raw

Container

Search

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

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

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

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
population_by_age.tsv	11/14/2024, 7:31:03 ...	Hot (Inferred)		Block blob	25.77 KiB	Available

Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

dbsaibdp

New

Workspace
Recents
Catalog
Workflows
Compute

SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses

Data Engineering
Job Runs
Data Ingestion
Delta Live Tables

Machine Learning

Untitled Notebook 2024-11-14 19:37:21

Python

File Edit View Run Help Last edit was 3 minutes ago

Run all | Riya Kashyap's Cluster | Schedule | Share

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

07:42 PM (<1s) 1

```
from pyspark.sql import functions as F
```

07:37 PM (1s) 2

```
spark.conf.set(
    "fs.azure.account.key.saibdp.dfs.core.windows.net",
    "masPB9KN5d+Chc72FEZ82AyYyDPWR/rStH2n58VsN3qL7eG7yBjJI9c2YLSPLXET5FDcXCAAmSFU+AStScYfrw==")
```

07:39 PM (1s) 3

```
dbutils.fs.ls("abfss://raw@saibdp.dfs.core.windows.net/population_by_age.tsv")
```

```
[FileInfo(path='abfss://raw@saibdp.dfs.core.windows.net/population_by_age.tsv', name='population_by_age.tsv', size=26387, modificationTime=1731630663000)]
```

07:41 PM (10s) 4

```
df = spark.read.csv('abfss://raw@saibdp.dfs.core.windows.net/population_by_age.tsv',sep='\t',header= True)
```

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

Untitled Notebook 2024-11-14 19:37:21

Python

File Edit View Run Help Last edit was 4 minutes ago

Run all | Riya Kashyap's Cluster | Schedule | Share

09:38 PM (<1s) 8

```
df1.createOrReplaceTempView("population_data")

# Use SQL to filter and clean the data
result_df = spark.sql("""
    SELECT REGEXP_REPLACE(age_group,'^PC_', '') AS age_group,
           country_code,
           regexp_replace(population_2019, '[^0-9.]', '') AS population_2019
    FROM population_data
    WHERE length(country_code) <= 2
""")
display(result_df)
```

(1) Spark Jobs

result_df: pyspark.sql.dataframe.DataFrame = [age_group: string, country_code: string ... 1 more field]

Table

	age_group	country_code	population_2019
21	Y0_14	HR	14
22	Y0_14	HU	14
23	Y0_14	IE	20
24	Y0_14	IS	19
25	Y0_14	IT	13
26	Y0_14	LI	14

Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

dbsaibdp

New

Workspace
Recents
Catalog
Workflows
Compute

SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses

Data Engineering
Job Runs
Data Ingestion
Delta Live Tables

Machine Learning

Untitled Notebook 2024-11-14 19:37:21

Python

File Edit View Run Help Last edit was 4 minutes ago

Run all | Riya Kashyap's Cluster | Schedule | Share

09:38 PM (1s) 10

```
# Pivot the DataFrame
pivot_df = df1.groupBy("country_code").pivot("age_group").sum("population_2019")

# Show the pivoted DataFrame
display(pivot_df)
```

(7) Spark Jobs

pivot_df: pyspark.sql.dataframe.DataFrame = [country_code: string, Y0_14: long ... 5 more fields]

Table

	country_code	Y0_14	Y15_24	Y25_49	Y50_64	Y65_79	Y80_MAX
34	GE	20	11	33	19	11	3
35	SK	15	10	38	19	12	3
36	UK	17	11	32	19	13	5
37	LV	15	9	33	20	14	5
38	XK	24	17	35	14	6	1
39	HU	14	10	35	19	14	4
40	CY	16	12	37	17	12	3
41	SI	15	9	34	21	14	5
42	IE	20	12	35	17	10	3
43	EL	14	10	33	20	14	7
44	BE	16	11	32	20	13	5

Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

dbsaibdp

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Untitled Notebook 2024-11-14 19:37:21

Python

File Edit View Run Help Last edit was 4 minutes ago

Run all | Riya Kashyap's Cluster | Schedule | Share

46	LU	16	11	38	19	10	4
47	IS	19	13	35	17	10	3
48	DK	16	12	31	19	15	4

48 rows | 1.16 seconds runtime

Refreshed 6 minutes ago

4 minutes ago (2s) 11 Python

```
result_df.write.json("abfss://processed@saibdp.dfs.core.windows.net/population_by_age.json")
```

(1) Spark Jobs

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

processed

Container

Search

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

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

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: processed / population_by_age.json

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
_committed_4314253494006253260	11/14/2024, 9:41:03 ...	Hot (Inferred)		Block blob	113 B	Available
_started_4314253494006253260	11/14/2024, 9:41:03 ...	Hot (Inferred)		Block blob	0 B	Available
_SUCCESS	11/14/2024, 9:41:03 ...	Hot (Inferred)		Block blob	0 B	Available
part-00000-tid-4314253494006253260-d9bfc19c-...	11/14/2024, 9:41:03 ...	Hot (Inferred)		Block blob	17.97 KiB	Available