Documento de Arquitetura

Pipeline de Big Data para Análise de Churn em Telecomunicações

---

Equipe: Leonardo Azevedo, Walter Barreto, Mariana Belo

Data: 13 de Outubro de 2024

Disciplina: Fundamentos de Big Data

1. Visão Geral do Projeto

O Desafio

Nossa empresa de telecomunicações enfrenta um desafio crítico: estamos perdendo clientes a uma taxa alarmante de 26,54%. Isso representa uma perda mensal de R$ 139.130,85 em receita - recursos que poderiam ser investidos em melhorias e expansão dos serviços.

Para enfrentar este problema, desenvolvemos uma solução completa de Big Data que nos permite identificar padrões comportamentais e fatores que levam os clientes a cancelar seus serviços. Com essas informações em mãos, podemos implementar estratégias de retenção personalizadas e eficazes.

26,54%
Taxa de Churn

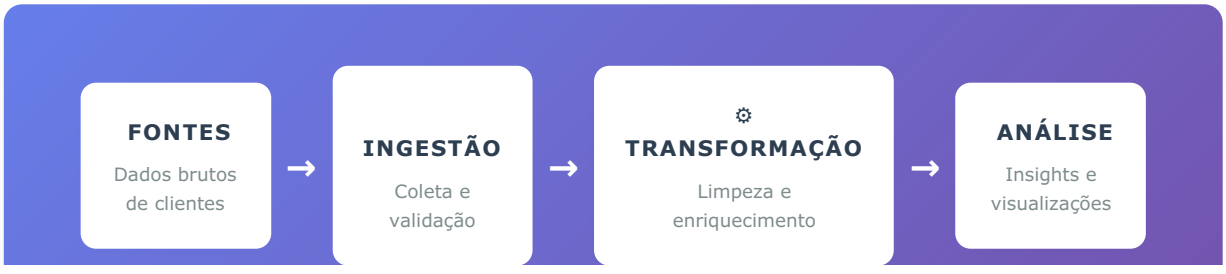
R$ 139.130
Perda Mensal em Receita

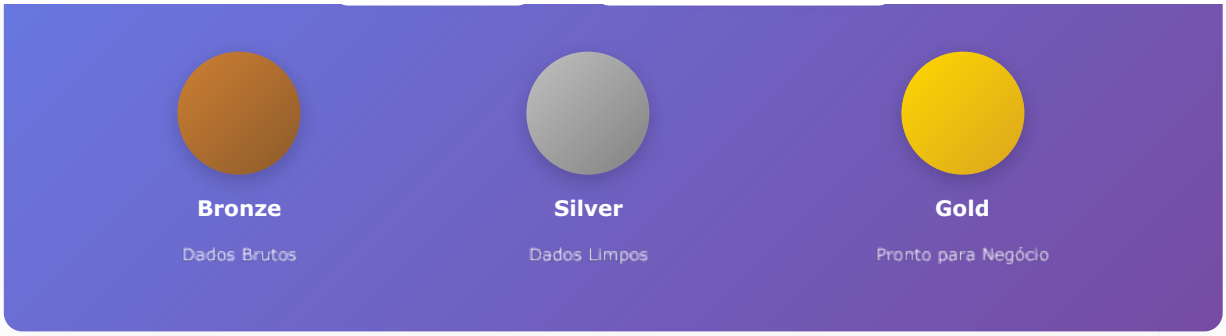
7.043
Clientes Analisados

23
Variáveis de Análise

2. Arquitetura do Pipeline de Dados

Nosso pipeline foi desenvolvido seguindo as melhores práticas da indústria, implementando a arquitetura Medallion (Bronze-Silver-Gold). Essa abordagem garante a qualidade dos dados em cada etapa do processamento, desde a ingestão bruta até a geração de insights estratégicos.





3. Detalhamento das Etapas

3.1 Camada Bronze: Ingestão dos Dados

Na primeira camada, realizamos a coleta dos dados brutos de 7.043 clientes. Utilizamos um dataset público de alta qualidade proveniente do Kaggle, que contém informações demográficas, contratuais e de serviços dos clientes no período de 2019-2020.

> #### Principais Validações Realizadas
>
> - Verificação de integridade dos dados
> - Identificação e remoção de duplicatas (0 encontradas)
> - Validação de tipos de dados
> - Registro de metadados de ingestão

3.2 Camada Silver: Transformação Inteligente

Esta é a etapa onde a mágica acontece. Realizamos uma limpeza profunda dos dados e criamos 10 novas features inteligentes que nos ajudam a entender melhor o comportamento dos clientes. Algumas das features mais importantes incluem:

- **ChurnRiskScore:** Um score de 0 a 9 que indica a probabilidade de um cliente cancelar o serviço
- **SatisfactionScore:** Métrica de 0 a 10 baseada no comportamento e engajamento do cliente
- **TenureGroup:** Classificação do tempo de permanência (Novo, Médio, Longo prazo)
- **NumServicos:** Total de serviços contratados por cada cliente
- **IsPremium:** Identificação de clientes premium com alto valor

Além disso, implementamos normalização dos valores monetários para facilitar análises comparativas e garantir que todas as variáveis tenham a mesma escala de importância.

> #### Resultados da Transformação
>
> - ✅ 7.043 registros processados com sucesso
> - ✅ 33 colunas finais (10 novas features criadas)
> - ✅ 0 valores ausentes após tratamento
> - ✅ Redução de 80,7% no tamanho dos arquivos usando formato Parquet

3.3 Camada Gold: Insights Estratégicos

Na camada final, geramos datasets agregados e visualizações que respondem às perguntas críticas do negócio. Criamos quatro conjuntos de dados especializados:

1. **Métricas por Contrato:** KPIs segmentados por tipo de contrato (mensal, anual, bienal)
2. **Churn por Segmento:** Análise detalhada de cancelamentos por grupos de clientes

3. **Perfil de Alto Risco:** Características dos clientes mais propensos a cancelar
4. **Correlações com Churn:** Features que mais influenciam a decisão de cancelamento

4. Tecnologias Utilizadas

Selecionamos um conjunto robusto de tecnologias Python que são padrão na indústria de Data Science e Big Data:

Tecnologia	Propósito	Por que escolhemos
Python 3.x	Base do projeto	Linguagem padrão para Data Science com vasto ecossistema
Pandas	Manipulação de dados	Biblioteca mais popular para análise de dados tabulares
NumPy	Computação numérica	Operações vetorizadas de alta performance
Parquet/PyArrow	Armazenamento eficiente	Redução de 80% no tamanho e acesso mais rápido aos dados
Matplotlib/Seaborn	Visualizações	Gráficos profissionais e estatísticos de alta qualidade
Scikit-learn	Machine Learning	Normalização e pré-processamento de dados

5. Nossa Equipe e Responsabilidades

O projeto foi desenvolvido com uma divisão clara de responsabilidades, permitindo que cada membro da equipe se especializasse em uma área específica do pipeline:

Leonardo Azevedo

Arquitetura e Ingestão

- Estrutura Medallion
- Pipeline de ingestão
- Validações de qualidade
- Documentação técnica

Walter Barreto

Transformação de Dados

- Limpeza de dados
- 10 novas features
- Normalização
- Otimização de performance

Mariana Belo

Análise e Visualização

- Análise exploratória
- 7 visualizações
- Dashboard executivo
- Insights estratégicos

6. Próximos Passos

Embora o pipeline atual esteja completo e funcional, já identificamos oportunidades de evolução para torná-lo ainda mais robusto e escalável:

- **Cloud Computing:** Migração para AWS, Azure ou GCP para maior escalabilidade
- **Processamento em Tempo Real:** Implementação de Apache Kafka para streaming de dados
- **Orquestração Automatizada:** Uso do Apache Airflow para agendamento e monitoramento
- **Dashboards Interativos:** Criação de painéis dinâmicos com Tableau ou Power BI
- **Machine Learning Avançado:** Desenvolvimento de modelos preditivos para antecipar churn

Conclusão

Desenvolvemos com sucesso um pipeline completo de Big Data que transforma dados brutos em insights acionáveis. Nossa arquitetura Medallion garante qualidade e rastreabilidade em cada etapa, enquanto as tecnologias escolhidas proporcionam eficiência e preparação para escalabilidade futura.

Status atual: Pipeline totalmente implementado e operacional!