# CSP554—Big Data Technologies

## Assignment #8

## Oduri Sai ram

## A20522183

1Ans. Companies employing ETL processes required new data to make choices. MapReduce was implemented as a batch processing layer in the Lambda architectural setup to assess tweet impressions for ad placement algorithms. ETL business intelligence utilized a day-old data, which added delay. Even the best case logs were always at least a few hours old. We all know that old data is an issue in real-time data analytics. ETL pipelines were also challenging to manage. The answer to the stated issue was to increase the frequency. However, raising the frequency would put strain on the pipes, and the breakpoint was reached.

2Ans. The Lambda architecture was an ideal tool for batch processing since there was no risk of a specific dictionary becoming larger than the available memory. The disk would be spilled by the Lambda architecture. In addition, if memory overflows during real-time processing, it creates an issue.

The article describes a rapid transient demand for 10 minutes of log data in one case. In this situation, the storm architecture tends to overlook certain logs during real-time processing, but when batch processing by Lambda architecture begins, they will be displayed again in the system. Because persistence is an explicit design goal, logging pipelines often take a different code route than the real-time processing layer and are frequently more resilient.

3Ans. To begin with, the Lambda design delayed the logged data by a few hours. It could not handle real-time data with a negligible processing latency. As a result, Storm architecture was designed to remedy the problem, but it came at a great cost. Coordinating Lambda architecture with Storm and Summing Bird created complications. The merger necessitated several sacrifices, but it did not meet Twitter's standards.

4Ans. Streams of data are processed via the Kappa architecture. In addition, the article has a paragraph on Kappa architecture that says, "In the kappa architecture, everything is a stream." And if everything is a stream, you only need a stream processing engine." Lambda architecture, on the other hand, processes data through batch processing.

5Ans. Apache beam has a rich API that recognizes the difference between event times, that is the time when an event occurred, its processing time, and the time when the event is observed in the system. For example, an event that occurred at 3:17(event time) is not observed till 3:20(processing time) because of delays in the logging pipeline.