Charley Ferrari

Week 14 Assignment

Lets assume that I wanted to solve the same problem as I did in the quiz: I want to aggregate the number of bytes downloaded for each language (lets also say I want to count the number of records per language as well.)

The Map stage would be able to take each hour’s log and perform its own aggregation. Within each node, the log files sent to each node would be separately aggregated. Using the rough numbers of the example we have, lets say we have 60,000 log files, and 600 nodes. 100 files would be moved to each node, and then summed up there.

I would be interested to see a project like this run. One issue I didn’t notice in the quiz was what the relationship is between the number of nodes and the type of aggregation you’re doing. In this case, we have a relatively small number of languages, but a large number of nodes. If there are more categories you’re aggregating into than nodes, you could imagine multiple categories being processed per node. In this case, multiple nodes are processing a different group of the same category, which would then have to be combined in the reduce phase. I imagine that part of the magic that Hadoop works involves working with different cases of nodes and aggregation categories. As long as we’re dealing with aggregations like count and sum, aggregation of a split category across several nodes should be easy.