# CSP554—Big Data Technologies

## Assignment #6

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?**

In general, the problem with ETL jobs was, they were difficult to build and maintain and also they introduced latency as these were mostly nightly jobs. So the data provided by the ETL jobs was a day old which impacted the business decisions. The problem at twitter was logging pipeline which is responsible to capture the data introduced a delay of hours.

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

Real time analytics for insights generation is a use case at twitter where lambda architecture was appropriate. As, lambda architecture provides a batch processing layer and a transient real-time processing layer, plus a ,merging layer on top.

1. **(2 points) What did Twitter find were the two of the limitations of using the lambda architecture?**

The two limitations of using lambda architecture are:

1. With lambda architecture everything needs to be done twice. Two separate implementations have to be maintained all the changes from the batch have to be propagated to real time and vice-versa.
2. Even though the lambda function architecture is working fine the semantics of computation are unclear.
3. **(1 point) What is the Kappa architecture?**

Kappa Architecture is a software architecture pattern. Kappa Architecture is a simplification of Lambda Architecture. A Kappa Architecture system is like a Lambda Architecture system with the batch processing system removed. To replace batch processing, data is simply fed through the streaming system quickly.

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

Apache Beam is an open source, unified model for defining and executing both batch and streaming data-parallel processing pipelines, as well as a set of language-specific SDKs for constructing pipelines and runtime-specific Runners for executing them.

Apache Beam presents 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.