Tutorial on Indexes & Filters

Database System Technology

Niv Dayan

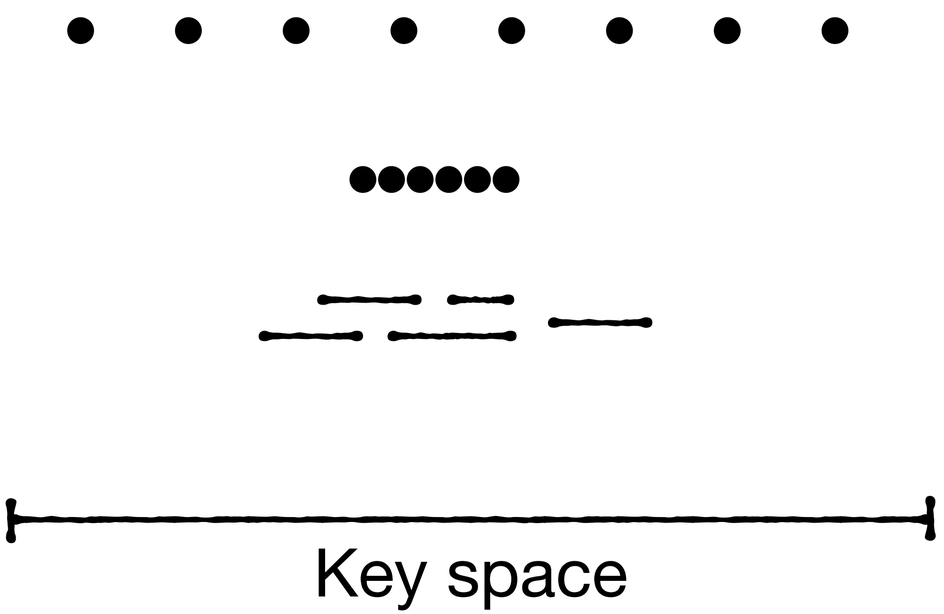
Question 1

Consider an LSM-tree and the three query workloads below. For each workload, comment on the usefulness of having Bloom filters and/or a buffer pool (block cache).

(1) uniformly randomly distributed get queries

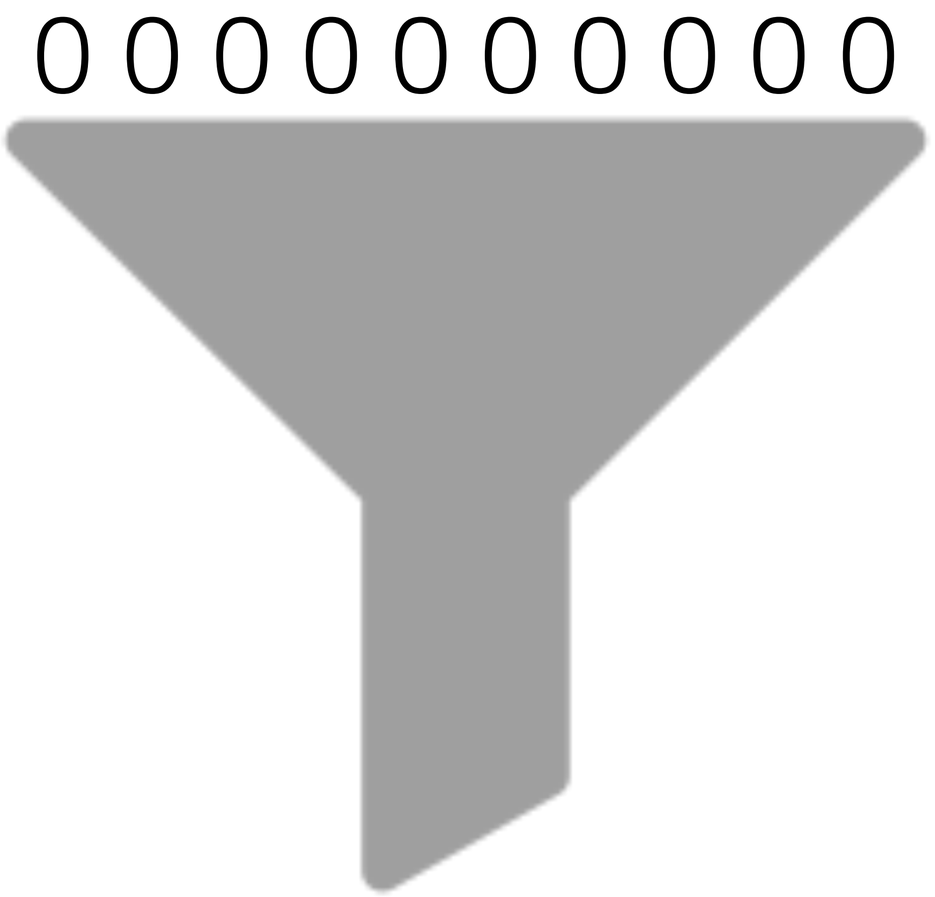
(2) get queries with lots of spatial locality

(3) scan queries with lots of spatial locality



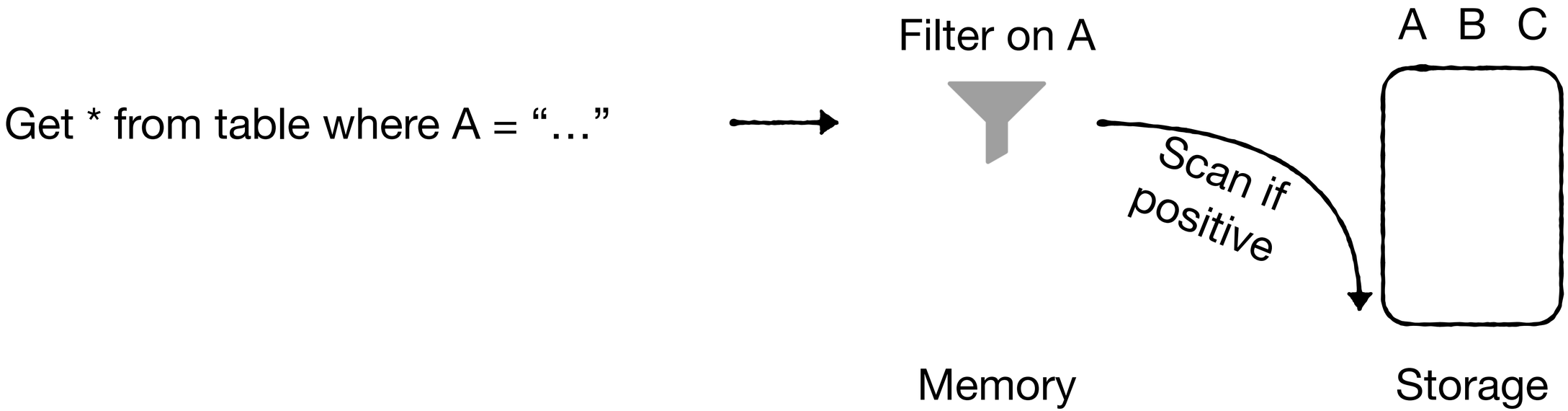
Question 2

Can a Bloom filter handle deletes? Why or why not?



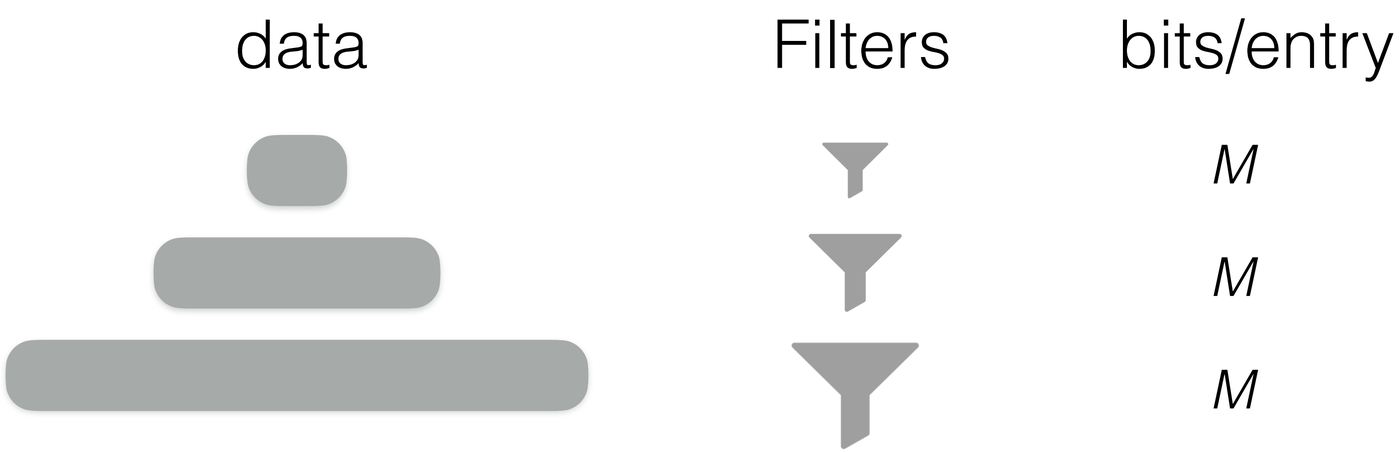
Question 3

Suppose we apply a Bloom filter on a given column in a DB table. What's the impact on the filter as more non-unique vs. unique values of A are added. What can be done to address any issues?



Question 4

We have a leveled LSM-tree with a Bloom filter with M bits/entry for every level. Our query workload consists of point queries to keys that exist, usually at the largest level. Suppose we are near our memory capacity and must free some space used by the filters. Option 1 is to reduce the number of bits/entry for each filter. Option 2 is to drop the filter for the largest level. Compare and contrast these approaches.



Question 5

What is the cost of building Bloom filters for a leveled LSM-tree, measured in worst-case memory accesses per insertion? Assume the same number of bits per entry *M* is assigned to filters at all levels.