# 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

Prepare:

* You should be informed of your project team assignment early the week of 10/18. Or you should already know you registered to do a research paper.
* If you are on in a project team , the “voice of the team” should coordinate a virtual meeting to discuss the specific project you want to pursue
* In either case (if you are doing a project or research paper) start preparing a half page description of what you intend to do. Make sure to include citations for at least two references you will use.
* The proposal must be submitted by no later than Saturday night 11/6. I will have an “assignment” for this on the blackboard to enable you to upload your proposal. For a project team the proposal should be uploaded only one by the “voice of the team.”

Make sure to read the article referenced below before the mid-term. But you can submit the assignment itself up to Sunday 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?
2. (1 point) What example is mentioned about Twitter of a case where the lambda architecture would be appropriate?
3. (2 points) What did Twitter find were the two of the limitations of using the lambda architecture?
4. (1 point) What is the Kappa architecture?
5. (1 point) Apache Beam is one framework that implements a kappa architecture. What is one of the distinguishing features of Apache Beam?