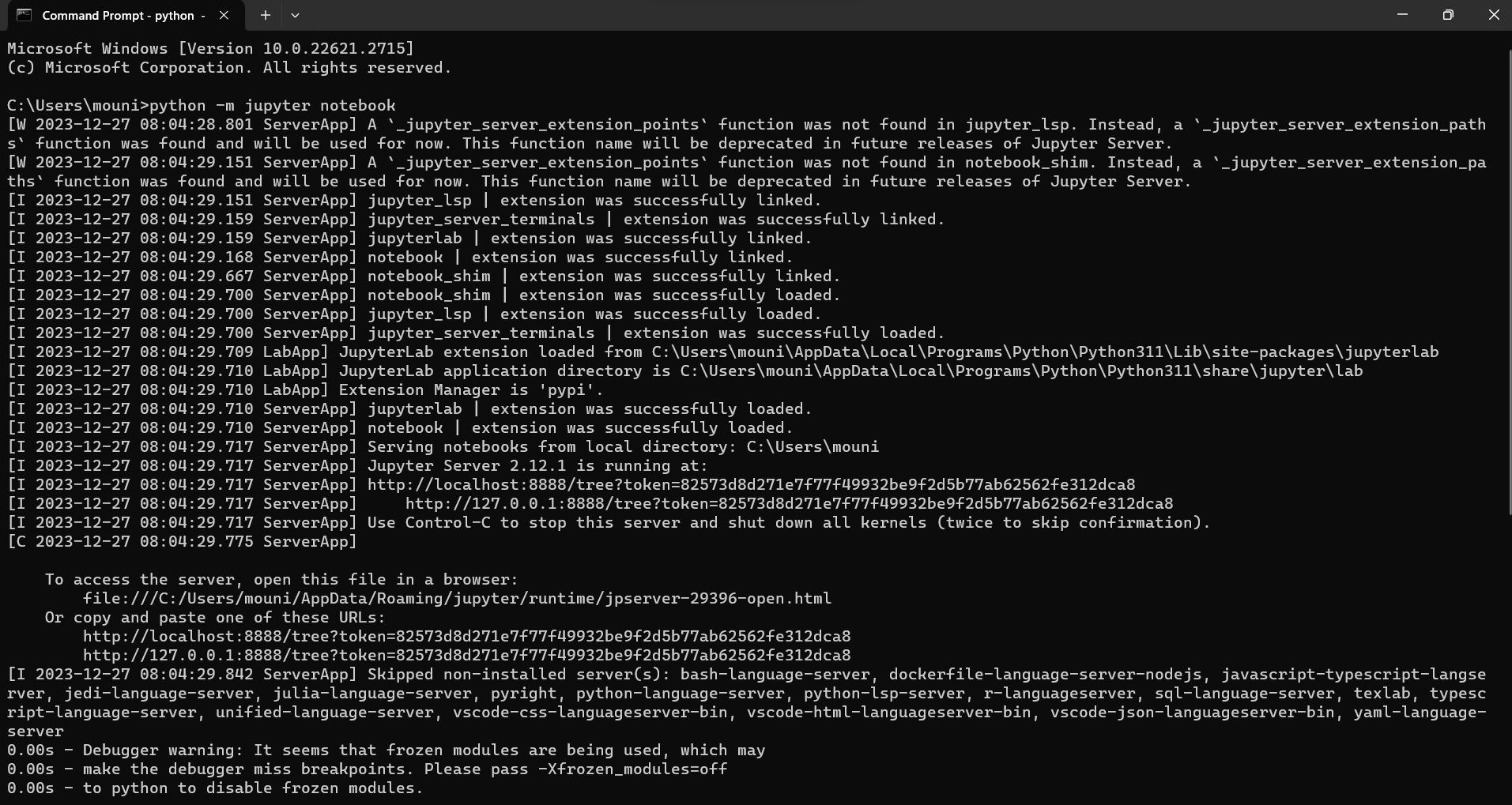
DATE: 27-12-23

Spark Assignment-4

Initializing jupyter notebook:



A screenshot of a computer

Description automatically generated

A computer screen shot of a computer screen

Description automatically generated

**ETL:**

ETL means extract transform and load.

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.

**Pyspark ETL:**

PySpark, the Python API for Apache Spark, has revolutionized the way we handle Big Data. It’s an ETL powerhouse that combines the simplicity of Python with the scalability and performance of Spark.

Extracting data:

Extract: Read data from an external CSV file  
df = spark.read.csv(source\_path, header=True,schema = ‘cust\_id int, first\_name string,last\_name string,cust\_order int,cust\_status string’)

Transforming data:

 Group by Age and Calculate Average Salary  
avg\_salary\_by\_age = df.groupBy(“age”).agg({“net\_salary”: “avg”}).withColumnRenamed(“avg(salary)”, “avg\_salary”).

Transforming data includes sorting,removing duplicates,merging etc. It changes the data to be able perform analytical process.

Loading data:

Save the transformed data to an external CSV file  
df.write.csv(target\_path, mode=”overwrite”, header=True)

Loading the data is nothing but storing it in any database oor dataware house to access the data easily and efficiently.