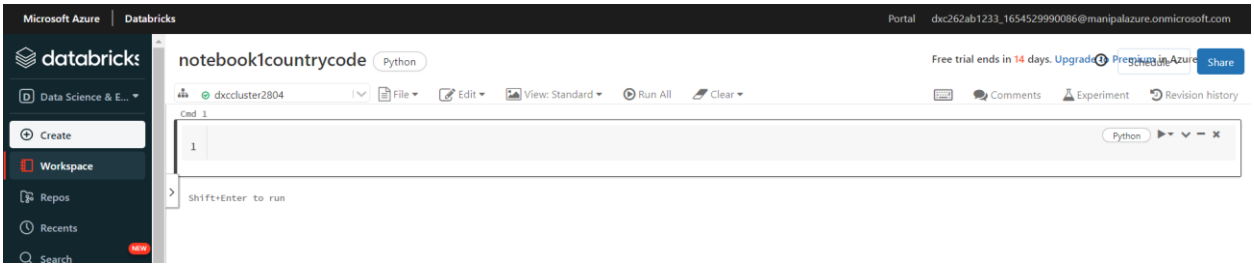
NAME: SHAIK ABDUL KHADAR JILANI ROLLNO: DXC262AB12038

BATCH: DXC-262-Analytics-B12-Azure SUBMISSION: 15-6-2022

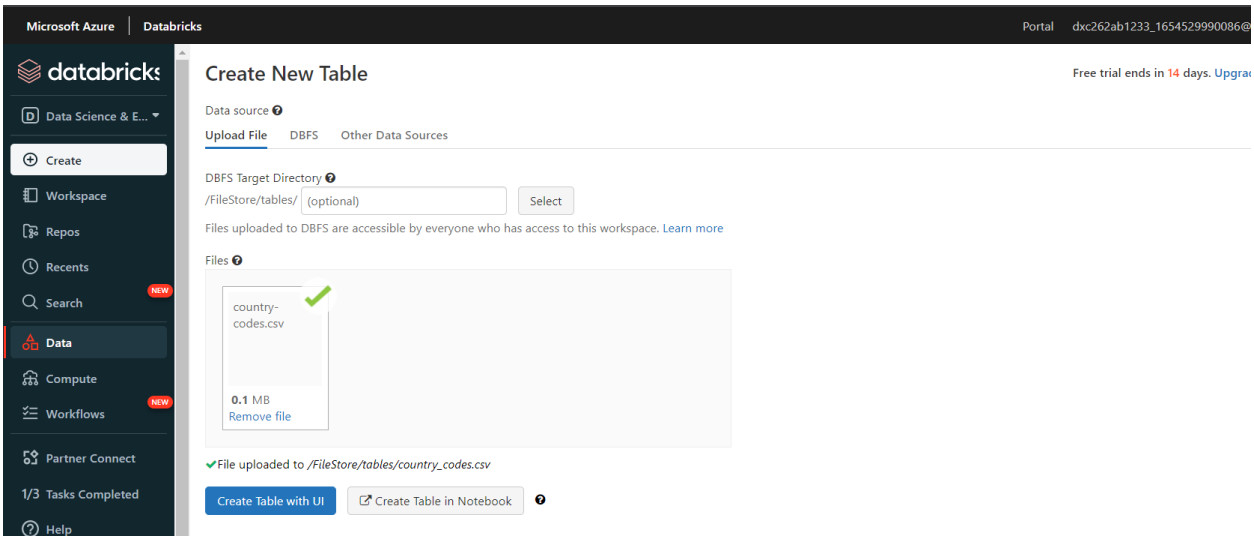
COMPANY: DXC TECHNOLOGY

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

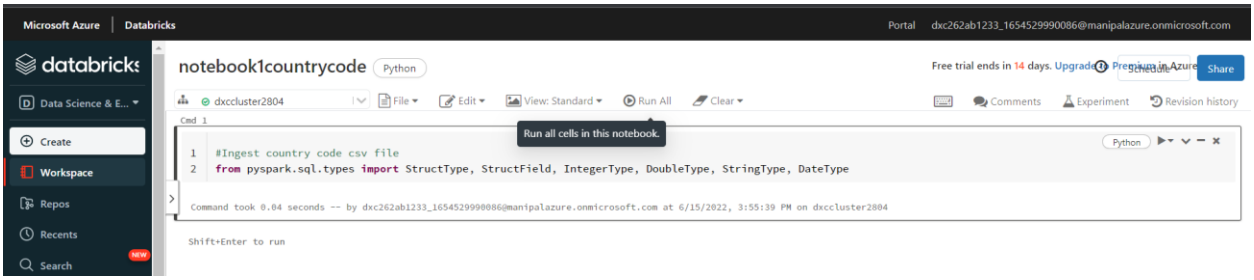
After creating the data bricks come to launch workspace and in that create a cluster and go to workspace and create a dataanalytics_project in that create a notebook and



Now 1st we need to upload the country code file so go to Data tab and click on create table



Now we need to ingest the county code csv file



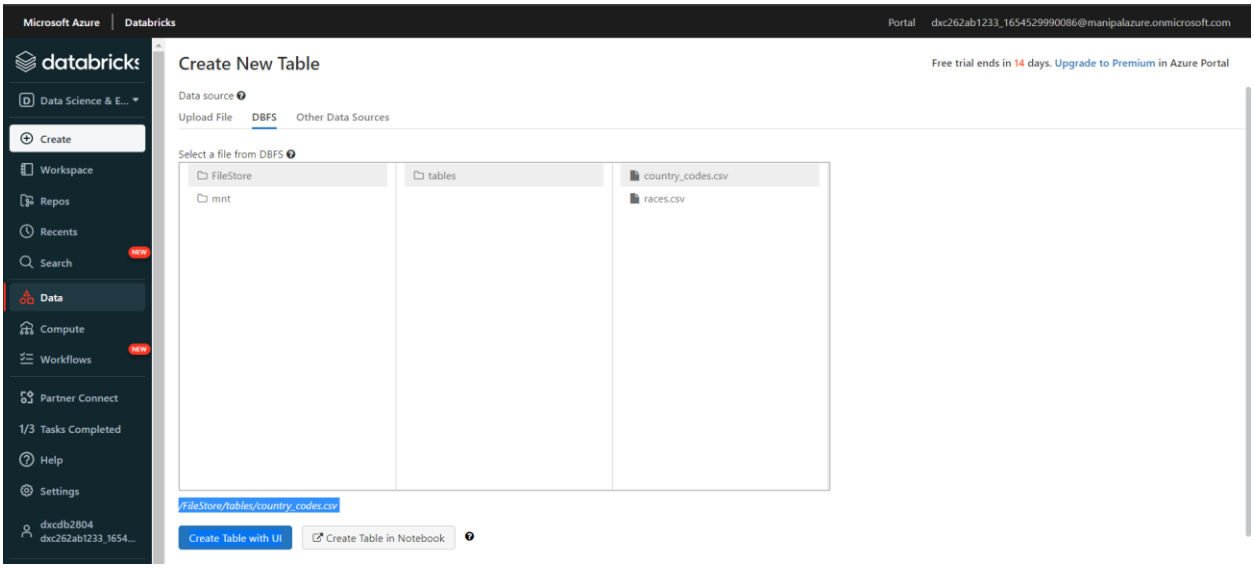
We have included schema

```
Cmd 2
#include the schema
country_schema = StructType(fields=[StructField("marc",StringType(),True),
                                    StructField("Capital",StringType(),True),
                                    StructField("M49",IntegerType(),True),
                                    StructField("Regioncode",IntegerType(),True),

                                   ])
```

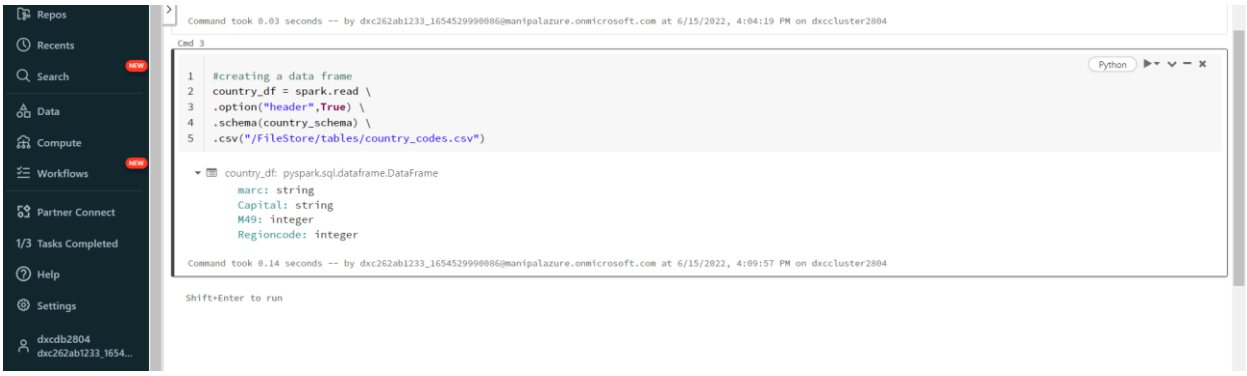
Command took 0.03 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:04:19 PM on dxccluster2804

Shift+Enter to run

We need to copy the path of the csv file before creating a data frame to do that go to Data click on create table and click on DBFS



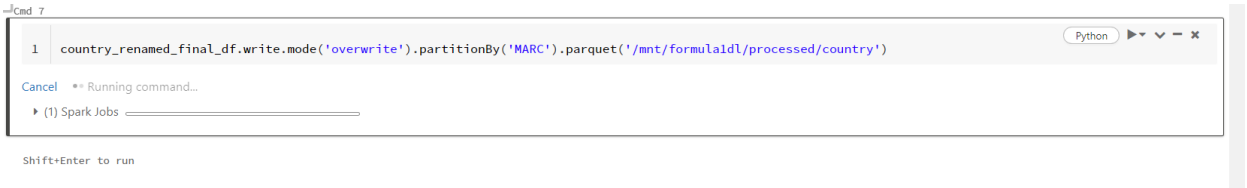
We need to create a data frame



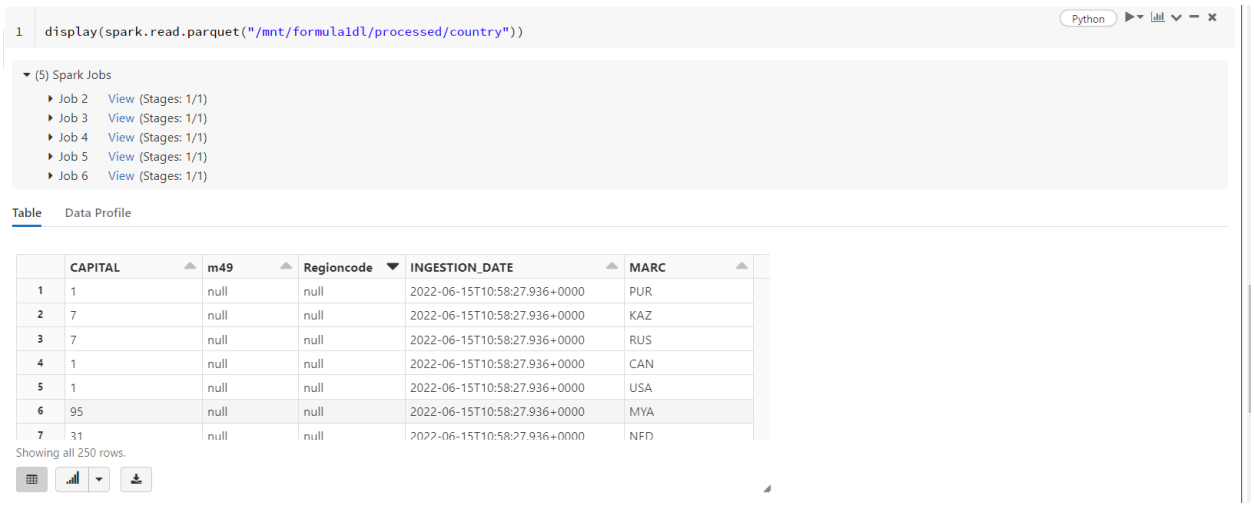
To add ingestion date

To rename the columns we need

write the output to processed container in parquet format



To display the data

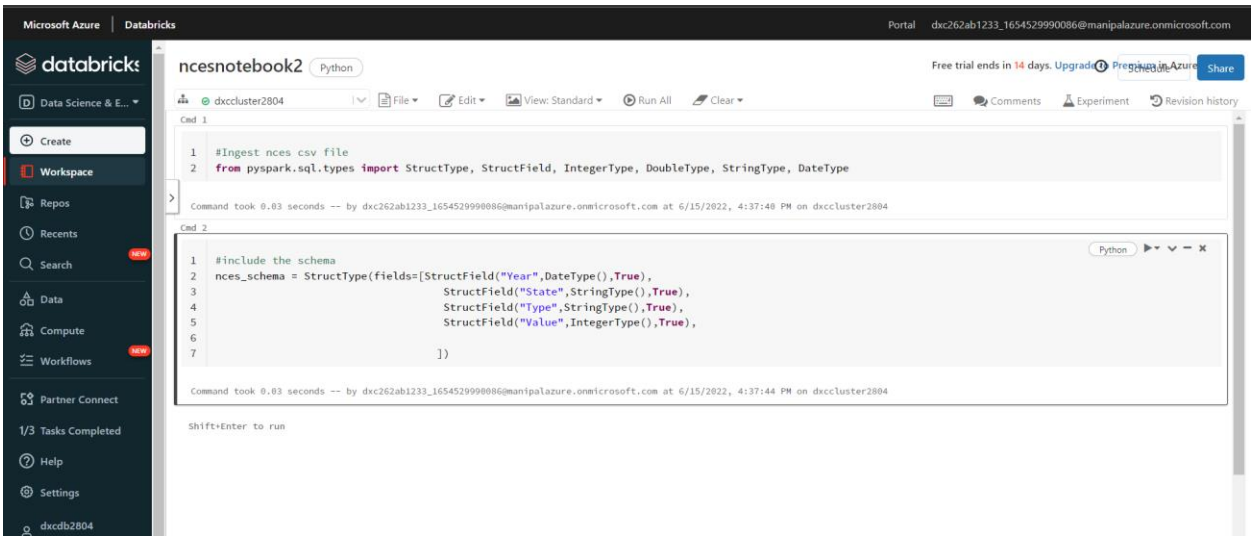


	CAPITAL	m49	Regioncode	INGESTION_DATE	MARC
1	1	null	null	2022-06-15T10:58:27.936+0000	PUR
2	7	null	null	2022-06-15T10:58:27.936+0000	KAZ
3	7	null	null	2022-06-15T10:58:27.936+0000	RUS
4	1	null	null	2022-06-15T10:58:27.936+0000	CAN
5	1	null	null	2022-06-15T10:58:27.936+0000	USA
6	95	null	null	2022-06-15T10:58:27.936+0000	MYA
7	31	null	null	2022-06-15T10:58:27.936+0000	NED

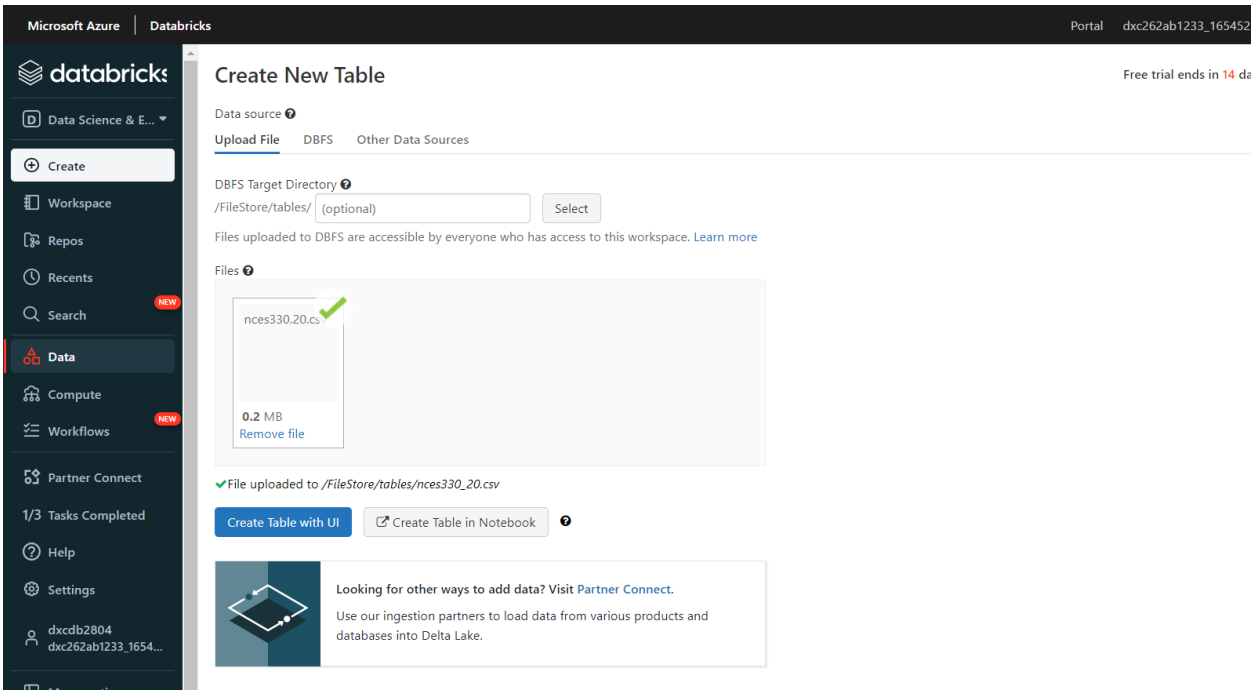
2) Using archive2.zip file - please ingest data into data bricks DBFS path & query the data display with notebooks accordingly

After creating the data bricks come to launch workspace and in that create a cluster and go to workspace and create a dataanalytics_project in that create a notebook and

Ingest the data and schema is created



To create a data frame we need to upload the file in data create table and upload by drag and drop



Now creating the data frame we need to copy the file path to include in the data frame

```
Cmd 3
#creating a data frame
nces_df = spark.read \
.option("header",True) \
.schema(nces_schema) \
.csv("/FileStore/tables/nces330_20.csv")

nces_df: pyspark.sql.dataframe.DataFrame
    Year: date
    State: string
    Type: string
    Value: integer

Command took 0.15 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:41:53 PM on dxccluster2804
```

Now ingestion date

```
from pyspark.sql.functions import current_timestamp

Command took 0.02 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:42:49 PM on dxccluster2804

Cmd 5
#add ingestion date to the data frame
nces_ingestiondate_df = nces_df.withColumn("ingestion_date",current_timestamp())

nces_ingestiondate_df: pyspark.sql.dataframe.DataFrame
    Year: date
    State: string
    Type: string
    Value: integer
    ingestion_date: timestamp

Command took 0.04 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:43:52 PM on dxccluster2804
```

Renaming the selected columns

```
Cmd 6
nces_renamed_final_df = nces_ingestiondate_df.withColumnRenamed("Year","YEAR") \
.withColumnRenamed("State","STATE") \
.withColumnRenamed("Type","What_TYPE") \
.withColumnRenamed("ingestion_date","INGESTION_DATE")

nces_renamed_final_df: pyspark.sql.dataframe.DataFrame
    YEAR: date
    STATE: string
    What_TYPE: string
    Value: integer
    INGESTION_DATE: timestamp

Command took 0.04 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:46:07 PM on dxccluster2804
```

To check the job

```
Cmd 7
nces_renamed_final_df.write.mode('overwrite

(1) Spark Jobs
  Job 7  View (Stages: 1/1)

Command took 1.83 seconds -- by dxc262ab1233_1654529990
```

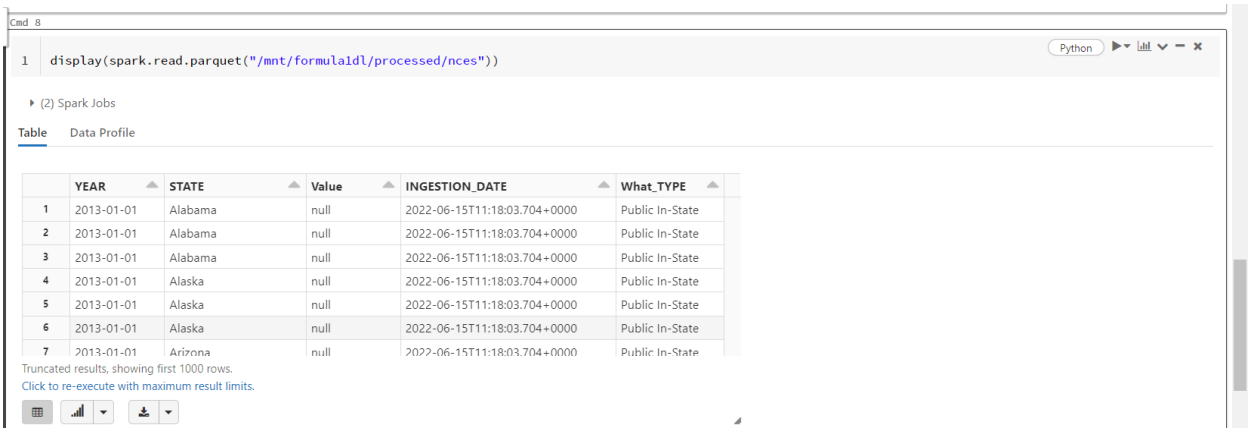
Details for Job 7

Status: SUCCEEDED
Submitted: 2022/06/15 11:18:03
Duration: 1 s
Associated SQL Query: 43
Job Group: 1875153522092939229_5869502220099066471_3e7c8ee38b4f4b5f8e0b202ad0c1d9ff
Completed Stages: 1

Completed Stages (1)

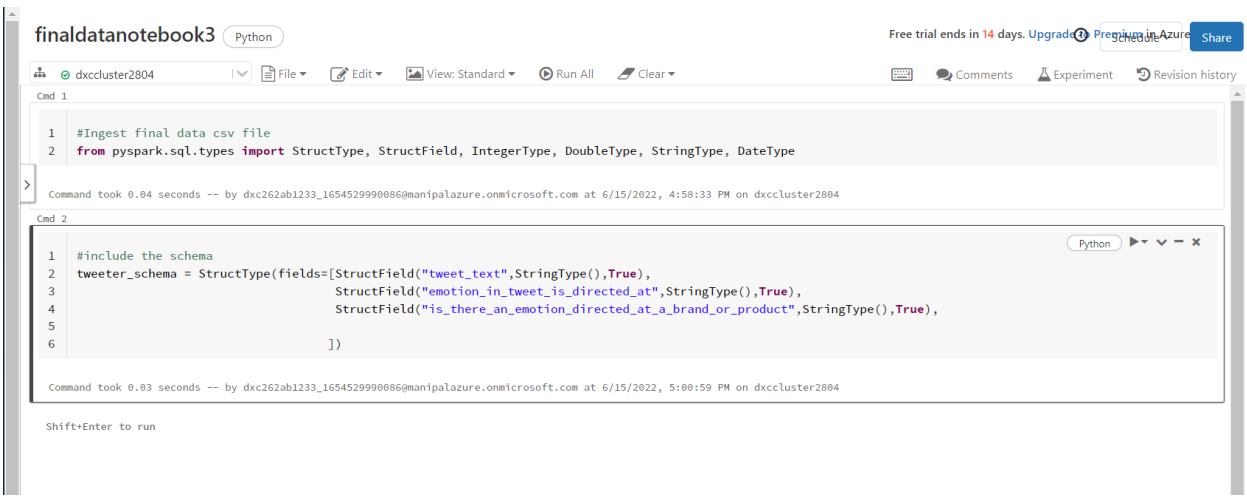
Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read
7	1875153522092939229	nces_renamed_final_df.write.mode('overwrite').p... parquet at NativeMethodAccessorImpl.java:0 +details	2022/06/15 11:18:03	1 s	1/1	174.6 KiB	7.2 KiB	

To display the data

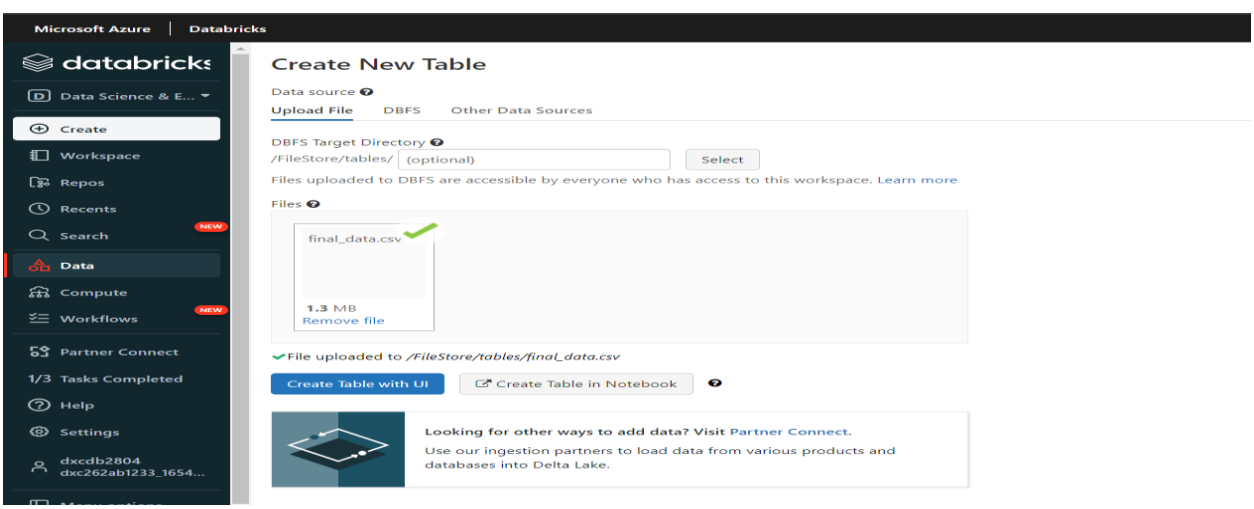


3) Using archive3.zip file - please ingest data into data bricks DBFS path & query the data display with notebooks accordingly

After creating the data bricks come to launch workspace and in that create a cluster and go to workspace and create a dataanalytics_project in that create a notebook



Now to create data frame again go to Data create table and drag and drop the data in it



Creating data frame

Cmd 3

```python
#creating a data frame
tweeter_df = spark.read \
.option("header",True) \
.schema(tweeter_schema) \
.csv("/FileStore/tables/final_data.csv")
```

tweeter_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 0.13 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 5:05:45 PM on dxccluster2804

Creating ingestion date

Cmd 4

```python
from pyspark.sql.functions import current_timestamp
```

Command took 0.03 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 5:11:08 PM on dxccluster2804

Cmd 5

```python
#add ingestion date to the data frame
tweeter_ingd_df = tweeter_df.withColumn("ingestion_date",current_timestamp())
```

tweeter_ingd_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
ingestion_date: timestamp
```

Renaming selected columns

Cmd 6

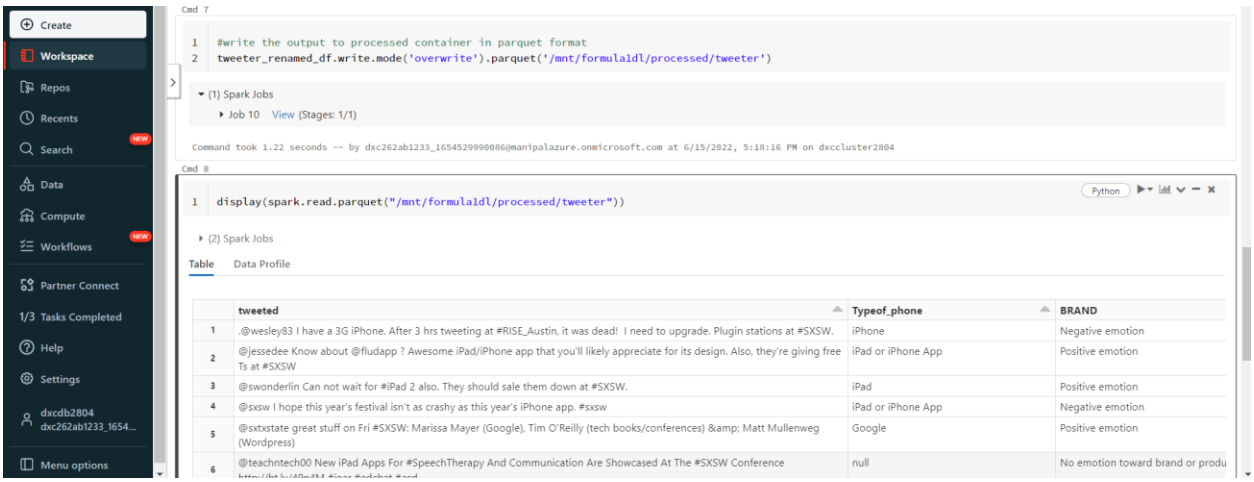
```python
tweeter_renamed_df = tweeter_ingd_df.withColumnRenamed("tweet_text","tweeted") \
.withColumnRenamed("emotion_in_tweet_is_directed_at","Typeof_phone") \
.withColumnRenamed("is_there_an_emotion_directed_at_a_brand_or_product","BRAND") \
.withColumnRenamed("ingestion_date","INGESTION_DATE")
```

tweeter_renamed_df: pyspark.sql.dataframe.DataFrame

```
tweeted: string
Typeof_phone: string
BRAND: string
INGESTION_DATE: timestamp
```

Command took 0.04 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 5:16:12 PM on dxccluster2804

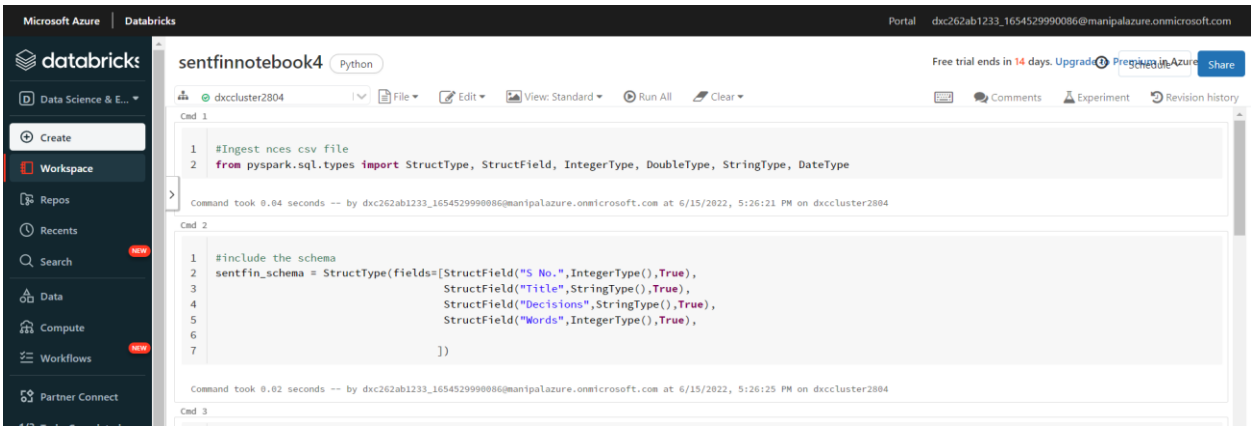
writing the output to processed container in parquet format and displaying it



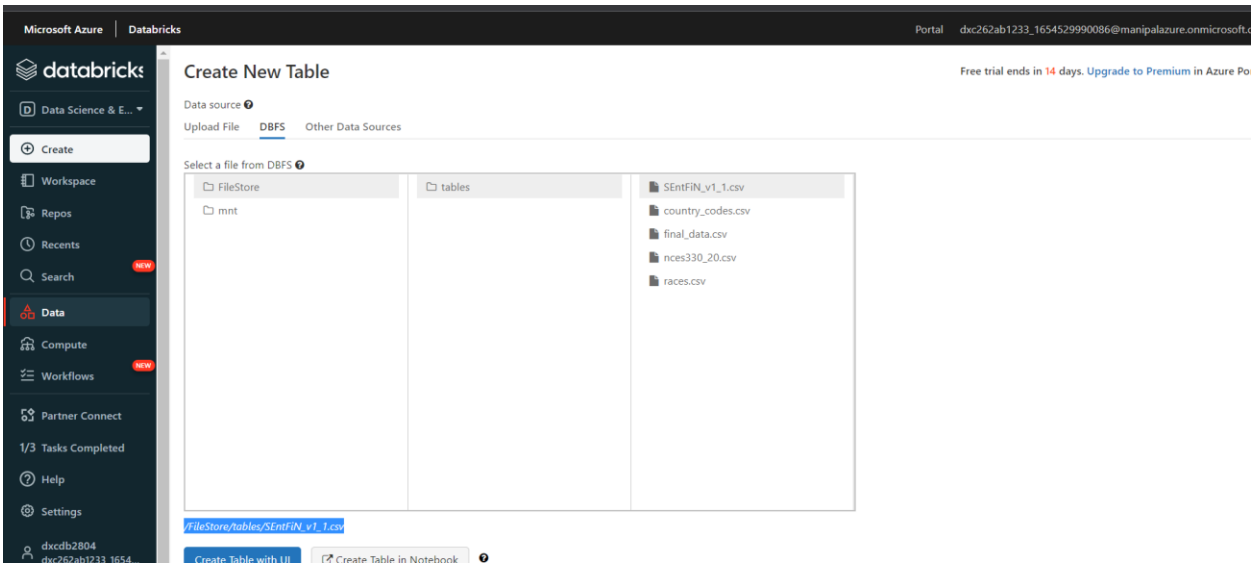
4) Using archive4.zip file - please ingest data into data bricks DBFS path & query the data display with notebooks accordingly

After creating the data bricks come to launch workspace and in that create a cluster and go to workspace and create a dataanalytics_project in that create a notebook

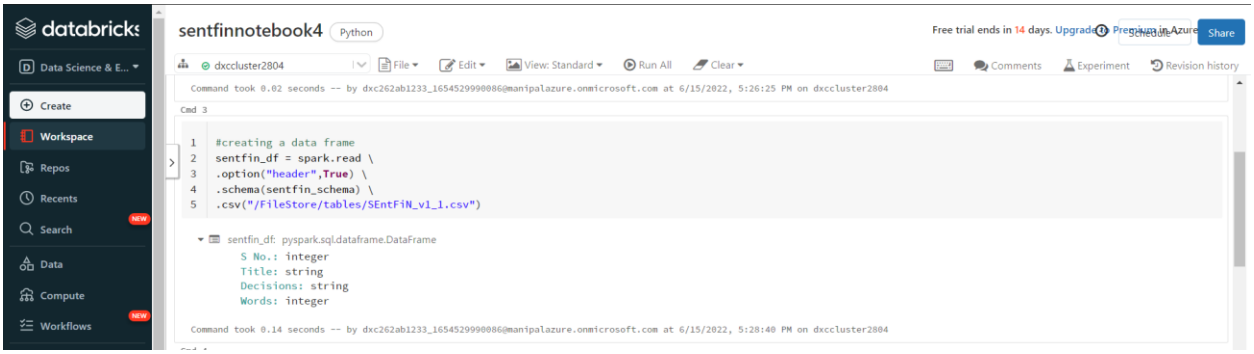
Ingest the data and giving schema for the data



Creating a data frame first upload the file in data create table and drag and drop the file and copy the location and paste in the data frame we create



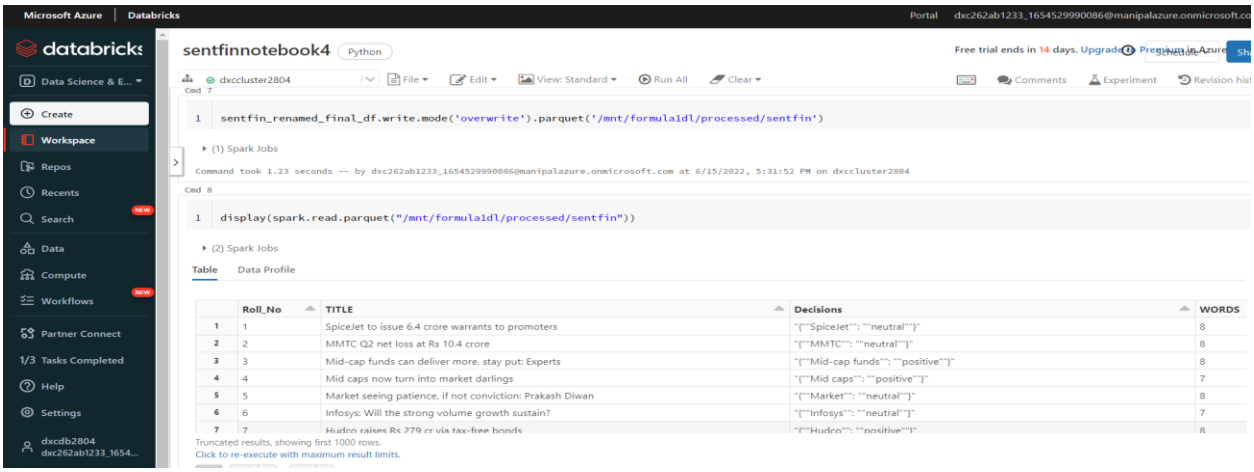
Creating a data frame



Creating ingestion date

Renaming the selected columns

writing the output to processed container in parquet format and displaying it



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

After creating the data bricks come to launch workspace and in that create a cluster and go to workspace and create a dataanalytics_project in that create a notebook

Ingest and including schema

```
#Ingest nces csv file
from pyspark.sql.types import StructType, StructField, IntegerType, DoubleType, StringType, DateType
```

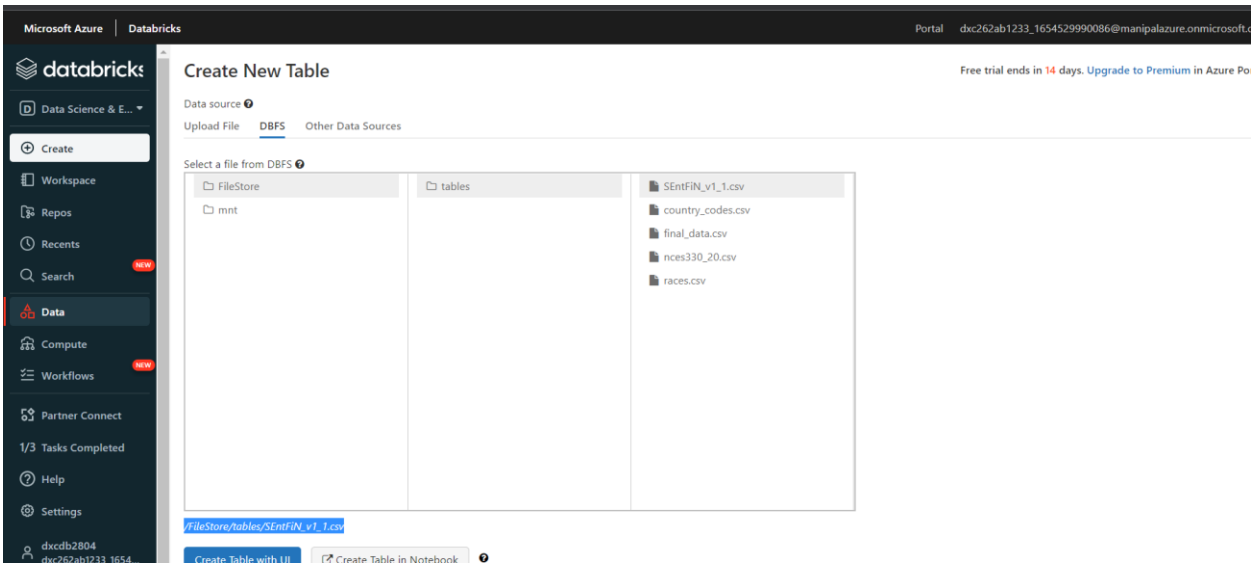
Command took 0.03 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 5:43:31 PM on dxccluster2804

Cmd 2

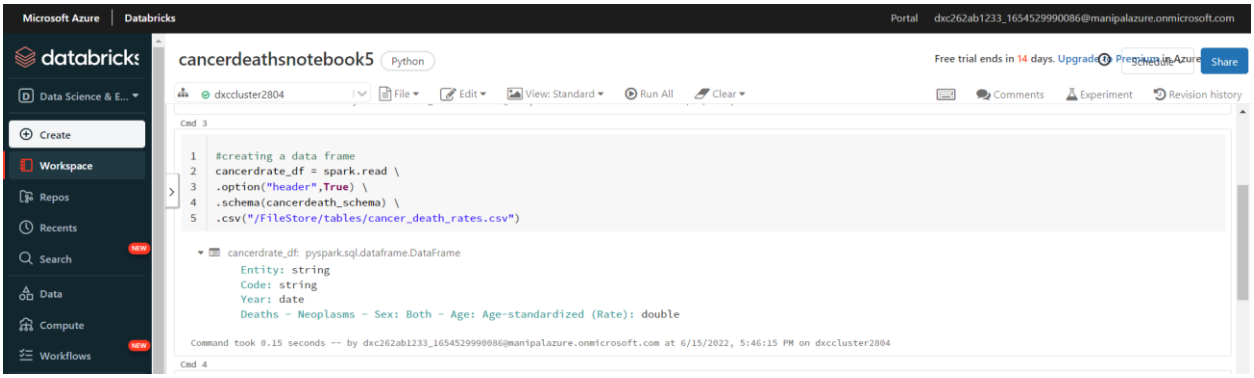
```
#include the schema
cancerdeath_schema = StructType(fields=[StructField("Entity",StringType(),True),
                                        StructField("Code",StringType(),True),
                                        StructField("Year",DateType(),True),
                                        StructField("Deaths - Neoplasms - Sex: Both - Age: Age-standardized (Rate)",DoubleType(),True),

                                       ])
```

Command took 0.04 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 5:46:05 PM on dxccluster2804

Creating a data frame first upload the file in data create table and drag and drop the file and copy the location and paste in the data frame we create



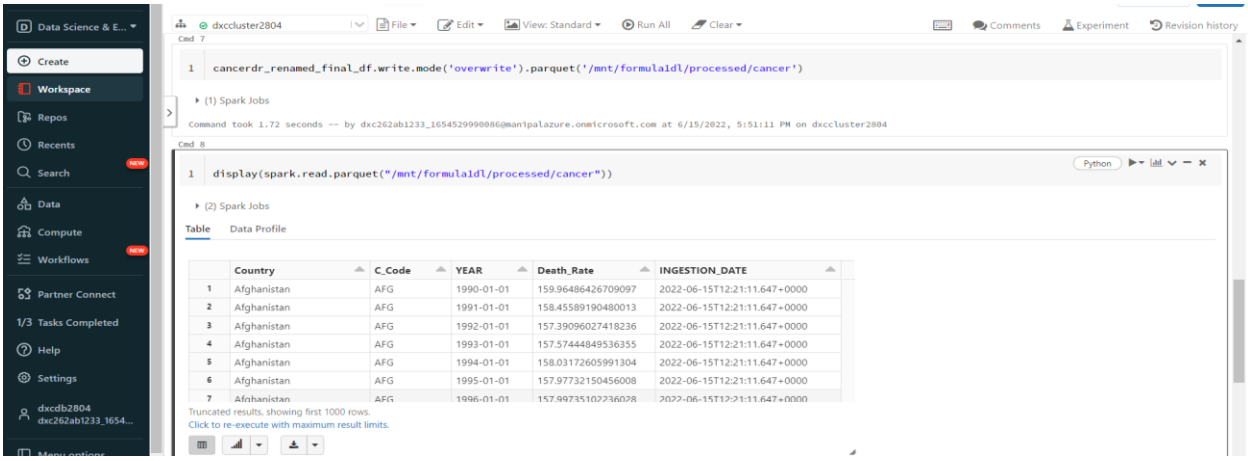
Creating data frame where we need to paste the location of the file



Creating ingestion date

Renaming the selected columns

writing the output to processed container in parquet format and displaying it



	Country	C_Code	YEAR	Death_Rate	INGESTION_DATE
1	Afghanistan	AFG	1990-01-01	159.96486426709097	2022-06-15T12:21:11.647+0000
2	Afghanistan	AFG	1991-01-01	158.45589190480013	2022-06-15T12:21:11.647+0000
3	Afghanistan	AFG	1992-01-01	157.39096027418236	2022-06-15T12:21:11.647+0000
4	Afghanistan	AFG	1993-01-01	157.57444849536355	2022-06-15T12:21:11.647+0000
5	Afghanistan	AFG	1994-01-01	158.03172605991304	2022-06-15T12:21:11.647+0000
6	Afghanistan	AFG	1995-01-01	157.97732150456008	2022-06-15T12:21:11.647+0000
7	Afghanistan	AFG	1996-01-01	157.99735102236028	2022-06-15T12:21:11.647+0000

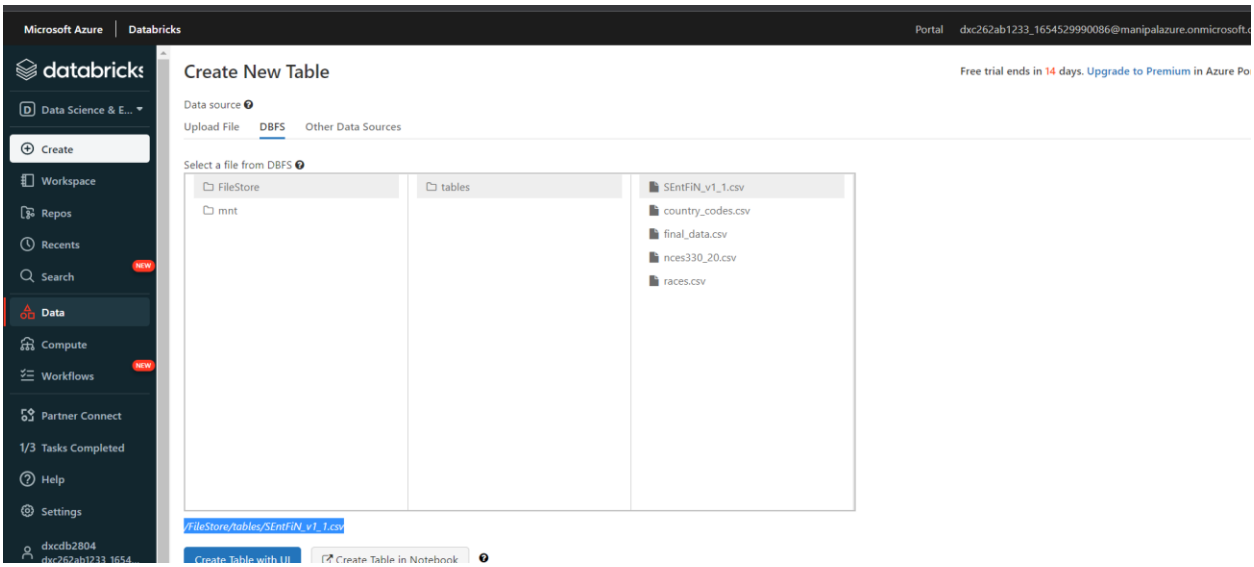
6) Using archive6.zip file - please ingest data into data bricks DBFS path & query the data display with notebooks accordingly

After creating the data bricks come to launch workspace and in that create a cluster and go to workspace and create a dataanalytics_project in that create a notebook

Ingest and including schema

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

inflation_schema = StructType(fields=[StructField("Country", StringType(), True),
                                      StructField("Country Code",StringType(),True),
                                       StructField("Year",IntegerType(),True),
                                       StructField("Inflation",DoubleType(),True),

                                      ])
```

For Creating a data frame first upload the file in data create table and drag and drop the file and copy the location and paste in the data frame we create



Creating data frame

```
inflation_df = spark.read \
.option("header", True) \
.schema(inflation_schema) \
.csv("/FileStore/tables/inflation-gdp.csv")
```

```
from pyspark.sql.functions import current_timestamp
```

#add ingestion date to the data frame

```
inflation_final_df = inflation_df.withColumn("ingestion_date",current_timestamp())
```

Renaming the selected columns

```
inflationgdp_renamed_final_df =
inflation_final_df.withColumnRenamed("Country",”COUNTRY") \

.withColumnRenamed("Country Code","COUNTRY_CODE") \

.withColumnRenamed("Year","IN_THEYEAR") \

.withColumnRenamed("ingestion_date","INGESTION_DATE")

.withColumnRenamed("Inflation","Inflation_gdb_ratio")
```

writing the output to processed container in parquet format and displaying it

```
inflationgdp_renamed_final_df.write.mode('overwrite').partitionBy('IN_THEYEAR').parquet('/
mnt/formula1dl/processed/inflation')
```

displaying it

```
display(spark.read.parquet("/mnt/formula1dl/processed/inflation"))
```