SHRADDHA PATEL A20499171

CSP554—Big Data Technologies

Assignment #8

Worth: 6 points

Due by Sunday after the mid-term

Assignments can be uploaded via the Blackboard portal.

Read (From the Free Books and Chapters section of our blackboard site):

- Learning Spark, Ch. 4 <- read this before the mid-term
- Kafka: The Definitive Guide, Ch. 1 <- read this for the first class after the mid-term

Make sure to read the article referenced below before the mid-term. But you can submit the assignment itself up to Friday after the mid-term without penalty.

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?
   Answer:
   The Problem encountered with the ETL process that impacted Data analytics:
   Firstly, ETL pipeline were difficult to build and maintain, second problem is Latency as data provided by ETL was old data, so processing was done on old data which impacted on Twitter's business decisions. lastly, need for fresh data in real time was not possible.

2. (1 point) What example is mentioned about Twitter of a case where the lambda architecture would be appropriate?
   Answer:
   Real time analytics for insights generation is a use case at twitter where lambda architecture was appropriate. The example which mentions by about twitter where Lambda architecture would be used was Count of tweet impressions in real-time using both real-time and historic data of a certain tweet.

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

Answer:
limitations of using the lambda architecture
1: complexity of cost:

The lambda architecture basically needs everything must be done twice: once for the batch processing and again for the stream processing.

Both different implementation to be done in parallel which often needs to be maintained by two teams.

2. Semantics of computations are unclear:

Although, Lambda architecture is working as intended. Unpredictable fluctuation aggregate values as missing data in real time due to some loss during stream processing but these losses are reflected in batch processing in after some time.

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

Answer:
In the kappa architecture, everything's a stream so only one processing engine is required. Kappa Architecture is a simplification of Lambda Architecture with the batch processing system removed where data is simply fed by the streaming system.

5. (1 point) Apache Beam is one framework that implements a kappa architecture. What is one of the distinguishing features of Apache Beam?

Answer:

Apache beam has an API that explicitly recognizes the difference between event time, the time when an event occurred, and processing time, the time when the event is observed in the system.