Streaming Algorithms

- Streaming Model of Computation
- Streaming Algorithms and DFA
- Stream: Motivation and Applications
- Synopsis: Sliding Window, Histogram, Wavelets
- Sampling from Stream: Reservoir Sampling
- Linear Sketch
- Count-Min Sketch
- AMS Sketch

Imdad ullah Khan

Synopsis

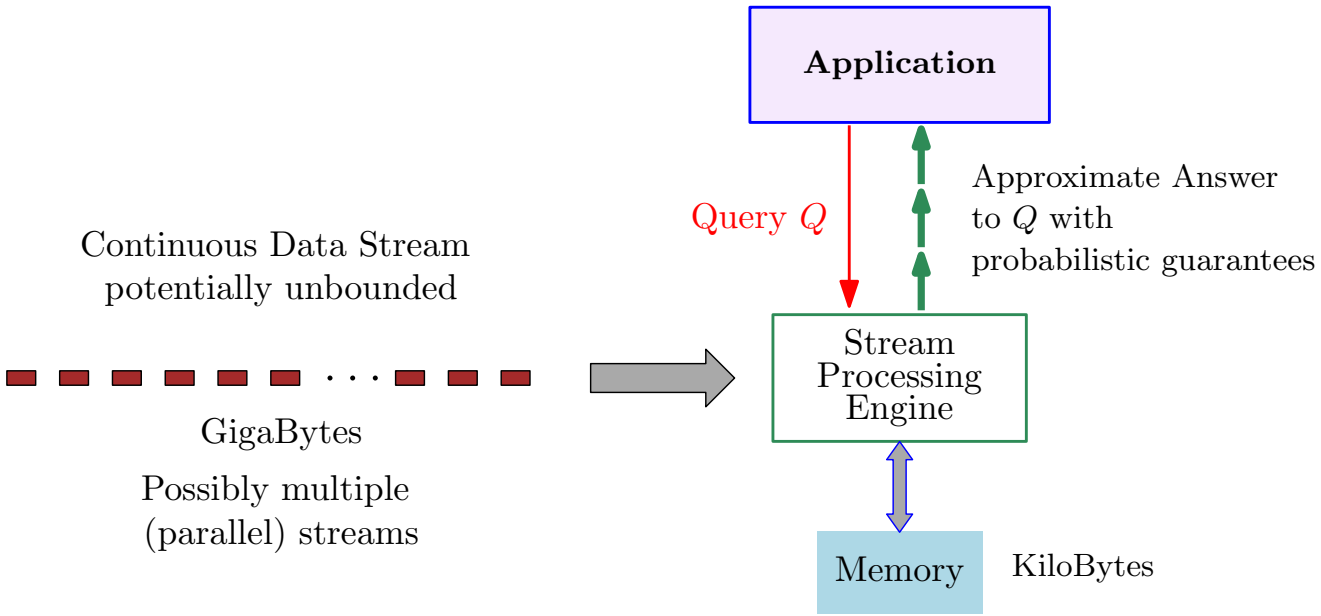
Stream Computation: Synopsis

Fundamental Methodology: Keep a synopsis of the stream and answer query based on it. Update synopsis after examining each item in $O(1)$

Synopsis: Succinct summary of the stream (so far) (poly-log bits)

Families of Synopsis

- Sliding Window
- Random Sample
- Histogram
- Wavelets
- Sketch



Synopsis Based Exact Stream Computation

- Length of $\mathcal{S}$ (m): Computed by storing a running counter
- Sum of $\mathcal{S}$: Computed by storing a running sum
- Mean of $\mathcal{S}$: Computed from sum and length of $\mathcal{S}$
- Variance of $\mathcal{S}$: Computed from sum, sum of square, and length of $\mathcal{S}$

$$Var(X) = E(X^2) - (E(X))^2$$

Synopsis Based Exact Stream Computation

Missing Element:

$n-1$ unique integers are streamed in from $[n]$

Find the missing integer?

- Trivial to find it if we use n bits
- A better solution is to save sum S of the stream ▷ $O(\log n)$ bits

  The missing integer is $n(n+1)/2 - S$
- Can do it in exactly $\log n$ bits by storing the parity sum of each bits

  The final parity sum is the missing integer

Synopsis Based Exact Stream Computation

Two Missing Elements

$n-2$ unique integers are streamed in from $[n]$

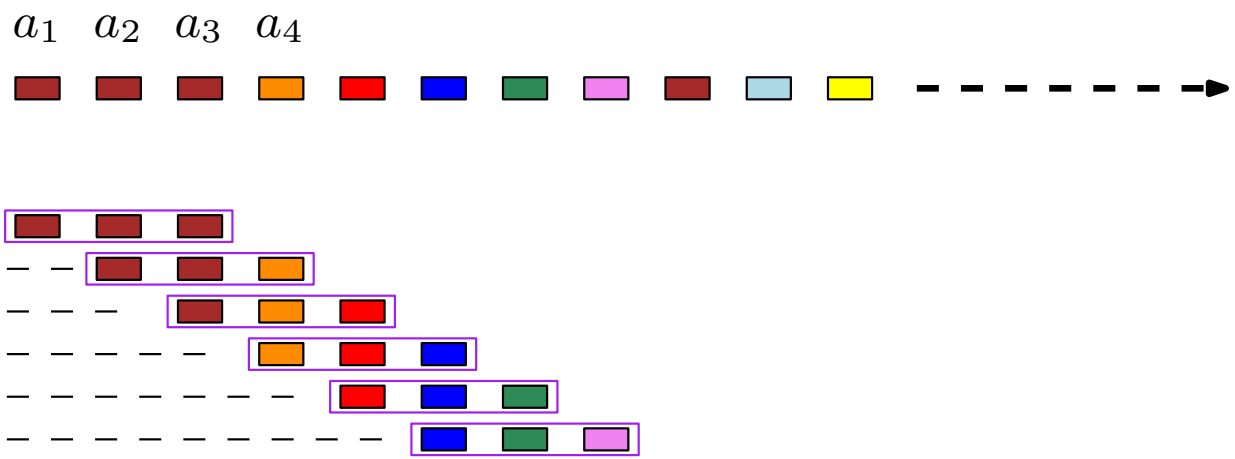
Find the missing integers?

- Trivial to find it if we use n bits
- Save sum of 1st and 2nd powers of stream elements $\triangleright$ $O(\log n)$ bits

  The missing integers are solution to 2 unknowns and two equations
- Readily generalizes to k missing elements

Synopsis: Sliding Window

- Keep the last w elements as synopsis (w is length of window)
- On input a_i ($i \geq w$), a_{i-w} expires and a_i added to window
- Can be used for queries like mean, sum, variance, count of pre-specified element(s) (e.g. non-zero, even)
- Extended to compute approximate median, and k-median



Synopsis: Histogram and Wavelets

Histogram

- The synopsis is some summary statistics (e.g. frequency, mean) of groups (subsets, buckets) in streams values
  - Equi-width histogram
  - Equidepth histogram
  - V-optimal histogram
  - Multi-dimensional histogram

Wavelets

- Essentially histograms of features (coefficients) in the frequency domain representation of the stream