

SCD TYPE 1 LOGIC ON DELTA TABLE IN DATABRICKS

Step 1:

I have uploaded my day 1 and day 2 csv files in my blob storage account.

The screenshot shows the Azure portal interface for a container named 'scdtype1'. The left sidebar displays the 'Overview' tab, which includes options for 'Diagnose and solve problems', 'Access Control (IAM)', and 'Settings'. The main content area shows the '2025/03/30/Day1.csv' blob file. The file's metadata is displayed, including the 'Authentication method' (Access key) and 'Location' (scdtype1 / 2025 / 03 / 30). The file is listed in a table with columns for 'Name' and 'Size'. The file 'Day1.csv' is shown with a size of 1 KB. The file is also listed in a table with columns for 'Name' and 'Size'.

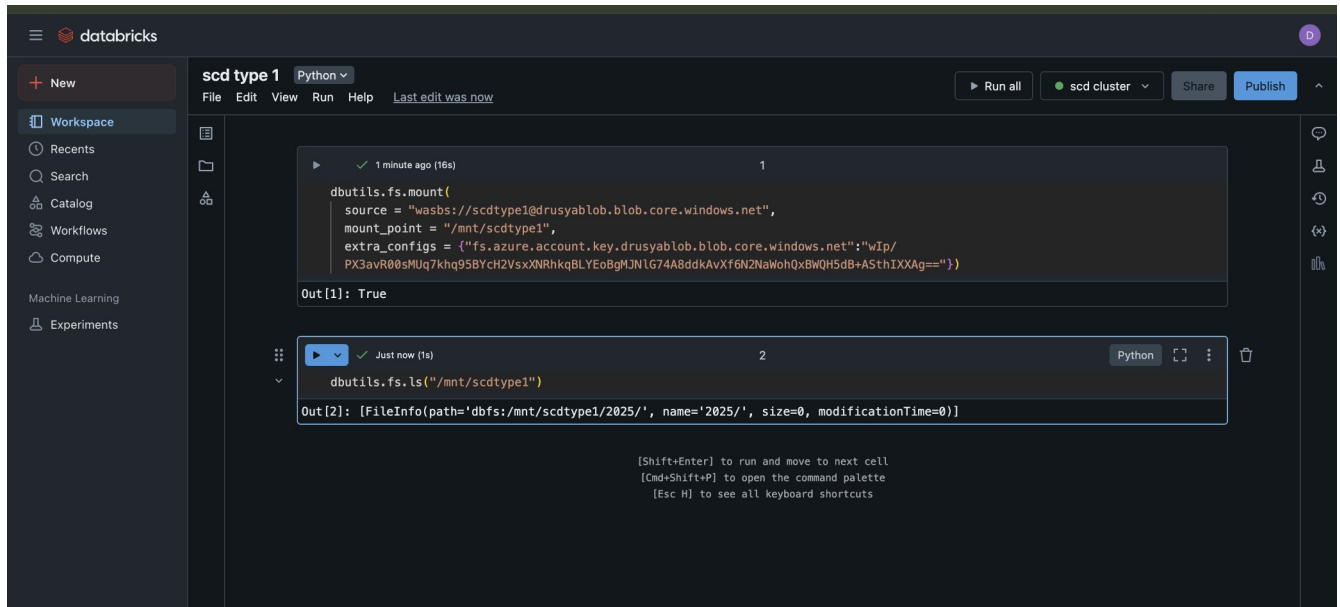
ID	NAME	CITY	PHONENUMBER
1	DRUSYA	KOTTAYAM	2200338812
2	RIYA	ERNAKULAM	2345617890
3	ARIYA	CHEMMAI	3210985432
4	PRIYA	BLR	1092345672

The screenshot shows the Azure portal interface for a container named 'scdtype1'. The left sidebar displays the 'Overview' tab, which includes options for 'Diagnose and solve problems', 'Access Control (IAM)', and 'Settings'. The main content area shows the '2025/03/31/Day2.csv' blob file. The file's metadata is displayed, including the 'Authentication method' (Access key) and 'Location' (scdtype1 / 2025 / 03 / 31). The file is listed in a table with columns for 'Name' and 'Size'. The file 'Day2.csv' is shown with a size of 1 KB. The file is also listed in a table with columns for 'Name' and 'Size'.

ID	NAME	CITY	PHONENUMBER
1	DRUSYA	TRIVANDRUM	2200338812
3	SANIYA	HYD	9901234231

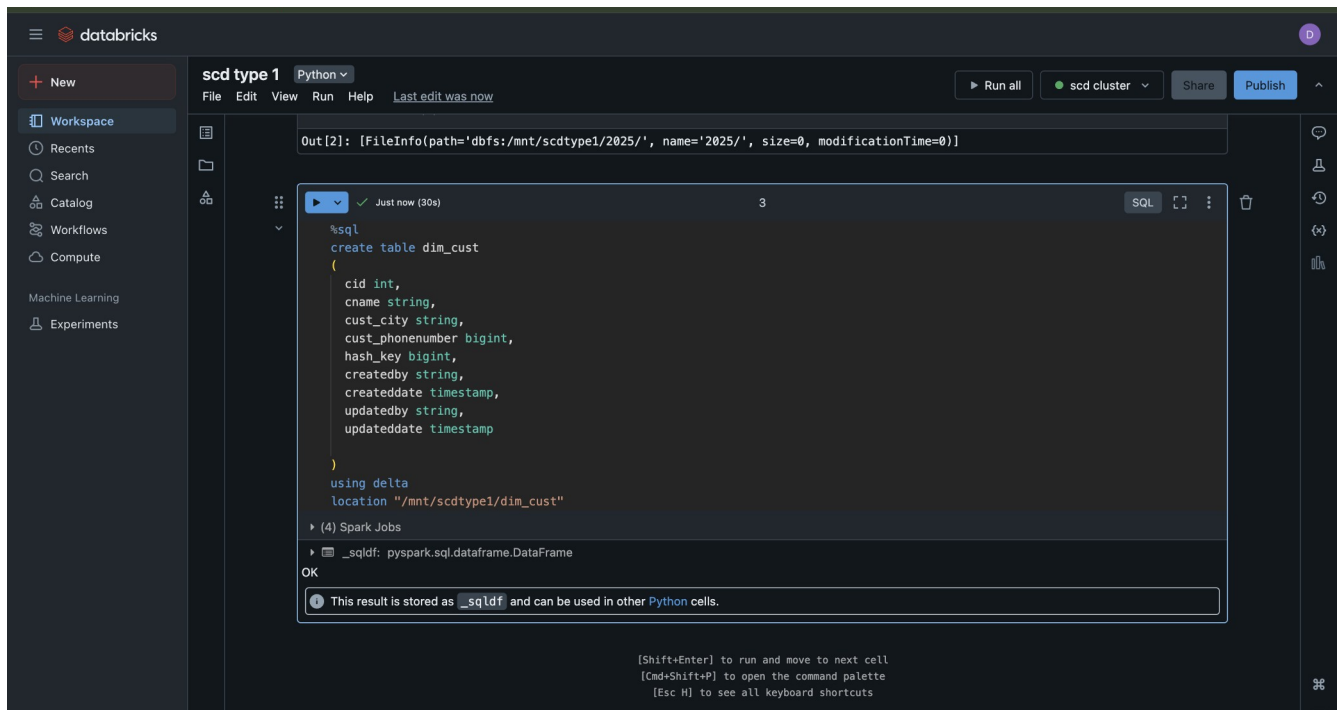
Step 2:

Mounted my storage account using the access key to the databricks community edition my account.



Step 3:

Creating a delta table in my account for storing the data.



Home > drusyablob | Containers >

scdtype1 Container

Search

Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots Create snapshot Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: scdtype1 / dim_cust

Search blobs by prefix (case-sensitive) ☐ Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> [-]						...
<input type="checkbox"/> _delta_log						...
<input type="checkbox"/> _delta_log	3/30/2025, 8:33:43 PM	Hot (Inferred)		Block blob	0 B	Available

Step 4:

Created a widget as file date.

scdtype1 Python

File Edit View Run Help Last edit was now

Run all scd cluster Share Publish

workspace

Recents

Search

Catalog

Workflows

Compute

Machine Learning

Experiments

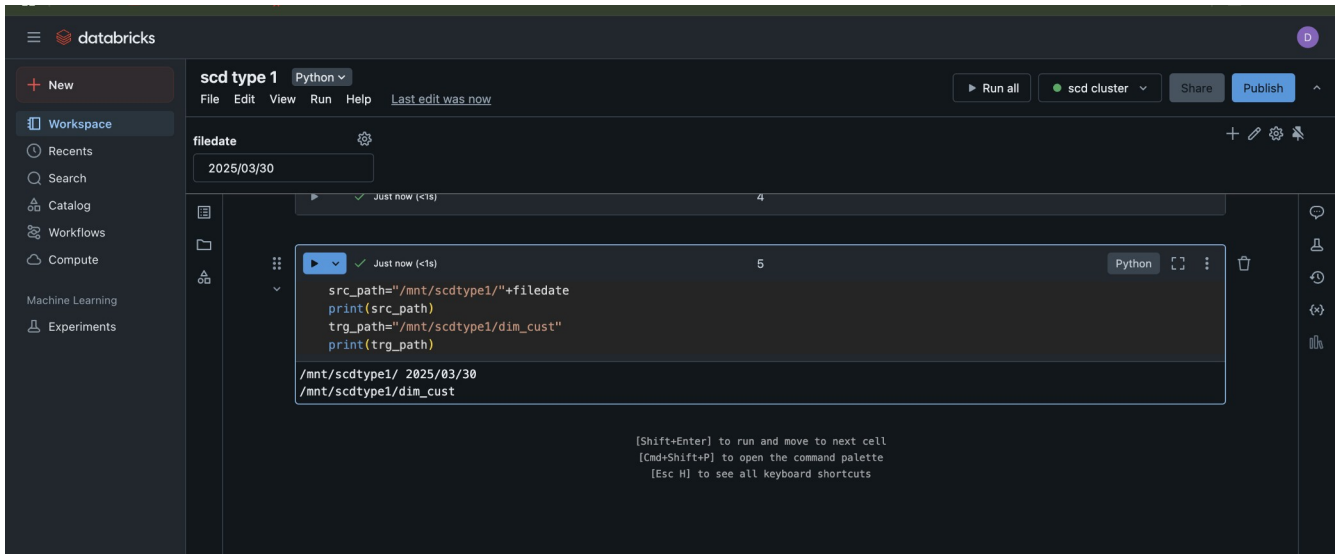
filedate

```
from pyspark.sql.functions import *
dbutils.widgets.text('filedate',' ')
filedate=dbutils.widgets.get('filedate')
```

[Shift+Enter] to run and move to next cell
[Cmd+Shift+P] to open the command palette
[Esc H] to see all keyboard shortcuts

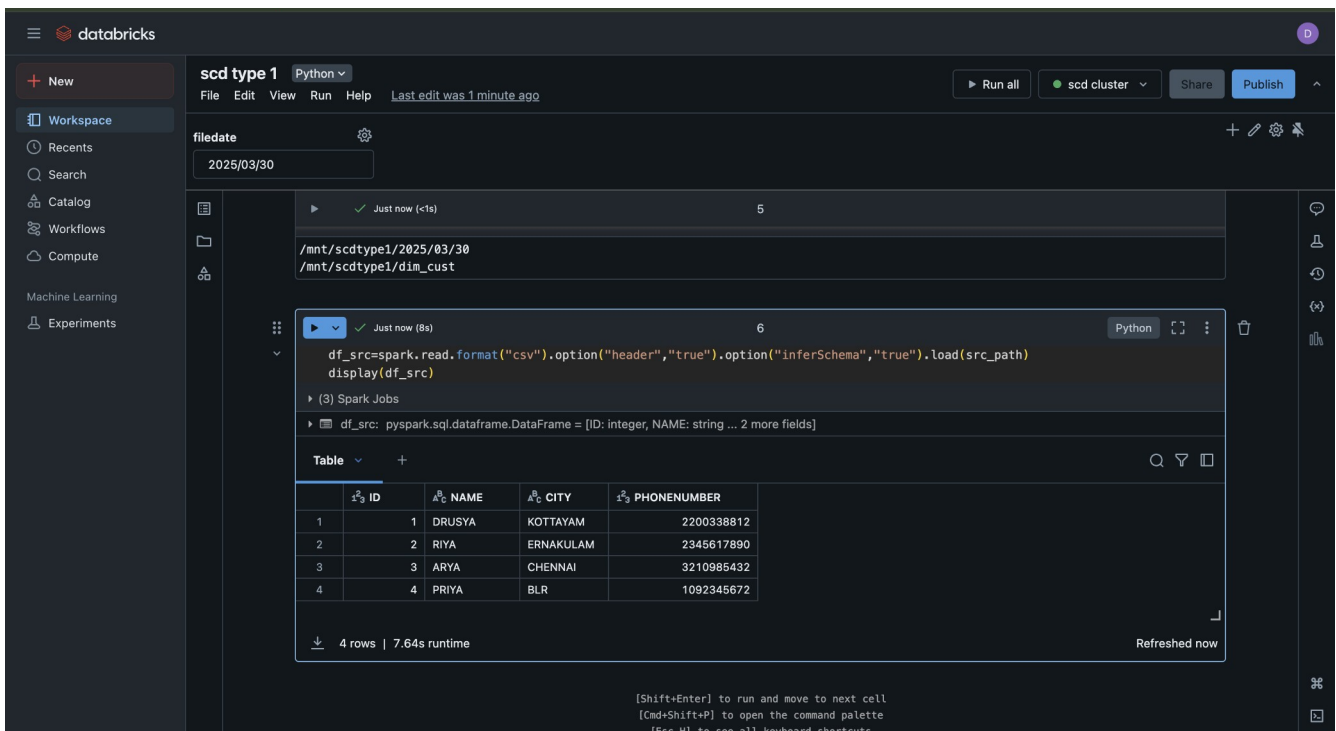
Step 5:

Now I have entered my source path and target path using the file date that I created earlier.



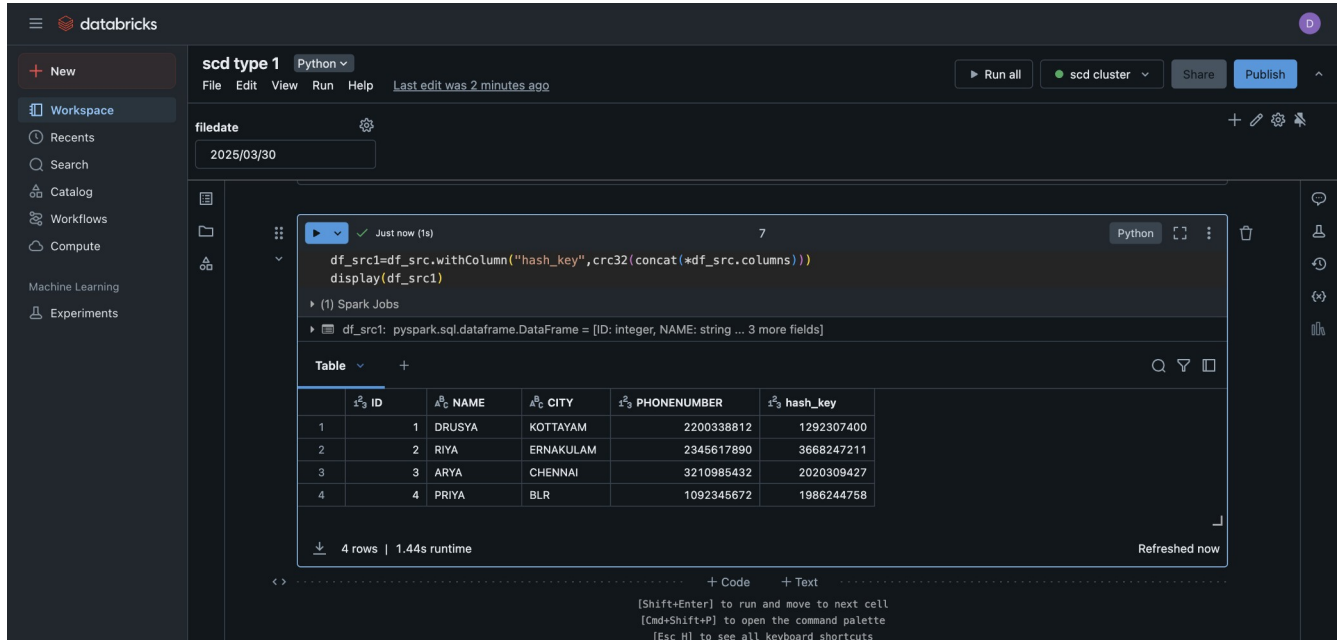
Step 6:

To display my source day 1 file I use the read command.



Step 7:

Now I add a new column haskey so that we can compare the data with other data files to check if it is a new one or updated one, with concating all the column values together to integer. For that I am using the crc32 command.



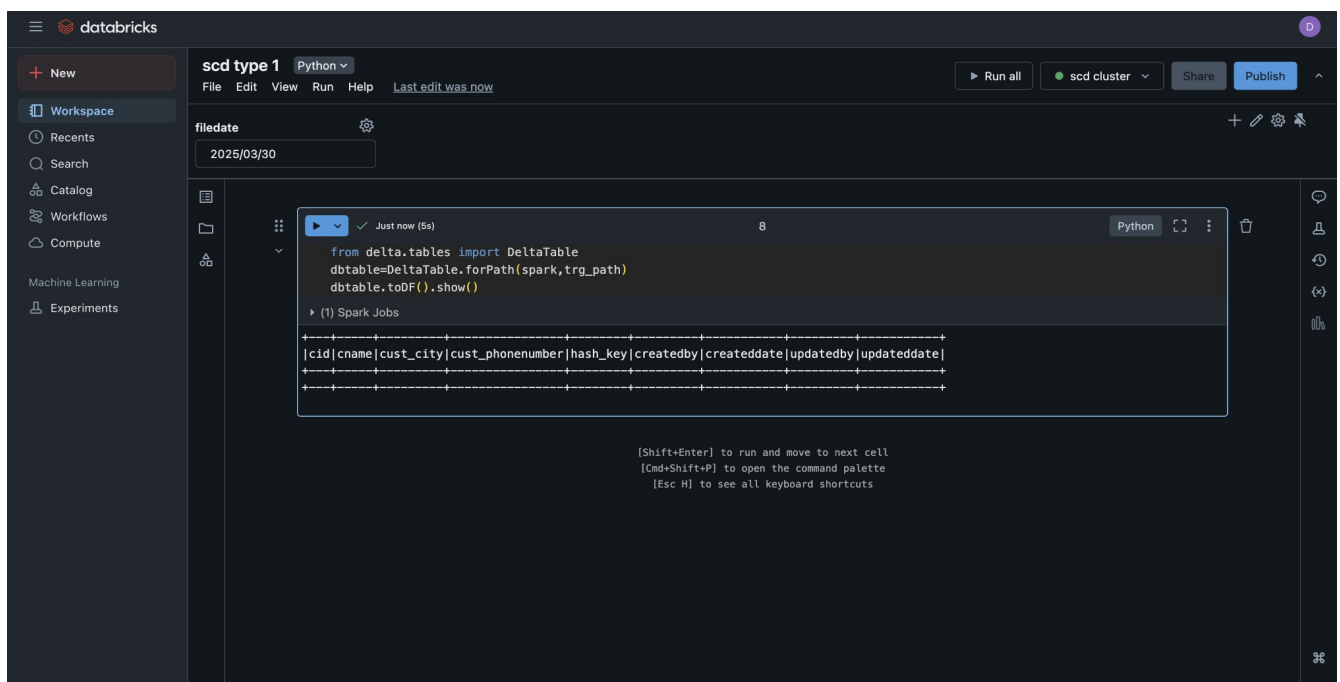
The screenshot shows a Databricks workspace with a Python code cell. The code adds a 'hash_key' column to a DataFrame using the crc32 function on concatenated column values. The output shows a table with 4 rows and 6 columns: ID, NAME, CITY, PHONENUMBER, and hash_key.

```
df_src1=df_src.withColumn("hash_key",crc32(concat(*df_src.columns)))
display(df_src1)
```

ID	NAME	CITY	PHONENUMBER	hash_key
1	DRUSYA	KOTTAYAM	2200338812	1292307400
2	RIYA	ERNAKULAM	2345617890	3668247211
3	ARYA	CHENNAI	3210985432	2020309427
4	PRIYA	BLR	1092345672	1986244758

Step 8:

Now we are converting our target delta table into a dataframe and is displaying it.



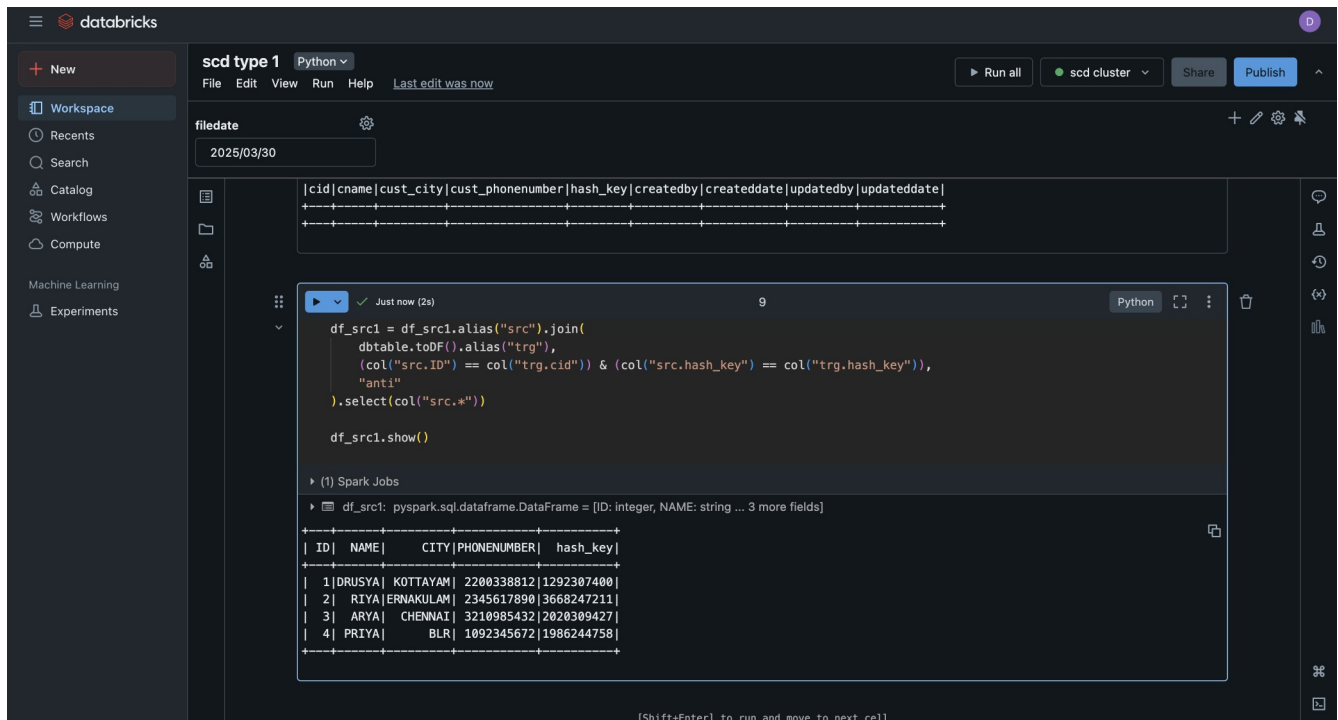
The screenshot shows a Databricks workspace with a Python code cell. The code imports DeltaTable and displays the contents of a target delta table. The output shows a table with 8 columns: cid, cname, cust_city, cust_phonenumber, hash_key, createdby, createddate, and updateddate.

```
from delta.tables import DeltaTable
dtable=DeltaTable.forPath(spark,trg_path)
dtable.toDF().show()
```

cid	cname	cust_city	cust_phonenumber	hash_key	createdby	createddate	updateddate
-----	-------	-----------	------------------	----------	-----------	-------------	-------------

Step 9:

Now I join my source file and target based on the ID so that in the next step I can compare the hashkey and ID to determine it is a insert or update.



The screenshot shows the Databricks workspace interface. The notebook is titled "scd type 1" and is in Python mode. The file date is set to 2025/03/30. The notebook cell contains the following Python code:

```
df_src1 = df_src1.alias("src").join(
    dbtable.toDF().alias("trg"),
    (col("src.ID") == col("trg.cid")) & (col("src.hash_key") == col("trg.hash_key")),
    "anti"
).select(col("src.*"))

df_src1.show()
```

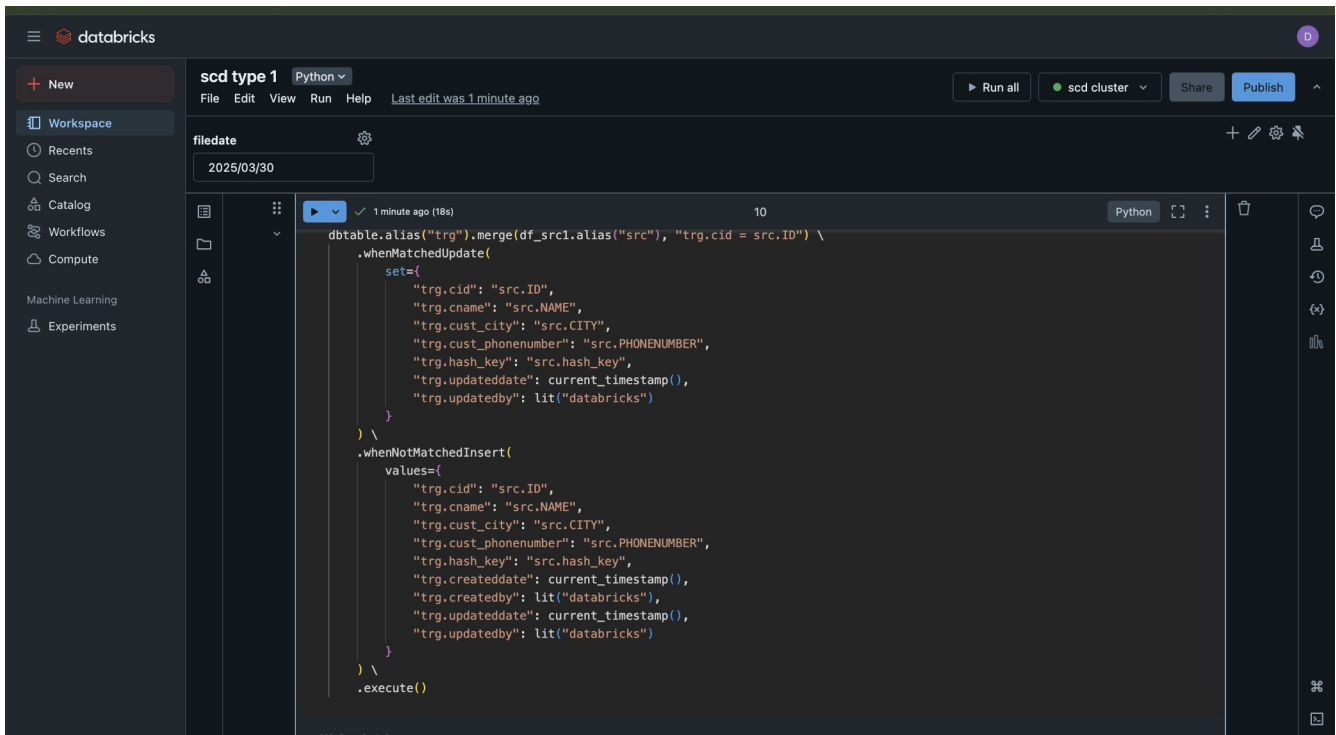
The output of the cell shows the Spark Jobs and the resulting DataFrame:

```
> (1) Spark Jobs
> df_src1: pyspark.sql.dataframe.DataFrame = [ID: integer, NAME: string ... 3 more fields]
```

ID	NAME	CITY	PHONENUMBER	hash_key
1	DRUSYA	KOTTAYAM	2200338812	1292307400
2	RIYA	ERNAKULAM	2345617890	3668247211
3	ARYA	CHENNAI	3210985432	2020309427
4	PRIYA	BLR	1092345672	1986244758

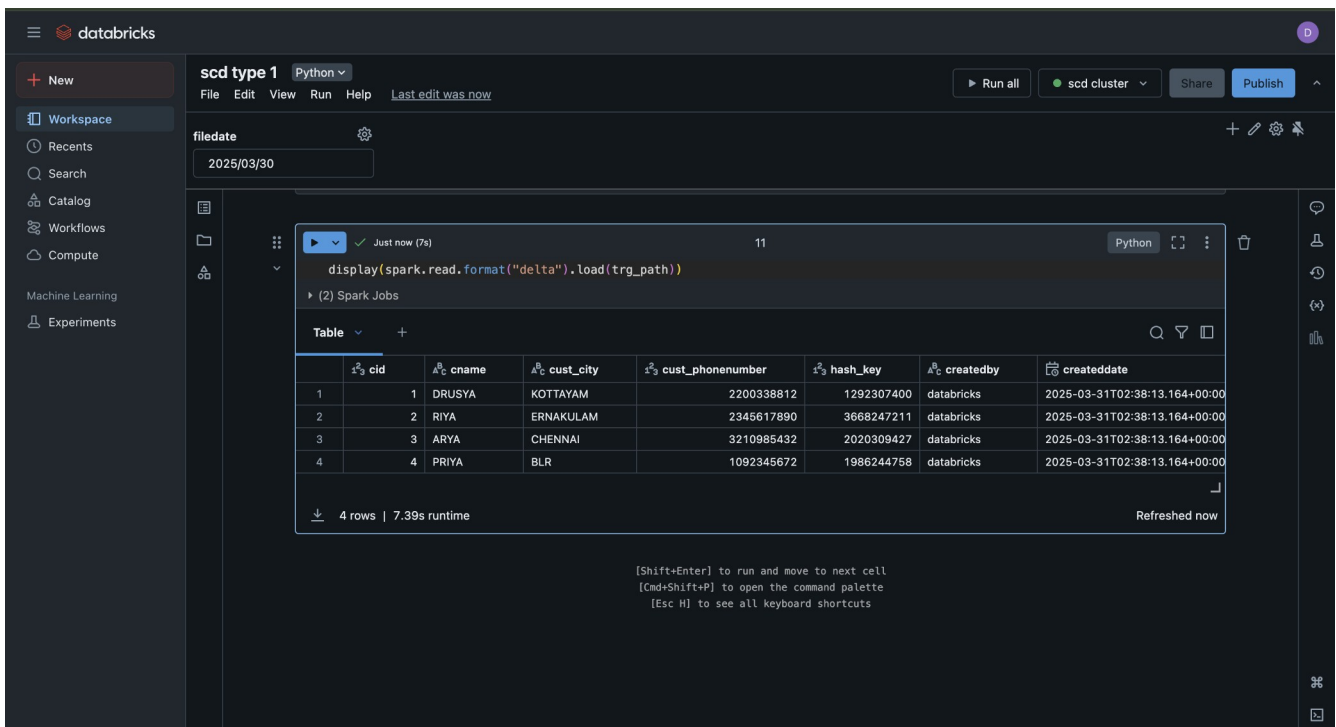
Step 10:

Based on update or insert new records condition merge the files:



Step 11:

Now reading the target file.



Step 12:

Exit the notebook when success.

New

Workspace

Recents

Search

Catalog

Workflows

Compute

Machine Learning

Experiments

scd type 1Python

File Edit View Run Help Last edit was now

Run allscd clusterSharePublish

filedate2025/03/30

(2) Spark Jobs

Table

	cid	cname	cust_city	cust_phonenumber	hash_key	createdby	createddate
1	1	DRUSYA	KOTTAYAM	2200338812	1292307400	databricks	2025-03-31T02:38:13.164+00:00
2	2	RIYA	ERNAKULAM	2345617890	3668247211	databricks	2025-03-31T02:38:13.164+00:00
3	3	ARYA	CHENNAI	3210985432	2020309427	databricks	2025-03-31T02:38:13.164+00:00
4	4	PRIYA	BLR	1092345672	1986244758	databricks	2025-03-31T02:38:13.164+00:00

4 rows | 7.39s runtime

Refreshed 1 minute ago

Just now (<1s)

12

Python

dbutils.notebook.exit("Success")

Notebook exited: Success

[Shift+Enter] to run and move to next cell

[Cmd+Shift+P] to open the command palette

[Esc H] to see all keyboard shortcuts

Step 13:

Passing day 2 file and check the output.

New

Workspace

Recents

Search

Catalog

Workflows

Compute

Machine Learning

Experiments

scd type 1Python

File Edit View Run Help Last edit was 5 minutes ago

Run allscd clusterSharePublish

filedate2025/03/31

1 minute ago (23s)

10

(10) Spark Jobs

Just now (5s)

11

Python

display(spark.read.format("delta").load(trg_path))

(3) Spark Jobs

Table

	cid	cname	cust_city	cust_phonenumber	hash_key	createdby	createddate
1	2	RIYA	ERNAKULAM	2345617890	3668247211	databricks	2025-03-31T02:38:13.164+00:00
2	3	ARYA	CHENNAI	3210985432	2020309427	databricks	2025-03-31T02:38:13.164+00:00
3	4	PRIYA	BLR	1092345672	1986244758	databricks	2025-03-31T02:38:13.164+00:00
4	1	DRUSYA	TRIVANDRUM	2200338812	1810742831	databricks	2025-03-31T02:38:13.164+00:00
5	5	SANIYA	HYD	9901234231	1390187022	databricks	2025-03-31T02:48:29.715+00:00

5 rows | 5.32s runtime

Refreshed now

Skipped

12

Python

dbutils.notebook.exit("Success")

Command skipped