# CSP554—Big Data Technologies

## Assignment #9

## Readings

Read Chapters 1, 2 and 3 from our next book: Pramod J. Sadalage and Martin Fowler. 2012. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley.(PS)

## Worth: 5 points + 5 points extra credit

## Due by the start of the next class period

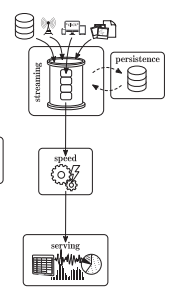
Assignments should be uploaded via the Blackboard portal

Exercise 1) 5 points

Read the article “Real-time stream processing for Big Data” available on the blackboard in the ‘Articles’ section and then answer the following questions:

1. (1.25 points) What is the Kappa architecture and how does it differ from the lambda architecture?

Kappa architecture does all the computation in the stream processing system and only perform recomputation when the business logic changes by replaying historical data. It employs a powerful stream processor capable of coping with data at a far greater rate than it Is incoming and a scalable streaming system for data retention.



The main difference between lambda and kappa is that the latter is a simpler architecture that does not recompute all data in the batch layer.

1. (1.25 points) What are the advantages and drawbacks of pure streaming versus micro-batch real-time processing systems?

Pure streaming systems provide very low latency but is more costly.

Micro-batch systems provide higher throughput but a reduced latency.

1. (1.25 points) In few sentences describe the data processing pipeline in Storm.

In this topology, data is ingested from the streaming layer and then passed between Storm components, until the final output reaches the serving layer. Spouts, the nodes that ingest data, emits tuples to the nodes downstream, bolts. Spouts and bolts are distributed in a round-robing method. It could also be done via a scheduler to account for scenarios in which a certain processing step has to be executed on a particular node. Storm provides the option of at-least-once processing through an acknowledgement feature that tracks the processing status of every single tuple on its way through the topology.

1. (1.25 points) How does Spark streaming shift the Spark batch processing approach to work on real-time data streams?

It chunks the stream of incoming data items into small batches, transforming them into RDDs and processing them as usual. It further takes care of data flow and distribution automatically.

Exercise 2) 5 points extra credit

Follow the document “Instructions for setting up a VM with Kafka” included with this assignment and execute the demo code. Provide enough screen shots to indicate you have completed the document through section 4. Then remember to terminate your VM.

