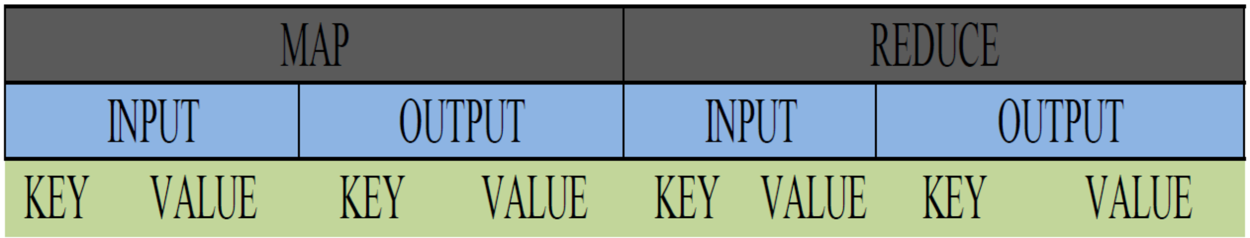
MapReduce Key Value pairs

Compute the key value pairs for the following problems.



MAP				REDUCE			
INPUT		OUTPUT		INPUT		OUTPUT	
KEY	VALUE	KEY	VALUE	KEY	VALUE	KEY	VALUE

1) You are the vendor of a popular website. User session information is available in the server logs. (Assume that the server log has information about an event that was done by a user. For instance – Location, userId, EventType, channelUsed, TimeStamp) You want to find the mean session length per geography which will be helpful in understanding geography based user behavior.

2) A client of yours runs an e-commerce business. The set of products on sale is fixed. The client is interested in product purchase co-occurrence pattern. If we are to implement this in MR framework, Elaborate the Key and value spaces.

3) Facebook has a list of friends. They also have lots of disk space and they serve hundreds of millions of requests every day. They've decided to pre-compute calculations when they can to reduce the processing time of requests. One common processing request is the "You and john have 230 friends in common" feature. When you visit someone's profile, you see a list of friends that you have in common. This list doesn't change frequently so it'd be wasteful to recalculate it every time you visited the profile. If we are going to use map-reduce so that we can calculate everyone's common friends once a day and store those results and do a quick look up. How would the MR key spaces vary?

4) New York times is interested in finding the Word Length distribution of all articles published since their inception. Can we use MR framework to solve the same?

5) In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus.

   An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "di-gram"); size 3 is a "trigram". Larger sizes are sometimes referred to by the value of n, e.g., "four-gram", "five-gram", and so on.

   If we are to implement n-gram statistics over the text corpus of all literature collections in digital libraries in MR, how would the Key-value spaces behave?

6) Ability to identify a failure in a physical component has enormous value in maintaining SLA's and reducing infrastructure downtime. An infrastructure service provider wants to utilize the power of Hadoop to analyze the server events and logs to determine potential failure of a component. Consider that server logs are continuously getting generated and this data has to be monitored to catch a symptom of a potential failure. If this application needs to be built on MR, describe the key value pairs needed for this processing. Can we extend this thought to also include type of failure?

7) There are bunch of numbers in a file, derive key-value space to compute the central imputations ?