Assignment 8

CSP 554

Ashmita Gupta (A20512498)

CSP554—Big Data Technologies

Assignment #8

Worth: 6 points

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. Businesses using the ETL process required new data in order to make decisions. The Lambda architecture setup used MapReduce as a batch processing layer that analyzed tweet impressions for ad placement algorithms. ETL used older data of the day, that introduced latency. It used to be that even the best-case logs were always at least a few hours old. As a result, a MapReduce-powered dashboard of tweet impressions was constantly several hours old, and old data causes issues for real-time data analytics. Additionally, managing ETL pipelines was a little challenging so increasing the frequency was the best way to address the problem at hand. However, increasing the frequency would also stress the pipelines and the breakpoint could be hit

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

   Ans. Lambda architecture was an appropriate tool for batch processing as there were no worries about a particular dictionary growing larger than the amount of memory available. Lambda architecture would automatically spill to disk. However, if memory overflows during real-time processing, it can cause issues.

   The article describes in one of the examples, a sudden transient load of log data for ten minutes in one case. These logs are typically missed by the storm architecture during real-time processing in this kind of situation, but they will now be seen again once batch processing by Lambda architecture. Logging pipelines typically take a different code path than the real-time processing layer and are usually more robust because persistence is an explicit design goal. In this way, 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. Two of the limitations of using the lambda architecture:

1) Lambda architecture delayed the logged data by a few hours. It was not able to handle real-time data with almost insignificant processing delay. Hence, Storm architecture was introduced to resolve the issue, but it also resulted in high costs.
2) Managing Lambda architecture with Storm and Summing bird resulted in complexity issues. The amalgamation required tradeoffs in many aspects, but it could not satisfy the requirements of twitter.

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

Ans. Data is processed in the form of streams in Kappa architecture. Additionally, there is a statement in the article about Kappa architecture that reads, “In the kappa architecture, everything’s a stream. And if everything’s a stream, all you need is a stream processing engine”. On the other hand, Lambda architecture uses batch processing to process the data.

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 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. For example, an event that occurred at 2:17(event time) is not observed till 2:20(processing time) because of delays in the logging pipeline.