Mining Data Streams

The Stream Model
Sliding Windows
Counting 1's

Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

Data Management Vs. Stream Management

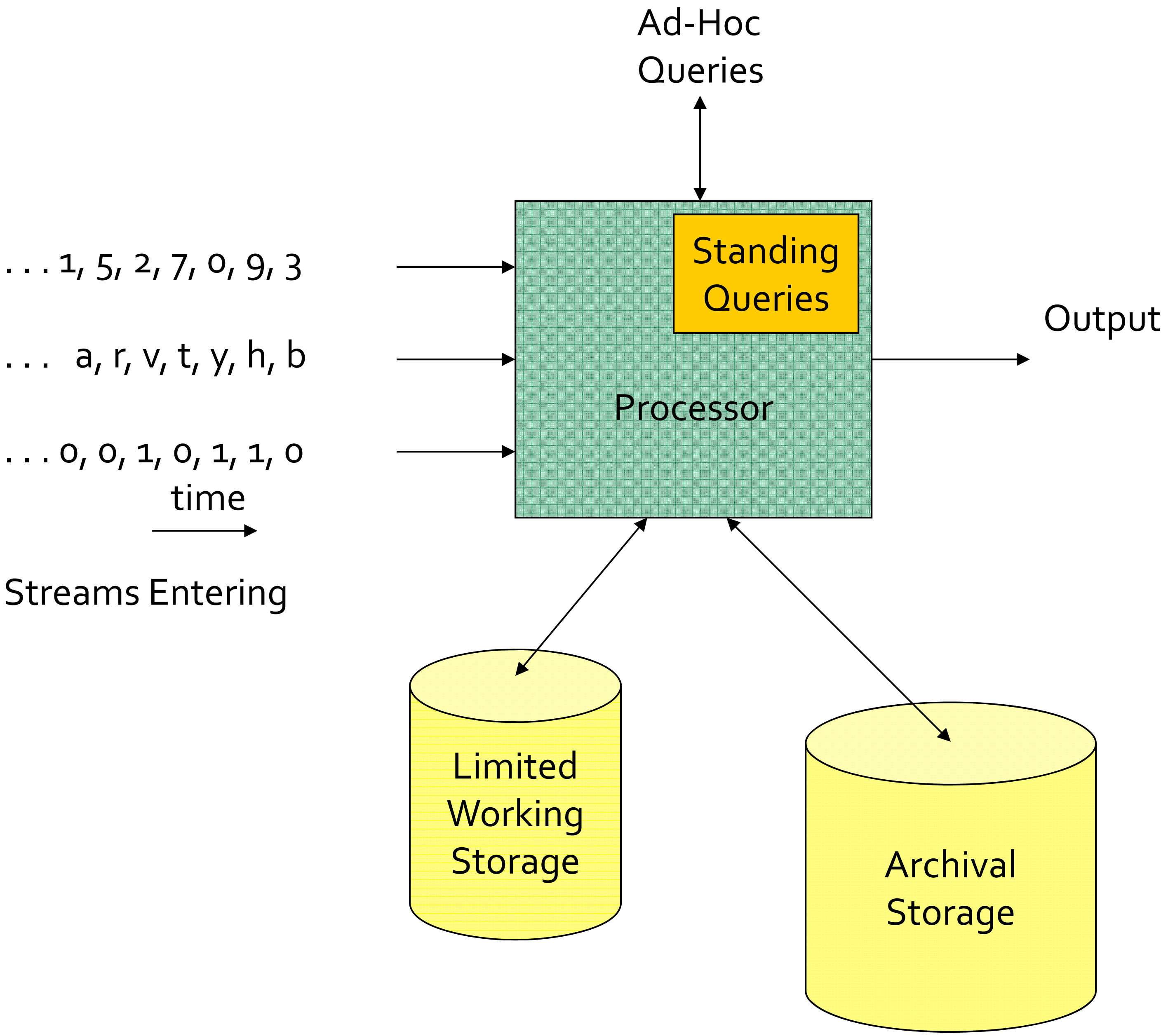
- In a DBMS, input is under the control of the programming staff.
  - SQL INSERT commands or bulk loaders.
- Stream Management is important when the input rate is controlled externally.
  - Example: Google search queries.

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*.

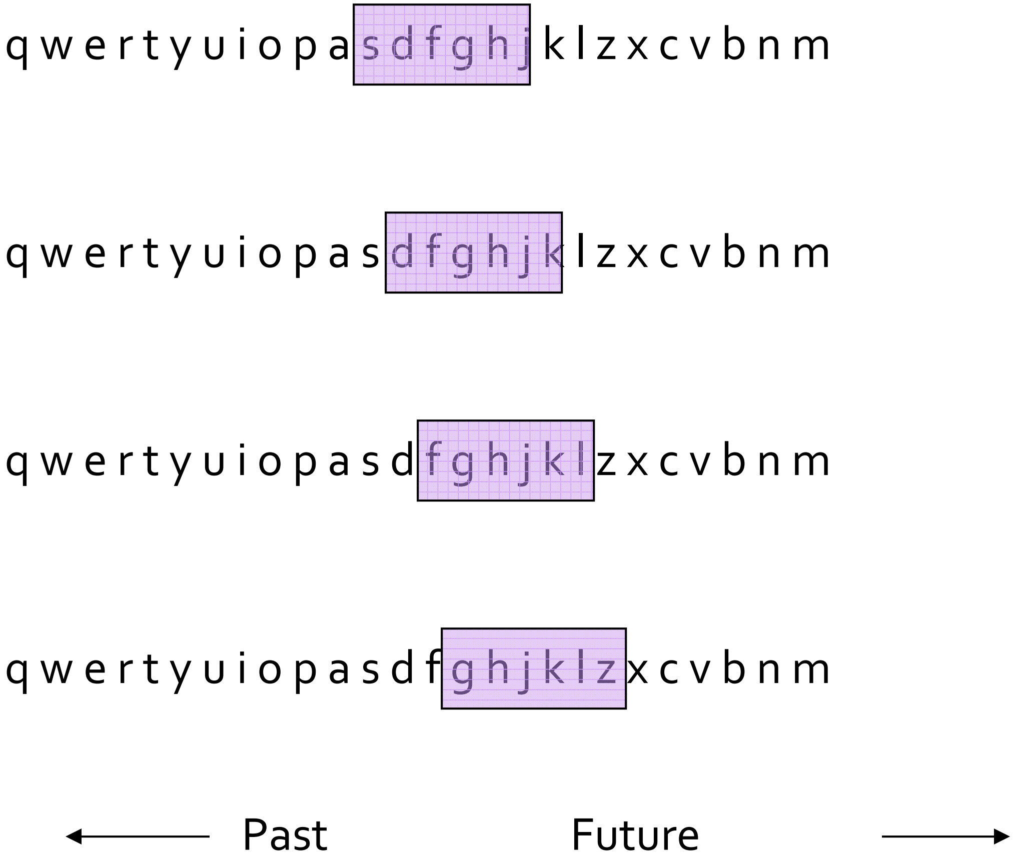

Ad-Hoc Queries
Standing Queries
Output
Processor
. . . 1, 5, 2, 7, 0, 9, 3
. . . a, r, v, t, y, h, b
. . . 0, 0, 1, 0, 1, 1, 0
time
Streams Entering
Limited Working Storage
Archival Storage

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.
- 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 cannot be stored.



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.