# ARQUITETURAS OTIMIZADAS PARA PIPELINES DE DADOS NA AWS

**André Cassulino Araújo Souza[1], Glauco Todesco[2], Deivison Shindi Takatu [3], Cainã Antunes Silva[4], Gabriel Claro da Silva[5], Lucas Miguel Leal da Silva[6]**

*[1] Professor Me. no Centro Universitário SENAI São Paulo – UniSENAI-SP, Campus Sorocaba – Santa Rosália. E-mail: andre.souza@sp.senai.br*
*[2]Professor Dr. no Centro Universitário SENAI São Paulo – UniSENAI-SP, Campus Sorocaba – Santa Rosália. E-mail: glauco.todesco@sp.senai.br*
*[3] Professor Me. no Centro Universitário SENAI São Paulo – UniSENAI-SP, Campus Sorocaba – Santa Rosália. E-mail: deivison.takatu@sp.senai.br*
*[4] Professor Esp. no Centro Universitário SENAI São Paulo – UniSENAI-SP, Campus Sorocaba – Santa Rosália. E-mail: caina.silva@sp.senai.br*
*[5] Professor Esp. no Centro Universitário SENAI São Paulo – UniSENAI-SP, Campus Sorocaba – Santa Rosália. E-mail: gabriel.csilva@sp.senai.br*
*[6] Coordenador Esp. no Centro Universitário SENAI São Paulo – UniSENAI-SP, Campus Sorocaba – Santa Rosália. E-mail: lucas.msilva@sp.senai.br*

*Resumo:* Este estudo analisa arquiteturas otimizadas para pipelines de dados na Amazon Web Services (AWS), com foco comparativo entre os paradigmas de processamento em lotes (*batch processing*) e processamento em fluxo contínuo (*stream processing*). A pesquisa discute as características, vantagens e limitações de cada abordagem, destacando a relevância do processamento em lotes para grandes volumes de dados históricos e a importância do processamento em fluxo para aplicações que exigem baixa latência e respostas em tempo real. No contexto da computação em nuvem, são examinados serviços gerenciados da AWS, como AWS Glue, Amazon Kinesis Data Streams, Amazon Data Firehose, AWS Lambda, Amazon S3, DynamoDB e Amazon SageMaker. A partir de uma abordagem aplicada e descritiva, o trabalho apresenta possibilidades de implementação de pipelines serverless para ingestão, transformação, armazenamento e enriquecimento de dados com inteligência artificial. Conclui-se que arquiteturas híbridas e orientadas a eventos permitem maior escalabilidade, flexibilidade operacional e otimização de custos, constituindo alternativas relevantes para organizações que buscam transformar dados em valor de negócio com maior agilidade.

*Palavras-chave*: Processamento de dados; AWS; Arquitetura serverless; Pipelines de dados; Computação em nuvem.

## INTRODUÇÃO

A expansão da economia digital e da Indústria 4.0 intensificou a geração de dados em ambientes corporativos, industriais e conectados. Transações financeiras, redes sociais, sensores de Internet das Coisas (IoT), sistemas de monitoramento e aplicações web produzem fluxos contínuos de informações que precisam ser coletados, processados e analisados com segurança, desempenho e escalabilidade. Nesse cenário, a arquitetura de processamento de dados torna-se uma decisão estratégica para organizações que dependem de informações confiáveis para apoiar decisões operacionais e analíticas.

Entre as principais abordagens utilizadas para esse fim, destacam-se o processamento em lotes, voltado ao tratamento de grandes volumes de dados acumulados, e o processamento em fluxo, orientado à análise contínua de eventos à medida que são produzidos. O processamento em lotes mantém relevância em atividades como consolidação de bases históricas, geração de relatórios, rotinas ETL e treinamento de modelos analíticos. Já o processamento em fluxo atende aplicações que exigem baixa latência, como detecção de fraudes, monitoramento de máquinas, recomendação em tempo real e acionamento automatizado de alertas.

A computação em nuvem amplia as possibilidades de implementação dessas arquiteturas ao oferecer serviços gerenciados, escaláveis e cobrados conforme o uso. No ecossistema da Amazon Web Services

(AWS), serviços como AWS Glue, Amazon Kinesis, AWS Lambda, Amazon Data Firehose, Amazon S3, DynamoDB e Amazon SageMaker permitem construir pipelines de dados com menor necessidade de gerenciamento de infraestrutura física. Dessa forma, torna-se possível combinar ingestão, transformação, armazenamento e análise em arquiteturas serverless e orientadas a eventos.

O objetivo deste artigo é analisar arquiteturas otimizadas para pipelines de dados na AWS, comparando as abordagens batch e streaming e discutindo sua aplicação em soluções híbridas. Busca-se demonstrar como serviços gerenciados podem apoiar a criação de pipelines automatizados, escaláveis e integrados a recursos de inteligência artificial, contribuindo para a redução de latência, otimização de custos e aumento da capacidade analítica das organizações.

## REVISÃO DE LITERATURA

A literatura sobre engenharia de dados evidencia que o processamento em lotes constitui uma abordagem consolidada para operações periódicas sobre grandes volumes de dados. Nesse modelo, os dados são acumulados durante determinado intervalo e processados em conjunto, normalmente em horários definidos ou fora dos períodos de maior demanda. Essa característica favorece análises históricas, consolidações contábeis, processamento de arquivos e cargas massivas em ambientes de data lake ou data warehouse.

Apesar de sua robustez, o processamento em lotes apresenta limitações quando aplicado a cenários que demandam respostas imediatas. Como os resultados dependem da conclusão do lote processado, a latência pode inviabilizar aplicações sensíveis ao tempo. Além disso, falhas em uma etapa do processamento podem comprometer todo o conjunto de dados, exigindo mecanismos de reprocessamento, monitoramento e controle de qualidade.

O processamento em fluxo, por sua vez, opera continuamente sobre eventos à medida que são gerados. Essa abordagem é adequada para aplicações baseadas em eventos, nas quais a rapidez na identificação de padrões, anomalias ou mudanças de estado é determinante. Entre os exemplos recorrentes estão sistemas antifraude, monitoramento industrial, análise de comportamento de usuários, telemetria e aplicações de IoT.

A adoção de arquiteturas híbridas permite combinar as vantagens dos dois modelos. Enquanto a camada batch fornece visão histórica e suporte a análises de maior profundidade, a camada streaming possibilita respostas quase imediatas. Essa combinação é especialmente útil em soluções de manutenção preditiva, análise operacional e integração com modelos de aprendizado de máquina, nas quais dados históricos e dados em tempo real se complementam.

## METODOLOGIA

Este estudo possui natureza aplicada, descritiva e qualitativa, fundamentada na análise de conceitos e serviços utilizados na construção de pipelines de dados na AWS. A investigação considera duas abordagens principais de processamento: batch e streaming, examinando suas características técnicas, vantagens, limitações e possibilidades de combinação em arquiteturas híbridas.

A análise foi organizada em três etapas. Na primeira, foram descritos os fundamentos de processamento em lotes e em fluxo, considerando requisitos de volume, latência, custo e complexidade operacional. Na segunda, foram examinados serviços da AWS associados a cada abordagem, com destaque para AWS Glue em rotinas ETL, Amazon Kinesis Data Streams e Amazon Data Firehose em ingestão e entrega de eventos, AWS Lambda em processamento serverless e Amazon SageMaker em consumo de modelos de inteligência artificial. Na terceira etapa, discutiu-se a integração desses serviços em uma arquitetura orientada a eventos capaz de processar, enriquecer e armazenar dados de forma automatizada.

## RESULTADOS E DISCUSSÕES

A análise evidencia que o AWS Glue se apresenta como uma solução adequada para pipelines batch, especialmente em rotinas de extração, transformação e carga. Seus componentes principais incluem o Data Catalog, responsável pelo armazenamento de metadados; os Crawlers, que inferem esquemas a partir de fontes de dados; os ETL Jobs, que executam transformações com base em Apache Spark; e os Triggers, que permitem o acionamento sob demanda, por agendamento ou por eventos. Essa estrutura favorece a padronização de metadados e a automação de fluxos de dados históricos.

No processamento em fluxo, o Amazon Kinesis oferece serviços voltados à ingestão e ao tratamento contínuo de eventos. O Kinesis Data Streams permite a criação de fluxos em tempo real consumidos por aplicações ou funções Lambda. O Amazon Data Firehose simplifica a entrega automatizada de dados para destinos como Amazon S3, Redshift ou OpenSearch. Já o Kinesis Data Analytics possibilita consultas e transformações em tempo real, ampliando as alternativas para análise de streams.

A integração entre Kinesis e AWS Lambda permite construir pipelines orientados a eventos. Nesse modelo, novos registros inseridos em uma stream acionam automaticamente funções responsáveis por validar, transformar ou encaminhar dados para

serviços de destino. Essa característica reduz a necessidade de provisionamento permanente de servidores e favorece o modelo pay-as-you-go, no qual os custos acompanham o uso efetivo dos recursos computacionais.

Uma arquitetura serverless para streaming data processing pode ser organizada a partir de uma API de entrada integrada a uma função Lambda, responsável por publicar eventos no Kinesis Data Streams. Em seguida, o Amazon Data Firehose recebe os dados, aciona outra função Lambda para pré-processamento e grava os resultados em um bucket S3. Quando necessário, a função de processamento pode chamar um endpoint do Amazon SageMaker para realizar predições em tempo real, enriquecendo os dados antes de seu armazenamento definitivo.

Os resultados discutidos indicam que a combinação entre serviços gerenciados e arquitetura orientada a eventos permite reduzir a complexidade de infraestrutura e aumentar a elasticidade da solução. Entretanto, a adoção de pipelines em tempo real exige atenção à governança, ao monitoramento, à consistência dos dados, ao controle de custos e à segurança. Dessa forma, a escolha entre batch, streaming ou uma solução híbrida deve considerar não apenas requisitos técnicos, mas também maturidade operacional, criticidade da aplicação e objetivos de negócio.

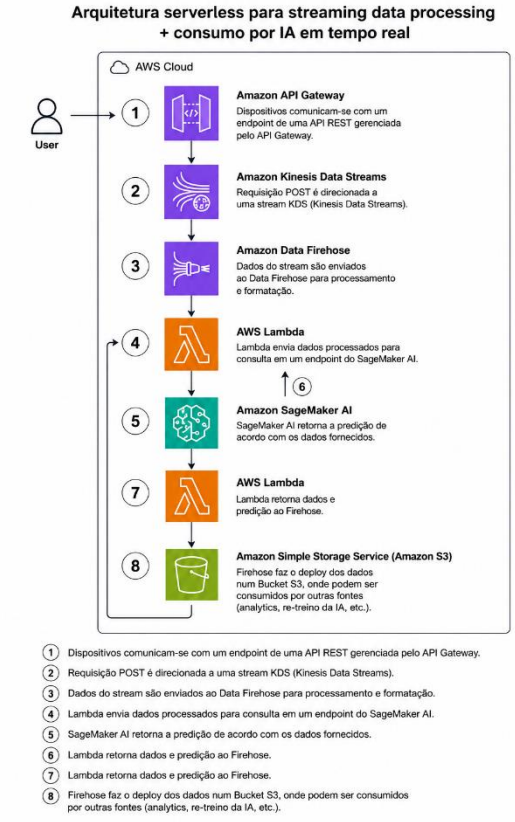


Figura 1 – Arquitetura serverless para processamento de dados em fluxo com integração a serviços AWS.
Fonte: Elaboração do autor.

## CONCLUSÃO

Este artigo analisou arquiteturas otimizadas para pipelines de dados na AWS, considerando as diferenças e complementaridades entre processamento em lotes e processamento em fluxo. Constatou-se que o modelo batch permanece essencial para grandes volumes de dados históricos, consolidações e rotinas analíticas programadas, enquanto o modelo streaming se destaca em aplicações que exigem baixa latência, respostas imediatas e processamento contínuo de eventos.

A utilização de serviços gerenciados da AWS, como AWS Glue, Amazon Kinesis, AWS Lambda, Amazon Data Firehose, Amazon S3, DynamoDB e Amazon SageMaker, demonstra a viabilidade de construir soluções escaláveis e automatizadas sem gerenciamento direto de servidores. A arquitetura híbrida apresentada evidencia que dados podem ser ingeridos, processados, enriquecidos por modelos de inteligência artificial e armazenados de forma integrada, favorecendo agilidade analítica e eficiência operacional.

Conclui-se que a convergência entre engenharia de dados, computação em nuvem e arquiteturas serverless oferece caminhos consistentes para transformar dados brutos em valor de negócio. Como trabalhos futuros, recomenda-se aprofundar a análise comparativa de custos entre arquiteturas serverless e clusters provisionados, investigar práticas de segurança e governança de dados em pipelines automatizados e explorar a integração com computação de borda para reduzir ainda mais a latência em ambientes IoT.

## REFERÊNCIAS

AMAZON WEB SERVICES. AWS Glue: key concepts. Amazon Web Services, 2025.

AMAZON WEB SERVICES. Streaming data solutions on AWS with Amazon Kinesis. Amazon Web Services, 2021.

AMAZON WEB SERVICES. Lambda architecture for batch and real-time processing on AWS. Amazon Web Services, 2020.

AMAZON WEB SERVICES. Serverless applications lens: AWS Well-Architected Framework. Amazon Web Services, 2025.

JOHNSON, M. Batch processing vs. stream processing: what is best for real-time analytics? BMC Blogs, 2022.

REMALA, R. Leveraging AWS serverless architecture for efficient data processing and analytics. 2021.

RIVERY. Batch vs. stream processing: pros and cons. Rivery.io, 2023.

SOUZA, A. B.; SILVA, C. D. Stream processing vs. batch processing: determining the best fit for machine learning pipelines. Journal of Data Engineering, v. 12, n. 3, p. 45-62, 2023.