Makeup Quiz - Ch. 6

Please answer all questions in complete sentences. Any responses not given in the form of a complete sentence will be marked incorrect.

1. Assume we have data comprised of records of the form <url, timestamp>. For example, one such record might be <www.facebook.com, Oct. 10, 2017 12:45:23>. Say we wish to run queries to calculate the number of pageviews by hourly ranges. For instance, calculate the number of pageviews of www.facebook.com from Oct. 1, 2017 01:00:00 to Oct. 1, 2017 03:00:00. What batch view should the batch layer precompute in order to make this query efficient to compute in real time?
    a. The batch view produced should precompute the number of pageviews in hourly buckets for each URL, like the following:

URL	Hour	# Pageviews
foo.com/blog	2012/12/08 15:00	876
foo.com/blog	2012/12/08 16:00	987
foo.com/blog	2012/12/08 17:00	762
foo.com/blog	2012/12/08 18:00	413
foo.com/blog	2012/12/08 19:00	1098
foo.com/blog	2012/12/08 20:00	657
foo.com/blog	2012/12/08 21:00	101

2. Why can the batch view computed by an incremental algorithm be substantially larger than the batch view computed by a recomputation algorithm?
    a. The batch view produced by an incremental algorithm can be substantially larger than that view produced by a recomputation algorithm because the recomputation algorithm has access to the entire master dataset whereas the incremental algorithm only has access to new data and the previous batch view. For instance, to compute an average, an incremental algorithm will need to also store the previous sample size used to compute the average in the batch view. To compute the number of unique visitors to a site, an incremental algorithm will need to store the id of each user in the batch view.

3. What two functions must be implemented to perform a batch computation job in Hadoop?
    a. In order to perform a batch computation job using MapReduce, you must implement a *Map* function and a *Reduce function.*