**UE19CS344: Database Technologies (DBT)**
**Jan-May 2022**

# COURSE PROJECT

| | |
|---|---|
| Max Allotted: | 15 marks |
| Professional Report Submission Due Date: | 29/04/22 |
| Team Presentations Dates: | 29/04/22 – 6/05/22 |

Team Size: 3-4 students (fill Google form) – complete formation before 8/04/22.

[A] Technologies / Frameworks to be exercised:

1. **Apache Spark Streaming** [have to execute multiple Spark SQL queries to manipulate the input data]
2. **Kafka Streaming** [have to publish/subscribe the results or produce/consume], Zookeeper.
3. Any other tools as required

[B] Run the same queries in a **batch mode**. Compare the results with the streaming mode of execution.

Language: Java / Python

Example of streaming input data: Twitter feed (tweets)
Computation example: Count of tweets within the window, grouping by #hashtags, etc.
Min, max or other aggregate functions on numeric data within each tumbling windows.
Consumption: Storage of tweets into a database for further processing like batch mode processing.
Note: Window size should be significant and suitable. E.g.: 15-30 mins of tweets.

All the details need to be documented in the project report.