Instituto Infnet	**TRABALHO FINAL DE CURSO** **DISCIPLINA KAFKA**
Prof. Leandro Maia Gonçalves	
Aluno: Leandro Leite	
Data: 11/03/24	

1. Escolha 5 conceitos fundamentais sobre o Apache Kafka e os descreva.

Os cinco principais conceitos fundamentais do Kafka são:

- Mensagem: é a unidade de dados representada dentro do Kafka como um array de bytes. Não possui um formato ou significado específico para o Kafka.
- Tópico: onde as mensagens são agrupadas em função de uma determinada categoria (nome). Seria análogo ao conceito de pasta num sistema de armazenamento de arquivos ou tabelas num modelo RDBMS.
- Partição: são as divisões dentro dos tópicos que ordenam as mensagens para serem consumidas.
- Produtor: são os responsáveis por publicar (gravar) as mensagens dos tópicos no broker (servidor) Kafka.
- Consumidor: são os responsáveis por consumirem (lerem) as mensagens de um determinado tópico a partir do broker Kafka.

2. Descreva como é a arquitetura do Apache Kafka.

A arquitetura do Apache Kafka pode ser descrita como uma plataforma distribuída para a transmissão de dados, onde o fluxo de dados tem a sua origem nos nós produtores (*producers*), que enviam suas mensagens organizadas em tópicos para os servidores Kafka (*brokers*) responsáveis por ordenar e armazenar esses dados (mensagens) em partições, onde as mensagens serão lidas pelos nós consumidores (*consumers*). A figura abaixo ilustra de forma simplificada esse conceito:

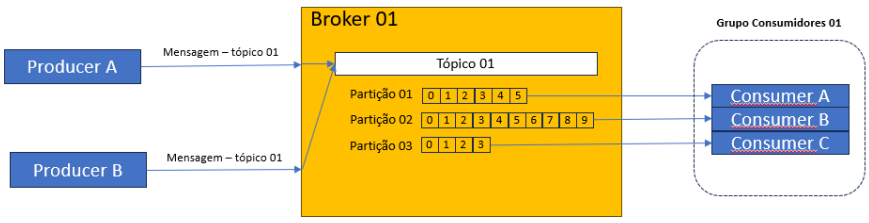


Figura 1- Arquitetura Kafka

Cabe destacar que poderão ser criados diversos tópicos e estes terão as suas respectivas partições associadas de forma independente das demais partições dos outros tópicos. Também, importante ressaltar que a partição de um tópico estará associada a um único consumidor do grupo a qual ele pertence, isto é, um consumidor poderá estar associado a uma ou mais partições de um mesmo tópico, porém, uma partição de um mesmo tópico estará associada a um único consumidor de um determinado grupo.

A arquitetura permite ainda que haja diversos servidores (brokers) dendro de um cluster Kafka, bem como diversos grupos de consumidores. A figura abaixo representa uma visão simplificada desta variação na arquitetura:

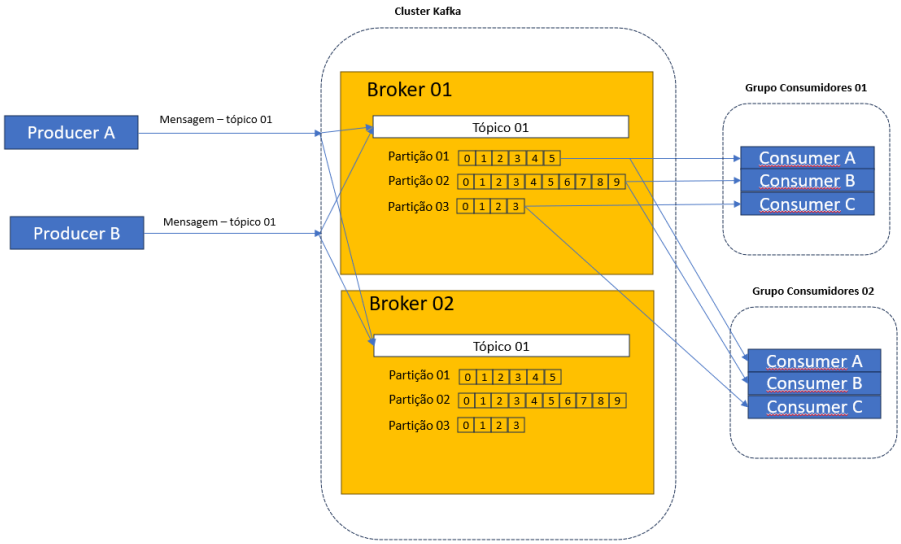


Figura 2 - Arquitetura cluster Kafka

3. Apresente exemplos de utilização do Apache Kafka em bases NoSQL e SQL.

Como exemplo de utilização do Apache Kafka em bases NoSQL podemos citar a sua utilização em pipelines de transmissão para compartilhamento de dados entre sistemas e aplicações desenvolvidas pela Red Hat e na integração entre a plataforma de marketplace e bases de dados SQL nas soluções disponíveis pela Salesforce, dentre outros casos.

Na pesquisa por exemplo de empresas e aplicações que utilizam o Apache Kafka em base de dados NoSQL encontramos: a utilização em sistema de processamento de dados em tempo real e integração com banco de dados NoSQL no aplicativo de música online do Spotfy e no processamento de dados em tempo real para atualização de notícias aos usuários do Twitter.

4. Descrevas os principais benefícios em utilizar o Apache Kafka.

Os principais benefícios no uso do Apache Kafka que podemos destacar são:

- Quebra da dependência entre os sistemas de origem dos dados e os sistemas destino dos dados, eliminando a complexidade e dependências existentes entre múltiplas integrações de sistemas;
- Arquitetura distribuía, resiliente e tolerante a falhas;
- Escalabilidade horizontal (possibilidade de escalar centenas de brokers simultâneos);
- Alta performance (latência menor que 10ms de processamento); e
- Solução robusta e estável utilizada por grandes empresas no mercado.

5. O que é um pipeline de dados?

Pipeline de dados pode ser definido como uma série de etapas que envolvem o processamento de dados para a sua análise e utilização para fins úteis. Um tipo especial de pipeline de dados é conhecido como ETL (Extação - Extract, Transformação – Transform e Carregamento - Load), que consiste numa etapa inicial de obtenção de dados brutos (extract) que serão tratados e transformados em dados úteis (transform) para a etapa final de processamento e carregamento (load) dessas informações em algum sistema de informação para a auxiliar na análise e na tomada de decisão dentro de determinado contexto.

6. Dê 2 (dois) exemplos de aplicações onde os pipelines de dados são utilizados em seu dia-a-dia

Como exemplo de aplicações do dia a dia que utilizam pipeline de dados, podemos citar:

- Uber: utiliza o pipeline de dados para transmissão dos dados em tempo real da localização do motorista para o cliente usuário do aplicativo;
- Netflix: as recomendações em tempo real para os usuários enquanto assistem determinados programas e séries de Tv são baseados no uso de pipeline de dados baseado em Kafka.

7. Tema:

Utilizaremos neste trabalho uma base de dados contendo exemplos de logs dos usuários da Netflix que pode ser utilizada para a personalização das sugestões de conteúdo dentro da plataforma.

Esta base de dados pública está disponível no Kaggle através da seguinte url:

https://www.kaggle.com/datasets/arjunajn/netflix-watch-log/data?select=All_ViewingActivity.csv

O dataset original disponibilizado já foi processado de forma a remover todas informações pessoais dos usuários reais que foram utilizados para a geração dessa massa de dados. Ele é composto por 05 (cinco) arquivos de dados do tipo csv.

Para fins de elaboração do presente trabalho, concentramos nossa análise no arquivo que contém os dados com os logs das atividades de visualização de cada título pelo usuário. O respectivo arquivo csv é denominado 'All_ViewingActivity .csv' e possui 1,35 MB de tamanho.

A estrutura dos dados armazenados pode ser verificada conforme a tabela abaixo:

Attribute	Description	Type
Profile Name	User profile name	String
Start Time	Start timestamp of the content playback	timestamp
Duration	Duration of the content played	timestamp
Attrtibutes	Interaction of the user to the content playback	String
Title	Title of the contente	String
Supplemental Video Type	Content type indicator. Blank means a movie/show	String
Device Type	Device the platform was accessed from	String
Bookmark	Last viewed timestamp during the playback	timestamp
Latest Bookmark	Recent bookmark timestamp	String
Country	Country the platform was accessed from	String

7.1. Analytics:

A partir da ingestão dos dados pretendemos responder as seguintes questões:

1. Qual a duração média que cada usuário assisti aos títulos na plataforma?
2. Qual a duração total dos títulos assistidos na plataforma pelos seus usuários ao longo dos dias da semana?

Nossas hipóteses iniciais para estas questões são que dado o grande volume de opções existentes na plataforma de vídeo e pela praticidade e facilidade de escolha, é provável que o tempo médio gasto por vídeo para cada um dos usuários seja bem baixa, isto é, algo em torno de 5 – 10 minutos.

Com relação a distribuição do tempo em que os usuários assistem aos programas na Netflix ao longo dos dias da semana, acreditamos que os finais de semana (sábado e domingo) representem mais de 80% do tempo total gasto.

Neste trabalho utilizaremos um pipeline de dados baseado na tecnologia Kafka. A figura abaixo representa de forma simplificada a estrutura do pipilene de dados utilizado:

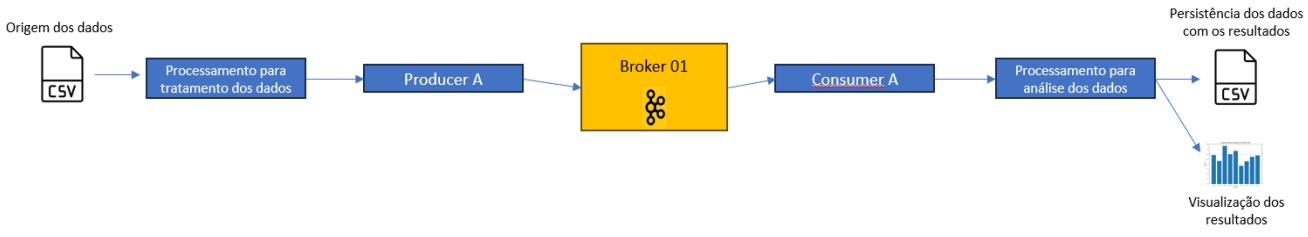


Figura 3 - Pipeline de dados do projeto

Todo o código fonte necessário para configuração e execução do pipeline de dados foi desenvolvido em linguagem Python utilizando o ambiente virtual do Google Colab. O notebook *Infnet_Kafka_LeandroLeite.ipynb* contendo todo o código está disponibilizado no Github e pode ser acessado pelo link abaixo:

https://github.com/leandroleite77/infnet_kafka

7.1.1. Resultados

A partir da análise dos 9992 registros de logs das atividades de visualização dos 09 usuários da nossa base de dados, pudemos verificar que o tempo médio de duração que cada usuário assiste a um determinado título é de 14.9 minutos.

```
36]  1 # Analisando os resulatdos:
     2 ViewActivity_df.describe()
```

	Duration Hour	Duration Minute	Duration Second	Total in Minutes
count	9992.000000	9992.000000	9992.000000	9992.000000
mean	0.030624	12.646417	25.066553	14.901664
std	0.197732	14.184310	17.710278	19.259993
min	0.000000	0.000000	0.000000	0.016700
25%	0.000000	0.000000	8.000000	0.533300
50%	0.000000	8.000000	23.000000	8.550000
75%	0.000000	21.000000	40.000000	21.616700
max	2.000000	59.000000	59.000000	176.666700

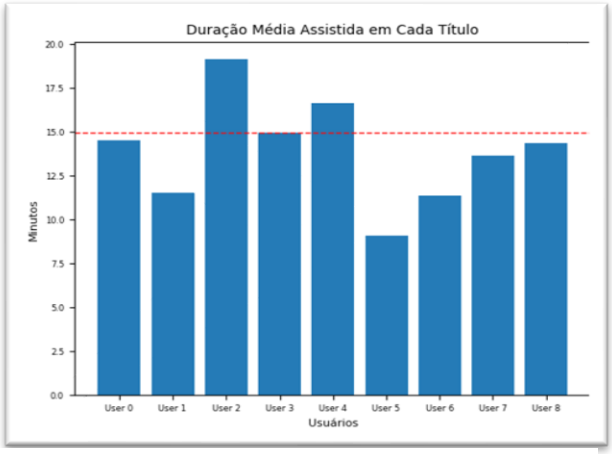
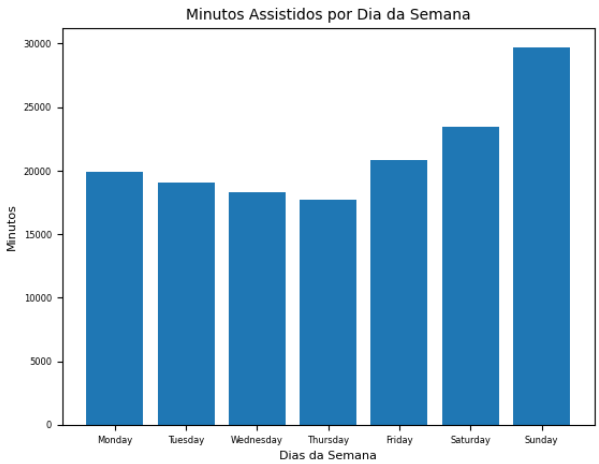


Figura 4 - Resultado análise duração média.

Com relação a distribuição do tempo em que os usuários assistem aos programas na Netflix ao longo dos dias da semana, podemos constatar que os dias em que os usuários mais assistem aos programas na Netflix são durante os finais de semana (sábado e domingo).



	weekday	Total in Minutes
0	Monday	19891.1663
1	Tuesday	19062.1355
2	Wednesday	18268.6690
3	Thursday	17695.0657
4	Friday	20804.5845
5	Saturday	23455.8840
6	Sunday	29719.9180

```
1 total_weekend = grouped_weekly['Total in Minutes'][5:7].sum()
2 total_week = grouped_weekly['Total in Minutes'][0:5].sum()
3
4 print(f'Percentual duração nos finais de semana: {total_weekend/total_week: .4f}')

Percentual duração nos finais de semana:  0.5555
```

Figura 5 - Tempo de duração distribuída nos dias da semana.

Porém, ao contrário do que imaginávamos, o tempo total assistido nos finais de semana representa 55,55% da duração total, ao invés dos 80% que considerávamos inicialmente antes da análise.

Como conclusão final, os resultados da análise foram gravados em dois arquivos no formato csv, finalizando o pipilene de dados proposto:

- *usergroup_duration.csv* – dados da análise da duração média
- *grouped_weekly.csv* – dados da análise de duração distribuída pelos dias da semana