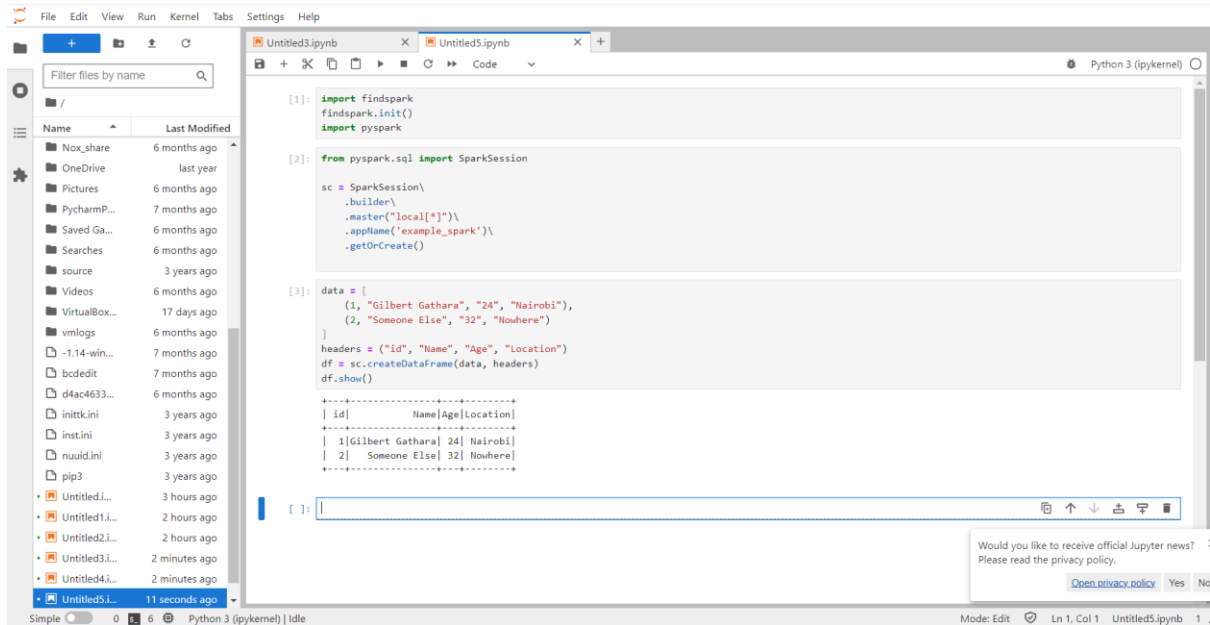


# ASSESSMENT

NAME-SWARNA LATHA GUVVALA

DATE-27/12/2023



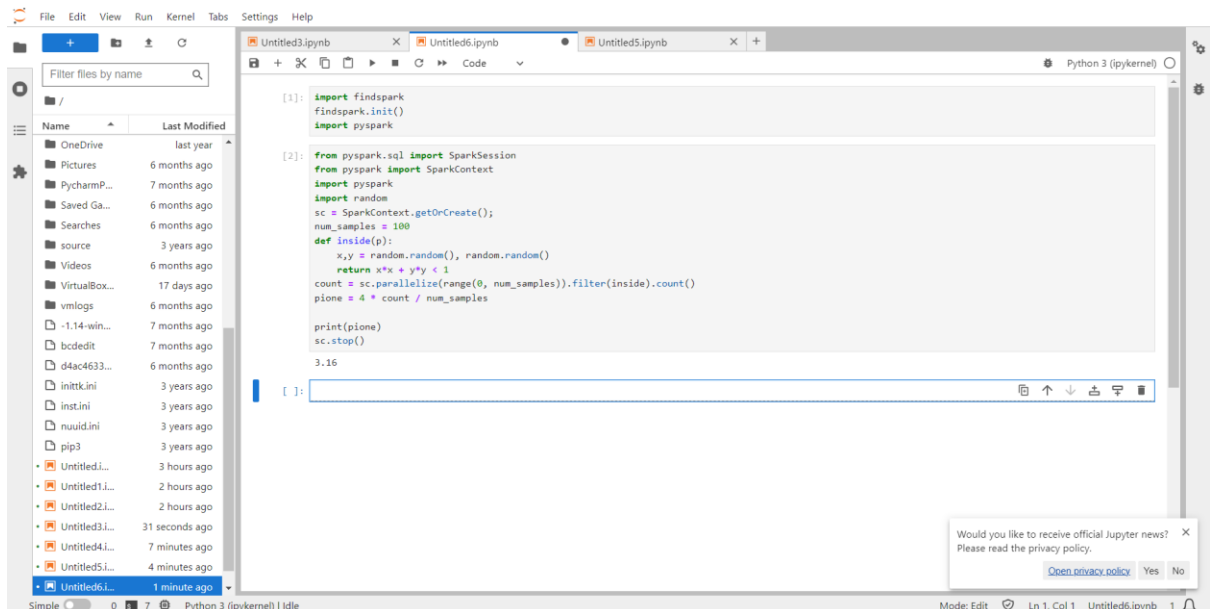
```
[1]: import findspark
findspark.init()
import pyspark

[2]: from pyspark.sql import SparkSession

sc = SparkSession\
    .builder\
    .master("local[*]")\
    .appName('example_spark')\
    .getOrCreate()

[3]: data = [
    (1, "Gilbert Gathara", "24", "Nairobi"),
    (2, "Someone Else", "32", "Nowhere")
]
headers = ("id", "Name", "Age", "Location")
df = sc.createDataFrame(data, headers)
df.show()
```

id	Name	Age	Location
1	Gilbert Gathara	24	Nairobi
2	Someone Else	32	Nowhere



```
[1]: import findspark
findspark.init()
import pyspark

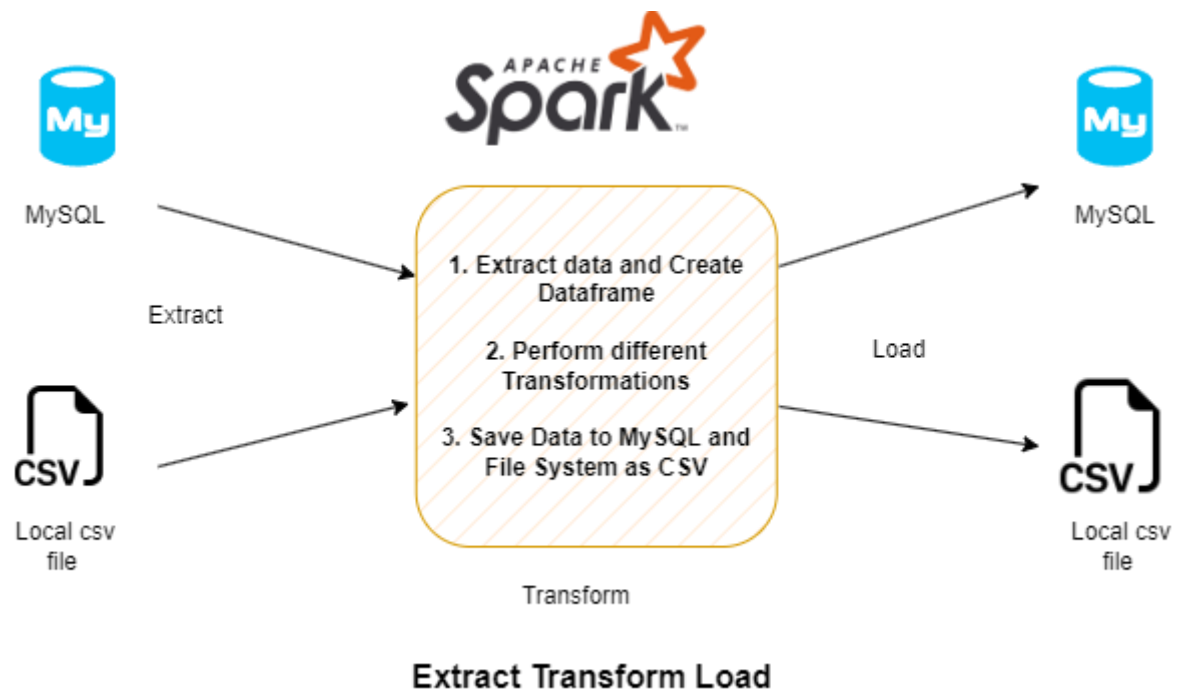
[2]: from pyspark.sql import SparkContext
from pyspark import SparkContext
import random

sc = SparkContext.getOrCreate();
num_samples = 100
def inside(p):
    x,y = random.random(), random.random()
    return x*x + y*y < 1
count = sc.parallelize(range(0, num_samples)).filter(inside).count()
pione = 4 * count / num_samples

print(pione)
sc.stop()

3.16
```

## ETL WITH PYSPARK:



### The PySpark ETL Workflow:

- ◇ Extract: Retrieve data from various sources like databases, files, or APIs.
- ◇ Transform: Clean, aggregate, and manipulate data to fit your analysis needs.
- ◇ Load: Store the transformed data into a database or data warehouse for analysis.