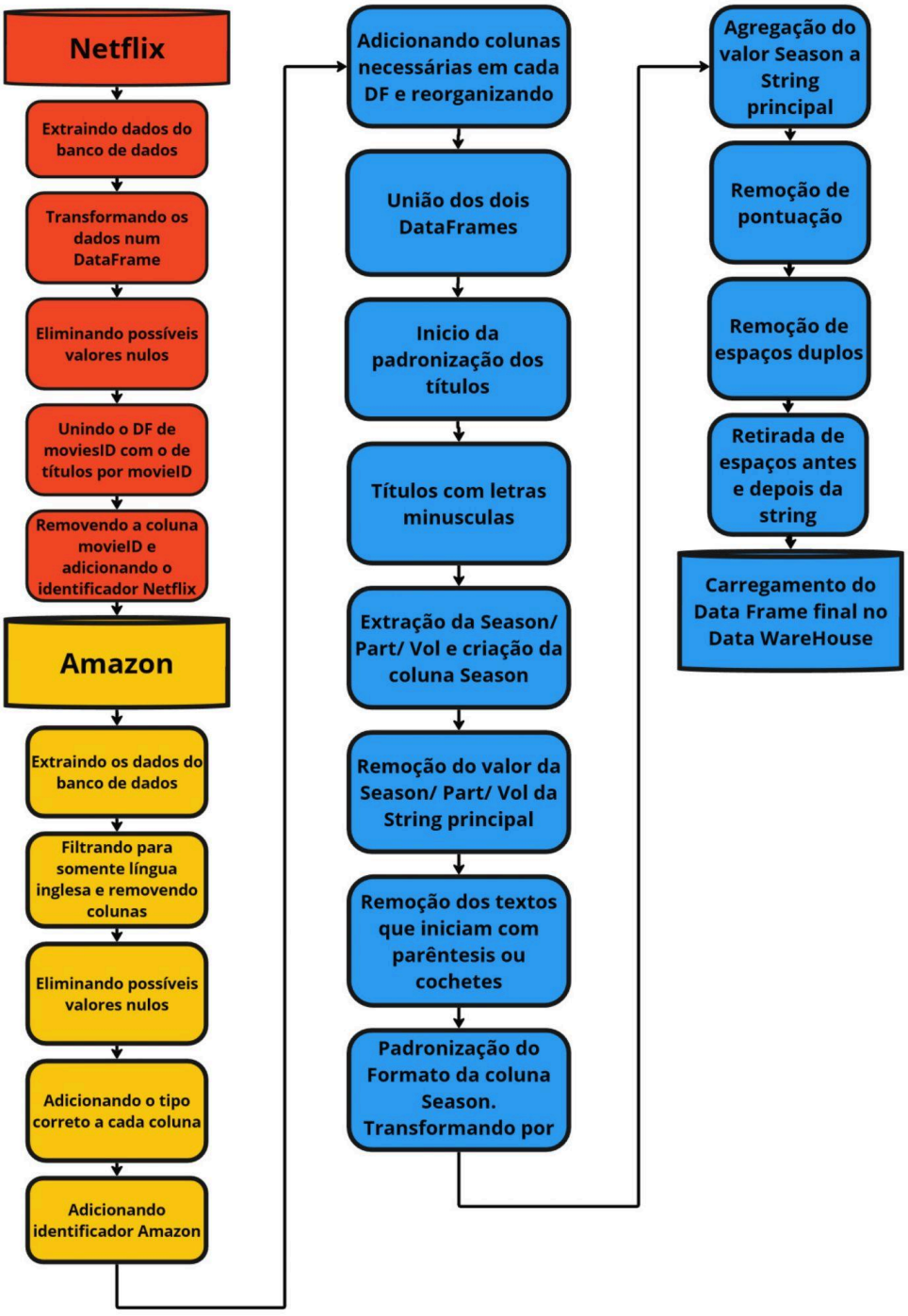
# Desafio E-Core Engenharia Dados

Marcelo Yuri Sampaio de Freitas

## Diagrama lógico de processamento dos dados




Obs: Nesse diagrama é mostrado a leitura e tratamento dos dados da Amazon seguindo a sequência lógica após o tratamento dos dados da Netflix, foi colocado dessa forma somente para seguir o formato do arquivo notebook de transformação disponibilizado, mas esses processos são independentes e podem ser feitos de forma paralela.

# E-mail para o Arquiteto de dados

Prezado arquiteto,

finalizei o processamento dos dados referentes aos dados históricos da Netflix e da Amazon. Disponibilizo em anexo neste email o link para o arquivo final transformado já no formato que melhor atende a necessidade do cliente e um arquivo de consultas SQL para responder às perguntas de negócio da 5Gflix. Tenho também alguns pontos para serem alinhados, são eles:

- ➔ O SGBD/solução de banco de dados escolhida para a consulta do time de BI foi: **Data Warehouse que pode ser Amazon Redshift, Google BigQuery ou Snowflake**.
- ➔ Formato dos arquivos e motivo da escolha: **.parquet**

Sobre a escolha da solução de banco de dados escolhida, realizei uma análise das opções que temos disponíveis, na qual comparei as vantagens de desvantagens de cada uma e para qual tipo de problema cada solução é indicada. As opções que temos com as suas características, vantagens e desvantagens são:

1. **Banco de dados Relacional** (MySQL, PostgreSQL, Microsoft SQL Server, Oracle Database)
    a. **Característica**: Adequado para dados estruturados e normalizados.
    **b. Vantagens**: Bom para consultas complexas e relatórios detalhados.
    **c. Desvantagens**: Dificuldade para grande volume de dados e alta escalabilidade horizontal.
    **d. Conclusão:** O volume dos dados que estamos trabalhando é considerável, são **25 anos** de dados históricos, então uma solução como essa apresentaria uma limitação de desempenho para o cliente, o que piora consideravelmente a sua experiência.
2. **Data WareHouse** (Amazon Redshift, Google BigQuery ou Snowflake)
    **a. Característica:** Projetado para análise de grande volume de dados.
    **b. Vantagens:** Ideal para análises BI devido a escalabilidade, integração dos dados com processo de ETL, compatibilidade com diversas ferramentas de BI como Tableau, Power BI, etc. Também permite escalabilidade horizontal, processamento paralelo de consultas e otimização para grande volume de dados.
    **c. Desvantagens:** Pode ser caro e difícil de configurar se comparado a bancos relacionais.
    **d. Conclusão:** Dado que o volume de dados é considerável e que serão realizadas consultas e visualizações pelo time de BI da 5GFlix, essa solução de banco de dados é a mais recomendada, pois entregará uma boa performance para o cliente e pode ser integrado ao sistema de BI que eles utilizam devido a boa compatibilidade com diversas ferramentas.
3. **Bancos de dados NoSQL** (MongoDB, Cassandra, Couchbase)
    **a. Características:** Ideal para dados não estruturados ou semiestruturados.
    **b. Vantagens:** Flexibilidade na estrutura dos dados e escalabilidade horizontal para grande volume de dados

    c. **Desvantagens:** Menos suporte a consultas complexas. Necessário um esforço adicional para modelagem dos dados e otimização das consultas.
    d. **Conclusão:** Não é adequado para o nosso caso, pois os dados da Netflix e Amazon são estruturados e o time de BI precisará realizar muitas consultas e análises.
4. **Armazenamento em Nuvem e DataLakes** (Amazon S3 + AWS Glue, Azure Datalake, Google Cloud Storage)
    a. **Características:** Permite armazenamento de grande volume de dados.
    b. **Vantagens:** Escalabilidade praticamente ilimitada e suporte para diferentes tipos de dados
    c. **Desvantagens:** Necessita de ferramentas adicionais para processamento e análise como: Apache Spark, AWS Athena, entre outros.
    d. **Conclusão:** Aqui são armazenados dados de forma bruta com diversos formatos, para análise BI esses dados precisam ser filtrados e transformados utilizando outra ferramenta. Além disso, o processo de consulta não é tão otimizado quanto em Data WareHouse, o que iria prejudicar o desempenho.

Quanto à escolha do formato do arquivo, assim como o banco de dados, existem diversas soluções disponíveis, a seguir as características, vantagens e desvantagens de cada uma delas.

1. **CSV**
    a. **Vantagens:** Simplicidade e ampla compatibilidade com diversas ferramentas e sistemas
    b. **Desvantagens:** Não é eficiente para grande volume de dados devido a falta de compressão e suporte limitado a dados complexos como estruturas aninhadas, por exemplo, JSON.
    c. **Conclusão:** Conforme já mencionado, o volume de dados que será trabalhado é considerável, então escolher CSV é desfavorável ao desempenho do sistema.
2. **Parquet**
    a. **Vantagens:** É um formato binário desenvolvido pela Apache para melhorar a eficiência de leitura e compressão dos dados. Tem formato em coluna e é otimizado para grande volume de dados. Também é otimizado para consultas rápidas e permite dados complexos.
    b. **Desvantagens:** Mais difícil de ser lido diretamente por humanos.
    c. **Conclusão:** Dado que nosso objetivo é trabalhar com um grande volume de dados e eles precisarão ser consultados diversas vezes pelo time de BI, esse é o formato que tem o melhor resultado para a necessidade do cliente.
3. **ORC**
    a. **Vantagens:** Otimizado para compressão, processamento paralelo e leitura eficiente.
    b. **Desvantagens:** Menos suportado fora do ecossistema Hadoop e ferramentas relacionadas.
    c. **Conclusão:** Utilizar esse formato poderia resultar em um bom desempenho, mas a incompatibilidade com as ferramentas de BI pode ser um gargalo para o cliente.
4. **Avro**
    a. **Vantagens:** Compacto e eficiente para leitura e escrita;

   b. **Desvantagens:** Menos eficiente para consultas em comparação com os formatos colunar.
   c. **Conclusão:** Esse formato tem pior performance para consultas, o que não é desejado na solução requerida
5. **JSON**
   a. **Vantagens:** Flexível e fácil de ser lido por humanos. Boa performance para dados semi estruturados
   b. **Desvantagens:** Menos eficiente para armazenamento e leitura de dados para grande volume de dados.
   c. **Conclusão:** Os dados trabalhados são estruturados e o objetivo é a leitura e armazenamento rápido, então não é adequado.

Fico à disposição para eventuais dúvidas ou necessidades.

Atenciosamente,
Marcelo Yuri Sampaio de Freitas
Data Engineer

## E-mail para o Cliente

Prezado Alan,

Conforme solicitado, foi realizada a análise da arquitetura recomendada tanto para responder às perguntas de negócio, quanto para auxiliar os times de BI. Os resultados finais serão explicitados aqui neste email. Antes de responder cada uma das perguntas de negócio solicitadas, descreveremos brevemente a arquitetura recomendada.

O objetivo principal da arquitetura é responder aos desafios de negócio e prover as melhores condições de utilização dos dados para os times de BI da 5Gflix. Tendo isso em vista, foi desenvolvido o processo de ETL dos dados que tem como resultado arquivos contendo os dados de ambas as concorrentes de modo padronizado, a fim de que seja possível a comparação entre elas. Após esse processo de ETL, os arquivos resultantes devem ser disponibilizados num Data Warehouse para o time de BI no formato .parquet. Essa arquitetura permitirá a utilização de diferentes ferramentas de BI e conferirá uma boa performance para consultas e leituras.
Seguem as respostas para os desafios de negócio:

- ➔ **Quantos filmes estão disponíveis na Amazon? E na Netflix?**
  - ◆ Amazon: 328.205
  - ◆ Netflix: 17.240

- ➔ **Dos filmes disponíveis na Amazon, quantos % estão disponíveis na Netflix?**
  - ◆ 4,2677% (14007 títulos)

- ➔ **O quão perto a média das notas dos filmes disponíveis na Amazon está dos filmes disponíveis na Netflix?**
  - ◆ 0,65, sendo a nota da Amazon maior na média.

- **Qual ano de lançamento possui mais filmes na Netflix?**
  - 2004 com 1435 filmes lançados

- **Quais filmes que não estão disponíveis no catálogo da Amazon foram *melhor avaliados? *Melhores notas são as 4 e 5**
  - São no total 3233 títulos, segue em anexo a lista completa. Alguns exemplos com nota 5:
    i. Copper Mountain
    ii. Bears Imax
    iii. Black Scorpion 2 ground zero
    iv. Caillou caillous summertime and other adventures
    v. Chef 3
    vi. Chihwaseon painted fire

- **Quais filmes que não estão disponíveis no catálogo da Netflix foram *melhor avaliados? *Melhores notas são as 4 e 5**
  - São no total 277.443 títulos, segue em anexo a lista completa. Alguns exemplos:
    i. Barney animal abc
    ii. The object
    iii. Janis ian live from grand center
    iv. Unbeatable harold
    v. Surprise beginnings

- **Quais os 10 filmes que possuem mais avaliações nas plataformas?**
  - **Amazon**:
    i. Pilot
    ii. Bosch 1
    iii. Downtown Abbey 1
    iv. Downtown Abbey 3
    v. Transparent 1
    vi. Justified 1
    vii. Frozen
    viii. Under the Dome 1
    ix. Downtown Abbey 4
    x. Downtown Abbey 2
  - **Netlflix**:
    i. Miss Congeniality
    ii. Independence Day
    iii. The Patriot
    iv. The GodFather
    v. The day after tomorrow
    vi. Pirates of Caribean
    vii. Pretty Woman
    viii. Twister
    ix. Gone in 60 seconds
    x. Forrest Gump

➔ **Quais são os 5 clientes que mais avaliaram filmes na Amazon e quantos produtos diferentes eles avaliaram? E na Netflix?**

- **Amazon:**
  i. ID:43430756, 6381 títulos
  ii. ID: 18116317, 3992 títulos
  iii. ID: 52287429, 2978 títulos
  iv. ID: 52496677, 2895 títulos
  v. ID: 20018062, 2786 títulos
- **Netflix:**
  i. ID: 305344, 17129 títulos
  ii. ID: 387418, 16919 títulos
  iii. ID: 2439493, 16087 títulos
  iv. ID: 1664010, 15380 títulos
  v. ID: 2118461, 14404 títulos

➔ **Quantos filmes foram avaliados na data de avaliação mais recente na Amazon?**

- 17.613
-

➔ **Quais os 10 filmes mais bem avaliados nesta data?**

| Título | Avaliação |
|---|---|
| Digging for fire | 5 |
| Mad dogs 1 | 5 |
| Hotel transylvania | 5 |
| Downton abbey 3 | 5 |
| The good wife 3 | 5 |
| Immortals | 5 |
| Shaun the sheep 4 | 5 |
| Grantchester 1 | 5 |
| The switch | 5 |
| Cloud atlas | 5 |

➔ **Quantos filmes foram avaliados na data de avaliação mais recente na Netflix?**
- 6.838

➔ **Quais os 10 filmes mais bem avaliados nesta data?**

| Título | Avaliação |
|---|---|
| The alamo | 5 |
| Ikiru | 5 |
| The twilight zone vol 39 | 5 |
| Tom and Jerry's greatest chases | 5 |
| 20 | 5 |
| Friends 1 | 5 |
| Strictly ballroom | 5 |
| Finding neverland | 5 |
| The green mile | 5 |
| The man with the golden gun | 5 |

Fico à disposição para eventuais dúvidas.

Atenciosamente,
Marcelo Yuri Sampaio de Freitas
Data Engineer