CSP 554 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?**

Sol) Lambda architecture setup at Twitter was having the MapReduce as batch processing layer which analysis the tweet impressions which later used for ad placement algorithms. First step of ETL is to introduce a logging pipeline delay due to the inherent flow of ETL by design. Even in the best possible case logs were always at least a few hours old. This implies that a dashboard of tweet impressions powered by MapReduce would always be a few hours out of date. Considering old data is a problem in real-time data analytics.

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

Sol) The Lambda architecture was an appropriate tool for batch processing as there was no worries about a particular dictionary growing larger than the amount of memory available. This is because the framework will automatically spill to disk. Furthermore, when it comes for real time processing, if the memory overflows, it will be a disaster.

In one example the article explains about a sudden transient load for 10 minutes of log data. In this type of case, the real-time processing, Storm tends to miss those logs but at the moment when batch processing by Lambda architecture begins, it will visible back into the system. Logging pipelines typically form a different code path than the real-time processing layer and are usually more robust because persistence is an explicit design goal. This is how it will support and ensure that no data is lost.

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

Sol) Below are the two major limitations discussed in the article:

The Lambda architecture delayed the logged data by a few hours. It could not handle the real-time data with non-noticeable processing delay. To resolve this issue, Storm architecture was adopted which resulted in more cost.

Second limitation was of managing complexity occurred, when handling Lambda Architecture with Storm as well as Summingbird. The integration required tradeoff in many aspects, but it could suffice the requirements of Twitter.

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

Sol) kappa architecture processes data like stream, which is very different to Lambda architecture which uses batch processing to process data. Also, in the given article a famous line was mentioned “In the kappa architecture, everything’s a stream. And if everything’s a stream, all you need is a stream processing engine”

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

Sol) Apache Beam owns a rich API that exactly recognizes the difference between event time. The time when an event actually occurred, processing time, and the time when the event is observed in the system.
For example, an event occurring at 2:17 (event time) isn’t observed until 2:20 (processing time) due to delays in the logging pipeline.