**Analytics Engineer – Take Home Assignment**

Main Question's Solution:

**1)** Executing the PYSPARK job on YARN.

**2)** Place "User\_visit\_process.py" and "utils.py" at same location ( S3 , HDFS ), i considering s3 location

--> The Main Program is User\_visit\_process.py which requires only one input source file hdfs location .

--> HDFS LOCATION OF SOURCE CSV FILE :- /tmp/truecaller/data.csv or /tmp/truecaller/

--> Run Below spark-submit command on a cluster.

./bin/spark-submit --verbose\

--deploy-mode client \

--driver-memory 5G \

--executor-memory 10G \

--num-executors 2 \

--executor-cores 4 \

--conf spark.pyspark.python=/usr/bin/python3 \

--conf spark.pyspark.driver.python=/usr/bin/python3 \

--py-files s3://<BUCKETS>/TRUECALLER/utils.py \

s3://<BUCKETS>/TRUECALLER/User\_visit\_process.py <HDFS LOCATION OF SOURCE CSV FILE>

**3)** Task divided for the above job.

--> Since Source file is a CSV File and holds billion records , this requires some changes to lessen the execution overall time.

--> So the raw CSV from HDFS is converted first to Parquet format (under 'parquet\_src\_files' folder ) without changing anything with respect to block size or partition, so that it can be easily be loaded via spark job and can be processed further with less execution time.

--> The source parquet file is then been loaded in spark and evaluation of a user stay time on a URL done on it .

--> The output is again a parquet files created under 'parquet\_output\_files' which will hold the required data.

--> logger is used here as a log file (uservisit.log) under /tmp/ location of unix master node .

**4)** File location information example

HDFS SRCFILES :- /tmp/truecaller/data.csv

HDFS Parquet SRCFILES :- /tmp/truecaller/parquet\_src\_files/part-\*-.parquet

HDFS Parquet OUTFILES :- /tmp/truecaller/parquet\_output\_files/part-\*-.parquet

**5)** The further optimization can be done based on file size , the spark configuration .

Kafka Follow-up question:

1. Kafka producer
2. Spark Streaming ( Spark as Consumer)

🡪 Here we transform data in a stream fashion and write it in memory / hdfs as we need user last time visit data inorder to calculate the stay duration , as we are expecting data to come in different user order.

3) HDFS for Checkpoint / Sink

