CSP – 554 Big Data Technologies

Assignment #8

Rutul Mehta
A20476293

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?

   Ans: In the lambda architecture at Twitter, the batch processing layer was MapReduce to analyze the tweet impressions which can be used for ad placement algorithms. The very first step of ETL i.e., logging pipeline introduced a delay due to the inherent flow of ETL by design. Logs were always at least a few hours old even in the best possible case. This meant that a dashboard of tweet impressions powered by MapReduce alone would always be a few hours out of date. In the world of real-time data analytics, such an old data is considered as problem.

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

   Ans: The Lambda architecture was proved to be an appropriate tool for batch processing as there is no need to worry about a particular dictionary growing larger than the amount of memory available because the framework will automatically spill to disk. Whereas for real time processing, if the memory overflows, it will be a disaster.

   One example the article explains is about a sudden transient load for 10 minutes of log data. In such case, the real-time processing with Storm might miss those logs but the moment batch processing by Lambda architecture happens, it will appear 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?

   Ans: Below are the two major limitations discussed in the article:

   - As explained in the answer to question 1, the Lambda architecture delayed the logged data by a few hours. It could not handle the real-time data with negligible processing delay. To overcome this issue, Storm architecture was adopted which resulted in more cost.
   - Another limitation was of handling complexity involved by 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?

   Ans: Unlike Lambda architecture which processes data in batch, Kappa architecture processes data like stream. As the article quotes, “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?

   Ans: Apache Beam is an API to recognize difference between event time (the time when an event occurred) and processing time (time when the event is observed in the system).

   For example, an event occurring at 1:35 (event time) isn’t observed until 1:38 (processing time) due to delays in the logging pipeline. Apache Beam has a feature to recognize such differences.