Rakshith Churchagundi Amarnath
A20424771
CSP 554 – Assignment 6

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?

   Answer: The problems encountered with ETL process at Twitter was that ETL pipeline introduced latency, which means that business intelligence was being conducted on a previous day data. When the organizations demanded for fresher data to help in the decision making, changing it to an hourly frequency was a solution but it also stressed ETL pipelines even more, often past the breaking point.

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

   Answer: The example used – to get several tweets (count) impressions, in real – time as users were tapping, swiping and clicking right now, but also historic counts dating back to moment a tweet was posted

   Example: Donald Trump's last year tweet that's receiving a new burst of engagement

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

   - Complexity - The lambda architecture basically means that everything must be written twice: once for the batch platform and again for the real-time platform.
   - Two Separate implementations need to be indefinitely maintained in parallel, sometimes by separate teams. Also, the semantics of the computations were unclear.

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

Answer: In the Kappa Architecture, everything is a stream – we only need a stream processing engine. Unlike the lambda, where it was 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?

Answer: It presents a rich API that explicitly recognizes the difference between event time, the time when an event occurred, and the processing time, the time when the event is observed in the system