

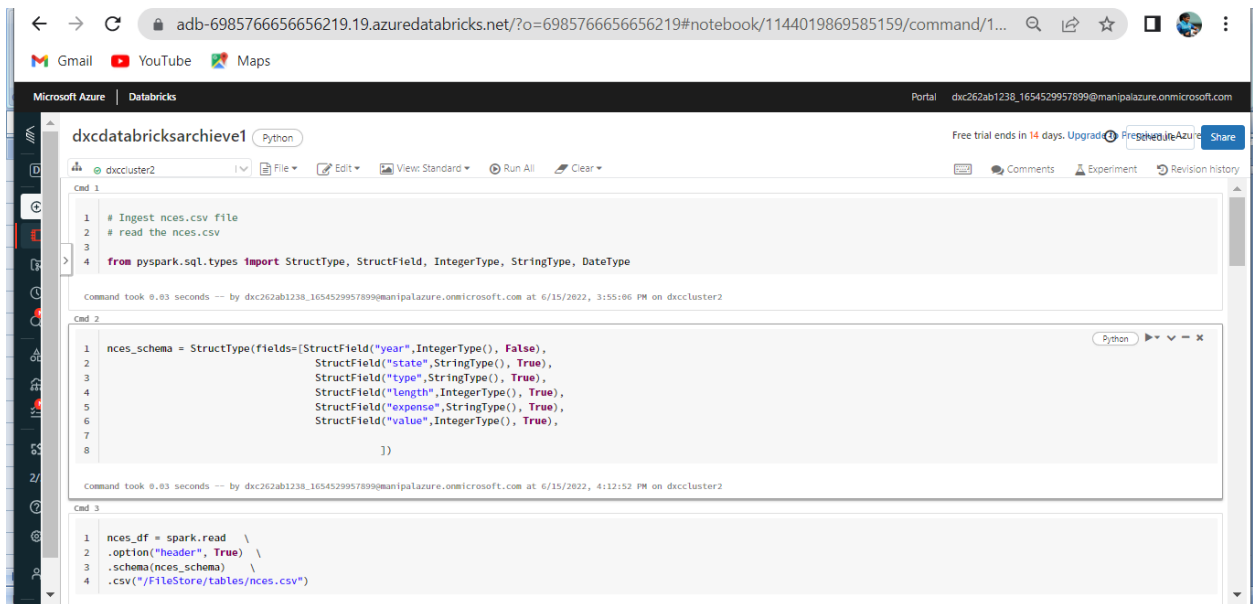
NAME:S.MAHALAKSHMI

REGNO:DXC262AB1229

DXC-262-ANALYTICS-B12-AZURE

## ASSIGNMENT 10

2.Using archive2.zip file - please ingest data into databricks DBFS path & query the data, redesign columns accordingly using dataframe commands - display with notebooks accordingly



The screenshot shows a Databricks notebook interface with the following code in the first three cells:

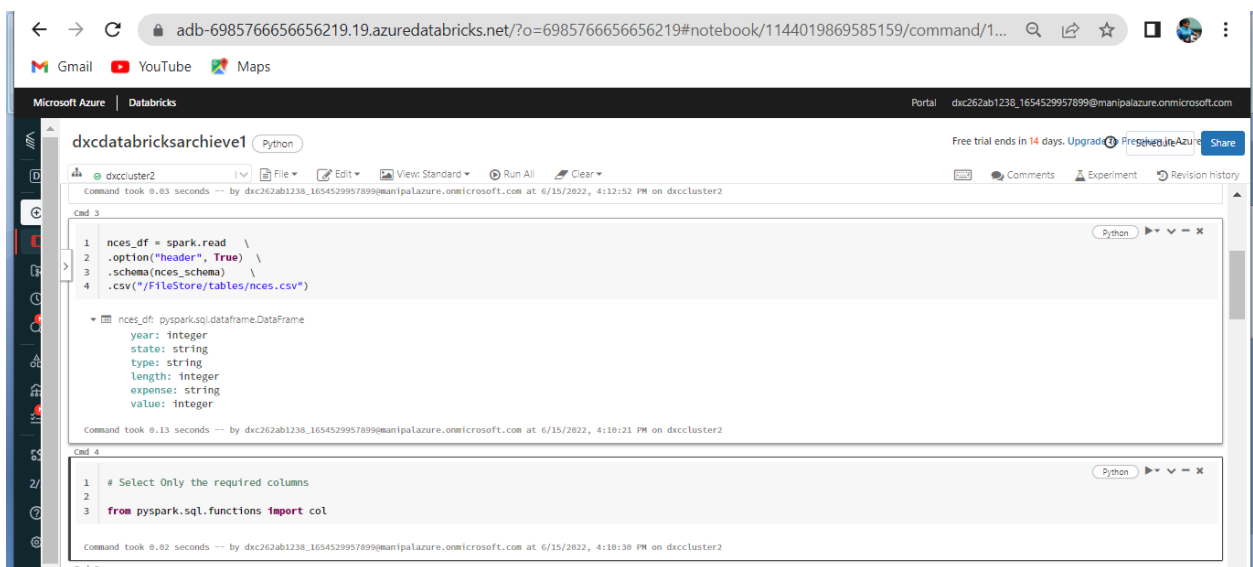
```
cd 1
1 # Ingest nces.csv file
2 # read the nces.csv
3
4 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DateType

Command took 0.83 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 3:55:06 PM on dxcccluster2

cd 2
1 nces_schema = StructType(fields=[StructField("year",IntegerType(), False),
2                                     StructField("state",StringType(), True),
3                                     StructField("type",StringType(), True),
4                                     StructField("length",IntegerType(), True),
5                                     StructField("expense",StringType(), True),
6                                     StructField("value",IntegerType(), True),
7                                     ])
8

Command took 0.83 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:12:52 PM on dxcccluster2

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



The screenshot shows the continuation of the Databricks notebook with the following code in the next two cells:

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

neces_df: pyspark.sql.dataframe.DataFrame
year: integer
state: string
type: string
length: integer
expense: string
value: integer

Command took 0.13 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:10:21 PM on dxcccluster2

cd 4
1 # Select Only the required columns
2
3 from pyspark.sql.functions import col

Command took 0.82 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:10:30 PM on dxcccluster2

cd 5
```

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free tier on Azure

Share

dxcccluster2

View Standard Run All Clear

Comments Experiment Revision history

cd 6

```
1 # rename the columns as required
2
3 nces_renamed_df = nces_selected_df.withColumnRenamed("year", "YEAR") \
4 .withColumnRenamed("state", "STATE") \
5 .withColumnRenamed("type", "TYPE") \
6 .withColumnRenamed("length", "LENGTH") \
7 .withColumnRenamed("expense", "EXPENSE") \
8 .withColumnRenamed("value", "VALUE")
```

nces\_renamed\_df: pyspark.sql.dataframe.DataFrame

YEAR: integer  
STATE: string  
TYPE: string  
LENGTH: integer  
EXPENSE: string  
VALUE: integer

Command took 0.84 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:16:14 PM on dxcccluster2

cd 7

```
1 #Add ingestion date to the dataframe
2
3 from pyspark.sql.functions import current_timestamp
```

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free tier on Azure

Share

dxcccluster2

View Standard Run All Clear

Comments Experiment Revision history

cd 8

```
1 nces_final_df = nces_renamed_df.withColumn("ingestion_date", current_timestamp())
```

nces\_final\_df: pyspark.sql.dataframe.DataFrame

YEAR: integer  
STATE: string  
TYPE: string  
LENGTH: integer  
EXPENSE: string  
VALUE: integer  
ingestion\_date: timestamp

Command took 0.83 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:20:18 PM on dxcccluster2

cd 9

```
1 # Write data to datalake as parquet
2
3
4 nces_final_df.write.mode("overwrite").parquet("/mnt/formula1dl/processed/nces")
5
```

(1) Spark Jobs

Job 32 View (Stages: 1/1)

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

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

dxcdatabricksarchive1 Python

dxcluster2

VALUE: integer  
Ingestion\_date: timestamp

Command took 0.93 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

cd 9

```
1 # Write data to datalake as parquet
2
3 nces_final_df.write.mode("overwrite").parquet("/mnt/formulaId1/process
4
5
```

(1) Spark Jobs

Job 32 View (Stages: 1/1)

Command took 1.92 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

cd 10

```
1 display(spark.read.parquet("/mnt/formulaId1/processed/nces"))
```

(2) Spark Jobs

Table Data Profile

YEAR	STATE	TYPE	LENGTH	EXP
1	2013	Alabama	Private	null

Details for Job 32

Status: SUCCEEDED

Submitted: 2022/06/15 10:52:11

Duration: 1 s

Associated SQL Query: 146

Job Group: 2868111045302864447\_7230163325472196676\_132d153465694e55a47c89da3e684f4d

Completed Stages: 1

Event Timeline

DAG Visualization

Completed Stages (1)

Page: 1 1 Pages. Jump to 1 - Show 100 items in a page. Go

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle
32	2868111045302864447	# Write data to datalake as parquet nces_fina... parquet at NativeMethodAccessorImpl.java0 +details	2022/06/15 10:52:11	1 s	1/1	174.6 KiB	15.0 KiB	

Page: 1 1 Pages. Jump to 1 - Show 100 items in a page. Go

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

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

dxcdatabricksarchive1 Python

dxcluster2

cd 10

```
1 display(spark.read.parquet("/mnt/formulaId1/processed/nces"))
```

(2) Spark Jobs

Job 33 View (Stages: 1/1)

Job 34 View (Stages: 1/1)

Table Data Profile

YEAR	STATE	TYPE	LENGTH	EXPENSE	VALUE	Ingestion_date
1	2013	Alabama	Private	null	Fees/Tuition	13983
2	2013	Alabama	Private	null	Room/Board	8503
3	2013	Alabama	Public In-State	null	Fees/Tuition	4048
4	2013	Alabama	Public In-State	null	Fees/Tuition	8073
5	2013	Alabama	Public In-State	null	Room/Board	8473
6	2013	Alabama	Public Out-of-State	null	Fees/Tuition	7736
7	2013	Alabama	Public Out-of-State	null	Fees/Tuition	20380

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

Command took 0.71 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:22:57 PM on dxcluster2

cd 11



Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

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

Python

cdxcluster2

cd 12

```
1 neces_selected_df.show()
```

(1) Spark Jobs

year	state	type	length	expense	value
2013	Alabama	Private	null	Fees/Tuition	13983
2013	Alabama	Private	null	Room/Board	8593
2013	Alabama	Public In-State	null	Fees/Tuition	4048
2013	Alabama	Public In-State	null	Fees/Tuition	8073
2013	Alabama	Public In-State	null	Room/Board	8473
2013	Alabama	Public Out-of-State	null	Fees/Tuition	7736
2013	Alabama	Public Out-of-State	null	Fees/Tuition	20380
2013	Alabama	Public Out-of-State	null	Room/Board	8473
2013	Alaska	Private	null	Fees/Tuition	21496
2013	Alaska	Private	null	Room/Board	8923
2013	Alaska	Public In-State	null	Fees/Tuition	3972
2013	Alaska	Public In-State	null	Fees/Tuition	6317
2013	Alaska	Public In-State	null	Room/Board	9098
2013	Alaska	Public Out-of-State	null	Fees/Tuition	4150
2013	Alaska	Public Out-of-State	null	Fees/Tuition	18790
2013	Alaska	Public Out-of-State	null	Room/Board	9098
2013	Arizona	Private	null	Fees/Tuition	11658
2013	Arizona	Private	null	Room/Board	8744

Command took 0.23 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:28:05 PM on dxccluster2

cd 13

3. Using archive3.zip file - please ingest data into databricks DBFS path & query the data  
redesign columns accordingly using dataframe commands - display with notebooks accordingly

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

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

Python

cdxcluster2

cd 1

```
1 # Ingest final.csv file
2 # read the final.csv
3
4 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DateType
```

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:34:43 PM on dxccluster2

cd 2

```
1 final_schema = StructType(fields=[StructField("tweet_text",StringType(), True),
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                                     ])
5
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:37:22 PM on dxccluster2

cd 3

```
1 final_df = spark.read \
2     .option("header", True) \
3     .schema(final_schema) \
4     .csv("/FileStore/tables/final.csv")
```

final\_df: pyspark.sql.dataframe.DataFrame

```
tweet_text: string
emotion in tweet is directed at: string
```

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free tier on Azure. Share

dxcccluster2 | File Edit View: Standard Run All Clear

6

```
final_renamed_df: pyspark.sql.dataframe.DataFrame
  TWEET: string
  EMOTION: string
  BRAND: string
```

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:43:43 PM on dxcccluster2

Cmd 7

```
1 #Add ingestion date to the dataframe
2
3 from pyspark.sql.functions import current_timestamp
```

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:44:44 PM on dxcccluster2

Cmd 8

```
1 final_final_df = final_renamed_df.withColumn("ingestion_date", current_timestamp())
```

```
final_final_df: pyspark.sql.dataframe.DataFrame
  TWEET: string
  EMOTION: string
  BRAND: string
  ingestion_date: timestamp
```

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:45:17 PM on dxcccluster2

Cmd 9

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free tier on Azure. Share

dxcccluster2 | File Edit View: Standard Run All Clear

Cmd 3

```
1 final_df = spark.read \
2 .option("header", True) \
3 .schema(final_schema) \
4 .csv("/FileStore/tables/final.csv")
```

```
final_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 dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:39:24 PM on dxcccluster2

Cmd 4

```
1 # Select Only the required columns
2
3 from pyspark.sql.functions import col
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:39:53 PM on dxcccluster2

Cmd 5

```
1 final_selected_df = final_df.select(col("tweet_text"), col("emotion_in_tweet_is_directed_at"), col("is_there_an_emotion_directed_at_a_brand_or_product"))
```

```
final_selected_df: pyspark.sql.dataframe.DataFrame
  tweet_text: string
```

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free tier or Azure

Share

dxcdatabricksarchive2 Python

dxcccluster2

Command took 0.83 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

```
1 # Write data to datalake as parquet
2
3 final_final_df.write.mode("overwrite").parquet("/mnt/formalaidl/processed/final")
4
5
```

(1) Spark Jobs

Job 37 View (Stages: 1/1)

Command took 1.32 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

cd 10

```
1 display(spark.read.parquet("/mnt/formalaidl/processed/final"))
```

(2) Spark Jobs

Table Data Profile

TWEET
1 @wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to
2 @jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate
3 Ts at #SXSW
3 @swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.

Details for Job 37

Status: SUCCEEDED

Submitted: 2022/06/15 11:16:02

Duration: 0.7 s

Associated SQL Query: 162

Job Group: 5476384096003780053\_8862941412296893274\_9490f26128004f86827cc91fa9903a6

Completed Stages: 1

Event Timeline

DAG Visualization

Completed Stages (1)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read
37	5476384096003780053	# Write data to datalake as parquet final_fin... parquet at NativeMethodAccessorImpl.java:0 + details	2022/06/15 11:16:02	0.7 s	1/1	1242.0 KB	553.2 KB	

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

iPad Positive emotion 2022-06-15T11:16:02.193+00

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free tier or Azure

Share

dxcdatabricksarchive2 Python

dxcccluster2

Command took 2.42 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:46:29 PM on dxcccluster2

```
1 display(spark.read.parquet("/mnt/formalaidl/processed/final"))
```

(2) Spark Jobs

Table Data Profile

TWEET	EMOTION	BRAND	ingestion_date
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	Negative emotion	2022-06-15T11:16:02.193+00
2 @jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate for its design. Also, they're giving free Ts at #SXSW	iPad or iPhone App	Positive emotion	2022-06-15T11:16:02.193+00
3 @swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.	iPad	Positive emotion	2022-06-15T11:16:02.193+00
4 @sxsw I hope this year's festival isn't as crashy as this year's iPhone app. #sxsw	iPad or iPhone App	Negative emotion	2022-06-15T11:16:02.193+00
5 @ixststate great stuff on Fri #SXSW: Marissa Mayer (Google), Tim O'Reilly (tech books/conferences) & Matt Mullenweg (WordPress)	Google	Positive emotion	2022-06-15T11:16:02.193+00
6 @teachntech00 New iPad Apps For #SpeechTherapy And Communication Are Showcased At The #SXSW Conference https://ht.ly/49n4M #lear-edchat #asd	null	No emotion toward brand or product	2022-06-15T11:16:02.193+00
7 null	null	No emotion toward brand or product	2022-06-15T11:16:02.193+00
- #SXSW is just starting. #CTIA is around the corner and #ooodoio is only a hop skip and a jump from there. good time to be an	Android	Positive emotion	2022-06-15T11:16:02.193+00

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

Command took 2.42 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:46:29 PM on dxcccluster2

cd 11

Microsoft Azure | Databricks

dxcdatabricksarchive2 Python

Free trial ends in 14 days. Upgrade to Preset Azure Share

dxcluster2 File Edit View: Standard Run All Clear

cmd 18

```
1 display(spark.read.parquet("/mnt/forulalld/processed/final"))
```

(2) Spark Jobs

Table Data Profile

Sort by: Feature order Reverse order Feature search (regex enabled)

Features: float(1) string(3)

Numeric Features (1)

count	missing	mean	std dev	zeros	min	median	max	custom
9,261	0%	1,668	0	0%	1,668	1,668	1,668	data type: timestamp min: 2022-06-15T11:16:02.193Z max: 2022-06-15T11:16:02.193Z

Categorical Features (3)

count	missing	unique	top	freq top	avg len	custom
9,260	0.01%	8,797	Google t...	46	103.01	data type: string
3,306	64.3%	20	iPad	945	9.65	data type: string

EMOTION

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

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585174/command/1...

Microsoft Azure | Databricks

dxcdatabricksarchive2 Python

Free trial ends in 14 days. Upgrade to Preset Azure Share

dxcluster2 File Edit View: Standard Run All Clear

cmd 11

```
1 display(final_selected_df)
```

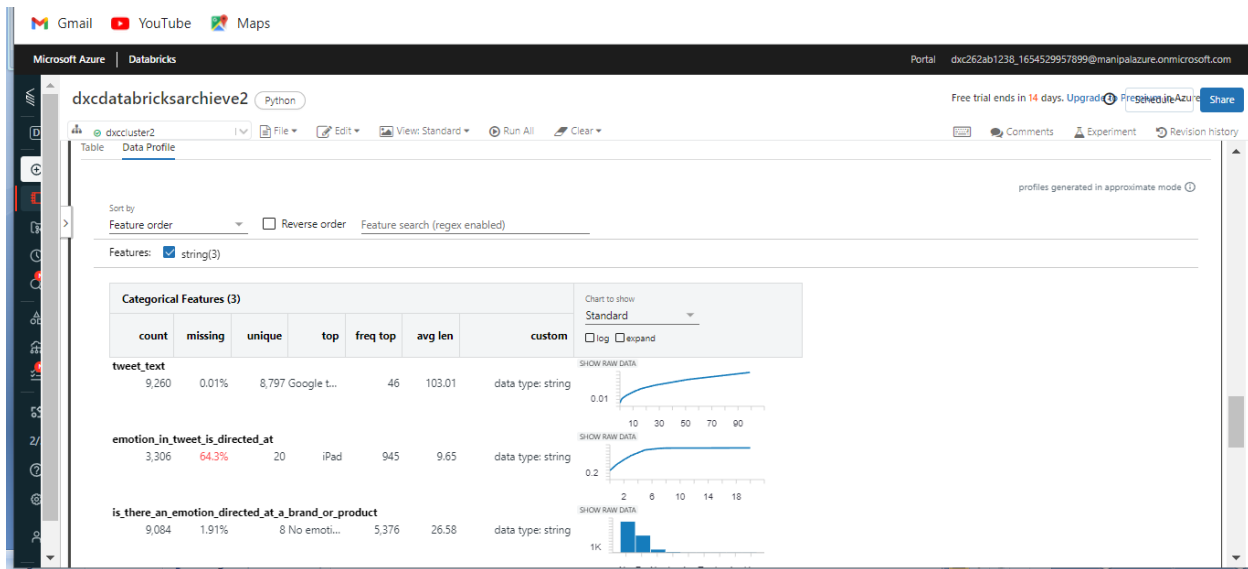
(1) Spark Jobs

Table Data Profile

tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_product
1 J@wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to upgrade. Plug in stations at #SXSW.	iPhone	Negative emotion
2 @jessedee Know about @fudapp ? Awesome iPad/iPhone app that you'll likely appreciate for its design. Also, they're giving free Ts at #SXSW	iPad or iPhone App	Positive emotion
3 @swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.	iPad	Positive emotion
4 @sxsw I hope this year's festival isn't as crashy as this year's iPhone app. #sxsw	iPad or iPhone App	Negative emotion
5 @ixbstaste great stuff on Fri #SXSW: Marissa Mayer (Google), Tim O'Reilly (tech books/conferences) & Matt Mullenweg (WordPress)	Google	Positive emotion
6 @teachntech00 New iPad Apps For #SpeechTherapy And Communication Are Showcased At The #SXSW Conference http://ht.ly/49n4M #lear #edchat #asd	null	No emotion toward brand or product
7 null	null	No emotion toward brand or product
8 #SXSW is just starting. #CTIA is around the corner and #googleio is only a hop skip and a jump from there. good time to be an	Android	Positive emotion

Command took 0.24 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:47:25 PM on dxcluster2





```
adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585174/command/1...
```

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free on Azure

dxcc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Python

dxcc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

cmd 12

```
1 final_selected_df.show()
```

(1) Spark Jobs

tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_product
[@wesley83 I have...]	iPhone	Negative emotion
[@jessedee Know ab...]	iPad or iPhone App	Positive emotion
[@swonderlin Can n...]	iPad	Positive emotion
[@ssw I hope this...]	iPad or iPhone App	Negative emotion
[@stxtstate great ...]	Google	Positive emotion
[@teachntech00 New...]	null	No emotion toward...
[ null]	null	No emotion toward...
[#SKSM is just sta...]	Android	Positive emotion
[Beautifully smart...]	iPad or iPhone App	Positive emotion
[Counting down the...]	Apple	Positive emotion
[Excited to meet t...]	Android	Positive emotion
[Find & Start ...]	Android App	Positive emotion
[Foursquare ups th...]	Android App	Positive emotion
[Gotta love this #...]	Other Google prod...	Positive emotion
[Great #ssw ipad ...]	iPad or iPhone App	Positive emotion
[haha, awesomely r...]	iPad or iPhone App	Positive emotion
[Holler Gram for i...]	null	No emotion toward...
[I just noticed DS...]	iPhone	Negative emotion

Command took 0.28 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:47:49 PM on dxcc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

cmd 13

4. Using archive4.zip file - please ingest data into databricks DBFS path & query the data redesign columns accordingly using daframe commands - display with notebooks accordingly

adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Preset Azure Share

dxcdatabricksarchie3 Python

dxcluster2

Cmd 1

```
1 # Ingest company.csv file
2 # read the company.csv
3
4 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DateType
```

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:02:30 PM on dxcluster2

Cmd 2

```
1 company_schema = StructType(fields=[StructField("Sno",IntegerType(), True),
2                                     StructField("Title",StringType(), True),
3                                     StructField("Decision",StringType(), True),
4                                     StructField("Words",IntegerType(), True),
5
6                                     ])
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:04:46 PM on dxcluster2

Cmd 3

```
1 company_df = spark.read \
2 .option("header", True) \
3 .schema(company_schema) \
4 .csv("/FileStore/tables/company.csv")
```

company\_df: pyspark.sql.dataframe.DataFrame  
Sno: Integer

adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Preset Azure Share

dxcdatabricksarchie3 Python

dxcluster2

Cmd 3

```
1 company_df = spark.read \
2 .option("header", True) \
3 .schema(company_schema) \
4 .csv("/FileStore/tables/company.csv")
```

company\_df: pyspark.sql.dataframe.DataFrame  
Sno: Integer  
Title: string  
Decision: string  
Words: Integer

Command took 0.13 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:04:53 PM on dxcluster2

Cmd 4

```
1 # Select Only the required columns
2
3 from pyspark.sql.functions import col
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:05:14 PM on dxcluster2

Cmd 5

```
1 company_selected_df = company_df.select(col("Sno"), col("Title"), col("Decision"), col("Words"))
```

company\_selected\_df: pyspark.sql.dataframe.DataFrame  
Sno: Integer

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

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

dxcdatabricksarchive3 Python

dxcluster2

Cell 5

```
1 company_selected_df = company_df.select(col("Sno"), col("Title"), col("Decision"), col("Words"))
```

company\_selected\_df: pyspark.sql.dataframe.DataFrame

Sno: integer  
Title: string  
Decision: string  
Words: integer

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:11:30 PM on dxcluster2

Cell 6

```
1 # rename the columns as required
2
3 company_renamed_df = company_selected_df.withColumnRenamed("Sno", "SNO") \
4 .withColumnRenamed("Title", "TITLE") \
5 .withColumnRenamed("Decision", "DECISION") \
6 .withColumnRenamed("Words", "WORDS")
```

company\_renamed\_df: pyspark.sql.dataframe.DataFrame

SNO: integer  
TITLE: string  
DECISION: string  
WORDS: integer

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:12:41 PM on dxcluster2

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

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

dxcdatabricksarchive3 Python

dxcluster2

Cell 7

```
1 TITLE: string
2 DECISION: string
3 WORDS: integer
```

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:12:41 PM on dxcluster2

Cell 7

```
1 #Add ingestion date to the dataframe
2
3 from pyspark.sql.functions import current_timestamp
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:13:31 PM on dxcluster2

Cell 8

```
1 company_final_df = company_renamed_df.withColumn("ingestion_date", current_timestamp())
```

company\_final\_df: pyspark.sql.dataframe.DataFrame

SNO: integer  
TITLE: string  
DECISION: string  
WORDS: integer  
ingestion\_date: timestamp

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:14:17 PM on dxcluster2

Cell 9

```
1 # Write data to datalake as parquet
```

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Azure Share

Python

dxcdatabricksarchive3

dxcccluster2

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure

```
1 # Write data to dataLake as parquet
2
3
4 company_final_df.write.mode("overwrite").parquet("/mnt/formulaidl/processed/company")
```

(1) Spark Jobs

Job 55 View (Stages: 1/1)

Command took 1.22 seconds -- by dxc262ab1238\_1654529957899@manipalazure

Cad 10

```
1 display(spark.read.parquet("/mnt/formulaidl/processed/company"))
```

(2) Spark Jobs

Table Data Profile

Details for Job 55

Status: SUCCEEDED

Submitted: 2022/06/15 11:44:54

Duration: 0.7 s

Associated SQL Query: 177

Job Group: 5355714283519937126\_4996741183705593270\_2f09b4fc3694eb6b8b39a4bc6df426f

Completed Stages: 1

Event Timeline

DAG Visualization

Completed Stages (1)

Page: 1

1 Pages. Jump to 1. Show 100 items in a page. Go

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output
61	5355714283519937126	# Write data to dataLake as parquet company_f... parquet at NativeMethodAccessorImpl.java:0	2022/06/15 11:44:54	0.7 s	1/1	1163.7 KiB	54 KiB

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Azure Share

Python

dxcdatabricksarchive3

dxcccluster2

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure

```
1 display(spark.read.parquet("/mnt/formulaidl/processed/company"))
```

(2) Spark Jobs

Job 56 View (Stages: 1/1)

Job 57 View (Stages: 1/1)

Table Data Profile

	SNO	TITLE	DECISION	WORDS
1	1	SpiceJet to issue 6.4 crore warrants to promoters	["SpiceJet"; "neutral"]	8
2	2	MMTC Q2 net loss at Rs 10.4 crore	["MMTC"; "neutral"]	8
3	3	Mid-cap funds can deliver more, stay put: Experts	["Mid-cap funds"; "positive"]	8
4	4	Mid caps now turn into market darlings	["Mid caps"; "positive"]	7
5	5	Market seeing patience, if not conviction: Prakash Diwan	["Market"; "neutral"]	8
6	6	Infosys: Will the strong volume growth sustain?	["Infosys"; "neutral"]	7
7	7	Hudco raises Rs 279 cr via tax-free bonds	["Hudco"; "positive"]	8

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

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...

GmailYouTubeMaps

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. UpgradePreschedule AzureShare

dxcdatabricksarchive3Python

dxcccluster2FileEditView: StandardRun AllClear

1display(company\_selected\_df)

(1) Spark Jobs

TableData Profile

	Sno	Title	Decision	Words
1	1	SpiceJet to issue 6.4 crore warrants to promoters	"["SpiceJet": "neutral"]"	8
2	2	MMTC Q2 net loss at Rs 10.4 crore	"["MMTC": "neutral"]"	8
3	3	Mid-cap funds can deliver more, stay put: Experts	"["Mid-cap funds": "positive"]"	8
4	4	Mid caps now turn into market darlings	"["Mid caps": "positive"]"	7
5	5	Market seeing patience, if not conviction: Prakash Diwan	"["Market": "neutral"]"	8
6	6	Infosys: Will the strong volume growth sustain?	"["Infosys": "neutral"]"	7
7	7	Hudco raises Rs 279 cr via tax-free bonds	"["Hudco": "positive"]"	8

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

Command took 0.27 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:15:43 PM on dxcccluster2

Cmd 12

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...

GmailYouTubeMaps

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. UpgradePreschedule AzureShare

dxcdatabricksarchive3Python

dxcccluster2FileEditView: StandardRun AllClear

TableData Profile

profiles generated in approximate mode

Sort by  
Feature orderReverse orderFeature search (regex enabled)

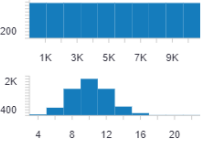
Features: ☒ int(2) ☒ string(2)

Numeric Features (2)

	count	missing	mean	std dev	zeros	min	median	max	custom
Sno	10.8k	0%	5,377	3,104.27	0%	1	5,376	10.8k	data type: int
Words	7,897	26.56%	9.74	2.38	0%	3	10	23	data type: int

Chart to show  
Standard

☐ log☐ expand



adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free Trial Azure Share

dxcdatabricksarchive3 Python

dxcccluster2

Command took 0.27 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:13:13 PM on dxcccluster2

cd 12

```
1 company_selected_df.show()
```

(1) Spark Jobs

Sno	Title	Decision	Words
1	SpiceJet to issue...	["SpiceJet": "...]	8]
2	MMTC Q2 net loss ...	["MMTC": "neu...]	8]
3	Mid-cap funds can...	["Mid-cap funds...]	8]
4	Mid caps now turn...	["Mid caps": "...]	7]
5	Market seeing pat...	["Market": "...]	8]
6	Infosys: Will the...	["Infosys": "...]	7]
7	Hudco raises Rs 2...	["Hudco": "po...]	8]
8	HOEC could retest...	["HOEC": "neu...]	7]
9	Gold shines on se...	["Gold": "pos...]	null]
10	Genpact appoints ...	["Genpact": "...]	7]
11	EXL beats profit ...	["EXL": "posi...]	7]
12	Wait and watch on...	["Bharti Airtel...]	8]
13	Would stick to ba...	["banking": "...]	null]
14	MSCI adds Aurobin...	["Aurobindo Pha...]	null]
15	Ashok Leyland rai...	["Ashok Leyland...]	8]
16	At Wipro, growth ...	["Wipro": "ne...]	6]
17	Why Chinese stock...	["us": "negat...]	null]
18	US stocks finish...	["tech": "neg...]	null]

Command took 0.29 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:16:31 PM on dxcccluster2

5. Using archive5.zip file - please ingest data into databricks DBFS path & query the data redesign columns accordingly using dataframe commands - display with notebooks accordingly

adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Free Trial Azure Share

dxcdatabricksarchive4 Python

dxcccluster2

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:26:32 PM on dxcccluster2

cd 1

```
1 # Ingest cancer.csv file
2 # read the cancer.csv
3
4 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DateType
```

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:30:10 PM on dxcccluster2

cd 2

```
1 cancer_schema = StructType(fields=[StructField("Entity", StringType(), True),
2                                     StructField("Code", StringType(), True),
3                                     StructField("Year", IntegerType(), True),
4                                     StructField("Death", IntegerType(), True),
5
6                                     ])
```

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:30:10 PM on dxcccluster2

cd 3

```
1 cancer_df = spark.read \
2 .option("header", True) \
3 .schema(cancer_schema) \
```

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

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

dxcccluster2 File Edit View: Standard Run All Clear

Cmd 3

```
1 cancer_df = spark.read \
2   .option("header", True) \
3   .schema(cancer_schema) \
4   .csv("/FileStore/tables/cancer.csv")
```

cancer\_df pyspark.sql.dataframe.DataFrame  
Entity: string  
Code: string  
Year: integer  
Death: integer

Command took 0.11 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:32:12 PM on dxcccluster2

Cmd 4

```
1 # Select Only the required columns
2
3 from pyspark.sql.functions import col
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:32:45 PM on dxcccluster2

Cmd 5

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

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

dxcccluster2 File Edit View: Standard Run All Clear

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:32:45 PM on dxcccluster2

Cmd 5

```
1 cancer_selected_df = cancer_df.select(col("Entity"), col("Code"), col("Year"), col("Death"))
```

cancer\_selected\_df pyspark.sql.dataframe.DataFrame  
Entity: string  
Code: string  
Year: integer  
Death: integer

Command took 0.05 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:33:49 PM on dxcccluster2

Cmd 6

```
1 #Add ingestion date to the dataframe
2
3 from pyspark.sql.functions import current_timestamp
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:34:23 PM on dxcccluster2

Cmd 7

adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade | Fresh Start | Share

dxcccluster2

View: Standard | Run All | Clear

Comments | Experiment | Revision history

Card 7

```
1 cancer_final_df = cancer_df.withColumn("ingestion_date", current_timestamp())
```

▼ cancer\_final\_df: pyspark.sql.dataframe.DataFrame

- Entity: string
- Code: string
- Year: integer
- Death: integer
- ingestion\_date: timestamp

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:35:09 PM on dxcccluster2

Card 8

```
1 # Write data to datalake as parquet
2
3
4 cancer_final_df.write.mode("overwrite").parquet("/mnt/formulaidl/processed/company")
```

▼ (1) Spark Jobs

- Job 65 View (Stages: 1/1)
- Stage 73: 1/1

adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade | Fresh Start | Share

dxcccluster2

View: Standard | Run All | Clear

Comments | Experiment | Revision history

Card 7

```
1 cancer_final_df = cancer_df.withColumn("ingestion_date",
```

▼ cancer\_final\_df: pyspark.sql.dataframe.DataFrame

- Entity: string
- Code: string
- Year: integer
- Death: integer
- ingestion\_date: timestamp

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure

Card 8

```
1 # Write data to datalake as parquet
2
3
4 cancer_final_df.write.mode("overwrite").parquet("/mnt/fo
```

▼ (1) Spark Jobs

- Job 65 View (Stages: 1/1)
- Stage 73: 1/1

Jobs | Stages | Storage | Environment | Executors | SQL | JDBC/ODBC Server | Structured Streaming

Details for Job 65

Status: SUCCEEDED

Submitted: 2022/06/15 12:06:40

Duration: 0.4 s

Associated SQL Query: 195

Job Group: 7110440155128268836\_5436340371620126902\_d40a2cb1033949afb0930b008360950c

Completed Stages: 1

Event Timeline

DAG Visualization

Completed Stages (1)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output
73	7110440155128268836	# Write data to datalake as parquet cancer_fi... parquet at NativeMethodAccessorImpl.java:0	2022/06/15 12:06:40	0.3 s	1/1	254.5 KIB	6.2 Kil



adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Azure Share

dxcdatabricksarchie4 Python

dxccuster2 File Edit View Standard Run All Clear

1 display(spark.read.parquet("/mnt/formula1/processed/company"))

(2) Spark Jobs

Table Data Profile

	Entity	Code	Year	Death	ingestion_date
1	Afghanistan	AFG	1990	null	2022-06-15T12:06:40.356+0000
2	Afghanistan	AFG	1991	null	2022-06-15T12:06:40.356+0000
3	Afghanistan	AFG	1992	null	2022-06-15T12:06:40.356+0000
4	Afghanistan	AFG	1993	null	2022-06-15T12:06:40.356+0000
5	Afghanistan	AFG	1994	null	2022-06-15T12:06:40.356+0000
6	Afghanistan	AFG	1995	null	2022-06-15T12:06:40.356+0000
7	Afghanistan	AFG	1996	null	2022-06-15T12:06:40.356+0000

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

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Azure Share

dxcdatabricksarchie4 Python

dxccuster2 File Edit View Standard Run All Clear




Table Data Profile

Sort by Feature order Reverse order Feature search (regex enabled)

Features: ☒ int(1) ☒ float(1) ☒ string(2) ☒ unknown(1)

profiles generated in approximate mode

Numeric Features (3)									Chart to show Standard
count	missing	mean	std dev	zeros	min	median	max	custom	<input type="checkbox"/> log <input type="checkbox"/> expand
Year	6,840	0%	2,004.5	8.66	0%	1,990	2,004	2,019	data type: int
Death	0	100%	0	0	0%	0	0	0	data type: int
ingestion_date	6,840	0%	1.668	0	0%	1.668	1.668	1.668	data type: timestamp min: 2022-06-15T12:06:40.356Z



adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Preset Azure Share

dxcdatabricksarchive4 Python

dxcluster2 File Edit View Standard Run All Clear 10 30 50 70 90

Command took 9.58 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:48:09 PM on dxcluster2

cd 10

```
1 display(cancer_selected_df)
```

(1) Spark Jobs

Table Data Profile

	Entity	Code	Year	Death
1	Afghanistan	AFG	1990	null
2	Afghanistan	AFG	1991	null
3	Afghanistan	AFG	1992	null
4	Afghanistan	AFG	1993	null
5	Afghanistan	AFG	1994	null
6	Afghanistan	AFG	1995	null
7	Afghanistan	AFG	1996	null

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

Command took 9.24 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:48:47 PM on dxcluster2

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585206/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade to Preset Azure Share

dxcdatabricksarchive4 Python

dxcluster2 File Edit View Standard Run All Clear

cd 11

```
1 cancer_selected_df.show()
```

(1) Spark Jobs

	Entity	Code	Year	Death
1	Afghanistan	AFG	1990	null
2	Afghanistan	AFG	1991	null
3	Afghanistan	AFG	1992	null
4	Afghanistan	AFG	1993	null
5	Afghanistan	AFG	1994	null
6	Afghanistan	AFG	1995	null
7	Afghanistan	AFG	1996	null
8	Afghanistan	AFG	1997	null
9	Afghanistan	AFG	1998	null
10	Afghanistan	AFG	1999	null
11	Afghanistan	AFG	2000	null
12	Afghanistan	AFG	2001	null
13	Afghanistan	AFG	2002	null
14	Afghanistan	AFG	2003	null
15	Afghanistan	AFG	2004	null
16	Afghanistan	AFG	2005	null
17	Afghanistan	AFG	2006	null
18	Afghanistan	AFG	2007	null

Command took 9.27 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:41:18 PM on dxcluster2

6.Using archive6.zip file - please ingest data into databricks DBFS path & query the data redesign columns accordingly using dafarame commands - display with notebooks accordingly

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585220

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Azure Share

dxcdatabricksarchive5 Python

dxcccluster2

cd 1

```
1 # Ingest consumer.csv file
2 # read the consumer.csv
3
4 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DateType
```

Command took 0.03 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:49:33 PM on dxcccluster2

cd 2

```
1 consumer_schema = StructType(fields=[StructField("Country",StringType(), True),
2                                     StructField("CountryCode",StringType(), True),
3                                     StructField("Year",IntegerType(), True),
4                                     StructField("Inflation",IntegerType(), True),
5
6                                     ])
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:51:19 PM on dxcccluster2

cd 3

```
1 consumer_df = spark.read \
2 .option("header", True) \
3 .schema(consumer_schema) \
4 .csv("/FileStore/tables/consumer.csv")
```

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585220/command/1...

Microsoft Azure | Databricks

Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Azure Share

dxcdatabricksarchive5 Python

dxcccluster2

cd 3

```
1 consumer_df = spark.read \
2 .option("header", True) \
3 .schema(consumer_schema) \
4 .csv("/FileStore/tables/consumer.csv")
```

▼ consumer\_df: pyspark.sql.dataframe.DataFrame  
Country: string  
CountryCode: string  
Year: integer  
Inflation: integer

Command took 0.14 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:52:49 PM on dxcccluster2

cd 4

```
1 # Select Only the required columns
2
3 from pyspark.sql.functions import col
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:53:09 PM on dxcccluster2

← → ↻ 🔒 adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585220/command/1... 🔍 📁 ☆ 🌐 ⋮

Gmail YouTube Maps

Microsoft Azure | Databricks Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Life Azure Share

dxcdatabricksarchive5 Python

dxcccluster2 | File Edit View: Standard Run All Clear

cmd 7

```
1 consumer_final_df = consumer_df.withColumn("ingestion_date", current_timestamp())
```

▼ consumer\_final\_df: pyspark.sql.dataframe.DataFrame

Country: string  
CountryCode: string  
Year: integer  
Inflation: integer  
ingestion\_date: timestamp

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:55:03 PM on dxcccluster2

cmd 8

```
1 # Write data to datalake as parquet
2
3
4 consumer_final_df.write.mode("overwrite").parquet("/mnt/formula1dl/processed/consumer")
```

▼ (1) Spark Jobs

Job 75 View (Stages: 1/1)

Command took 1.00 second -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:55:43 PM on dxcccluster2

cmd 9

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

← → ↻ 🔒 adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585220/command/1... 🔍 📁 ☆ 🌐 ⋮

Gmail YouTube Maps

Microsoft Azure | Databricks Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Life Azure Share

dxcdatabricksarchive5 Python

dxcccluster2 | File Edit View: Standard Run All Clear

cmd 5

```
1 # Select Only the required columns
2
3 from pyspark.sql.functions import col
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:53:09 PM on dxcccluster2

cmd 5

```
1 consumer_selected_df = consumer_df.select(col("Country"), col("CountryCode"), col("Year"), col("Inflation"))
```

▼ consumer\_selected\_df: pyspark.sql.dataframe.DataFrame

Country: string  
CountryCode: string  
Year: integer  
Inflation: integer

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:54:17 PM on dxcccluster2

cmd 6

```
1 #Add ingestion date to the dataframe
2
3 from pyspark.sql.functions import current_timestamp
```

Command took 0.02 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:54:33 PM on dxcccluster2

← → ↻ 🔒 adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585220/command/1... 🔍 📄 ⭐ 🌐

Gmail YouTube Maps

Microsoft Azure | Databricks Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Life Azure Share

Python

dxcdatabricksarchive5

dxccuster2

File Edit View Standard Run All Clear

Jobs Stages Storage Environment Executors SQL JDBC/ODBC Server Structured Streaming

cad 7

```
1 consumer_final_df = consumer_df.withColumn("ingestion_date", current_timestamp())
```

▼ consumer\_final\_df: pyspark.sql.dataframe.DataFrame

Country: string  
CountryCode: string  
Year: Integer  
Inflation: Integer  
ingestion\_date: timestamp

Command took 0.04 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

cad 8

```
1 # Write data to datalake as parquet
2
3
4 consumer_final_df.write.mode("overwrite").parquet("/mnt/formulaId1/processed/consumer")
```

▼ (1) Spark Jobs

▶ Job 75 View (Stages: 1/1)

Command took 1.00 second -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

cad 9

```
1 display(spark.read.parquet("/mnt/formulaId1/processed/consumer"))
```

Details for Job 75

Status: SUCCEEDED  
Submitted: 2022/06/15 12:25:43  
Duration: 0.6 s  
Associated SQL Query: 208  
Job Group: 8228088374590975492\_7796316010674541875\_956308015a824b52b2fd5c335b5e246d  
Completed Stages: 1

▶ Event Timeline  
▶ DAG Visualization

▼ Completed Stages (1)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read
85	8228088374590975492	# Write data to datalake as parquet consumer_... parquet at NativeMethodAccessorImpl.java:0 +details	2022/06/15 12:25:43	0.5 s	1/1	433.1 KB	7.9 KB	

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

← → ↻ 🔒 adb-6985766656656219.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585220/command/1... 🔍 📄 ⭐ 🌐

Gmail YouTube Maps

Microsoft Azure | Databricks Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

Free trial ends in 14 days. Upgrade Preset Life Azure Share

Python

dxcdatabricksarchive5

dxccuster2

File Edit View Standard Run All Clear

Jobs Stages Storage Environment Executors SQL JDBC/ODBC Server Structured Streaming

cad 9

```
1 display(spark.read.parquet("/mnt/formulaId1/processed/consumer"))
```

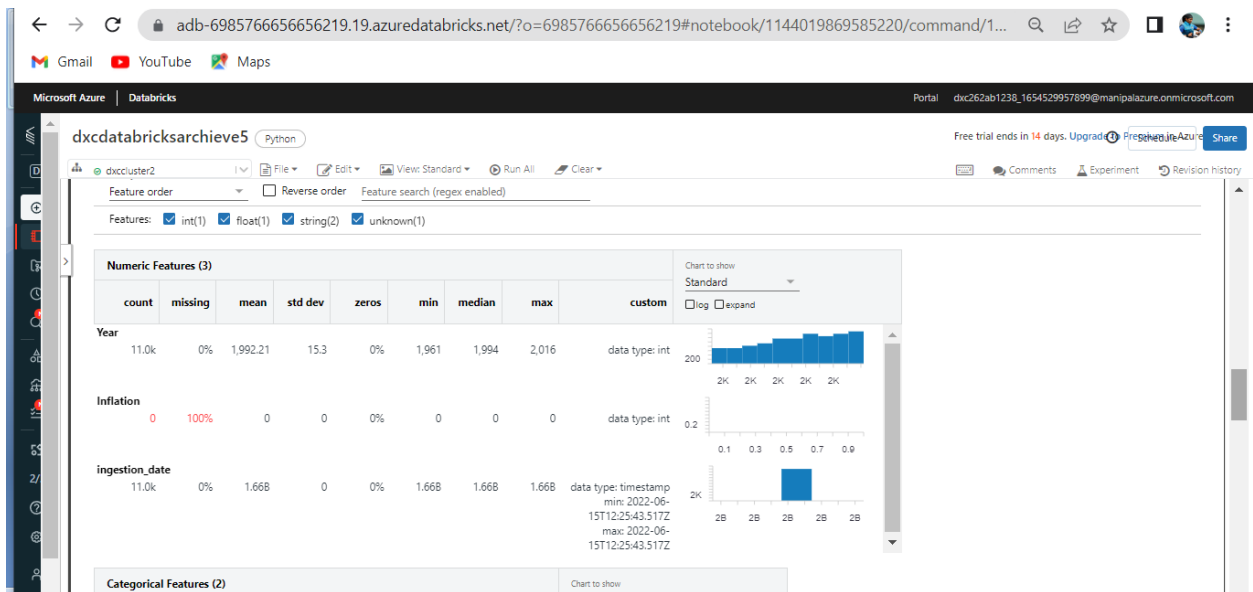
▼ (2) Spark Jobs

Table Data Profile

	Country	CountryCode	Year	Inflation	ingestion_date
1	Arab World	ARB	1969	null	2022-06-15T12:25:43.517+0000
2	Arab World	ARB	1970	null	2022-06-15T12:25:43.517+0000
3	Arab World	ARB	1971	null	2022-06-15T12:25:43.517+0000
4	Arab World	ARB	1972	null	2022-06-15T12:25:43.517+0000
5	Arab World	ARB	1973	null	2022-06-15T12:25:43.517+0000
6	Arab World	ARB	1974	null	2022-06-15T12:25:43.517+0000
7	Arab World	ARB	1975	null	2022-06-15T12:25:43.517+0000

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

Command took 0.41 seconds -- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:56:12 PM on dxccuster2



← → ↻ 🔒 adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585220/command/1... 🔍 📄 ⭐ 🌐

📧 Gmail 📺 YouTube 📍 Maps

Microsoft Azure | Databricks Portal dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com

**dxcdatabricksarchive5** Python Free trial ends in 14 days. [Upgrade](#) [Refresh](#) [Share](#)

🏠 dxccluster2 | 📄 File | ✎ Edit | 🖨️ View: Standard | ⚙️ Run All | 🗑️ Clear

Cell 10

```
1 display(consumer_selected_df)
```

▶ (1) Spark Jobs

Table Data Profile

	Country	CountryCode	Year	Inflation
1	Arab World	ARB	1969	null
2	Arab World	ARB	1970	null
3	Arab World	ARB	1971	null
4	Arab World	ARB	1972	null
5	Arab World	ARB	1973	null
6	Arab World	ARB	1974	null
7	Arab World	ARB	1975	null

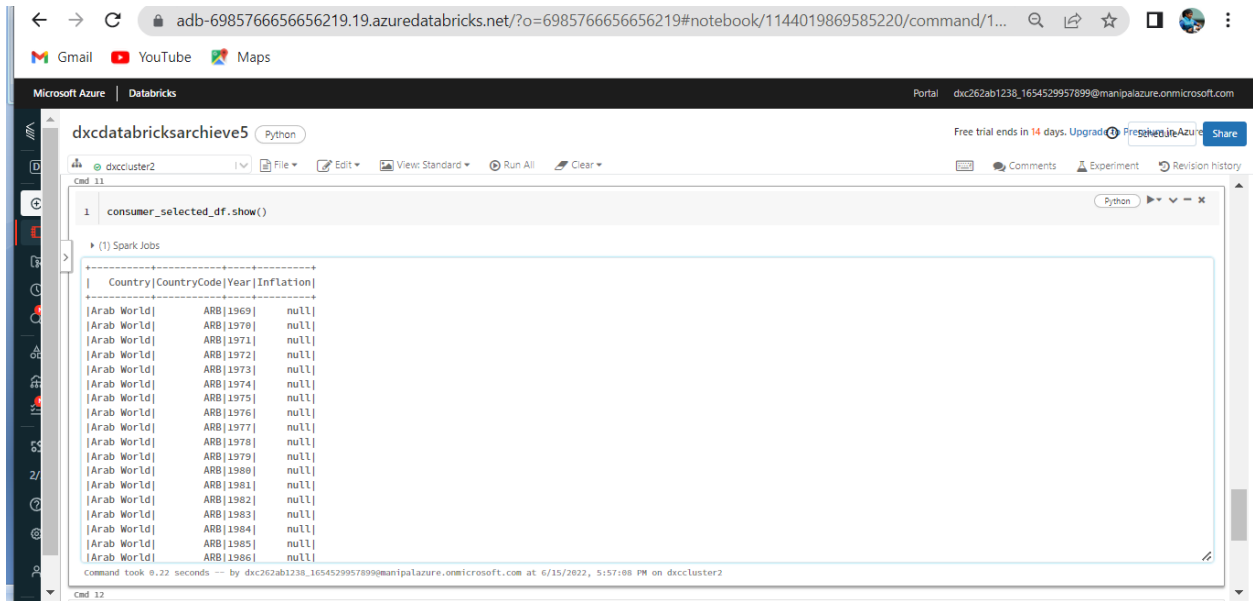
Truncated results, showing first 1000 rows.  
[Click to re-execute with maximum result limits.](#)

📄 📊 📄 📄

Command took 0.25 seconds --- by dxc262ab1238\_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:16:44 PM on dxccluster2

Cell 11

```
1 consumer_selected_df.show()
```



1. Using archive1.zip file - please ingest data into databricks DBFS path & query the data redesign columns accordingly using dataframe commands - display with notebooks accordingly

Command1:

#Ingest Countrycode.csv file

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

Command 2:

```
Countrycode_schema = StructType(fields=[StructField("FIFA", StringType(), False),
StructField("dial", IntegerType(), True),
StructField("MARC", StringType(), True),
StructField("FIPS", StringType(), True),])
```

Command 3:

```
Countrycode_df = spark.read\
.option("header", True) \
.schema(Countrycode_schema) \
.csv("/FileStore/tables/Countrycode.csv")
```

Command 4:

```
#Select only the required columns
```

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

Command 5:

```
Countrycode_selected_df = circuit_df.select(col("FIFA"), col("dial"), col("MARC"), col("FIPS"))
```

Command 6:

```
#Add ingestion date to the dataframe
```

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

Command 7:

```
Countrycode_final_df = Countrycode_df.withColumn("ingestion_date", current_timestamp())
```

Command 8:

```
# Write data to datalake as parquet
```

```
Countrycode_final_df.write.mode("overwrite").parquet("/mnt/formuladl/processed/Countrycode")
```

Command 9:

```
display(spark.read.parquet("/mnt/formuladl/processed/Countrycode"))
```

Command 10:

```
display(Countrycode_selected_df)
```

Command 11:

```
Countrycode_selected_df.show()
```