CSP554—Big Data Technologies

Student ID: A20549927
Name : Prachi Kotadia

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?**

   Ans. Similar to several corporations handling extensive data, Twitter encountered various ETL obstacles that affected their data analytics. One problem was the overwhelming amount of data, which made conventional ETL procedures sluggish and ineffective and prevented fast data analysis. Another issue was the complexity of data transformations brought about by the diversity of data sources and formats, which made it more difficult to combine the data into a form that could be used for analytics. Furthermore, there were substantial operational hurdles in keeping ETL operations scalable and reliable in the face of frequently changing data and high velocity.

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

   Ans. when calculating the quantity of impressions received by a tweet. Additionally, if they want to monitor the historical counts from the moment the tweet was sent. One tweet from Donald Trump, for instance, caused a spike in conversation.

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

   Ans. Developers have to implement twice due to complexity.
   Uncertain Semantics: There is uncertainty in semantics.
   It is a strain for developers to have to go through the implementation process twice because of the task's complexity, and this extra layer of complexity makes their job even more difficult. Furthermore, as the semantics of the system or data in question are still unknown and susceptible to future changes or ambiguities, the uncertainty around these semantics exacerbates the challenges already present in the project.

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

   Ans. By treating everything as a stream, this data processing architecture makes batch and real-time processing simpler. A single stream processing engine is used by the Kappa architecture for batch and real-time processing.

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

   Ans. The advanced API offered by Apache Beam clearly distinguishes between event time—the actual time an event happens—and processing time—the time the event is noticed by the system.
   To assert the completeness of observed data in terms of event times, Apache Beam makes use of the notion of "watermarks".