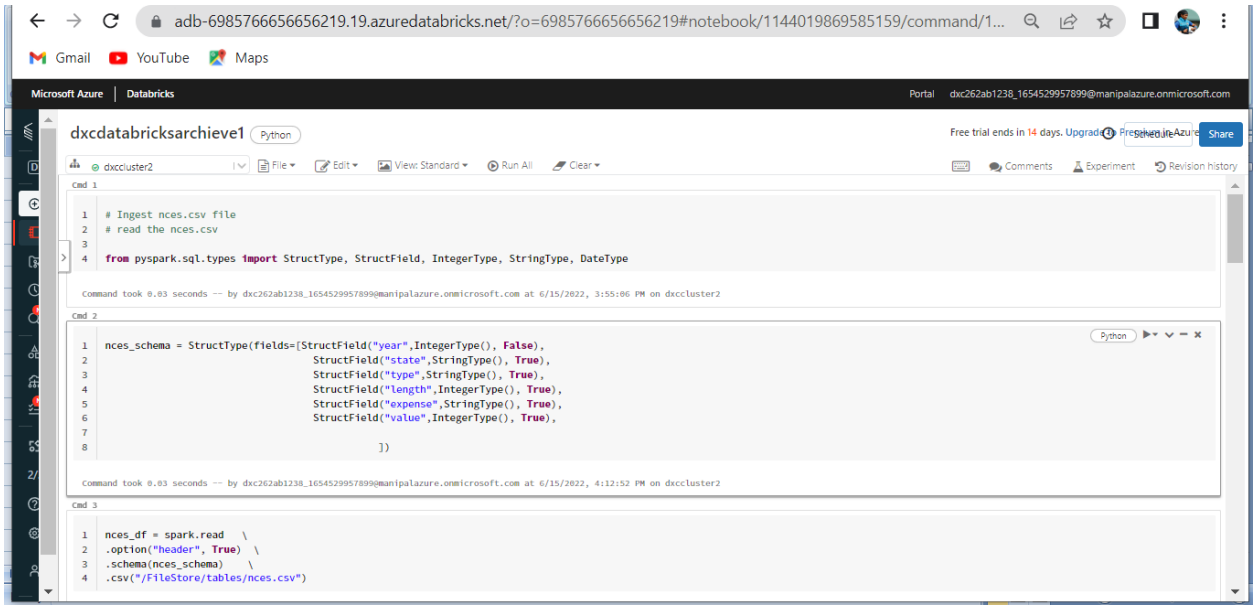
NAME:S.MAHALAKSHMI

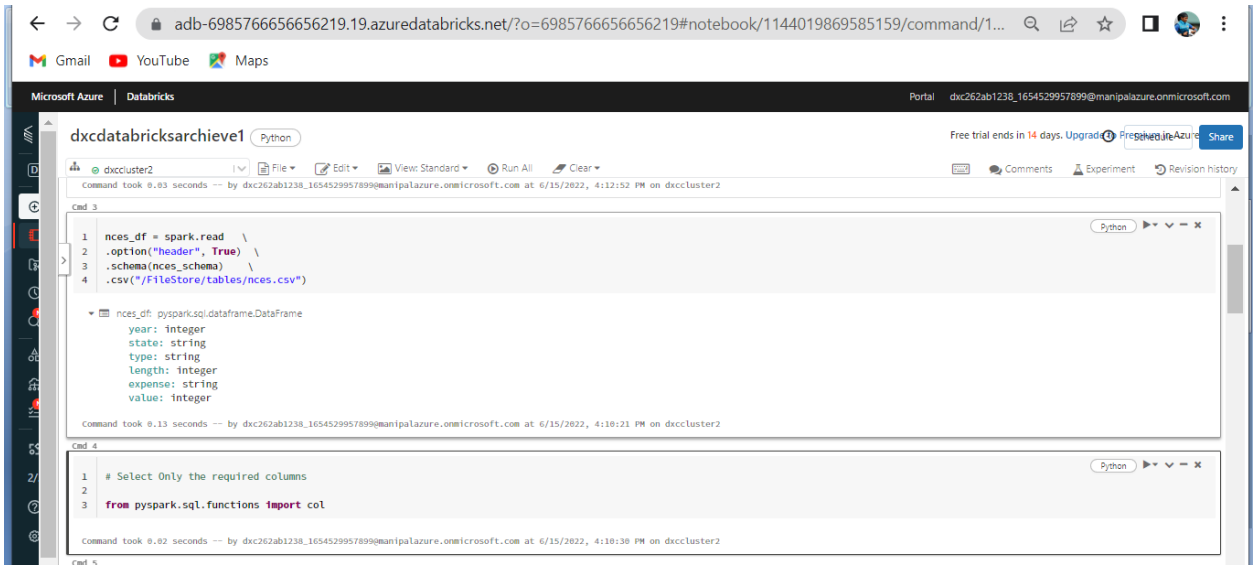
REGNO:DXC262AB1229

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 dafarame commands - display with notebooks accordingly





Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve1 (Python)

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

Schedule Share

dxccluster2 | File | Edit | View: Standard | Run All | Clear | Comments | Experiment | Revision history

Cmd 6

```python
# rename the columns as required

nces_renamed_df = nces_selected_df.withColumnRenamed("year" , "YEAR") \
.withColumnRenamed("state" , "STATE") \
.withColumnRenamed("type" , "TYPE") \
.withColumnRenamed("length" , "LENGTH") \
.withColumnRenamed("expense" , "EXPENSE") \
.withColumnRenamed("value" , "VALUE")
```

nces_renamed_df: pyspark.sql.dataframe.DataFrame

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

Command took 0.04 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:16:14 PM on dxccluster2

Cmd 7

```python
#Add ingestion date to the dataframe

from pyspark.sql.functions import current_timestamp
```

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve1 (Python)

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

Schedule Share

dxccluster2 | File | Edit | View: Standard | Run All | Clear | Comments | Experiment | Revision history

Command took 0.02 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:20:24 PM on dxccluster2

Cmd 8

```python
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 0.03 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:20:48 PM on dxccluster2

Cmd 9

```python
# Write data to datalake as parquet


nces_final_df.write.mode("overwrite").parquet("/mnt/formulaidl/processed/nces")

```

(1) Spark Jobs

Job 32 View (Stages: 1/1)

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve1 (Python)

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

Share

dxccluster2 | File | Edit | View: Standard | Run All | Clear | Comments | Experiment | Revision history

```
VALUE: integer
ingestion_date: timestamp
```

Command took 0.03 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

Cmd 9

```
# Write data to datalake as parquet


nces_final_df.write.mode("overwrite").parquet("/mnt/formulaidl/process
```

(1) Spark Jobs

Job 32 View (Stages: 1/1)

Command took 1.92 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

Cmd 10

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

(2) Spark Jobs

Table | Data Profile

	YEAR	STATE	TYPE	LENGTH	EXP
1	2013	Alabama	Private	null	Fees

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

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 Read
32	2868111045302864447	# Write data to datalake as parquet nces_fina... parquet at NativeMethodAccessorImpl.java:0 +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

dxcdatabricksarchieve1 (Python)

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

Share

dxccluster2 | File | Edit | View: Standard | Run All | Clear | Comments | Experiment | Revision history

Cmd 10

```
display(spark.read.parquet("/mnt/formulaidl/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	2022-06-15T10:52:11.472+0000
2	2013	Alabama	Private	null	Room/Board	8503	2022-06-15T10:52:11.472+0000
3	2013	Alabama	Public In-State	null	Fees/Tuition	4048	2022-06-15T10:52:11.472+0000
4	2013	Alabama	Public In-State	null	Fees/Tuition	8073	2022-06-15T10:52:11.472+0000
5	2013	Alabama	Public In-State	null	Room/Board	8473	2022-06-15T10:52:11.472+0000
6	2013	Alabama	Public Out-of-State	null	Fees/Tuition	7736	2022-06-15T10:52:11.472+0000
7	2013	Alabama	Public Out-of-State	null	Fees/Tuition	20380	2022-06-15T10:52:11.472+0000

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 dxccluster2

Cmd 11

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve1 (Python)

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Command took 0.71 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:22:57 PM on dxccluster2

Cmd 11

```
1 display(nces_selected_df)
```

(1) Spark Jobs

Table Data Profile

	year	state	type	length	expense	value
1	2013	Alabama	Private	null	Fees/Tuition	13963
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 0.28 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:27:27 PM on dxccluster2

Cmd 12

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve1 (Python)

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

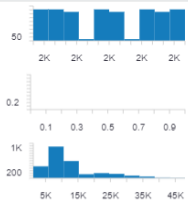
Sort by: Feature order — Reverse order — Feature search (regex enabled)

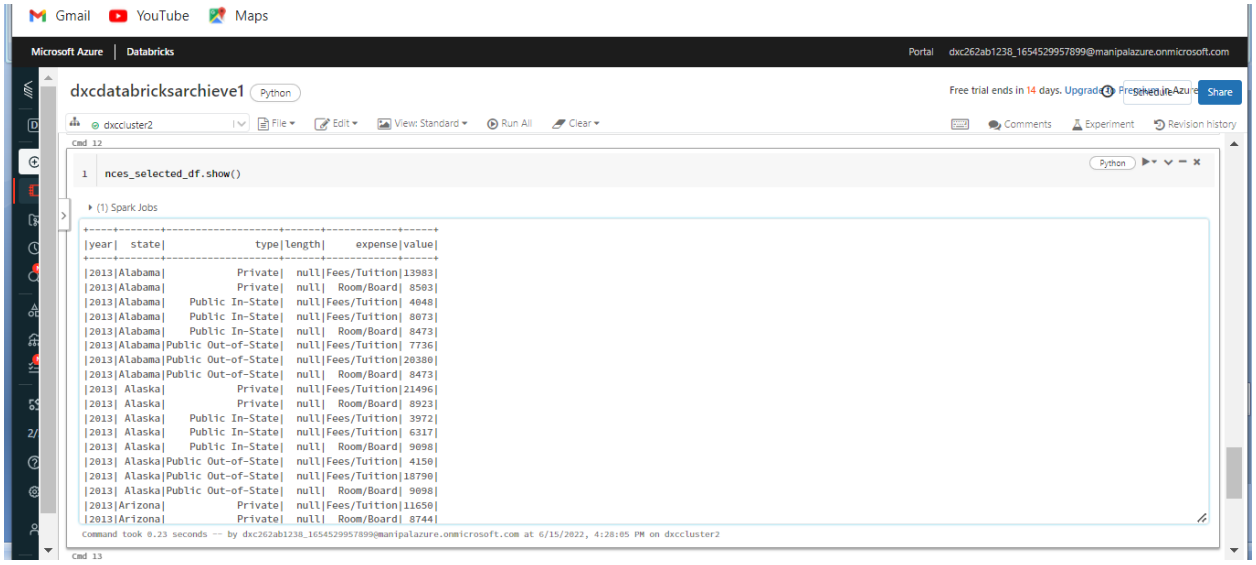
Features: int(2) string(3) unknown(1)

Numeric Features (3)

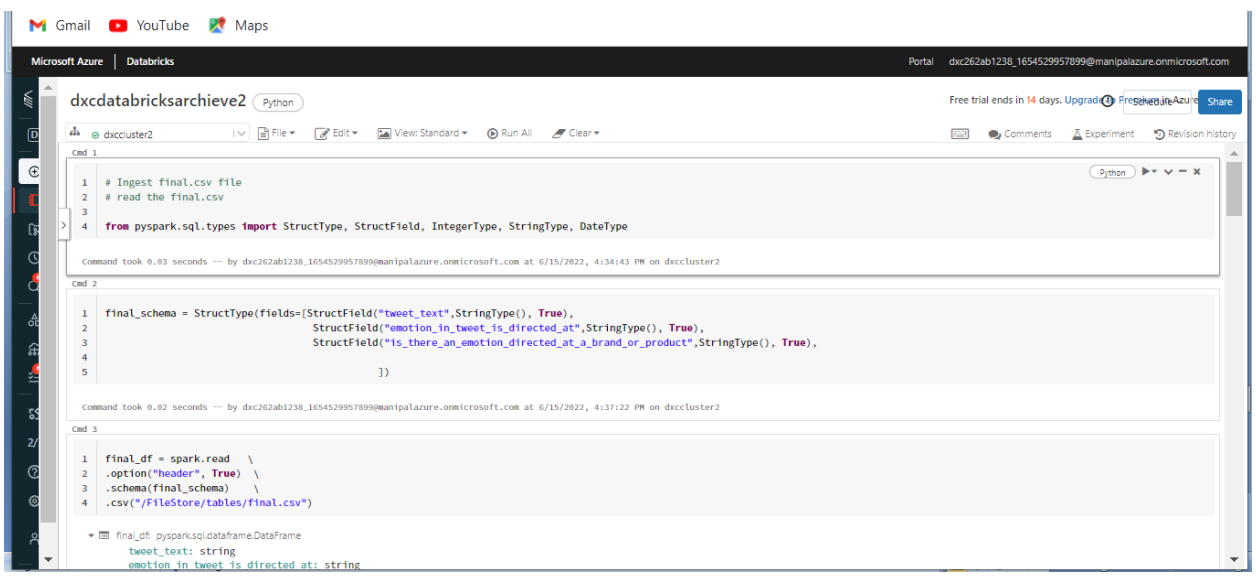
Chart to show: Standard — log expand

	count	missing	mean	std dev	zeros	min	median	max	custom
year	3,203	0%	2,016.48	2.29	0%	2,013	2,016	2,020	data type: int
length	0	100%	0	0	0%	0	0	0	data type: int
value	3,203	0%	12.8k	8,528.45	0%	1,225	10.1k	48.0k	data type: int





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



Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve2 Python

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

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

```
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 dxccluster2

Cmd 7

```
#Add ingestion date to the dataframe

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 dxccluster2

Cmd 8

```
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 dxccluster2

Cmd 9

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve2 Python

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

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 3

```
final_df = spark.read   \
.option("header", True)  \
.schema(final_schema)    \
.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 dxccluster2

Cmd 4

```
# Select Only the required columns

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 dxccluster2

Cmd 5

```
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
```

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve2 (Python)

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

Share

dxccluster2 | File | Edit | View: Standard | Run All | Clear | Comments | Experiment | Revision history

Command took 0.03 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:45:17 PM on dxccluster2

Cmd 9

```
# Write data to datalake as parquet

final_final_df.write.mode("overwrite").parquet("/mnt/formulaidl/proces
```

(1) Spark Jobs

Job 37 View (Stages: 1/1)

Command took 1.32 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

Cmd 10

```
display(spark.read.parquet("/mnt/formulaidl/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 t
2	@jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate Ts at #SXSW
3	@swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.

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

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_9490f26128004f86827cc91faa9903a6
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 KiB	553.2 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

dxcdatabricksarchieve2 (Python)

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

Share

dxccluster2 | File | Edit | View: Standard | Run All | Clear | Comments | Experiment | Revision history

Cmd 10

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

(2) Spark Jobs

Table | Data Profile

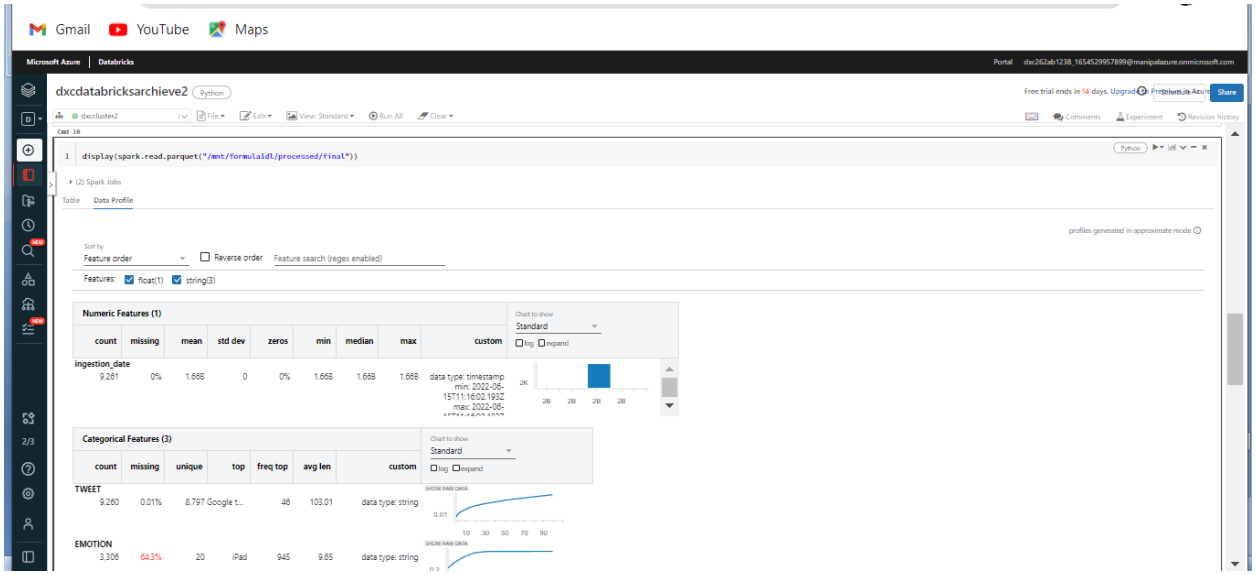
	TWEET	EMOTION	BRAND	ingestion_date
1	.@wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to upgrade. Plugin stations at #SXSW.	iPhone	Negative emotion	2022-06-15T11:16:02.193+00
2	@jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate 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	@sxtxstate 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 http://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
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	2022-06-15T11:16:02.193+00

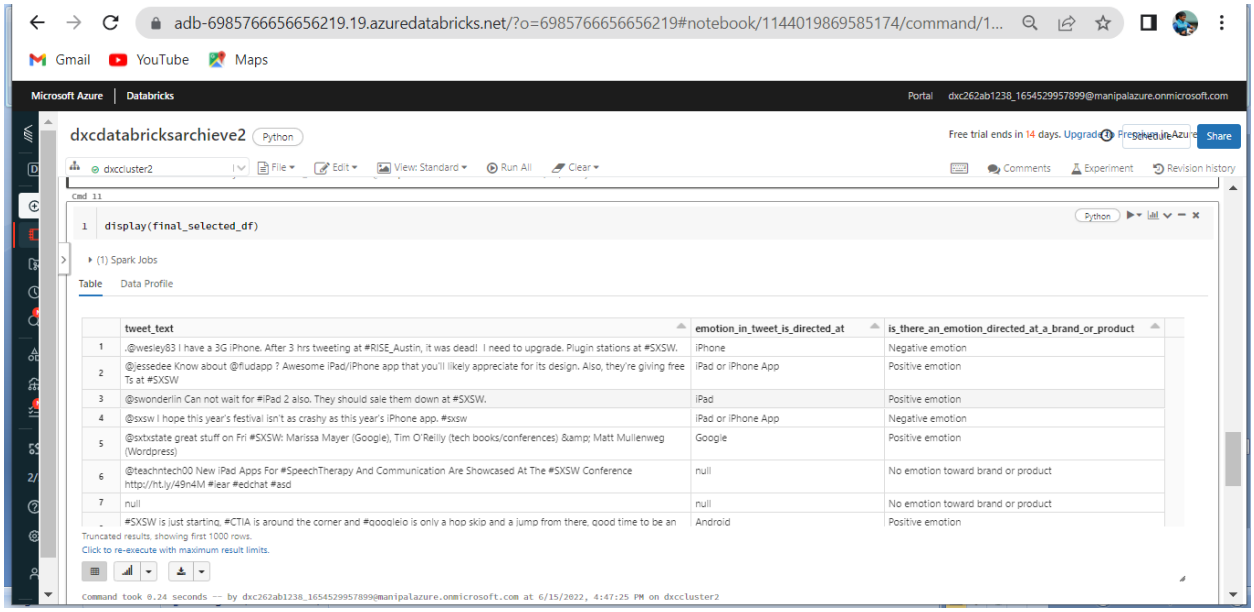
Truncated results, showing first 1000 rows.

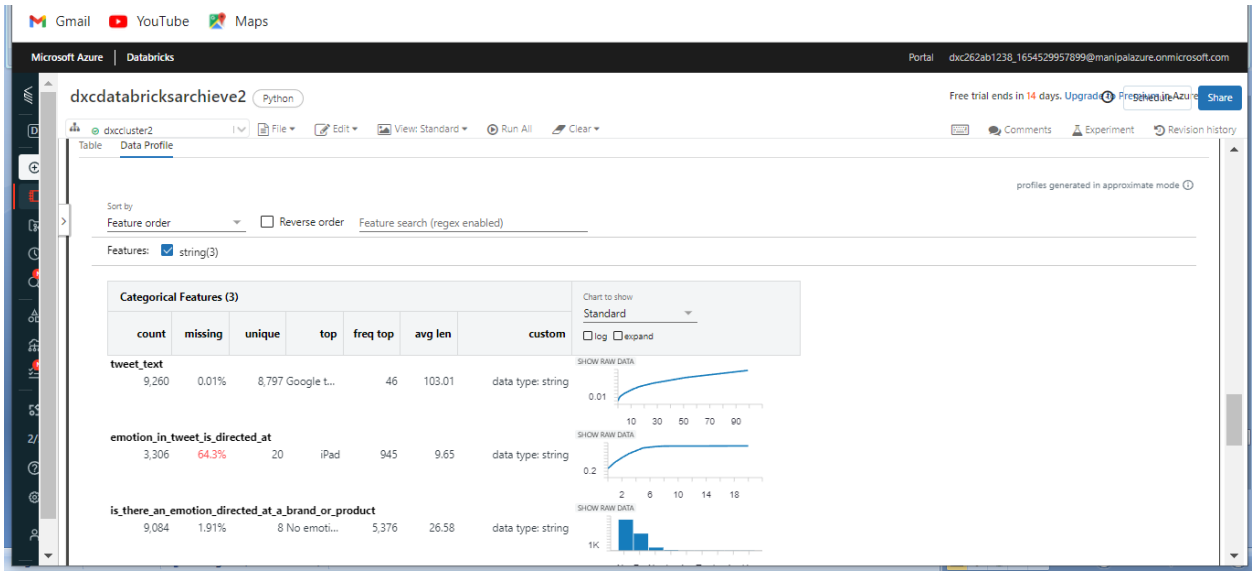
Click to re-execute with maximum result limits.

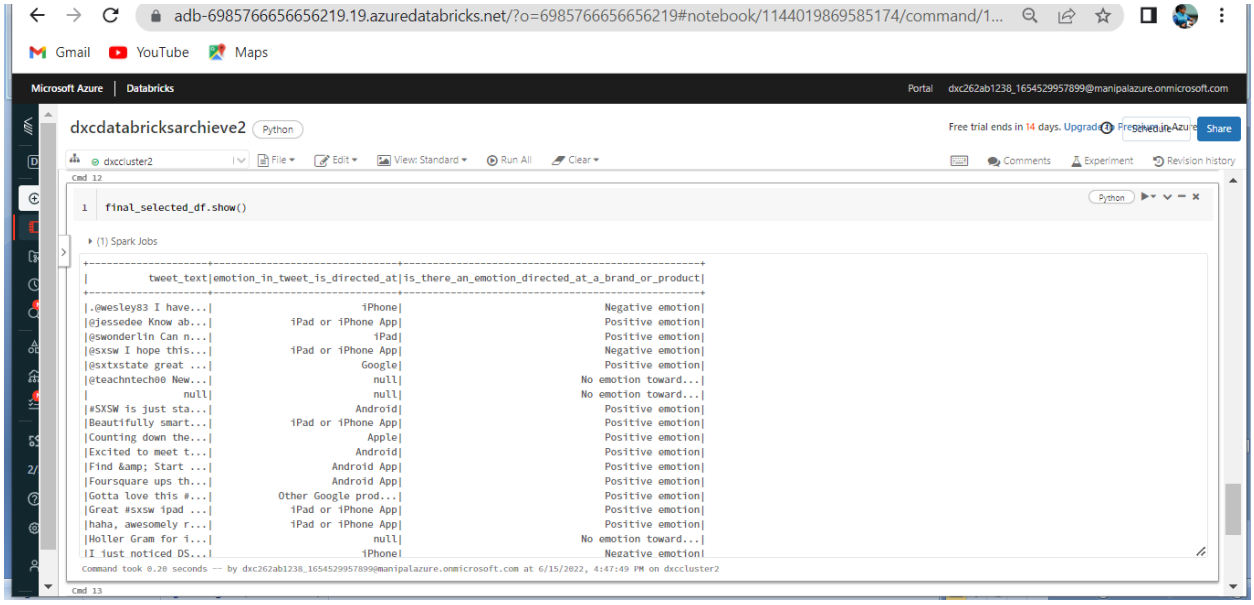
Command took 2.42 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:46:29 PM on dxccluster2

Cmd 11


Gmail
YouTube
Maps
Microsoft Azure | Databricks
dxcdatabricksarchieve2
Python
Share
dxccluster2
File
Edit
View: Standard
Run All
Clear
Comments
Experiment
Revision history
Cmd 10
display(spark.read.parquet("/mnt/formulaldl/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
ingestion_date
9,261 0% 1.66B 0 0% 1.66B 1.66B 1.66B
Categorical Features (3)
count missing unique top freq top avg len custom
TWEET
9,260 0.01% 8,797 Google t... 46 103.01 data type: string
EMOTION
3,306 64.3% 20 iPad 945 9.65 data type: string


adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585174/command/1...
Gmail
YouTube
Maps
Microsoft Azure | Databricks
Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com
dxcdatabricksarchieve2
Python
Free trial ends in 14 days.
Share
dxccluster2
File
Edit
View: Standard
Run All
Clear
Comments
Experiment
Revision history
Cmd 11
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 .@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
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
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 @sxtxstate 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 #iear #edchat #asd null No emotion toward brand or product
7 null null No emotion toward brand or product
#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
Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.
Command took 0.24 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 4:47:25 PM on dxccluster2





4.Using archive4.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/1144019869585191/command/1...

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve3 Python

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 1

```python
# Ingest company.csv file
# read the company.csv

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 dxccluster2

Cmd 2

```python
company_schema = StructType(fields=[StructField("Sno",IntegerType(), True),
                                    StructField("Title",StringType(), True),
                                    StructField("Decision",StringType(), True),
                                    StructField("Words",IntegerType(), True),

                                    ])
```

Command took 0.02 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:04:46 PM on dxccluster2

Cmd 3

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

company_df: pyspark.sql.dataframe.DataFrame
Sno: integer

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve3 Python

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 3

```python
company_df = spark.read   \
.option("header", True)  \
.schema(company_schema)    \
.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 dxccluster2

Cmd 4

```python
# Select Only the required columns

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 dxccluster2

Cmd 5

```python
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...

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve3 Python

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

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 5

```python
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 dxccluster2

Cmd 6

Python

```python
# rename the columns as required

company_renamed_df = company_selected_df.withColumnRenamed("Sno" , "SNO") \
.withColumnRenamed("Title" , "TITLE") \
.withColumnRenamed("Decision" , "DECISION") \
.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 dxccluster2

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve3 Python

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

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

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 dxccluster2

Cmd 7

```python
#Add ingestion date to the dataframe

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 dxccluster2

Cmd 8

Python

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

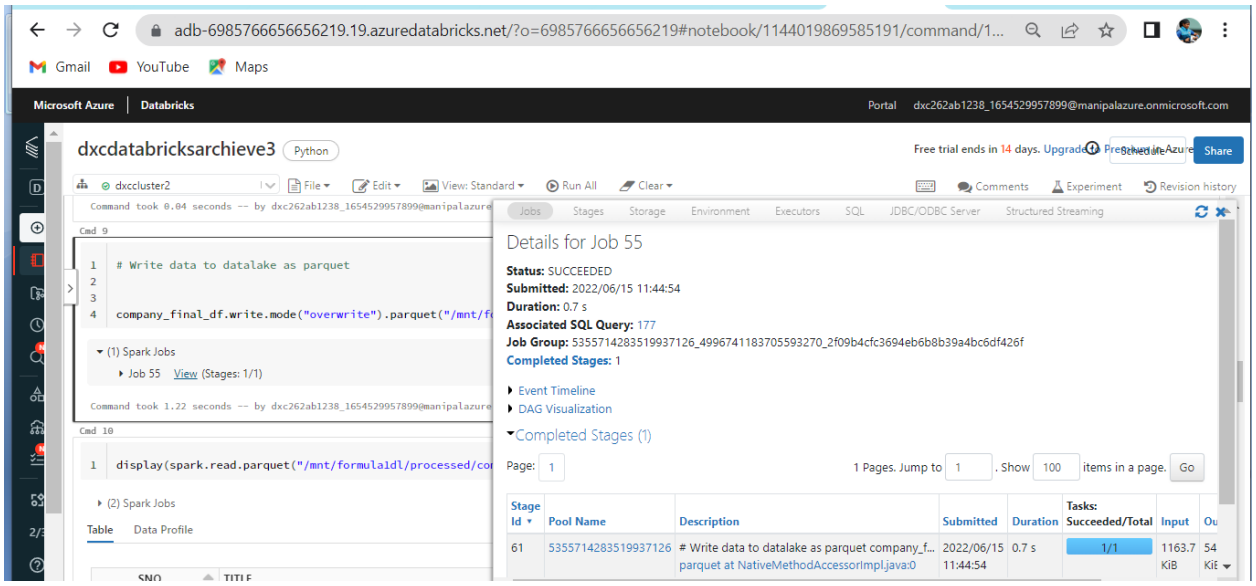
company_final_df: pyspark.sql.dataframe.DataFrame

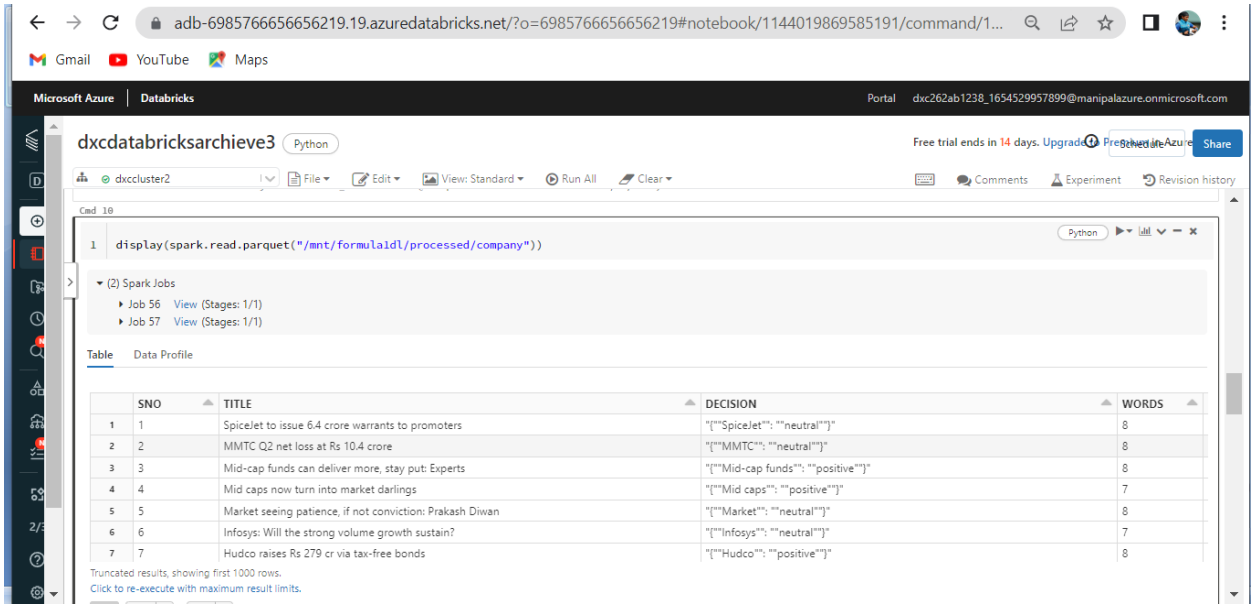
```
SNO: integer
TITLE: string
DECISION: string
WORDS: integer
ingestion_date: timestamp
```

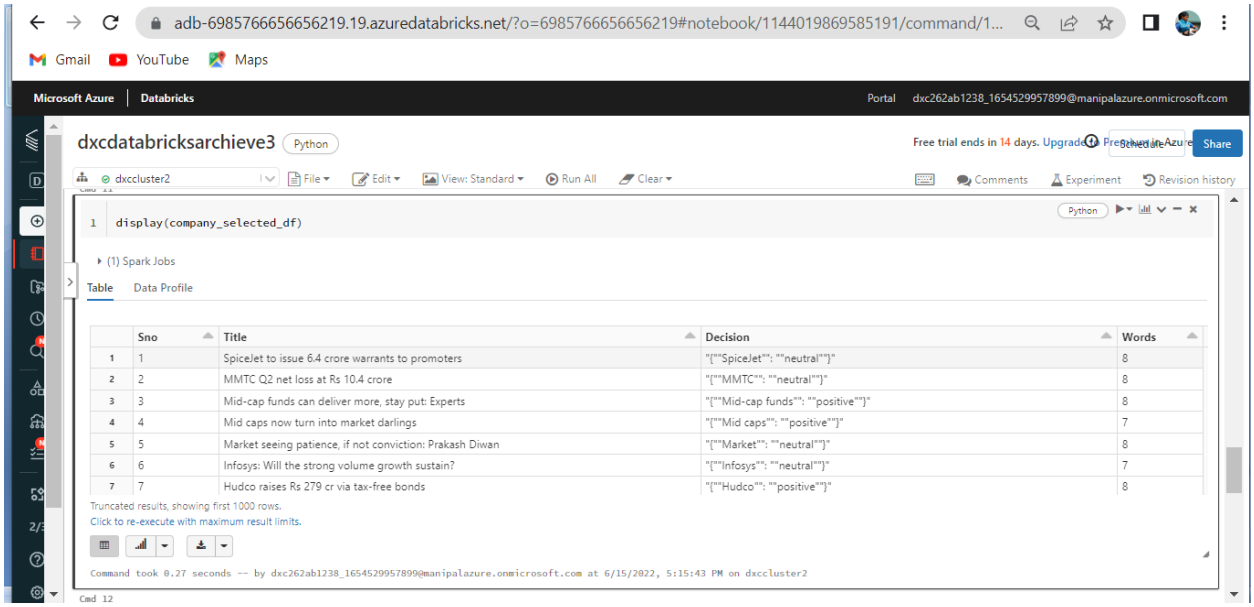
Command took 0.04 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:14:17 PM on dxccluster2

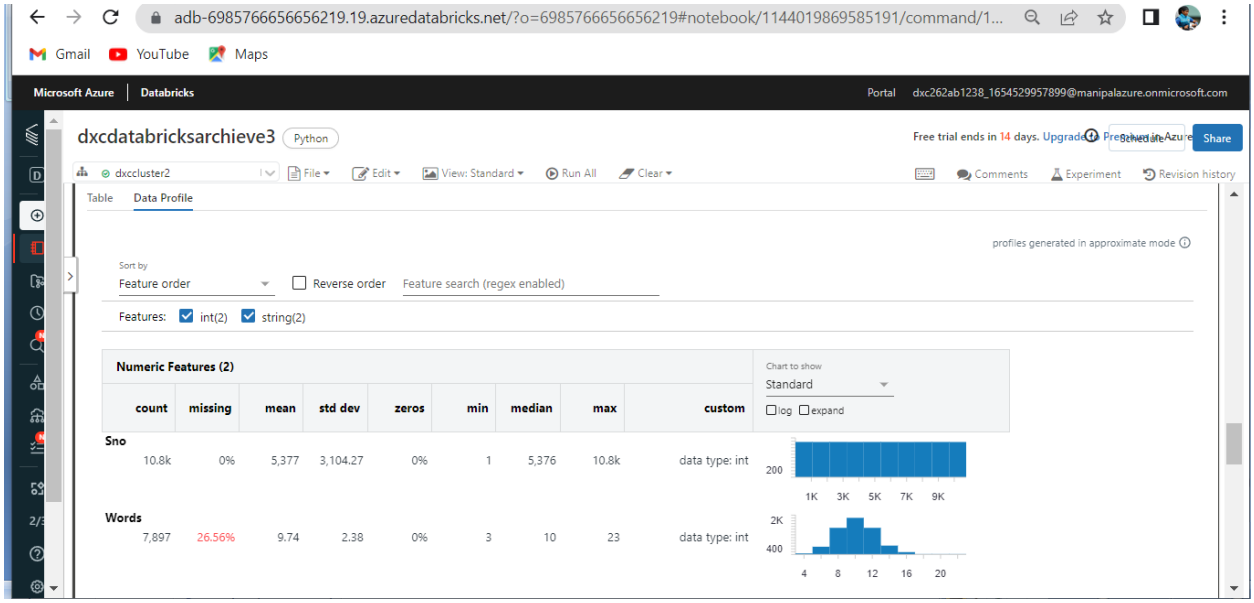
Cmd 9

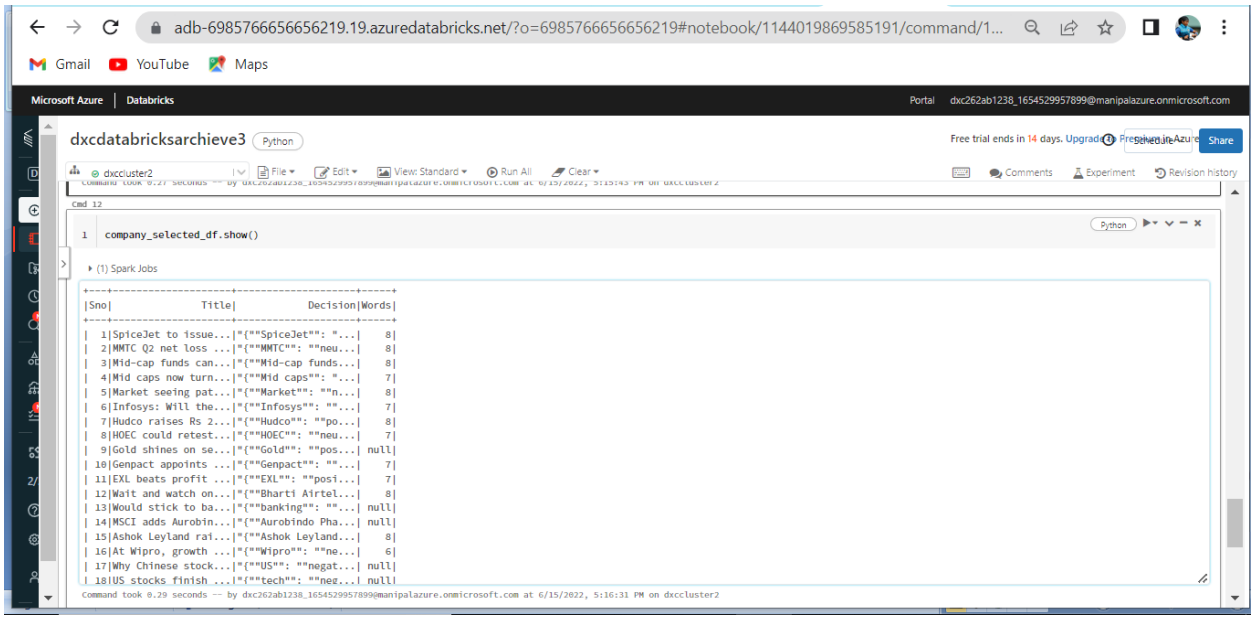
```python
# Write data to datalake as parquet
```


adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...
Gmail
YouTube
Maps
Microsoft Azure
Databricks
Portal
dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com
dxcdatabricksarchieve3
Python
Free trial ends in 14 days. Upgrade
Share
dxccluster2
File
Edit
View: Standard
Run All
Clear
Comments
Experiment
Revision history
Command took 0.04 seconds -- by dxc262ab1238_1654529957899@manipalazure
Cmd 9
Write data to datalake as parquet
company_final_df.write.mode("overwrite").parquet("/mnt/f
(1) Spark Jobs
Job 55 View (Stages: 1/1)
Command took 1.22 seconds -- by dxc262ab1238_1654529957899@manipalazure
Cmd 10
display(spark.read.parquet("/mnt/formula1dl/processed/co
(2) Spark Jobs
Table
Data Profile
SNO
TITLE
Jobs
Stages
Storage
Environment
Executors
SQL
JDBC/ODBC Server
Structured Streaming
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_2f09b4cfc3694eb6b8b39a4bc6df426f
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
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

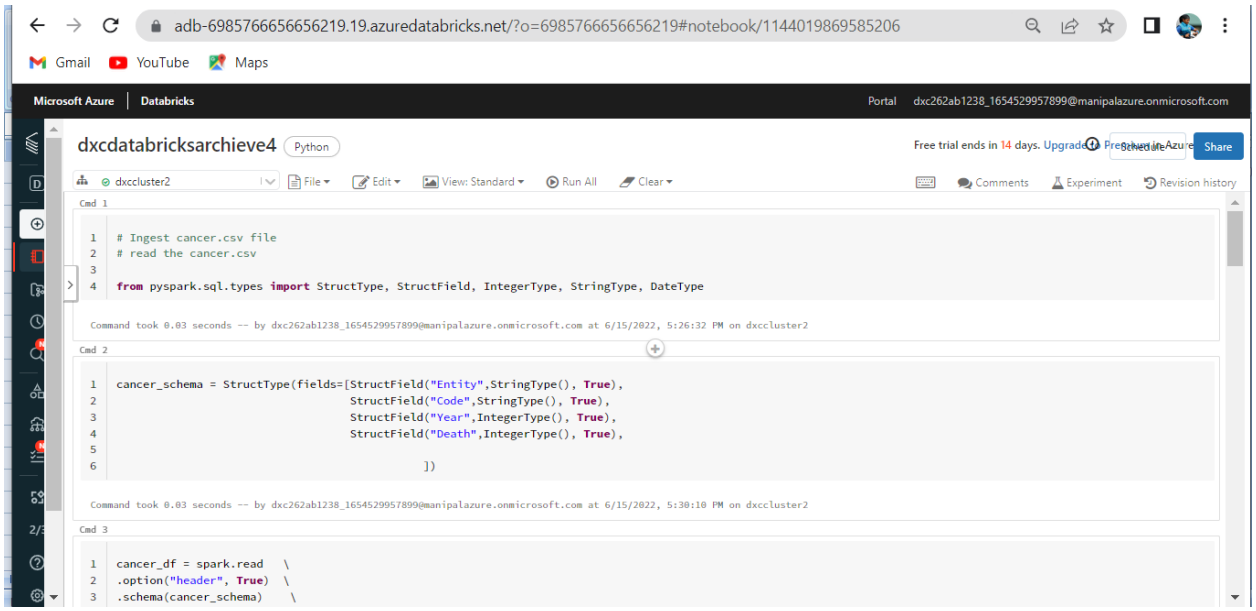

adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...
Gmail
YouTube
Maps
Microsoft Azure
Databricks
Portal
dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com
dxcdatabricksarchieve3
Python
Free trial ends in 14 days. Upgrade
Share
dxccluster2
File
Edit
View: Standard
Run All
Clear
Comments
Experiment
Revision history
Cmd 10
display(spark.read.parquet("/mnt/formula1dl/processed/company"))
Python
(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...
Gmail
YouTube
Maps
Microsoft Azure
Databricks
Portal
dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com
dxcdatabricksarchieve3
Python
Free trial ends in 14 days. Upgrade
Share
dxccluster2
File
Edit
View: Standard
Run All
Clear
Comments
Experiment
Revision history
display(company_selected_df)
(1) Spark Jobs
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.
Command took 0.27 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:15:43 PM on dxccluster2
Cmd 12


adb-6985766656656219.19.azuredatabricks.net/?o=6985766656656219#notebook/1144019869585191/command/1...
Gmail
YouTube
Maps
Microsoft Azure
Databricks
Portal
dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com
dxcdatabricksarchieve3
Python
Free trial ends in 14 days. Upgrade
Share
dxccluster2
File
Edit
View: Standard
Run All
Clear
Comments
Experiment
Revision history
Table
Data Profile
profiles generated in approximate mode
Sort by
Feature order
Reverse order
Feature search (regex enabled)
Features: int(2) string(2)
Numeric Features (2)
Chart to show
Standard
log
expand
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
200
1K 3K 5K 7K 9K
Words
7,897 26.56% 9.74 2.38 0% 3 10 23 data type: int
2K
400
4 8 12 16 20



5.Using archive5.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/1144019869585206/command/1...

Gmail YouTube Maps

Microsoft Azure | Databricks Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve4 (Python) Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 3

```python
cancer_df = spark.read   \
.option("header", True)  \
.schema(cancer_schema)   \
.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 dxccluster2

Cmd 4

```python
# Select Only the required columns

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 dxccluster2

Cmd 5

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

Gmail YouTube Maps

Microsoft Azure | Databricks Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve4 (Python) Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Command took 0.02 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:32:45 PM on dxccluster2

Cmd 5

```python
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 dxccluster2

Cmd 6

```python
#Add ingestion date to the dataframe

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 dxccluster2

Cmd 7

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve4 Python

Free trial ends in 14 days. Upgrade Preserve in Azure Schedule Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 7

```
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 dxccluster2

Cmd 8

```
# Write data to datalake as parquet


cancer_final_df.write.mode("overwrite").parquet("/mnt/formula1dl/processed/company")
```

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

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve4 Python

Free trial ends in 14 days. Upgrade Preserve in Azure Schedule Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 7

```
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

Cmd 8

```
# Write data to datalake as parquet


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

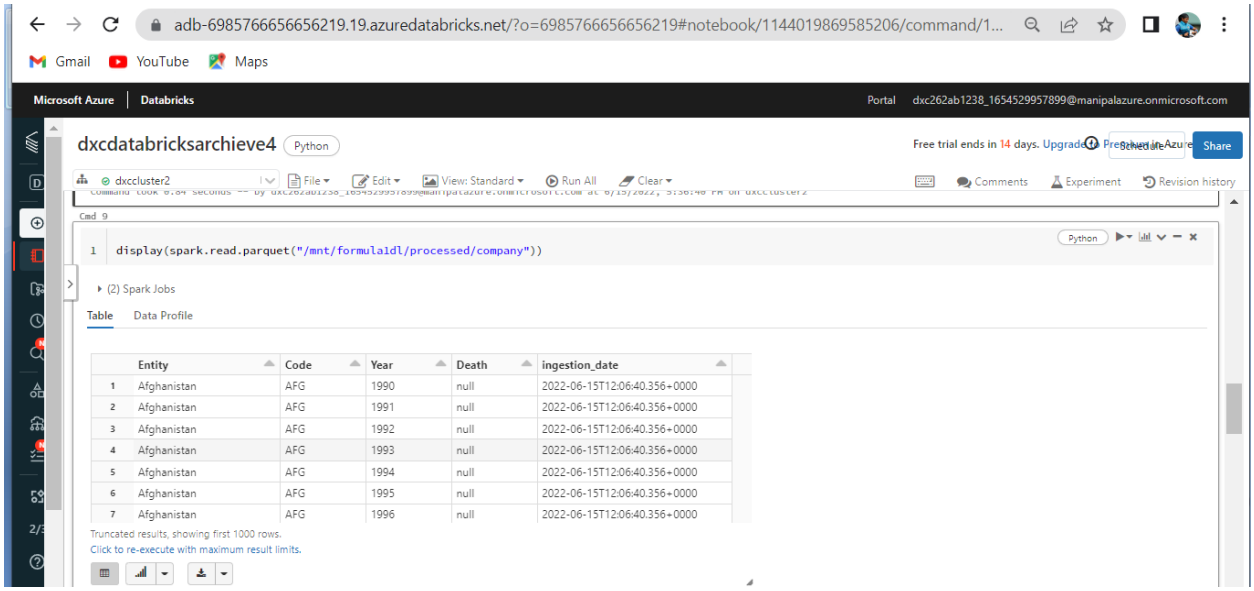
Status: SUCCEEDED
Submitted: 2022/06/15 12:06:40
Duration: 0.4 s
Associated SQL Query: 195
Job Group: 7110440155128268836_5436340371620126902_d40a2cb1033949afb0930b008360950c
Completed Stages: 1

- Event Timeline
- DAG Visualization

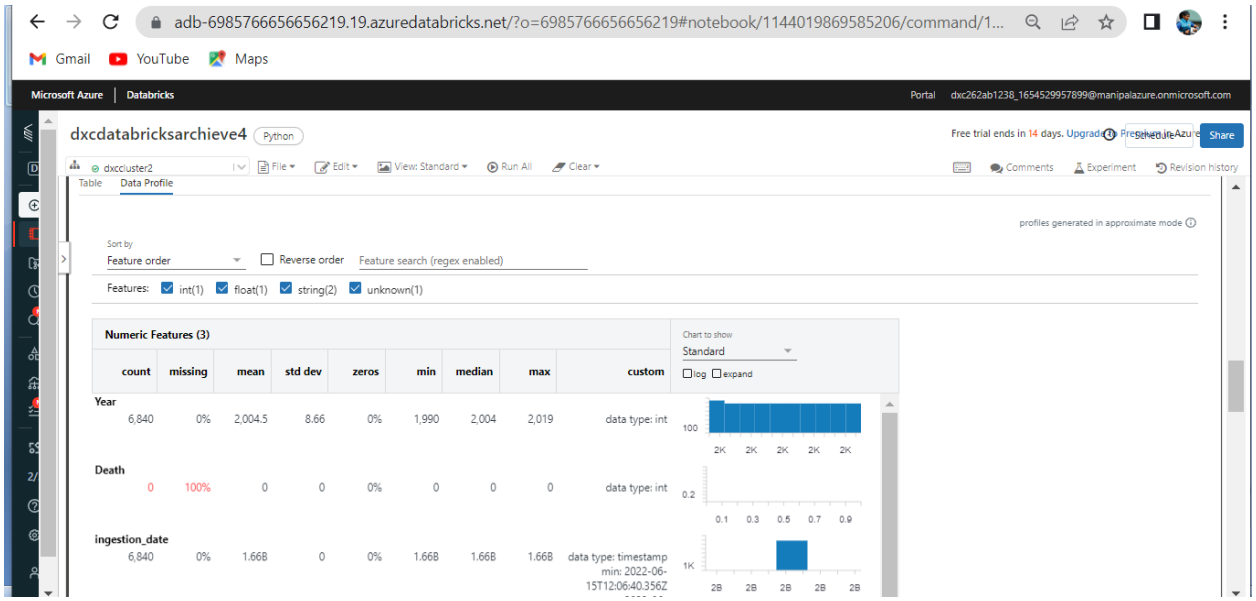
Completed Stages (1)

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

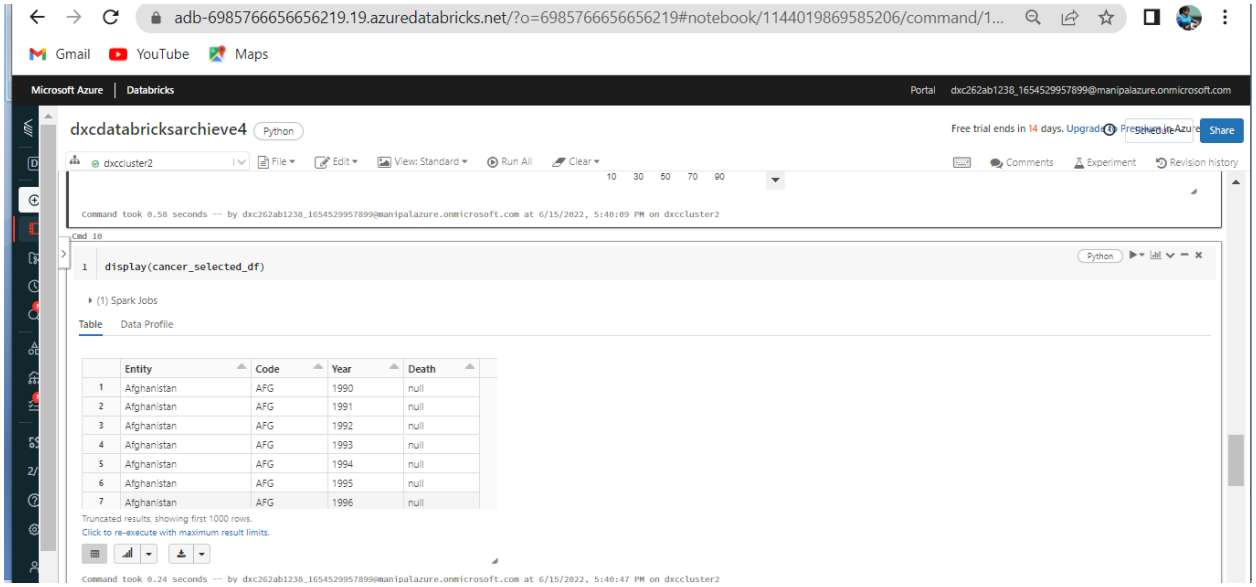
Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Outpu
73	7110440155128268836	# Write data to datalake as parquet cancer_fi... parquet at NativeMethodAccessorImpl.java:0	2022/06/15 12:06:40	0.3 s	1/1	254.5 KiB	6.2 Kil

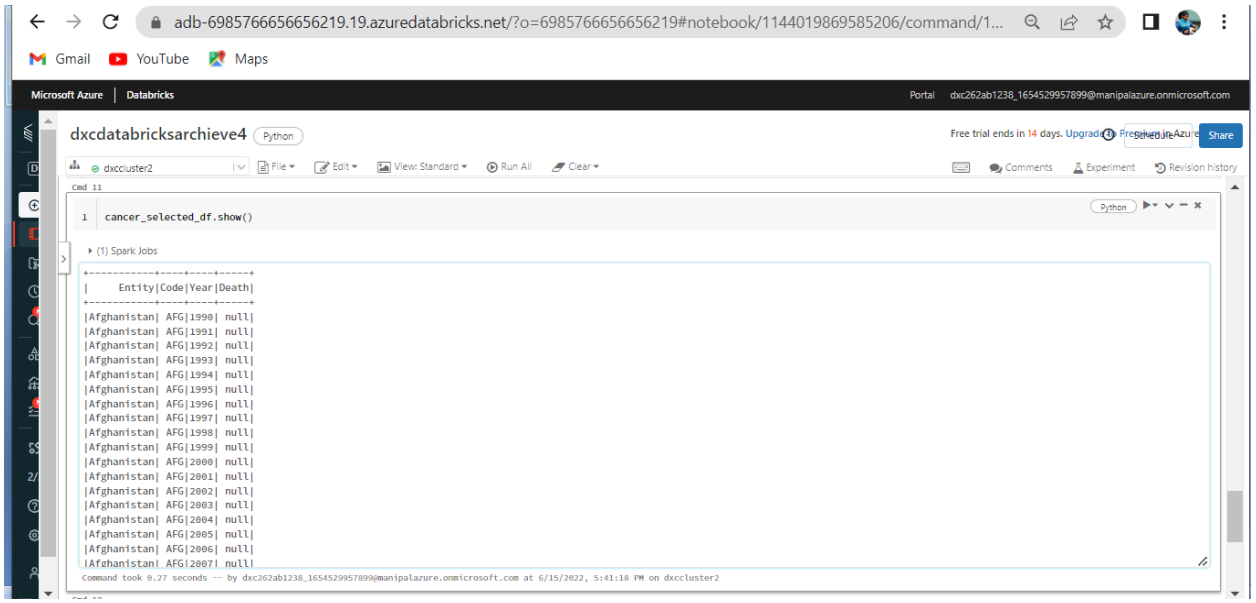


	Entity	Code	Year	Death	ingestion_date
1	Afghanistan	AFG	1990	null	2022-06-15T12:06:40.356+0000
2	Afghanistan	AFG	1991	null	2022-06-15T12:06:40.356+0000
3	Afghanistan	AFG	1992	null	2022-06-15T12:06:40.356+0000
4	Afghanistan	AFG	1993	null	2022-06-15T12:06:40.356+0000
5	Afghanistan	AFG	1994	null	2022-06-15T12:06:40.356+0000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



	count	missing	mean	std dev	zeros	min	median	max	custom
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%	1.66B	0	0%	1.66B	1.66B	1.66B	data type: timestamp min: 2022-06-15T12:06:40.356Z





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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve5 Python

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 1

```python
# Ingest consumer.csv file
# read the consumer.csv

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 dxccluster2

Cmd 2

```python
consumer_schema = StructType(fields=[StructField("Country",StringType(), True),
                                     StructField("CountryCode",StringType(), True),
                                     StructField("Year",IntegerType(), True),
                                     StructField("Inflation",IntegerType(), True),

                                     ])
```

Command took 0.02 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:51:19 PM on dxccluster2

Cmd 3

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

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve5 Python

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

```python
                                     ])
```

Command took 0.02 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:51:19 PM on dxccluster2

Cmd 3

```python
consumer_df = spark.read   \
.option("header", True)  \
.schema(consumer_schema)   \
.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 dxccluster2

Cmd 4

```python
# Select Only the required columns

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 dxccluster2

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve5 Python

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 7

```python
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 dxccluster2

Cmd 8

```python
# Write data to datalake as parquet


consumer_final_df.write.mode("overwrite").parquet("/mnt/formulaidl/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 dxccluster2

Cmd 9

```python
display(spark.read.parquet("/mnt/formulaidl/processed/consumer"))
```

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve5 Python

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

```python
# Select Only the required columns

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 dxccluster2

Cmd 5

```python
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 dxccluster2

Cmd 6

```python
#Add ingestion date to the dataframe

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 dxccluster2

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve5 Python

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

Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 7

```
consumer_final_df = consumer_df.withColumn("ingestion_date", current_t
```

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

Cmd 8

```
# Write data to datalake as parquet

consumer_final_df.write.mode("overwrite").parquet("/mnt/formulaidl/pro
```

(1) Spark Jobs

Job 75 View (Stages: 1/1)

Command took 1.00 second -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

Cmd 9

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

Details for Job 75

Status: SUCCEEDED
Submitted: 2022/06/15 12:25:43
Duration: 0.6 s
Associated SQL Query: 208
Job Group: 8228088374590975492_779631601067454187 5_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 KiB	7.9 KiB	

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

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve5 Python

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

Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Command took 1.00 second -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:55:43 PM on dxccluster2

Cmd 9

```
display(spark.read.parquet("/mnt/formulaidl/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 dxccluster2

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve5 Python

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Feature order Reverse order Feature search (regex enabled)

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

Numeric Features (3)

	count	missing	mean	std dev	zeros	min	median	max	custom
Year	11.0k	0%	1,992.21	15.3	0%	1,961	1,994	2,016	data type: int
Inflation	0	100%	0	0	0%	0	0	0	data type: int
ingestion_date	11.0k	0%	1.66B	0	0%	1.66B	1.66B	1.66B	data type: timestamp min: 2022-06-15T12:25:43.517Z max: 2022-06-15T12:25:43.517Z

Chart to show Standard log expand

Categorical Features (2)

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

Gmail YouTube Maps

Microsoft Azure | Databricks

Portal dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com

dxcdatabricksarchieve5 Python

Free trial ends in 14 days. Upgrade Share

dxccluster2 File Edit View: Standard Run All Clear Comments Experiment Revision history

Cmd 10

```
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

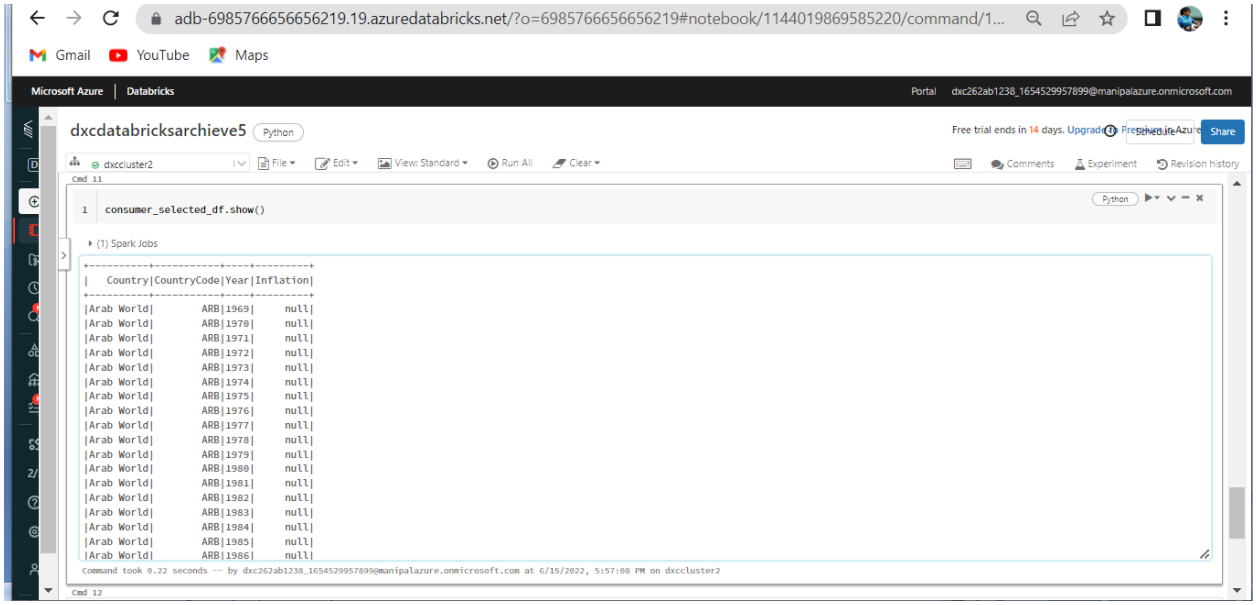
Truncated results, showing first 1000 rows.

Click to re-execute with maximum result limits.

Command took 0.25 seconds -- by dxc262ab1238_1654529957899@manipalazure.onmicrosoft.com at 6/15/2022, 5:56:44 PM on dxccluster2

Cmd 11

```
consumer_selected_df.show()
```



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

Command1:

#Ingest Countrycode.csv file

from pyspark.sql.types import StructType, StructField, Integer Type, 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/formulaldl/processed/Countrycode
")
```

Command 9:

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

Command 10:

```
display(Countrycode_selected_df)
```

Command 11:

```
Countrycode_selected_df.show()
```