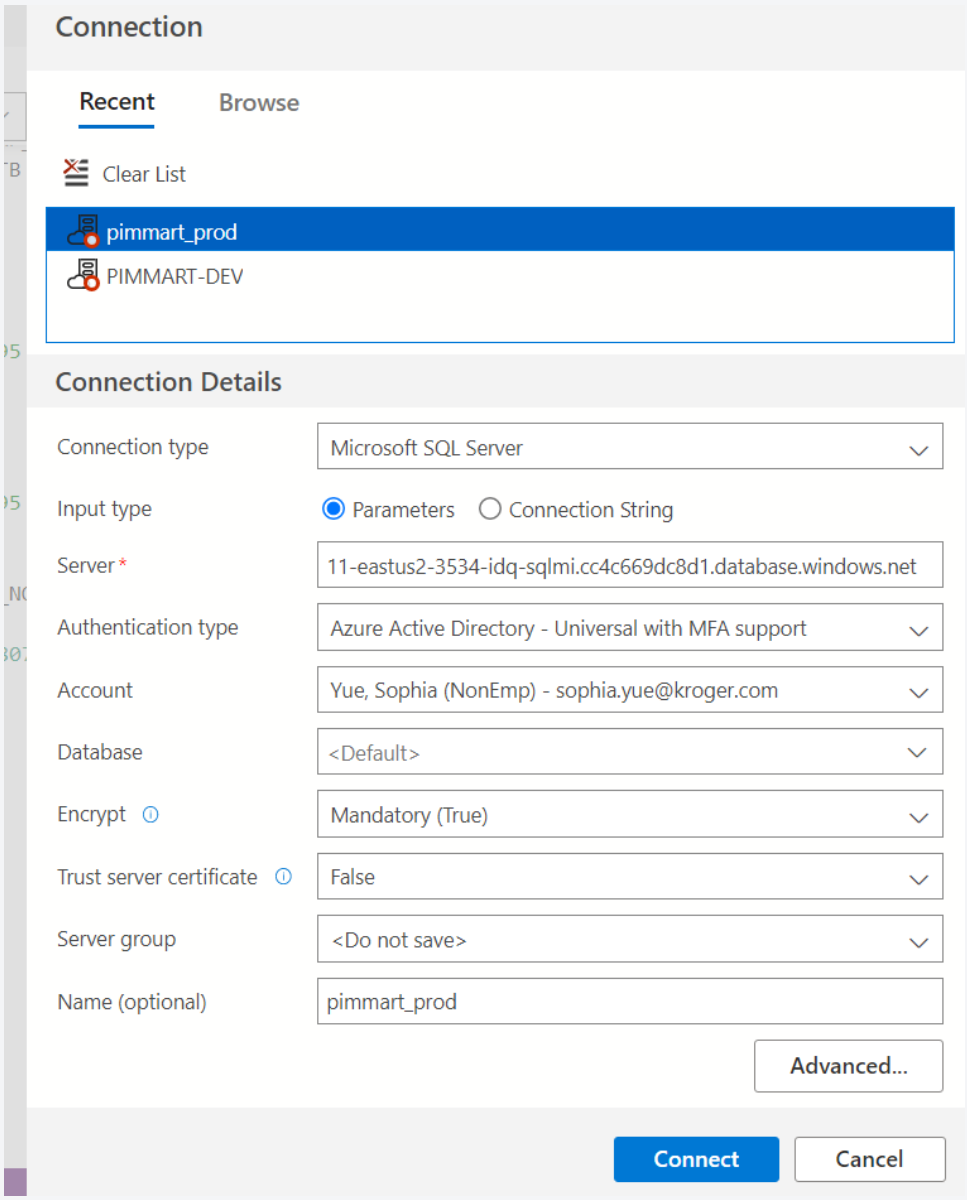
---

### 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**

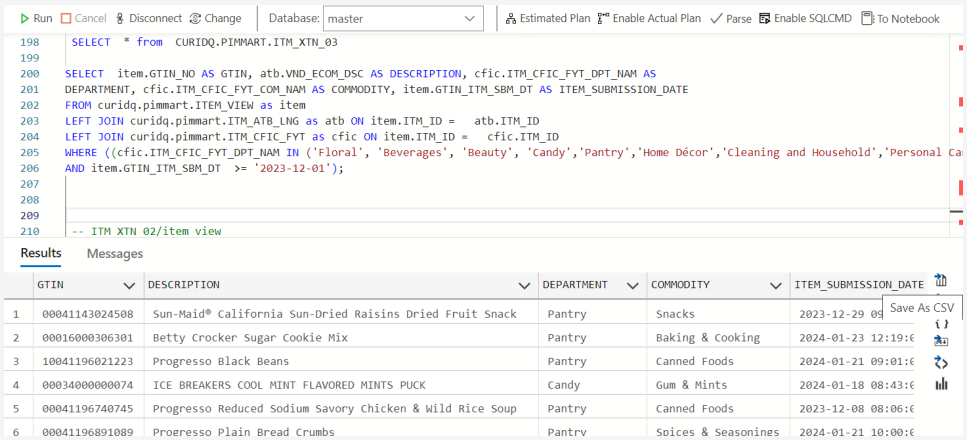



- For the query below, replace yyyy-mm-dd with last run date and run the query. If the last run date is 2023-12-01, the yyyy-mm-dd should be 2023-12-01

  ```
  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.ITM_ATB_LNG as atb ON item.ITM_ID =   atb.ITM_ID
  LEFT JOIN curidq.pimmart.ITM_CFIC_FYT as cfic ON item.ITM_ID =   cfic.ITM_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');
  ```

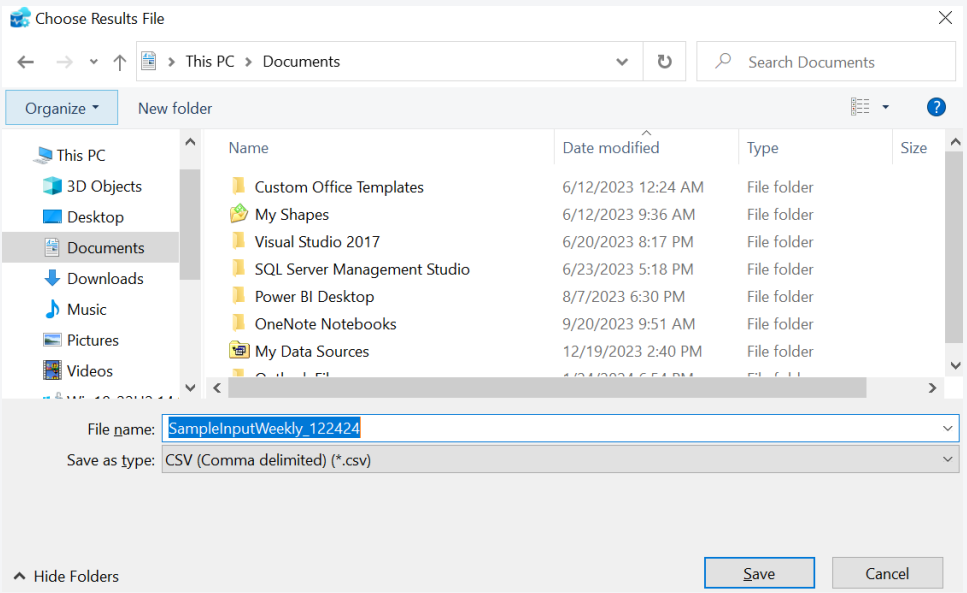
- Save the result to a CSV file with naming convention as SampleInputWeekly_mmddyy.csv. If run date is 012424, the mmddyy should be 012424.
  - Run the SQL, once you have the query result displayed, mouse over the top icon of right sidebar, 'save as CSV' will pop up on the screen from the 'Results' window.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

**Screenshot1-2**




- 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**

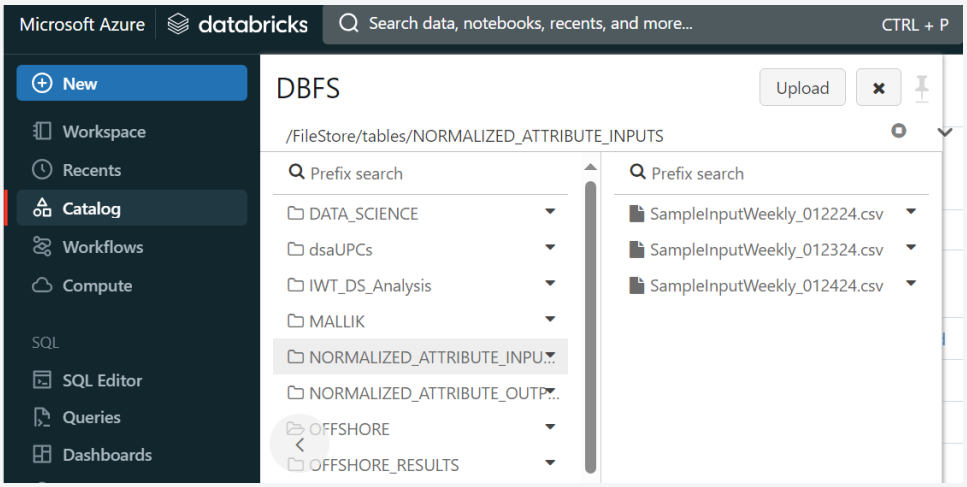



## Step 2: Upload the CSV file to Azure Databricks

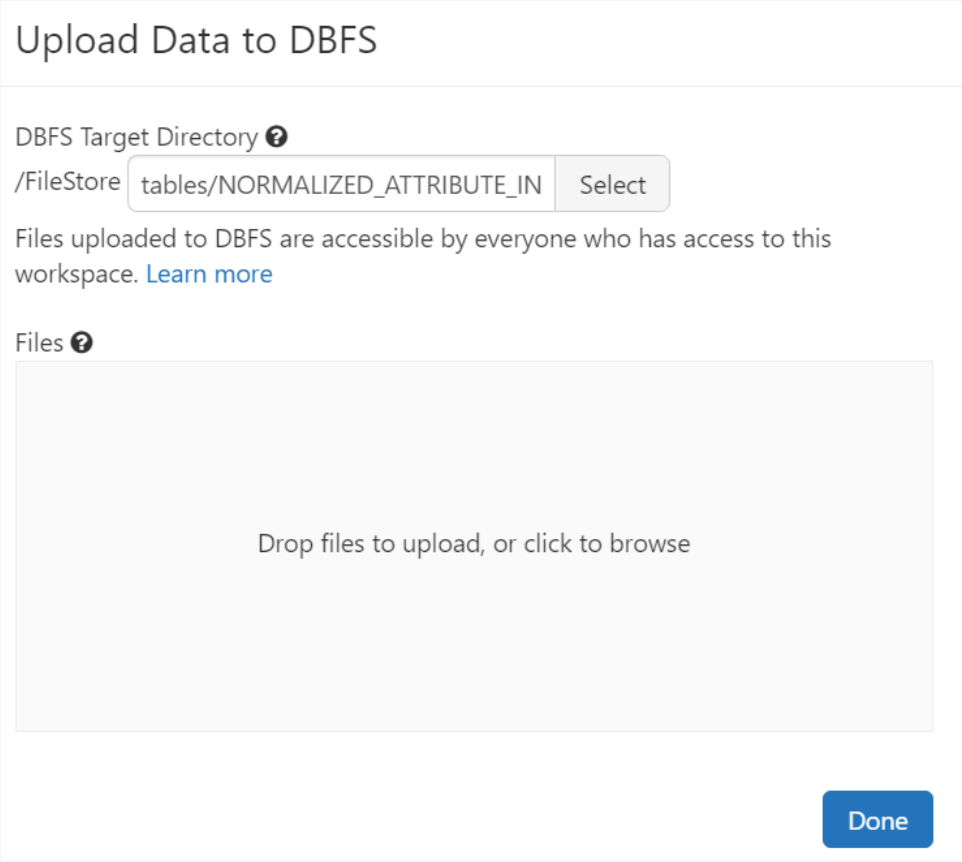
- The input file is required to upload to Azure Databricks and save it in: /dbfs/FileStore/tables/NORMALIZED_ATTRIBUTE_INPUTS
- Upload the input file
  - It requires to have Azure Databricks access to upload the input file. Please raise the APT for the access: **jcs000-mx-insights-dev-8666-domain name.** 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.

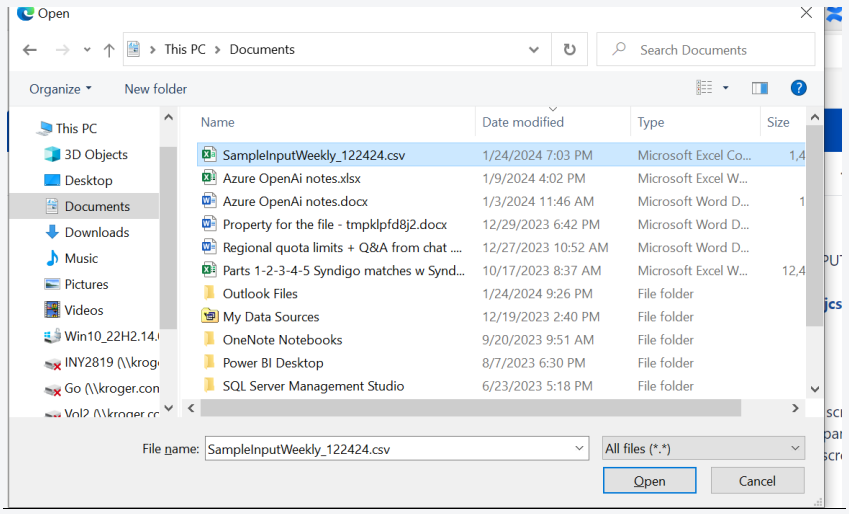
**Screenshot2-1**



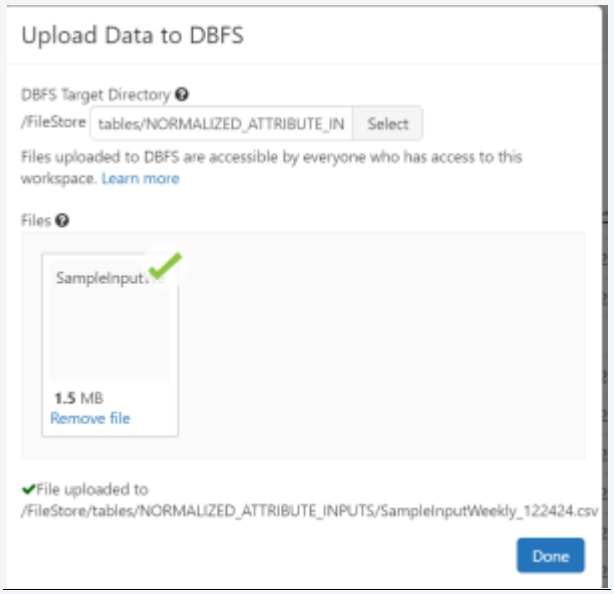

**Screenshot2-2**




**Screenshot2-3**

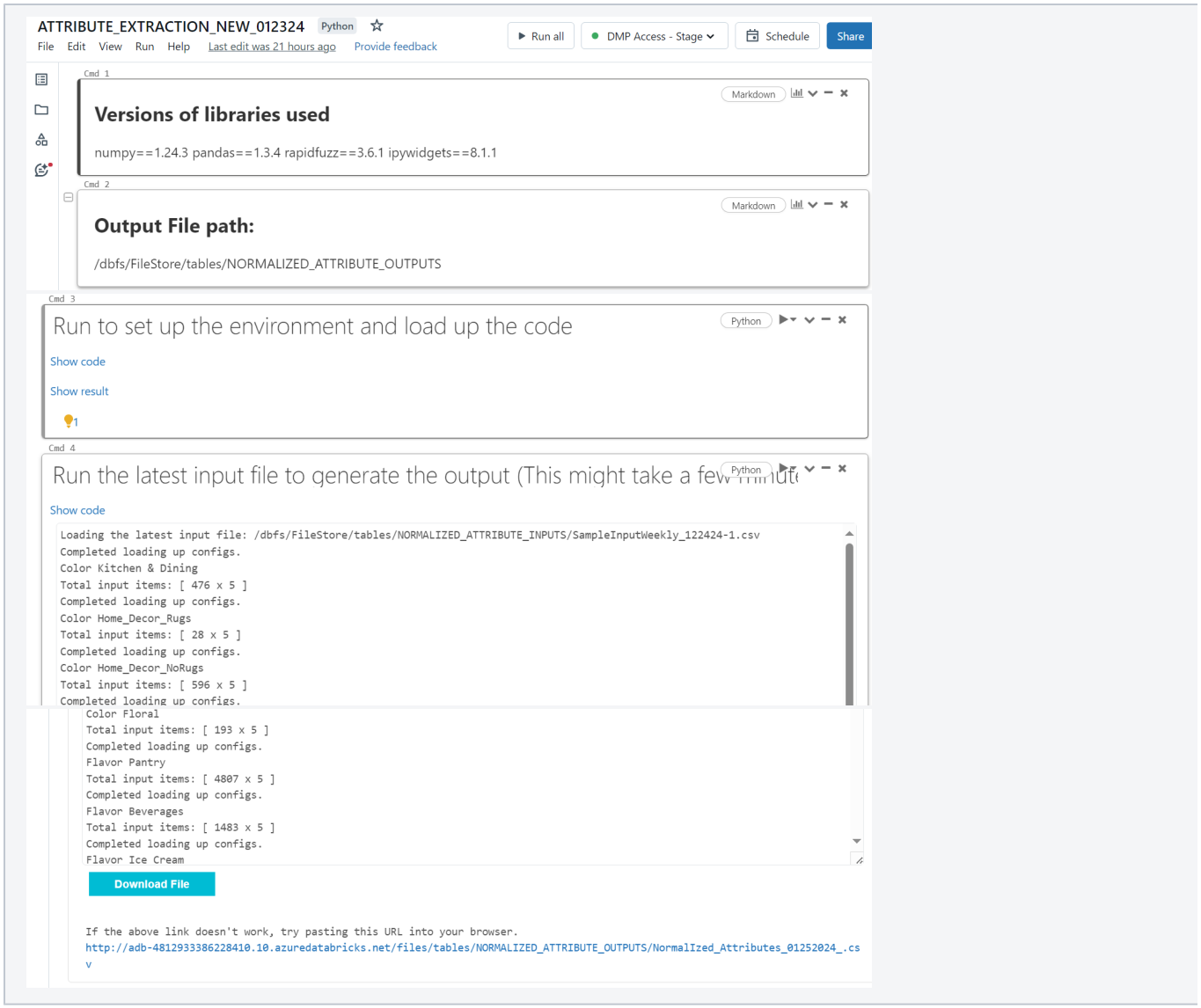



**Screenshot2-4**




## 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.




## Reference

Automation of normalized attribute extraction for new items.