**Data Engineering Assessment’s Solution**

***Introduction***

* Based on the assessment description, I understand that given the files provided stores in different folders in different file formats (CVS, JSON, and AVRO). I am tasked with combining the files and create a usable dataset in which I can then perform analyses to answer the questions below.
* But most importantly, I need to consider the fact that I am working for a customer that is primarily used to working with low code tools. As a result, I will solve the assignment using low code tool, namely Azure Data Factory.

***Terminologies***

* JSON file format are files that stores key-value pairs.
* AVRO:
  + is not a file format but it is a framework for serializing data. Where serialization mean converting the data into binary format so that the data is machine readable.
  + In AVRO, there is a collection of header and blocks.
  + Where header contains:
    - a title, which a four ASCII characters (O, b, j, 1)
    - file metadata scheme in JSON format
    - specification that compression will be done by codec
    - sync maker, which helps facilitate file splitting for MapReduce
  + A block contains:
    - Count of objects
    - Block size (byte) of serialized objects
    - Serialized data (compressed or not)
    - Sync maker

***Workflow Steps using ADF***

* *Ingestion:* I will use ADF to ingest the relevant files from their respective locations.
* *Transformation:* I will use ADF to create data flows which allow me to perform data transformation processes in a graphical manner.
  + Notably, with the data flows, I will remove duplicates based on unique Order IDs,
  + handle JSON parsing, especially for nested JSON columns, using functions like OPENJSON,
  + and convert AVRO and JSON formats into a structured format that can be analyzed alongside CSV data.
* *Serving:* After I have cleaned and transformed the data accordingly, I will need to load the data into an analytical data store, such as Azure SQL Database where SQL queries can be run for ad-hoc analysis.

***The expected usable dataset***

* There are 3 different folders.
* One folder contains files about sales orders in the US stored in AVRO format.
* The other folder contains files about sales orders in the US stored in JSON format.
* Another folder contains files about sales orders in other countries in csv format.
* Notably, each files in the AVRO and the JSON format contains 4 main columns:
  + Date,
  + Zip,
  + OrderID,
  + OrderLines (a list of dictionaries)
* Whereas each files in the CSV format contains 5 main columns:
  + Date,
  + Zip,
  + Country,
  + OrderID,
  + OrderLines (a list of dictionaries)
* Where the ‘OrderLines’ column is a list of dictionaries. Where each dictionary contains a specific order line information that has 5 key-value pair. Notably, the keys are:
  + ProductID,
  + Date,
  + Units,
  + Revenue,
  + OrderLine
* Furthermore, there is also another excel file that contains information about a product portfolio and a manufacturer portfolio.
* Based on this, this is what I need to do regarding combining the data:
  + First, merge the files from three different folders into one single data set (I think through concatenation; and I need to fill in the “United States” value for the Country column for the JSON and AVRO files because they don’t have that column at the moment),
  + Second, merge the extra excel file using the ProductID sub-column from the OrderLines column of the merged dataset.