CSP554—Big Data Technologies

Assignment #8

Worth: 6 points

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?**
   - Building and maintaining ETL pipelines has proven to be difficult. For some personal experience, developers spent two years grappling with this piece of the data stack on Twitter. Another problem was that the ETL pipelines induced latency in the day-to-day operations of business intelligence. Organizations started to demand more and more recent data to inform decision-making as the speed of business increased. Ad hoc data mining is a technique that is used to mine data on the fly. 2. Machine Learning on a Large Scale.

2. **(1 point) What example is mentioned about Twitter of a case where the lambda architecture would be appropriate?**
   - An example of how to count tweet impressions in practice. We also want historic counts from the time the tweet was released, not just real-time notifications, because users are tapping, swiping, and clicking right now.

3. **(2 points) What did Twitter find were the two of the limitations of using the lambda architecture?**
   - Costs are complicated, because they're done in batches. All must be written twice, once for the batch platform and again for the real-time platform, according to the lambda architecture. In batch processing, you don't have to worry about a dictionary outgrowing the available memory because the framework will immediately spread to the disk. On a real-time site, though, bad things will happen if the dictionary fills up too quickly.

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

- It's all a stream in Kappa's architecture. Lambda's architecture has been simplified. This framework functions similarly to a Lambda Architecture system that is no longer part of the batch processing system. To substitute batch processing, data is simply fed through the streaming system quickly. It simplifies database migration and reorganization by allowing you to simply remove your serving layer database and fill a new copy with data from a canonical store. Since there isn't a batch processing layer, just one set of code has to be kept up to date.

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 has a rich API that understands the difference between event time (the time when an event occurred) and processing time (the time when the event is observed in the system). Owing to delays in the logging pipeline, an event that occurs at 4:17 p.m. isn't noticed until 4:20 p.m., for example.