Akash Tanwani **A20448831**

CSP554—Big Data Technologies

Assignment #8

Exercise 1: Read the article "The Lambda and the Kappa" found on our blackboard site in the "Articles" section and answer the following questions using between 1-3 sentences each. Note this, article provides a real-world and critical view of the lambda pattern and some related big data processing patterns:

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?
   - For Decision making, the organizations were demanding for fresher and fresher data, but the ETL was using day older data, which introduced Latency. Also, the ETL pipelines were difficult to build and maintain. Increasing the frequency was the obvious solution. However, if the frequency was increased to even Hourly, the pipelines would be stressed and reach the break point.

2. (1 point) What example is mentioned about Twitter of a case where the lambda architecture would be appropriate?
   - Suppose we want to count the number of tweet impressions; the count should reflect the real time updates as well as the previous counts since the tweet was posted. Once the (delayed) results from the batch layer arrived, results from the real-time layer could be discarded. In other words, the batch computations provided truth, while the Realtime results were transient.

3. (2 points) What did Twitter find were the two of the limitations of using the lambda architecture?
   - 1. The cost of the complexity increased since two separate implementations need to be maintained in parallel, sometimes by separate teams. This means that changes need to be propagated from one to the other, or else the results will be suspect.
     2. The semantics of the computations were unclear.

4. (1 point) What is the Kappa architecture?
   - 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?
   - Apache Beam presents a rich API that explicitly recognizes the difference between event time, the time when an event actually occurred, and processing time, the time when the event is observed in the system. Also, the Apache Beam API provides one possible abstraction for managing these complexities.