Makeup Quiz - Ch. 8

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

1. One query we wish to compute is the number of pageviews for each URL by time slices. In order to speed up this query, we precompute page views into hourly buckets that look 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

   However, this query becomes slower as you increase the size of the time range. Why does this query become slower?
   a. This query becomes slower as you increase the size of the time range because more values will need to be retrieved from disk in order to complete the computation.

2. What is one way of speeding up the query described above?
   a. One way of speeding up the above computation is by precomputing pageview counts at coarser granularities such as 1-day, 7-day, and 28-day granularities. For each year, this only adds 430 additional rows to the batch view.

3. Another query we wish to compute is the number of unique visitors to a URL by time slices. Assume that we have a batch view where each row contains <url, hourly bucket, number of unique visitors> like so:

URL	Hour	# Unique Visitors
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

Can we resolve our query by summing the number of unique visitors in different hourly buckets? Why or why not?

a. We cannot resolve this query by summing the number of unique visitors in different hourly buckets because the same user may appear as a unique visitor in different hourly buckets. If we simply added rows, then we would be counting that same user multiple times.