Logotipo

Descrição gerada automaticamente

**PLANEJAMENTO DO PROJETO**

**DATA DRIVEN**

02 de agosto de 2021

[1. Objetivo 2](#_Toc79617136)

[2. Estrutura Organizacional 2](#_Toc79617137)

[3. Definições e Siglas 3](#_Toc79617138)

[4. Escopo do Projeto 3](#_Toc79617139)

[5. Ferramentas 4](#_Toc79617140)

[6. Desafios e Riscos do Projeto 5](#_Toc79617141)

[7. Planejamento 5](#_Toc79617142)

[7.1. Cronograma 5](#_Toc79617143)

[7.2. Reunião de Kickoff 6](#_Toc79617144)

[7.3. Requisitos e Recursos 6](#_Toc79617145)

[7.4. Equipe 6](#_Toc79617146)

[7.5. Metodologia 6](#_Toc79617147)

[7.5.1. Documentação 6](#_Toc79617148)

[7.5.2. Fontes de dados 6](#_Toc79617149)

[8. Design 9](#_Toc79617150)

[8.1. Entrevistas de negócio 9](#_Toc79617152)

[8.2. Fontes de dados 9](#_Toc79617153)

[8.3. Modelo dimensional 9](#_Toc79617154)

[8.4. Dashboards 10](#_Toc79617155)

[9. Implementação 10](#_Toc79617156)

[9.1. Dados 10](#_Toc79617158)

[9.1.1. Ingestão 10](#_Toc79617162)

[9.1.2. Modelos 10](#_Toc79617163)

[9.1.3. Testes 10](#_Toc79617164)

[10. Deploy 11](#_Toc79617165)

[11. Referências 11](#_Toc79617166)

# Objetivo

Segundo a Pesquisa Perfil do Ciclista de 2018 (TRANSPORTE ATIVO, 2018), a maior parte das pessoas que usam a bicicleta como meio de locomoção no Brasil a utilizam praticamente todos os dias. De todos os entrevistados na pesquisa, 82% afirmaram usá-la em seus deslocamentos semanalmente.

Essa cultura é responsável por fazer do Brasil um país que produz mais bicicletas do que carros.

Em 2018, foram produzidos no país 3,8 milhões de unidades de acordo com a Associação Brasileira dos Fabricantes de Motocicletas, Ciclomotores, Motonetas, Bicicletas e Similares (Abraciclo). Esse desempenho coloca o Brasil atrás apenas da China no ranking mundial de produção.

De acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE), a indústria da bicicleta no país gerou mais de R$ 1 bilhão em valor de produção.

Desde o início da pandemia, a população vem procurando meios alternativos de transporte para evitar aglomerações em transportes coletivos, esse movimento impulsionou a fabricação e venda de bicicletas no país. Em janeiro de 2021, o Polo Industrial de Manaus produziu 56.981 unidades de bicicletas, aumento de 44,6% em relação ao mês de dezembro e 1% superior ao registrado em janeiro de 2020.

A Adventure Works é uma indústria de bicicletas em franco crescimento, alavancada principalmente pelo momento aquecido do mercado. Atualmente a companhia possui mais de 500 produtos distintos no seu portifólio, com uma carteira de 20 mil clientes e 31 mil pedidos.

Para aproveitar o momento aquecido do mercado, crescer ainda mais e se diferenciar da concorrência, a Adventure Works quer utilizar seus dados de forma estratégica, norteando suas decisões baseadas em dados.

# Estrutura Organizacional

O projeto foi idealizado e patrocinado pela Diretoria de Inovações e tem o apoio do CEO, que vê no uso de dados um diferencial estratégico em longo prazo. A Diretoria de TI é responsável por garantir o acesso aos dados.

Diagrama

Descrição gerada automaticamente

# Definições e Siglas

**Data Warehouse** – Um *data* *warehouse* centraliza e consolida grandes quantidades de dados de várias fontes. Seus recursos analíticos permitem que as organizações obtenham informações de negócios úteis de seus dados para melhorar a tomada de decisões.

**ELT** – O ELT é um processo de dados usado para replicar dados de uma fonte para um banco de dados de destino, sendo uma evolução do mais comumente conhecido ETL (extrair, transformar e carregar) pois torna o processo de replicação de dados muito menos complexo, uma vez que o passo de transformação é realizado após os dados estarem no destino.

**Github** – GitHub é uma plataforma de hospedagem de código-fonte e arquivos com controle de versão usando o Git.

**BI** – Processo orientado pela tecnologia para analisar dados e apresentar informações acionáveis para ajudar executivos, gerentes e outros usuários finais corporativos a tomar decisões de negócios bem informadas.

**Data Driven** – É um adjetivo que qualifica processos orientados por dados, ou seja, embasados na coleta e análise de informações.

**Opensource** – Software de código aberto é o software de computador com o seu código fonte disponibilizado e licenciado com uma licença de código aberto no qual o direito autoral fornece o direito de estudar, modificar e distribuir o software de graça para qualquer um e para qualquer finalidade.

**Dimensão** – Coleção de informações de referência sobre um evento mensurável.

**Fato** - Tabela dominante de um esquema de modelagem tipo estrela, em um modelo multidimensional, e tem como característica principal a presença de dados altamente redundantes para se obter um melhor desempenho.

**Stitch** – Ferramenta utilizada para movimentar dados dos sistemas de origens e centralizar em um determinado sistema de origem.

**DBT** – Ferramenta para transformar os dados e armazenar em determinado sistema de destino, utilizando um modelo dimensional.

**Pipeline** - Etapas envolvidas na movimentação de dados do sistema de origem para o sistema de destino. Essas etapas incluem copiar dados, transferi-los de um local para a nuvem e organizá-los ou combiná-los com outras fontes de dados.

# Escopo do Projeto

Este projeto de *Data Driven* se refere aos processos organizacionais orientados a dados, habilitando a Adventure Works basear a tomada de decisão e o planejamento estratégico na coleta e na análise de informações, e não em intuições ou simples experiências.

O projeto será executado em 90 dias e contempla os seguintes itens:

* **Diagrama Conceitual e lógico do *Data Warehouse*:** criação das tabelas de fatos e dimensões necessárias para responder as perguntas de negócio apresentadas abaixo. Apresentação resumida de quais tabelas fonte são necessárias para criar as tabelas fatos e dimensões;
* **Configuração do *Data Warehouse*:** criação de DW utilizando *cloud*;
* ***Pipeline* de EL (*Extract, load*):** criação do pipeline para popular o DW com os dados do Banco de Dados da Adventure Works;
* **Transformação de dados – T (*transform*):** transformação dos dados brutos para o modelo de dados criado no DW. Documentação e testes de dados armazenados em repositório do *Github*;
* **Painel de *Bussines Analitycs* (BI):** criação de painel para responder as seguintes perguntas de negócio:
  + Qual o número de pedidos, quantidade comprada, valor total negociado por produto, tipo de cartão, motivo de venda, data de venda, cliente, status, cidade, estado e país?
  + Quais os produtos com maior ticket médio por mês, ano, cidade, estado e país?
  + Quais os 10 melhores clientes por valor total negociado filtrado por produto, tipo de cartão, motivo de venda, data de venda, status, cidade, estado e país?
  + Quais as 5 melhores cidades em valor total negociado por produto, tipo de cartão, motivo de venda, data de venda, cliente, status, cidade, estado e país?
  + Qual o número de pedidos, quantidade comprada, valor total negociado por mês e ano?
  + Qual produto tem a maior quantidade comprada para o motivo “*Promotion*”?
* **Apresentação do projeto Data Driven:** gravação de apresentação de todas as etapas do projeto (DW, EL, transformação e B) compartilhado no *google drive*.

Diagrama

Descrição gerada automaticamente

# Ferramentas

Para este projeto as ferramentas utilizadas farão a coleta dos dados de diversas fontes, tanto internas quanto externas. Elas cruzarão informações de modo a oferecer um panorama mais claro do negócio, clientes e produtos, além da própria organização para que os executivos possam agir estrategicamente.

|  |  |  |
| --- | --- | --- |
| **FERRAMENTA** | **CUSTO (R$)** | **APLICAÇÃO** |
| **STITCH** | USD 1.000/yr  5 linhas/mês | O Stitch será utilizado para executar o pipeline de dados, extraindo dos sistemas e fontes internas e carregando para o DW. |
| **DBT** | Free - $0,00 | O DBT será utilizado para transformar os dados do DW em *data marts* para consultas analíticas. |
| **GCP – BIG QUERY** | - |  |
| **GOOGLE DATA STUDIO** | - | O Google Data Studio será utilizado como ferramenta de construção dos painéis analíticos. |

# Desafios e Riscos do Projeto

Durante a execução deste projeto Data Driven, alguns desafios e riscos deverão ser considerados para não comprometer o escopo aprovado pela Adventure Works.

* Construir um planejamento de implantação alinhado há um cronograma que atenda as expectativas dos executivos;
* Utilizar ferramentas *Opensource* para minimizar o custo de implantação e elevar o *Return over Investment* – ROI do projeto;
* Criar modelos de dados que entreguem valor e respondam as perguntas de negócios que os executivos da Adventure Works levantaram durante a reunião de *kickoff*;
* Identificar as fontes de dados que armazenam os dados de forma integrada em seus processos e operações, as fontes de dados não integradas (excel, pdf e arquivos); e
* Processar o volume de dados dentro de uma velocidade, variedade, veracidade e valor aceitável para que os executivos tomem suas decisões.

# Planejamento

## Cronograma

|  |  |
| --- | --- |
| **DATA** | **EVENTO** |
| **13/07/2021** | **REUNIÃO DE KICKOFF E PLANEJAMENTO** |
| **15/07/2021** | **ENTREVISTAS COM AS ÁREAS DE NEGÓCIO** |
| **19/07/2021** | **LEVANTAMENTO DE REQUISITOS PARA ANÁLISE DOS RELATÓRIOS DE SISTEMAS INTERNOS** |
| **21/07/2021** | **DEFINIÇÃO DOS CONCEITOS E REGRAS DE NEGÓCIO SOBRE AS MÉTRICAS, DADOS DA EMPRESA E TESTES QUALITATIVOS DOS DADOS** |
| **22/07/2021** | **CRIAÇÃO E DEFINIÇÃO DO MODELO CONCEITUAL E DIMENSIONAL DO DW** |
| **23/07/2021** | **APROVAÇÃO DO MODELO CONCENTUAL E DIMENSIONAL DO DW** |
| **26/07/2021** | **CONFIGURAÇÃO E CRIAÇÃO DO PROCESSO DE EXTRAÇÃO E CARREGAMENTO DE DADOS** |
| **28/07/2021** | **CONFIGURAÇÃO E TRANSFORMAÇÃO DOS DADOS PARA ARMAZENAR NO DW** |
| **02/08/2021** | **TESTES DO PIPELINE DE DADOS STITCH/DBT** |
| **04/08/2021** | **CRIAÇÃO E ENTREGA DA DASHBOARD** |
| **06/08/2021** | **GO LIVE** |
| **11/08/2021** | **ENTREGA E FINALIZAÇÃO DO PROJETO** |

## Reunião de Kickoff

Foi realizado a reunião de *kickoff* para entendermos o momento atual da companhia e suas expectativas. Durante a reunião de *kickoff* a diretoria da Adventure Works listou perguntas às quais deseja que o projeto responda através dos cruzamentos de dados.

Durante a reunião de *Kickoff* foi possível diagnosticar alguns sistemas que a Adventure Works utiliza, e que geram dados relevantes para o negócio. Esses sistemas em algum momento farão parte da infraestrutura de dados.

|  |  |
| --- | --- |
| **FUNÇÃO** | **SISTEMA** |
| **ERP** | **SAP** |
| **CRM** | **SALESFORCE** |
| **WEB ANALYTICS** | **GOOGLE ANALYTICS** |
| **SITE** | **WORDPRESS** |

## Requisitos e **Recursos**

Para iniciar o projeto e obter resultados rápidos, a opção foi iniciar pela área de vendas (*sales*), mas algumas tabelas de outras áreas podem ser necessárias para conseguir as informações desejadas.

O acesso ao Banco de Dados foi autorizado pela Diretoria de TI que indicou um analista o qual podemos recorrer diretamente.

## Equipe

A Diretoria de Inovação é o patrocinador do projeto *Data Driven*, além desta Diretoria o Diretor de TI é responsável por subsidiar os acessos e recursos tecnológicos necessário para implantação do projeto, desta maneira ele delegou ao Analista Gabriel Santos está tarefa.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| - | NOME | TELEFONE | E-MAIL | ASSUNTO |
| 1º | GABRIEL SANTOS | 41 98843-4586 | joao.santos@adventureworks.com | BANCO DE DADOS/DADOS |
| 2º | NILSON RAMOS | 41 98889-0886 | nilson.ramos@adventureworks.com | TI/INFRAESTRUTURA/ACESSOS |
| 3º | JOÃO MULLER | 41 99802-4190 | joao.muller@adventureworks.com | NEGÓCIO/FINANCEIRO/PROJETO |

## Metodologia

### Documentação

A documentação deste projeto pode ser encontrada na conta de acesso a ferramenta DBT.

### Fontes de dados

A Adventure Works possui um banco de dados transacional (PostgreSQL) que armazena os dados de suas diferentes áreas. Esses dados estão distribuídos em 68 tabelas divididas em 5 *schemas*: *humanresources* (recursos humanos), *sales* (vendas), *production* (produção), *purchasing* (compras) e person (pessoa).

Gráfico, Esquemático

Descrição gerada automaticamente

O Analista de TI da Adventure Works disponibilizou um acesso para você realizar a extração de dados para o projeto:

Host: 35.239.223.162

Port: 5432

User: (cofre de senhas)

Password: (cofre de senhas)

Database: Adventureworks

Para facilitar o entendimento dos dados, a Adventure Works disponibilizou um dicionário de dados necessários para seu projeto com algumas definições que eles utilizam internamente. Como atualmente se utiliza relatórios exportados pelo sistema, podem ser necessárias outras tabelas para obter as informações, e cabe a você encontrar esses relacionamentos.

|  |  |  |
| --- | --- | --- |
| **DIMENSÃO** | **COLUNA NO BD** | **TABELAS** |
| **CLIENTE** | **PERSON.NAME** | **CUSTOMER/PERSON** |
| **MOTIVO DA VENDA** | **SALESREASON.NAME** | **SALES REASON** |
| **VENDEDOR** | **PERSON.NAME** | **SALESPERSON/EMPLOYEE/PERSON** |
| **ESTADO** | **STATEPROVINCE.NAME** | **STATE PROVINCE** |
| **CIDADE** | **ADDRESS.CITY** | **ADDRESS** |
| **PAÍS** | **COUNTRYREGION.NAME** | **COUNTRY REGION** |
| **DATA DE VENDA** | **SALESORDERHEADER.ORDERDATE** | **SALES ORDER HEADER** |
| **PRODUTO** | **PRODUCT.NAME** | **SALES ORDER DETAIL/PRODUCT** |

|  |  |  |
| --- | --- | --- |
| **MÉTRICA** | **CÁLCULO** | **TABELA PRIMÁRIA** |
| **NÚMERO DE PEDIDOS** | **CONTAGEM DISTINTA DE SALESORDERID** | **SALES ORDER HEADER** |
| **QUANTIDADE COMPRADA** | **SOMA DE ORDERQTY** | **SALES ORDER DETAIL** |
| **VALOR TOTAL NEGOCIADO** | **SOMA DE UNITPRICE** | **SALES ORDER DETAIL** |
| **TICKET MÉDIO** | **MÉDIA DE UNITPRICE** | **SALES ORDER DETAIL** |

# Design



## Entrevistas de negócio

Foi entrevistado as áreas de negócio para obter o entendimento e consenso sobre as métricas e informações utilizadas dentro da Adventure Works. O processo de entrevista foi fundamental para chegarmos ao entendimento comum sobre os dados e a partir desta definição foi possível iniciar a análise de requisitos para se obter a origem da fonte de dados para se trabalhar.

## Fontes de dados

Durante as entrevistas para levantamento do escopo foi possível identificar os relatórios utilizados e com base nos relatórios e Diagrama de Entidade e Relacionamento (DER), foi identificado a entidades que armazenam os dados dos relatórios.

Considerando que não serão utilizadas todas as entidades da fonte de dados, foi optado por separar e aplicar o processo de *extract and load* apenas para essas entidades:

|  |  |
| --- | --- |
| **SCHEMA** | **TABLE** |
| **Sales** | **salesorderheader** |
| **sales** | **customer** |
| **person** | **person** |
| **person** | **address** |
| **person** | **stateprovince** |
| **person** | **countryregion** |
| **sales** | **salesorderdetail** |
| **production** | **product** |
| **sales** | **salesorderheadersalesreason** |
| **sales** | **salesreason** |

## **Modelo** dimensional

Entendendo as fontes de dados e escopo de entrega deste projeto, foi possível construir um modelo dimensional (estrela) com as seguintes dimensões: *customer, product*, entrega (*shipp*), *creditcard*; e fato *salesorderdetail*.

Diagrama

Descrição gerada automaticamente

## Dashboards

A proposta inicial deste projeto era utilizar o Power BI como ferramenta de visualização. Como houve dificuldade de conectar o PBI com *Google BigQuery*, optou-se pelo uso Google Data Studio.

# Implementação



## Dados



### Ingestão

Os dados deste projeto serão carregados do ERP, cujo dados do Banco de Dados foi disponibilizado para área de Tecnologia da Adventure Works.

Após o entendimento do banco de dados relacional e identificado as entidades necessárias para o processo de ingestão, foi utilizado o Stitch para extrair e carregar diariamente os dados para dentro do DW.

### Testes

Durante o processo de entrevista alguns controles e testes dos dados foram sugeridos pelas áreas de negócios e foram aplicadas no processo de transformação de dados para garantir a integridade e qualidade dos dados utilizados nas análises internas.

1. A quantidade de pedidos entre as datas 07/06/2011 e 07/07/2014 deve ser igual a 64.477.

2. A quantidade total de produtos vendidos anterior a data de modificação 08/02/2014 deve ser igual a 274.238.

3. A soma de todos os produtos vendidos anterior a data de modificação 08/02/2014 deve ser igual 56410918.7987

Além das regras de negócio acima, alguns dados não podem ser nulos, enquanto outro devem ser únicos.

    tables:

      - name: address

        columns: addressid

     - name: countryregion

        columns: countryregioncode

     - name: creditcard

        columns: creditcardid

     - name: customer

        columns: customerid

      - name: person

        columns: businessentityid

      - name: product

        columns: productid

      - name: salesorderdetail

        columns: salesorderdetailid

      - name: salesorderheader

        columns: salesorderid

      - name: salesreason

        columns: salesreasonid

      - name: stateprovince

        columns: stateprovinceid

            tests:

              - unique

              - not\_null

      - name: salesorderheadersalesreason

        columns: salesorderid, salesreasonid

            tests:

              - not\_null

# Deploy

O código deste projeto está disponível no *Github* e pode ser encontrado no repositório <https://github.com/lenowds/adventureworks-indicium.git>.

# Referências

<https://www1.folha.uol.com.br/mercado/2021/02/governo-bolsonaro-reduz-tarifa-de-importacao-de-bicicletas.shtml>

<https://www.mobilize.org.br/midias/pesquisas/revista-fabricacao-e-montagem-de-bicicletas.pdf>