Write a mapreduce program to use CombineFileInputFormat. Pass a directory containing multiple small files. The program should be able to pick multiple files in a single mapper.

To use CombineFileInputFormat you need to implement three classes. The class CombineFileInputFormat is an abstract class with no implementation, so you must create a subclass to support it; we’ll name the subclass CFInputFormat. The subclass will initiate a delegate CFRecordReader that extends RecordReader; this is the code that does the file processing logic. We’ll also need a class for FileLineWritable, which replaces LongWritable normally used as a key to file lines.

CFInputFormat.java

The CFInputFormat.java doesn’t do much. You implement createRecordReader to pass in the record reader that does the combine file logic, that’s all. Note that you can call setMaxSplitSize in the initializer to control the size of each chunk of files; if you don’t want to split files into half, remember to return false in isSplitable method, which defaults to true.

FileLineWritable.java

This file is very simple: store the file name and offset, and override the compareTo method to compare the file name first, then compare the offset.







