

Aula 01

Introdução ao Data Stream

Prof. Julio Cezar Estrella
jcezar@icmc.usp.br

Roteiro

- O que é?
- Introdução
- Aplicações de Data Streams
- Dados Relacionais x Dados em Streams
- Processamento em Lote x Streams
- Arquitetura de Dados em Streams
- Processamento de Consultas de Fluxo
- Desafios

O que é?

- Um fluxo de dados é uma sequência (potencialmente ilimitada) de tuplas
- Cada tupla consiste em um conjunto de atributos, semelhante a uma linha na tabela do banco de dados

O que é?

- **Modelo de fluxos de dados:**

- Os dados entram em uma taxa de alta velocidade
- O sistema não pode armazenar todo o fluxo, mas apenas uma pequena fração
- Como você faz cálculos críticos sobre o fluxo usando uma quantidade limitada de memória?

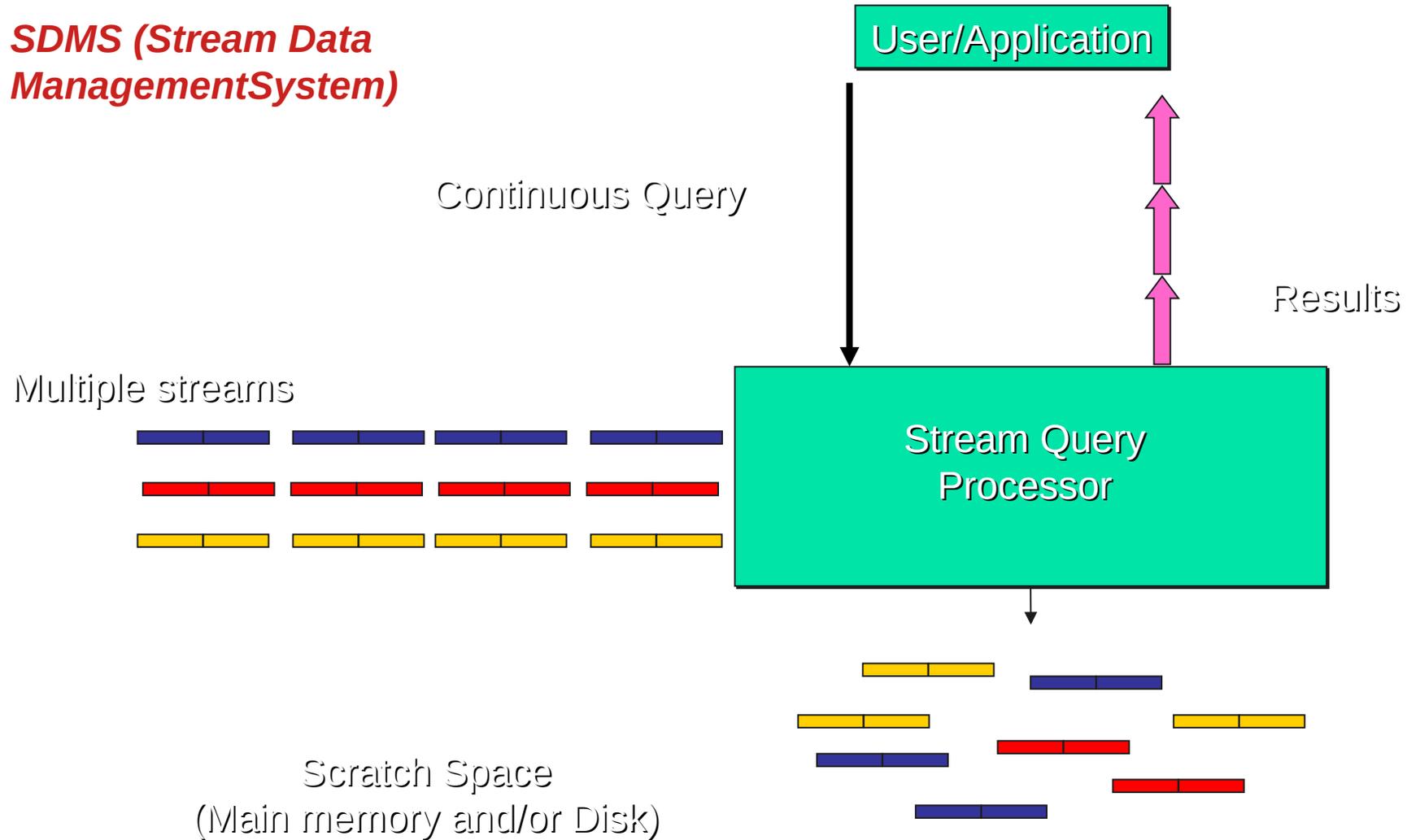
O que é?

- **Características**

- Enormes volumes de dados contínuos, possivelmente infinitos
- Mudança rápida e requer resposta rápida e em tempo real
- O acesso aleatório é caro - algoritmos de varredura única (só podem dar uma olhada)

Introdução

SDMS (Stream Data Management System)



Aplicações de Data Streams

- Registros de chamadas de telecomunicações
- Negócios: fluxos de transações de cartão de crédito
- Monitoramento de rede e engenharia de tráfego
- Mercado financeiro: bolsa de valores
- Engenharia e processos industriais: fornecimento de energia e fabricação

Aplicações de Data Streams

- Sensor, monitoramento e vigilância: fluxos de vídeo, RFIDs
- Logs da Web e fluxos de cliques de páginas da Web
- Conjuntos de dados maciços (mesmo salvos, mas o acesso aleatório é muito caro)

Dados Relacionais x Dados em Streams

- **Dados relacionais**

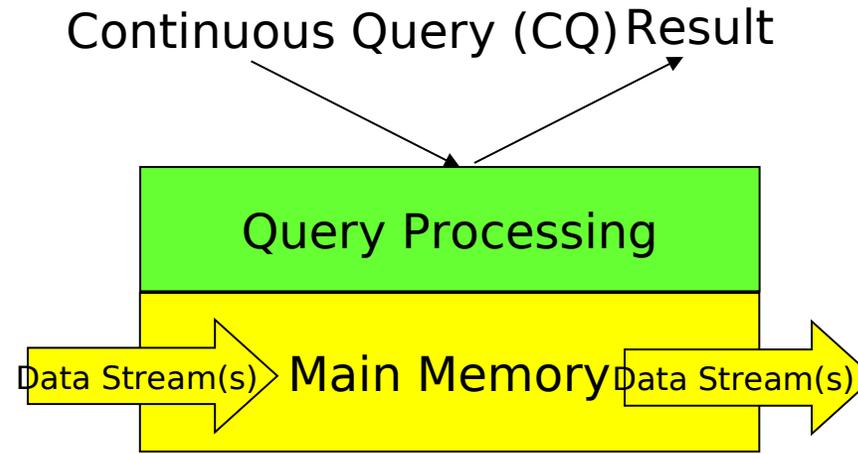
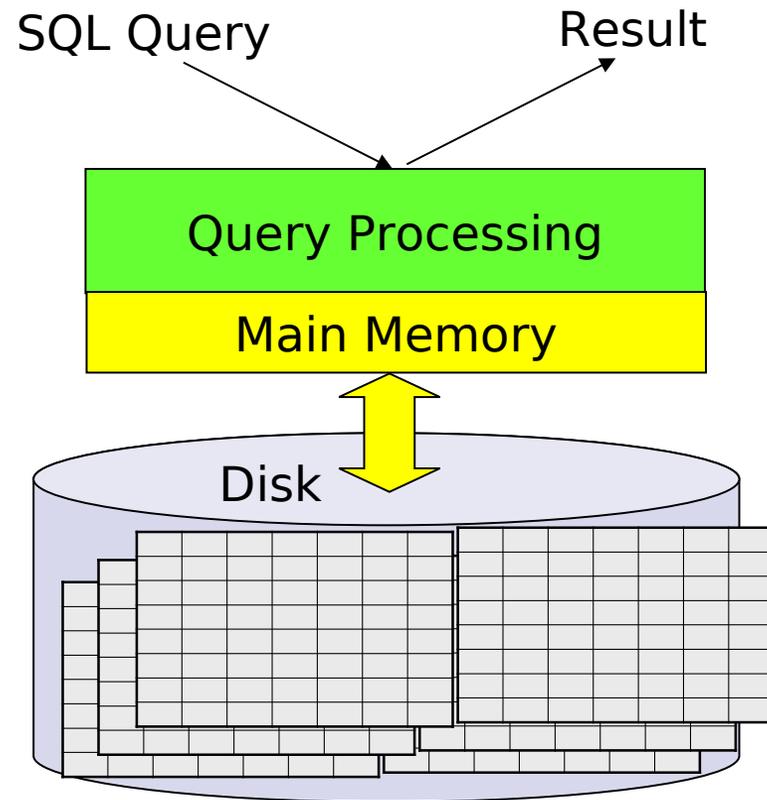
- Relações persistentes
- Consultas únicas
- Acesso aleatório
- Armazenamento em disco “ilimitado”
- Apenas o estado atual importa
- Sem serviços em tempo real
- Taxa de atualização relativamente baixa
- Dados em qualquer granularidade
- Assumir dados precisos
- Plano de acesso determinado pelo processador de consulta, design de banco de dados físico

Dados Relacionais x Dados em Streams

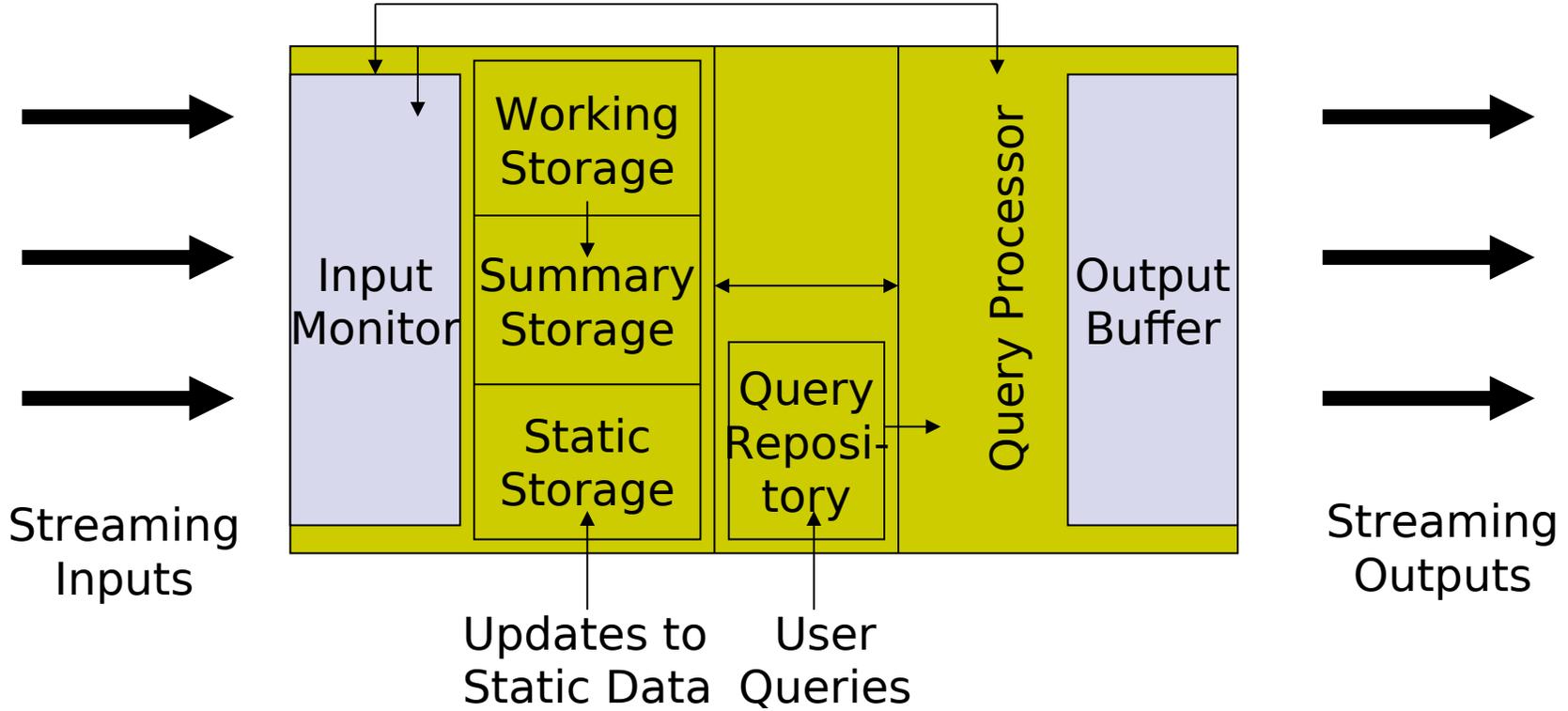
- **Dados em Streams**

- Fluxos transitórios
- Consultas contínuas
- Acesso sequencial
- Memória principal limitada
- Dados históricos são importantes
- Requisitos em tempo real
- Possivelmente taxa de chegada de vários GB
- Dados com granularidade fina
- Dados desatualizados/imprecisos
- Chegada e características de dados imprevisíveis/variáveis

Dados Relacionais x Dados em Streams



Arquitetura de Dados em Streams



Desafios

- As consultas costumam ser complexas
 - Além do processamento de um elemento por vez
 - Além do processamento de fluxo por vez
 - Além das consultas relacionais (científicas, mineração de dados)
 - Processamento multinível/multidimensional e mineração de dados
 - A maioria dos dados de fluxo são de baixo nível ou de natureza multidimensional

Processamento de Consultas de Fluxo

- Tipos de consulta
 - Consulta única versus consulta contínua (sendo avaliada continuamente à medida que o fluxo continua a chegar)
 - Consulta predefinida x consulta ad-hoc (emitido on-line)

Processamento de Consultas de Fluxo

- Requisitos de memória ilimitados
 - Para resposta em tempo real, o algoritmo da memória principal deve ser usado

Processamento de Consultas de Fluxo

- Resposta aproximada da consulta
 - Com memória limitada, nem sempre é possível produzir respostas exatas
 - Respostas aproximadas de alta qualidade são desejadas
 - Métodos de redução de dados e construção de sinopse
 - Esboços, amostragem aleatória, histogramas

Desafios

- Fluxos ordenados múltiplos, contínuos, rápidos, variáveis no tempo
- Cálculos da memória principal
- As consultas geralmente são contínuas
- Avaliado continuamente à medida que os dados de fluxo chegam
- Resposta atualizada ao longo do tempo

Referências

Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)

Atividade

- Disponível no Moodle conforme consta no cronograma da disciplina

Próxima Aula

- Introdução aos Containers e ao Docker