Assignment – Day 13

-Sarthak Niranjan Kulkarni (Maverick)

- sarthakkul2311@gmail.com - (+91) 93256 02791

20/11/2024 (Wednesday)

Summary of View in Spark:-

1. **Spark Session Creation**:

- The SparkSession is created using .builder.appName("SparkByExamples.com").enableHiveSupport().getOrCreate(). This initializes a Spark session that can interact with Hive if needed.

2. **Data and Schema Setup**:

- A list of tuples (data) is created, containing sample personal information such as first name, last name, country, and state.
- A list of column names (columns) is defined: "firstname", "lastname", "country", and "state".

3. **DataFrame Creation**:

- The data is converted into a DataFrame (sampleDF) by using spark.sparkContext.parallelize(data).toDF(columns). The parallelize function distributes the data across the Spark cluster, and toDF(columns) converts the list of data into a structured DataFrame with specified columns.

4. **Creating Temporary Views**:

- The sampleDF DataFrame is registered as two temporary SQL views: "Person" and "mydata", using createOrReplaceTempView(). These views allow Spark SQL queries to be executed against the DataFrame.

5. **Displaying Data**:

- sampleDF.show() is used to display the contents of the DataFrame in a tabular format.
- spark.sql("select * from person").show() and spark.sql("select * from mydata").show() run SQL queries on the two views and display the same data since both views are based on the same sampleDF DataFrame.

Views Practice: -

1. **Creating a Spark DataFrame and registering it as temporary SQL views for querying.**

```
from pyspark.sql import SparkSession
# Create spark session
spark = SparkSession \
.builder \
.appName("SparkByExamples.com") \
.enableHiveSupport() \
.getOrCreate()
data = [("Sarthak","Kulkarni","IND","MH"),
("Lakshita","Sathe","IND","MP"),
("Harsh","Choudhari","USA","COL"),
("Pratik","Pathak","IRE","DUB")]
columns = ["firstname","lastname","country","state"]
# Create dataframe
sampleDF = spark.sparkContext.parallelize(data).toDF(columns)
sampleDF.createOrReplaceTempView("Person")
sampleDF.createOrReplaceTempView("mydata")
sampleDF.show()
```

```
▸ (5) Spark Jobs
▸ sampleDF: pyspark.sql.dataframe.DataFrame = [firstname: string, lastname: string ... 2 more fields]
+---------+---------+-------+-----+
|firstname| lastname|country|state|
+---------+---------+-------+-----+
|  Sarthak| Kulkarni|    IND|   MH|
| Lakshita|    Sathe|    IND|   MP|
|    Harsh|Choudhari|    USA|  COL|
|   Pratik|   Pathak|    IRE|  DUB|
+---------+---------+-------+-----+
```

---

2. Executing SQL queries on temporary views in Spark to display data

```
spark.sql("select * from person").show()
spark.sql("select * from mydata").show()
```

▸ (6) Spark Jobs

```
+---------+---------+-------+-----+
|firstname| lastname|country|state|
+---------+---------+-------+-----+
|  Sarthak| Kulkarni|    IND|   MH|
| Lakshita|    Sathe|    IND|   MP|
|    Harsh|Choudhari|    USA|  COL|
|   Pratik|   Pathak|    IRE|  DUB|
+---------+---------+-------+-----+

+---------+---------+-------+-----+
|firstname| lastname|country|state|
+---------+---------+-------+-----+
|  Sarthak| Kulkarni|    IND|   MH|
| Lakshita|    Sathe|    IND|   MP|
|    Harsh|Choudhari|    USA|  COL|
|   Pratik|   Pathak|    IRE|  DUB|
+---------+---------+-------+-----+
```