

# Fundamentos de *Big Data*

**PROFESSOR**

Dr. Flavio Ceci



ACESSE AQUI O SEU  
LIVRO NA VERSÃO  
**DIGITAL!**

# **EXPEDIENTE**

## **DIREÇÃO UNICESUMAR**

**Reitor** Wilson de Matos Silva **Vice-Reitor** Wilson de Matos Silva Filho **Pró-Reitor de Administração** Wilson de Matos Silva Filho **Pró-Reitor Executivo de EAD** William Victor Kendrick de Matos Silva **Pró-Reitor de Ensino de EAD** Janes Fidélis Tomelin **Presidente da Mantenedora** Cláudio Ferdinandi

## **NEAD - NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Diretoria Executiva** Chrystiano Mincoff, James Prestes, Tiago Stachon **Diretoria de Design Educacional** Débora Leite **Diretoria de Graduação e Pós-graduação** Kátia Coelho **Diretoria de Cursos Híbridos** Fabricio Ricardo Lazilha **Diretoria de Permanência** Leonardo Spaine **Head de Curadoria e Inovação** Tania Cristiane Yoshiie Fukushima **Head de Produção de Conteúdo** Franklin Portela Correia **Gerência de Contratos e Operações** Jislaine Cristina da Silva **Gerência de Produção de Conteúdo** Diogo Ribeiro Garcia **Gerência de Projetos Especiais** Daniel Fuverki Hey **Supervisora de Projetos Especiais** Yasminn Talyta Tavares Zagonel **Supervisora de Produção de Conteúdo** Daniele C. Correia

## **FICHA CATALOGRÁFICA**

### **Coordenador(a) de Conteúdo**

Flavia Lumi Matuzawa

### **Projeto Gráfico e Capa**

André Morais, Arthur Cantareli e  
Matheus Silva

### **Editoração**

Matheus Silva de Souza

### **Design Educacional**

Amanda Peçanha dos Santos

### **Revisão Textual**

Cindy Mayumi Luca

### **Ilustração**

André Azevedo, Welington Oliveira

### **Fotos**

Shutterstock

### **C397 CENTRO UNIVERSITÁRIO DE MARINGÁ.**

Núcleo de Educação a Distância. **CECI**, Flavio.

### **Fundamentos de Big Data.**

Dr. Flavio Ceci.

Maringá - PR.: UniCesumar, 2021.

**200 p.**

"Graduação - EaD".

1. Fundamentos 2. Big 3. Data. 4. EaD. I. Título.

---

Impresso por:

CDD - 22 ed. 005.7

CIP - NBR 12899 - AACR/2

ISBN 978-65-5615-435-0

---

Bibliotecário: João Vivaldo de Souza CRB-9-1679



NEAD - Núcleo de Educação a Distância

Av. Guedner, 1610, Bloco 4 Jd. Aclimação - Cep 87050-900 | Maringá - Paraná

www.unicesumar.edu.br | 0800 600 6360

# BOAS-VINDAS

A UniCesumar celebra os seus 30 anos de história avançando a cada dia. Agora, enquanto Universidade, ampliamos a nossa autonomia e trabalhamos diariamente para que nossa educação à distância continue como uma das melhores do Brasil. Atuamos sobre quatro pilares que consolidam a visão abrangente do que é o conhecimento para nós: o intelectual, o profissional, o emocional e o espiritual.

A nossa missão é a de "Promover a educação de qualidade nas diferentes áreas do conhecimento, formando profissionais cidadãos que contribuam para o desenvolvimento de uma sociedade justa e solidária". Neste sentido, a UniCesumar tem um gênio importante para o cumprimento integral desta missão: o coletivo. São os nossos professores e equipe que produzem a cada dia uma inovação, uma transformação na forma de pensar e de aprender. É assim que fazemos juntos um novo conhecimento diariamente.

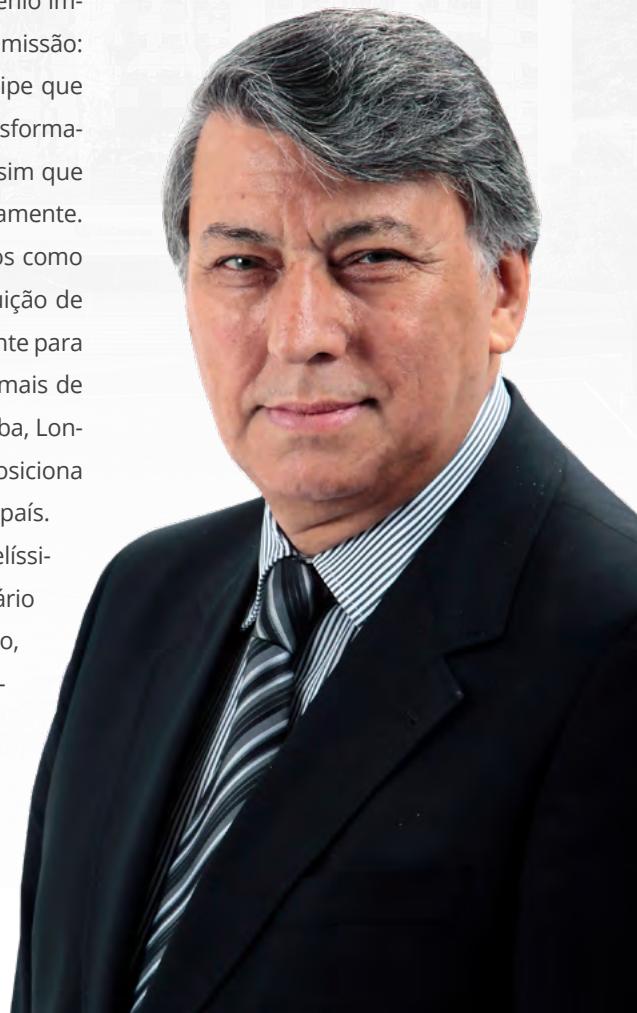
São mais de 800 títulos de livros didáticos como este produzidos anualmente, com a distribuição de mais de 2 milhões de exemplares gratuitamente para nossos acadêmicos. Estamos presentes em mais de 700 polos EAD e cinco campi: Maringá, Curitiba, Londrina, Ponta Grossa e Corumbá), o que nos posiciona entre os 10 maiores grupos educacionais do país.

Aprendemos e escrevemos juntos esta belíssima história da jornada do conhecimento. Mário Quintana diz que "Livros não mudam o mundo, quem muda o mundo são as pessoas. Os livros só mudam as pessoas". Seja bem-vindo à oportunidade de fazer a sua mudança!

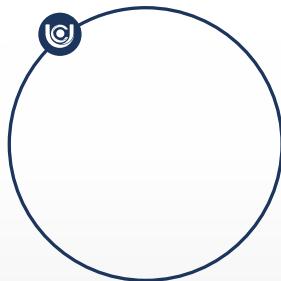
## Reitor

Wilson de Matos Silva

*Tudo isso para honrarmos a nossa missão, que é promover a educação de qualidade nas diferentes áreas do conhecimento, formando profissionais cidadãos que contribuam para o desenvolvimento de uma sociedade justa e solidária.*



# MINHA HISTÓRIA MEU CURRÍCULO



Aqui você pode  
conhecer um  
pouco mais sobre  
mim, além das  
informações do  
meu currículo.

Olá, aluno(a), vou contar um pouco a minha história! Nasci em uma cidade do interior da área continental da grande Florianópolis. Sempre fui uma criança curiosa e queria entender como as máquinas funcionavam. Ao ter contato com os computadores no início da década de 90, tive a certeza de que queria trabalhar com eles e, se possível, torná-los cada vez mais inteligentes. Em paralelo, descobri outra paixão: a música! Aprendi a tocar violão, guitarra e baixo, e pude participar de várias bandas de rock. Foi em consequência da música que encontrei os meus grandes amigos.

Como esperado, iniciei a graduação em Ciência da Computação. Aos finais de semana, era muito fácil de me encontrar: geralmente, estava tocando violão com os meus amigos nos bancos da pista de skate de minha cidade. A música me acompanhou durante toda a minha graduação e durante todo o meu primeiro emprego. Com 17 anos, tive a minha carteira de trabalho assinada pela primeira vez. A minha função era a de técnico de informática e, com o salário, eu pagava os custos da graduação e ajudava em casa. No entanto, quando estava no quarto ano, decidi que era o momento de buscar novos desafios e comecei a estagiar como programador. Na área de desenvolvimento de sistemas, ocupei várias cadeiras.

Ainda na graduação, descobri outra paixão: a área da Inteligência Artificial (IA). Quando estava no último ano da graduação, fui selecionado para trabalhar como desenvolvedor em um instituto de pesquisa que desenvolvia soluções baseadas em IA e em dados para vários setores. Foram quase nove anos de experiência naquele ambiente. Também cursei o mestrado e o doutorado, a fim aprofundar os meus estudos em relação ao uso de dados (principalmente, os não estruturados) para a tomada de decisão e para a extração de conhecimento. Quando iniciei o mestrado, comecei a namorar com quem seria a minha esposa. Ao finalizar o mestrado, dei início a carreira de docente. Quando terminei o doutorado, mergulhei na área da ciência de dados, atuando como cientista de dados. Nesse período, auxiliei na construção de processos e na divulgação da cultura de dados na organização em que eu fazia parte.

Na sequência, fui chamado para ser gestor da área de Data Science em uma instituição financeira e, no mesmo ano recebi, a notícia que mudou a minha vida: eu seria pai! Atualmente, além de me aventurar como diretor de tecnologia em uma empresa focada no desenvolvimento de soluções analíticas, meu foco e meu desafio estão na criação e no desenvolvimento do meu amado filho... Joaquim!

# **PROVOCAÇÕES INICIAIS**

## **FUNDAMENTOS DE BIG DATA**

Com o advento das plataformas de conteúdo e o movimento da chamada Web 2.0, todo usuário da Internet deixou de ser apenas consumidor e passou a ser, também, produtor de conteúdo, o que causou um enorme crescimento dos dados publicados. Esse fenômeno foi potencializado com a evolução das tecnologias móveis e diante da evolução das redes de dados, como a 3G, a 4G e, agora, a 5G.

As organizações perceberam que muitos dados são disponibilizados na Internet pelos próprios consumidores e usuários. Esses dados, em muitos casos, representam as suas opiniões, comportamentos, preferências e dentre outras informações que podem ser muito valiosas para o processo de tomada de decisão.

O armazenamento e o processamento dessas informações externas à organização são um desafio, sobretudo quando se exige que elas sejam cruzadas e somadas às informações internas. Além disso, é sabido que um ambiente de Big Data tem complexidades computacionais e de uso, o que faz com que as organizações tenham que evoluir em sua maturidade analítica.

Outras questões importantes são: quais seriam os profissionais que poderiam apoiar uma organização em um cenário como esse? Quais são as principais competências e habilidades que eles devem ter? Como é possível organizar tudo isso de forma que as organizações possam aproveitar esses dados e informações?

As organizações têm aberto cada vez mais vagas para cientistas de dados. Não só, mas também buscam estruturar as áreas focadas em dados, a fim de construir ambientes de Big Data que possam ser governados e tenham profissionais que consigam fazer uso desse ambiente. Assim, dados internos são cruzados e insights e análises são gerados para a camada tomadora de decisão.

Outro aspecto que deve ser focalizado em uma área de dados é a adequação dos processos internos da organização, com o objetivo de respeitar a Lei Geral de Proteção de Dados (LGPD), garantindo que sejam desenvolvidas as soluções focadas nos usuários e nos clientes, mas sem infringir a legislação.

As questões expostas até este momento são muito comuns. São poucas as organizações que têm maturidade analítica para terem profissionais e processos que trabalhem em um ambiente de Big Data. Diante disso, neste livro, conheceremos a jornada de um jovem empreendedor chamado Anderson, que objetiva ter um negócio que envolva o contexto da música. Acompanharemos a implantação dos primeiros sistemas de informação para a estruturação dos dados internos e a disponibilização das ferramentas analíticas.

Outros dois personagens muito importantes nessa história são Joaquim, que é um estudante de Ciência de Dados o qual acompanhará e apoiará toda a jornada, e Lara, que é a principal responsável pela área de tecnologia da organização. A narrativa sobre a empresa de Anderson e as ações tomadas por Joaquim e Lara são inspiradas em casos e em situações reais, já vividas e relatadas por outros profissionais que atuam nessa área.

É visível que muitos são os desafios para as organizações fazerem uso de um ambiente de Big Data e que, para se chegar até esse aspecto, é preciso aperfeiçoar as ferramentas e as soluções de tecnologia, e obter maturidade analítica. Tudo isso ficará bem claro mediante a presença e a participação de Joaquim e Lara durante toda a história.

O que será que já foi publicado sobre os principais desafios para o uso de ambientes de Big Data nas organizações? Agora, é o momento de você pesquisar na Internet e em artigos alguns cases de implantação de soluções para ambientes de Big Data. Conheça, sobretudo, os desafios encontrados.

# **RECURSOS DE IMERSÃO**



## **REALIDADE AUMENTADA**

Sempre que encontrar esse ícone, esteja conectado à internet e inicie o aplicativo Unicesumar Experience. Aproxime seu dispositivo móvel da página indicada e veja os recursos em Realidade Aumentada. Explore as ferramentas do App para saber das possibilidades de interação de cada objeto.



## **RODA DE CONVERSA**

Professores especialistas e convidados, ampliando as discussões sobre os temas.



## **PÍLULA DE APRENDIZAGEM**

Uma dose extra de conhecimento é sempre bem-vinda. Posicionando seu leitor de QRCode sobre o código, você terá acesso aos vídeos que complementam o assunto discutido.



## **PENSANDO JUNTOS**

Ao longo do livro, você será convidado(a) a refletir, questionar e transformar. Aproveite este momento.



## **EXPLORANDO IDEIAS**

Com este elemento, você terá a oportunidade de explorar termos e palavras-chave do assunto discutido, de forma mais objetiva.



## **NOVAS DESCOBERTAS**

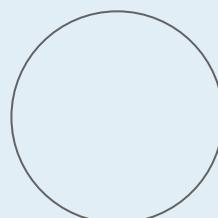
Enquanto estuda, você pode acessar conteúdos online que ampliam a discussão sobre os assuntos de maneira interativa usando a tecnologia a seu favor.



## **OLHAR CONCEITUAL**

Neste elemento, você encontrará diversas informações que serão apresentadas na forma de infográficos, esquemas e fluxogramas os quais te ajudarão no entendimento do conteúdo de forma rápida e clara

Quando identificar o ícone de QR-CODE, utilize o aplicativo **Unicesumar Experience** para ter acesso aos conteúdos on-line. O download do aplicativo está disponível nas plataformas:



# CAMINHOS DE APRENDIZAGEM

1 11  
**SISTEMAS DE  
INFORMAÇÃO E  
CONHECIMENTO**

2 53  
**INTRODUÇÃO  
À CIÊNCIA DE  
DADOS**

3 89  
**INTRODUÇÃO  
AO *BIG DATA***

4 125  
**TECNOLOGIAS  
DE APOIO AO  
*BIG DATA***

5 163  
***BIG DATA* E AS  
ORGANIZAÇÕES**



# Sistemas de Informação e Conhecimento

Dr. Flávio Ceci

## OPORTUNIDADES DE APRENDIZAGEM

Oportunidades de aprendizagem: Nesta unidade, saberemos o que são os Sistemas de Informação. Para tanto, conheceremos a sua importância para a gestão de uma empresa e explicaremos de que maneira eles podem contribuir para um ambiente de dados. Não só, mas também trabalharemos os principais tipos de Sistemas de Informação e saberemos quando aplicá-los. Por fim, estudaremos o Business Intelligence (BI) e a sua contribuição para o processo de tomada de decisão.

Anderson é um jovem empresário recém-graduado em Administração. Ele é proveniente de uma família muito grande e que vive no interior do estado do Paraná. Mudou-se para Maringá quando iniciou o seu curso de graduação e, encantado com o local, decidiu não voltar para o interior, mas se desenvolver profissionalmente na cidade.

Desde muito novo, Anderson teve contato com o mundo das vendas e do comércio. O seu pai, seu Jorge, assim como é conhecido pelos moradores do bairro em que nasceu, tem uma mercearia há mais de 40 anos no mesmo lugar e, tradicionalmente, todos os filhos trabalharam em algum momento de sua vida em seu estabelecimento. Anderson é o filho mais novo e trabalhou meio período com o seu Jorge durante a sua adolescência, principalmente enquanto cursava o ensino médio em sua cidade natal.

Seu Jorge sempre incentivou os seus filhos a trabalharem em uma área relacionada ao comércio e que proporcionasse contato com os clientes. Anderson acredita que esse foi um dos principais motivos pelos quais escolheu empreender, a fim de ter a história de sucesso de seu pai como exemplo. No entanto, tinha uma certeza: queria escrever a sua história e ter o seu próprio negócio. Esse foi o grande motivo de ter escolhido o curso de Administração.

Próximo de concluir o curso e munido de ferramentas estratégicas, financeiras e de gestão desenvolvidas durante a sua graduação, constatou que era o momento de planejar e implantar o seu negócio, cujo foco estaria relacionado à música, que era outra paixão que Anderson tinha. Assim, gostaria de abrir uma loja de instrumentos musicais e de itens que estivessem ligados a esse universo. Com a ajuda de seu Jorge e de Marta, sua irmã mais velha que assumiu a mercearia de seu Jorge depois de sua aposentadoria, levantou um capital para iniciar o seu negócio e comprar um conjunto de instrumentos e produtos para começar as suas atividades.



Em menos de um ano, a loja já era um sucesso. Já contava com vários clientes e tinha um fluxo de venda considerável. Anderson sempre foi muito organizado no que diz respeito às contas, lição que aprendeu com o seu pai e a sua irmã mais velha. Todos eles realizavam o controle a partir de um conjunto de cadernos e nunca tiveram problemas em relação a isso. Anderson iniciou o seu negócio utilizando planilhas eletrônicas para apoiar o seu processo de gestão e, com isso, conseguiu se organizar. Entretanto, o dia a dia se tornou cada vez mais difícil de ser controlado, pois havia muitos fornecedores e o controle de fluxo de caixa e o crediário aos clientes eram extensos, o que gerou uma verdadeira bagunça nas planilhas. Diante disso, Anderson teve que extinguir o seu contato com os clientes para focar na devida organização de suas planilhas.

Diante de sua situação, Anderson se recordou de uma disciplina do seu curso que explicava os Sistemas de Informação. Naquele momento, sabia que esse seria o caminho mais fácil para ter os dados e as informações sobre o seu negócio organizados, o que lhe permitiria voltar a interagir com os seus clientes. Para Anderson, esse era um diferencial em seu processo de vendas, já que, quando permaneceu em seu escritório, houve uma queda significativa nas vendas e nos serviços prestados. Outro aspecto que tornava essa decisão ainda mais urgente é o fato de que Anderson queria ampliar o seu negócio, mas não de maneira física, e sim de modo virtual, construindo um comércio eletrônico. Dessa forma, seria possível se conectar em *marketplaces*, tais como Mercado Livre, Amazon e Magalu. Tudo isso tornava a necessidade de formação de vínculos ainda mais latente.

Perante o cenário elaborado, foram levantados todos os desafios que seriam enfrentados nos próximos meses para a organização dos seus dados e para a ampliação dos seus canais de venda. Assim, foi redigida a seguinte lista:

- Ter uma visão unificada dos dados de vários setores, já que são necessárias algumas áreas, tais como a financeira, a comercial, a de controle de estoque, a de logística e a de marketing.
- É sabido que conhecer o seu cliente é muito importante para o seu negócio. Nesse sentido, é preciso pensar em estratégias que mantenham esse relacionamento, mesmo quando a base de clientes seja ampliada.
- Em detrimento do fato de que a empresa está ampliando os seus canais de venda e a sua base de cliente, é necessário ter uma estratégia bem definida e indicadores para acompanhar o sucesso ou as falhas de suas ações cotidianas.



Anderson ficou bastante preocupado com os desafios que tem pela frente, mas entende que é um caminho sem volta, ainda mais depois de experienciar um momento de pânico como o vivido durante a pandemia ocasionada pelo vírus Covid-19. Anderson sabe que o digital é um caminho sem volta e precisa se preparar para isso. Assim, buscou as anotações que fez em seus cadernos durante a disciplina que trabalhou os Sistemas de Informação e relembrou alguns aspectos que podem auxiliar em sua proposta de solução:

- É importante conhecer as diferenças entre dados, informação e conhecimento, a fim de direcionar os seus próximos passos para a transformação digital de sua empresa.
- Os Sistemas de Informação chamados de *Enterprise Resource Planning* (ERPs) podem auxiliar no tratamento dos dados operacionais e transacionais. Além disso, podem proporcionar uma visão unificada dos seus vários setores e departamentos (inclusive os novos, que terão que ser elaborados para atenderem aos desafios do digital).
- Os sistemas de *Customer Relationship Management* (CRM) objetivam melhorar o relacionamento com o cliente.
- Os sistemas de apoio à decisão e o *Business Intelligence* (BI) visam apresentar indicadores para apoiar as tarefas gerenciais e a tomada de decisão.

Diante do exposto, Anderson já sabe quais são os assuntos a serem aprofundados e os tipos de ferramentas que podem apoiá-lo nos desafios que está enfrentando.

Será que existem soluções grátis de ERP, CRM e BI no mercado? Qual é a relação existente entre essas ferramentas? É possível apenas realizar o download da ferramenta e utilizá-la? Essas são questões muito importantes a serem levantadas. Auxilie Anderson a responder essas perguntas por intermédio de uma pesquisa na Internet. Leia tudo o que encontrar em relação à essas indagações.

É perceptível que o entendimento em relação ao que é **dado, informação e conhecimento** é primordial para entender como cada um dos tipos de **Sistemas de Informação** podem apoiar os **processos** de uma **organização**. No caso dos sistemas **ERPs**, é preciso ter um processo de configuração para que, de fato, ele “espelhe” os processos das organizações. Já no caso dos **CRMs**, é necessário repensar toda a **estratégia** organizacional, para que ela seja **focada no cliente**. Já nos sistemas de **BI**, é preciso considerar os **indicadores** que, de fato, **representem** os elementos vinculados às **perguntas estratégicas da organização**.

Apanhe um bloco de papel, uma caneta e anote os aspectos que lhe chamaram a atenção a partir da pesquisa feita. Você pode anotar as características das soluções que encontrou e complementar os seus apontamentos de acordo com o aprofundamento que realizaremos a partir de agora. Ao final, verificaremos se as soluções pesquisadas também ajudariam Anderson em sua jornada. Aperte os cintos, porque há muita emoção pela frente!

## DIÁRIO DE BORDO

Agora que já conhecemos um pouco mais a história de nosso jovem empreendedor e sabemos os principais desafios e soluções que ele tem pela frente, mergulharemos um pouco mais nos conceitos necessários para atuar nas frentes a serem enfrentadas.

## Dados, informações e conhecimento

Muito se fala em *Big Data* e em todos os possíveis recursos e benefícios que são obtidos com o seu uso para fins de negócio, saúde, finanças e educação, por exemplo. Contudo, do que é composta uma estrutura de *Big Data*?

Geralmente, é afirmado que um *Big Data* é constituído por um grande conjunto de dados. No entanto, você pode estar se perguntando: o que é um dado? Quando há diferença entre um dado e uma informação? Quando nos referimos à conhecimento, trata-se da mesma coisa?

O primeiro passo é entendermos o que é um dado e qual é a sua natureza. Para tornar esse processo mais fácil, compreendamos o comercial do chinelo Havaianas com a atriz Susana Vieira. Essa propaganda foi ao ar em 2015 na TV aberta.



No vídeo, um vendedor adivinho encontrava os seus possíveis clientes e os informava de seu grande dom, que é o de adivinhar o tamanho do pé das Havaianas de seus clientes. Inicialmente, há um grupo de mulheres e, uma a uma, o vendedor acerta os tamanhos de seus pés. No entanto, repentinamente, Susana Vieira se faz presente. A atriz, que é reconhecida por ser uma mulher muito vaidosa, aproxima-se do grupo de amigas e o adivinho fala: "37!?" Prontamente, Susana Vieira responde: "Fofo! Parece".

Quando o adivinho falou o número 37, ele estava fora de contexto, o que fez com que Susana Vieira interpretasse o **dado** da melhor forma que lhe pareceu, ou seja, que ele estava se referindo à sua idade. Esse exemplo já nos apresenta uma noção do que se pode entender como dado.

Fonte: o autor.

Segundo Rezende (2011, p. 34), “o dado é um conjunto de letras, números ou dígitos que, tomados isoladamente, não transmitem nenhum conhecimento, ou seja, não contém um significado claro”. A definição apresentada descreve exatamente o que aconteceu durante a propaganda dos chinelos Havaianas. Nesse sentido, é possível utilizar a definição e o exemplo expostos para sabermos quando estamos falando de “dados”.

Aprendemos o primeiro conceito importante para o nosso caminho dentro da área de *Big Data*. O nome, por si só, já nos apresenta pistas do que está sendo tratado. Em outras palavras, se estou trabalhando em um cenário de *Big Data*, tenho uma base massiva de dados que podem ser processados para auxiliar algum elemento.

Quando falamos em dados, devemos ter ciência de que eles podem ter diferentes tipos de acordo com a sua organização. Nesse contexto, os tipos de dados são: dados estruturados, dados semiestruturados e dados não estruturados. Segundo Castro e Ferrari (2016), os **dados estruturados** são aqueles que têm uma “estrutura” previamente determinada, como os dados de uma planilha eletrônica ou de uma tabela de banco de dados. Observe o quadro a seguir:

Matrícula	Nome	Telefone	E-mail
98411	Aluno da Silva	(41) 1234-3212	aluno.silva@uol.com
98412	Aluno de Matos	(41) 4321-1234	aluno.matos@uol.com

**Quadro 1 - Exemplo de dados estruturados em uma tabela / Fonte: o autor.**

Assim como você pode perceber, os dados estruturados carregam delimitadores para especificar onde o dado se inicia e onde se termina. No exemplo “Aluno da Silva”, também é possível saber do que se trata o dado, que, nesse caso, está relacionado ao nome de uma pessoa.

Agora que já sabemos o que é um dado estruturado, estudaremos os dados não estruturados. Assim como o próprio nome sugere, os **dados não estruturados** são compreendidos como um conjunto de dados que não tem delimitadores. São exemplos: relatórios, notícias, livros, cartas e *logs* de chats (CASTRO; FERRARI, 2016).

## Destaque

Leia um trecho da música “Faroeste Caboclo” da banda de rock nacional Legião Urbana:

*Foi quando conheceu uma menina  
E de todos os seus pecados ele se arrependeu  
Maria Lúcia era uma menina linda  
E o coração dele pra ela o Santo Cristo prometeu.*

Tanto o trecho apresentado quanto toda a letra da música são dotados de dados não estruturados. Entretanto, na prática, o que isso quer dizer? Quando lemos o trecho exposto, o interpretamos e sabemos o que os termos e as suas delimitações representam, temos ciência de que “Maria Lúcia” é uma pessoa e que o seu nome se trata de um nome composto. Já um computador não sabe se o termo é “Maria”, “Maria Lúcia”, “Maria Lúcia era”, “Lúcia” ou “Lúcia era”. Assim são os dados não estruturados, tais como os dados relacionados ao diálogo do comercial dos chinelos Havaianas.

Por fim, para Castro e Ferrari (2016), os **dados semiestruturados** correspondem à combinação dos elementos dos dados estruturados com os não estruturados. Diante disso, você pode estar se perguntando: como isso é possível? É possível apresentar um exemplo? A resposta é: sim! Considere os e-mails: eles possuem alguns dados estruturados, tais como o assunto, o destinatário e o título. Depois, há um grande campo em que se pode escrever qualquer coisa. Os campos “assunto”, “destinatário” e “título” são dados estruturados, mas o campo aberto é não estruturado. Portanto, o e-mail é entendido como um exemplo de dados semiestruturados.

Agora que você já sabe tudo sobre dados, podemos passar para o estudo da informação. Ainda em relação à propaganda dos chinelos Havaianas, saber que a Susana Vieira usa Havaianas tamanho 37 é uma informação. Para se chegar a essa informação, foi necessário utilizar os dados de base e processá-los levando em consideração os elementos do domínio (cenário). Segundo Melo (2002), para se ter uma **informação**, é preciso realizar uma análise do fato envolvido a partir dos dados, o que gera um conjunto de operações que abrange desde a síntese até o processamento dos dados.

É perceptível que, para se transformar um dado em uma informação, é necessário correlacionar os fatos e as suas implicações para as pessoas e para uma organização. Para um melhor entendimento, analise a figura a seguir:



Figura 1 - Transformando dados em informação / Fonte: Fialho et al. (2006, p. 80).

**Descrição da Imagem:** na figura, é apresentada a transformação de um dado em uma informação a partir do processamento, que utiliza os fatos e as suas relações com os indivíduos e com a organização.

Estamos transformando dados em informação a todo momento, ao considerarmos os fatos que nos são apresentados (ou experienciados) e ao relacionarmos com os indivíduos e com os elementos de uma organização ou ambiente. No caso da tecnologia, esse processamento pode ser feito a partir do cruzamento dos dados relacionados e por intermédio da aplicação das regras de negócio envolvidas e previamente mapeadas. Além disso, as informações são dotadas de algumas características que são apresentadas a seguir com base em Rezende (2011):

- A informação tem um conteúdo único. Em outras palavras, a cada momento ou evento, a informação tem um conteúdo.
- A informação exige mais de duas palavras. Em outras palavras, não é possível ter uma informação relacionada apenas à palavra “saldo”, por exemplo, mas à “saldo de alguma coisa”.
- A informação não pode ser generalizada, mas deve ser sempre expressa em seu detalhe.
- A informação não pode ser abstrata, de compreensão difícil, vaga ou irreal.

Segundo Ceci (2012, p. 19, grifos nossos), “para transformar informação em **conhecimento** não basta apenas a aplicação de uma etapa de processamento (como no caso dos dados para a informação), é necessário um processo de síntese por parte de quem está consumindo a informação”.

Os dados são formas de representação dos fatos a partir de letras e números. Quando são processados (e correlacionados), geram as informações. Ao trazerem elementos do domínio em conjunto com o cruzamento de fatos não explícitos e contextualizados, permitem a obtenção de novos conhecimentos. De acordo com Rezende (2011, p. 35), “quando a informação é trabalhada por pessoas e pelos recursos computacionais, possibilitando a geração de cenários, simulações e oportunidades, pode ser chamado de conhecimento”.



Em relação ao cenário vivido por Anderson (nossa jovem empreendedor), qual é a sua relação com os dados, as informações e o conhecimento?

Anderson tinha vários dados armazenados em planilhas eletrônicas que eram usadas em seu dia a dia para gerenciar o seu negócio. Com a implantação dos Sistemas de Informação em sua empresa, ele terá acesso a várias informações a partir de relatórios e de indicadores formulados por meio do processamento e do cruzamento dos dados armazenados. Ao acompanhar a evolução dos indicadores e dos relatórios, tomará algumas decisões e aprenderá com elas, gerando, assim, novos conhecimentos para o seu processo de tomada de decisão.

Agora que já sabemos a importância e a diferença entre o uso dos dados, das informações e do conhecimento, trabalharemos os conceitos relacionados aos Sistemas de Informação. Eles são a base da proposta de solução para o negócio do Anderson. Nesse sentido, conheceremos os seus conceitos, tipos e, sobretudo, a sua aplicabilidade. Sigamos, pois há muitos assuntos a serem estudados.

## Sistemas de Informação

Os Sistemas de Informação já fazem parte do nosso dia a dia. Basta irmos até um mercado de médio ou de grande porte ou até uma farmácia que os encontraremos. Eles também estão presentes em postos de gasolina, cinemas, restaurantes e em vários outros setores. Todos esses exemplos de Sistemas de Informação fazem parte de um tipo de sistema em especial, o chamado de Sistema de Informação Operacional (ou transacional). Assim como o próprio nome sugere, trata-se de

um sistema que está focado no processo de controle e de gestão das operações do dia a dia, tratando cada uma das transações realizadas.

Entretanto, o que é um sistema? Talvez, quando é mencionada a palavra “sistema”, você pense em sistemas computacionais. No entanto, e os sistemas respiratório e digestivo? É possível considerar os sistemas como todos os conjuntos de processos que têm uma entrada, um processamento e uma saída. De acordo com Silva, Ribeiro e Rodrigues (2004), um sistema deve ter elementos inter-relacionados e a correlação entre os elementos (partes) é o que gera o efeito de “sistema”.

Diante da definição de sistema apresentada, é possível estabelecermos uma relação com os Sistemas de Informação. Assim, é plausível afirmar que é preciso ter um objetivo para o seu uso e levar em consideração os processos das áreas de negócio, de tal forma que sejam acolhidos na função interna do sistema. Além disso, é visível que não existem Sistemas de Informação sem as pessoas envolvidas. Para Silva, Ribeiro e Rodrigues (2004, p. 52), “um sistema de informação (SI) pode ser definido como um conjunto de procedimentos organizados que, quando executados, provêem informação de suporte à organização”.



#### PENSANDO JUNTOS

O que foi exposto era um dos aspectos que Anderson mais buscava. Com a sua proposta de solução, percebemos que ele está no caminho certo. Contudo, os Sistemas de Informação podem apoiar os gestores em quais aspectos?

De acordo com Laudon e Laudon (2001), o uso dos Sistemas de Informação reflete diretamente na forma com que os gestores decidem, planejam e determinam quais produtos e serviços são produzidos ou ofertados. Para Gouveia e Ranito (2004), um dos principais objetivos de um Sistema de Informação é apoiar no processo de tomada de decisão, permitindo a geração de dados estruturados e de informações de maneira adequada, considerando o seu custo, tempo e formato.



#### PENSANDO JUNTOS

Será que, na prática, basta que uma organização compre um Sistema de Informação para obter todos os benefícios provenientes?

A pergunta exposta foi feita por Anderson, quando estava refletindo sobre a compra de Sistemas de Informação para a sua empresa. Em detrimento do fato de que não obteve resposta a partir das suas anotações relacionadas à disciplina cursada durante a graduação, Anderson passou a pesquisar alguns sistemas na Internet que apoiam as mais diversas áreas. A primeira área para a qual ele resolveu buscar um sistema foi a parte de controle de estoque, já que era um elemento que não estava finalizado em sua loja e, com a inserção do canal via *e-commerce*, seria ainda mais importante ter total controle sobre os seus processos.

Contudo, Anderson não encontrou nenhum sistema que atendesse exatamente os seus processos. Diante disso, chegou à seguinte conclusão: “tenho duas alternativas: ou contrata alguma empresa para desenvolver o meu sistema de controle de estoque ou adapto os meus processos internos e, se possível, a forma com que o sistema está desenvolvido”.

A reflexão feita por Anderson foi muito acertada. Hoje, um dos grandes problemas que se tem durante a adoção de um Sistema de Informação é justamente a falta de sinergia entre a forma com que o sistema foi desenvolvido e os processos internos da organização. Realmente, as duas alternativas pensadas pelo jovem empreendedor são as mais pragmáticas. No entanto, não é permitido um sistema que não se relacione com os processos internos, pois, assim, não haverá uma coleta eficiente dos dados de entrada e um suporte efetivo aos processos.

Anderson continuou a tecer a sua reflexão e entendeu que o ideal seria solicitar a construção de um sistema feito especificamente para atender às suas demandas e ao seu processo de controle de estoque. Todavia, quando realizou os orçamentos com as empresas de desenvolvimento de software da região, ficou bastante preocupado, pois o valor era muito mais alto do que poderia pagar. Havia, ainda, um tempo considerável de desenvolvimento. Diante da situação em questão, Anderson cogitou a possibilidade de buscar sistemas “de prateleira”, a fim de reorganizar os seus processos para que eles estivessem em sintonia com os elementos do sistema adquirido.

Logo após a implantação do sistema de controle de estoque, Anderson já notou os grandes benefícios que ele trouxe para a sua operação. Com o sistema, era muito fácil fazer o processo de coleta e armazenamento dos dados e das informações relacionadas ao estoque e disseminar essas informações aos demais funcionários da empresa. De fato, esse era o caminho correto para levar a sua empresa até a Internet e ampliar os seus canais de venda.

A percepção apresentada está em total sintonia com as principais funções de um Sistema de Informação de acordo com Gouveia e Ranito (2004):

- **Coleta de dados:** permite a entrada de dados de maneira fácil e organizada (geralmente, é organizada em formulários e em telas).
- **Armazenamento de dados:** persistência dos dados já organizados em estruturas baseadas em banco de dados.
- **Processamento:** aplica as validações e as regras do negócio previamente programadas.
- **Representação de dados e informações:** são utilizados recursos gráficos para apresentar as informações e os dados armazenados nas bases do sistema.
- **Distribuição e disseminação:** garante o fluxo de dados no sistema e permite que várias pessoas tenham acesso às informações e aos dados já processados e centralizados.

Anderson ficou realmente feliz com a implantação do seu primeiro Sistema de Informação focado no controle de estoque. Agora, ele entendia que precisava focar na implantação de um sistema para apoiar no processo de contas a pagar e a receber. Foi nesse momento que se perguntou: será que, para cada área do negócio, eu terei que fazer todo esse processo? Se são sistemas distintos, como conseguirei ter uma visão centralizada dos dados da organização, a fim de fazer cruzamentos de dados?

Depois disso, percebeu que estudar apenas os conceitos de sistemas, Sistemas de Informação e sistemas operacionais (transacionais) não era o suficiente para resolver os seus problemas. Era necessário mergulhar um pouco mais nos tipos de Sistemas de Informação, com o objetivo de saber qual ou quais tipos poderiam melhor lhe atender. Feito isso, Anderson se lembrou de uma figura de uma pirâmide com os níveis hierárquicos que podem existir em uma organização. Assim, perguntou-se se existia alguma relação entre o tipo do Sistema de Informação e o nível hierárquico em questão. A figura a seguir apresenta a pirâmide em questão:



Figura 2 - Níveis organizacionais em relação ao tipo de decisão tomada

Fonte: O'Brien (2004 apud OLIVEIRA; CARREIRA; MORETI, 2009, p. 155).

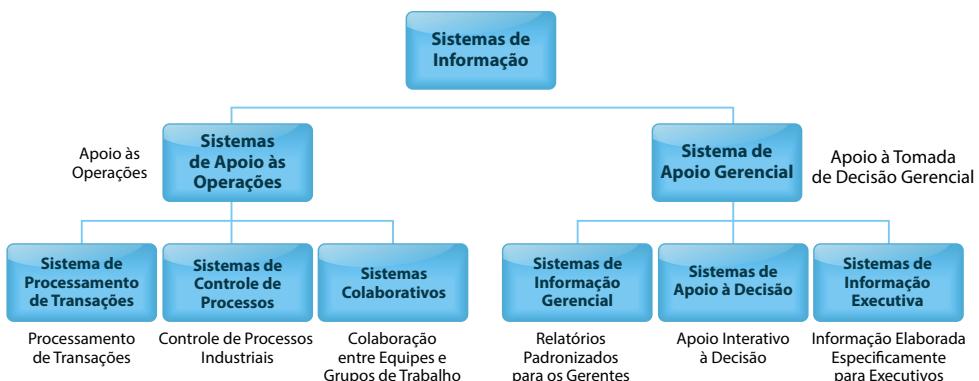
**Descrição da Imagem:** na figura, são apresentados os três níveis organizacionais no que diz respeito ao tipo de tomada de decisão na forma de uma pirâmide. Na base da pirâmide, estão as decisões de apoio às operações e aos processos. Na camada do meio da pirâmide, está o apoio à tomada de decisão gerencial. Por fim, no topo da pirâmide, há o apoio às estratégicas e às decisões executivas.

Anderson, mais uma vez, chegou a uma conclusão correta: realmente, os Sistemas de Informação têm tipos específicos para apoiar os processos que podem estar mais focados em uma camada em especial. Contudo, a primeira grande divisão é realizada entre os sistemas de apoio a operações e os sistemas de apoio à decisão gerencial. O sistema de controle de estoque é um sistema focado em apoiar processos e transações, ou seja, tem foco no apoio à camada de base da pirâmide, em que estão a maioria dos funcionários e gestores com cargos de supervisores, coordenadores ou líderes técnicos.

Esses sistemas apoiam diretamente as principais operações das organizações. Um tipo muito conhecido e utilizado de sistemas de apoio a operações são os sistemas *Enterprise Resource Planning* (ERP), que apresentam uma visão unificada de toda a organização, propõem soluções para vários setores, mas também há a possibilidade de visualização única, cruzando várias dimensões do negócio. Quando Anderson descobre os sistemas ERPs, entende que essa pode ser uma

ótima solução para os seus problemas. No entanto, pelo fato de que não deseja cometer o mesmo equívoco, avança um pouco mais nos estudos em relação aos tipos de Sistemas de Informação.

Ainda em relação aos sistemas de apoio a operações, há os seguintes tipos: sistemas de processamento de transações (que é o caso dos sistemas ERP), sistemas de controle de processos e sistemas colaborativos. A figura a seguir apresenta mais detalhadamente os tipos de Sistemas de Informação:



**Figura 3 - Tipos de Sistemas de Informação**

Fonte: O'Brien (2004 apud OLIVEIRA; CARREIRA; MORETI, 2009, p. 157).

**Descrição da Imagem:** a figura apresenta, em forma de organograma, os tipos dos Sistemas de Informação. Na imagem, constam dois tipos: os sistemas de apoio às operações, que, assim como o próprio nome já sugere, apoiam as operações organizacionais, e os sistemas de apoio gerencial, cujo foco recai no apoio à tomada de decisão gerencial da organização. Abaixo dos sistemas de apoio às operações, há três outros subtipos: o sistema de processamento de transações, os sistemas de controle de processos (geralmente apoiam processos industriais) e os sistemas colaborativos (em que a colaboração se dá entre equipes e grupos de trabalho). Já abaixo dos sistemas de apoio gerencial, há mais três subtipos: os sistemas de informação gerencial (que geram relatórios padronizados para os gerentes), os sistemas de apoio à decisão e os sistemas de informação executiva, cujo foco está na geração de informações estratégicas para os executivos da organização.

Os **sistemas de processamento de transações** estão diretamente ligados aos tipos de Sistemas de Informação que estamos acostumados a lidar em nosso cotidiano, tais como o sistema de controle de estoque. No que diz respeito aos **sistemas de controle de processos**, eles estão relacionados aos sistemas que monitoram a execução de processos e de etapas dentro de fábricas e indústrias. Por fim, os **sistemas colaborativos**, geralmente, são focados no compartilhamento de casos conhecidos, tais como os sistemas que gerenciam perguntas mais frequentes ou comunidades de práticas.

Os **Sistemas de Apoio Gerencial (SAG)** estão focados no processo de apoio à tomada de decisão, ou seja, o seu foco se dá nas duas camadas superiores da pirâmide apresentada pela Figura 2: as camadas dos gerentes e dos diretores da organização. O primeiro tipo de SAG são os **sistemas de informação gerencial**, que objetivam gerar relatórios para que seja possível acompanhar os processos de sua área e podem detalhar a transação. Os **sistemas de apoio à decisão** podem ser utilizados por toda a organização, mas, tradicionalmente, são mais utilizados pelos gerentes e pelos diretores, tendo em vista que, com ele, é possível visualizar os vários indicadores de gestão da organização. Por fim, há os **sistemas de informação executiva**, que apresentam uma visão de alto nível da organização, geralmente, partindo dos indicadores mais macros e que estão diretamente ligados à evolução das metas e a todos os demais indicadores que demonstrem a saúde do negócio e a efetividade da estratégia definida.

Anderson ficou muito feliz com tudo o que estudou até agora. Ele compreendeu ainda mais os tipos de Sistemas de Informação e selecionou três tipos que julgou serem extremamente alinhados às suas necessidades. Os tipos escolhidos foram: ERPs, CRMs e soluções de *Business Intelligence*. Esses serão os próximos destinos a serem explorados durante a sua viagem em busca das melhores soluções para o seu negócio.

## ***Enterprise Resource Planning (ERP)***

Os sistemas chamados *Enterprise Resource Planning* (em português, Planejamento de Recursos Empresariais) ou simplesmente ERP, assim como são popularmente conhecidos, são a evolução de uma técnica que se popularizou nos anos 60, a denominada *Material Requirements Planning* (MRP). Ela tinha, como foco, a gestão do estoque de matéria-prima para o processo de produção industrial. A sua evolução aconteceu durante a década de 80, com os chamados *Manufacturing Resources Planning* (MRP-II). Nessa versão, também eram feitos o gerenciamento da distribuição e o controle de produção e do chão de fábrica.

Já na década de 90, em conjunto com o lançamento da Internet em seu formato comercial, o termo evoluiu mais uma vez para “ERP” e passou a adquirir as características dos principais processos e tarefas empresariais. Diante do exposto, é possível perceber que a evolução do termo só foi possível em detrimento do aumento da capacidade de processamento computacional durante as décadas mencionadas:



O sistema ERP tem como objetivo permitir que as empresas possuam uma maior integração entre os processos da organização, pois quanto mais preciso e ágil o fluxo das informações, maior vai ser a velocidade com que esta informação será processada, o que é essencial para atender a velocidade do mercado globalizado. Integrar estes processos de uma maneira que permita que a informação flua rapidamente, sem o auxílio da tecnologia de informação é humanamente impossível (GONÇALVES; LIMA, 2010, p. 61).

A partir dos anos 2000, com a progressão do poder computacional e da banda larga da Internet, os sistemas ERPs puderam ser facilmente plugáveis em bases externas e com fornecedores, constituindo, desse modo, um sistema único para fazer todo o processo de apoio à operação da organização. Ao ler sobre a evolução dos sistemas ERPs, Anderson estava mais do que convencido de que esse seria o substituto ideal para o seu sistema de controle de estoque e para as demais planilhas que eram mantidas para o gerenciamento de vários setores, inclusive os que a sua empresa teria a partir da sua inserção no *e-commerce*.

Anderson ficou muito empolgado com a possibilidade de ter um sistema que pudesse apoiá-lo em todos os processos que vinha fazendo manualmente. Todavia, continuou buscando informações sobre o que deveria ser feito ou qual estratégia deve ser adotada para escolher e implantar um sistema de ERP. Foi nesse momento em que encontrou o trabalho de Breternitz (2004), que apresenta duas etapas para a seleção da ferramenta de ERP:

- **Etapa 1:** nesta etapa, deve-se buscar informações sobre os principais fornecedores de ERPs, selecionando os mais qualificados e limitando, se possível, de três a quatro.
- **Etapa 2:** na sequência, é preciso pedir uma proposta técnica detalhada, a fim de verificar a aderência das funcionalidades e dos processos com os das áreas de negócio da organização. Tudo isso objetiva uma menor quantidade de personalização das funções, o que implica no valor da solução final.

Anderson ficou ainda mais empolgado com tudo e, rapidamente, iniciou a primeira fase. Entretanto, foi alertado de que era importante considerar a infraestrutura computacional para executar o ERP, ou seja, se essa infraestrutura estaria disponibilizada de maneira interna ou na nuvem. Depois de realizar mais pesquisas, Anderson optou pelas ferramentas de ERP que fossem disponibilizadas como serviço (SaaS), ou seja,

sem a necessidade de instalação. Diante disso, a primeira etapa já havia sido concretizada. Na sequência, deu início ao mapeamento dos processos internos existentes, a fim de avaliar qual ERP estaria mais aderente ou que apresentaria um melhor processo a ser implantado.

Nesse momento, Anderson chamou todos os funcionários para conversar sobre a nova implantação e sobre os processos que haviam sido mapeados. Repassou todos os processos para a equipe, falou sobre o sistema ERP, explicou que teriam os processos internos já desenvolvidos e preservados pela ferramenta e, para as áreas que ainda não havia processos definidos, o fornecedor apresentaria sugestões de processos para tirar vantagens da ferramenta contratada. Por fim, falou sobre a grande descoberta que foi a aquisição do sistema ERP como serviço (SaaS) e sustentou que ninguém precisaria se preocupar com a infraestrutura de servidores, visto que o sistema já era integrado com grande parte dos fornecedores.

Havia um cliente dentro da loja que assistiu a todo o discurso do Anderson. O cliente o chamou em um canto da loja e lhe disse: “Muito legal o sistema ERP que você adquiriu. Muito inteligente a abordagem de tê-lo contratado como serviço, mas, uma dúvida: você terá acesso a todos os dados armazenados, certo?”. Anderson ficou sem chão, pois ainda não havia pesquisado em relação ao acesso à base de dados. Diante da situação, pensou que o cliente não sabia muito sobre o assunto e tentou alterar o rumo da conversa: “Joaquim, os vinis do Audioslave que você encomendou já chegaram! Vou buscá-los para você!”.

Anderson andou até o depósito e pensou: o Joaquim é um garoto muito inteligente. Ele deve estar fazendo graduação em alguma área relacionada aos dados. Uma vez, disse-me que o seu pai trabalhava como executivo em uma empresa de dados e teve toda a sua carreira na área de tecnologia. Acredito que vale a pena entender um pouco mais o motivo de sua pergunta.

Anderson colocou o vinil do Audioslave dentro da sacola em conjunto com um jogo de palhetas de brinde. Foi até o rapaz e lhe disse: “Joaquim, está aqui a sua compra. Também estou te dando algumas palhetas de presente. Contudo, diga-me: qual graduação você cursa?”. Joaquim conferiu as palhetas, ficou muito agradecido com o presente e falou: “Muito obrigado! Sobre a minha graduação, estou no segundo ano do curso de Ciência de Dados e estou gostando muito”.

Anderson não sabia exatamente para qual área o curso era voltado e ficou com vergonha de perguntar. Assim, foi direto ao assunto: “Por que você me perguntou sobre acesso à base de dados do ERP?”. Joaquim adorou a sua pergun-

ta e respondeu: "Anderson, o sistema ERP te ajudará a gerenciar todos os seus processos e operações de maneira simples. Você conseguirá ter várias visões e cruzamentos dentro do próprio sistema, o que é muito legal. No entanto, e se você quiser utilizar esses dados para compor outros sistemas analíticos? E se você desejar a realização de uma análise exploratória dos seus dados? Da forma com que comentou, parece-me que não é possível".

Os aspectos levantados por Joaquim eram muito relevantes e, de fato, poderiam representar um grande problema no futuro, caso não tivesse acesso aos dados coletados e processados pelos sistemas ERP. Anderson ficou muito surpreso, visto que esse fato não foi comentado em nenhum dos artigos que leu, mas não perderia tempo. Nesse contexto, acionou o fornecedor e exigiu mais esse requisito. O fornecedor ficou surpreso com o questionamento e falou: "Você está pensando em utilizar uma ferramenta para melhorar o seu relacionamento com o cliente e medir as metas e os indicadores da organização?". De imediato, Anderson respondeu positivamente. Na sequência, o fornecedor afirmou que eles disponibilizam acesso às bases de dados e têm todo um módulo de integração de dados com as ferramentas de CRM.

Anderson voltou a ficar muito empolgado com tudo. Ele sabia que estava dando um importante passo para alcançar o seu objetivo, ao implantar a ferramenta de ERP. Ficou mais feliz ainda, visto que ela já teria a possibilidade de acesso aos seus dados e, ainda, teria um módulo de integração com as ferramentas de CRM. Nesse sentido, assim que a operação digital estivesse disponível e gerando lucro, ele passaria a estudar mais detalhes sobre esse tipo de ferramenta, para que serve e como ela pode auxiliar o seu negócio.

## ***Customer Relationship Management (CRM)***

Passado um ano de implantação do ERP e mais de 20.000 clientes conquistados por meio dos canais digitais, a empresa de Anderson deu um grande salto e passou de apenas uma loja física para uma organização que conta com mais duas lojas além da matriz e um galpão muito grande para o estoque de toda a sua operação. Anderson estava muito feliz com o salto que a empresa deu, mas estava preocupado com um fato: estava perdendo um dos pilares para a gestão do seu negócio e que havia aprendido com o seu pai, que era conhecer o seu cliente e proporcionar uma experiência personalizada.

De fato, Anderson percebeu que, dos 20 mil clientes que fizeram compras em sua loja on-line, apenas 100 mantêm um relacionamento com a sua loja. Ele conseguiu fazer esse levantamento por meio do ERP implantado. Outro aspecto que também o deixou muito preocupado é o de que, com a ampliação das lojas físicas, esse relacionamento também foi perdido com os novos clientes, principalmente nas lojas novas. Foi então que se lembrou de que, logo depois que havia estudado os sistemas ERP e iniciado o seu processo de implantação, havia anotado que estudaria o CRM.

Anderson foi até a livraria mais próxima e buscou livros que trabalhassem o CRM. Lá, descobriu que a sigla significa *Customer Relationship Management* e, em português, pode ser traduzida como Gerenciamento do Relacionamento com o Cliente. Já se empolgou apenas pela tradução do nome, pois era exatamente isso que precisava implantar em seu negócio. Algo que Anderson sempre acreditou e confirmou ao observar o negócio de seu pai é o de que, mantendo um relacionamento com o cliente, há um processo de fidelização e a receita recorrente é muito mais garantida. Segundo Pinheiro (2008), o CRM não diz respeito apenas à tecnologia, mas a uma estratégia de negócio focada no entendimento das necessidades atuais do cliente e de possíveis compras futuras. Assim, é composto por um processo de captura e armazenamento dos dados do cliente ao longo do tempo. As informações são unificadas e de fácil acesso.

Ao ler a definição exposta, Anderson entendeu mais uma oportunidade que o CRM estaria lhe proporcionando. Pela primeira vez, comprehendeu o sentido da recomendação que o seu cliente, Joaquim, havia lhe dado: Anderson tem total acesso aos dados armazenados por intermédio da ferramenta de ERP implantada, o que facilitaria muito o trabalho de implantação de uma ferramenta de CRM.

De acordo com Pinheiro (2008, p. 18), algumas ações constituem um ambiente de gerenciamento de relacionamento. São elas:



- Identificação dos melhores clientes de uma organização;
- Ter o controle sobre campanhas com objetivos e metas claras;
- Produzir indicadores para as equipes de vendas;
- Apoiar as estratégias de transformação de potenciais clientes;
- Recuperar clientes perdidos;
- Aumentar a lucratividade;
- Otimizar os processos de vendas;
- Permitir a personalização do atendimento a partir de um relacionamento individualizado.

Em relação às funcionalidades de um CRM, observe a figura a seguir:



Figura 4 - Funcionalidades de um CRM / Fonte: Sanchez (2014 apud ALVES, 2018, p. 125).

**Descrição da Imagem:** a figura apresenta as funcionalidades de uma solução de CRM a partir de quatro dimensões: vendas, pedidos, suporte e marketing. Em vendas, são apresentadas as funcionalidades de: 1) contatar, qualificar e converter público-alvo em clientes; 2) rastrear oportunidades; e 3) fechar pedidos. Já no pilar “pedidos”, há: 1) entregar demanda; e 2) faturar. No pilar “suporte”, há: 1) gerenciar atendimentos; 2) conduzir treinamentos; 3) fornecer serviços; e 4) desenvolver base de conhecimento. Por fim, no pilar “marketing”, constam: 1) projetar, desenvolver e executar campanhas; 2) definir público-alvo; e 3) criar base de dados.

A Figura 4 expõe a visão de Alves (2018), que sustenta que o CRM tem funcionalidades para as áreas de marketing, de vendas, de pedidos e de suporte, o que demonstra ser uma solução muito versátil e, de fato, pode contribuir para a estratégia da organização. Anderson adorou saber que poderia estar apoiando diretamente as quatro áreas mencionadas e pensou no que seria necessário fazer para implantar uma solução como essa. Para se ter uma solução de CRM efetiva e de sucesso, é preciso lançar um novo olhar sobre a empresa e a sua estratégia. Em outras palavras, segundo Pinheiro (2008), é necessário:

- Repensar a estrutura organizacional.
- Reavaliar os sistemas de informação.
- Revisitar os processos.
- Reavaliar os orçamentos.

O **repensar a estrutura organizacional** está diretamente ligado a uma premissa básica para se trabalhar com CRM, que é mudar o foco estratégico da organização e centralizá-lo no cliente. Em outras palavras, colocar o cliente no centro da estratégia é fazer com que todos os processos e ações da organização estejam sempre considerando o fornecimento de uma melhor experiência para ele, uma vez que essa atitude por si só fará com a organização tenha um aumento em seu faturamento. Logo, a estrutura organizacional deve ser repensada para dar esse novo foco estratégico.

No processo de **reavaliar os sistemas de informação**, é preciso verificar quais sistemas têm informações relevantes sobre o histórico do cliente com a organização e, diante disso, acessar às suas bases de dados. No caso do Anderson, ele tem um sistema ERP que centraliza o controle de todos os processos em uma única base de dados e já estava previsto, em contrato, o acesso total à sua base.

O **revisitar os processos** está relacionado com a ideia de tornar a organização centrada no cliente. Nesse sentido, processos devem ser revistos e criados, para que seja possível, a partir dos dados extraídos das ferramentas de CRM, promover uma melhor experiência para o cliente final. Por fim, **revisitar o orçamento** é natural de se considerar, já que a estrutura e os processos podem ser alterados e, com toda a certeza, serão desenvolvidas novas campanhas que permitirão gerar novas experiências para os clientes, demandando mais investimentos.

Após revisitar os aspectos propostos e disseminar a cultura do cliente aos seus colaboradores, Anderson buscou um fornecedor de CRM e, para a sua surpresa, encontrou até ferramentas grátis. Em detrimento do fato de que precisou elaborar uma área da tecnologia para apoiá-lo nos processos relacionados aos canais digitais, pediu para que a equipe fizesse a implantação da ferramenta de CRM em um dos seus servidores e fossem definidos os processos de migração e atualização dos dados relacionados aos clientes das bases do ERP para a base do CRM.

Depois da implantação da ferramenta e do carregamento da base, foi possível visualizar os recursos dentro da ferramenta de CRM, o que permitiu que as jornadas de comunicação fossem gerenciadas facilmente. Para isso, foram utilizadas, como base, as interações do cliente com os canais da organização (físico ou digital), possibilitando a criação de alertas e recomendações de produtos baseados em seus gostos. Anderson estava pressentindo uma nova evolução em sua empresa.

## ***Business Intelligence (BI)***

Oito meses se passaram desde que foi implantada a solução de CRM. Os processos foram revistos para estarem sempre voltados ao cliente, a fim proporcionar a melhor experiência. Além do mais, foram criadas áreas de apoio, tais como a área de experiência do cliente (internamente chamada de UX), com o objetivo de analisar todos os processos com o cliente no centro. A área de marketing ganhou o auxílio do marketing digital e, assim, aumentou a recorrência de compra dos clientes que apenas visitavam os canais digitais e conquistou novos fregueses, chegando a uma base de, aproximadamente, 100 mil clientes. Anderson estava muito feliz com toda a evolução do seu negócio e abriu a sua primeira loja fora do estado do Paraná, em Florianópolis, Santa Catarina. Em sua percepção, ter implantando uma solução de CRM realmente o auxiliou no processo de relacionamento com o seu cliente, uma vez que permitiu que ele o conhecesse melhor e mantivesse um canal de comunicação sempre aberto.

Era uma terça-feira à tarde. A sua loja matriz estava com muitos clientes e Anderson reconheceu um que estava testando uma guitarra: era o Joaquim. Anderson quis lhe contar os seus últimos passos e relatar como a dica que ele havia dado sobre ter acesso às bases de dados do sistema de ERP facilitou muito o processo de implantação da ferramenta de CRM. Joaquim ficou muito feliz em ajudar Anderson, sobretudo porque realmente havia sentido uma diferença no relacionamento da loja com ele, ao receber comunicações sobre promoções e descontos no dia do seu aniversário, por exemplo. Eles conversaram muito sobre os sistemas implantados e como o uso de dados e de informações levaram o negócio de Anderson a crescer tanto e, agora, contar com quase 100 funcionários e 100 mil clientes. Em um determinado momento, Joaquim perguntou a Anderson qual era a tecnologia ele estava utilizando para a sua solução de *Business Intelligence (BI)*. Anderson não soube ao certo responder e falou que não havia uma solução de BI. Diante disso, parou por um minuto para refletir.

Joaquim estava dedilhando uma guitarra Fender Jazz Master e Anderson perguntou: “Como uma solução de BI poderia ajudar o meu negócio? Eu já tenho um sistema ERP que me permite ter uma visão de toda a minha operação de maneira centralizada e tenho uma solução de CRM que permite que eu me relacione melhor com o meu cliente. O que uma solução de BI iria me agregar?”. Joaquim elogiou o timbre da guitarra e começou a expressar a sua visão sobre o assunto. Sob a ótica de Joaquim, a empresa de Anderson precisa controlar mais de perto a execução de sua estratégia e garantir uma forma fácil de acompanhar as metas das áreas de negócios e dos funcionários a partir do uso de um sistema de apoio à decisão. Nesse caso, uma solução de BI seria ótima, já que ela deve partir da formulação de um conjunto de perguntas estratégicas que guiarão a modelagem de dados e a construção dos indicadores que serão acompanhados pelas áreas de negócios. Joaquim reforçou que, no caso dessas soluções, não há nada “pronto”, mas é necessário desenvolver algo com base especificamente no seu negócio e respeitando a estratégia organizacional desenvolvida.

Anderson ficou muito agradecido por toda a explicação que Joaquim deu e perguntou se poderia contar com ele na construção desse projeto, já que os seus colaboradores do setor de TI tinham uma visão muito mais focada em infraestrutura de tecnologia do que propriamente em dados. Joaquim ficou muito feliz com a possibilidade de participar de um projeto como esse. Ele já havia participado de outros projetos de dados junto de seu pai, mas seria a primeira vez que ele estaria à frente de um projeto desses sozinho. Joaquim deixou bem claro para Anderson a sua falta de experiência em conduzir projetos como esse, mas explicou que coletaria todo o material que tem estudado em sua graduação e os compartilharia com ele, para que possam, juntos, pensar nas etapas do projeto.

No que diz respeito a uma solução de BI, Joaquim sabe que não se trata apenas de tecnologia. Uma solução de BI está diretamente ligada às perguntas estratégicas que têm indicadores como respostas. Esses indicadores auxiliam no controle e na medição de como a organização está “caminhando”, além de possibilitarem o acompanhamento, em tempo real, da eficiência das estratégias e das ações tomadas pelos dirigentes da organização. Essa primeira fase pode ser considerada uma etapa de análise de requisitos do negócio e é composta pelos seguintes feitos:

- **Definição das perguntas estratégicas da organização:** como perguntas estratégicas, são entendidos os principais questionamentos que os gestores têm e que estão relacionados com a estratégia definida, a fim de se chegar aos objetivos principais.

- **Definição dos indicadores (ou medidas) como respostas às perguntas:** são os valores numéricos que medem algo para apoiar o acompanhamento da estratégia organizacional.
- **Identificação dos possíveis filtros vinculados com cada indicador:** assim como o próprio nome sugere, são os possíveis filtros e agrupamentos que podem ser aplicados para visualizar o indicador.
- **Prototipação de painéis de indicadores para a avaliação dos gestores:** formulação de painéis com os indicadores e os filtros representados em forma de gráfico para tornar mais fácil o processo de avaliação com os gestores da organização.

Diante da revisão realizada, Anderson e Joaquim iniciaram o processo a partir da definição das perguntas estratégicas da organização. A primeira reflexão foi: aonde quero chegar com o meu negócio? Depois de pensar bem, Anderson chegou à conclusão de que o seu objetivo comercial é o de ter clientes em todo o Brasil. Assim, pode abrir lojas físicas em cidades estratégicas para atender a todo o território nacional e, com isso, dobrar o seu faturamento. Com os objetivos bem definidos, foram formuladas as seguintes perguntas estratégicas:

- Como estão distribuídos os meus clientes pelo Brasil?
- Devo abrir lojas físicas em quais cidades para atender o meu cliente cada vez melhor?
- Qual é o mix de produto ideal para o negócio?

Depois de levantadas as perguntas estratégicas, é necessário **identificar** quais **indicadores** podem auxiliar na resposta às perguntas, a fim de ajudar os gestores a saberem o quanto estão se aproximando dos seus objetivos. Nesse sentido, cada pergunta estratégica foi analisada isoladamente e, na sequência, para cada indicador, houve a **identificação dos possíveis filtros**:

- Como estão distribuídos os meus clientes pelo Brasil?
  - ◊ Percentual de clientes por estado e cidades.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade.
  - ◊ Quantidade de novos clientes por estado e cidades.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade.
  - ◊ Quantidade de clientes por estado e cidades.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade.

- Em quais cidades devo abrir lojas físicas para atender o meu cliente cada vez melhor?
  - ◊ Quantidade de clientes com, pelo menos, uma compra por mês por estado e cidade.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade.
  - ◊ Quantidade média de compras por cliente por estado e cidade.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade;
  - ◊ Quantidade de clientes com apenas uma compra por estado e cidade.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade.
  - ◊ Gasto médio dos clientes por estado e cidade.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade.
- Qual é o mix de produto ideal para o negócio?
  - ◊ Quantidade de vendas de produto por dia.
    - Filtros: produto, fabricante, tipo de produto, dia, mês, ano, estado, cidade e loja.
  - ◊ Quantidade de vendas de produto por estado e cidade.
    - Filtros: produto, fabricante, tipo de produto, dia, mês, ano, estado, cidade e loja.
  - ◊ Média de vendas de um produto por dia e por estado/cidade.
    - Filtros: produto, fabricante, tipo de produto, dia, mês, ano, estado, cidade e loja.

O processo de prototipagem dos painéis é um importante passo na fase de análise dos requisitos do negócio, pois é a partir dos protótipos que é possível avaliar o que foi especificado até o momento. Isso torna o trabalho de construção da solução mais fácil. Para esse processo, não é necessária a utilização de qualquer ferramenta específica: podem ser usados apenas papel e caneta ou gráficos elaborados em planilhas eletrônicas com dados fictícios para demonstrar as possíveis representações gráficas dos indicadores, assim como é demonstrado na figura a seguir:



Figura 5 - Exemplo de prototipação de painéis

**Descrição da Imagem:** a figura apresenta um protótipo de um painel. O seu objetivo não é apresentar dados reais, mas demonstrar como as informações podem ser apresentadas, ao deixar em evidência os elementos gráficos e visuais que podem ser utilizados para essa representação. Na figura, são expostos alguns indicadores em forma de cartões e a partir do uso de tabelas, gráficos de pizza, gráficos de área e de linha.

Quando se trata das soluções tecnológicas para BI, há o que é chamado de arquitetura típica de BI, que diz respeito a um conjunto de componentes com tarefas bem definidas, ligadas entre si e que respeitam um padrão do ciclo de vida do dado para a informação. É possível ter uma arquitetura típica de BI pelo fato de existirem características comuns de utilização das soluções de BI, o que permite a existência de um conjunto de componentes que atendam a essas características. O quadro a seguir apresenta mais detalhes:

Foco	Ambiente interno	Ambiente externo
Objetivos de análise	<ul style="list-style-type: none"> <li>• Operações do negócio.</li> <li>• Cadeia de suprimentos.</li> <li>• Gestão de relacionamento com os clientes.</li> <li>• Clientes e fornecedores.</li> </ul>	<ul style="list-style-type: none"> <li>• Segmentação, preferências e comportamentos dos clientes.</li> <li>• Economia.</li> <li>• Aspectos regulatórios.</li> <li>• Concorrência: <ul style="list-style-type: none"> <li>• segmentação;</li> <li>• líderes.</li> </ul> </li> <li>• Perfil de compra.</li> </ul>
Objetivos	<ul style="list-style-type: none"> <li>• Eficiência</li> </ul>	<ul style="list-style-type: none"> <li>• Posicionamento no mercado</li> </ul>
Utilização	<ul style="list-style-type: none"> <li>• Análise, refinamento e reengenharia do desempenho do mercado.</li> </ul>	<ul style="list-style-type: none"> <li>• Modelagem e previsão do comportamento do mercado.</li> <li>• Posicionamento no mercado.</li> <li>• Aprendizagem das tendências de consumo.</li> <li>• Identificação de riscos, tecnologias e regulação.</li> </ul>

Quadro 2 - Características de utilização das soluções de BI / Fonte: Sell (2006 apud CECI, 2012, p. 56).

Na arquitetura tradicional de BI, existem três componentes principais:

- **ETL (*extraction, transformation and loading*)**: o processo de ETL tem como responsabilidade **extrair** os dados dos sistemas de informações transacionais (sistemas ERPs, por exemplo), fazer toda a **transformação** necessária para tornar o dado confiável e realizar a **carga** do dado na camada do repositório de dados analíticos.
- **Repositório de dados analíticos**: os repositórios analíticos de dados são tradicionalmente desenvolvidos como *Data Warehouse*. Quando fazem parte de uma arquitetura de BI, geralmente, os dados organizados utilizam a modelagem de dados dimensional.
- **Visualizador de dados (DataViz)**: camada em que os usuários finais da solução de BI terão acesso. Geralmente, é desenvolvida em forma de painel ou *dashboards*. Tem como responsabilidade apresentar os indicadores para os tomadores de decisão, permitindo que eles possam combinar várias dimensões e aplicar vários tipos de filtros.

Existem outros elementos envolvidos na arquitetura tradicional de BI. A figura a seguir apresenta mais detalhes sobre os vários elementos que podem estar relacionados com essa arquitetura:

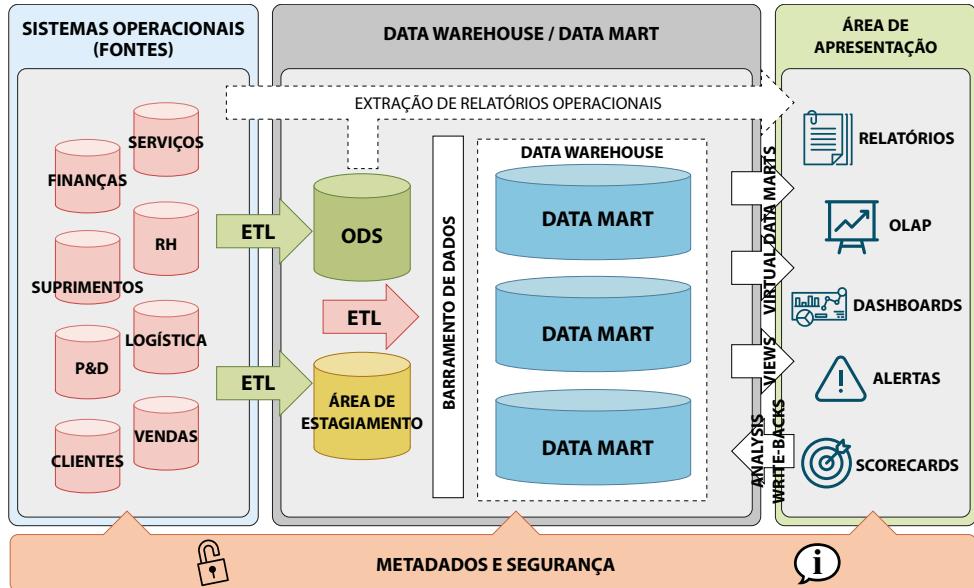


Figura 6 - Arquitetura tradicional de BI com os principais componentes  
Fonte: Silva (2011, p. 34).

**Descrição da Imagem:** a figura apresenta a arquitetura tradicional de BI. Na imagem, ela está dividida em três quadrantes: o que está à esquerda é o das fontes de dados dos sistemas operacionais. Nele, estão representados vários bancos de dados separados, um de cada área de negócio, tais como finanças, RH, suprimentos, logísticas, dentre outros. O primeiro quadrante está ligado ao segundo pelo processo de ETL. No segundo quadrante, encontram-se as bases de estagiamentos e o ODS, que, a partir de um barramento de dados, permite a adição, via processo de ETL, ao Data Warehouse, que é composto por vários Data Marts. Por fim, o segundo quadrante é ligado ao terceiro pelas informações que são consumidas. O terceiro quadrante está focado em ser a área de apresentação, que pode utilizar alguns recursos, tais como relatórios, OLAP, dashboards, alertas ou scorecards. Todos os quadrantes são apoiados por uma camada transversal de metadados e segurança.

Ao observarmos a Figura 6, no primeiro quadrante, estão os sistemas operacionais (fontes), que são os Sistemas de Informação que lidam com as operações e as transações do dia a dia. No contexto da empresa de Anderson, tratam-se dos sistemas de ERP e CRM. A área central representa o repositório de dados analíticos, que, tradicionalmente, é desenvolvido na forma de *Data Warehouse* (DW). Entretanto, ainda existem outras bases que podem ser utilizadas como apoio, tais como:

- **Áreas de estagiamento:** bancos de dados que podem ser utilizados para fazer um pré-processamento do dado ou facilitar o processo de transformação do dado em questão. São bases de apoio e não são consumidas por outros sistemas ou processos.
- **ODS (Operational Data Store):** é uma base de dados que centraliza os conceitos que podem estar distribuídos entre vários sistemas operacionais, de modo que o processo de geração de relatórios seja simples. Nessas bases, o dado está disponível na forma transacional. No caso de organizações que tenham apenas o sistema de ERP, esse tipo de base não é muito utilizado.

Para que os dados sejam migrados às bases dos sistemas operacionais para o *Data Warehouse*, é necessária a aplicação de ferramentas de ETL. Comumente, as ferramentas de ETL são desenvolvidas após a modelagem do *Data Warehouse*, tendo em vista que devem levar em consideração os modelos de dados das fontes de origem e de destino. A seguir, são detalhados os processos principais da **ETL**:

- **Extração:** o processo de extração consiste em ir até as fontes de origem, que podem estar disponíveis em um banco de dados ou em arquivos, ler o seu conteúdo e deixá-los disponíveis para as próximas etapas.
- **Transformação:** é o coração do processo de ETL. Nessa etapa, todos os dados extraídos das fontes de origem são tratados e limpos. Além do mais, é nessa fase que são aplicadas as regras de negócio para garantir a geração do novo dado ou informação relacionada às características do negócio. Por fim, o dado é disponibilizado e transformado para ser gravado (persistido) na fonte de dados do destino. No caso da arquitetura tradicional, diz respeito ao *Data Warehouse*.
- **Carga:** o processo de carga consiste em apanhar o dado ou informação já transformada e salvá-la nas fontes de dados de destino.

As ferramentas de ETL são muito utilizadas para a construção de soluções de BI e de *Big Data*. A etapa de transformação, quando vinculada à arquitetura tradicional de BI, tem algumas responsabilidades específicas e necessárias, pelo fato de se estar salvando os dados em um *Data Warehouse*. Um exemplo é o processo de padronização da informação, assim como é apresentado na figura a seguir:

Cliente	Sexo	Telefone
Cliente 1	Masculino	555-6532
Cliente 2	Feminino	555-4567

Cliente	Sexo	Telefone
Cliente 171	M	555-7485
Cliente 127	F	555-5478



Identificador	Sexo
M	Masculino
F	Feminino

Cliente	Sexo	Telefone
Cliente 781	0	555-7475
Cliente 947	1	555-7412

Figura 7 - Exemplo de padronização das informações / Fonte: Ceci (2012, p. 68).

**Descrição da Imagem:** a figura demonstra o processo de padronização de dados. Assim, são apresentadas três tabelas de origem. A primeira representa o sexo, que é dividido em "masculino" e em "feminino". A segunda representa o "M" ou "F". A terceira expressa "0" ou "1". Após o processo de padronização, é identificado "M" para "masculino" e "F" para "feminino". Todas as demais tabelas do Data Warehouse que utilizarem essa informação o farão a partir do que já padronizado.

Agora que já entendemos o que é e quais são as principais etapas do processo de ETL, aprofundemos o nosso entendimento sobre os **Data Warehouse (DW)**. Os DW são repositórios de dados analíticos que carregam características bem definidas, assim como é apresentado a seguir:

- **Base de dados integrada:** repositório único dos dados analíticos da organização. Pode ter domínios de aplicação distintos, mas todos podem ser cruzados e relacionados, permitindo uma visão sistêmica da organização.
- **Orientado por assunto:** os dados são organizados de acordo com a visão da empresa. Cada assunto é uma temática do domínio que será analisado, de tal forma que tudo se dá ao redor dos assuntos selecionados.
- **Volatilidade:** os dados que são adicionados ao *Data Warehouse* representam dados consolidados e agregados referentes a uma unidade de tempo (a ser definida pelo projeto). O valor adicionado não pode sofrer alteração, ou seja, um dado já inserido não pode sofrer alteração ao longo do tempo.

- **Variável com o tempo:** o tempo é uma das dimensões mais importantes quando se trabalha com DW. É muito comum fazer uma analogia com fotografias, ou seja, o DW é composto por um conjunto de fotografias que retratam as situações do domínio de aplicação distribuídas pelo tempo. A Figura 8 apresenta mais detalhes sobre essa característica.
- **Integração:** processo de integração das várias bases de dados. Realiza os devidos tratamentos ao aplicar rotinas de qualidade de dados e ao fazer as transformações e os processamentos necessários.

Cliente	Valor venda	Data
Cliente 1	1.000,00	08/02/2012
Cliente 18	500,00	08/02/2012
Cliente 47	375,00	08/02/2012
Cliente 12	745,00	08/02/2012
Cliente 1	524,00	09/02/2012
Cliente 13	85,00	09/02/2012
Cliente 53	45,00	10/02/2012



Data	Total Venda
08/02/2012	2.620,00
09/02/2012	609,00
10/02/2012	45,00

Figura 8: Analogia da “fotografia” no tempo do DW / Fonte: Ceci (2012, p. 67).

**Descrição da Imagem:** a figura apresenta um exemplo de como se pode entender a analogia das fotografias distribuídas pela dimensão “tempo” de um Data Warehouse. Assim, na tabela de origem, encontra-se todo o detalhamento de vendas por cliente por dia. Já na tabela destino, há apenas o valor total vendido (já calculado) por dia.

É perceptível que os dados relacionados com o cliente foram desconsiderados, ou seja, é feita uma agregação dos valores a nível do dia. Portanto, para cada dia, há um valor total de vendas. Pode-se dizer que o “grão” de informação tratado é a nível de dia. Entende-se como “grão”, a menor unidade de informação trabalhada no DW. Já o DW trabalha com dados consolidados em forma de fotografias distribuídas pelo tempo. Em um sistema de informação operacional, o grão é a nível de transação. Por isso, eles também são chamados de sistemas transacionais.

Os DWs são tradicionalmente desenvolvidos a partir de um tipo de modelagem de dados chamada de **modelagem dimensional**. A modelagem dimensional, tradicionalmente, utiliza bancos de dados relacionais e apresenta uma forma diferente de organização dos dados, ao usar os conceitos de dimensões, fatos e medidas para tornar o acesso e o consumo dos dados já consolidados mais versáteis. Sell (2006) faz uma analogia entre a modelagem dimensional e um cubo, que é composto por três ou mais dimensões que podem ser definidas de acordo com os elementos do negócio. A modelagem dimensional também é conhecida como modelo estrela. O modelo estrela ganhou esse nome pelo fato de, esteticamente, as tabelas estarem organizadas em um formato parecido com uma estrela: no centro, encontram-se uma ou várias tabelas fato e, em volta, estão as tabelas de dimensões.

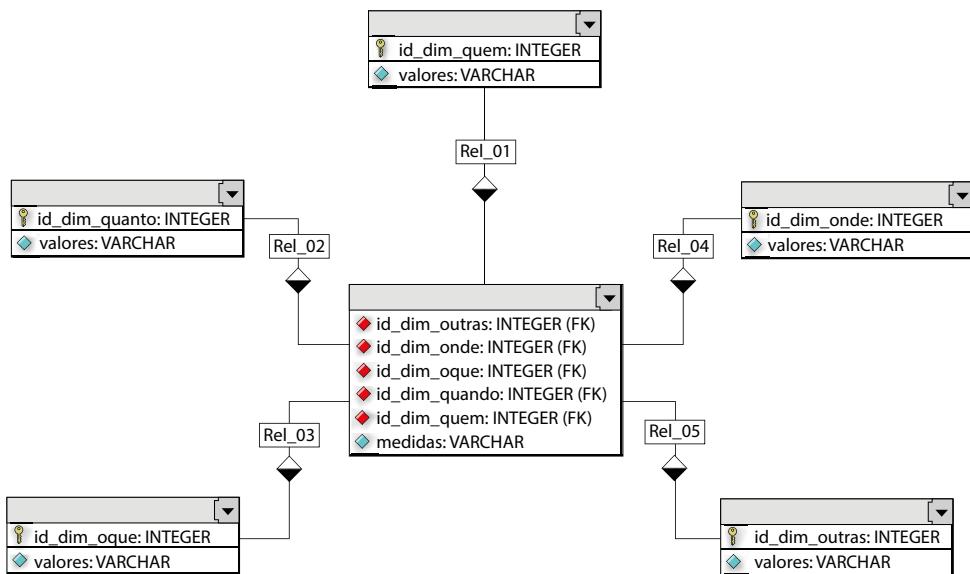


Figura 9 - Exemplo de modelo estrela / Fonte: Ceci (2012, p. 83).

**Descrição da Imagem:** a figura apresenta como se dá a organização de um modelo dimensional. Assim, há várias tabelas que representam as dimensões em questão. Geralmente, as tabelas respondendo às perguntas: o quê, quando, quem e onde. Para cada tabela de dimensão, há uma chave primária e, na sequência, os valores que representam a dimensão em questão. Ao centro da figura, está a tabela fato, que está diretamente ligada a um assunto. Ela é composta por todas as chaves primárias das dimensões relacionadas em forma de uma chave primária composta e tem, como demais campos, todas as medidas (sumarizadas) que foram calculadas.

Os DWs são orientados a assuntos e cada assunto será convertido em uma **tabela fato**. Essa tabela tem **medidas**, que são os valores numéricos que somam, totalizam ou consolidam os valores que estão organizados e distribuídos em **dimensões**. As tabelas permitem fazer filtros e combinações entre as visões para as medidas consolidadas dentro das tabelas fato.

De acordo com Ceci (2012), as tabelas de dimensões geralmente estão ligadas a quatro perguntas básicas: quando? Quem? Onde? O quê? Para complementar o seu entendimento sobre os fatos, as dimensões e as medidas, observe o quadro a seguir:

	<b>Fato</b>	<b>Dimensões</b>	<b>Medidas</b>
Escopo	Representam um item, uma transação ou um evento de negócio.	Determinam o contexto de um assunto de negócios, como por exemplo, uma análise da produtividade dos grupos de pesquisa.	São os atributos numéricos que representam um fato e são determinados pela combinação das dimensões que participaram desse fato.
Objetivo	Refletem a evolução dos negócios.	São os balizadores de análise de dados.	Representam o desempenho de um indicador de negócios relativo às dimensões que participam de um fato.
Tipo de Dado	São representados por conjuntos de valores numéricos (medidas) que variam ao longo do tempo.	Normalmente não possuem atributos numéricos, pois são somente descritivas e classificatórias dos elementos que participam de um fato.	Podem possuir uma hierarquia de composição de seu valor.

Quadro 3 - Tabela descritiva de fatos, medidas e dimensões / Fonte: Sell (2006, p. 31).

Ao finalizar os estudos sobre ETL, DW e modelagem dimensional, Joaquim iniciou o processo de modelagem dimensional para o DW em relação à primeira pergunta estratégica levantada para seus indicadores e filtros, assim como é apresentado a seguir:

- Como estão distribuídos os meus clientes pelo Brasil?
  - ◊ Percentual de clientes por estado e cidades.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade.
  - ◊ Quantidade de novos clientes por estado e cidades.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade.
  - ◊ Quantidade de clientes por estado e cidades.
    - Filtros: idade, sexo, profissão, loja da última compra, estado e cidade;

Joaquim e Anderson identificaram que o assunto principal da pergunta estratégica é cliente, logo, tem-se uma tabela: **fato\_cliente**. Outro aspecto observado é o de que a medida a ser trabalhada é sempre relacionada a quantidade de clientes, logo, tem-se a medida **quantidade** na tabela fato\_cliente. As dimensões trabalhadas para fazer os cruzamentos e filtros são: estado, cidade, sexo, profissão, loja e idade. Nesse sentido, essas serão as dimensões do modelo dimensional do DW referente à primeira pergunta estratégica. A Figura 10 apresenta o modelo dimensional desenvolvido:

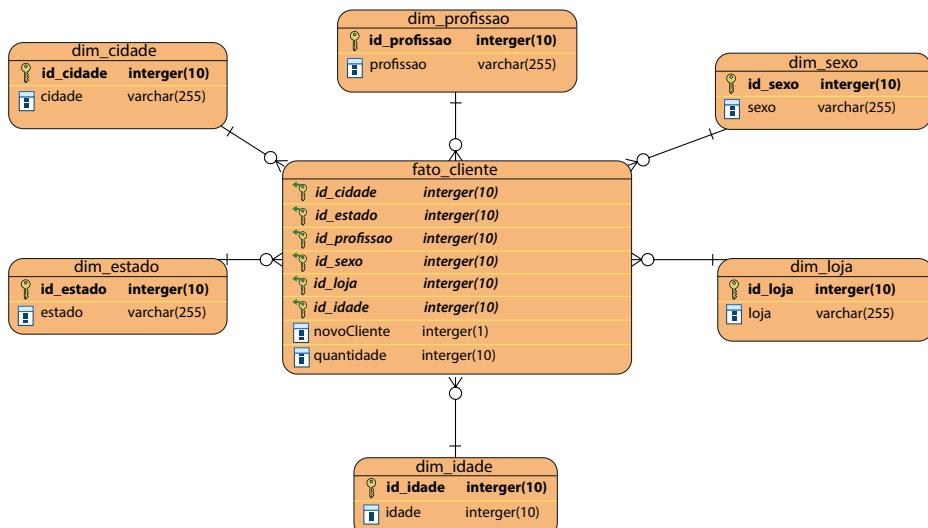
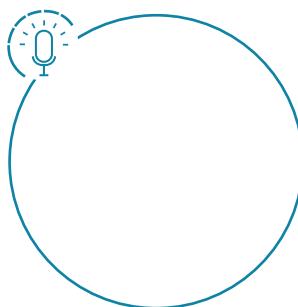


Figura 10 - Modelagem dimensional desenvolvida / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta a representação de uma modelagem dimensional. Há as seguintes dimensões envolvidas: cidade, estado, profissão, sexo, loja e idade. Nessas tabelas, existe apenas uma chave primária do tipo INTEGER e um valor que representa a descrição de cada uma das dimensões. A tabela fato, cujo assunto é cliente e está ao centro da figura, tem todas as chaves primárias das dimensões relacionadas, além de um valor numérico que determina se o cliente é novo ou não. Esses são os campos que fazem parte da chave primária composta. Por fim, há o campo quantidade do tipo INTEGER, que apresenta as quantidades de clientes que respeitam a combinação das dimensões em questão.

A partir da modelagem dimensional proposta, foi possível desenvolver as rotinas de ETL para fazer a carga das fontes de origem para o DW (que utiliza essa modelagem). Depois de o processo de carga ser implementado, foi realizado um *dashboard* para se obter a visualização dos indicadores propostos (Figura 5).

O *dashboard* está ligado à última parte da arquitetura tradicional de BI e tem como foco a visualização e a apresentação das informações geradas. É a partir dessa etapa que os usuários podem consumir as informações.



Quando falamos do processo de tomada de decisão, há as soluções de Business Intelligence (BI), que representam recursos importantes para que os gestores sejam apoiados em seu processo decisório. Como será que está a demanda por profissionais dessa área? O que é necessário para se tornar um analista de BI? Quais são os principais desafios que essa área tem? Essas e outras questões serão abordadas em nosso podcast! Não deixe de escutá-lo!



Tivemos a oportunidade de conhecer os Sistemas de Informação e saber como eles podem apoiar as organizações em seus processos internos e na captura e uso dos dados. Conhecemos Anderson, um jovem empreendedor que se viu desafiado pelo fato de não ter as ferramentas adequadas para gerir o seu negócio.

Anderson estava com uma empresa que era gerida por meio de controles baseados em planilhas eletrônicas. À medida que o seu negócio foi crescendo, ele percebeu a ausência de uma ferramenta que apresentasse uma visão unificada das várias áreas e que armazenasse esses dados de modo que fossem facilmente recuperados. A solução adotada foi a implantação de um sistema ERP, que atende todos os requisitos e gera visão de futuro da empresa.

Um cliente chamado Joaquim sugeriu a Anderson que buscasse um sistema ERP que liberasse o acesso às suas bases de dados para que elas possam ser utilizadas por outros sistemas. Essa é uma prática bastante utilizada pelas organizações.

Anderson, ao ver o seu negócio ser ampliado, sentiu a necessidade de ter recursos para apoiar os processos de fidelização e relacionamento com seus clientes. Assim, foi implantada uma solução de CRM, que permitiu gerenciar todos os processos de comunicação e relacionamento com os clientes. Por fim, à medida que o negócio tomou proporções ainda maiores, foi necessário desenvolver uma solução de *Business Intelligence* (BI), que permitiu a criação de indicadores para avaliar a condução do seu negócio, a eficiência da estratégia organizacional e a distância dos seus objetivos enquanto negócio.

Essa combinação de soluções é bastante comum em novas empresas baseadas em dados e em instituições de médio e grande porte. Esses são os primeiros passos para se ter uma empresa dirigida por dados. Sem uma base de dados interna acessível e organizada, e indicadores estratégicos bem definidos, não há motivos para investir em uma abordagem mais complexa ou baseada em *Big Data*. Por isso, o caminho percorrido pela empresa de Anderson foi correto até este momento.



# AGORA É COM VOCÊ



- Leia a descrição a seguir:

São sistemas que apresentam uma visão unificada dos dados transacionais das áreas da empresa, permitindo uma visão sistêmica e integrada.

Assinale a alternativa que apresenta o elemento ao qual a descrição faz menção:

- a) CRM.
- b) Sistemas operacionais.
- c) Mineração de dados.
- d) BI.
- e) ERP.

- A camada gerencial de uma organização necessita do apoio de dados e de informações para o processo de tomada de decisão.

Qual alternativa apresenta a aplicabilidade principal dos Sistemas de Apoio à Decisão (SAD)?

- a) Os SAD são utilizados principalmente para apoiar a decisão robótica em sistemas de automação.
- b) Os SAD são a base para os sistemas transacionais de controle de estoque, por exemplo.
- c) Os SAD são utilizados para a construção de Sistemas Operacionais.
- d) Os SAD são sistemas teóricos e impossíveis de serem desenvolvidos.
- e) Os SAD são mais encontrados no nível estratégico da organização, mas atualmente são utilizados em todos os níveis.

- O *Business Intelligence* (BI) representa uma importante solução para que as organizações possam acompanhar os seus indicadores mais de perto.

Assinale a alternativa que não se relaciona com o conceito de *Business Intelligence* (BI):

- a) É um sistema para apoio à decisão. Um exemplo é o PowerBI.
- b) Fornece KPIs para a camada tomadora de decisão.
- c) Pode utilizar *dashboards* para apoiar o processo de visualização.
- d) Pode utilizar a modelagem dimensional para apoiar a construção das suas bases.
- e) Auxilia o processo de tomada de decisão.



4. Os dados utilizados pelas soluções de BI, em sua grande maioria, fazem uso de processos para a limpeza e para a transformação, de modo que os indicadores sejam sempre confiáveis.

Assinale a alternativa correta em relação às bases de dados utilizadas nas soluções de BI:

- a) São implementadas exclusivamente em MySQL.
- b) As bases de dados utilizadas pelas soluções de BI são baseadas na modelagem dimensional.
- c) São desenvolvidas com base em arquivos texto criptografado.
- d) São prioritariamente desenvolvidos em solução NoSQL.
- e) Precisam de uma infraestrutura de Big Data para apoiar o processo de BI.

5. O processo de ETL está diretamente ligado ao sucesso de um *Data Warehouse*.

Assinale a alternativa que apresenta as principais etapas de um processo de ETL:

- a) Extração, tratamento e limpeza.
- b) Entrada, troca e liberação.
- c) Extração, transformação e carga.
- d) Entrada, transformação e carga.
- e) Encontrar, tratar e colar.

# CONFIRA SUAS RESPOSTAS



1. E.

A explicação se encaixa perfeitamente nos ERPs.

2. E.

Os SADs foram construídos inicialmente para apoiar a camada tomadora de decisão, ao garantir que tivessem os principais indicadores para guiar as suas decisões. No entanto, hoje, são utilizados para todos os níveis, pois trazem informações sumarizadas para apoio, inclusive a operação.

3. A.

O Business Intelligence (BI) não se trata de uma tecnologia específica, mas um conjunto de técnicas, metodologias e tecnologias que auxiliam no processo de tomada de decisão inteligente.

4. B.

Na arquitetura tradicional de uma solução de Business Intelligence (BI), é utilizada a modelagem dimensional como a base para as bases de dados.

5. C.

O processo de ETL é o responsável por extrair os dados das fontes transacionais, transformá-los e, depois, carregá-los às bases dimensionais.

# REFERÊNCIAS



ALVES, E. B. **Sistemas de Informações em Marketing:** uma visão 360 das informações mercadológicas. Curitiba: InterSaberes, 2018.

BRETERNITZ, V. J. A seleção de sistemas ERP (Enterprise Resource Planning) para pequenas e médias empresas. **Revista Análise**, v. 5, n. 10, p. 57-71, 2004.

CASTRO, L. N. de; FERRARI, D. G. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CECI, F. **Business Intelligence.** Palhoça: Editora Unisul Virtual, 2012.

FIALHO, F. A. P. et al. **Gestão do conhecimento e aprendizagem:** as estratégias competitivas da sociedade pós-industrial. Florianópolis: Visual Books, 2006.

GONÇALVES, G.; LIMA, I. A. de. Implantação de um sistema de informação - Enterprise Resource Planning (ERP): estudo de caso em uma indústria eletrônica. **Revista de Engenharia e Tecnologia**, v. 2, n. 1, p. 57-68, 2010.

GOUVEIA, L. B.; RANITO, J. **Sistemas de informação de apoio à gestão.** Porto: Sociedade Portuguesa de Inovação, 2004.

LAUDON, K. C.; LAUDON, J. P. **Gerenciamento de sistemas de informação.** Rio de Janeiro: LTC, 2001.

MELO, I. S. **Administração de sistemas de informação.** São Paulo: Pioneira, 2002.

OLIVEIRA, A. L. B. de; CARREIRA, M. L.; MORETI, T. M. Aprimorando a gestão de negócios com a utilização de tecnologia de informação. **Revista de Ciências Gerenciais**, v. 13, n. 17, p. 141-159, 2009.

PINHEIRO, C. A. R. **Inteligência analítica:** mineração de dados e descoberta de conhecimento. Rio de Janeiro: Ciência Moderna, 2008.

REZENDE, D. A. **Planejamento de sistemas de informação e informática:** guia prático para planejar a tecnologia da informação integrada ao planejamento estratégico das organizações. São Paulo: Atlas, 2011.

SELL, D. **Uma arquitetura para business intelligence baseada em tecnologias semânticas para suporte a aplicações analíticas.** 2006. Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Catarina, Florianópolis, 2006.

SILVA, A.; RIBEIRO, A.; RODRIGUES, L. **Sistemas de informação na administração pública:** modelos em UML. Rio de Janeiro: Renavan, 2004.

SILVA, D. C. da. **Uma arquitetura de business intelligence para processamento analítico baseado em tecnologias semânticas e em linguagem natural.** 2011. Dissertação (Mestrado em Engenharia do Conhecimento) – Universidade Federal de Santa Catarina, Florianópolis, 2011.





# Introdução à Ciência de Dados

Dr. Flávio Ceci

## OPORTUNIDADES DE APRENDIZAGEM

Esta unidade objetiva apresentar uma visão introdutória sobre a ciência de dados. Assim, serão evidenciadas as principais áreas envolvidas e esclarecidos os principais métodos e técnicas disponíveis para uso. Também estudaremos a importância da Governança de Dados (GD), ao reconhecermos a sua aplicabilidade e os seus principais desafios e soluções. Por fim, conheceremos os modelos de maturidade analítica relacionados à Governança de Dados e os seus principais níveis.

O projeto de implantação de uma solução de *Business Intelligence* (BI) para a empresa de Anderson foi um sucesso. Após a conclusão do projeto, foram disponibilizados os principais indicadores relacionados às áreas de negócio, o que permitiu que os gestores tomassem suas decisões com base em dados.

Diante disso, a empresa cresceu totalmente alinhada à estratégia organizacional proposta. Isso permitiu um mapeamento rápido dos resultados de várias ações e, em muitos casos, foram tomadas decisões com base nos indicadores em tempo real. Já se passaram seis meses desde a conclusão da implantação do projeto de BI e os resultados foram melhores que o esperado. Além disso, as perguntas estratégicas foram imprescindíveis para o norteamento dos passos da organização, que se aproximaram de todos os objetivos definidos pelos gestores.

A implantação do projeto de *Customer Relationship Management* (CRM) permitiu gerenciar o processo de relacionamento com os clientes, a fim de garantir que eles recebam e-mails promocionais em datas comemorativas, como em seus aniversários. Também foi assegurado o acionamento da empresa quando os seus clientes ficarem por um período sem comprarem na loja. Além do mais, em conjunto com a solução de BI, foi possível manter um canal ativo com o cliente e entender melhor os padrões de compras por região, a fim de que sejam criadas campanhas específicas. Tanto a solução de BI quanto a de CRM só foram possíveis pelo fato de haver um processo de coleta de dados padronizado, que, nesse caso, é feito por intermédio do sistema ERP.

Anderson está muito feliz com o resultado de sua empresa e é um grande defensor do uso de dados e de informações para a transformação e o alavancamento dos negócios. Desde a sua primeira ação, com o objetivo de migrar do controle das planilhas eletrônicas para a utilização dos Sistemas de Informação, a sua empresa só cresceu e chegou a um tamanho inimaginável. Evidentemente, Anderson sempre está pensando em qual passo deve dar em sua organização. Na atualidade, é CEO da empresa e faz realmente o trabalho de executivo. Além disso, há alguns gerentes que o auxiliam em várias áreas de negócio. Em consequência de toda a evolução realizada nos canais digitais, foi contratada uma nova gerente de tecnologia e desenvolvimento. Ela se chama Lara e está otimizando todos os processos de tecnologias, formando equipes e organizando a infraestrutura.

Certo dia, uma senhora entrou na loja matriz e pediu para conversar com Anderson. Ao chegar, ficou surpreso, pois não a conhecia. Ela disse que era a mãe do Joaquim e explicou que ele se formaria naquele final de semana e precisava comprar cordas para a sua guitarra, já que tocaria na formatura com a sua banda. Depois que o projeto de BI foi finalizado, Joaquim focou em finalizar o seu curso de graduação, por isso, não continuou atuando na empresa com Anderson. No entanto, enviou um convite de formatura pela sua mãe. A mãe do Joaquim não sabia qual corda o filho utilizava, mas sabia que ele comprava três jogos de corda todo mês e comentou esse fato com Anderson, que rapidamente foi até o sistema e, em poucos segundos, já sabia qual era o jogo de corda utilizado. Todavia, houve uma surpresa: o jogo de corda não estava em estoque!

Depois de 15 segundos em silêncio, Anderson, envergonhado, explicou para a mãe de Joaquim que, infelizmente, estava sem o jogo de cordas em estoque, mas que ela poderia adquiri-lo em uma loja da concorrência que estava próxima. Anderson anotou o nome do jogo de cordas e o endereço da outra loja. Para ele, esse fato foi marcante: como aconteceu algo assim, ainda mais com um cliente tão antigo como o Joaquim? O que adianta realizar um excelente trabalho em relação ao relacionamento com os clientes, se o mix de produto ainda não está adequado por lojas? Quantos clientes deixaram de comprar em suas lojas físicas pelo fato de não haver o produto esperado?

Durante a formatura de Joaquim, Anderson pensou em como seria importante ter dois recursos analíticos para o seu negócio:

- Saber quais clientes estão propensos a deixar de comprar em sua loja.
- Saber qual seria o mix ideal de produto por loja.

Joaquim foi até a mesa onde Anderson estava para conversar e perguntar o que ele achou do show de sua banda. Anderson deu os parabéns pela formatura e pelo excelente show que fizeram. Ele ficou muito orgulhoso de ter visto que a maioria dos instrumentos e equipamentos da banda de Joaquim foram fornecidos por ele. Diante disso, Anderson pediu desculpas a Joaquim no que diz respeito à falta do jogo de corda que ele sempre compra e falou sobre dois recursos analíticos que gostaria de ter em sua loja. Depois, perguntou se Joaquim conhecia algo que poderia ajudá-lo.



Joaquim adorou os questionamentos de Anderson e disse: “Existe um tipo de modelo estatístico muito utilizado, o chamado ‘Modelo de Churn’. Nesse modelo, é possível atribuir um *score* (valor numérico) em relação à propensão de um cliente deixar de comprar em sua loja naquele momento com base em seu histórico. Eu trabalhei na construção de um modelo desses quando era funcionário de uma empresa do setor financeiro. Tivemos excelentes resultados com a implantação dessa ferramenta. Sobre o segundo aspecto, a solução de BI implantada fornece uma ideia de mix de produtos por loja com base nos indicadores de compras feitas pelos clientes. É possível fazer uma análise exploratória e, depois, aplicar algumas técnicas, como a clusterização, a fim de gerar segmentações se pautando no perfil de compras e outras características dos usuários. Depois, é verificado quais segmentos estão mais presentes em cada loja”.

Anderson ficou feliz pelo fato de Joaquim ter uma ideia de como solucionar os aspectos que havia comentado e perguntou se Joaquim trabalharia com o pai dele. Ele explicou que ele e seu pai tem um acordo: Joaquim só trabalharia na empresa de seu pai quando já tivesse trabalhado em outras, a fim de ter experiências diferentes. Portanto, estava à procura de um emprego. Nesse momento, o Anderson disse: “Parabéns, você é o primeiro cientista de dados contratado em minha empresa! Você pode começar na próxima semana?”.

Joaquim ficou feliz com a proposta e, obviamente, aceitou! Ele já sabia que não faltavam oportunidades para profissionais da área de ciência de dados e, com ele, não foi diferente.

Você achou interessante o Modelo de Churn? Faça uma pesquisa na Internet e leia artigos para saber como funciona uma ferramenta como essa. Joaquim iniciará o seu trabalho como cientista de dados: quais são as principais habilidades e competências que ele deve ter para desempenhar um excelente trabalho em relação à sua profissão?

Faça uma pesquisa sobre o questionamento apresentado para que Joaquim esteja preparado para iniciar o seu trabalho. Uma maneira de planejar é saber quais pontos deverão ser desenvolvidos enquanto estiver ocupando a função que lhe foi determinada na empresa de Anderson.

A ciência de dados está diretamente ligada às áreas de negócio, visto que é uma fonte geradora de estudos, ferramentas e suporte analítico. O cientista de dados deve ser uma pessoa curiosa e inquieta, pois precisa querer entender os dados em detalhes e compreender as correlações que existem entre eles. Para isso, deve ter conhecimento sobre matemática/estatística, computação e negócio (alguns autores falam em ciência). Dentro da área, ele também pode decidir se deseja aprofundar os seus conhecimentos em subáreas vinculadas aos três pilares mencionados. Esses são os aspectos que nortearão o desenvolvimento desta unidade.

## DIÁRIO DE BORDO

## O que é ciência de dados?

A expressão “ciência de dados” não é recente. Existem registros do seu uso a partir da década de 60. Contudo, a sua conotação obtida nos últimos anos é completamente diferente da inicial. Para compreender o que é ciência de dados, é necessário, primeiramente, entender o que é ciência. Segundo o Dicionário Michaelis, “ciência” é um “conhecimento sistematizado como campo de estudo” e um “conjunto de conhecimentos teóricos e práticos canalizados para um determinado ramo de atividade” (MICHAELIS, [2021], on-line)<sup>1</sup>.

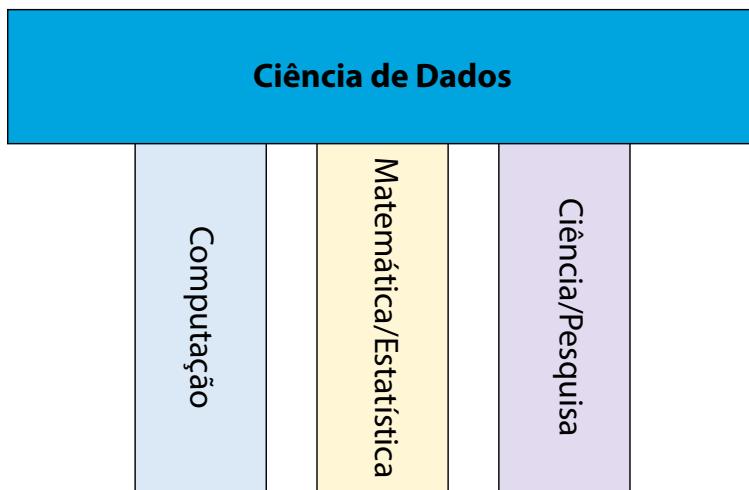
Dante do exposto, é perceptível que o termo “ciência” está totalmente relacionado ao contexto proposto pela ciência de dados, que objetiva agrupar e relacionar os conhecimentos de várias áreas levando em consideração todo o ciclo de vida do dado, que pode ser observado a seguir:

- **Captura:** estrutura as formas e os métodos de captura de dados de maneira estruturada, semi ou não estruturada, a fim de que sejam utilizados em seus estudos.
- **Tratamento:** realiza transformações e cruzamentos necessários para tornar o dado passível de ser analisado posteriormente.
- **Armazenamento:** refere-se à persistência do dado, que pode ter passado pela etapa de tratamento, ou não, em bases relacionais ou *Not only SQL* (NoSQL).
- **Processamento:** consiste na transformação do dado em uma informação a partir da aplicação de regras de negócio, por meio do uso de elementos de contextualização do dado ou por intermédio da construção de modelos, por exemplo.
- **Análise:** etapa em que são feitos cruzamentos, estudos e aplicação de técnicas para a mineração e a exploração de dados e informações.
- **Apresentação:** diz respeito ao modo com que as análises são apresentadas ou entregues para os usuários ou sistemas. Geralmente, são utilizados recursos visuais ou numéricos para apoiar a entrega das análises.
- **Descarte:** há uma falsa percepção de que, quanto mais dados armazenados, melhor. Esse fato não é verdade, tendo em vista que os dados que não são relevantes geram custos de armazenamento e processamento. Portanto, devem ser descartados.

Ao observar os processos apresentados, é fácil perceber o motivo pelo qual a área tem

o termo “ciência” em seu nome. Para cada processo, existe um conjunto de competências e habilidades interdisciplinares envolvidas, o que a torna uma área bastante abrangente e com possibilidades distintas de focos. Segundo Amaral (2016, p. 4), “normalmente, a Ciência de Dados é associada de forma equivocada apenas aos processos de análise de dados, onde com o uso de estatística, aprendizado de máquina ou a simples aplicação de filtro se produz informação e conhecimento”.

Para dar suporte a todos os processos, existem três pilares principais que fundamentam a ciência de dados. Eles são apresentados na figura a seguir:



**Figura 1 - Pilares da ciência de dados / Fonte: o autor.**

**Descrição da Imagem:** a figura apresenta um retângulo horizontal com o título “Ciência de dados”. Abaixo, estão outros três retângulos verticais, a fim de representar a ideia de serem os pilares que suportam a ciência de dados. Os retângulos verticais são: «Computação», «Matemática/Estatística» e «Ciência/Pesquisa».

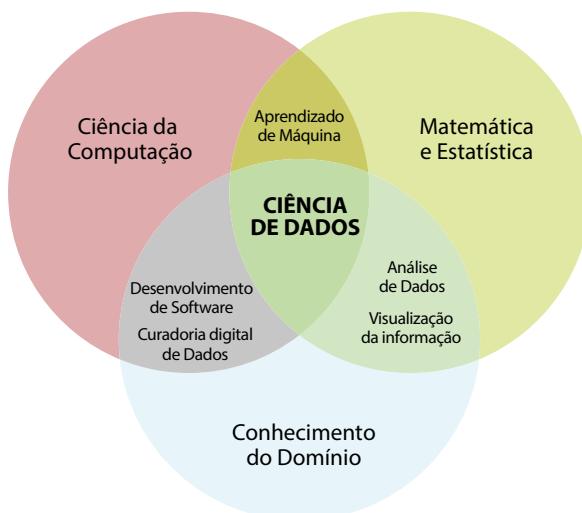
A ciência de dados utiliza muito os elementos da metodologia da pesquisa, muito utilizada pela ciência em geral. Nesse sentido, quando se trabalha com o processo de análise e, sobretudo, de definição e elaboração do estudo a ser feito, geralmente, é preciso estabelecer o problema ou a problemática, as hipóteses, as variáveis, as etapas metodológicas e os objetivos, para que os resultados obtidos tenham significância e relevância.

A matemática e a estatística apoiam todo o processo de transformação, processamento e, principalmente, análise do ciclo de dados. O apoio acontece desde o processo de amostragem e abrange a geração de escores (índices numéricos), a classificação,

a recomendação, o reconhecimento de padrões, a identificação de *outliers* (exceções e pontos fora de uma curva), dentre várias outras possibilidades.

Diariamente, são gerados e armazenados muitos dados nos repositórios das organizações. Assim, a computação disponibiliza todo o ferramental de apoio para que seja possível aplicar os métodos e técnicas necessários em um ambiente massivo de dados como o *Big Data*. A computação apoia todos os processos do ciclo de vida de dados, por isso, existem muitos profissionais que migraram dessa área para a ciência de dados, tendo em vista a vasta possibilidade de apoio aos processos de ciclo de dados.

Quando Joaquim falou de seu pai, comentou que, inicialmente, ele integrava a área da computação. Todavia, em um determinado momento de sua carreira, migrou para a ciência de dados. Isso também acontece com profissionais com formação em cursos, tais como Estatística, Economia, Matemática, Física e algumas engenharias. Quando nos referimos às habilidades necessárias para um especialista em ciência de dados, o uso da ciência já é algo natural. Portanto, é muito forte o conhecimento em relação ao negócio. Observe a figura a seguir:



**Figura 2 - Interdisciplinaridade da ciência de dados**

Fonte: Conaway (2010 apud RAUTENBERG; CARMO, 2019, p. 59).

**Descrição da Imagem:** a figura apresenta três círculos e há uma intersecção no centro. Os três círculos têm os seguintes títulos: "Ciência da Computação", "Matemática e Estatística" e "Conhecimento de Domínio". Na intersecção entre "Ciência da Computação" e "Matemática e Estatística", há o título "Aprendizado de Máquina". Já na intersecção entre "Ciência da Computação" e "Conhecimento do Domínio", há os títulos "Desenvolvimento de Software" e "Curadoria Digital de Dados". Na intersecção entre "Matemática e Estatística" e "Conhecimento de Domínio", há os títulos "Análise de Dados" e "Visualização da Informação". Por fim, na intersecção entre as três grandes áreas, há o termo "Ciência de Dados".

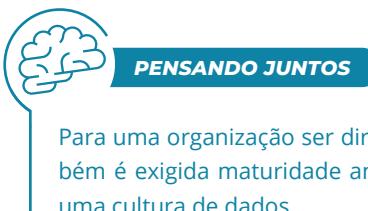
A visão exposta pela Figura 2 foi originalmente desenvolvida por Conaway (2010) e é muito interessante, pois demonstra como se dá a combinação entre as áreas e expõe quais são as intersecções entre as áreas e as disciplinas relacionadas. Um exemplo é o fato de que a aprendizagem de máquina se relaciona diretamente com a computação, a matemática e a estatística.

As empresas que têm uma área de dados ou ciência de dados realmente têm uma equipe multidisciplinar. No caso de Joaquim, ele é formado em Ciência de Dados, portanto, já é dotado dessa interdisciplinaridade construída durante a sua formação. Hoje, ele é um profissional mais versátil para a empresa de Anderson, pois, de fato, conhece um pouco sobre cada uma das áreas envolvidas. O que falta para Joaquim exercer o seu papel enquanto cientista de dados na empresa em questão é saber sobre o negócio. Em detrimento do fato de que é apaixonado por música, o processo é um pouco mais fácil. Entretanto, de qualquer forma, deve conhecer os elementos da estratégia organizacional, saber como se dá os principais processos das áreas do negócio e ter ciência de quais são os principais desafios e oportunidades já mapeados. Desse modo, Joaquim conseguirá exercer melhor o seu papel e gerará análises realmente relevantes para o negócio.

Em seu primeiro dia de trabalho, Joaquim pediu para que Anderson se aprofundasse em seu negócio, com o objetivo de preencher a lacuna que foi exposta no parágrafo anterior. Na perspectiva da organização da empresa, Joaquim foi inserido na área de tecnologia e desenvolvimento, que é gerenciada por Lara, uma cientista da computação com MBA em Gestão Estratégica e especialização em Engenharia de Projeto de Software. Ela era muito compreensiva e gostava bastante da área de dados. Depois de ter mergulhado nos processos e nos detalhes do domínio do negócio, Joaquim foi levado até a sua mesa. Em relação ao seu trabalho, estaria alocado na área de Lara, mas também responderia diretamente à Anderson, a fim de que pudesse ser acessado pela presidência e auxiliasse Lara nas questões relacionadas aos dados.

Esse arranjo foi interessante para a organização, pois potencializou a área da tecnologia e desenvolvimento, ao incluir uma pessoa preocupada com os dados, independentemente do projeto ou do contexto. Além disso, permitia que Joaquim acompanhasse as questões estratégicas e as principais dores do negócio, para que pudesse negociar com Lara quando atuaria em cada uma dessas demandas e projetos. O grande foco de Joaquim é o de tornar a empresa de Anderson totalmente dirigida por dados (*Data-Driven Organization*).

Segundo Anderson (2015), muitas organizações são dirigidas por dados, pelo fato de terem vários relatórios e a necessidade de tomada de decisões a partir deles. O mesmo processo ocorre nas instituições que carregam vários modelos de predição ou previsão, mas que não têm nenhuma relação com os seus processos de negócio. Nesse sentido, quando dizemos a respeito de uma organização dirigida por dados, estamos nos referindo à construção de ferramentas analíticas, habilidades e, principalmente, de uma cultura em que o foco esteja no dado.



Para uma organização ser dirigida por dados, não basta ter apenas tecnologia, mas também é exigida maturidade analítica por parte dos seus colaboradores e a existência de uma cultura de dados.

Segundo Rautenberg e Carmo (2019), em relação à perspectiva tecnológica da tomada de decisão com base em dados, o foco da ciência de dados está em apoiar os gestores em suas tarefas intensivas. Os referidos autores também detalham as tarefas que são potencializadas:

- **Associação:** geralmente, é associada a uma situação de causa e efeito. Também pode estar ligada à correlação entre dois eventos (sair para comprar fralda e cerveja).
- **Avaliação:** busca analisar um caso e apresentar as opções enquanto propostas de solução. Um exemplo é averiguar um perfil e verificar se concederá crédito ou não.
- **Diagnóstico:** objetiva identificar o estado de um objeto ou de uma situação para, em seguida, tomar uma decisão.
- **Monitoramento:** aplicação de diagnósticos de tempos em tempos (ciclos) para identificar evoluções e tomar decisões.
- **Predição:** estima um evento futuro a partir dos dados históricos armazenados.

Lara pediu para Joaquim fazer um levantamento dos principais métodos e técnicas que utilizará no seu dia a dia, para que fosse possível dimensionar os recursos computacionais necessários para executar as suas tarefas e automatizar as execuções em um ambiente de produção.

## Métodos e técnicas para a ciência de dados

Joaquim está muito empolgado com o seu trabalho e, principalmente, com o fato de ser o primeiro cientista de dados da empresa de Anderson. Assim, iniciou cedo o processo de organização dos principais métodos e técnicas que utilizará em seus estudos. Em consequência de todo o ciclo de dados ser muito amplo, Joaquim focará nas técnicas e nos métodos relacionados aos processos de tratamento, armazenamento e processamento quando for estruturar os processos relacionados ao *Big Data*, a fim de formular o processo de apresentação e revisitar as ferramentas e as técnicas de visualização de dados.

O desafio atual está no processo de análise dos dados já armazenados nas bases transacionais e analíticas, que é o foco do levantamento a ser feito. Nesse contexto, Joaquim está organizando o seu levantamento a partir de três tipos de análise de dados: análise exploratória, análise implícita e análise explícita.

Quando se fala em análise explícita, o foco está em averiguar os dados que já estão minimamente explicitados a partir de *Data Warehouses*, de bases transacionais (como os sistemas de operação ou ERPs), dos painéis e *dashboards* disponíveis, por exemplo. Joaquim comentou que, basicamente, precisaria ter acesso ao banco de dados e, a partir das consultas com o uso do *Structured Query Language* (SQL), (ver Figura 3) realizaria a análise dos dados. Outro aspecto seria utilizar os *dashboards* criados em *Data Studio* para analisar os dados com a aplicação de alguns filtros e navegar nas informações armazenadas nos *Data Warehouse*. Para fazer a análise explícita, não é necessário adicionar nenhuma tecnologia nova ou processo da organização, o que permite que sejam feitos os primeiros estudos.

The screenshot shows a PostgreSQL IDE interface. The top pane contains several SQL statements:

```

<PostgreSQL - AWS - Akropolis> Script x experience class_driver_name
select max(datatime) from timeline;
select * from cliente_lgpd;
select * from category order by id_category, id_father;
select * from item;
select id_item, c.name as category, i.name, i.properties from item i
join category c on (i.id_category = c.id_category);
SELECT id_experience, id_category, id_item, environment, class_driver_name
from item;
SELECT e.id_experience, e.id_category, c.name , e.id_item, environment,
       e.environment, e.class_driver_name
FROM item e
JOIN category c ON (e.id_category = c.id_category)
JOIN experience x ON (e.id_experience = x.id_experience)
JOIN driver d ON (x.class_driver_name = d.name)
WHERE e.id_item = 1
      
```

The bottom pane displays the results of the last query in a table format:

	id_item	category	name	properties
1	1	Alerta	Recebou SMS	[NULL]
2	2	Transação Financeira	Recebimento de Fornecedor	[NULL]
3	3	Transação Financeira	Compra com cartão Mastercard	[NULL]
4	4	Transação Financeira	Compra com cartão Visa	[NULL]
5	5	Promocional	Recebou SMS	[NULL]
6	6	Recomendação	Recebou SMS	[NULL]
7	7	Pesquisa de satisfação	Recebou SMS	[NULL]
8	8	NPS	Recebou SMS	[NULL]
9	9	Problema	Abriu e-mail	[NULL]

Details at the bottom of the interface include: Record: 1 / 9, Rows: 1, 9 row(s) fetched - 260ms, BRT pt\_BR Gravável, Inserção inteligente, 40 : 1 [122], Sel: 122 | 2.

Figura 3 - Exemplo de IDE para interação com banco de dados / Fonte: o autor.

**Descrição da Imagem:** a figura demonstra uma ferramenta de interação com um banco de dados. Na parte superior, podem ser inseridas as instruções SQL, enquanto, na parte inferior, é apresentado o resultado em forma de tabela.

A partir de algumas ferramentas, tais como o DBeaver, Joaquim pôde se conectar a bases de dados distintas (Postgres, MySQL, SQLServer e Oracle, por exemplo) e, a partir de instruções SQL, pôde cruzar dados e informações, de tal forma que seja fácil explicar alguns fenômenos que podem não estar tão explícitos em indicadores.



As bases de dados apresentadas (Postgres e MySQL) são exemplos de Sistemas Gerenciadores de Banco de Dados (SGBD). Tratam-se de sistemas de armazenamento de dados em forma de entidade e relacionamentos que são disponibilizados em tabelas. Hoje, representam o principal tipo de banco de dados usado e é a forma mais indicada para os apoiar sistemas de informações transacionais e operacionais.

Fonte: o autor.

O processo de análise com base em indicadores também pode ser feito a partir dos *dashboards* disponíveis na organização ou por intermédio de consultas, com a aplicação de filtros e cruzamentos nas bases do *Data Warehouse*. Nos *dashboards*, é possível navegar nos indicadores de maneira gráfica e com o uso de gráficos e séries históricas, a fim de facilitar os processos de análise. Esse recurso permite que analistas que não tenham tanto conhecimento técnico possam fazer as suas análises sem a necessidade de conhecer o banco de dados ou a linguagem SQL. A figura a seguir apresenta um exemplo de um *dashboard* utilizado para fazer análises explícitas:



Figura 4 - Exemplo de *dashboard* / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta a visão de um painel estilo *dashboard*. Nele, há vários gráficos que apresentam visões distintas para os indicadores relacionados a um mesmo domínio.

A análise exploratória de dados objetiva conhecer os dados antes de analisá-los, verificando as suas propriedades numéricas e a sua distribuição ou dispersão para, depois, aplicar as análises implícitas e explícitas. Quando nos referimos às propriedades numéricas (ou técnicas quantitativas), estamos falando das medidas de dispersão e posição, tais como média, mediana, amplitude e desvio padrão (AMARAL, 2016). Para esse tipo de análise, Joaquim deve importar os seus *datasets* (conjuntos de dados) para trabalhar a partir de uma linguagem como R ou Python. Assim, consegue ter muito facilmente as informações para cada uma das variáveis envolvidas.

A análise exploratória utiliza muito os recursos visuais para auxiliar no entendimento das características do dado em questão. Nesse contexto, um recurso muito utilizado é o gráfico de dispersão, que, pela sua natureza, projeta os dados em dois eixos (x e y). Dessa forma, demonstra o quanto dispersa ou unida está a distribuição dos dados e evidencia se existe algum tipo de padrão de agrupamento. A figura a seguir apresenta um exemplo de gráfico de dispersão:

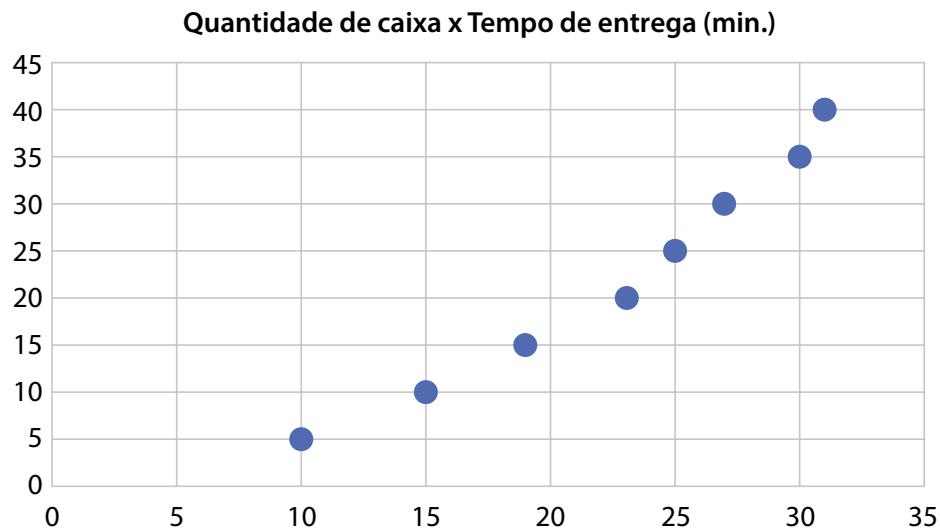


Figura 5 - Exemplo de gráfico de dispersão / Fonte: o autor.

**Descrição da Imagem:** a figura exemplifica um gráfico de dispersão, que objetiva apresentar os valores distribuídos por um plano cartesiano em conjunto com os dados distribuídos por eixos verticais e horizontais. Na figura, é demonstrada a relação do tempo de entrega pela quantidade de caixas. O gráfico evidencia que, quanto mais caixas, mais tempo é utilizado. No entanto, à medida que o número de caixas aumenta, a razão da progressão diminui, ou seja, o valor aumentado sofre uma redução gradativa.

Em seu levantamento, até o presente momento, Joaquim está calmo, pois, para desenvolver os seus estudos com base nas técnicas e nos métodos apresentados, não foi necessária nenhuma ferramenta robusta, um supercomputador ou até mesmo a criação de um ambiente na nuvem para trabalhar com dados massivos. O gráfico de dispersão apresentado na Figura 5 foi desenvolvido pela planilha eletrônica Excel, mas poderia ser gerado por meio das linguagens R, Python ou Java. Joaquim comen-

tou com Lara que acredita que, nesse estágio da empresa, seja importante que os seus colaboradores tenham acesso às ferramentas de planilha eletrônica, uma vez que é uma forma de democratizar a tomada de decisão por dados, estimular, evoluir a cultura de dados da organização e dar liberdade para as áreas de negócio fazerem as suas análises em relação ao que não está materializado em indicadores em soluções de BIs.

Ainda sobre os recursos visuais que podem auxiliar no processo de análise exploratória, há os diagramas de caixa e os histogramas. No caso do diagrama de caixa, é possível observar como os dados estão distribuídos nos quatro quartis, averiguar o valor da mediana e conhecer os dados que fogem do intervalo esperado (*outliers*). Esses gráficos são muito utilizados para acompanhar as evoluções das ações na bolsa de valores, que, em muitos casos, omite as informações dos quartis e da mediana. A figura a seguir apresenta um exemplo de uso do diagrama de caixa:

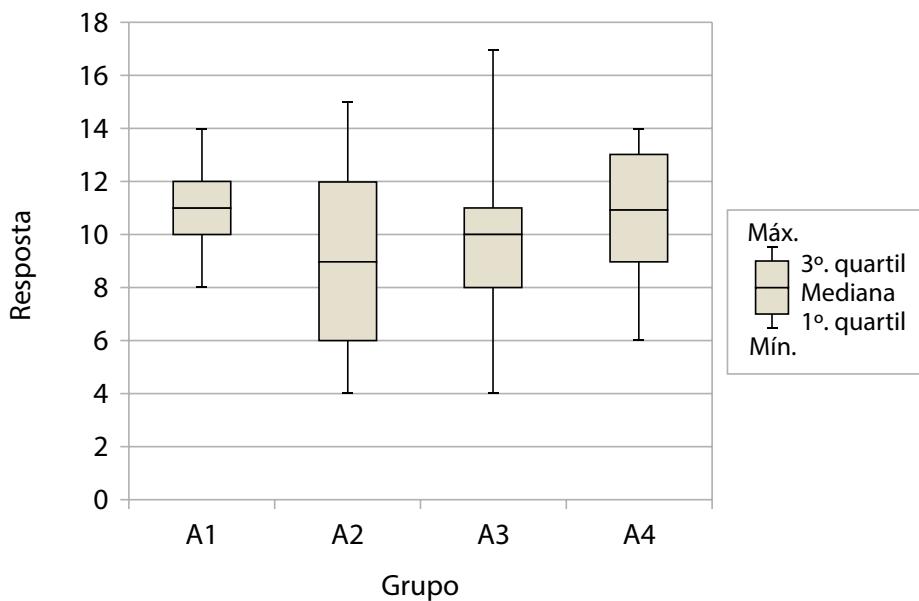


Figura 6 - Exemplo de diagrama de caixa (Ibovespa) / Fonte: Capela e Capela (2011, p. 363).

**Descrição da Imagem:** a figura apresenta um diagrama de caixa. Nele, há quatro grupos: A1, A2, A3 e A4. Para cada grupo, há valores expressos na caixa. Assim, na parte superior da linha, há o máximo. Na parte superior da caixa, há o terceiro quartil. Na metade da caixa, há a mediana. Na parte inferior da caixa, há o primeiro quartil e, na parte inferior da linha, há o mínimo. Esse diagrama busca apresentar a distribuição estatística dos valores referentes a uma variável ou grupo.

No contexto da análise exploratória, é muito importante utilizar todas as informações quantitativas relacionadas ao dado, pois, dessa forma, será possível ter uma melhor visão sobre a sua distribuição e comportamento. Em detrimento do fato de que a ideia desse tipo de análise é a de levantar características e gerar *insights* voltados à necessidade, ou não, do mergulho nas características do domínio de aplicação, quanto mais elementos e visualizações diferentes dos dados, melhor é para você.

Ao finalizar o relato sobre o diagrama de caixa, Joaquim pensou: e quando eu tiver que analisar dados não estruturados? O que eu poderia fazer? Existem recursos visuais que me auxiliariam em uma análise exploratória? Assim, avançou em suas anotações e percebeu que a resposta era positiva. Ele pode utilizar técnicas simples, como a geração de nuvens de termos a partir da frequência com que essas palavras aparecem nos textos. Isso o auxiliaria a entender os dados mesmo antes de navegar pelo domínio de aplicação. A figura a seguir exemplifica uma nuvem de termos:



**Figura 7 - Exemplo de nuvem de termos / Fonte: o autor.**

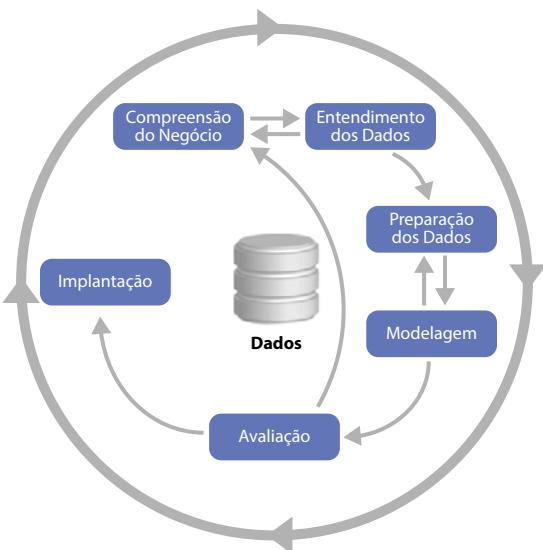
**Descrição da Imagem:** na figura, há uma nuvem de termos, que representa um conjunto de palavras que são representativas em um texto ou em uma coleção de documentos. Além disso, na nuvem de termos, o tamanho das palavras (tamanho da fonte que está sendo utilizada) é diretamente ligado a frequência com que a palavra aparece. Dessa forma, é possível visualizar quais são os principais termos.

A nuvem de termos é um recurso simples, mas proporciona uma noção clara dos principais assuntos abordados em documentos não estruturados a partir dos termos mais frequentes, que são representados por meio tamanho da fonte. Nesse sentido, quanto mais frequente, maior será o tamanho do termo na nuvem. Ao observarmos a Figura 7, é perceptível que os termos mais frequentes são “cidade”, “mundo” e “copa do mundo”. Cores também podem ser utilizadas para auxiliar na análise de elementos relacionados aos documentos.

Joaquim estava satisfeito e avançou para o levantamento das técnicas e dos métodos relacionados à **análise implícita**, que, assim como o próprio nome sugere, utiliza os dados que não têm significados explícitos. Esse fato exige a utilização de formas de mineração dos dados. Assim, Joaquim se lembrou da área de **aprendizado de máquina** (*machine learning*).

De acordo com Amaral (2016, p. 81), “aprendizado de máquina computacional é aplicação de técnicas computacionais na tentativa de encontrar padrões ocultos em dados” e, “enquanto o aprendizado de máquina trata de algoritmos que buscam reconhecer padrões em dados, a **mineração de dados** é a aplicação destes algoritmos em grandes quantidades de dados em busca de informação e conhecimento”.

Para trabalhar com a mineração de dados, é possível utilizar uma metodologia chamada CRISP-DM, que apresenta um modelo de processos para facilitar o entendimento das etapas relacionadas à aplicação da mineração de dados. A figura ao lado expõe as suas fases.



**Figura 8 - Fases da metodologia CRISP-DW**  
Fonte: Shearer (2000 apud RAMOS et al., 2020, p. 1094).

**Descrição da Imagem:** a figura apresenta a ideia de um fluxo e, no centro, estão os dados de uma organização. A primeira fase é a “Compreensão do Negócio”. Ligada a ela, há uma seta que vai até a fase de “Entendimento dos dados”. A partir dessa fase, pode-se voltar para a etapa de “Compreensão do Negócio” ou avançar para a fase de “Preparação dos dados”. Depois, há etapa de “Modelagem”, que permite retornar para a fase de “Preparação dos dados” ou avançar para a fase de “Avaliação”. A partir da “Avaliação”, é possível retornar para a etapa inicial de “Compreensão do Negócio” ou seguir para a última etapa, que é a “Implantação”.

As etapas do método são descritas com mais detalhes a seguir:

- **Compreensão do negócio:** busca entender quais são os objetivos do projeto e qual é o domínio de aplicação no qual ele será implantado.
- **Entendimento dos dados:** permite aplicar técnicas de análise exploratórias para entender os dados que estão envolvidos no projeto e relacioná-los com os elementos do domínio, a fim de formular hipóteses e analisar variáveis.
- **Preparação dos dados:** um dos processos mais importantes e custosos do ciclo. Consiste em deixar os dados limpos e em um formato que é possível utilizar durante os estudos.
- **Modelagem:** identificação dos modelos que melhor se adequam ao problema e a sua aplicação nos dados já preparados.
- **Avaliação:** identificação do grau de assertividade do modelo utilizado e já treinado. Caso não tenha sucesso, é preciso retornar para a etapa de compreensão do negócio, a fim de se buscar uma nova estratégia.
- **Implantação:** ao constatar que o resultado tem um grau de assertividade aceitável, pode ser implantado no ambiente de produção e disponibilizado para a organização.

Joaquim se lembrou de suas aulas de inteligência artificial e constatou: o aprendizado de máquina é dividido em três principais tarefas, a **classificação**, o **agrupamento** e a **associação**, que são detalhadas a seguir:

- **Classificação:** consiste em selecionar novos itens ou elementos para as classes já conhecidas. Além disso, exige a construção de um modelo que deve ser treinado (processo de aprendizagem) para que as classificações ocorram.
- **Agrupamento:** consiste em organizar itens ou elementos em grupos levando em consideração as suas características, sem a necessidade de ter um grupo previamente definido.
- **Associação:** também conhecida como “regras de associação”, é baseada nas descobertas de relações entre as variáveis.

Além das três atividades apresentadas, há a classificação das tarefas em supervisadas e não supervisionadas. A seguir, há as particularidades de cada uma:

- **Tarefa supervisionada:** são tarefas que precisam de uma etapa de treinamento para que possam ser executadas. Por exemplo, a classificação é uma tarefa supervisionada.

- **Tarefa não supervisionada:** são tarefas que não precisam de uma base já classificada como treinamento para que possam ser executadas. Por exemplo, o agrupamento e a associação são tarefas não supervisionadas.

Amaral (2016) desenvolveu uma tabela que explicita quais algoritmos estão relacionados a cada uma das tarefas de aprendizado de máquina:

Tarefas	Tipos de Algoritmos	Algoritmos
Classificação	Bayes	Naive Bayes BaysNet
	Regras	Party Decision Table
	Árvore de decisão	Random Forest J48
	Redes neurais	Back-Propagation
Agrupamento	Por densidade	DBSCAN
	Baseado em protótipo	K-means K-medoids
	Redes neurais	Mapas de Kohonen
Associação		Apriori FP Growth

Quadro 1 - Algoritmos e tarefas de aprendizado de máquina / Fonte: adaptado de Amaral (2016).

Joaquim entende que ainda não é o momento de detalhar cada um dos algoritmos e focará apenas em sua descrição. Nesse contexto, os algoritmos chamados “Bayes” utilizam cálculos simples de probabilidade, mais especificamente baseados na teoria de Thomas Bayes. Segundo Amaral (2016, p. 101), “dos valores de cada atributo, o algoritmo vai avaliar o quanto ele contribui para classificar a instância como boa ou ruim, construindo uma tabela de probabilidade”. Depois, é feita a soma dos índices com os valores, a fim de classificá-los como bons ou maus. Ao final, o valor de classe que tiver o maior índice é o selecionado.

Os tipos de algoritmos baseados em regras têm uma etapa de treinamento. Nela, são identificados padrões nas variáveis presentes e, na sequência, é gerado um conjunto de regras a serem utilizadas na classificação de novos casos e elementos. Muitas ve-

zes, esses modelos são transformados em árvores de decisões: assim como o próprio nome sugere, os atributos e as instâncias são organizados em forma de nós e arestas, a fim de representarem a naveabilidade que existe entre as regras construídas. Ao final da árvore, há uma classificação, ou seja, quando é chegada em uma folha (nó que não tem mais nenhum filho), existe uma classificação.

As redes neurais artificiais dizem respeito a uma área muito recorrente, mas que não é nova. Os primeiros estudos realizados pela matemática ocorreram entre a década de 60 e 70. No entanto, as redes foram deixadas em segundo plano durante um tempo, pelo fato de exigirem um poder computacional considerado alto para a época. Nas últimas décadas, com a evolução dos computadores e, principalmente, com o surgimento dos ambientes de alta performance na nuvem, as redes neurais voltaram aos holofotes, ao proporem soluções viáveis para serem utilizadas pelas empresas.

Nessa área, o objetivo é reproduzir a forma de funcionamento do cérebro. Desse modo, neurônios artificiais aprendem e são organizados em camadas. Na prática, o processo de andamento dentro da rede de neurônio é constantemente ajustado e, em consequência disso, o aprendizado é constante e evolutivo. É muito utilizado para o reconhecimento facial, por exemplo, que, apesar de cortarmos o cabelo ou trocarmos de óculos, permite que o processo ainda aconteça de maneira assertiva. No Quadro 1, é observável que as redes neurais carregam algoritmos focados em mais de uma tarefa e são bastante versáteis no que se refere à aplicabilidade.

Os algoritmos de agrupamento por densidade são formas de agrupamento não parametrizadas que fazem os processos de identificação de elementos próximos. Nesse caso, o fato de não serem parametrizadas significa que não é necessário expor previamente a quantidade de cluster a ser gerada. Já em relação aos algoritmos de agrupamento baseados em protótipos, é preciso informar quantos clusters devem ser gerados e, a partir desse número, os elementos são organizados por proximidade, ou seja, pelas suas características comuns. Contudo, assim como já afirmamos, não é necessário o processo de treinamento.

Ao finalizar a fase do levantamento, Joaquim estava com todos os aspectos mapeados. Ele apresentou o resultado do seu levantamento para Lara e comentou que, naquele momento, não era necessário adquirir novos computadores, nem a compra de serviço na nuvem para executar os processos de mineração de dados. Entretanto, à medida que a massa de dados aumentasse, seria preciso realizar uma expansão na infraestrutura. Lara ficou muito satisfeita com o trabalho de Joaquim e fez mais uma solicitação: que fossem levantadas quais ferramentas, além das linguagens de

programação R, Python e Java e das planilhas eletrônicas, poderiam ser utilizadas para a execução do seu trabalho enquanto cientista de dados. Lara explicou que, se fosse necessário adquirir alguma licença, era importante mapear agora, para que ela fizesse a solicitação de compra.

Joaquim se lembrou de uma ferramenta que usou durante a sua graduação, a chamada **Weka** (<https://www.cs.waikato.ac.nz/ml/weka/>). Ela é uma plataforma construída em Java, grátis e de código aberto. Foi desenvolvida pela *University of Waikato*, uma universidade da Nova Zelândia e tem uma plataforma visual para interagir com os seus algoritmos. Além disso, carrega uma estrutura de construção de processos baseados em fluxos, que podem ser modelados ao arrastar os componentes considerando as etapas de um projeto de mineração de dados. A Weka também permite que os seus algoritmos implementados sejam embarcados e invocados por novas soluções. Além da sua disponibilização na versão em Java, permite que as suas bibliotecas sejam acessadas por aplicações em Python, R e Scala.

A figura a seguir apresenta um exemplo de como é a interface gráfica da Weka:

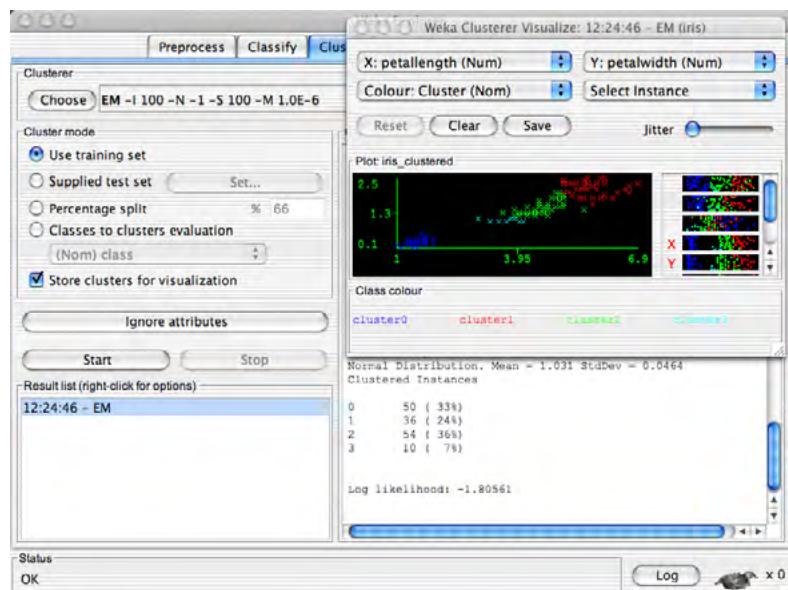


Figura 9 - Interface gráfica da ferramenta Weka / Fonte: Weka ([2021], on-line)<sup>2</sup>.

**Descrição da Imagem:** a figura apresenta uma tela de trabalho da ferramenta Weka. No menu, à esquerda, há as configurações para a seleção de uma etapa de treino de um modelo de clusterização. Na parte superior, é apresentado um gráfico de dispersão com a organização dos elementos por cores. Assim, cada cor é um cluster gerado. Na parte inferior, à direita, é exposta a distribuição dos dados por cluster, demonstrando o percentual de cada um.

Joaquim ficou muito empolgado com a possibilidade de desenvolver as soluções a partir das implementações disponíveis no Weka. Todavia, ficou preocupado com a possibilidade de fornecer essa ferramenta para analistas menos técnicos, pois a sua interface visual não é a mais simples e intuitiva. Em detrimento do fato de que Joaquim pretende desenvolver uma cultura analítica baseada em dados na organização, deveria buscar alternativas que fossem mais compreensíveis, para que as áreas de negócio pudessem desenvolver as suas soluções e executassem as suas análises.

Ao navegar pela Internet, encontrou a ferramenta **Orange** (<https://orange.biolab.si>), que também é gráts e de código aberto. Trata-se de uma ferramenta escrita em Python que também permite acessar os algoritmos implementados. No entanto, de fato, o foco de Joaquim está no uso de sua ferramenta visual, que é bastante intuitiva e permite que, com pouco conhecimento sobre os processos e os algoritmos envolvidos, sejam geradas tarefas de mineração de dados e texto.

A figura a seguir apresenta um exemplo de fluxo de extração de termos de uma coleção de documentos não estruturados. Assim, são expressos o termo e a frequência com que ele ocorre no texto. Depois, o resultado é salvo em uma planilha eletrônica.

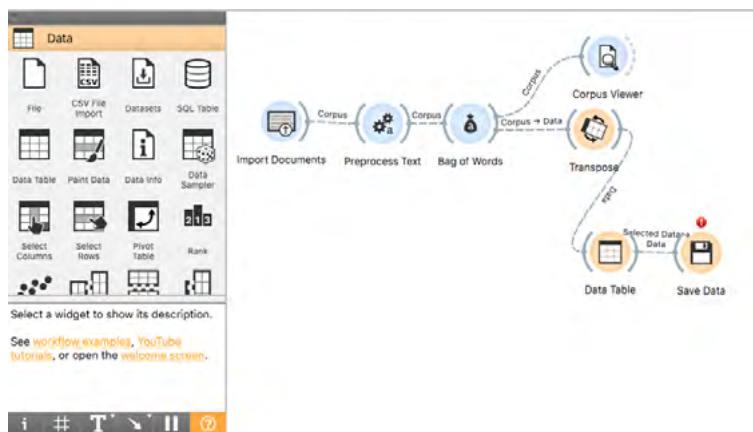


Figura 10 - Exemplo de uso da ferramenta Orange / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta a ferramenta Orange e como ela está organizada. Desse modo, no menu, à esquerda, estão disponíveis as opções de processamento e os algoritmos disponíveis e organizados pela tarefa em questão. Também são demonstrados os ícones referentes aos processos de manipulação de dados. Cada ícone pode ser arrastado para a área de workflow, que fica à direita. Nessa área, é preciso organizar os processos de maneira linear, ligando as entradas e as saídas. Ao final, os valores podem ser visualizados ou salvos em algum tipo de arquivo escolhido.

Essa ferramenta é ideal para quem precisa fazer estudos simples e está iniciando os estudos relacionados à área de mineração e análise de dados.

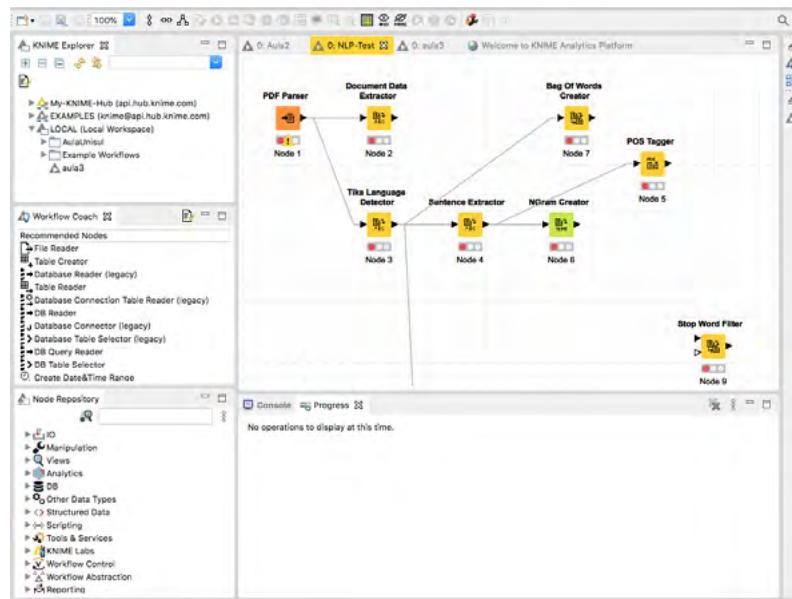


Figura 11 - Exemplo de uso do Knime / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta a ferramenta Knime. A sua organização é parecida com a da ferramenta Orange. Ao lado esquerdo, há três painéis empilhados. Na parte superior, é possível navegar pelas pastas dos projetos, que estão salvos no sistema de arquivo do computador. Já no segundo painel, estão as opções de componentes e as ferramentas que podem ser adicionadas ao workflow em questão. No terceiro painel, são organizados os tipos de componentes e as ferramentas, considerando a sua natureza de uso (como manipulação de dados e visualizações). À direita, há um espaço para a construção do workflow que será utilizado como processo e, na parte inferior, há a console com os logs de execução do processo criado.

Joaquim também procurou uma ferramenta mais completa e com recursos para análises mais complexas. Em sua pesquisa, encontrou uma que é chamada **Knime** (<https://www.knime.com>). A Figura 11 apresenta um exemplo de sua interface gráfica. A Knime é uma plataforma que tem uma versão gratuita e uma versão paga, com foco no uso profissional. Também é baseada em fluxos, mas apresenta um nível de personalização e configuração bastante grande, o que a torna uma ferramenta muito poderosa, mas de complexidade maior.

Ao finalizar o levantamento das ferramentas, foi verificado que não era necessário adquirir uma nova licença, já que as três ferramentas podem ser utilizadas à medida que as necessidades analíticas se tornem mais complexas ou a maturidade analítica da organização seja elevada. Lara ficou muito feliz com o resultado de todo o levantamento, inclusive das ferramentas. Com a parceria com o Joaquim, já tem em mente qual será o novo desafio que demandará ao cientista de dados.

## Governança de Dados (GD)

Lara e Joaquim apresentaram a Anderson todos os aspectos levantados e especificados, além dos que foram aqui narrados. Eles desenvolveram um desenho para demonstrar como ficaram as ferramentas instaladas e os demais computadores envolvidos nas temáticas analíticas. Anderson ficou muito feliz com o que os dois fizeram em apenas uma semana e percebeu que realmente essa dupla revolucionaria o seu negócio.

Ao terminar a reunião, Lara chamou Joaquim e disse que já tinha uma nova missão para ele. Nesse caso, não estaria ligada apenas aos novos processos a serem desenvolvidos, mas também implicaria no que já estava implantado e em uso na organização. Joaquim disse que seria um prazer poder contribuir e adorou saber que a sua nova missão estava diretamente ligada com a organização dos dados de toda a organização. Contudo, quis saber mais detalhes sobre as dores que estavam enfrentando ou se seria só uma medida de precaução.

Lara perguntou se Joaquim gostaria de tomar um café, para que pudessem conversar em um lugar mais calmo. Ele aceitou o seu convite e foram até uma padaria que ficava próxima ao escritório da matriz. Logo depois de fazerem os seus pedidos, Lara explicou algumas situações que estavam ocorrendo dentro da empresa. Assim, relatou que a implantação da ferramenta de BI foi um sucesso, mas muitas áreas gostariam de ter visões diferentes.

Pelo fato de que não havia um pessoal responsável pela continuidade e manutenção da solução, todos passaram a utilizar novamente as planilhas eletrônicas com dados extraídos de vários lugares. Essa situação está gerando muito desconforto nas reuniões de diretoria, já que é comum que gestores apresentem o mesmo indicador com valores distintos, o que gera uma discussão em relação ao número, e não questões estratégicas.

Essa situação se intensificou, porque as áreas do negócio não sabem onde podem recuperar os dados para fazerem os seus estudos e elaborarem as suas estratégias quando eles não estão disponíveis nas soluções de BI corporativas. Outra dor que Lara relatou à Joaquim foi o fato de não ter controle de acesso e permissão às fontes de dados, o que permite que todos os funcionários tenham acesso a dados que são sensíveis. Joaquim escutou e anotou tudo atentamente. Depois de analisar a situação com calma, comentou que seria importante a implantação de uma Governança de Dados na empresa.

Por governança, entende-se o conjunto de decisões, operações e atribuições que cada membro de uma organização tem. A Governança de Dados engloba um conjunto de processos, decisões, operações e atribuições que cada agente de dados possui, levando em consideração todos os processos do ciclo de vida dos dados. Já como agente de dados, compreende-se qualquer pessoa da organização que tenha participação em alguma etapa do processo, mesmo que seja apenas realizando o consumo dos dados e das informações.



**Figura 12 - Visão geral sobre Governança de Dados com 5W e 2H**  
Fonte: adaptada de Barbieri (2011).

**Descrição da Imagem:** a figura apresenta a sigla "GD" no centro e há um círculo em volta do termo com os seguintes elementos: i) "o quê?", que sustenta que o foco da governança corporativa se dá sobre os recursos dos dados, das informações e do conhecimento, que são considerados ativos empresariais; ii) "por quê?", que inclui mercado, clientes, regulações: aderência/compliance, reputação, qualidade, segurança e fontes variáveis de dados: ERP, SCM, SFA e PLM; iii) "onde?", que abrange as áreas sensíveis da empresa, tais como clientes, fornecedores e produtos, as áreas recentemente juntadas e as áreas de Master Data (MD); iv) "quando?", que está dentro de um planejamento estratégico em vários anos e inclui um ciclo de projetos dentro do Programa de GD; v) "quem?", que se refere às pessoas, à comunicação, aos papéis envolvidos nas áreas de negócios sensíveis, ao gestor de dados da área de negócio e TI, ao escritório de dados e ao comitê gestor de Governança de Dados; vi) "como?", que diz respeito ao processo de Governança de Dados sobre domínios (qualidade, MDM, BI, analytics), às políticas sobre direitos, padrões, responsabilidade, controle de segurança, privacidade, regras de negócio e riscos, às medições de qualidade de dados e ao programa de GD, tais como projetos com retornos claros e imediatos; e vii) "quanto?", que abrange ROI, recursos, custos e ganhos intangíveis, e custo negativo - Data Flaws.

Em seu livro, Carlos Barbieri (2011) explica a visão geral da Governança de Dados (GD) a partir do 5W e 2H. A seguir, são detalhados cada item da Figura 12:

**O quê? (what):** é preciso esclarecer quais artefatos, elementos relacionados aos dados e informações da organização devem ser organizados. Além disso, é preciso definir “o seu uso controlado, a sua qualidade e as diretrizes para o seu consumo e produção” (BARBIERI, 2011, p. 28).

**Por quê? (why):** quais são os motivos para se ter uma Governança de Dados instalada? É possível estabelecer um paralelo com as dores mapeadas por Joaquim e Lara.

**Onde? (where):** é necessário definir quais áreas devem ter prioridade de ação e os seus dados governados.

**Quando? (when):** é preciso planejar a implantação dos processos e das ações de Governança de Dados em ciclos, ao desenhar o momento em que cada interação deve acontecer, considerando as áreas de negócio.

**Quem? (who):** quem são os agentes de dados envolvidos em cada uma das iterações planejadas, para que sejam incluídos no processo de implantação? É necessário envolver os agentes no projeto, a fim de que apoiem os processos de especificação, execução e avaliação.

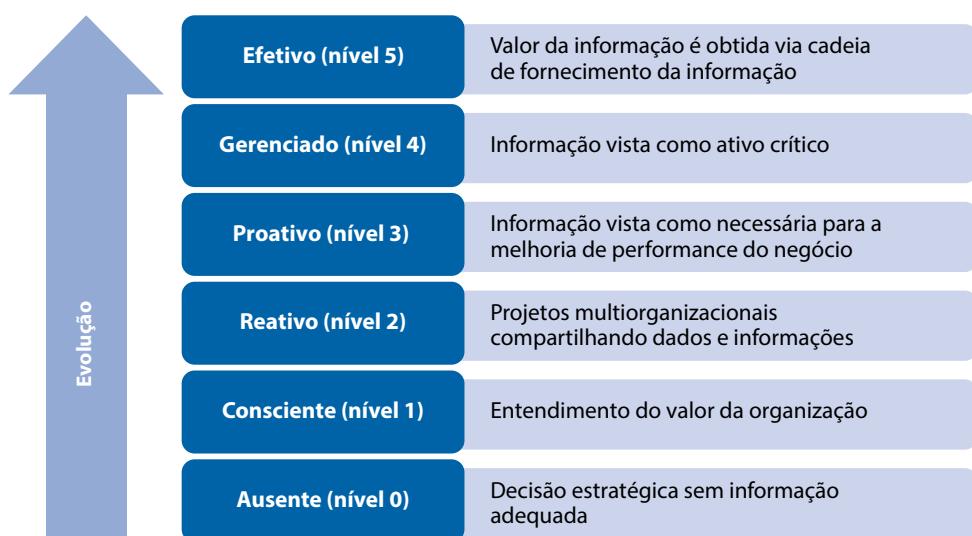
**Como? (how):** definição de regras e políticas que serão as norteadoras do modo de execução dos processos estabelecidos. Além disso, é gerada uma ideia em relação aos padrões e às responsabilidades dos associados ao uso.

**Quanto? (how much):** identificação dos custos dos projetos relacionados à execução para se chegar a uma definição de quanto se terá de retorno com essa frente.

Joaquim e Lara partiram para o levantamento das pessoas que eram os agentes de dados da organização. A partir dessa lista, foram identificados os possíveis Responsáveis Técnicos (RTs) de cada área de negócio, que serão os envolvidos no processo de levantamento e especificação das bases de dados e indicadores. Após o levantamento das pessoas, foi feita uma reunião para informar o importante papel que eles terão no processo de implantação da Governança de Dados da empresa. Os dois evidenciaram alguns problemas que a organização

enfrenta e como a implantação pode ajudá-los a não terem que conviver com esses problemas novamente.

Ao final da apresentação, uma garota chamada Carolina levantou a mão e perguntou: “Vocês sabem em qual nível de maturidade de governança nós estamos?”. Naquele momento, Joaquim ficou assustado. Ele havia esquecido de considerar algum modelo de maturidade em Governança de Dados para posicionar o atual e os futuros estados. Por sorte, lembrou-se de um modelo proposto pela *Gartner Group*, o chamado *Enterprise Information Management* (EIM), que tem níveis de maturidade definidos e pode ajudar as organizações a saberem onde pretendem chegar. Assim, é possível gerar um plano estratégico para o alcance das metas estabelecidas. A Figura 13 apresenta mais detalhes:



**Figura 13 - Maturidade em Governança de Dados - Modelo Gartner EIM**

Fonte: adaptada de Barbieri (2011).

**Descrição da Imagem:** a figura apresenta uma seta que aponta para cima e é intitulada “Evolução”, a fim de demonstrar a evolução entre os níveis: 0 - Ausente (decisão estratégica sem informação adequada); 1 - Consciente (entendimento do valor da organização); 2 - Reativo (projetos multiorganizacionais compartilhando dados e informações); 3 - Proativo (a informação é compreendida como necessária para a melhoria de performance do negócio); 4 - Gerenciado (a informação é entendida como ativo crítico); e 5 - Efetivo (o valor da informação é obtido via cadeia de fornecimento da informação).

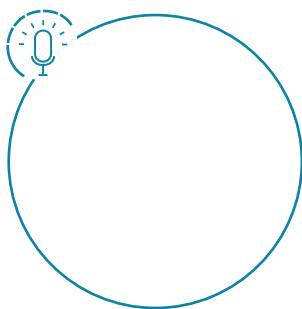
Com base em Barbieri (2011), os níveis são detalhados a seguir:

- **Nível 5 – Efetivo:** o valor da informação está vinculado aos processos e aos níveis de acordo de serviços. Nele, a Governança de Dados é corporativa e institucionalizada, e há total controle, acesso e gestão dos ativos de dados e informações da organização.
- **Nível 4 – Gerenciado:** os dados são entendidos como recursos críticos. Assim, políticas, padrões e procedimentos são definidos e acordados com toda a organização.
- **Nível 3 – Proativo:** a informação é compreendida como algo valioso e estratégico para a organização. Dessa forma, é criada e mantida uma arquitetura de informação e há a definição dos papéis. Além do mais, existe um processo mapeado dentro dos processos de desenvolvimento de software (ou implantação de tecnologia) da organização.
- **Nível 2 – Reativo:** há o compartilhamento de dados e informações entre os departamentos. É defendida a ideia de que é preciso analisar os dados organizacionais, mas não existe uma visão de dados uniformizada.
- **Nível 1 – Consciente:** o dado e a informação têm valor, mas não existem processos para a gestão do dado. Além disso, é exigida a formulação de uma arquitetura corporativa de dados e informações.
- **Nível 0 – Ausente:** assim como o próprio nome sugere, não existe nenhuma iniciativa de Governança de Dados. Há, no máximo, algumas iniciativas isoladas nas áreas de negócio.

O primeiro passo de Joaquim foi classificar o estado atual da empresa de Anderson. Nesse contexto, considerou os seguintes aspectos: a empresa tem uma arquitetura de informação, pois foi instituído um *Data Warehouse* com componentes de ETL para garantir a transformação e a consolidação do dado, armazenando-o de maneira centralizada. Também tem as bases transacionais organizadas (tanto a do ERP quanto a do CRM). Entretanto, apesar de as áreas de negócio terem evoluído, o projeto de BI permaneceu estagnado. Logo, os indicadores que estavam concentrados nos *dashboards* não são mais o suficiente para os gestores, porque a empresa cresceu, novas áreas foram criadas e algumas estratégias foram revisitadas, o que deveria gerar um processo de manutenção nos indicadores e no DW.

Em detrimento do fato de que as novas áreas não tinham indicadores elaborados, as soluções de BI passaram a gerar os seus próprios indicadores a partir de planilhas eletrônicas, que acessavam diretamente as bases de dados transacionais. Contudo, pelo fato de que não há uma documentação em relação à qual campo deve ser utilizado e não se tem clareza das regras de negócio utilizadas pelas áreas, para que gerem os seus indicadores, as planilhas expressam valores conflitantes em reuniões de comitê com os diretores.

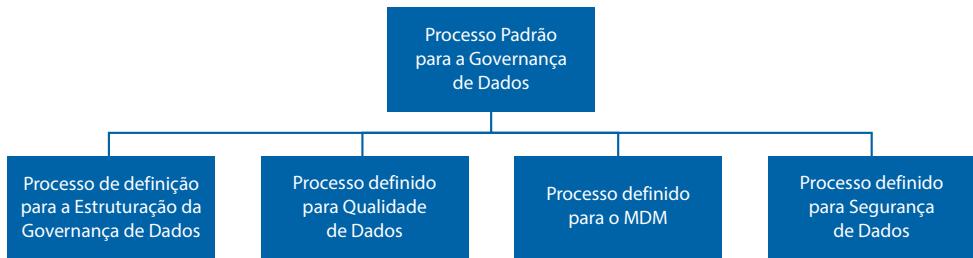
Inicialmente, Joaquim acreditava que a empresa estava no nível 3 de maturidade (proativo), mas, em consequência de não existirem processos, papéis e responsabilidades relacionadas ao ciclo de dados, Joaquim classificou a empresa em nível 2 e já sabia exatamente o que fazer para subir na hierarquia dos níveis de maturidade da Governança de Dados. Para isso, bastava se localizar entre os níveis e observar o que havia de diferente no próximo nível. Desse modo, teria um *roadmap* de evolução do projeto até chegar ao nível 5. Joaquim logo percebeu que era necessário pensar nos processos relacionados aos ciclos de vida do dado, aos papéis, às responsabilidades e ao repositório para documentação.



Um dos processos mais importantes para uma empresa dirigida por dados ou que deseja implantar uma área de ciência de dados é a implantação da Governança de Dados (GD). Nesse contexto, quais são os desafios conhecidos para se implantar uma GD em uma organização? Qual seria o primeiro passo? Como a GD pode apoiar o dia a dia da ciência de dados? Essas e outras questões serão abordadas em nosso podcast! Não deixe de escutá-lo!

Para facilitar o processo de implantação do projeto de Governança de Dados, mais uma vez, Joaquim teve de recorrer às anotações e aos conteúdos estudados durante a sua graduação. Um processo padrão, para a Governança de Dados, pode ser especializado em processos menores, tais como as definições de estruturação da Governança de Dados, da qualidade de dados, de *Master Data Management* (MDM) e de segurança de dados.

A figura a seguir apresenta como se dá a especialização do processo padrão de GD:



**Figura 14 - Especialização do processo padrão para a GD / Fonte: adaptada de Barbieri (2011).**

**Descrição da Imagem:** a figura apresenta uma hierarquia. Nela, há um processo maior, o chamado “Processo padrão para a Governança de Dados”, que se especializa em quatro processos inferiores: “Processo de definição para a estruturação da Governança de Dados”, “Processo definido para qualidade de dados”, “Processo definido para o MDM” e “Processo definido para a segurança de dados”.

Inicialmente, foram organizados os grupos de agentes de dados que participarão do processo de mapeamento das fontes de dados, do levantamento de requisitos e das regras de negócio e da validação da infraestrutura de dados e indicadores propostos. Também foi instituído um comitê de dados para que, de maneira colaborativa, tenha a priorização e a avaliação de todos os artefatos de governança desenvolvida. Essa é uma forma de apoiar o processo de estruturação da GD.

Foram definidas algumas ferramentas para auxiliar na documentação dos processos, nas decisões tomadas no comitê de dados, na organização dos indicadores, na declaração das regras de negócio, no dicionário de dados, nos possíveis selos de maturidades e dentre outros artefatos da governança. Foi selecionada uma ferramenta do tipo Wiki para ser o ponto central no armazenamento de todos os artefatos, a fim de que seja possível manter os dados e as informações organizadas para que possam ser utilizadas por todos os seus agentes de dados durante a execução das suas tarefas no dia a dia.

O processo de construção de uma camada de dados MDM é justamente o que garante a edificação de uma base unificada em relação aos conceitos de negócio e de dados que já passaram por uma camada de validação. Esse tipo de base geralmente está em uma camada de dados chamada “camada de acesso”. Trata-se de uma camada de dados já processada para ser consumida pelas áreas que queiram gerar os seus relatórios e deve ser totalmente aderente e governada por um processo de Governança de Dados.



## EXPLORANDO IDEIAS

O Master Data Management (MDM) é um conjunto de processos, técnicas e ferramentas que organizam os dados da organização em um banco de dados unificado. Nele, os dados já estão validados e transformados, e expressam conceitos únicos para todos os domínios da organização.

Fonte: o autor.

A organização lógica das bases de dados permite que o processo de governança tenha sucesso durante os processos de documentação, validação e controle dos dados organizados. A figura a seguir apresenta mais detalhes sobre essa visão em camadas:

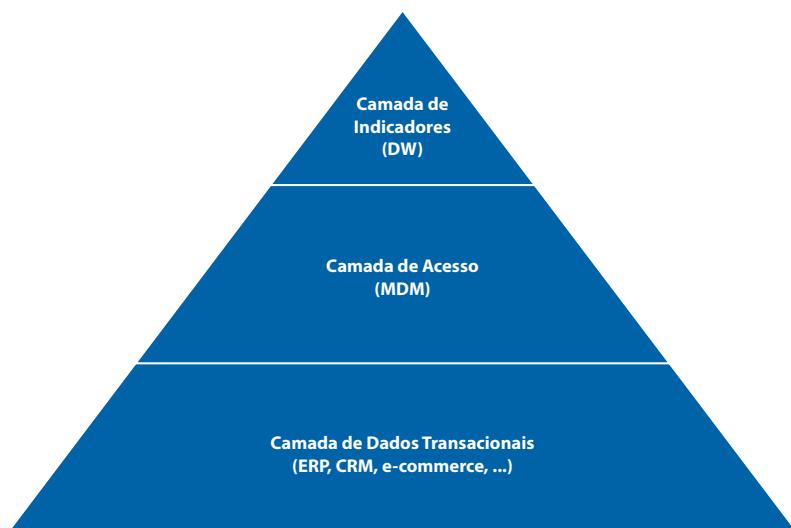


Figura 15 - Camadas lógicas das bases de dados para a GD / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta uma pirâmide com três camadas. A base tem o seguinte título: "Camada de Dados Transacionais (ERP, CRM e e-commerce, por exemplo)". A camada central é chamada de "Camada de Acesso (MDM)", enquanto a camada superior é intitulada "Camada de Indicadores (DW)".

Na camada de base, está a **Camada de Dados Transacionais**, que abrange as bases de dados utilizadas pelos sistemas operacionais e transacionais. Essas bases não devem ser acessadas por nenhum agente humano, apenas pelos sistemas de carga, para que seja possível aplicar as regras de negócio e as transformações necessárias para a obtenção de dados íntegros e aderentes aos processos e regras definidas pela GD. Os agentes

de dados têm acesso à **Camada de Acesso**, que, geralmente, é implementada com base no conceito de MDM. Ela é uma camada totalmente governada e deve ter processos de gestão de acesso e segurança de dados. Em outras palavras, para um usuário, só são visíveis os dados que ele tem permissão para visualizar. Por fim, a **Camada de Indicadores** é, de modo geral, implementada em forma de *Data Warehouse* e é consumida por *dashboards*, que compõem uma solução de BI. Nesse caso, todos os indicadores são governados e submetidos às políticas de acesso e segurança da informação.

Como se pode perceber, o **processo de definição da segurança da informação** é muito importante e deve ser contemplado quando se fala em GD. Dois aspectos iniciais são o processo de gestão de acesso e formação dos perfis, que se refere ao mapeamento do que cada agente de dado pode acessar, e a obtenção de ferramentas tecnológicas que garantam esse acesso (e seus filtros). A GD deve ter processos para a inclusão, manutenção e exclusão de privilégios de acesso por perfil dos agentes de dados. Além do mais, é importante que toda a decisão de acesso seja colegiada entre os gestores e, em muitos casos, concedidas apenas com a autorização dos diretores.

Posteriormente, Joaquim e Lara passaram a definir os processos de **qualidade de dados**, que abrangem uma das etapas mais importantes para a implantação da GD. A qualidade de dados pode ser definida com base em várias dimensões, assim como Barbieri (2019) sustenta:

- **Integridade de dados:** dado confiável que passou por um processo de avaliação e pelas regras de negócio (se houver), garantindo a sua sinergia com os conceitos definidos.
- **Segurança e privacidade dos dados:** garantia de que não sejam inseridos, nas bases, dados sensíveis ou que venham a ferir alguma legislação, como a Lei Geral de Proteção de Dados.
- **Documentação:** ter clareza sobre a natureza dos dados, o seu ciclo de vida, atualização, regras de negócio e conceitos vinculados. Essa documentação deve estar acessível para todos os envolvidos.
- **Dados qualificados para a tomada de decisão:** está relacionado ao fato de os dados analíticos estarem organizados em repositórios do tipo DW, MDM ou outros que tenham processos de transformação e limpeza atrelados.

Por fim, Joaquim e Lara buscaram algumas ferramentas para apoiá-los no processo de garantia da qualidade de dados e conheceram em cinco ferramentas. De acordo com Barbieri (2011), são elas:

- **Profiling e visualização:** análise de atributos mínimo, máximo, distribuição de frequência e nível de preenchimento, permitindo que os dados sejam visualizados e armazenados.
- **Parsing:** análise dos dados textuais para verificar se existem dados que estão violando alguma legislação ou não seguindo um padrão em especial (por exemplo, o padrão de CNPJ).
- **Matching:** verificação de dados a partir do seu conteúdo, de maneira que seja de rápido acesso e para a contabilização.
- **Cleansing:** processos de limpeza e padronização de dados a partir das regras de negócio estabelecidas.
- **Monitoração:** monitoramento do resultado das ferramentas anteriores, a fim de garantir que os processos relacionados à qualidade de dados estejam sendo executados e para mensurar alguns indicadores sobre a qualidade do dado armazenado.

Depois de todo o processo de implantação da GD foi possível realizar as análises que Anderson havia solicitado. Relembremos as solicitações:

- Saber quais clientes estão propensos a deixar de comprar em sua loja.
- Qual seria o mix ideal de produto por loja.

A implantação da GD garantiu a implantação de processos e políticas que asseguraram qualidade em todo o ciclo de dados. A organização da infraestrutura de dados em três camadas fez com que todas as áreas consumissem os dados do mesmo repositório, além de ter sido elaborada uma documentação sobre as regras de negócio e demais itens relacionados à natureza dos dados armazenados. Tendo uma base de dados confiável, foi possível que Joaquim, a partir da análise de compra dos clientes, ao aplicar alguns métodos e técnicas existentes em seu relatório, chegasse a um modelo que permitisse calcular o índice de propensão de um cliente deixar de comprar na loja, potencializando ainda mais o trabalho da área de CRM.

A partir do processo de agrupamento (clusterização), foi possível segmentar os clientes das lojas por grupos (a partir do seu histórico de compra) e verificar quais são mais rentáveis para a organização. Essa informação também pode ser usada no processo de construção do mix ideal de produto por loja, além, é claro, de permitir a ciência de quais produtos saem de maneira recorrente, a fim de que seja possível garantir a sua disponibilidade na loja para a compra dos clientes.

# AGORA É COM VOCÊ



1. A área da ciência de dados é multidisciplinar. Dessa forma, o profissional deve conhecer métodos e técnicas de várias áreas distintas.

Quais são os três principais pilares da ciência de dados?

- f) Matemática/estatística, computação e engenharia.
- g) Computação, engenharia e negócio.
- h) Estatística, negócio e direito.
- i) Matemática/estatística, computação e negócio.
- j) Ciência da informação, engenharia e matemática.

2. No contexto do aprendizado de máquina, existem as chamadas tarefas, que podem ser supervisionadas ou não supervisionadas.

Assinale a alternativa que apresenta a principal diferença entre as tarefas supervisionadas e não supervisionadas:

- a) As tarefas supervisionadas precisam de uma etapa de treinamento e de uma base de casos conhecidos para que possam ser executadas. As não supervisionadas não precisam de treinamento.
- b) As tarefas não supervisionadas não precisam de um agente humano, que desempenha o papel de supervisor.
- c) As tarefas supervisionadas se baseiam em ferramentas estatísticas. Já as não supervisionadas são baseadas em um ferramental computacional.
- d) As tarefas supervisionadas precisam ser executadas em uma infraestrutura de *Big Data*, enquanto as não supervisionadas não precisam.
- e) As tarefas não supervisionadas não precisam de um agente computacional que desempenha o papel de supervisor.

3. A Governança de Dados (GD) pode ser medida a partir de um modelo de análise da maturidade. Diante disso, a Gartner propôs um modelo dividido em seis níveis.

Assinale a alternativa que apresenta o nome de, pelo menos, três níveis desse modelo:

- a) Ausente, presente e consciente.
- b) Ausente, presente e reativo.
- c) Reativo, proativo e presente.
- d) Consciente, inconsciente e gerenciado.
- e) Gerenciado, proativo e reativo.

# CONFIRA SUAS RESPOSTAS



1. D.

A ciência de dados é fundamentada em três pilares: matemática/estatística, tecnologia/computação e negócio.

2. A.

Tarefa supervisionada: diz respeito às atividades que precisam de uma etapa de treinamento para que possam ser executadas. Classificação é uma tarefa supervisionada.

Tarefa não supervisionada: refere-se às tarefas que não precisam de uma base já classificada como treinamento para que possam ser executadas. Agrupamento e associação são tarefas não supervisionadas.

3. E.

Os nomes das seis fases são: ausente, consciente, reativo, proativo, gerenciado e efetivo.

# REFERÊNCIAS



AMARAL, F. **Introdução à Ciência de Dados:** mineração de dados e big data. Rio de Janeiro: Alta Books, 2016.

ANDERSON, C. **Creating a Data-Driven Organization:** practical advice from the trenches. United States of America: O'Reilly Media, 2015.

BARBIERI, C. **BI2 - Business Intelligence:** modelagem & qualidade. São Paulo: Elsevier, 2011.

BARBIERI, C. **Governança de dados:** práticas, conceitos e novos caminhos. Rio de Janeiro: Alta Books, 2019.

CAPELA, M. V.; CAPELA, J. M. V. Elaboração de gráficos *box-plot* em planilhas de cálculo. In: CONGRESSO DE MATEMÁTICA APLICADA E COMPUTACIONAL, 1., 2011, Uberlândia. **Anais** [...]. Uberlândia: CMAC, 2011.

CONAMAY, D. The data science venn diagram. **Drew Conway**, 30 set. 2010. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. Acesso em: 17 fev. 2021.

RAMOS, J. L. C. et al. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31., 2020, Natal. **Anais** [...]. Natal: SBC, 2020.

RAUTENBERG, S.; CARMO, P. R. V. do. Big Data e Ciência de Dados: complementariedade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Science: Research Trends**, v. 13, n. 1, p. 56-67, 2019.

## REFERÊNCIAS ON-LINE

<sup>1</sup>Em: <https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/ci%C3%A3ncia/>. Acesso em: 17 fev. 2021.

<sup>2</sup>Em: [https://weka.sourceforge.io/explorer\\_screenshots/ClusterPanel.png](https://weka.sourceforge.io/explorer_screenshots/ClusterPanel.png). Acesso em: 18 fev. 2021.

# 3

# Introdução ao *Big Data*

Dr. Flávio Ceci

## OPORTUNIDADES DE APRENDIZAGEM

Nesta unidade, entenderemos como se caracteriza um cenário de Big Data, conheceremos as suas principais características e analisaremos como se dá o seu processo de formação. Dessa forma, saberemos como e quando podemos trabalhar com um recurso como esse em uma organização. Por fim, estudaremos os principais métodos e técnicas utilizados na construção de uma solução baseada em Big Data, ainda sem abordarmos as questões relacionadas às tecnologias e às ferramentas.

Ao adicionar a ciência de dados nas entregas analíticas, foi dado um salto muito importante, visto que foi possível conhecer melhor os clientes, o que potencializou ainda mais as soluções de CRM. Nesse contexto, foram desenvolvidos os Modelos de Churn, que demonstram a probabilidade de um cliente já fidelizado deixar de comprar na loja. Também foi entendido o comportamento de compra dos clientes, o que permitiu a proposição de um mix de produtos por loja considerando o mês do ano, o que diminui ao máximo a possibilidade de um cliente não encontrar o que procura em uma loja física. Esse mesmo estudo foi estendido para o ambiente digital, já que foi elaborada uma série de indicadores e alertas para a área de compra, de modo que os clientes do *e-commerce* também não tivessem problema em encontrar os produtos que necessitavam.

Tanto a área de CRM quanto a área de compras estavam em êxtase com as novas entregas analíticas feitas pela área de Lara. As equipes comentavam que a chegada de Joaquim complementou muito a área da tecnologia, que já vinha fazendo um excelente trabalho, mas que, claramente, obteve um ganho ainda maior com o ingresso do novo integrante, que tem a sua mente totalmente dirigida por dados e sempre levanta questões relacionadas ao ciclo de vida do dado e, principalmente, à sua utilização para o negócio.

A área de logística também estava muito satisfeita com as novas ferramentas desenvolvidas para apoiar o processo de descentralização do seu estoque em forma de vários centros de distribuição (CD) espalhados

pelo país. Os estudos feitos foram essenciais para saber quais seriam as cidades mais estratégicas para a instalação dos centros e para conhecer o ciclo de vida de cada um dos produtos durante a distribuição entre os CDs, levando em consideração o mix ideal de produtos por loja e a frequência de vendas com



base no mês em questão. A área de marketing também estava dotada de excelentes indicadores e ferramentas analíticas que permitiam fazer uma comunicação baseada no horário em que o cliente estava mais propenso a abrir o e-mail ou o SMS, a fim de realizar uma entrega personalizada enquanto recomendação de produtos e serviços, além de promoções exclusivas.

A empresa de Anderson estava consolidada como um dos cinco maiores *e-commerces* de instrumentos e itens musicais do Brasil, além de ter uma rede de lojas físicas espalhadas por todo o país, o que permitia que os seus clientes tivessem uma experiência única. A empresa já contava com mais de 500 funcionários e um faturamento nunca imaginado. No entanto, houve um fato que deixou Anderson muito preocupado: há menos de 15 dias, houve o lançamento de uma linha de guitarras feitas por um *luthier* catarinense que chamou a atenção de grandes músicos do Brasil. Em consequência disso, houve uma procura inesperada por essas guitarras. A empresa de Anderson não estava preparada para essa situação, ou seja, muitos clientes fizeram uma busca em seu site, mas não encontraram o produto solicitado. Depois de refletir durante um tempo sobre a situação, chegou à seguinte questão: é possível prever novas situações ou tendências, de modo a preparar as empresas para futuras demandas?

Joaquim havia criado um alerta com base nos termos de busca utilizados no site e no aplicativo da loja, referentes aos produtos desconhecidos ou que não tivessem presentes no estoque. Nesse caso, quando foi obtido o número de cinco pessoas, em regiões distintas do Brasil, que estavam pesquisando o mesmo nome de um produto desconhecido, foi emitido um alerta para a equipe da Lara e para o setor de compras. Quando Joaquim recebeu o alerta, mais ou menos três horas depois de o primeiro cliente buscar pelo nome da guitarra, foi logo pesquisá-lo na Internet, com o objetivo de entender do que se tratava aquele produto e verificou a situação mencionada. Assim que entendeu o que estava acontecendo, foi até Lara e explicou o ocorrido. Os dois organizaram os dados e os fatos e foram até a sala de Anderson para explicitar a situação.

Anderson escutou tudo com muita atenção e permaneceu em silêncio em conjunto com Joaquim e Lara durante um minuto. Após um prolongado suspiro, disse que era muito grato por tudo que os dois fizeram pela empresa e que, se não fosse toda a infraestrutura tecnológica inicialmente desenvolvida por Lara e, depois, mantida pela sua equipe, em paralelo com as ferramentas de alerta desenvolvidas por Joaquim, ele só descobriria uma situação como essa depois de perder muitas vendas, e não em menos de cinco horas após a primeira busca. Anderson telefonou para a sua equipe de compras entrar em contato com o luthier, a fim de adquirir algumas peças ou até fazer uma parceria, com o objetivo de que a loja de Anderson fosse o canal oficial de venda de suas guitarras. Dos cinco usuários que haviam feito a busca pelo produto, três estavam logados no sistema e foi possível contatá-los para informá-los de que, em poucos dias, a empresa teria o produto disponível e, caso ainda não o tivessem comprado, teriam um desconto no valor final.

O aspecto que estava incomodando Lara e Joaquim é o de que, mesmo tendo identificado a situação muito rapidamente, ela poderia ter sido ainda mais antecipada, o que permitiria a realização do contato com o fornecedor antes mesmo de o primeiro cliente ter feito a busca em sua plataforma. Quando ambos explicaram esse fato para Anderson, ele deu um salto de sua cadeira e falou: o que precisamos fazer para conseguirmos isso? Joaquim explicou que, ao receber o alerta, navegou pela Internet, com o objetivo de compreender o que estava acontecendo e, em poucos cliques, encontrou notícias relacionadas ao lançamento dos instrumentos e a sua adoção por músicos famosos. Em poucos dias, comentou que a empresa poderia coletar os dados publicados em redes sociais, blogs e sites especializados, a fim de compor uma base de tendência. Também sustentou que eles poderiam avaliar as tendências a partir do acompanhamento de termos de busca em ferramentas globais, como o motor de busca do Google. Lara completou a fala de Joaquim, ao afirmar que, para isso, seria necessário coletar, armazenar e processar uma grande quantidade de dados a partir da Internet. Essa atitude, certamente, impactaria a infraestrutura computacional utilizada, já que seria iniciada uma abordagem de Big Data para a empresa.

Você achou interessante a proposta de Joaquim de avaliar as possíveis tendências encontradas a partir da evolução do uso dos termos em motores de busca, tais como o Google? No entanto, como essa atitude é feita?

A empresa Google disponibiliza uma ferramenta que permite que os seus usuários acompanhem a evolução das buscas feitas a partir do Google Trends (<https://trends.google.com.br>).

Aproveite para explorar um pouco mais os recursos disponíveis nessa ferramenta. Uma sugestão de termo seria “Big Data”: observe como está a evolução de interesse nesse termo, verifique quais são as palavras relacionadas e qual estado brasileiro tem feito mais consultas em relação a essa expressão.

Em detrimento do fato de que existem muitos dados de valor externos à organização, as redes sociais, os blogs, os sites e demais recursos permitem que clientes e consumidores em geral apresentem os seus interesses, necessidades, percepções e opiniões. Ao trazer os dados dessa natureza para as bases da organização, passa-se a ter um ambiente massivo de dados, que também é conhecido como Big Data. Em consequência disso, essa é uma das áreas que mais está em alta nos últimos anos e demanda muitos profissionais qualificados para as empresas. Nesta unidade, serão apresentados mais detalhes sobre essa temática.

Apanhe um bloco de papel e uma caneta para anotar os aspectos que lhe chamaram a atenção a partir da pesquisa feita. Você também pode anotar curiosidades sobre os dados que foram analisados pelo Google Trends. Não deixe de complementar as anotações com tudo o que estudaremos a partir de agora. Ao final, refletiremos sobre o modo como o Big Data pode auxiliar as organizações. Continuemos os nossos estudos!

## DIÁRIO DE BORDO

## O que é *Big Data*?

Anderson, mais uma vez, ficou muito animado com a alternativa apresentada por Lara e Joaquim. Ele se impressionou com a possibilidade de identificar tendências, com o objetivo de antecipar compras ou parcerias e estar preparado para as futuras demandas de seus clientes. Evidentemente, Anderson sabe que esse tipo de abordagem carrega uma taxa de erro embutida, já que diz respeito aos processos de previsões e predições.

A empresa teve um crescimento muito grande e Anderson decidiu vender parte dela para um grupo de empresários. Diante disso, precisa justificar o seu novo investimento em infraestrutura para suportar as novas soluções baseadas em Big Data e pediu para que Joaquim e Lara desenvolvessem um relatório para todos os executivos da empresa. Nesse documento, era preciso esclarecer o que é Big Data, como ele pode beneficiar a organização, como se dá a sua formação e quais são os principais métodos e técnicas. Dessa forma, seria possível adquirir espaço na pauta do conselho, para que, na próxima reunião, fossem tratadas as questões relacionadas ao valor ou ao conjunto de despesas necessárias para a implantação da solução.

Inicialmente, é necessário entender o que é Big Data. O primeiro equívoco é acreditar que se trata apenas de um conjunto gigantesco de dados armazenados ou crer que há uma definição ou entendimento único para esse termo. Para tornar o seu entendimento mais fácil, são apresentadas algumas características incontestáveis a seu respeito:

- Tem bases com grandes quantidades de dados, estruturados ou não estruturados. Não há a possibilidade de armazenamento desses dados por meio de abordagens tradicionais.
- São utilizadas tecnologias, geralmente, com processamento paralelo e distribuído, para que seja possível manipular essa grande quantidade de dados.
- Pode prover importantes repositórios para a ciência de dados desenvolver os seus modelos ou ferramentas analíticas, com o objetivo de apoiar o processo de tomada de decisão.

Alguns autores, tais como Taurion (2013) e Hurwitz et al. (2016), explicam as características comuns dos ambientes de Big Data. Um exemplo são os 3Vs, que abrangem:

- **Volume:** grandes quantidades de dados de qualquer natureza que são armazenados e processados.
- **Velocidade:** os dados são gerados em uma velocidade muito grande. Assim, em um segundo, há vários registros produzidos.
- **Variedade:** múltiplas fontes de origem que apresentam vários formatos para o dado disponibilizado.

Voltamos ao cenário em que Lara e Joaquim estão inseridos. A partir da necessidade de se observar as temáticas relacionadas à música e aos instrumentos musicais, é preciso capturar os dados produzidos nas redes sociais e demais plataformas que tenham a música como objeto de discussão. No caso em questão, o **volume** que será armazenado é muito grande, levando em consideração a quantidade de pessoas que discorrem sobre essa temática nos mais diferentes canais. Além do mais, a **velocidade** de produção de conteúdo é extremamente alta e a **variedade** de formatos produzidos também. Portanto, o cenário em questão é um cenário de Big Data.

Segundo Rautenberg e Carmo (2019), levando em consideração a evolução das Tecnologias da Informação e Comunicação (TICs), é preciso adicionar novos “V’s” ao modelo de 3Vs, passando para o chamado modelo de 6Vs. A figura ao lado apresenta mais detalhes:

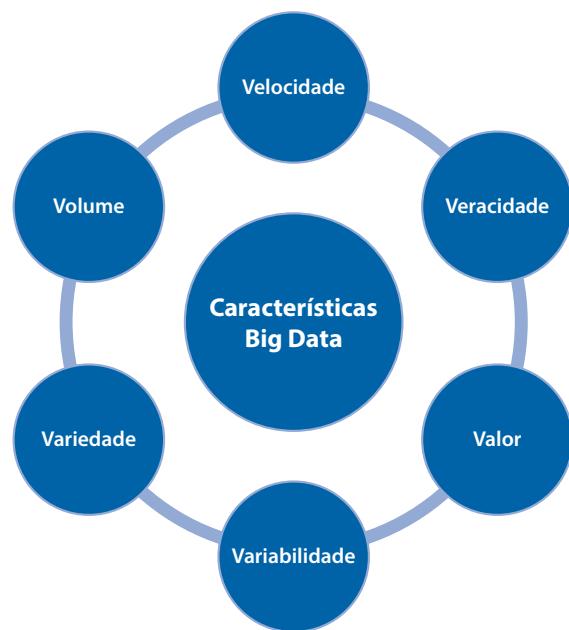


Figura 1 - Características 6Vs do Big Data / Fonte: adaptada de Rautenberg e Carmo (2019) e Akhtar (2018).

**Descrição da Imagem:** a figura tem um círculo central com a seguinte frase: "Características do Big Data". Ao seu redor, estão outros seis círculos, cada um com uma característica, a fim de demonstrar que eles estão relacionados e são pilares do Big Data. As seis características apresentadas são: velocidade, veracidade, valor, variabilidade, variedade e volume.

A seguir, são detalhados os três novos “V’s propostos ao modelo dos 3Vs inicial. Essa proposta foi feita por Akhtar (2018) e aprofundada por Rautenberg e Carmo (2019):

- **Veracidade:** está diretamente ligada a qualidade dos dados que estão sendo inseridos na base de dados massiva da organização. Portanto, é preciso evitar a inserção de dados que não são relevantes ao domínio de aplicação.
- **Variabilidade:** entendimento dos eventos sazonais ou das situações que possam influenciar os dados armazenados. Essa informação é importante para que não sejam gerados ruídos e para que se possa tratar o fenômeno de maneira isolada e adequada.
- **Valor:** é um dos aspectos principais de um projeto de Big Data, que entende que os dados ali armazenados são valiosos para a organização.

Por que foi necessário adicionar novas características às soluções de Big Data?



É interessante refletir sobre a inclusão dos novos 3Vs. Essa atitude demonstra claramente uma mudança de mentalidade nos processos de construção e de uso das soluções de Big Data. Em um primeiro momento, havia uma preocupação muito grande em relação à infraestrutura computacional, para que fosse possível armazenar e processar grandes quantidades de dados produzidos por segundo. Por outro lado, em um segundo momento, tornou-se visível uma preocupação com a obtenção e o armazenamento de um dado de melhor qualidade. Essa alteração ocorreu, visto que foi constatado que dados que não são relevantes geram custos de armazenamento e de processamento e, em muitos casos, levam a tomada de decisão de forma equivocada.

As características “veracidade”, “variabilidade” e “valor” estão muito vinculadas à ideia de se ter um maior controle e conhecimento dos dados. Além disso, evidenciam o quanto valiosos esses dados podem ser para a tomada de decisão, o que se trata de uma informação extremamente importante para justificar os investimentos dessa natureza, já que permite mensurar o ganho de faturamento a partir das novas fontes e ferramentas analíticas.

Fonte: o autor.

Levando em consideração os três novos “V’s utilizados para caracterizar um ambiente de Big Data, Lara e Joaquim fizeram um exercício para caracterizar o seu ambiente com base nessas propriedades. A seguir, são apresentadas as constatações obtidas:

- No que diz respeito à **veracidade**, no contexto do projeto, a ideia é coletar as menções e as opiniões sobre os produtos publicadas por consumidores em redes sociais. Entende-se que a opinião é algo subjetivo, mas representa a percepção pessoal de um indivíduo. Quando essa opinião passa a ser compartilhada por muitos, é estabelecido um padrão. Assim, esse tipo de informação é muito importante para identificar tendências e mudanças de percepção em relação a um produto em especial. Para garantir um pouco mais de veracidade na opinião emitida, são considerados apenas textos e comentários de usuários identificados (as informações pessoais do usuário não são armazenadas).
- Em relação à **variabilidade**, são considerados os eventos sazonais já conhecidos e há uma pretensão em utilizar a análise dos dados coletados para identificar a probabilidade de novos eventos (uma nova feira musical em um determinado mês do ano, por exemplo). Ter essa preocupação desde o início do projeto permite que sejam projetados recursos e ferramentas que monitorem esse tipo de evento e sejam utilizados a favor do processo de tomada de decisão.
- Por fim, no que se refere ao **valor**, é possível relatar o cenário recentemente vivido pela empresa de Anderson, com base nos produtos que foram buscados e não foram encontrados. Dessa forma, pode-se calcular o quanto de dinheiro se deixou de ganhar, além de mensurar quantos casos como esse acontecem e não são divulgados.

Os três itens apresentados estão focados na situação-problema trabalhada, mas é possível levantar inúmeros benefícios diante da coleta e do uso dos dados advindos de redes sociais, blogs e sites sobre produtos musicais. Essas opiniões também podem ser utilizadas para melhorar o processo de mix perfeito por loja e otimizar todo o processo de compra, por exemplo.

Lara revisou todos os aspectos apresentados por ela e por Joaquim até o momento e fez uma provocação: “Joaquim, você não acredita que os diretores podem não visualizar o real benefício dessa abordagem e defenderem a ideia de que ela é equivalente a uma solução de BI?”. Joaquim não havia pensado sob essa ótica e, de fato, era necessário evidenciar quais são os tipos de decisão tomados a partir de cada uma das soluções. A figura a seguir ilustra essa diferença:

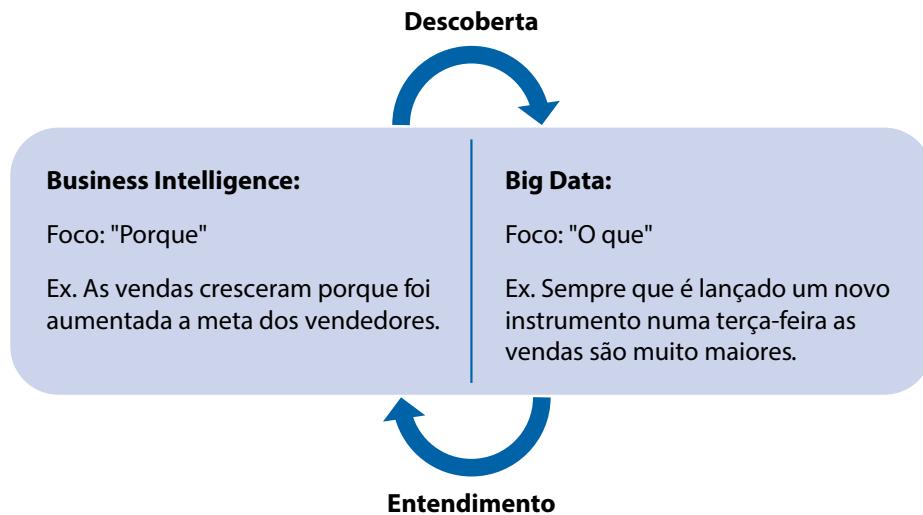


Figura 2 - Diferença de BI para Big Data / Fonte: o autor.

**Descrição da Imagem:** a figura é composta por duas áreas. A primeira, que está no lado esquerdo, refere-se ao Business Intelligence (termo que está de destaque) e, logo abaixo, está escrito: "Foco: 'porque'". Na sequência, é dado o seguinte exemplo: "As vendas cresceram porque foi aumentada a meta dos vendedores". Entre o quadro do Business Intelligence e o próximo, que está no lado direito, há uma seta que contém o termo "Descoberta". O segundo quadro tem o termo em destaque: "Big Data". Na sequência, está escrito: "Foco: 'o que'" e é apresentado o seguinte exemplo: "Sempre que é lançado um novo instrumento em uma terça-feira, as vendas são muito maiores". Também há uma seta que parte do quadro dois para o quadro um e carrega o seguinte termo: "Entendimento".

A primeira grande diferença que existe se dá na mudança de paradigma, uma vez que, nas soluções tradicionais de BI, foca-se no **"porque"**. Já no caso dos ambientes de Big Data, o foco está no **"o que"**, o que retrata uma mudança muito importante. Assim como é possível observar na Figura 2, uma solução de BI pode conviver harmonicamente com um ambiente de Big Data. Além disso, um ambiente de Big Data é propício para a descoberta de novos padrões, fatos e situações, enquanto uma solução de BI pode auxiliar no entendimento de um padrão, fato ou situação.

Nem todas as instituições estão preparadas para tomarem as suas decisões a partir de ambientes de Big Data. Quando se fala em uma mudança de paradigma de "porque" para "o que", há, na verdade, um conjunto de alterações no processo de tomada de decisão que deve ser considerado pelos gestores. Ao se trabalhar com uma solução de BI, pressupõe-se que o dado apresentado na forma de indicador foi processado, validado e é exato, situação que não é encontrada em ambientes de Big Data, que não gera precisão em seus valores. Além do mais, é comumente defendido que as soluções de BI têm uma entrega de valor quantitativa, enquanto as soluções de Big Data proporcionam um entrega de valor qualitativa.

Primeiramente, pode parecer loucura deixar de ter um valor exato para se ter apenas um fato. Todavia, diariamente, todos nós utilizamos abordagens como essa. Um exemplo é o motor de busca do Google. Sempre que é feita uma busca por um termo comum, como “Big Data”, será apresentada uma listagem de páginas e, no rodapé do site, há um número muito grande, o qual representa a quantidade de páginas aproximadas que foram encontradas. A Figura 3 apresenta esse exemplo. Um fato importante a se considerar é o de que sempre que fazemos uma consulta como essa, temos o resultado muito rapidamente. Se fossemos apresentar o número exato de páginas que possuem o termo Big Data em seu conteúdo, muito provavelmente, não teríamos a resposta de nossa consulta em um dia. Em outras palavras, nós aceitamos um número gigantesco enquanto um indicador qualitativo, que nos expressa que existem muitas páginas sobre o assunto, embora saibamos que o valor não é exato.



**Figura 3 – Busca do termo “Big Data” no Google / Fonte:** o autor.

**Descrição da Imagem:** a figura apresenta uma captura da tela do motor de busca do Google. Nela, é exposto o termo: “Big Data”. O resultado da busca tem a seguinte descrição: “Aproximadamente 7.210.000.000 resultados (0,89 segundos)”.

Sempre que nos sentimos confortáveis com um tipo de decisão que não temos o valor exato, mas aceitamos a real confusão dos dados, estamos preparados para utilizar um ambiente de Big Data a nosso favor. Essa foi uma reflexão que Lara e Joaquim fizeram sobre o cenário que estavam vivendo e decidiram, mais uma vez, que deveriam trabalhar desse modo para evoluir a cultura de dados da empresa. Em um curto espaço de tempo, os gestores também se sentiriam confortáveis com a utilização desse tipo de abordagem. Com o objetivo de estabelecermos uma comparação, apresentamos um quadro desenvolvido por Davenport (2014):

	Big Data	Analítico Tradicional
Tipo de dado	Formato não estruturado	Dados formatados em linhas e em colunas

	Big Data	Analítico Tradicional
Volume de dados	100 terabytes a petabytes	Dezenas de terabytes ou menos
Fluxo de dados	Fluxo constante de dados	Pool estático de dados (ETL)
Métodos de análise	Aprendizado de máquina	Baseado em hipóteses
Objetivo principal	Produtos baseados em dados	Supporte ao processo decisório interno

Quadro 1 - Comparativo entre o Big Data e os recursos analíticos tradicionais

Fonte: adaptado de Davenport (2014).

Davenport (2014) atribui uma visão muito computacional de agregação de inteligência aos produtos. Hoje, é possível trabalhar com ambientes de Big Data sem a necessidade de formar uma base de conhecimento ou comportamento inteligente para produtos. Um exemplo é o uso de repositórios analíticos de dados para o estudo e para as pesquisas na área da saúde.

Depois de Lara e Joaquim terem conhecido de forma profunda o conceito e as características de um ambiente de Big Data, eles entenderam que precisavam saber quais seriam as principais bases de origem que eles poderiam coletar para compor o seu novo repositório massivo de dados. Outro aspecto importante que deveria ser trabalhado é o de como que se forma um ambiente de Big Data. Nesse contexto, foram levantadas algumas questões:

- Como se forma um ambiente de Big Data?
- Quais são as principais origens de dados?
- Como posso me beneficiar atuando em elementos que façam a geração de novos dados?

Essas perguntas servirão de guia para a caminhada de Lara e Joaquim.

## Formação de um ambiente de Big Data

O termo “Big Data” começou a aparecer com mais frequência em 2010, visto que passou a ser percebido pelas organizações enquanto uma estratégia para a sua evolução e

posicionamento no mercado. Desde o surgimento da Internet, em suas perspectivas comercial e pessoal, em meados da década de 90, até os dias de hoje, mais de 25 anos depois, muito se evoluiu quanto à Internet e conectividade, de tal forma que só temos os atuais ambientes de Big Data em consequência desses avanços.

Uma importante mudança que ocorreu com o uso da Internet, de modo geral, deu-se com o surgimento das plataformas de publicação de conteúdos, já que os usuários da Internet deixaram de ser consumidores de dados e de informações e passaram a ser produtores também. Quem é mais novo, talvez não se lembre ou não tenha vivenciado, mas, durante os anos da primeira década da Internet (com o seu uso comercial), só era possível ter um conteúdo publicado se você soubesse desenvolver páginas em HTML, tivesse um servidor de aplicação disponível e fizesse todo o processo de *deploy*, ou seja, deveria ser uma pessoa com conhecimento técnico. No entanto, com o surgimento das plataformas de conteúdo, foi democratizado o uso da Internet e permitida a publicação de textos, imagens e músicas de forma fácil.

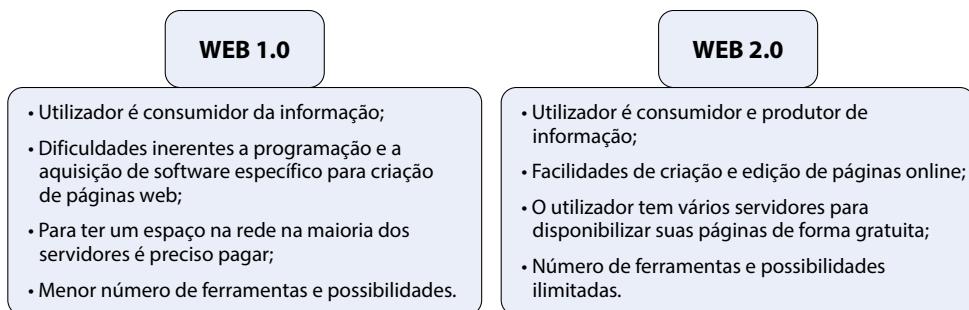
O surgimento dessas plataformas foi tão importante que chegou a ser cunhado o termo “Web 2.0”, que retrata essa mudança de papel dos usuários da Internet, os quais são dotados de muito mais ação e podem facilmente postar e compartilhar os seus conteúdos. A figura a seguir apresenta as suas principais características:



**Figura 4 - Características da Web 2.0 / Fonte:** adaptada de Blattmann e Silva (2007).

**Descrição da Imagem:** a figura apresenta um modelo em que há um retângulo no centro com o seguinte texto: “Posicionamento estratégico (a web como plataforma). Competências-chave: arquitetura de participação e aproveitamento da inteligência coletiva”. Ao redor desse retângulo, existem sete retângulos com os seguintes conteúdos: “Ênfase nos usuários, e não na tecnologia”; “O comportamento do usuário não está predeterminado”; “O software melhora à medida que mais pessoas o utilizam”; “Fim do ciclo de liberação do software”; “Modelos leves de programação”; “Experiência e conhecimento pessoal dos usuários”; e “Confiar em seus usuários”.

Assim como é visualizado no centro da Figura 4, a web passou a ser tratada como plataforma. Por isso, foi possível formular uma arquitetura de participação em que recursos, tais como blogs e wikis, permitiam a publicação de conteúdos e que outros usuários também pudessem dar a sua opinião, construindo, assim, uma cultura de colaboração e de construção coletiva. A partir da Web 2.0, os usuários passaram a produzir muito conteúdo em seus blogs, a postar fotos e, principalmente, emitir opiniões sobre produtos, serviços e empresas em geral. São exemplos atuais de plataformas de Web 2.0: Facebook, Twitter, Instagram e LinkedIn. A figura a seguir ilustra as principais diferenças entre os recursos da web clássica (ou Web 1.0) em relação à Web 2.0:



**Figura 5 - Diferenças entre a Web 1.0 e Web 2.0 / Fonte: adaptada de Silva e Leão (2009).**

**Descrição da Imagem:** a figura apresenta dois quadros. O da esquerda tem o seguinte título: "Web 1.0" e apresenta as seguintes características em seu interior: o utilizador é consumidor da informação; dificuldades inerentes à programação e à aquisição de um software específico para a criação de páginas web; para ter um espaço na rede, na maioria dos servidores, é preciso pagar; menor número de ferramentas e possibilidades. Já no quadro localizado no lado direito, o título é: "Web 2.0" e as características que são apresentadas em seu interior são: o utilizador é consumidor e produtor de informação; facilidades de criação e na edição de páginas on-line; o utilizador tem vários servidores para disponibilizar as suas páginas de forma gratuita; número de ferramentas e possibilidades ilimitadas.

A partir da implantação das plataformas de conteúdo na Web 2.0, o número de publicações e conteúdos gerados cresceu exponencialmente e rapidamente. Diante disso, as organizações passaram a constatar o valor que estava sendo depositado nas bases de dados dessas plataformas e muitas delas passaram a utilizar essas informações desenvolvidas para apoiar o seu processo de tomada de decisão. Com essa nova percepção de valor, surgiram áreas de pesquisa e desenvolvimento, como as chamadas "Análise de Sentimentos", "Business Intelligence 2.0", dentre outras.

**EXPLORANDO IDEIAS**

**Análise de sentimento:** a área de análise de sentimento é uma evolução (ou especialização) da área de mineração de opiniões. Objetiva, a partir de textos escritos em uma linguagem natural, analisar o seu conteúdo, a fim de concluir se o texto é positivo ou negativo. As aplicações de análise de sentimento evoluíram para a identificação de produtos e de serviços pelo texto e apresentam a classificação (positiva ou negativa) sob a ótica de quem publicou o texto. Essa é uma importante informação para apoiar na compreensão da imagem de um produto ou de um serviço e para se verificar como está a imagem organizacional.

**Business Intelligence 2.0 (BI 2.0):** evolução da arquitetura tradicional de BI. Nela, são coletados dados não estruturados (internos ou externos) da organização, com o objetivo de gerar indicadores e análises complementares que auxiliem no processo de tomada de decisão por parte dos gestores.

Fonte: o autor.

Nesse momento da pesquisa, Lara e Joaquim já estavam muito empolgados com as possibilidades de uso e aplicação de dados com a utilização dos recursos da Web 2.0. Ambos claramente entenderam que muitos ambientes de Big Data são formados dentro das organizações por intermédio do uso de dados dessas fontes, já que eles são produzidos em uma velocidade gigantesca. Imagine a possibilidade de coletar dados sobre música e instrumentos musicais levando em consideração um conjunto de aplicações da Web 2.0 simultaneamente por segundo? Ao final de um ano, teríamos uma massa de dados muito grande e com a possibilidade de aplicações como as que já foram apresentadas.

Outro importante avanço que aconteceu em paralelo à da Web e contribuiu muito para a explosão dos dados publicados foi a evolução das tecnologias móveis. É possível pensar nessa evolução sob duas óticas: a do dispositivo móvel e a do padrão de comunicação móvel (tipo de redes). Quando observamos a evolução dos dispositivos móveis, lembramo-nos do uso do celular durante a década de 90. Naquela época, ele era dotado apenas da função de telefone. Entretanto, no final da década, surgiram aparelhos com jogos, recursos de agenda e a possibilidade de envio de mensagens SMS. A partir dos anos 2000, os celulares já tinham telas maiores e mais algumas funcionalidades. Em 2005, já eram encontrados aparelhos que utilizavam sistemas operacionais multitarefas, ou seja, que permitiam fazer duas tarefas ao mesmo tempo, com aplicativos para acesso

aos e-mails e outros recursos. Com a evolução e o surgimento do iPhone (Apple) e do sistema de operação Android (Google), foi proporcionada uma experiência mais rica no que diz respeito ao uso de aplicações e naveabilidade na web. Esses recursos já nasceram com a premissa de que o celular estaria conectado a uma rede de dados e faria o uso dela.

A partir do momento em que você tem um celular multitarefa, com poder de processamento e conectado à uma rede rápida de dados, o processo de produção de conteúdo passa a ser muito maior: você tira uma foto e a posta na Internet em conjunto com um texto. Você pode ir até a um restaurante e, durante a visita, relatar a sua experiência em uma plataforma especializada. Você também pode ir de um ponto a outro da cidade, observar as informações sobre o trânsito e publicar informações sobre o seu trajeto. Todos esses recursos proporcionaram um grande aumento no processo de geração de conteúdo, o que potencializou ainda mais os dados armazenados pelas plataformas da Web 2.0 e permitiu que as organizações constatassem ainda mais o valor de uso desses tipos de dados. A figura a seguir apresenta uma estimativa feita para o crescimento dos dados até 2020:

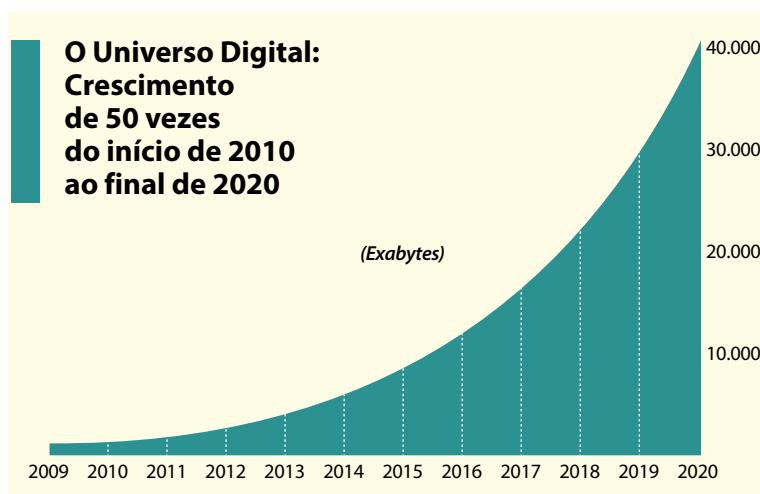


Figura 6 - Crescimento de dados produzidos de 2010 a 2020

Fonte: Isotani e Bittencourt (2015, p. 24).

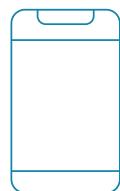
**Descrição da Imagem:** na figura, é apresentado um gráfico de barras. Nas linhas, estão presentes os anos entre 2009 e 2020. O gráfico demonstra que a quantidade de dados aumentou 50 vezes de 2010 a 2020.

A Figura 6 expressa uma grande crescente na produção de dados na Internet. É possível afirmar que os principais motivos para esse crescimento já foram apresentados até aqui, mas existem outros movimentos que também apoiaram esse fenômeno, como a publicação de dados abertos. A expressão **dados abertos** não é nova, mas apareceu em 1995 em um relatório emitido pela agência científica americana. Contudo, de fato, ganhou popularidade nos anos 2000, quando o presidente norte americano Barack Obama emitiu um memorando que tratava dos dados abertos governamentais, o que fez com que muitos países pelo mundo seguissem a mesma ideia: a de transparecer os seus números e indicadores (ISOTANI; BITTENCOURT, 2015). O objetivo dos dados abertos, assim como o próprio nome diz, é dar total acesso e transparência aos dados. São muito utilizados pela ciência, a fim de permitir com que outros pesquisadores reproduzam experimentos.

Lara e Joaquim, ao terminarem a leitura sobre dados abertos, ficaram curiosos e se questionaram onde poderiam encontrar dados abertos sobre o Brasil e quais tipos de dados estariam disponíveis. Assim, descobriram as seguintes informações: em 2011, o Brasil aprovou a sua legislação sobre a transparência dos dados públicos em forma de dados abertos. No entanto, antes disso, em 2004, já havia sido lançado o Portal de Transparência do Governo Federal, inicialmente focado na publicação dos gastos públicos (NEVES, 2013).



Hoje, mediante o Portal Brasileiro de Dados Abertos (<https://dados.gov.br>), é possível acessar dados de várias naturezas distintas. São exemplos: economia, finanças, educação, saúde, meio ambiente, geologia, indústria, dentre outros.



De posse dessas informações, Lara e Joaquim pensaram na possibilidade de coleta e uso dos dados públicos. Nesse sentido, chegaram à seguinte lista:

- **PIB (Produto Interno Bruto):** essa informação pode apoiá-los no processo de entendimento das áreas que já têm lojas instaladas, além de dar indicadores de onde abrir as próximas com base no perfil dos clientes mapeados.

- **Escolas de música:** dados disponibilizados pelo Ministério da Cultura. Podem ser utilizados para várias ações, desde a abertura de lojas até a definição do mix de produtos.
- **Bandas de música:** dados disponibilizados pelo Ministério da Cultura. Podem ser utilizados para várias ações, desde a abertura de lojas até a definição do mix de produtos.
- **Festivais e eventos:** dados disponibilizados pelo Ministério da Cultura. Podem ser utilizados para várias ações, desde a abertura de lojas até a definição do mix de produtos.
- **Espaços musicais:** dados disponibilizados pelo Ministério da Cultura. Podem ser utilizados para várias ações, desde a abertura de lojas até a definição do mix de produtos.
- **Associações e sociedades de música:** dados disponibilizados pelo Ministério da Cultura. Podem ser utilizados para várias ações, desde a abertura de lojas até a definição do mix de produtos.

Há um conjunto de arquivos sobre os dados relacionados aos indicadores macroeconômicos. Eles também podem ser utilizados para apoiar o processo de tomada de decisão, mas, neste momento, foram essas as bases selecionadas para a coleta e a utilização de dados nos processos internos da empresa. Lara e Joaquim formularam um roadmap para adicionar as bases disponíveis em forma de dados abertos que podem ser utilizadas nos processos internos da empresa. Os dois estavam satisfeitos com o que já haviam mapeado no que diz respeito aos dados externos. Foi assim que se depararam com uma matéria de uma revista sobre economia que chamou a atenção de ambos: ela dizia a respeito da Internet das Coisas (em inglês, “*Internet of Things*” - IoT) e a sua relação com o Big Data.

Segundo Sinclair (2018), a Internet das Coisas ou simplesmente IoT é uma evolução da Internet. Nela, as questões relacionadas à disponibilidade e à velocidade estão controladas, o que permite que sejam adicionados inúmeros sensores e atuadores que apoiam o monitoramento e a execução de ações. A figura a seguir apresenta detalhes dos elementos envolvidos em uma solução de IoT:

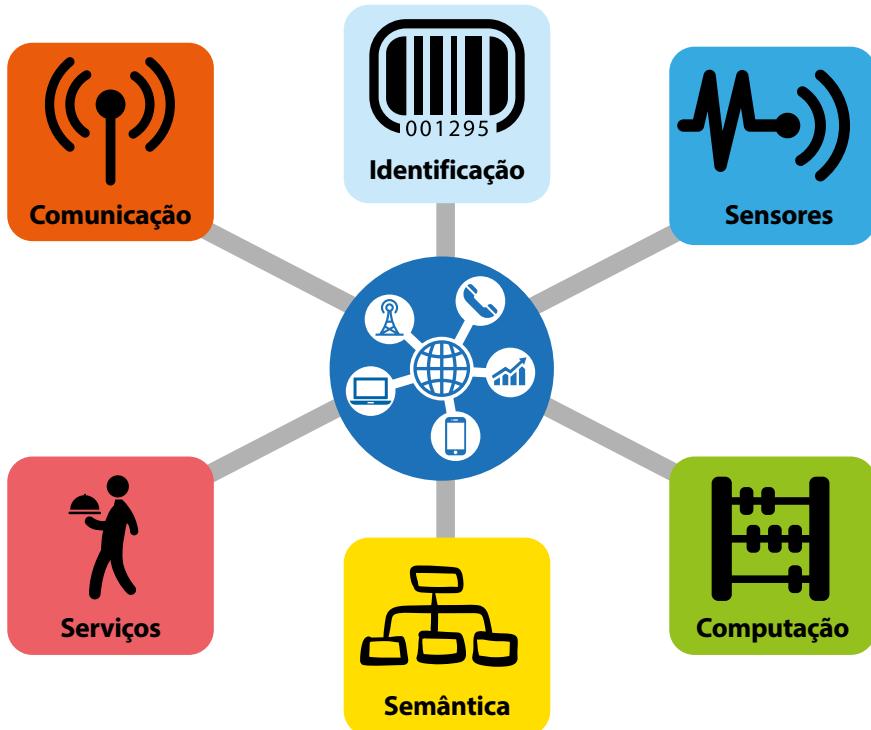


Figura 7 - Blocos básicos da IoT / Fonte: Santos et al. (2016, p. 6).

**Descrição da Imagem:** na figura, há, no centro, uma solução de IoT que está ligada a alguns ícones com os seguintes dizeres: comunicação, identificação, sensores, computação, semântica e serviço.

Santos et al. (2016) descrevem os principais blocos básicos para a construção de soluções da IoT. São eles:

- **Identificadores:** tecnologias que permitem identificar recursos, pessoas, sensores e demais itens relacionados a uma solução de IoT. Sem a identificação dos elementos envolvidos, não é possível ter comunicação.
- **Sensores/atuadores:** os sensores são os responsáveis por fazer o processo de coleta de dados do ambiente em que se está inserido. Os atuadores podem interagir com o ambiente em que se está inserido por intermédio da execução de uma ação.
- **Comunicação:** está relacionado às diversas técnicas utilizadas para manter a conexão dos objetos inteligentes que estão envolvidos na solução de IoT em questão.

- **Computação:** responsável por fazer todo o processamento de dados e gerar as ações a partir do uso de técnicas de inteligência artificial, por exemplo. É nesse bloco que estão as bases massivas de dados de sensores que alimentam uma série de algoritmos focados no apoio à tomada de decisão.
- **Serviços:** são os serviços que, a partir do bloco da computação, podem fazer as entregas de informação e de valor aos usuários e demais recursos que estejam conectados a essa solução.
- **Semântica:** está diretamente ligado à capacidade de extração de conhecimento dos elementos envolvidos na solução de IoT, a fim de apoiar os processos de contextualização e dentre outros recursos semânticos.

As soluções de IoT representam um meio para uma nova Revolução Industrial, a chamada “Indústria 4.0”. Com ela, a linha de produção passará a ser inteligente, já que tem vários sensores e atuadores para tratar as várias situações possíveis de ocorrerem dentro de uma linha de produção. Além disso, muitos dados são coletados e gerados por segundo, o que permite a aplicação de algoritmos de aprendizagem de máquina para apoiar na descoberta de novos padrões de produção, a recomendação de melhorias no processo de produção e a formulação de alertas inteligentes. Para que todos esses recursos possam ser viáveis para a indústria, é necessário que todos os dados sejam armazenados em repositórios massivos, a fim de que possam ser recuperados e utilizados a partir de soluções de Big Data.

Com o avanço das soluções de IoT, o conceito de cidades inteligentes (smart cities) ficou muito mais próximo da realidade. A ideia preconizada pelas cidades inteligentes é a da constituição de cidades vivas, conectadas, visto que existem vários sensores espalhados e, a partir da colaboração dos cidadãos, muitos dados são coletados por segundos e muitas ações podem ser tomadas, com o objetivo de melhorar o bem-estar de todos. Existem outros pilares para o conceito de cidades inteligentes, mas eles não serão tratados aqui. O fato é que muitos desses dados também são compartilhados no formato de dados abertos, o que pode apoiar ainda mais as organizações na construção dos seus repositórios massivos de dados.

Lara e Joaquim perceberam todo o potencial que a implantação de soluções de IoT terão para a indústria, para o varejo e para as organizações em geral, mas entendem que, hoje, não é o momento ideal para iniciar uma frente de desenvolvimento dentro da organização, pois existem questões mais prioritárias a serem focadas. Entretanto, compreendem que, muito em breve, as soluções de IoT serão uma realidade

em seu cotidiano e o seu ambiente deve estar preparado para isso, o que reforça o argumento a ser apresentado para a diretoria no que diz respeito à importância de se desenvolver uma infraestrutura para suportar um ambiente de Big Data.

Carolina, que era a gerente de projetos de aplicativos para dispositivos móveis da empresa e já havia contribuído com o projeto, ao perguntar para Joaquim e Lara sobre a utilização de modelos de maturidade analítica, contribui ainda mais com o trabalho de nossa querida dupla: Carolina se lembrou de que um dos recursos previstos para o aplicativo da loja é o de registrar toda a interação do cliente, a fim de complementar o seu perfil, ao levantar informações sobre interesses. Outro aspecto que será registrado se refere às coordenadas de GPS de uso do aplicativo. Essas informações podem ser utilizadas para gerar alertas de promoções personalizadas e para identificar os locais em que os clientes passam e se lembram da loja, o que pode ser uma informação utilizada para a área de expansão das lojas físicas.

Com a observação feita por Carolina, ficou ainda mais evidente a necessidade de se ter uma infraestrutura de Big Data, tendo em vista que muitos dados serão coletados pelo aplicativo e devem estar disponíveis para as áreas analíticas da organização, para que, assim, complementem a construção de novas ferramentas e análises. Lara e Joaquim já tinham todos os argumentos a serem apresentados para a diretoria e sabiam que seria solicitada uma prévia do investimento. Para isso, seria necessário entender melhor quais são os métodos e as técnicas que são aplicados em um ambiente de Big Data. Em um futuro próximo, eles poderiam levantar os principais fornecedores de tecnologia para apoiá-los na implantação desse ambiente. Esse é o próximo destino que os dois percorrerão. Estão cada vez mais próximos a aprovação e o desafio da implantação de todo o ambiente de Big Data.

## Métodos e técnicas para Big Data

Lara e Joaquim levantaram materiais sobre os possíveis métodos e técnicas utilizados em soluções de Big Data. O primeiro desafio foi o de não selecionar tecnologias ou ferramentas, pois muitos materiais que estão disponíveis na Internet sobre esse tema focam demasiadamente nas ferramentas, o que pode ser um grande risco, tendo em vista que a área está em constante evolução. Nesse sentido, as ferramentas, caso não tenham um ciclo de evolução rápido, podem não ser mais

relevantes para a problemática. Já os métodos e as técnicas dão subsídios para se pensar em uma proposta de solução mais robusta e genérica, o que permite que, em um segundo momento, seja feita a seleção das ferramentas tecnológicas.

Para guiar o entendimento em relação aos principais métodos e técnicas utilizados nas soluções de Big Data, Lara e Joaquim se basearam na chamada “stack de Big Data”. A figura a seguir apresenta mais detalhes:

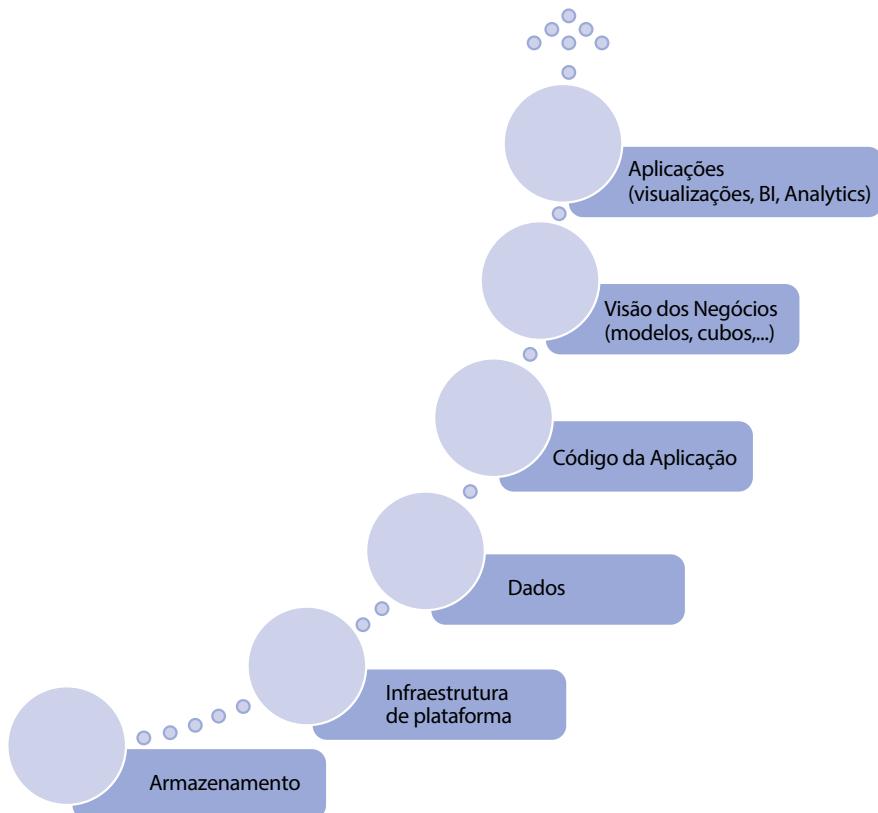


Figura 8 - Stack de Big Data / Fonte: Davenport (2014, p. 117).

**Descrição da Imagem:** a figura apresenta as camadas de uma solução de Big Data. A camada de base está relacionada com o armazenamento. Na sequência, há a camada de infraestrutura de plataforma, seguida pela camada de dados. Depois, há a camada de código da aplicação, seguida pela camada de visão dos negócios. Nela, há, entre parênteses, os seguintes exemplos: modelos e cubos. Por fim, na camada superior, há as aplicações e, entre parênteses, estão: visualizações, BI e analytics.

A figura apresenta as várias camadas que compõem uma solução de Big Data. A seguir, elas são explicadas detalhadamente:

- **Armazenamento:** geralmente, trabalha com modelos de dados NoSQL e com sistemas de arquivos distribuídos para garantir performance.
- **Infraestrutura de plataforma:** é a camada de software que gerencia os vários serviços e recursos computacionais. Normalmente, eles estão hospedados na nuvem.
- **Dados:** processos relacionados ao ciclo de vida do dado. São exemplos: a inclusão e a consulta de dados armazenados.
- **Código da aplicação:** são os códigos e os algoritmos utilizados para transformar os dados armazenados.
- **Visão dos negócios:** concentra e aplica as regras de negócio. É possível disponibilizar os dados com o uso de elementos do domínio de aplicação para apoiar o consumo dos dados pelas áreas de negócio.
- **Aplicações:** são os vários serviços e dashboards que podem ser ligados ao ambiente de Big Data para se obter a entrega de valor direto para as áreas de negócio e para os tomadores de decisão.

Ao analisarem a figura e as camadas apresentadas, Lara e Joaquim decidiram iniciar o seu mapeamento pelos métodos e pelas técnicas que envolvem o armazenamento de dados. A primeira dúvida que tiveram foi a seguinte:



Será que a utilização de Data Warehouse é a melhor alternativa para receber os dados coletados de várias fontes distintas e inseridos em uma solução de Big Data?

Nesse momento, a nossa dupla começou a refletir sobre as premissas de uma solução baseada em Data Warehouse em relação a uma solução de Big Data. Um dos aspectos principais na construção de um Data Warehouse é a utilização de um processo de ETL e uma das grandes responsabilidades da fase de “transformação” é justamente validar os dados e aplicar as regras de negócio, a fim de garantir que os dados inseridos estejam limpos e validados. Quando consideramos um ambiente de Big Data, milhares de novos dados podem ser gerados por segundo, o que tornaria inviável um processo de limpeza e aplicação de regras de negócio durante a fase de ingestão dos dados.

Além disso, lembraram-se de que, em sua grande maioria, os Data Warehouses são organizados a partir do uso da chamada “modelagem dimensional”, que procura trabalhar com os dados agregados e organizados em dimensões, assuntos (fatos) e indicadores (medidas). Ao se recordarem das características de um ambiente de Big Data, perceberam que, mais uma vez, em detrimento do grande volume que pode ser inserido por segundo, não seria possível sumarizar e processar todos os dados para compor um modelo dimensional. Portanto, essa realmente não seria uma solução factível para o cenário em questão.

Ao buscarem por repositórios massivos de dados enquanto uma alternativa ao uso dos Data Warehouse, chegaram ao conceito de **Data Lake**, que, traduzido do inglês, significa “lago de dados”. Quando pensamos em um lago, a primeira característica que vem até a nossa mente é a de sua amplitude. Geralmente, um lago é composto por estruturas que armazenam uma grande quantidade de água e podem ter profundidades medianas. Ao estabelecermos um paralelo com um ambiente de dados, podemos considerar os seguintes aspectos:

- Um Data Lake tem uma grande quantidade de dados armazenados.
- Não existe um padrão bem definido para os dados inseridos e, em consequência disso, existe uma amplitude muito grande de assuntos e domínios tratados.
- Dados de diferentes tipos e natureza são inseridos sem um tratamento prévio.
- É possível aplicar camadas para dar contexto e aplicabilidade aos dados armazenados.

Segundo Pires (2017), o Data Lake corresponde a uma arquitetura moderna de dados que objetiva complementar os repositórios de dados das organizações, ao armazenar todas as informações de negócio e ao lidar com vários tipos de dados a partir do uso de técnicas distribuídas e de processamentos massivos. Além do mais, representa uma importante evolução nas arquiteturas de dados existentes.

Nossa dupla percebeu que, na verdade, um Data Lake não é uma alternativa a um Data Warehouse, mas, na verdade, ambos carregam várias diferenças. Portanto, devem ser utilizados em situações distintas. Pires (2017) apresenta um quadro comparativo entre as duas soluções levando em consideração dez dimensões:

Dimensões	Data Warehouse	Data Lake
Utilizadores	Centenas até milhares de utilizadores em concorrência – Utilizadores de empresas	Alguns utilizadores – Cientistas de dados
Workload	Processamento Batch	Processamento Batch em larga escala
Esquema	Schema on Write – esquema definido antes dos dados serem armazenados, para posteriormente serem escritos	Schema on Read – esquema definido depois dos dados serem armazenados
Escala	Pode ser expansível para grandes volumes de dados mas com um custo moderado	Desenhado para obter um baixo custo de armazenamento
Métodos de acesso	Os dados são acedidos através de SQL e ferramentas de BI, suportadas por sistemas de reporting e analytics	Os dados são acedidos através de programas desenvolvidos por programadores, sistemas baseados em SQL e outros métodos
SQL	ANSI SQL, Propriedades ACID	Programação flexível envolvendo SQL
Dados	Dados estruturados / transformados	Dados “as-is” sem qualquer transformação. Estruturados, semiestruturados, não estruturados
Acesso	Seeks	Scans
Complexidade	Joins complexos	Processamento complexo
Segurança	Madura	Em desenvolvimento
Custo / Eficiência	Uso eficiente de CPU/IO	Baixo custo de armazenamento e processamento

Quadro 2 - Diferenças entre um Data Warehouse e um Data Lake / Fonte: Pires (2017, p. 10-11).

É evidente que as soluções de Data Lake e Data Warehouse podem ser utilizadas em conjunto e de maneira sinérgica, de modo que cada uma focará em um tipo de dados e será utilizada por processos analíticos distintos. Essa abordagem pode ser chamada de solução híbrida e a figura a seguir apresenta mais detalhes sobre esse tipo de solução:



Figura 9 - Soluções híbridas de Data Warehouse e Data Lake / Fonte: Pires (2017, p. 12).

**Descrição da Imagem:** a figura tem dois quadrados. O primeiro é intitulado “Data Warehouse” e possui as seguintes características: dados estruturados, suporte em tempo real, capacidades analíticas e governança de dados. No segundo quadrado, há o título “Data Lake” e as seguintes características: dados multi estruturados, conjunto de dados completos e perspectiva histórica/contexto. Os dois quadrados apontam para um círculo central, intitulado “Solução híbrida”. Ele tem as seguintes características: dados multi estruturados, conjunto de dados completos, capacidades analíticas e governança de dados.

Lara e Joaquim entendem que precisam ter as duas soluções em uma perspectiva híbrida para apoiar os processos analíticos que já estão instalados e funcionando, como as soluções de Business Intelligence, que já têm uma grande entrega de valor para a empresa, mas estão abertas para os dados massivos e externos à organização. A partir do entendimento da chamada **solução híbrida**, que mantém as características principais das soluções baseadas em Data Warehouse e Data Lake, ambos passaram a desenhar uma arquitetura para a infraestrutura de dados que suportasse o legado analítico, permitindo a construção de novas dimensões e fatos a partir dos dados vindos do Data Lake e a disponibilização dessa essa infraestrutura aos processos de análises exploratórias e rotinas de ciências de dados.

A partir de algumas discussões, foi obtido um desenho que demonstra as camadas lógicas do repositório analíticos de dados. Eles o chamaram de Data Lake, que é composto por três camadas: uma chamada “estagiamento” (*staging*), outra denominada “acesso” e última “de indicadores”. A figura a seguir apresenta a organização das camadas lógicas:

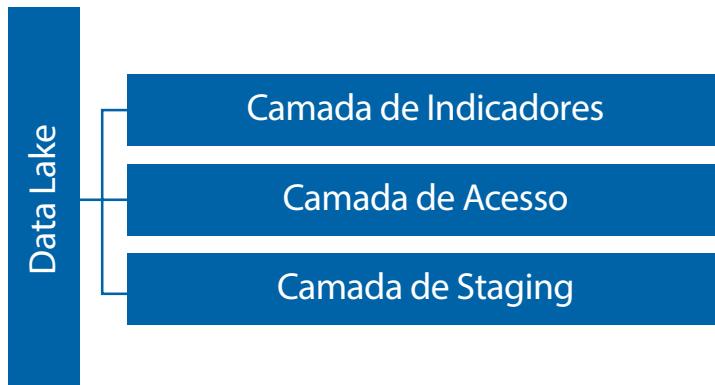


Figura 10 - Camadas lógicas de uma solução de Data Lake / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta um retângulo na vertical com o título “Data Lake”. Esse retângulo está ligado a outros três retângulos horizontais. O inferior tem o título “Camada de Staging”, o do meio é chamado de “Camada de Acesso” e o superior é intitulado “Camada de Indicadores”.

Cada camada do Data Lake tem uma responsabilidade e uma forma de implementação bem definida. Lembre-se de que essa é uma visão lógica da proposta de solução. Quando partir para a implementação, é possível ter mais de uma tecnologia em uma camada e uma tecnologia sendo compartilhada por camadas lógicas distintas. A seguir, as camadas são descritas com detalhes:

- **Camada de Staging:** é a camada inferior da arquitetura lógica proposta. Nela, são armazenados os dados brutos à medida que são coletados. Ela não é acessível por agentes de dados humanos, apenas por sistemas de carga e pelo administrador do ambiente. Além disso, pode ser desenvolvida utilizando criptografia dos dados de base ou por compactação (dependendo do tipo de tecnologia escolhida). É possível ter todos os dados externos coletados e uma réplica dos dados gerados a partir dos sistemas internos da organização.
- **Camada de Acesso:** reside os dados coletados e armazenados na camada de staging, já processados e modelados de acordo com os domínios da empresa. Eles podem ser modelados utilizando o conceito de Master Data Management (MDM). Essa é a camada que os analistas de dados consumem para gerar os seus relatórios de operação e gerenciais. Nela, é possível gerar bases exclusivas para a área da ciência de dados construir os seus estudos e modelos de apoio à decisão.
- **Camada de Indicadores:** representa a camada superior da arquitetura de dados proposta. Tradicionalmente, é desenvolvida utilizando uma mode-

lagem dimensional e, geralmente, é onde fica o Data Warehouse da organização. Caso ela já tenha um Data Warehouse criado, é preciso adequar os processos de ETL para suportar as validações feitas entre as camadas do Data Lake e considerar os novos dados coletados para as novas dimensões, medidas e fatos. É comumente acessada pelas soluções de Business



### EXPLORANDO IDEIAS

Master Data Management (MDM) ou gerenciamento de dados mestre é a capacidade de organizar os principais conceitos que existem dentro de um negócio, de modo que sejam claras as tabelas e as colunas que são as principais fontes para a representação desses conceitos. A partir desse mapeamento, é possível ter uma base de dados única para representar os principais conceitos a partir da retirada dos dados dos locais corretos e da realização das devidas aplicações de regras de negócios.

Fonte: o autor.

Intelligence.

Para se ter uma solução de MDM na camada de acesso de um Data Lake, é necessário ter uma governança de dados implantada. Outra observação importante é a de que o MDM pode ser utilizado para apoiar os processos de ETL para a carga da camada de indicadores.

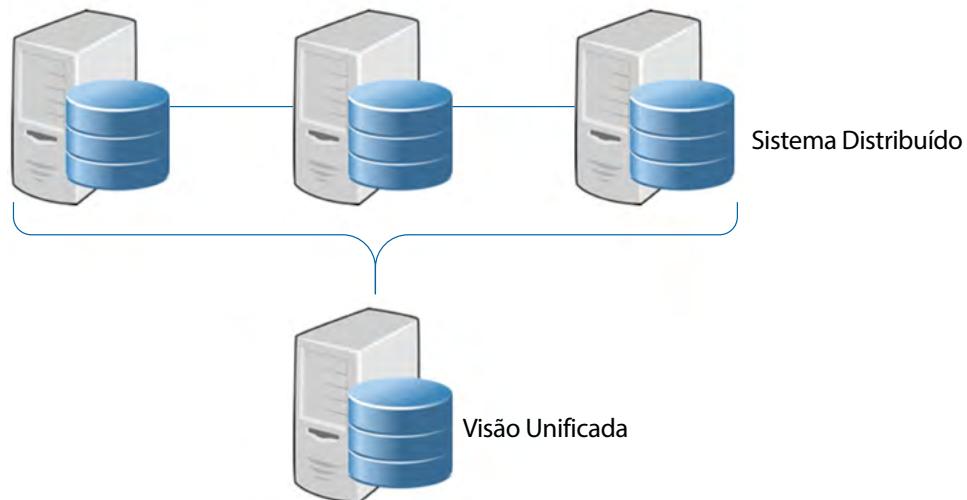
Obtidas as conclusões expostas e diante do desenho da arquitetura lógica da infraestrutura de dados formulado, surgiram mais algumas dúvidas:

- Como garantir que a estrutura da camada de estagiamento mantenha quantidades de dados realmente grandes?
- Como garantir que todo os dados sejam processados em tempo hábil para a tomada de decisão?

Lara e Joaquim estavam muito próximos de concluir o seu levantamento em relação aos métodos e às técnicas que devem ser aplicados em um ambiente de Big Data. Com base nos elementos apresentados na Figura 8, eles ainda tinham aspectos a serem verificados enquanto armazenamento e infraestrutura de plataforma, a fim de garantir o processamento massivo dos dados armazenados. Exploraram alguns livros e realizaram pesquisas em sites especializados até encontrarem um

conceito-chave: soluções distribuídas. Nesse contexto, são aplicados tanto os conceitos de processamento distribuído quanto o sistema de arquivo distribuído.

Quando falamos em um sistema distribuído, referimo-nos ao fato de ele rodar em máquinas (nós) distintas, ou seja, existem vários recursos rodando uma mesma solução em paralelo. Isso permite que a infraestrutura cresça à medida que for necessário, pois basta adicionar mais máquinas para que haja um crescimento horizontal, considerando tanto o contexto de armazenamento quanto de processamento. Os sistemas distribuídos apresentam uma visão unificada, como se fossem apenas uma máquina e um sistema de arquivos. Isso facilita todo o processo de execução e gestão do dado na base de dados em questão. A figura a seguir apresenta um exemplo desse cenário:



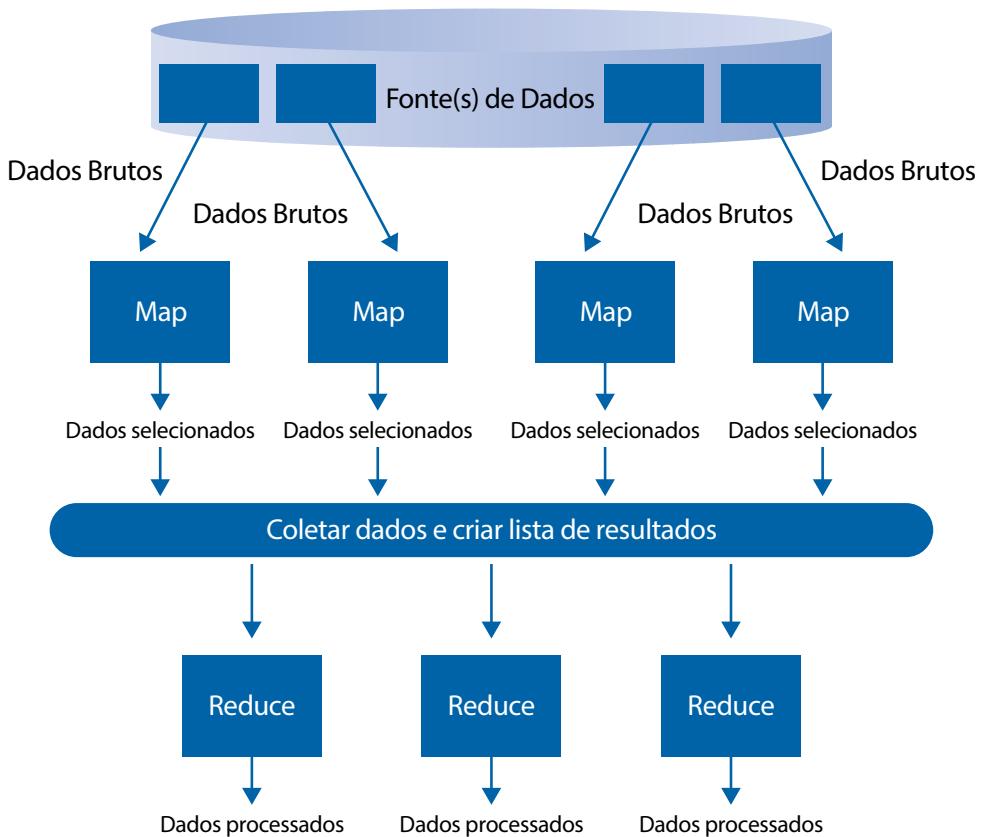
**Descrição da Imagem:** a figura mostra três servidores com a representação de um banco de dados em cada um deles, os três servidores estão ligados entre si, e existe um rótulo: Sistema Distribuído, esses três servidores estão dentro de uma chave que dá a ideia que as três máquinas estão contidas no quarto servidor que possui o rótulo: Visão Unificada.

**Figura 11 - Exemplo de uma arquitetura distribuída / Fonte: o autor.**

A figura mostra um conjunto de servidores que tem estruturas físicas distintas, mas que são utilizadas a partir de uma visão unificada, ou seja, como se houvesse apenas um servidor com somente um sistema de arquivo. Para se obter essa visão unificada, é necessário conhecer a ideia de sistemas de arquivos distribuídos. Quando falamos de um sistema de arquivo, referimo-nos a uma estrutura básica dentro de um sistema operacional de um computador que faz a gestão dos dados a nível de armazenamento do disco rígido (ou outra tecnologia de armazenamento utilizada).

Ele sabe o local em que o arquivo está armazenado, como fazer o seu acesso facilitado e demais recursos. Já quando nos referimos a um sistema de arquivo distribuído, estamos tratando de um sistema que já tem todos os recursos de um sistema de arquivo, mas analisa vários sistemas de arquivos distintos armazenados em servidores diferentes e que são acessíveis como se fosse apenas um sistema de arquivo.

Ao se aprofundarem ainda mais nessa temática, Lara e Joaquim chegaram a um importante conceito: o MapReduce. A figura 12 apresenta mais detalhes sobre o fluxo de dados no MapReduce:



**Descrição da Imagem:** a figura apresenta, em sua parte superior, uma fonte de dados. A partir dos seus dados brutos, é passada por um processo de Map (na imagem, são apresentados quatro processos de Map em paralelo), cuja saída são os dados selecionados que são agregados em uma barreira com o nome: "Coletar dados e criar lista de resultado". A partir dessa barra, três processos de reduce consomem dados em paralelo, gerando dados processados.

Figura 12 - Fluxo de dados em MapReduce / Fonte: Hurwitz et al. (2016, p. 106).

Na figura 12, foi apresentada a entrada do fluxo e as múltiplas bases de dados que são encontradas, uma espécie de grande repositório. Dessas bases, são extraídos os dados brutos e, a partir do processo de “Map”, são selecionados os que fazem sentido para o domínio em questão. A partir desse processo, os dados são inseridos em uma barra e consumidos pelos processos de reduce, em que são feitas as possíveis seleções e aplicações de regras de negócio. Depois, há a obtenção dos dados já processados.

Segundo Hurwitz et al. (2016), a disposição de aplicativos em soluções distribuídas é um dos principais desafios de um ambiente de Big Data. Nesse contexto, o MapReduce é uma solução que permite a realização de uma distribuição confiável em escala a partir de uma estrutura de software que possibilita que programas processem quantidades massivas de dados desestruturados de maneira paralela e distribuída. O MapReduce foi proposto por engenheiros do Google no início dos anos 2000 com os seguintes requisitos:

- O processamento deve ser elástico, ou seja, deve ser capaz de expandir e contrair, quando necessário.
- O processamento deve ser capaz de ser executado, mesmo que ocorram problemas em um dos computadores (nós) durante a execução, a fim de realizar o processo de distribuição para os outros recursos disponíveis.
- A solução baseada nessa técnica deve ser capaz de ser executada independentemente do local das fontes de origem e dos recursos computacionais que serão utilizados como ambiente de execução.

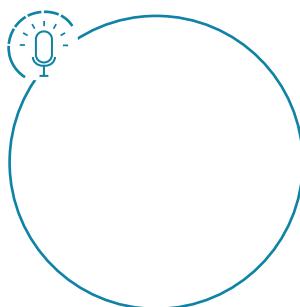
Assim como o próprio nome sugere, o MapReduce é composto por dois principais processos: o de Map e o de Reduce. Eles são detalhados a seguir:

- **Map:** a função Map, assim como o próprio nome já sugere, objetiva mapear. Para quem tem experiência em programação, é muito similar à ideia dos mapas de algumas linguagens de programação, que são estruturas do tipo: chave e valor que abstraem fisicamente o dado armazenado. Basta apenas pedir o dado que a função se encarregará de encontrá-lo e entregá-lo. No entanto, é importante reforçar que esse dado pode vir combinado com outros dados relacionados.

- **Reduce:** a função reduce atua na redução da lista de resultados encontrados para o dado solicitado, além de selecionar apenas os dados relevantes. É possível aplicar outros tipos de filtros ou processamentos de regras de negócio para entregar o dado já processado conforme esperado.

Lara e Joaquim entenderam que o processo de levantamento dos métodos e das técnicas necessários para o desenvolvimento do seu ambiente de Big Data tinha chegado ao fim. Por mais que existam muito mais elementos que poderiam fazer parte do levantamento, ambos sabiam que já tinham o suficiente para iniciar o desenvolvimento do seu ambiente e que esse deveria ser o caminho de outros negócios que estejam no mesmo momento.

A apresentação e a defesa do projeto para a construção do ambiente de Big Data foram um sucesso. Anderson ficou mais que satisfeito com tudo o que viu e escutou. Por já conhecer a dupla, sabia que encontraria muita coisa boa pela frente! Lara e Joaquim ficaram aliviados, pois conseguiram a aprovação do projeto e sabiam que estavam embasados para realizar o levantamento das tecnologias e das ferramentas que viabilizariam a proposta de solução. No entanto, essa é uma etapa que será analisada na próxima unidade.



Será que a criação de um Data Lake é uma tarefa fácil? O que você acha de escutar um pouco dessa experiência sob a ótica de quem já construiu vários tipos de repositórios analíticos? Não deixe de escutar o nosso podcast.

Lara e Joaquim fizeram um mergulho sobre o que é Big Data e conheceram as suas principais características, a sua formação e os principais métodos e técnicas que são possíveis de serem utilizados em uma perspectiva de armazenamento e processamento de dados. Nossa dupla entendeu que a empresa estava no momento exato para investir na coleta de dados externos para complementar o seu ferramental analítico, de tal forma que fosse formado um ambiente de Big Data para suportar a coleta e o processamento por anos.

Eles perceberam que identificar o momento do negócio é fundamental. De nada adianta uma organização que não tem indicadores básicos ou um sistema para apoiar as suas operações investir em soluções de Big Data. Outro aspecto muito importante é o de que as empresas devem observar o tipo de dados externos que eles coletarão e armazenarão. Na empresa de Anderson, nossa dupla pôde facilmente defender o investimento, porque todos os executivos observaram o quanto a empresa cresceu pelo fato de ser apoiada por dados, indicadores e ferramentas de apoio à decisão. Assim, foi definido o desenvolvimento de um repositório massivo, um Data Lake.

Muitas empresas não sabem por onde começar a construção de um Data Lake. A experiência de Lara e Joaquim partiu pelo desejo da arquitetura lógica, que é composta por três camadas que permitem que todos os indicadores sejam aproveitados e isolados em um primeiro momento. Todos os dados externos coletados são facilmente inseridos no Data Lake, já que não são tratados, nem transformados. Dessa forma, são armazenados em uma camada de staging e, no centro da arquitetura, está a camada de acesso, que é onde os cientistas de dados, analistas e demais pessoas podem gerar relatórios, fazer experimentos e atestar hipóteses.

O embasamento teórico visto até aqui é de extrema importância para a construção de uma proposta de solução e para a seleção das ferramentas e tecnologias que devem ser utilizadas no ambiente de Big Data. Esse é um importante aspecto: nenhum ambiente é 100% igual a outro. Por isso, muito mais que conhecer várias ferramentas, deve-se, primeiro, conhecer os métodos e as técnicas disponíveis, exatamente como Lara e Joaquim fizeram. Agora, eles estão preparados para uma nova missão: construir o ambiente.



1. O termo Big Data não se refere a uma tecnologia específica, mas a um cenário, momento e um ambiente.

Assinale a alternativa que apresenta apenas as características relacionadas ao Big Data:

- f) Dados estruturados, repositórios analíticos e a utilização apenas do Data Warehouse.
- g) Dados não estruturados, corpus textuais e a aplicação de aprendizado de máquina.
- h) Mineração de dados, dados não estruturados e dados validados na inclusão.
- i) Dados multi estruturados, repositórios massivos e a possibilidade de uso do Data Lake.
- j) Dados integrados, métodos estatísticos e dados governados.

2. Assinale a alternativa que apresente apenas os elementos que contribuíram para a mudança de paradigma na produção de dados responsáveis pela criação de ambientes de Big Data:

- a) Plataformas da Web 2.0 e IoT.
- b) Internet das Coisas e Dados de ERP.
- c) Dados de CRM e Dados de BI.
- d) Dados de BI e IoT.
- e) Plataformas da Web 2.0 e Data Warehouse.

3. Assinale a alternativa que apresenta corretamente a definição de um Data Lake:

- a) Trata-se de repositórios analíticos baseados em uma modelagem dimensional. Comumente, são utilizados como fonte de dados de soluções de BI.
- b) Trata-se de repositórios massivos de dados multi estruturados que podem ser organizados em camadas para facilitar o processo de inserção dos dados.
- c) Trata-se de uma importante etapa no processo de descoberta de conhecimento em um banco de dados.
- d) Trata-se de tecnologias que apoiam o processo de visualização de dados e de informações na Web 2.0.
- e) Trata-se de um software baseado em MapReduce que objetiva apresentar um sistema de arquivo distribuído.

# CONFIRA SUAS RESPOSTAS



1. D.

Ambientes de Big Data utilizam dados multi estruturados e os armazenam em repositórios massivos. Também podem fazer uso de Data Lakes.

2. A.

O surgimento das plataformas de publicação de conteúdo na Internet (Web 2.0) permitiu que os usuários deixassem de ser apenas consumidores de dados para também serem produtores. Isso fez com que a quantidade de dados publicados aumentasse muito. Outro fenômeno foi o surgimento da chamada Internet das Coisas, que possibilitou o uso de muitos sensores, a fim de gerar muitos dados por segundo.

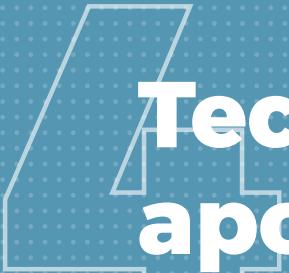
3. B.

São repositórios massivos de dados multi estruturados que podem ser organizados em camadas para facilitar o processo de inserção dos dados. São exemplos: camada de staging, camada de acesso e camada de indicadores.

# REFERÊNCIAS



- AKHTAR, S. M. F. **Big Data Architect's Handbook**: a guide to building proficiency in tools and systems used by leading big data experts. Birmingham: Packt Publishing, 2018.
- BLATTMANN, U.; SILVA, F. C. C. da. Colaboração e interação na Web 2.0 e Biblioteca 2.0. **Revista ACB**, v. 12, n. 2, p. 191-215, 2007.
- DAVENPORT, T. H. **Big Data no trabalho**: derrubando mitos e descobrindo oportunidades. São Paulo: Elsevier, 2014.
- HURWITZ, J. et al. **Big Data para leigos**. Rio de Janeiro: Alta Books, 2016.
- ISOTANI, S.; BITTENCOURT, I. I. **Dados abertos conectados**. São Paulo: Novatec, 2015.
- NEVES, O. M. de C. Evolução das políticas de Governo Aberto no Brasil. In: CONGRESSO CON-SAD DE GESTÃO PÚBLICA, 6., 2013, Brasília. **Anais** [...]. Brasília: CONSAD, 2013.
- PIRES, F. M. M. **Data Lake em Viticultura**: big data management na agricultura. 2017. Dissertação (Mestrado em Gestão da Informação) – Universidade Nova de Lisboa, Lisboa, 2017.
- RAUTENBERG, S.; CARMO, P. R. V. do. Big Data e ciência de dados: complementariedade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Science**, v. 13, n. 1, p. 56-67, 2019.
- SANTOS, B. P. et al. Internet das coisas: da teoria à prática. In: SIQUEIRA, F. A. et al. (org.). **Livro de Minicursos SBRC 2016**. Porto Alegre: SBC, 2016. p. 1-50.
- SILVA, B. L.; LEÃO, M. A contribuição da web 2.0 no processo de ensino e aprendizagem de Química. **Enseñanza de las ciencias**: revista de investigación y experiencias didácticas, n. extra 0, p. 3107-3113, 2009.
- SINCLAIR, B. **Como usar a Internet das Coisas para alavancar seus negócios**. São Paulo: Autêntica Business, 2018.
- TAURION, C. **Big Data**. Rio de Janeiro: Brasport, 2013.



# Tecnologias de apoio ao *Big* *Data*

Dr. Flávio Ceci



## OPORTUNIDADES DE APRENDIZAGEM

Nesta unidade, estudaremos as tecnologias e as ferramentas utilizadas em ambientes de Big Data, a fim de garantir o armazenamento, o processamento e a recuperação de dados em tempo hábil de negócio. As ferramentas e as tecnologias estão organizadas a partir de alguns pilares: o armazenamento massivo de dados, o processamento paralelo e distribuído e a visualização de dados e informação.

Com o investimento aprovado para a criação de um ambiente de *Big Data*, Lara e Joaquim tinham a missão de construir um ambiente que suportasse analiticamente as áreas de negócio por, pelo menos, mais cinco anos, levando em consideração os dados externos disponíveis na Web ou produzidos por sensores.

Diante disso, nossa dupla fez um inventário que levantou os principais métodos e técnicas que são utilizados em um processo de construção de um ambiente de *Big Data*. Esse levantamento é importante, pois evidencia quais são os principais recursos e desafios que podem ser encontrados em um ambiente massivo de dados. Agora é o momento de buscar tecnologias e ferramentas que possam apoiar a implantação do ambiente pretendido e um dos principais desafios será o de elencar as tecnologias que podem suportar o armazenamento massivo de dados.

Os modelos tradicionais de armazenamento não atendem aos requisitos do projeto. Portanto, será necessário navegar em outras tecnologias de base, tais como os bancos de dados relacionais, que suportam o desenvolvimento de sistemas de informação operacionais e transacionais, e os sistemas ERPs, os quais também suportam a contração de repositórios analíticos baseados em modelagem dimensional, como os *Data Warehouse*.

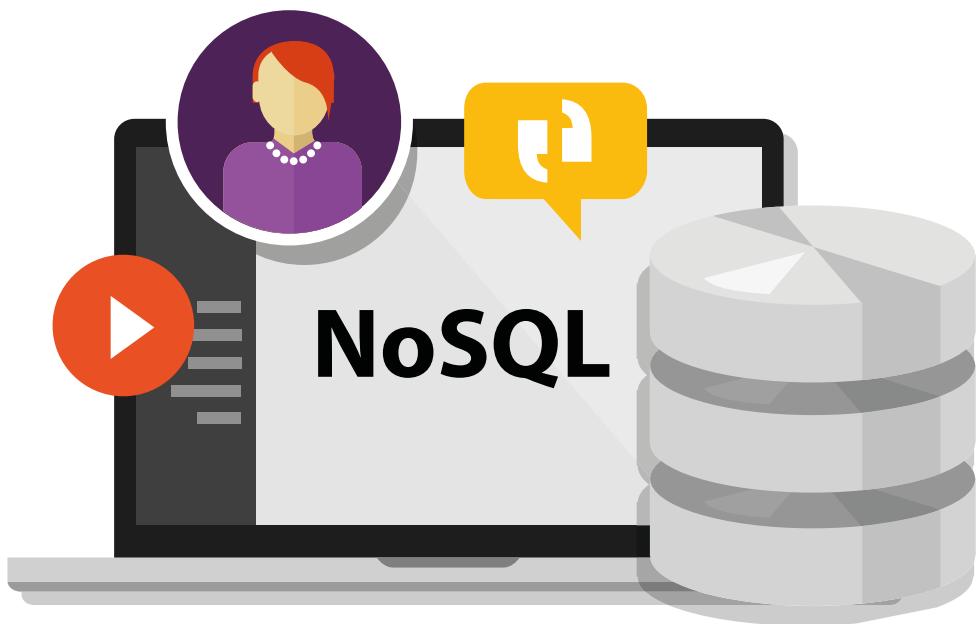
Lara e Joaquim entenderam que, além de terem que explorar outras formas de armazenamento que não sejam os baseados nos modelos relacionais tradicionais, também seria necessário repensar a forma de fazer os processos de extração, transformação e carga (ETL), já que o volume de dados que pode ser produzido por segundo é muito grande e não seria possível trabalhar com uma arquitetura em que o processo de ETL seja executado uma vez por dia.

Depois de terem os dados inseridos no repositório da camada de *staging* do *Data Lake*, deverão encontrar formas de processamento dos dados massivos armazenados. Portanto, terão que estudar tecnologias que implementem e sejam compatíveis com os conceitos relacionados ao *Map Reduce*, a fim de que seja possível extrair valor dos dados armazenados.

Outro desafio que terão será o de encontrar meios de consumir (visualizar) os dados e os indicadores gerados. Nesse contexto, o *Data Lake* será muito utilizado para a realização de análises exploratórias e para a construção de estudos a serem expostos aos executivos e aos gestores da empresa. Quão complexa será a geração de narrativas que facilitem o processo de comunicação e apresentação dos resultados dos estudos realizados?

Realmente, nossa dupla tem muito trabalho pela frente. Agora que já têm os recursos que precisavam, eles devem elaborar o que foi proposto e, além disso: gerar o maior resultado para a empresa. O aspecto positivo dessa situação é o fato de Lara e Joaquim serem motivados pelos desafios e, certamente, mergulharão de cabeça mais uma vez.

Quando falamos em *Big Data*, a primeira relação que fazemos é justamente com a grande quantidade de ferramentas e tecnologias disponíveis para apoiar os processos vinculados a esse ambiente. Todavia, esse fato também é um desafio, visto que é preciso responder a seguinte pergunta: quais são as ferramentas mais indicadas para o meu cenário?



É sabido que os bancos de dados chamados *Not only SQL* (NoSQL, que, em português, significa “não somente SQL”) são fortemente utilizados para apoiar os ambientes massivos de dados. Esse tipo de solução não faz uso do modelo entidade-relacional baseado em tabelas e em relacionamentos com as chaves primária e estrangeira. Logo, conhecer esse tipo de solução pode ser um importante passo para quem deseja ingressar nesse universo de soluções de *Big Data*. Assim, Lara e Joaquim deverão conhecer os vários tipos de soluções NoSQL, saber qual é a diferença entre as soluções de NoSQL de primeira e segunda geração, e explicar o que seria uma abordagem NoSQL híbrida.

Que tal apoiar Lara e Joaquim nessa jornada? Faça uma pesquisa na Internet para saber um pouco mais sobre:

- Quais são os tipos de bancos de dados NoSQL?
- O que significa um banco de dados NoSQL de primeira e de segunda geração?
- Como seria um banco de dados NoSQL híbrido?

Muitas são as ferramentas e as tecnologias que podem apoiar os ambientes de *Big Data*. Nesse contexto, uma das principais missões para quem for projetar a arquitetura física do ambiente é conhecer cada opção, para que seja possível implantar um ambiente extremamente otimizado, extensível e robusto. Nesta unidade, exploraremos as principais tecnologias que podem apoiar o cenário de nossa querida dupla.

## DIÁRIO DE BORDO

## Armazenamento massivo de dados

Joaquim ficou responsável por fazer um levantamento dos principais tipos de bancos de dados não relacionais que poderiam apoiá-lo no projeto de construção do *Data Lake*. Para a camada de indicadores (*Data Warehouse*), ele continuaria utilizando o banco de dados relacional com modelagem de dados dimensional, que já tinha os dados sumarizados e estava atendendo perfeitamente às demandas de negócios. Para as camadas de *staging* e de acesso, a definição ainda era uma incógnita, mas Joaquim decidiu seguir a lógica de ingestão de dados no *Data Lake*, que se dá a partir da camada de *staging*. Assim, essa seria a sua primeira missão: buscar uma ou mais tecnologias que auxiliassem na construção do *Data Lake* da empresa.

O termo NoSQL se refere a um conjunto de banco de dados que pode, ou não, utilizar a linguagem SQL para interagir com as suas estruturas ou com os seus dados. Segundo Alvarez, Ceci e Gonçalves (2016, p. 8), “as bases de dados NoSQL (*Not Only Structured Query Language*) vem ganhando popularidade na era da Web 2.0. Pois, a promessa de alto desempenho quando se envolve dados altamente interconectados, atraiu a atenção dos consumidores de tecnologia”. Os bancos de dados NoSQL também são conhecidos como “banco de dados não relacionais”. Pelo fato de utilizarem outras formas de organização dos dados, são mais flexíveis quanto estruturação. Além disso, têm alta escalabilidade para gerenciar grandes quantidades de dados, o que gera uma grande disponibilidade dos dados (LÓSCIO; OLIVEIRA; PONTES, 2011). Os referidos autores explicam as principais características dos bancos de dados NoSQL:

- **Escalabilidade horizontal:** está diretamente relacionada à necessidade de ampliar a capacidade de armazenamento de maneira horizontal, ou seja, adicionando (ou retirando) novas máquinas que compõem a solução de armazenamento.
- **Ausência de esquema ou esquema flexível:** quando se trabalha com banco de dados relacionais, há o conceito de tabela, o qual defende que a estrutura é fixa para todas as tuplas (linhas) que a compõem. Por outro lado, quando se tem um esquema flexível (também chamado de “livre de esquema”), é possível haver registros com colunas a mais ou a menos, ou seja, não há uma fixação em detrimento do tipo de estrutura vigente.

- **Suporte nativo a replicação:** para que seja possível obter a escalabilidade horizontal, deve haver a possibilidade de replicação dos dados entre as máquinas que compõem a infraestrutura de armazenamento.
- ***Application Programming Interface* (API ou, em português, “interface de programação de aplicação”) simples para acesso de dados:** em detrimento do fato de que esse tipo de banco de dados é focado em soluções massivas de dados, o processo de acesso deve ser facilitado. Nesse sentido, as APIs permitem a realização de integrações fáceis e rápidas com as soluções Web, o que favorece o processo de gestão e consumo dos dados.
- **Consistência eventual:** para ter performance, escalabilidade horizontal, replicação e demais recursos, é necessário dispensar o controle rigoroso sobre a consistência dos dados (assim como acontece nos bancos de dados relacionais). Em outras palavras, eventualmente, é possível haver dados duplicados ou desatualizados em algum dos nós de armazenamento.

Os bancos de dados NoSQL ou não relacionais proporcionam outras formas de organização dos dados e, em consequência disso, é necessário conhecer os modelos de dados disponíveis e as suas características, para que, na sequência, seja possível entender a aplicabilidade de cada um. Caso seja aplicado um modelo de dados equivocado em uma determinada situação, é possível haver um efeito adverso ao que era esperado, ou seja, ao contrário de apoiar o processo massivo de dados, o modelo pode tornar os processos de armazenamento, recuperação e processamento mais custosos e complexos.

Um aspecto importante que se deve ter em mente é o fato de que os bancos de dados NoSQL, no que se refere às suas transações, garantem apenas três das quatro propriedades ACID, que são:

- **Atomicidade:** constituída por apenas uma unidade de trabalho.
- **Consistência:** em caso de falhas, os dados retornam ao estado inicial.
- **Isolamento:** uma transação não deve influenciar outra.
- **Durabilidade:** após uma instrução de confirmação (*commit*), os dados são persistidos.

O primeiro modelo de dados NoSQL que Lara e Joaquim se depararam foi o chamado “**chave/valor**”. Ele tem uma forma de armazenamento muito parecida com a estrutura de dados chamada “*hash*” (tabelas de espalhamento), em que os objetos são indexados por uma única chave que remete a um valor de formato qualquer (inclusive um objeto). Ao ler essa informação, Joaquim ficou mais tranquilo: ele não sabia da existência de bancos apoiados em chave/valor, mas conhecia o conceito de tabelas *hash*, que foi estudado em algumas disciplinas, tais como a de estrutura de dados.

As **tabelas hash** são estruturas focadas no processo de busca de valores a partir de uma chave. Para isso, todos os valores são “espalhados” em endereços que têm chaves vinculadas a ela, diminuindo o tempo de busca necessário. No caso dos bancos baseados em chave/valor, a ideia é a mesma, o que os caracterizam como bancos focados na busca por objetos a partir de um valor-chave que pode ser definido para a estrutura.



#### EXPLORANDO IDEIAS

Indexação é um processo muito utilizado em banco de dados e em tecnologias de buscas textuais feitas em documentos completos. O foco desse processo é o de garantir que seja facilitado o processo de busca, criando, muitas vezes, estruturas paralelas para apoiar essa necessidade.

Fonte: o autor.

Os valores armazenados nas estruturas dos bancos baseados em chave/valor são normalmente estruturados como objetos *JavaScript Object Notation* (JSON), um formato muito utilizado para a interoperabilidade entre sistemas na Internet. Os objetos JSON facilitam a estruturação dos valores, já que apresentam o nome dos atributos que o compõem. Além disso, são fáceis de serem processados tanto computacionalmente quanto por seres humanos.

Para facilitar o seu entendimento, considere um objeto que represente um aluno e as características que você deseja armazenar são as seguintes: nome, idade, matrícula e curso. Suponha que o valor que devemos armazenar está relacionado ao aluno Miguel, que tem 23 anos, sua matrícula é 9812 e o curso em que está matriculado é Medicina Veterinária. Nesse contexto, o objeto ficaria da seguinte forma:

```
{  
    "nome": "Miguel",  
    "idade": 23,  
    "matrícula": 9812,  
    "curso": "Medicina Veterinária"  
}
```

O próximo registro armazenado poderia ser:

```
{  
    "nome": "Lua",  
    "idade": 23,  
    "matrícula": 9813,  
    "curso": "Design",  
    "turno": "Noturno"  
}
```

Perceba que o segundo objeto armazenado tem um atributo turno que não havia sido inserido. Essa atitude é possível, em consequência de os bancos não relacionais serem livres de esquema. Dessa forma, o modelo chave/valor pode utilizar qualquer um dos atributos como chave para encontrar os objetos (por exemplo, a chave pode ser o atributo “matricula”). Segundo Lóscio, Oliveira e Pontes (2011, p. 7), “as operações disponíveis para manipulação de dados são bem simples, como o get( ) e o set( ), que permitem retornar e capturar valores, respectivamente. A desvantagem deste modelo é que não permite a recuperação de objetos por meio de consultas mais complexas”.

O objeto JSON pode ser armazenado como um valor literal, mas não é possível executar um processo de filtro ou busca por um atributo em especial. Por esse motivo, Lóscio, Oliveira e Pontes (2011) comentam que há uma desvantagem em relação à recuperação de objetos a partir de consultas mais complexas, dado que, de fato, é necessário fazer uma busca exata pela chave em questão. Essa estrutura é muito utilizada por bancos de dados em memória, com o objetivo de fazer cache de aplicação Web, ou por bancos que armazenam estados de objetos durante um processo. A figura a seguir apresenta uma representação gráfica de como os dois objetos apresentados estariam organizados em um banco chave/valor, considerando o valor da matrícula como a chave utilizada:

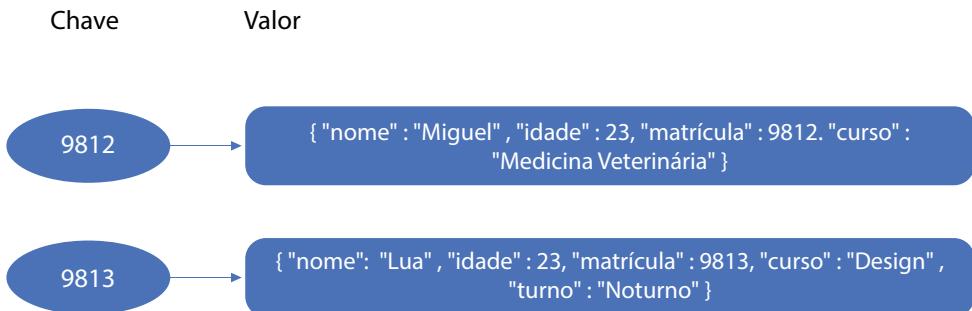


Figura 1 - Exemplo do modelo chave/valor / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta duas colunas. Uma tem o título "Chave" e a outra tem o título "Valor". Abaixo da coluna "Chave", há uma elipse com o valor 9812 que aponta para um retângulo, o qual está abaixo da coluna "Valor". O conteúdo do retângulo é o seguinte: {"nome": "Miguel", "idade": 23, "matrícula": 9812, "curso": "Medicina Veterinária"}. Abaixo, há o mesmo exemplo, mas com a chave 9813, que aponta para os valores: {"nome": "Lua", "idade": 23, "matrícula": 9813, "curso": "Design", "turno": "Noturno"}.

Existem algumas opções de banco de dados chave/valor:

- **DynamoDB:** banco de dados chave/valor desenvolvido pela Amazon e está disponível em sua plataforma AWS.
- **Redis:** banco de dados de código aberto e de licença livre, o que permite ser implantado sem custo de licenciamento. Muito utilizado como banco em memória e está disponível nas principais nuvens.

Ao terminar os seus estudos sobre os bancos chave/valor, Joaquim começou a estudar os **bancos de dados colunar** ou **banco de dados de família de colunas**. Nesse modelo, os dados que seguem o mesmo tipo são agrupados em famílias de colunas. De acordo com Lóscio, Oliveira e Pontes (2011, p. 8), “neste modelo os dados são indexados por uma tripla (linha, coluna e *timestamp*), onde linhas e colunas são identificadas por chaves e o *timestamp* permite diferenciar múltiplas versões de um mesmo dado. Vale ressaltar que operações de leitura e escrita são atômicas”. Os bancos de dados colunares são muito utilizados para tratar dados temporais, pelo fato de haver o *timestamp* (que é um valor que representa o exato milissegundo de algo que está acontecendo) em um dos campos utilizados no processo de indexação dos valores.

Além do mais, os bancos colunares são calcados em um conceito desenvolvido pelo Google no início dos anos 2000, o chamado de *Big Table*. O seu objetivo é fazer o processo de desnormalização dos dados, ou seja, ao contrário de se dividir os dados em várias tabelas, com vários relacionamentos entre elas, os dados são organizados em forma de colunas, que são estruturadas por famílias e combinadas em uma grande tabela. Esse processo facilita a construção de soluções distribuídas. A figura a seguir apresenta um exemplo desse modelo de dados:



Figura 2 - Exemplo de organização do modelo colunar / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta um grande retângulo com o título “Tabela”, que busca agrregar todos os retângulos que estão abaixo. Assim, abaixo do retângulo “Tabela”, há outros dois retângulos. O primeiro, que é mais estreito, é chamado de “Família de coluna 1”. Abaixo, há apenas uma coluna chamada “Nome”, que possui valores armazenados. Ao lado do retângulo “Tabela”, está um retângulo mais largo, o “Família de coluna 2”, que tem duas colunas de retângulos abaixo, as denominadas: “Celular” e “Telefone”. Abaixo de cada uma das colunas, há os valores que as correspondem.

Existem vários bancos de dados que implementam o modelo colunar. A seguir, são apresentados alguns exemplos:

- **Apache Cassandra:** banco de dados *open source* e de licença grátis mantido pela Fundação Apache. É uma solução escalável e bastante utilizada no mercado.
- **Amazon Redshift:** exclusivo para uso dentro da infraestrutura da AWS, da Amazon. Apresenta ótima performance e exige pagamento.
- **Google BigQuery:** banco que implementa os principais conceitos do *BigTable*. É proprietário da Google e exclusivo para uso no Google Cloud.

Joaquim achou interessante a possibilidade de utilizar um banco de dados colunar para apoiar as aplicações em que a dimensão temporal é de extrema importância para a análise. Ao relembrar o que havia estudado sobre IoT, chegou à conclusão de que esse tipo de banco de dados pode apoiar as soluções dessa natureza, em que os dados vindos dos sensores são organizados e indexados, levando em consideração o momento exato (*timestamp*) da captura.

O terceiro tipo de banco de dados NoSQL escolhido foi o **banco de dados orientado a documento**. Esse tipo de banco de dados não relacional é, com certeza, o mais conhecido e utilizado desde os bancos de dados dessa natureza. Segundo Diana e Gerosa (2010), os documentos dos bancos de dados orientados a documentos são coleções de atributos e valores em que um atributo pode ser multivlorado. Além disso, em geral, esses bancos de dados não têm esquema. Os documentos são tradicionalmente representados por objetos JSON e, “como o próprio nome diz, este modelo armazena coleções de documentos. Um documento, em geral, é um objeto com um identificador único e um conjunto de campos, que podem ser strings, listas ou documentos aninhados” (LÓSCIO; OLIVEIRA; PONTES, 2011, p. 8).

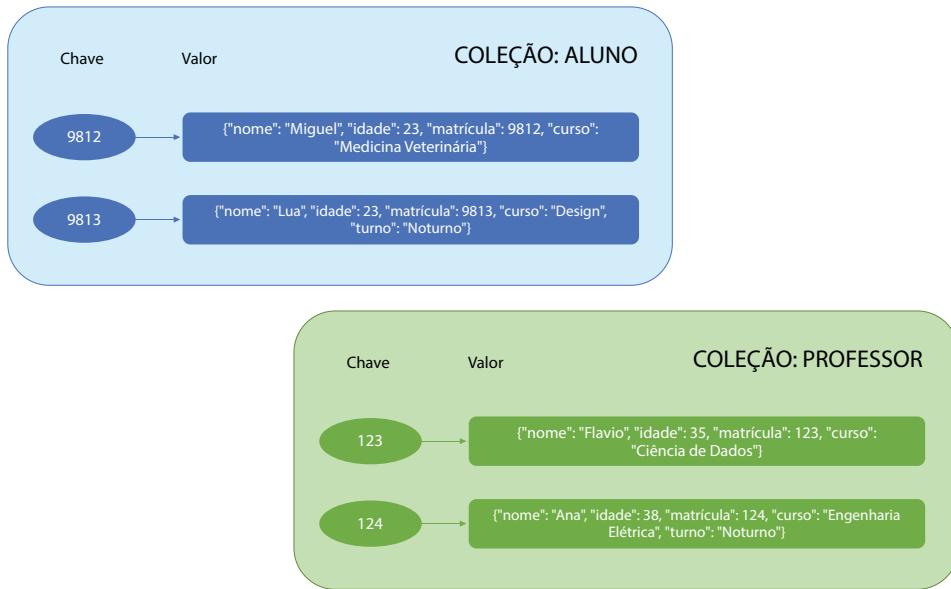


Figura 3 - Modelo orientado a documentos e suas coleções / Fonte: o autor.

**Descrição da Imagem:** na figura, são demonstradas duas formas e ambas carregam um conjunto de elementos. Uma forma é chamada de “Coleção: aluno” e a outra é intitulada “Coleção: professor”. Dentro das duas formas, há elementos muito parecidos, mudando apenas os valores dos atributos. Na “Coleção: aluno”, estão duas colunas: uma é chamada de “Chave” e a outra é denominada “Valor”. Abaixo da coluna “Chave”, há uma elipse com o valor 9812, que aponta para um retângulo que está abaixo da coluna “Valor” e é dotado do seguinte conteúdo: {"nome": "Miguel", "idade": 23, "matrícula": 9812, "curso": "Medicina Veterinária"}. Abaixo, há o mesmo exemplo, mas com a chave 9813 e com o valor {"nome": "Lua", "idade": 23, "matrícula": 9813, "curso": "Design", "turno": "Noturno"}. Já na “Coleção: professor”, estão duas colunas: uma é intitulada “Chave” e a outra é denominada “Valor”. Abaixo da “Chave”, existe uma elipse com o valor 123, que aponta para um retângulo que está abaixo da coluna “Valor” e carrega o seguinte conteúdo: {"nome": "Flavio", "idade": 35, "matrícula": 123, "curso": "Ciéncia de Dados"}. Abaixo, há o mesmo exemplo, mas com a chave 124 e com o valor: {"nome": "Ana", "idade": 38, "matrícula": 124, "curso": "Engenharia Elétrica", "turno": "Noturno"}.

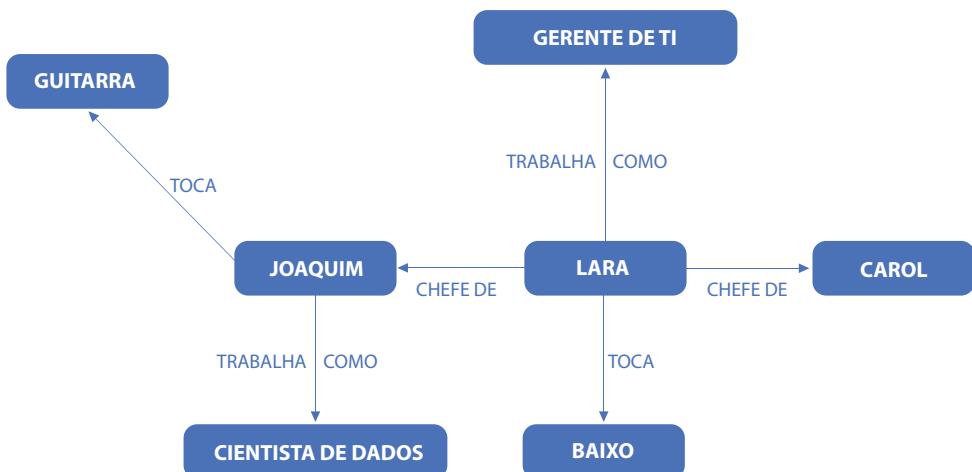
A estrutura do banco de dados orientado a documento é semelhante à estrutura chave/valor. Lóscio, Oliveira e Pontes (2011, p. 8) explicam que, “no modelo chave-valor, apenas uma única tabela hash é criada para todo o banco. No modelo orientado a documentos temos um conjunto de documentos e em cada documento temos um conjunto de campos (chaves) e o valor deste campo”. Esse conjunto de documentos é chamado de “coleção”, ou seja, em bancos orientados a documentos, trabalha-se com coleções de documentos e a Figura 3 apresentou um exemplo.

Os bancos de dados orientados a documentos são muito utilizados em plataformas sociais na Web, em soluções de IoT e para o desenvolvimento de aplicações Web, geralmente, baseadas em *Javascript*. Em ambientes de *Big Data*, são vastamente utilizados e, dentre as soluções SQL, eles são os que mais dispõem de profissionais habilitados para uso.

Existem várias implementações de banco de dados orientados a documentos. A seguir, são apresentadas algumas opções disponíveis:

- **MongoDB**: é o banco de dados orientado a documentos mais conhecido e utilizado pela comunidade de desenvolvedores. Ele tem código aberto e licença livre. Além do mais, há muita documentação disponível, casos de uso e permite consulta textual em campos armazenados dentro dos documentos.
- **Apache CouchDB**: é um banco de dados livre e mantido pela Fundação Apache. Busca ser muito fácil de se utilizar e foi construído para apoiar o desenvolvimento de sistemas na Web.

Joaquim observou muita sinergia entre as necessidades futuras da empresa e o uso dos bancos orientados a documentos. Diante disso, estava ansioso para conhecer o último tipo de banco NoSQL, o chamado **banco de dados orientados a grafo**. Os bancos orientados a grafo carregam esse nome justamente pelo fato de organizarem os dados em forma de um grafo. Um grafo tem uma estrutura de dados que se assemelha a de uma árvore ou de uma rede. A figura a seguir apresenta mais detalhes:



**Figura 4 - Modelo de dados de um banco orientado a grafo / Fonte: o autor.**

**Descrição da Imagem:** a figura apresenta um grafo. Há um nó central com o título “Lara” e uma aresta cujo título é “Trabalha como”. Essa aresta está ligada a uma seta que vai até outro nó, que é chamado de “Gerente de TI”. O nó “Lara” também tem uma aresta com título “toca” ligado ao nó “Baixo”. Também há uma aresta chamada “chefe de” ligada aos nós “Joaquim” e “Carol”. A partir do nó “Joaquim”, há uma aresta “trabalha como” ligada ao nó “Cientista de Dados”. Por outro lado, a aresta “toca” está ligada ao nó “Guitarra”.

No modelo orientado a grafo, as entidades podem ser representadas por nós e as relações podem ser ilustradas por arestas. Em muitos casos, os atributos também são formados por nós, que têm arestas ligadas ao nó referente à entidade em questão. Esse modelo é totalmente livre de esquema e pode ter todas as características de um grafo tradicional aplicadas, tais como a busca pelo menor caminho. Além disso, esse tipo de modelo de dados é muito utilizado para armazenar elementos que podem ser organizados em forma de rede, como o relacionamento dos usuários em uma rede social. Assim, é possível aplicar algoritmos de menor caminho para saber quem seria a(s) melhor(es) pessoa(s) para se apresentar a uma pessoa distinta e fechar um negócio.

Existem alguns bancos de dados que são reconhecidos por implementar o modelo de dados orientado a grafos. A seguir, são apresentados dois exemplos:

- **Neo4J:** banco de dados orientado a grafo e desenvolvido em Java. Tem código aberto e licença livre. Suas consultas são feitas por intermédio da linguagem Cypher e é um dos bancos de dados orientados a grafo mais utilizados.
- **OrientDB:** é um banco de dados multi-modelo que implementa, além da orientação à grafo, a ideia de documentos e de chave/valor. Em consequência disso, é reconhecido como banco NoSQL de segunda geração. Também foi desenvolvido em Java, tem código aberto e licença livre.

Os bancos de dados NoSQL de segunda geração, comumente, utilizam abordagens multi-modelo. No entanto, entre o Neo4J e o OrientDB, há uma diferença de performance. Alvarez, Ceci e Gonçalves (2016) fizeram uma análise comparativa utilizando o conceito de busca em profundidade em grafos e o Neo4J teve um desempenho e uma robustez muito maiores do que o OrientDB. Em contrapartida, o OrientDB permitiu o desenvolvimento de soluções muito mais robustas, ao combinar recursos de modelos distintos. Esse é um aspecto muito importante e deve ser levado em consideração na escolha do banco NoSQL: quando é necessário utilizar um recurso muito específico da estrutura de dados que baseou o modelo de dados em questão, um banco NoSQL de primeira geração (que tem apenas um modelo implementado) tende a ter uma performance e uma robustez maiores.

Joaquim tinha vontade de testar todos os bancos de dados NoSQL que havia levantado em sua pesquisa. Todavia, sabia que, com o tempo disponível, não seria possível. Assim, chegou até o trabalho desenvolvido por Faraon (2018), que expõe uma lista de vantagens e desvantagens em relação ao uso de soluções NoSQL. A seguir, são apresentadas as vantagens:

- **Esquema flexível:** também conhecido como “livre de esquema”, representa a ideia de não haver uma estrutura rígida para armazenar dados sobre um mesmo assunto ou temática. Assim, há flexibilidade no armazenamento e simplicidade na organização dos domínios de dados.
- **Escalonamento e desempenho:** o processo de crescimento dessas soluções é horizontal, ou seja, são soluções distribuídas que permitem que novos nós computacionais sejam adicionados (ou retirados), o que torna a infraestrutura computacional elástica e optimiza os recursos utilizados.
- **Replicação de dados:** a forma de replicação dos dados feitos pelos bancos NoSQL permite o compartilhamento e a distribuição dos dados de forma automática e transparente. Isso faz com que o processo de substituição de um servidor seja simplificado.
- **Disponibilidade:** diante do processo de replicação dos dados, há uma grande disponibilidade, já que os dados estão replicados e distribuídos. Mesmo que haja um problema em um recurso, ele pode ser facilmente substituído, sem a necessidade de tornar o ambiente indisponível.

Após apresentar as vantagens no que diz respeito ao uso dos bancos NoSQL, Faraon (2018) expõe as desvantagens da utilização desse tipo de banco de dados:

- **Linguagem:** não existe uma linguagem padrão, assim como há no SQL, para interagir com os bancos NoSQL. Isso gera curvas de aprendizagem e personalização de aplicações para a troca de soluções desse tipo.
- **Consistência:** é sabido que, para se ter dados e informações distribuídas com performance, flexibilidade e disponibilidade, não é possível garantir a consistência de todos os dados armazenados. Desse modo, é possível haver dados desatualizados e duplicados em um nó.
- **Esquema flexível:** em consequência do esquema flexível, todo o controle de integridade e de consistência é de responsabilidade da camada de aplicação.

As estruturas NoSQL não são recomendadas para os sistemas transacionais/operacionais, tendo em vista que, nesses casos, a consistência dos dados armazenados é de extrema importância. Joaquim se deparou com outro tipo de banco de dados, o chamado **banco NewSQL**, que objetiva ser uma solução que esteja entre os bancos de dados relacionais e os bancos de dados NoSQL. Segundo Faraon (2018, p. 30), o “NewSQL é uma classe de sistemas de gerenciamento de bancos de dados relacionais, que procura oferecer o mesmo desempenho escalável do modelo NoSQL, para cargas de trabalho de leitura e gravação no processamento de transações on-line, mantendo as garantias ACID”.

Para Knob *et al.* (2019, p. 3), “eles são geralmente BDs em memória principal que maximizam a vazão de dados, prevenindo custosos acessos em disco aos dados. Eles também possuem mecanismos de controle de concorrência que evitam o bloqueio de dados, possibilitando alta disponibilidade de dados”. Segundo Faraon (2018), os bancos NewSQL são bancos relacionais desenvolvidos para processar milhões de transações por segundo, de maneira que possam ser escalados horizontalmente e executados quase totalmente em memória. O estudosso ainda complementa o conceito apresentado a partir da exposição das seguintes vantagens:

- **Banco totalmente relacional:** aceita todos os pontos previstos pela linguagem SQL.
- **Conformidade ACID:** tem as garantias de integridade previstas para os bancos relacionais, suportando as transações ACID.
- **Latência de milissegundos:** execução de milhares de transações por segundo.
- **Tolerância a falhas:** modelos distribuídos com alta disponibilidade.
- **No local ou na nuvem:** permite a sua implantação tanto em bases locais ou a partir de soluções na nuvem.

Existem algumas implementações de bancos de dados classificados como NewSQL disponíveis no mercado. A seguir, são apresentadas algumas opções:

- **VoltDB:** tem uma versão *enterprise* e outra de licença grátils. Trabalha com a serialização no acesso aos dados. Há uma arquitetura em *cluster*, com replicação em vários servidores.

- **NuoDB:** tem versões pagas e de licença livre, mas com restrições de escalabilidade. Internamente, é estruturado em uma arquitetura de duas camadas que facilita o seu processo de distribuição e uso.
- **CockroachDB:** converte as instruções SQL em instruções chave/valor pela velocidade de manipulação. Tem versão paga e licença livre.

Depois de conhecer todas as ferramentas apresentadas, Joaquim se sentia preparado para continuar a sua jornada. Ele sabia que a sua proposta de solução utilizaria um ou mais de um banco de dados apresentados, mas começou a refletir: apesar de esses bancos de dados suportarem grandes massas de dados, como farei para processar os dados armazenados? Joaquim constatou que as soluções estudadas o apoiarão na rápida recuperação dos dados armazenados. No entanto, e quando for necessário fazer uma análise mais aprofundada, o que exige o acesso a um volume realmente grande dos dados armazenados? Essas são as perguntas norteadoras do próximo caminho que ele percorrerá.

## Processamento massivo de dados

Joaquim chamou Lara e apresentou o resultado de sua pesquisa sobre as várias tecnologias e ferramentas que são utilizadas em ambientes de *Big Data*, a fim de suportar o armazenamento massivo de dados. Lara achou muito legal o inventário produzido e comentou que já trabalhou com alguns bancos de dados expostos pelo inventário. Também afirmou que agora era o momento de pensarem em formas que garantissem o processamento massivo de dados em tempo real, pois esse pode ser um requisito a ser apresentado nos próximos meses. Quando pensamos em processamento em tempo real, devemos levar em consideração o recebimento de vários sinais ou mensagens por segundo, os quais vêm dos mais diferentes canais, tais como celulares, sites, ERP nas lojas e sensores. Dessa forma, o ambiente elaborado deve estar preparado para receber vários estímulos por segundo e precisa recuperar e processar uma grande massa de dados, para que seja gerada uma ação, que pode ser desde uma mensagem, uma recomendação ou um conjunto de instruções para os setores internos, por exemplo.

Joaquim anotou praticamente tudo o que Lara comentou e conectou os itens que já havia estudado sobre o assunto em sua mente. O primeiro conceito lembrado foi o do MapReduce: ele sabia que, de alguma forma, esse modelo poderia auxiliá-lo nessa missão. Entretanto, ainda faltava um pouco mais de clareza em relação ao modo como as coisas estão conectadas, quais seriam os locais em que ficariam os bancos de dados, como eles poderiam se conectar ao modelo MapReduce, dentre outras dúvidas. Para tentar esclarecer essas questões, Joaquim foi buscar ferramentas que implementassem o modelo MapReduce e chegou até a ferramenta Hadoop.

O **Hadoop** foi desenvolvido para que as organizações administrassem grandes volumes de dados com facilidade. Essa ideia permitiria que grandes problemas fossem divididos em elementos menores e a análise fosse feita de modo mais rápido e de maneira paralela. Em outras palavras, foi construído para parallelizar o processamento dos dados por meio de nós computacionais, acelerando o processamento e escondendo a latência (HURWITZ *et al.*, 2016). Segundo Amaral (2016), o Hadoop foi desenvolvido com licença livre e é mantido pela Fundação Apache. Assim, há uma versão mantida pela fundação e algumas versões com implementações proprietárias de fornecedores de nuvem, como a Microsoft. Além da implantação do modelo MapReduce, o Hadoop tem o acréscimo de um sistema de arquivo distribuído, para que seja possível distribuir os dados em vários nós. Essa solução leva o nome de HDFS.

O **HDFS** é uma sigla para *Hadoop Distributed File System* (em português, “Sistema de Arquivo Distribuído Hadoop”), que diz respeito a “um sistema de arquivos distribuído em uma estrutura mestre/escravo. O nó mestre é chamado de NameNode, e é responsável pelos metadados: nomes de arquivos, permissões e localização de cada bloco” (AMARAL, 2016, p. 58). A seguir, com base no trabalho de Hurwitz *et al.* (2016), detalhamos os nós mestres (*Name Nodes*) e os nós de dados (*Data Node*):

- **Name Nodes:** é o responsável por distribuir e mapear onde os blocos de dados (que foram divididos) estão armazenados. Também age como orquestrador, gerenciando o acesso aos arquivos, incluindo a leitura, a gravação, a criação, a exclusão e a réplica dos blocos de dados. Não existe acoplamento com os nós de dados, somente comunicação e troca de mensagens e blocos de dados. Deve se replicar durante as execuções para se proteger.
- **Data Node:** é orquestrado pelo *name nodes* e é resiliente. Dentro do agrupamento do HDFS, os blocos de dados são replicados a partir de múltiplos nós. Também carrega um identificador, que é utilizado como “ID do Rack”.

A figura a seguir apresenta um exemplo do *cluster Hadoop* e os elementos do HDFS:

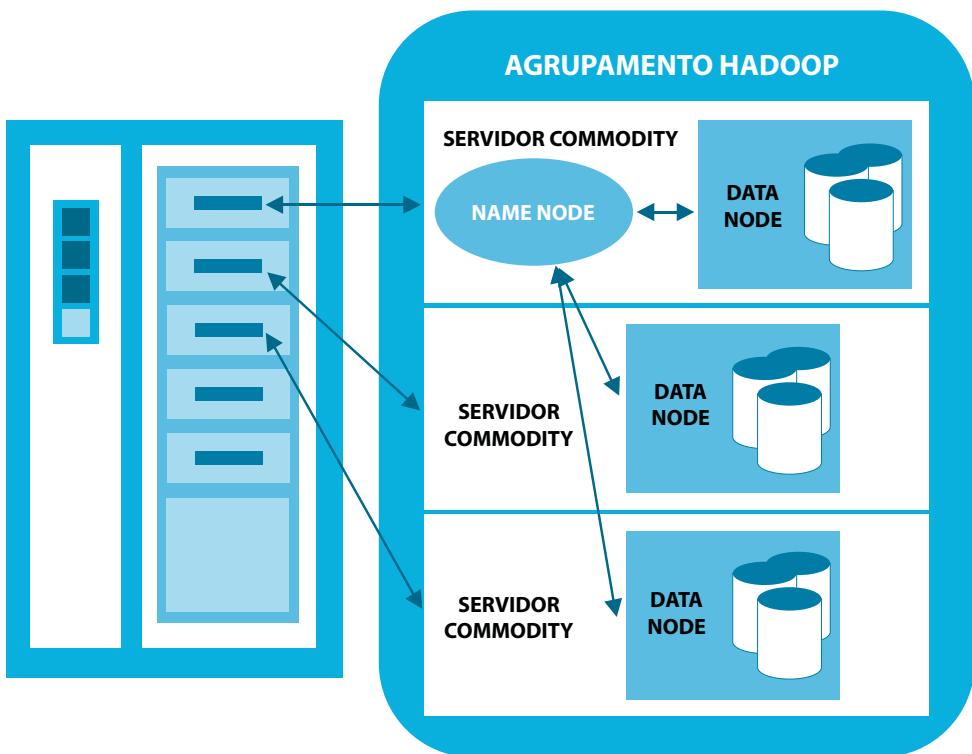


Figura 5 - Exemplo de *cluster Hadoop* / Fonte: Hurwitz et al. (2016, p. 113).

**Descrição da Imagem:** na figura, é apresentado um armário com vários servidores armazenados. Assim, é exposto o detalhamento de três servidores. Um é chamado de “Name Node”, que está em conjunto com uma instância do “Data Node”. Já nos outros dois outros servidores existe apenas uma instância de Data Node.

O Hadoop é uma solução de processamento distribuído que é classificada como uma abordagem de **processamento em lote**. Segundo Oliveira (2019, p. 33), “o processamento em lote é uma forma eficiente de processar grandes volumes de dados, onde os dados são coletados, processados e então produzidos os resultados em lote, visando principalmente a alta vazão do sistema, mesmo que em detrimento da latência”. O autor também explica o processamento em *batch*, que tem a não interação com o usuário como característica. Dessa forma, o processo é feito com base em áreas de entrada e de saída predefinidas.

A segunda abordagem é a de **processamento de streams**. Oliveira (2019, p. 34) defende que “o processamento de streams de dados deve ser realizado no momento

em que os dados chegam ao sistema com uma baixa latência de resposta. [...] Dado uma sequência de dados (um stream), uma série de operações é aplicada a cada elemento no stream". Há algumas ferramentas que implementam o conceito de processamento de *streams* e alguns exemplos são apresentados a seguir:

- **Apache Storm:** framework de computação distribuída para o processamento de *streaming*. Desenvolvido em Clojure, é de código aberto e tem licença livre. As aplicações são desenhadas em forma de grafo acíclico dirigido.
- **Spark Streaming:** pode ler dados a partir de HDFS, Kafka e outras fontes de dados. Isso permite que o processamento seja feito tanto em *streaming* quanto em *batch*, ao utilizar todo o potencial do *Spark*.
- **Hadoop Streaming:** solução de *streaming* para trabalhar dentro do ecossistema Hadoop.
- **Apache Flink:** escrito em *Java* e em *Scala*, é um framework de processamento distribuído que trabalha tanto com *streaming* quanto com processamento em *batch*. Não tem um sistema de armazenamento e é compatível com HDFS, Kafka, *Apache Cassandra*, dentre outros.
- **Apache Kafka:** escrito em *Java* e em *Scala*, é uma plataforma de código aberto para processamento distribuído. Já tem uma camada de armazenamento e é baseado no conceito de fila de mensagens.

A terceira abordagem para o processamento distribuído é o **processamento em tempo real**. Oliveira (2019, p. 34) assevera que ele “permite que o usuário possa executar suas próprias análises em segundos sobre dados estruturados e não estruturados”. Como exemplo de implementação, há o *Apache Drill*, que utiliza o HDFS como armazenamento e o MapReduce para a análise em lote. Contudo, também há a possibilidade de executar as consultas em sistemas de arquivos distribuídos, o que retorna os dados em formato de colunas.

Nesse momento, Joaquim parou, respirou e começou a refletir. Depois de conhecer todas as ferramentas e os conceitos de processamento distribuído e massivo de dados, pensou que, em uma arquitetura moderna para um ambiente de *Big Data*, é preciso combinar uma série de ferramentas e técnicas para que seja possível formar uma infraestrutura de dados. Será que é necessário programar todas as camadas de integração entre as ferramentas? Não teria uma forma mais natural para organizar os fluxos de dados e processamentos?

Ao navegar pelas ferramentas de busca, chegou ao termo “***dataflows***”, que, em sua essência, diz respeito aos *workflows* (fluxos de trabalho) orientados a dados, cujo foco é modelar os fluxos e automatizar as execuções dos processos relacionados à manipulação e à análise de dados. O que torna o uso de *dataflows* interessante é o fato de poder encadear uma série de processos de maneira serial ou paralela, o que faz com que as etapas sejam tratadas de maneira muito especializada. A figura a seguir apresenta um exemplo:

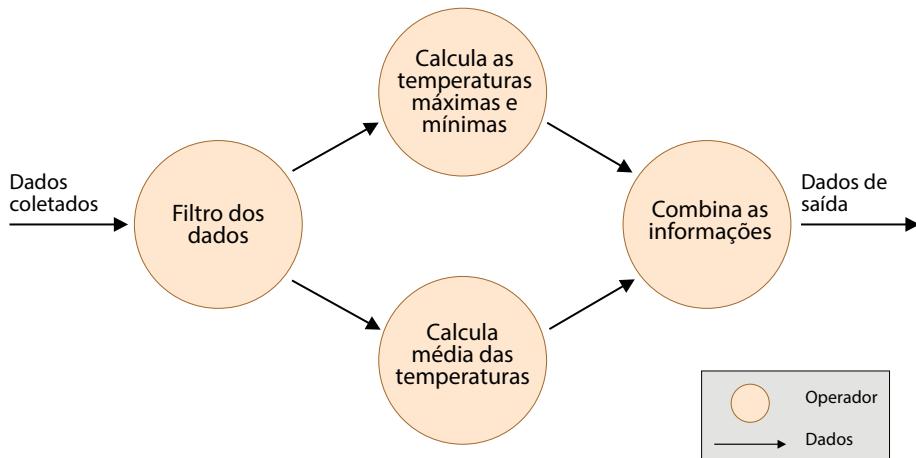


Figura 6 - Exemplo de *dataflow* / Fonte: Magano (2020, p. 14).

**Descrição da Imagem:** na figura, há a primeira etapa, chamada de “Dados coletados”. Assim, é visualizada uma seta que vai até um operador (processo) intitulado “Filtro de Dados”. A partir dessa etapa, há duas setas que saem em paralelo: a superior vai até o operador “Calcula as temperaturas máximas e mínimas” e a segunda segue para o operador “Calcula média das temperaturas”. Ambos os operadores apontam para o próximo operador no fluxo, cujo título é “Combina as informações”. Ao final, uma seta informa: “Dados de saída”.

Os *dataflows* são compostos por duas estruturas básicas: os processos e os *pipelines*. Segundo Magano (2020, p. 13):



- **Processo:** é o componente mais simples. Trata-se de uma transformação aplicada a um dado de entrada, gerando uma saída.
- **Pipeline:** é resultante da combinação sequencial de transformações, produzindo resultados intermediários utilizados pelo seu sucessor, sendo uma extensão do modelo de processo, mas considerada uma estrutura única, por ser recorrente em modelos de workflows.

A figura a seguir apresenta os principais tipos de estruturas que podem ser utilizados em um *workflow*:

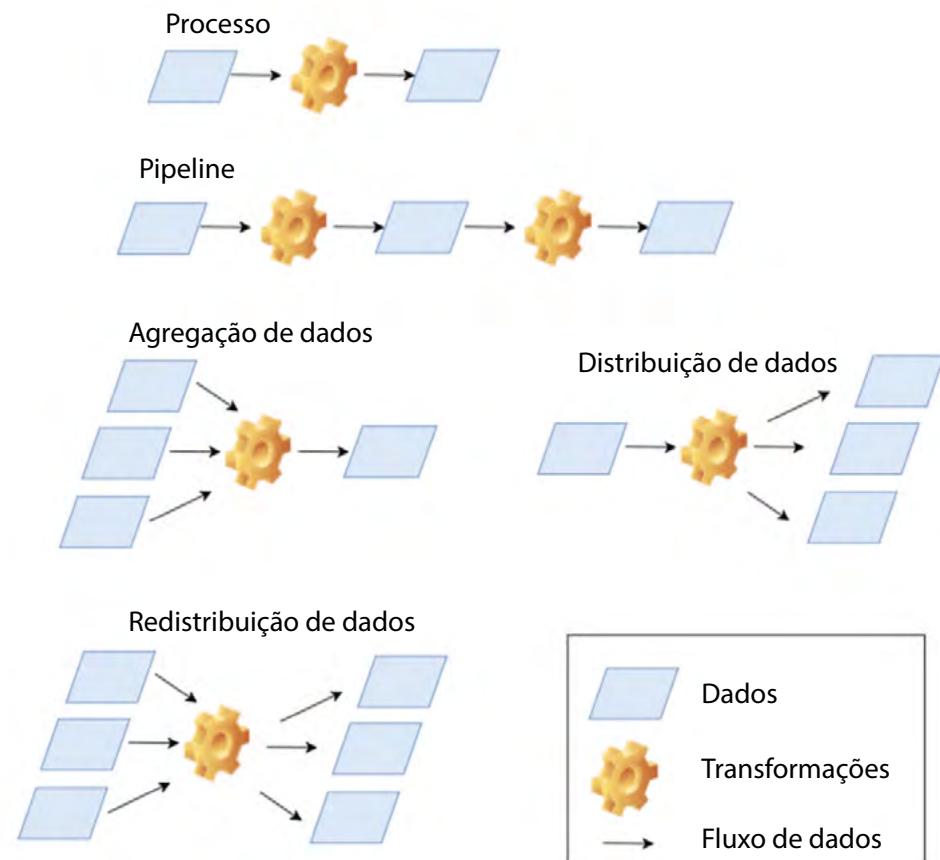


Figura 7 - Tipos de estruturas de *workflows* / Fonte: Magano (2020, p. 15).

**Descrição da Imagem:** a figura apresenta vários tipos de estruturas que podem existir em um workflow. Inicialmente, é apresentada a ideia de processo. Assim, há um dado de entrada, uma transformação e uma saída. Na sequência, é exposto um exemplo de pipeline, em que dois processos estão encadeados. Também é apresentado o conceito de agregação de dados. Nele, há três conjuntos de dados de entrada, uma transformação e um dado de saída. Depois, é apresentado o conceito de distribuição de dados, o qual é dotado de um dado de entrada, uma transformação e três dados de saída. Por fim, é exposto o conceito de redistribuição, que carrega três dados de entrada, uma transformação e três dados de saída.

Os conceitos de processo e *pipeline* já foram expressos. Na sequência, são apresentados outros três tipos de estruturas possíveis em um *workflow*, a saber:

- **Agregação de dados:** o processo de agregação se dá a partir da entrada de vários dados distintos. Dessa forma, é feita uma transformação ou ação e há um dado como saída.
- **Distribuição de dados:** nesse processo, há um dado como entrada, é realizada a aplicação de alguma transformação ou ação e, como saída, é gerado mais do que um dado.
- **Redistribuição de dados:** nesse processo, é possível aplicar transformações e ações, fazendo o processo de orquestração dos dados de saída a partir de um conjunto de regras que podem alimentar outros processos paralelos.

Existem algumas soluções que permitem implementar a ideia do *dataflow*. O Apache Flink é um exemplo de uma solução. Outra solução bastante utilizada para esse cenário é o Apache Nifi, que foi desenvolvido para automatizar o fluxo de informações entre sistemas, ou seja, pode interagir com várias aplicações distintas, a fim de realizar a orquestração dos dados de entrada e saída de cada uma das camadas de sistemas e bancos de dados. Além do mais, funciona tanto para o processamento em lote quanto para o processamento em tempo real. Tem uma interface gráfica Web, para que seja possível desenhar os fluxos de dados entre os sistemas, o que permite acompanhar os processos em execução.

Nesse momento, Joaquim ficou mais tranquilo, pois sabia que era possível orquestrar todas as camadas de sistemas envolvidas no ambiente de *Big Data* de maneira simples, a partir de uma ferramenta que permite que os fluxos e os *pipelines* sejam modelados de maneira simplificada, sem a necessidade de programar todas as interações a partir de uma linguagem de programação como *Java* ou *Python*. Agora, de fato, já havia um conjunto de ferramentas que permitia a edificação de toda a infraestrutura de dados para a elaboração do *Data Lake*. O desafio era: como fazer a entrega desses dados? Esse era um aspecto que ainda estava sem resposta e era extremamente importante para esse momento.

## Análise e visualização dos dados e das informações

A empresa de Anderson, agora, conta com uma infraestrutura para armazenamento e processamento massivo de dados, o que é ideal para suportar um ambiente de *Big Data*, que combina os dados internos e externos da organização. As soluções de CRM podem ser beneficiadas a partir das novas bases formadas com os dados externos, os quais são coletados das redes sociais e dos demais canais disponíveis da Internet, o que permite complementar o entendimento em relação ao perfil, ao comportamento e ao interesse dos clientes.

A partir da coleta e do processamento de dados publicados em fóruns ou em sites especializados em notícias, é possível fazer análises de tendências a partir do ferramental disponível na ciência de dados. A soluções de BI também são potencializadas, pois, a partir do processo de estruturação dos dados coletados, eles podem compor novas dimensões e tabelas fato dos modelos dimensionais. Isso faz com que a camada tomadora de decisão tenha ainda mais elementos analíticos a serem utilizados na construção das estratégias e no acompanhamento dos indicadores organizacionais.

As soluções de BI e CRM possuem recursos gráficos para facilitar o consumo e a análise das informações e dos dados armazenados em suas bases de dados. Em outras palavras, essas soluções já têm formas predefinidas para fazer a entrega de valor às áreas de negócio e não exigem a presença de um profissional da área de dados para apoiar o consumo das informações presentes nessas soluções. Um ambiente de *Big Data* pode corroborar com as soluções analíticas já implantadas e utilizadas pela organização, mas ele terá muito mais valor enquanto fonte para a construção de novos modelos analíticos e para a realização estudos aprofundados com base em todo o ferramental da ciência de dados.

Todavia, nesse caso, não há uma estrutura formalizada para o processo de visualização das informações, a fim de entregá-las à camada tomadora de decisão. O processo de entrega e apresentação das análises feitas é tão importante quanto o próprio método utilizado para a execução da análise. De nada adianta fazer um estudo aprofundado utilizando várias técnicas estatísticas baseadas em inteligência artificial, se não for possível se conectar aos elementos do negócio e transmitir as principais ideias, *insights* e mensagens aos tomadores de decisão.

De acordo com Pereira (2015), a visualização de dados (*dataviz*) tem como objetivo a transformação de um dado em conhecimento. Ela é utilizada no processo de entendimento dos dados heterogêneos organizacionais, em que a prática da vi-

sualização de dados e de informações melhora o processo de tomada de decisão, aumentando o poder de análise da organização. Além disso, o “poder explicativo e exploratório da visualização de dados permite que o utilizador baseie as suas decisões em recursos visuais. Os meios visuais são mais eficazes do que os dados brutos” (PEREIRA, 2015, p. 24).

O uso da visualização gráfica pode trazer muitos benefícios. Aguilar *et al.* (2017, p. 10) expõem mais detalhes a seguir:



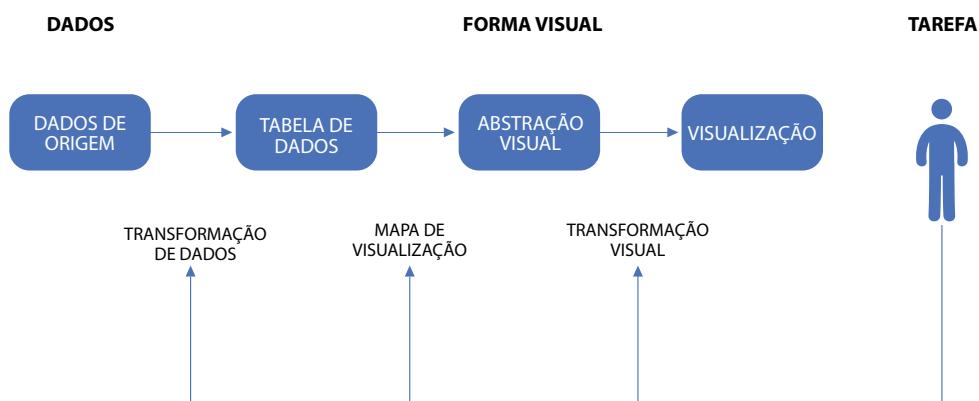
- Permite que se possa lidar com estruturas complexas de uma maneira ilustrativa e de fácil consumo, se fazendo valer de estruturas gráficas e cores para facilitar o entendimento do domínio e da mensagem que se deseja passar;
- Ajudar na percepção das propriedades existentes nos dados mapeados que não foram previstas inicialmente;
- Destacar problemas nos dados em sua coleção. Com um visualizador adequado, os erros nos dados são rapidamente detectados;
- Recolher simultaneamente propriedades grandes ou pequenas de dados;
- Redução dos recursos cognitivos mobilizados pelo usuário para processar e analisar as informações (alta interação do usuário, tendo suas percepções conduzidas para um fácil acesso à riqueza dos dados);
- Simplificação da busca por informação (uma grande quantidade de dados em um pequeno espaço, tendo a possibilidade de agrupamento de dados por critérios);
- Aumento das estruturas de possibilidade de detecção (relações de dados com a consolidação de se reagrupamento);
- Interferência perceptiva utilizando a percepção visual (alguns problemas são óbvios com uma representação visual, como um mapa de localização de metrô).

São muitas as vantagens ao se trabalhar com as visualizações gráficas a partir de dados armazenados. Os aspectos positivos são potencializados quando existem repositórios massivos de dados, já que não será necessário acessar todos os dados armazenados para realizar um estudo ou uma aplicação. De acordo com Aguilar *et al.* (2017, p. 11), a visualização facilita o processamento da informação de três formas:



- Permite um processamento mais rápido de informação relacional, seja por analogia direta, quando se trata de relações espaciais, seja por analogia metafórica, com base no poder de processamento da percepção visual;
- Representa externamente um conjunto de informações que já não são necessárias para manter a memória do trabalho, mas que pode ser acessado;
- Permite perceber tratamentos feitos ao dado, seja em formato simbólico ou via inferência.

No que se refere à visualização da informação, existem modelos de referência que fazem parte do processo contínuo de entendimento. A figura a seguir apresenta o modelo desenvolvido por Card *et al.* (1999 apud AGUILAR *et al.*, 2017, p. 11; PEREIRA, 2015, p. 32):



**Figura 8 - Modelo de processamento de informação**

Fonte: Card (1999 apud AGUILAR *et al.*, 2017, p. 11; PEREIRA, 2015, p. 32).

**Descrição da Imagem:** a figura apresenta a ideia de um fluxo. No primeiro passo, estão os «Dados de Origem», que são passados pela tarefa “Transformação de dados” na forma de “Tabela de Dados”. A partir da tarefa “Mapa de Visualização”, os dados são gerados como “Abstração Visual”. Por fim, a partir da tarefa “Transformação Visual”, é gerada uma “Visualização” a ser analisada por um usuário.

A figura apresenta todo o fluxo de maneira prática, desde o seu início, que se dá pelo dado, até a visualização por parte do usuário, além de explicitar a tarefa que está envolvida na migração de um passo ao outro. Um aspecto importante é o de que, a partir da tarefa “Transformação de dados”, é gerada uma “Tabela de dados”. Isso demonstra que existe um processo de estruturação dos dados, para que eles possam ser utilizados em um processo de abstração visual, o que só será

possível a partir do mapa de visualização, a fim de se alcançar a visualização pretendida pelo usuário final. Outro aspecto significativo é o de que cada pessoa pode perceber o recurso visual de maneira distinta, mas existem alguns padrões de percepção que são comuns aos seres humanos. Segundo Pereira (2015, p. 33), “os estudiosos descobriram que organizamos tudo o que vemos de forma a fazer sentido, chamando a este acontecimento de comportamento visual”.

O estudo mencionado pelo autor, o qual foi desenvolvido pela *Gestalt School of Psychology*, gerou, como resultado, princípios que podem apoiar o processo de visualização a partir dos grupos ilustrados na figura a seguir:

Proximidade	Objetos estão próximos uns dos outros que podem ser percebidos como grupos	
Similaridade	Objetos com atributos similares podem ser percebidos como grupos	
Caixa	Objetos que aparecem limitados por uma zona em comum podem ser percebidos como grupo	
Continuidade	Objetos aparecem alinhados ou com uma continuação do anterior podem ser percebidos como grupo	
Conexão	Objetos que aparecem interligados podem ser percebidos como um grupo	

Figura 9 - Princípios de Gestalt / Fonte: adaptado de Pereira (2015).

**Descrição da Imagem:** a figura apresenta os princípios de Gestalt. Cada princípio é apresentado em uma linha e existem três colunas. Na primeira, há o princípio, na segunda, é apresentada uma explicação e, na terceira, há uma representação gráfica. No total, são apresentados cinco princípios. O primeiro princípio é «Proximidade» e, na segunda coluna, está a seguinte explicação: “Objetos que estão próximos uns aos outros e podem ser percebidos como grupos”. A representação gráfica demonstra exatamente o que é descrito. O segundo princípio é “Similaridade” e a descrição para esse princípio é: “Objetos com atributos similares que podem ser percebidos como grupos”. O terceiro princípio é “Caixa” e a descrição apresentada é: “Objetos que aparecem limitados por uma zona em comum e podem ser percebidos como grupo”. O quarto princípio diz respeito à “Continuidade” e a sua definição é a seguinte: “Objetos que aparecem alinhados como uma continuação do anterior e podem ser percebidos como grupo”. Por fim, o princípio “Conexão” é apresentado e a sua descrição é a seguinte: “Objetos que aparecem interligados e podem ser percebidos como um grupo”.

A Figura 9 apresenta os cinco principais princípios de Gestalt, que são: proximidade, similaridade, caixa, continuidade e conexão. Eles são muito importantes e devem ser levados em consideração no momento de criação de um *dashboard*, apresentação ou infográfico.

Segundo Aguilar *et al.* (2017), existem vários tipos de gráficos que podem ser utilizados no processo de transcrição dos dados para uma forma gráfica. A seguir, são apresentados os principais tipos:



- **Barra:** um gráfico de barras (também chamado gráfico de colunas) permite a exibição ou comparação de vários conjuntos de dados. Os gráficos de barras mais úteis são: histograma e gráfico de barras empilhadas;
- **Curvas:** exibe os dados como um conjunto de pontos conectados por uma linha. Este tipo de diagrama é particularmente adequado para que apresentem os melhores dados em forma de vários grupos como, por exemplo, as vendas totais ao longo de vários anos;
- **Área:** exibe os dados como áreas ou superfícies, cada zona é reforçada pelas cores ou padrões diferentes. Este tipo de gráfico é o mais adequado para apresentações de dados para um número limitado de grupos;
- **Setores:** apresenta dados na forma de um gráfico de pizza com fatias diferentes ou seções, são enfatizados por cores ou padrões diferentes. Este tipo de diagrama pode mostrar apenas um grupo de dados;
- **Anel:** um diagrama de anel se assemelha a um gráfico de pizza que exibe dados por seções de um círculo ou de um anel. Com este tipo de gráfico é possível selecionar vários diagramas de anéis para vários conjuntos de dados;
- **Colunas 3D:** exibe dados em séries de objetos tridimensionais, dispostos lado a lado em planos tridimensionais. Também exibe valores das relações extremas, por exemplo, as diferenças de vendas por cliente e por país;

- **Superfície 3D:** apresenta uma visão topográfica de vários conjuntos de dados, por exemplo, é possível mostrar o número de vendas por clientes por país;
- **Nuvem XY:** um gráfico de dispersão XY é geralmente uma coleção de pontos plotados que representam dados específicos e que possuem uma riqueza de informação. Este diagrama permite que o usuário considere um tamanho de dados maior e determine sua tendência;
- **Radar:** um grupo de dados em posições gráficas em forma de radar ou teia de aranha. A inserção de diagramas de radar trabalha com valores numéricos de inputs e expõe seus valores do centro do radar para o extremo;
- **Bolha:** exibe dados como uma série de bolhas cujo tamanho é proporcional à sua qualidade. Quanto maior for a bolha maior será o número de produtos vendidos em uma região, por exemplo;
- **Bitola:** um indicador que mostra valores como pontos em um mostrador. Geralmente, é medido como um diagrama, em que os setores são usados para um grupo de dados, por exemplo, o percentual de vendas para todas as unidades populacionais;
- **Gantt:** é um gráfico de barras horizontais, muito utilizado para fornecer uma ilustração de um planejamento na linha do tempo, o eixo horizontal representa a linha do tempo e o eixo vertical a sequência de atividades;
- **Funil:** geralmente, um funil é utilizado para representar as fases de um processo de vendas, um funil é semelhante a um gráfico de barras empilhadas;
- **Histograma:** é um tipo de gráfico de barras usado para descrever a variação das medidas em relação a um valor médio. Pode ajudar a identificar a causa de um problema de um processo através do exame de sua distribuição, bem como a largura da distribuição (AGUILAR *et al.*, 2017, p. 22-24).

Joaquim ficou surpreso ao conhecer os aspectos a serem considerados na construção de recursos visuais a partir de dados. Entretanto, o que achou mais interessante foi o fato de que os itens por ele estudados poderiam ser utilizados tanto para a construção de *dashboards* quanto para a definição de outros recursos visuais para a entrega da informação. Joaquim sabia que, quando se trata de ciência de dados, nem sempre se tem um *dashboard* como forma de representação gráfica para as informações geradas. Além disso, em muitos casos, é necessário construir os recursos de maneira artesanal, considerando os elementos do domínio e, principalmente, a melhor forma de representar tal informação. A entrega da informação pode se dar em forma de relatório, de apresentação ou, em muitos casos, em forma de um infográfico.



#### EXPLORANDO IDEIAS

Os infográficos são representações visuais que podem combinar figuras, textos e elementos gráficos para representar uma ideia, notícia ou análise, por exemplo, a fim de gerar um maior engajamento ao leitor, de modo que ele interaja com o conteúdo apresentado, realize a sua reflexão e o entenda de maneira rápida (e, muitas vezes, descontraída).

Fonte: o autor.

Joaquim ficou se questionou se haveria uma forma de trabalhar com a visualização gráfica, de modo que ela fosse melhor aproveitada pelos tomadores de decisão, já que não há uma estrutura previamente construída, assim como é no caso dos *dashboards*. Motivado, pesquisou sobre boas práticas para a construção de apresentações gráficas a partir de dados e chegou até um conceito: ***storytelling* com dados**.

A prática de *storytelling* com dados (em inglês, “*Storytelling with data*”) foi cunhada por Cole Nussbaumer Knaflic, uma analista de dados que foi gerente da área de *People Analytics* do Google. Knaflic é reconhecida pelas apresentações e representações gráficas que fazia e é graduada e mestre em Matemática Aplicada pela Universidade de Washington. A estudiosa sempre se preocupou com o modo de apresentação dos dados e, assim, compilou um conjunto de boas práticas para a representação gráfica dos dados. Em consequência disso, teve a oportunidade de viajar por vários países dando palestras sobre esse conteúdo, que também foi transscrito em forma de um livro que recebe o mesmo nome.

Quando nos referimos ao *storytelling*, estamos falando da prática de formular narrativas enquanto histórias. Quando pensamos em histórias, geralmente, é feita uma introdução ou contextualização para apresentar os elementos do cenário ou do contexto. Depois, é desenvolvida a história e, ao final, é apresentado um fechamento, que pode ser resultante de uma conclusão, reflexão ou provocação. Essa capacidade de contar histórias pode ser utilizada em vários contextos, como em uma apresentação ou durante a elaboração um infográfico. Além do mais, a prática de contar histórias a partir de dados é uma importante ferramenta para o processo de disseminação e direcionamento dos resultados de análises, estudos e pesquisas feitas pela ciência de dados, ou seja, é uma prática que pode apoiar muito um ambiente de *Big Data*.

Segundo Knafllic (2017), existem alguns passos que são necessários para a visualização dos dados com sucesso. São eles:

- Descobrir o objetivo da apresentação.
- Aplicar uma análise exploratória dos dados em questão.
- Qual é o público-alvo que pretende atingir?
- O que quero que essas pessoas entendam ao final?

Em relação ao “descobrir o objetivo da apresentação”, precisamos refletir sobre o que desejamos apresentar. No contexto de Joaquim, ele pode ter feito uma análise exploratória dos dados coletados nas redes sociais e identificado que existe uma forte tendência de compra de guitarras Jackson de cor rosa, pelo fato de o artista John Mayer ter participado de várias gravações ao lado do Ed Sheeran tocando essa guitarra. Assim, é possível considerarmos que o objetivo da apresentação do Joaquim é sensibilizar o setor de compras para a aquisição de algumas peças do instrumento, a fim de que sejam vendidas na loja.

Para se chegar a esse resultado, foi necessário aplicar técnicas de análise exploratória nos dados em questão (KNAFLIC, 2017). Para que Joaquim consiga sensibilizar as pessoas do setor de compras, ele deve refletir sobre quem seriam as pessoas que assistiriam a sua apresentação ou receberiam o seu relatório. Ele se lembrou de que a maioria dos profissionais do setor são contadores e, em consequência disso, não haveria problema em analisar números e gráficos. Contudo, também se recordou de que o gerente da área é daltônico, o que exigiria cuidado com o uso das cores e, talvez, fizesse sentido utilizar textos e sombreamentos para realizar os destaques necessários.

Além da análise dos textos extraídos das redes sociais, Joaquim explorou os indicadores de vendas que já estavam disponíveis na solução de BI e apresentou várias informações em relação às vendas de guitarras Jackson e, sobretudo, as que tinham a cor rosa. Outra base utilizada foi criada por ele, na qual foram relacionadas as aparições de músicos ao instrumento utilizado. Além do mais, foi possível correlacionar essas informações com as vendas e verificar quais artistas impulsionam a venda de instrumentos. Agora, era o momento de apanhar esses dados e informações, a fim de organizá-los como uma espécie de história. Para isso, formulou uma sequência de tópicos e mensagens que queria apresentar:

- Comentários sobre a aparição de John Mayer e a guitarra Jackson rosa.
- Formular o engajamento dos usuários e as quantidades de comentários positivos.
- Evidenciar a evolução das vendas de guitarras da marca Jackson.
- Apresentar o aumento na venda de instrumentos a partir da aparição do artista em programas.

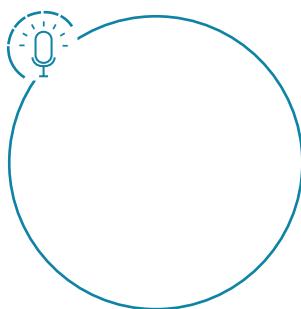
Joaquim já tinha um roteiro que continha os passos e a história que pretendia contar. Faltava apenas o último item apresentado por Knaflic (2017), que era: o que quero que essas pessoas entendam ao final? A mensagem que Joaquim queria transmitir era a de que o setor de compras deveria adquirir algumas guitarras Jackson de cor rosa. Foi dessa forma que concluiu a sua história e o resultado foi positivo. Todos os que assistiram à apresentação ou receberam o relatório (que seguia a mesma narrativa) elogiaram a forma com que ele apresentou os dados e as informações e concordaram em comprar alguns exemplares da guitarra para tê-las no estoque.

Joaquim ficou pensativo sobre o gestor da área de compras, que era daltônico, e constatou a importância de se pensar nesse fato antes de gerar os gráficos e os recursos visuais. Diante disso, perguntou-se se não teria um conjunto de boas práticas em relação à construção de gráficos e recorreu mais uma vez ao livro da Knaflic (2017). Nele, extraiu algumas dicas e fez questão de listá-las e enviá-las a todos os colegas da área de tecnologia. São elas:

- Quando há apenas um ou dois números a serem apresentados, utilize um texto simples e que dê destaque aos números em questão.
- Quando for apresentar tabelas com muitos dados numéricos, o uso de cor ou tons de uma cor podem apoiar a visualização (mapa de calor).

- Evite gráficos de pizza e procure utilizar gráficos de barra. Sempre ordene do maior valor para o menor.
- Cuidado com o uso de mais de uma escala lateral em gráficos compostos, pois pode gerar confusão no consumo da informação.
- Combine texto com gráficos e cores, a fim de guiar o entendimento de um conjunto de informações. Além disso, direcione a conclusão no título do gráfico.
- Procure não desenvolver gráficos muito carregados, visto que pode dificultar a leitura.

Foram obtidas todas as informações necessárias no que diz respeito às ferramentas de armazenamento e processamento, e às boas práticas para consumo, representação e apresentação dos dados extraídos do ambiente de *Big Data*. Joaquim apresentou tudo para Lara, que ficou feliz com o resultado. Depois disso, ambos foram apresentar os dados para Anderson, que parabenizou pelo excelente trabalho e disse que aguardava pelos resultados nas vendas. Se houvesse sucesso, haveria novos desafios para a nossa querida dupla. O que será que o destino estava reservando?



Quantas ferramentas interessantes Joaquim teve contato durante essa jornada! Como será a experiência de colocar um ambiente de Big Data em funcionamento? Será que um ambiente massivo de dados pode apoiar realmente o processo de tomada de decisão? Como deve ser a aplicação de storytelling em um ambiente corporativo?

Ficou curioso pelas respostas das perguntas? Não deixe de escutar o nosso podcast.

O primeiro grande desafio que a nossa dupla teve na etapa de implantação do ambiente de *Big Data* foi selecionar o tipo de banco de dados a ser utilizado em cada camada do *Data Lake*. A camada mais básica é a “camada de estagiamento” ou “camada de *staging*”. Nela, é preciso armazenar os dados brutos oriundos das mais distintas fontes de dados. Os dados internos da organização estavam todos estruturados em forma de banco de dados relacional, ou seja, havia uma visão em forma de tabelas. Já os dados coletados da Web, em sua maioria, eram recuperados no formato JSON.

Ao constatar a necessidade de uma possível expansão do ambiente, de tal forma que seja fácil o crescimento horizontal da infraestrutura de dados proposta, optou-se pela utilização de dois tipos de banco de dados NoSQL para compor a camada de *staging*: um banco de dados orientado a coluna (ou colunar), para se trabalhar com os dados estruturados e vindos dos bancos de dados relacionais, e um banco de dados orientado a documento, a fim de armazenar os dados não estruturados e coletados da *Web*.

Como banco de dados orientados à coluna, optou-se pelo uso do *Apache Cassandra* e, como banco de dados orientados a documento, optou-se pelo *MongoDB*. Esses dois bancos de dados foram configurados para serem utilizados de maneira distribuída e carregavam vários nós computacionais disponíveis em um serviço da nuvem. Para o processo de processamento distribuído, foi determinada a implantação da ferramenta que implementa o *MapReduce*, a chamada *Apache Hadoop*. Como ferramenta para a ingestão de dados em tempo real (que foi um requisito apresentado durante a execução do projeto), o *Apache Nifi* foi o escolhido.

As organizações, ao iniciarem o processo de construção de um ambiente de *Big Data*, sempre são surpreendidas com a grande quantidade de ferramentas disponíveis para atuar em aspectos muito específicos do projeto. O caminho mais natural para se chegar a uma solução robusta, mas simples, é navegando pelas características principais das ferramentas e relacionando-as com o cenário atual da organização. Além do mais, é necessário considerar o crescimento esperado para os próximos cinco anos e, a partir dessas informações, desenvolver um desenho arquitetural da solução, apresentando as principais ferramentas que comporão a infraestrutura básica de armazenamento e processamento de dados massivos.

As práticas de visualização de dados e de *storytelling* são de extrema importância para uma pessoa que se importa em apresentar os dados e as informações de maneira relevante. No entanto, para quem trabalha em um ambiente de *Big Data*, ambos são ainda mais importantes. Em consequência disso, Joaquim começou a fazer algumas apresentações internas para os setores da organização, disseminando os conhecimentos adquiridos sobre essa temática. Essa é uma prática que ajuda a evoluir a maturidade analítica das organizações. Em muitos casos, é reservada uma hora por semana apenas para que os profissionais que trabalham com ciência de dados compartilhem experiências e boas práticas. É somente assim que uma organização se torna dirigida por dados (*Data-Driven Organization*).



1. Os bancos de dados não relacionais são uma importante solução para os ambientes massivos de dados e os bancos NoSQL podem ser classificados em quatro tipos principais.

Assinale a alternativa correta:

- a) Orientados a *BigTable*, orientados a documentos, orientados a tabelas e orientados a grafos.
- b) Orientados a *BigTable*, orientados a documentos, orientados a tabelas e chave/valor.
- c) *MapReduce*, orientado a documentos, *BigTable* e chave/valor.
- d) Orientados a grafos, orientados a documentos, colunares e chave/valor.
- e) Orientados a tabelas, orientados a documentos, orientados a tuplas e chave/valor.

2. Em um ambiente de Big Data, para que seja possível ter acesso aos dados de maneira rápida, é necessário trabalhar com ambientes de processamento distribuído.

Assinale a alternativa que apresenta os três principais tipos de processamento distribuídos:

- a) Processamento em lote, em paralelo e em tempo real.
- b) Processamento em lote, de *streams* e em paralelo.
- c) Processamento em lote, de *streams* e em tempo real.
- d) Processamento em paralelo, de *streams* e em tempo real.
- e) Processamento em série, de *streams* e em tempo real.

3. Um ambiente de Big Data pode proporcionar um grande aumento na capacidade analítica de uma organização. Como existem dados de naturezas distintas e é possível gerar resultados que não seguem o padrão de apresentação de um *dashboard*, a forma de visualização dos dados e das informações são muito importantes.

Assinale a alternativa que se relaciona corretamente ao contexto apresentado:

- a) A aplicação de *storytelling* com dados apoia o processo de entrega de resultados e de entendimento por parte de quem está assistindo. Além disso, permite atingir o objetivo da apresentação.
- b) A prática de *storytelling* com dados é focada na construção de narrativas textuais. Entretanto, informações numéricas não podem ser apresentadas.
- c) A ideia principal da visualização de dados é a de otimizar ao máximo o espaço em tela. Quanto mais informações são apresentadas na tela, melhor é.
- d) A visualização dos dados não é relevante aos projetos que utilizam ambientes de *Big Data*.
- e) *DataViz*, como também é conhecida a visualização de dados, pode ser aplicada exclusivamente em *dashboards* relacionados às soluções de BI.

# CONFIRA SUAS RESPOSTAS



1. D.

Os quatro principais tipos de bancos de dados NoSQL são: orientados a documentos, orientados a grafos, colunares (ou de família de colunas) e chave/valor.

2. C.

Quando falamos em processamento distribuído, podemos classificá-lo em três principais tipos:

- Processamento em lote.
- Processamento de *streams*.
- Processamento em tempo real.

3. A.

A aplicação de *storytelling* com dados apoia o processo de entrega de resultados, permite o entendimento por parte de quem está assistindo e possibilita o alcance do objetivo da apresentação. Isso acontece, pois a estrutura de uma narrativa de fatos exige a exposição de uma conclusão, que deve ser a mensagem principal a ser entregue. Essa técnica pode ser aplicada em vários momentos distintos, tais como em uma apresentação, em um relatório ou até mesmo durante a organização de painéis em um *dashboard*.

# REFERÊNCIAS



AGUILAR, A. G. et al. **Visualização de dados, informação e conhecimento**. Florianópolis: Editora UFSC, 2017.

ALVAREZ, G. M.; CECI, F.; GONÇALVES, A. L. Análise comparativa dos bancos orientados a grafos de primeira e segunda geração: uma aplicação na análise social. In: ENCONTRO DE INOVAÇÃO EM SISTEMAS DE INFORMAÇÃO, 3., 2016, Florianópolis. **Anais** [...]. Florianópolis: EISI, 2016.

AMARAL, F. **Introdução à ciência de dados**: mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016.

DIANA, M. de; GEROSA, M. A. Nosql na web 2.0: um estudo comparativo de bancos não-relacionais para armazenamento de dados na web 2.0. In: WORKSHOP DE TESES E DISSERTAÇÕES EM BANCO DE DADOS, 9., 2010, [S. I.]. **Anais** [...]. [S. I.: s. n.], 2010.

FARAON, R. **Bancos de dados NewSQL**: conceitos, ferramentas e comparativo para grandes quantidades de dados. 2018 Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade de Caxias do Sul, Caxias do Sul, 2018.

HURWITZ, J. et al. **Big Data para leigos**. Rio de Janeiro: Alta Books, 2016.

KNAFLIC, C. N. **Storytelling com dados**: um guia sobre visualização de dados para profissionais de negócios. Rio de Janeiro: Alta Books, 2017.

KNOB, R. et al. Uma análise de soluções NewSQL. In: ESCOLA REGIONAL DE BANCO DE DADOS, 15., 2019, Porto Alegre. **Anais** [...]. Porto Alegre: SBC, 2019.

LÓSCIO, B. F.; OLIVEIRA, H. R. de; PONTES, J. C. de S. NoSQL no desenvolvimento de aplicações Web colaborativas. In: SIMPÓSIO BRASILEIRO DE SISTEMAS COLABORATIVOS, 8., 2011, Paraty. **Anais** [...]. Paraty: SBC, 2011.

MAGANO, F. de C. **Dataflows de tempo real como abstração para ferramentas de processamento de Big Data**. 2020 Tese (Doutorado em Ciência da Computação) – Universidade de São Paulo, São Paulo, 2020.

OLIVEIRA, S. S. T. de. **Explorando paralelismo em big data no processamento de séries temporais de imagens de sensoriamento remoto**. 2019. Tese (Doutorado em Computação) – Universidade de Goiás, Goiânia, 2019.

PEREIRA, F. P. A. **Big Data e Data Analysis**: visualização de informação. 2015. Dissertação (Mestrado em Engenharia e Gestão de Sistemas de Informação) – Universidade do Minho, Portugal, 2015.



# 5

# ***Big Data e as organizações***

Dr. Flávio Ceci

## ***OPORTUNIDADES DE APRENDIZAGEM***

Nesta unidade, acompanharemos o processo de implantação de uma área focada em dados em uma organização. Após a implementação de algumas iniciativas isoladas e relacionadas à ciência de dados, muitas empresas percebem a sua potencialidade e decidem estruturar uma área focada em dados (ou ciência de dados). Essa estruturação não é simples e demanda uma reflexão por parte de toda a camada tomadora de decisão. Assim, além de estudarmos as possibilidades de estruturação de uma área de dados, também discutiremos os papéis envolvidos e as suas principais atividades e responsabilidades.

Depois de alguns meses de trabalho constante, o ambiente de *Big Data* foi finalmente implantado na empresa de Anderson. Nesse sentido, foi possível estruturar um *Data Lake* em três camadas, ao migrar o *Data Warehouse* existente para a camada de indicadores do *Data Lake*. As bases operacionais e transacionais de todos os sistemas de informação foram replicadas na camada de *staging* e, além desses dados internos, coletores foram configurados para extrair os dados de redes sociais, sites de notícias e blogs especializados e armazená-los na camada em questão.

A camada intermediária entre a camada de *staging* e a de indicadores é a chamada camada de acesso. Nela, há uma visão unificada e já tratada dos dados da organização, sejam eles internos, sejam eles externos, mediante a utilização dos conceitos de dados mestres (MDM). Em paralelo ao desenvolvimento da infraestrutura de armazenamento massivo de dados, também foi configurado todo o ambiente para o processamento distribuído.

Quando foi iniciada a implantação, Lara ficou muito preocupada, pois os profissionais que atuavam na área de tecnologia não dominavam a maioria das tecnologias selecionadas por ela e por Joaquim. Para isso, tiveram que buscar novos profissionais. A maioria das pessoas selecionadas eram colegas de curso do próprio Joaquim. Em detrimento de o contexto de trabalho desses profissionais serem muito focados em ambientes de *Big Data* e de ciência de dados, Lara achou que seria mais correto pedir para que o próprio Joaquim fosse o gestor direto dessa nova equipe. Anderson adorou a ideia e disse que seria um excelente desafio.

Em paralelo ao processo de implantação de um ambiente de *Big Data*, Joaquim buscou outros profissionais com perfil analítico para compor a sua equipe. O novo gestor tomou essa atitude, visto que sabia que, a partir do momento em que essa nova estrutura fosse entregue, seria necessário mostrar os benefícios de maneira clara, sobretudo o retorno obtido a partir do investimento feito, ou seja, qual foi o real benefício financeiro obtido.

Passaram-se três meses desde a implantação do *Data Lake*. Joaquim e os membros de sua equipe conseguiram gerar modelos de previsão de *churn* e modelos de apoio à área de CRM, à área de marketing e à área comercial. Os resultados foram ótimos! Mesmo em um curto espaço de tempo, já eram perceptíveis os resultados da empresa e o investimento foi pago em pouco mais de um mês. Joaquim estava muito feliz em poder conduzir um time voltado aos dados e tinha total apoio de sua gestora direta. Entretanto, ele não sabia que Anderson e os demais diretores tinham outros planos para a nossa querida dupla.

Era uma segunda chuvosa e Joaquim estava monitorando os processos de carga que rodaram durante a madrugada e o final de semana para o *Data Lake*. Lara entrou na sala onde ele e a equipe estavam e disse a Joaquim que os diretores gostariam que os dois comparecessem até a sala de reuniões da presidência. Era um assunto urgente. Ao entrarem na sala, foram recebidos com uma salva de palmas. Os dois sorriram, mas não sabiam o que aconteceria ali. Anderson tomou a frente e iniciou o seu pronunciamento. Primeiramente, comentou toda a trajetória de Lara dentro da empresa, incluindo os seus feitos e como ela foi importante para a empresa. Depois, disse que, naquele momento, ela deixaria de ser a gerente de tecnologia e passaria a ter uma cadeira entre os diretores enquanto diretora de tecnologia.

Joaquim ficou muito feliz por Lara, uma vez que ele sempre a admirou e aprendeu muito com ela. Continuando o seu pronunciamento, Anderson começou a falar sobre Joaquim. Assim, explicou o modo como ele potencializou as entregas da área de tecnologia e defendeu que ele era a pessoa que se preocupava com os dados. Além disso, enfatizou a garantia dada por Joaquim de que novas tecnologias analíticas pudessem ser desenvolvidas, a fim de apoiar todos os processos estratégicos da organização.

Depois disso, Anderson olhou nos olhos de Joaquim e falou: “Lara assumirá novas tarefas como executiva da empresa. Dividiremos a área de tecnologia em três gerências e, infelizmente, a parte voltada à ciência de dados não estará em nenhuma delas”. Joaquim abaixou a cabeça nesse momento, mas continuou escutando o que Anderson dizia: “Joaquim, você foi um dos grandes fomentadores da inovação dessa organização. Você e Lara foram os responsáveis por elevar a maturidade analítica da organização a níveis que eu desconhecia e, em consequência disso, queremos te oferecer uma cadeira de gerente para uma nova área de dados que será criada. Além do mais, entendendo a importância dessa temática, a sua gerência responderá direto à presidência. Se tudo ocorrer assim como imaginamos, ela se tornará uma diretoria em um ano”.

Joaquim vibrou de alegria, uma vez que sempre sonhou em ser gestor de uma área de dados. Ao voltar para a sua mesa, constatou que tinha um mês para desenhar, organizar, estruturar e iniciar a área, além de formular a sua equipe e desenhar as principais atividades e responsabilidades. Diante disso, surgiu a seguinte questão: como fazer todas essas ações? Como as empresas elaboram uma área de dados e *analytics*?

Diferentemente de áreas, tais como a de desenvolvimento de software, a área de dados que utiliza os ambientes de *Big Data* e demais ferramentas da ciência de dados é muito recente e não evidencia os papéis, as atividades, as responsabilidades e os processos envolvidos. Além disso, as áreas de dados podem ser evoluções de áreas mais tradicionais, como a área de *Business Intelligence*, a área de modelagem estatística (muito comum em instituições financeiras) e a área de banco de dados.

O que faz a área de dados ser complexa em sua estruturação é a sua natureza multidisciplinar. Já sabemos que a ciência de dados tem três pilares: computação, estatística/matemática e negócio. Assim, essa área inclui pessoas com formações multidisciplinares e trajetórias profissionais distintas. No entanto, esses profissionais devem formular mesmo objetivo estratégico para a organização.

Quando nos referimos à área de dados, dizemos respeito à área que concentra os principais processos analíticos de uma organização e as questões relacionadas à infraestrutura de dados. A construção de uma área de dados deve ser feita e orientada com base nos objetivos estratégicos do negócio, na infraestrutura de dados disponível e na maturidade analítica da organização. Alicerçado nesses três aspectos, é possível desenhar uma área que seja capaz de guiar e apoiar a organização em seus desafios e possibilidades em relação ao envolvimento de dados.

Joaquim concluiu que tinha um grande desafio pela frente. Com toda a certeza, seria o maior de sua carreira até o prezado momento. Ele sabia

que estruturar essa nova área não seria fácil, mas, assim como era de costume, o primeiro passo seria realizar uma pesquisa para entender mais detalhes sobre os elementos do projeto. Nesse contexto, Joaquim buscou relatos e estudos de caso voltados a esse tipo de tarefa.



O que você acha de ajudar o nosso querido Joaquim em mais essa pesquisa? Na Internet, procure relatos ou estudos de casos de pessoas ou empresas que já tenham passado pelo processo de implementação de uma área relacionada à ciência de dados. Um aspecto importante é identificar os papéis envolvidos em uma área de ciência de dados.

É visível que não há um único caminho para a construção de uma área de dados e não há muitas informações disponíveis na Internet. Todavia, qualquer caso de sucesso é obtido a partir de uma área de base, tais como tecnologia, de dados, BI, modelagem, dentre outras. O importante é que os dirigentes estejam preparados para ter uma organização dirigida por dados e, para isso, a área deve ter autonomia em relação às outras áreas de negócio ou de tecnologia para elaborar e implantar as suas ferramentas e processos.

Chegou o momento de fazer as suas anotações acerca dos aspectos que lhe chamaram a atenção a partir da pesquisa feita. Não deixe de anotar as principais ideias identificadas na criação de uma área de dados, os papéis envolvidos e as suas principais atividades. Esses são os elementos que nos apoiarão durante esta unidade, a fim de sabermos como Joaquim lidará com esse novo desafio. Aperte os cintos e sigamos em frente!

## DIÁRIO DE BORDO

## Área de dados

Enfim, chegou o primeiro dia de estruturação da área de dados de Joaquim. Como de costume, Joaquim abriu o seu navegador de Internet e, a partir de um motor de busca, encontrou blogs e páginas que explicavam como se dá a elaboração de equipes e de áreas relacionadas à ciência de dados. O primeiro aspecto que ele deveria pensar é: haveria uma **estrutura totalmente centralizada ou distribuída pela organização?** Em outras palavras, todas as análises, modelos e indicadores deveriam sair, ou não, de sua área?

O objetivo da empresa de Anderson é o de se tornar uma organização dirigida por dados (*Data-Driven Organization*). Hoje, os executivos já fazem uso de indicadores e estudos para apoiar o seu processo de tomada de decisão. O mesmo procedimento acontece em algumas áreas do negócio, nas quais são solicitadas evoluções nos *dashboards*, para que eles sejam utilizados nas decisões gerenciais. Contudo, ainda existem áreas que não se beneficiam de indicadores e há uma longa fila de solicitações de novos indicadores e estudos. Esses pedidos eram gerenciados por Lara e repassados à sua equipe de tecnologia.

Diante do exposto, é perceptível que há um obstáculo na evolução da maturidade analítica da empresa de Anderson em consequência da centralização de diversas atividades relacionadas a uma mesma área. Joaquim estava muito confuso em relação à qual caminho seguir: se ele simplesmente migrasse as pessoas da diretoria de tecnologia para a área de dados e mantivesse os mesmos processos, a tendência era manter o obstáculo em relação à evolução da maturidade analítica da organização. Esse problema poderia ser resolvido a partir do aumento do número de funcionários e da equipe. Todavia, será que, ao entregar os recursos analíticos mais rapidamente, não haveria um aumento na velocidade das solicitações por novos indicadores e estudos?

Joaquim fez uma retrospectiva de tudo o que foi criado no contexto dos dados dentro da área de tecnologia desde a sua chegada. Foi preciso muito esforço para estruturar as bases analíticas da empresa e foi necessário desenvolver processos e políticas de governança de dados, a fim de apoiar o desenvolvimento das novas soluções analíticas. Quando foi definida a implantação de um ambiente de *Big Data*, foi exigida a contratação de profissionais com compe-

tências distintas. O cenário que a empresa se encontra hoje só foi possível em consequência dessa trajetória, ou seja, quando nos referimos à estruturação de uma área de dados em uma organização que tem uma maturidade analítica muito baixa, faz sentido que a primeira abordagem seja centralizada, pois, dessa forma, é possível desenvolver os repositórios analíticos massivos de maneira organizada e governada. Além disso, é possível criar os primeiros processos e entregáveis analíticos para a organização, o que garante que as entregas sejam feitas com qualidade e mediante o uso de dados já validados e reais.

Depois de realizar a sua retrospectiva, Joaquim desenhou uma figura para facilitar o seu entendimento sobre as principais características de uma abordagem centralizada e distribuída. Você pode estar se perguntando: afinal, como seria cada uma dessas abordagens na prática? Essa é uma importante questão para entender como se dá cada um desses modelos. O primeiro passo é entender o conceito de **agente de dados**. Um agente de dados, em uma organização, diz respeito a qualquer colaborador que faz uso de dados e de ferramentas analíticas para executar as suas análises ou para fazer a construção de painéis, relatórios, modelos e dentre outros recursos de apoio à decisão.

Em organizações mais tradicionais, como as do mercado financeiro, é muito comum encontrar os analistas de *Management Information System* (MIS), que são os profissionais responsáveis por analisar os dados a partir dos sistemas de informações gerenciais e das bases de dados dos sistemas de informações transacionais, a fim de gerar relatórios e fazer várias análises mediante o uso de diversas planilhas eletrônicas. Esses profissionais, geralmente, são distribuídos por áreas de negócio em uma organização. Outro exemplo são os analistas de BI, que também podem estar alocados em áreas de negócio distintas e geram indicadores e painéis analíticos gerenciais, para apoiar os gerentes em sua tomada de decisão. Esses são dois exemplos de agentes de dados distribuídos pela organização.

Quando há um negócio com baixa maturidade analítica e os agentes de dados estão distribuídos entre as áreas de negócio, problemas com indicadores podem ocorrer, uma vez que eles podem ser gerados com o mesmo nome e apresentar valores diferentes. Isso acontece justamente pela falta de bases de dados, processos e políticas governadas. A figura a seguir apresenta mais detalhes sobre as características das duas abordagens:

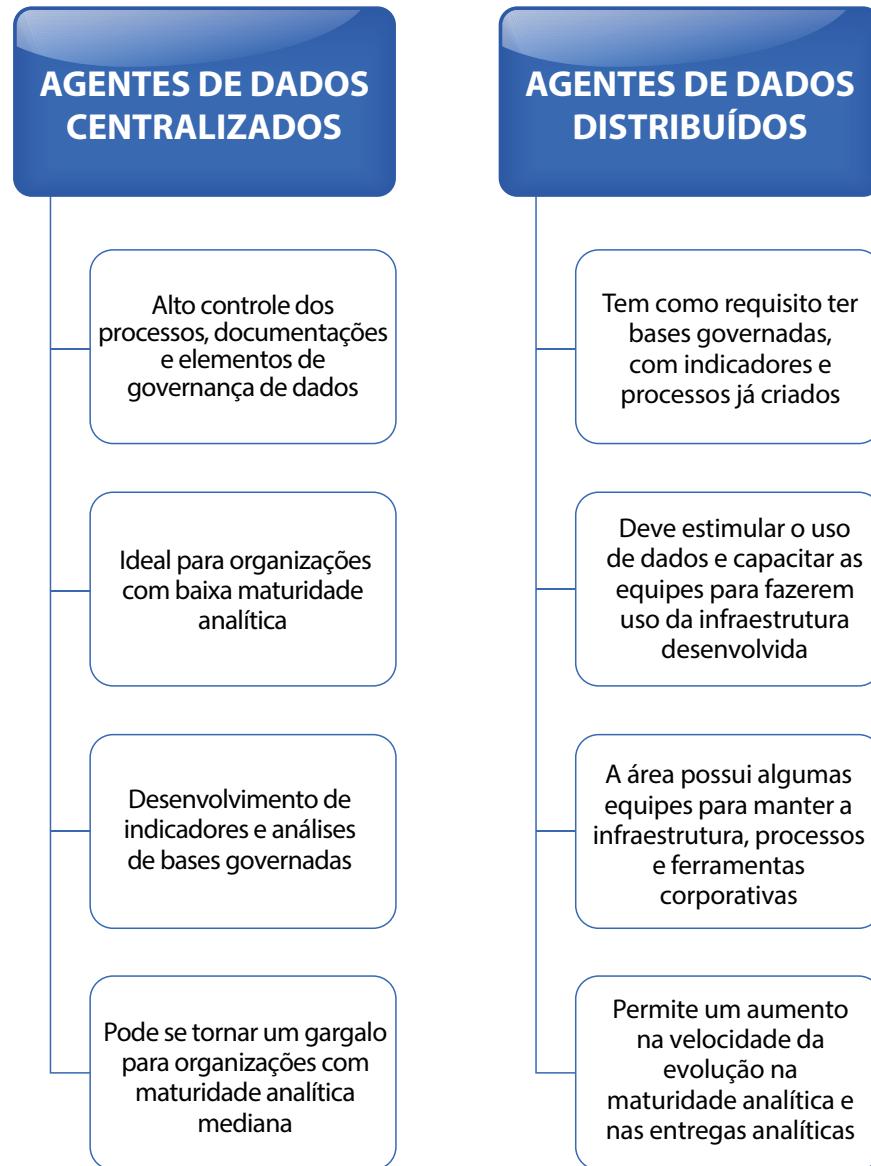


Figura 1 - Características das áreas de dados centralizadas e distribuídas / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta duas colunas. A primeira diz respeito aos "Agentes de dados centralizados", que têm as seguintes características: "Alto controle dos processos, das documentações e dos elementos de governança de dados", "Ideal para as organizações com baixa maturidade analítica", "Desenvolvimento de indicadores e análises de bases governadas" e "Pode se tornar um gargalo para as organizações com maturidade analítica mediana". Já a segunda coluna se refere aos "Agentes de dados distribuídos", que têm as seguintes características: "Exige a existência de bases governadas, com indicadores e processos já criados", "Deve estimular o uso de dados e capacitar as equipes para que façam o uso da infraestrutura desenvolvida", "A área tem algumas equipes para manter a infraestrutura, os processos e as ferramentas corporativas" e "Aumenta a evolução da maturidade das entregas analíticas".

Joaquim entendeu que a empresa já tem maturidade analítica mediana, o que permite que a área seja construída com os agentes de dados de maneira distribuída. Em outras palavras, Joaquim deve estimular o desenvolvimento dos analistas das áreas de negócio, para que tenham condições de desenvolver as suas visões analíticas sem depender de pessoas da sua área. Para isso, devem se pautar no *Data Lake* desenvolvido e nas ferramentas corporativas. Além disso, serão necessárias a realização de treinamentos e a elaboração de guildas de temáticas relacionadas à área. Dessa forma, será possível compartilhar as questões técnicas, as boas práticas e os conceitos para todas as áreas da empresa, disseminando, assim, a cultura de dados, o que proporcionará a evolução das competências analíticas dos agentes de dados.



### EXPLORANDO IDEIAS

O conceito de **guilda** é muito comum em organizações que trabalham com times multidisciplinares e autogerenciáveis que se organizam em **squads**. As squads tem como foco um produto, um projeto, um módulo ou um serviço específico e todos os integrantes da equipe devem estar 100% concentrados no objetivo. Considerando o fato que, na equipe, existem profissionais com várias competências (programadores, designers, analistas de negócio, testers, dentre outros), a troca de conhecimento técnico entre os profissionais com competências similares é garantida nas guildas. Para exemplificar, as guildas permitem que todos os programadores de squad distintas possam se reunir para que troquem experiências e apresentem boas práticas.

Fonte: o autor.

Quando Joaquim pensou em desenvolver uma guilda para os agentes de dados da organização, o seu objetivo era o de garantir a realização de um encontro formal entre todos os profissionais envolvidos, a fim de disseminar experiências, boas práticas e apresentar novas funcionalidades, ferramentas e serviços desenvolvidos pela área de dados. A guilda recebeu o nome “Guilda de Dados”. Além do mais, foi acordado com os diretores que haveria uma reunião uma vez por semana, com duração de duas horas.

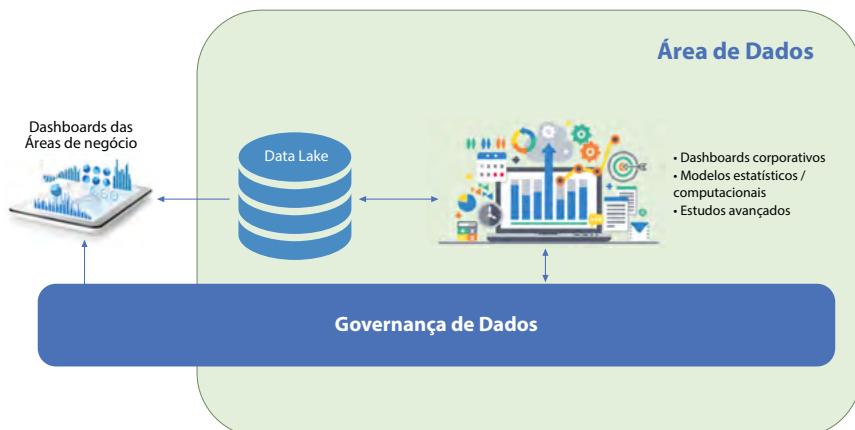
A área de dados já teve duas definições importantes: ela trabalhará com os agentes de dados distribuídos entre as áreas de negócio da empresa e deve garantir a disseminação de conhecimento entre os profissionais que estão na área de dados e os que estão distribuídos nas áreas de negócio. No entanto, de fato, como se deve organizar a área de dados?

Após algumas pesquisas, Joaquim leu nas obras de Aiken e Gorman (2014) e Anderson (2015), que a área de dados deve responder diretamente ao presidente da empresa. Isso significa que ela deve ser autossuficiente, ou seja, não deve depender de profissionais de outras equipes, mesmo que tenham competências similares. Por esse motivo, Joaquim não iria mais responder diretamente a Lara e seria necessário mover alguns profissionais da área de tecnologia para a sua nova área. Assim, o primeiro passo foi fazer uma lista dos recursos e das ferramentas desenvolvidas pela área de tecnologia que deveriam ser levados para a sua área. Lara o ajudou a pensar nos elementos presentes na lista, que é apresentada com mais detalhes a seguir:

- **Data Lake:** todas as bases analíticas estão concentradas nas três camadas do *Data Lake*. Essa base deve ser gerenciada e mantida na área de dados, pois é nela que estão concentrados todos os indicadores e os dados já processados e validados em forma de MDM. O *Data Lake* deve ser acessível para todas as áreas de negócio da organização, sempre respeitando as políticas de acesso e segurança para que cada agente de dados acesse somente as informações previamente acordadas entre os gestores das áreas de negócio.
- **Solução de BI:** a construção de *dashboards* deve ser descentralizada, ou seja, cada área de negócio deve ter a capacidade de construir as suas visualizações a partir dos indicadores já modelados dentro da camada de *Data Warehouse* do *Data Lake*. Se for necessário elaborar novos indicadores, é preciso solicitar à área de dados, para que sejam respeitadas todas as dimensões de governança. A área de dados é a responsável por gerar os indicadores executivos, que são consumidos pelos diretores e pela presidência da empresa. Dessa forma, há um único aspecto para a validação dos principais indicadores que são utilizados pelos executivos.
- **Governança de dados:** deve ser acessível a toda a organização, de modo a apoiar todas as operações relacionadas aos dados das áreas de negócio. A área de dados é a responsável por orquestrar as definições dos processos de governança e fazer o processo de monitoramento e validação dos recursos analíticos que são desenvolvidos na organização. Além disso, pode chancelar selos de maturidade analítica (ouro, prata e bronze) para as bases, os indicadores e os *dashboards* desenvolvidos, a fim de saber em qual camada da organização é possível utilizar essas informações para o processo de tomada de decisão.

- **Modelagem estatística e computacional:** os processos relacionados à construção de modelos para classificação, *score* e previsão demandam profissionais versados em ferramentas e em técnicas estatísticas, matemáticas e computacionais avançadas. Em organizações com maturidade analítica muito alta, talvez, esses recursos também sejam distribuídos, mas, no contexto da empresa de Anderson, é mais seguro concentrá-los na área de dados, para que os processos de governança sejam respeitados e para que haja agilidade no processo de construção de bases de dados para a realização de estudos e para a construção de modelos.
- **Estudos avançados:** os estudos avançados, geralmente, são parte de um processo de construção de modelos em que são aplicadas técnicas de análise exploratória a partir de uma massa de dados. Esse recurso pode ser distribuído em empresas com alta maturidade analítica, mas, na empresa de Anderson, faz sentido a sua permanência na área de dados.

A figura a seguir apresenta os elementos que são de responsabilidade da área de dados e das áreas de negócio:

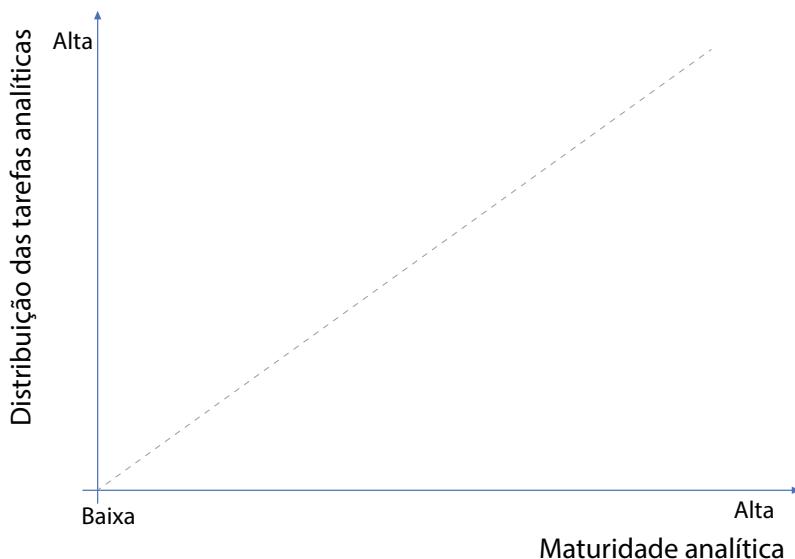


**Figura 2 - Relação entre os elementos da área de dados e as demais áreas de negócio**  
Fonte: o autor.

**Descrição da Imagem:** a figura apresenta uma área delimitada que tem todos os elementos que são de responsabilidade da área de dados. Assim, há uma figura que representa o Data Lake, que é acessado por uma imagem que ilustra as ferramentas analíticas. Nessa ilustração, estão descritos três elementos: dashboards corporativos, modelos estatísticos/computacionais e estudos avançados. Também há uma camada em que a sua maior parte está inclusa na área de dados, que é a governança de dados. Por fim, há uma figura que demonstra os indicadores em um dashboard com o seguinte título: "Dashboards das áreas de negócio". Além de ele estar ligado à governança de dados, também tem acesso ao Data Lake.

A figura apresentada demonstra como as ferramentas herdadas da área de tecnologia e os recursos estão dispostos a partir da criação da área de dados. É perceptível que, no atual momento da organização, somente há a geração de relatórios e a distribuição dos *dashboards* pelas áreas de negócio. Vale ressaltar que, à medida que a maturidade analítica organizacional cresce, a distribuição das tarefas analíticas também é repassada, porque os agentes de dados estarão cada vez mais proficientes em métodos e em técnicas analíticas. Além do mais, as bases e os processos estarão muito mais maduros. Em paralelo, há a governança de dados, que é orquestrada e gerida pela área de dados e executada por toda a organização.

Para demonstrarmos a relação existente entre a maturidade analítica organizacional e a autonomia das áreas de negócio, observe a figura a seguir:



**Figura 3 - Relação entre a distribuição das tarefas analíticas e a maturidade organizacional**  
Fonte: o autor.

**Descrição da Imagem:** a figura apresenta um gráfico de linha. No eixo vertical, está escrito: "Distribuição das tarefas analíticas". Em sua extremidade inferior, está redigido "baixa" e, na extremidade superior, está escrito "alta". No eixo horizontal, está escrito "Maturidade analítica" e, em seu início, está redigido "baixa" e, no fim, "alta". Entre os eixos, existe uma linha tracejada que demonstra a relação linear entre as duas dimensões. Assim, quanto maior for a maturidade analítica da organização, maior poderá ser a distribuição das tarefas analíticas nas áreas de negócio.

A figura apresentada demonstra a relação direta que existe entre a maturidade analítica e a possibilidade de distribuição das tarefas. Isso só é possível pelo fato de haver uma governança de dados robusta e de fácil acesso por toda a organização.

Joaquim e Lara já haviam obtido uma proposta de solução em relação ao que deveria ser movido da área de tecnologia para a nova área de dados. O próximo passo era transferir os profissionais para a nova área em questão, mas muitos deles não teriam mais a mesma função. Diante disso, foi necessário revisitar as contratações. Além dos profissionais que fazem parte da área de tecnologia, seriam necessários novos perfis profissionais para suportar os novos tipos de demandas que a área de dados atenderá. Nesse contexto, Joaquim realizou uma pesquisa em livros e em redes sociais profissionais para obter uma visão mais ampla do perfil dos membros de uma área de dados.

## Perfis dos membros de uma área de dados

Em consequência de a área de dados ou de ciência de dados ser uma estrutura nova na maioria das organizações, não há uma definição única no que diz respeito aos profissionais envolvidos e à descrição dos cargos. Nesse sentido, Joaquim fez um levantamento levando em consideração as empresas de grande ou de médio porte residentes no Brasil e chegou a um conjunto composto por seis cargos (ou perfis) existentes em áreas dessa natureza. Os cargos são: engenheiro de dados, analista de BI e negócios, analista de dados, cientista de dados e estatístico e engenheiro de aprendizado de máquina (*machine learning*).

O **engenheiro de dados** é o profissional responsável por toda a infraestrutura de dados. Pelo fato de haver muitos elementos envolvidos, é possível encontrar subdivisões ou, ainda, especializações a partir desse cargo. De maneira geral, o engenheiro de dados é o responsável por implantar e configurar toda a infraestrutura física do banco de dados relacionais ou não relacionais. Além disso, pode formular ambientes de *Big Data* levando em consideração o armazenamento e o processamento distribuído. Também é o responsável pela construção de processos de ETL (extração, transformação e carga) entre as várias bases de dados da organização e pela edificação de *Data Warehouses*.

Os profissionais que, hoje, exercem a função de engenheiros de dados, geralmente, ocupavam o cargo de programadores ou de *Database Administrator* (DBA ou, em português, “administrador de banco de dados”). Isso aconteceu pela necessidade desse profissional ter um conhecimento avançado em relação ao banco de dados e à programação. As principais áreas de especialização que esse profissional pode escolher são:

- **Infraestrutura:** toda a infraestrutura de um banco de dados relacionando às bases analíticas pode ficar aos cuidados do engenheiro de dados. Vale lembrar que, aqui, são considerados os bancos de dados de sistemas de informações transacionais e operacionais, e apenas os repositórios de dados analíticos, pensando em soluções que possam ser escaláveis de maneira distribuída.
- **Big Data:** o profissional será o responsável pela construção de ambientes de *Big Data*, ao envolver as ferramentas para armazenamento e processamento massivo de dados, geralmente, utilizando ambientes distribuídos e escaláveis. Além da construção, deve garantir que o ambiente esteja produtivo mediante o seu monitoramento e gestão dos recursos.
- **Nuvem (cloud):** a necessidade de ambientes escaláveis faz com que as abordagens em nuvem sejam uma excelente opção, a fim de se aproveitar ao máximo os serviços disponíveis a um custo de operação muito menor por fornecedor.
- **ETL:** todo o processo de ETL, independentemente do ambiente disponível, é o responsável por operar as ferramentas, monitorar os processos de carga e otimizá-los.
- **Segurança:** o processo de acesso e disponibilização dos dados é bastante sensível, principalmente em um ambiente na nuvem, o que exige processos e rotinas de segurança dentro das atividades dos engenheiros de dados. Em muitas empresas, há uma área especial para tratar dos assuntos relacionados à segurança. Nesses casos, basta que os engenheiros de dados sigam os protocolos e os procedimentos de segurança. Caso eles não existam, é uma grande tarefa que deve ser conduzida.

Outro aspecto que os engenheiros de dados são responsáveis é pela modelagem das bases de dados para fins analíticos. É muito comum que as áreas de negócio que já possuem agentes de dados precisem de bases para estudo (ou de analistas internos da área de dados) e cabe a eles fazer a modelagem e a carga dessas bases. Quando o destino da modelagem são os modelos dimensionais, a fim de compor as estruturas de um *Data Warehouse*, é necessária a presença de um outro importante perfil da área de dados: o analista de BI e de negócio, que pode estar alocado nas áreas de negócio ou fazer parte da área de dados. A sua principal responsabilidade no processo da modelagem das bases é apresentar os requisitos e os elementos do negócio, para que eles possam ser modelados seguindo as regras de negócio da organização e do domínio em questão.

O **analista de BI e de negócios** é o agente de dados responsável por entender e mapear as regras de negócio de cada domínio. Além disso, auxilia na construção da modelagem dimensional (*Data Warehouse*) e na formulação de painéis, relatórios e *dashboards*. Em grandes empresas ou em segmentos mais tradicionais, como o mercado financeiro, os analistas de BI e de negócio são, geralmente, profissionais que desempenhavam a função de analista de MIS. Além do mais, em algumas instituições, essa função é dividida em analista de negócio e em analista de BI. Na empresa de Anderson, levando em consideração o seu porte e o seu grau de maturidade analítica, a criação de um único cargo foi optada. A figura a seguir introduz as principais competências desse profissional:



**Figura 4 - Principais competências de um analista de BI e de negócio / Fonte: o autor.**

**Descrição da Imagem:** a figura apresenta um funil com três círculos. Cada círculo apresenta uma competência do analista de BI e de negócio, a saber: banco de dados, negócio e ferramentas de análise. Na saída do funil, está escrito “Analista de BI e de negócio”, a fim de gerar a ideia de junção das três competências na composição do profissional em questão.

A figura apresenta as três principais competências que o analista de BI e de negócios deve ter para exercer o seu trabalho dentro de uma área de negócio ou de uma área de dados. A seguir, são detalhadas cada uma das competências:

- **Negócio:** o analista deve conhecer os conceitos e as regras relacionadas ao domínio em questão. Quando alocado em uma área de negócio, ele já está inserido nos principais processos e conhece bem as bases de dados utilizadas. Por outro lado, quando alocado na área de dados, é necessário que se aprofunde no domínio em questão, de tal forma que seja a principal interface entre esses dois mundos (a área de negócio e a de dados).
- **Banco de dados:** é esperado tanto o conhecimento voltado à modelagem de bases de dados levando em consideração os elementos do negócio quanto o conhecimento avançado em relação à construção de consultas (SQL), para que seja possível extrair dados e informações e desenvolver estudos e análises.
- **Ferramentas de análise:** os analistas devem dotar de conhecimento em relação às ferramentas de análise e visualização de dados. Eles podem ser acionados para gerar relatórios que apresentem informações gerenciais ou para a construção de painéis e *dashboards* em uma arquitetura de BI. O profissional deve conhecer todos os elementos relacionados à arquitetura tradicional de BI e pode desenvolver algumas ETLs simples para a construção de estudos, sempre respeitando a governança de dados.

A governança de dados (GD) é transversal a todas as áreas da organização. Em muitos casos, é o analista de BI e de negócio o responsável por garantir a aplicabilidade dos processos e das políticas de GD dentro das áreas de negócio. Quando fazem parte da área de dados, geralmente, são envolvidos no processo de construção das políticas e na execução dos processos. Assim, podem apoiar todo o processo de monitoramento dos indicadores da própria área, o qual também pode ser sustentado pelos analistas de dados.

Os **analistas de dados**, assim como os analistas de BI e de negócio, podem estar distribuídos pelas áreas de negócio ou podem estar alocados na área de dados. De modo geral, o foco desses profissionais está ligado ao acompanhamento dos indicadores existentes nas soluções de BI. Entretanto, também desenvolvem relatórios e estudos avançados a partir das bases de dados. Esses analistas, normalmente, têm, como formação, cursos com base forte em matemática. São exemplos: Estatística, Matemática, Economia, Engenharia, Contabilidade e Computação.

O analista de dados tem certa similaridade com o analista de BI e de negócio. A diferença entre ambos se dá no fato de que, geralmente, o analista de dados trabalha mais com os métodos estatísticos e matemáticos. A figura a seguir apresenta mais detalhes:

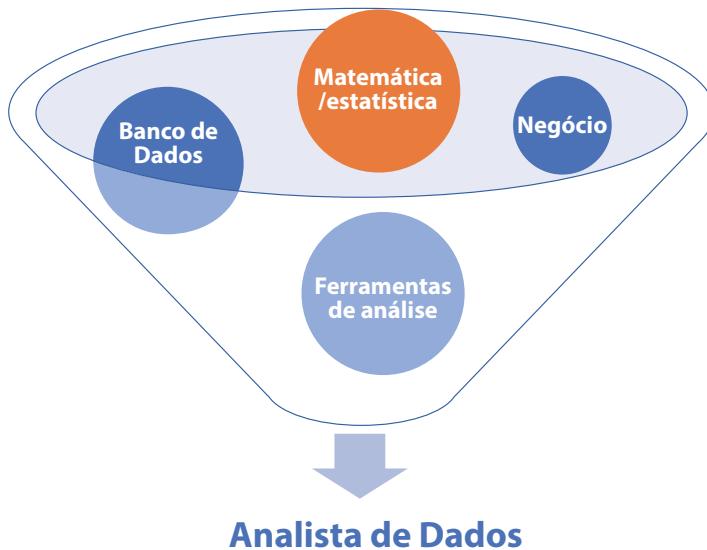


Figura 5 - Principais competências de um analista de dados / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta um funil com quatro círculos dentro. Dentre todos, três círculos são apresentados no funil do analista de BI e de negócio, que são: banco de dados, negócio e ferramentas de análise. Contudo, o último círculo, que é intitulado “Matemática/estatística”, carrega uma cor diferente, a fim de evidenciar que se trata de uma competência nova.

Na grande maioria das empresas, o analista de dados tem a competência de negócio (regras de negócio) menor que as demais, porque ele não objetiva tanto as regras de negócio, mas os conceitos relacionados. Nesse sentido, esse profissional trabalha muito próximo dos analistas de BI e de negócio. A competência “matemática/estatística” apoia o seu trabalho, ao proporcionar uma análise aprofundada sobre determinadas situações ou cenários. Desse modo, o analista de dados pode identificar um comportamento incomum a partir da análise dos indicadores existentes em *dashboards*. Depois, pode realizar uma análise exploratória dos dados, a fim de se obter conclusões, o que faz com que esse tipo de analista também esteja em parceria com os cientistas de dados. Portanto, os analistas de dados podem desenvolver estudos utilizando a análise exploratória, mas, quando é necessário desenvolver algum modelo ou realizar um aprofundamento nas análises, os cientistas de dados também são alocados.

O **cientista de dados** é um profissional com competências multidisciplinares e o seu objetivo é ser o arquiteto de soluções analíticas em uma organização. Para tanto, deve ter condições de analisar um problema ou problemática, desenhar as etapas metodológicas para se chegar à proposta de solução, envolver os demais profissionais no desenho da proposta, além de implementar e testar os resultados. Em muitas empresas, o trabalho do analista de dados é dividido entre cientista de dados e analista de BI e de negócio. Essa classificação é dependente do tamanho da organização e do grau de maturidade analítica que a empresa se encontra.

Já sabemos que o cientista de dados é um arquiteto de soluções analíticas que deve conhecer todos os processos já apresentados nesta unidade, mesmo que de maneira superficial, para que possa construir desenhos de funcionalidades, soluções para estudos ou novas ferramentas analíticas. A figura a seguir apresenta as principais competências desse profissional:

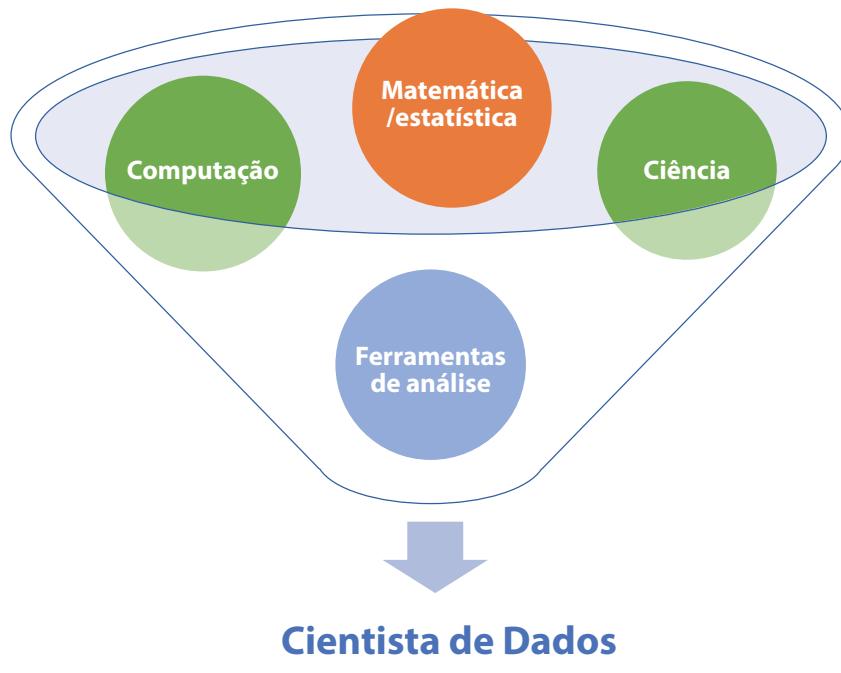


Figura 6 - Principais competências de um cientista de dados / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta a mesma analogia do funil presente na imagem anterior, ao concentrar quatro círculos que são combinados à expressão “Cientista de dados”, que está presente no fim do funil. Os quatro círculos são: computação, ciência (ambos adicionados em um círculo de cor verde), matemática/estatística e negócio.

A seguir, são apresentados mais detalhes em relação à cada capacidade multidisciplinar de um cientista de dados:

- **Computação:** tanto o analista de dados quanto o analista de BI e de negócio devem ter uma competência voltada ao banco de dados. No caso do cientista de dados, além de conhecer o banco de dados, é necessário que o profissional entenda programação, conheça os ambientes de execução e tenha noções de integração de plataformas e de ambientes de *Big Data*, por exemplo.
- **Matemática/estatística:** a estatística é uma importante ferramenta para o cientista de dados, visto que o auxilia principalmente na construção de uma amostragem significativa e na realização de análises de domínios para a construção de modelos estatísticos. A matemática, sobretudo a álgebra linear e geometria analítica, apoiam o processo de construção de algoritmos utilizando a proximidade vetorial ou a recomendação, por exemplo.
- **Ciência:** o cientista de dados guia os seus estudos a partir de métodos científicos, ao identificar do problema, gerar hipóteses, selecionar variáveis e validar as hipóteses. Além do mais, estrutura os seus estudos a partir de etapas metodológicas previamente definidas e documentadas.
- **Negócio:** o cientista de dados precisa conhecer o negócio para entender o comportamento das variáveis e projetar os seus modelos estatísticos e computacionais aderentes às questões da organização. Esse conhecimento é muito importante para os estudos avançados e o aprofundamento das regras de negócio pode acontecer em conjunto com o analista de BI e de negócio.

Existem muitas vagas para os cientistas de dados e esse fato se dá por dois motivos: primeiro, pela falta de mão de obra qualificada para atuar nos projetos analíticos e, depois, em detrimento de que os cientistas de dados têm competência para atuar em várias frentes, o que facilita para uma empresa que deseja iniciar uma área de dados, tendo em vista que esse profissional poderá apoiar grande parte das análises. Na prática, foi o que acompanhamos na empresa de Anderson: Joaquim foi contratado como cientista de dados e estruturou vários processos e ambientes em conjunto com a equipe de TI de Lara, o que fez com que eles também evoluíssem em suas carreiras.

Dependendo do nível de maturidade analítica da organização, talvez, seja necessária a contratação de profissionais dotados de conhecimentos mais específicos sobre modelagem estatística ou computacional. Nesse caso, esses profissionais trabalham em conjunto com os cientistas de dados e são chamados de **estatísticos** (especialistas em modelagem) e **engenheiros de aprendizagem de máquina** (*machine learning*).

A profissão de **estatístico** é bastante conhecida e tradicional. Em uma área de dados, esse profissional pode se concentrar na confecção de modelos estatísticos para a geração de *scores* e classificações. Outro aspecto que o estatístico pode apoiar é a construção e a validação de amostras que sejam significativas para o universo de dados em questão, além de auxiliar na confecção de instrumentos para pesquisa e análises de variáveis.

Em organizações, tais como bancos, e em empresas de varejo, que trabalham com a concessão de crédito, esse profissional é bastante comum, pois é o responsável por desenvolver os modelos que classificarão a possibilidade de conceder, ou não, o crédito a um cliente, a partir de um *score*. Depois da inserção do cliente na base, alicerçado no seu histórico de relacionamento com a empresa, é possível atribuir um *score* do quanto bom pagador é esse cliente, permitindo, assim, a redução ou o aumento de limite. Geralmente, esse tipo de modelagem é feito por estatísticos ou por cursos que estudam a estatística com profundidade, tais como Matemática, Física e Economia. O estatístico convive com o cientista de dados, pois, em muitos casos, é o cientista de dados que transpõe o modelo estatístico em questão para um elemento que possa ser implantado e utilizado.

No caso do **engenheiro de aprendizagem de máquina** (*machine learning*), a sua participação em uma área de dados é muito similar à do estatístico. No entanto, o seu foco recai na construção de modelos computacionais normalmente desenvolvidos a partir de técnicas e métodos de inteligência artificial, que podem ser incorporados em plataformas, produtos e aplicativos. Dessa forma, esses profissionais podem apoiar o trabalho dos cientistas de dados, ao apanhar os modelos inicialmente desenvolvidos em forma de protótipos, a fim de desenvolvê-los em um ambiente de produção. Esses engenheiros têm mais competência em computação que os cientistas de dados, o que proporciona uma interação natural com equipes da área de tecnologia.

Da mesma forma que os estatísticos, esses engenheiros podem ser contratados em empresas que não possuem uma área de dados constituída. Além disso, de modo geral, trabalham em áreas de produto ou de tecnologia justamente por trazerem “inteligência” aos sistemas computacionais desenvolvidos. Em organizações que já tem uma área de dados, o cientista de dados é um grande orquestrador do trabalho desse profissional, ao apresentar requisitos dos projetos, demandar otimizações em modelos computacionais e apoiar a implantação de modelos desenvolvidos.

## Colocando a área de dados em funcionamento

Depois de Joaquim finalizar o mapeamento das competências necessárias para a área e conhecer os cargos que a constituem, foi iniciado o processo de migração dos profissionais da área de tecnologia para a área de dados. O próximo passo seria mapear o ambiente de dados e de *Big Data* que já existe na organização. Não só, mas era preciso pensar nos principais processos e atividades que a nova área ofertará para as demais, já que, da mesma forma que a área de tecnologia é um “meio” de apoiar as áreas de negócio, no contexto da empresa de Anderson, a área de dados também é.

Relembremos os sistemas e os recursos que compõem a infraestrutura de dados disponível na organização:

- **Sistema ERP:** a solução de ERP é a porta de entrada dos dados operacionais e transacionais da empresa. Nele, são controlados os principais processos das áreas de negócio. Após uma evolução da ferramenta junto ao fornecedor, já há uma integração completa com a plataforma on-line da loja, o que garante que todos os pedidos e serviços sejam registrados e geridos em apenas um local. Esse sistema é de responsabilidade da área de tecnologia, mas a sua base de dados é replicada pela área de dados para dentro do *Data Lake*.
- **CRM:** a solução de CRM se tornou uma área de negócio. Nela, são combinados processos, uma ferramenta de CRM e pessoas. A ferramenta de CRM é mantida pela área de tecnologia, mas os dados de sua base de dados são replicados no *Data Lake* pela área de dados.

- **Solução de Business Intelligence:** a solução de *Business Intelligence* passou a ser de inteira responsabilidade da área de dados, desde a gestão do contrato da ferramenta até a licença do banco de dados utilizado para armazenar a estrutura de *Data Warehouse*. Nesse caso, toda a gestão do ambiente e os processos de produção são de responsabilidade da área de dados. As áreas de negócio poderão desenvolver os seus *dashboards* a partir das bases já modeladas dentro do *Data Warehouse*.
- **Ambiente de Big Data:** todo o ambiente de *Big Data* é de responsabilidade da área de Joaquim, desde a gestão do ambiente em nuvem utilizado até o monitoramento e a manutenção do ambiente de instâncias (máquina ou serviços instanciados na nuvem). Além disso, abrange o desenvolvimento e o acompanhamento de todos os sistemas de carga e dos sistemas para consulta e utilização. Outro aspecto que também deve ser acompanhado e mantido são as ferramentas que fazem a coleta dos dados de redes sociais e de sites da Internet.
- **Data Lake:** por se tratar de um repositório de dados e de informações analíticas, o *Data Lake* é de total responsabilidade da área de dados. Além do mais, deve garantir a execução dos processos e das políticas de governança de dados em todas as suas camadas. Assim, exige a aplicação de rotinas de monitoramento e processos de *data quality* (qualidade dos dados), a fim de garantir que os dados disponíveis nas camadas de acesso e de indicadores (*Data Warehouse*) estejam validados e governados.

A partir da releitura dos ambientes de dados da organização, é possível verificar quais são as principais atividades e tarefas que a área de dados deve exercer. Desse modo, Joaquim construiu uma lista dos principais objetivos da área de dados levando em consideração o estado atual da organização. A lista dos **objetivos da área** é composta pelos seguintes itens:

- Gerenciar o ciclo de vida dos dados analíticos.
- Gerenciar as cargas de dados internos e externos.
- Disponibilizar as bases analíticas para consumo da organização.
- Disponibilizar informações inerentes ao negócio.
- Desenvolver estudos baseados em dados.

- Difundir a cultura de dados na companhia.
- Melhorar o processo de tomada de decisão e descoberta de *insights*.
- Direcionar os colaboradores às melhores práticas de consumo e visualização dos dados.

Entendendo o momento atual da organização Joaquim optou por trabalhar com quatro cargos: analista de BI/negócios, analista de dados, cientista de dados e engenheiro de dados. Alguns cientistas de dados teriam focos estatísticos, enquanto outros teriam focos computacionais. Na sequência, Joaquim levantou as principais atribuições de cada cargo em forma de listas:

**Principais atribuições dos analistas de BI e de negócio:**

- Levantar as necessidades de negócio das áreas e apresentar soluções baseadas em *dashboards* e indicadores.
- Monitorar a disponibilidade dos *dashboards* das áreas de negócio.
- Garantir a aplicação das políticas da governança de dados para os indicadores organizacionais.
- Chancelar os selos de qualidade (maturidade) nos recursos passíveis de governança.
- Construir os *dashboards* corporativos ou de áreas de negócio (via solicitação de projeto).
- Propor melhorias e acompanhar os *dashboards* das áreas de negócio com base em boas práticas.
- Fomentar a cultura de dados na organização.

**Principais atribuições dos analistas de dados na área de dados:**

- Analisar *dashboards* e relatórios, a fim de identificar tendências e *insights*, e comunicar as áreas de negócio.
- Explorar as bases de dados para a identificação de padrões e para apoiar os estudos exploratórios a partir das demandas da organização.
- Acompanhar campanhas e pesquisas com clientes a fim de garantir uma análise aprofundada sobre o domínio em questão.
- Interagir com os analistas de BI/negócio para entender as demandas.
- Garantir que pesquisas e campanhas sejam executadas de maneira correta.
- Monitorar e acompanhar os indicadores corporativos para garantir que sempre sejam entregues de maneira confiável.

### Principais atribuições dos **cientistas de dados (com foco computacional)**:

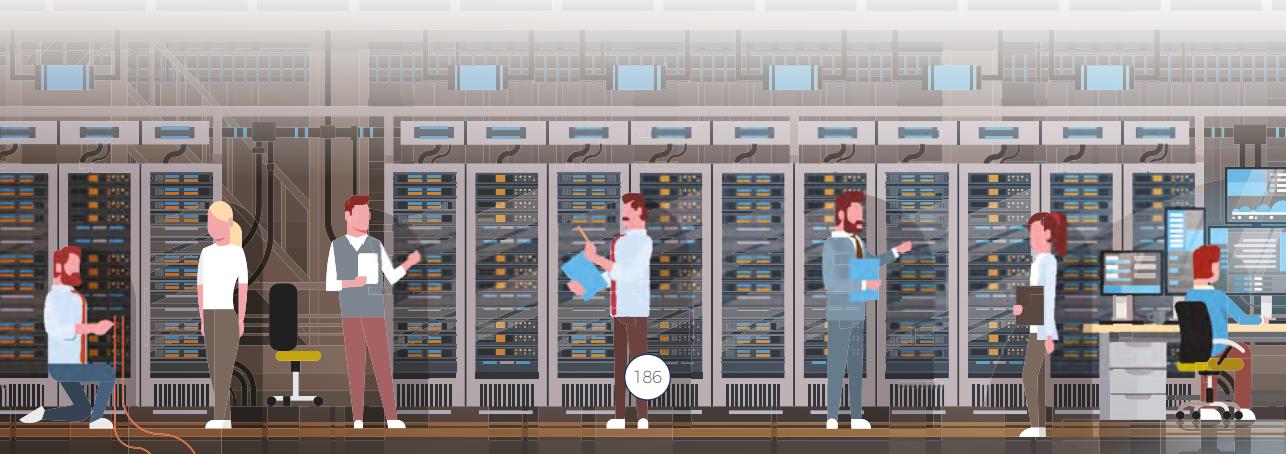
- Implementar algoritmos para a mineração de dados estruturados e não estruturados.
- Desenvolver componentes e serviços de apoio a tomada de decisão baseados em Inteligência Artificial.
- Automatizar processos computacionais com fins analíticos (gerador de amostras estratificadas, por exemplo).
- Monitorar a disponibilidade dos serviços desenvolvidos.
- Pesquisar (e desenvolver – P&D) soluções para problemas de natureza analítica das áreas de negócio.
- Auxiliar na implementação de códigos e garantir a aplicação de boas práticas e versionamento desses recursos.

### Principais atribuições dos **cientistas de dados (com foco estatístico)**:

- Desenvolver modelos estatísticos para apoiar as necessidades das áreas de negócio.
- Monitorar a performance dos modelos que estão em produção.
- Dar suporte estatístico a estudos, pesquisas e testes.
- Avaliar modelos estatísticos de terceiros.
- Identificar oportunidades para a criação de modelos que apoiem as demandas e as necessidades das áreas de negócio.
- Desenhar experimentos para projetos estatísticos com embasamento científico.

### Principais atribuições dos **engenheiros de dados** na área:

- Manter a integridade e a consistência dos dados nos bancos de dados analíticos.



- Agregar novas tecnologias e conceitos para garantir que o processo de armazenamento e recuperação dos dados sejam sempre garantidos com eficiência.
- Monitorar e zelar pela infraestrutura disponibilizada.
- Comunicar e recomendar as mudanças necessárias nessas estruturas de dados e na infraestrutura de suporte baseada no monitoramento.
- Processos de cargas (ETL).
- Disseminar conhecimento técnico e boas práticas relacionadas ao uso e à construção de banco de dados.

Ao descrever as principais atribuições dos integrantes da área, é necessário entender como se dará o fluxo de solicitações para a área de dados. Nesse contexto, dois tipos de solicitações são trabalhados:

- **Sustentação (via chamado):** as solicitações de sustentação são aquelas que têm relação com os elementos que já estão em produção, tais como bases de dados, *dashboards*, novas cargas e correções de problemas. As solicitações são feitas a partir do sistema de controle de chamados utilizado na organização (muito usado pelas áreas de tecnologia).
- **Projeto:** todo recurso a ser desenvolvido deve ser demandado como um projeto. Essa demanda deve ser apresentada em um fórum com os gestores da área de dados, a fim de obter prioridade, ou não, no processo de desenvolvimento. Assim que o projeto ganha prioridade, é submetido um analista de BI e de negócio, que analisa a demanda, faz o mapeamento do domínio de aplicação e o apresenta ao gestor da área de dados, que realiza a alocação da equipe e inicia o processo de desenvolvimento.



A figura a seguir apresenta um fluxo de trabalho entre os envolvidos na área de dados:

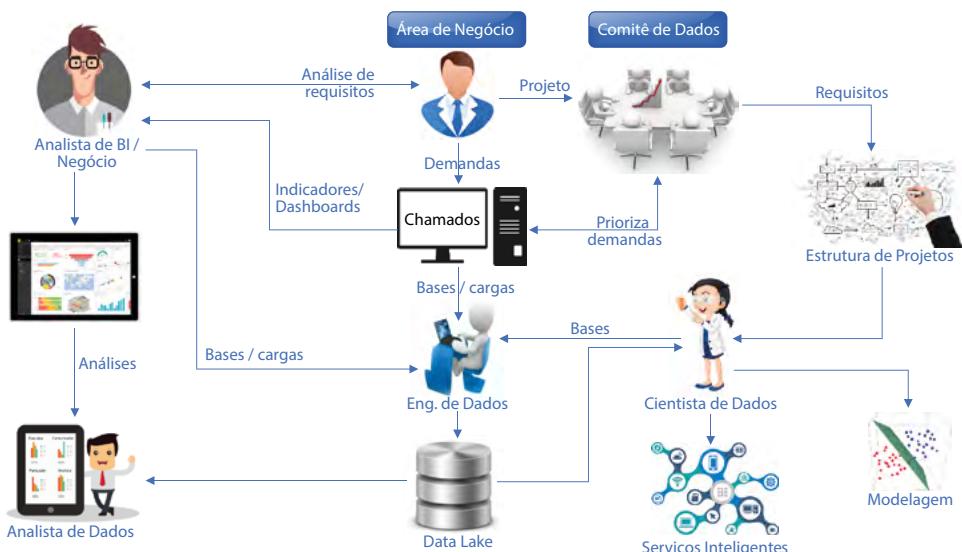


Figura 7 - Fluxo de atividades da área de dados / Fonte: o autor.

**Descrição da Imagem:** a figura apresenta os atores envolvidos nos fluxos de atividades da área de dados. Inicialmente, há o demandante de uma área de negócio, que, caso tenha uma solicitação de sustentação, realiza a abertura de um chamado. Por outro lado, caso tenha a solicitação de um novo projeto, a transfere para o comitê de dados. O analista de BI e de negócios faz a análise de requisitos e, caso eles estejam relacionados aos indicadores, são entregues em forma de dashboards. Os analistas de dados fazem análises a partir dos dashboards e do Data Lake. O comitê de dados prioriza o desenvolvimento dos projetos e dos chamados abertos pelas áreas de negócio. Os engenheiros de dados controlam as cargas e disponibilizam as bases analíticas para os demais profissionais da área. Os cientistas de dados analisam as demandas de projetos e desenvolvem serviços inteligentes, estudos ou modelos estatísticos a partir dos dados do Data Lake.

A tendência é a de que, quanto mais a cultura de dados for disseminada dentro da organização, maior será a demanda por recursos analíticos. Portanto, a fila para se atender às solicitações e aos novos projetos será muito grande, o que evidencia a importância do comitê de dados. O **comitê de dados** é um encontro que pode ser semanal (a sugestão é que se tenha, pelo menos, um por mês) e conta com a presença de representantes de todas as áreas de negócio e dos gestores que possam negociar a priorização das demandas da área.

O comitê de dados também pode ser utilizado para apresentar boas práticas no uso de dados, ministrar treinamentos de ferramentas, divulgar novas bases de dados e indicadores, validar novas políticas de governança de dados, apresentar e chancelar mudanças nos selos de maturidade das bases de dados ou dos recursos analíticos, além de ser um ponto de encontro para a divulgação da cultura de dados, tentando tornar a organização cada vez mais dirigida por dados.

A governança de dados é um dos importantes pilares de uma organização dirigida por dados e é uma das grandes responsabilidades da área de dados. A GD é a grande orquestradora das políticas é a responsável por conduzir o seu processo de criação, manutenção e monitoramento. Todas as áreas devem aplicar e respeitar as políticas e a área de dados fornece o apoio para a melhor execução das tarefas, além de expressar os processos de acompanhamento e monitoria, o que garante que todo o ambiente de dados esteja devidamente disponível e governado.

Joaquim finalizou o processo de desenho da área de dados. Ele já tinha definido os principais objetivos da área, quais eram os cargos e as funções, delimitado o fluxo de trabalho e expresso todo o processo de priorização. Além disso, fez a montagem de sua equipe, ao transferir alguns profissionais da área de TI e ao contratar novos profissionais. De acordo com o combinado, assumiu a solução de BI, o ambiente de *Big Data* e todo o controle e gestão do *Data Lake*, além de ofertar novos serviços e possibilidades de novos projetos para as áreas de negócio.

## Lei Geral de Proteção de Dados (LGPD)

Em meados de 2020, entrou em vigor a Lei nº 13.709/2018, a chamada “Lei Geral de Proteção de Dados” (LGPD). Ela foi construída à luz da chamada *General Data Protection Regulation* (GDPR), que é a regulamentação de dados desenvolvida na União Europeia. A proposta da LGPD é a de dar mais controle e transparência em relação ao uso de dados pessoais ao seu dono, deixando claro quais dados serão armazenados, que tipo de ação será feita a partir dos dados e dando a possibilidade do titular (dono dos dados) aceitar, ou não, que a organização faça essas ações com os seus dados pessoais.

Joaquim já vinha acompanhando a evolução da LGPD desde 2018, ano em que a lei foi publicada. Assim, realizou todas as adequações necessárias para garantir que a empresa estivesse em conformidade com a lei. A LGPD tem alguns agentes que são considerados na legislação. De acordo com Miranda (2019, p. 12), são eles:



**Titular:** Pessoa natural a quem se refere os dados pessoais que são o objeto de tratamento;

**Controlador:** Pessoa natural ou jurídica, de direito público ou privado, a quem competem as decisões referentes ao tratamento dos dados pessoais;

**Operador:** Pessoa natural ou jurídica, de direito público ou privado, que realiza Tratamento de dados pessoais em nome do controlador;

**Encarregado:** Pessoa natural, indicada pelo controlador, que atua como canal de comunicação entre o controlador e os titulares e a autoridade nacional.

A figura a seguir apresenta os atores envolvidos na aplicação da LGPD:



Figura 8 - Atores envolvidos na aplicação da LGPD / Fonte: A2C (2019).

**Descrição da Imagem:** na figura, é apresentado o termo “Titular” com a seguinte descrição: “Pessoa a quem se referem os dados pessoais que são objeto de tratamento”. Depois, é apresentada a expressão: “Agentes de tratamento” e são expostos dois tipos: “Controlador: é quem decide como serão tratados os dados pessoais” e “Operador: quem realiza o tratamento de dados em nome do controlador”. Também é apresentado o: “Encarregado da proteção de dados pessoais”, que possui a seguinte descrição: “Pessoa indicada pelo controlador para atuar como canal de comunicação entre o controlador, os titulares dos dados e a ANPD”. Por fim, é exposta a “ANPD - Autoridade Nacional de Proteção de Dados”, que carrega a seguinte descrição: “Órgão da administração pública responsável por zelar, implementar e fiscalizar o cumprimento da LGPD”.

A LGPD prevê que toda organização tenha um técnico responsável pelas transformações dos dados pessoais. Esse profissional é o responsável por garantir que a lei está sendo executada corretamente. Em caso de uma vistoria por parte da ANPD, é esse profissional que será chamado para acompanhar o processo e ser responsabilizado. O mercado tem se referido a esse profissional como o *Data Protection Officer* (DPO), que pode ser traduzido como o responsável pela proteção de dados.



#### EXPLORANDO IDEIAS

A missão do DPO é receber as reclamações e comunicações dos titulares dos dados, bem como prestar os devidos esclarecimentos e garantir que sejam tomadas as medidas necessárias ao cumprimento das regras e das boas práticas de proteção de dados. Deverá, ainda, receber comunicações da autoridade nacional de proteção de dados (ANPD) e adotar as providências eventualmente exigidas, bem como orientar os funcionários e os contratados da entidade a respeito das práticas a serem tomadas em relação à proteção de dados pessoais.

Fonte: OAB (2019, p. 79).

Ao entender toda a responsabilidade que se tem enquanto DPO, Joaquim decidiu que ele seria o próprio DPO da organização, o que faria que ele não se desligasse das questões relacionadas à LGPD e ao tratamento de dados pessoais. Os dados pessoais podem ser entendidos como qualquer dado que permita a identificação de uma pessoa. Sob a ótica da LGPD, existem quatro tipos de dados, os quais são apresentados por Miranda (2019, p. 12):



**Dados identificados:** são aqueles que você consegue saber quem é o titular, nome, identidade, CPF, etc.

**Dados identificáveis:** são dados que você não consegue diretamente saber quem é o titular, mas em contato com outras informações você consegue atingir seu objetivo: o número do cartão de crédito, o IP do computador, nome da empresa que a pessoa trabalha, etc.

**Dados sensíveis:** classificados pela nova legislação, requerem ainda mais cuidado na sua guarda, acesso e manuseio, pois estão relacionados à origem étnica ou racial, crenças religiosas, filiação sindical, direcionamento político, orientação sexual e especialmente, informações relativas à saúde, genética ou biométrica.

**Dados anonimizados:** são aqueles onde não é possível identificar a pessoa como por exemplo uma pesquisa do IBGE.

Na prática, Joaquim tomou as seguintes ações para se adequar à LGPD e ter o seu ambiente de *Big Data*:

- Foi enviado um comunicado solicitando o “de acordo” por parte dos clientes já cadastrados nas bases de dados, deixando claro o que estava sendo armazenado e para que seriam utilizados esses dados. Caso o cliente não aceitasse ou não respondesse, os seus dados passariam por um processo de anonimização.
- Foram desenvolvidas rotinas para mapear todos os dados de uma pessoa da base, a fim de dar a possibilidade de fornecer todos os seus dados pessoais e sensíveis armazenados. Também foi desenvolvida uma rotina em que são excluídos e anonimizados os dados, se solicitado pelo titular.
- Foi solicitada à área de tecnologia que todos os sistemas que tenham cadastros já apresentem as informações sobre os dados e solicitem o “de acordo” do titular. Deve ser evidenciado o que será armazenado e de que forma será utilizado, dando total liberdade e transparência sobre o tratamento e o uso dos seus dados.

Diante dos parâmetros exigidos pela LGPD já direcionados, a área de dados estava em plena operação. Joaquim, em conjunto com o seu time, conseguiu implantar novos modelos para as áreas de CRM, compras, vendas e parcerias. Além do mais, apoiou a construção de uma nova área de crédito, o que permite que os clientes utilizem o crediário próprio da loja.

A empresa cresceu muito e, hoje, conta com, pelo menos, cinco lojas físicas em cada estado do Brasil, expandindo para o Uruguai e Argentina. Os diretores entendem que essa evolução só foi possível por intermédio da realização de estudos e análises, e dos indicadores disponibilizados, os quais sempre permitiram a obtenção de uma visão correta sobre o estado atual da empresa, além de mostrarem projeções e previsões de comportamento de mercados e indicadores.

Era uma sexta-feira, às oito horas da manhã. Joaquim estava sentado ao redor de sua mesa. Tomava o seu café e conferia os e-mails que havia recebido nas últimas horas. Nesse momento, Lara entrou em sua sala e pediu para que

ele a acompanhasse, porque Anderson queria conversar. Joaquim chegou até a sala da presidência e lá estavam todos os diretores e conselheiros reunidos. Anderson o cumprimentou e fez uma pergunta: Joaquim, hoje, quem é o seu braço direito? Quem é a pessoa que conseguiria dar continuidade a área de dados se, porventura, você não pudesse mais ocupar esse cargo?

Joaquim ficou pálido e muitos pensamentos vieram até a sua mente. Contudo, ele se concentrou e respondeu: Carolina é, hoje, a pessoa mais preparada! Anderson agradeceu a informação e pediu para que ele fosse buscá-la. No entanto, antes, gostaria de falar algo. Assim, caminhou até Joaquim e disse que Lara gostaria de lhe dar uma notícia. Lara olhou para Joaquim e disse: Joaquim, Carolina será a nova gerente de dados da empresa. Preciso que você retire as suas coisas da mesa, para que ela possa ocupar o posto. Por favor, pegue as suas coisas e as coloque nessa mesa, ao lado da minha, porque, a partir de hoje, você será o nosso diretor de dados! Bem-vindo à diretoria!

## Diretoria de dados

Joaquim, já posicionado em sua nova mesa, próxima dos demais diretores, entendeu que teria um orçamento para a sua diretoria e seria necessário ampliar ainda mais a capacidade de atendimento das demandas das áreas de negócio. Entretanto, a sua primeira missão era entender o que exatamente um diretor de dados (ou CDO - *Chief Data Officer*) faz em uma organização.

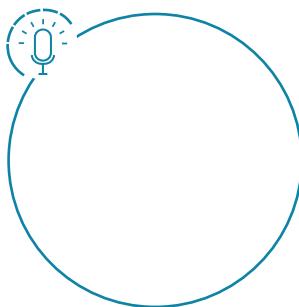
O CDO é o diretor que está focado nos processos e nas áreas relacionadas à infraestrutura de dados. Ele está diretamente ligado com o CAO (*Chief Analytics Officer*) e, no caso da empresa em que Joaquim está atuando, ele desempenhará o papel das duas diretorias, o que é comum de se encontrar em organizações que ainda têm pontos de evolução na cultura e nos processos relacionados aos dados. O entendimento é o de que o CDO tem uma perspectiva mais computacional e atua muito próximo do diretor de tecnologia (CTO - *Chief Technology Officer* ou CIO - *Chief Information Officer*), a fim de cuidar das áreas que trabalham com a infraestrutura de dados. Já no caso do CAO, ele está mais ligado ao uso dos dados para a tomada de decisão. Em organizações menores, é comum encontrar apenas um desses dois profissionais agrupando as duas frentes (ANDERSON, 2015; AIKEN; GORMAN, 2014).

Agora, é o momento de Joaquim olhar para a sua área, que cresceu cinco vezes desde que foi montada e quebrada em outras áreas. Vejamos:

- Área de **infraestrutura de dados**: está focada nos processos relacionados ao *Data Lake* da organização, garantindo que a governança de dados sempre seja aplicada.
- Área de **Business Intelligence e Governança de Dados**: é a área que está focada na construção dos indicadores e dos painéis corporativos (utilizado pela presidência). Além do mais, gerencia todos os processos de criação e manutenção das políticas de governança de dados.
- Área de **ciência de dados**: focada na construção de modelos e estudos avançados e no desenvolvimento de componentes inteligentes para uso das áreas de negócio.
- Área de **projetos e pesquisa**: área responsável por fazer pesquisas de novos produtos e serviços de dados. Além disso, estrutura pesquisas com usuários e especifica novos projetos da diretoria.

Carolina teve a liberdade de escolher a área que gostaria de gerenciar e foram chamados outros colaboradores para liderar as três novas áreas. Joaquim estava completamente realizado, pois teve a oportunidade de acompanhar toda a evolução analítica da empresa de Anderson e pode aplicar os seus conhecimentos. Era desafiado diariamente e chegou até a maior cadeira da organização no contexto de dados e apoio a tomada de decisão.

Agora, tinha o desafio de fazer a empresa crescer ainda mais e disseminar cada vez mais a cultura de dados entre todos os colaboradores. Qual será o próximo desafio que a vida está preparando para Joaquim? Bom, esse já é assunto para um novo livro!



Chegamos ao final de nossa aventura! Como será que é o dia a dia de uma área de dados? Será que é fácil trabalhar com uma equipe multidisciplinar e com cargos de natureza tão distintas? Será que é possível ter um comitê de dados que funcione na prática? Se você ficou interessado em saber as respostas dessas perguntas, não deixe de escutar o nosso podcast!

A criação de uma área de dados (ou de ciência de dados) representa uma importante evolução para uma organização que deseja ser centrada no cliente e/ou dirigida por dados. Entende-se que essa área deve ter autonomia para desenvolver os seus projetos, respondendo diretamente à presidência ou a uma diretoria específica. Além do mais, deve funcionar desacoplada da área de tecnologia e deve integrar os profissionais necessários para fazer com que os processos sejam executados e as ferramentas sejam disponibilizadas sem a necessidade de intervenção dos profissionais da tecnologia.

Existem especializações de antigos cargos para atender às demandas dessa área, o que garante que os profissionais estejam ligados aos novos processos e aos focos que a área demandará. Também existem novos cargos que demandam por profissionais com competências e habilidades multidisciplinares. Essa área deve ser gerida por um profissional que tenha conhecimento em relação aos principais processos e atividades, para que seja possível avaliar e guiar o desenvolvimento de propostas de soluções, cujo problema esteja ligado aos dados.

A LGPD apresenta um conjunto de processos e responsabilidades para toda empresa que faz uso de dados em seu marketing, relacionamento ou serviço de inteligência. Diante disso, uma área de dados pode apoiar a operacionalização e a adequação da organização à lei. Joaquim conseguiu facilmente adequar a empresa às questões da legislação, pelo fato de ter bases governadas e processos muito bem construídos. Isso facilita muito toda a adequação e permite que a empresa faça o uso correto e consciente dos dados dos seus clientes.

A criação de uma diretoria de dados e *analytics* é um caminho natural para toda empresa que deseja ser dirigida por dados. Para a criação de uma diretoria dessa natureza, é necessário que a organização já tenha, pelo menos, uma área de dados com processos e *roadmap* de evolução da maturidade analítica já desenvolvidos. O CDO e CAO representam os novos cargos executivos nas organizações e são os diretores responsáveis por garantir que sejam desenvolvidos recursos analíticos, demonstrando o seu impacto nos negócios e deixando clara a relação entre custo e benefício. Não só, mas também guiam as outras áreas para que sejam cada vez orientadas a dados e para que tenham maior assertividade em suas decisões.



1. A área de dados representa uma importante evolução para as organizações que pretendem ser dirigidas por dados.

Assinale a alternativa correta:

- A área de dados pode ser 100% distribuída, mesmo não tendo processos formalizados.
  - A governança de dados é de responsabilidade da área de tecnologia.
  - As áreas de dados e de tecnologia representam a mesma coisa. Elas apenas carregam o nome diferente.
  - A área de dados pode ser distribuída, mas, para isso, é exigida certa maturidade analítica.
  - A governança de dados deve ser respeitada pela área de dados e orquestrada pela área de tecnologia.
2. Existem muitos cargos que podem ser necessários para uma área de dados. Em relação à principal atribuição de um engenheiro de dados, assinale a alternativa correta:
    - Levanta os requisitos de negócio para a construção de projetos da área de dados.
    - Monitora os recursos ligados ao **Data Lake** e desenvolve os **dashboards** das soluções de BI.
    - Desenvolve estudos avançados e monitora o uso das bases de dados.
    - Gerencia o ambiente de **Big Data** e desenvolve modelos estatísticos.
    - Zela pela infraestrutura de dados e atua na construção de ETLs e recursos de um ambiente de **Big Data**.
  3. A Lei Geral de Proteção de Dados entrou em vigor em meados de 2020. Sobre a referida lei, assinale a alternativa correta:
    - A LGPD permite que as organizações tenham mais liberdade para processar os dados pessoais dos seus clientes.
    - A LGPD objetiva dar mais transparência e controle ao titular dos dados.
    - A LGPD só pode ser aplicada em empresas ligadas ao Estado ou que sejam públicas.
    - A LGPD está focada apenas na inibição de **Fake News**.
    - A LGPD está focada em não permitir a emissão de **Fake News** a partir de ferramentas colaborativas.

# CONFIRA SUAS RESPOSTAS



1. D.

O processo de distribuição dos agentes de dados pelas áreas de negócio tem relação direta com o grau de evolução analítica que da organização. Em outras palavras, quanto maior for a maturidade analítica, maior pode ser a distribuição das tarefas e recursos para as áreas de negócio.

2. E.

O engenheiro de dados deve zelar pela infraestrutura de dados e atuar na construção de ETLs e recursos de um ambiente de *Big Data*.

3. B.

A LGPD objetiva empoderar o titular dos dados, ao evidenciar quando os seus dados serão armazenados o que se pretende fazer com eles.

# REFERÊNCIAS



- A2C. **LGPD no marketing digital:** como adequar a sua estratégia. São Paulo: Blog Digital, 2019.
- AIKEN, P.; GORMAN, M. **A função do chief data officer:** reorganizando os cargos executivos para alavancar o seu mais valioso ativo. Rio de Janeiro: Campus 2014.
- ANDERSON, C. **Creating a data-driven organization:** practical advice from the trenches. Sebastopol: O'Reilly Media, 2015.
- MIRANDA, M. G. **Lei Geral de Proteção de Dados - LGPD.** [S. l.: s. n.], 2019.
- OAB. **O que estão fazendo com os meus dados?** A importância da Lei Geral de Proteção de Dados. Recife: OAB Pernambuco, 2019.



