Asmita Porwal
Batch-1
Day-15
10/2/2024
Data engineering

Assignment-15

Hand written notes during the session :

NOTES

Date : ..../..../.....

empDF.join(deptDF, empDF.emp_dept_id == deptDF.dept_id, "inner").show()

10/2/24

→ Spark SQL

It is a component on top of Spark core that introduces a new data abstraction called Schema RDD

→ challenges

. Perform ETL to and from various data sources
. Perform advanced analytics that are hard to express in relational systems

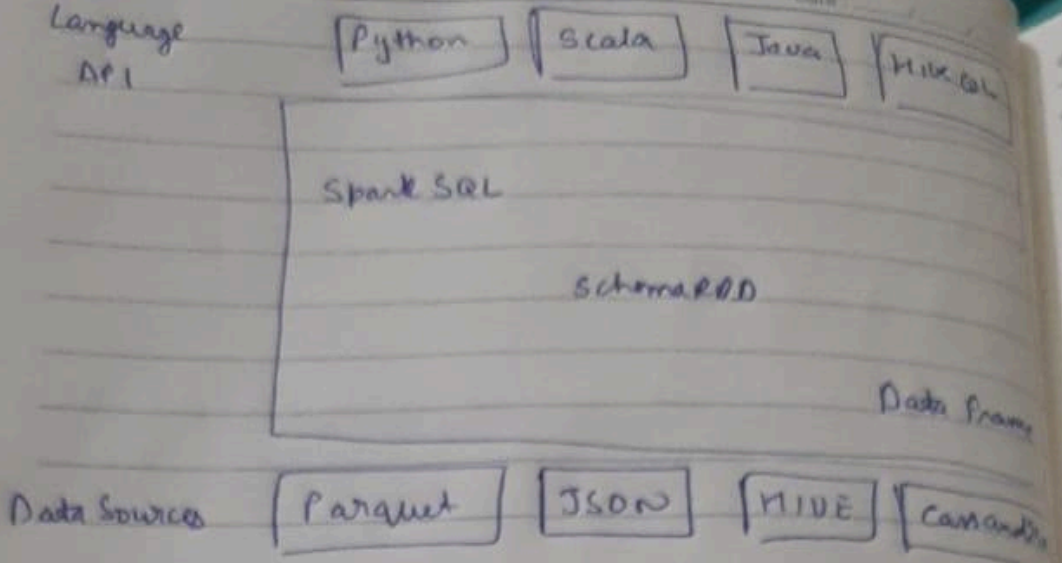
→ Solutions.

. A DF API that can perform relational operations on both external data sources and spark built in RDD.

- Spark SQL Architecture

NOTES

Date



- Features of Spark SQL

1. Integrated
   - Seamlessly mix SQL queries with Spark programs.
   - Spark SQL let you query structured.

3 Hive compatibility
   - Run unmodified Hive queries on existing warehouse.

4. Standard Connectivity
   Connect through JDBC or O

NOTES

Date : ...../...../.....

- User Defined functions

- UDF: plug in your own procession code

Database

```
In [53]: spark = SparkSession.builder.appName("Practice").config("spark.sql.catalogImplementation", "hive").getOrCreate()

In [56]: spark.sql("CREATE DATABASE customer_db;")
Out[56]: DataFrame[]

In [64]: spark.sql("CREATE DATABASE IF NOT EXISTS customer_db COMMENT 'This is customer database'WITH DBPROPERTIES (ID=1, Name='John')");

In [70]: spark.sql("DESCRIBE DATABASE customer_db")
Out[70]: DataFrame[info_name: string, info_value: string]
```

```
In [73]: from pyspark.sql import SparkSession

         spark = SparkSession \
             .builder \
             .appName("Python Spark SQL basic example") \
             .config("spark.some.config.option", "some-value") \
             .getOrCreate()

In [84]: df = spark.read.json("E:\\downloads\\test.json" ,multiLine=True)

In [85]: df
Out[85]: DataFrame[employee: struct<married:boolean,name:string,salary:bigint>]

In [86]: df.show()
         +--------------------+
         |            employee|
         +--------------------+
         |{true, sonoo, 56000}|
         +--------------------+

In [ ]:
```