Chirag Khandhar

A20438926 | CSP 554 - Big Data Technologies | Fall 2020 | Assignment 8

Exercise 1:

1. 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?

- The companies were demanding fresher and fresher data for decision-making, but the ETL used the older data of the day, that introduced Latency. The ETL pipelines were also difficult to construct and manage. The simple solution was to raise the frequency. However, if the frequency was raised to hourly, the pipelines would be stressed, and the break point would be hit.

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

- Suppose we try to count the amount of tweet impressions; the count should represent current time updates as well as past counts before the tweet was posted. If the (delayed) results of the batch layer have arrived, the results of the real-time layer could be discarded. In other words, the batch computations produced the reality, while the real-time results were temporary.

3. What did Twitter find were the two of the limitations of using the lambda architecture?

- The cost of the complexity increased since two separate implementations need to be maintained in parallel, sometimes by separate teams. This means that changes need to be propagated from one to the other, or else the results will be suspect.
- The semantics of the computations were unclear.

4. What is the Kappa architecture?
- In the Kappa architecture, everything's a stream. And if everything's a stream, all you need is a stream processing engine.

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

- Apache Beam provides a rich API that clearly understands the difference between the time of event, the time when the event actually occurred, and the time of processing, the time when the event is detected in the system. The Apache Beam API also offers one potential abstraction for handling these complexities.