Sensor Data Streams

Redes de Sensores Sem Fio

Helen Peters de Assunção

Jeferson Moreira dos Anjos

Data Stream Systems

- Nova classe de aplicações:
  - Dados chegando rapidamente, em intervalos variáveis e com fluxo ilimitado.
  - Dados são melhores modelados não como relações persistentes, mas como streams de dados transientes.
  - SGBS´s tradicionais não foram desenvolvidos para armazenar dados de forma contínua e rápida e não suportam consultas contínuas.
  - Exemplos: Monitoração de redes, redes de sensores, aplicações web, etc.

Modelo de Stream de Dados

- Um fluxo contínuo de dados.
- Não existe controle da ordem de chegada de cada elemento a ser processado.
- Stream de dados tem tamanho ilimitado.
- Uma vez processado, um elemento é geralmente descartado. Ele não é recuperado a não ser que seja armazenado em memória, que é tipicamente pequena para o tamanho dos dados que chegam.
- Consultas sobre essas streams precisam ser processadas quase que em tempo real (por representar um evento do mundo real ou por ser caro armazenar os dados).

Redes de Sensores Sem Fio

- RSSFs consistem, tipicamente, em alguns milhares de sensores que coletam e comunicam, continuamente, seus dados para uma estação-base.
- Devido ao seu “baixo custo”, espera-se que os dispositivos sensores se tornem pervasivos.

  ➡ Uma das principais **fontes de informação** para banco de dados.

- Streams de dados que necessitam ser agregadas, monitoradas e analisadas.

Limitações das RSSFs

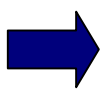
- Energia.
- Largura de banda.
- Capacidade de processamento.
- Capacidade de armazenamento.
- Perda de pacotes (Falhas de conexão, suavização do sinal, colisão de pacotes, interferência, …).
  - > 10% dos links sofrem uma perda média > 50%.
- Topologia mudando continuamente (falha de nós, mobilidade).

Qualidade dos Dados

- Essas limitações causam problemas:
  - Perda de informações relevantes.
  - Recursos limitados: *O nó poderá sensoriar?*
  - Condições do ambiente incontroláveis: ruído.
  - Problemas de HW e rádio.
  - Tecnologia atual: sensores baratos e de baixa qualidade.
  - Nós maliciosos.

- Dados incompletos ou imprecisos

Falso positivo/negativo



Problema para tomada imediata de decisões

Sensor Streaming

- Dados são apenas amostras de um conjunto de dados. A taxa de amostragem varia de aplicação para aplicação.
- Datas geralmente imprecisos e com ruído.
- Tamanho moderado

  (aplicações atuais).
- Aquisição de dados tem um custo.

x

Traditional Streaming

- Todos os dados estão disponíveis (ex: log).
- Dados exatos e livres de erro.
- Grande quantidade de dados a serem armazenados e processados.
- Aquisição de dados sem custo adicional.

Sensor Stream System

- Sistemas que extraem dados dos sensores e permitem que usuários observem, analisem e consultem estes dados.
  - Eficientes em energia
  - Escaláveis
  - Auto-organizáveis e robustos contra falha de nós e mudanças de topologia.

Sensor Stream System

- Prover armazenamento persistente e realizar consultas como um sistema centralizado provê é impossível para uma RSSF.
  - In-network Storage
  - In-network Aggregation

In-network Storage

- Armazenamento:
  - Externo: Dados enviados continuamente ao ponto de acesso - custo com transmissão.
  - Local: Dados são armazenados no nó de origem – custo com consulta.
  - Data Centric Storage (DCS): Um nó armazena os dados de um conjunto de nós – custo com transmissão e com consulta amenizados.

Energia x Armazenamento

Modelo de Armazenamento Baseado em Predição

- PREMON (PREdiction-based MONitoring)
  - A estação base prediz dados futuros baseados em snapshots dos dados sensoriados e envia essas predições aos nós sensores.
  - Os sensores não enviam dados se o valor estiver coletado for próximo ao previsto, dado um limite predefinido.
  - Reduz custo de comunicação

In-network Aggregation

- É um mecanismo para reduzir a quantidade total de energia e banda necessárias para processar uma consulta de um usuário, permitindo que os nós façam agregação intermediária dos dados.
- Uma consulta é enviada a rede ou a uma área específica, e a resposta é roteada por uma árvore onde é possível realizar a agregação de dados.

Direct Delivery

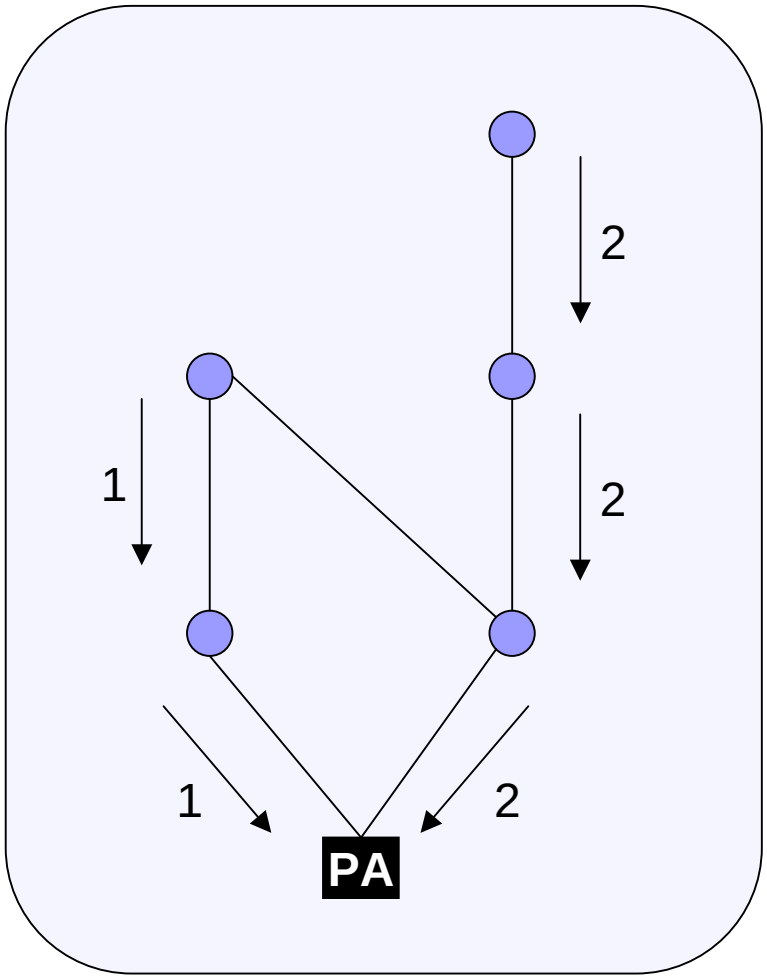
- Direct Delivery:
  - Cada nó responde a uma consulta requisitando dados. O ponto de acesso agrega esses dados e entrega a um usuário.
- Desvantagens:
  - Grande número de pacotes trafegando na rede.

In-network Aggregation - Vantagens

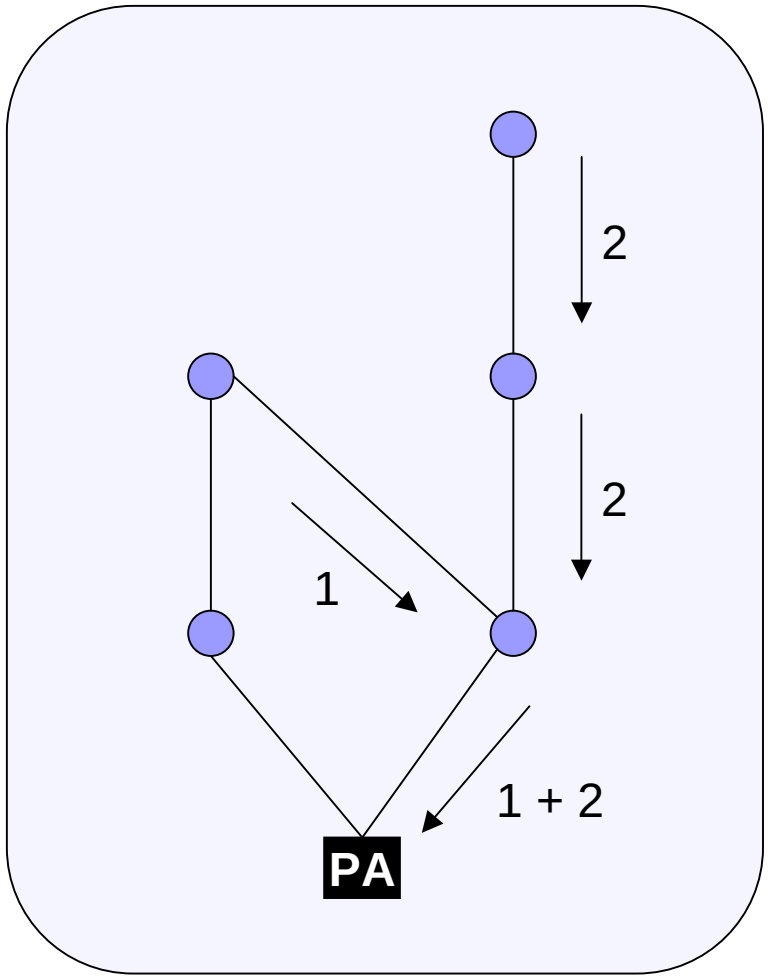
- Redução do número de pacotes enviados pela rede.
- Redução da probabilidade de colisão de pacotes.
- Redução de dados redundantes recebidos no ponto de acesso.

Roteamento

- Abordagem centrada no endereço:
  - Encontrar rotas curtas entre pares de nós endereçáveis.
- Abordagem centrada em dados
  - Encontrar rotas de múltiplos nós para um único destino que permita a agregação de dados redundantes dentro da rede.



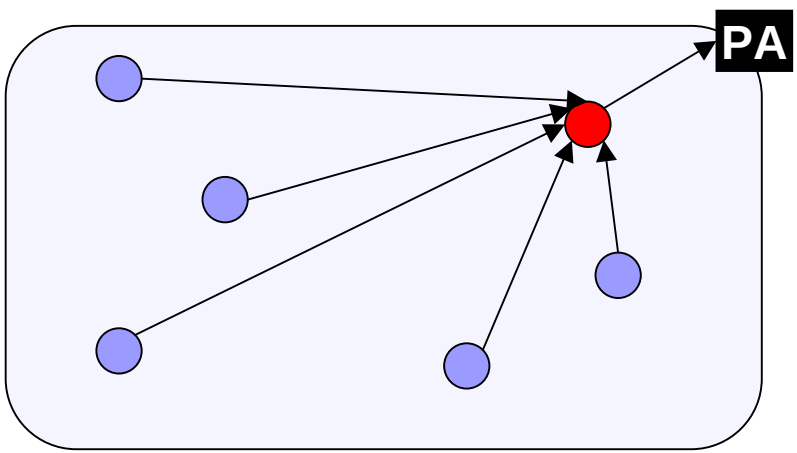
Address-Centric Routing

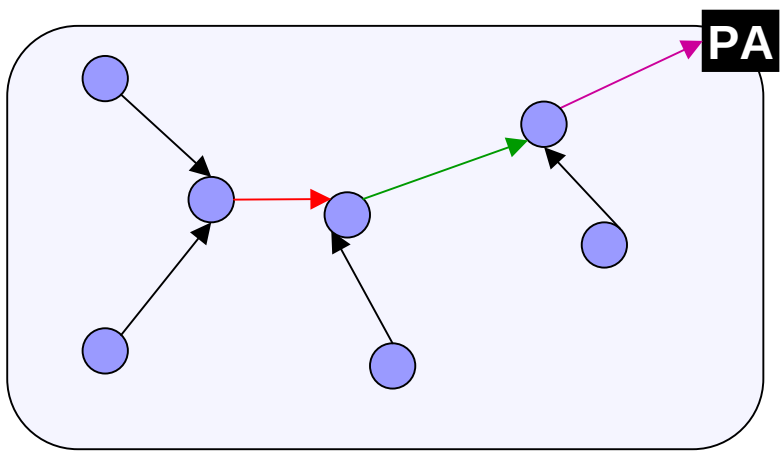


Data-Centric Routing

Algoritmos

- Center at Nearest Source (CNS)
  - Nó mais próximo ao sink é o ponto de agregação.
- Shortest Paths Tree
  - Cada nó utiliza uma rota de caminho mínimo. Rotas que se sobrepõem têm dados agregados.





Funções de Agregação

- Funções de Agregação
  - MIN, MAX, MEDIAN, AVERAGE, COUNT, …

- Propriedades:
  - Duplicate sensitivity.

  A função de agregação retorna o mesmo resultado quando os conjuntos apresentam valores duplicados (MEDIAN, AVERAGE, COUNT).

Funções de Agregação

- Propriedades:
- Exemplary/Summary.
  - Exemplary retorna um valor representativo do conjunto de dados (MIN, MAX, MEDIAN).
  - Sumary realiza algum calculo sobre todo o conjunto de dados e retorna o valor calculado (AVERAGE, COUNT) .
- Monotonic aggregates.
  - Permite teste de predicados na rede antes do envio dos dados
  - (Ex.: Enquanto nós enviam seus valores para uma consulta MAX, outros nós só enviam seus próprios valores se forem maiores que MAX corrente).

Performance

- Fatores que influenciam a performance dos métodos de agregação:
  - Posição dos nós
  - Número de nós
  - Topologia de comunicação da rede
- Economia de energia x Atraso

RSSF como um Banco de Dados

- RSSFs podem ser vistas como um banco de dados.
- Da mesma forma que tabelas de bancos de dados são consultadas, nós sensores podem ser consultados.

```
SELECT nodeid,light,temp FROM sensors

WHERE light > 400

SAMPLE PERIOD 1024
```

RSSF como um Banco de Dados

- Consultas a Data Streams são sempre aproximadas

  Data Streams não tem tamanho limitado, logo quantidade de armazenamento necessário cresce também de forma ilimitada.

  - Sliding Windows: consulta a dados recentes da stream.
  - Batch Processing: Dados são armazenados em um buffer e a consulta é realizada de tempos em tempos.
  - Sampling: Mais dados do que é possível processar – Faz uma amostragem dos dados e realiza a consulta.
  - Synopisis: Consulta realizada sobre uma aproximação dos dados (sinopse).

TinyDB

- TinyDB é um sistema processador de consultas para extração de informações de uma rede de sensores que utilizam o TinyOS.
- Interface SQL-like para especificar os dados a serem extraídos, especificando a taxa que essa consulta deve ser refeita.
- Data uma consulta a dados de interesse o TinyDB coleta os dados dos nós, os filtra, agrega e dissemina para um PC.

Query Constructor

Graphical Interface | Text Interface

Sample Period: 1024

Available Attributes

- nodeid
- light
- temp
- parent
- accel_x
- accel_y

None

>>>

<<<

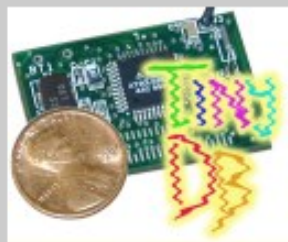
Projected Attributes

- nodeid
- light

Send Query

Display Topology

Magnet Demo



```
SELECT nodeid, light FROM sensors
WHERE light > 200
SAMPLE PERIOD 1024
TRIGGER ACTION SetSnd(500)
```

☐ GROUP BY nodeid

New Predicate

WHERE light > 200 –

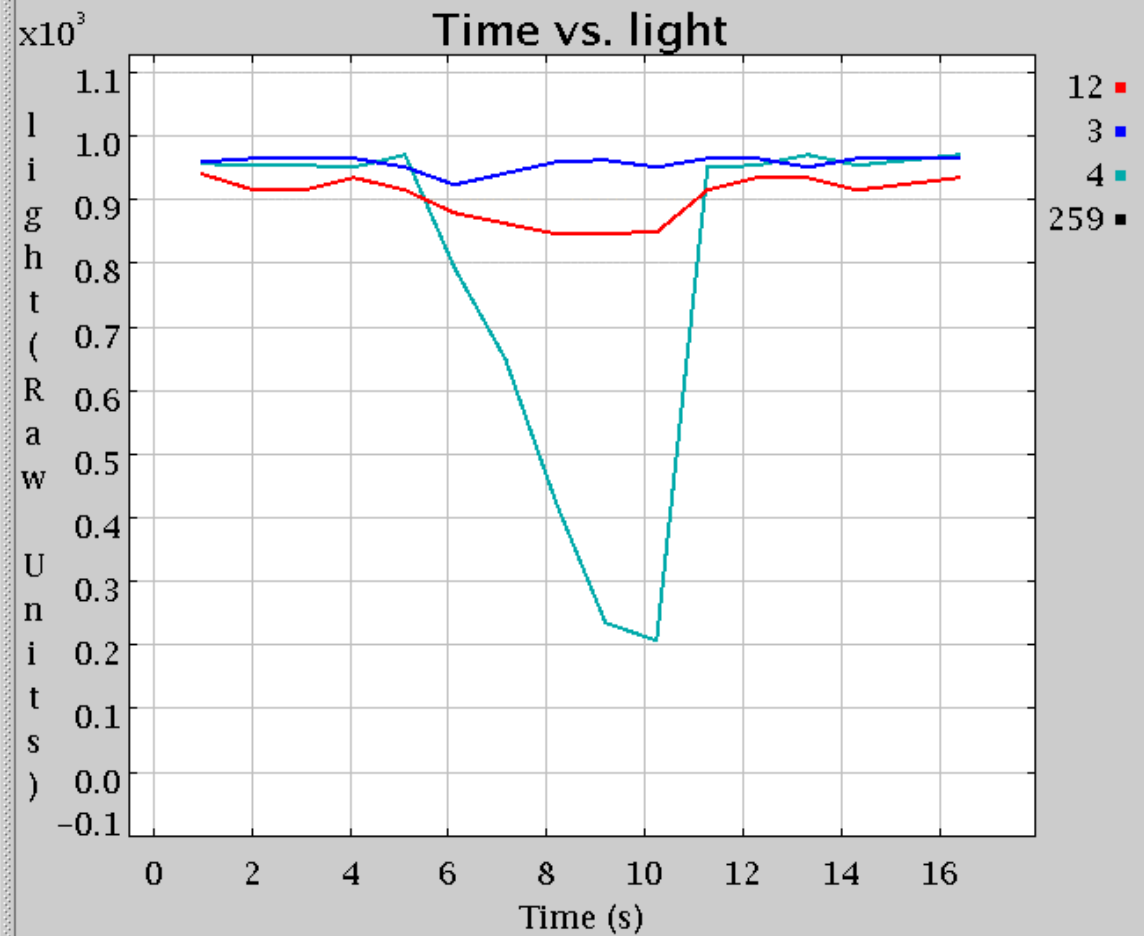
☑ TRIGGER ACTION Sounder (500ms)

☐ Log to Database

Query 0

```
SELECT nodeid, light FROM sensors
WHERE light > 200
SAMPLE PERIOD 1024
```

Epoch	nodeid	light
3	3	964
3	12	915
4	4	952
4	12	934
4	3	964
5	4	970
5	12	915
5	3	952
6	12	878
6	3	923
6	4	793
7	4	648
8	12	846
8	3	959
8	4	423
9	4	234
9	3	962
10	4	207
10	12	847
10	3	950
11	12	915
11	3	964
11	4	952
12	4	953
12	3	964
12	12	934
13	4	970
13	3	951
13	12	934
14	3	964
14	12	915
14	4	953
15	259	951
16	4	969
16	12	934



light | Reset Graph | Clear Graph

Stop Query | Resend Query

Data Streams: Aspectos Formais

- Modelos
- Paradigmas de Programação
- Áreas Relacionadas

Modelos de Data Stream

Um fluxo de entrada a_1,a_2...chega seqüencialmente, item por item, e descreve um sinal *A*, criado por uma função *f(x).* Os modelos de Fluxo de Dados diferem em como os $a_{i's}$ descrevem o sinal *A*.

Modelos de Data Stream

- Modelo de Série Temporal
- Modelo de Agregação
- Modelo *Turnstile*

Modelo de Série Temporal

- Neste modelo os eventos que chegam são armazenados em um vetor cujo índice é incrementado a cada atribuição.
- Este é um modelo apropriado para fluxo de dados, onde está sendo monitorado por exemplo o dados recebido pelo *P.A* em um determinado intervalo de tempo.

$$a_i = A[i]$$

Modelo de Agregação

- Neste modelo os dados que chegam são armazenados em uma túpla <a_i,valor>.
- Para todo evento que chega é atribuído um valor a este e incremento sempre que o mesmo ocorrer.

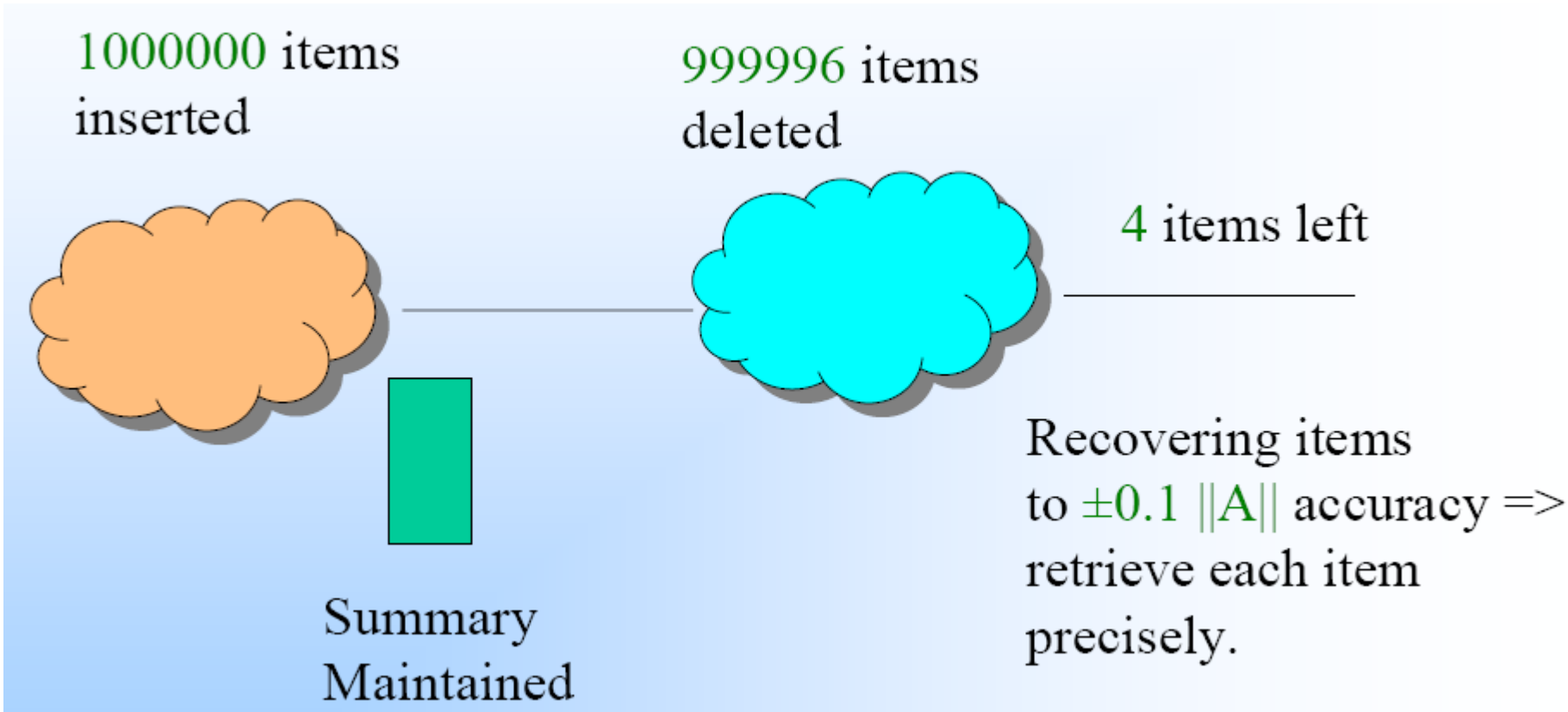
Modelo de Agregação

- Exemplo:
  - Data Stream <nó,pacotes>
    - <2, 10> <4; 15> <3, 13> <2, 23> <4,3>
  - Resultado
    - <2,33> <4,18> <3,13>

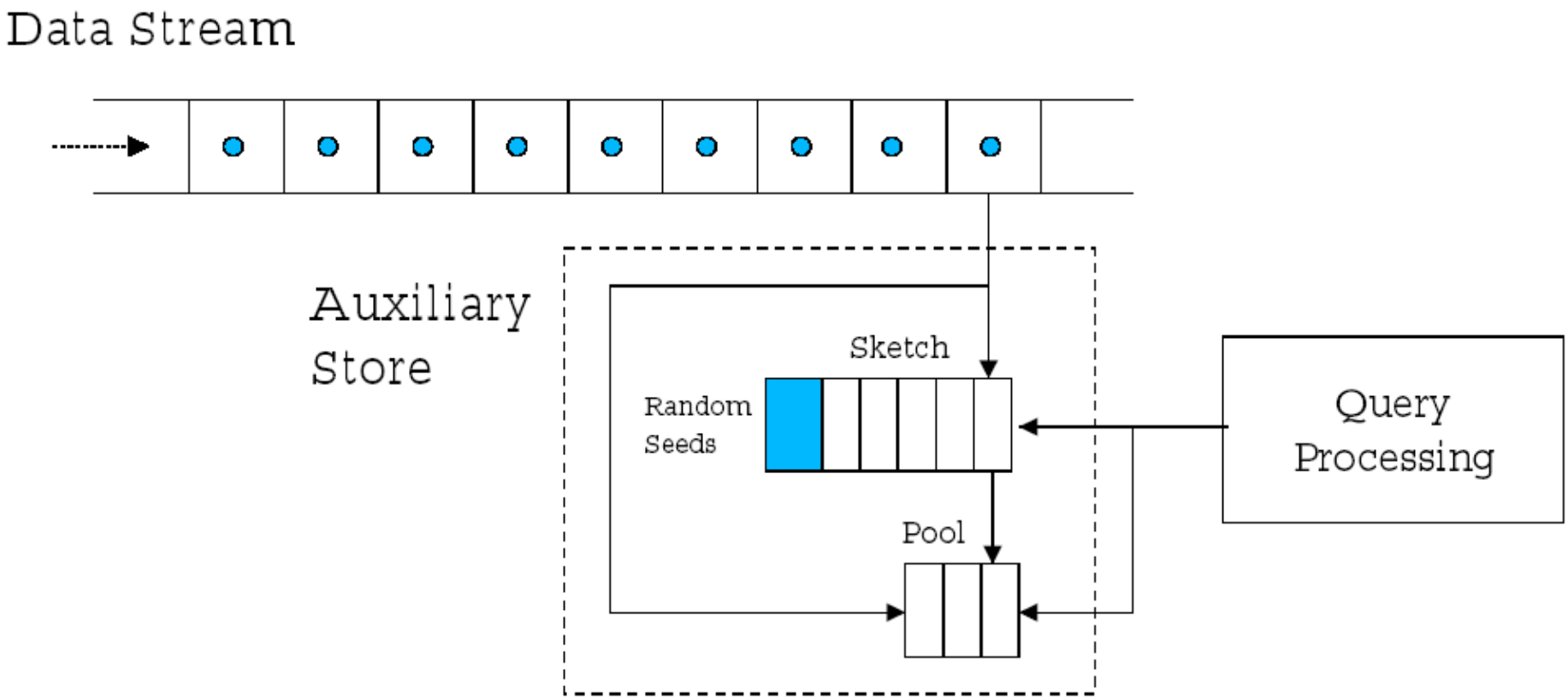
Modelo *Turnstile*

- Este é o modelo geral, onde os a_i's são atualizações de A[j]'s
- É o modelo apropriado para estudar as interações inteiramente dinâmicas.

Modelo *Turnstile*



Modelo de Processamento do Fluxo de Dados



Paradigmas de Programação

- Existe vários paradigmas de programação:
  - Busca binária;
  - Programação dinâmica;
  - Estratégia Gulosa;
  - Divisão e Conquista, etc;
- As que se aplicam diretamente no contexto de Data Stream, normalmente utilizam de amostras e projeções de dados.

Amostras Aleatórias

- É uma estrutura que pode ser utilizada em cenários onde espera-se capturar as características essenciais da série de dados através de uma pequena amostra deste.
- É uma estrutura fácil de implementar em um DSMS.
- A amostragem é uniforme.

Amostras Aleatórias

- Amostragem recente:
  - Reduzir erro devido a variação dos dados;
  - Reduzir erro para grupos de *Queries.*
- Processar uma amostra aleatória sobre um fluxo de dados é relativamente fácil.

Histogramas

- São estruturas freqüentemente utilizadas para capturar a distribuição dos valores em um conjunto de dados.
- Utilizados para varias tarefas como:
  - Estimação do tamanho das *Queries;*
  - Resposta aproximada das *Queries;*
  - Mineração de Dados.

Histogramas

- Há diferentes tipos de histogramas propostos na literatura, entre eles temos:
  - *V-Optimal Histogram*
  - *Equi-Width Histograms*
  - *End-Biased Histograms*

Wavelets

- São ondas pequenas com determinadas propriedades que as tornam adequadas para servirem de base para decomposição de outras funções.
- A analise de sinais com wavelets permite a extração de dados coerentes tanto no domínio da freqüência quanto no do tempo (ou espaço, para imagens).
- Utilizados devido a facilidade de computação.

Wavelets

- Pode ser utilizados para diferentes tarefas como:
  - Predição
  - Aproximação de dados
  - Agregação multi-dimensional

Áreas Relacionadas

- *Probably Approximately Correct (PAC) learning*
- Teste de propriedade
- Métodos *Markov*

PAC *Learning*

- Preocupa-se em decidir quantos dados são necessários coletar para que um “classificador” possa fazer predições corretas em testes futuros .
- É necessário ser feito uma comparação entre aprendizagem utilizando Wavelets com PAC learning.

Teste de Propriedade

- Utiliza algoritmos que focam na amostragem e em processar somente a quantidade necessária de dados
- Para que o tempo de execução cresça lentamente em relação ao tamanho do problema;
- Dá somente um resposta aproximada ou provavelmente correta;

Métodos de Markov

- Utiliza cadeias de Markov para determinar o fluxo de dados em determinado estados;
- Trabalha com verossimilhança;
- É uma área nova para tratamento de fluxo de dados que necessita ser estudada.

Comentários