

****Arquitetando o Universo Big Data da Melhores Compras****

Andre Luiz Sazana Waleczki | RM:559685

Guilherme Vinícius dos Santos | RM:560564

Henrique Caproni Siqueira | RM:560105

Renan Thiago Aviz e Silva | RM:560849

Thiago Evangelista Dias | RM:559403

Versão 3

**HISTÓRICO DE VERSÕES**

|  |  |  |  |
| --- | --- | --- | --- |
| Versão | Data | Responsável | Descrição |
| 1 | 18/06/2024 | Patrícia Maura Angelini | Versão Inicial Template PBL Fase 6 - CAP 01 - ARQUITETANDO O UNIVERSO BIG DATA DA MELHORES COMPRAS |
| 2 | 27/06/2024 | Rita de Cássia Rodrigues | Revisão acadêmica |
| 3 | 21/04/2025 | Andre Luiz Sazana Waleczki | Criação de Conteúdo academico |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

FICHA CATALOGRÁFICA   
**[NÃO PREENCHER - PARA USO DO DEPTO DE EAD E BIBLIOTECA]**

A000a Sobrenome, Nome

Título [livro eletrônico] / Nome Sobrenome. -- São Paulo : Fiap, 2016.

x MB ; ePUB

Bibliografia.

ISBN 000-00-00000-00-0

Categoria. 2. Subcategoria. S., Nome. II. Título.

CDU 000.000.00

RESUMO

Template para atividade de PBL fase 6 1º ano TSC.

Palavras-chave: PBL. FASE 6. TEMPLATE

LISTA DE Figuras

[Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente Figura 1 – Backlog e Atividades no Quadro Kanban 12](#_Toc196250548)

[Figura 1 – Detalhe item de backlog 12](#_Toc196250549)

[Figura 2 - Detalhe item de backlog 13](#_Toc196250550)

LISTA DE QUADROS

[Quadro 1 – Quadro resumo das tarefas do PBL **Erro! Indicador não definido.**](#_Toc169645325)

LISTA DE TABELAS

[Tabela 1 – Origem de Dados 16](#_Toc169645329)

[Tabela 2 - Dicionário de dados de colunas de tabelas 17](#_Toc169645330)

LISTA DE CÓDIGOS-FONTE

No table of figures entries found.LISTA DE ComandoS de prompt do sistema operacional

No table of figures entries found.

Sumário

[ARQUITETANDO O UNIVERSO BIG DATA DA MELHORES COMPRAS 11](#_Toc196250937)

[1 desafio(s) enfrentado(s) pela melhores compras 11](#_Toc196250938)

[1.1 Contextualização do Problema 11](#_Toc196250939)

[2 Planejamento das atividades 12](#_Toc196250940)

[2.1 Kanban das Atividades 12](#_Toc196250941)

[2.2 Detalhamento das Atividades 12](#_Toc196250942)

[3 origem dos dados 16](#_Toc196250943)

[3.1 Panorama geral das fontes de dados 16](#_Toc196250944)

[3.3 Detalhamento das fontes de dados 17](#_Toc196250945)

[4 Arquitetura de Solução Big Data / Pipeline de Dados 17](#_Toc196250946)

[4.1 Desenho da Arquitetura 17](#_Toc196250947)

[Diagrama

Descrição gerada automaticamente 17](#_Toc196250948)

[4.1 Justificativa da Arquitetura 18](#_Toc196250949)

[4.3 Detalhamento da Arquitetura 18](#_Toc196250950)

[GLOSSÁRIO 19](#_Toc196250951)

# ****ARQUITETANDO O UNIVERSO BIG DATA DA MELHORES COMPRAS****

# 1 desafio(s) enfrentado(s) pela melhores compras

## 1.1 Contextualização do Problema

Após análise dos principais desafios enfrentados pela Melhores Compras, a equipe Breaking Data optou por trabalhar em cima do problema do abandono de carrinho de compras em larga escala.

Essa escolha foi feita porque o abandono de carrinho impacta diretamente o faturamento da empresa, sendo uma perda concreta de receita. Além disso, é um problema que permite o uso intenso de tecnologias de Big Data, como análise em tempo real e machine learning para recomendação e retargeting.

Identificar padrões de abandono e agir rapidamente são fatores críticos para melhorar a taxa de conversão, a experiência do cliente e a competitividade da Melhores Compras frente ao mercado.

# 2 Planejamento das atividades

### 2.1 Kanban das Atividades

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente Figura 1 – Backlog e Atividades no Quadro Kanban

Fonte: Elaborado pelo autor (2025)

### 2.2 Detalhamento das Atividades

Interface gráfica do usuário, Texto, Site

Descrição gerada automaticamente

Figura 2 – Detalhe item de backlog

Fonte: Elaborado pelo autor (2025)

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

Figura 3 - Detalhe item de backlog

Fonte – Elaborado pelo autor (2025)

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

Figura 3 - Detalhe item de backlog

Fonte – Elaborado pelo autor (2025)

Interface gráfica do usuário, Texto, Site

Descrição gerada automaticamente

Figura 4 - Detalhe item de backlog

Fonte – Elaborado pelo autor (2025)

Interface gráfica do usuário, Aplicativo, Site

Descrição gerada automaticamente

Figura 5 - Detalhe item de backlog

Fonte – Elaborado pelo autor (2025)

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

Figura 6 - Detalhe item de backlog

Fonte – Elaborado pelo autor (2025)

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

Figura 7 - Detalhe item de backlog

Fonte – Elaborado pelo autor (2025)

# 3 origem dos dados

## 3.1 Panorama geral das fontes de dados

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Origem | Formato | Velocidade | Volume | Horário Coleta | Localização | Proprietário |
| Logs de navegação do Site | JSON | Streaming (Kafka) | Alto | Contínua (real-time) | AWS / Cloud Infra | Equipe de Engenharia de Software |
| Base de Cadastro de Clientes | Tabelas SQL | Batch (diário) | Médio | 02:00 AM (diário) | Oracle Cloud | Departamento de CRM |
| Base de Produtos | Tabelas SQL | Batch (diário) | Médio | 02:00 AM (diário) | Oracle Cloud | P&D / Desenvolvimento de Produto |
| Redes Sociais (comentários) | JSON | Batch (diário) | Variável | 03:00 AM (diário) | APIs externas | Equipe de Marketing Digital |

Tabela 1 – Origem de Dados

Fonte -Elaborado pelo autor (2025)

3.2 Justificativas das Fontes de Dados

• Logs de Navegação: São fundamentais para entender o comportamento do usuário durante sua jornada de compra, ajudando a identificar o momento e o motivo do abandono do carrinho.

• Base de Cadastro de Clientes: Permite associar comportamentos de navegação com perfis de usuários, enriquecendo a análise de padrões de abandon

• Base de Produtos: Ajuda a entender se há correlação entre certos produtos e o abandono do carrinho (ex.: preço elevado, falta de estoque).

• Redes Sociais: Podem trazer feedbacks e menções que expliquem de maneira externa o comportamento dos consumidores, como reclamações de preço, usabilidade, ou prazos de entrega.

## 3.3 Detalhamento das fontes de dados

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tabela | Apelido | Descrição | Interessados | Dono da Informação | Retenção |
| T\_LOGS\_NAV | Logs | Dados de navegação dos usuários no site | TI, Marketing, Experiência do Cliente | Engenharia de Software | Retenção de 12 meses para análise comportamental |
| T\_CLIENTES | Clientes | Dados cadastrais dos usuários registrados | Vendas, CRM | Departamento de CRM | Backup diário completo, retenção de 2 anos |
| T\_PRODUTOS | Produtos | Dados sobre produtos ofertados | P&D, Vendas | Desenvolvimento de Produto | Produtos vendidos nos últimos 300 dias são mantidos |
| API\_SOCIAL\_FEEDBACK | Sociais | Comentários e menções públicas nas redes sociais | Marketing | Equipe de Marketing | Coleta diária, retenção máxima de 6 meses |

Tabela 2 - Dicionário de dados de colunas de tabelas

Fonte - Material da fase – Elaborado pelo autor

# 4 Arquitetura de Solução Big Data / Pipeline de Dados

## 4.1 Desenho da Arquitetura

## Diagrama Descrição gerada automaticamente

Figura 1 – Big Data Pipeline Architecture

Fonte – Elaborado pelo autor (2025)

## 4.1 Justificativa da Arquitetura

Escolhemos essa arquitetura por ela ser escalável, flexível e capaz de trabalhar tanto com dados em tempo real quanto com dados históricos.

O uso de Kafka e Spark garante a capacidade de lidar com grandes volumes de dados e gerar insights em tempo adequado para as necessidades de e-commerce.

## 4.3 Detalhamento da Arquitetura

• Camada de Captura (Data Source): Responsável por coletar eventos do site, atualizações de clientes e produtos, além de feedback social.

• Camada de Ingestão: Processa a entrada dos dados usando Kafka para eventos em tempo real e processos de ETL para bases tradicionais.

• Camada Data Lake:

• Raw: Dados originais armazenados para histórico.

• Trusted: Dados limpos e normalizados.

• Curated: Dados prontos para consumo analítico.

• Camada de Análise (Compute): Spark é usado para análises pesadas e machine learning; Kafka Streams para análises imediatas em real-time.

• Camada de Consumo: Dashboards para visualização de KPIs e alertas automáticos para áreas de vendas e marketing.

# GLOSSÁRIO

|  |  |
| --- | --- |
| **Termo** | Explicação. |
| **Big Data** | Conjunto de tecnologias e práticas voltadas ao tratamento de grandes volumes de dados com variedade e velocidade. |
| **Data Lake** | Repositório de dados brutos, estruturados e não estruturados, usado para armazenar e processar grandes quantidades de dados. |
| **Kafka** | Plataforma de streaming distribuído utilizada para ingestão e processamento de dados em tempo real. |
| **Apache Spark** | Ferramenta de processamento distribuído usada para análise de grandes volumes de dados em batch ou streaming. |
| **HDFS** | Hadoop Distributed File System - sistema de arquivos distribuído utilizado para armazenar grandes quantidades de dados. |
| **ETL** | Processo de Extração, Transformação e Carga de dados, usado para integração entre fontes e repositÃ³rios. |
| **Dashboards** | Painéis visuais que apresentam indicadores, gráficos e métricas de negÃ³cio em tempo real. |
| **Abandono de carrinho** | Quando o usuário adiciona produtos ao carrinho no e-commerce mas não finaliza a compra. |
| **Trusted / Curated** | Camadas do Data Lake onde os dados passam por validações (trusted) e são preparados para análises (curated). |
| **Machine Learning** | Campo da inteligência artificial voltado Ã  criação de modelos preditivos baseados em dados históricos. |