CSP 554 BIG DATA TECHNOLOGIES
MOHAMMED WASIM R D(A20497053)

ASSIGNMENT 8

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?**

   Due to the latency that ETL pipelines introduced, business intelligence was being performed on day-old data because of nightly jobs. Organizations started asking for increasingly recent data to inform decision-making as the business accelerated, but this only put more strain on the rickety ETL pipeline, frequently to its breaking point.

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

   The counting of tweet impressions serves as the example. We also want historical counts going back to the time a tweet was posted, not just real-time updates as people are now tapping, swiping, and clicking.
   Example: Think about a tweet from Elon Musk from last year that is currently experiencing a surge in interaction.

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

   Complexity – In essence, the lambda architecture requires that everything be written twice, once for the batch platform and again for the real-time platform. It is necessary to maintain two distinct implementations concurrently for all time, occasionally by different teams. The semantics of the computations were also ambiguous.

   Semantics were unclear - It implies that overall values may occasionally change in an unpredictable manner. For instance, nobody would be aware of it until the batch layer analyzed the logs later if the Storm cluster experienced a brief demand spike and lost 10 minutes' worth of log data.

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

   Everything is a stream in the Kappa Architecture; all we need is a stream processing engine. In contrast to the lambda, it wasn't batch processing.

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

   It offers a robust API that explicitly distinguishes between event time, which is the time at which an event occurred, and processing time, which is the time at which the event is noticed by the system.