

WEEK 12 FAQS

W12:1 Renshi is getting error as shown in below picture

```
import org.apache.spark.sql.SparkSession
import org.apache.log4j.Level
import org.apache.log4j.Logger
object DataFrame9_ColumnString_Exp extends App {
  val spark = SparkSession.builder()
    .appName("This is program for reading the file")
    .enableHiveSupport()
    .master("local[1]")
    .getOrCreate()
  Logger.getLogger("org").setLevel(Level.ERROR)

  val baseDf = spark.read
    .format("csv")
    .option("header", true)
    .option("inferSchema", true)
    .option("path", "/c:/BigDataFactory/BigData/WK12/DataSet/orders.csv")
    .load

  //baseDf.select("order_id", "order_date", "order_customer_id", "order_status").show
  import spark.implicits._

  baseDf.select(column("order_id"), col("order_date"), $"order_customer_id", 'order_status).show()

}
```

Ans: Add import statement, import org.apache.spark.sql.functions._

W12:2 <DF>.write.format("avro").partitionBy("country").mode(SaveMode.Overwrite).save(sink_path_avro)

during the above write operation... i am facing an Exception in thread "main" org.apache.spark.sql.AnalysisException: Failed to find data source: avro. Avro is a built-in but external data source module since Spark 2.4. Please deploy the application as per the deployment section of "Apache Avro Data Source Guide". At org.apache.spark.sql.execution.datasources.DataSource\$.lookupDataSource(DataSource.scala:665) at org.apache.spark.sql.DataFrameWriter.save(DataFrameWriter.scala:265) at org.apache.spark.sql.DataFrameWriter.save(DataFrameWriter.scala:249) at main.scala.com.week11.structuredapi.assignment.ProblemOne\$.delayedEndpoint\$main\$scala\$com\$week11\$structuredapi\$assignment\$ProblemOne\$1(ProblemOne.scala:69) at

WEEK 12 FAQS

main.scala.com.week11.structuredapi.assignment.ProblemOne\$delayedInit\$body.apply(ProblemOne.scala:14) did I miss any dependency?

Ans: We have to add spark avro dependency .

https://mvnrepository.com/artifact/org.apache.spark/spark-avro_2.11/2.4.5 .

Note : Do download jar of proper version.

W12:3 Cannot find the "json" files for orders and customers data which is discussed in Structured Api Session-19. Please help as all the others dataset were there in "Download: Week12 Practice Datasets". But json files are not there.

Ans: You can create it, by writing the spark program, take input as csv & using standard dataframe writer api save results in json.

W12:4 We have column object expr, column string expr example. While using column object expression we are not using the "column or col" keyword. Raj was confused as in session 13 while using column object we do need to mention "column or col" keyword.

Ans: Session covers both the approaches a programmatic and SQL.
In programmatic approach , if we look at the function signatures of sum , avg, count, etc they can accept both string as well as column object as input (Method Overloading).

So, we can use either sum("Quantity") or sum(col("Quantity")) here

W12:5 Ritu is getting error. Error: value toDS is not a member of org.apache.spark.rdd.RDD[Orders]

WEEK 12 FAQS

```
assignment.scala      assignment1.scala      DataFrameEx.scala      regexeg.scala      explicitSchema.scala
import org.apache.spark.SparkConf
import org.apache.spark.sql.SparkSession

object regexeg {
  def main(args: Array[String]){

    val myregex = """^(\S+) (\S+)\t(\S+),(\S+)"""
    case class Orders(orderid:Int,cusid:Int,status:String)

    def parser(line:String) ={
      line match{
        case myregex(orderid,date,cusid,status) =>
          Orders(orderid.toInt, cusid.toInt, status)
      }
    }

    val sparkConf = new SparkConf
    sparkConf.set("spark.app.name","regexeg")
    sparkConf.set("spark.master","local[*]")

    val spark = SparkSession.builder()
      .config(sparkConf)
      .getOrCreate()

    val line = spark.sparkContext.textFile("H:/Big Data/trendytech/Week12_Spark_Structured_API_2/unstructured")
    import spark.implicits._

    val inpDS = line.map(parser)
    val Orderds= inpDS.toDS()

    Orderds.printSchema()
    Orderds.groupBy("order_status").count()
  }
}
```

Ans: Case class should be defined above main function. So put the case class definition above main

W12:6 While writing data to local file using dataframe , Tisha is getting an errors Caused by: org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 2.0 failed 1 times, most recent failure: Lost task 0.0 in stage 2.0 (TID 3, localhost, executor driver): ExitCodeException exitCode=-1073741515: at org.apache.hadoop.util.Shell.runCommand(Shell.java:582) code snippet ordersDf.write .format("csv") .mode(SaveMode.Overwrite) .option("path","C:\\Users\\Trendy Tech\\Week-12-Spark-API-2\\output3") .save()

Ans: This is now resolve, Windows required MSVCR100.dll file to be installed follow this link to install it
<https://www.microsoft.com/en-hk/download/confirmation.aspx?id=13523>

WEEK 12 FAQS

W12:7 While writing data is there a way we can decide the output filename? Like some specific name with date n time information when the file is created?

Ans: In spark, filename prefix is hardcoded hence cant be changed. If you need the filename of your own, you can

- Rename the file
- Implement a custom FileOutputStream and use one of Spark's save methods that accept a FileOutputStream class

W12:8 Rosy ran the below code and getting error on unix_timestamp.. She has added the req imports as in video.. still issue persists

The screenshot shows a Jupyter Notebook interface with a Scala code cell. The code sets up a SparkSession, creates a list of tuples representing orders, and then creates a DataFrame from this list. It adds a timestamp column using `unix_timestamp` and prints the schema and content of the DataFrame.

```
sparkConf.set("spark.app.name", "my first application")
sparkConf.set("spark.master", "local[2]")

val spark = SparkSession.builder()
  .config(sparkConf)
  .getOrCreate()

val myList = List((1,"2013-07-25",1159,"CLOSED"),
  (2,"2014-07-25",256,"PENDING_PAYMENT"),
  (3,"2013-07-25",11599,"COMPLETE"),
  (4,"2019-07-25",8827,"CLOSED"))

import spark.implicits._

val ordersDF = spark.createDataFrame(myList).toDF("order_id","orderdate","customerid","status")

val newDF = ordersDF.withColumn("orderdate", unix_timestamp(col("orderdate").cast(DateType)))

newDF.printSchema()
newDF.show()

spark.stop
```

Output:

```
-- orderdate: string (nullable - true)
-- customerid: integer (nullable - false)
-- status: string (nullable - true)
```

order_id	orderdate	customerid	status
1	2013-07-25	1159	CLOSED
2	2014-07-25	256	PENDING_PAYMENT
3	2013-07-25	11599	COMPLETE
4	2019-07-25	8827	CLOSED

Ans: import org.apache.spark.sql._

WEEK 12 FAQS

W12:9 While watching videos Shakshi noticed that in order to print dataframe we have used show in different ways for example ordersDF.show ordersDF.show() So, She is assuming that braces after show is optional. Also, I am wondering if it is the case with other methods as well

Ans: In Scala you can omit parentheses () on methods which have no arguments

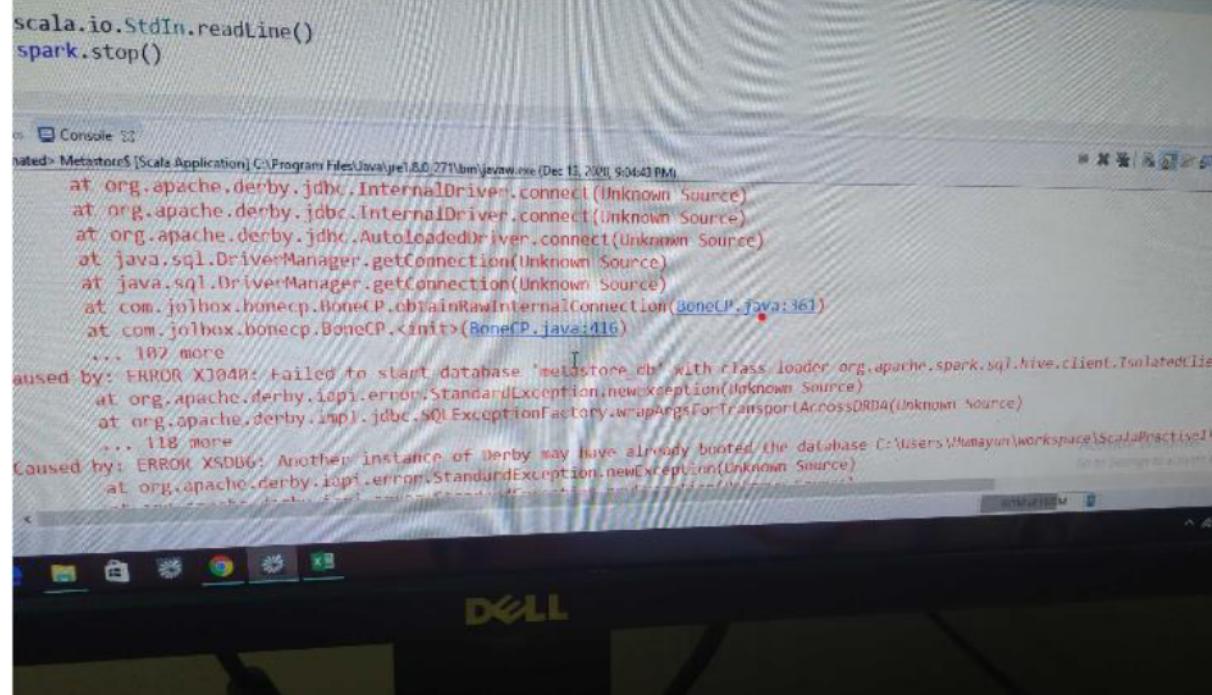
so you can write ordersDF.show() as ordersDF.show

However, this is not universal for all methods. It can only be used when the method has no side-effects that means it should be purely-functional.

So, we can omit parentheses when calling df.show()

but we can't when calling println() - because it has side effects.

W12:10 Yash is getting below error.. pls help in resolving below error



```
scala.io.StdIn.readLine()
spark.stop()

scala> Metastore [Scala Application] C:\Program Files\Java\jre1.8.0_271\bin\javaw.exe (Dec 15, 2020, 9:04:43 PM)
        at org.apache.derby.jdbc.InternalDriver.connect(Unknown Source)
        at org.apache.derby.jdbc.InternalDriver.connect(Unknown Source)
        at org.apache.derby.jdbc.AutoLoadedDriver.connect(Unknown Source)
        at java.sql.DriverManager.getConnection(Unknown Source)
        at java.sql.DriverManager.getConnection(Unknown Source)
        at com.jolbox.bonecp.BoneCP.obtainRawInternalConnection(BoneCP.java:361)
        at com.jolbox.bonecp.BoneCP.<init>(BoneCP.java:416)
        ... 107 more
Caused by: ERROR XJ04R: Failed to start database 'metastore_db' with class loader org.apache.spark.sql.hive.client.ThriftClient$Client
        at org.apache.derby.iapi.error.StandardException.newException(Unknown Source)
        at org.apache.derby.impl.jdbc.SQLExceptionFactory.wrapArgsForTransportAcrossDRDA(Unknown Source)
        ... 118 more
Caused by: ERROR XSDB6: Another instance of Derby may have already booted the database C:\Users\Humayun\workspace\ScalaPractices\1
        at org.apache.derby.iapi.error.StandardException.newException(Unknown Source)
        at org.apache.derby.iapi.error.StandardException.newException(Unknown Source)
```

Ans: try deleting metastore_db folder in your project first and then restart the eclipse again, it will be created again with next execution.

WEEK 12 FAQS

W12:11 Try to run dataframe writer using hive meta store.download and add hive jar file , add enableHiveSupport() in spark session.create the database and use it in saveAsTable. No error is displaying in coding , during run the program metastore-db is created but warehouse dir was not created and below error is showing in console:

Caused by: java.lang.RuntimeException: java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be writable. Current permissions are: rw-rw-rw-

At

org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.java:522)

at

org.apache.spark.sql.hive.client.HiveClientImpl.newState(HiveClientImpl.scala:183)

at

org.apache.spark.sql.hive.client.HiveClientImpl.<init>(HiveClientImpl.scala:117)

at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)

at sun.reflect.NativeConstructorAccessorImpl.newInstance(Unknown Source)

at

sun.reflect.DelegatingConstructorAccessorImpl.newInstance(Unknown Source)

at java.lang.reflect.Constructor.newInstance(Unknown Source)

Please someone suggest where the issue is?

Ans: you try this

<https://medium.com/@Zhuinden/i-got-this-error-c57236af17ee>

this C:\BigDataLocalSetUp\hadoop\bin\winutils.exe chmod 777

c:/tmp/hive command is working for me in win 10 else can u try giving full control to winutils.exe using properties

W12:12 Ashish is getting this error. ("Could not find or load main class <objectname>)

WEEK 12 FAQS

```
import org.apache.spark.sql.SparkSession
import org.apache.log4j.Level
import org.apache.log4j.Logger
import org.apache.spark.sql.types.DateType

object CaseStudy extends App {

    Logger.getLogger("org").setLevel(Level.ERROR)

    val sparkConf = new SparkConf()
    sparkConf.set("spark.app.name", "converting rdd to df")
    sparkConf.set("spark.master", "local[*]")

    val spark = SparkSession.builder()
        .config(sparkConf)
        .getOrCreate()

    val myList = List((1, "2013-07-25", 11599, "CLOSED"),
                    (2, "2014-07-25", 256, "PENDING_PAYMENT"),
                    (3, "2013-07-25", 11599, "COMPLETE"),
                    (4, "2019-07-25", 8827, "CLOSED"))

    import spark.implicits._

    val ordersDF = spark.createDataFrame(myList)
        .toDF("orderid", "order_date", "customer_id", "status")

    val newDF = ordersDF.withColumn("order_date", unix_timestamp(col("order_date").cast(DateType)))
}
```

Ans: import "org.apache.spark.sql.functions._" was missing.

W12:13 In Hive - Data is stored on HDFS and metadata is stored in Mysql. But when we want to save as a table in spark.....we are using hive metastore....so hive internally again stores in mysql ?

Ans: By default, spark sql uses the embedded Derby database to store metadata.

The default embedded Derby deployment mode is not recommended for production environments due to limitation of only one active SparkSession at a time.

So we use enableHiveSupport() method to store metadata in Hive Metastore.

Hive will store Metadata in a backend RDBMS, it may be mysql or any RDBMS database configured by the admin.

W12:14 Raj need some clarification on the below assignment task(step-5)

Problem 2:

step 1: Create spark session

step 2: Set the logging level to error

WEEK 12 FAQS

step3: Load the data file windowdata.csv as a rdd

step 4: Create a dataframe from this RDD by defining case class

step 5: Save this dataframe in JSON format in 8 files. Save this dataframe in JSON format in 8 files

?

Ans: he is asking you to use df.repartition(8).

W12:15 Abhishek checked the spark UI after reading the dataset - orders_data.csv. This file has size around 44 MB and the number of records including header is 541783. But when checked in UI, the input Size/records are doubled =88.1MB/1083566. My question is that, is spark in the local system creating partitions with duplicate data?

The screenshot shows the Apache Spark Web UI interface. At the top, there's a navigation bar with links for 'MapPartitionsRDD[2/2]' and 'Input of 20 RddSampleRate 10'. Below the navigation, there are two buttons: 'Show Additional Metrics' and 'Event Timeline'. The main area is titled 'Summary Metrics for 4 Completed Tasks'. It displays summary statistics for tasks completed on executor 0:

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	2 s	2 s	3 s	3 s	2 s
GC Time	0.1 s				
Input Size / Records	20.0 MB / 245121	20.0 MB / 245121	24.1 MB / 296662	24.1 MB / 296662	24.1 MB / 296662

Below this, there's a section titled 'Aggregated Metrics by Executor' showing data for executor 0. The table includes columns for Executor ID, Address, Task Time, Total Tasks, Failed Tasks, Killed Tasks, Succeeded Tasks, Input Size / Records, and Blocked Tasks.

Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Input Size / Records	Blocked Tasks
0	Geometric:53951	11 s	4	0	0	4	88.1 MB / 1083566	0

Finally, the 'Tasks (4)' table provides detailed information for each task. It lists task index, ID, Attempt, Status, Locality Level, Executor ID, Host, Launch Time, Duration, GC Time, Input Size / Records, and Errors. The tasks are all marked as 'SUCCESS'.

Index	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration	GC Time	Input Size / Records	Errors
0	1	0	SUCCESS	PROCESS_LOCAL	0@geometric	localhost	2023/03/16 13:31:17	3 s	0.1 s	24.1 MB / 296662	
1	2	0	SUCCESS	PROCESS_LOCAL	0@geometric	localhost	2023/03/16 13:31:17	3 s	0.1 s	24.1 MB / 296662	
2	3	0	SUCCESS	PROCESS_LOCAL	0@geometric	localhost	2023/03/16 13:31:17	3 s	0.1 s	20.0 MB / 245121	
3	4	0	SUCCESS	PROCESS_LOCAL	0@geometric	localhost	2023/03/16 13:31:17	3 s	0.1 s	20.0 MB / 245121	

Ans: It is not duplicate data nor it is about input file size. When you load data spark load it in memory in the deserialized form which generally holds more space because it also includes overhead memory for intermediate results.

WEEK 12 FAQS

W12:16 While creating a dataframe using spark.createDataFrame(myList), Erick is getting an error as shown below

```
object DFUSeCase extends App {
  Logger.getLogger("org").setLevel(Level.ERROR)

  //setting up spark Conf
  val sparkConf = new SparkConf()
  sparkConf.set("spark.app.name", "my 1st Application")
  sparkConf.set("spark.master", "local[2]")

  //creating spark session
  val spark = SparkSession.builder()
    .config(sparkConf)
    .getOrCreate()

  val myList =((1,"2013-07-25",11599,"CLOSED"),
  (2,"2013-07-25",256,"PENDING PAYMENT"),
  (3,"2013-07-25",12111,"COMPLETE"),
  (4,"2013-07-25",8827,"CLOSED"))

  import spark.implicits._

  overloaded method value createDataFrame with alternatives: [A <: Product](data: Seq[A])(implicit evidence$3: reflect.runtime.universe.TypeTag[A])org.apache.spark.sql.DataFrame or
  [A <: Product](rdd: org.apache.spark.rdd.RDD[A])(implicit evidence$2: reflect.runtime.universe.TypeTag[A])org.apache.spark.sql.DataFrame cannot be applied
}
```

Ans: One more thing i notice that you have not mentioned List when you created the myList.

it should be:

```
val myList = List (.....)
```

Change this it will work do not forget to do a show

W12:17 object Assignment1_Week12 extends App {

```
Logger.getLogger("org").setLevel(Level.ERROR)
val sparkConf = new SparkConf()
sparkConf.set("spark.app.name", "Assignment1_Week12")
sparkConf.set("spark.master", "local[2]")
val spark = SparkSession.builder()
  .config(sparkConf)
  .getOrCreate()
val deptDf = spark.read
  .format("json")
  .option("path","/Users/minisha/Desktop/dept")
  .load()
//deptDf.show()
//deptDf.printSchema()
val employeeDf = spark.read
```

WEEK 12 FAQS

```
.format("json")
.option("path","/Users/minisha/Desktop/employee")
.load()
employeeDf.show()
employeeDf.printSchema()
val joinCondition = deptDf.col("deptid") === employeeDf.col("deptid")
val joinType = "left"
val joinedDf = deptDf.join(employeeDf, joinCondition, joinType)
val joinedDfNew = joinedDf.drop(employeeDf.col("deptid"))
joinedDfNew.groupBy("deptid").agg(count("empname").as("empcount"),first("deptName").as("deptName")).dropDuplicates("deptName").show()
spark.stop()
```

Ans:

- use comma instead of dot ex-->
.option("path","/Users/minisha/Desktop/employee")
- Also import necessary packages - it is not visible in your screen shot.

W12:18 Same is getting column and col not found while using column Object.

```
//column Object
import spark.implicits._
orderDF.select(column("order_id"),col("order_status"),
```

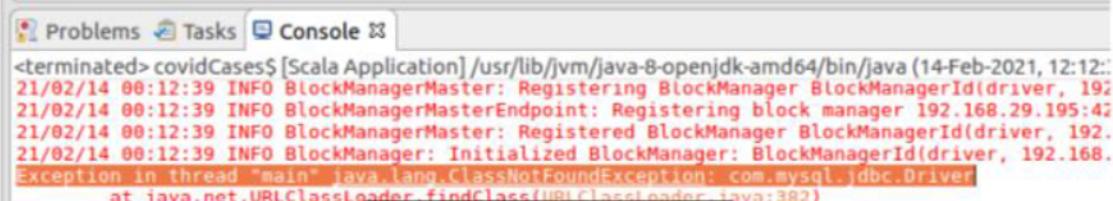
Ans: Add the below import and check
import org.apache.spark.sql.functions._

W12:19 Russel is trying to write data to the mysql table from Spark, but looks like I have a jar missing for jdbc driver and hence getting the error. Can anyone point me to the correct jar file to include in order

WEEK 12 FAQS

to connect to mysql from Spark ?

```
.getOrCreate(); // If available get it for us, if not available then create  
  
val url = "jdbc:mysql://localhost";  
val properties = new Properties()  
properties.put("user", "root")  
properties.put("password", "root")  
Class.forName("com.mysql.jdbc.Driver");  
  
val countriesDF = spark.read  
.format("csv")  
.option("header", "true")  
.option("inferSchema", "true")  
.option("path", "/home/rsi/BIGDATA/demo/countries-aggregated.csv")  
//.csv("/home/rsi/BIGDATA/demo/countries-aggregated.csv");  
.load();  
  
val table = "spark_practise.country";  
countriesDF.write.mode(SaveMode.Overwrite).jdbc(url, table, properties);  
  
//countriesDF.createOrReplaceTempView("countries");  
//val resultDF = spark.sql("select country, sum(Confirmed) as ConfirmedCount, sum(Recovered)  
//resultDF.show();  
// countriesDF.printSchema();  
  
scala.io.StdIn.readLine() //just to hold on to screen so we are able to see DAG  
spark.stop();  
}
```



Ans: download "mysql-connector-java-5.1.25-bin.jar" file from <https://jar-download.com/artifacts/mysql/mysql-connector-java/5.1.25/source-code> and importing it in eclipse.

W12:20 Rupa has created a sql table for csv file using ordersDF.createOrReplaceGlobalTempView("orders"). When querying on the table I'm getting an exception. Please find the screenshots attached for code and error.

```
val ordersDF = spark.read.format("csv").option("header", true).option("inferSchema", true).option("path", "C:/Users/varun/Downloads/orders-201019-002101.csv").load()  
ordersDF.createOrReplaceGlobalTempView("orders")  
// returns dataframe  
val resultDF = spark.sql("select order_status, count(*) as status_count from orders group by order_status order by status_count")  
resultDF.show
```

Ans: Use createOrReplaceTempView not global temp view

WEEK 12 FAQS

W12:21 when Raj applies the 'date_format' function on the 'datetime' column I'm getting null values. Please find the screenshots attached for code and output.

```
30  val MyList = List(
31    "WARN, 2016-12-31 04:19:32",
32    "FATAL, 2016-12-31 03:22:34",
33    "WARN, 2016-12-31 03:21:21",
34    "INFO, 2015-4-21 14:32:21",
35    "FATAL, 2015-4-21 19:23:20")
36
37  val rdd1 = spark.sparkContext.parallelize(MyList)
38  val rdd2 = rdd1.map(mapper)
39  val df1 = rdd2.toDF()
40
41  df1.createOrReplaceTempView("loggin_table")
42
43  spark.sql("select datetime as month from loggin_table").show
44  spark.sql("select date_format(datetime,'MMMM') as month from loggin_table").show
45 }
```

month
2016-12-31 04:19:32
2016-12-31 03:22:34
2016-12-31 03:21:21
2015-4-21 14:32:21
2015-4-21 19:23:20

month
null

Ans: `date_format(current_timestamp(),"yyyy MM dd E").as("date1")`
Try this , if it works that means our datetime column has some issue.

W12:22 Did anyone find FileA and FileB (cricket datasets) of the assignment problem3. I didn't find them

Ans: Create LIST of cricket datasets and then Use [.parallelize (mylist)] Then convert it to D.F.

WEEK 12 FAQS

W12:23 shumaila having error below, kindly help me know what She's doing wrong

```
import org.apache.spark.sql.Dataset
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._
import org.apache.spark.sql.Row
import org.apache.spark.sql.types.DateType

object SpkBigLogLes23 extends App {

    case class Logging(level:String, datetime: String)
    def mapper(line:String): Logging = {
        val fields = line.split(',')
        val logging:Logging = Logging(fields(0), fields(1))
        return logging
    }

    /** Our main function where the action happens*/
    def main(args: Array[String]) {

        //Setting the logging level
        Logger.getLogger("org").setLevel(Level.ERROR)

        //Set up spark conf
        val sparkConf = new SparkConf()
        sparkConf.set("spark.app.name", "my first application")
        sparkConf.set("spark.master", "local[2]")

        //Creating spark session
        val spark = SparkSession.builder()
            .appName("sparkSQL")
            .master("local[*]")
            .config(sparkConf)
            .getOrCreate()

        import spark.implicits._

        //Demo 1: Create a Scala list then create DataFrame with column names
        val mylist = List(
            "DEBUG,2015-2-6 16:24:07",
            "WARN,2016-7-26 18:54:43",
            "INFO,2012-10-18 14:35:19",
            "DEBUG,2012-4-26 14:26:50",
            "DEBUG,2013-9-28 20:27:13",
            "INFO,2017-8-20 13:17:27",
            "INFO,2015-4-13 09:28:17",
            "DEBUG,2015-7-17 00:49:27")

        val rdd1 = spark.sparkContext.parallelize(mylist) /* we are creating a base rdd using parallelize*/
        val rdd2 = rdd1.map(mapper) /* this mapper will be cast to the case class*/
        val df1 = rdd2.toDF()
        df1.show()

        //Logger.getLogger(getClass.getName).info("Job Completed Successfully")
        scala.io.StdIn.readLine()

        //Stopping the Spark session
        spark.stop()
    }
}
```

Ans: you are using "extends App" then no need to give "main method"

WEEK 12 FAQS

W12:24 Jack is using Spark version 2.1.0 and Scala version 2.12
When jack added a jar in intellij and He is trying to run the program, it throws an error and says "Unable to instantiate SparkSession with Hive support because Hive classes are not found." The JAR that i have added is "spark-hive_2.11-2.1.0.jar" since Jack don't see any version having Scala 2.12 and Spark 2.1 match

Ans: Spark version 2.1.0 is compatible with Scala version 2.11 or 2.10. download jar with proper version ,
<https://mvnrepository.com/artifact/org.apache.spark/spark-hive>

W12:25 Elon is getting errors in df2 like type annotation required unit definition.

```
//case class person(name: String, age: Int, city: String)
Logger.getLogger("org").setLevel(Level.ERROR)

def ageCheck(age: Int): String ={
  if (age > 18) "Y" else "N"
}

val sparkconf = new SparkConf()

sparkconf.set("spark.app.name", "Week 12 program")
sparkconf.set("spark.master", "local[*]")

val spark = SparkSession.builder().config(sparkconf).getOrCreate()

val input = spark.read
  .format("csv")
  .option("InferSchema", true)
  .option("path", "/Users/dheeraj.kachhap/Desktop/Big_Data/spark/datasets.txt")
  .load()

val df: Dataset[Row] = input.toDF("name", "age", "city") // we can give our own column name because inferSchema is false

//import spark.implicits._

//val ds = df.as[person] // converting dataframe to datasets using case class
//val df1 = ds.toDF() // converting dataset to dataframe again

val ageGreater = udf(ageCheck(_:Int): String) // Registering UDF
val df2 = df.withColumn("adult", ageGreater(col("age"))) // calling function with age column
```

Ans: Type annotation is not an error , as you are explicitly typing the type. I believe you won't get any runtime error once you ran this code If you really want to get rid of this message, You can give val df2:Dataset[Row] = df.withColumn("adult", ageGreater(col("age")))

WEEK 12 FAQS

W12:26 Please help to solve this

The screenshot shows a Java IDE interface with two tabs open. The top tab contains a Scala script named `DataFramesTable.scala`. The script sets up a SparkConf, creates a SparkSession, reads an orders CSV file into a DataFrame, writes it back to a CSV file, and then stops the spark session. The bottom tab is a terminal window titled "Console" showing the execution of the script. The terminal output shows various INFO logs from the Spark framework, including broadcast storage and task submission details. It ends with a fatal error from the `FileOutputCommitter` class, indicating a failure to create a file in HDFS at the path `/C:/Users/dell3207480/workspace/DataFrames/spark-warehouse/orders/_temporary/0/_temporary/_tmp_10000`.

```
val sparkConf = new SparkConf()
sparkConf.set("spark.app.name", "my first application")
sparkConf.set("spark.master", "local[2]")

val spark = SparkSession.builder()
.config(sparkConf)
.getOrCreate()

val ordersDf = spark.read
.format("csv")
.option("header",true)
.option("inferSchema",true)
.option("path","C:/Users/dell 320/Desktop/orders.csv") //output is always a directory
.load()

ordersDf.write
.format("csv")
.mode(SaveMode.Overwrite)
.saveAsTable("orders1") //by default it will create in default database

scala.io.StdIn.readLine()
spark.stop()
```

```
21/04/15 15:03:21 INFO DAGScheduler: Final stage: ResultStage 2 (saveAsTable at DataFramesTable.scala:27)
21/04/15 15:03:21 INFO DAGScheduler: Parents of final stage: List()
21/04/15 15:03:21 INFO DAGScheduler: Submitting ResultStage 2 (MapPartitionsRDD[10] at saveAsTable at DataFramesTable.scala:27), which has no parents
21/04/15 15:03:21 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 147.1 KB, free 884.4 MB)
21/04/15 15:03:21 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 53.2 KB, free 884.4 MB)
21/04/15 15:03:21 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on DESKTOP-9PU25BB:8085 (size: 53.2 KB, free: 885.4 MB)
21/04/15 15:03:21 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (MapPartitionsRDD[10] at saveAsTable at DataFramesTable.scala:27)
21/04/15 15:03:21 INFO TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
21/04/15 15:03:21 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 3, localhost, executor driver, partition 0, PROCESS_LOCAL, 8263 bytes)
21/04/15 15:03:21 INFO Executors: Running task 0.0 in stage 2.0 (TID 3)
21/04/15 15:03:21 INFO FileOutputCommitter: File Output Committer algorithm version is 1
21/04/15 15:03:21 INFO SQLHadoopMapReduceCommitProtocol: Using output committer class org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
21/04/15 15:03:21 ERROR Executor: Exception in task 0.0 in stage 2.0 (TID 3)
java.io.IOException: mkdirs failed to create file:/C:/Users/dell3207480/workspace/DataFrames/spark-warehouse/orders/_temporary/0/_temporary/_tmp_10000
    at org.apache.hadoop.fs.ChecksumFileSystem.create(ChecksumFileSystem.java:455)
    at org.apache.hadoop.fs.ChecksumFileSystem.create(ChecksumFileSystem.java:446)
    at org.apache.hadoop.fs.FsDataset.create(FsDataset.java:911)
    at org.apache.hadoop.fs.FsDataset.create(FsDataset.java:892)
    at org.apache.hadoop.fs.FsDataset.create(FsDataset.java:789)
    at org.apache.spark.sql.execution.datasources.CodecStreams$createOutputStream(CodecStreams.scala:81)
    at org.apache.spark.sql.execution.datasources.CodecStreams$createOutputStreamWriter(CodecStreams.scala:93)
    at org.apache.spark.sql.execution.datasources.CsvOutputWriter.(CsvFileFormat.scala:177)
    at org.apache.spark.sql.execution.datasources.CsvFileFormat$$anon$1.newInstance(CsvFileFormat.scala:85)
    at org.apache.spark.sql.execution.datasources.SingleDirectoryDataWriter.mroOutputWriter(FileFormatDataWriter.scala:120)
    at org.apache.spark.sql.execution.datasources.SingleDirectoryDataWriter.(FileFormatDataWriter.scala:100)
    at org.apache.spark.sql.execution.datasources.FileFormatWriter.org$apache$spark$sql$execution$datasources$FileFormatWriter$$executeTask()
    at org.apache.spark.sql.execution.datasources.FileFormatWriter$$anonfun$write$1.apply(FileFormatWriter.scala:179)
    at org.apache.spark.sql.execution.datasources.FileFormatWriter$$anonfun$write$1.apply(FileFormatWriter.scala:187)
    at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:98)
    at org.apache.spark.scheduler.Task.run(Task.scala:122)
    at org.apache.spark.executor.Executor$TaskRunner$$anonfun$10.apply(Executor.scala:406)
    at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1360)
    at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:101)
```

Ans: Have you set Hadoop_home environment variable?

WEEK 12 FAQS

W12:27 keep receiving Column and expr and not found error

The screenshot shows a Java IDE interface. The code editor contains Scala code for reading a CSV file and selecting specific columns. The terminal window below shows the command to run the application and its output, which includes several stack trace entries indicating that 'Column' and 'expr' are not found.

```
val sparkConf = new SparkConf()
sparkConf.set("spark.app.name", "testspark2 code")
sparkConf.set("spark.master", "local[2]")

val spark = SparkSession.builder()
.setConf(sparkConf)
.getOrCreate()

val ordersDF = spark.read
    .option("header", true)
    .option("inferSchema", true)
    .option("path", "/Users/karthy/Downloads/Bigdatafile12/orders-201025-223922.csv")
    .load

// column string
ordersDF.select("order_id", "order_date", "order_customer_id", "order_status").show

// column object
import spark.sql._

// ordersDF.select(column("order_id"), col("order_date"), column("order_customer_id"), column("order_status")).show

ordersDF.select(column("order_id"), col("order_date"), col("order_customer_id"), col("order_status")).show()
```

```
com.intellij.execution.ExecutionException: Error running class OrderProgram at C:\Program Files\Apache Software Foundation\Apache Spark\3.2.1-bin-hadoop3 (17-Apr-2021, 4:05:11 PM)
at OrderProgram.main(OrderProgram$1.run(OrderProgram.java:12)
at OrderProgram$1.run(OrderProgram.java:12)
at OrderProgram.main(OrderProgram.java:12)
```

Ans: import org.apache.spark.sql.functions.{ col, column, expr }

//OR

```
import org.apache.spark.sql.functions.col
import org.apache.spark.sql.functions.column
import org.apache.spark.sql.functions.expr
import org.apache.spark.sql.functions._
```

Remember this import , and then use ctrl+shift+O is working ;-

Import org.apache.spark.sql.functions._

WEEK 12 FAQS

W12:28 Frazz is unable to understand why this error is coming in spark-sql.

```
import org.apache.log4j.Level
import org.apache.log4j.Logger
import org.apache.spark.SparkConf
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._

@ object DataFrameExamples16 extends App{
    Logger.getLogger("org").setLevel(Level.ERROR)
    val sparkConf = new SparkConf()
        sparkConf.set("spark.app.name","simpleAggregation")
        sparkConf.set("spark.master","local[2]")
    val spark = SparkSession.builder()
        .config(sparkConf)
        .getOrCreate

    val input = spark.read
        .format("csv")
        .option("header",true)
        .option("inferSchema",true)
        .option("path","file:///E:/Trendytech/week-12/order_data.csv")
        .load

    input.createOrReplaceTempView("sales")
    val new1 = spark.sql("select count(*) as No_Of_rows ,sum(Quantity) as total_quantity ,avg(UnitPrice) ,"+
        "count(distinct(InvoiceNo)) as uniqueInvoice from sales")
    new1.show()

}

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Exception in thread "main" org.apache.spark.sql.catalyst.parser.ParseException:
extraneous input 'sales' expecting <EOF>(line 1, pos 128)

-- SQL --
select count(*) as No_Of_rows ,sum(Quantity) as total_quantity ,avg(UnitPrice) ,count(distinct(InvoiceNo)) as uniqueInvoice from sales
-- 
-- 
at org.apache.spark.sql.catalyst.parser.ParseException.withCommand(ParseDriver.scala:241)
at org.apache.spark.sql.catalyst.parser.AbstractSqlParser.parse(ParseDriver.scala:117)
at org.apache.spark.sql.execution.SparkSqlParser.parse(SparkSqlParser.scala:48)
at org.apache.spark.sql.catalyst.parser.AbstractSqlParser.parsePlan(ParseDriver.scala:69)
at org.apache.spark.sql.SparkSession.sql(SparkSession.scala:642)
at DataframeExamples16$.delayedEndpoint$main$1(DataframeExamples16.scala:25)
at DataframeExamples16$$anonfun$delayedInit$body$$anonfun$delayedInit$body$1(DataframeExamples16.scala:8)
at scala.Function0$class.apply$mcV$sp(Function0.scala:34)
at scala.runtime.AbstractFunction0.apply$mcV$sp(AbstractFunction0.scala:12)
at scala.App$$anonfun$main$1.apply(App.scala:76)
at scala.App$$anonfun$main$1.apply(App.scala:76)
at scala.collection.immutable.List.foreach(List.scala:392)
at scala.collection.generic.TraversableForwarder$class.foreach(TraversableForwarder.scala:35)
at scala.App$class.main(App.scala:76)
at DataframeExamples16$.main(DataframeExamples16.scala:8)
at DataframeExamples16.main(DataframeExamples16.scala)
```

Ans: Space is missing before from

Error window did not show space hence I recommended adding space.

W12:29 when to use createdataframe() and toDF() to create dataframes?

WEEK 12 FAQS

Ans: toDF() method, We don't have the control over schema customization. Used only for local testing.

createDataFrame() method, we have complete control over the schema customization. Used for local testing as well as in production.

W12:30 What's the use of implicits then when we are importing everything manually?even after removing implicits. The program worked fine

Ans: It used to convert rdd to DS or DF , Whenever scala finds wrong expression it tries to search for implicit. Implicit helps in converting rdd to DS or DF using encoders. <https://jaceklaskowski.gitbooks.io/mastering-spark-sql/content/spark-sql-SparkSession-implicits.html>

W12:31 Sachin is unable to understand how the below program is column Object expression (But here we don't use col or column etc.) Please explain ?

```
val groupDf = input.groupBy("Country","InvoiceNo")
    .agg(sum("Quantity").as("Total_quantity"),sum(expr("Quantity * UnitPrice")).as("invoice value"))
As we know these are used
["Column" , "Col" , "$"Country" , 'Invoice'] , for column Object expression. He is confused ?
```

Ans: There are two ways we could refer to a column in a data frame. 1. String 2. Object.. We use the String only if we need to get the result of the column, if we need to modify or write expression we refer it with object as the changes should be reflected in DF ,
orderDF.select(count("*").as("total_count"),sum("Quantity").as("avg_qty"),
 avg("UnitPrice").as("avg_unitPrice"),countDistinct("InvoiceNo").as("unique_Inv")).show

As you know the above is an expression containing aggregate functions like sum, avg and count. Here we pass column objects as reference for these functions to perform the operation. Hence we call this as column object expression in the above expression we were passing column values as object to functions.. But if we have to be referring any columns then as you said we would be using (column, col, \$, ')

W12:32 Could anyone help in understanding in rdd/df/ds in terms of memory. Consider If in my spark job I am having 5 rdd's/ DF's/ DS's.

WEEK 12 FAQS

Means Raja is storing data either in RDD or Data frames or Datasets. Once the action is triggered, it starts loading all the transformations. Raja understands at the stage level, the computed rdd's data is written into a buffer on disk and read on the next stage (Exchange) but, considering in my stage 1, Raja has computed 3 transformations , meaning 3 RDD's are created.

Raja Question is

Does the spark start garbage collection, immediately after the 2nd transformation is completed and free up the rdd1 memory?
If not then how does it handle memory, if total data size of 3 rdd's in stage 1 exceeds the allocated memory size?(Note: I am not using any caching mechanism)

Ans: Spark application has huge memory as it stores data in memory. GC works at regular intervals and clears memory in old generation(long lived objects) . Latest objects are stored in the younger generation (short lived objects). It does not depend on rdd1 completed or so. As per my understanding, The scenario you are sharing should cause an OOM out of memory exception. This issue is discussed in later weeks

Spark does not immediately free up rdd/df/ds once they are computed. Blocks are evicted from spark memory based on LRU (Least Recently Used) mechanism therefore, the blocks which are least used will be removed first. If you are using the 1st RDD to derive other RDDs then it will be in the memory within a stage

WEEK 12 FAQS

W12:33 JSON file creation giving below issue. Any clue?

The screenshot shows a Java IDE interface with a code editor and a terminal window. The code editor contains Scala code for writing data to Avro and JSON files. The terminal window shows the execution of the application and its termination, followed by a stack trace indicating an exception related to bucketing.

```
//===== AVRO file Format =====
// DF.write.format("avro").partitionBy("Country_Name").mode(SaveMode.Overwrite).option("path", "/D:/tst/Avro").save()

//===== JSON file Format =====
// DF.write.json("/D:/tst/Json")
DF.write.format("json").bucketBy(8, "Week_Number").mode(SaveMode.Overwrite).option("path", "/D:/tst/Json").save()

Logger.getLogger(getClass().getName()).info("My Application Completed Successfully !!")
spark.stop()

}
}

Console 22
<terminated> WnswrData$ [Scala Application] C:\Program Files\Java\jre1.8.0_211\bin\java.exe (13-May-2021, 12:32:57 AM)
Using Spark's default log4j profiles: org/apache/spark/log4j-defaults.properties
Exception in thread "main" org.apache.spark.sql.AnalysisException: 'save' does not support bucketBy right now;
    at org.apache.spark.sql.DataFrameWriter.assertNotBucketed(DataFrameWriter.scala:357)
    at org.apache.spark.sql.DataFrameWriter.save(DataFrameWriter.scala:243)
    at sparkDataFrame.WindowData$.delayedEndpoint$sparkDataFrame$WindowData$1(WindowData.scala:52)
    at sparkDataFrame.WindowData$delayedInit$body.apply(WindowData.scala:18)
    at scala.Function0$class.apply$mcV$sp(Function0.scala:34)
    at scala.runtime.AbstractFunction0.apply$mcV$sp(AbstractFunction0.scala:12)
    at scala.App$$anonfun$main$1.apply(App.scala:76)
    at scala.App$$anonfun$main$1.apply(App.scala:76)
    at scala.collection.immutable.List.foreach(List.scala:392)
    at scala.collection.generic.TraversableForwarder$class.foreach(TraversableForwarder.scala:35)
    at scala.App$class.main(App.scala:76)
    at sparkDataFrame.WindowData$.main(WindowData.scala:18)
    at sparkDataFrame.WindowData.main(WindowData.scala)
```

Ans: Buckets does not supported on save mode , we need to use a table to store using SPARK SQL

W12:34 There are 2 ways to read json/csv files : spark.read.json OR spark.read.format("json") [Similar for csv]. Is there any significant difference between the 2 approaches ?

Ans: Format is generalized way and json is specialized way

W12:35 Rupali has a dataset folder with 2 part files in it. While loading with Dataframe Reader API & when Spark config has spark.master as local[2], 2 tasks get created. While loading again when spark.master is set as local[*], 4 tasks get created. Shouldn't it have been 2 tasks irrespective of this?

Ans: It should be 2 tasks in parallel or 4 tasks in parallel based on the number of cores available. The number of parallel tasks in each executor depends on the number of cores. The number of tasks itself is dependent on the number of partitions within the RDD.

W12:36 Udit is running Spark on a single node (Windows PC) trying to load a CSV file , and using inferSchema option while loading the data to create an RDD. The file size while on Disk is 2 GB. But when Udit check the Job details, He is see the Input size to be 4 GB and 32 tasks processing around 128 MB partitions, what is causing this

WEEK 12 FAQS

difference in size as when loaded into Spark

Fraction ID		Address	Task Time	Total Tasks	Failed Tasks	Skipped Tasks	Succeeded Tasks	Input Size / Records		
driver		DepthLearn34875	0.1 min	32	0	0	32	4.8 GB / 317126		
- Tasks (32)										
Index	ID	Attempt	Status	Locality Level	Execution ID	Host	Launch Time	Duration	GC Time	Input Size / Records
0	1	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 984000
1	2	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 973155
2	3	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 988321
3	4	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 1011126
4	5	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 1023897
5	6	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 1034261
6	7	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 1038048
7	8	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 952012
8	9	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.2 s	128.1 MB / 1031600
9	10	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 991129
10	11	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 993189
11	12	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 952095
12	13	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 952048
13	14	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 991127
14	15	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 988311
15	16	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 984006
16	17	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 973136
17	18	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 999221
18	19	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 1011126
19	20	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 1023897
20	21	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 1034261
21	22	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 1044940
22	23	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 950313
23	24	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 1031688
24	25	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 991126
25	26	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 903116
26	27	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 952096
27	28	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2021/09/18 20:57:58	16 s	0.1 s	128.1 MB / 962016

Ans: When you process data , it's loaded in memory and is in deserialize form.

W12:37 I am trying to write the data to the target. but the output file is blank and Pinky is not getting any data in the output folder...

```

val orderInput=spark.read
    .format("csv")
    .option("header", true)
    .schema(schemaEval)
    // .option("inferSchema", true)
    .option("path", "C:/Users/Sunilkumar DV/Desktop/orders-20101"
    .load
    //orderInput.show()

orderInput.write
    .format("csv")
    .mode(SaveMode.Overwrite)
    .option("path", "C:/Users/Sunilkumar DV/Desktop/orders-output3")
    .save()

spark.stop()

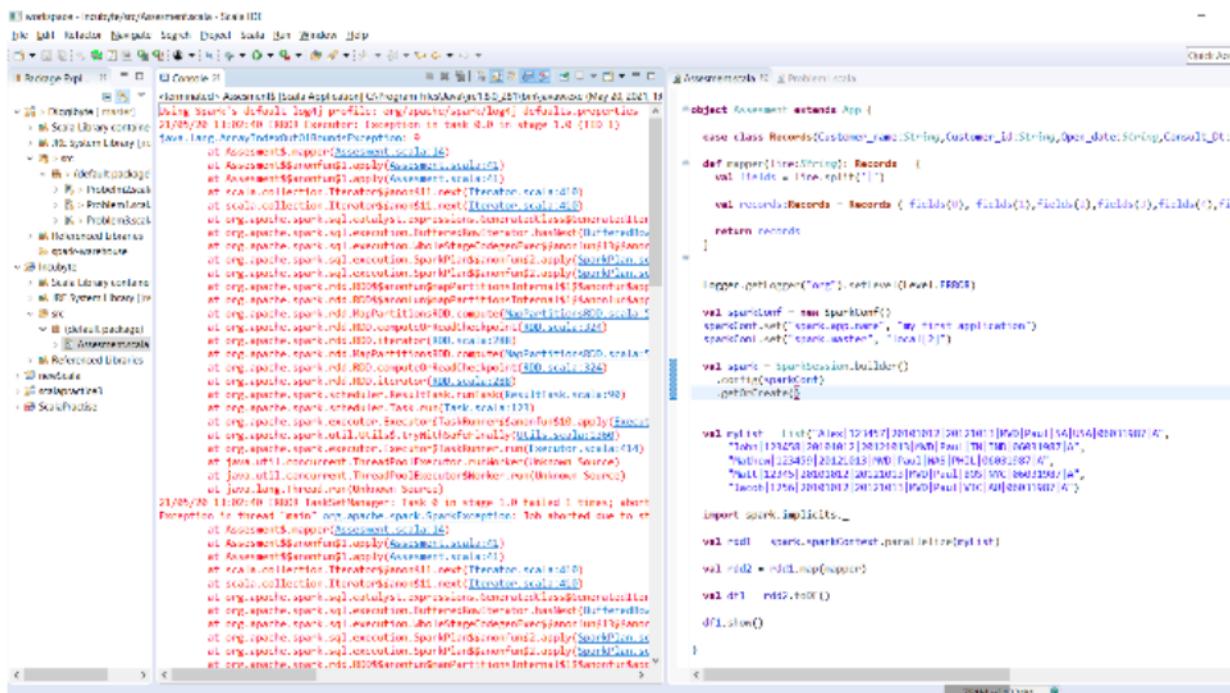
```

WEEK 12 FAQS

```
21/05/20 06:37:48 INFO FileOutputCommitter: File output committer algorithm ver
21/05/20 06:37:48 INFO SQLHadoopMapReduceCommitProtocol: Using output committer
21/05/20 06:37:48 INFO CodeGenerator: Code generated in 11.5769 ms
21/05/20 06:37:48 INFO CodeGenerator: Code generated in 11.3571 ms
21/05/20 06:37:48 INFO CodeGenerator: Code generated in 20.6048 ms
21/05/20 06:37:48 ERROR Utils: Aborting task
ExitCodeException exitCode=-1073741515:
    at org.apache.hadoop.util.Shell.runCommand(Shell.java:582)
    at org.apache.hadoop.util.Shell.run(Shell.java:479)
    at org.apache.hadoop.util.Shell$ShellCommandExecutor.execute(Shell.java)
    at org.apache.hadoop.util.Shell.execCommand(Shell.java:866)
    at org.apache.hadoop.util.Shell.execCommand(Shell.java:849)
```

Ans:<https://stackoverflow.com/questions/45947375/why-does-starting-a-streaming-query-lead-to-exitcodeexception-exitcode-1073741>
ON window-10, Install VC++ 2010 redistributable package

W12:38 Can anyone please help me with this error



Ans: This is an `ArrayIndexOutOfBoundsException`. It occurs when you are trying to access the index which does not exist in the array , array index starts from 0, cross check the number of values in your dataset and index referenced in code.

W12:39 Mkdir failed to create error in Windows.

WEEK 12 FAQS

```
C:\Program Files\Java\jre1.8.0_291\bin\javaw.exe (28/05/2021, 11:08:40 अपराह्न)
file: org/apache/spark/log4j-defaults.properties
mitter: Mkdirs failed to create file:/C:/Users/karamjeet%20kaur/workspace/scala/spark-warehouse/orders/_tempor
ception in task 0.0 in stage 1.0 (TID 1)
ed to create file:/C:/Users/karamjeet%20kaur/workspace/scala/spark-warehouse/orders/_tempor
ChecksumFileSystem.create(ChecksumFileSystem.java:455)
ChecksumFileSystem.create(ChecksumFileSystem.java:440)
.FileSystem.create(FileSystem.java:911)
.FileSystem.create(FileSystem.java:892)
        ...
```

Ans : This is because you have given the Profile user name with space(karamjeet<space>kaur). Move your scala IDE workspace path to C:\workspace.

NOTE: Files created path should always be without space so that you won't get error. that is why I have given the simple path as direct C drive TO change the workspace path:

file-->switch workspace--->other-->browse-->give C:\ drive now you can create new projects and run it or you import projects from previous workspace.