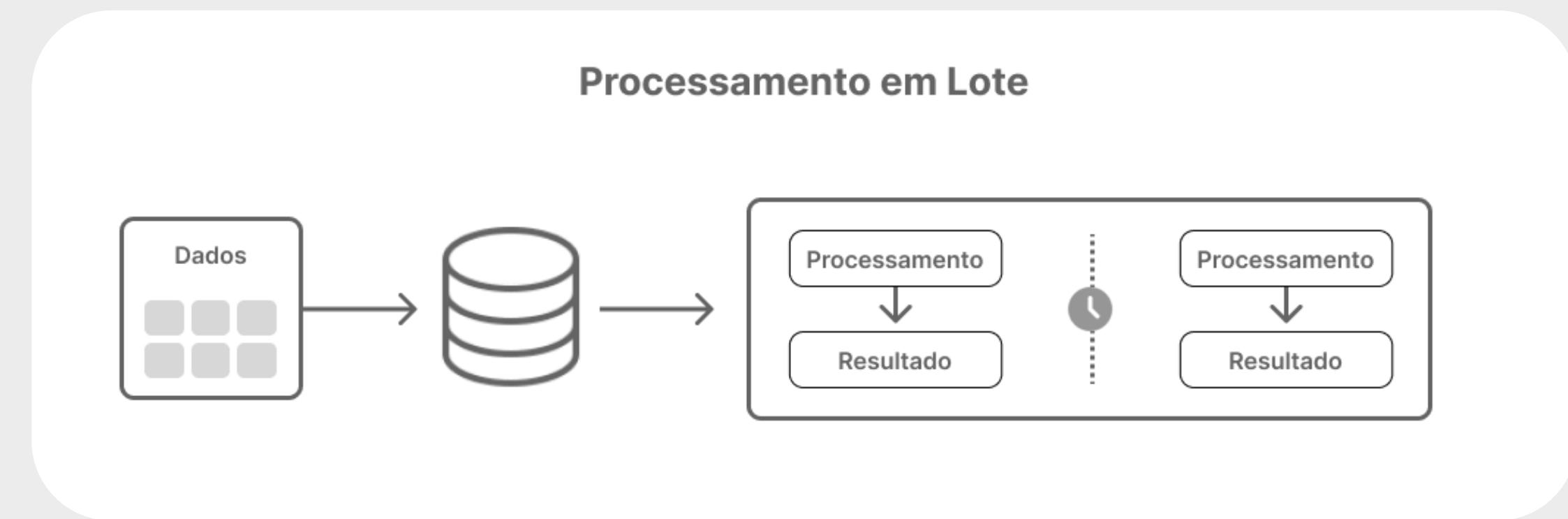


DATA STREAMING EM TEMPO REAL COM APACHE KAFKA

Grupo: Eduardo Braga, Henrique Franca, Isabela Medeiros e Júlia Vilela

O MUNDO DO LOTE (BATCH)

- Processamento em lote (Batch): processa grandes volumes de dados de uma vez só, em intervalos fixos
- Os dados ficam parados e só são acessados quando necessário, o que gera um atraso significativo
- Limitações e atrasos: impossibilidade de uma reação imediata a tendências ou a problemas urgentes.

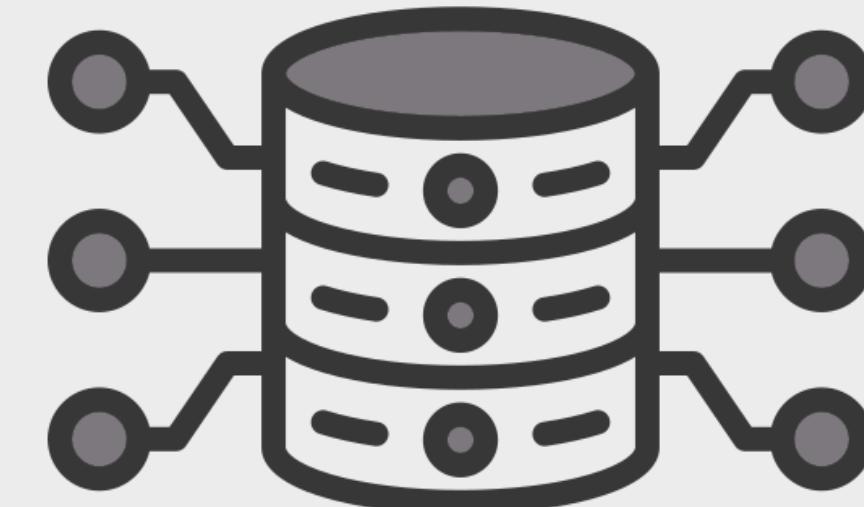


NECESSIDADE DE TOMADA DE DECISÃO EM TEMPO REAL

- O mundo digital exige respostas imediatas. A informação de hoje é a mais valiosa.
- A rapidez é crucial para a sobrevivência de muitas empresas e serviços, como por exemplo: Serviços Financeiros, E-commerce, Manutenção de Sistemas, entre muitos outros.
- Limitações e atrasos: O modelo de processamento em lote (batch) cria uma lacuna que impede a agilidade, impossibilita reação imediata a tendências ou a problemas urgentes.

DATA STREAMING

- É o processamento contínuo de dados à medida que eles são gerados, permitindo uma resposta quase instantânea.
- O streaming elimina o atraso entre a geração dos dados e a sua utilização.
- Principais vantagens:
 - Tomada de Decisão em Tempo Real
 - Maior Agilidade e Competitividade
 - Novas Oportunidades de Negócio
 - Experiência do Cliente Aprimorada



APACHE KAFKA

- É uma plataforma de streaming de eventos distribuída, ele foi feito para lidar com um volume massivo de dados em tempo real, de forma confiável e escalável.
- Criado para Escalar no LinkedIn
- Tornado Open Source em 2011
- Benefícios:
 - Alta Performance e Escalabilidade
 - Tolerância a Falhas e Durabilidade
 - Desacoplamento de Sistemas
 - Versatilidade



APACHE KAFKA

- Plataforma de streaming de eventos distribuída. Sua principal função é atuar como um message broker de alta performance. Ele é projetado para:
 - Coletar e Ingerir Dados
 - Armazenar Dados
 - Transportar Dados

APACHE FLINK

- Framework e mecanismo de processamento de fluxo distribuído. Seu objetivo é executar cálculos complexos e análises sobre fluxos de dados em tempo real
 - Computar Dados
 - Analisar Dados
 - Executar Aplicações

PRODUCER

- É o aplicativo que escreve ou envia mensagens para o Kafka.
- Atua como a fonte do fluxo de dados. Ele cria a mensagem com as informações do evento e a publica em um tópico específico no Kafka.

TOPIC

- É uma categoria, nome ou canal onde as mensagens são publicadas.
- É o elemento central para a organização dos dados no Kafka. Todas as mensagens relacionadas a um mesmo tipo de evento ficam em um único tópico.

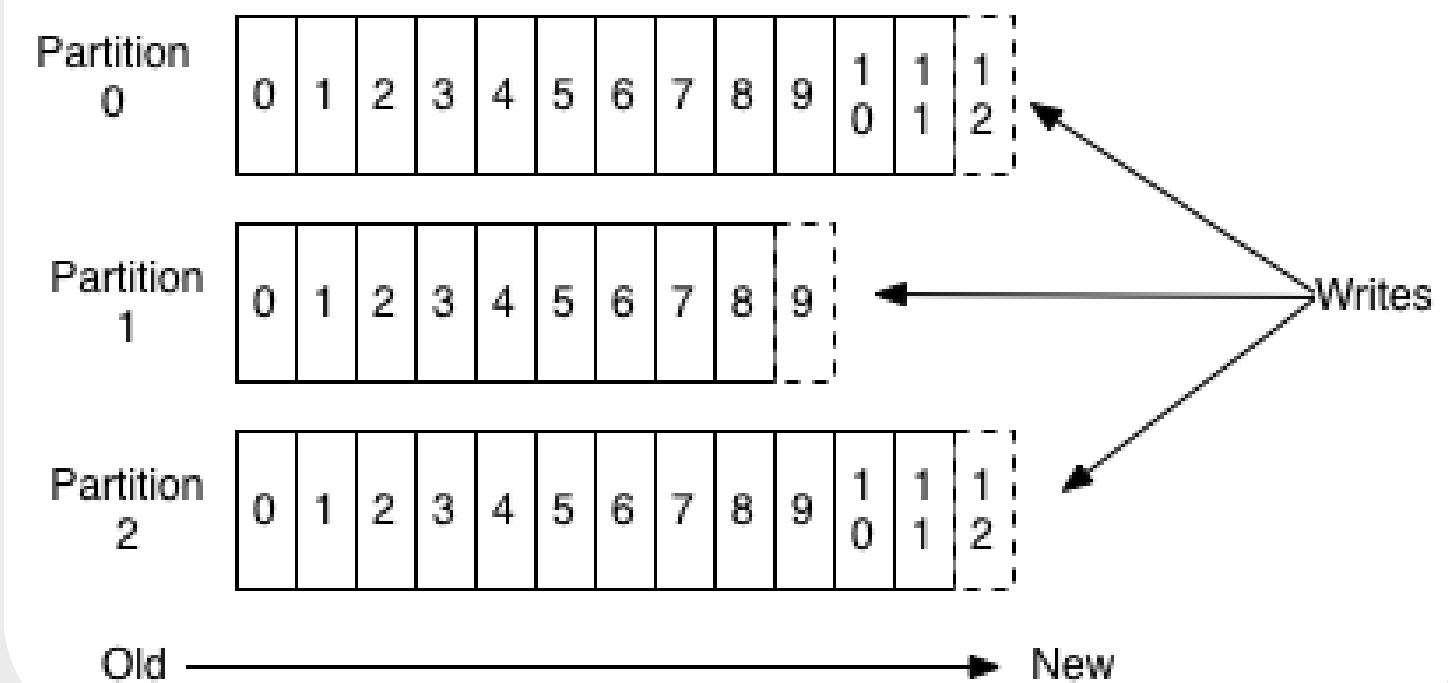
CONSUMER

- É o aplicativo que lê e processa as mensagens de um ou mais tópicos.
- É o responsável por agir com base nos dados. Um consumidor se inscreve em um tópico de seu interesse para obter as mensagens.

PARTIÇÕES

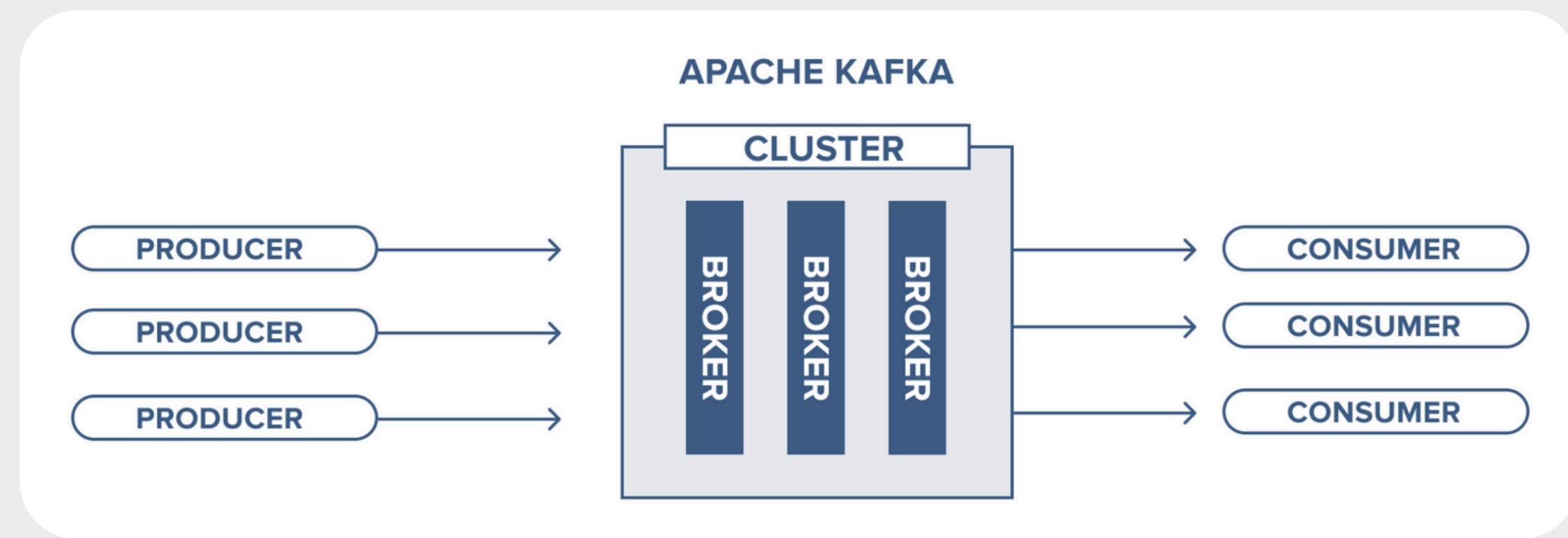
- São a unidade básica de paralelismo no Kafka.
- Cada tópico é dividido em uma ou mais partições, permitindo que um único tópico seja processado por múltiplos consumidores ao mesmo tempo.
- O principal objetivo das partições é permitir a escalabilidade horizontal do Kafka.
- É crucial entender que a ordem das mensagens é garantida apenas dentro de uma partição.

Anatomy of a Topic

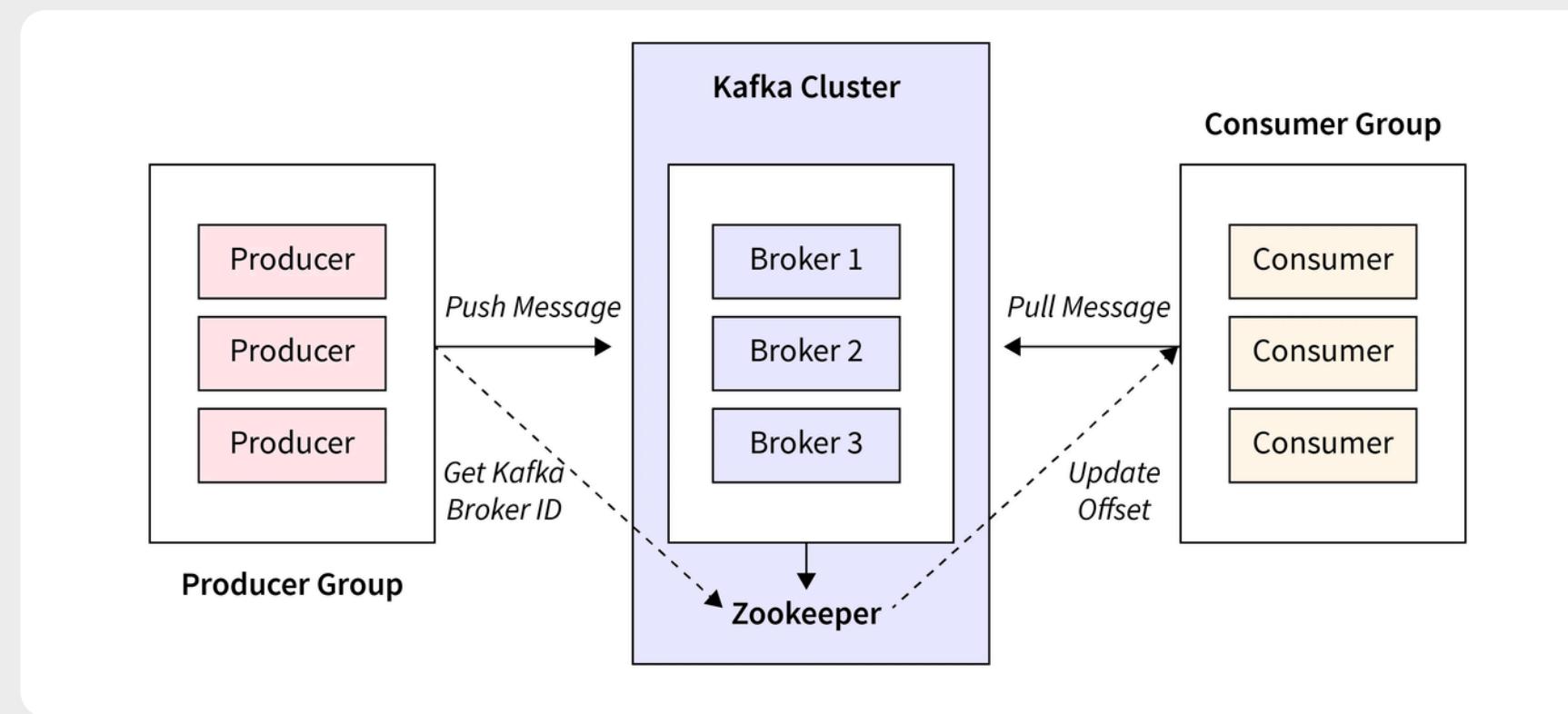


CLUSTER E BROKERS

- Broker: É um servidor único em um ambiente Kafka. É responsável por armazenar os dados das partições dos tópicos.
- Cluster: É um conjunto de corretores que trabalham em conjunto. Ele permite que o Kafka opere em grande escala e com alta disponibilidade.
- A arquitetura de cluster é a chave para a tolerância a falhas e a alta disponibilidade do Kafka.



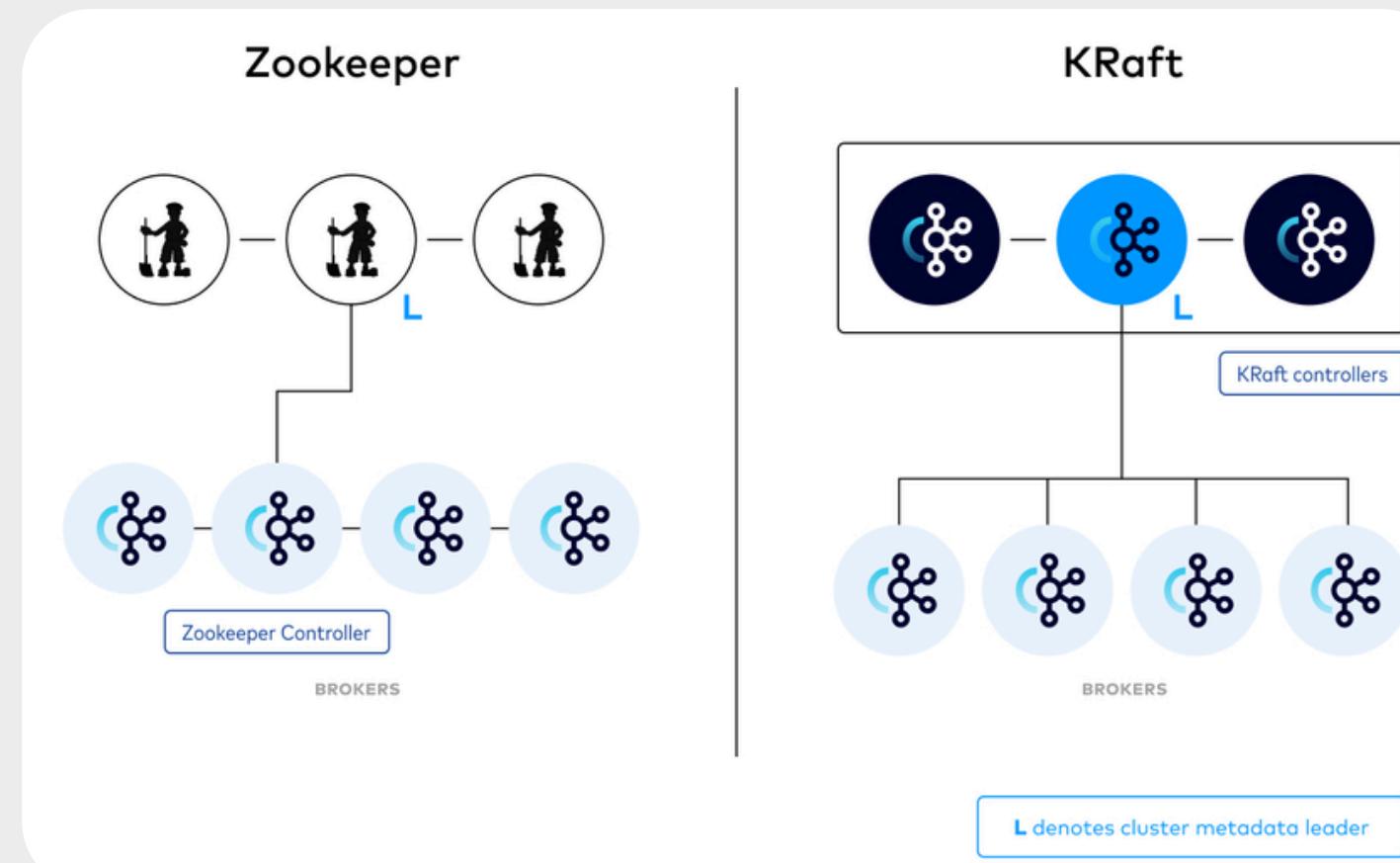
ZOOKEEPER



- Gerencia metadados, controle de brokers e coordenação do cluster.
 - Mantém informações sobre tópicos, partções e líderes.
 - Faz a eleição de líderes de partções.
 - Garante consistência entre os brokers.
- Kafka depende de um sistema externo (ZooKeeper), tornando a arquitetura mais complexa.

KRAFT

- Substitui o ZooKeeper, integrando a gestão de metadados no próprio Kafka.
 - Usa o protocolo Raft para replicação e consenso de metadados.
 - Brokers elegem um "controller" líder sem depender de sistema externo.
 - Simplifica a administração e aumenta a resiliência.
- Kafka se torna autossuficiente (sem ZooKeeper).
- Melhor desempenho e menor latência na coordenação.



APLICAÇÃO PRÁTICA EM UM E-COMMERCE

Monitoramento em tempo real de um site de compras online.

- Produtores: O site de e-commerce atua como o produtor. Cada ação do usuário gera um evento que é enviado para o Kafka.
 - Exemplos de eventos: um clique em um produto, a adição de um item ao carrinho, ou a conclusão de uma transação de compra.
- Tópicos: Cada tipo de evento é direcionado para um tópico específico.
 - Tópicos usados: cliques_do_site, adicoes_ao_carrinho, transacoes_concluidas.
 - Isso garante que os dados sejam organizados e que os consumidores se inscrevam apenas nos fluxos de dados de seu interesse.
- Passo 3: Consumidores: Diferentes sistemas internos agem como consumidores, lendo os dados em tempo real para diversas finalidades.
 - Análise em Tempo Real, recomendação de Produtos, detecção de Fraudes:

ONDE O KAFKA É USADO

- Netflix: Usa Kafka para coletar dados de eventos dos usuários (reproduções, buscas, etc.). Esses dados alimentam em tempo real o sistema de recomendações de conteúdo e a personalização da interface, melhorando a experiência do usuário.
- PayPal: Utiliza Kafka para monitorar e processar transações e logs de segurança em tempo real. Isso permite que a empresa detecte e previna fraudes em milissegundos, garantindo a segurança.
- Pinterest: A plataforma usa Kafka para coletar todas as ações dos usuários (cliques, buscas) e construir um feed personalizado e dinâmico que aumenta a relevância do conteúdo.

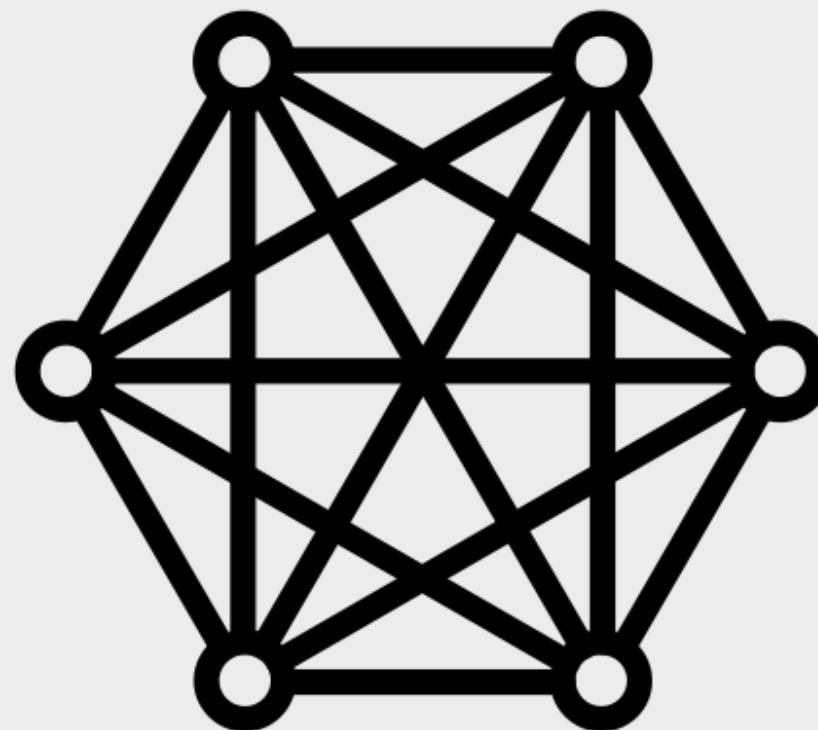


VANTAGENS DO KAFKA

- Escalabilidade: Adicionar mais corretores facilmente.
- Alta Performance: Milhões de mensagens por segundo.
- Durabilidade: Dados persistidos em disco.
- Tolerância a Falhas: Replicabilidade das mensagens.



LIMITAÇÕES E DESAFIOS



- Complexidade Operacional: A gestão de um cluster em produção pode ser desafiadora.
- Ordenação: A ordem só é garantida por partição.
- Curva de Aprendizagem: Requer tempo para dominar os conceitos.

SÍNTESE E RELEVÂNCIA

- O streaming é a nova forma de processar dados, permitindo que as empresas reajam a eventos em tempo real, em vez de esperarem pelo processamento em lote.
- O Kafka é a base de toda arquitetura moderna de streaming.
 - Ele garante que os dados sejam coletados, transportados e armazenados com segurança, de forma escalável e tolerante a falhas.
- Kafka é uma plataforma completa que sustenta todo um ecossistema de dados, viabilizando análises e ações em tempo real.