



# Assignment

**Week11: Apache Spark - Structured API**  
**Part-1**

# IMPORTANT

## Self-assessment enables students to develop:

1. A sense of responsibility for their own learning and the ability & desire to continue learning,
2. Self-knowledge & capacity to assess their own performance critically & accurately, and
3. An understanding of how to apply their knowledge and abilities in different contexts.

All assignments are for self assessment. Solutions will be released on every subsequent week. Once the solution is out, evaluate yourself.

No discussions/queries allowed on assignment questions in slack channel.

**Note:** You can raise your doubts once the solution is released

## Problem 1:

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.

### Note:

*The schema should be provided using StructType (do not use infer schema)*

**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.



TRENDYTECH 9108179578

## Problem 2:

**step 1:** Create spark session

**step 2:** Set the logging level to error

**step 3:** 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.

*UPLIFT YOUR CAREER!*



**5** Star Google Rated  
Big Data Course

**LEARN FROM THE EXPERT**



**9108179578**

**Call for more details**