Week 2 Lecture Summary

**Lecture**

Generators are functions that return an iterable object similar to a list; however, unlike a list, the elements of an iterator aren’t stored in memory and are generally more memory efficient when processing large datasets.

Functional programming is a programming paradigm where functions are treated as first-class objects. Depending on the language, such functions have no side effects; therefore are easier to reason about. Some popular functions within this paradigm are map(), filter(), and reduce().

MRJob is a python library for the MapReduce framework. The framework operates using key value pairs. The three common functions of the framework are mapper(), reducer(), and combiner(). Mapper functions maps values from the input to variables. The reducer function combines input via the provided key. The combiner function is used as an optimization where data locality processing is enforced.

**Readings**

Chapter 2 in Data Intensive Processing with MapReduce expands on the topics discussed in the lecture. The key idea behind MapReduce is similar to divide-and-conquer, where large problems are divided into several smaller ones so that parallelism can be used for optimization. Its roots stem from functional programming where higher-order functions can accept (and return) other functions as arguments. Such functions within the MapReduce framework depend on key-value pairs that are the most atomic data structure within the framework. MapReduce supports a level of abstraction where the *what* is separated from the *how*. Developers submit jobs composed of mapper, reducer, partitioner, and combiner functions to a node of a cluster and the Hadoop execution framework takes care of everything else from scheduling, data and code locality, synchronization, and error handling.

The MRJob documentation details what one would expect.. the library itself. The authors argue that this library is the easiest route to generate Python programs for Hadoop without being a Hadoop expert; however, the library also lets one run their code without Hadoop at all. It is encouraged that all MapReduce code for one job remains in a single class and that command line filters be placed before and/or after the Python code. MRJob supports cloud integration using Amazon Elastic and relatively easy debugging using the console via “obscure” log files.