Luis: RM565324

Adeilson: RM566282

SUMÁRIO

Sumário

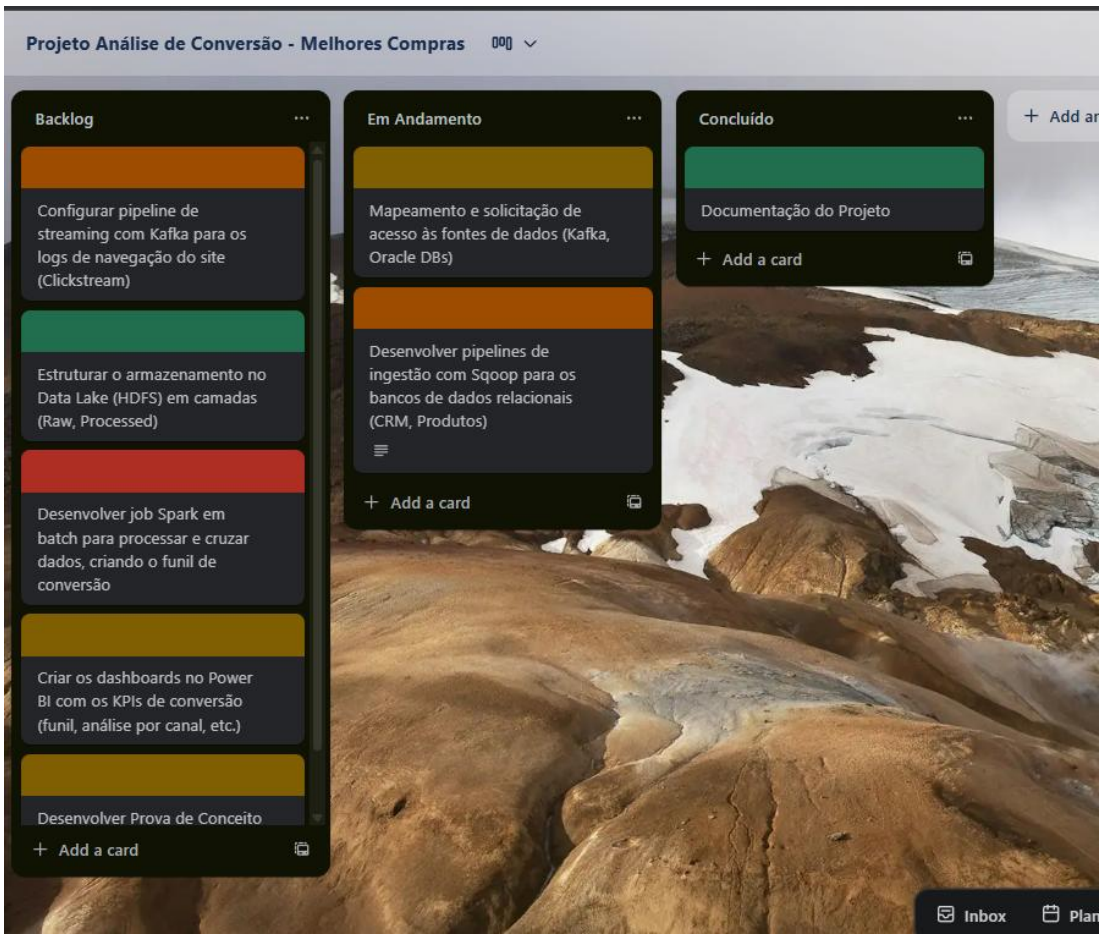
2 desafio(s) enfrentado(s) pela melhores compras

2.1 Contextualização do Problema

O problema escolhido foi o 3 - Diminuição da taxa de conversão (usuário anônimo para cliente) – em razão de que se caracteriza como um dos principais desafios das empresas atualmente. Em um cenário de mercado altamente competitivo e com custos de aquisição de tráfego (via anúncios digitais) cada vez mais elevados, a capacidade de transformar um visitante em um cliente pagador não é apenas um indicador de performance, mas uma questão de sustentabilidade financeira.

Para diagnosticar e reverter essa tendência, a conexão com Big Data é direta e essencial. É preciso analisar os dados de comportamento de milhões de usuários para encontrar a verdadeira causa da baixa conversão, algo que só essa tecnologia permite.

3 Planejamento das atividades



https://trello.com/invite/b/68e4851d7d7d4725ed666b4b/ATTIdae74bf7d017fa3838c6c35dd0d5432cBAFE48FF/projeto-analise-de-conversao-melhores-compras

4 origem dos dados

Para realizar um diagnóstico preciso sobre a queda na taxa de conversão, foram selecionadas quatro fontes de dados primárias. Elas nos permitirão cruzar informações de marketing, comportamento de navegação e dados cadastrais, fornecendo uma visão completa da jornada do usuário.

4.1 Panorama geral das fontes de dados

Origem	Formato	Velocidade	Volume	Horário Coleta	Localização	Proprietário
Logs de Navegação (Clickstream)	JSON	Streaming (Real-time)	Alto (Milhões de eventos/dia)	Contínuo	Servidores Web da Aplicação	Equipe de E-commerce
BD de Campanhas de Marketing	Tabela (CSV/API)	Batch (Diário)	Médio (Milhares de registros/dia)	Diariamente às 02:00	Plataformas (Google/Meta Ads)	Equipe de Marketing
BD de Clientes (CRM)	Tabela (SQL)	Batch (Diário)	Médio (Novos clientes/dia)	Diariamente às 23:00	Banco Oracle (On-premise)	Equipe de Vendas
BD de Produtos	Tabela (SQL)	Batch (Diário)	Baixo (Atualizações pontuais)	Diariamente às 23:30	Banco Oracle (On-premise)	Equipe de Produto (P&D)

4.2 Justificativas das Fontes de Dados

- Logs de Navegação (Clickstream): Essencial para mapear a jornada do usuário passo a passo e descobrir os pontos exatos do site onde eles desistem da compra.
- BD de Campanhas de Marketing: Serve para medir a eficiência dos anúncios, identificando quais campanhas atraem visitantes que realmente compram e quais apenas geram tráfego de baixa qualidade.
- BD de Clientes (CRM): Usado como base de comparação para entender o perfil e o comportamento de um "cliente de sucesso", contrastando com os visitantes que não convertem.

- BD de Produtos: Permite investigar se a baixa conversão é um problema geral ou se está concentrada em produtos ou categorias específicas do site.

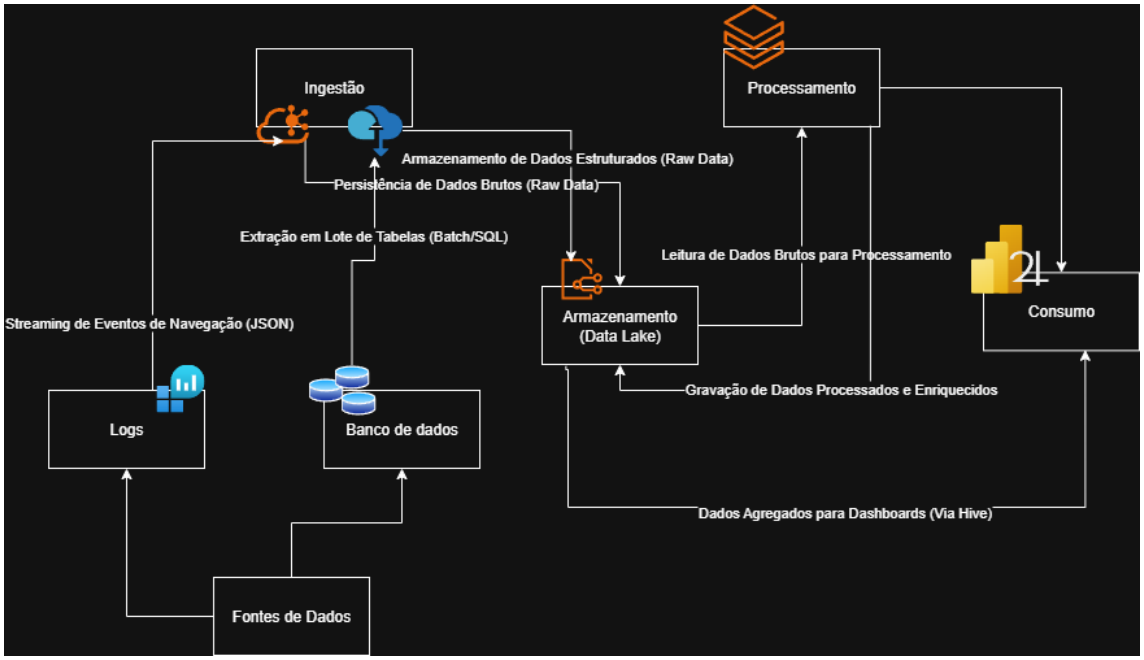
4.3 Detalhamento das fontes de dados

Tabela	Apelido	Descrição	Interessados	Dono da informação	Retenção
T_PRODUTO	Prod	Armazena todos os dados cadastrais e de inventário dos produtos comercializados pela "Melhores Compras".	Marketing, Vendas, Logística, BI	Diretor de Produto (P&D)	Produtos são mantidos por até 2 anos após a última venda. Backup diário.

Coluna	Tipo de Dado	Descrição	Exemplo
ID_PRODUTO	NUMBER(10)	Identificador único e sequencial para cada produto. Chave primária.	837492
NM_PRODUTO	VARCHAR2(255)	Nome comercial do produto exibido no site.	Smart TV LED 50" 4K Samsung
DS_CATEGORIA	VARCHAR2(100)	Categoria principal à qual o produto pertence.	Eletrônicos
VL_PRECO_ATUAL	NUMBER(10,2)	Preço de venda atual do produto em Reais (R$).	2199.90
QT_ESTOQUE	NUMBER(5)	Quantidade de unidades do produto disponíveis em estoque.	150
DT_CADASTRO	DATE	Data em que o produto foi cadastrado no sistema pela primeira vez.	05/03/2024

5 Arquitetura de Solução Big Data / Pipeline de Dados

5.1 Desenho da Arquitetura



https://drive.google.com/file/d/1b88hmb2UuLKvpQj2dDB5jLT0H8Co63ON/view?usp=sharing

5.2 Justificativa da Arquitetura

A arquitetura proposta, baseada no ecossistema open-source Apache Hadoop e Spark, foi escolhida por ser uma solução de padrão industrial que garante alta escalabilidade horizontal e baixo Custo Total de Propriedade (TCO), evitando o aprisionamento tecnológico (vendor lock-in). Tecnicamente, ela implementa um padrão Lambda simplificado, utilizando o Spark como um motor unificado para o processamento tanto em **lote (batch)**, para análises históricas de funil, quanto em **tempo real (streaming)**, para análise de eventos correntes. Essa capacidade dual, combinada com o suporte nativo do Spark para Machine Learning (MLlib), oferece uma plataforma flexível e robusta, capaz de suportar desde a construção de dashboards operacionais até futuras demandas de análise preditiva.

5.3 Detalhamento da Arquitetura

A arquitetura de dados opera em um fluxo contínuo, iniciando com a camada de **ingestão**, onde **Apache Kafka** captura o streaming de logs de navegação em tempo real, enquanto **Apache Sqoop** transfere em lote (batch) os dados dos bancos de dados relacionais. Todos esses dados brutos, de fontes heterogêneas, são centralizados e persistidos no **HDFS**, que atua como o Data Lake da solução. A camada de **processamento** é unificada no **Apache Spark**, cujo motor em memória executa jobs de batch para análises históricas (ETLs, funis), streaming para análises de eventos correntes e Machine Learning via biblioteca MLlib. Finalmente, na camada de **consumo**, os dados já processados são expostos para ferramentas de Business Intelligence através de uma interface SQL provida pelo **Apache Hive**, e para a análise aprofundada de cientistas de dados via **Jupyter Notebooks** utilizando a API PySpark.

REFERÊNCIAS

FIAP. **Fase 6: Arquitetura de Big Data e Soluções Paralelas e Distribuídas**. São Paulo: FIAP ON, 2025. Material didático do curso de Data Science. Disponível em: https://on.fiap.com.br/local/salavirtual/conteudo-digital.php.