## Week 5: Daily Morning Challenge

## Day 1: Tuesday 21st January 2020

## Adeyeye Ayobami (Awhy)

**Question 1:** Give an overview of what you understand as the MapReduce technology. Explain how it is similar to the Split-Apply-Combine technology created by Hadley Wickham.

**MapReduce** is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster

A MapReduce program is composed of a map procedure (or method), which performs filtering and sorting, and a reduce method, which performs a summary operation. The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

The model is a specialization of the split-apply-combine strategy for data analysis. It is inspired by the map and reduce functions commonly used in functional programming, although their purpose in the MapReduce framework is not the same as in their original forms. The key contributions of the MapReduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved for a variety of applications by optimizing the execution engine. As such, a single-threaded implementation of MapReduce is usually not faster than a traditional (non-MapReduce) implementation; any gains are usually only seen with multi-threaded implementations on multi-processor hardware. The use of this model is beneficial only when the optimized distributed shuffle operation (which reduces network communication cost) and fault tolerance features of the MapReduce framework come into play. Optimizing the communication cost is essential to a good MapReduce algorithm.

The split-apply-combine strategy is similar to the map-reduce strategy for processing large data, recently popularized by Google. In map-reduce, the map step corresponds to split and apply, and reduce corresponds to combine, although the types of reductions are much richer than those performed for data analysis. Map-reduce is designed for a highly parallel environment, where work is done by hundreds or thousands of independent computers, and for a wider range of data processing needs than just data analysis. Just recognizing the split-apply-combine strategy when it occurs is useful, because it allows you to see the similarly between problems that previously might have appeared unconnected. This helps suggest appropriate tools and frees up mental effort for the aspects of the problem that are truly unique. This strategy can be used with many existing tools: APL’s array operators, Excel’s pivot tables, the SQL group by operator, and the by argument to many SAS procedures. However, the strategy is even more useful when used with software specifically developed to support it; matching the conceptual and computational tools reduce cognitive impedance.

Software framework architecture adheres to open-closed principle where code is effectively divided into unmodifiable frozen spots and extensible hot spots. The frozen spot of the MapReduce framework is a large distributed sort. The hot spots, which the application defines, are:

an input reader

a Map function

a partition function

a compare function

a Reduce function

an output writer

## Question 2: Briefly explain three effective method for field research

**Case Study**

A case study research is an in-depth analysis of a person, situation or event. This method may look difficult to operate; however, it is one of the simplest ways of conducting research as it involves a deep dive and thorough understanding the data collection methods and inferring the data.

**Participant Observation**

In this method of field research, the researcher is deeply involved in the research process, not just purely as an observer, but also as a participant. This method too is conducted in a natural environment but the only difference is the researcher gets involved in the discussions and can mould the direction of the discussions. In this method, researchers live in a comfortable environment with the participants of the research, to make them comfortable and open up to in-depth discussions.

**Direct Observation**

In this method, the data is collected via an observational method or subjects in a natural environment. In this method, the behaviour or outcome of situation is not interfered in any way by the researcher. The advantage of direct observation is that it offers contextual data on people, situations, interactions and the surroundings. This method of field research is widely used in a public setting or environment but not in a private environment as it raises an ethical dilemma.