

13 – Introdução

- No atual cenário tecnológico, a grande maioria das empresas já informatizaram praticamente todas as suas operações (vendas, contabilidade, financeiro, etc.) por meio de **sistemas integrados de gestão empresarial (ERP)**.
- Os sistemas ERP possuem módulos específicos para cada setor da empresa, possibilitando que todos os departamentos organizacionais estejam integrados.
- Por exemplo: Quando uma venda é realizada, o ERP automaticamente faz a baixa no estoque, gera recebimento no financeiro, estabelece o controle contábil, efetiva a comissão dos vendedores, entre muitas outras operações relacionadas a venda.
- Considerando que todas estas informações inseridas no ERP ficam armazenadas em um banco de dados, torna-se possível estabelecer uma série de processos de **análise de dados**, objetivando auxiliar os gestores na tomada de decisões.

13 – Introdução

- A imagem abaixo mostra de forma resumida como acontece o fluxo de dados. Basicamente, os usuários inserem os dados no sistema, para que ele trate e envie estas informações para armazenamento definitivo no banco de dados.

Cadastro de Clientes

Preencha os campos e clique em Gravar Dados

Nome:

Endereço:

Bairro:

Estado:

Telefone:

Celular:

Email:

Gravar Cadastro Novo Cadastro Ver Cadastros



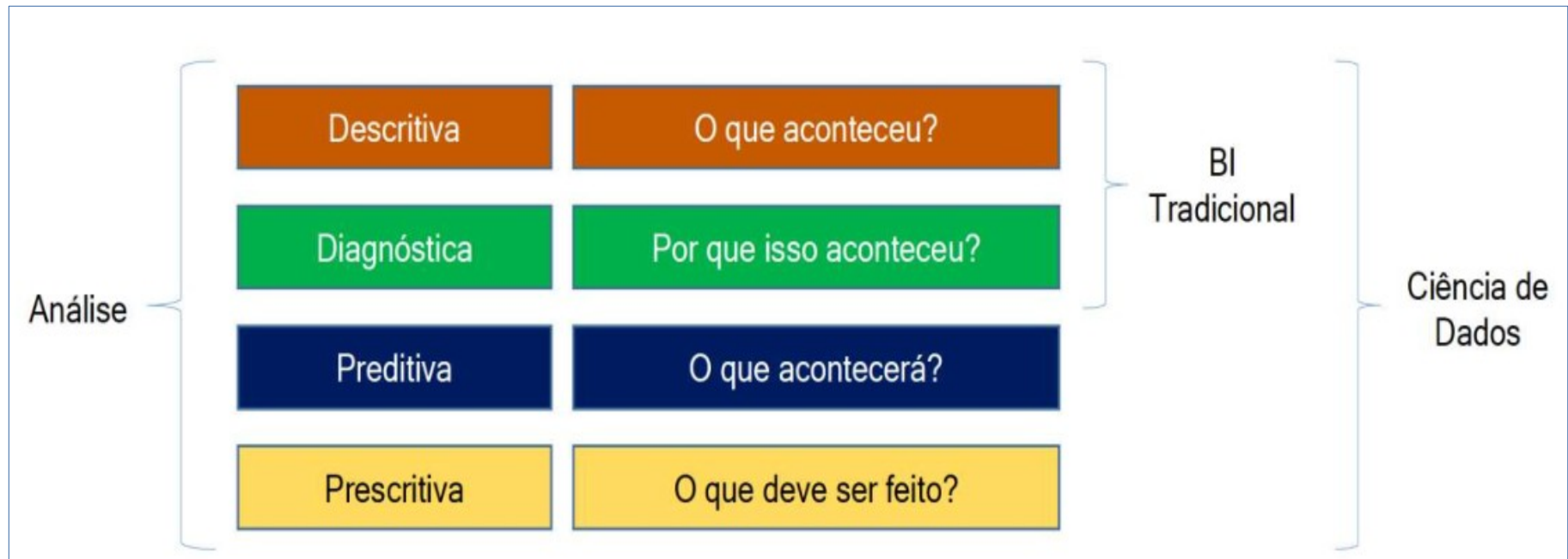
id_cliente	nome	endereço	bairro	estado	telefone	celular	e-mail
1	Paulo Soares	Rua Central, 42	Zona 7	PR	999623432	999877665	paulo@ph.com

13 – Introdução

- A partir do momento que a empresa possui um ERP armazenado todas as suas informações em um banco de dados, torna-se possível cruzar os registros dos diferentes departamentos. Com isso, uma análise de dados mais robusta pode ser feita.
- É neste ponto que atua a ciência de dados (data science), uma área de grande destaque dentro do cenário de TI atualmente.
- De forma resumida, o objetivo da ciência de dados é analisar os dados, aplicando uma série de técnicas estatísticas, e gerar conhecimento (muitas vezes gráficos ou dashboards) para auxiliar os gestores.
- Um bom cientista de dados deve ser capaz de compactar todo o conhecimento extraído no processo de análise de dados em **dashboards de fácil compreensão**.

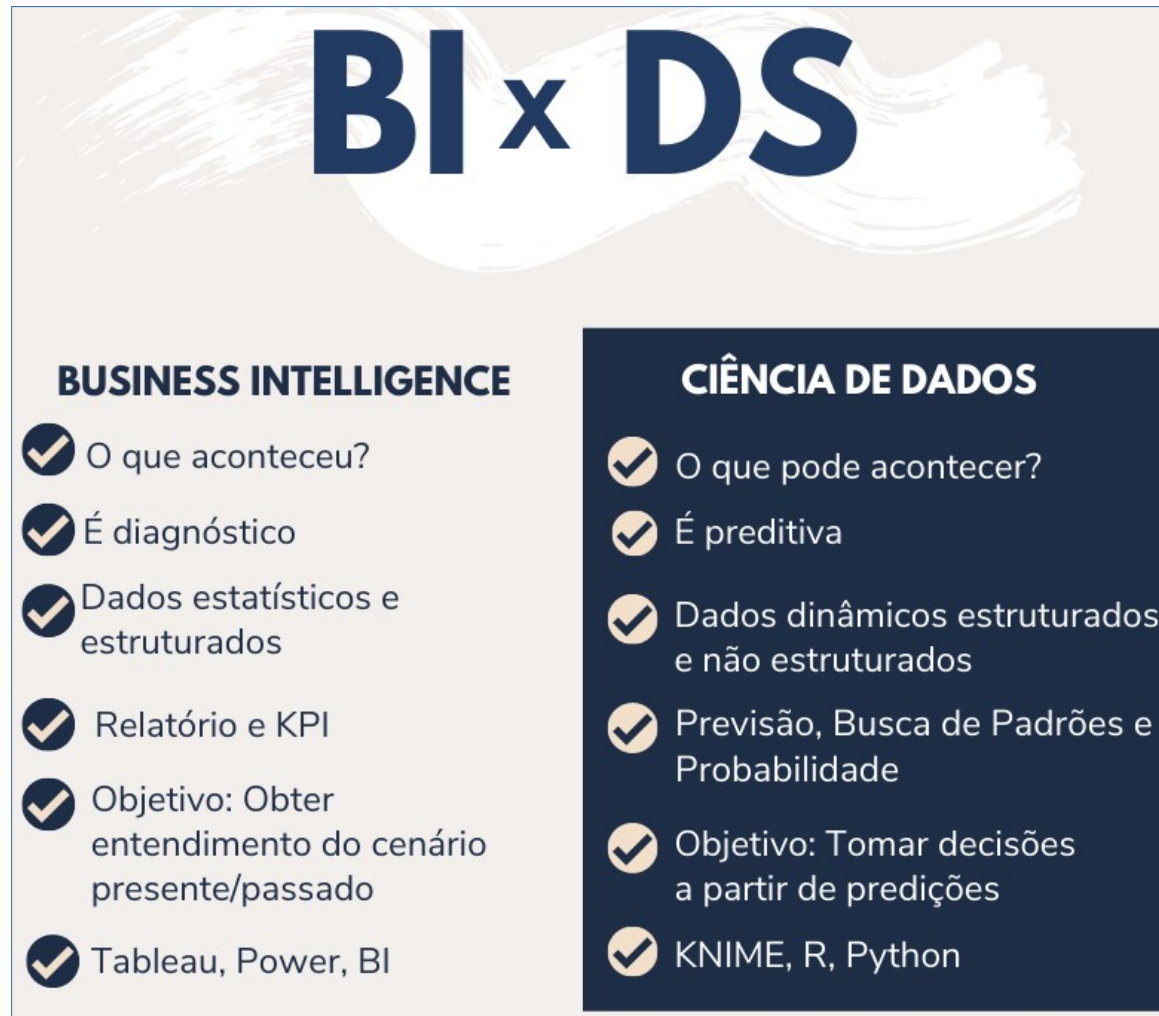
13.1 – Níveis da Ciência de Dados

- A ciência de dados é uma grande área, que pode ser dividida em níveis distintos. O profissional que atua fazendo análises **descritiva** e **diagnóstica**, é denominado **analista de dados**.
- Já o profissional que atua em processos de análise **preditiva** e **prescritiva**, muitas vezes trabalhando com algoritmos de machine learning, é denominado **cientista de dados**.



13.1 – Níveis da Ciência de Dados

- A imagem abaixo mostra as principais diferenças entre análise de dados (business intelligence) e ciência de dados:



13.1 – Níveis da Ciência de Dados

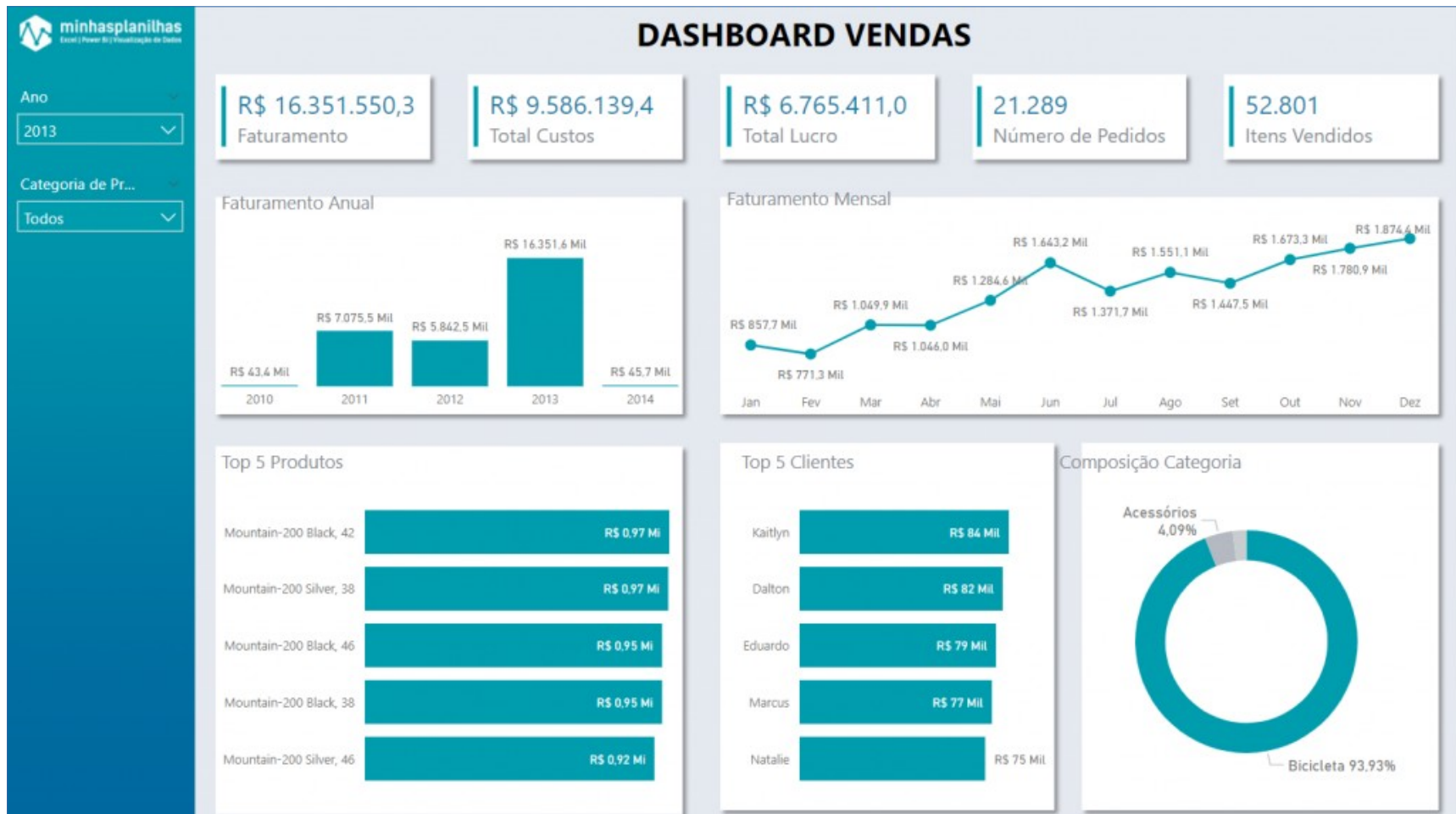
- O analista de dados utiliza amplamente a linguagem SQL, além de software específicos de business intelligence, tais como Power BI, Tableau, Excel, entre muitos outros. Seu objetivo é responder questões como:

1. Produtos mais vendidos durante o mês
2. Vendedores que mais venderam no ano
3. Produtos com estoque baixo
4. Produtos com prazo de validade finalizando



13.1 – Níveis da Ciência de Dados

- Ao final do processo de análise de dados, o objetivo é gerar **gráficos** e **dashboards interativos**, como da imagem abaixo, visando auxiliar os gestores a entender o atual cenário da empresa, bem como tomar as melhores decisões estratégicas.



13.1 – Níveis da Ciência de Dados

- Já o cientista de dados tem como objetivo central estabelecer padrões, e estimar **movimentos futuros (análise preditiva)** a partir de técnicas de machine learning. Para isso, trabalha com grande bases de dados (Big Data), muitas vezes extremamente heterogêneas.

- O cientista de dados costuma trabalhar com tecnologias como **Python, R, Spark**, entre muitas outras. Seu objetivo é responder questões como:

1. Quais produtos são vendidos de forma conjunta (cliente que compra o produto X, também costuma levar o produto Y e Z)

2. Quais cidades terão o maior número de clientes nos próximos anos

3. Qua



alunos que



ciplina de banco de da



13.2 – Aplicações da Ciência de Dados

➤ Saúde:

1. Planos de saúde tem utilizado ciência de dados para estabelecer o risco de cada cliente, para determinar o valor da mensalidade.
2. Por meio de sensores implantados, algoritmos determinam padrões de risco no estado de saúde dos pacientes, e podem alertar a equipe médica.
3. Robô Laura: Faz conexão com prontuários e sensores, auxiliando diagnósticos e alertando em caso de piora do paciente. Reduziu o número de mortes em 25%, reduziu tempo de internação em 10% e gerou economia de mais de 5 milhões em alguns hospitais. Auxiliou no combate a pandemia, monitorando pacientes infectados e alertando quando deve procurar o hospital.



13.2 – Aplicações da Ciência de Dados

➤ **Supermercado:**

1. Determinar produtos que são vendidos conjuntamente.
2. Determinar padrões de compra dos clientes, possibilitando ofertas exclusivas, e proporcionando aumento nas vendas.
3. Previsão de vendas dos produtos, auxiliando no gerenciamento de estoques (impedindo falta e excesso de produtos).



13.2 – Aplicações da Ciência de Dados

➤ **Instituições Bancárias:**

1. Analise de riscos, visando estabelecer os limites de crédito que podem ser cedidos para cada um dos clientes que abre conta no banco.
2. Estabelecer o perfil de cada cliente, com o objetivo de oferecer os produtos que mais se adequam as suas características.
3. Objetivando aumento na segurança, movimentações na conta fora do padrão geram um alerta para bloqueio.



13.2 – Aplicações da Ciência de Dados

➤ **Marketing:**

1. Direcionar propagandas para o público que tenham as características mais prováveis de consumirem o produto anunciado.
2. Sites de venda pela internet utilizam algoritmos de machine learning para prever vendas futuras e produtos associados, direcionando as buscas e os anuncios. A Amazon foi pioneira neste segmento.



13.2 – Aplicações da Ciência de Dados

➤ **Agricultura:**

1. Utilizando drones com cameras de alta resolução, é possível tirar fotos da plantação, que serão analisadas pelos algoritmos de machine learning. Após o processo de análise, é possível determinar a situação da lavoura, além de detectar doenças e pragas.



13.2 – Aplicações da Ciência de Dados

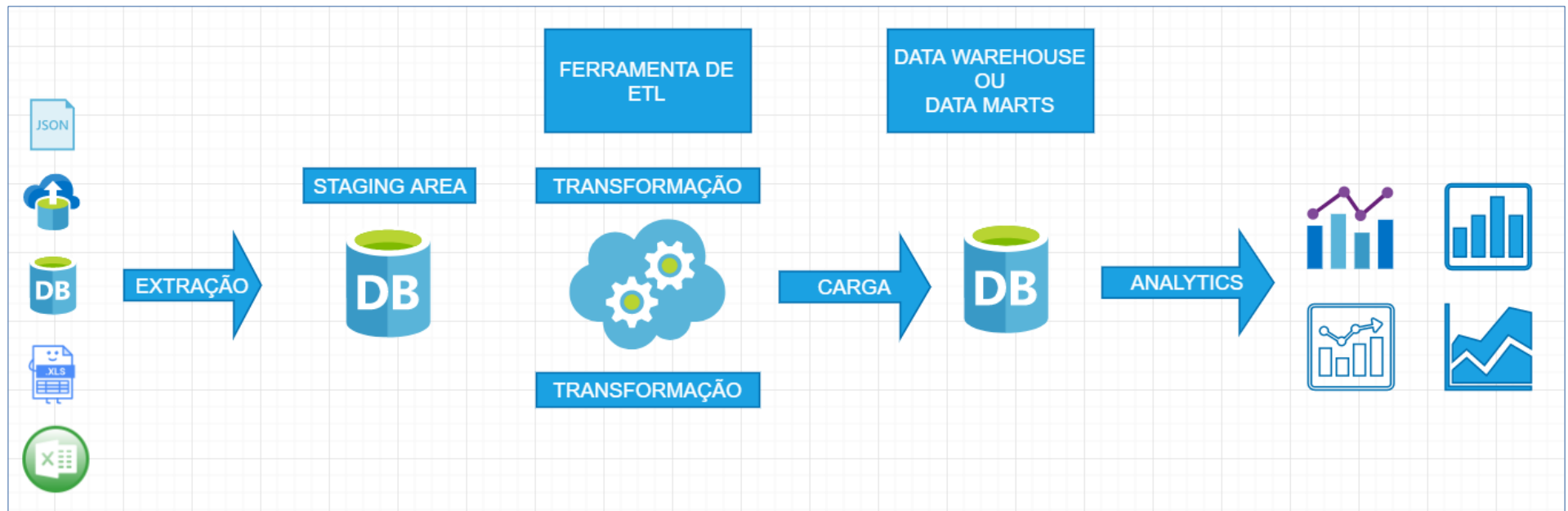
➤ **Mercado Financeiro:**

1. Os chamados robôs de investimento utilizam técnicas de machine learning para determinar os melhores investimentos. Baseados em milhares de parâmetros de entrada, eles estabelecem padrões de subida e descida de determinados segmentos de ações, auxiliando os investidores.



13.3 – Etapas da Análise de Dados

- A imagem abaixo mostra as etapas da análise de BI (ou da análise de dados):



13.3 – Etapas da Análise de Dados

➤ Vamos a uma breve descrição das cinco etapas do processo de BI:

1. **Fontes de dados:** Este processo consiste em capturar os dados que serão analisados. Pode ocorrer da fonte dos dados serem heterogêneas (bancos NoSQL, relacionais, arquivos, sensores, etc.). Neste caso, processo de captura e tratamento tende a ser mais complexo.

2. **Staging Area:** É um banco de dados intermediário que receberá os arquivos do banco de dados de origem (que contém os dados brutos, gerados pelas transações de negócio). Esta etapa é importante, pois as consultas iniciais para análise, que seriam feitas no banco de dados original, poderão ser feitas na Staging Area, evitando a concorrência de acessos.

3. **ETL (Extrair, Transformar e Carregar):** Consiste em fazer extração dos dados de uma ou várias fontes e unificá-las. Em seguida, esses dados podem ter várias transformações, como: Alterar os campos em novos formatos, criar campos calculados, tratar nulos, fazer concatenações e limpar toda a sujeira ou dados incorretos proveniente da fonte original. E por fim, fazer a carga dos dados limpos em

13.3 – Etapas da Análise de Dados

4. **Data Warehouse (DW):** É um banco de dados dimensional projetado especificamente para armazenar grandes volumes de dados e ter performance nas consultas. O Data Mart é um subconjunto de um Data Warehouse, podendo ser agrupado por departamentos – Data Mart Financeiro, Data Mart Vendas, e assim por diante. Ao final do processo de ETL, os dados devem ser carregados para o Data Warehouse.

5. **Análise de Dados:** Após os dados estarem devidamente tratados e armazenados no DW, eles estão prontos para serem analisados. Caberá ao analista gerar gráficos e dashboards de forma que possam facilitar o entendimento das informações presentes no banco de dados. Em complemento a este processo, um cientista de dados poderá aplicar algoritmos de machine learning para fazer análises preditivas e prescritivas.

13.3 – Etapas da Análise de Dados

- Na imagem abaixo temos um exemplo de dados que precisam ser tratados no processo de ETL:

Projeto 4 - Editor do Power Query

Página Inicial Transformar Adicionar Coluna Exibição Ferramentas Ajuda

Nova Fonte Recentes Inserir Dados Configurações da fonte de dados Gerenciar Parâmetros Atualizar Visualização Editor Avançado Gerenciar Consultas Gerenciar Colunas Reduzir Linhas Classificar Dividir Coluna Agrupar por Tipo de Dados: Texto Usar a Primeira Linha como Cabeçalho Substituir Valores

ltas [1] $\times$ $\checkmark$ f_x = Table.ReplaceValue("#Tipo Alterado2","07/03/2017","2017-02-39",Replacer.ReplaceText,{"Data"})

	AB_C Data	AB_C Produto	AB_C Serial number	AB_C Valor de Venda	1 ² 3 Preço Custo	1 ² 3 Duração V
1	19/03/2017	AX101	GF54309	6871		3436
2	02/03/2017	AX101	GF54381	6871		3779
3	28/03/2017	BX101	GF54527	3006		1353
4	03/03/2017	AX101	GF54695	6871		3573
5	19/03/2017	BX102	GF54484	5357		2839
6	05/03/2017	AX103	GF54240	535		2515
7	06/03/2017	DX101	GF54319	1762		916
8	27/03/2017	AX102	GF54236	5128		2718
9	2017-02-39	AX102	GF54473	5128		2564
10	25/03/2017	AX103	GF54256	535		2515
11	27/03/2017	CX101	GF54609	453		2175
12	29/03/2017	DX103	GF54556	Fail		2339
13	19/03/2017	BX102	GF54434	5357		2625
14	27/03/2017	BX102	GF54418	5357		2625
15	23/03/2017	AX103	GF54594	535		2568
16	2017-02-39	DX103	GF54315	Fail		1913
17	14/03/2017	CX101	GF54568	453		2129
18	27/03/2017	AX101	GF54354	6871		3161
19	13/03/2017	DX102	GF54634	6246		3061
20	2017-02-39	BX103	GF54492	6893		3584
21	31/03/2017	AX103	GF54314	535		2782
22	19/03/2017	DX102	GF54527	6246		3061
23	24/03/2017	AX102	GF54245	5128		2462
24						

AS, 100 LINHAS Criação de perfil de coluna com base nas primeiras 1000 linhas

13.4 – Conceito de Big Data

- Segundo a Oracle, **big data** pode ser definido como um conjunto maior e mais complexo de dados, especialmente de novas fontes de dados (dados de sensores, tweets, dados de sistemas, vídeos, sons, etc.).
- Esses conjuntos de dados são tão volumosos que o software tradicional de processamento de dados simplesmente não consegue gerenciá-los. No entanto, esses grandes volumes de dados podem ser usados para resolver problemas de negócios que você não conseguiria resolver antes.
- Com o imenso volume de dados presente no big data, tornou-se possível a utilização de técnicas de análise dados, inteligência artificial e machine learning, para resolver problemas complexos e gerar informações que antes não era possível.
- Este conceito de aplicar técnicas analíticas (utilizando IA, machine learning, entre outras) em big data, é conhecido como **Big Data Analytics**.

13.4 – Conceito de Big Data

- Segundo a Oracle, big data são dados com maior variedade que chegam em volumes crescentes e com velocidade cada vez maior. Isso também é conhecido como os **três Vs** do big data, como mostra a imagem abaixo:

Volume	A quantidade de dados importa. Com o big data, você terá que processar grandes volumes de dados não estruturados de baixa densidade. Podem ser dados de valor desconhecido, como feeds de dados do Twitter, fluxos de cliques em uma página web ou em um aplicativo para dispositivos móveis, ou ainda um equipamento habilitado para sensores. Para algumas empresas, isso pode utilizar dezenas de terabytes de dados. Para outras, podem ser centenas de petabytes.
Velocidade	Velocidade é a taxa mais rápida na qual os dados são recebidos e talvez administrados. Normalmente, a velocidade mais alta dos dados é transmitida diretamente para a memória, em vez de ser gravada no disco. Alguns produtos inteligentes habilitados para internet operam em tempo real ou quase em tempo real e exigem avaliação e ação em tempo real.
Variedade	Variedade refere-se aos vários tipos de dados disponíveis. Tipos de dados tradicionais foram estruturados e se adequam perfeitamente a um banco de dados relacional . Com o aumento de big data, os dados vêm em novos tipos de dados não estruturados. Tipos de dados não estruturados e semiestruturados, como texto, áudio e vídeo, exigem um pré-processamento adicional para obter significado e dar suporte a metadados.

13.4 – Conceito de Big Data

- Nos últimos anos, mais **dois Vs** surgiram: **valor** e **veracidade**. Via de regra os dados possuem valor, mas ele nem sempre está evidente, é necessário descobri-lo para que seja útil.
- Tão importante quanto o valor, é a veracidade dos dados, determinar o quanto eles são confiáveis é fundamental no processo de análise. Pior do que não ter dados, é ter dados inconsistentes, que podem induzir ao erro.
- Atualmente, as empresas que possuem dados confiáveis, e possuem ferramentas para analisá-los, possui grande vantagem sobre seus concorrentes. As imagens abaixo mostram alguns cases de sucesso de instituições que utilizam Big Data Analytics:



13.5 – Conceito de Inteligência Artificial

O que é Inteligência Artificial (IA)?

- A Inteligência Artificial (IA) é um campo da computação que busca criar sistemas capazes de realizar tarefas que normalmente exigiriam inteligência humana
- Isso inclui reconhecer padrões, tomar decisões, entender linguagem natural e até mesmo aprender com a experiência (aprendizagem por reforço).
- É importante reforçar que para uma IA funcionar corretamente, é fundamental que ela seja treinada com a maior quantidade de dados possíveis.
- E esses dados precisam estar equalizados e limpos (sem ruídos ou valores inconsistentes). Uma vez que a IA é treinada com dados imprecisos, implicará em resultados ineficientes.

13.5 – Conceito de Inteligência Artificial

Principais subáreas da IA

- **Aprendizado de Máquina (Machine Learning – ML):** Ensina máquinas a aprender a partir de dados sem serem programadas explicitamente. Exemplo: Algoritmos que recomendam filmes na Netflix com base no que você já assistiu.
- **Redes Neurais Artificiais e Deep Learning:** Inspiradas no funcionamento do cérebro humano, essas redes são usadas para resolver problemas complexos. Exemplo: Reconhecimento facial em celulares.
- **Processamento de Linguagem Natural (NLP - Natural Language Processing):** Permite que máquinas compreendam e gerem linguagem humana. Exemplo: Assistentes virtuais como Alexa e ChatGPT.
- **Visão Computacional:** Ensina máquinas a interpretar imagens e vídeos. Exemplo: Sistemas de reconhecimento de placas de carro em pedágios automáticos.
- **Sistemas Especialistas:** São programas que imitam a tomada de decisão de especialistas humanos em áreas específicas. Exemplo: Softwares médicos que ajudam no diagnóstico de doenças.
- **Robótica e Automação:** Aplica IA no controle de robôs e sistemas automatizados. Exemplo: Carros autônomos da Tesla. IA
- **IA Generativa:** Cria novos conteúdos, como imagens, textos, músicas e vídeos. Exemplo: Geradores de imagens como o DALL·E.