On finer details of Hadoop

How does Map and Reduce phase gets executed in an overlapped fashion?
The map phase consists of the function defined in the mapper. On the other hand the reduce phase also includes the shuffle and sort process. While the actual reducer function can not start until the mapper has finished processing all records, the shuffle step can start beforehand and this overlapping of computation appears in the logs of map reduce jobs as concurrent processing of map and reduce phase.

InputFormat and OutputFormat for Map job?
Some of you might have noticed that the map job has LongWritable as the data type set for the key. This data type primarily depends on the InputFormat that you are using in your program, in addition to the data in the file. For example, in the lab, you were using the InputFormat of TextInputFormat in which case the key is the byte offset of the line in the file (thus LongWritable). Another format is the KeyValueInputFormat which breaks a line into key and value based on the first tab character that it finds. Finally there is a SequenceInputFileFormat allows you to read special binary files that are specific to Hadoop.

The corresponding OutputFormats are :
1. *TextOutputFormat:* Default; writes lines in "key \t value" form
2. *SequenceFileOutputFormat:* Writes binary files suitable for reading into subsequent MapReduce jobs
3. *NullOutputFormat:* Disregards its inputs