Frameworks de Big Data

UNIDADE 05

Explorando o oceano de dados: introdução aos fluxos de dados na era da informação em tempo real

Bem-vindo à nossa aula sobre *stream* de dados!

Durante este encontro, mergulharemos no fascinante mundo dos fluxos de dados, compreendendo sua definição e explorando suas diversas fontes. Os dados estão em constante movimento, e entender como lidar com essa dinâmica é fundamental para qualquer profissional que busque dominar o universo da análise de dados em tempo real.

Vamos explorar não apenas o que são fluxos de dados, mas também as diferentes fontes que provêm esses dados, desde sensores IoT até redes sociais, passando por transações financeiras e registros de eventos, examinaremos como esses fluxos de dados são gerados e como podem ser aproveitados para *insights* valiosos.

Nessa jornada, também discutiremos as tecnologias e ferramentas essenciais para capturar, processar e analisar fluxos de dados em tempo real. Ao longo do caminho, destacaremos estudos de caso e exemplos práticos que ilustram a importância e o impacto dos fluxos de dados em diversos setores e cenários.

Esse é apenas o começo de uma jornada emocionante rumo ao entendimento profundo dos fluxos de dados. Juntos, vamos desvendar os segredos por trás dessas correntes de informações em constante movimento e explorar as infinitas possibilidades que elas oferecem para *insights* e inovação.

Hoje, exploraremos um tema interessante sobre o que é ***stream* de dados**. Já pararam para pensar em como as grandes empresas lidam com a enorme quantidade de dados que flui constantemente em suas operações? Imagine, por exemplo, um sistema bancário que não apenas processa transações, mas também oferece notificações personalizadas aos clientes com base em seus hábitos financeiros e preferências individuais.

Como isso é possível? A resposta está na eficiente utilização de *stream* de dados. Esse será tema fundamental que abordaremos nesta aula. Prepare-se para descobrir como os bancos e outras instituições financeiras aproveitam as abordagens de *stream* de dados para otimizar suas operações e proporcionar experiências sob medida aos clientes.

O que é stream de dados?

O conceito de *data streaming* representa uma evolução do *big data*, viabilizando a prospecção e análise dinâmica e contínua de dados. EsSe paradigma se destaca em plataformas de áudio, como Spotify, Deezer, Apple Music, YouTube Music, Amazon Music, Tidal, entre outras, e em plataformas de vídeo, como YouTube, Netflix, Vimeo, DailyMotion, Twitch, entre outras. Além disso, o *streaming* transcende esses contextos, penetrando em diversas áreas da internet e se tornando uma expressão comum no vocabulário dos desenvolvedores.

Esta tendência conecta o mundo de forma abrangente, impulsionando a disseminação de informações em tempo real. O crescimento do *stream* de dados é notável, emergindo como um dos formatos mais proeminentes recentemente. No entanto, muitas empresas ainda não adotaram plenamente essa abordagem, desconhecendo o vasto potencial estratégico que ela oferece.

Na **figura 1**, apresentamos um exemplo como a abordagem de *Stream* de Dados aplicado no processo de compra em um e-commerce.

Figura 1: Fluxo de coleta de dados para recomendação de produtos em *e-commerce*. Fonte: Adaptado de INSIGHTLAB UFC (2023).

Agora você compreende por que recebe notificações de produtos em sua caixa de *e-mail* ou aquelas mensagens no celular que te lembram dos itens deixados no carrinho de compra de uma loja virtual. Hoje, graças às técnicas de *stream* de dados, é possível mapear toda a jornada do usuário ao acessar um serviço *on-line*.

Quais são os componentes do streaming de dados em tempo real?

Na *Figura 2,* temos um exemplo dos elementos de uma estrutura que se utiliza de stream de dados no ambiente da Amazon Web Services.

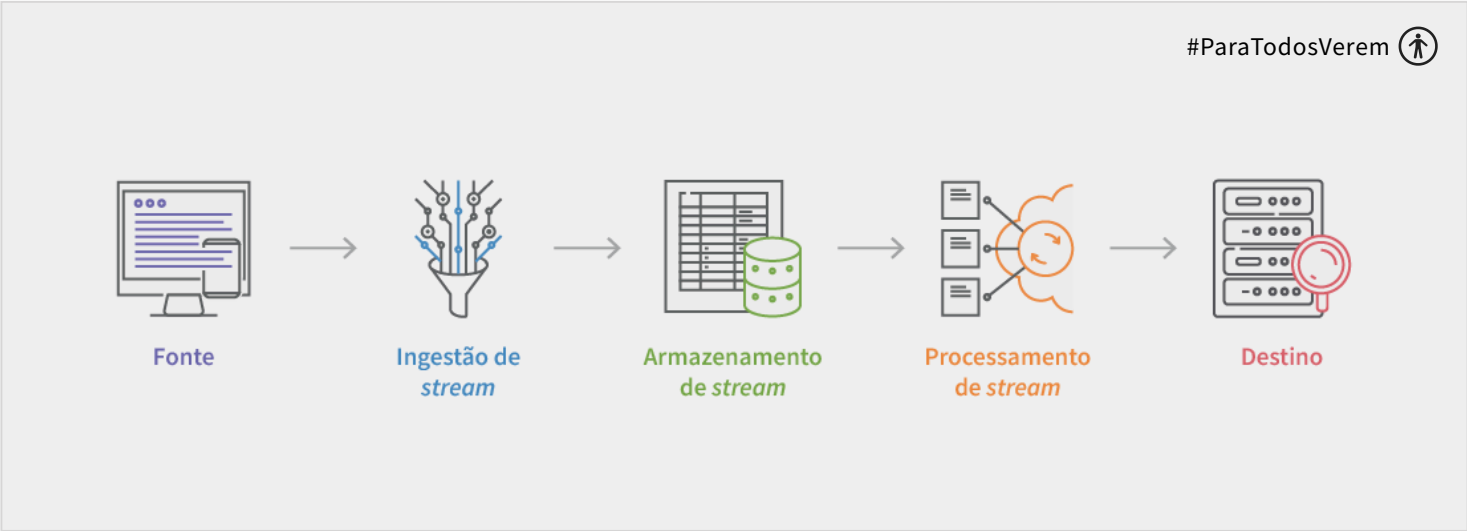

Figura 2: Componentes de uma estrutura streaming de dados. Fonte: Adaptado de Amazon Web Services (2023).

+ Fonte

Representa uma ampla variedade de dispositivos e aplicações que estão gerando grandes volumes de dados continuamente em alta velocidade. Exemplos incluem dispositivos móveis, aplicativos da *web* (*clickstream*), *logs* de aplicativos, sensores de IoT, dispositivos inteligentes e aplicativos de jogos.

+ Ingestão de *streaming*

Estrutura que permite capturar dados contínuos de milhares de dispositivos de forma durável e segura. Isso pode ser representado

por uma entrada de dados fluindo para dentro de um funil ou canal de ingestão.

+ Armazenamento de *stream*

Opções para armazenamento de dados em *streaming*, permitindo que os dados sejam armazenados com base em requisitos de escalabilidade, latência e processamento.

+ Processamento de *streaming*

Uma seleção de serviços que podem ser usados para processar os dados em *streaming*. Isso inclui soluções simples que transformam e entregam os dados a um destino.

+ Destino

Dados de streaming podem ser entregues a uma variedade de destinos para análise adicional ou armazenamento de longo prazo. Isso pode incluir *data lakes*, *data warehouses* e serviços de análise totalmente integrados.

Em seguida, no **quadro 1** apresentamos algumas características do uso de dados via *stream.*

Quadro 1: Características do uso de dados via *stream*

Característica	Descrição	Exemplo

Significância cronológica	Os elementos individuais possuem carimbos de data e hora, e o fluxo de dados pode ter sensibilidade ao tempo.	Recomendações de restaurantes com base na localização atual do usuário.
Fluidez contínua	A coleta de dados ocorre constante e continuamente.	*Logs* de atividade do servidor se acumulam enquanto o servidor está em execução.
Exclusividade	A retransmissão é desafiadora devido à sensibilidade ao tempo.	Processamento preciso de dados em tempo real é fundamental.
Não homogeneidade	Diversos formatos estruturados (JSON, Avro, CSV) e tipos de dados (*string*, número, data, binário).	Sistemas de processamento de *streaming* devem lidar com essas variações.
Imperfeição	Erros temporários na fonte podem resultar em elementos danificados ou ausentes.	Validação de dados para mitigar ou minimizar erros.

Fonte: Adaptado de Amazon Web Services (2023).

Como apresentado no **quadro 1**, temos um resumo claro e conciso das principais características dos dados de *streaming*. As cinco características descritas – significância cronológica, fluidez contínua, exclusividade, não homogeneidade e imperfeição – são cruciais para entender a natureza dos dados de *streaming* e os desafios que eles apresentam para o processamento e análise.

Exemplos de fontes de Stream de Dados

Os dados em *streaming* têm origem em diversas fontes, e a IoT está desempenhando um papel fundamental ao trazer dados de lugares inimagináveis há alguns anos, como óculos de realidade virtual, relógios inteligentes e até mesmo geladeiras. À medida que os dados provenientes da IoT se tornam mais acessíveis economicamente, presenciamos um aumento significativo na geração de dados em tempo real, demandando o surgimento de novas aplicações para processá-los. É verdadeiramente fascinante estarmos no epicentro dessa revolução!

As fontes de *streaming* de dados são diversas e em constante expansão. Vamos explorar algumas delas:

+ **Monitoramento de operações**

O monitoramento em tempo real de toda a infraestrutura de *software* e *hardware* tornou-se crucial para empresas de médio e grande porte, especialmente com a explosão de aplicações nas últimas décadas. Sistemas de monitoramento de operações emitem alertas instantâneos em caso de falhas de disco, erros de execução de aplicações e outros problemas. A análise da imensidão de dados gerados por máquinas permite a geração de métricas de performance, garantindo operações de TI com baixo custo e alta adaptabilidade a falhas e mudanças na infraestrutura.

+ ***Web analytics***

O comércio eletrônico e o *marketing* digital impulsionaram a necessidade de monitoramento em tempo real de cliques em *sites* e *banners*. Essa análise permite direcionar campanhas e ofertas de forma personalizada, além de acompanhar outras métricas, como volume de acessos, correlação entre cliques, reações a campanhas e muito mais.

Esse grande volume de dados, conhecido como *big data*, gerado em tempo real, levou à criação de sistemas de recomendação e à realização de experimentos como testes A/B e outras técnicas analíticas em tempo real. A análise instantânea dos dados da interação dos usuários com *websites* é crucial para o sucesso de empresas, pois a espera de dias para análise pode levar à perda de clientes para a concorrência.

+ **Mídias sociais**

As redes sociais, como o X, que processa mais de 500 milhões de postagens por dia (150 mil por segundo), são uma enorme fonte de dados em tempo real. Essa informação, junto com dados de outras plataformas como Facebook, Foursquare e YouTube, torna-se crucial para a tomada de decisões.

O desafio, no entanto, está em processar e analisar esses dados não estruturados em tempo real. Técnicas de processamento de linguagem natural são necessárias para transformar a imensidão de dados em informações precisas para a tomada de decisões. Essa nova realidade exige sistemas automatizados e eficientes para acompanhar o ritmo acelerado da geração de dados.

+ ***Mobile***

A popularização do iPhone em 2007, com sua tela *touchscreen* e acesso à internet, intensificou a geração de dados em tempo real. *Smartphones* permitem que empresas detectem a localização de usuários, enviem promoções personalizadas e realizem vendas em minutos. A análise instantânea desses dados é crucial para o sucesso, pois esperar 3 dias pode significar perder oportunidades de negócio.

Wearables, como *smartwatches*, sensores em roupas e óculos de realidade virtual, também geram dados em tempo real, expandindo ainda mais o universo do *streaming* de dados.

Esses são apenas alguns exemplos de como o *streaming* de dados está se tornando o padrão, possibilitando novas aplicações analíticas e decisões mais rápidas e precisas. As empresas que se adaptam a essa nova realidade estarão na vanguarda do mercado.

Problemas relacionados com o Stream de Dados

Os problemas relacionados aos fluxos de dados podem surgir de diversas fontes e apresentar desafios únicos. Um dos principais problemas é a latência, que se refere ao atraso entre a geração e o processamento dos dados. Em aplicações que requerem respostas em tempo real, como sistemas de monitoramento de tráfego ou detecção de fraudes financeiras, a latência pode ser crítica e afetar diretamente a eficácia da solução. Além disso, a escalabilidade é uma preocupação constante, especialmente em cenários em que a quantidade de dados aumenta rapidamente. Garantir que os sistemas de *streaming* sejam capazes de lidar com o crescimento contínuo dos dados sem comprometer o desempenho é essencial. Outro desafio é a integridade e a confiabilidade dos dados, pois a transmissão em tempo real pode resultar em perda de pacotes ou corrupção de informações. Portanto, é fundamental implementar mecanismos robustos de garantia de qualidade para assegurar a precisão e a consistência dos dados em ambientes de *streaming*.

Problemas e Soluções

No **quadro 2**, apresentamos os principais problemas relacionados ao *streaming* de dados, suas causas e possíveis soluções:

Quadro 2: Problemas relacionados ao uso de *stream* de dados

Problema	**Causa**	**Solução**
Volume de dados	Múltiplas fontes, alta frequência de envio.	Esquemas de dados flexíveis, técnicas de amostragem, compressão de dados.
Diversidade de dados	Fontes e dispositivos heterogêneos.	Padronização de dados, mapeamento de esquemas, ferramentas de integração.
Falta de profissionais qualificados	Área de conhecimento emergente.	Treinamento e capacitação, programas de educação continuada, comunidades de prática.
Processamento em tempo real	Restrições de memória e tempo.	Algoritmos eficientes, técnicas de aproximação, processamento paralelo.
Escalabilidade	Aumento no número de *streams*.	Arquiteturas distribuídas, soluções em nuvem, automação de provisionamento.
Durabilidade	Armazenamento e persistência de dados.	Soluções de *big data*, bancos de dados NoSQL, replicação de dados.

Fonte: O autor (2024).

Os problemas representam desafios significativos na implementação e manutenção de sistemas de *streaming* de dados, e suas soluções demandam abordagens técnicas e estratégicas cuidadosamente planejadas.

Em suma, o *streaming* de dados oferece um enorme potencial para diversas áreas, mas exige uma abordagem cuidadosa e proativa para superar os desafios que apresenta.

Investir em soluções robustas, no desenvolvimento de profissionais qualificados e no acompanhamento das inovações é fundamental para aproveitar ao máximo as vantagens dessa tecnologia.

Ao superar esses desafios, as empresas podem desbloquear um mundo de oportunidades e *insights* valiosos, impulsionando a tomada de decisões mais inteligente e ágil, a otimização de processos e a criação de novos produtos e serviços.

Conclusão

Na aula de hoje, exploramos um tema fascinante: o que é *stream* de dados e como essa tecnologia revolucionou a forma como as grandes empresas lidam com a enorme quantidade de informações que fluem constantemente em suas operações. Por meio de exemplos, como um sistema bancário que oferece notificações personalizadas aos clientes com base em seus hábitos financeiros, pudemos compreender como o uso eficiente de stream de dados pode transformar radicalmente a maneira como as organizações interagem com seus clientes e otimizam suas operações.

Ao longo da aula, discutimos as diversas fontes de dados em streaming, desde dispositivos móveis e sensores IoT até aplicativos da web, e exploramos as soluções disponíveis para ingestão, armazenamento, processamento e destino desses dados. Vimos como as empresas podem aproveitar essas tecnologias para fornecer experiências personalizadas, detecção de fraudes em tempo real, insights preditivos e muito mais.

No entanto, é importante ressaltar que, embora o *stream* de dados ofereça inúmeras oportunidades e vantagens, também apresenta desafios significativos, como garantir a segurança e privacidade dos dados, lidar com a complexidade e escalabilidade das soluções e garantir a qualidade e integridade das informações em tempo real.

Portanto, ao refletir sobre o que aprendemos hoje, podemos concluir que o *stream* de dados é uma ferramenta poderosa e transformadora, capaz de impulsionar a inovação, otimizar processos e melhorar a tomada de decisões. No entanto, seu uso eficaz requer não apenas conhecimento técnico, mas também uma compreensão profunda das necessidades e expectativas dos clientes, bem como um compromisso com a ética e a responsabilidade no uso dos dados. Ao integrar esses elementos, as empresas podem verdadeiramente aproveitar o potencial do *stream* de dados para criar valor e promover um futuro mais inteligente e conectado.

Referências Bibliográficas

AMAZON WEB SERVICES (AWS). **O que são dados de *streaming*?** Amazon Web Services (AWS), Seattle, WA, 2023. Disponível em: https://aws.amazon.com/pt/what-is/streaming-data/. Acesso em: 20 mar. 2024.

GBTECH. **Solução de *streaming analytics* do Grupo Boticário**. Medium, 2022. Disponível em: https://medium.com/gbtech/solu%C3%A7%C3%A3o-de-streaming-analytics-do-grupo-botic%C3%A1rio-d9f3c5ccc0c9. Acesso em: 20 mar. 2024.

FERREIRA, B. *et al*. Arquitetura de dados com *event streams*. **Medium**, 2022. Disponível em: https://brenocferreira.medium.com/arquitetura-de-dados-com-event-streams-7361dd69438d. Acesso em: 20 mar. 2024.

GOMES, M.; MENEZES, M. *Event streaming*, Apache Kafka e Kafka Streams em uma arquitetura de microsserviços. **Medium**, 2022. Disponível em: https://medium.com/@marcelomg21/event-streaming-apache-kafka-kafka-streams-em-uma-arquitetura-de-micro-servi%C3%A7os-136c28d7c9c8. Acesso em: 20 mar. 2024.

INSIGHTLAB UFC. **Entenda como funciona *streaming* de dados em tempo real** 2. InsightLab UFC, Fortaleza, CE, 2023. Disponível em: https://www.insightlab.ufc.br/entenda-como-funciona-streaming-de-dados-em-tempo-real-2/. Acesso em: 20 mar. 2024.

PEREIRA, M. A.; NEUMANN, F. B.; MILANI, A. M. P. *et al*. ***Framework* de *big data***. Porto Alegre: Grupo A, 2020.

REICHERT JR., I. Análise de dados de sensores em tempo real com Spark Streaming e Apache Kafka. **Medium**, 2022. Disponível em: https://medium.com/@ingoreichertjr/an%C3%A1lise-de-dados-de-sensores-em-tempo-real-com-spark-streaming-e-apache-kafka-8e1548949f44. Acesso em: 20 mar. 2024.