
# Projeto Final da Disciplina Infraestrutura Kafka – Pós Graduação MIT em Engenharia de dados: Big Data

Rafael Diniz Ramos

O Apache Kafka é um sistema de mensageria, utilizado em pipeline de dados, funcionando como um middleware entre a fonte de dados, seja ela um banco de dados relacional, uma rede social ou um streaming, até o cliente onde consumirá essa informação.

Podemos descrever 5 conceitos fundamentais do Kafka, são eles:

- Tópicos
- Partições
- Produtores e Consumidores
- Brokers
- Grupos de Consumidores

1. **Tópicos**

Os tópicos são categorias de mensagens que vão sendo armazenadas, na medida que são produzidas pelos Produtores. Elas são associadas de acordo com cada tópico, que posteriormente serão consumidas pelo cliente.

2. **Partições**

As partições são subdivisões dos tópicos, onde as mensagens ficam armazenadas, e podem ser acessadas por paralelismo, aumentando assim a eficiência do Kafka. O número dessas partições, determina o número máximo de consumidores que podem consumir informações daquele tópico, de maneira efetiva.

3. **Produtores e Consumidores**

Produtores são responsáveis em gerar as mensagens, que serão enviadas aos Brokers (servidores), onde serão gravadas no tópico correspondente, podendo ser um ou mais deles.

Os Consumidores irão ler essas mensagens, também poder ser em um ou mais tópicos.

4. **Brokers**

Os Brokers são servidores que comandam o cluster do Kafka, eles gerenciam as partições dos tópicos, recebem as mensagens dos Produtores e entregam para os Consumidores.

5. **Grupos de Consumidores**

Como o nome já diz, são grupos de consumidores que compartilham a leitura de um tópico. Através desses grupos, é feito o balanceamento de carga, onde a carga de leitura é distribuída entre os consumidores do grupo.

**Arquitetura Kafka**

Após a explicação acima dos conceitos fundamentais do Kafka, podemos descrever que a arquitetura do sistema começa com uma fonte de dados, os Produtores, onde enviam suas mensagens até os brokers para armazená-las em tópicos. Esses servidores, gerenciam os tópicos e suas partições para que as mensagens fiquem disponíveis aos Consumidores ou Grupos de consumidores, que irão lê-las.

**Utilização do Apache Kafka em bases NoSQL e SQL**

Apache Cassandra

O Kafka poder ser utilizado para ingestão de dados em tempo real para o Cassandra, para que seja feita análise de séries temporais, por exemplo.

SQL Server

O Kafka pode ser utilizados em tempo real para alimentar bases do SQL Server, de transações financeiras.

**Principais benefícios em utilizar o Apache Kafka**

- Alta taxa de transferência de dados --> o Kafka é capaz de lhe dar com alto volume de dados, com baixa latência e streaming de dados em tempo real.

- Escalabilidade horizontal --> o Kafka foi projetado para ser facilmente escalável, sendo necessário apenas adicionar nós ao cluster para que assim aumente sua capacidade de processamento e armazenamento.

- Tolerância a falhas --> os dados no sistema Kafka são replicados entre os nós do clusters, garantindo assim problemas contra falha de hardware.

- Integração com Big Data --> por ter alta capacidade de processamento de grandes volumes de dados e fácil escalabilidade, o Kafka tem fácil integração com Big Data, fazendo o papel de middleware no pipeline de dados.

**Pipeline de Dados**

Pipeline de dados é uma infraestrutura, um tunelamento, com diversas tecnologias empregadas, para fazer a extração do dado, o tratamento e o seu carregamento (ETL).

O Kafka pode ser utilizado fazendo a extração do dado de alguma fonte (Produtores) e posteriormente fazendo a carga dos dados para os consumidores finais.

Em meu dia a dia, o pipeline de dados seria nosso sistema ERP da empresa que registra as informações por exemplo de compra e venda. Eu extraio essas informações para um CSV com seus dados brutos, faço o tratamento, limpeza e enriquecimento com Python, e utilizo o Power BI para fazer análise exploratória de dados e criação de dashboard, para assim gerar insights de compra e venda e apresentar para a Diretoria.

## Abaixo segue o link do Colab para a parte prática do projeto.

https://colab.research.google.com/drive/1aM4AfwpXWRrO0c_VodD3vS8sV_mzHTyG?usp=sharing