**This document is meant to assess your technical skills and is classified as "Publicis Sapient confidential". This document by any means shall not be used/shared without permission from Publicis Sapient.**

**Big Data Assignments:**

**QUESTION -1 (General Coding)**

**Description:** From the sample data given below, remove duplicates on the combination of Name and Age and print results.

* Please do not use high level API/Framework like pandas /spark-sql etc.
* Solve this problem by using simple data structures given in a programming language.
* Please try to optimize the solution for efficiency in terms of space and time.

**Given Dataset:**

|  |  |  |
| --- | --- | --- |
| Name | Age | Location |
| Amit | 21 | London |
| Cynthia | 28 | Belfast |
| Wendy | 26 | Manchester |
| Gareth | 21 | Cardiff |
| Charles | 29 | Edinburgh |

**QUESTION - 2**

**Description:** Given a time series data which is a clickstream of user activity is stored in any flat flies, ask is to enrich the data with session id.

**Session Definition:**

* Session expires after inactivity of 30 mins, because of inactivity no clickstream record will be generated.
* Session remains active for a total duration of 2 hours

**Steps:**

* Load Data in any flat file format.
* Read the data and use spark batch (pyspark/scala) to do the computation.
* Save the results in parquet with enriched data.

**Note**: Please do not use direct spark-sql.

**Given Dataset:**

|  |  |
| --- | --- |
| **timestamp** | **userid** |
| **2018-01-01T11:00:00Z** | u1 |
| **2018-01-01T12:00:00Z** | u1 |
| **2018-01-01T11:00:00Z** | u2 |
| **2018-01-02T11:00:00Z** | u2 |
| **2018-01-01T12:15:00Z** | u1 |

**QUESTION 3**

**Description:** In addition to the problem statement given in question 2 assume below scenario as well and design schema based on it:

* Get Number of sessions generated in a day.
* Total time spent by a user in a day
* Total time spent by a user over a month.

Here are the guidelines and instructions for the solution of above queries:

* Design the table in any flat file format
* Write the script to create the file
* Load data into file
* Write all the queries in spark-sql
* Think in the direction of using partitioning, bucketing, etc.