

Assignment 15 June 2022

NAME: ASHOK KUMAR

ROLL NUMBER: DXC-262-AB-1233

BATCH: DXC-262-ANALYTICS-B12-AZURE

COMPANY: DXC TECHNOLOGY

EMPLOYEE DOMAIN: AZURE ANALYTICS

TRAINING UNDER: MANIPAL PRO LEARN

TRAINER NAME: MR. AJAY KUMAR

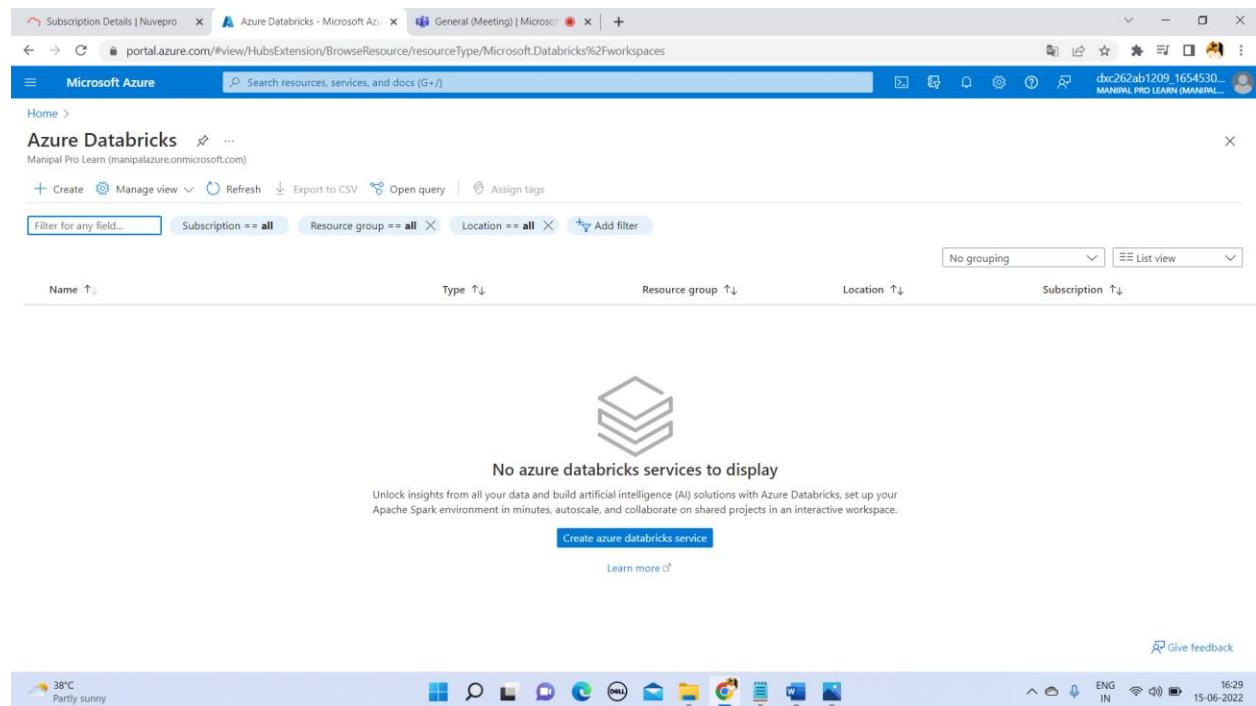
DATE OF SUBMISSION: 15 JUNE 2022

NO. OF CASES: 6

CASE 1.Using archive1.zip file - please ingest data into databricks DBFS path & query the data, display with notebooks accordingly

To Create azure databricks workspace Go to azure databricks on azure portal

<https://portal.azure.com/#home>



Click on create to create azure databricks workspace

Create an Azure Databricks workspace

Project Details

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

Subscription *

Resource group * [Create new](#)

Workspace name *

Region *

Pricing Tier *



Validation Succeeded

Summary

Basics

Workspace name	ashokdatabricks31
Subscription	Azure-DXC262AB12Lab
Resource group	ashokrg31
Region	East US
Pricing Tier	trial

Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP) No

Deploy Azure Databricks workspace in your own Virtual Network (VNet) No

Advanced

Enable Infrastructure Encryption No

Create < Previous Download a template for automation

The screenshot shows the Microsoft Azure portal with the URL [portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id%2Fsubscriptions%2F442b51c3-1fed-40a7-9d05-105684c1217c%2FresourceGroups%2Fashokrg31...MANUAL PRO LEARN \(MANUAL\)](#). The page title is "ashokrg31_ashokdatabricks31 | Overview". The main content area displays a green checkmark icon and the message "Your deployment is complete". Below this, deployment details are listed: Deployment name: ashokrg31_ashokdatabricks31, Subscription: Azure-DXC262A12Lab, Resource group: ashokrg31. To the right, deployment metadata is shown: Start time: 6/15/2022, 4:32:17 PM, Correlation ID: d7e97df6-ba15-4488-bc25-acf2476f1bf7. A "Go to resource" button is at the bottom. On the left, a sidebar lists "Overview", "Inputs", "Outputs", and "Template". At the top, there are navigation icons for "Delete", "Cancel", "Redeploy", and "Refresh". A feedback message "We'd love your feedback!" is also present.

Subscription Details | Nuvelo x ashokdatabricks31 - Microsoft A x General (Meeting) | Microsoft x +

portal.azure.com/#/resource/subscriptions/442b51c3-1fed-40a7-9d05-105684c1217c/resourceGroups/ashokrg31/providers/Microsoft.Databricks/...

Microsoft Azure Search resources, services, and docs (G+)             

Home > ashokrg31_ashokdatabricks31 >

ashokdatabricks31

Azure Databricks Service

Search (Ctrl+ /)  Delete 

 Overview

 Activity log

 Access control (IAM)

 Tags

 Settings

 Virtual Network Peers

 Encryption

 Properties

 Locks

 Automation

 Tasks (preview)

 Export template

 Support + troubleshooting

 New Support Request

 Essentials

Status : Active

Resource group : [ashokrg31](#)

Location : East US

Subscription : [Azure-DXC262AB12Lab](#)

Subscription ID : 442b51c3-1fed-40a7-9d05-105684c1217c

Tags [\(edit\)](#) : [Click here to add tags](#)

Managed Resource Group : [databricks-rg-ashokdatabricks31-qjpyzhtlgrs](#)

URL : <https://adb-1482461170098795.azuredatabricks.net>

Pricing Tier : Trial (Premium - 14-Days Free DBUs)


[Launch Workspace](#) 
[Upgrade to Premium](#) 

 Documentation 

 Getting Started 

 Import Data from File 

 Import Data from Azure Storage 

38°C Partly sunny  ENG IN  15-06-2022  16:35 

The screenshot shows the Microsoft Azure Databricks Data Science & Engineering portal. The left sidebar contains navigation links for Data Science & Engineering, Create, Workspace, Repos, Recents, Search, Data, Compute, Workflows, Partner Connect, Tasks Completed, Help, Settings, and a user profile for ashokdatabricks31. The main content area features a "Data Science & Engineering" section with icons for Notebook (Create a new notebook for querying, data processing, and machine learning), Data import (Quickly import data, preview its schema, create a table, and query it in...), and Guide: Quickstart tutorial (Spin up a cluster, run queries on preloaded data, and display results in 5 minutes). Below this is a "Recents" section showing a message "There are no recents yet". At the bottom, there are sections for Documentation (Get started guide, Runtime release notes, Azure Databricks preview releases), Release notes, and Blog posts (Building ETL pipelines for the cybersecurity lakehouse with Delta Live Tables, June 3, 2022). The status bar at the bottom indicates the weather as 38°C partly sunny, the time as 16:36, and the date as 15-06-2022.

Create new cluster (go to create and click on cluster)

The screenshot shows the Microsoft Azure Databricks Create Cluster - Databricks portal. The left sidebar is identical to the previous screenshot. The main content area is titled "Clusters / New Compute" and shows a "New Cluster" form. The "Cluster name" field is filled with "ashokcluster31". The "Cluster mode" dropdown is set to "Single Node". The "Databricks runtime version" dropdown is set to "Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1)". A promotional discount message states "Promotional discount applied to Photon during preview". Under "Autopilot options", the "Terminate after" field is set to "30 minutes of inactivity". The "Node type" dropdown is set to "Standard_DS3_v2" (14 GB Memory, 4 Cores). The "DBU / hour: 0.75" field is also set to "Standard_DS3_v2". An "Advanced options" button is visible. The status bar at the bottom indicates the weather as 38°C partly sunny, the time as 16:39, and the date as 15-06-2022.

This screenshot shows the Microsoft Azure Databricks Cluster configuration page. The cluster is named "ashokcluster31". The configuration tab is selected, showing the following details:

- Policy:** Unrestricted
- Cluster mode:** Single Node
- DataBricks Runtime Version:** 10.4 LTS (includes Apache Spark 3.2.1, Scala 2.12)
- Autopilot options:** Use Photon Acceleration (Preview), Autopilot options, Terminate after 30 minutes of inactivity.
- Node type:** Standard_DS3_v2 (14 GB Memory, 4 Cores)
- DBU / hour:** 0.75
- Standard_DS3_v2** is highlighted as the selected node type.

The UI and JSON tabs are available at the top right. The status bar at the bottom indicates it's 16:42 on 15-06-2022.

This screenshot shows the Microsoft Azure Databricks Compute page. The cluster "ashokcluster31" is listed in the All-purpose clusters section. The table includes the following columns: Name, Policy, Runtime, Active memory, Active cores, Active DBU / h, Source, and Creator. The cluster "ashokcluster31" has the following values:

Name	Policy	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator
ashokcluster31	-	10.4	14 GB	4 cores	0.75	UI	dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com

The status bar at the bottom indicates it's 16:50 on 15-06-2022.

Create Folder (workspace → create → Folder)

The screenshot shows the Microsoft Azure Databricks workspace interface. On the left, there's a sidebar with various navigation options like Data Science & Engineering, Create, Workspace, Repos, Recents, Search, Data, Compute, Workflows, Partner Connect, Tasks Completed, Help, Settings, and a user profile. The main workspace area has tabs for Home, Compute, and General (Meeting). A context menu is open over a cluster, with the 'Create' option expanded. Under 'Create', the 'Folder' option is highlighted. The cluster details shown are Active memory: 10.4, Active cores: 14 GB, Active DBU / h: 4 cores, Source: dxc262ab1209..., and Creator: dxc262ab1209... The status bar at the bottom indicates it's 36°C and Cloudy, and the date is 15-06-2022.

This screenshot shows the same Databricks workspace as the previous one, but with a modal dialog box in the foreground. The dialog is titled 'New Folder Name' and contains a single input field with the value 'ashok_dataanalytics_project1'. Below the input field are two buttons: 'Cancel' and 'Create Folder'. The background of the workspace is dimmed, and the status bar at the bottom remains the same as in the first screenshot.

Create Notebook in folder **ashok_dataanalytics_project1** (workspace → ashok_dataanalytics_project1 → create → Notebook)

The screenshot shows the Microsoft Azure Databricks workspace interface. On the left, there's a sidebar with options like Data Science & Engineering, Create, and Workspace. Under Workspace, a folder named "ashok_dataanalytics_project1" is expanded, showing sub-folders like Shared and Users. A context menu is open over this folder, with "Create" selected, revealing options: Notebook, Library, Folder, and MLflow Experiment. In the main workspace area, there's a cluster configuration card with details: Active memory 14 GB, Active cores 4 cores, Active DBU / h 0.75, Source UI, and Creator dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com. The status bar at the bottom shows the date 15-06-2022 and time 16:54.

This screenshot shows the "Create Notebook" dialog box overlaid on the workspace. The dialog has fields for Name (set to "notebook1"), Default Language (set to Python), and Cluster (set to "ashokcluster31"). There are "Cancel" and "Create" buttons at the bottom. The background workspace is identical to the one in the previous screenshot, showing the same cluster configuration and sidebar.

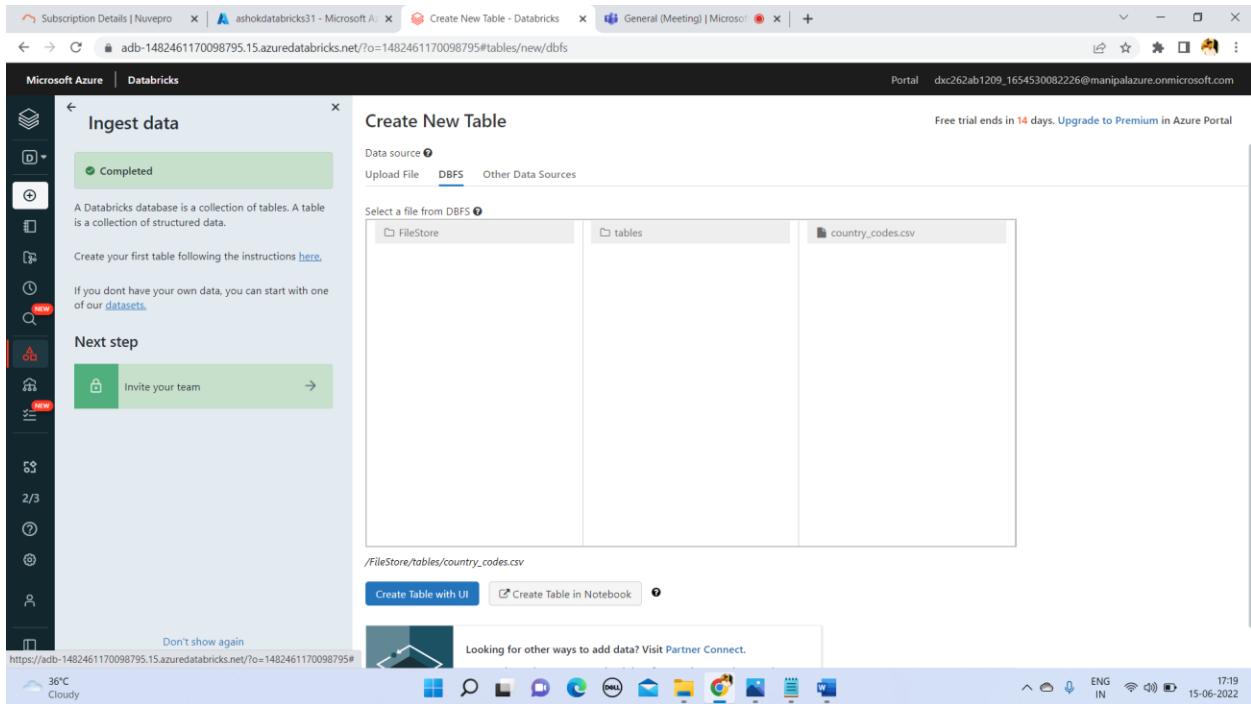
This screenshot shows the Microsoft Azure Databricks interface. On the left, there's a sidebar titled 'Get started' with sections for 'Create a cluster', 'Ingest data', and 'Invite your team'. Below it are 'Next steps' for 'Explore Notebook gallery' and 'Read documentation'. The main area is titled 'notebook1 - Python' and contains a code editor with the following Python code:

```
1 x = 100
2 y = 500
3 print(x+y)
```

The output of the code is '600'. A message at the bottom says 'Command took 0.04 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 4:57:34 PM on ashokcluster31'. The status bar at the bottom right shows the date as 15-06-2022 and the time as 17:11.

Create table (Data → create table) and upload file from archive1 (country-codes)

This screenshot shows the 'Create New Table' interface in Microsoft Azure Databricks. The left sidebar is identical to the previous screenshot. The main area is titled 'Create New Table' and shows a 'Upload File' section. A file named 'country-codes.csv' is listed with a size of '0.1 MB'. Below the file list are two buttons: 'Create Table with UI' and 'Create Table in Notebook'. A note at the bottom says 'Looking for other ways to add data? Visit Partner Connect.' and 'Use our ingestion partners to load data from various products and databases into Delta Lake.' The status bar at the bottom right shows the date as 15-06-2022 and the time as 17:18.



```
#ingest country-codes.csv file
```

```
from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
```

```
country_schema = StructType(fields=[StructField("FIFA",StringType(), True),
                                    StructField("Dial",IntegerType(), True),
                                    StructField("MARC",StringType(), True),
                                    StructField("is_independent",StringType(), True),
                                    StructField("ISO3166-1-numeric",IntegerType(), True),
                                    StructField("GAUL",IntegerType(), True),
                                    StructField("FIPS",StringType(), True),
                                    StructField("WMO",StringType(), True),
                                    StructField("ISO3166-1-Alpha-2",StringType(), True),
                                    StructField("ITU",StringType(), True),
                                    StructField("IOC",StringType(), True),
                                    StructField("DS",StringType(), True),
                                    StructField("UNTERM Spanish Formal",StringType(), True),
                                    StructField("Global Code",StringType(), True),
```

```

    StructField("Intermediate Region Code", IntegerType(), True),
    StructField("official_name_fr", StringType(), True),
    StructField("UNTERM French Short", StringType(), True),
    StructField("ISO4217-currency_name", StringType(), True),
    StructField("Global", StringType(), True),
    StructField("Developed / Developing Countries", StringType(), True),
    StructField("UNTERM Russian Formal", StringType(), True),
])

```

The screenshot shows the Microsoft Azure Databricks interface. On the left, there's a sidebar with various icons and a 'Completed' message for the 'Ingest data' step. The main area is titled 'notebook1' and contains a Python code cell with the following content:

```

1 #ingest country-codes.csv file
2 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
3
4 country_schema = StructType(fields=[StructField("FIFA", StringType(), True),
5                                     StructField("Dial", IntegerType(), True),
6                                     StructField("MARC", StringType(), True),
7                                     StructField("is_independent", StringType(), True),
8                                     StructField("ISO3166-1-numeric", IntegerType(), True),
9                                     StructField("GAUL", IntegerType(), True),
10                                    StructField("FIPS", StringType(), True),
11                                    StructField("WMO", StringType(), True),
12                                    StructField("ISO3166-1-Alpha-2", StringType(), True),
13                                    StructField("ITU", StringType(), True),
14                                    StructField("IOC", StringType(), True),
15                                    StructField("DOS", StringType(), True),
16                                    StructField("UNTERM Spanish Formal", StringType(), True),
17                                    StructField("Global Code", StringType(), True),
18                                    StructField("Intermediate Region Code", IntegerType(), True),
19                                    StructField("official_name_fr", StringType(), True),
20                                    StructField("UNTERM French Short", StringType(), True),
21                                    StructField("ISO4217-currency_name", StringType(), True),
22                                    StructField("Global", StringType(), True),
23                                    StructField("Developed / Developing Countries", StringType(), True),
24                                    StructField("UNTERM Russian Formal", StringType(), True),
25])

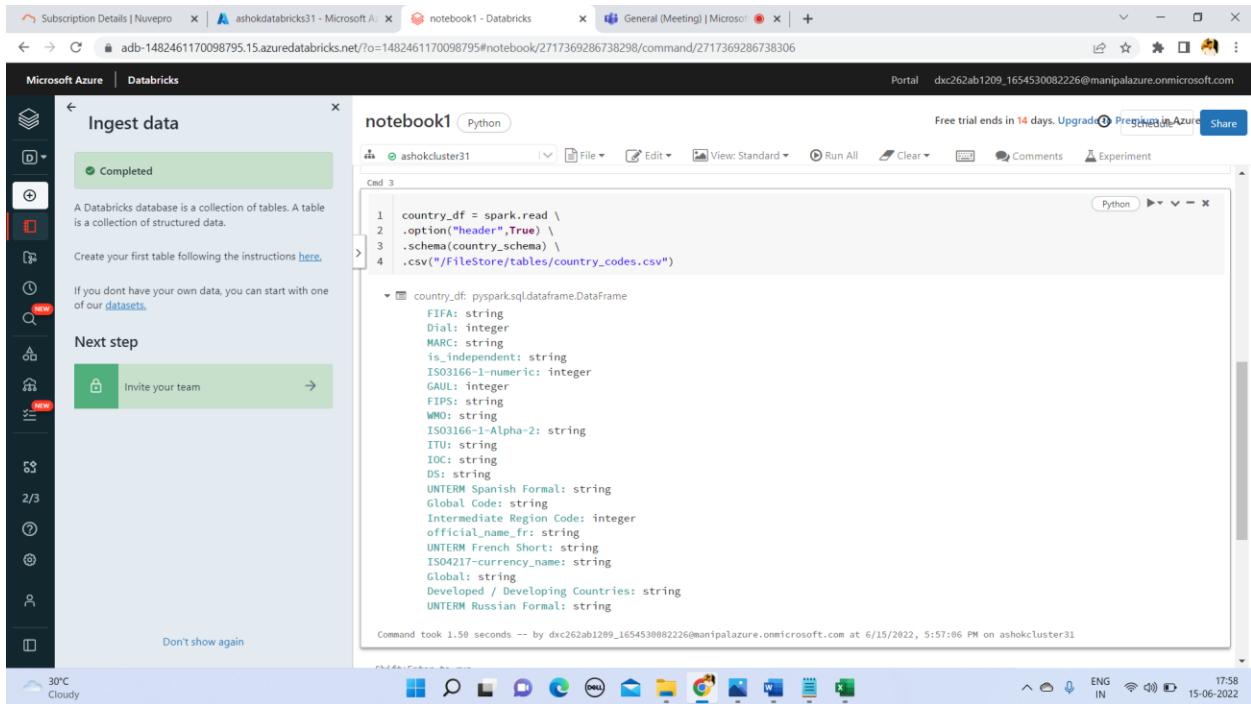
```

The notebook tab bar shows other open tabs like 'Subscription Details | Nuvelio', 'ashokdatabricks31 - Microsoft A...', and 'notebook1 - Databricks'. The status bar at the bottom indicates it's running on 'ashokcluster31'.

```

country_df = spark.read \
.option("header",True) \
.schema(country_schema) \
.csv("/FileStore/tables/country_codes.csv")

```



Select only required columns

```
from pyspark.sql.functions import col
```

```
country_df =  
country_df.select(col('FIFA'),col('Dial'),col('MARC').alias('GAUL'),col('FIPS').alias('WMO'),col('ITU').alias('IOC'),col('DS'),col('Global Code'),col('Global'))\
```

The screenshot shows the Microsoft Azure Databricks interface. On the left, there's a sidebar with various icons. In the center, a modal window titled "Ingest data" is open, showing a green "Completed" status bar. Below it, there's some text about Databricks databases and tables, followed by a "Next step" section with a "Invite your team" button. The main workspace is titled "notebook1 - Python". It contains several code cells (Cmd 4, Cmd 5, Cmd 6) and their execution logs. The logs show the execution of PySpark SQL code to select specific columns from a DataFrame named "country_df". The logs indicate the command took approximately 1.48 seconds, 0.03 seconds, and 0.10 seconds respectively. The status bar at the bottom right shows the date as 15-06-2022 and the time as 21:02.

display(country_df)

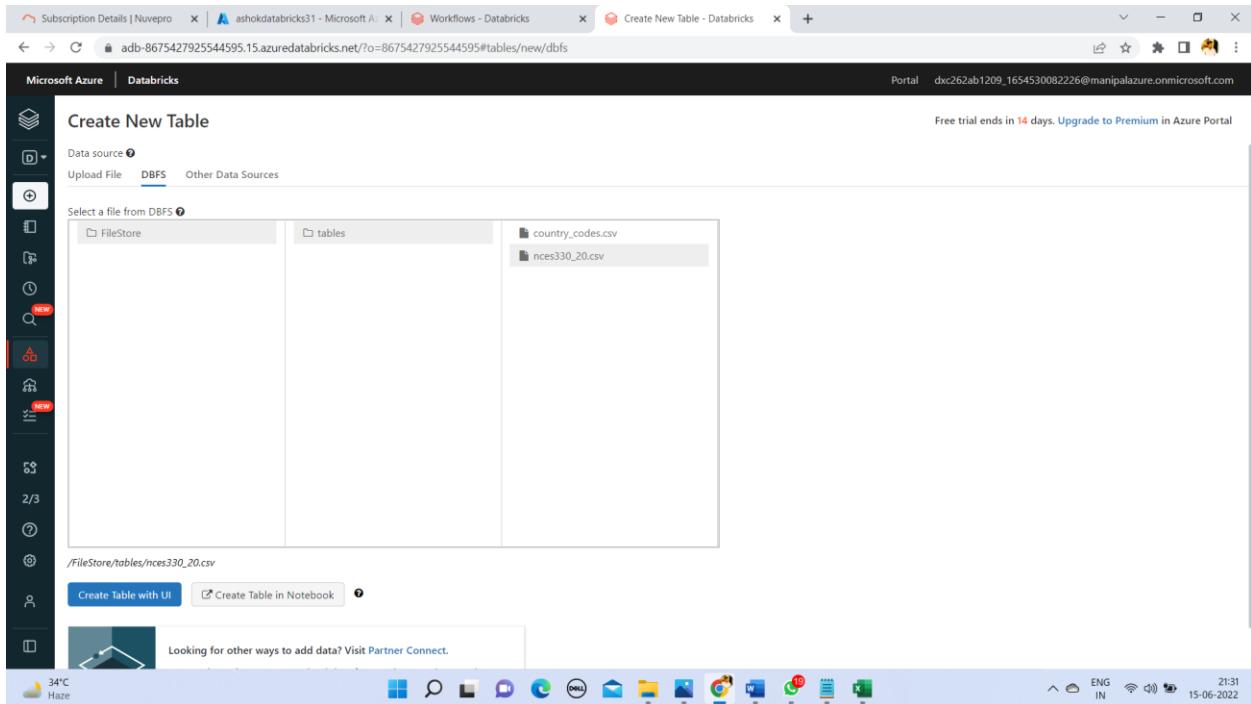
This screenshot shows the same Microsoft Azure Databricks environment as the previous one. The "Ingest data" modal is still open. In the main workspace, the command `display(country_df)` has been run in Cmd 6. The resulting DataFrame is displayed as a table with columns: FIFA, Dial, GAUL, WMO, IOC, DS, and Global Code. The table lists 250 rows of data, including entries for countries like TPE, AFG, ALB, ALG, ASA, AND, and ANG, along with their corresponding codes and names. The status bar at the bottom right shows the date as 15-06-2022 and the time as 21:03.

FIFA	Dial	GAUL	WMO	IOC	DS	Global Code
1 TPE	886	TWN	925	TW	TPE	null
2 AFG	93	AFG	1	AF	AFG	República Islámica del Afganistán (la)
3 ALB	355	ALB	3	AL	ALB	la República de Albania
4 ALG	213	DZA	4	DZ	ALG	la República Argelina Democrática y Popular
5 ASA	null	ASM	5	AS	ASA	null
6 AND	376	AND	7	AD	AND	el Principado de Andorra
7 ANG	244	AGO	8	AO	ANG	la República de Annaña

CASE 2.Using archive2.zip file - please ingest data into databricks DBFS path & query the data display with notebooks accordingly

Create Notebook

Upload the file from archive2 **nces330.20.csv** in upload file section at create table



from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType

```
nces_schema = StructType(fields=[StructField("year",IntegerType(),False),
                                 StructField("State",StringType(),True),
                                 StructField("Type",StringType(),True),
                                 StructField("Length",StringType(),True),
                                 StructField("Expense",StringType(),True),
                                 StructField("Value",IntegerType(),True),
                                 ])
```

The screenshot shows a Microsoft Azure Databricks notebook titled "notebook2 - Python". The notebook interface includes a header bar with tabs for "Subscription Details | Nuvepro", "ashokdatabricks31 - Microsoft A...", "Workflows - Databricks", "notebook2 - Databricks", and "notebook2 - Databricks". The main area contains three command cells (Cmd 1, Cmd 2, Cmd 3) and a sidebar with various icons. The code in Cmd 1 imports PySpark SQL types. Cmd 2 defines a schema for "nces_schema" and Cmd 3 reads a CSV file into "nces_df" using the schema. The status bar at the bottom shows system information like battery level, language, and date.

```
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
2
3 nces_schema = StructType(fields=[StructField("year", IntegerType(), False),
4                                 StructField("State", StringType(), True),
5                                 StructField("Type", StringType(), True),
6                                 StructField("Length", StringType(), True),
7                                 StructField("Expense", StringType(), True),
8                                 StructField("Value", IntegerType(), True)])
9
10 nces_df = spark.read \
11     .option("header", True) \
12     .schema(nces_schema) \
13     .csv("/FileStore/tables/nCES330_20.csv")
```

```
nces_df = spark.read \
.option("header", True) \
.schema(nces_schema) \
.csv('/FileStore/tables/nCES330_20.csv')
```

```
from pyspark.sql.functions import col
nces_selected_df = nces_df.select(col('Type'), col('Length'), col('Value'))
```

A screenshot of a Microsoft Azure Databricks notebook titled "notebook2" in Python. The notebook interface includes a sidebar with cluster management tools like "ashokcluster31", a toolbar with file operations, and a main workspace for running code. The workspace shows two command cells:

```
1 nces_df = spark.read \
2 .option("header", True) \
3 .schema(nces_schema) \
4 .csv("./FileStore/tables/nces330_20.csv")
```

and

```
1 from pyspark.sql.functions import col, lit
2 nces_selected_df = nces_df.select(col('Type'),
3                                     col('Length'), col('Value'))
```

Both cells have run successfully, indicated by the output below them.

`display(nces_selected_df)`

A screenshot of the same Microsoft Azure Databricks notebook after running the `display(nces_selected_df)` command. The workspace now shows the resulting DataFrame as a table:

	Type	Length	Value
1	Private	4-year	13983
2	Private	4-year	8503
3	Public In-State	2-year	4048
4	Public In-State	4-year	8073
5	Public In-State	4-year	8473
6	Public Out-of-State	2-year	7736
7	Public Out-of-State	4-year	20380

The table has a header row and 7 data rows. A note at the bottom indicates that the results are truncated.

CASE 3.Using archive3.zip file - please ingest data into databricks DBFS path & query the data display with notebooks accordingly

Type	Length	Value
1 Private	4-year	13983
2 Private	4-year	8503
3 Public In-State	2-year	4048
4 Public In-State	4-year	8073
5 Public In-State	4-year	8473
6 Public Out-of-State	2-year	7736
7 Public Out-of-State	4-year	20380

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

Command took 0.41 seconds -- by dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com at 6/15/2022, 9:40:37 PM on ashokcluster31

Upload file final_data.csv

Ingest data

Completed

A Databricks database is a collection of tables. A table is a collection of structured data.

Create your first table following the instructions [here](#).

If you don't have your own data, you can start with one of our [datasets](#).

Next step

Upload file DBFS Other Data Sources

DBFS Target Directory /FileStore/tables/ (optional) Select

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

final_data.csv

1.3 MB Remove file

✓ File uploaded to /FileStore/tables/final_data.csv

Create Table with UI Create Table in Notebook

Looking for other ways to add data? Visit [Partner Connect](#). Use our ingestion partners to load data from various products and databases into Delta Lake.

Don't show again

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A | notebook3 - Databricks | notebook2 - Databricks | +

Microsoft Azure | Databricks

Ingest data

Completed

A Databricks database is a collection of tables. A table is a collection of structured data.

Create your first table following the instructions [here](#).

If you don't have your own data, you can start with one of our [datasets](#).

Next step

Invite your team

Don't show again

34°C Haze

notebook3 Python

ashokcluster31 File Edit View Standard Run All Clear Comments Experiment

Cmd 1

```
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
```

Command took 0.84 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 9:44:40 PM on ashokcluster31

Cmd 2

```
1 final_data_schema = StructType(fields=[StructField("tweet_text",StringType(),False),  
2 StructField("emotion_in_tweet_is_directed_at",StringType(),True),  
3 StructField("is_there_an_emotion_directed_at_a_brand_or_product",StringType(),True),  
4 ])
```

Command took 0.83 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 9:45:08 PM on ashokcluster31

Cmd 3

```
1 final_data_df = spark.read \  
2 .option("header", True) \  
3 .schema(final_data_schema) \  
4 .csv("FileStore/tables/final_data.csv")
```

final_data_df: pyspark.sql.dataframe.DataFrame
tweet_text: string
emotion_in_tweet_is_directed_at: string
is_there_an_emotion_directed_at_a_brand_or_product: string

Command took 0.17 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 9:46:00 PM on ashokcluster31

Shift+Enter to run

Python

21:46 15-06-2022

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A | notebook3 - Databricks | notebook2 - Databricks | +

Microsoft Azure | Databricks

Ingest data

Completed

A Databricks database is a collection of tables. A table is a collection of structured data.

Create your first table following the instructions [here](#).

If you don't have your own data, you can start with one of our [datasets](#).

Next step

Invite your team

Don't show again

33°C Haze

notebook3 Python

ashokcluster31 File Edit View Standard Run All Clear Comments Experiment

Cmd 4

```
1 .csv("FileStore/tables/final_data.csv")
```

final_data_df: pyspark.sql.dataframe.DataFrame
tweet_text: string
emotion_in_tweet_is_directed_at: string
is_there_an_emotion_directed_at_a_brand_or_product: string

Command took 0.17 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 9:46:00 PM on ashokcluster31

Cmd 5

```
1 from pyspark.sql.functions import col
```

Command took 0.83 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 9:47:42 PM on ashokcluster31

Cmd 6

```
1 final_data_selected_df =  
final_data_df.select(col('tweet_text'),col('emotion_in_tweet_is_directed_at').alias('emotion_towards'),col('is_there_an_emotion_d  
irected_at_a_brand_or_product').alias('is_there_a_brand'))
```

final_data_selected_df: pyspark.sql.dataframe.DataFrame
tweet_text: string
emotion_towards: string
is_there_a_brand: string

Command took 0.66 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 9:48:58 PM on ashokcluster31

display(final_data_selected_df)

Python

21:50 15-06-2022

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A... | notebook3 - Databricks | notebook2 - Databricks | +

ashb-8675427925544595.15.azuredatabricks.net/?o=8675427925544595#notebook/4234722532109343/command/4234722532109355

Microsoft Azure | Databricks

Ingest data

Completed

A Databricks database is a collection of tables. A table is a collection of structured data.

Create your first table following the instructions [here](#).

If you don't have your own data, you can start with one of our [datasets](#).

Next step

Invite your team

Don't show again

notebook3 Python

ashokcluster31

File View Standard Run All Clear Comments Experiment

Free trial ends in 14 days. Upgrade Premium in Azure Share

1 display(final_data_selected_df)

(1) Spark Jobs

Table Data Profile

tweet_text emotion_towards

1 @wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to upgrade. Plugin stations at #SXSW. iPhone

2 @jessedee Know about @fludapp? Awesome iPad/iPhone app that you'll likely appreciate for its design. Also, they're giving free iPad or iPhone App

3 @swonderlin Can not wait for #iPad 2 also. They should sell them down at #SXSW. iPad

4 @sxsw I hope this year's festival isn't as crashy as this year's iPhone app. #sxsw iPad or iPhone App

5 @sxtxstate great stuff on Fri #SXSW: Marissa Mayer (Google), Tim O'Reilly (tech books/conferences) & Matt Mullenweg (Wordpress) Google

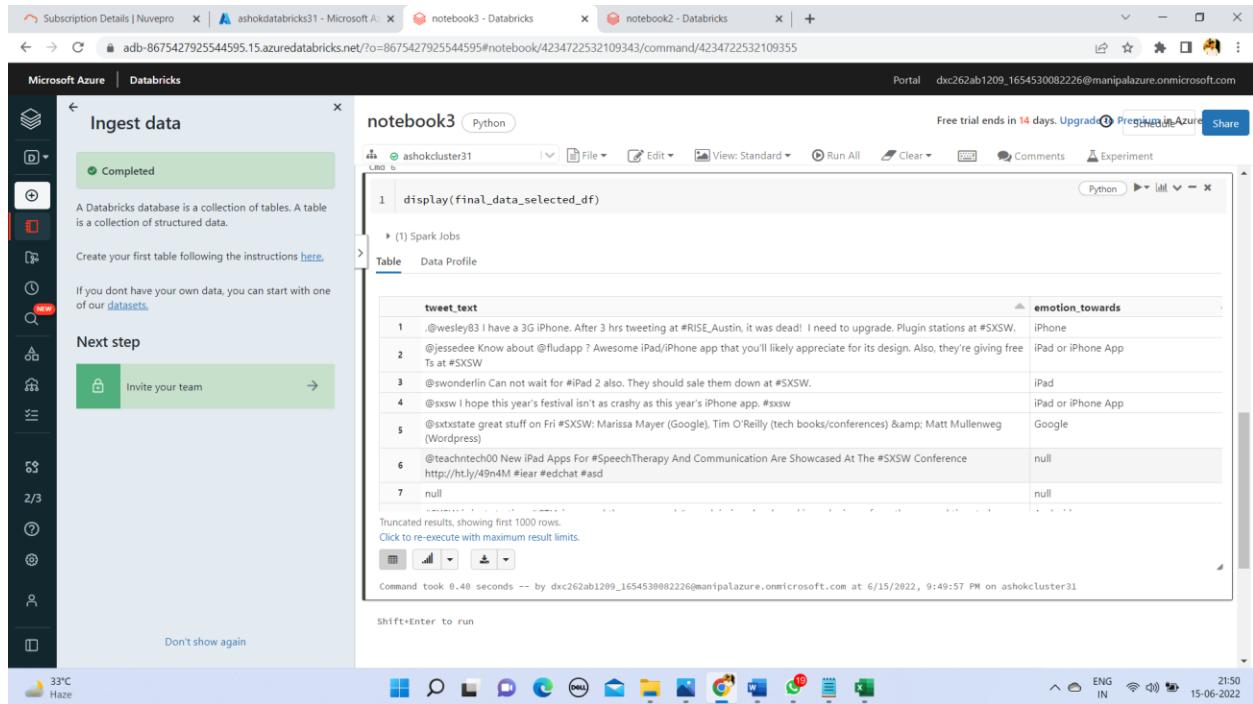
6 @teachtech00 New iPad Apps For #SpeechTherapy And Communication Are Showcased At The #SXSW Conference null

7 null

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

Shift+Enter to run

Windows Taskbar: 33°C Haze, ENG IN, 21:50, 15-06-2022



CASE 4.Using archive4.zip file - please ingest data into databricks DBFS path & query the data display with notebooks accordingly

The screenshot shows the Microsoft Azure Databricks interface. A notebook titled "notebook3" is open, displaying a command to ingest data from a CSV file into a table named "emotion_towards". The command output shows the first few rows of the data:

```

emotion_towards
iPhone
iPad or iPhone App
iPad
iPad or iPhone App
Google
null
null

```

The interface includes a sidebar for "Ingest data" and a "Create Notebook" dialog.

Upload SEntFiN-v1.1.csv file.

The screenshot shows the Microsoft Azure Databricks interface with the "Create New Table" dialog open. A file named "SEntFiN-v1.1.csv" is selected for upload to the DBFS target directory "/FileStore/tables/". The dialog also shows a message about files being accessible to everyone with access to the workspace.

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A | notebook4 - Databricks | +

ADB URL: adb-8675427925544595.15.azuredatabricks.net/?o=8675427925544595#notebook/4234722532109356/command/4234722532109365

Microsoft Azure | Databricks

notebook4 Python

ashokcluster31 File Edit View Standard Run All Clear

Free trial ends in 14 days. Upgrade Premium in Azure Share

Portal dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com

Cmd 1

```
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
```

Command took 0.04 seconds -- by dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com at 6/15/2022, 9:53:37 PM on ashokcluster31

Cmd 2

```
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType
```

Command took 0.03 seconds -- by dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com at 6/15/2022, 9:54:31 PM on ashokcluster31

Cmd 3

```
1 sent_schema = StructType(fields=[StructField("S No.", IntegerType(), False),
                                    StructField("Title", StringType(), True),
                                    StructField("Decisions", StringType(), True),
                                    StructField("Words", IntegerType(), True),
                                    ])
```

Command took 0.05 seconds -- by dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com at 6/15/2022, 9:56:54 PM on ashokcluster31

Shift+Enter to run

33°C Haze 21:57 15-06-2022

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A | notebook4 - Databricks | +

ADB URL: adb-8675427925544595.15.azuredatabricks.net/?o=8675427925544595#notebook/4234722532109356/command/4234722532109365

Microsoft Azure | Databricks

notebook4 Python

ashokcluster31 File Edit View Standard Run All Clear

Free trial ends in 14 days. Upgrade Premium in Azure Share

Portal dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com

sent_selected_df=sent_df.select(col('Title'),col('Decisions'),col('Words'))

SyntaxError: EOL while scanning string literal

Command took 0.08 seconds -- by dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com at 6/15/2022, 10:05:18 PM on ashokcluster31

Cmd 7

```
1 display(sent_df)
```

(1) Spark Jobs Job 3 View (Stages: 1/1)

Table Data Profile

S No.	Title	Decisions	Words
1	SpiceJet to issue 6.4 crore warrants to promoters	{"Spicejet": "neutral"}	8
2	MMTC Q2 net loss at Rs 10.4 crore	{"MMTC": "neutral"}	8
3	Mid-cap funds can deliver more, stay put: Experts	{"Mid-cap funds": "positive"}	8
4	Mid caps now turn into market darlings	{"Mid caps": "positive"}	7
5	Market seeing patience, if not conviction: Prakash Diwan	{"Market": "neutral"}	8
6	Infosys: Will the strong volume growth sustain?	{"Infosys": "neutral"}	7
7	Hurree raises Rs 279 cr via tax-free bonds	{"Hurree": "neutral"}	8

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

33°C Haze 22:06 15-06-2022

CASE 5.Using archive5.zip file - please ingest data into databricks DBFS path & query the data display with notebooks accordingly

The screenshot shows the Microsoft Azure Databricks interface. On the left, there's a notebook titled "notebook4" in Python. The code cell contains:

```
1 sent_selected_df=sent_df.select(col("Title"),col("Decisions"))
2
3 SyntaxError: EOL while scanning string literal
4 Command took 0.08 seconds -- by dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com
5
6 display(sent_df)
7
8 (1) Spark Jobs
9   Job 3 View (Stages: 1/1)
```

A modal dialog box titled "Create Notebook" is open, prompting for:

- Name: notebook5
- Default Language: Python
- Cluster: ashokcluster31

On the right, a table titled "Data Profile" is displayed with the following data:

S No.	Title	Decisions	Words
1	SpiceJet to issue 6.4 crore warrants to promoters	{"SpiceJet": "neutral"}	8
2	MMTC Q2 net loss at Rs 10.4 crore	{"MMTC": "neutral"}	8
3	Mid-cap funds can deliver more, stay put: Experts	{"Mid-cap funds": "positive"}	8
4	Mid caps now turn into market darlings	{"Mid caps": "positive"}	7
5	Market seeing patience, if not conviction: Prakash Diwan	{"Market": "neutral"}	8
6	Infosys: Will the strong volume growth sustain?	{"Infosys": "neutral"}	7
7	Hudson raises Rs 279 cr via tax-free bonds	{"Hudson": "positive"}	8

Below the table, it says "Truncated results, showing first 1000 rows. Click to re-execute with maximum result limits." The command took 0.48 seconds.

The screenshot shows the Microsoft Azure Databricks interface for creating a new table. The "Upload File" tab is selected. A file named "cancer-death_rates.csv" is uploaded to the "DBFS Target Directory" located at "/FileStore/tables/".

The file details are shown:

- File: cancer-death_rates.csv
- Size: 0.3 MB
- Status: Remove file

Below the file list, a message indicates the file was uploaded successfully: "File uploaded to /FileStore/tables/cancer_death_rates.csv".

At the bottom, there are two buttons: "Create Table with UI" and "Create Table in Notebook".

On the right, a sidebar provides information about adding data:

- Looking for other ways to add data? Visit Partner Connect.
- Use our ingestion partners to load data from various products and databases into Delta Lake.

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A | notebook5 - Databricks

notebook5 - Python

ashokcluster31

File Edit View Standard Run All Clear

Free trial ends in 14 days. Upgrade Premium in Azure Share

Comments Experiment Revision history

```
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType, FloatType
```

Command took 0.02 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:12:23 PM on ashokcluster31

```
1 cancer_death_rates_schema = StructType(fields=[StructField("Entity",StringType(),False),
                                                 StructField("Code",StringType(),True),
                                                 StructField("Year",IntegerType(),True),
                                                 StructField("Deaths - Neoplasms - Sex: Both - Age: Age-standardized (Rate)",FloatType(),True),
                                                 ])
2 ])
```

Command took 0.03 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:12:25 PM on ashokcluster31

```
1 Cancer_death_rates_df = spark.read \
2 .option("header", True) \
3 .schema(cancer_death_rates_schema) \
4 .csv("./Filestore/tables/cancer_death_rates.csv")
5 
```

Cancer_death_rates_df: pyspark.sql.dataframe.DataFrame

- Entity: string
- Code: string
- Year: integer
- Deaths - Neoplasms - Sex: Both - Age: Age-standardized (Rate): float

Command took 0.13 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:12:56 PM on ashokcluster31

33°C Haze 22:14 ENG IN 15-06-2022

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A | notebook5 - Databricks

notebook5 - Python

ashokcluster31

File Edit View Standard Run All Clear

Free trial ends in 14 days. Upgrade Premium in Azure Share

Comments Experiment Revision history

```
1 cancer_death_rates_selected_df = cancer_death_rates_df.select(col('Entity'),
                                                               col('Year'),col('Deaths - Neoplasms - Sex: Both - Age: Age-standardized (Rate)').alias('Deaths'))
```

SyntaxError: EOL while scanning string literal

Command took 0.09 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:13:31 PM on ashokcluster31

```
1 display(Cancer_death_rates_df)
```

(1) Spark Jobs

Table Data Profile

Entity	Code	Year	Deaths - Neoplasms - Sex: Both - Age: Age-standardized (Rate)
Afghanistan	AFG	1990	159.96486
Afghanistan	AFG	1991	158.45589
Afghanistan	AFG	1992	157.39096
Afghanistan	AFG	1993	157.57445
Afghanistan	AFG	1994	158.03172
Afghanistan	AFG	1995	157.97733
Afghanistan	AFG	1996	157.99734

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

33°C Haze 22:14 ENG IN 15-06-2022

CASE 6.Using archive6.zip file - please ingest data into databricks DBFS path & query the data display with notebooks accordingly

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A... | notebook5 - Databricks | Create New Table - Databricks | +

adb-8675427925544595.15.azureddatabricks.net/?o=8675427925544595#tables/new

Microsoft Azure | Databricks

Create New Table

Data source

Upload File DBFS Other Data Sources

DBFS Target Directory /FileStore/tables/ (optional)

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

Files

inflation-gdp.csv

0.4 MB Remove file

✓ File uploaded to /FileStore/tables/inflation_gdp.csv

Looking for other ways to add data? Visit [Partner Connect](#).
Use our ingestion partners to load data from various products and databases into Delta Lake.

33°C Haze 22:15 15-06-2022

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A... | notebook5 - Databricks | Create New Table - Databricks | +

adb-8675427925544595.15.azureddatabricks.net/?o=8675427925544595#tables/new

Microsoft Azure | Databricks

Create New Table

Data source

Upload File DBFS Other Data Sources

Select a file from DBFS

FileStore	tables	
-----------	--------	--

/FileStore/tables/inflation_gdp.csv

Looking for other ways to add data? Visit [Partner Connect](#).

33°C Haze 22:16 15-06-2022

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A | notebook5 - Databricks | notebook6 - Databricks | +

ashokdatabricks31 - Microsoft A | notebook5 - Databricks | notebook6 - Databricks | +

notebook6 - Databricks

notebook6 - Databricks

Microsoft Azure | Databricks

notebook6 Python

ashokcluster31

File Edit View: Standard Run All Clear

Free trial ends in 14 days. Upgrade Premium in Azure Share

Portal dxc262ab1209_165453008226@manipalazure.onmicrosoft.com

Comments Experiment Revision history

Cmd 1

```
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType, FloatType
```

Command took 0.03 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:17:22 PM on ashokcluster31

Cmd 2

```
1 inflation_gdp_schema = StructType(fields=[StructField("Country",StringType(),False),  
                                             StructField("Country Code",StringType(),True),  
                                             StructField("Year",IntegerType(),True),  
                                             StructField("Inflation",FloatType(),True),  
                                             ])
```

Command took 0.03 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:24:46 PM on ashokcluster31

Cmd 3

```
1 inflation_gdp_df = spark.read \  
2 .option("header", True) \  
3 .schema(inflation_gdp_schema) \  
4 .csv("/FileStore/tables/inflation_gdp.csv")
```

inflation_gdp_df: pyspark.sql.dataframe.DataFrame

Country: string
Country Code: string
Year: integer
Inflation: float

Command took 0.14 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:24:49 PM on ashokcluster31

Cmd 4

Python

33°C Haze

22:25 15-06-2022

Subscription Details | Nuvepro | ashokdatabricks31 - Microsoft A | notebook5 - Databricks | notebook6 - Databricks | +

ashokdatabricks31 - Microsoft A | notebook5 - Databricks | notebook6 - Databricks | +

notebook6 - Databricks

notebook6 - Databricks

Microsoft Azure | Databricks

notebook6 Python

ashokcluster31

File Edit View: Standard Run All Clear

Free trial ends in 14 days. Upgrade Premium in Azure Share

Portal dxc262ab1209_165453008226@manipalazure.onmicrosoft.com

Comments Experiment Revision history

Cmd 4

```
1 from pyspark.sql.functions import col
```

Command took 0.14 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:24:49 PM on ashokcluster31

Cmd 5

```
1 inflation_gdp_selected_df = inflation_gdp_df.select(col('Country'), col('Year'), col('Inflation'))
```

inflation_gdp_selected_df: pyspark.sql.dataframe.DataFrame

Country: string
Year: integer
Inflation: float

Command took 0.06 seconds -- by dxc262ab1209_165453008226@manipalazure.onmicrosoft.com at 6/15/2022, 10:24:54 PM on ashokcluster31

Cmd 6

```
1 display(inflation_gdp_df)
```

(1) Spark Jobs

Table Data Profile

Country

33°C Haze

22:25 15-06-2022

Subscription Details | Nuvepro | adb-8675427925544595.15.azuredatabricks.net - Microsoft Azure | notebook5 - Databricks | notebook6 - Databricks | +

Microsoft Azure | Databricks

notebook6 Python

ashokcluster31 | File | Edit | View Standard | Run All | Clear | Command took 0.86 seconds -- by dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com at 6/15/2022, 10:24:54 PM on ashokcluster31

Free trial ends in 14 days. Upgrade to Premium in Azure | Share | Schedules

Cmd 6

```
1 display(inflation_gdp_df)
```

(1) Spark Jobs

Table Data Profile

Country	Country Code	Year	Inflation
1 Arab World	ARB	1969	1.3037902
2 Arab World	ARB	1970	2.6022408
3 Arab World	ARB	1971	6.884719
4 Arab World	ARB	1972	2.4960809
5 Arab World	ARB	1973	11.555281
6 Arab World	ARB	1974	26.922678
7 Arab World	ARR	1975	5.599144

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

Command took 0.38 seconds -- by dxc262ab1209_1654530082226@manipalazure.onmicrosoft.com at 6/15/2022, 10:23:20 PM on ashokcluster31

Shift+Enter to run

33°C Haze

ENG IN 22:25 15-06-2022

Thank You!

ASHOK KUMAR