case study  :

- solution should be developed using OOPS and have 100% unit test coverage and error handling as required. Rerun without modification should always generate same value. Solution will run every month and corresponding data will be stored in partitioned by Month

Input file (.csv) contains Feature values(FEATURE\*) and Feature score values(FEATURE\_SCORE\* value from 1 to 5), originator and beneficiary

TOTAL\_Score = sum of all \_Score\*

MONTH = YYYYMM format with the data of different month

columns for input file as below : (REF\_ID (unique\_key) string, ORIG and BENEF string, PAYMENT\_DATE = Timestamp, rest double,

REF\_ID, ORIG, BENEF, FEATURE1, FEATURE1\_Score, FEATURE2, FEATURE2\_Score, FEATURE3, FEATURE3\_Score, FEATURE4, FEATURE4\_Score, FEATURE5, FEATURE5\_Score, TOTAL\_Score, PAYMENT\_DATE, MONTH

As a operator, i would like to have the events generated on the feature data with following requirement :

- only select TOTAL\_Score above 15 (filter criteria)

- group the records to generate a single event using connected components of ORIG and BENEF,

- grouped alert will have a ALERT\_KEY(originator) and top3 feature (ordered descending by score) this is corresponding to the highest Total Score and minimum payment\_date

output file format :

REF\_ID, ORIG, BENEF, top\_feat1, top\_feat1\_value, top\_feat1\_score, top\_feat2, top\_feat2\_value, top\_feat2\_score,top\_feat3, top\_feat3\_value, top\_feat3\_score, TOTAL\_Score, PAYMENT\_DATE, MONTH, group, ALERT\_KEY, alert\_top\_feat1, alert\_top\_feat1\_value, alert\_top\_feat1\_score, alert\_top\_feat2, alert\_top\_feat2\_value, alert\_top\_feat2\_score,alert\_top\_feat3, alert\_top\_feat3\_value, alert\_top\_feat3\_score

values before group corressponds to the input record, columns from group onwards are related to the grouped alert

connected components

ref, orig, benef , group

1, Jack, Harry, 0

2, John, Jack, 0

3, Harry, Tom, 0

4, Tom, Jack, 0

5, Bruce, Jackie,1

As the orig, benef are connected to each other from ref 1 to 4 they are grouped together, ref 5 orig and benef are not involved so it is seperate