Synopsis for Massive Data

Carlos Martinez and Nazrin Najafzade

Introduction

Massive data gives new opportunities, higher profits, smarter business moves to the business. In order to answer these questions, traditional methods are not enough. For instance, full data can be stored in a data warehouse and can be indexed and available, however working with full data will be slow and costly. Approximate Query Processing (AQP) is one of the solutions to that kind of problems by its cost-saving and fast techniques. It gives an approximate answer for a query and there is no need to deal with all data when an approximate answer may be accepted. Using synopses and working with it is more practical, cost-effective.

Sampling

Given a set of data (that we are going to call population) and an aggregation simple query (like SUM or AVG) that we want to estimate, we are going to select a small random set of the population and try to estimate statistics (mean and variance) and using this statistics for estimating the final result of the query and provide the accuracy.

The most typical sampling method are:

A) Bernoulli sampling - all members of the population have equal probability be selected

B) Simple random sampling - all elements in a dataset have the equal chance to be selected.
- Simple random sampling with replacement (SRWR) - all values can occur only once.
- Simple random sampling without replacement(SRWoR) - removes restriction to be selected only once. Values can be selected one or more times.

C) Stratified sampling - dataset is divided into m subgroups called strata. For each strata, SRSWoR is applied.

Sketches

Sketches is one of the newest methods for data synopses. Nowadays sketches has a real impact in streaming processing of large structured data and they have been very useful for answer count distinct queries. The main idea, is that we have a sketch matrix that is going to multiply a column data vector, in order to get the sketch vector.

An example:
We have the following Stream ->
[A, B, C, A, A, B, D, …]
We habe 4 hash function, the next table summarize the hash function for the characteres A, B, C and D.

Character \ Hash function	H1	H2	H3	H4
A	1	6	3	1
B	3	2	4	6
C	1	4	1	6
D	6	2	4	1

So, for each character we are going to execute all the hash functions and increas one cell of the matrix, when the character D arrive, the matrix would be like this:

Hash Function \ Hash Value	0	1	2	3	4	5	6
H1	0	4	0	2	0	0	1
H2	0	0	3	0	1	0	3
H3	0	1	0	3	3	0	0
H4	0	4	0	0	0	0	3

Therefore, the summarized query for counting A, we need the cell in the matrix where A vaue is stored:
H1 (A) = 1 H2 (A) = 6 H3 (A) = 3 H4 (A) = 1
So the result would be the vector [4,3,3,4] and the solution would be the minimum value.

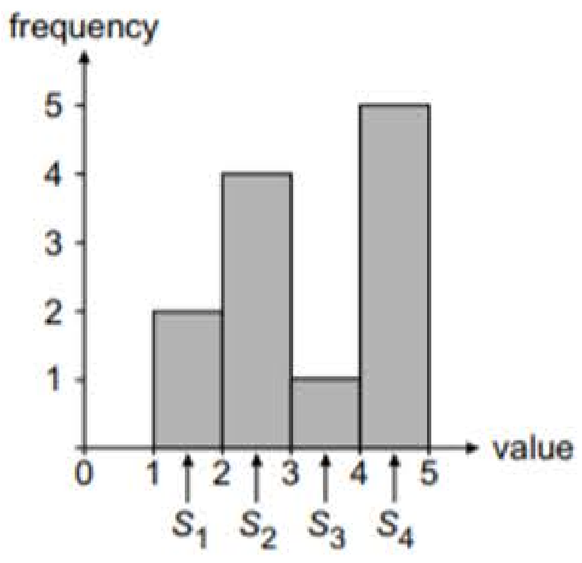
Changing data is not a problem with sketches because it is flexible and this update will affect only an entry in the data vector. Moreover, the answer for a query is fast and additionally it takes less storage and the reduction of the bandwidth are benefits of the sketching. For the both. Batch and real-time data the sketch can be used.

Histograms

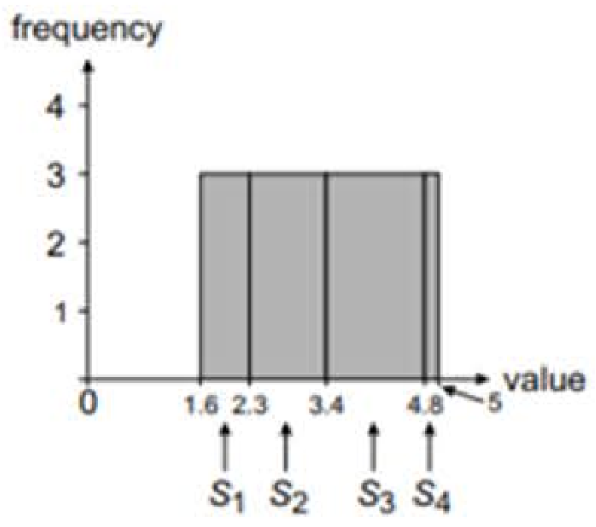
Histograms are going to summarize the dataset by grouping the different data values in buckets (small subsets). Then, Some statistics is used to estimate the result or for trying to reconstruct the data of the bucket. Like sampling, the histograms have been studied for many years and they have been included in query optimizers.

Example of a data set:
D={1.61, 1.72, 2.23, 2.33, 2.71, 2.90, 3.41, 4.21, 4.70, 4.82, 4.85, 4.91}

Equi-width histograms: we have a set of points and each point is assigned into one specific bucket. The size of the buckect are predifined and an example data set with equi-width hitograms would be like this:



Equi-depth histograms – it is for bound the worst-case estimation error. We are going to insert the same amount of point in each bucket. The data set in this way would be like this.



Histograms are very used for answering count or average queries. However, histograms are very bad for multimensional queries. In the case of equi-width the problem is if the are a lot of points in one boundary of the bucket (increase a lot the error).

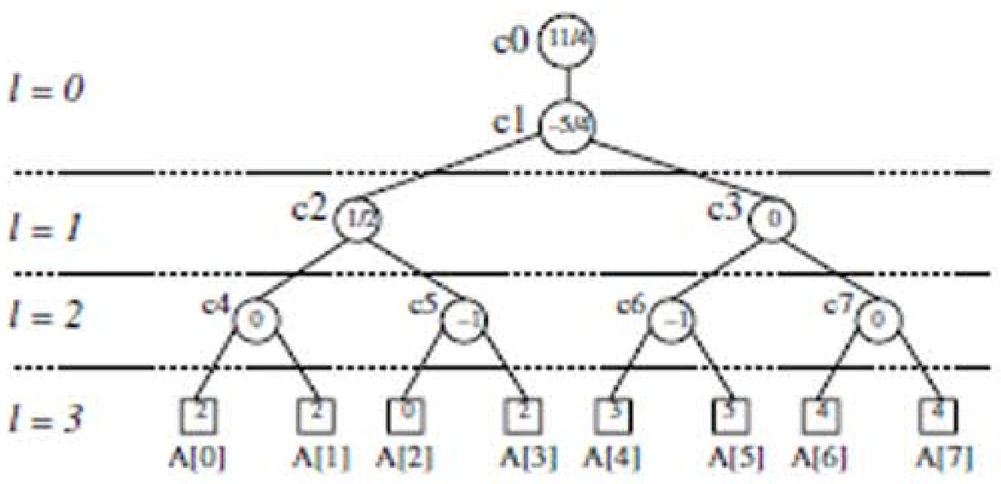
Wavelets

However wavelets is originally introduced for image and signal processing, it is applied to various fields nowadays, such as network management, computer vision, data mining, fingerprint verification, DNA analysis, protein analysis, speech recognition, data compression and so on. It provides effective, efficient, accurate solutions. The main idea here is that wavelet transformation is applied to the given data vector with M elements and coefficients are founded. If the coefficient is lower than the threshold, then it is eliminated. Wavelets is efficient and simple to use.

Assume that there is a one-dimensional data vector A with 8 elements.
A = [2,2,0,2,3,5,4,4]

Resolution	Averages	Detail Coefficients
3	[2, 2, 0, 2, 3, 5, 4, 4]	—
2	[2, 1, 4, 4]	[0, −1, −1,0]
1	[3/2, 4]	[1/2, 0]
0	[11/4]	[−5/4]

Haar Wavelet Transform(HWT) is a tool to apply discrete wavelet transformation. The one-dimensional HWT of A is
WA = [11/4, −5/4, 1/2, 0, 0, −1, −1, 0].



The Future of Data Synopses

Are information management systems going to increase the usage of approximate queries, or will they only be interested in exact answer?

The main challenge:

How present the result to the user? The user are used to exact queries, data synopses must find the most comprehensive way to write the result.

Take advantage of the current technologies and architectures, with parallelism we could reduce the cost.

The integration of the approximate queries is not only going focus on answering queries of end user but integrate synopsis into processing engine.

We have seen approximate queries in exact and determnisitic data. Nowadays, uncertained data set have appeared, so it would be interested to apply approximate queries over uncerained data sets.