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_ColmnString_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

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]

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+)""".r

    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("saprk.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()

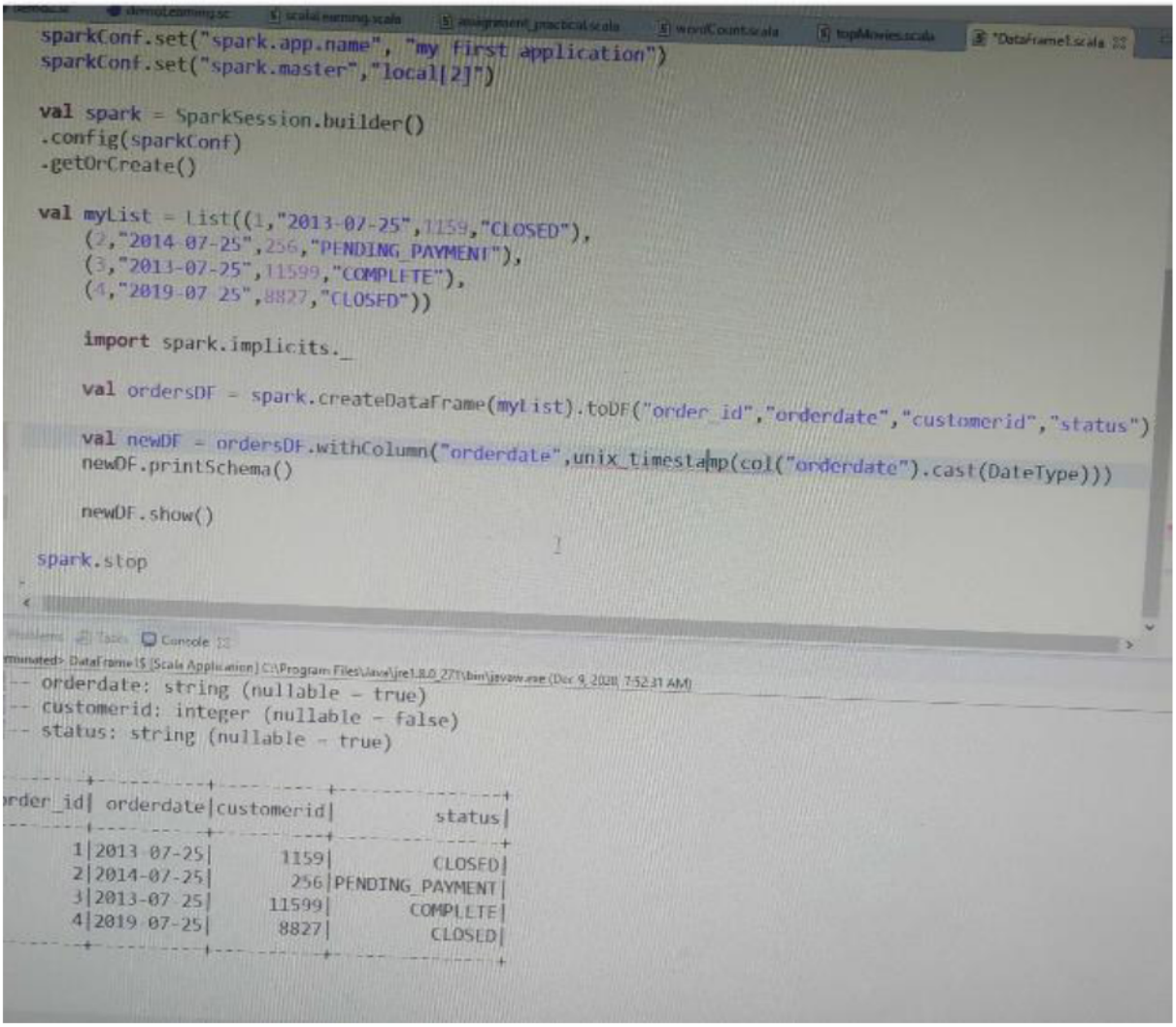
Ans: This is now resolve, Windows required MSVCR100.dll file to be installed follow this link to install it
https://www.microsoft.com/en-hk/download/confirmation.aspx?id=13523

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 FileOutputFormat and use one of Spark's save methods that accept a FileOutputFormat 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

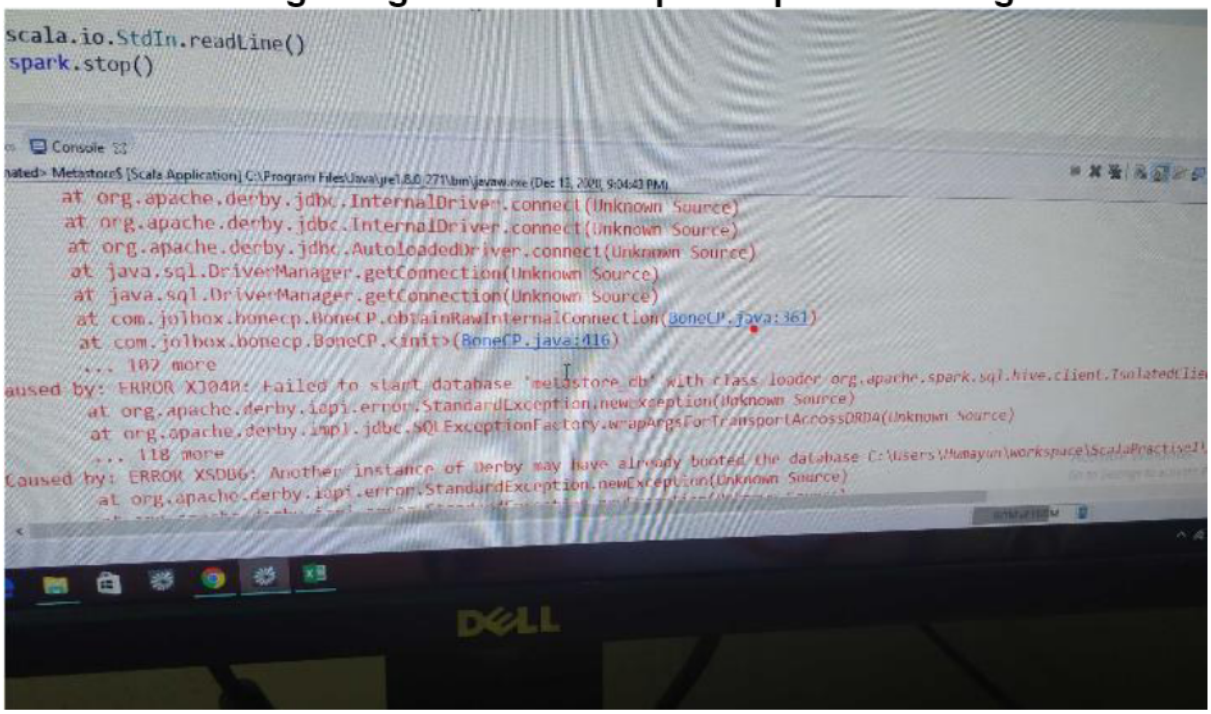


Ans: import org.apache.spark.sql._

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



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

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>)

```
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 ?

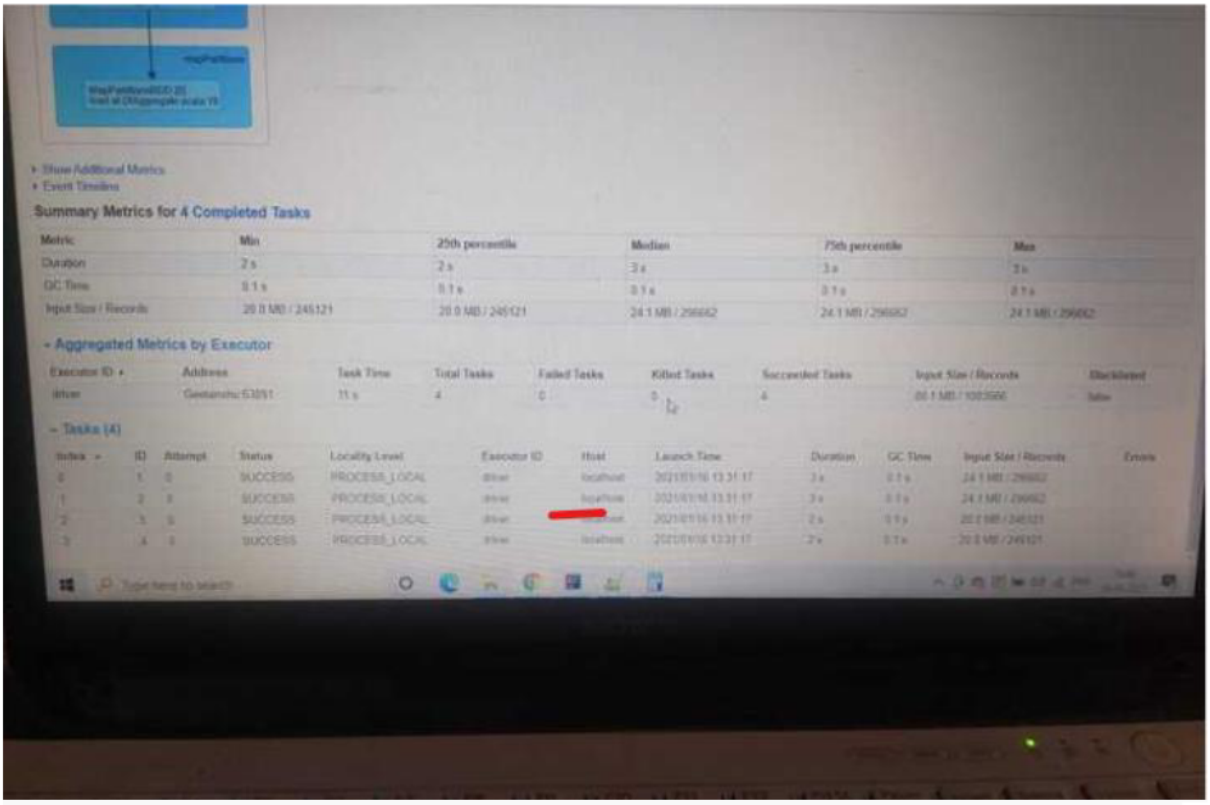
Ans: By default, spark sql uses the embedded Derby database to store metadata.
The default embedded Derby deployment mode is not recommended for production environments due to limitation of only one active SparkSession at a time.
So we use enableHiveSupport() method to store metadata in Hive Metastore.
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

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?



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.

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.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
```

```
.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.

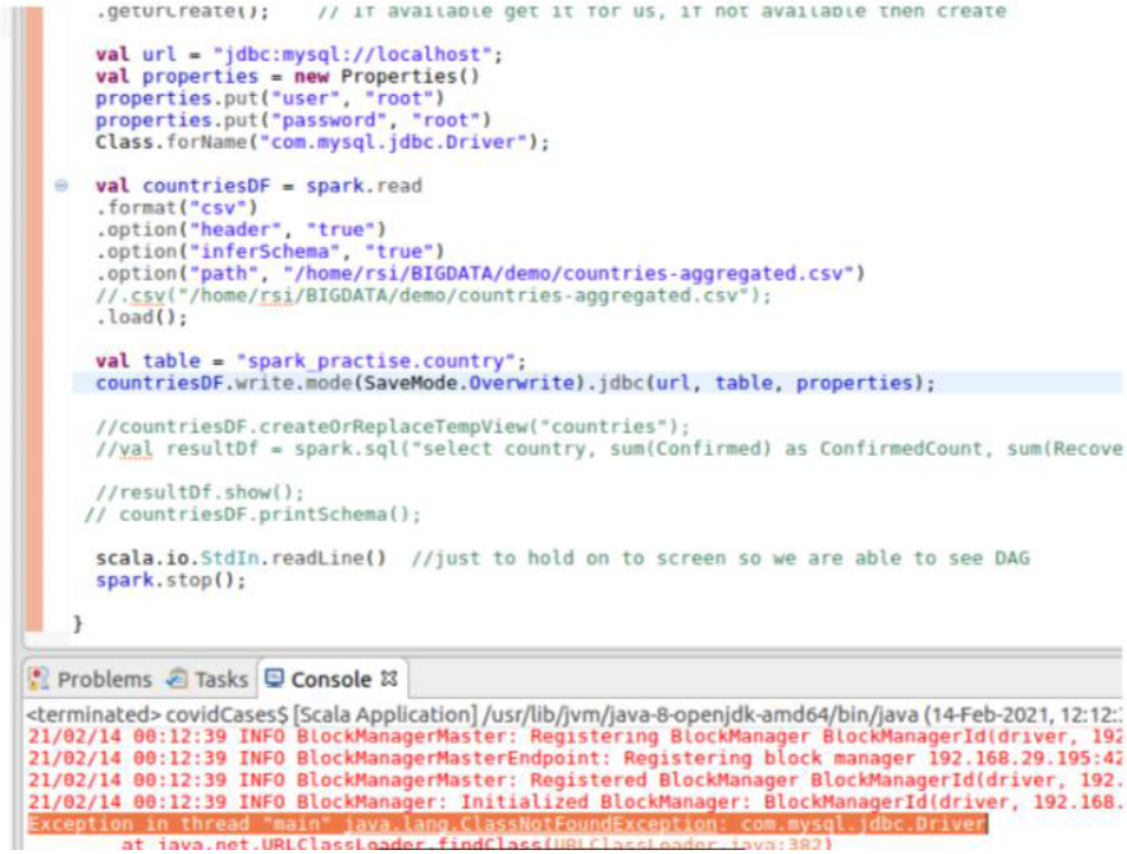
W12:18 Saame is getting column and col not found while using column Object.

```
//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._

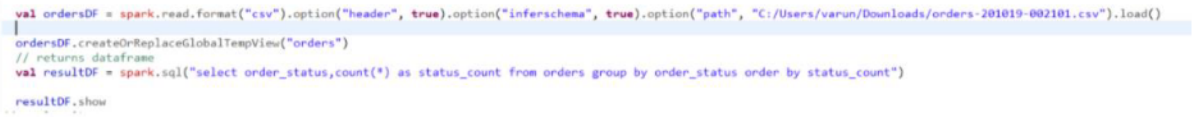
W12:19 Russel is trying to write data to the mysql table from Spark, but looks like I have a jar missing for jdbc driver and hence getting the error.Can anyone point me to the correct jar file to include in order

to connect to mysql from Spark ?



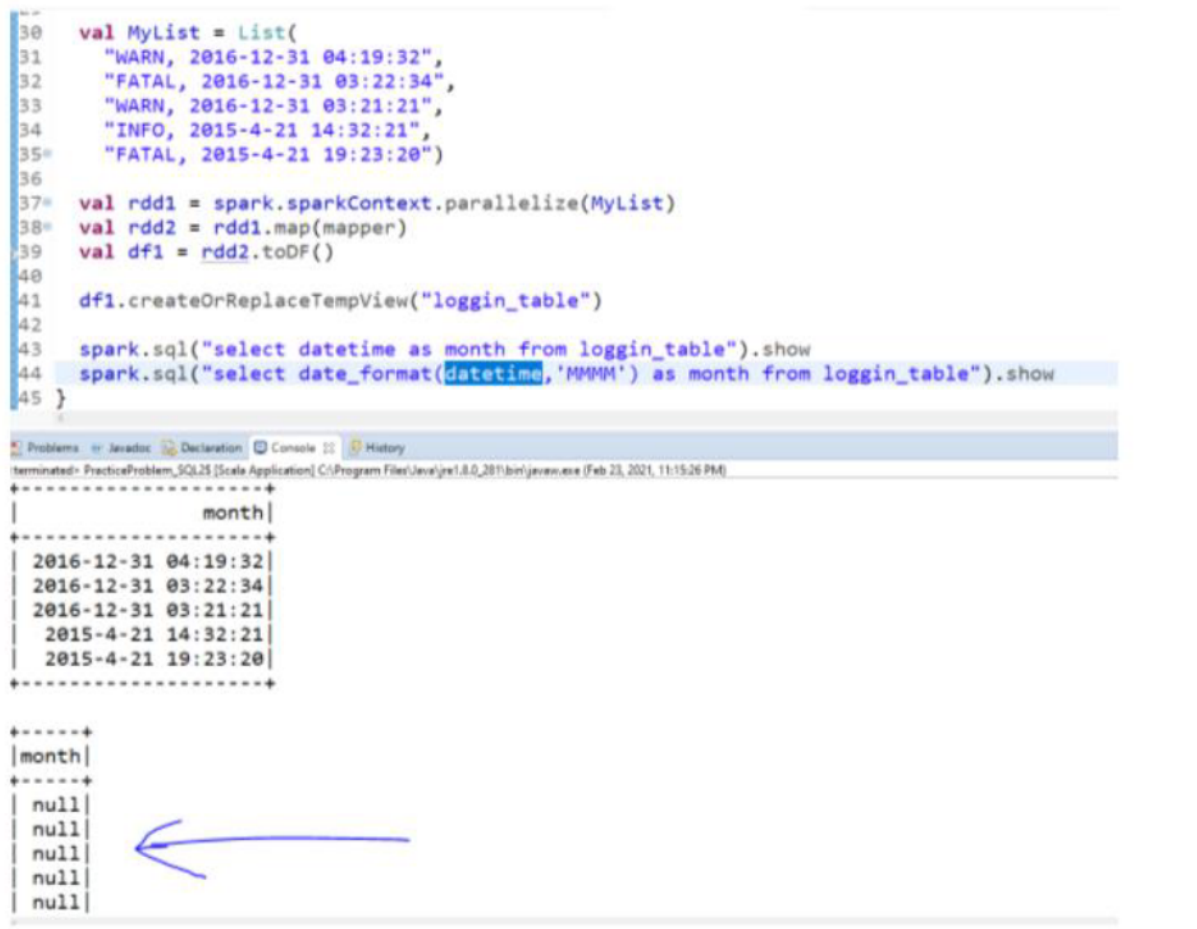
Ans: download "mysql-connector-java-5.1.25-bin.jar" file from https://jar-download.com/artifacts/mysql/mysql-connector-java/5.1.25/source-code and importing it in eclipse.

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.



Ans: Use creatOrReplaceTempView not global temp view

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|
| null|
| null|
| null|
| 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.

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"

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

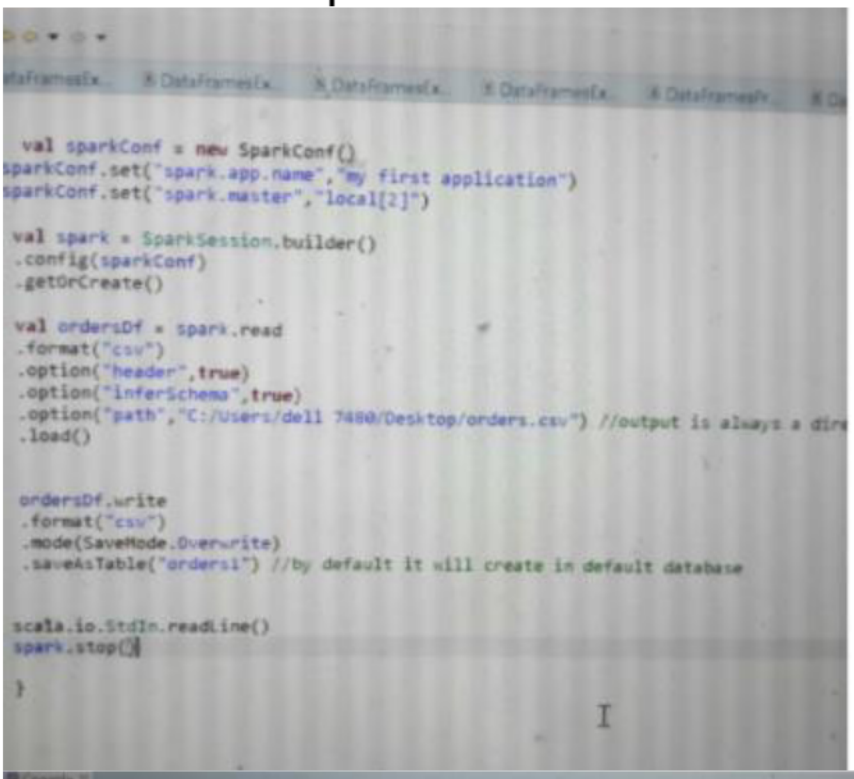
Ans: Spark version 2.1.0 is compatible with Scala version 2.11 or 2.10. download jar with proper version , https://mvnrepository.com/artifact/org.apache.spark/spark-hive

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

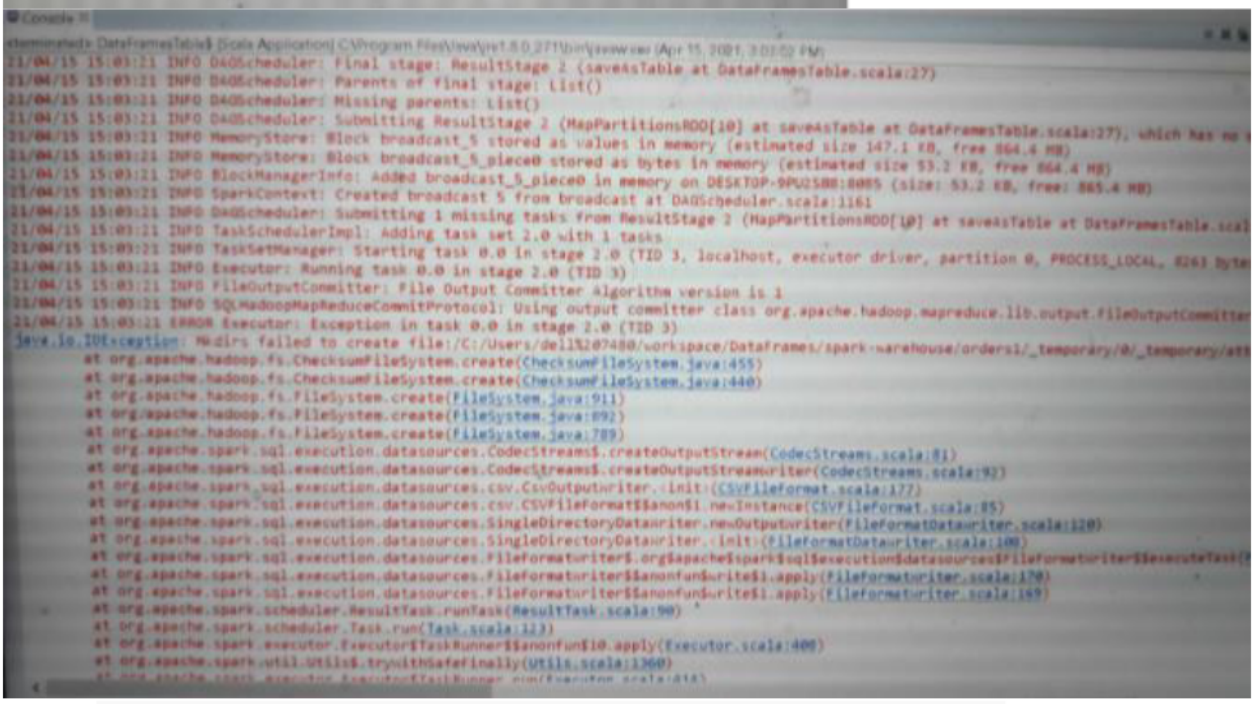
```
//case class person(name: String, age: Int, city: String)

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

//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")))

W12:26 Please help to solve this



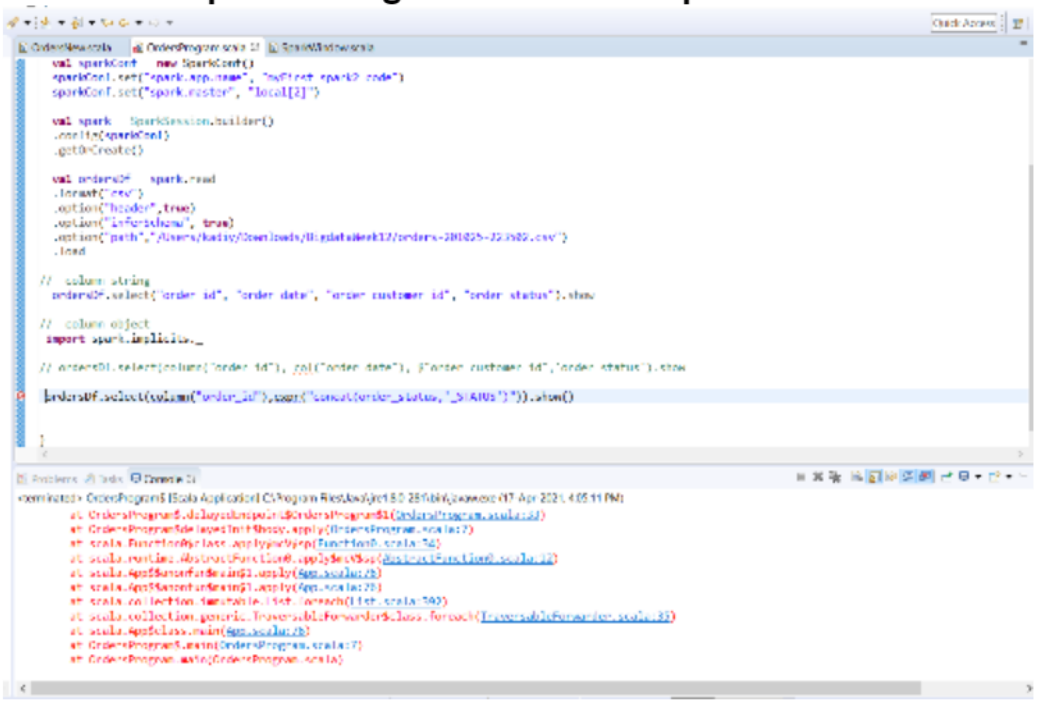
```
val sparkConf = new SparkConf()
sparkConf.set("spark.app.name","my first application")
sparkConf.set("spark.master","local[2]")

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

val ordersDf = spark.read
.format("csv")
.option("header",true)
.option("inferSchema",true)
.option("path","C:/Users/dell 7480/Desktop/orders.csv") //output is always a dir
.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: Missing parents: List()
21/04/15 15:03:21 INFO DAGScheduler: Submitting ResultStage 2 (MapPartitionsRDD[10] at saveAsTable at DataFramesTable.scala:27), which has no
21/04/15 15:03:21 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 147.1 KB, free 864.4 MB)
21/04/15 15:03:21 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 53.2 KB, free 864.4 MB)
21/04/15 15:03:21 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on DESKTOP-9PU2588:8085 (size: 53.2 KB, free: 865.4 MB)
21/04/15 15:03:21 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:1161
21/04/15 15:03:21 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (MapPartitionsRDD[10] at saveAsTable at DataFramesTable.scal
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 byte
21/04/15 15:03:21 INFO Executor: 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/dell%207480/workspace/DataFrames/spark-warehouse/orders1/_temporary/0/_temporary/att
        at org.apache.hadoop.fs.ChecksumFileSystem.create(ChecksumFileSystem.java:455)
        at org.apache.hadoop.fs.ChecksumFileSystem.create(ChecksumFileSystem.java:440)
        at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:911)
        at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:892)
        at org.apache.hadoop.fs.FileSystem.create(FileSystem.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:92)
        at org.apache.spark.sql.execution.datasources.csv.CsvOutputWriter.<init>(CSVFileFormat.scala:177)
        at org.apache.spark.sql.execution.datasources.csv.CSVFileFormat$$anon$1.newInstance(CSVFileFormat.scala:85)
        at org.apache.spark.sql.execution.datasources.SingleDirectoryDataWriter.newOutputWriter(FileFormatDataWriter.scala:120)
        at org.apache.spark.sql.execution.datasources.SingleDirectoryDataWriter.<init>(FileFormatDataWriter.scala:108)
        at org.apache.spark.sql.execution.datasources.FileFormatWriter$.org$apache$spark$sql$execution$datasources$FileFormatWriter$$executeTask(
        at org.apache.spark.sql.execution.datasources.FileFormatWriter$$anonfun$write$1.apply(FileFormatWriter.scala:170)
        at org.apache.spark.sql.execution.datasources.FileFormatWriter$$anonfun$write$1.apply(FileFormatWriter.scala:169)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:123)
        at org.apache.spark.executor.Executor$TaskRunner$$anonfun$10.apply(Executor.scala:408)
        at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1360)
```

Ans: Have you set Hadoop_home environment variable?

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



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._
```

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 uniqueInvoicefrom 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$.delayedEndpoint$DataFrameExamples16$1(DataFrameExamples16.scala:25)
        at DataFrameExamples16$delayedInit$body.apply(DataFrameExamples16.scala:8)
        at scala.Function0$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?

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.

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.

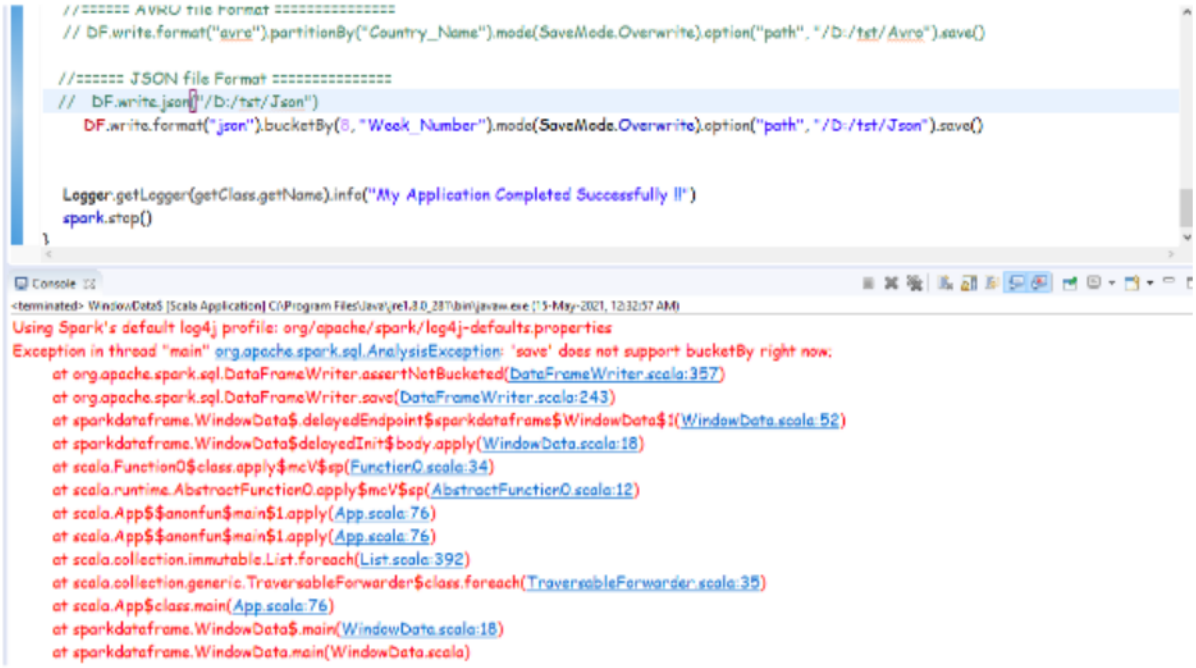
Raja Question is

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

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



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

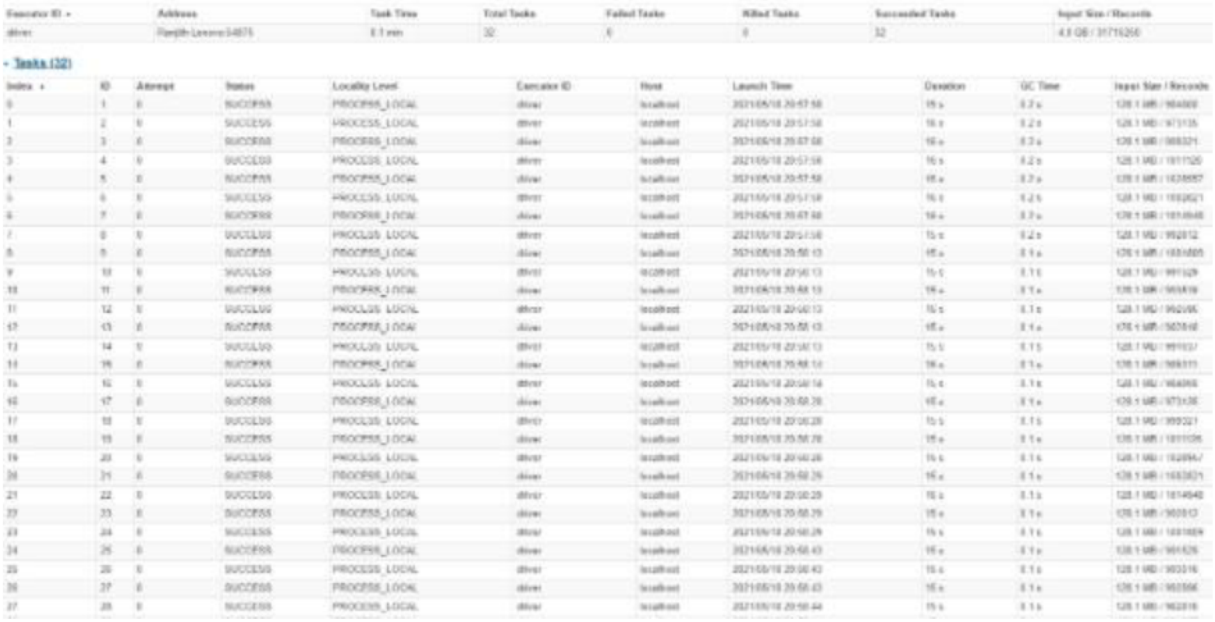
W12:34 There are 2 ways to read json/csv files : spark.read.json OR spark.read.format("json") [Similar for csv]. Is there any significant difference between the 2 approaches ?

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

difference in size as when loaded into Spark



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...

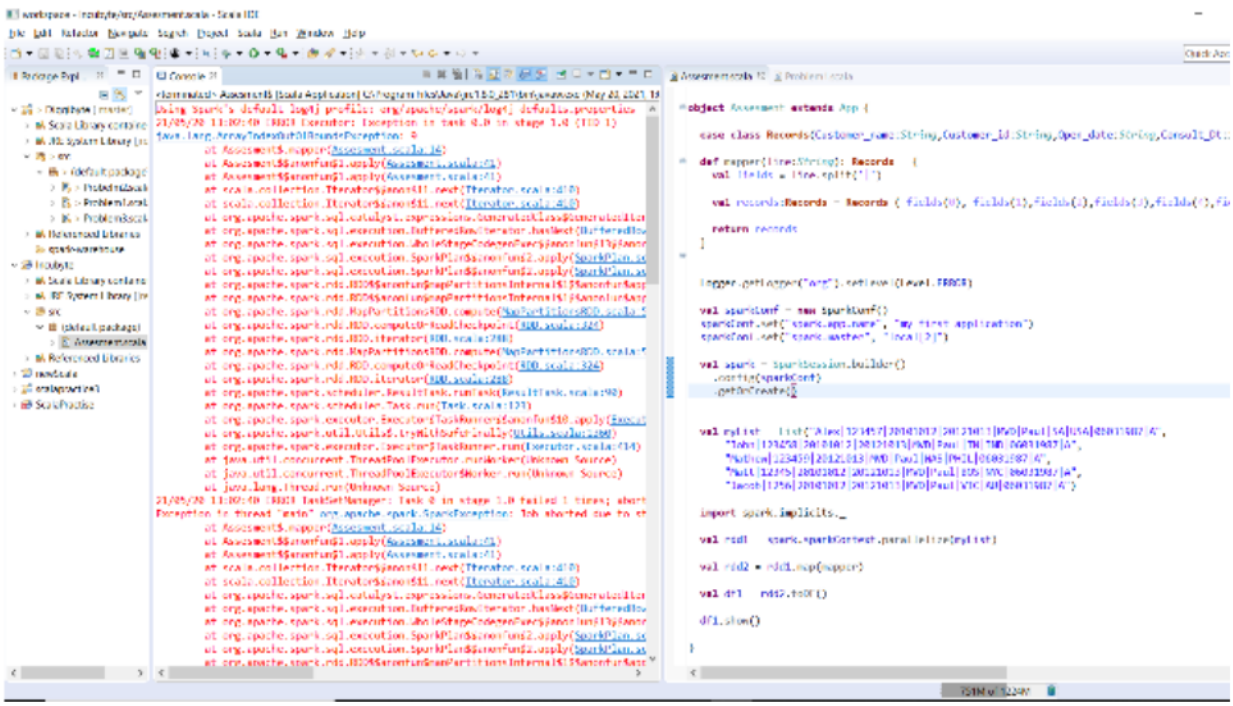
```
val orderInput=spark.read
              .format("csv")
              .option("header", true)
              .schema(schemaEval)
              //.option("inferSchema", true)
              .option("path", "C:/Users/Sunilkumar DV/Desktop/orders-20101
              .load
        //orderInput.show()

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

spark.stop()
```

```
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

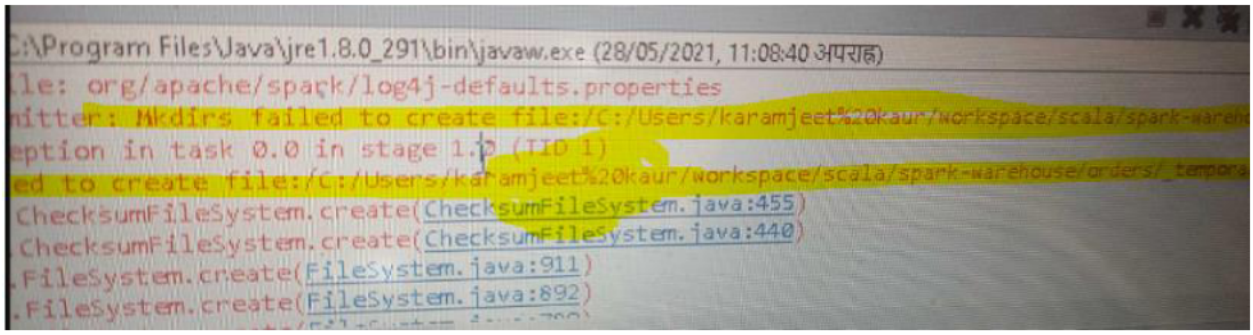
ON window-10, Install VC++ 2010 redistributable package

W12:38 Can anyone please help me with this error



Ans: This is an ArrayIndexOutOfBounds error. 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.



Ans : This is because you have given the Profile user name with space(karamjeet<space>kaur). Move you 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 work space.