Mining Data Streams

The Stream Model
Sliding Windows
Counting 1's

Cloud and Big Data Summer School, Stockholm, Aug. 2015

Jeffrey D. Ullman

Data Management Vs. Stream Management

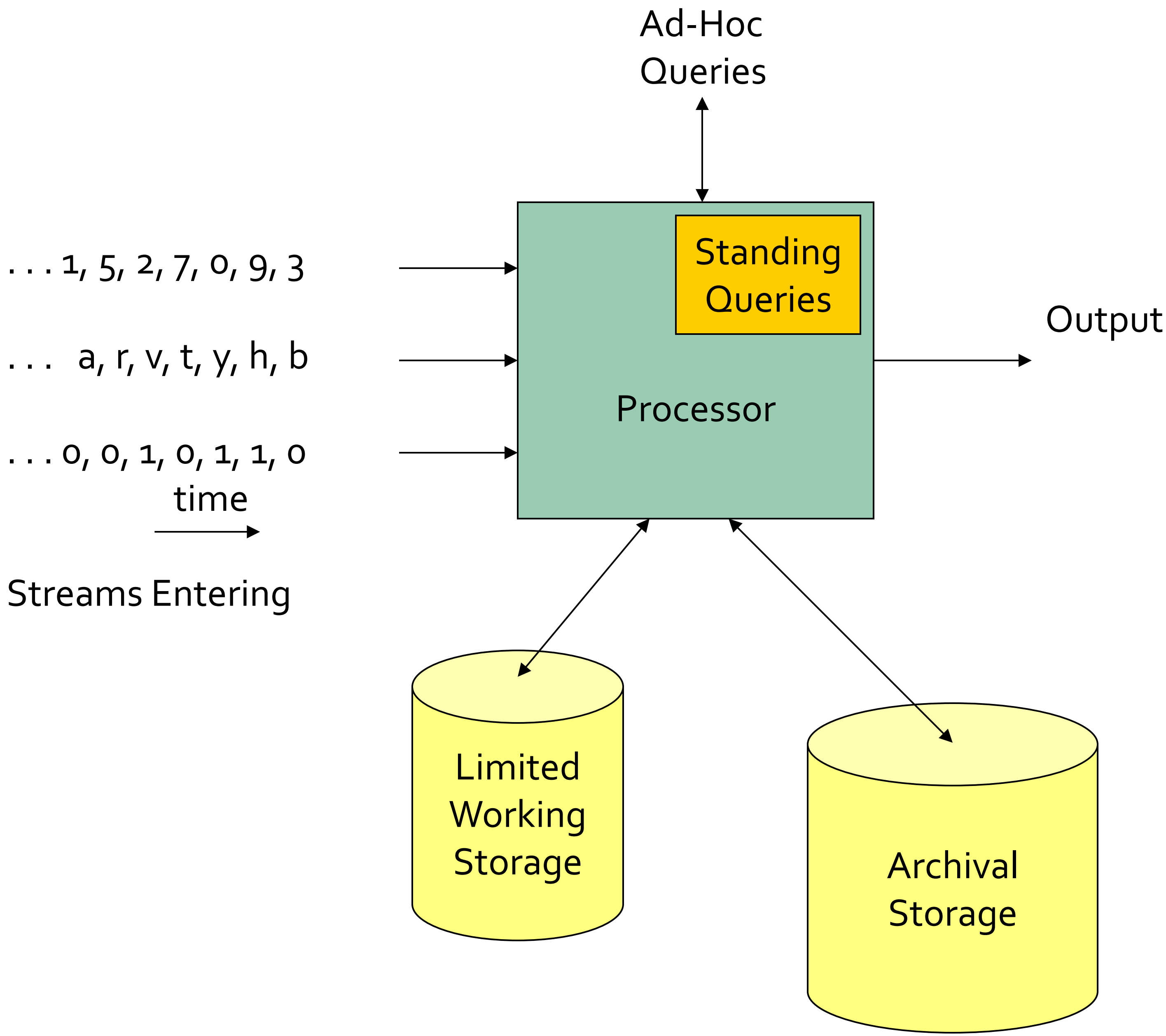
- In a DBMS, input is under the control of the programming staff.
  - SQL INSERT commands or bulk loaders.
- Stream management is important when the input rate is controlled externally.
  - Example: Google search queries.

The Stream Model

- Input tuples enter at a rapid rate, at one or more input ports.
- The system cannot store the entire stream accessibly.
- How do you make critical calculations about the stream using a limited amount of (primary or secondary) memory?

Two Forms of Query

1. *Ad-hoc queries*: Normal queries asked one time about streams.
   - Example: What is the maximum value seen so far in stream *S*?
2. *Standing queries*: Queries that are, in principle, asked about the stream at all times.
   - Example: Report each new maximum value ever seen in stream *S*.


Ad-Hoc
Queries
Standing
Queries
Output
Processor
. . . 1, 5, 2, 7, 0, 9, 3
. . . a, r, v, t, y, h, b
. . . 0, 0, 1, 0, 1, 1, 0
time
Streams Entering
Limited
Working
Storage
Archival
Storage

Applications

- Mining query streams.
  - Google wants to know what queries are more frequent today than yesterday.
- Mining click streams.
  - Yahoo! wants to know which of its pages are getting an unusual number of hits in the past hour.
    - Often caused by annoyed users clicking on a broken page.
- IP packets can be monitored at a switch.
  - Gather information for optimal routing.
  - Detect denial-of-service attacks.

Sliding Windows

- A useful model of stream processing is that queries are about a *window* of length N – the N most recent elements received.
  - Alternative: elements received within a time interval T.
- Interesting case: N is so large it cannot be stored in main memory.
  - Or, there are so many streams that windows for all do not fit in main memory.

q w e r t y u i o p a s d f g h j k l z x c v b n m

q w e r t y u i o p a s d f g h j k l z x c v b n m

q w e r t y u i o p a s d f g h j k l z x c v b n m

q w e r t y u i o p a s d f g h j k l z x c v b n m

← Past Future →

Example: Averages

- Stream of integers, window of size *N*.
- Standing query: what is the average of the integers in the window?
- For the first *N* inputs, sum and count to get the average.
- Afterward, when a new input *i* arrives, change the average by adding (i - j)/N, where *j* is the oldest integer in the window.
- Good: O(1) time per input.
- Bad: Requires the entire window in memory.

Counting 1's

Approximating Counts
Exponentially Growing Blocks
DGIM Algorithm

Counting Bits

- Problem: given a stream of 0's and 1's, be prepared to answer queries of the form "how many 1's in the last k bits?" where $k \leq N$.
- Obvious solution: store the most recent N bits.
- But answering the query will take $O(k)$ time.
  - Very possibly too much time.
- And the space requirements can be too great.
  - Especially if there are many streams to be managed in main memory at once, or N is huge.

Example: Bit Counting

- Count recent hits on URL’s belonging to a site.
- Stream is a sequence of URL’s.
- Window size N = 1 billion.
- Think of the data as many streams – one for each URL.
- Bit on the stream for URL x is 0 unless the actual stream has x.

DGIM Method

- Name refers to the inventors:
  - Datar, Gionis, Indyk, and Motwani.
- Store only $O(\log^2 N)$ bits per stream (N = window size).
- Gives approximate answer, never off by more than 50%.
  - Error factor can be reduced to any fraction > 0, with more complicated algorithm and proportionally more stored bits.

Timestamps

- Each bit in the stream has a *timestamp*, starting 0, 1, …
- Record timestamps modulo N (the window size), so we can represent any *relevant* timestamp in $O(\log_2 N)$ bits.

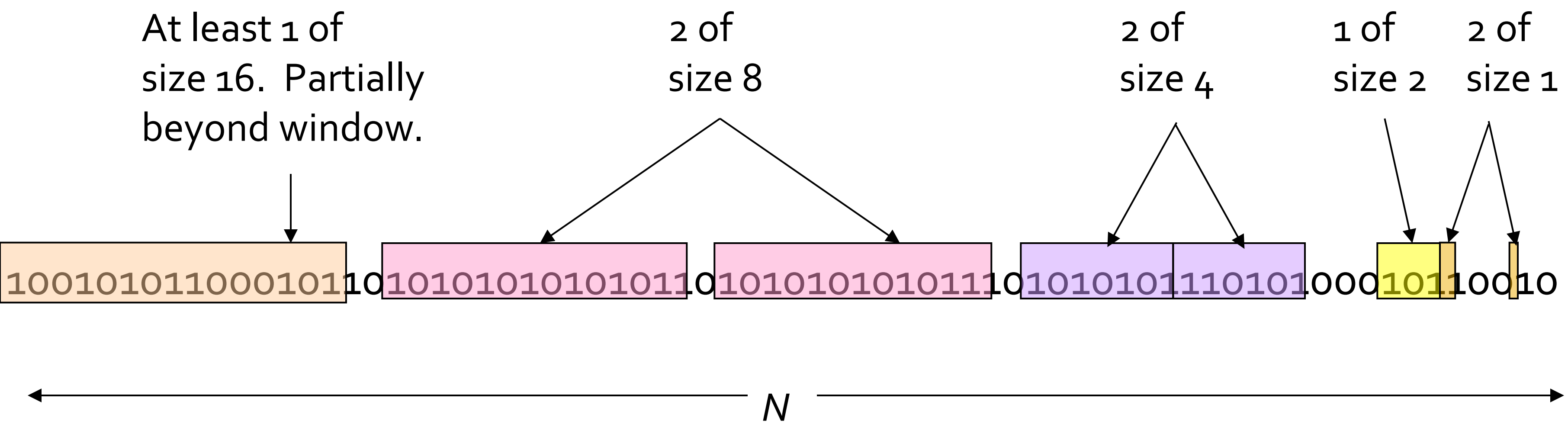
Buckets

- A *bucket* is a segment of the window; it is represented by a record consisting of:
  1. The timestamp of its end [O(log N) bits].
  2. The number of 1's between its beginning and end.
     - Number of 1's = *size* of the bucket.
- Constraint on bucket sizes: number of 1's must be a power of 2.
  - Thus, only O(log log N) bits are required for this count.

Representing a Stream by Buckets

- Either one or two buckets with the same power-of-2 number of 1's.
- Buckets do not overlap.
- Buckets are sorted by size.
  - Older buckets are not smaller than newer buckets.
- Buckets disappear when their end-time is > *N* time units in the past.

Example: Bucketized Stream



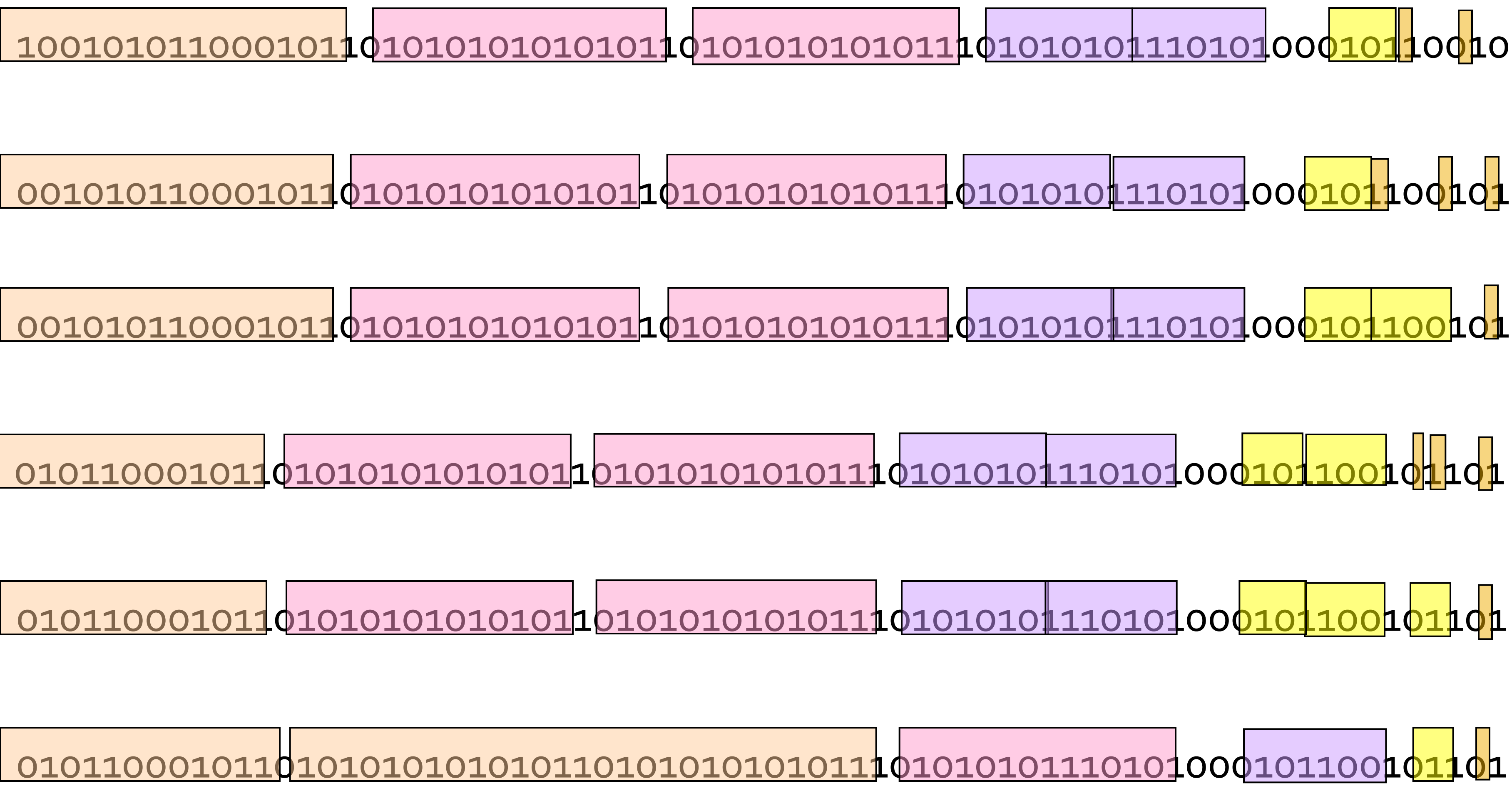
Updating Buckets

- When a new bit comes in, drop the last (oldest) bucket if its end-time is prior to N time units before the current time.
- If the current bit is 0, no other changes are needed.

Updating Buckets: Input = 1

- If the current bit is 1:
  1. Create a new bucket of size 1, for just this bit.
     - End timestamp = current time.
  2. If there are now three buckets of size 1, combine the oldest two into a bucket of size 2.
  3. If there are now three buckets of size 2, combine the oldest two into a bucket of size 4.
  4. And so on …

Example: Managing Buckets



Querying

- To estimate the number of 1's in the most recent $k \leq N$ bits:
  1. Restrict your attention to only those buckets whose end time stamp is at most k bits in the past.
  2. Sum the sizes of all these buckets but the oldest.
  3. Add half the size of the oldest bucket.
- Remember: we don't know how many 1's of the last bucket are still within the window.

Error Bound

- Suppose the oldest bucket within range has size 2^i.
- Then by assuming 2^{i-1} of its 1's are still within the window, we make an error of at most 2^{i-1}.
- Since there is at least one bucket of each of the sizes less than 2^i, and at least 1 from the oldest bucket, the true sum is no less than 2^i.
- Thus, error at most 50%.