Big Data Technologies- CSP 554

Assignment-8

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?**

   The companies implementing the process of ETL were demanding fresh data for making 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. Even the best possible case logs used to be always at least a few hours old. Hence, a dashboard of tweet impressions powered by MapReduce was always a few hours out of date and we know that old data is a problem in real-time data analytics. ETL pipelines were also a little difficult to manage. The best solution to the above problem was to increase the frequency. But 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?**

   The 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 itself spill the disk. Also, in real-time processing, if the memory overflows, it results in a problematic situation.

   In one example the article explains a sudden transient load for 10 minutes of log data. In this type of case, in real-time processing, the storm architecture tends to miss those logs but now when batch processing by Lambda architecture begins, it will be visible back into the system. 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?**

   Firstly, 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.

   Secondly, 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?**

   Kappa architecture processes the data in the form of streams. Also, the article has a line about Kappa architecture that states that “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?**

   Apache beam has a rich API that recognizes the difference between event times, that is the time when an event occurred, its processing time, and the time when the event is observed in the system. For example, an event that occurred at 3:17(event time) is not observed till 3:20(processing time) because of delays in the logging pipeline.

Submitted By: -
Aastha Dhir
CWID- A20468022
adhir2@hawk.iit.edu