PYSPARK CONNECTION WITH S3 BUCKET

Q . Read file 'test.csv' from s3 bucket "sample1", select the records having age > 45 and gender "male". The records is outdated by 5 years, So update the age of employee by adding 5 years. Update the value of gender "Male" with "M". Save the csv file in s3 bucket "sample2".

Create spark session

```
from pyspark.sql import SparkSession
spark=SparkSession.builder.\
        config('spark.master','local').\
        config('spark.app.name','S3app').\
        config('spark.jars.packages','org.apache.hadoop:hadoop-aws:3.3.3,org.apache.hadoop:hadoop-
          common:3.3.3').\
        getOrCreate()
```

```
spark
```

Configure aws connection with access key and secret key

```
spark.sparkContext._jsc.hadoopConfiguration().set('fs.s3a.access.key','AHPNEOVAWENBCPOEV')
spark.sparkContext._jsc.hadoopConfiguration().set('fs.s3a.secret.key','jfde/apenbcutkshggndl')
spark.sparkContext._jsc.hadoopConfiguration().set('fs.s3a.endpoint','s3.amazonaws.com')
```

#note: here you need to use your own access key and secret key.

Read file from s3 bucket "zaki80"

```
df=spark.read.format('csv').load('s3a://sample1/test.csv',header=True,inferSchema=True)
df.show()
```

Filter age and gender.

```
df1=df.filter((col(" age")>45) & (col(" gender")=="Male"))
```

Update the age by 5 years

```
df2=df1.withColumn('updated_age', df1[' age']+5)
```

Update the gender value "Male" to "M"

➔ df3=df2.withColumn(" gender", when(col(" gender")=="Male", "M").otherwise(col(" gender")))

Save the file in S3 bucket "sample2"

➔ output_path="s3a://sample2/test"

```
df3.write \
.format("csv") \
.option("header", "True") \
.save(output_path)
```

➔ df3.show()

#NOTE: We can get hadoop-aws & hadoop-common from maven repositories . Here, I tried with older version(3.2) of hadoop-aws and hadoop-common, but didn't work. So, I tried with 3.3.3 version and it worked.