Kajol Tanesh Shah
A20496724 Fall 2022

CSP554—Big Data Technologies

Assignment #8

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 were encountered with the ETL process at Twitter (and more generally) that impacted data analytics?

Ans. ETL pipelines were difficult to build and maintain. The issue encountered with ETL process at Twitter was that ETL pipelines introduced latency—nightly jobs (the norm) meant that business intelligence was being conducted on day-old data. This implied that the dashboard of tweet impressions would always be a few hours out of date. This impacted data analytics and as the pace of business quickened, organizations began demanding fresher and fresher data to guide decision-making.

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

Ans. If the Storm cluster suffered a transient load spike and 10 minutes' worth of log data got dropped, no one would notice this until the logs are processed by the batch layer sometime later. 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. In this scenario, the missing data reappear and the aggregate values change suddenly. In such a case, lambda architecture would be appropriate.

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

Ans. The two limitations of using the lambda architecture are:-
With the lambda architecture, new capabilities were gained but at a cost of increased complexity.
The lambda architecture exemplified deploying the right tool for the job — a platform for batch processing and a separate platform for real-time processing — and then dealing with the complexities of munging two sets of results together. So, Summingbird was a more suitable option than this.

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

Ans. In the kappa architecture, everything's a stream. And as everything's stream, all you need is a stream processing engine. The kappa architecture represents a swing of the pendulum back to a one-size-fits-all solution.

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 (a distinction, for example, not captured in Spark Streaming, but recently introduced in Spark's Structured Streaming API).