

# PYSPARK CODING ASSESSMENT

NAME-SWARNA LATHA GUVVALA

DATE-27/12/2023

## EXPLAIN ETL (EXTRACT, TRANSFORM, LOAD) WITH PYSPARK?

ETL, which stands for Extract, Transform Load, is a common process in data warehousing and data integration .It involves retrieving data from various sources, transforming it into a suitable format ,and loading it into a database or data warehouse.

The ETL Workflow in PYSPARK:

- ◇ 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.

## USING SPARK SQL - CREATING DATABASES, TABLES ?

Spark & PySpark SQL allows you to create a database and table either directly from DataFrame

### CREATE DATABASESE:

Creates a database with the specified name. If database with the same name already exists, an exception will be thrown.

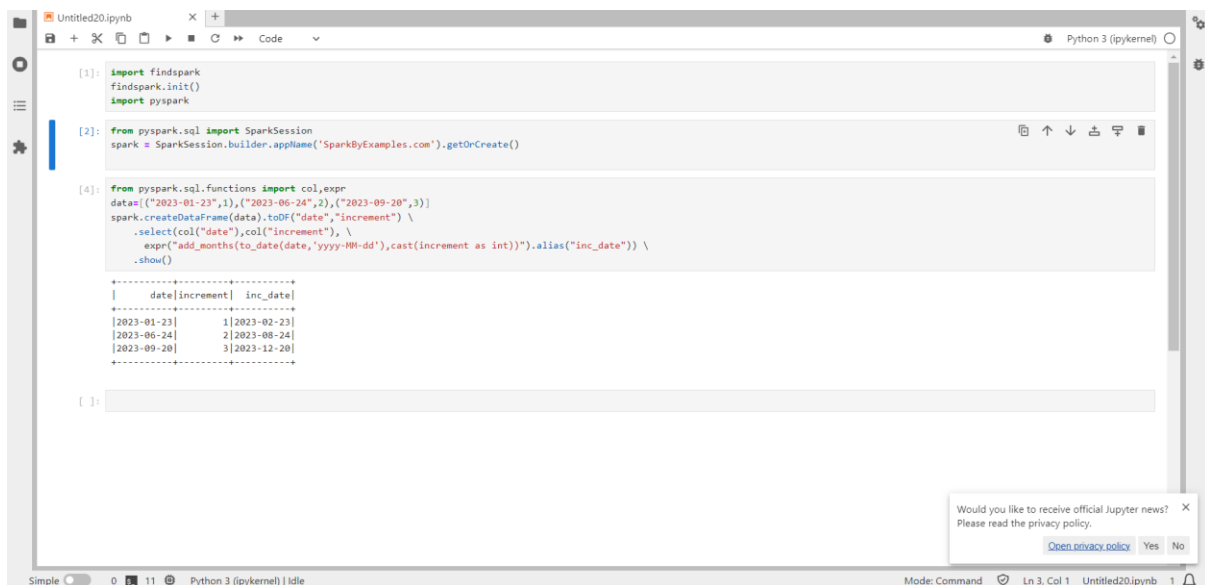
### Syntax

CREATE {DATABASE | SCHEMA} [ IF NOT EXISTS ] database\_name

[ COMMENT database\_comment ]

[ LOCATION database\_directory ]

[ WITH DBPROPERTIES (property\_name=property\_value [ , ...]) ]



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

[2]: from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()

[4]: from pyspark.sql.functions import col,expr
data=[("2023-01-23",1),("2023-06-24",2),("2023-09-20",3)]
spark.createDataFrame(data).toDF("date","increment") \
.select(col("date"),col("increment"), \
expr("add_months(to_date(date,'yyyy-MM-dd'),cast(increment as int))").alias("inc_date")) \
.show()
```

date	increment	inc_date
2023-01-23	1	2023-02-23
2023-06-24	2	2023-08-24
2023-09-20	3	2023-12-20

[ ]:

Would you like to receive official Jupyter news? Please read the privacy policy. [Open privacy policy](#) Yes No

## **USING SPARK SQL - TRANSFORMATIONS SUCH AS FILTER, JOIN, SIMPLE AGGREGATIONS, GROUPBY?**

### **Filtering data – where clause:**

We can use where clause to filter the data.

- One by using class.attributeName and comparing with values
- Make sure both orders and orderItems data frames are created

### **Aggregations using group by :**

Many times we want to perform aggregations such as sum, average, minimum, maximum etc with in each group. We need to first group the data and then perform aggregation.

- group by is the function which can be used to group the data on one or more columns
- Once data is grouped we can perform all supported aggregations – sum, avg, min, max etc