**Databricks Assignment Number 2**

**USE CASE :**

1. Read a positional file (input - headerless data file) from s3 bucket (bronze layer) - .TXT (at least 25 columns) -- create a normal txt file with header and multiple cols.

2. Parse the file and apply schema on it(create structType and schema).

3. convert the extracted timestamp in zulu format as a column in all rows extracted from the file.

4. Write the intermediate data into silver layer in avro format.

6. Check the de-dup and remove them and add any more transformations if seemed fit.

7. write the data into gold layer in parquet format.

8. Store the data into unity catalog.

10. Read the incremental data using step #1-#8(Everyday we are getting a new file, choose latest file only)

11. Perform SCD-1 using Pyspark in unity catalog ( insert overwrite should happen on old record 5. Handle null values to have default values as ' '

**SOLUTIONS** :

1)Created connection with S3 buckets (assuming Bronze container has file coming in from upstream everyday )



2) Define schema for the file

my\_schema = StructType([

StructField("Playerid", IntegerType(), True),

StructField("Tournament\_ID", StringType(), True),

StructField("Matches", IntegerType(), True),

StructField("Batting\_Innings", IntegerType(), True),

StructField("Not\_Out", IntegerType(), True),

StructField("Runs\_Scored", IntegerType(), True),

StructField("Highest\_core", StringType(), True),

StructField("Batting\_average", FloatType(), True),

StructField("Balls\_Faced", FloatType(), True),

StructField("Batting\_Strike\_rate", FloatType(), True),

StructField("100", IntegerType(), True),

StructField("50", IntegerType(), True),

StructField("0", IntegerType(), True),

StructField("4s", IntegerType(), True),

StructField("6s", IntegerType(), True),

StructField("Bowling\_Innings", FloatType(), True),

StructField("Overs\_Bowled", IntegerType(), True),

StructField("Maidens\_Bowled", IntegerType(), True),

StructField("Runs\_Conceded", IntegerType(), True),

StructField("Wickets\_Taken", IntegerType(), True),

StructField("Bowling\_Average", StringType(), True),

StructField("Bowling\_Economy\_Rate", StringType(), True),

StructField("Bowling\_Strike\_Rate", FloatType(), True),

StructField("Catches\_Taken", IntegerType(), True),

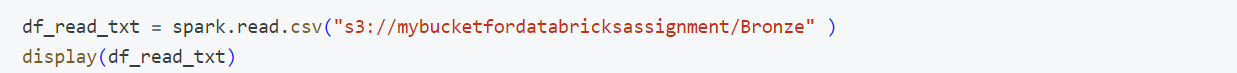
StructField("Team\_id", StringType(), True),

])

3) Define the columns of the fixed length file A screen shot of a computer code

Description automatically generated

4) Reading the file



5)Currently the positional file looks like below, which is not clear to understand

A screenshot of a computer

Description automatically generated

6)Defining a function which takes the fieldname we defined earlier ,starting position and the length of the file and creating a column for the same

A screen shot of a computer code

Description automatically generated

7)We can drop the original column

A screenshot of a computer

Description automatically generated

8) Display the data

A screenshot of a computer

Description automatically generated

9) Using input\_file\_name functions we are taking the name of the file and filtering out the timestamp from file name

A computer screen shot of a computer code

Description automatically generated

10)Converting the Timestamp into Zulu format, also dropping the columns which are not required and replacing the null values with blank. Also removing the duplicate values.

A close-up of a computer code

Description automatically generated

A screenshot of a chat

Description automatically generated

11)Renaming the columns & saving the intermediate results in silver layer

A screenshot of a computer program

Description automatically generated

12)Saving the final results in gold layer

A screenshot of a computer code

Description automatically generated

13) using Rank functions to filter the latest data from the table according to the time stamp

A screenshot of a computer program

Description automatically generated

A close-up of a computer screen

Description automatically generated

14)Overwriting the unity\_table with the latest data inserted

A screenshot of a computer

Description automatically generated

**SNIPPET OF FINAL RESULT :**

A screenshot of a computer

Description automatically generated