Create an Azure Databricks workspace

...

Basics | Networking | Encryption | Tags | Review + create

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * : Azure for Students

Resource group * : covid-reporting-rg

Create new

Instance Details

Workspace name * : covid-reporting-dbk-ws

Region * : UK South

Pricing Tier * : Standard (Apache Spark, Secure with Azure AD)

Managed Resource Group name : Enter name for managed resource group

Review + create | < Previous | Next : Networking >

Compute › New compute › UI preview Provide feedback

covid-reporting-cluster

UI | JSON

Multi node Single node

Access mode

Single user

Single user access

Chinta, Avinash

Performance

Databricks runtime version

Runtime: 12.2 LTS (Scala 2.12, Spark 3.3.2)

Use Photon Acceleration

Node type

Standard_D4a_v4 16 GB Memory, 4 Cores

Terminate after 30 minutes of inactivity

Tags

Add tags

Key Value Add

› Automatically added tags

Create Cluster Cancel

Summary

1 Driver 16 GB Memory, 4 Cores

Runtime 12.2.x-scala2.12

Standard_D4a_v4 0.75 DBU/h

covid-reporting-cluster

More ··· | Terminate | Edit

UI | JSON

Multi node | Single node

Access mode | Single user access

Single user | Chinta, Avinash

Performance

Databricks Runtime Version

12.2 LTS (includes Apache Spark 3.3.2, Scala 2.12)

Use Photon Acceleration

Node type

Standard_DS3_v2 14 GB Memory, 4 Cores

Terminate after 30 minutes of inactivity

Tags

No custom tags

> Automatically added tags

Advanced options

Summary

1 Driver 14 GB Memory, 4 Cores

Runtime 12.2.x-scala2.12

Standard_DS3_v2 | 0.75 DBU/h

mount_storage Python

File Edit View Run Help Last edit was 7 minutes ago Provide feedback

Run all | covid-reporting-cluster | Schedule | Share

```
#Linking the datalake storage to workspace by using access keys
```

Cmd 4

Python

```
storage_account_name = "ercovidreportingdatalake"
storage_account_key = "9S43wACOwOhKl9Nsu6USZtFKuAo5kmjyVZA+qh30QVZPWxUVRYqmR5wBsyaovbVDMX7ZrRk8qxlD+ASt3rQO/A=="

spark.conf.set(
    f"fs.azure.account.key.ercovidreportingdatalake.dfs.core.windows.net",
    f"9S43wACOwOhKl9Nsu6USZtFKuAo5kmjyVZA+qh30QVZPWxUVRYqmR5wBsyaovbVDMX7ZrRk8qxlD+ASt3rQO/A==")

container_name = "raw" # make sure to use the right container
df_raw_population = spark.read.csv(f"abfss://raw@ercovidreportingdatalake.dfs.core.windows.net/population", sep=r'\t', header=True)
```

▸ (1) Spark Jobs

▸ df_raw_population: pyspark.sql.dataframe.DataFrame = [indic_de,geo\time: string, 2008 : string ... 11 more fields]

Command took 8.63 seconds -- by avinashchinta@my.unt.edu at 7/2/2023, 1:11:24 PM on covid-reporting-cluster

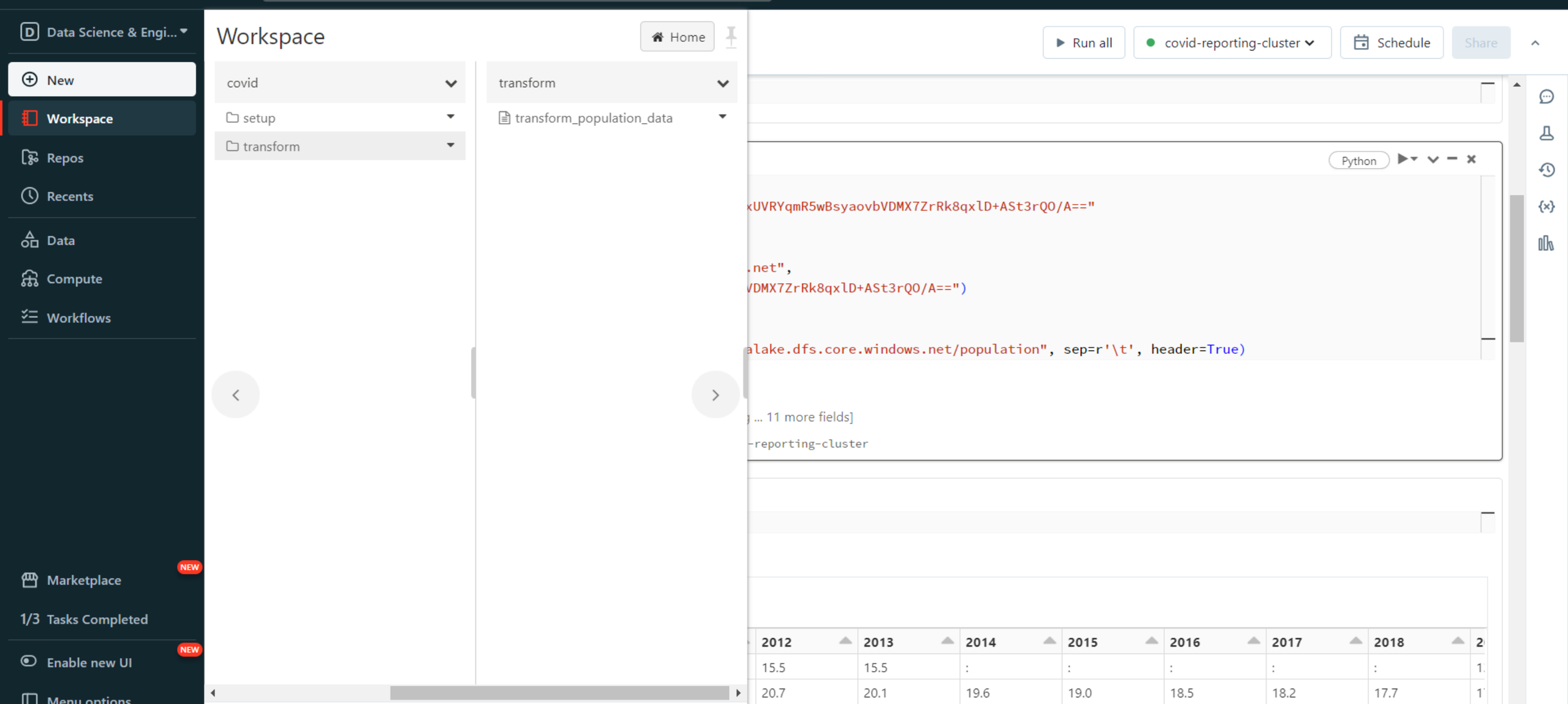
Cmd 5

```
display(df_raw_population.head(5))
```

▸ (3) Spark Jobs

Table +

	indic_de,geo\time	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2
1	PC_Y0_14,AD	14.6	14.5	14.5	15.5	15.5	15.5	:	:	:	:	:	1
2	PC_Y0_14,AL	24.1	23.3	22.5	21.6	20.7	20.1	19.6	19.0	18.5	18.2	17.7	1
3	PC_Y0_14,AM	19.0	18.6	18.3	:	:	:	:	19.4	19.6	20.0	20.2	2


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
NEW
1/3 Tasks Completed
Enable new UI
NEW
Menu options
Workspace
Home
covid
setup
transform
transform
transform_population_data
Run all
covid-reporting-cluster
Schedule
Share
Python
xUVRYqmR5wBsyaovbVDMX7ZrRk8qxlD+ASt3rQO/A=="
.net",
VDMX7ZrRk8qxlD+ASt3rQO/A==")
alake.dfs.core.windows.net/population", sep=r'\t', header=True)
... 11 more fields]
-reporting-cluster
2012
2013
2014
2015
2016
2017
2018
15.5
15.5
:
:
:
:
:
20.7
20.1
19.6
19.0
18.5
18.2
17.7

transform_population_data Python

File Edit View Run Help Last edit was 28 minutes ago Provide feedback

Run all Connect Schedule Share

Replace storage account name with your storage account name before executing.

Cmd 3

```
from pyspark.sql.functions import *
```

Cmd 4

Connect datalake storage and read csv file

Cmd 5

```
storage_account_name = "ercovidreportingdatalake"
storage_account_key = "9S43wACOwOhKl9Nsu6USZtFKuAo5kmjyVZA+qh30QVZPWxUVRYqmR5wBsyaovbVDMX7ZrRk8qxlD+ASt3rQO/A=="

spark.conf.set(
    f"fs.azure.account.key.ercovidreportingdatalake.dfs.core.windows.net",
    f"9S43wACOwOhKl9Nsu6USZtFKuAo5kmjyVZA+qh30QVZPWxUVRYqmR5wBsyaovbVDMX7ZrRk8qxlD+ASt3rQO/A==")

container_name = "raw" # make sure to use the right container
df_raw_population = spark.read.csv(f"abfss://raw@ercovidreportingdatalake.dfs.core.windows.net/population", sep=r'\t', header=True)
```

Cmd 6

Read the population data & create a temp view

Cmd 7

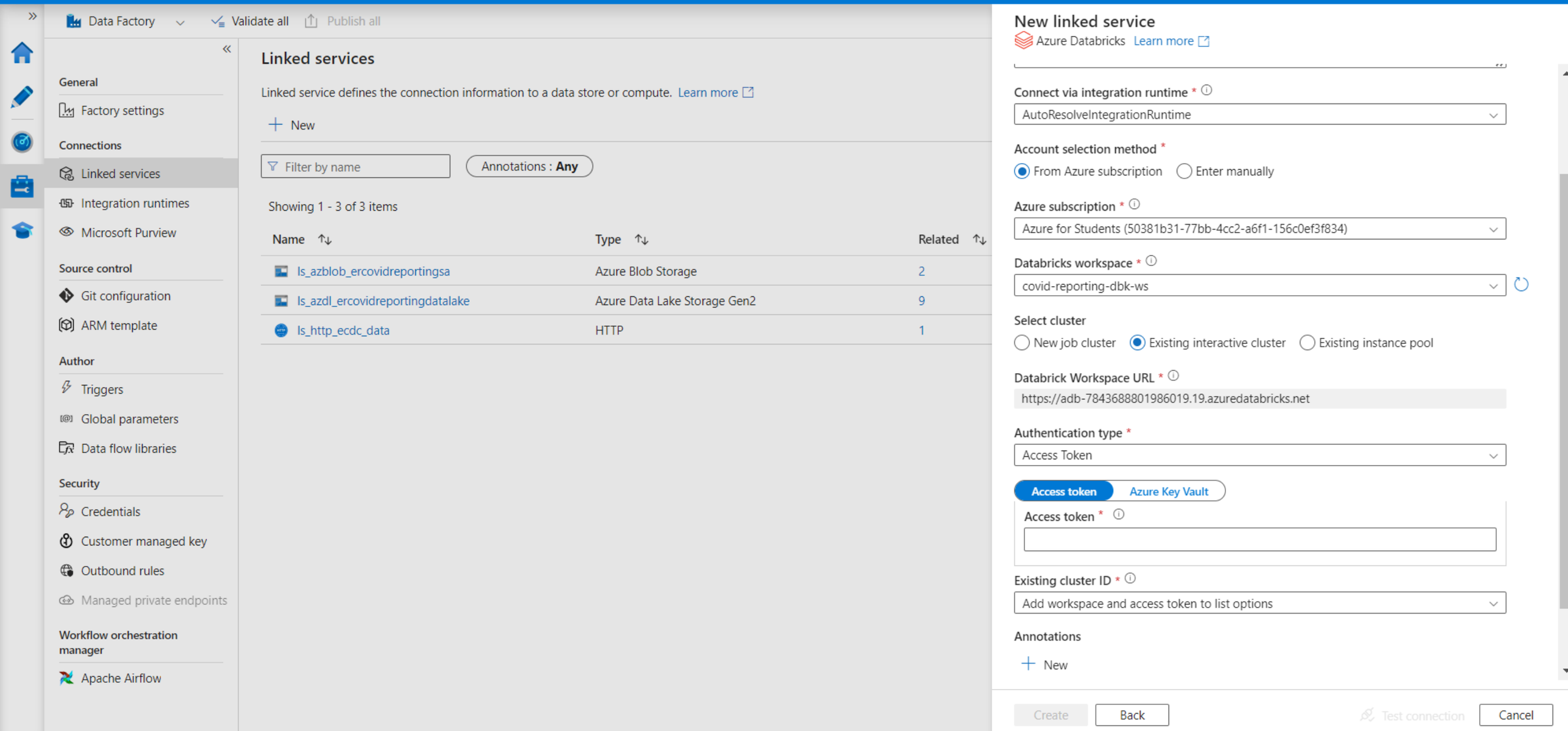
```

df_raw_population = df_raw_population.withColumn('age_group', regexp_replace(split(df_raw_population['indic_de,geo\\time'], ',')[0], 'PC_', '')).withColumn('country_code', split(df_raw_population['indic_de,geo\\time'], ',')[1])
df_raw_population = df_raw_population.select(col("country_code").alias("country_code"),
                                             col("age_group").alias("age_group"),
                                             col("2019 ").alias("percentage_2019"))
df_raw_population.createOrReplaceTempView("raw_population")
```

Cmd 8

Pivot the data by age group

Cmd 9

```
# Create a data frame with pivoted percentages
df_raw_population_pivot = spark.sql("SELECT country_code, age_group, cast(regexp_replace(percentage_2019, '[a-z]', '') AS decimal(4,2)) AS percentage_2019 FROM raw_population WHERE length(country_code) = 2").groupBy("country_code").pivot("age_group").sum
("percentage_2019").orderBy("country_code")
df_raw_population_pivot.createOrReplaceTempView("raw_population_pivot")
```


Data Factory
Validate all
Publish all
General
Factory settings
Connections
Linked services
Integration runtimes
Microsoft Purview
Source control
Git configuration
ARM template
Author
Triggers
Global parameters
Data flow libraries
Security
Credentials
Customer managed key
Outbound rules
Managed private endpoints
Workflow orchestration manager
Apache Airflow
Linked services
Linked service defines the connection information to a data store or compute. Learn more
New
Filter by name
Annotations : Any
Showing 1 - 3 of 3 items
Name
Type
Related
ls_azblob_ercovidreportingsa
Azure Blob Storage
2
ls_azdl_ercovidreportingdatalake
Azure Data Lake Storage Gen2
9
ls_http_ecdc_data
HTTP
1
New linked service
Azure Databricks Learn more
Connect via integration runtime *
AutoResolveIntegrationRuntime
Account selection method *
From Azure subscription
Enter manually
Azure subscription *
Azure for Students (50381b31-77bb-4cc2-a6f1-156c0ef3f834)
Databricks workspace *
covid-reporting-dbk-ws
Select cluster
New job cluster
Existing interactive cluster
Existing instance pool
Databrick Workspace URL *
https://adb-7843688801986019.19.azuredatabricks.net
Authentication type *
Access Token
Access token
Azure Key Vault
Access token *
Existing cluster ID *
Add workspace and access token to list options
Annotations
New
Create
Back
Test connection
Cancel

User Settings

Access tokens | Git integration | Editor settings | Email preferences | Language settings

Personal access tokens can be used for secure authentication to the Databricks API instead of passwords.

Generate new token

Comment	Creation	Expiration
demo	2023-07-02 13:54:41 CDT	2023-07-22 13:54:41 CDT

1/3

Data Factory | Validate all | Publish all

General
- Factory settings

Connections
- Linked services
- Integration runtimes
- Microsoft Purview

Source control
- Git configuration
- ARM template

Author
- Triggers
- Global parameters
- Data flow libraries

Security
- Credentials
- Customer managed key
- Outbound rules
- Managed private endpoints

Workflow orchestration manager
- Apache Airflow

Linked services

Linked service defines the connection information to a data store or compute. Learn more

+ New

Filter by name | Annotations : **Any**

Showing 1 - 3 of 3 items

Name	Type	Related
ls_azblob_ercovidreportingsa	Azure Blob Storage	2
ls_azdl_ercovidreportingdatalake	Azure Data Lake Storage Gen2	9
ls_http_ecdc_data	HTTP	1

New linked service

Azure Databricks Learn more

Account selection method *
- (•) From Azure subscription
- ( ) Enter manually

Azure subscription * (i)
Azure for Students (50381b31-77bb-4cc2-a6f1-156c0ef3f834)

Databricks workspace * (i)
covid-reporting-dbk-ws

Select cluster
- ( ) New job cluster
- (•) Existing interactive cluster
- ( ) Existing instance pool

Databrick Workspace URL * (i)
https://adb-7843688801986019.19.azuredatabricks.net

Authentication type *
Access Token

Access token | Azure Key Vault

Access token * (i)
••••••••••••••••••••••••••••••••••••

Choose from existing clusters * (i)
covid-reporting-cluster

Annotations
+ New

> Parameters

> Advanced (i)

Create | Back | Test connection | Cancel

Data Factory | Validate all | Publish all 1 | Preview experience | Off

Factory Resources

Filter resources by name

- Pipelines 5
  - pl_process_pop_data
  - pl_cases_deaths_data_processed
  - pl_ingest_ecdc_data
  - pl_ingest_population_data
  - pl_processed_hospitaladmissions_data
- Change Data Capture (preview) 0
- Datasets 12
- Data flows 2
- Power Query 0

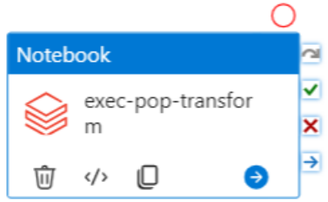
pl_process_pop_data

Activities

Search activities

- Move & transform
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
  - Notebook
  - Jar
  - Python
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Validate | Debug | Add trigger



General | Azure Databricks | Settings | User properties

Databricks linked service * ls_dbk_covid_cluster | Test connection | Edit | New

Properties

General | Related

Name * pl_process_pop_data

Description

Annotations

New

Data Factory | Validate all | Publish all 1 | Preview experience | Off

Factory Resources

Filter resources by name

- Pipelines 5
  - pl_process_pop_data
  - pl_cases_deaths_data_processed
  - pl_ingest_ecdc_data
  - pl_ingest_population_data
  - pl_processed_hospitaladmissions_data
- Change Data Capture (preview) 0
- Datasets 12
- Data flows 2
- Power Query 0

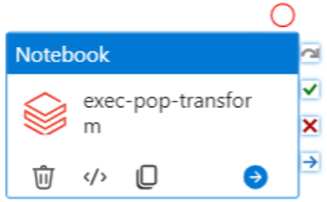
pl_process_pop_data

Activities

Search activities

- Move & transform
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
  - Notebook
  - Jar
  - Python
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Validate | Debug | Add trigger



General | Azure Databricks | Settings | User properties

Notebook path * /covid/transform/transform_population_... Browse Open

Base parameters

Append libraries

Properties

General | Related

Name * pl_process_pop_data

Description

Annotations

New

Dashboards

Runs

Pipeline runs

Trigger runs

Change Data Capture (previ...

Runtimes & sessions

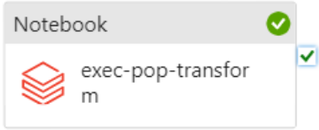
Integration runtimes

Data flow debug

Notifications

Alerts & metrics

Rerun | Refresh | Update pipeline | List | Gantt



Activity runs

Pipeline run ID 6d31e9d6-80bf-4e21-98aa-932caa574725

All status | Export to CSV

Showing 1 - 1 items

Activity name	Status	Activity type	Run start	Duration	Log	Integration runtime	User properties	Activity run ID
exec-pop-transform	Succeeded	Notebook	7/2/2023, 2:06:53 PM	35s		AutoResolveIntegratior		3cd97361-9c10-4c35-af0c-9ec76

Microsoft Azure Storage Explorer

File Edit View Help

EXPLORER

Search for resources

Collapse all Refresh all

- Quick Access
- Emulator & Attached
  - Storage Accounts
- Azure for Students (AvinashChinta@my.unt.edu
  - Storage Accounts
    - cs71003200143e43ad6
    - dbstorageocb4jzhwww7hy (ADLS Ger
    - dlspractise (ADLS Gen2)
    - ercovidreportingdatalake (ADLS Gen2
      - Blob Containers
        - $logs
        - lookup
        - processed
        - raw
      - File Shares
      - Queues
      - Tables
    - ercovidreportingsa
      - Blob Containers
        - $logs
        - configs
        - population
        - scripts
      - File Shares
      - Queues
      - Tables
  - Disks
    - cloud-shell-storage-southcentralus
    - covid-reporting-rg
    - databricks-rg-dbs-handson-bj47lln3b
    - NetworkWatcherRG
    - rg-avinash-handson

processed | $logs | population | configs | lookup | scripts

Upload | Download | Open | Preview | New Folder | Select All | Copy | Paste | Clone | Rename | Move | Manage ACLs | Properties | Delete | Undelete | Folder Statistics | Refresh

Active blobs (default) | processed > population | Filter by prefix (case-sensitive)

Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size
_SUCCESS	Hot (inferred)		7/2/2023 2:07 PM	Block Blob	application/octet-stream	0 B
_committed_210491613074435726	Hot (inferred)		7/2/2023 2:07 PM	Block Blob	application/octet-stream	111 B
_started_210491613074435726	Hot (inferred)		7/2/2023 2:07 PM	Block Blob	application/octet-stream	0 B
part-00000-tid-210491613074435726-82273a7e-bfa2-4a29-b412-838dae369b66-16-1-c000.csv	Hot (inferred)		7/2/2023 2:07 PM	Block Blob	application/octet-stream	2.70 K

Showing 1 to 4 of 4 cached items

Activities

Clear completed Clear successful

Transfer of 'C:\Users\Dheeraji Bajjuri\Downloads\testing\testing\000000_0' to 'processed/ecdc/testing/' complete: 1 item transferred (used SAS, discovery completed) Copy AzCopy Command to Clipboard
Started at: 7/1/2023 7:35 PM, Duration: 5 seconds

Successfully created folder 'ecdc/testing'

Transfer of 'C:\Users\Dheeraji Bajjuri\Downloads\covid_transform_testing.hql' to 'scripts/hql/' complete: 1 item transferred (used SAS, discovery completed) Copy AzCopy Command to Clipboard
Started at: 7/1/2023 7:32 PM, Duration: 7 seconds

ADF_er-covid-reporting-adf_pl_process_pop_data_exec-pop-transform_3cd97361-9c10-4c35-af0c-9ec765c36c0f run

Delete job run

Output

Hide code | Export as HTML

```
df_processed_population = spark.sql("""SELECT c.country,
       c.country_code_2_digit,
       c.country_code_3_digit,
       c.population,
       p.Y0_14  AS age_group_0_14,
       p.Y15_24 AS age_group_15_24,
       p.Y25_49 AS age_group_25_49,
       p.Y50_64 AS age_group_50_64,
       p.Y65_79 AS age_group_65_79,
       p.Y80_MAX AS age_group_80_max
  FROM raw_population_pivot p
  JOIN dim_country c ON p.country_code = country_code_2_digit
ORDER BY country""")
```

▸ df_processed_population: pyspark.sql.dataframe.DataFrame = [country: string, country_code_2_digit: string ... 8 more fields]

Command took 0.19 seconds

Write output to the processed mount point

```
df_processed_population.write.format("com.databricks.spark.csv").option("header","true").option("delimiter", ",").mode
("overwrite").save("abfss://processed@ercovidreportingdatalake.dfs.core.windows.net/population")
```

Command took 3.80 seconds

Task run details

Job ID	283344686361047
Task run ID	665
Run as	Chinta, Avinash
Started	2023-07-02 14:06:54 CDT
Ended	2023-07-02 14:07:15 CDT
Duration	20s
Status	Succeeded

Notebook

/covid/transform/transform_population_data

Compute

- 0702-022001-r0adqqmj

Permissions

No permissions