Spark Dataframe – Scala (Part II)

Use case 1 :

Read a Json File in spark Dataframe :

scala> val df = spark.read.json("/home/thirra4/Documents/sampledata.json")

```
scala> val df = spark.read.json("/home/thirra4/Documents/sampledata.json")
df: org.apache.spark.sql.DataFrame = [deptno: string, designation: string ... 5 more fields]

scala> df.show(5);
+------+-----------+-----+------+----------+-------+----+
|deptno|designation|empno| ename| hire_date|manager| sal|
+------+-----------+-----+------+----------+-------+----+
|    20|      CLERK| 7369| SMITH|12/17/1980|   7902| 800|
|    30|   SALESMAN| 7499| ALLEN| 2/20/1981|   7698|1600|
|    30|   SALESMAN| 7521|  WARD| 2/22/1981|   7698|1250|
|    20|    MANAGER| 7566|TURNER|  4/2/1981|   7839|2975|
|    30|   SALESMAN| 7654|MARTIN| 9/28/1981|   7698|1250|
+------+-----------+-----+------+----------+-------+----+
only showing top 5 rows
```

Use case II :

Read a Multiline Json :

scala> val json_data1 = spark.read.option("multiline","true").json("/home/thirra4/Downloads/multiline.json")

```
scala> val json_data1 = spark.read.option("multiline","true").json("/home/thirra4/Downloads/multiline.json")
json_data1: org.apache.spark.sql.DataFrame = [City: string, RecordNumber: bigint ... 3 more fields]
```

Scala > json_data1.show();

```
scala> json_data1.show(2);
+--------------------+------------+-----+-----------+-------+
|                City|RecordNumber|State|ZipCodeType|Zipcode|
+--------------------+------------+-----+-----------+-------+
|PASEO COSTA DEL SUR|           2|   PR|   STANDARD|    704|
|        BDA SAN LUIS|          10|   PR|   STANDARD|    709|
+--------------------+------------+-----+-----------+-------+


scala>
```

Use Case III

Reading all json files from a folder:

Example : I kept two files in my directory

/home/thirra4/Documents/json_sample/sample_json1

/home/thirra4/Documents/json_sample/sample_json2

scala> val json_allfiles = spark.read.json("/home/thirra4/Documents/json_sample")

```
scala> val json_allfiles = spark.read.json("/home/thirra4/Documents/json_sample")
json_allfiles: org.apache.spark.sql.DataFrame = [deptno: string, designation: string ... 5 more fields]

scala> json_allfiles.show(15);
+------+-----------+-----+--------------+----------+-------+----+
|deptno|designation|empno|         ename| hire_date|manager| sal|
+------+-----------+-----+--------------+----------+-------+----+
|    20|      CLERK| 6369|         SMITH|12/17/1980|   7902| 800|
|    30|   SALESMAN| 5499|         ALLEN| 2/20/1981|   7698|1600|
|    30|   SALESMAN| 4421|christo ronaldo| 2/22/1981|   7698|1250|
|    20|    MANAGER| 3566|     aln TURNER|  4/2/1981|   7839|2975|
|    30|   SALESMAN| 2694|  MARTIN cruise| 9/28/1981|   7698|1250|
|    30|    MANAGER| 1598|   MILLER david|  5/1/1981|   7839|2850|
|    10|    MANAGER| 9482|     CLARK mike|  6/9/1981|   7839|2450|
|    20|    ANALYST|57188|   SCOTT andrew| 12/9/1982|   7566|3000|
|    10|  PRESIDENT|68639|          KING|11/17/1981|   NULL|5000|
|    20|      CLERK| 7369|         SMITH|12/17/1980|   7902| 800|
|    30|   SALESMAN| 7499|         ALLEN| 2/20/1981|   7698|1600|
|    30|   SALESMAN| 7521|          WARD| 2/22/1981|   7698|1250|
|    20|    MANAGER| 7566|         TURNER|  4/2/1981|   7839|2975|
|    30|   SALESMAN| 7654|         MARTIN| 9/28/1981|   7698|1250|
|    30|    MANAGER| 7698|         MILLER|  5/1/1981|   7839|2850|
+------+-----------+-----+--------------+----------+-------+----+
only showing top 15 rows


scala>
```

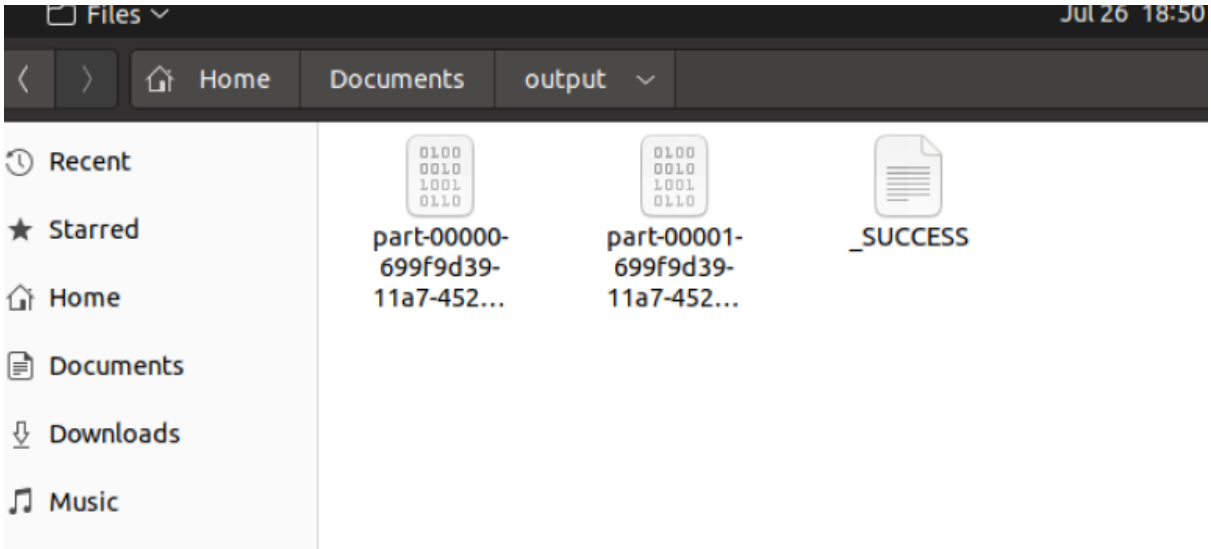
Use Case IV

To write json data into CSV file or parquet file:

scala>
json_allfiles.write.mode("overwrite").csv("/home/thirra4/Documents/output")

scala>
json_allfiles.write.mode("overwrite").parquet("/home/thirra4/Documents/output")

```
scala> json_allfiles.write.mode("overwrite").csv("/home/thirra4/Documents/output")
scala> json_allfiles.write.mode("overwrite").parquet("/home/thirra4/Documents/output")
scala>
```



Use Case V

How to access the Hive table and write data to ORC format using HiveContext in Scala:

scala> val hiveContext = new
org.apache.spark.sql.hive.HiveContext(spark.sparkContext)

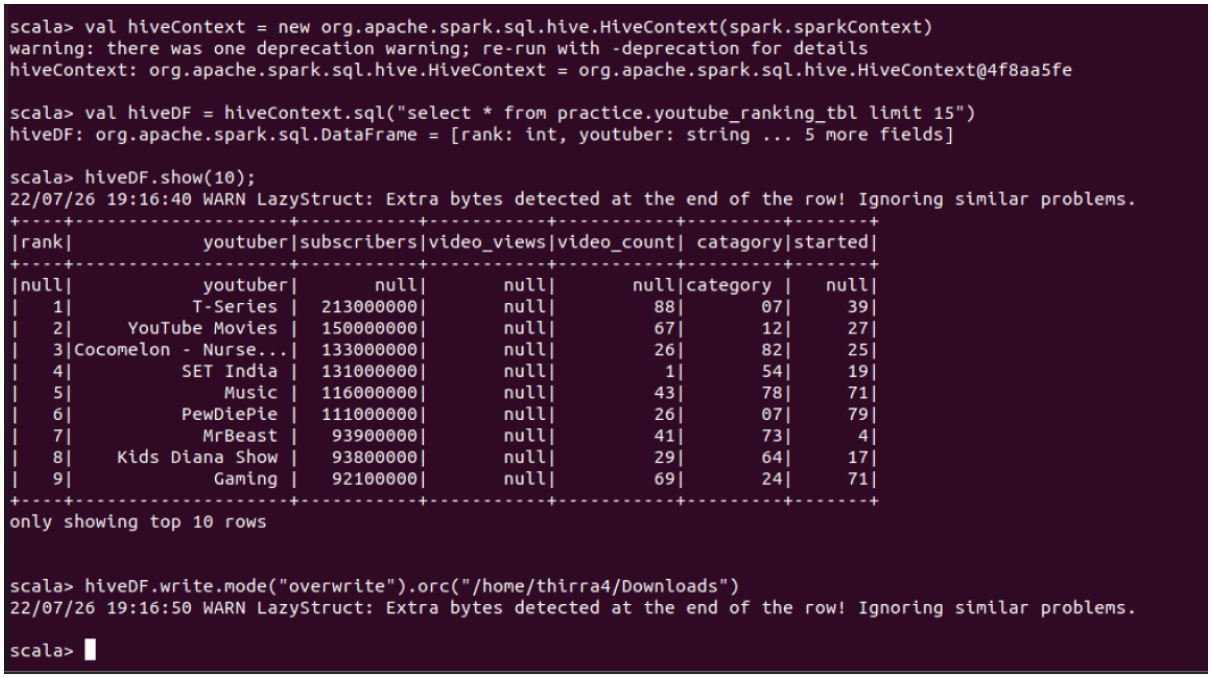
warning: there was one deprecation warning; re-run with -deprecation for details

hiveContext: org.apache.spark.sql.hive.HiveContext = org.apache.spark.sql.hive.HiveContext@4f8aa5fe

scala> val hiveDF = hiveContext.sql("select * from practice.youtube_ranking_tbl limit 15")

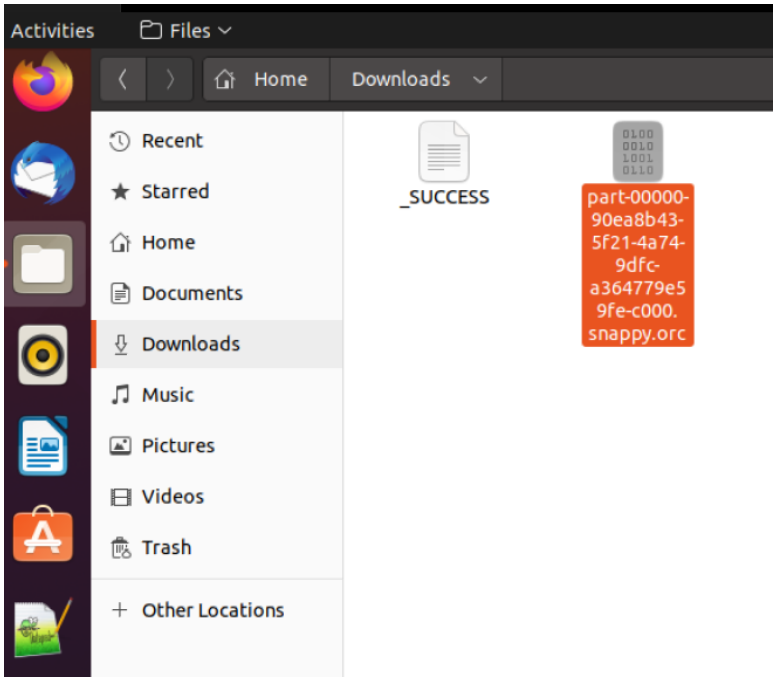
hiveDF: org.apache.spark.sql.DataFrame = [rank: int, youtuber: string ... 5 more fields]

scala> hiveDF.show(10);



```
scala> val hiveContext = new org.apache.spark.sql.hive.HiveContext(spark.sparkContext)
warning: there was one deprecation warning; re-run with -deprecation for details
hiveContext: org.apache.spark.sql.hive.HiveContext = org.apache.spark.sql.hive.HiveContext@4f8aa5fe

scala> val hiveDF = hiveContext.sql("select * from practice.youtube_ranking_tbl limit 15")
hiveDF: org.apache.spark.sql.DataFrame = [rank: int, youtuber: string ... 5 more fields]

scala> hiveDF.show(10);
22/07/26 19:16:40 WARN LazyStruct: Extra bytes detected at the end of the row! Ignoring similar problems.
+----+-------------------+-----------+-----------+-----------+---------+-------+
|rank|            youtuber|subscribers|video_views|video_count| catagory|started|
+----+-------------------+-----------+-----------+-----------+---------+-------+
|null|            youtuber|       null|       null|       null|category |   null|
|   1|          T-Series |  213000000|       null|         88|       07|     39|
|   2|     YouTube Movies |  150000000|       null|         67|       12|     27|
|   3|Cocomelon - Nurse...|  133000000|       null|         26|       82|     25|
|   4|          SET India |  131000000|       null|          1|       54|     19|
|   5|              Music |  116000000|       null|         43|       78|     71|
|   6|          PewDiePie |  111000000|       null|         26|       07|     79|
|   7|            MrBeast |   93900000|       null|         41|       73|      4|
|   8|    Kids Diana Show |   93800000|       null|         29|       64|     17|
|   9|             Gaming |   92100000|       null|         69|       24|     71|
+----+-------------------+-----------+-----------+-----------+---------+-------+
only showing top 10 rows


scala> hiveDF.write.mode("overwrite").orc("/home/thirra4/Downloads")
22/07/26 19:16:50 WARN LazyStruct: Extra bytes detected at the end of the row! Ignoring similar problems.

scala>
```

scala> hiveDF.write.mode("overwrite").orc("/home/thirra4/Downloads")


Activities
Files
Home
Downloads
Recent
Starred
Home
Documents
Downloads
Music
Pictures
Videos
Trash
Other Locations
_SUCCESS
part-00000-90ea8b43-5f21-4a74-9dfc-a364779e59fe-c000.snappy.orc