

Custom Learnings

Day 20

Databricks

Structured Streaming: Modes:

1. Append: Optional when not using aggregation.
2. Complete: Mandatory in case of aggregation.
3. Update: On the go will check for the updating if required it will do.

This involves more cost than ADF.

Parquet is columnar format which goes well with big data. But it doesn't support deletion operation.

Delta Table: Whenever we create a table in databricks, delta table is created.

We can do time-travelling in delta table. We can either mimic the table for that time period or rollback to that particular version. Inside the delta.log folder, the json file will store these versions. Vacuum command is used to clean the versions in periodic intervals.

The screenshot shows the Databricks web interface. The top navigation bar includes the Microsoft Azure logo, the Databricks logo, a search bar, and user information. The left sidebar contains a 'New' button and a 'Workspace' section with links to Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, and SQL Warehouses. The main area displays an 'Untitled Notebook' with two commands:

```
Cmd 1
1 dbutils.fs.ls("dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log")

[FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000000.crc', name='00000000000000000000.crc', size=3404, modificationTime=1695788488000),
 FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000000.json', name='00000000000000000000.json', size=2325, modificationTime=1695788488000),
 FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000001.crc', name='00000000000000000001.crc', size=5077, modificationTime=1695788497000),
 FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000001.json', name='00000000000000000001.json', size=2155, modificationTime=1695788495000),
 FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/___tmp_path_dir/', name='___tmp_path_dir/', size=0, modificationTime=1695788488000),
 FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/_copy_into_log/', name='_copy_into_log/', size=0, modificationTime=1695788491000)]

Command took 0.14 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 10:08:00 AM on Shellunext's Cluster
```

```
Cmd 2
1 %sql
2 describe history zipcodes_1_csv

(1) Spark Jobs
  _sqlid: pyspark.sql.dataframe.DataFrame = [version: long, timestamp: timestamp ... 13 more fields]
```

Below the commands, a table view is partially visible with a 'Table' dropdown and a '+' button.

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs

Untitled Notebook 2023-09-27 10:04:20 Python

File Edit View Run Help Last edit was 7 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

```
1 %sql
2 describe history zipcodes_1_csv
```

(1) Spark Jobs

_sqlidf: pyspark.sql.dataframe.DataFrame = [version: long, timestamp: timestamp ... 13 more fields]

Table

	version	timestamp	userId	userName
1	1	2023-09-27T04:21:35.000+0000	4753776711292834	shellunext_1693422208756@npunext.onmicrosoft.com
2	0	2023-09-27T04:21:20.000+0000	4753776711292834	shellunext_1693422208756@npunext.onmicrosoft.com

2 rows | 1.91 seconds runtime Refreshed 5 minutes ago

This result is stored as PySpark data frame _sqlidf and in the IPython output cache as Out[3]. Learn more

Command took 1.91 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 10:09:18 AM on Shellunext's Cluster

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

Untitled Notebook 2023-09-27 10:04:20 Python

File Edit View Run Help Last edit was 7 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

```
1 %sql
2 select * from zipcodes_1_csv
```

(1) Spark Jobs

_sqlidf: pyspark.sql.dataframe.DataFrame = [RecordNumber: integer, Zipcode: integer ... 18 more fields]

Table

	RecordNumber	Zipcode	ZipCodeType	City	State	LocationType	Lat
1	1	704	STANDARD	PARC PARQUE	PR	NOT ACCEPTABLE	17.96
2	2	704	STANDARD	PASEO COSTA DEL SUR	PR	NOT ACCEPTABLE	17.96
3	10	709	STANDARD	BDA SAN LUIS	PR	NOT ACCEPTABLE	18.14
4	61391	76166	UNIQUE	CINGULAR WIRELESS	TX	NOT ACCEPTABLE	32.72
5	61392	76177	STANDARD	FORT WORTH	TX	PRIMARY	32.75
6	61393	76177	STANDARD	FT WORTH	TX	ACCEPTABLE	32.75
7	4	704	STANDARD	IJRR FI IGFINF RICE	PR	NOT ACCEPTABLE	17.96

21 rows | 1.77 seconds runtime Refreshed 3 minutes ago

This result is stored as PySpark data frame _sqlidf and in the IPython output cache as Out[4]. Learn more

Command took 1.77 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 10:11:49 AM on Shellunext's Cluster

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

Untitled Notebook 2023-09-27 10:04:20 Python File Edit View Run Help Last edit was 7 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

Cmd 4

```
1 %sql
2 delete from zipcodes_1_csv where RecordNumber = 61391
```

(6) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [num_affected_rows: long]

num_affected_rows
1

1 row | 4.83 seconds runtime Refreshed 1 minute ago

This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[5]. Learn more

Command took 4.83 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 10:13:30 AM on Shellunext's Cluster

Cmd 5

```
1
```

Shift+Enter to run
Shift+Ctrl+Enter to run selected text

After deleting:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

Untitled Notebook 2023-09-27 10:04:20 Python File Edit View Run Help Last edit was 5 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

Cmd 1

```
1 dbutils.fs.ls("dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log")
```

[FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000000000000000000.crc', name='00000000000000000000000000000000.crc', size=3404, modificationTime=1695788488000),
FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000000000000000000.json', name='00000000000000000000000000000000.json', size=2325, modificationTime=1695788488000),
FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000000000000000001.crc', name='00000000000000000000000000000001.crc', size=5077, modificationTime=1695788497000),
FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000000000000000001.json', name='00000000000000000000000000000001.json', size=2155, modificationTime=1695788495000),
FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000000000000000002.crc', name='00000000000000000000000000000002.crc', size=5076, modificationTime=1695789814000),
FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/00000000000000000000000000000002.json', name='00000000000000000000000000000002.json', size=2852, modificationTime=1695789812000),
FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/_tmp_path_dir/', name='_tmp_path_dir/', size=0, modificationTime=1695788480000),
FileInfo(path='dbfs:/user/hive/warehouse/zipcodes_1_csv/_delta_log/_copy_into_log/', name='_copy_into_log/', size=0, modificationTime=1695788491000)]

Command took 0.16 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 10:18:11 AM on Shellunext's Cluster

Cmd 2

```
1 %sql
2 describe history zipcodes_1_csv
```

(1) Spark Jobs

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion

Untitled Notebook 2023-09-27 10:04:20 Python

File Edit View Run Help Last edit was 5 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

```
1 %sql
2 describe history zipcodes_1_csv
```

(1) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [version: long, timestamp: timestamp ... 13 more fields]

	version	timestamp	userId	userName
1	2	2023-09-27T04:43:32.000+0000	4753776711292834	shellunext_1693422208756@npunext.onmicrosoft.com
2	1	2023-09-27T04:21:35.000+0000	4753776711292834	shellunext_1693422208756@npunext.onmicrosoft.com
3	0	2023-09-27T04:21:20.000+0000	4753776711292834	shellunext_1693422208756@npunext.onmicrosoft.com

3 rows | 1.23 seconds runtime Refreshed now

This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[7]. Learn more

Command took 1.23 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 10:18:49 AM on Shellunext's Cluster

Cmd 3

Dataframe Activity:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

dataframe_activity Python

File Edit View Run Help Last edit was 5 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

Free trial ends in 14 days. Upgrade to Premium in Azure Portal

```
Cmd 1
1 df = spark.read.option("header", True).option("inferSchema", True).csv("/mnt/input/zipcodes.csv")
```

(2) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [RecordNumber: integer, Zipcode: integer ... 18 more fields]

Command took 1.09 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 11:27:54 AM on Shellunext's Cluster

```
Cmd 2
1 df.display()
```

(1) Spark Jobs

	RecordNumber	Zipcode	ZipCodeType	City	State	LocationType	Lat
1	1	704	STANDARD	PARC PARQUE	PR	NOT ACCEPTABLE	17.96
2	2	704	STANDARD	PASEO COSTA DEL SUR	PR	NOT ACCEPTABLE	17.96
3	10	709	STANDARD	BDA SAN LUIS	PR	NOT ACCEPTABLE	18.14
4	61391	76166	UNIQUE	CINGULAR WIRELESS	TX	NOT ACCEPTABLE	32.72
5	61392	76177	STANDARD	FORT WORTH	TX	PRIMARY	32.75
6	61393	76177	STANDARD	FT WORTH	TX	ACCEPTABLE	32.75

The image displays two screenshots from a Microsoft Azure environment. The top screenshot shows the Databricks workspace interface. The left sidebar contains navigation options like Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, and Experiments. The main area shows a notebook titled 'dataframe_activity' with Python code. Command 3 shows a PySpark job saving a table. Command 4 shows a SQL job vacuuming a table. Below the code, a table view shows a single row with the path 'dbfs:/mnt/input/table'. The bottom screenshot shows the Azure Storage 'input' container. The left sidebar has options like Overview, Diagnose and solve problems, Access Control (IAM), Settings, Shared access tokens, Manage ACL, Access policy, Properties, and Metadata. The main area shows the container's contents, including a table with columns: Name, Modified, Access tier, Archive status, and Blob type. The table lists three items: a directory '[.]', a directory '_delta_log', and a file 'part-00000-ee029548-f7aa-4ca2-beff-3017f5a14f4...'.

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

dataframe_activity Python ☆ File Edit View Run Help Last edit was 6 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

Cmd 3

```
1 df.write.option("path", "/mnt/input/table/").saveAsTable("e_zip_tbl")
```

▶ (4) Spark Jobs

Command took 5.61 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 11:29:47 AM on Shellunext's Cluster

Cmd 4

```
1 %sql
2 VACUUM e_zip_tbl RETAIN 169 HOURS
```

▶ (18) Spark Jobs

▶ _sqldf: pyspark.sql.dataframe.DataFrame = [path: string]

path
dbfs:/mnt/input/table

1 row | 50.45 seconds runtime Refreshed 12 minutes ago

This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[5]. Learn more

Command took 50.45 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 11:31:55 AM on Shellunext's Cluster

Cmd 5

Microsoft Azure Search resources, services, and docs (G+/) Shellunext_1693422208... UNEXT (NPUNEXT.ONMICROSOFT...)

Home > idashellsa03 | Containers >

input Container

Search Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview Diagnose and solve problems Access Control (IAM) Settings Shared access tokens Manage ACL Access policy Properties Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: input / table

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

Name	Modified	Access tier	Archive status	Blob type
[.]				
_delta_log				
part-00000-ee029548-f7aa-4ca2-beff-3017f5a14f4...	9/27/2023, 11:29:49 ...	Hot (Inferred)		Block blob

We optimize the partitions so that each partitioned file consists of the same no. of records.

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

dataframe_activity Python ☆ File Edit View Run Help Last edit was 6 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

Cmd 5

```
1 %sql
2 OPTIMIZE e_zip_tbl ZORDER BY (RecordNumber)
```

▶ (4) Spark Jobs

▶ _sqlidf: pyspark.sql.dataframe.DataFrame = [path: string, metrics: struct]

	path	metrics
1	dbfs/mnt/input/table	{ "numFilesAdded": 0, "numFilesRemoved": 0, "filesAdded": { "min": null, "max": null, "avg": 0, "totalFiles": 0, "totalSize": "filesRemoved": { "min": null, "max": null, "avg": 0, "totalSize": 0, "partitionsOptimized": 0, "zOrderStats": { "strategyName": "minCubeSize(107374182400)", "inputCubeFiles": { "num": 0, "size": 0, "inputOtherFiles": { "num": 1, "inputNumCubes": 0, "mergedFiles": { "num": 0, "size": 0, "numOutputCubes": 0, "mergedNumCubes": null, "numBatch": "totalConsideredFiles": 1, "totalFilesSkipped": 1, "preserveInsertionOrder": false, "numFilesSkippedToReduceWriteAmplification": 0, "startTimes": 1695794624517, "endTimes": 1695794624965, "totalClusterParallelism": 4, "totalScheduledTasks": 0, "autoCompactParallelismStats": null, "deletionVectorStats": { "numDeletionVectorsRemoved": 0, "numDeletionVectorRowsRemoved": 0, "numTableColumns": 20, "numTableColumnsWithStats": 20, "totalTaskExecutionTimeMs": 0, "skippedArchivedFiles": 0, "clusteringMetrics": null}} <p>1 row 1.34 seconds runtime Refreshed 12 minutes ago</p> <p>This result is stored as PySpark data frame _sqlidf and in the IPython output cache as Out[6]. Learn more</p> <p>Command took 1.34 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 11:33:43 AM on Shellunext's Cluster</p> <p>Cmd 6</p>

By Partition creating table:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

dataframe_activity Python ☆ File Edit View Run Help Last edit was 5 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

Cmd 6

```
1 df.write.partitionBy("State", "City").option("path", "/mnt/input/table_1/").saveAsTable("ida_zip_tbl")
```

▶ (4) Spark Jobs

Command took 11.63 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 11:41:25 AM on Shellunext's Cluster

Cmd 7

```
1
```

Shift+Enter to run
Shift+Ctrl+Enter to run selected text

Microsoft Azure

Search resources, services, and docs (G+)

Shellunext_1693422208...
UNEXT (NPUNEXT.ONMICROSOFT...)

Home > idashellsa03 | Containers >

input
Container

Search

UploadAdd DirectoryRefreshRenameDeleteChange tierAcquire leaseBreak leaseGive feedback

OverviewDiagnose and solve problemsAccess Control (IAM)

SettingsShared access tokensManage ACLAccess policyPropertiesMetadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: input / table_1

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

Name	Modified	Access tier	Archive status	Blob type
<input type="checkbox"/> [-]				
<input type="checkbox"/> _delta_log				
<input type="checkbox"/> State=AL				
<input type="checkbox"/> State=AZ				
<input type="checkbox"/> State=FL				
<input type="checkbox"/> State=NC				
<input type="checkbox"/> State=PR				
<input type="checkbox"/> State=TX				

Microsoft Azure

Search resources, services, and docs (G+)

Shellunext_1693422208...
UNEXT (NPUNEXT.ONMICROSOFT...)

Home > idashellsa03 | Containers >

input
Container

Search

UploadAdd DirectoryRefreshRenameDeleteChange tierAcquire leaseBreak leaseGive feedback

OverviewDiagnose and solve problemsAccess Control (IAM)

SettingsShared access tokensManage ACLAccess policyPropertiesMetadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: input / table_1 / State=AL

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

Name	Modified	Access tier	Archive status	Blob type
<input type="checkbox"/> [-]				
<input type="checkbox"/> City=SPRING GARDEN				
<input type="checkbox"/> City=SPRINGVILLE				
<input type="checkbox"/> City=SPRUCE PINE				

Restore previous version:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables

dataframe_activity Python ☆ File Edit View Run Help Last edit was 2 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

Cmd 7

```
1 %sql
2 describe history e_zip_tbl
```

(1) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [version: long, timestamp: timestamp ... 13 more fields]

Table +

	version	timestamp	userid	userName
1	2	2023-09-27T06:02:44.000+0000	4753776711292834	shellunext_1693422208756@npunext.onmicrosoft.com
2	1	2023-09-27T06:02:32.000+0000	4753776711292834	shellunext_1693422208756@npunext.onmicrosoft.com
3	0	2023-09-27T05:59:50.000+0000	4753776711292834	shellunext_1693422208756@npunext.onmicrosoft.com

3 rows | 1.38 seconds runtime Refreshed 2 minutes ago

This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[9]. Learn more

Command took 1.38 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 12:09:22 PM on Shellunext's Cluster

Cmd 8

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

dataframe_activity Python ☆ File Edit View Run Help Last edit was 1 minute ago Provide feedback Run all Shellunext's Cluster Schedule Share

Cmd 8

SQL

```
1 %sql
2 select * from e_zip_tbl version as of 0
```

(1) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [RecordNumber: integer, Zipcode: integer ... 18 more fields]

Table +

	RecordNumber	Zipcode	ZipCodeType	City	State	LocationType	Lat
1	1	704	STANDARD	PARC PARQUE	PR	NOT ACCEPTABLE	17.96
2	2	704	STANDARD	PASEO COSTA DEL SUR	PR	NOT ACCEPTABLE	17.96
3	10	709	STANDARD	BDA SAN LUIS	PR	NOT ACCEPTABLE	18.14
4	61391	76166	UNIQUE	CINGULAR WIRELESS	TX	NOT ACCEPTABLE	32.72
5	61392	76177	STANDARD	FORT WORTH	TX	PRIMARY	32.75
6	61393	76177	STANDARD	FT WORTH	TX	ACCEPTABLE	32.75
7	4	704	STANDARD	LIRR FIELDHOUSE	PR	NOT ACCEPTABLE	17.96

21 rows | 0.98 seconds runtime Refreshed 2 minutes ago

This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[10]. Learn more

Command took 0.98 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 12:10:38 PM on Shellunext's Cluster

Cmd 9

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

dataframe_activity Python Run all Shellunext's Cluster Schedule Share

```
1 %sql
2 restore e_zip_tbl version as of 0
```

(9) Spark Jobs

_sqlidf: pyspark.sql.dataframe.DataFrame = [table_size_after_restore: long, num_of_files_after_restore: long ... 4 more fields]

	table_size_after_restore	num_of_files_after_restore	num_removed_files	num_restored_files	removed_files_size
1	6482	1	0	0	0

1 row | 11.84 seconds runtime Refreshed 1 minute ago

This result is stored as PySpark data frame _sqlidf and in the IPython output cache as Out[11]. Learn more

Command took 11.84 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/27/2023, 12:11:09 PM on Shellunext's Cluster

Cmd 10 Python

Shift+Enter to run
Shift+Ctrl+Enter to run selected text

Medallion/ Multi Hop Architecture

We follow three folder architecture:

1. Raw/Bronze: All raw data from all the sources are kept in this folder.
2. Silver: Cleansing and formatting the data files into one particular file format.
3. Gold: Performing some aggregations and joins and keeping the data files needed according to the requirements.

Scheduling the task:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments

Workflows > Jobs > dataframe_notebook Run now

Runs Tasks

dataframe_notebook

...t.com/ida_shell_db/dataframe_activity

+ Add task

Task name* dataframe_notebook

Type* Notebook

Source* Workspace

Job details

Job ID 560098516364433

Creator Shellunext

Run as Shellunext

Tags +Tag

Git

Not configured

Add Git settings

Schedule

Every minute

Edit schedule Pause Delete

Compute

dataframe_notebook_cluster

Driver: Standard_DS3_v2 - Worker: Standard_DS3_v2 - 0

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladb shellunext_1693422208756@npun...

New

- Workspace
- Recents
- Catalog
- Workflows
- Compute

SQL

- SQL Editor
- Queries
- Dashboards
- Alerts
- Query History
- SQL Warehouses

Data Engineering

- Job Runs**
- Data Ingestion
- Delta Live Tables

Machine Learning

- Experiments

Workflows

Free trial ends in 14 days. Upgrade to Premium in Azure Portal

Jobs Job runs Delta Live Tables

☐ Only my job runs Run status 2023-09-25 14:30:00 → 2023-09-27 14:30:00 Create job

1

1

0 26 Sep, 12 AM 26 Sep, 12 PM 27 Sep, 12 AM 27 Sep, 12 PM

Failed Skipped Succeeded

Start time	Job	Run as	Launched	Duration	Status	Run parameters
Sep 27, 2023, 01:47 PM	dataframe_noteb...	Shellunext	By scheduler	2s	Skipped	
Sep 27, 2023, 01:46 PM	dataframe_noteb...	Shellunext	By scheduler	1m 25s	Pending	
Sep 27, 2023, 01:43 PM	dataframe_noteb...	Shellunext	By scheduler	2s	Skipped	
Sep 27, 2023, 01:42 PM	dataframe_noteb...	Shellunext	By scheduler	2s	Skipped	
Sep 27, 2023, 01:41 PM	dataframe_noteb...	Shellunext	By scheduler	3s	Skipped	

< Previous Next >

Reading the data from azure sql:

AzureSQL_Connection_Notebook Python ☆

File Edit View Run Help Last edit was 26 minutes ago Provide feedback ▶ Run all ● Shellunext's Cluster ⌄ Schedule Share ^

Cmd 1

```
1 jdbc_url = "jdbc:sqlserver://idaservernew.database.windows.net:1433; databaseName=idadb02"
2 connection_properties = {
3     "user": "sqladmin",
4     "password": "IDAshe11@123",
5     "driver": "com.microsoft.sqlserver.jdbc.SQLServerDriver"
6 }
```

Command took 0.11 seconds -- by shellunext_169342229799@npunext.onmicrosoft.com at 9/27/2023, 2:21:03 PM on Shellunext's Cluster

Cmd 2

```
1 df = spark.read.jdbc(url=jdbc_url, table = "SalesLT.Customer",
2                       properties=connection_properties)
```

df: pyspark.sql.dataframe.DataFrame = [CustomerID: integer, NameStyle: boolean ... 13 more fields]

Command took 2.39 seconds -- by shellunext_169342229799@npunext.onmicrosoft.com at 9/27/2023, 2:21:11 PM on Shellunext's Cluster

Cmd 3

```
1 df.display()
```

(1) Spark Jobs

Table +

- For the same workspace, there can be multiple users that work on different notebooks.
- For development, production, UAT env there are different workspaces.
- To share notebooks present under different users in different workspaces, the admin must provide access separately to every user for any file present in different workspaces.
- This is a very inefficient and time-consuming process.
- With unity catalog, workspace sharing can be enabled.
- This provides data access without needing admin permission.
- Azure Access Connector is used to provide connection to different workspace files.