INTELIGÊNCIA ARTIFICIAL

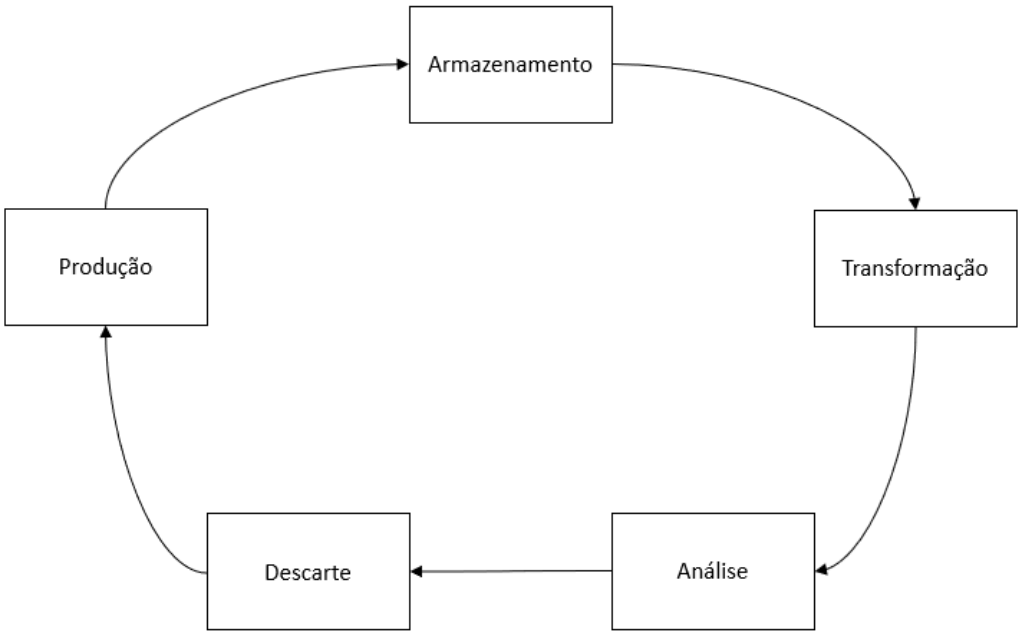
Prof. Luciano Frontino de Medeiros

TEMA 1 – INTRODUÇÃO À CIÊNCIA DE DADOS

Com o aumento substancial do uso de redes sociais, o comércio eletrônico em rápida expansão e a ascensão da internet, a web está criando um tsunami de dados estruturados e não estruturados, gerando bilhões em gastos em tecnologia da informação (TI). Evidências recentes também indicam um declínio no uso de plataformas de inteligência de negócios-padrão, na medida em que as empresas são forçadas a considerar uma massa de dados não estruturados com valor incerto no mundo real (Battiti; Brunato, 2017).

Ainda que a ciência de dados já venha sendo mencionada desde 1960, pode-se afirmar que ela é uma ciência relativamente nova. Equivocadamente, é associada apenas com os processos relacionados à análise com estatística, aprendizado de máquina ou ainda filtragem para produzir informação ou conhecimento (Amaral, 2016). No entanto, é necessário compreender que o dado possui um ciclo de vida (Figura 1) e que, em função disso, deve-se constatar a abrangência do que vem a significar a real ciência de dados.

Figura 1 – Ciclo de vida do dado



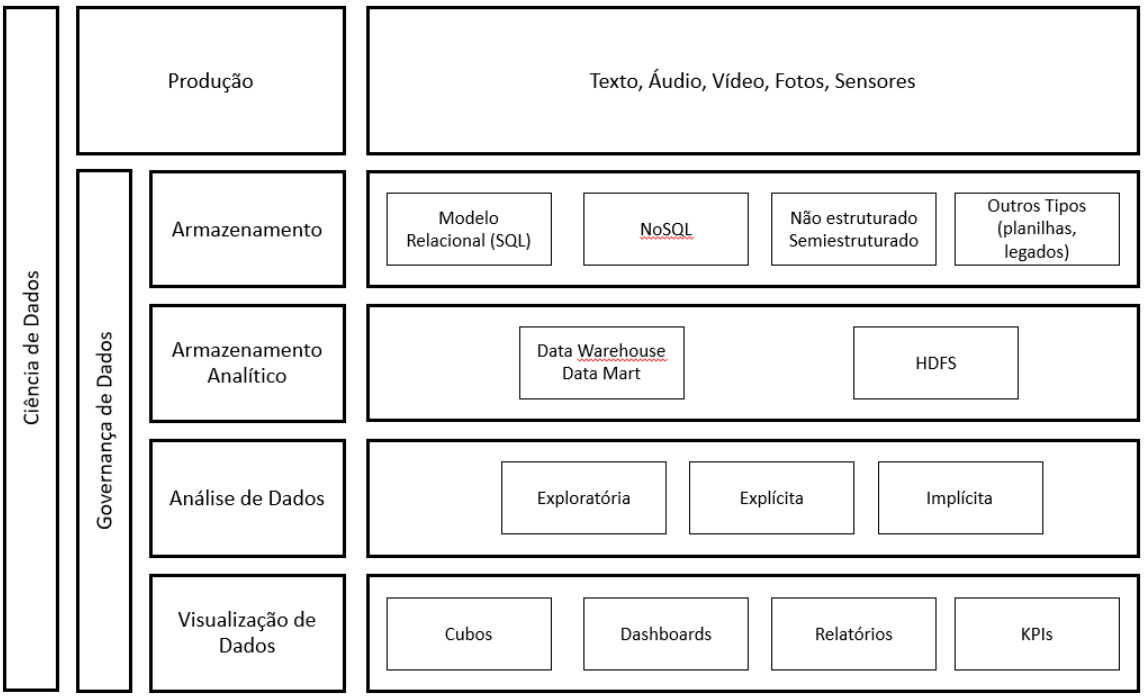
Fonte: Adaptado de Amaral, 2016, p. 6.

Em sua forma digital, o dado é produzido por algum dispositivo, como um computador ou um celular, enquanto se digita um texto; ou com base no sensoriamento de um veículo ou no uso de uma câmera digital para produzir fotos

e vídeos. A partir daí, o dado precisa ser armazenado em algum tipo de mídia. Existindo em meios de armazenamento, o dado pode passar por processos de transformação, tal como no modelo intitulado *extração, transformação e carga* (ETL), para a construção de *data warehouses*. Segue-se então a etapa de análise dos dados, com a execução de qualquer operação para se extrair informação e conhecimento, utilizando uma simples consulta em Structured Query Language (SQL) ou por meio de um processo de classificação de redes neurais artificiais. Ao final, o dado precisa passar por um processo de descarte, de acordo com a necessidade de sua retenção por razões organizacionais ou atendendo ao especificado em legislação pertinente (Amaral, 2016, p. 5).

Dessa forma, pode-se definir a ciência de dados como o conjunto de processos e tecnologias que estudam os dados durante todas as fases do seu ciclo de vida. A Figura 2 mostra a ciência de dados dividida nas fases de produção e governança, permitindo identificar uma série de elementos relacionados ao ciclo de vida e os exemplos de tecnologias que lidam com cada uma das fases.

Figura 2 – *Framework* da ciência de dados



Fonte: Adaptado de Amaral, 2016, p. 7.

Na fase da análise de dados, a **análise exploratória** se incumbe do estudo inicial de um conjunto de dados, na busca da identificação de categorias nas quais os dados se encaixem e que possam empregar técnicas quantitativas ou ainda visuais. Na **análise explícita**, a informação e o conhecimento estão disponíveis

de forma explícita nos dados, sendo necessária alguma operação de baixa complexidade para ressaltar a característica-alvo de análise e produzir a informação. Na **análise implícita**, a informação não está disponível de forma clara, devendo ser produzida com o uso de alguma técnica sofisticada (Amaral, 2016, p. 61).

Assim, a inteligência artificial (IA) fornece como subsídio para a ciência de dados uma série de tecnologias que tornam possíveis as análises implícitas, intimamente relacionadas com o aprendizado de máquina: classificação, agrupamentos (*clustering*) ou regras de associação. A Figura 3 mostra alguns algoritmos geralmente utilizados para cada atividade de aprendizado de máquina.

Quadro 1 – Atividades de aprendizado de máquina

Atividade	Tipos de Algoritmos	Algoritmos
Classificação	Bayes	NaiveBayes
	Rules	Party, DecisionTable
	Árvores de Decisão	Random Forest J48 ID3
	Redes Neurais Artificiais	Perceptron Multicamada (MLP) Máquina de Vetores de Suporte (SVM)
Agrupamentos	Por Densidade	DBSCAN
	Baseado em Protótipo	K-Means K-Medoids
Regras de Associação		Apriori FP Growuth

Fonte: Adaptado de Amaral, 2016, p. 99.

TEMA 2 – *BIG DATA*

Desde o início da computação pessoal, nos anos 1980, até a onipresença da web, atualmente, na vida cotidiana, os dados foram produzidos em quantidades cada vez maiores de fotos, vídeos, sons, textos, registros de todos os tipos. Desde a democratização da internet, há volumes impressionantes de dados criados diariamente por indivíduos, empresas e agora também objetos e máquinas conectados. O termo *big data*, traduzido literalmente por *dados grandes* ou *dados massivos*, indica essa eclosão dos dados. Também falamos de *massa de dados*, em analogia com a biomassa, um ecossistema complexo e de larga escala (Lemberger; Morel; Raffaelli, 2015, p. 3).

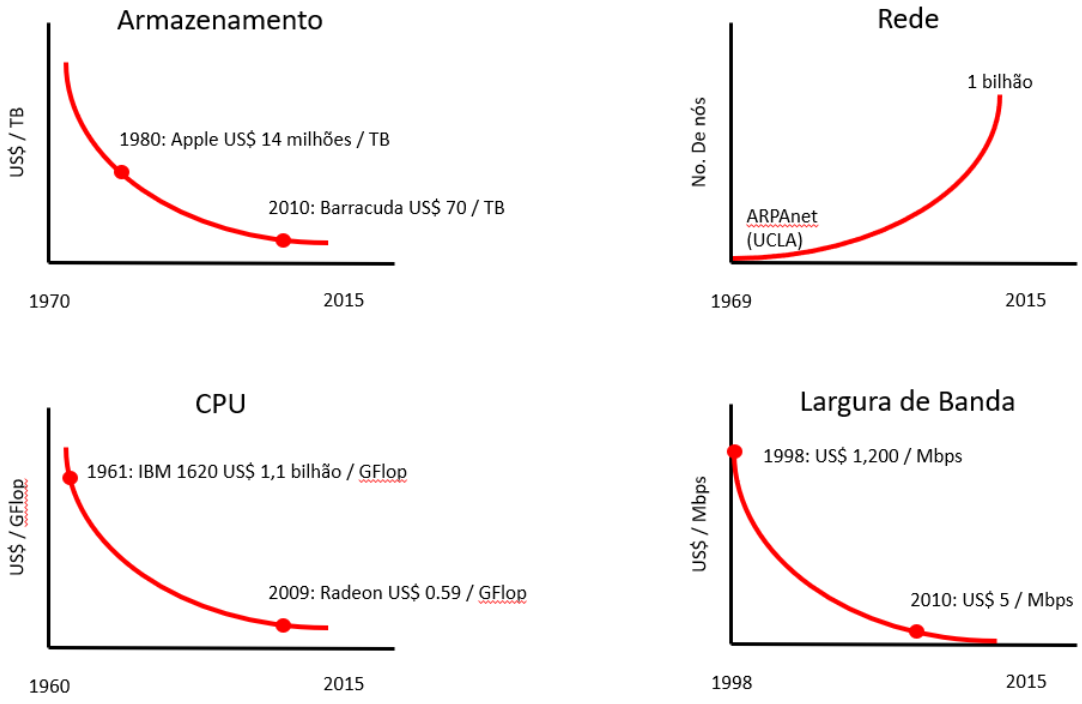
O conceito de *big data* está associado a grandes volumes de dados e sua definição envolve um conjunto de três a cinco vês: volume, velocidade, variedade e, ainda, veracidade e valor. Na sua acepção profunda, *big data* é o fenômeno em que dados são produzidos em vários formatos e armazenados em uma grande quantidade de diferentes dispositivos ou equipamentos. Como motor desse fenômeno, os insumos de tecnologia (processadores, memórias, locais de armazenamento) têm se tornado cada vez mais baratos. O baixo custo leva à disseminação de tais equipamentos, os quais produzem, nas mãos dos usuários, a quantidade massiva de dados (Amaral, 2016, p. 7).

Outro detalhe a respeito da geração massiva de dados é que o *big data*, além do próprio volume, se refere também à diversidade. Os dados são gerados nas mais variadas formas, línguas, formatos de arquivos e representação. Ele não se refere apenas a processos que permitem o aparecimento de grandes repositórios de dados, em servidores ou *clusters*, para serem posteriormente analisados. *Big data* se constitui numa mudança social, cultural, mesmo numa nova fase da revolução industrial (Amaral, 2016, p. 9).

Com relação ao aspecto tecnológico, o conceito-chave associado ao *big data* é o registro de qualquer fenômeno, natural ou não, e sua transformação em dados. Esses dados são reproduzidos ou analisados, imediatamente ou no futuro. Esse fenômeno é conhecido como *datafication*: o registro eletrônico de um fenômeno qualquer, como o movimento de um celular, o acionamento de um freio de um carro, uma foto ou gravações de câmeras de segurança. O *big data* permite que a miríade de eventos provocados pelos seres humanos seja armazenada e, consequentemente, reproduzida ou analisada (Amaral, 2016, p. 10).

Para se beneficiar desses colossais recursos de armazenamento, os gigantes da *web* tiveram que desenvolver novas tecnologias para suas próprias necessidades, particularmente em termos de paralelismo de processamento, operando em volumes de dados que totalizam várias centenas de *terabytes*. Um dos principais avanços nessa área veio do Google, que, para processar os dados recuperados pelos rastreadores de seu mecanismo de pesquisa e indexar toda a *web*, desenvolveu um modelo de design que automatiza a paralelização de uma grande classe de tratamentos. É o famoso modelo **MapReduce**. Muitos avanços na engenharia de software desenvolvidos nessa ocasião se espalharam posteriormente para a comunidade de código aberto, que oferecia sistemas equivalentes gratuitos. O sistema de processamento paralelo **Hadoop Apache** é o principal exemplo dessa transferência de tecnologia para o mundo do código aberto. E, diante dos novos requisitos de escalabilidade e disponibilidade, surgiu uma nova classe de sistema de gerenciamento de banco de dados não relacional. Esses sistemas são geralmente designados pelo acrônimo **NoSQL** (Lemberger; Morel; Raffaelli, 2015, p. 6).

Figura 3 – Evolução dos preços dos recursos de TI ao longo das décadas passadas



Fonte: Adaptado de Lemberger; Morel; Raffaelli, 2015, p. 6.

Outro fator determinante para as empresas com relação à adoção do *big data* se refere à vantagem competitiva. Indo de uma abordagem tradicional da análise dos dados para melhorar o que está relacionado diretamente ao negócio da empresa, o *big data* vai além e permitirá o uso do dado de forma a torná-la mais competitiva e eficiente, coletando dados e analisando também aqueles que não estão relacionados diretamente ao seu negócio. Internamente, a empresa melhorará o processo de seleção, contratando profissionais mais comprometidos e com o perfil exigido para o cargo, além de proporcionar produtividade para seus colaboradores. As linhas de produção tenderão a ser mais eficientes, com menos paradas e menor custo de produção. Externamente, a empresa será capaz de entender as necessidades dos clientes, atuar na prevenção de perdas por *recalls* ou comprometimento de imagem e tenderá a ter consumidores mais fiéis. Por outro lado, as empresas que não souberem usar *big data* irão desaparecer do mercado (Amaral, 2016, p. 11).

TEMA 3 – INTERNET DAS COISAS

Uma das razões para o aumento crescente de dados disponíveis para o *big data* é devida à realidade da internet das coisas (de *internet of things* – IoT). Quando se pensa em produzir dados, costuma-se lembrar do teclado de um computador pessoal, que é aceito como um padrão para entrada de dados. Entretanto, desde os primórdios da computação, outros meios de entrada de dados já estavam à disposição, como interruptores e cartões perfurados. Os dados, hoje em dia, podem ser gerados por uma série de dispositivos tais como mouses, telas *touch screen*, leitores de códigos de barras e QRCode, identificação por radiofrequência, mesas digitalizadoras, entre outros. Podem ser enquadrados ainda como dispositivos aqueles que não operam necessariamente conectados de alguma forma a um computador: câmeras de vídeo, máquinas fotográficas e dispositivos médicos portáteis (Amaral, 2016, p. 18).

A internet das coisas se refere à integração, por meio dos protocolos da internet, de todo ou qualquer dispositivo, bem como torná-los mais inteligentes, capazes de coletar e processar informações do ambiente ou das redes às quais estejam conectados. A implantação da internet das coisas está revolucionando a forma como as pessoas se relacionam com as coisas ao redor, em relação à segurança, energia, meio ambiente, trânsito, mobilidade ou logística (Oliveira, 2017, p. 16).

Assim como no *big data* a queda dos preços foi um fator determinante para a produção massiva de dados, microcontroladores e componentes eletrônicos mais baratos também influenciaram na proliferação dos dispositivos conectados à internet. O conceito de IoT não é novo. Com a popularização da internet, na década de 1990, já se vislumbravam novas formas de interligar equipamentos de uso diário à internet (Oliveira, 2017, p. 16).

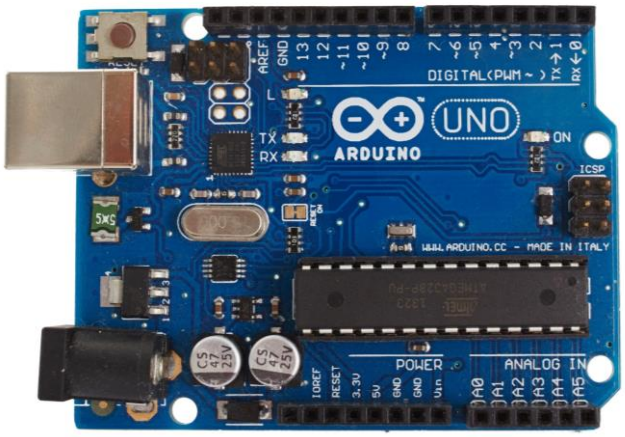
As tecnologias de comunicação e as redes de computadores se desenvolveram e popularizaram, desde o seu surgimento. A internet, por meio da família de protocolos TCP/IP, seguida pelas redes wi-fi, tornou possível a mobilidade e dispensou a fiação típica da conectividade de rede. Aliado a esse fato, as redes de telefonia celular 2G/3G/4G foram fundamentais para aumentar o rol de equipamentos a serem conectados. Dessa forma, a comunicação de dados se tornou possível, reduzindo o custo e o tempo de integração de soluções (Oliveira, 2017, p. 17).

Os telefones celulares são dispositivos que, a cada versão ou melhoria de tecnologia, se tornam mais inteligentes. A designação de *smartphones* tipifica a diversidade de sensores que são disponibilizados ao longo da sua evolução. O *smartphone*, além de ser telefone, pode incluir acelerômetro, *touch screen*, Global Positioning System (GPS), giroscópio e magnetômetro, além dos dispositivos voltados à comunicação: *bluetooth*, wi-fi, entre outros (Amaral, 2016, p. 21).

Os sensores não apenas coletam dados. Eles podem também acionar os efetuadores, responsáveis por algum movimento a ser feito no ambiente. A tendência do crescimento da IoT é que a tecnologia irá se disseminar, tornando as casas e os escritórios autônomos, repletos de sensores e efetuadores, que irão, por sua vez, abrir ou fechar toldos ou cortinas, ligar a iluminação, a irrigação de plantas e jardins, com comandos controlados por *smartphones* conectados a tais dispositivos de IoT (Amaral, 2016, p. 22).

Entretanto, pode-se afirmar que o sucesso dos dispositivos relacionados com IoT reside nos microcontroladores. Um microcontrolador é um tipo de processador, uma espécie de pequeno computador inserido em um único chip (Figura 4). Os microcontroladores possuem tudo o que havia nos primeiros computadores pessoais e ainda são dotados de outras tecnologias. Eles contêm um processador, memória RAM e flash para armazenamento, além de pinos de entrada e saída que ligam o controlador a outros componentes eletrônicos, bem como aos demais dispositivos externos (Monk, 2012, p. 6).

Figura 4 – Placa do Arduino Uno, contendo um microcontrolador Atmel Atmega 328P-PU

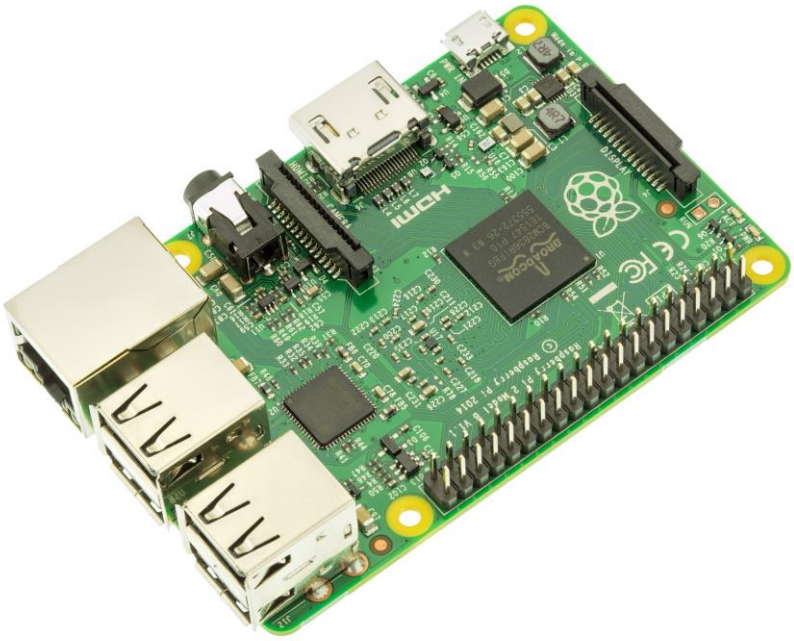


Crédito: Zossia/Shutterstock.

Os microcontroladores permitiram que placas de uso geral fossem popularizadas, tais como a plataforma Arduino. Um Arduino nada mais é do que uma placa com um microcontrolador adaptada com *plug* USB, que permite a conexão com um computador. O Arduino contém diversos terminais que permitem a conexão com dispositivos externos variados: diodos emissores de luz (LEDs), sensores, motores, diodos a laser, alto-falantes etc. Dessa forma, o Arduino consiste em uma plataforma de microcontrolador a qual permite o que é denominado de *computação física*. Um programa é desenvolvido em uma interface de desenvolvimento (IDE) e é feito o upload no microcontrolador, que pode executar então uma série de ações: ligar ou desligar lâmpadas, motores, fazer a medição de sensores etc. O programa não roda apenas em um computador, mas depende de dispositivos externos e neles manifesta os seus resultados (Monk, 2012, p. 7).

O objetivo principal do desenvolvimento da plataforma Arduino era voltado para o ensino de estudantes. Em 2005, ela foi lançada comercialmente por Massimo Banzi e David Cuartielles. Acabou tornando-se um produto extremamente bem-sucedido entre fabricantes, estudantes e artistas. Os projetos de Arduino estão disponíveis gratuitamente sob licença da Creative Commons. Há, no mercado, muitas placas alternativas que “clonam” o Arduino.

Figura 5 – Placa do Raspberry PI



Crédito: Denis Linine/Shutterstock.

Outras plataformas permitem um poder de processamento mais robusto, tal como a Raspberry Pi (Figura 5). Essa placa funciona com o sistema operacional Linux, contendo mais interfaces que as proporcionadas pelo Arduino, tais como HDMI, USB, *drivers* para muitos periféricos. No entanto, são um pouco mais caros e requerem um consumo maior de energia (Oliveira, 2017, p. 209). Projetos em IoT que requerem que o dispositivo efetue análises explícitas ou mesmo implícitas deverão prever a utilização de plataformas tais como a Raspberry Pi, permitindo a implementação de programas de IA escritos em linguagem Python, por exemplo.

TEMA 4 – ARMAZENAMENTO ANALÍTICO

Como visualizado no tema referente ao *big data*, o armazenamento dos dados é um dos componentes principais para a adoção de análises de dados baseadas em ferramentas de aprendizado de máquina. Com relação à estrutura em que os dados são armazenados, o **modelo relacional**, surgido na década de 1970 com Edgar Codd, se mostrou com alta eficiência em aplicações empresariais, permitindo que os dados fossem armazenados com integridade e possibilitando operações de inclusão, alteração e exclusão de dados (Amaral,

2016, p. 25). Um exemplo de um modelo relacional simples pode ser visto na Figura 6.

Figura 6 – Modelo relacional de dados para um conjunto de pedidos, clientes e itens

CLIENTES

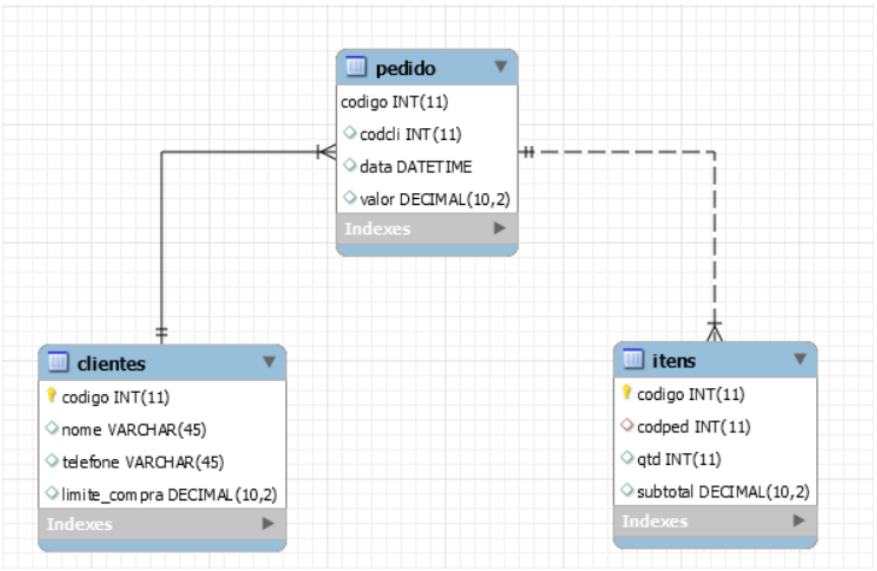
codigo	nome	telefone	limite compra
202	Ernesto	222-0809	1500,00
221	Amélia	233-2474	2000,00
233	Luis Alberto	323-0071	1500,00
282	José Antonio	343-9021	800,00
295	Carlos Silva	224-5678	800,00

PEDIDO

codigo	codcli	data	valor
1005	202	10/03/2004	1200,00
1008	221	13/03/2004	960,00
1010	233	17/03/2004	1020,00
1015	282	20/03/2004	755,00
1023	202	21/03/2004	900,00

ITENS

codigo	codped	qtd	subtotal
001	1005	10	1200,00
002	1008	12	720,00
003	1008	2	240,00
004	1010	1	500,00
005	1010	2	335,00
006	1010	1	185,00
007	1015	1	125,00
008	1015	1	630,00
009	1023	1	400,00
010	1023	1	300,00
011	1023	1	200,00



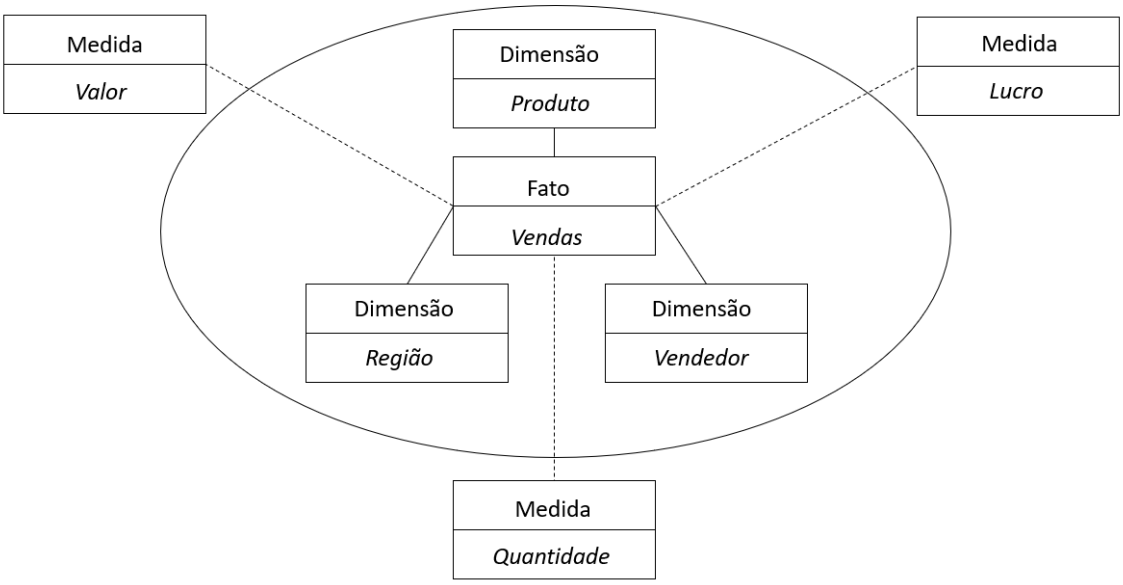
Fonte: Medeiros, 2007.

Apesar de o modelo relacional manter o armazenamento com integridade, com pouca ou mesmo nenhuma redundância, a tarefa de normalizar os dados se mostra difícil para utilização em um contexto de análise de dados. No caso de se recuperar informações para saber, por exemplo, vendas por mês, por produto ou por região, a complexidade de consolidação dos dados é maior, requerendo junções para cada tipo de informação analisada. Além disso, há um custo computacional alto para o sistema gerenciador do banco de dados (Amaral, 2016, p. 29).

Na década de 1990, começa a se popularizar o conceito de *data warehouses*. Estes se constituíam em repositórios de dados, estruturados com base nos bancos de dados relacionais. A sua ideia principal é facilitar as análises de dados, mantendo informações calculadas previamente e dados que não seguem um padrão de normalização tal como no modelo relacional. Outro elemento relativo à construção de *data warehouses* é que os dados podem ser repetidos de acordo com a necessidade de análise, sendo permitida a redundância de dados (Amaral, 2016, p. 40).

De maneira diferenciada do modelo relacional, um *data warehouse* utiliza um modelo multidimensional, estando no centro desse modelo um elemento central denominado de *fato*, que se refere à informação nuclear que se quer analisar. Um fato contém medidas refletidas em valores a serem analisados ou, ainda, calculados previamente. Um fato possui **dimensões**, que são as diferentes características pelas quais se quer analisar o fato. Exemplos de dimensões são tempo, região, tipo de produto, tipo de cliente, categoria etc. (Figura 7).

Figura 7 – Exemplo de modelo multidimensional



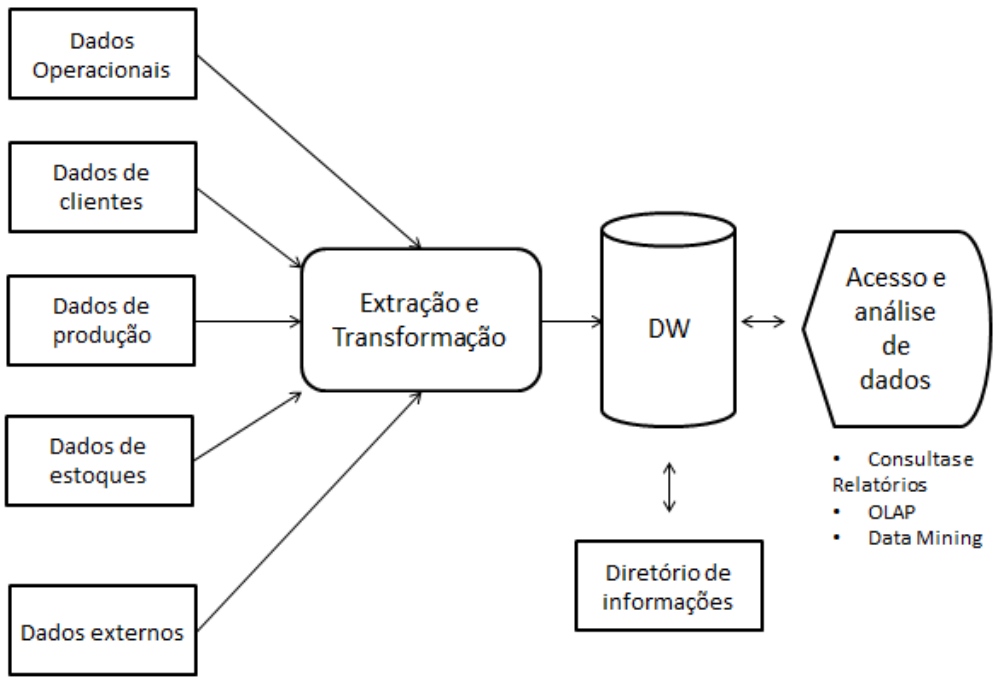
Fonte: Adaptado de Amaral, 2016, p. 42.

Um mesmo fato pode corresponder a várias perguntas, olhando-se conforme a medida para diferentes dimensões. Por exemplo, na Figura 7, pode ser perguntado: “qual produto vendeu mais no ano?”. Ou, ainda: “qual o vendedor que vendeu mais no semestre?”. Ou, então: “qual produto vendeu mais na Região Norte?” e “qual a região que mais proporcionou lucro?”.

Data marts são subconjuntos de um *data warehouse*, em que existe uma porção resumida ou bem focalizada dos dados da empresa em um banco de dados separado, geralmente destinado a um grupo específico de usuários (Figura 8). Por exemplo, *data marts* específicos para as áreas de vendas e marketing podem ser gerados para análise de informações sobre compras dos clientes (Laudon; Laudon, 2010).

Uma vez que os dados sejam organizados em *data warehouses* e *data marts*, eles ficam disponíveis para que sejam feitas análises posteriores. O analista ou usuário pode contar com diversas ferramentas para fazer essas análises, descobrir padrões de comportamento dos dados, correlações e captura de insights úteis para a orientação na tomada de decisão. Dessa forma, as ferramentas de **inteligência de negócios** (*business intelligence* – BI) são ferramentas que permitem consolidar, analisar e acessar massivas quantidades de dados. Algumas das principais ferramentas de BI são *softwares* para consultas de banco de dados, ferramentas para análise multidimensional de dados (ou processamento analítico *on-line*, do inglês *online analytical processing* – Olap) e *data mining*.

Figura 8 – Componentes genéricos de um *data warehouse*



Fonte: Laudon; Laudon, 2010.

O **Olap** é um modelo construído na perspectiva de um banco de dados multidimensional que é associado à construção de cubos de dados. **Cubos** são

representações multidimensionais que normalmente requerem um único fato, em que, por meio de operações denominadas de *drill down* e *drill up*, o usuário pode expandir ou colapsar o nível de detalhes apresentado. Apesar de a representação ser textual, o Olap pode ser utilizado para mostrar os dados de forma gráfica (Amaral, 2016, p. 50).

Figura 9 – Exemplo de Olap, mostrando os dados em uma representação multidimensional (acima) e em um cubo produzido (abaixo)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	OrderID	CustomerI	Employee	OrderDate	Year	RequiredDate	ShippedDate	ShipDiff	OrderDiff	ShipVia	Freight	ShipName	ShipAddre
2	10264	FOLKO	6	24/8/1994	1994	21/9/1994	23/9/1994	(2)	28	3	R$ 3,67	Folk och fe	Åkergatan
3	10264	FOLKO	6	24/8/1994	1994	21/9/1994	23/9/1994	(2)	28	3	R$ 3,67	Folk och fe	Åkergatan
4	10271	SPLIR	6	1/9/1994	1994	29/9/1994	30/9/1994	(1)	28	2	R$ 4,54	Split Rail E	P.O. Box 5
5	10280	BERGS	2	14/9/1994	1994	12/10/1994	13/10/1994	(1)	28	1	R$ 8,98	Berglunds	Berguvsvä
6	10280	BERGS	2	14/9/1994	1994	12/10/1994	13/10/1994	(1)	28	1	R$ 8,98	Berglunds	Berguvsvä
7	10280	BERGS	2	14/9/1994	1994	12/10/1994	13/10/1994	(1)	28	1	R$ 8,98	Berglunds	Berguvsvä
8	10302	SUPRD	4	11/10/1994	1994	8/11/1994	9/11/1994	(1)	28	2	R$ 6,27	Suprêmes	Boulevard
9	10302	SUPRD	4	11/10/1994	1994	8/11/1994	9/11/1994	(1)	28	2	R$ 6,27	Suprêmes	Boulevard
10	10302	SUPRD	4	11/10/1994	1994	8/11/1994	9/11/1994	(1)	28	2	R$ 6,27	Suprêmes	Boulevard
11	10309	HUNGO	3	20/10/1994	1994	17/11/1994	23/11/1994	(6)	28	1	R$ 47,30	Hungry Ov	8 Johnstov
12	10309	HUNGO	3	20/10/1994	1994	17/11/1994	23/11/1994	(6)	28	1	R$ 47,30	Hungry Ov	8 Johnstov
13	10309	HUNGO	3	20/10/1994	1994	17/11/1994	23/11/1994	(6)	28	1	R$ 47,30	Hungry Ov	8 Johnstov
14	10309	HUNGO	3	20/10/1994	1994	17/11/1994	23/11/1994	(6)	28	1	R$ 47,30	Hungry Ov	8 Johnstov
15	10309	HUNGO	3	20/10/1994	1994	17/11/1994	23/11/1994	(6)	28	1	R$ 47,30	Hungry Ov	8 Johnstov
16	10320	WARTH	5	3/11/1994	1994	17/11/1994	18/11/1994	(1)	14	3	R$ 34,57	Wartian Hi	Torikatu 3
17	10380	HUNGO	8	12/1/1995	1995	9/2/1995	16/2/1995	(7)	28	3	R$ 35,03	Hungry Ov	8 Johnstov
18	10380	HUNGO	8	12/1/1995	1995	9/2/1995	16/2/1995	(7)	28	3	R$ 35,03	Hungry Ov	8 Johnstov
19	10380	HUNGO	8	12/1/1995	1995	9/2/1995	16/2/1995	(7)	28	3	R$ 35,03	Hungry Ov	8 Johnstov
20	10380	HUNGO	8	12/1/1995	1995	9/2/1995	16/2/1995	(7)	28	3	R$ 35,03	Hungry Ov	8 Johnstov
21	10423	GOURL	6	23/2/1995	1995	9/3/1995	27/3/1995	(18)	14	3	R$ 24,50	Gourmet L	Av. Brasil,
22	10423	GOURL	6	23/2/1995	1995	9/3/1995	27/3/1995	(18)	14	3	R$ 24,50	Gourmet L	Av. Brasil,
23	10427	PICCO	4	27/2/1995	1995	27/3/1995	3/4/1995	(7)	28	2	R$ 31,29	Piccolo un	Geislweg
24	10433	PRINI	3	6/3/1995	1995	3/4/1995	4/4/1995	(1)	28	3	R$ 73,83	Princesa I	Estrada d
25	10451	QUICK	4	22/3/1995	1995	5/4/1995	12/4/1995	(7)	14	3	########	QUICK-Sto	Taucherst

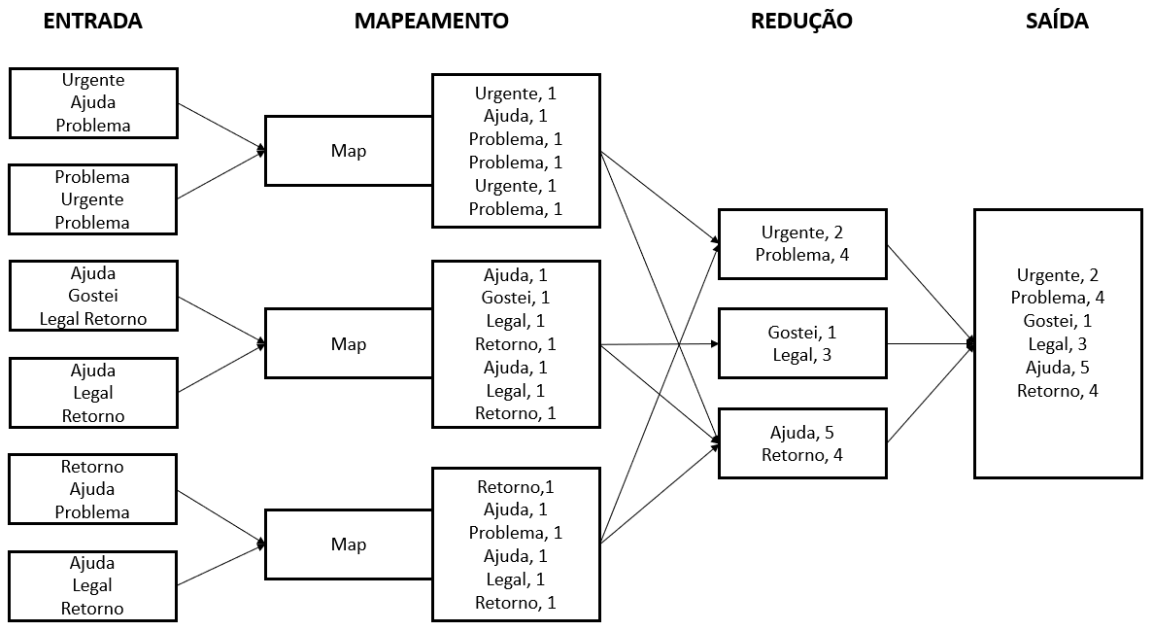
Soma de Total		Year							
Customers_C	CustomerID	1994		1995		1996		Total geral	
Argentina	CACTU			R$	676,50	R$	3.016,10	R$	3.692,60
	OCEAN			R$	748,40	R$	11.253,00	R$	12.001,40
	RANCH			R$	3.004,80	R$	4.076,10	R$	7.080,90
Argentina Total				R$	4.429,70	R$	18.345,20	R$	22.774,90
Austria	ERNSH	R$	40.693,84	R$	155.664,98	R$	202.267,48	R$	398.626,30
	PICCO	R$	17.186,56	R$	32.303,94	R$	9.900,00	R$	59.390,50
Austria Total		R$	57.880,40	R$	187.968,92	R$	212.167,48	R$	458.016,80
Belgium	MAISD			R$	5.706,00	R$	18.145,52	R$	23.851,52
	SUPRD	R$	18.920,10	R$	25.397,14	R$	44.761,80	R$	89.079,04
Belgium Total		R$	18.920,10	R$	31.103,14	R$	62.907,32	R$	112.930,56
Brazil	COMMI	R$	8.676,00	R$	1.344,00	R$	919,50	R$	10.939,50
	FAMIA	R$	3.257,68	R$	9.393,86			R$	12.651,54
	GOURL			R$	19.848,03	R$	2.347,95	R$	22.195,98
	HANAR	R$	8.992,20	R$	12.581,08	R$	39.013,20	R$	60.586,48
	QUEDE	R$	2.388,56	R$	12.161,06	R$	4.690,32	R$	19.239,94
	QUEEN			R$	67.606,98	R$	21.519,83	R$	89.126,81
	RICAR	R$	3.156,00	R$	9.481,94	R$	21.041,06	R$	33.679,00
	TRADH	R$	1.296,00	R$	3.581,20	R$	12.416,52	R$	17.293,72
	WELLI	R$	1.035,60	R$	11.089,26	R$	5.953,28	R$	18.078,14
Brazil Total		R$	28.802,04	R$	147.087,41	R$	107.901,66	R$	283.791,11
Canada	BOTTM			R$	28.568,75	R$	32.021,43	R$	60.590,18
	LAUGB			R$	949,00	R$	561,00	R$	1.510,00
	MEREP	R$	15.422,64	R$	69.183,14			R$	84.605,78
Canada Total		R$	15.422,64	R$	98.700,89	R$	32.582,43	R$	146.705,96
Denmark	SIMOB	R$	705,20	R$	47.550,80	R$	12.003,32	R$	60.259,32
	VAFFE	R$	3.336,80	R$	32.356,45	R$	15.772,08	R$	51.465,33
Denmark Total		R$	4.042,00	R$	79.907,25	R$	27.775,40	R$	111.724,65

Em face dos problemas de armazenamento e processamento distribuído de dados, foi necessário o desenvolvimento de novos modelos de armazenamento. Um deles é o **MapReduce**, proposto por funcionários do Google (Jeffery Dean e Sanjay Ghemawat) para o processamento de volumes muito

grandes de dados de maneira simplificada. O modelo permite dividir o processamento entre vários computadores de uma rede. Cada computador é designado de *nó* e todos esses computadores, juntos, formam o que se chama de *agrupamentos* ou *clusters*. A carga do processamento é distribuída e balanceada, com a possibilidade de tolerância a falhas (Amaral, 2016, p. 57).

O MapReduce opera em uma representação de dados chamada de *chave-valor*. Ele contém duas funções principais: o **mapeamento** e a **redução**. O mapeamento vai encontrando os dados nos nós de um *cluster*, conforme a função de mapeamento. A redução recebe o resultado do mapeamento e consolida os dados. Os *n* nós que efetuam as funções de mapeamento entregam o resultado do seu processamento para os *n* nós que perfazem a operação de redução (Amaral, 2016, p. 57). Uma implementação bastante popular do modelo MapReduce é o **Hadoop**, mantido pela Fundação Apache (Amaral, 2016, p. 59).

Figura 10 – Exemplo do modelo MapReduce



Fonte: Adaptado de Amaral, 2016, p. 58.

Quando os dados são armazenados por centenas ou mesmo milhares de computadores, é necessária a existência de um sistema de arquivos que consiga fazer o gerenciamento desses dados. O *hadoop distributed file system* (**HDFS**) é um tipo de sistema de arquivos distribuídos estruturado em uma arquitetura mestre/escravo. Um dos nós, denominado de *nó mestre*, contém os metadados, nomes de arquivos, permissões de acesso e localização de cada bloco de armazenamento. Os nós escravos, que, por padrão, contêm, cada um, 64 MB,

armazenam, por sua vez, os dados. Existe a possibilidade de armazenamento redundante: no caso de falha de um nó principal, outro nó secundário, que contenha os mesmos dados, pode substituí-lo (Amaral, 2016, p. 58-59).

Para dar conta dos novos requisitos de análise de dados, o modelo relacional deu lugar a uma nova geração de sistemas de gerenciamento de banco de dados conhecidas como *NoSQL*. O termo *NoSQL* indica que os bancos de dados estão armazenados em uma estrutura que não é a do modelo relacional. O modelo mais tradicional de NoSQL utiliza o conceito de chave-valor, tal como visto no MapReduce. Ao invés de incluir um conjunto de atributos, a operação inclui somente uma chave com o seu valor respectivo. Alguns tipos de sistemas NoSQL são orientados a documentos, tal como o XML e o Json.

Figura 11 – Exemplos de representação NoSQL orientados a documentos: XML e Json

XML

```
<?xml version="1.0" encoding="utf-8" ?>
<peixes>
        <peixe>
            <id>1</id>
                <nome>Abotoado</nome>
                <nomecientifico>Pterodoras granulosus</nomecie
                <habitat><![CDATA[O Abotoado habita águas de g
                <alimento><![CDATA[É uma espécie omnívora, ali
                <caracteristicas><![CDATA[O peixe Abotoado é u
                <url>http://www.cpt.com.br/artigos/peixes-de-a
                <urlfoto>http://cptstatic.s3.amazonaws.com/ima
                <urlicone>http://qcsimulator.com.br/peixes/pei
                <latitude>-2.155676</latitude>
                <longitude>-55.0486666</longitude>
        </peixe>
        <peixe>
            <id>2</id>
                <nome>Corvina</nome>
                <nomecientifico>Plagioscion squamosissimus</nc
                <habitat><![CDATA[O Corvina é um peixe que hab
                <alimento><![CDATA[É um peixe piscívoro, alime
                <caracteristicas><![CDATA[O peixe Corvina é um
                <url>http://www.cpt.com.br/artigos/peixes-de-a
                <urlfoto>http://cptstatic.s3.amazonaws.com/ima
                <urlicone>http://qcsimulator.com.br/peixes/pei
                <latitude>-28.4800072</latitude>
                <longitude>-49.0074554</longitude>
        </peixe>
```

JSON

```
{
  "peixes": {
    "peixe": [
      {
        "id": "1",
        "nome": "Abotoado",
        "nomecientifico": "Pterodoras granulosus",
        "habitat": "O Abotoado habita águas de grande profundidade, como rios,
        "alimento": "É uma espécie omnívora, alimentando-se preferencialmente
        "caracteristicas": "O peixe Abotoado é um peixe de couro. Seu corpo é
        "url": "http://www.cpt.com.br/artigos/peixes-de-agua-doce-do-brasil-ab
        "urlfoto": "http://cptstatic.s3.amazonaws.com/imagens/enviadas/materia
        "urlicone": "http://qcsimulator.com.br/peixes/peixe_abotoado.png",
        "latitude": "-2.155676",
        "longitude": "-55.0486666"
      },
      {
        "id": "2",
        "nome": "Corvina",
        "nomecientifico": "Plagioscion squamosissimus",
        "habitat": "O Corvina é um peixe que habita poços, remansos e reserva
        "alimento": "É um peixe piscívoro, alimentando-se de outros peixes e
        "caracteristicas": "O peixe Corvina é um peixe de escamas, com colora
        "url": "http://www.cpt.com.br/artigos/peixes-de-agua-doce-do-brasil-co
        "urlfoto": "http://cptstatic.s3.amazonaws.com/imagens/enviadas/materia
        "urlicone": "http://qcsimulator.com.br/peixes/peixe_corvina.png",
        "latitude": "-28.4800072",
        "longitude": "-49.0074554"
      },
      {
```

TEMA 5 – ANÁLISE DE DADOS

A análise de dados é a identificação de algum tipo de transformação nos dados, em busca de informação ou conhecimento. Em termos de análise, pode-se aplicar uma gama de técnicas ou ferramentas, dependendo do objetivo de análise em questão. De maneira a propor uma divisão nas abordagens de análise de dados, um cientista de dados pode utilizar três formas: **exploratória**, **explícita** ou **implícita**.

Uma primeira abordagem aos dados, dos quais se tem pouco ou nenhum conhecimento, começa pela análise **exploratória**. O objetivo aqui é buscar conhecer os dados antes de tentar fazer uma análise deles, utilizando-se técnicas

explícitas ou implícitas. A análise exploratória pode-se utilizar tanto de técnicas quantitativas quanto visuais (Amaral, 2016, p. 64).

Técnicas quantitativas podem fazer uso de medidas de tendência central, tais como média, mediana e moda; e de medidas de dispersão, por exemplo variância e desvio-padrão. O uso de softwares estatísticos, tais como o R, é recomendável nessa fase, de maneira a tornar produtiva a análise exploratória. A título de exemplo, será utilizado um banco de dados muito popular, que auxilia na compreensão, de maneira didática, das técnicas que podem ser aplicadas a diferentes contextos: o banco de dados **íris** (Dua; Graff, 2019). Esse banco de dados foi coletado nos anos 1930, sendo composto de um conjunto de 150 amostras da planta íris, separadas em três espécies diferentes. Assim, são 50 amostras de cada espécie (setosa, versicolor e virgínica). Cada amostra se refere a uma planta, em que se obtiveram as medições de quatro atributos distintos: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. Na Figura 12 encontra-se um sumário de informações quantitativas a respeito do banco de dados íris.

Figura 12 – Sumário de informações quantitativas do banco de dados íris, gerado no software R

```
Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Outra maneira é explorar os dados visualmente. A Figura 13 mostra um exemplo de um diagrama de dispersão dos dados. O diagrama de caixa (*boxplot*) também é bastante utilizado por integrar, em um único gráfico, uma série de medidas, proporcionando a comparação visual de diferentes grandezas em análise. Na Figura 14, pode ser visto um exemplo de diagrama de caixa com quatro grandezas consideradas para o banco de dados íris.

Histogramas são gráficos que permitem identificar como os dados se distribuem, utilizando-se uma grandeza de cada vez. No histograma, pode-se identificar as frequências dos dados em cada intervalo de valores possíveis considerados. Na Figura 15 pode ser visualizado um exemplo de histograma.

Figura 13 – Exemplo de diagrama de dispersão de dados

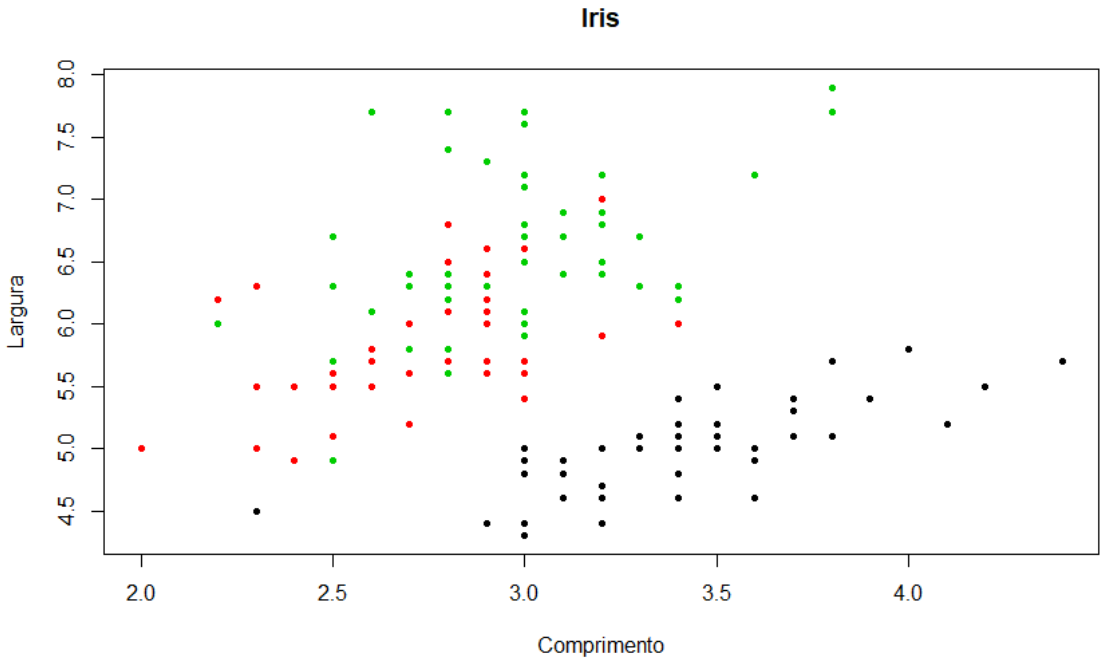


Figura 14 – Exemplo de diagrama de caixa (*boxplot*)

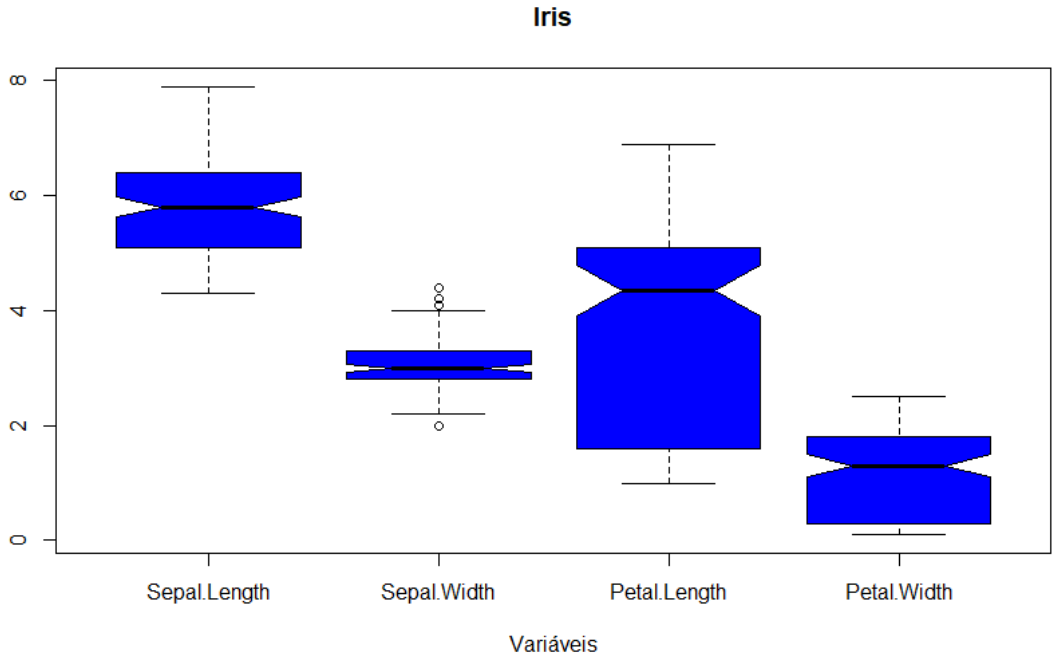
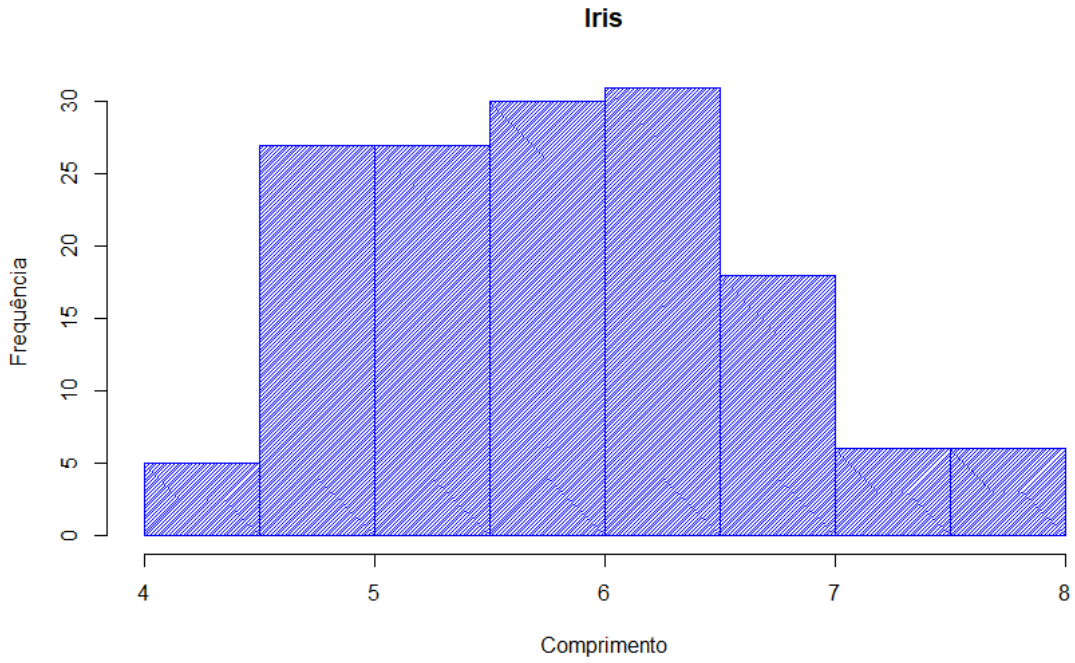


Figura 15 – Exemplo de histograma para uma grandeza do banco de dados íris



Quando se considera mineração de textos, como as que acontecem em análises de redes sociais, pode-se utilizar a nuvem de palavras (*tag cloud*). A ideia é considerar as palavras que aparecem com mais frequência no texto com a fonte maior em relação às outras. A Figura 16 mostra um exemplo de uma nuvem de palavras.

Figura 16 – Exemplo de nuvem de palavras

Fonte: Adaptado de Amaral, 2016, p. 70.

As **análises explícitas** são técnicas simples, nas quais se busca destacar informações existentes nos dados. A diferença com relação à análise exploratória reside mais nos objetivos do que nas técnicas. A análise explícita busca exemplos mais específicos como resumir vendas em um mês, verificar notas que faltam, verificar cálculo de imposto a pagar etc. (Amaral, 2016, p. 73). Basicamente, as técnicas utilizadas na análise explícita se resumem às operações que podem ser feitas com os dados, tais como junções ou antijunções que permitam relacionar dados de tabelas diferentes por meio de chaves comuns (no jargão de bancos de dados, *chaves primárias e estrangeiras*). Assim, uma ferramenta útil para executar análises explícitas é um sistema gerenciador de bancos de dados, o qual fornece uma interface de linguagem de dados, tais como SQL, para permitir a extração das informações desejadas.

As **análises implícitas** se referem àquelas nas quais se deseja conhecer em profundidade características que existem nos dados, as quais não são facilmente observadas, quando se olha para os dados diretamente. Aqui, entram as diversas técnicas estudadas na IA para a descoberta de padrões existentes nos dados. Com base em técnicas de aprendizado de máquina (*machine learning*), pode-se encontrar padrões que estejam ocultos em um conjunto de dados (Amaral, 2016, p. 81).

O aprendizado de máquina também está intimamente ligado a técnicas estatísticas e à mineração de dados. Enquanto o *machine learning* está relacionado a algoritmos que buscam o reconhecimento de padrões em dados, a mineração de dados se refere à aplicação desses algoritmos em conjuntos de dados massivos, em busca de informação e conhecimento (Amaral, 2016, p. 81). Dessa forma, tem-se a relação entre o *big data* e a mineração de dados, pois esta pode ser operacionalizada por intermédio das técnicas de *machine learning*.

REFERÊNCIAS

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

BATTITI, R.; BRUNATO, M. **The Lion way**: Machine Learning plus Intelligent Optimization – version 3.0. Trento: LionLab, 2017.

LAUDON, K.; LAUDON, J. **Sistemas de informação gerenciais**. 9. ed. São Paulo: Pearson Prentice Hall, 2010.

LEMBERGER, P.; MOREL, M. B. M.; RAFFAELLI, J. L. **Big data et machine learning**: manuel du data scientist. Paris: Dunod, 2015.

MEDEIROS, L. F. **Banco de dados**: princípios e prática. Curitiba: Ed. Ibpex, 2007.

MONK, S. **Programação com Arduino**: começando com sketches. Porto Alegre: Bookman, 2012.

OLIVEIRA, S. **Internet das coisas com ESP8266, Arduino e Raspberry Pi**. São Paulo: Novatec, 2017.