CSP554—Big Data Technologies

Assignment #8

Worth: 6 points

Due by Sunday after the mid-term

Assignments can be uploaded via the Blackboard portal.

Read (From the Free Books and Chapters section of our blackboard site):

- Kafka: The Definitive Guide, Ch. 1 <- read this for the first class after the mid-term

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?

Ans 1: In Twitter, frontend logs are used to collect and store Impression data. However, there was a delay brought about by the logging pipeline; even in the best situation, logs were a few hours old. In other words, real-time data is not being used for the analytics.

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

Ans 2: Organizations like Twitter want real-time capabilities generation and business insights without sacrificing analytics over a large volume of historic data. Therefore, the lambda architecture would be appropriate. Additionally, Lambda architecture consists of a batch processing layer, a transient real-time processing layer, and a merging layer on top which not only provide real-time insights as users are tweeting, clicking, and sharing but also insights on historic data.

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

Ans 3:

Limitation 1: Everything must be written twice, once for the batch platform and once for the real time platform. In many cases, the implementations are completely different. Two separate implementations need to be maintained in parallel: Changes need to be propagated from one to another.

Limitation 2: Even when the lambda architecture is working as intended, the semantics of the computations are not clear. For example, aggregate values can sometimes fluctuate unpredictably. Let's say Storm cluster lost 10 minutes of log data due to sudden load spike. It goes unnoticed until the logs are processed by batch layer sometime later and aggregate values change suddenly.

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

Ans 4: Everything is a stream in Kappa Architecture and all you need is a stream processing engine. In fact, batch processing of lambda architecture is equivalent to streaming through historic data. The initial execution of this vision was a processing framework called Samza, which used Kafka as the underlying messaging framework. The most recent version is Kafka streams which is used on top of Kafka.

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

Ans 5: Apache beam does not distinguish between batch and streaming computations.
It explicitly recognizes the difference between event time and processing time. The notion of watermark captures the relationship between the two and tries to make a statement about the completeness of observed data with respect to event times.