CSP 554 Big Data Technologies

Assignment – #8

Shiva Sankar Modala(A20517528)

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 problem that was encountered with the ETL(Extract-transform-load) process at Twitter that impacted data analytics are:

- Developers soon began to realize that ETL pipelines were difficult to build and maintain.
- ETL pipelines introduced latency - nightly jobs (the norm) meant that business intelligence was being conducted on day-old data.

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

The example that mentioned about Twitter of a case where the lambda architecture would be appropriate is to count tweet impressions. Furthermore, not only do we want real-time updates as users are tapping, swiping, and clicking right now, but we want historic counts dating back to the moment a tweet was posted.

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

The two of the limitations of using the lambda architecture found by twitter are:

- The lambda architecture basically means that everything must be written twice: once for the batch platform and again for the real-time platform. In many cases, the implementations are completely different.
- Two separate implementations need to be indefinitely maintained in parallel, sometimes by separate teams. This means that changes need to be propagated from one to the other, or else the final results will be suspect.

4. What is the Kappa architecture?

Kappa architecture is in which the data is processed as a stream, and it has a stream processing engine to process the data. It was proposed in a 2014 blog post by Jay Kreps,10 one of the original authors of Kafka and a data architect at LinkedIn at the time.

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

Apache Beam is one framework that implements a kappa architecture. The distinguishing features of Apache beam is “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”.