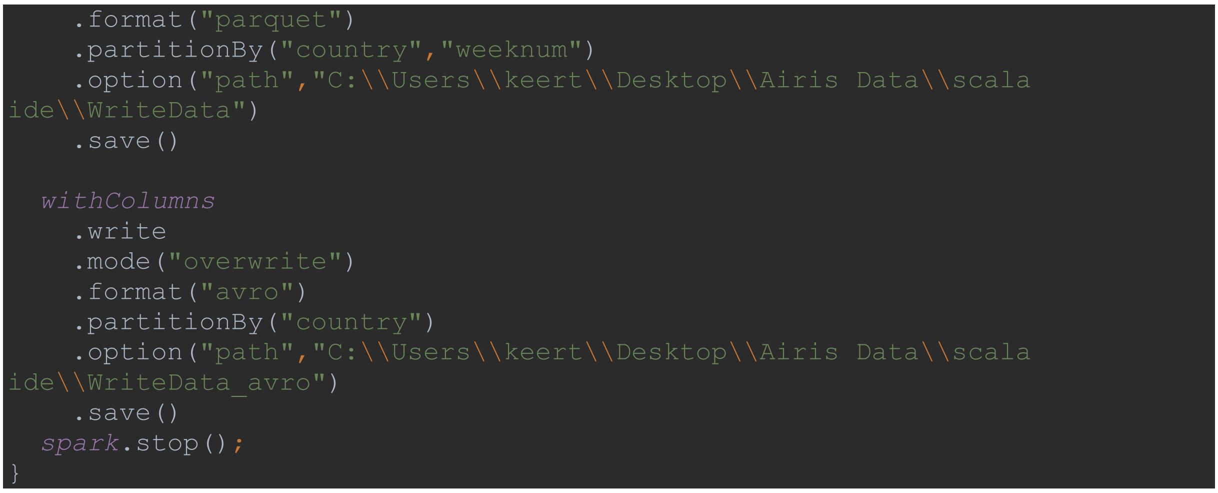
ASSIGNMENT:

We have a file windowdata.csv and the field names are country, weeknum, numinvoices, totalquantity, invoicevalue

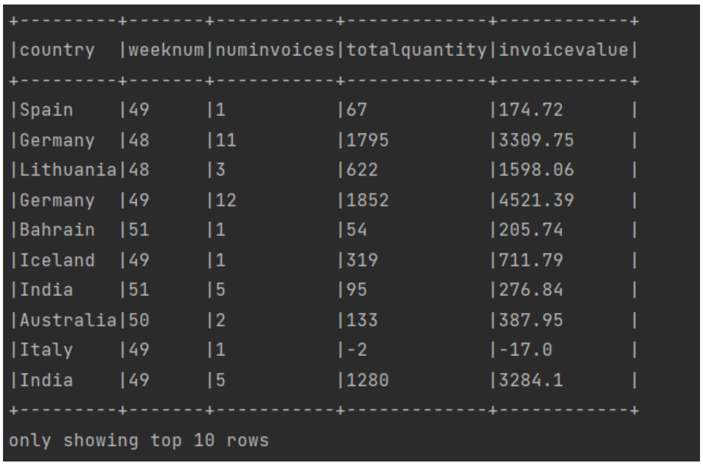
Step 1: create spark session
Step 2: set the logging level to error
Step 3:   Using the standard dataframe reader API load the file and create a dataframe.
Step 4:   Use the standard dataframe writer api to save it in parquet format. While saving make sure data is stored where we should have a folder for each country, weeknum (combination)
Step 5: Also use the dataframe write api to save the data in Avro format. While saving make sure data is stored where we should have a folder for each country.
Step 6: Apply header
Step 7: Convert dataframe to dataset(Specific type)

```
import org.apache.log4j.{Level, Logger}
import org.apache.spark.sql.{SaveMode, SparkSession}

object Assignment extends App{
  Logger.getLogger("org").setLevel(Level.ERROR)
  case class
Customer(country:String,weeknum:Int,numinvoices:Int,totalquantity:Int,invoi
cevalue:Double)
  val spark = SparkSession
    .builder.appName("Window Data")
    .master("local[*]")
    .getOrCreate()

  val df = spark
    .read
    .option("inferSchema","true")
    .csv("C:\\Users\\keert\\Desktop\\Airis Data\\scala
ide\\windowdata.csv")

  val withColumns = df.toDF("country", "weeknum", "numinvoices",
"totalquantity", "invoicevalue");

  withColumns.show(10,false);
  // Reading the data with headers
  val dfHeader = spark
    .read
    .option("header","true")
    .option("inferSchema","true")
    .csv("C:\\Users\\keert\\Desktop\\Airis Data\\scala
ide\\windowdata_header.csv")

  import spark.implicits._
  val datasetHeader = dfHeader.as[Customer]

  datasetHeader.show(10,false);

  //country, week num combination
  withColumns
    .write
    .mode("overwrite")
```

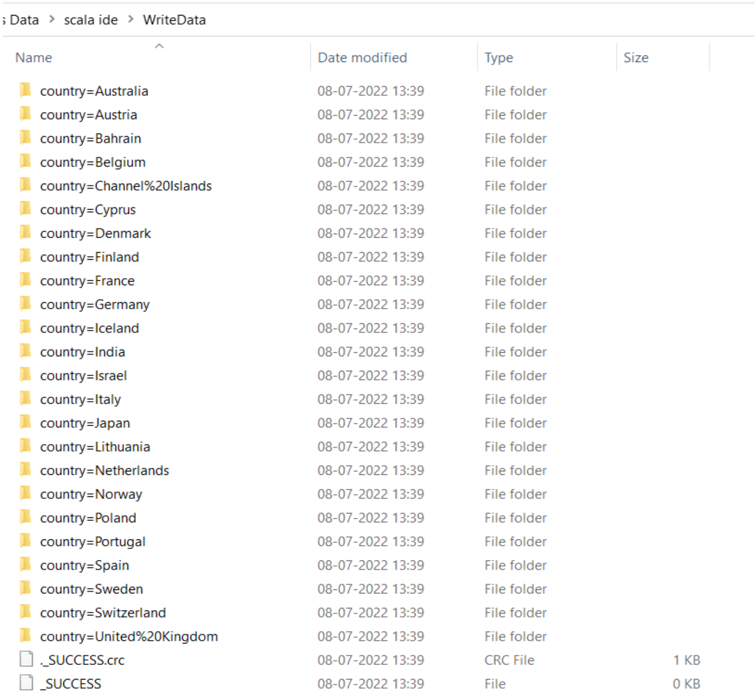


```
    .format("parquet")
    .partitionBy("country","weeknum")
    .option("path","C:\\Users\\keert\\Desktop\\Airis Data\\scala
ide\\WriteData")
    .save()

  withColumns
    .write
    .mode("overwrite")
    .format("avro")
    .partitionBy("country")
    .option("path","C:\\Users\\keert\\Desktop\\Airis Data\\scala
ide\\WriteData_avro")
    .save()
  spark.stop();
}
```

OUTPUT:

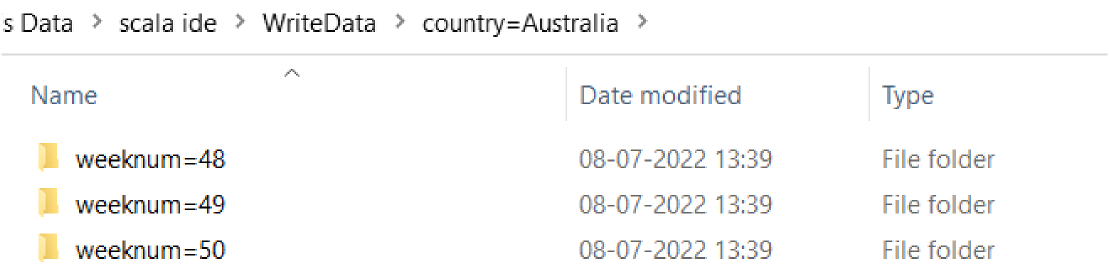


```
+---------+-------+-----------+-------------+------------+
|country  |weeknum|numinvoices|totalquantity|invoicevalue|
+---------+-------+-----------+-------------+------------+
|Spain    |49     |1          |67           |174.72      |
|Germany  |48     |11         |1795         |3309.75     |
|Lithuania|48     |3          |622          |1598.06     |
|Germany  |49     |12         |1852         |4521.39     |
|Bahrain  |51     |1          |54           |205.74      |
|Iceland  |49     |1          |319          |711.79      |
|India    |51     |5          |95           |276.84      |
|Australia|50     |2          |133          |387.95      |
|Italy    |49     |1          |-2           |-17.0       |
|India    |49     |5          |1280         |3284.1      |
+---------+-------+-----------+-------------+------------+
only showing top 10 rows
```



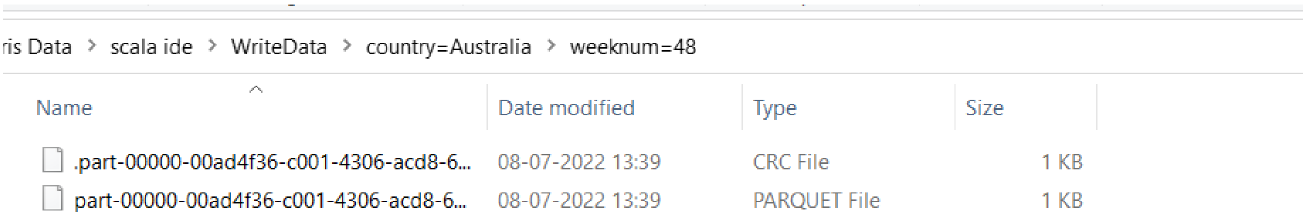
; Data > scala ide > WriteData

Name	Date modified	Type	Size
country=Australia	08-07-2022 13:39	File folder	
country=Austria	08-07-2022 13:39	File folder	
country=Bahrain	08-07-2022 13:39	File folder	
country=Belgium	08-07-2022 13:39	File folder	
country=Channel%20Islands	08-07-2022 13:39	File folder	
country=Cyprus	08-07-2022 13:39	File folder	
country=Denmark	08-07-2022 13:39	File folder	
country=Finland	08-07-2022 13:39	File folder	
country=France	08-07-2022 13:39	File folder	
country=Germany	08-07-2022 13:39	File folder	
country=Iceland	08-07-2022 13:39	File folder	
country=India	08-07-2022 13:39	File folder	
country=Israel	08-07-2022 13:39	File folder	
country=Italy	08-07-2022 13:39	File folder	
country=Japan	08-07-2022 13:39	File folder	
country=Lithuania	08-07-2022 13:39	File folder	
country=Netherlands	08-07-2022 13:39	File folder	
country=Norway	08-07-2022 13:39	File folder	
country=Poland	08-07-2022 13:39	File folder	
country=Portugal	08-07-2022 13:39	File folder	
country=Spain	08-07-2022 13:39	File folder	
country=Sweden	08-07-2022 13:39	File folder	
country=Switzerland	08-07-2022 13:39	File folder	
country=United%20Kingdom	08-07-2022 13:39	File folder	
._SUCCESS.crc	08-07-2022 13:39	CRC File	1 KB
_SUCCESS	08-07-2022 13:39	File	0 KB



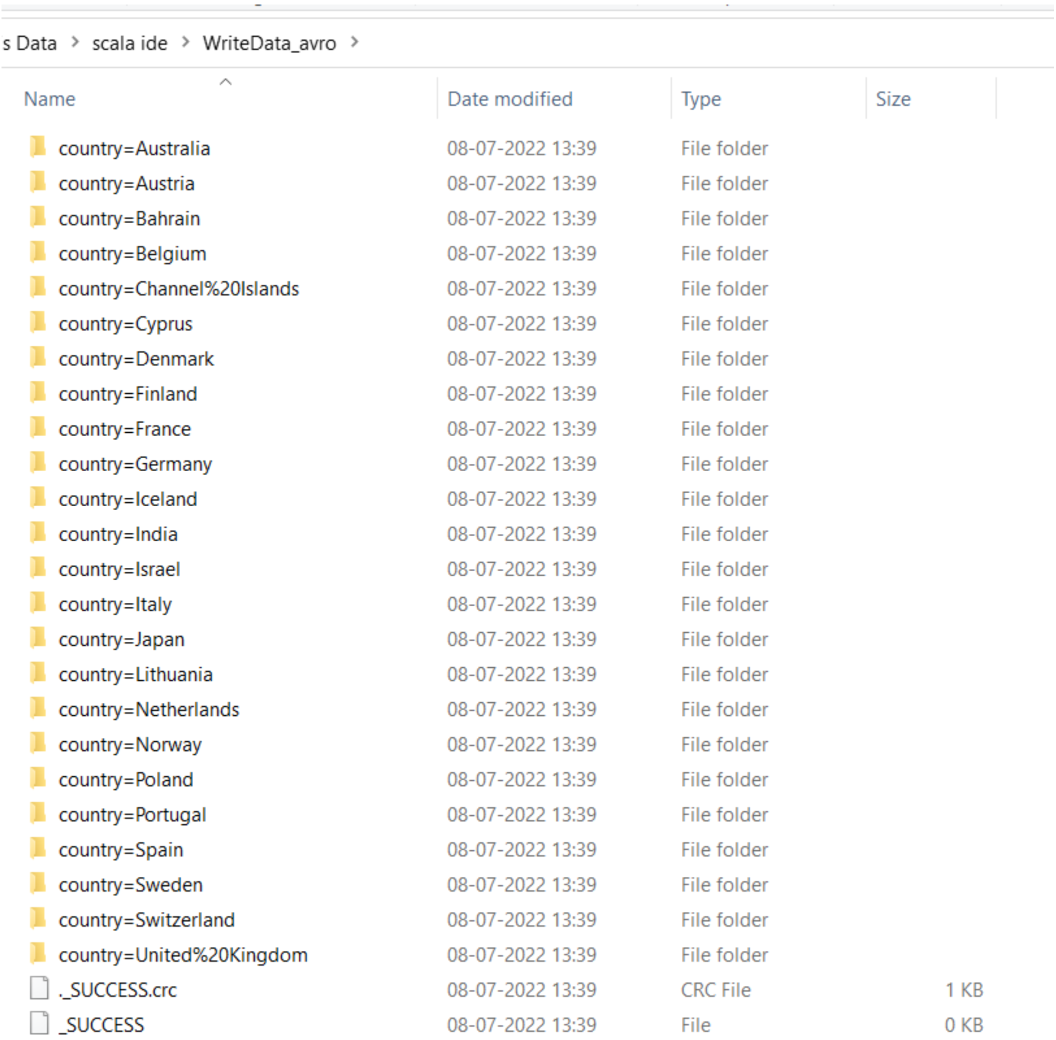
s Data › scala ide › WriteData › country=Australia ›

Name	Date modified	Type
weeknum=48	08-07-2022 13:39	File folder
weeknum=49	08-07-2022 13:39	File folder
weeknum=50	08-07-2022 13:39	File folder



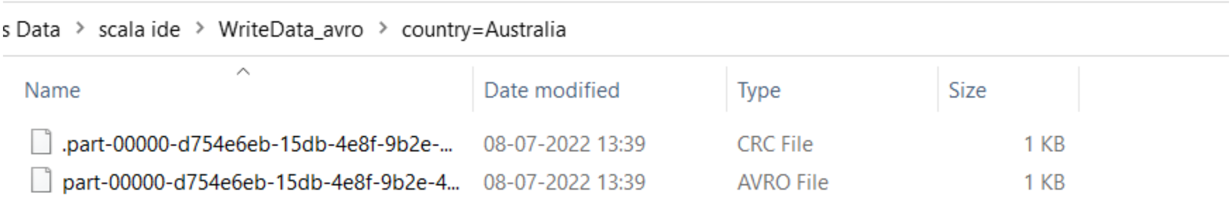
ris Data › scala ide › WriteData › country=Australia › weeknum=48

Name	Date modified	Type	Size
.part-00000-00ad4f36-c001-4306-acd8-6...	08-07-2022 13:39	CRC File	1 KB
part-00000-00ad4f36-c001-4306-acd8-6...	08-07-2022 13:39	PARQUET File	1 KB



s Data › scala ide › WriteData_avro ›

Name	Date modified	Type	Size
country=Australia	08-07-2022 13:39	File folder	
country=Austria	08-07-2022 13:39	File folder	
country=Bahrain	08-07-2022 13:39	File folder	
country=Belgium	08-07-2022 13:39	File folder	
country=Channel%20Islands	08-07-2022 13:39	File folder	
country=Cyprus	08-07-2022 13:39	File folder	
country=Denmark	08-07-2022 13:39	File folder	
country=Finland	08-07-2022 13:39	File folder	
country=France	08-07-2022 13:39	File folder	
country=Germany	08-07-2022 13:39	File folder	
country=Iceland	08-07-2022 13:39	File folder	
country=India	08-07-2022 13:39	File folder	
country=Israel	08-07-2022 13:39	File folder	
country=Italy	08-07-2022 13:39	File folder	
country=Japan	08-07-2022 13:39	File folder	
country=Lithuania	08-07-2022 13:39	File folder	
country=Netherlands	08-07-2022 13:39	File folder	
country=Norway	08-07-2022 13:39	File folder	
country=Poland	08-07-2022 13:39	File folder	
country=Portugal	08-07-2022 13:39	File folder	
country=Spain	08-07-2022 13:39	File folder	
country=Sweden	08-07-2022 13:39	File folder	
country=Switzerland	08-07-2022 13:39	File folder	
country=United%20Kingdom	08-07-2022 13:39	File folder	
._SUCCESS.crc	08-07-2022 13:39	CRC File	1 KB
_SUCCESS	08-07-2022 13:39	File	0 KB



s Data › scala ide › WriteData_avro › country=Australia

Name	Date modified	Type	Size
.part-00000-d754e6eb-15db-4e8f-9b2e-...	08-07-2022 13:39	CRC File	1 KB
part-00000-d754e6eb-15db-4e8f-9b2e-4...	08-07-2022 13:39	AVRO File	1 KB