# CSP554—Big Data Technologies

## Assignment #8 A20512400

1. (1 point) Extract-transform-load (ETL) is the process of taking transactional business data (think of data collected about the purchases you make at a grocery store) and converting that data into a format more appropriate for reporting or analytic exploration. What problems was encountering with the ETL process at Twitter (and more generally) that impacted data analytics?

- The main problem was that ETL pipelines displayed inertness - daily positions(the standard) imply that business insight is being derived from day-old data.
- Due to the accelerated pace of business, associations began requesting fresher and fresher information for decision-making.

2. (1 point) What example is mentioned about Twitter of a case where the lambda architecture would be appropriate?

- **Counting tweet impressions** is an appropriate case for lambda architecture. Furthermore, we need memorable counting that traces back all the way to the very second a tweet was posted, not only constant updates as clients tap, swipe, and click.

3. (2 points) What did Twitter find were the two of the limitations of using the lambda architecture?

- **Issue 1:** The lambda design essentially requires that everything be composed twice: once for the cluster stage, and again for the constant stage.
    - **Eg :** In group preparing, it's not difficult to figure the cardinality of a set (for instance, the number of tweets in 60 minutes), and this worth can be utilized for an assortment of calculations and advancements. This is absurd in a continuous stage in case you're preparing input steadily.
It is simple to calculate the cardinality of a set in group preparation (for example, the number of tweets in 60 minutes), and this value can be used for a variety of calculations and developments. If you are steadily preparing input, this is nonsensical in a continual stage.
- **Issue 2:** The total quality can vary erratically from time to time.
    - **Eg :** Let's say the Tempest group encountered a brief increase in load and lost 10 minutes of log data. Before the group layer in a short while prepares the logs, nobody would notice this. Logging pipelines frequently use different code than the ongoing handling layer and are generally more robust because creativity is a clear plan objective. In this instance, the missing details are revealed, and the overall characteristics unexpectedly shift.

4. (1 point) What is the Kappa architecture?

• In the kappa architecture, everything's a stream. Furthermore, if everything is a stream, you just need a stream preparing motor. In the lambda design, group handling essentially means spilling through noteworthy data.

5. (1 point) Apache Beam is one framework that implements a kappa architecture. What is one of the distinguishing features of Apache Beam?

• Using Apache Beam, you are presented with a rich Programming interface that clearly distinguishes between occasion time, when an event truly occurred, and handling time, when it is actually seen in the framework.