





## Create New Table

DBFS Target Directory ?

/FileStore/tables/ (optional) Select

Files uploaded to DBFS are accessible by everyone who has access to this workspace. [Learn more](#)

Files ?

Financial\_Complaints\_Product.csv

4.4 MB

Remove file

Financial\_Complaints\_Issue.csv

5.9 MB

Remove file

Financial\_Complaints\_Region.csv

1.7 MB

Remove file

Financial\_Complaints\_Product.csv

8.7 MB

Cancel upload

- ✓File uploaded to /FileStore/tables/Financial\_Complaints\_Product.csv
- ✓File uploaded to /FileStore/tables/Financial\_Complaints\_Issue.csv
- ✓File uploaded to /FileStore/tables/Financial\_Complaints\_Region.csv

Create Table with UI Create Table in Notebook ?

Product

Cmd 2

```
# Mount from Azure Blob
dbutils.fs.mount(
    source = "wasbs://<container-name>@<storage-account-name>.blob.core.windows.net",
    mount_point = "/mnt/<mount-name>",
    extra_configs = {"<conf-key>":dbutils.secrets.get(scope = "<scope-name>", key = "<key-name>")})
```

java.lang.IllegalStateException: Secrets API is not enabled for this workspace.

Command took 3.54 seconds -- by soumitkar.5@gmail.com at 10/24/2021, 8:42:57 PM on Tableau\_Demo

Cmd 3

```
# File location and type
file_location1 = "/mnt/<mount-file/Financial_Complaints_Product.csv"
file_location2 = "/mnt/<mount-file/Financial_Complaints_Issue.csv"
file_location3 = "/mnt/<mount-file/Financial_Complaints_Response.csv"
file_location4 = "/mnt/<mount-file/Financial_Complaints_Region.csv"
file_type = "csv"

# The applied options are for CSV files. For other file types, these will be ignored.
df1 = spark.read.csv(file_location1, header=True, inferSchema=True)
df2 = spark.read.csv(file_location2, header=True, inferSchema=True)
df3 = spark.read.csv(file_location3, header=True, inferSchema=True)
df4 = spark.read.csv(file_location4, header=True, inferSchema=True)
```

Product



```
# create merged dataframe
from pyspark.sql.functions import col
df = df1.join(df2, ["Complaint ID"]) \
    .join(df3, ["Complaint ID"]) \
    .join(df4, ["Complaint ID"])
```

df: pyspark.sql.dataframe.DataFrame = [Complaint ID: integer, \_c0: integer ... 19 more fields]

Command took 0.23 seconds -- by soumitkar.5@gmail.com at 10/24/2021, 8:44:05 PM on Tableau\_Demo

Cmd 6



```
df.printSchema()
```

```
root
|-- Complaint ID: integer (nullable = true)
|-- _c0: integer (nullable = true)
|-- Date Submitted: string (nullable = true)
|-- Product: string (nullable = true)
|-- Sub-product: string (nullable = true)
|-- _c0: integer (nullable = true)
|-- Issue: string (nullable = true)
|-- Sub-issue: string (nullable = true)
|-- Company: string (nullable = true)
|-- Company public response: string (nullable = true)
|-- Company: string (nullable = true)
|-- Consumer consent provided?: string (nullable = true)
|-- Submitted via: string (nullable = true)
|-- Tags: string (nullable = true)
```

## 2021-10-24 - DBFS Example (Python)

Tableau\_Demo | File Edit View: Standard Permissions Run All Clear

Publish Comments Experiment

Product

```
#convert string to date format
from pyspark.sql.functions import unix_timestamp, from_unixtime
df = df.select(
    'Date_Sumbited',
    from_unixtime(unix_timestamp('Date_Sumbited', 'MM/dd/yyyy')).alias('Date_Sumbited'), 'Date_Received',
    from_unixtime(unix_timestamp('Date_Received', 'MM/dd/yyyy')).alias('Date_Received'))
```

df: pyspark.sql.dataframe.DataFrame = [Date\_Sumbited: string, Date\_Sumbited: string ... 2 more fields]

Command took 0.14 seconds -- by soumitkar.5@gmail.com at 10/24/2021, 10:19:37 PM on Tableau\_Demo

Cmd 8

```
#selected columns

column_list = ['Complaint ID', 'Date_Sumbited', 'Product', 'Sub-product', 'Issue',
               'Sub-issue', 'Company public response', 'State', 'ZIP code',
               'Tags', 'Consumer consent provided?', 'Submitted via', 'Date_Received',
               'Company response to consumer', 'Timely response?',
               'Consumer disputed?']

df = df.select([column for column in df.columns if column in column_list])
```

df: pyspark.sql.dataframe.DataFrame = [Date\_Sumbited: string, Date\_Received: string]

Command took 0.10 seconds -- by soumitkar.5@gmail.com at 10/24/2021, 10:20:34 PM on Tableau\_Demo

Cmd 9

```
#Load to ADLS
df.write.csv("/mnt/<mount-name>/Trans_Financial_Complaints, header = True, mode='overwrite')
```