

Research Day July 2006

ANIMA: Adaptive Netting In
stream dAta

Ben Schroeder

Presentation Outline

- Data Stream Processing Challenges
- Abstract Reference Data Stream Processing Architecture
- The current state of ANIMA in the stream processing paradigm
- Next Steps

Data Streams vs. Static Data

- Static Data:
 - Fixed, finite datasets
 - Whole dataset is available and supposed to be exact
 - Random access is possible
 - Most processing techniques for static data require *multiple passes* over the data or at the very least the *whole dataset* to be available
- Data Streams:
 - Potentially infinite; end is unknown until it is actually reached
 - Cannot store whole data stream
 - Cannot randomly access data stream
 - Variable and possibly very high data rate

Data Stream Processing: Challenges

- Large amounts of data in short time
- Limited resources: cannot store whole data stream
- Limited access: cannot randomly access data stream
- Often real-time requirements

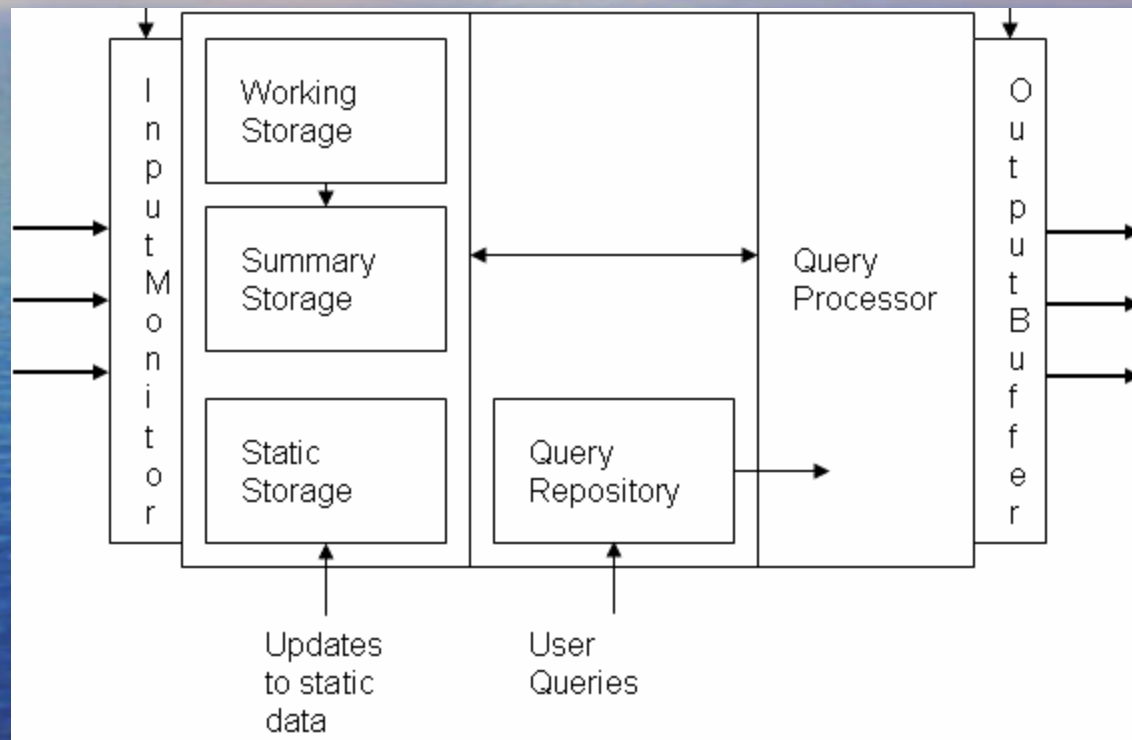
Data Stream Processing: Requirements

- Small time per record (ideally constant)
 - Fast processing algorithms
- Use fixed amount of main memory
 - Memory bound processing algorithms
- Only a single sequential scan of the data is possible
 - Incremental approach
 - Sliding window approach
- Cope with high and changing data rates
 - Input buffers
 - Dropping packets/records
- Two types of queries
 - Continuous Queries
 - Ad-hoc queries

Data Stream Processing

- Two approaches:
 - General Purpose Stream Processing (DSMS)
 - STREAM, CQL (Stanford, halted)
 - AURORA (MIT, Brandies, Brown)
 - Application Specific Stream Processing
 - Traderbot (real time stock market analysis tool able to find emerging patterns)
 - Network Intrusion, Clickstreams, sensor monitoring, weather forecasts,...

Abstract Reference Data Stream Processing Architecture



- L. Golab and M.T. Özsu. Issues In Data Stream Management. *SIGMOD Record*, Vol.32, No.2, June 2003.

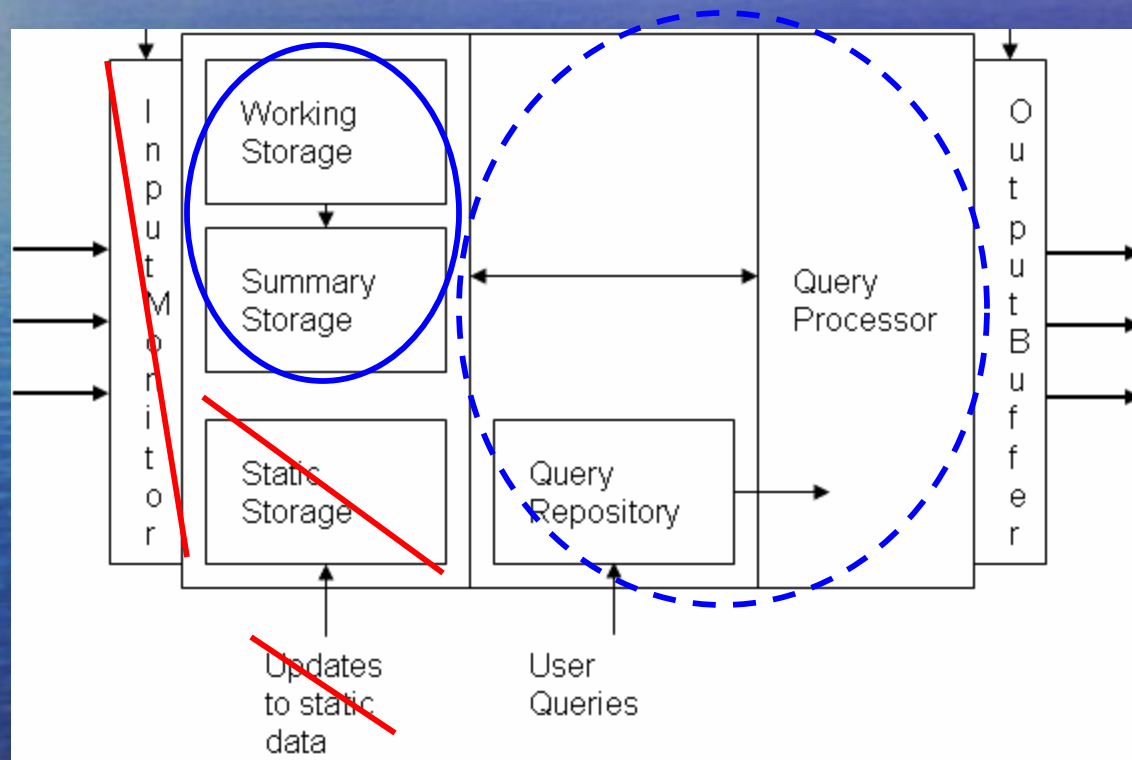
Abstract Reference Data Stream Processing Architecture

- Input Monitor: regulate input, dropping packets/transactions, preprocessing, buffering...
- Temporary Working Storage: window queries
- Summary Storage: stream synopses
- Static Storage: Meta-data about the stream, user knowledge, expert input...

Abstract Reference Data Stream Processing Architecture

- Query Repository: for long-running, continuous queries
- Query Processor: processes continuous queries and ad-hoc queries
- Output Buffer: streams query results to the user

ANIMA in the Stream Processing Architecture



ANIMA in the Stream Processing Architecture

- ANIMA Network = Hybrid of Working Storage and Summary Storage
- More intuitive than sliding window (Working Storage)
 - Sliding window discriminates based on time
 - ANIMA discriminates based on importance/frequency of occurrence and time
- More precise than Full Summary Storage
 - A full summary may contain a lot of unimportant/uninteresting/outdated information

ANIMA in the Stream Processing Architecture

- Queries
 - currently queries are built in
 - full custom user queries are not possible
 - User can change only certain parameters of queries

Current State of ANIMA

- Working prototype of Short-term/working memory
- Built-in query: stream top nodes and connections in “real-time” to the user
- Takes file as input
- User can set functions to update weights of nodes and connections freely

Current State of ANIMA

Demonstration

Next Steps

- Implementation Related (short term)
 - Thoroughly test current implementation with different sets of data
 - Extend implementation to sequences and directed networks
 - Experiment with adaptive weights and introduce parameters allowing to specify the desired size of the network (#nodes, #connections)
 - Provide a dynamic visual representation of the network

Next Steps

- ANIMA:
 - Integration of ANIMA in other projects
 - Finding strongly connected subgraphs + Long-term memory => Finding temporal patterns, detecting concept drift,...
 - Find a specific application area for ANIMA

Next Steps

- Redesign WIKI pages (content & layout) and keep them up to date
- Prepare Teaching for Winter 2006/2007
- Write document on data streams and ANIMA and keep it up to date

References

- B. Babcock, S. Babu, M. Datar, R. Motwani and J. Widom. **Models and issues in data stream management.** *Proceedings of PODS*, 2002.
- L. Golab and M.T. Özsu. **Issues In Data Stream Management.** *SIGMOD Record*, Vol.32, No.2, June 2003.
- J. Hsu. **Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st century.** *The Proceedings of ISECON*, v19, San Antonio, 2002.
- S. Mutukishnan. **Data Streams: Algorithms and Applications.** *Proceedings of the 14th annual ACM-SIAM symposium on discrete algorithms*, 2003.
- P. Domingos and G. Hulten. **Catching Up with the Data: Research Issues in Mining Data Streams.** *Workshop on Research Issues in Data Mining and Knowledge Discovery*, Santa Barbara, 2001.