UE20CS343: Database Technologies (DBT)
Jan-May 2023

COURSE PROJECT

Professional Report Submission Due Date:	Mon 24/04/23
Team Presentations Dates:	21/04/23 – 28/05/23

Team Size: 4 students (fill Google form) – complete formation on/before 5/04/23.

[A] Technologies / Frameworks to be exercised:

1. **Apache Spark Streaming, *Spark SQL*** [execute multiple workloads e.g., *Spark SQL* queries to carry out action, transformation or aggregation on the input data]
2. **Apache Kafka Streaming** [have to publish/subscribe the results or produce/consume choosing >=3 topics].
3. Store the data in a DBMS of your choice like *postgres, MySQL*.
4. Make use of any other tool/s as required like *Zookeeper*.

[B] Run the same queries in a **batch mode** on the same/whole data from the database (#3 above).

[C] Compare the above results/accuracy/performance with the streaming mode of execution.

Language: Java / Python

Example of streaming input data:	Twitter feed (tweets), newsfeed, etc.
Computation examples:	Count of tweets within the window, grouping by #hashtags, etc. Min, max or other aggregate functions on numeric data within each tumbling window.
Consumption:	Storage of tweets into a database for further processing like batch mode processing.

Note: Window size should be significant and suitable to your chosen domain problem. E.g.: 15-30 mins of tweets as one window.

- Explain in detail the problem selected and the corresponding solution including the techniques applied like sliding or tumbling windows.
- All the details like code (SQL, java/py), input and output data snapshots, etc. need to be documented in the project report (std template will be provided) which should be in PDF format.
- Before taking snapshots of screens, ensure that the background color is white.