

Announcements

■ Today's Agenda

- Welcome Chase Basich, who is lecturing in my absence.
- Work through the Scrabble API example posted this [past Wednesday](#).
- Discuss the MapReduce Programming Model, what a mapper is, what a reducer is, and how they can be chained together in a single pipeline of processes to analyze and process large data sets.
 - We'll present the map and reduce executables associated with the most canonical of MapReduce jobs: word counts. The slides present code in Python, since it's very short and easy to follow (even if you don't know Python). My lecture won't focus on the code, but rather on the general idea (which is fairly straightforward, imo.)
 - We'll discuss how very large data sets can be partitioned into many, many chunk files and processed by a large number of simultaneously executing map and reduce jobs on hundreds or even thousands of machines.
 - We'll discuss the **group-by-key** algorithm (very straightforward, actually) which is run on the full accumulation of mapper output files to generate the full set of reducer input files.