1 – Introdução

- A grande maioria das empresas já informatizaram praticamente todas as suas operações por meio de **sistemas integrados de gestão empresarial (ERP)**.

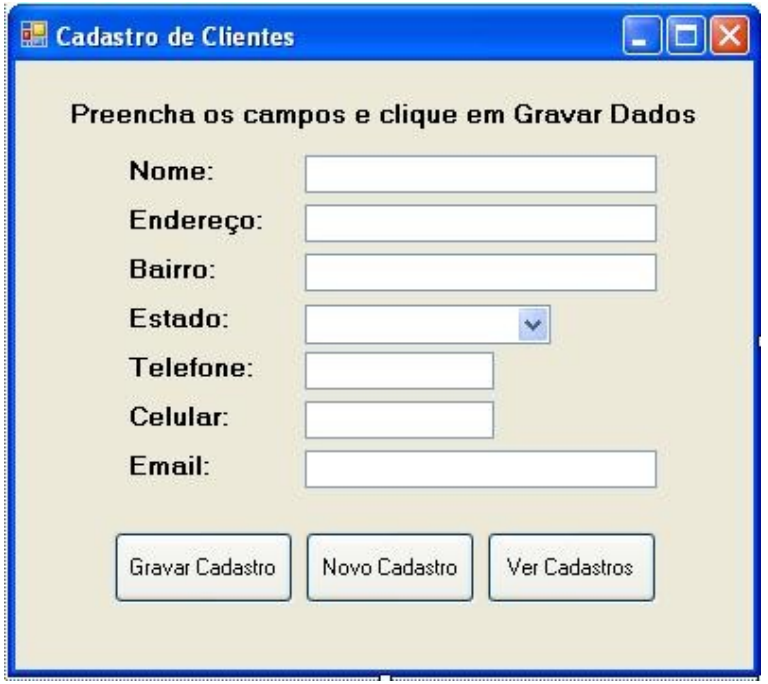
- Os sistemas ERP possuem módulos específicos para cada setor da empresa, possibilitando que todos os departamentos organizacionais estejam integrados.

- 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.

1 – Introdução

- A imagem abaixo mostra de forma resumida como acontece o fluxo de dados em um sistema:



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

1 – 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.

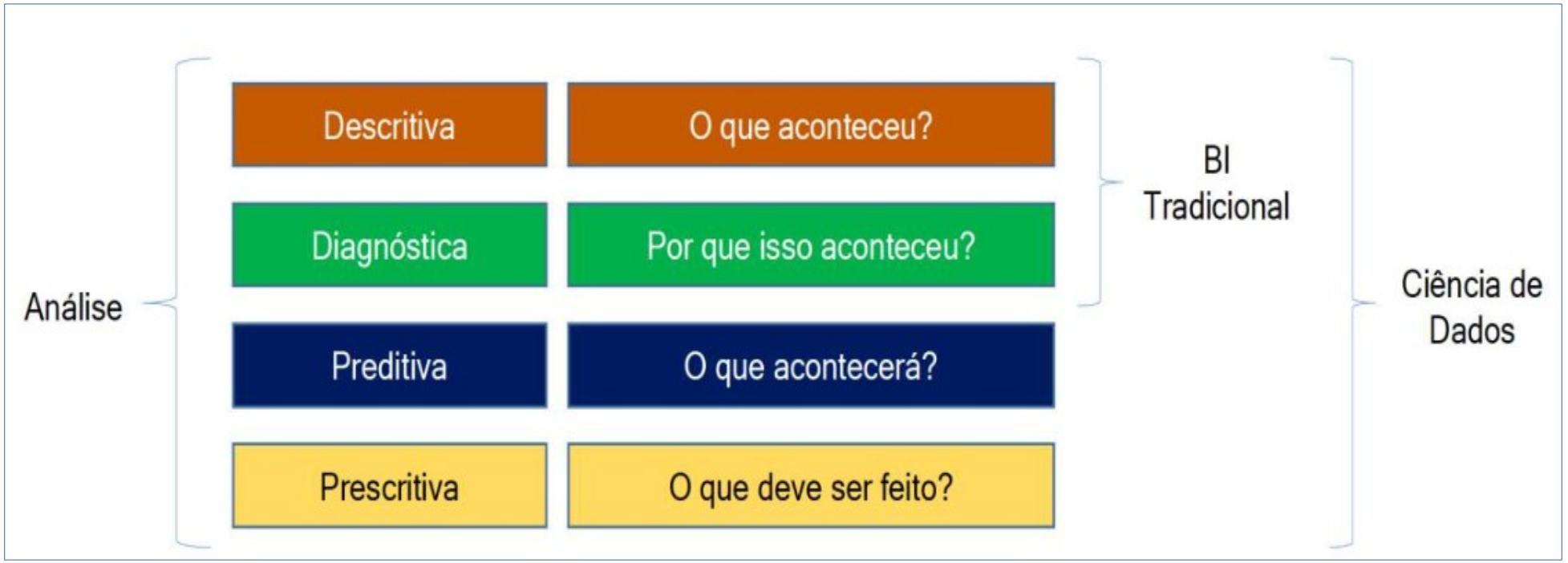
- É neste ponto que atua a ciência de dados (data science), uma área de grande destaque dentro do cenário de TI atualmente.

- De forma resumida, o objetivo da ciência de dados é analisar os dados, aplicando uma śerie 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**.

1.1 – Níveis da Ciência de Dados

- A ciência de dados é uma grande área, que pode ser dívidida em níveis distintos:



1.1 – Níveis da Ciência de Dados

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

1.1 – 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

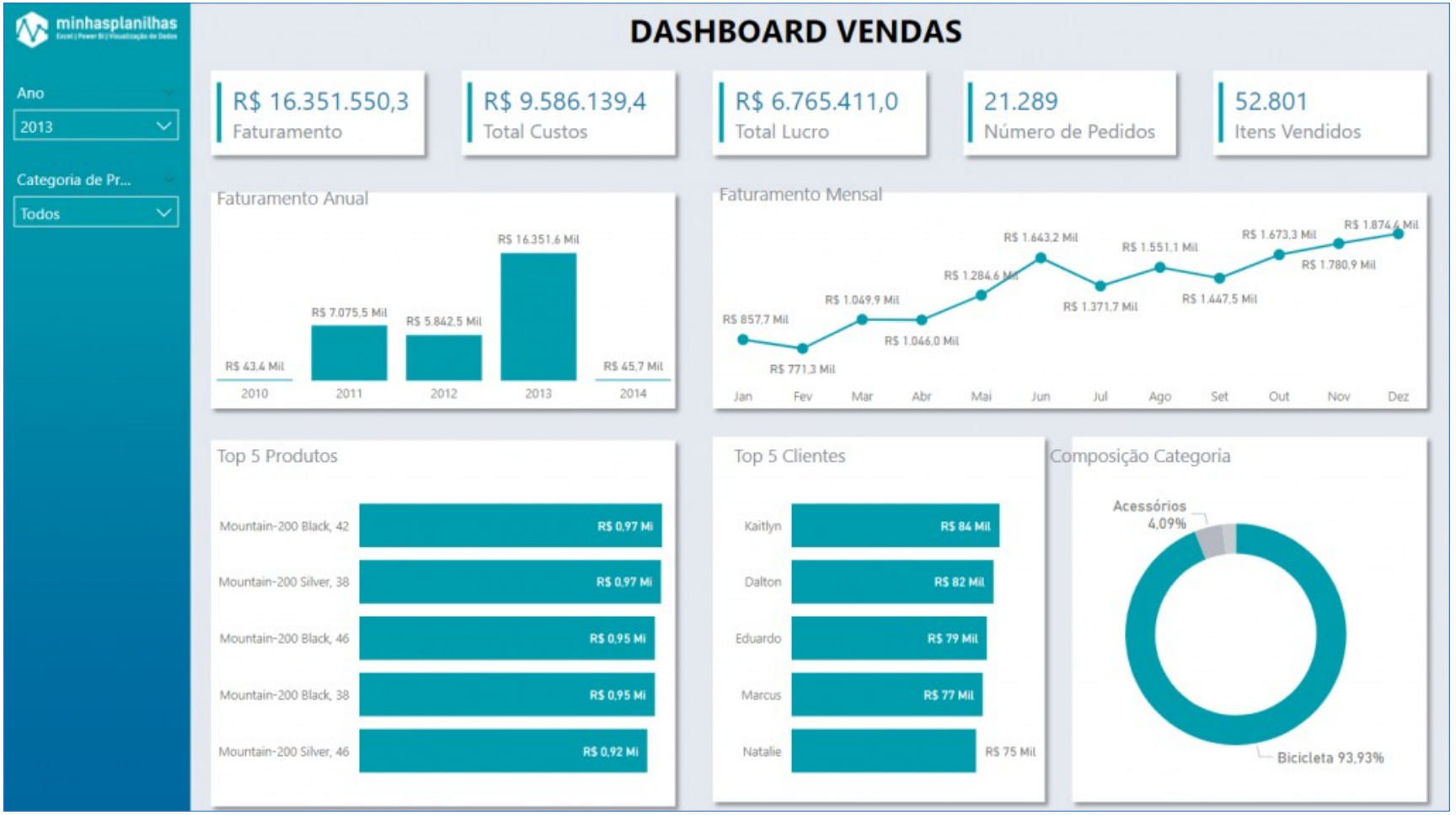
2. Vendedores que mais venderam no ano

3. Produtos com estoque baixo

4. Produtos com prazo de validade finalizando

1.1 – 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.



1.1 – 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 

1.2 – 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 ecônomia de mais de 5milhões em alguns hospitais. Auxiliou no combate a pandemia, monitorando pacientes infectados e alertando quando deve procurar o hospital.

1.2 – 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).

1.2 – 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 adéquam as suas características.

3.Objetivando aumento na segurança, movimentações na conta fora do padrão geram um alerta para bloqueio.

1.2 – 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.

1.2 – 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.

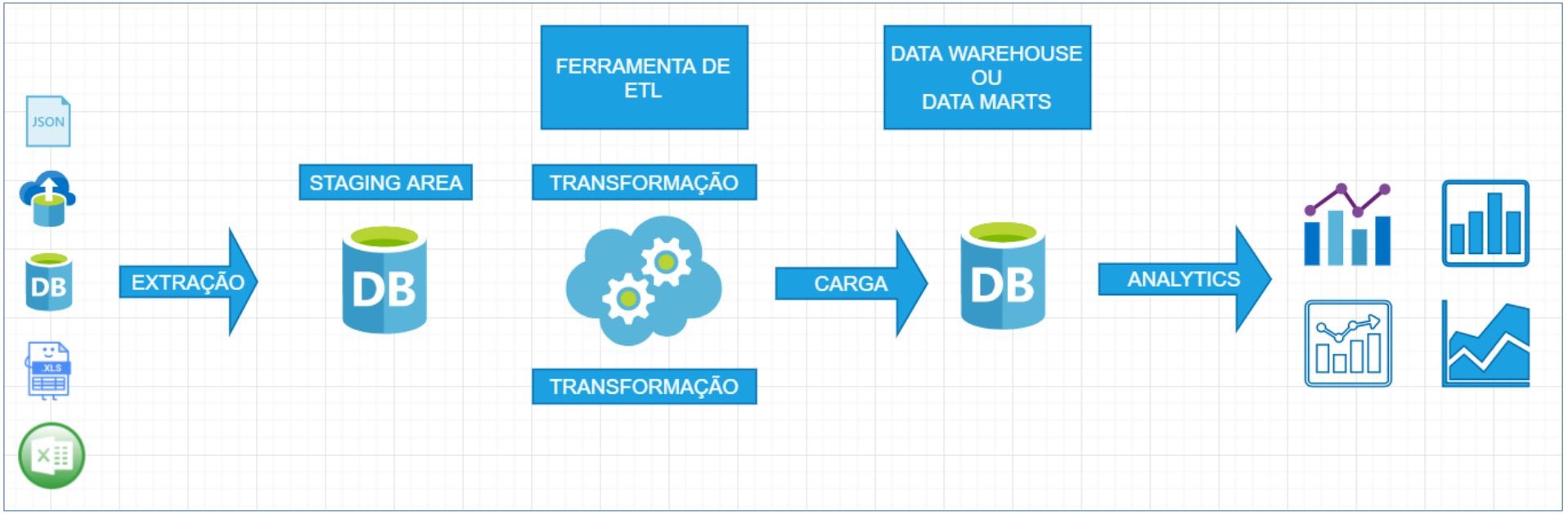
1.2 – Aplicações da Ciência de Dados

- **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.

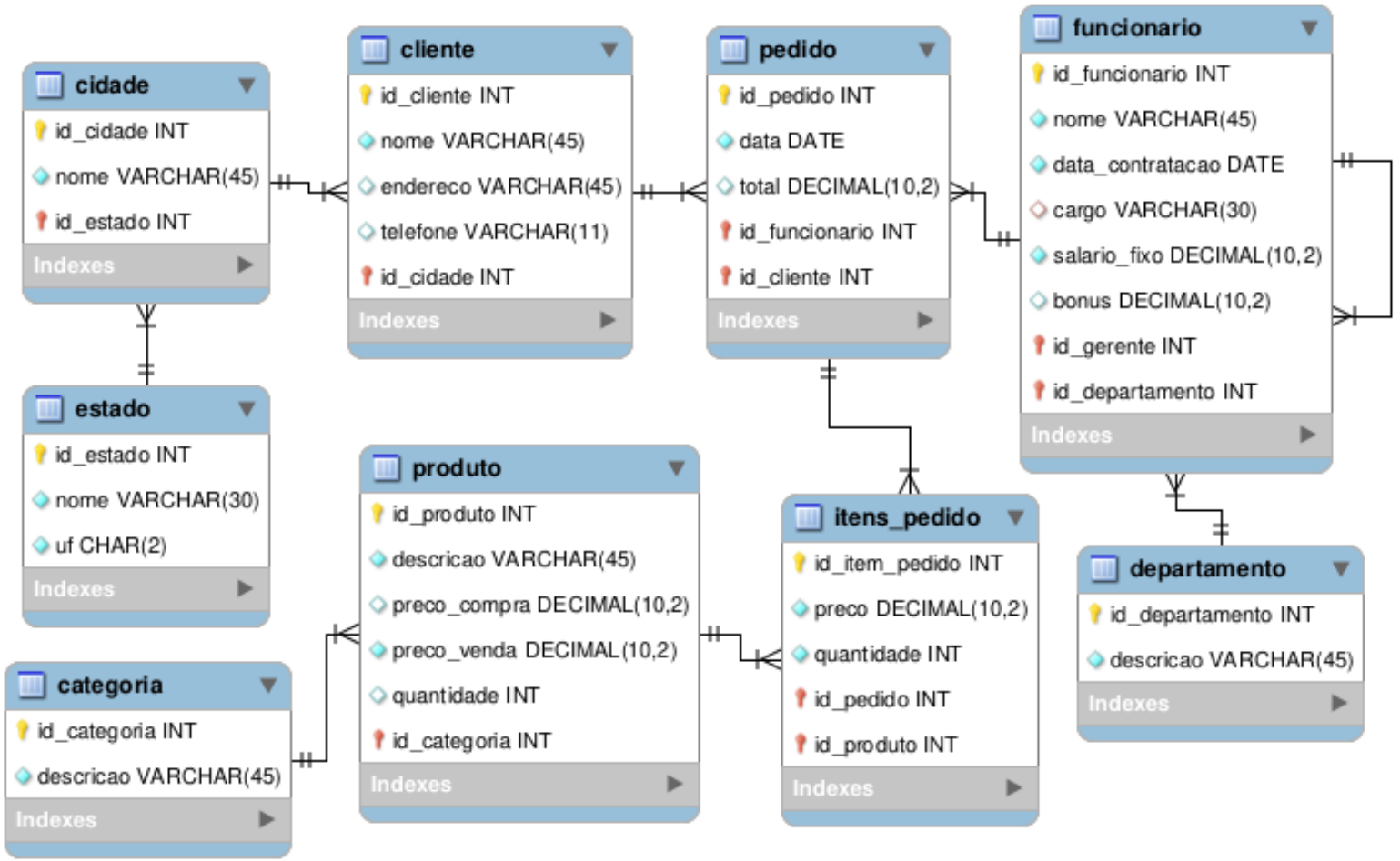
1.3 – Etapas da Análise de Dados

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



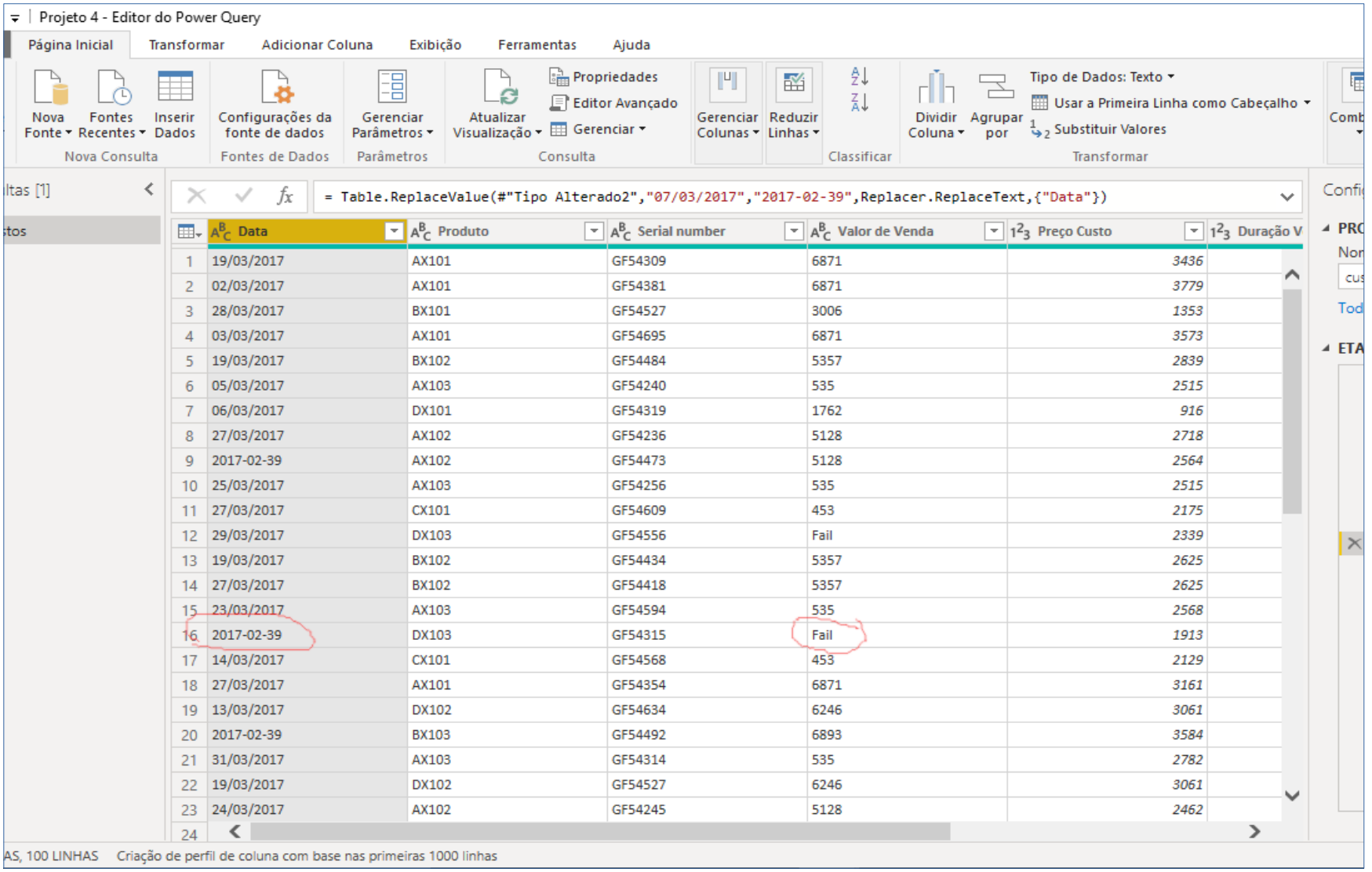
1.3 – Etapas da Análise de Dados

- O processo de extração é um dos mais complexos dentro do BI, pois os dados raramente estarão prontos para serem utilizados. Eles geralmente estarão segmentados em bases e tabelas distintas, como mostra a imagem abaixo:



1.3 – Etapas da Análise de Dados

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



1.4 – 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**.

1.4 – 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.

1.4 – 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:

1.5 – SQL no Processo de Análise de Dados

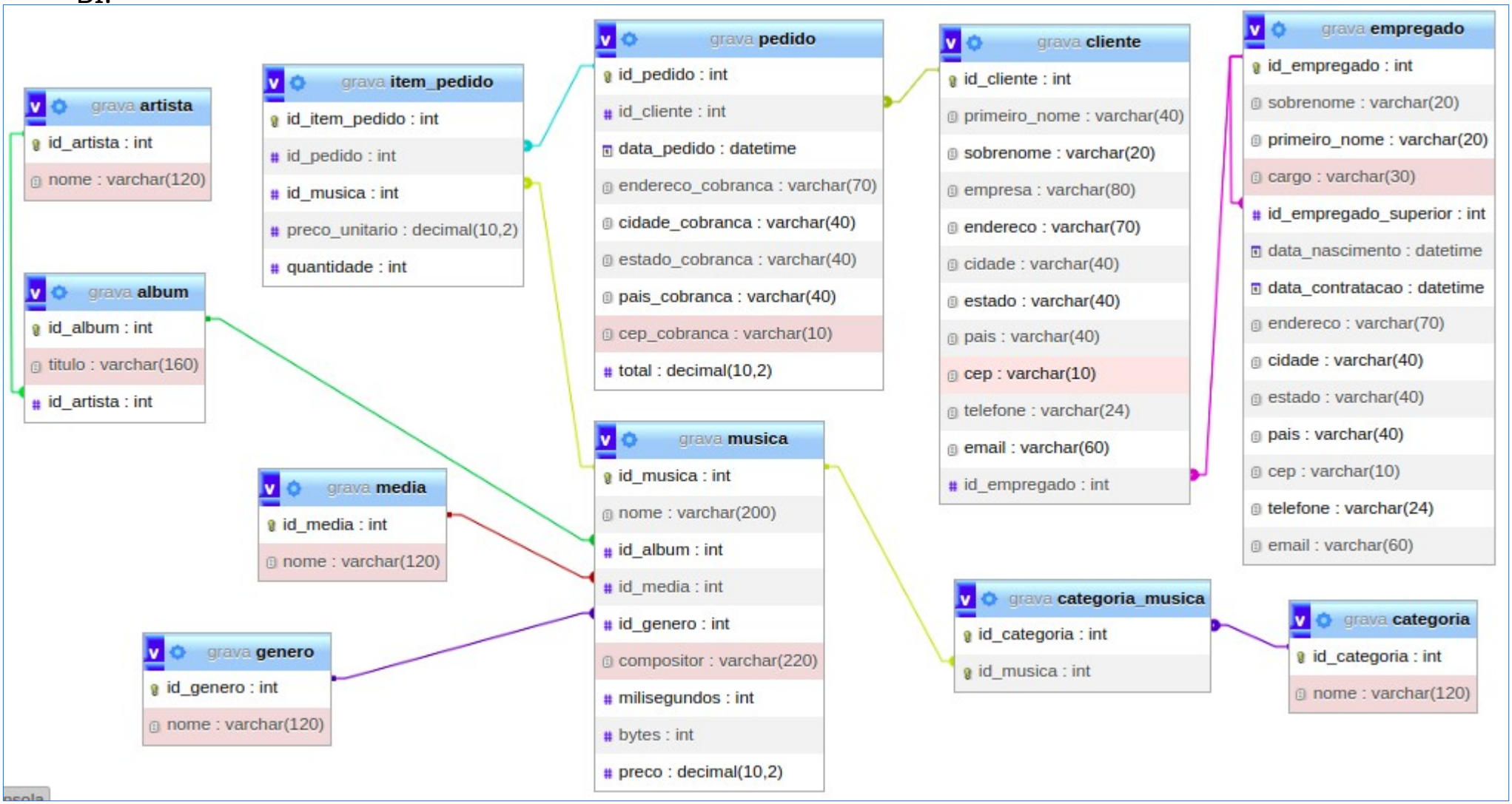
- A linguagem SQL é a principal forma de comunicação com o banco de dados, com ela conseguimos inserir, alterar e excluir registros, e também fazer diversos tipos de consultas (simples e complexas).

- Desta forma, SQL é muito útil para explorar os dados de forma que seja possível gerar informações relevantes para gestores da instituição proprietário do banco de dados.

- Além disso, antes de qualquer processo de análise e ciência de dados, é necessário conhecer profundamente o banco de dados, e a linguagem SQL possibilita fazer esta atividade de forma relativamente fácil.

1.5 – SQL no Processo de Análise de Dados

- Utilizaremos um banco e dados de uma gravadora (imagem abaixo) para exemplificar um processo de BI:



1.5 – SQL no Processo de Análise de Dados

- Determine qual é o faturamento da gravadora em cada um dos 12 meses de 2009, bem como o valor total faturado neste ano:

```
select month(data_pedido) as 'Mês', sum(total) as 'Faturamento' from pedido
where year(data_pedido)=2009 group by(month(data_pedido))
UNION
select 'Total', sum(total) from pedido where year(data_pedido)=2009
order by 1 asc;
```

Mês	Faturamento
1	35.64
10	37.62
11	37.62
12	37.62
2	37.62
3	37.62
4	37.62
5	37.62
6	37.62
7	37.62
8	37.62
9	37.62
Total	449.46

1.5 – SQL no Processo de Análise de Dados

- Determine quais são os 10 artistas que mais faturaram com venderas de músicas em 2009. Mostre o nome dos artistas, e o total faturado em ordem decrescente (mais vendeu para menos):

```
select artista.nome, sum(total) as 'Faturamento' from artista join album using(id_artista) join musica
using(id_album) join item_pedido using(id_musica)
join pedido using(id_pedido) where year(data_pedido)=2009 group by(id_artista) order by sum(total)
desc limit 0,10;
```

nome	Faturamento
Iron Maiden	310.86
Led Zeppelin	161.37
Metallica	152.46
Deep Purple	126.72
Pearl Jam	99.99
Chico Science & Nação Zumbi	74.25
Red Hot Chili Peppers	71.28
Eric Clapton	69.30
Various Artists	69.30
Queen	69.30