Mallu Sravana Sandhya Assessment-19 26/12/23

**Jupyterlab**

In jupyterlab web page we can execute the spark commands

To do this practice first we have to open command prompt and install jupyterlab by hitting the below comand

Python -m pip install jupyterlab

Then after successful installation of jupyterlab we have hit the other command to open jupyterlab as below

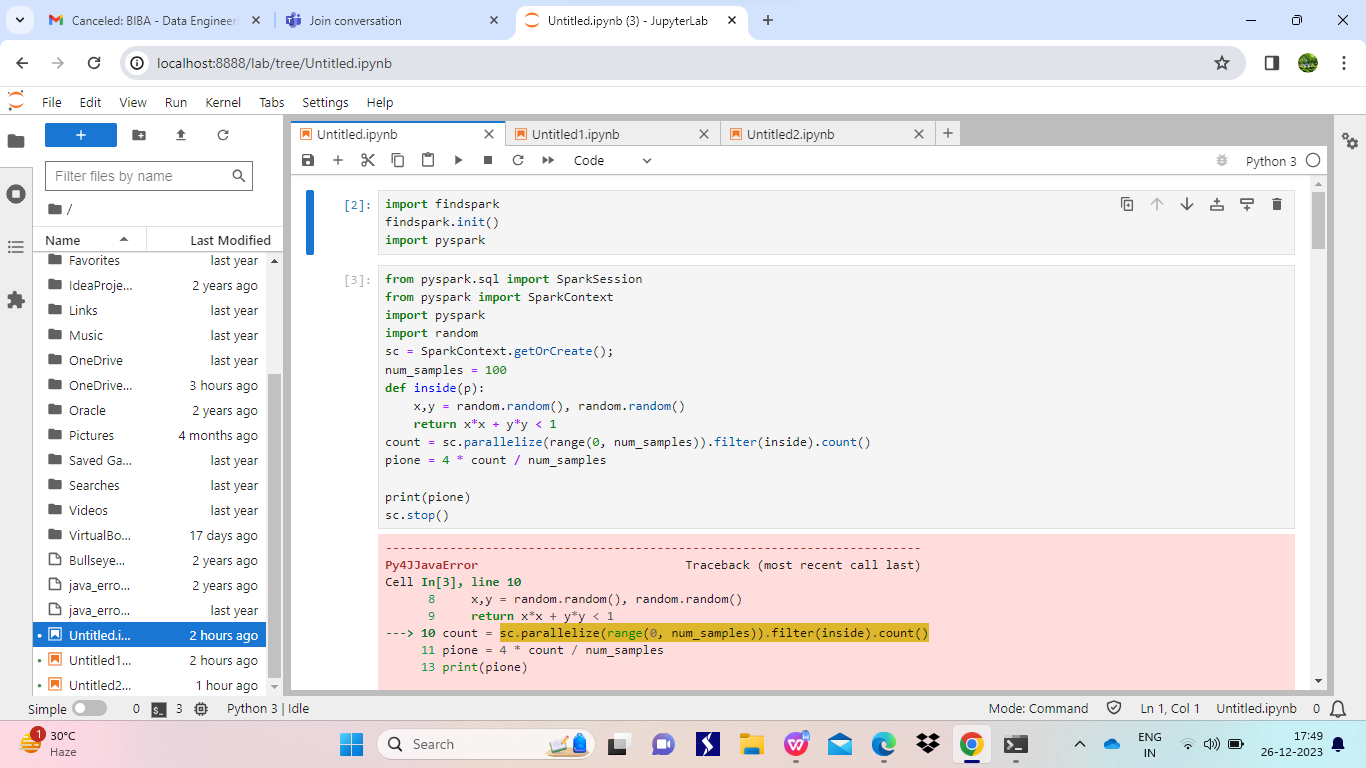
Python -m jupyterlab

Now it will open in the browser now we can select the notebook to write the spark commands

To import the pyspark and other modules first we have to import findspark which can be imported by installing findspark in command prompt by hitting the command pip install findspark

Now all set to run the pyspark code in the jupyterlab

By executing the code we can get the output below



**ELT-Extract load Transform**

It is a process which is helpful to deal with the huge databases and bigdata

To perform ELT process pyspark is very helpful because it is

Easy to use: as it contains a python code mostly which is easy to read and understand

Performance: It is 100 times faster than any other traditional machine

Scalable:it can handle very large amount of data and datasets

Rich Ecosystem:pyspark consists of different tools and libraries

**ETL work flow**

Extract : it extracts the input data from databases or local systems or APIs

Transform: It transforms the data as required by the user for analysis as per the requirement

Load: It saves the data that is transformed inorder to do analysis further

To initialize the spark first we have to import some modules

From pyspark.sql import SparkSession

To perform some actions or transformations we have to import some more functions from spark.sql.functions

The command to initialize spark session is

Spark=SparkSession.builder.appName(“name of session”).getOrCreate()

To extract data we use below command

df = spark.read.csv(source\_path, header=True,schema =’col\_name1 datatype,col\_name2 datatype,….’)

Here source\_path is the path of the file from where we have to extract the data

To transform data we have many functions to transform data some among them are

df=df.filter(col(age)>29)

To get data whose age is more than 29

df=df.orderBy(“salary”)

To get data in the ascending order of salary

And so on

To load or save the data we use the below command

df.write.csv(target\_path,mode=”overwrite”,header=True)

Where target\_path is the path where we have to save the transformed data