databricks Mini_10_PySpark

(https://databricks.com)

```
from pyspark.sql import SparkSession

# Initialize Spark session
spark = SparkSession.builder.appName("AzureBlobStorage").getOrCreate()

# Define storage account information
storage_account_name = "climateb"
storage_account_access_key =
"7C7yeUwkQdDEoz7dWJRFnRTN8CMVzb2LFLbM1I/7S228JYsltNtCIHMKas5nPrcMJXw/4gNizBh++ASt3Q
P8zg=="
container_name = "mini10climatebigdata"
file_name = "ghcnd_daily.csv"

# Set up configuration
spark.conf.set(
    f"fs.azure.account.key.{storage_account_name}.blob.core.windows.net",
    storage_account_access_key
)

# Read dataset
df =
spark.read.csv(f"wasbs://{container_name}@{storage_account_name}.blob.core.windows.
net/{file_name}", header=True, inferSchema=True)

# Show the DataFrame
df.show()
```

```
+-----------+----+-----+-------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------+--
----+------+------+------+------+------+------+------+------+------+------+-----
-+------+------+------+------+------+------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+
|         id|year|month|element|value1|mflag1|qflag1|sflag1|value2|mflag2|qflag2
|sflag2|value3|mflag3|qflag3|sflag3|value4|mflag4|qflag4|sflag4|value5|mflag5|qf
lag5|sflag5|value6|mflag6|qflag6|sflag6|value7|mflag7|qflag7|sflag7|value8|mflag
8|qflag8|sflag8|value9|mflag9|qflag9|sflag9|value10|mflag10|qflag10|sflag10|valu
e11|mflag11|qflag11|sflag11|value12|mflag12|qflag12|sflag12|value13|mflag13|qfla
g13|sflag13|value14|mflag14|qflag14|sflag14|value15|mflag15|qflag15|sflag15|valu
e16|mflag16|qflag16|sflag16|value17|mflag17|qflag17|sflag17|value18|mflag18|qfla
g18|sflag18|value19|mflag19|qflag19|sflag19|value20|mflag20|qflag20|sflag20|valu
```

```
df.describe()
```

```
DataFrame[summary: string, id: string, year: string, month: string, element: stri
ng, value1: string, mflag1: string, qflag1: string, sflag1: string, value2: strin
g, mflag2: string, qflag2: string, sflag2: string, value3: string, mflag3: string
, qflag3: string, sflag3: string, value4: string, mflag4: string, qflag4: string,
sflag4: string, value5: string, mflag5: string, qflag5: string, sflag5: string, v
alue6: string, mflag6: string, qflag6: string, sflag6: string, value7: string, mf
lag7: string, qflag7: string, sflag7: string, value8: string, mflag8: string, qfl
ag8: string, sflag8: string, value9: string, mflag9: string, qflag9: string, sfla
g9: string, value10: string, mflag10: string, qflag10: string, sflag10: string, v
alue11: string, mflag11: string, qflag11: string, sflag11: string, value12: strin
g, mflag12: string, qflag12: string, sflag12: string, value13: string, mflag13: s
tring, qflag13: string, sflag13: string, value14: string, mflag14: string, qflag1
4: string, sflag14: string, value15: string, mflag15: string, qflag15: string, sf
lag15: string, value16: string, mflag16: string, qflag16: string, sflag16: string
, value17: string, mflag17: string, qflag17: string, sflag17: string, value18: st
ring, mflag18: string, qflag18: string, sflag18: string, value19: string, mflag19
: string, qflag19: string, sflag19: string, value20: string, mflag20: string, qfl
ag20: string, sflag20: string, value21: string, mflag21: string, qflag21: string,
sflag21: string, value22: string, mflag22: string, qflag22: string, sflag22: stri
ng, value23: string, mflag23: string, qflag23: string, sflag23: string, value24:
string, mflag24: string, qflag24: string, sflag24: string, value25: string, mflag
```

```
# create a new temporary df
df.createOrReplaceTempView("my_temp_table")


# select all records where a certain column, say mflag1, is greater than a value1
result_df = spark.sql("SELECT * FROM my_temp_table WHERE mflag1 > value1")


# add a new column that is the result of some operation on an existing column
transformed_df = df.withColumn("year", df["month"] * 2)


result_df.show()  # This will display the result of SQL query
transformed_df.show()  # This will display the DataFrame after transformation
```

```
+---+----+-----+-------+------+------+------+------+------+------+------+------+
------+------+------+------+------+------+------+------+------+------+------+---
---+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+
| id|year|month|element|value1|mflag1|qflag1|sflag1|value2|mflag2|qflag2|sflag2|
value3|mflag3|qflag3|sflag3|value4|mflag4|qflag4|sflag4|value5|mflag5|qflag5|sfl
ag5|value6|mflag6|qflag6|sflag6|value7|mflag7|qflag7|sflag7|value8|mflag8|qflag8
|sflag8|value9|mflag9|qflag9|sflag9|value10|mflag10|qflag10|sflag10|value11|mfla
g11|qflag11|sflag11|value12|mflag12|qflag12|sflag12|value13|mflag13|qflag13|sfla
g13|value14|mflag14|qflag14|sflag14|value15|mflag15|qflag15|sflag15|value16|mfla
g16|qflag16|sflag16|value17|mflag17|qflag17|sflag17|value18|mflag18|qflag18|sfla
g18|value19|mflag19|qflag19|sflag19|value20|mflag20|qflag20|sflag20|value21|mfla
```

```
df.show(n=20)  # Shows the first 20 rows
```

```
+-----------+----+-----+-------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------+--
----+------+------+------+------+------+------+------+------+------+------+-----
-+------+------+------+------+------+------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+----
---+-------+-------+-------+
|         id|year|month|element|value1|mflag1|qflag1|sflag1|value2|mflag2|qflag2
|sflag2|value3|mflag3|qflag3|sflag3|value4|mflag4|qflag4|sflag4|value5|mflag5|qf
lag5|sflag5|value6|mflag6|qflag6|sflag6|value7|mflag7|qflag7|sflag7|value8|mflag
8|qflag8|sflag8|value9|mflag9|qflag9|sflag9|value10|mflag10|qflag10|sflag10|valu
e11|mflag11|qflag11|sflag11|value12|mflag12|qflag12|sflag12|value13|mflag13|qfla
g13|sflag13|value14|mflag14|qflag14|sflag14|value15|mflag15|qflag15|sflag15|valu
e16|mflag16|qflag16|sflag16|value17|mflag17|qflag17|sflag17|value18|mflag18|qfla
g18|sflag18|value19|mflag19|qflag19|sflag19|value20|mflag20|qflag20|sflag20|valu
```

```
Row(id='AE000041196', year=1966, month=7, element='TMAX', value1=-9999, mflag1=N
one, qflag1=None, sflag1=None, value2=-9999, mflag2=None, qflag2=None, sflag2=No
ne, value3=-9999, mflag3=None, qflag3=None, sflag3=None, value4=411, mflag4=None
, qflag4=None, sflag4='I', value5=-9999, mflag5=None, qflag5=None, sflag5=None,
value6=-9999, mflag6=None, qflag6=None, sflag6=None, value7=-9999, mflag7=None,
qflag7=None, sflag7=None, value8=-9999, mflag8=None, qflag8=None, sflag8=None, v
alue9=372, mflag9=None, qflag9=None, sflag9='I', value10=-9999, mflag10=None, qf
lag10=None, sflag10=None, value11=-9999, mflag11=None, qflag11=None, sflag11=Non
e, value12=-9999, mflag12=None, qflag12=None, sflag12=None, value13=-9999, mflag
13=None, qflag13=None, sflag13=None, value14=-9999, mflag14=None, qflag14=None,
sflag14=None, value15=422, mflag15=None, qflag15=None, sflag15='I', value16=-999
9, mflag16=None, qflag16=None, sflag16=None, value17=-9999, mflag17=None, qflag1
7=None, sflag17=None, value18=-9999, mflag18=None, qflag18=None, sflag18=None, v
alue19=-9999, mflag19=None, qflag19=None, sflag19=None, value20=-9999, mflag20=N
one, qflag20=None, sflag20=None, value21=-9999, mflag21=None, qflag21=None, sfla
g21=None, value22=-9999, mflag22=None, qflag22=None, sflag22=None, value23=-9999
, mflag23=None, qflag23=None, sflag23=None, value24=-9999, mflag24=None, qflag24
=None, sflag24=None, value25=-9999, mflag25=None, qflag25=None, sflag25=None, va
lue26=-9999, mflag26=None, qflag26=None, sflag26=None, value27=-9999, mflag27=No
ne, qflag27=None, sflag27=None, value28=372, mflag28=None, qflag28=None, sflag28
='I', value29=-9999, mflag29=None, qflag29=None, sflag29=None, value30=-9999, mf
```