



# **FUNDAMENTOS DE BIG DATA**

Autoria: Jenifer Vieira Toledo Tavares

1<sup>a</sup> Edição

**UNIASSELVI-PÓS**

Programa de Pós-Graduação EAD



**UNIASSELVI**

CENTRO UNIVERSITÁRIO LEONARDO DA VINCI  
Rodovia BR 470, Km 71, nº 1.040, Bairro Benedito  
Cx. P. 191 - 89.130-000 – INDAIAL/SC  
Fone Fax: (47) 3281-9000/3281-9090

Reitor: Prof. Hermínio Kloch

Diretor UNIASSELVI-PÓS: Prof. Carlos Fabiano Fistarol

Equipe Multidisciplinar da Pós-Graduação EAD:

Carlos Fabiano Fistarol

Ilana Gunilda Gerber Cavichioli

Jóice Gadotti Consatti

Norberto Siegel

Julia dos Santos

Ariana Monique Dalri

Marcelo Bucci

Revisão Gramatical: Equipe Produção de Materiais

Diagramação e Capa:

Centro Universitário Leonardo da Vinci – UNIASSELVI

**Copyright © UNIASSELVI 2019**

Ficha catalográfica elaborada na fonte pela Biblioteca Dante Alighieri  
UNIASSELVI – Indaial.

T231f

Tavares, Jenifer Vieira Toledo

Fundamentos de big data. / Jenifer Vieira Toledo Tavares. –  
Indaial: UNIASSELVI, 2019.

123 p.; il.

ISBN 978-85-7141-330-6  
ISBN Digital 978-85-7141-331-3

1. Big data. - Brasil. II. Centro Universitário Leonardo Da Vinci.

CDD 658

Impresso por:

# SUMÁRIO

APRESENTAÇÃO .....	07
CAPÍTULO 1	
INTRODUÇÃO A BIG DATA .....	07
CAPÍTULO 2	
TRABALHANDO COM BIG DATA .....	47
CAPÍTULO 3	
TECNOLOGIAS PARA BIG DATA .....	87



# APRESENTAÇÃO

Este livro foi escrito baseado em pesquisas sobre o que já se escreveu e vem sendo realizado sobre *Big Data*, com o objetivo de proporcionar ao leitor um entendimento teórico e prático do conceito *Big Data*, compreendendo tarefas como coleta, armazenamento, processamento e visualização de dados, métodos essenciais para profissionais que atuam analisando dados na área de Tecnologia da Informação.

No primeiro capítulo é apresentada uma introdução para contextualizar os principais aspectos de *Big Data*, o crescimento e evolução dos dados, a importância destes para os negócios, tipos e meios de obtê-los. Ainda trataremos de *Data Science* e Modelos Preditivos.

Já no segundo capítulo começaremos a trabalhar com o *Big Data*, compreendendo as maneiras para se obter dados, armazená-los e principalmente gerenciá-los e monitorá-los para posteriores análises na tomada de decisão.

Finalizamos com o terceiro capítulo, em que são tratadas algumas tecnologias e suas aplicações para *Big Data*, como *Internet das Coisas (IoT)* e *Machine Learning*. E assim chegamos ao final deste livro, mas não da sua jornada de aprendizagem. A partir desta leitura, coloque a mão na massa e pratique os novos conhecimentos adquiridos.

E saiba, é prazeroso trazer para você, leitor, um pouco mais de esclarecimentos e conhecimentos para você mergulhar e pesquisar sobre *Big Data* e quem sabe chegar a desenvolver um ótimo projeto de *Big Data*.

Boa leitura e espero contribuir para seu crescimento profissional e pessoal.

Prof<sup>a</sup>. Jenifer Vieira Toledo Tavares





# CAPÍTULO 1

## INTRODUÇÃO A *BIG DATA*

**A partir da perspectiva do saber-fazer, são apresentados os seguintes objetivos de aprendizagem:**

- ✓ conhecer os principais conceitos e desafios sobre Big Data;
- ✓ caracterizar os aspectos essenciais de Big Data, como: volume, variedade, velocidade, veracidade e valor dos dados;
- ✓ classificar, comparar e diferenciar os principais conceitos de Big Data para distinguir e debater sobre eles;
- ✓ explicar o cenário geral de Data Science;
- ✓ conhecer quais são os tipos de Modelos Preditivos e para que servem.



# 1 CONTEXTUALIZAÇÃO

Organizações das mais diversas áreas, como fabricantes, educadores, bancários, serviços públicos e privados, comércio, indústrias, entre tantos outros, buscam cada vez mais formas sistemáticas para identificar, gerenciar e integrar os dados dos seus consumidores que estão disponíveis nos mais diversos meios de comunicação.

São estes dados, abundantes e disponíveis em diversos meios, que executivos estão utilizando para obter *insights* importantes sobre suas organizações, sobre o mercado, seus consumidores e seus processos organizacionais. Porém, diariamente, as organizações recebem uma quantidade incalculável de dados, estando diante um grande desafio conseguir separar todos os dados relevantes dos dados não importantes para a tomada de decisões organizacionais.

Nesse contexto, as organizações precisam do apoio de profissionais, como equipes de *analytics* de forma integrada, para saber como captar, organizar, analisar e gerenciar dados dos consumidores e suas operações, trazendo de fato impactos decisivos para o mercado e a concorrência.

Esse grande volume de dados vem sendo denominado como *Big Data* – o processo da coleta, tratamento e refinamento de dados, que permite utilizar informações na busca de valores dos negócios, seja para pesquisas, controles governamentais e, claro, também para própria sociedade.

E você? Já notou que tudo que fazemos, todas as ações, como escolher em qual restaurante ir, qual caminho percorrer, com qual médico agendar uma consulta, tudo é direcionado por dados? Sejam experiências que tivemos, o que gostamos e o que não gostamos, o que desejamos ou nos motivamos, tudo se resume a tomar decisões, de alguma maneira, por meio de algum dado. Portanto, o mundo todo se baseia em análises de dados para tomada de decisões, concorda?

A partir desse questionamento, é importante que você reflita sobre esse cenário, para relacionar os conceitos que serão apresentados neste primeiro capítulo, características essenciais para compreensão e aplicação de *Big Data*.

Ótimos estudos!

# 2 A ERA DOS DADOS

O vasto uso de dispositivos móveis, de redes sociais e da *web*, em que vários dados são disponibilizados constantemente, gerou uma considerável elevação da quantidade de dados armazenados e trafegados no mundo. Há



também uma crescente abundância de dados originados por organizações. Essa exponenciação de dados tornou-se um dos principais desafios para Ciência da Computação (MCAFEE; BRYNJOLFSSON, 2012).

As organizações vêm utilizando a extração de dados para atrair mais consumidores. E é através da coleta dos dados, seja por dispositivos móveis, *sites*, aplicativos, redes sociais, entre outros, que o comportamento dos consumidores pode ser identificado, permitindo que as organizações desenvolvam estratégias para fornecer novos produtos e serviços mais inteligentes. Dados interessantes são apresentados pela pesquisa da *International Business Machines* (IBM) com trinta mil consumidores em treze países. Constatou-se que 78% a 84% dos consumidores se baseiam nas mídias sociais quando pensam em comprar algum produto; 45% pedem opinião aos amigos e parentes e 18% apenas se fundamentam nas informações de varejistas.

Apesar da Era dos Dados proporcionar oportunidades por meio dos dados, existe o desafio de conseguir organizá-los e identificar melhorias nas futuras interações entre organização e consumidor. Mesmo com as empresas investindo nesta área e possuindo dados significativos, ainda não conseguem gerar valores para estes dados e posteriormente para os seus negócios. Marcas centenárias estão enfrentando a expectativa dos seus consumidores por novos produtos e serviços, ao mesmo tempo em que precisam lidar com uma série de sistemas legados e culturas organizacionais e se relacionarem com empresas que já nasceram neste mundo digital.

A maioria das organizações começa a ter informações descentralizadas devido ao uso de diversos canais simultaneamente, *sites*, lojas físicas e redes sociais. Mas as empresas precisam que todos estes sistemas conversem entre si e todas as informações sejam consolidadas de forma única para tocar seu processo. Um cliente que acessa qualquer canal espera que ele seja visto como um cliente único. Se ontem foi feita uma interação do cliente com a organização utilizando um canal e hoje por meio de outro canal, todos devem saber dessas interações que foram realizadas pelo cliente para melhor atendê-lo. Esse tratamento da informação centralizada também é um desafio e possível diferencial competitivo, proporcionado por essa nova Era dos Dados, trazendo uma nova visão de modelos comerciais e gerenciais.

Portanto, o saneamento de dados ou a obtenção de uma informação apta a ser consumida pela organização é poder ter uma informação confiável para tomada de decisões. E muitas empresas acabam tendo decisões calcadas em informações não confiáveis ou duplicadas. Na Era dos Dados, esse saneamento de dados é tão importante quanto um ativo financeiro e deve ser tratado com o mesmo cuidado. Sanear os dados, ter dados confiáveis e atualizados é fator

crítico para proporcionar um serviço único diante os diversos canais de acesso para o cliente.

Inicialmente, o que já podemos concluir sobre a Era dos Dados? As empresas podem ficar imersas esperando que as mudanças passem por cima delas, ficando para trás, ou podem identificar oportunidades para aumentar seu espaço no mercado, sendo mais ágeis que os concorrentes?

Como primeiro passo para responder a essas questões, é fundamental buscar entender e identificar as oportunidades que a Era dos Dados pode trazer para os produtos e serviços oferecidos aos consumidores. Um exemplo para ajudar: a Tesco – multinacional varejista britânica, de pequenas lojas e supermercados, desenvolveu um aplicativo móvel que permite aos seus clientes em qualquer lugar, na casa dos amigos ou em suas próprias casas, escanear o código de barras dos produtos que queiram comprar e adicioná-los automaticamente a sua lista de compras. Vamos saber mais algumas informações nos próximos tópicos e descobrir as oportunidades da Era dos Dados que levaram ao surgimento do *Big Data*!

*“O sucesso das organizações depende das pessoas e da utilização inteligente das informações disponíveis.” Peter Drucker*

*“O sucesso das organizações depende das pessoas e da utilização inteligente das informações disponíveis.” Peter Drucker*

---

Neste curso que você está iniciando, a UNIASSELVI disponibiliza na sua plataforma virtual de aprendizagem um conjunto de materiais que irá auxiliar nos seus estudos. Acesse com seu *login* e senha no site da instituição: <[www.uniasselvios.com.br](http://www.uniasselvios.com.br)>.

---



## 2.1 SURGIMENTO DO *BIG DATA*

Lembra quando as pessoas eram público-alvo ou *targets*, audiências, como se só tivessem ouvidos? Nesta época do *broadcast*, os consumidores eram quase só números, uma grande massa, sem rosto ou voz. Até que esse cenário acabou! Agora com a comunicação em Rede, cada indivíduo ganha poder e voz. Pode ter sua mídia pessoal e falar de igual para igual com grandes empresas e marcas. Ao invés de receptoras de mensagens, as pessoas estão se tornando emissoras de informações.

Estamos conectados digitalmente, desde a hora que acordamos até a hora de dormir. E durante este intervalo diário, absorvemos e geramos um imenso



volume de conteúdo. Observe alguns números para entender melhor o que estou querendo mostrar: a famosa rede social *Facebook* recebe mais de 1 (um) bilhão de usuários por dia. Ressalto que, conforme a Organização das Nações Unidas (ONU), a Terra tem em média 7 (sete) bilhões de habitantes. Mais de quatro bilhões de *likes* e 300 milhões de fotos são publicadas todos os dias nesta rede. O Brasil tem em média 200 milhões de pessoas, portanto, é como se todo dia uma pessoa brasileira publicasse mais de uma foto no *Facebook*. Uma em cada cinco *pageviews* nos EUA é do *Facebook*, uma audiência surpreendente (ZEPHORIA, 2018). A *Netflix* também possui números relevantes, em horários de pico, ou seja, quando os usuários estão em casa, mais de 30% de toda banda de *Internet* dos EUA é do *Netflix*. O *Netflix* tem 75 milhões de assinantes e 15 mil títulos. Em média os usuários assistem dez bilhões de horas por mês (VARIETY, 2015).

Agora vamos sair um pouco do ambiente *web* e tratar o ambiente dos dispositivos móveis. Uma pesquisa da GSMA apresentou que dois terços da população mundial estão conectados utilizando dispositivos móveis, existem mais dispositivos conectados à *Internet* do que seres humanos na Terra. E cada um desses dispositivos pode gerar *gigabytes* de dados todos os dias, pode ser um sensor, um celular, um satélite, uma geladeira, são diversos dispositivos gerando dados invariavelmente, observe a Figura 1 e veja o que acontece em 30 segundos na *Internet*:

FIGURA 1 – DADOS EM 30 SEGUNDOS NA INTERNET



FONTE: <[digitalinformationworld.com](http://digitalinformationworld.com)>. Acesso em: 19 jan. 2019.

São diversos canais, sendo cada vez mais utilizados, seja por meio de uma rede social, um *site* de *e-commerce*, um dispositivo, um sensor, uma área científica e esses dados nos mostram um volume de dados aumentando significativamente. O volume de dados cresce constantemente, junto com suas fontes e naturezas, havendo dados heterogêneos. Compartilhamos o tempo todo o que estamos consumindo, por onde estamos passando, o que assistimos, o que lemos ou comemos, quem conhecemos, que jogos estamos vendo ou até mesmo jogando.

E, ao navegar pela Rede, deixamos rastros digitais, com inúmeras informações de valor, tais como informações de como consumimos a mídia e tomamos decisões de compras, como influenciamos nossos amigos, do que gostamos ou não das marcas, que assuntos pesquisamos nos sites de buscas, entre outras.

Antes da Era dos Dados, quando se queria investir em uma campanha de propaganda ou lançamento de produto, era preciso investir fortemente em pesquisas de opinião. Isso já não é mais preciso, as pessoas estão inundando as plataformas digitais de opiniões e informações, mais do que empresas e marcas conseguem administrar. A humanidade nunca produziu tantos dados, boa parte deles nos últimos anos, transformando a *Internet* em uma imensa plataforma de pesquisa, gerando um imenso manancial de dados e métricas, que precisam ser integrados a informações de outras fontes de mídias e contatos, além de dados de mercado e consumo para gerar valor as tomadas de decisões das organizações.

“Tão importante quanto gerar informação é a capacidade de processamento de dados volumosos em alta velocidade. Isso comprova de fato de que, nas últimas décadas, presenciamos o desenvolvimento de supercomputadores que atendam essa necessidade: quanto mais a tecnologia foi penetrando no meio social, mais informações as pessoas foram gerando e consumindo” (VOLPATO; RUFINO; DIAS, 2014, p. 5).

*“Tão importante quanto gerar informação é a capacidade de processamento de dados volumosos em alta velocidade. Isso comprova de fato de que, nas últimas décadas, presenciamos o desenvolvimento de supercomputadores que atendam essa necessidade: quanto mais a tecnologia foi penetrando no meio social, mais informações as pessoas foram gerando e consumindo” (VOLPATO; RUFINO; DIAS, 2014, p. 5).*

---

Surge uma dúvida, como transformar esse monte de dados desestruturados em inteligência de negócios e oportunidades? Para responder, tente lembrar um momento de sua vida em que utilizou algum recurso, de alguma empresa, inserido alguns dos seus dados e que depois surpreendentemente você recebeu um e-mail da mesma empresa lhe “presenteando” com um belo desconto. Como este fato aconteceu?

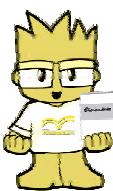
---



O uso das informações, depois de armazenadas em Banco de Dados e analisadas por algoritmos de estatísticas de software, está causando mudanças e progressos significativos ao redor do mundo. Elas capacitam pessoas, empresas e comunidades a criarem soluções para problemas reais e usam seus recursos de um modo mais eficaz. Vejamos alguns casos de sucesso:



- No Canadá, pesquisadores criaram mecanismos que possibilitaram que médicos melhorassem o atendimento monitorando bebês prematuros, rastreando mais de mil pontos de dados por segundo, antecipando eventuais problemas.
- Produtores de soja brasileiros aumentam a eficácia do controle de pragas, reduzem custos e aperfeiçoam os métodos de cultivo com o uso de software para análise de dados.
- Estocolmo instalou 1600 sistemas GPS em táxis para coletar dados sobre o tráfego, depois utilizou um software para analisar os dados e implantar planos para reduzir o congestionamento da cidade.



- **Dados:** são compostos por fatos coletados, estatísticas ou entradas aleatórias que detêm pouco valor.
- **Informação:** conjunto de dados analisados, é derivada de uma coleção de dados processados em que o contexto e o significado foram adicionados a fatos diferentes que permitem uma melhor análise e a interpretação/compreensão pelo receptor da informação.
- **Conhecimento:** é a informação refinada por meio da análise, informação interpretada e aplicada a um fim.

### 3 DESCOPRINDO *BIG DATA*

Muito se fala sobre *Big Data*, mas o que é *Big Data*?

Quando você quer muito uma coisa, por exemplo, um fone de ouvido sem fio. E vive pensando nesse fone, fala para seus amigos pelas redes sociais que quer muito esse fone, pesquisa no *Google*, em sites *e-commerce*, até que de repente o fone aparece em tudo em sua vida, todos os anúncios na *Internet* são sobre o fone sem fio, você acha isso coincidência? Não é! É *Big Data*.

*Big Data* é o nome que se dá à enorme quantidade de dados produzidos a cada segundo por milhares de aparelhos do mundo inteiro, desde o seu celular, passando por um avião pousando no aeroporto, as câmeras de trânsito, até satélites orbitando ao redor da Terra. Ainda pode ser classificado como o processamento e análise de conjuntos de dados extremamente grandes, que não podem ser processados utilizando-se ferramentas convencionais de processamento de dados. Imagine um

Banco de Dados convencional – *SQL Server* ou *Oracle Database*, por maior que seja um Banco de Dados utilizando essas tecnologias e por melhores que sejam as ferramentas de análise, *Big Data* se refere a algo ainda maior, que não pode ser processado utilizando ferramentas convencionais.



Na década de 50 surgiram os primeiros computadores universitários e militares. E como os dados eram guardados? Eram guardados em papel, ao se fazer um cadastro, este era preenchido em uma ficha, que era armazenada em uma pasta e posteriormente em um armário. Era a única maneira de se armazenar dados. Mas hoje já não funciona desta maneira, apesar de que, ainda arquivamos muita coisa no papel. Seguindo essa mesma linha de raciocínio das fichas, pastas e armários e trazendo para a área de Tecnologia da Informação (TI), temos os registros, tabelas e arquivos. Visto o meio de armazenamento da década de 50, um acúmulo gigante de papel estava ocorrendo. E o grande desafio entre a década de 50 e 60 era digitalizar todos esses dados, pois a computação já não pertencia apenas às universidades e aos militares, ela começava a ganhar o mundo empresarial, os gigantes computadores universitários e militares começavam a reduzir cada vez mais para atender a realidade das empresas, tornando ainda mais necessário guardar essa massa de dados de maneira digital. Mas calma! Ainda não surgiu o Banco de Dados por conta destes computadores, o que se fazia era basicamente pegar as fichas e cadastrá-las uma após a outra dentro de um arquivo sequencial, os arquivos eram guardados em fitas magnéticas ou cartões perfurados, meios de armazenamento sequenciais. E se quisesse encontrar um arquivo específico, deveria varrer toda a sequência das fichas cadastradas. Depois das fitas, surgiram os discos, disquetes e HDs, tipos de mecanismos que já armazenavam os dados de maneira direta e não sequencial. Até que os arquivos também evoluíram e através de mecanismos de armazenamento tornou-se possível guardar todos os registros e mantê-los dentro de uma espécie de tabelas, índices, numerações, guardar chaves identificadoras de cada um dos registros. E então, na década de 60, o Departamento de Defesa dos Estados Unidos buscou identificar uma maneira que também fosse segura e inteligente de se armazenar os dados. Para isso criou-se um evento, o CODASYL, um encontro reunindo militares, empresas e universidades para discutir tecnologias. E desse encontrou surgiu a Linguagem Cobol. Além da Linguagem, surgiu nesse evento uma nova tecnologia, o Banco de Dados. E até hoje o Banco de Dados é composto por quatro partes: a base de dados, Sistema Gerenciado de Banco de Dados (SGBD), a Linguagem de Exploração e Programas



Adicionais. Além do Departamento de Defesa dos Estados Unidos, a *International Business Machines* (IBM), igualmente, foi muito valiosa para construção e evolução do Banco de Dados. Inicialmente a IBM propôs a criação de dados hierárquicos, ou seja, dados armazenados por meio de uma hierarquia, havendo dados interligados de maneira bem simplista, chamando-se Modelo Hierárquico. Este modelo evoluiu para o Modelo em Rede, nele os dados não teriam a determinação de quem é superior ou inferior, eles seriam ligados em uma forma de Rede Inteligente. Esses modelos permitiam guardar dados de funcionários, dos clientes, dos serviços, fornecedores, entre outros sem problema nenhum. Todavia, não permitiam criar relacionamentos, ou seja, relacionar um dado ou um conjunto de registros/dados a outro. Surgindo um novo modelo, com dados armazenados e relacionados – o Modelo Relacional, utilizado até hoje para se criar Banco de Dados.

FONTE: <<https://en.wikipedia.org/wiki/CODASYL>>.

---

O termo *Big Data*, ou na sua tradução básica “Grandes Dados”, surgiu no início dos anos 2000 por um analista do *Gartner Group*. Desde então, diversos conceitos vêm sendo apresentados por autores, pesquisadores e organizações. Dentre os mais relevantes encontram-se:

*Big Data*, em geral, é definido como ativos de alto volume, velocidade, variedade de informação que exigem custo-benefício, de formas inovadoras de processamento de informações para maior visibilidade e tomada de decisão (GARTNER GROUPS, 2012).

[...] as tecnologias de *Big Data* descrevem uma nova geração de tecnologias e arquiteturas projetadas para extrair economicamente o valor de volumes muito grandes e de uma variedade de dados, permitindo alta velocidade de captura, descoberta e/ou análise (IDC, 2011).

*Big Data* é o termo utilizado para descrever grandes volumes de dados e que ganha cada vez mais relevância à medida que a sociedade se depara com um aumento sem precedentes no número de informações geradas a cada dia (IBM, 2014).

*Big Data*, são dados cuja escala, distribuição, diversidade e/ou atualidade exigem o uso de nova arquitetura técnica analítica para possibilitar compromissos que liberem novas fontes de valor comercial. Requer nova arquitetura de dados, banco analíticos; novas ferramentas; novos métodos analíticos; integrando múltiplas habilidades em novo papel de cientista de dados (EMC, 2012).

O importante a se observar nessas definições é que não conseguimos desassociar o termo *Big Data* do termo de negócios. E não faria nem sentido ter um volume gigante de informação e não vincular a isso uma necessidade efetivamente prática do dia a dia empresarial, ou corporativo, ou qualquer ambiente que se esteja analisando. A análise está vinculada ao negócio, negócio está vinculado à tomada de decisão. E precisamos prover a todo este mecanismo para tomada de decisão ferramentas adequadas. E o *Big Data* é uma dessas ferramentas que atende diversas demandas, algumas delas:

- Identificar Top formadores de opinião.
- Identificar o que falam da sua marca/produto.
- Conhecer melhor seus clientes/potenciais clientes.
- Conhecer melhor seus concorrentes.
- Melhorar cadeia de distribuição.
- *Active Operations Management* (AOM).
  - o ClickStream ou análise de cliques.
  - o Redes Sociais.
  - o Utilização de sensores (RFID, por exemplo) para *Pervasive Business Intelligence* (BI).
  - o Logística: distribuição de produtos e rotas.
  - o Sensores: manutenção de máquinas e equipamentos.

Além disso, a busca constante e utilização de *Big Data* está trazendo algumas tendências, a saber:

- *Big Data* deixar de ser *Buzzword* e vira realidade, ou seja, chegou e vai ficar por um longo tempo.
- Clientes possuem cada vez mais valor. O que mais as empresas querem é conhecer seu cliente. Apesar daquele velho ditado: “Cliente está em primeiro lugar”, as organizações estão sempre mais preocupadas se há o produto a ser vendido, se ele foi entregue ou se a distribuição está correta. E não estão de fato preocupados se é aquilo que o cliente quer. E é isso que está sendo possível descobrir utilizando *Big Data*. As organizações estão começando a entender o que o cliente quer, olhando de fato para ele, seu comportamento e tendo mecanismos para isso. Não olhando mais apenas para dentro da organização, para o que elas fazem, o que elas produzem, para o serviço que elas prestam.
- Maior utilização de *Big Data* para *Marketing* e o Cientista de Dados poderá estar dentro da empresa. O Cientista de Dados possui um *perfil* que alinha conhecimentos técnicos com os conhecimentos do negócio, falaremos mais adiante neste capítulo sobre os profissionais de *Big Data*.



- *Internet das Coisas versus Web das Coisas*, cada vez mais há integração, as coisas estão interligadas, sensores estão se espalhando pelas residências, com inteligências implementadas, conhecerá um pouco mais sobre *Internet das Coisas* no capítulo 03 deste livro.
- *Extreme Data*, mais dados, mais descobertas sem fim. *Analytics* conquistando espaço definitivo nas organizações. Ainda não há este espaço, há diversas empresas trabalhando com *Business Intelligence* (BI), tomando decisões a partir de *dashboards*. Mas o *analytics* ainda não é realidade em grande parte das empresas, só se tornará realidade através de um apoio maior de *Big Data*.

Nas palavras de Eric Siegel, em seu livro *Predictive Analytics*: “Os dados que coletamos atualmente, nos permitem ver coisas que até pouco tempo atrás eram grandes demais para enxergarmos”.

Nas palavras de Eric Siegel, em seu livro *Predictive Analytics*: “Os dados que coletamos atualmente, nos permitem ver coisas que até pouco tempo atrás eram grandes demais para enxergarmos”.

Já sabemos que coletar dados pode auxiliar fortemente a tomada de decisão das organizações, mas como coletar estes dados?

De uma forma simplória, veja a Figura 2 e imagine que você está construindo uma praça e precisa decidir qual o caminho mais utilizado pelas pessoas para construir uma passagem de concreto. Nesta missão, você precisará prever qual o caminho as pessoas mais usam. Para isso você pode verificar alguns dados, como: quais são os prédios mais visitados ao redor da praça, onde tem ponto de ônibus, comércio, entre outras informações relevantes para sua construção. Sabendo disso, os trajetos mais utilizados podem ser previstos para serem concretados.

FIGURA 2 – ILUSTRAÇÃO PARA COLETA DE DADOS



FONTE: <[http://mediapool.fabrico.com.br/index.php/games\\_laureate-01/15\\_banco-de-praca-cenario-arvores-shutterstock\\_158480780-Converted](http://mediapool.fabrico.com.br/index.php/games_laureate-01/15_banco-de-praca-cenario-arvores-shutterstock_158480780-Converted)>. Acesso em: 4 dez. 2018.

Provavelmente, depois de algum tempo, alguém vai desviar pela grama e deixar marcas de um novo caminho. Encontramos então outra forma de resolver a questão:

não construindo nada na praça, só o gramado. E deixando as pessoas passarem naturalmente por alguns meses, depois de um tempo, basta ver quais as partes mais marcadas pela grama, para se ter uma medida do que as pessoas preferem fazer. Esse processo que você está fazendo é coleta de dados, usar as marcas registradas pela grama para aprender e predizer por onde as pessoas mais passarão.

Outro exemplo sobre coleta de dados, se você já usou a função auto completar do *Google*, já viu como funciona uma das estratégias de coleta e análise de dados. Por mais que ele não possa prever exatamente o que você vai escrever, o *Google* usa automaticamente os termos de busca para aprender o que as pessoas estão mais interessadas em saber.

Contudo, o segredo maior é como coletar um volume enorme de dados. Um caso prático, por meio de quem aparece nas mesmas fotos que você, com reconhecimento facial o *Facebook* sabe com quem e onde você estava com alto grau de precisão. E juntando mais dados de interesses, como o que curtimos, que músicas ouvimos e vídeos a que assistimos, as revelações são mais profundas.

Já se questionou como são posicionadas as compras em um supermercado? Testes com diversas combinações são realizados e são mantidas as posições que mais funcionaram. Foi assim que empresas de cereais aprenderam onde colocar suas embalagens, em prateleiras na altura das crianças, com personagens olhando para os olhos delas, aumentando as vendas. Portanto, lojas podem nos predizer o que comprar antes mesmo de pensarmos nisso. Aquela fralda, de que não sabia que ia precisar, pode chegar muito mais rápido se os navios e trens se comunicarem automaticamente com as empresas de transporte para avisarem que estão a caminho no tempo certo. Tem aviões mandando informações de por onde estão voando e quais as condições climáticas, permitindo voos muito mais seguros e quase independentes dos pilotos. Até o seu carro pode transmitir quais são as condições da estrada em que você está dirigindo e acompanhar os dados do GPS de motoristas que ajudam a predizer os horários e locais que estarão mais congestionados, em parte é isso que você faz quando usa aplicativos para cortar caminhos.

---

“Após pesquisar panela de pressão no *Google*, mulher recebe visita da polícia – Família norte-americana foi surpreendida em casa por oficiais, após buscar por panela de pressão e mochilas.”



FONTE: <<http://g1.globo.com/mundo/noticia/2013/08/familia-recebe-visita-da-policia-apos-busca-por-panela-de-pressao-na-internet.html>>. Acesso em: 4 dez. 2018.

---



Outra pergunta que deve ser sempre levada em consideração ao se pensar em trabalhar com *Big Data* é: Como se preparar tecnologicamente? Para isso alguns critérios devem ser sempre colocados em prática, a saber:

- Análise de valor antes de iniciar o Projeto, pense primeiramente no valor antes do Projeto.
- Saiba o que será analisado antes de achar o dado, o que se quer extrair, o que dá para se fazer, o que será feito com os dados. Exemplo: “Quero analisar tweets buscando análise de sentimentos”, pense isso antes e não depois de se ter os dados.
- Entenda a diferença entre Busca e Análise. São bem diferentes, para fazer uma busca basta um *SELECT* e lhe serão retornados os dados. E a análise? A empresa precisa ter consciência do que é análise. O que é analisar dado? Que dado eu quero analisar? Para que eu preciso analisar este dado? O que isso vai mudar minha empresa?
- Prefira uma infraestrutura de *Big Data* que possa crescer com seu negócio, se vai trabalhar com uma *cloud* privada ou se vai trabalhar com meia infraestrutura em sua própria casa. Pode até começar com algo menor, mas tenha certeza que não será viável por muito tempo. Então, por que não já iniciar com um fornecedor bom, em que confie e que saiba que pode crescer, que terá escabilidade e condições diante desse mundo de constantes mudanças?
- Saiba como relacionar o que for descoberto com *Big Data* com aquilo que há disponível na organização. A sacada não é pegar qualquer dado não estruturado que esteja disponível em qualquer lugar, mas misturar com aquilo que já se possui e enriquecer sua tomada de decisões. Antes de partir para uma aventura em *Big Data*, pense naquilo que você já tem. Será que você tem informações do seu cliente que lhe permitam integrar com a base de dados do *Facebook* e efetivamente poder fazer algo com isso? São questões essenciais que precisam ser respondidas. Será que se você conseguir fazer uma análise de sentimentos, você vai conseguir atingir seu cliente da forma como ele quer? Com a resposta que ele precisa para o problema que apresentou? Você tem estrutura para isso? Você tem como atingir seu cliente? Ou será que seu problema ainda é saber quem é o seu cliente? Pois, quem nunca passou por várias atualizações de dados cadastrais em uma mesma empresa? Nesses casos, para que usar *Big Data*?
- Entenda que será necessário trabalhar com mais de um fornecedor, não há um fornecedor que tenha todas as soluções de que você precise.

- Identifique com clareza os usuários, ou seja, vai trabalhar com determinada informação, mas quem irá utilizar isso? Para que ele irá utilizar esses dados?
- Crie uma métrica que avalie o sucesso (ou não) da iniciativa. O insucesso também deve ser avaliado, o que é ruim e o que não deu certo ainda pode ensinar bastante.
- Permita e facilite a exploração dos dados.
- Invista em Metadados/Semântica, entenda o que é aquele dado. Por exemplo, dentro de uma empresa, podemos solicitar a margem líquida da empresa, mas em cada setor/departamento é atribuída uma nova variável, não há uma definição comum na empresa, como vai ser possível encontrar a margem líquida? Lembre-se, definir o dado e aquilo que se quer alcançar é o básico para quem trabalha com dados.



- **Projeto:** “É um processo único, consistindo de um grupo de atividades coordenadas e controladas com datas para início e término, compreendido para alcance de um objetivo conforme requisitos específicos, incluindo limitações de tempo, custo e recursos” (NBR 10006/ ABNT, 2000).

## 3.1 ANÁLISE DE DADOS

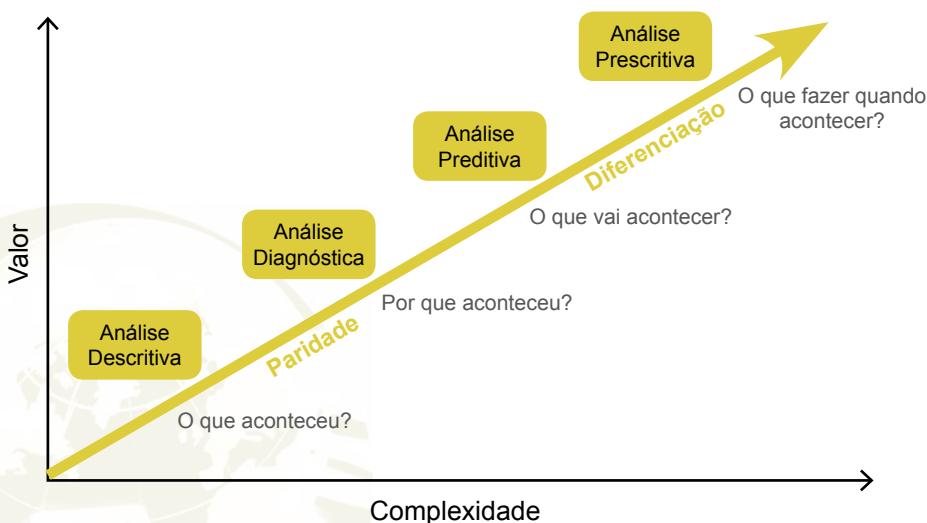
Esperamos que esteja ficando claro que, quando estamos falando de *Big Data*, estamos tratando a capacidade de análise dos dados. E nesse contexto há quatro tipos de análises em *Big Data*, apresentados na Figura 3, que se ressaltam pela usabilidade e potencialidade das suas implicações. Saiba um pouco sobre cada uma delas:

1. **Análise Descritiva (o que aconteceu?):** utilizada para perceber performances passadas e atuais de negócios, para tomada de decisões, categorizando, caracterizando, consolidando e classificando os dados em informação útil. Utilizam-se métricas e técnicas de estatística para gerar gráficos e relatórios sobre orçamentos, vendas, receitas, Processamento Analítico *Online* (OLAP), painéis/*scorecards* e visualização de dados, os quais podem ser tidos como exemplos de análise descritivas (WATSON, 2015). Através desta análise, uma organização pode avaliar dados sobre a queda das vendas de um produto ou faturamento da empresa nas últimas semanas ou meses, por exemplo.



2. **Análise Diagnóstica (por que aconteceu?)**: se preocupa com os dados passados, é utilizada com interesse em saber o motivo por que determinados eventos aconteceram na análise descritiva, na tentativa de minimizar eventuais problemas presentes. Corresponde a um tipo mais avançado de análise, em que são utilizadas técnicas, como mineração, correlações, detalhamento e descoberta de dados, em busca da descoberta das causas do problema.
3. **Análise Preditiva (o que vai acontecer?)**: avalia performances passadas, detectando padrões e relações entre os dados futuros. O objetivo é “prever” o futuro, por meio de mineração de dados, dados estatísticos e históricos. São utilizadas também para este modelo técnicas como *Machine Learning* e Inteligência Artificial. Um exemplo seria a previsão do faturamento para o próximo trimestre ou a quantidade de chamadas que poderão ocorrer em uma central de *call center* para próxima campanha publicitária.
4. **Análise Prescritiva (o que fazer se for acontecer?)**: utiliza a otimização de forma a identificar as melhores alternativas e maximizar ou minimizar algum objetivo (GANDOMI; HAIDER, 2015). Utilizam-se ferramentas estatísticas tanto de análise descritiva quanto preditiva, alinhadas à gestão de negócios, para gerar recomendações automáticas buscando aperfeiçoar estratégias. Basicamente é uma forma de definir qual escolha será mais efetiva em determinada situação.

FIGURA 3 – ANÁLISES DE DADOS



FONTE: <<https://thoughtworksinc.github.io/guia-de-desenvolvimento-tecnico/topics/Analytics.html>>. Acesso em: 4 dez. 2018.

A análise de dados, apesar de seus benefícios, apresenta desafios consideráveis, um deles é o de as organizações terem dificuldades para encontrar profissionais com o conjunto de habilidades específicas para sua real aplicação. Além disso, quando se encontra profissionais qualificados, estes apresentam dificuldades na comunicação junto aos gestores sobre as ideias referentes ao negócio para melhorias na tomada de decisão. Mais informações sobre análise de dados ou *Data Analytics* serão abordadas na disciplina *Big Data Analytics* e a Tomada de Decisão.

---

Imagine as respectivas situações:

- a) Uma organização deseja produzir especificações de empreendimentos imobiliários ou abrangência analítica de uma determinada situação.
- b) Além disso, a empresa também deseja determinar fatores com indicativos de sucesso em episódios já realizados em uma abrangência analítica de uma situação.
- c) Empresários desta mesma empresa, preocupados que são, desejam avaliar o resultado histórico provável para um evento ou probabilidade de uma situação se repetir.
- d) Já o setor financeiro da empresa gostaria de determinar as especificações de empreendimentos imobiliários ou a competitividade de uma situação.
- e) Por fim, gostariam de determinar o resultado provável futuro para um evento ou probabilidade de ocorrer uma situação.



Quais dessas opções representa a maneira como a técnica Análise Preditiva ajudará essa organização?

---

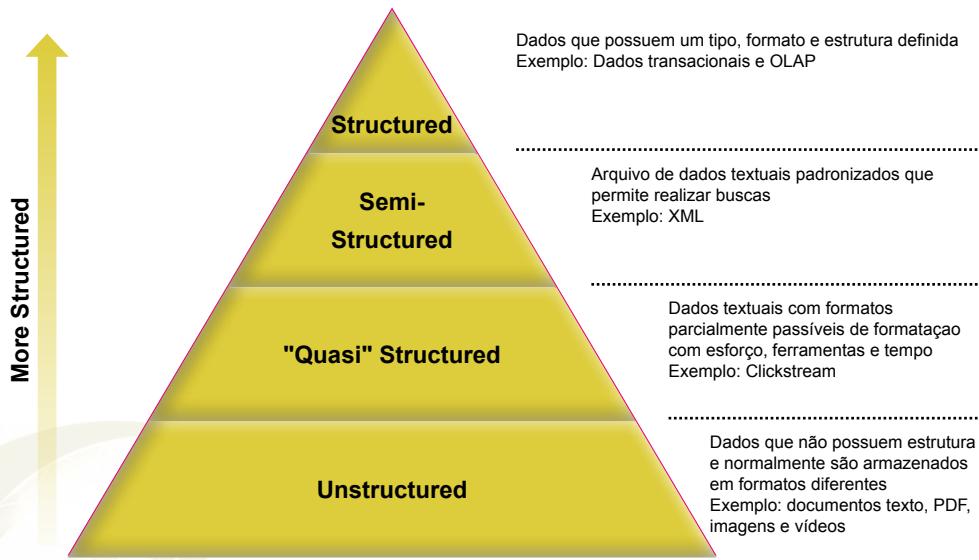
## 3.2 FONTES DE DADOS

Os dados que alimentam a ideia de *Big Data* podem provir de diversas fontes de dados. Na verdade, qual a fonte que não pode ter *Big Data*? Pois hoje na *Internet* encontramos um grande volume de dados com conteúdos relacionados a educação, ciência, varejo, indústria do entretenimento, governo, social, finanças, transporte, saúde. Todos esses dados são fontes de *Big Data*. Mas entender a diferença entre eles proporcionará uma melhor gestão do seu Projeto, portanto, atenção quanto às classificações dos dados a seguir e apresentados na Figura 4:



- **Estruturados:** dados armazenados em Banco de Dados tradicionais ou relacionais organizados em tabelas. São tabelas com informações contidas em linhas e colunas, na maior parte números, valores numéricos ou caracteres. Possui esquema fixo, formato bem definido (inteiro, string, data), conhecimento prévio da estrutura de dados, simplicidade para relacionar as informações (relacionar usuário, com entrega, com pedido), dificuldade para alterar o modelo. Grande parte das empresas trabalha com dados estruturados há anos.
- **Semiestruturados:** podem seguir diversos padrões, de forma heterogênea, dados embaralhados pela Web em arquivos HTML, XML, ou em Banco de Dados não relacionais.
- **Não-estruturados:** mescla de dados oriundos de várias fontes distintas, como vídeo, texto, áudio, imagens, XML, entre outros. Sem tipo predefinido (o dado vai sendo modelado conforme o tempo vai passando, com os campos adaptados), não possui estrutura regular, pouco ou nenhum controle sobre a forma, manipulação mais simplificada e facilidade de alteração.

FIGURA 4 – ESTRUTURA DOS DADOS CORPORATIVOS



Pesquisas apontam que cerca de 80% dos dados das empresas são dados não-estruturados e que precisam ser tratados, mas você pode estar se perguntando onde estão todas estas fontes de dados? Existe uma lista vasta para aplicações de *Big Data*, segue uma pequena amostragem:

- Monitoramento em redes sociais, o que monitorar em uma rede social em que milhares ou bilhões de usuários simultaneamente estão trocando informações.
- Netflix (recomendação de filmes): utiliza técnicas de *Big Data* para recolher dados dos milhões de usuários e descobrir que tipo de informação recomendar para cada um dos usuários.
- *Web Analytics* (sites de e-commerce).
- Dados provenientes de múltiplos sensores em sistemas e transporte.
- Análises de dados financeiros (para evitar fraudes), sistemas de cartão de crédito alinhados a um pouco de inteligência artificial, conseguem, por exemplo, identificar quando alguém faz uma transação de crédito que não é típica do usuário dono do cartão, podendo bloquear a transação e informar ao dono para verificar de fato o que está acontecendo.
- Análises de dados médicos.
- Análises de dados trafegados em redes.
- Publicidade e propaganda personalizadas.
- Uso de telefones celulares.
- *Tags RFID (Radio Frequency Identification)*.
- Informações sobre o tempo.
- Informações sobre trânsito e modelos de tráfego.

São inúmeras as aplicações em que *Big Data* que vem auxiliando na resolução de problemas, que levariam anos, meses ou até mesmo não seriam possíveis de analisar. Fora as grandes empresas que empregam *Big Data*, podemos citar: a IBM, Google, SAP, Cloudera, Teradata, New Relic, Salesforce, Microsoft, Tableau Software e tantas outras. São diversas empresas que vem desenvolvendo ou utilizando sistemas na área *Big Data*. Sugiro que verifique no site dessas empresas, lá você encontrará bastantes informações, muitos *papers*, estudo de caso e softwares para que você possa analisar e complementar seus conhecimentos sobre *Big Data*.

O grande desafio é descobrir como gerenciar e analisar o volume de dados, que cresce infinitamente todos os dias, mais que isso, entendê-los e criar ferramentas para gerar mais experiência, produtividade, consumo e novos serviços.

---

Pesquise e indique aplicações de *Big Data* por empresas brasileiras que trouxeram resultados significativos. Em seguida classifique quais tipos de análises foram realizadas e quais fontes/tipos de dados existiam.

---





### 3.2.1 Qualidade de dados

A integração de bases de dados diferentes pode apresentar ruídos, informações ambíguas, conflitantes ou mesmo errôneas. Portanto, a qualidade do processo de análise dos dados dependerá da qualidade dos dados armazenados nas bases. Algumas características são importantes para se garantir a qualidade dos dados (LOSHIN, 2010):

- **Integridade (completeness) Validez (validity)**: trata dos conjuntos de dados que refletem corretamente a realidade representada pela fonte de dados que são consistentes entre si e que, portanto, são dados válidos.
- **Granularidade (uniqueness)**: seleção dos dados no seu nível atômico, que trata de exibir a informação no seu nível mais detalhado (bairro de domicílio de um cliente) ao final da análise, a qual pode ter sido iniciada em uma visão agregada (clientes por estados ou municípios).
- **Tempestividade (timeliness)**: agilidade na obtenção das informações. A realização de qualquer análise que não tenha sido previamente definida, a qualquer tempo (oferecimento de informações do passado e na formação de séries históricas de valor inestimável para as previsões), pelo Cientista de Dados.
- **Precisão (accuracy)**: refere-se à precisão com que os dados são descritos por meio dos metadados.
- **Consistência (consistency)**: importância de a informação disponibilizada não conter erros.
- **Flexibilidade (flexibility)**: facilidade de elaboração de análises inéditas, manuseio e formatação de acordo com as necessidade e preferências.



Alguns mitos sobre *Big Data* e *Oportunidades* podem ser verificados no Livro de Rosângela Marquesone, 2017, disponível em: <[encurtador.com.br/nxL48](http://encurtador.com.br/nxL48)>. Leia os respectivos tópicos e crie um Mapa Mental com palavras-chaves relacionando os tópicos e itens vistos até o momento neste primeiro capítulo.

### 3.3 Os Vs de *Big Data*

Conforme os tópicos anteriores, podemos caracterizar *Big Data* quanto aos aspectos dos dados, são os conhecidos Vs de *Big Data*, a saber:

1. **Volume de Dados:** refere-se à enorme quantidade de dados disponível. Estima-se que, até 2020, existam cerca de 35 (trinta e cinco) ZB (*zettabytes*) de dados armazenados no mundo. Um ZB equivale a  $10^{21}$  bytes, ou 01 (um) bilhão de *terabytes*. Imagine que um HD de um computador costuma ter em média a capacidade de um *terabyte*. De acordo com o IDC (2011), a informação no mundo dobra a cada dois anos. É um grande motivador para se criar esse novo universo de tecnologias que são capazes de processar este grande volume de dados. No passado, armazenar tamanha quantidade de informações era um problema diante os recursos e custos computacionais escassos. Outras estatísticas que representam volume de dados (MARQUESONE, 2017):

- a cada segundo, certa de 40.000 buscas são realizadas no *Google*;
- a rede social *Instagram* recebe em média 80 milhões de fotos por dia;
- em 2016 o *Facebook* contabilizou uma média de 1.13 bilhão de usuários, 2.5 bilhões de compartilhamentos e 2.7 bilhões de “curtidas” por dia.

Uma questão frequente relacionada ao volume de dados é saber quando um determinado conjunto de dados pode ser considerado *Big Data*: somente se possuir *petabytes* de dados? Não necessariamente, o tamanho dos dados é algo relativo ao se tratar de *Big Data*. Uma clínica médica pode precisar de soluções de *Big Data* para visualizar imagens de 30 *gigabytes* de dados, por exemplo (MARQUESONE, 2017).

2. **Variedade de Dados:** os dados incluem não apenas dados transacionais (Banco de Dados comuns/ estruturados), mas também oriundos de outras fontes: páginas *web*, índices de pesquisa, arquivos *log*, fóruns (tente imaginar um fórum, com milhares de entradas e pessoas tratando diferentes assuntos), mídias sociais (imagine o volume de dados armazenados por dia ou por hora pelo *Facebook* e *Instagram*), *e-mails*, dados de sensores variados, IoT, áudio e vídeo, dados estes heterogêneos, modelados de formas diferentes. Nas organizações, os sistemas de software (ou hardware dedicado) comumente usam os mais diferentes tipos de dispositivos de armazenamento, sistemas de arquivos, tipos de codificação, estrutura de dados, entre outros (MANYIKA et al., 2011). Idealize que, se quisermos fazer uma agregação das informações relacionadas à renda por estado, teremos que ir a cada estado, fazer



a pesquisa, recuperar esse conjunto de dados, integrá-los, para poder disponibilizar as informações. Não é apenas a questão de tipos de dados difíceis de lidar e trabalhar, que também se entende como variedade, mas a variedade das origens, das fontes de dados. Os sistemas tradicionais não conseguem armazenar, processar e entender essa vasta gama de dados. Assim, deve-se utilizar novas tecnologias, algoritmos e técnicas para realizar a análise desses dados, tanto estruturados quanto não-estruturados, em conjunto.

3. **Velocidade dos Dados:** os dados são gerados em grande velocidade, definimos essa velocidade de acordo com o quanto rápido os dados são resgatados, armazenados e recuperados. Basicamente, falamos em taxa de fluxo de dados quando nos referimos a sua velocidade. Assim, o fluxo de (geração e transmissão) de dados pode se tornar tão elevado que os sistemas tradicionais de análise não conseguem manipulá-los. *Big Data* tem suas técnicas específicas para tratamento dessas informações. Uma prática do aspecto velocidade é o armazenamento das interações de um consumidor em um site de vendas de calçados, que armazena os “cliques” do mouse do usuário que navega pelos seus produtos e ofertas. Essa interação fornece uma grande quantidade de dados, que serão empregados para realimentar o site, sugerindo outros produtos ao usuário. Esse processo precisa ser rápido o suficiente para que o consumidor veja o novo produto, enquanto navega na página do item que o levou a consultar o site.

Vale destacar que estes são os três principais V's de *Big Data*, representados pela Figura 5. Mas ainda há os V's veracidade e valor.

FIGURA 5 – OS 3 VS DE BIG DATA



FONTE: Marquesone (2017)

1. **Veracidade:** refere-se à confiabilidade dos dados, que devem possuir características como: qualidade e consistência, origem conhecida/fonte dos dados, serem verdadeiros, e não fabricados/ oriundos de opinião. E verificar se são internos ou externos à organização. Uma empresa que vai aplicar técnicas de *Big Data* vai coletar dados internos ou externos também? São dados que estão dentro da validade ou vigência ou são dados desatualizados?
2. **Valor:** valor é uma aplicação do *Big Data* que permite aumentar receita, identificar novas oportunidades de operações, economizar custos, melhorar a qualidade do produto e a satisfação do cliente, garantindo melhores resultados e resolução de problemas. Se não for capaz de gerar valor, não valerá a pena.

Compreender os Vs de *Big Data* é fundamental para compreensão e disseminação das posteriores informações deste livro e aplicação de *Big Data*.

---

Em relação ao *Big Data* e às técnicas mais tradicionais como *Data Mining*, *Business Intelligence* (BI), como diferenciar? Essas técnicas anteriores não servem mais? Devemos utilizar o que está na “moda”? Como saber se é ou não *Big Data*?



É importante destacar que na área de Tecnologia da Informação, é notório que algumas coisas se repetem, só trocam o nome, mas se assemelham ao que já vem sendo desenvolvido há um tempo. Ou seja, o que temos é sempre uma evolução. Se pegarmos as ferramentas de análise de dados ou exploração de dados, há um conjunto de ferramentas que irá atende-lo dentro do BI e também de *Big Data* perfeitamente. Outro exemplo, ao se tratar sobre enriquecimento de dados, é um assunto já antigo em BI, mas o que é novo em *Big Data* é o problema da variedade. O BI nada mais é que um integrador de base de dados para que se possam tomar decisões. Um dado de um ERP, de um sistema transacional, de relacionamento com o cliente – CRM, uma outra informação em um sistema de suprimentos. Mas com quantas bases eu posso, eu consigo fazer isso dentro do ambiente tradicional de BI? 20, 30, 40, 50 com uma excelente equipe. Com *Big Data* é possível integrar um número de base muito maior. Outro critério importante de diferenciação do *Big Data* é a velocidade, ou seja, pode ser até que no BI e na mineração a variedade e complexidade do dado sejam os mesmos que o *Big Data*, mas ambos precisam da síntese dos dados para fazer sentido e ela é fornecida com uma velocidade muito maior em *Big Data*.



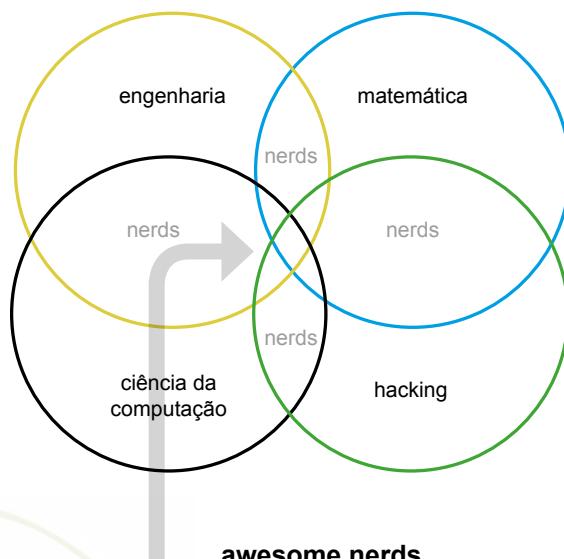
## 3.4 PROFISSIONAIS *BIG DATA*

Diante de tantos novos conceitos, novas profissões também estão surgindo. A maioria dos especialistas nomeava o profissional Cientista dos Dados com conhecimentos técnicos em estatística, *NoSQL*, *cloud computing*, mineração de dados (*data mining*), álgebra relacional, modelagem multidimensional, *MapReduce*, verticalização, entre outros. Tantos saberes vinculados a um único profissional que seria somente possível ao *Watson*, o megacomputador da IBM, executar tantas funções.

Mas atualmente o cenário indica um *perfil* profissional um pouco diferente desta perspectiva inicial. Na verdade, o mercado precisa de profissionais multidisciplinares, mas com conhecimento básico em tecnologia, modelos, conceitos, infraestrutura e negócios, observe a Figura 6.

FIGURA 6 – CONHECIMENTOS DO CIENTISTA DE DADOS

### Data science?



FONTE: A autora, traduzido de Hilary Mason.



Você quer aprender mais sobre *Data Science* e de forma gratuita, veja alguns sites interessantes com treinamentos gratuitos:

**Coursera:** <<http://www.coursera.org/specialization/jhudatascience/1>>.

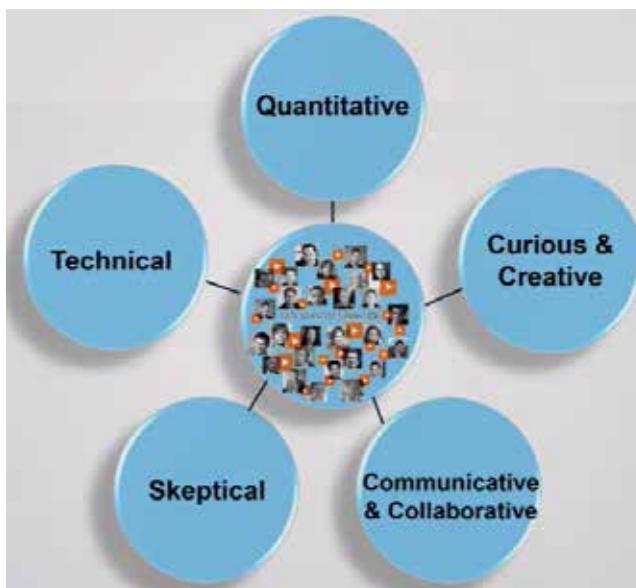
**Udemy:** <<https://www.udemy.com/courses/search/?q=data+science>>.

**MIT OpenCourseWare:** <<http://ocw.mit.edu/index.htm>>.

**Harvard Open Course:** <<http://extension.harvard.edu/courses/data-science>>.

O Cientista de Dados está entre *nerds*, engenheiros, *hackers*, matemáticos e Cientistas da Computação, esse é o *perfil* desejado para se trabalhar com *Big Data*. E tem mais, observe a Figura 7:

FIGURA 7 – ÁREAS DO CIENTISTA DE DADOS



FONTE: EMC (2012)

Segundo pesquisas dos Estados Unidos, um Cientista de Dados ganha em média 120 mil dólares por ano. E um analista, uma média de 85 mil.





O Cientista de Dados deve conhecer de tecnologia, de matemática (quantitative), ter curiosidade e criatividade. Um cientista curioso é aquele que não se conforma com qualquer coisa, qualquer informação. Agora conseguir um profissional com conhecimentos técnicos, quantitativos, curioso e colaborativo/ comunicativo é bem difícil. E ainda precisa ser cético, não podendo acreditar em tudo o que se ouve, deve ser pesquisador, correr atrás das melhores informações.

Ao mesmo tempo, este Cientista de Dados deve saber atuar como um líder, juntamente com profissionais especializados no que fazem, ou seja, é preciso de um Cientista de Dados junto com sua equipe para implementar soluções para os problemas apresentados, dentro dos custos, prazos e escopo determinado no Projeto. Mas que tipos de profissionais irão compor esta equipe? Conheça alguns dos profissionais aliados ao Cientista de Dados (AMARAL, 2016):

- **Data Engineer:** garante o fluxo contínuo entre as fontes de dados e os sistemas de armazenamento e processamento. São responsáveis por definir a arquitetura dos dados. Principais atribuições: transformação de dados, processamento paralelo, integração de sistemas heterogêneos, construir aplicações escaláveis, análise e performance. E conhecimentos em programação avançada, *design* de Banco de Dados e otimização.
- **Equipe de Extração:** geralmente são usuários que também atuam como DBAs (*Database Administrator*), programadores e especialistas em ETL. A principal função destinada à equipe de extração é checar se os dados extraídos são, de fato, os esperados, se estão completos, íntegros e atualizados, podendo ainda ter a incumbência de carregar dados para sistemas de arquivos distribuídos, como *Hadoop Distributed File System* (HDFS).
- **DBA:** função já desenvolvida em Banco de Dados Relacionais e Multidimensionais, mas que agora foi expandida para os Banco de Dados Não Relacionais e sistemas de arquivos distribuídos. Podendo prestar ainda auxílio na coleta de metadados, entendimento de estruturas, rotinas de replicação, integração, entre outros.
- **Programador:** nem sempre são utilizados softwares especializados em análise de dados e algumas empresas preferem implementar a análise programando *Stored Procedures* diretamente em gerenciadores de Banco de Dados ou em uma Linguagem de Programação como *Java* ou *Python*, sendo utilizadas ferramentas prontas e uma programação mais caseira.
- **Especialista no assunto:** competência mais importante, em que se conhecem as regras, a legislação envolvida e as exceções. São geralmente os profissionais mais difíceis de serem encontrados.

- **Estatístico e/ou Minerador de Dados:** testes de hipóteses, construção de modelos preditivos ou elementos de visualização são sempre necessários.
- **Especialistas em Ferramentas Específicas:** diversas ferramentas entre as fases do processo da criação de um produto são utilizadas, como ferramentas de extração, de visualização. E para cada uma dessas ferramentas deve-se ter um técnico especializado ou capacitarlo para o seu uso.
- **Arquiteto:** este papel é fundamental para definir padrões, *frameworks* e protocolos. Também deve propor a arquitetura concisa para o Projeto, desde CPUs, storages, licenças, entre outros.
- **Analistas de Negócios:** necessário para eliciar requisitos e especificar o escopo do projeto, normalmente atua com um Gerente de Projeto (GP).
- **Designer:** apresentar uma forma sofisticada de visualização é importantíssimo. Um especialista em visualização de dados ou até mesmo um designer que produza artefatos com alta qualidade visual deve sempre compor a equipe.

Destaca-se que fatores como questões técnicas, de negócio e ativos empresariais podem elencar particularidades de especialistas para atuação em cada projeto ou a cada organização. Não há um padrão determinado de profissionais atuantes em *Big Data*, alguns papéis e funções já estão obtendo seus destaques no mercado de trabalho como futuras profissões.

### Taxonomia dos Cientistas de Dados



**Taxonomy of Data Scientists**

Posted by Vincent Granville on November 20, 2013 at 8:00pm [View Blog](#)

This is a first attempt at classifying data scientists. I invite you to produce a more comprehensive, better solution.

	Analytics	Big Data	Data Mining	Machine Learning
DJ Patil	0,34	0,09	0,30	0,26
Dean Abbott	0,24	0,02	0,46	0,27
Eric Colson	0,28	0,38	0,27	0,08
Gregory Platetsky-Shapiro	0,21	0,24	0,40	0,16
Kirk Borne (1)	0,00	0,15	0,45	0,39
Marck Valsman	0,28	0,17	0,42	0,13
Milind Bhandarkar	0,09	0,54	0,12	0,25
Monica Rogati	0,09	0,17	0,31	0,43
Simon (Ximeng) Xhang (2)	0,53	0,14	0,32	0,00
Vincent Granville	0,38	0,18	0,34	0,10



Segundo a pesquisa de Vicent Granville (2013), realizada no *Linkedin* com os principais cientistas de dados, os perfis apontavam em comum as respectivas características: análise de dados, conhecer *Big Data*, Mineração de Dados e Aprendizagem de Máquina.

Ainda em sua pesquisa, Granville após o mapeamento, explorou um pouco mais e fez um *ranking* de quantas pessoas possuíam determinadas habilidades, conforme a tabela.

Skill	Association		
	# People	with DS	Percent
Data Mining	9	554,81	21,6%
Machine Learning	5	302,63	11,8%
Analytics	6	344,38	13,4%
Big Data	7	293,53	11,4%
Predictive Analytics	34	187,50	7,3%
Data Analysis	2	186,22	7,2%
Predictive Modeling	2	144,97	5,6%
Hadoop	1	65,86	2,6%
Text Mining	1	51,22	2,0%
Statistics	1	81,39	3,2%
Natural Language Processing	1	46,21	1,8%
Start-Ups	1	63,94	2,5%
Algorithms	1	64,34	2,5%
Distributed Systems	1	34,21	1,3%
Map Reduce	1	31,62	1,2%
Data Warehousing	1	24,82	1,0%
Business Intelligence	1	29,60	1,2%
SQL	1	25,46	1,0%
R	1	18,44	0,7%
Scalability	1	22,14	0,9%

### 3.5 BANCOS DE DADOS NÃO RELACIONAIS

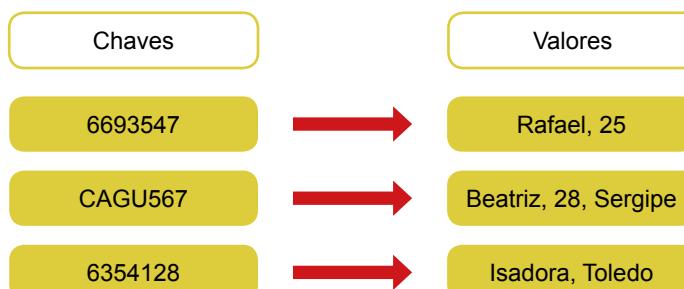
Já faz anos que conceitos como Banco de Dados, a importância de projetar Banco de Dados e a modelagem de dados são considerados fundamentais para Projetos, em especial Projetos de software. Sem o planejamento e análise precisa,

pode-se criar um Banco de Dados que omita alguns dados exigidos ou inconsistências em uma relação ao contexto de informações que ele deve refletir. Nesses Bancos de Dados (BDs), conhecidos como Banco de Dados Relacionais, as informações são armazenadas, manipuladas e recuperadas somente por meio de tabelas estruturadas. Apesar das suas vantagens como garantia da integridade dos dados, validação e verificação dos dados, controle de transações e consultas. Algumas desvantagens estão tendo mais destaque, como a dificuldade em conciliar o modelo com a demanda por escalabilidade e de organizar dados não estruturados em sistemas distribuídos.

Com a avalanche de dados, uma nova alternativa vem surgindo para se trabalhar com informações em grande escala e com dados não estruturados, o Banco de Dados Não Relacional, no qual grandes empresas do mercado elegem a escalabilidade mais necessária do que a confiabilidade e a consistência proporcionadas pelos BDs relacionais. Os Banco de Dados Não Relacionais não utilizam tabelas e ao invés disso, fazem o uso de chaves de identificação, em que os dados podem ser encontrados através dessas chaves (GYORÖDI et al., 2015). Conheça um pouco sobre os principais tipos de Banco de Dados Não Relacionais:

- **Banco de Dados chave-valor:** estes BDs mapeiam uma chave para um valor ou um conjunto de valores, constituindo chaves únicas e atômicas, classificadas como chaves indestrutíveis. São empregadas para consultar as entradas nos BDs de armazenamento chave-valor, observe a representação pela Figura 8:

FIGURA 8 – BANCO DE DADOS NÃO RELACIONAL CHAVE-VALOR



FONTE: A autora

Diante seu modelo de dados simplório, a *Application Programming Interface* (API) de armazenamento de chave-valor fornece apenas operações baseadas em *put*, *get* e *delete*. Caso seja preciso funcionalidades de consulta adicionais, é necessário implementar na camada de aplicativo, o que pode gerar complexidade e penalidades de desempenho. Contudo, não é aconselhável utilizar armazenamentos de chave-valor quando for necessário realizar consultas mais complexas (HECHT; JABLONSKI, 2011).



- **Banco de Dados orientado a documentos:** é uma opção aos BDs relacionais utilizada para armazenar dados semiestruturados existentes na forma de XML, Java Script Object Notation (JSON) ou em outros formatos parecidos. Cada documento pode ser comparado a uma linha de um BD relacional contendo as informações relacionadas ao documento. Apesar de o BD apresentar um *design* livre de esquema, os registros armazenados são semiestruturados em forma de hierarquia (SRIVASTAVA et al., 2015). O modelo ainda oferece uma API mais completa, com consultas de intervalos em valores, índices secundários, consulta de documentos alinhados e operações “e”, “ou”, “entre” (HECHT; JABLONSKI, 2011).
- **Banco de Dados orientado a colunas:** neste BD, conhecido como família de colunas, uma coluna inteira de uma tabela é armazenada em conjunto e mapeada para uma única chave, como exemplificado na Figura 9.

FIGURA 9 – BANCO DE DADOS NÃO RELACIONAL ORIENTADO A COLUNA.

Chaves	Colunas		
	Nome	Idade	Estado
Ubá	Rafael	25	Minas Gerais
São Luís	Quelita	29	Maranhão
Aracaju	Fernanda	34	Sergipe

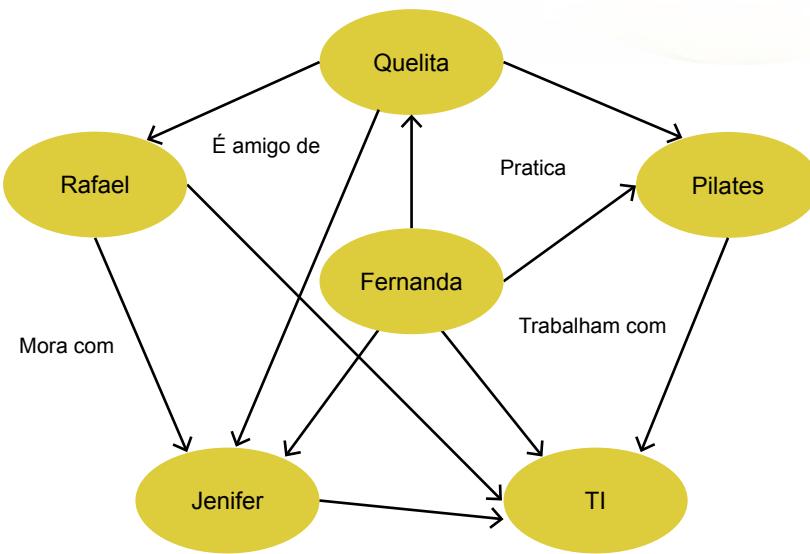
FONTE: A autora

Com todas as colunas possuindo índices, torna-se possível pesquisar apenas parte da tabela, além disso, uma coluna pode ter hierarquias de colunas dentro dela, tornando-se uma “super coluna”, fornecendo pesquisas fáceis e acesso rápido, o que evita gastos desnecessários para identificar a chave individual de um registro (SRIVASTAVA et al., 2015). Entretanto, apenas algumas consultas de intervalo e operações “in”, “e/ou” e expressão regular, são fornecidos se forem aplicadas em chaves de linha ou valores indexados.

- **Banco de Dados orientado a grafos:** são mais ajustados para percorrer e pesquisar aplicativos, como desvendar *links* relacionados ao *LinkedIn*, buscar amigos no *Facebook*. Dando mais relação entre itens e dados e não dados. São otimizados para percorrer rápido e fazer uso eficiente de algoritmos de grafos, como o caminho mais curto (SRIVASTAVA et al., 2015). Sendo definido como uma estrutura contida em nós, arestas e propriedades para representar e armazenar dados, veja a Figura 10. Os nós representam registros na analogia *Relational Database Management Systems* (RDBMS). Bordas ou relações são derivadas

de alguma coluna predefinida em um nó, permitindo o gerenciamento complexo de relacionamentos muitos-para-muitos que são difíceis de serem absorvidos por outros paradigmas de BD. Nesses BDs não são necessárias operações como junções em Sistemas Gerenciadores de Banco de Dados Relacionais (SGBDR), já que são altamente escaláveis e os dados *ad hoc* podem ser mais facilmente gerenciados com um BD baseado em grafo (JAYATHILAKE et al., 2012). SPARQL é uma linguagem de consulta popular, declarativa com uma sintaxe simples e correspondente ao padrão de grafos (HECHT; JABLONSKI, 2011).

FIGURA 10 – BANCO DE DADOS NÃO RELACIONAL ORIENTADO A GRAFOS.



FONTE: A autora.

Apesar de a teoria de grafos já existir há mais de dois séculos, o modelo de grafos dentro do paradigma de Banco de Dados ainda é novidade, considerado imaturo, mas já mostrando ser bem funcional em modelagens complexas para processamento, armazenamento e manipulação de grande volume de dados.

Cada modelo possui suas especificidades, vantagens e desvantagens, que devem ser verificadas para aplicação adequada às particularidades de cada projeto. Destaca-se também que o modelo de BD Não Relacional não visa substituir totalmente o modelo relacional, apenas agregar valor e prover resultados mais eficientes diante as necessidades atuais.



Há diversos sistemas de armazenamento para cada tipo de Banco de Dados Não Relacional, por exemplo, o *Redis* um sistema de armazenamento do tipo valor-chave mais rápido em operações na memória. *Mongo DB* é um BD de armazenamento de documentos que armazena dados semiestruturados escritos no formato *Binary JSON* (*BSON*). Já *Cassandra* é um BD *Open Source* orientado a colunas desenvolvido por *Apache* e o *Neo4J* é baseado em grafos altamente escalável, construído para alavancar não apenas dados, mas também seus relacionamentos.

Aprofunde seus conhecimentos e crie uma Tabela apontando alguns recursos que estes sistemas de armazenamento incluem e quando são mais recomendados ou adequados. Aproveite e acrescente quais empresas utilizam esses sistemas.

---

## 4 DATA SCIENCE E MODELOS PREDITIVOS

Apesar de ser uma expressão que vem sendo usada desde os anos de 1960, a Ciência dos Dados ou *Data Science* é associada equivocadamente apenas ao processo de análise dos dados, com o uso de estatística, aprendizado de máquina ou aplicações de filtro para produzir informação e conhecimento. Mas na verdade a Ciência de Dados é composta por várias outras ciências, modelos, tecnologias, processos e procedimentos relacionados ao dado.

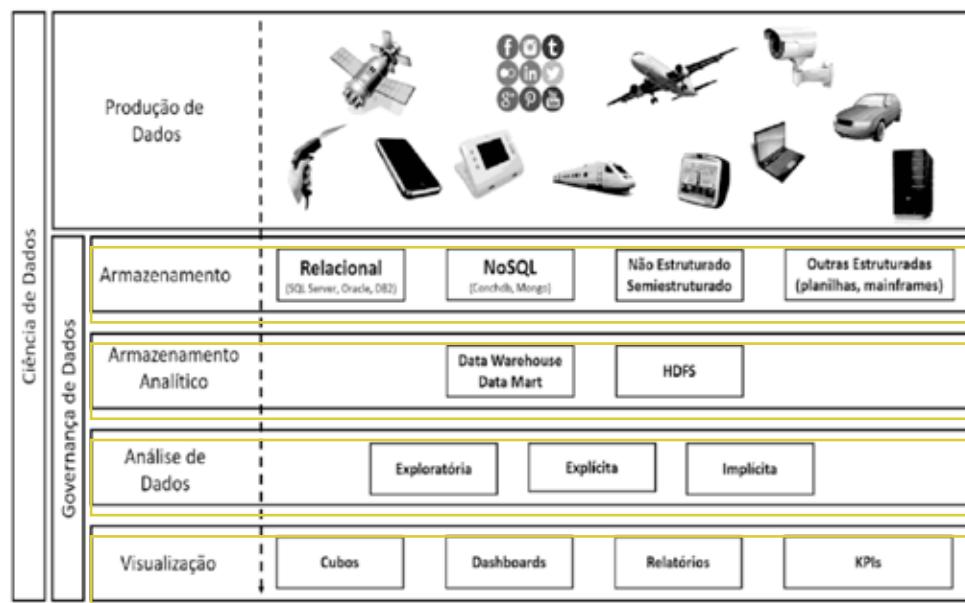
É uma ciência um pouco difícil de se definir, diante de sua natureza multidisciplinar e pelas múltiplas áreas da ciência a que ela se aplica. Portanto, para melhor compreensão, vamos a um bom exemplo: o filme estreado pelo ator Brad Pit, *Um homem que mudou o jogo*. É uma história real da aplicação de *Data Science* a um dos esportes mais amados nos EUA, mostrando como um gerente de esportes de um pequeno time de *baseball* americano desafiou o sistema, o conhecimento convencional do *baseball*, e foi forçado a construir sua equipe com poucos recursos financeiros. No começo, o projeto foi “boicotado” pelo próprio técnico do time, que não aceitava as ideias modernas das previsões geradas pelos modelos matemáticos feitas por um Cientista de Dados. E na primeira partida a equipe de *baseball* jogou bem mal, permitindo que todos os críticos dentro e fora da equipe afirmassem que o novo método era um verdadeiro fracasso. Mas, após a adesão e confiança nas previsões feitas, *Data Science* fez com que um dos

times de *baseball* com menor orçamento da liga profissional dos EUA disputasse os *playoffs* do campeonato nacional. E esse fato mudou totalmente a fase de contratação de todos os times americanos, o pequeno time venceu nada mais nada menos que 20 partidas, batendo recordes que não eram vencidos desde 1927. A chave para esse sucesso foi começar a utilizar dados e estatísticas de desempenho dos jogadores em campo e com isso achar jogadores mais baratos e mais preparados. E por meio de *Data Science*, os insights dos dados de todo o time previram uma equipe altamente competitiva e de baixo custo.

Então o que é *Data Science*? Basicamente estamos falando sobre técnicas, processos, alguns métodos científicos. Sobretudo, com a finalidade de extrair conhecimentos ou o que a gente chama de insight, percepções, revelações de informações importantes a partir dos dados. *Data Science* é uma ciência que pode ser aprendida e aplicada por qualquer um para obter percepções de fenômenos representados por dados, extraíndo valores altamente significativos.

Agora a pergunta: você tem dados? Eles precisam ser manipulados? Como empregar *Data Science* aos seus projetos? Vejamos um panorama da Ciência de Dados conforme a Figura 11:

FIGURA 11 – PANORAMA DA CIÊNCIA DE DADOS



FONTE: Amaral (2016).

A Figura 11 resume grande parte dos conceitos já foram introduzidos neste primeiro capítulo, como a Produção de Dados, o Armazenamento, a Análise



de Dados e Visualização. A Ciência dos Dados se resume, portanto, como os processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida, desde a produção, armazenamento, transformação, análise e descarte (AMARAL, 2016).

## 4.1 ÁREAS DE CONHECIMENTO DE *DATA SCIENCE*

Como abordamos anteriormente *Data Science* é uma mistura de áreas, e a primeira grande área é a Estatística. Alguns estudiosos até mencionam que *Data Science* é um nome moderninho para Estatística, outros dizem que Cientistas de Dados são estatísticos que vivem em São Francisco (EUA), mas na verdade estatística se define como:

“Estatística: é a ciência da coleta, análise, interpretação, apresentação e organização de dados” (SILVA, 2011).

Haja vista que a palavra dados e ciência aparecem na definição de estatística, você deve pensar que *Data Science* é de fato uma nova roupagem para estatística. Vejamos a definição de *Data Science* para melhor comparação:

“*Data Science*: é o estudo científico da criação, validação e transformação de dados para criar significados.” (JOSH WILLS)

Assim, as principais características que diferenciam estatística de *Data Science* é a capacidade de manipular enormes quantidades de dados, usar algoritmos de Inteligente Artificial e de Computação em larga escala, para obter aqueles insights necessários. *Data Science* usa em toda a sua amplitude a estatística, porém é mais ampla que isso. É como se ela fosse uma superextensão da estatística usando uma grande quantidade de dados. Quando falamos de grande quantidade voltamos a tratar sobre *Big Data*. *Data Science* e *Big Data* estão amplamente ligados, *Data Science* tem toda sua potencialidade maximizada para lidar com uma gigantesca massa de dados, que vai além de uma planilha Excel, que só suporta um milhão de linhas de informações. Já em *Big Data*, estamos falando de conjunto de dados de *Terabytes* de informações, que contém bilhões de linhas de informação, por isso são necessários métodos robustos de busca de dados. Portanto, temos a segunda grande área, que é combinada à estatística para formar *Data Science* – a Computação.

E da combinação entre Computação, Matemática, Probabilidade e estatística nasce a subárea da Inteligência Artificial (IA), que é uma das bases da Ciência de

Dados, que é *Machine Learning*, ou Aprendizado de Máquina, que veremos um pouco mais nos próximos capítulos deste livro.

FIGURA 12 – PANORAMA DA CIÊNCIA DE DADOS.



FONTE: A autora

A Ciência dos Dados, basicamente, estuda os dados, assim como a Ciência Social estuda a origem, o desenvolvimento e organização da sociedade. A Ciência dos Dados estuda a origem, o desenvolvimento e a organização dos dados. Dados são reais, têm propriedade reais, precisam ser entendidos e estudados, assim como a sociedade. Portanto, o profundo conhecimento e a experiência em dados a serem analisados é a terceira grande área necessária em *Data Science*.

Você pode ter técnicas matemáticas das mais apuradas para fazer previsões, avaliações e ter vasto desempenho computacional, e uma coleção enorme de algoritmos, sobretudo os métodos de otimização. Mas se não tiver conhecimento do domínio dos dados, nada disso vai adiantar, todo o recurso pode ir por água abaixo. Quem vai dar o contexto a que os dados pertencem? Quem vai dizer prioritariamente que tipo de fenômeno os dados representam? O conhecimento do domínio ou a falta dele é uma das maiores discussões sobre *Data Science*, então se você quer extraír conhecimento dos seus dados, conheça-os muito bem.

## 4.2 MODELOS PREDITIVOS

Como citado em um dos tópicos anteriores, a aprendizagem de máquina é um subcampo da Inteligente Artificial, que evolui a partir do estudo de reconhecimento



de padrões e teoria da aprendizagem computacional, dando ao computador a capacidade de aprender sem estar programado de forma explícita. E para isso acontecer precisamos criar um modelo, ou melhor, um modelo preditivo.

Em um resultado de qualquer processo de aprendizagem de máquina, o modelo preditivo possui uma função matemática aplicada, com conceito de estatística a depender do algoritmo que esteja sendo utilizado. Esses algoritmos operam a partir de um modelo, de um conjunto de dados, que são observações de entradas ou *inputs*, a fim de fazer previsões. Logo, considera-se um modelo uma função matemática que explica o relacionamento entre os dados. E quanto mais sofisticado for o algoritmo, mais sofisticado será o modelo preditivo.

O objetivo principal do modelo preditivo é ir além de saber o que aconteceu, ao fornecer uma melhor estimativa do que poderá acontecer no futuro, usando dados, algoritmos e métodos oriundos da estatística, aprendizado de máquinas e mineração de dados para se determinar as chances de resultados futuros, ou desconhecidos, com base em dados passados. O modelo preditivo apresenta três aspectos importantes e que precisam ser compreendidos, tais como:

1. **Coletar dados:** coletar informações ou características de diversas fontes distintas do produto ou serviço a ser analisado.
2. **Ensinar o modelo:** depois da coleta dos dados, é preciso identificar qual algoritmo é o mais indicado para ensinar o modelo a relacionar o que você quer que seja feito com os dados coletados. Vamos a um caso prático, suponha que precisamos criar um modelo preditivo que seja capaz de predizer se uma música (que ainda não foi lançada) fará sucesso ou não nas *playlists*, dadas algumas informações sobre ela. O algoritmo irá obter conhecimento a partir de dados da música que já foram lançados. Com base nesses conhecimentos o algoritmo irá treinar o modelo para predizer se a música será ou não um sucesso.
3. **Fazer previsões:** com o treinamento do modelo preditivo sendo realizado, é possível apresentar características de uma música (ainda não lançada) a ele, que será apto de predizer se a música será ou não sucesso nas *playlist*.

A análise preditiva, se formos pensar com mais critérios, se refere a algo que a humanidade em si já faz algum tempo. Nós, como seres humanos, sempre fazemos este processo, olhamos para nosso passado tentando entender algumas coisas, de forma a guiar os nossos passos futuros ou resolver questões ainda pendentes.

Atualmente esta análise encontra-se mais criteriosa, a avaliação dos dados tanto internos como externos de uma organização torna-se cada vez mais necessária, dada a escassez de tempo e a cobrança por agilidade e flexibilidade imposta pelo mercado na tomada de decisão. A palavra-chave ou conceito-chave relacionado à palavra preditiva é a tomada de decisão. A todo momento somos demandados a tomar decisões desde as mais corriqueiras, como que roupa vestir até decisões mais importantes, tal como guiar minha carreira?

E decisões afetam tanto indivíduos com grandes organizações, ser capaz de tomar decisões adequada é importante em qualquer esfera e o valor de modelagem preditiva está justamente em dar suporte à tomada de decisão.

---

Quer conhecer mais sobre Análise Preditiva? Leia o livro: **Análise preditiva**: o poder de prever quem vai clicar, comprar, mentir ou morrer, do autor Eric Siegel, Rio de Janeiro: Editora Alta Books, 2017.



Sugiro ainda os demais:

- **Data mining: practical machine learning tools and techniques.** WITTEN, I. H. et al.: Morgan Kaufmann, 2011.
  - **Machine Learning.** MITCHELL, Tom M.: McGraw Hill, 1997.
  - **Predictive data mining.** WEISS, Sholom M.; INDURKHYA; Nitin. Morgan Kaufmann: 1997.
- 

## 5 ALGUMAS CONSIDERAÇÕES

Vivemos em um cenário de competições e expectativas por novas experiências e inovações que visam realizar as ações de melhor forma e em menos tempo. Este capítulo expôs algumas informações sobre a exponenciação dos dados e em paralelo à busca pelo desenvolvimento de novas tecnologias que suportam e saibam lidar com este grande volume de dados para gerar uma melhor tomada de decisão, seja qual for o cenário trabalhado. Destacando, portanto, os principais conceitos de *Big Data*, os seus critérios, fontes e técnicas vinculadas a qualidade dos dados que visam atender essa nova demanda deste cenário da Era dos Dados.

*Big Data* vem ganhando destaque e compreender seus principais aspectos como os cinco V's o ajudará a definir com precisão quais tipos de dados devem ser considerados grandes volumes de dados, destacando que, em determinadas aplicações, um aspecto pode ter mais ênfase que o outro, ou seja, há casos em



que a variedade pode ser mais predominante que o volume ou a velocidade o ajudará compreender na formação da sua equipe de trabalho, na escolha de suas ferramentas e tecnologias.

Apesar de todos os desafios, o primeiro passo está sendo dado, buscar conhecer os conceitos, técnicas, ferramentas e métodos existentes para saber aplicá-los diante as suas reais necessidades diagnosticadas. Aprimore cada vez mais suas experiências sobre *Big Data*, antes de cair no impulso para utilizar *Big Data*, pois não basta apenas adquirir tecnologias, requer expertise para selecionar e validar os dados para apontar melhores modelos preditivos para cada projeto.

“Dado é o novo Petróleo! Precisamos encontrá-lo, extraí-lo, refiná-lo, distribuí-lo e monetizá-lo!” David Buckingam

## REFERÊNCIAS

ABNT. NBR ISO 10006. **Gestão da qualidade** – Diretrizes para a qualidade no Gerenciamento de Projetos. ABNT/CB-25 – Comitê Brasileiro de Qualidade, 2000.

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

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, 2015.

GYORÖDI, C. et al. A comparative study: MongoDB vs. MySQL. In: **13th International Conference on Engineering of Modern Electric Systems** (EMES), 2015.

HECHT, R.; JABLONSKI, S. NoSQL Evaluation a use case oriented survey. In: **International Conference on Cloud and Service Computing**, 2011.

JAYATHILAKE, D. et al. A study into the capabilities of NoSQL Databases in handling a highly heterogeneous tree. In: **99X Technology Colombo**, Sri Lanka, 2012.

LOSHIN, David. **The practitioner's guide to data quality improvement.**, MK, 2010.

MARQUESONE, Rosangela. **Big Data**: técnicas e tecnologias para extração de valor dos dados. 2017.

MCAFEE, A.; BRYNJOLFSSON, E. Big Data: The management revolution. **Harvard Business Review**, 2012.

SRIVASTAVA, P.; GOYAL, S.; KUMAR, A. **Analysis of various NoSql database**. In: Dept. of Computer Science and Engineering Mody University of Science and Technology Rajasthan, India, 2015.

VARIETY. **Netflix bandwidth usage climbs to nearly 37% of internet traffic at peak hours**, 2015. Disponível em: <<http://variety.com/2015/digital/news/netflix-bandwidth-usage-internet-traffic-1201507187>>. Acesso em: 12 out. 2018.

VOLPATO, T.; RUFINO, R.; DIAS, J. **Big Data**: transformando dados em decisões, Unipar, PR, 2014.

WATSON, H. Should you pursue a career in BI/Analytics? **Business Intelligence Journal**, 4-8. 2015

ZEPHORIA DIGITAL MARKETING. **The top 20 valuable Facebook statistics**, 2018. Disponível em: <<https://zephoria.com/top-15-valuable-facebook-statistics/>>. Acesso em: 15 out. 2018.





# CAPÍTULO 2

## TRABALHANDO COM *BIG DATA*

A partir da perspectiva do saber-fazer, são apresentados os seguintes objetivos de aprendizagem:

- ✓ Descrever, apresentar e explicar as formas existentes para obter, armazenar e gerenciar dados do Big Data.
- ✓ Manipular, analisar, avaliar, selecionar, aplicar, demonstrar e empregar técnicas para captura dos dados até a sua visualização.



# 1 CONTEXTUALIZAÇÃO

Não é novidade que a maioria dos dados disponíveis pelas organizações e consumidores é de dados desestruturados, mas até o momento as organizações ainda estão investindo mais em sistemas com dados estruturados, nos quais se implementa uma base de dados para atender uma necessidade específica de um único projeto e pronto. Porém, nos dias atuais, uma arquitetura de base de dados para de fato apoiar o crescimento das informações organizacionais diante dos dados desestruturados é muitas vezes necessária.

Conforme foi apresentado no primeiro capítulo, o processo para obtenção dos dados possui diversas etapas, desde a coleta do dado até o seu descarte. Acrescenta-se ainda a questão das **Legislações e Normas** corporativas, que podem decretar o período da permanência ou descarte dos dados. Logo, a natureza e a finalidade dos dados sempre irão influenciar parte desse processo de coleta e armazenamento. Visando auxiliar essas necessidades para captar e gerar dados e atribuir valor junto à tomada de decisões, há um processo genérico no qual é estruturado o ciclo de vida dos dados, por meio das respectivas etapas (AMARAL, 2016):

- **Produção:** basicamente é o meio em que se origina o dado, seja por meio de um computador, celular, dispositivo acoplado, como uma câmera fotográfica. Dados gerados por humanos ou máquinas.
- **Armazenamento:** os dados produzidos são mantidos em um dispositivo, seja de memória RAM, ou **não volátil**, como um disco de estado sólido.
- **Transformação:** em grande parte o dado para gerar informação passa por uma etapa intermediária de transformação, a qual é fundamental, visto que o dado é armazenado em um formato e para manter sua integridade deve ser atribuído um outro formato otimizado de armazenamento.
- **Análise:** a etapa de análise busca produzir informação e conhecimento a partir do dado.
- **Descarte:** o descarte pode ser gerado diante processos internos, visto não haver mais valor para as tomadas de decisões organizacionais.



- **Dispositivo Volátil:** aquele que mantém o estado dos dados enquanto houver fornecimento de energia elétrica, na falta dela, a informação se perde, como exemplo tem-se a memória RAM de um computador (AMARAL, 2016).
- **Dispositivo Não Volátil:** aquele que é capaz de manter o dado mesmo que o dispositivo não esteja conectado a uma fonte de energia, como por exemplo, um disco rígido, disco óptico, memória *flash*, entre outros (AMARAL, 2016).

Delimitar fases ou períodos em que são caracterizadas as necessidades e competências é indispensável para o acesso e uso dos dados, sendo ponto central os próprios dados, uma forma de evidenciar e gerenciar os diferentes momentos e fatores envolvidos nesse processo do ciclo de vida dos dados.

Uma vez que os dados passam por essas etapas, ferramentas e algoritmos de aprendizagem de máquina são agregados para realizar buscas, permitindo extrair conhecimentos e visualizá-los por meio de relatórios. A ferramenta *Hadoop* tem sido apontada como um dos *frameworks* mais utilizados, principalmente na etapa de processamento, por ser de código aberto, ter boa documentação e propiciar uma série de funcionalidades para o Analista de Dados. Você conhecerá mais sobre este *framework* em um dos próximos tópicos deste capítulo.

Diante de todo esse processo do ciclo de vida dos dados, nota-se que grande parte das empresas por meio dessas etapas busca reunir os dados, mas não sabe o que fazer com eles, não possui uma compreensão sobre as oportunidades de negócios e sobre os valores a serem obtidos por meio desses dados. Na maioria das vezes, isso ocorre por falta de um **Plano de Dados**, um conjunto claro e detalhado de casos de uso para os dados, além de uma lista de ferramentas e tecnologias que serão utilizadas para tirar o valor dos dados disponíveis. Além do que grande parte dos dados identificados acaba não contribuindo ou eles não servem de fato para serem aproveitados em posteriores análises, ocupando um maior armazenamento do Banco de Dados e prejudicando o desempenho ao serem processados. Contudo, entender o significado dos dados, incluindo os metadados (“representam informações que caracterizam a informação documentada, estes respondem o que, quem, quando, onde e como sobre cada faceta da informação, auxiliando a organização na sua publicação e suporte (DEVMEDIA, 2007), definir uma metodologia para comunicação e análise dos dados fará total diferença para que seu *Big Data* se transforme em um custo vantajoso para seu negócio. Espera-

se que você compreenda melhor este ciclo de vida dos dados e como criar um Plano de Dados adequado para trabalhar com *Big Data* através dos próximos tópicos deste capítulo.



Aprovado pelo plenário do Senador Federal, o **PLC 53/2018** dispõe sobre a proteção de dados pessoais e altera a **Lei nº 12.965/16 do Marco Civil da Internet**, sendo consolidada como a **Lei Geral de Proteção de Dados Brasileira (LGPD)**. No qual cria um novo regramento para o uso de dados pessoais no Brasil, tanto no âmbito on-line quanto off-line, nos setores privados e públicos, com os principais objetivos:

- **Direito à privacidade:** garantir o direito à privacidade e à proteção de dados pessoais dos cidadãos ao permitir um maior controle sobre seus dados, por meio de práticas transparentes e seguras, visando garantir direitos e liberdades fundamentais.
- **Regras claras para empresas:** estabelecer regras claras sobre coleta, armazenamento, tratamento e compartilhamento de dados pessoais para empresas.
- **Promover desenvolvimento:** fomentar o desenvolvimento econômico e tecnológico numa sociedade movida a dados.
- **Direito do consumidor:** garantir a livre iniciativa, a livre concorrência e a defesa do consumidor.
- **Fortalecer confiança:** aumentar a confiança da sociedade na coleta e uso dos seus dados pessoais.
- **Segurança jurídica:** aumentar a segurança jurídica como um todo no uso e tratamento de dados pessoais.

Portanto, as empresas que utilizam *Big Data*, que coletam informações diretamente do usuário, precisam mais que nunca garantir que estão avisando aos seus consumidores de todos os usos que serão dados às informações e que eles concordam com esses usos. Além disso, adotar todas as práticas necessárias para a proteção dos dados dos usuários delineados na Lei.

FONTE: <<http://dataprivacy.com.br/PLC53-18.pdf>>. Acesso em: 5 dez. 2018.



Conheça mais também sobre:

**Regulamentação europeia de proteção de dados/General Data Protection Regulation (GDPR):<<https://gdpr-info.eu/>>.**

**Lei do Cadastro Positivo/ Lei nº 12.414:**

<[http://www.planalto.gov.br/ccivil\\_03/\\_ato2011-2014/2011/lei/L12414.htm](http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/L12414.htm)>.

## 2 MANEIRAS PARA OBTENÇÃO DOS DADOS

Quando falamos em produzir dados, diversas fontes são imaginadas e de fato utilizadas como já vimos em nosso primeiro capítulo, o quantitativo de dados produzidos pela Internet em alguns segundos. Há ainda aquelas fontes mais intrigantes utilizadas em projetos como o Projeto *SETI* o qual busca vida extraterrestre, sendo considerado um dos maiores acontecimentos da computação distribuída no mundo, diante da utilização de processamento de computadores voluntários. E ainda outros projetos, como o *Climate Prediction*, que visa construir modelos de previsão meteorológica, e o Projeto *Rosetta*, que busca a cura de doenças.

Mas grande parte das empresas ainda não está de fato sabendo trabalhar com tamanha quantidade de dados, segundo a HP (2018), a maioria das empresas ainda não têm domínio completo de tecnologias *Big Data*, pois:

- 87% das empresas atualmente não conseguem explorar completamente as informações humanas;
- 85% das empresas acham que não são rápidas o suficiente para analisar os dados de máquinas; e
- 92% das empresas acham que não exploram os dados de *clickstream* on-line tão bem quanto deveriam.

Assim, conhecer o processo do ciclo de vida dos dados torna-se um fator crucial para trabalhar com *Big Data*. A primeira etapa desse processo se caracteriza pela produção ou coleta dos dados, então, como serão coletados os dados? Entre outras questões fundamentais desta fase, pode-se destacar (PAPO, 2013):

- Qual é o escopo da necessidade informacional para se obter esses dados?
- Que tipo de resultado se espera com a coleta desses dados?
- Com quais características deve-se obter o resultado?
- Quais são os dados necessários?
- Onde estão as fontes para estes dados?
- Como os dados podem ser coletados?
- Em que formato estão os dados?
- Quais são os tratamentos necessários para que os dados fiquem adequados ao que se precisa?
- A coleta desses dados não proporciona risco de privacidade para os indivíduos ou entidades referenciados a eles?
- Quais elementos que, em alguns casos, poderiam ser considerados como secundários, que permitem a integração entre os diversos dados coletados?
- Como avaliar sua integridade física e lógica, além de outros elementos que garantam sua qualidade?
- Como identificar a procedência dos dados?
- Você tem o direito ou permissão de coletar esses dados?
- Estão sendo coletados dados que permitam que estes venham a ser identificáveis e recuperáveis em um momento futuro?
- Estão sendo coletados dados que propiciem a manutenção e acesso a eles no futuro, caso venham a ser armazenados?

Por fim, uma das questões mais relevantes: como obter valor a partir dos dados existentes? É uma pergunta muito importante, pois nem sempre obter dados significa obter mais conhecimentos ou informações que acrescentem valor de negócio à tomada de decisões. Sendo preciso ir além de saber quais dados existem na sua empresa e quais ferramentas estão sendo ou precisarão ser utilizadas, mas ainda avaliar a capacidade dos dados, identificando oportunidades para atender as necessidades de negócio e de fato gerar valor. E para isso algumas questões básicas devem ser respondidas (GAVINO, 2018):

1. Quais dados sua empresa realmente precisa? Compreenda como seu negócio identifica prioridades para os dados melhorarem a tomada de decisões da empresa podendo impulsionar os seus negócios.
2. Quais dados existem em sua empresa? Verifique quais dados dos seus consumidores proporcionam maiores engajamentos em sua empresa. Como seus consumidores se comportam, pagam, informam sobre seus produtos e serviços pelos mais diversos canais? Há como aproveitar tecnicamente e legalmente estes dados?
3. Como é realizada a gestão dos dados da sua empresa? É preciso definir, esclarecer e treinar sobre as políticas, os processos, sobre a gestão e a segurança dos dados confiados a sua empresa.



4. Como os dados da sua empresa são obtidos, armazenados, organizados, integrados e utilizados? Preparar a arquitetura dos dados e documentar as fontes fazem parte do seu Plano de Dados.
5. Em sua empresa há pessoas capacitadas para trabalhar em estratégias de dados?
6. É possível encontrar valores nos dados da sua empresa de forma clara e objetiva?

Resumidamente diante uma variedade e a quantidade de dados infinitos em *Big Data*, aprimorar a qualidade na gestão não significa necessariamente aumentar o volume de dados coletados, mas sim, elaborar um programa de controle das informações, definindo o papel que os dados terão dentro da instituição.

A princípio é preciso considerar dados e suas distintas categorias, tais como: **dados internos**, ou aqueles de que a empresa é dona e sobre os quais possui controle, como: dados de Sistemas de Gestão da empresa, arquivos ou documentos escaneados, formulários de seguros, notas fiscais, documentos gerados por colaboradores, sensores e até registros de *logs*.

Há também os **dados externos**, textuais ou transacionais coletados por dados da Web, que tornam um desafio não somente capturar os dados, mas também o armazenamento e análise diante do volume, da velocidade e da variedade de dados disponíveis. Todavia, por meio das informações de mídias sociais, é possível detectar, por exemplo, quais aspectos são mais comentados pelos clientes sobre a sua organização. Esses dados de mídias sociais podem ser capturados por meio de uma *Application Programming Interface* (API), as principais mídias, como Facebook, Twitter, YouTube, já disponibilizam APIs para os usuários utilizarem os dados que circulam dentro das suas redes.



É relevante saber e conhecer as APIs para se trabalhar com *Big Data*, saiba mais sobre algumas delas em:

Instagram– <<https://www.instagram.com/developer/>>.

Facebook – <<https://developers.facebook.com/>>.

LinkedIn – <<https://developer.linkedin.com/>>.

Twitter – <<https://dev.twitter.com/>>.

YouTube– <<https://developers.google.com/youtube/>>.

Pinterest– <<https://developers.pinterest.com/>>.

Outra categoria de dados classificada é aquela gerada **por humanos** oriundos de mídias sociais, em que os usuários publicam o que pensam sobre algo, suas preferências e suas emoções por meio de textos, imagens, áudio ou vídeo, bem como através de avaliações sobre produtos e serviços, acrescentando também os dados gerados por aplicativos em troca de mensagens, como WhatsApp, Skype, entre outros (MARQUESONE, 2016).

Completam-se ainda os dados gerados **por máquinas**, tais como processos de computadores, aplicações e outros mecanismos que não precisam explicitamente de intervenção humana. Podendo também considerar o monitoramento dos dados de aplicações Web em servidores, em que são gerados registros de *logs* com milhares de registros com informações úteis aos provedores de serviços (MARQUESONE, 2016).

Por fim, há ainda aqueles **dados híbridos** entre máquinas e humanos, a saber (HURWITZ, 2016):

- **Dados de sensor:** podem ser classificados como etiquetas de radiofrequência (RFID, *Radio-Frequency IDentification*), medidores inteligentes, dispositivos médicos e dados de sistemas de posicionamento global (GPS, *Global Positioning System*).
- **Dados de web log:** servidores, aplicativos de redes e outros, operam e capturam tipos de dados sobre suas atividades.
- **Dados de ponto de venda:** código de barras de produtos sendo associados todos os dados do produto.
- **Dados financeiros:** sistemas financeiros atuais operam com base em regras predefinidas que automatizam processos. Dados de negociações de ações são um exemplo, contendo dados estruturados, o símbolo da empresa e valor do dólar.
- **Dados de entrada:** dados que uma pessoa possa inserir em um computador para entender o comportamento básico do cliente.
- **Dados de fluxo de clique:** dados gerados por cliques em um website para analisar determinados comportamentos de clientes.
- **Dados relacionados a jogos:** cada movimento feito em um jogo podendo ser gravado, tornando-se útil para entender como os usuários se movem.

Ao pensar em fontes de *Big Data*, deve-se pensar em todos os dados disponíveis para análise que são originários de diversos canais, com formatos e origens diferentes, como:



- **Formato:** estruturado, semiestruturado ou não estruturado.
- **Velocidade e Volume:** a velocidade em que os dados chegam e a taxa em que são entregues modificam-se conforme a fonte de dados.
- **Ponto de Coleta:** o ponto em que são coletados os dados, diretamente ou através de provedores de dados, em tempo real ou em modo em lote, podendo vir de uma fonte primária, com condições climáticas, ou de uma fonte secundária, como um canal de clima patrocinado pela mídia.

Mas é preciso obter todos esses dados destas categorias para obter oportunidades com as tecnologias de *Big Data*? A resposta é claro que não, mas oportunidades podem ser desperdiçadas pelo fato de (MARQUESONE, 2016):

- os dados não estarem integrados;
- os dados demoram para ser analisados;
  - os dados não estarem categorizados;
  - os dados estarem obscuros;
  - os dados não são usados para tomadas de decisões;
  - os dados não são visualizados com clareza;
  - os dados não são medidos.

Ter uma visão panorâmica da cadeia de evolução dos dados em sua empresa irá contribuir para uma gestão mais bem direcionada, sem perder o foco das estratégias do negócio. Se não houver um bom planejamento, a gestão dos dados atuará de modo impreciso, coletando dados incorretos e gerando, por fim, resultados inadequados.

Quer dizer que a coleta precisa e exata de dados permitirá reunir informações suficientes para a construção de um conhecimento sólido usado na tomada de decisões seguindo as metas da empresa. Ter uma visão panorâmica da cadeia de evolução dos dados em sua empresa irá contribuir para uma gestão mais bem direcionada, sem perder o foco das estratégias do negócio. Se não houver um bom planejamento, a gestão dos dados atuará de modo impreciso, coletando dados incorretos e gerando, por fim, resultados inadequados. Portanto, crie seu Plano de Dados e uma arquitetura de gestão segura, preparando uma estrutura para se trabalhar com os dados identificados. Vejamos um pouco sobre uma base arquitetônica no próximo tópico.



O processo para captar, gerar dados e atribuir valor junto à tomada de decisões não é fácil, demandando de etapas significativas. Descreva com suas palavras quais são essas etapas e como elas funcionam.

## 2.1 A BASE ARQUITETÔNICA

Suas necessidades dependerão da natureza dos dados, conhecer seus dados irá lhe permitir dar suporte ao desempenho pretendido. Você necessitará conhecer as categorias dos seus dados para identificar a quantidade do poder de velocidade computacional existente, visto que algumas das análises serão realizadas em tempo real. Entenda que sua empresa e suas necessidades influenciarão no planejamento e na atenção que deverão ser prestados para utilização de *Big Data*. Algumas questões iniciais para planejar sua arquitetura podem ser resumidas em (HURWITZ, 2016):

- Sua empresa precisará administrar quantos dados agora e futuramente?
- Sua empresa precisará administrar os dados em tempo real ou próximo ao tempo real? E com que frequência?
- Quais os riscos que sua organização consegue suportar?
- Qual o grau de importância da velocidade para sua gestão de dados?
- O quanto precisam ser seus dados?

Assim, uma variedade de serviços deve ser compreendida em uma arquitetura de gestão de *Big Data*, conforme a Figura 1.

FIGURA 1 – ARQUITETURA DE GESTÃO DE *BIG DATA*

Pilha de Tecnologia Big Data



FONTE: Hurwitz (2016, p. 18)



Note que interfaces e fontes estão entrando e saindo tanto dos dados internos administrados quanto dos dados de fontes externas, isto é, estão sendo coletados vários dados de diversas fontes. Entender essa necessidade de coletar os dados de diversas fontes é essencial para entender e trabalhar com *Big Data*. Nos próximos subtópicos será descrita cada uma das etapas da arquitetura de gestão de *Big Data* (HURWITZ, 2016).

## 2.1.1 Infraestrutura física redundante

*Big Data* não teria surgido sem a disponibilidade de uma infraestrutura física robusta que suporte um volume de dados em uma computação distribuída por meio de sistemas de arquivos distribuídos, ferramentas e aplicativos. Já a redundância ocorre diante das várias formas de replicação dos serviços, por exemplo, se a empresa deseja conter o crescimento interno de TI, pode usar serviços externos de nuvem aumentando seus recursos internos.

## 2.1.2 Infraestrutura de segurança

Proteger os dados organizacionais é de fundamental importância, tanto para atender às exigências das normas existentes quanto para proteger a privacidade dos consumidores dos produtos e serviços da organização. Buscando identificar quem tem permissão para consultar os dados, quais as circunstâncias desses acessos e ainda verificar a identidade dos usuários desde o início e não ser algo visto como um acréscimo ao uso dos dados.

## 2.1.3 Fontes de dados operacionais

Uma fonte de dados operacional consistia em dados estruturados e administrados por linha de código de negócios em uma base de dados relacional. Porém, atualmente dados operacionais devem apreender um conjunto mais amplo de fontes, incluindo dados não estruturados, já em base de dados NoSQL. Logo, é preciso mapear as arquiteturas de dados aos tipos de transações, buscando arquitetura de dados que suportem conteúdo desestruturado, incluindo bases relacionais e não relacionais.

## 2.1.4 Organizando serviços de dados e ferramentas

Tempos atrás grande parte das organizações não conseguia capturar ou armazenar grande quantidade de dados, ou por ser muito caro ou devido à quantidade de dados. Algumas organizações até conseguiam captar os dados, mas não havia ferramentas adequadas para auxiliar de melhor forma esse processo. Atualmente ainda há uma quantidade de dados crescente de informações com uma enorme variedade de fontes não organizadas, vindo de máquinas, sensores, fontes públicas e privadas. O importante é sempre lembrar que nem todos os dados utilizados são operacionais. E que diante do avanço das tecnologias de computação, já é possível administrar o imenso volume de dados que não era possível ser administrado anteriormente. As soluções resultantes estão transformando o mercado de gestão de dados com o desenvolvimento de ferramentas como o *MapReduce*, *Hadoop* e *Big Table*.

## 2.1.5 Análises tradicionais e avançadas

Formas abrangentes de administrar *Big Data* necessitam de abordagens diferentes a fim de auxiliar o negócio, seu planejamento e obter um retorno futuro de sucesso. Por isso, após captar todos os dados, de todas as formas, qual o sentido desses dados para o seu negócio? Qual abordagem de análise utilizar? Um armazém de dados tradicional ou análises preditivas avançadas? Esses armazéns de repositórios proveem compreensão, particionamento multinível e uma arquitetura massiva de processos paralelos. A análise em *Big Data*, diante de sua complexidade, já alveja a utilização de modelos preditivos que conectam os dados estruturados e não estruturados.

## 2.1.6 Relatórios e visualizações

As organizações contaram sempre com a elaboração de relatórios para compreender melhor os dados e as informações que estes apresentavam. *Big Data* muda a forma como os dados são administrados e utilizados, relatórios e visualização de dados se tornam ferramentas para visualizar o cenário de como os dados estão conectados e qual o impacto dessas relações no futuro.

## 2.1.7 Aplicativos de *Big Data*

Com o desenvolvimento de *Big Data* aplicativos estão sendo projetados para aproveitarem especificamente características do *Big Data*. Esses aplicativos



contam com enormes volumes, velocidades e variedades de dados para verificar o comportamento de um determinado segmento, por exemplo, um aplicativo de gestão de tráfego pode reduzir o número de engarrafamentos em rodovias.

Uma vez obtidos, os dados podem ser utilizados para um fim imediato e descartados. No entanto, pode ser necessário e útil manter esses dados disponíveis de alguma forma para acesso futuro, conheça um pouco sobre o armazenamento dos dados conforme o próximo tópico deste capítulo.



Quer conhecer mais sobre insights que *Big Data* proporciona para serviços financeiros? Consulte o e-book *Big Data em serviços financeiros*: conheça as aplicações e cases de sucesso. FONTE: <<http://offers.bigdatabusiness.com.br/big-data-servicos-financeiros>>.

## 3 ARMAZENAMENTO DOS DADOS

Os dados produzidos ou gerados devem, consequentemente, ser armazenados, garantindo que futuramente os dados possam ser recuperados para replicação de um determinado processo, ou para produzir informações e conhecimentos, levando sempre em consideração um conjunto de premissas, como a segurança da informação, a integridade, a diminuição de redundância, a concorrência, o espaço, entre outras questões.

Para mais, já sabemos que os dados podem estar em Banco de Dados diferentes, arquivos diferentes, logo devem existir processos que extraiam a informação, transformando-a em outro modelo e carregando-a para outro lugar (PAPO, 2013). Antes de tudo, questões fundamentais desta fase devem ser verificadas, a saber:

- Quais são os dados disponíveis?
- Quais destes dados serão armazenados?
- Qual estrutura (física e lógica) será utilizada para seu armazenamento?
- Como garantir a permanência dos dados complementares sobre a coleta para que se tenha garantido o contexto de sua obtenção?
- Estes dados podem representar um risco à privacidade dos indivíduos ou instituições neles referenciados de alguma forma?
- Como as partes de sua estrutura lógica serão interligadas e como serão mantidas as interligações com outros conjuntos de dados?

- Como garantir que os elementos que sustentam a sua qualidade sejam mantidos?
- Tem-se o direito de armazenar estes dados?
- Todos os aspectos que podem contribuir para sua encontrabilidade estão sendo armazenados?
- Todos os fatores para sua utilização ao longo do tempo estão sendo mantidos?

Arthur Chapman, em seu livro *Princípios de Qualidade de Dados* (2015), avalia como o armazenamento influencia na qualidade dos dados e destaca fatores que devem ser levados em conta na boa gestão do armazenamento dos dados, conheça alguns deles:

- **Cópias de segurança:** é vital que sejam realizadas regularmente e que as instituições mantenham um programa de recuperação de cópia, em casos de desastres, não haverá perdas significativas.
- **Arquivamento:** é um processo contínuo que inclui também o descarte de dados obsoletos e contribui para que os dados se mantenham facilmente acessíveis e prontos para serem analisados.
- **Integridade dos dados:** busca proteger os dados não só de perdas, mas também de que serem acessados por pessoas não autorizadas.

Outrossim, para gerir os dados com melhor qualidade, é necessário definir o Banco de Dados mais adequado ao tipo de dado e objetivos da empresa. É sempre importante destacar que *Big Data* não é tanto ter muita capacidade de disco, mas a forma de como se processa a informação. A grande dificuldade é conseguir dividir muita informação, obter várias porções de informações, tratar essas porções, obter resultados e depois conciliar estes resultados em um desfecho único, se tornando uma vertente a mais de análise e não simplesmente de capacidade de armazenamento. Portanto, armazene seus dados com cautela e permaneça seguindo seu Plano de Dados traçado inicialmente, conheceremos mais alguns esclarecimentos sobre o armazenamento de dados, conforme os próximos subtópicos.

---

Conheça 26 principais certificações internacionais de *Big Data*:  
<https://computerworld.com.br/2018/10/01/26-certificacoes-internacionais-de-big-data/>.

---





## 3.1 PREPARAÇÃO E ARMAZENAMENTO DOS DADOS

Para definir uma forma mais adequada para o armazenamento da informação, deve-se ter atenção aos seguintes aspectos:

- **Escalabilidade:** quantidade de usuários que acessa o Banco de Dados, pode crescer e decrescer rapidamente, portanto a empresa deve estar apta para uma solução escalável.
- **Alta disponibilidade:** o acesso à informação deve estar sempre disponível, o Banco de Dados sempre deve estar funcionando.
- **Flexibilidade:** a forma de armazenamento da informação deve ser flexível, podendo armazenar dados estruturados e não estruturados, que serão processados com a utilização de diversas tecnologias.

Além disso, para determinar a melhor estratégia para armazenar as informações, deve-se considerar: a estrutura do Banco de Dados, custo de equipamentos, custo de equipe de *Big Data*, aspectos de segurança da informação, como as informações serão processadas, quais os softwares e aplicativos envolvidos e como os dados serão gerenciados.

O ideal, na maioria das vezes, é criar um único repositório para que todas as informações estejam disponíveis a todos os usuários. Este repositório é chamado **Data Lake**, no qual as informações são armazenadas de forma bruta, ou melhor, na mesma forma em que foram coletadas na fonte de dados. O *Data Lake* pode ser criado na empresa, por meio da *Cloudera* uma das principais fornecedoras de soluções, suporte e serviços de software para *Big Data*. Para criação de um *Data Lake*, vários membros da empresa devem participar, como: equipe de TI, área de modelagem, área de negócios e diretores. Um *Data Lake* pode ser criado também na nuvem, a *Microsoft* disponibiliza serviços de armazenagem de dados, o *Azure Data Lake* possibilita armazenar dados de qualquer tamanho, forma e velocidade, bem como realizar todo o tipo de processamento e análise em diferentes plataformas e linguagens. *Data Lake* ainda remove as complexidades relacionadas a captura e armazenamento dos dados enquanto acelera a execução das análises.

Repare que *Big Data* trouxe inovação na forma de se armazenar as informações, utilizando Banco de Dados SQL e NoSQL a depender do objetivo do projeto da sua organização. E ainda se pode se utilizar Banco de Dados com

as características já vistas em nosso primeiro capítulo: orientado a chave-valor, orientado a colunas, orientado a documentos ou orientado a grafos.

## 3.2 CLOUD COMPUTING

De acordo com o *National Institute of Standards and Technology* (NIST), **Cloud Computing** é um modelo que permite um acesso sob demanda via Redes de Computadores a um conjunto compartilhado de recursos computacionais que podem ser rapidamente provisionados e liberados com um mínimo de esforço administrativo ou interação com o provedor de serviço.

Uma empresa pode optar por armazenar os dados em uma **Cloud Privada**, com uso exclusivo de uma empresa, quando deseja-se um nível muito alto de segurança e confidencialidade. Ou em uma **Cloud Pública** de uso público, sendo que uma organização é dona da infraestrutura e vende os serviços, podendo dimensionar a quantidade de servidores e serviços de acordo com a necessidade da empresa contratante, e o pagamento ocorre conforme a aplicação. Ainda há a **Cloud Híbrida**, a combinação do ambiente público com o ambiente privado, fazendo um modelo misto, de forma que uma empresa possa colocar a parte mais confidencial internamente dentro da empresa e as demais informações podem ser armazenadas em uma *Cloud Pública*.

Portanto, modelos de armazenamento em nuvem apresentam flexibilidade, permitindo avaliar uma abordagem melhor diante as necessidades do cliente e dos negócios. Enfim, os investimentos para a análise de *Big Data* podem ser significativos e demandam uma infraestrutura eficiente e econômica, com recursos para apoiar modelos de computação distribuída internamente. Nesse contexto, as nuvens privadas podem oferecer um modelo mais eficiente e econômico para implantar a análise de *Big Data*, enquanto permitem ampliar os recursos internos com serviços de nuvem pública. Já esse sistema híbrido de nuvem permite às empresas usar o armazenamento de espaço on-line de acordo com a demanda e capacidade de informatização, por meio de serviços de nuvem pública para algumas opções de análise (por exemplo, projetos de curto prazo) e fornecer capacidade adicional e escala conforme necessário (INTEL, 2012).

*Big Data* pode mixar fontes internas e externas, enquanto as empresas costumam manter dados mais sigilosos internamente (*in-house*). Grande parte do *Big Data* (de propriedade da empresa ou gerados por terceiros e prestadores públicos) pode ser alocada, externamente, já em ambiente de nuvem. Processar fontes de dados relevantes por trás do *firewall* pode ser um investimento significativo de recursos. Analisar os dados onde eles residem requer centros de



dados de nuvem internos ou públicos, ou em sistemas de ponta e dispositivos do cliente, muitas vezes faz mais sentido (INTEL, 2012).

Modelos de computação em nuvem podem ajudar a acelerar o potencial para soluções de análise em escala. Nuvens oferecem flexibilidade e eficiência de acesso aos dados, fornecendo insights e agregando valor, no entanto, análises de *Big Data* baseadas em nuvem não são solução para tudo.

A camada de armazenamento basicamente é responsável por adquirir dados das fontes e, se necessário, convertê-los para um formato adequado à maneira como devem ser analisados. Podemos citar a necessidade de converter uma imagem para armazená-la em um armazenamento *Hadoop Distributed File System* (HDFS) para processamento posterior, conforme será visto nos próximos tópicos.

## 4 PROCESSAMENTO DOS DADOS

Já sabemos que uma das dificuldades em trabalhar com *Big Data* é o fato de a maioria dos dados não ser estruturada e gerada de maneira organizada e padronizada, e poderem ser armazenados em Banco de Dados Relacional tradicional facilmente. Para resolver isso, geralmente se fala das soluções NoSQL (*NO SQL = NotOnly SQL*). O termo NoSQL denota que são Bancos de Dados que não estão apenas baseados no modelo relacional.

A arquitetura para o processamento de grandes volumes de dados apresenta características como: distribuição paralela para contornar o problema da pouca confiabilidade do hardware e redundante, pois a falha de qualquer nó (um processador e uma unidade de armazenamento) não deve causar nenhum efeito sobre o desempenho do sistema (GHEMAWAT; GOBIOFF; LEUNG, 2003). Resumidamente podemos destacar alguns pontos sobre a arquitetura de processamento (MACHADO, 2017):

- O tamanho típico dos arquivos utilizados por ela é sempre grande.
- O sistema deve ser intrinsecamente tolerante a falhas. Dado o tamanho dos *clusters* e o tipo de hardware utilizado, considerando a probabilidade de ocorrer uma falha em algum dos nós é, para todos os efeitos, de 1 para 1000, portanto, médio entre falhas (em inglês, *mean time between failures*, MTBF).
- A maior parte das alterações nos arquivos é de acréscimos (*appends*). Ou seja, as escritas em posições aleatórias são praticamente inexistentes. Ou seja, a grande maioria dos acessos é para leitura e, geralmente, essa leitura é sequencial.

- O tipo mais comum de processamento é em lote, ou *batch*, ao invés de pequenos ciclos de leitura e escrita. Assim, o sistema deve ser otimizado para garantir uma resposta contínua em termos de consumo de banda, ao invés de ter uma latência pequena.

O Google foi a primeira empresa a criar uma plataforma para o processamento massivamente paralelo denominada *Google File System* (*Google FS*), posteriormente surgiram outras arquiteturas, como a *Hadoop Distributed File System* (*HDFS*) (HADOOP, 2016), abordaremos um pouco sobre elas neste capítulo, em especial sobre o *HDFS*.

Na verdade, é difícil falar sobre *Big Data* sem mencionar o Google, pois, muitos dos seus estudos e trabalhos foram motivadores para as ferramentas que utilizamos hoje. Por exemplo, em 2003 o Google publicou um artigo *The Google File System*, com a ideia de criar um sistema de arquivos distribuídos e que fosse tolerante a falhas. Mas escalar em um universo de *Big Data* com muitos dados em uma única máquina é muito caro. Dessa forma eles estavam tentando escalar em muitas máquinas e buscando as mais baratas para funcionamento paralelo. Porém, máquinas mais simples possuem constantemente uma chance de falhar muito grande. Assim, criaram um sistema de arquivo tolerante a falhas e que aguentasse um grande volume de dados, o *Google File System*.

Já em 2004 foi publicado outro artigo, *MapReduce: Simplified Data Processing on Large Clusters*, uma técnica de programação paralela, que define um modelo em que seu programa paralelo tem que ser seguido a partir de duas operações – *Map* e *Reduce*. Se seu algoritmo se encaixasse neste modelo, ele conseguiria fazer uso de todo o *framework* que a Google tinha criado internamente. Foi algo revolucionário ao trazer para o universo de grandes dados uma programação paralela que fosse fácil de implementar, não que qualquer algoritmo se encaixasse no modelo *MapReduce*, mas nos que se encaixavam, era muito mais fácil de se implementar um algoritmo paralelo.

E em 2006 foi publicado, ainda pela Google, o artigo *Bigtable: a Distributed Storage System for Structured Data*, falando sobre um grande Banco de Dados para atender essa demanda de programação paralela – o *Bigtable*. A ideia era que se tivesse também um Banco de Dados distribuído e paralelo em que você pudesse acessar os dados com muitas atividades e um grande volume de dados.

Essas três ferramentas auxiliaram o Google a criar o mecanismo de busca que eles utilizam, índices diversos, mecanismos de varrer a Rede inteira, buscar palavras, indexar essas palavras, em quais sites estão cada uma. Toda essa operação em um grande volume de dados inicialmente foi viabilizada por meio destas tecnologias: um sistema de arquivo distribuído, um *framework* de



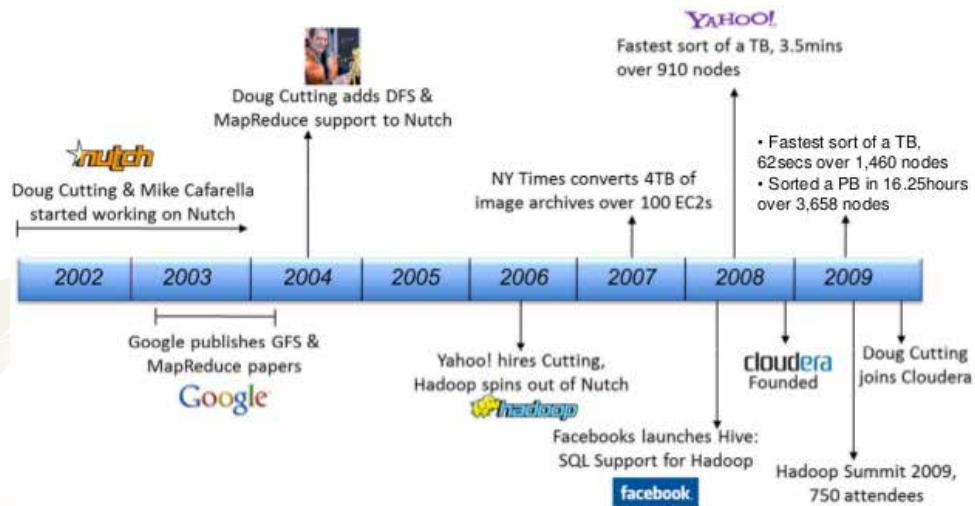
programação paralela e um Banco de Dados distribuído. Apesar desses artigos publicados, todas as ferramentas criadas eram utilizadas internamente somente pela Google. Até que um grupo de pesquisadores que trabalhavam e faziam mais ou menos o que a Google fazia começou a trabalhar no Yahoo e resolveu implementar o que estava sendo apresentado nos artigos da Google, em uma ferramenta *open source*, até que surgiu o *Hadoop*.

## 4.1 HADOOP

O *Hadoop* é um *framework open source* para armazenamento e processamento de dados em larga escala, composto por um sistema de arquivos distribuídos ou *Hadoop Distributed File System* (HDFS), por um sistema de gerenciamento de recursos distribuídos, que é o *Yet Another Resource Negotiator* (YARN) e um *framework* para executar *MapReduce*, permitindo criar um *cluster* de máquinas de processamento paralelo para trabalhar com *Big Data*. Apesar das suas semelhanças com os atuais sistemas de arquivos distribuídos, o *Hadoop* é altamente tolerante a falhas e é projetado para ser implantado em *hardware* de baixo custo (BORTHAKUR, 2007).

Este *framework* possui diversos eventos para sua evolução, como já foram citadas as publicações da Google, o seguimento pela Yahoo e ainda o surgimento da empresa Cloudera e as novas versões do *Hadoop*, conforme pode ser visto na Figura 2:

FIGURA 2 – EVOLUÇÃO DO HADOOP



FONTE:<<https://www.cloudera.com/>>. Acesso em: 9 dez. 2018.

Além da armazenagem de grande volume de dados de forma distribuída, o *Hadoop* apresenta um ecossistema de ferramentas e bibliotecas que auxilia em tarefas administrativas para o *cluster*, no processamento e análise de dados e no próprio armazenamento de dados, conforme a Figura 3, são apresentadas as ferramentas:

**D3:** é uma biblioteca *JavaScript* para visualização de dados – <<http://d3js.org/>>.

- **Tableau:** plataforma proprietária de visualização e análise de dados, entretanto, possui versões gratuitas para estudantes e para universidades – <<http://www.tableausoftware.com/pt-br>>.

FIGURA 3 – ECOSSISTEMA HADOOP



FONTE: A autora, adaptado de Bidoop Layes (2014).

- **Mahout:** é um projeto da *Apache Software Foundation* para produzir implementações livres de algoritmos de aprendizado de máquina escaláveis, focados principalmente nas áreas de filtragem colaborativa, *clustering* e classificação –<<https://mahout.apache.org/>>.
- **R:** ambiente para análises estatísticas –<<http://www.r-project.org/>>.



- **Java/Python/...**: Java é a linguagem oficial para criar programas em um *cluster Hadoop*, entretanto, é possível utilizar outras linguagens, como *Python e Ruby*.
- **Pig**: é uma plataforma para análise de dados que consiste de uma linguagem de alto nível para expressar análise de dados e a infraestrutura para executar essa linguagem –<<http://pig.apache.org/>>.
- **Hive**: fornece um mecanismo para projetar, estruturar e consultar os dados usando uma linguagem baseada em SQL, chamado *HiveQL* –<<http://hive.apache.org/>>.
- **MongoDB**: banco de dados orientado a documentos no formato JSON –<<http://www.mongodb.org/>>.
- **Kettle**: ferramenta de ETL (*Extract, Transform, Load*) que permite tratamento de dados construindo *workflows* gráficos – <<http://community.pentaho.com/projects/dataintegration/>>.
- **Flume**: ferramenta de coleta e agregação eficiente de *streams* de dados – <<http://flume.apache.org/>>.
- **Sqoop**: ferramenta que permite a transferência de dados entre bancos relacionais e a plataforma *Hadoop* – <<http://sqoop.apache.org/>>.
- **Chukwa**: sistema de coleta de dados para monitoramento de sistemas – <<https://chukwa.apache.org/>>.
- **Oozie**: é um sistema gerenciador de *workflows* para gerenciar tarefas no *Hadoop* –<<http://oozie.apache.org/>>.
- **Nagios**: ferramenta para monitorar aplicativos e redes – <<http://www.nagios.org/>>.
- **Zoo Keeper**: é um serviço centralizador para manter informações de configuração –<<http://zookeeper.apache.org/>>.



Quer conhecer mais sobre *Hadoop*, consulte o livro *Hadoop – The Definitive Guide: Storage and Analysis at Internet Scale*.

FONTE: <<http://barbie.uta.edu/~jli/Resources/MapReduce&Hadoop/Hadoop%20The%20Definitive%20Guide.pdf>>.

O Quadro 1 relata empresas que utilizam o *Hadoop* em sua infraestrutura:



QUADRO 1 – UTILIZAÇÃO DO HADOOP

UTILIZAÇÃO DO HADOOP		
Quem?	Onde?	Para que?
Adobe	Ferramentas e serviços para conteúdo digital	No armazenamento e processamento de dados internos e de redes sociais. Média de 80 nós de processamento
e-Bay	Comércio eletrônico.	Na otimização de buscas. Média 532 nós de processamento.
Facebook	Site que provê serviço de rede social.	Análise de log. Média de 1.400 nós de processamento.
Last.FM	Site que provê serviço de rádio online.	Análise de log, análise de perfil de usuário, teste A/B, outros. Média de 64 nós de processamento.
Linkedin	Site que provê serviço de rede social.	Análise e busca similaridade entre perfis de usuários. Média de 1.900 nós de processamento.
The New York Times	Ferramentas e serviços para conteúdo digital.	Em conversão de imagens e armazenamento de jornais digitais. Utiliza servidores da Amazon.
Twitter	Site que provê serviço de rede social	No armazenamento de mensagens e no processamento de informações.
Yahoo!	Site que provê serviço de buscas na web, notícias e e-mail.	No processamento de buscas, recomendações de publicidades, testes de escalabilidade. Média de 40.000 nós de processamento.

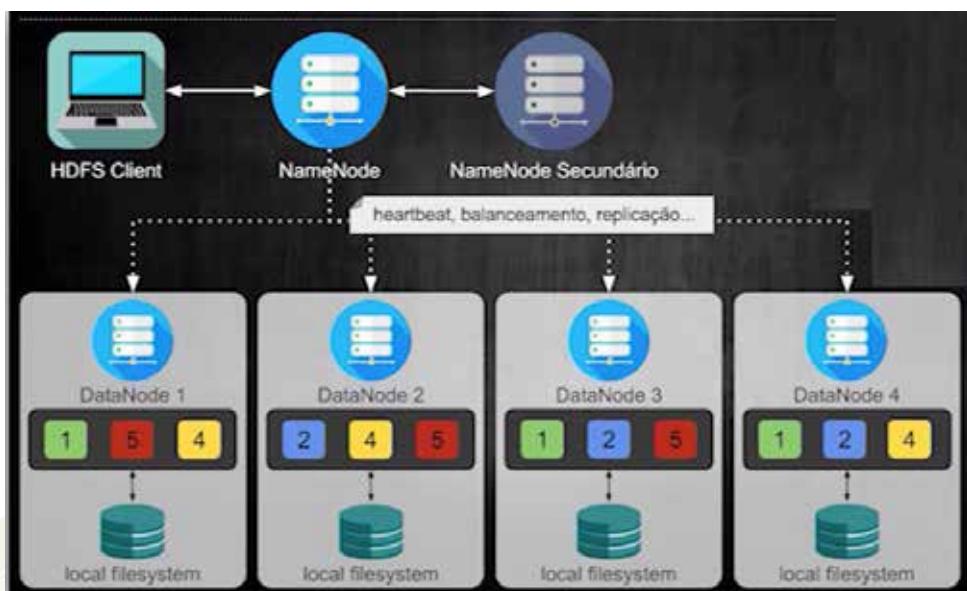
FONTE: Goldman et al. (2012)



## 4.1.1 HADOOP Distributed File System (HDFS)

Vamos compreender o funcionamento de cada um dos sistemas que se integram ao *Hadoop*, primeiramente o HDFS. A ideia inicial do HDFS é que temos arquivos muito grandes e queremos distribuí-los por máquinas em um *cluster*. Portanto, inicialmente vamos quebrá-los em blocos, pedaços que em geral possuem 64 *megabytes* ou podem ser configurados como desejar. E vamos distribuir esses pedacinhos pelas máquinas, só que, ao invés de colocar um pedacinho em cada máquina, vamos pegar cada bloco, cada pedacinho e vamos replicar nas diversas máquinas do nosso *cluster*. Normalmente este grau de replicação é três, digamos que temos um arquivo de 128 *megabytes* e cada bloco é de 64 *megabytes*, então eu tenho dois blocos e vou colocar cada um destes blocos em três máquinas diferentes, veja o exemplo da Figura 4 para melhor compreensão:

FIGURA 4 – HADOOP DISTRIBUTED FILE SYSTEM (HDFS)



FONTE: A autora, adaptado de <<https://codemphasic.wordpress.com/2012/09/27/big-data-hadoop-hdfs-and-mapreduce/>>.

A Figura 4 representa basicamente que temos um *DataNode* 1 possuindo os Blocos: 1, 5 e 4. O Bloco 1 também podemos encontrar no *DataNode* 3 e no *DataNode* 4. Já o Bloco 5 encontramos no *DataNode* 2 e no *DataNode* 3. E o Bloco 4 vamos encontrar no *DataNode* 2, *DataNode* 1 e *DataNode* 4. Portanto,

se uma máquina falhar, conseguimos recuperar os dados a partir de outras máquinas. Consegue compreender que o HDFS faz é criar estes arquivos e salvar no sistema de arquivo local?

Já sabemos que o *DataNode* é quem tem os dados, mas precisamos de algum lugar para consultar onde é que eles estão. Havendo um grande índice desses dados, ou seja, o *NameNode* organiza o meu sistema de arquivo. Ilustrativamente o meu código vai perguntar ao *NameNode* onde está o arquivo “tal” e ele vai retornar “ah, este arquivo tem tantos Blocos e cada um destes Blocos está nessa e nesta máquina”, trazendo todas as informações e permitindo meu código acessar os dados. Caso ocorra uma falha, o *NameNode* também precisa ser informado, assim os *DataNode* ficam constantemente encaminhando *heartbeat*, dizendo “olha, estou funcionando”. E uma vez que o *NameNode* perceber que uma máquina parou de mandar mensagem, parou de responder, ele tenta redistribuir imediatamente os dados que estavam naquele nó em outras máquinas. Ele faz a replicação novamente, colocando, por exemplo, os Blocos 1, 5 e 4 nas máquinas que não têm esses dados. O Bloco 1, ele poderia colocar no *DataNode* 2, o Bloco 5 no *DataNode* 4 e já o Bloco 4 no *DataNode* 3. Redistribuindo os dados em um grau de replicação 3, sendo muito importante o funcionamento deste mecanismo, senão, não conseguiremos balancear os nossos dados. Então o *NameNode* acaba sendo um ponto de falha crítico e único, sendo preciso que ele também tenha um *backup*, chamado *NameNode Secundário*, que assume caso o *NameNode* pare de funcionar.

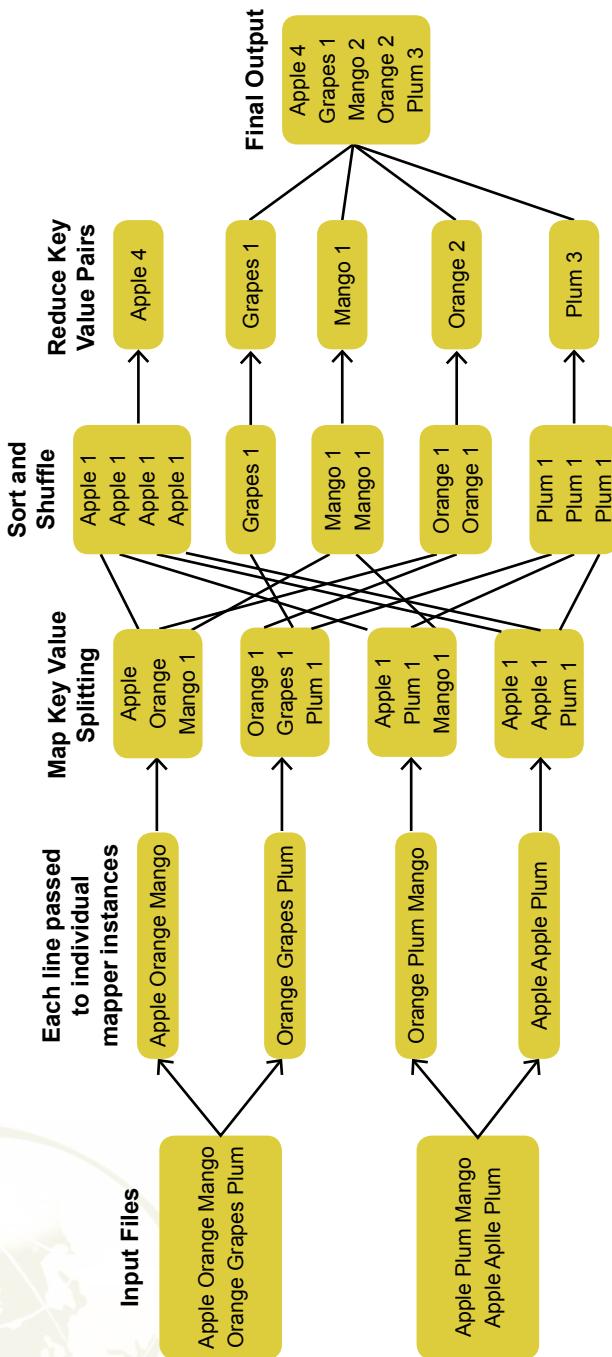
### 4.1.2 MapReduce

Distribuir os dados não é somente replicá-los, a grande ideia por trás destas ferramentas de processamento de grande volume de dados – ferramentas de *Big Data*, é que ao invés de pegar os dados e transferir para uma máquina que os processe, também sejam entregues resultados. Para isso, ao invés de ter custos para transferir o dado para onde quer que seja, vamos transferir a computação. Vamos pegar a nossa “computação” e jogar para onde os dados estão, transferir a execução do código para onde os dados estão, conhecido como o princípio da localidade. Quando o Google propôs este método *MapReduce*, eles estavam pensando exatamente isso, criar uma abstração em que pudesse encaixar o algoritmo de programação a ser executado em paralelo, em um formato de *Map* e *Reduce*.

Vamos entender como o *MapReduce* funciona, a partir do exemplo clássico do contador de palavras, representado pela Figura 5.



FIGURA 5 – CONTADOR DE PALAVRA – MAPREDUCE



FONTE: A autora, adaptado de Marquesone (2016)

Imagine que temos vários arquivos contendo textos e gostaríamos de saber quantas ocorrências acontecem de cada palavra. Na Figura 5 temos como entrada dois arquivos – *Apple Orange Mango Orange GrapesPlum* e o arquivo *Apple Plum Mango Apple ApplePlum*, cada um em uma máquina diferente. E vamos executar um código *MapReduce* que conte estas palavras. Então, a primeira operação de *Map* vai executar para cada linha de cada arquivo. O primeiro arquivo vai ter uma operação de *Map* para primeira linha – “*Apple Orange Mango*” e uma operação de *Map* para segunda linha – “*Orange GrapesPlum*”. Assim como o segundo arquivo também vai ter uma operação de *Map* para primeira linha – “*Apple Plum Mango*” e uma operação de *Map* para segunda linha – “*Apple ApplePlum*”. Em seguida, a operação de *Map* vai quebrar estas linhas, em palavras, por exemplo, a linha – “*Apple Orange Mango*”, o resultado da operação de *Map* será – “*Apple, 1*”, “*Orange, 1*”, “*Mango, 1*”. Observe que está sendo associado a cada palavra um número, para cada operação de *Map* serão atribuídos uma chave e um valor. O próprio *framework MapReduce* fará isso automaticamente. E a junção de tudo que tiver a mesma chave, chamado de *SortandShuffle*, terá em seguida uma operação de redução para cada chave igual – *Reduce Key Value Pairs*.

Digamos que temos a palavra *Apple*, a chave *Apple* com o valor 01(um), 04 (quatro) vezes. Fazemos uma operação que some estes quatro valores, este será meu *reduce*, a mesma coisa com as outras palavras. No final, os *reduces* vão produzir palavras, chaves, com seus respectivos valores e vai entregar resultado – *Final Output*.

Este é o modelo do algoritmo *MapReduce*, a ideia é que você tenha uma fase de *Map* onde é gerada uma chave com valor, ele faz uma transformação, uma associação em que cada chave fica agrupada e um *reduce* para cada uma dessas chaves e gera um resultado final. Esse modelo permitiu construir algoritmos com muito mais facilidade, visto que o *framework* já entrega muita coisa pronta, a maneira de como receber os dados, a parte intermediária entre associar e ordenar chaves e a maneira de guardar os dados de forma distribuída na fase *Reduce*.

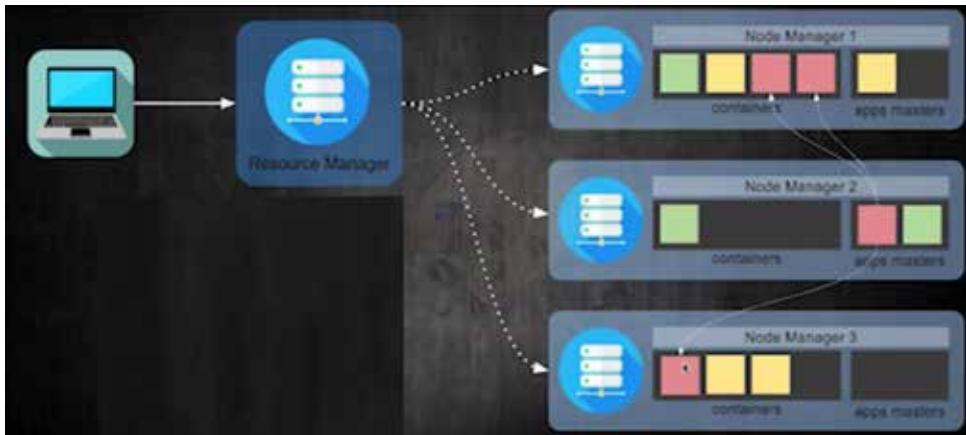
### 4.1.3 Yet Another Resorce Negotiator (YARN)

Como o YARN funciona? O YARN possui um *Resource Manager* global que conhece todos os recursos do *cluster*. E cada nó, cada máquina do meu *cluster* terá um *Node Manager* local, que irá gerenciar os recursos locais daquela máquina. Quando a aplicação quiser rodar, ela pede recursos para o *Resource Manager* e ele aloca em uma das máquinas um *Apps Masters*, que vai gerenciar,



controlar a “saúde” da aplicação, alocando recursos em cada uma das máquinas, tendo o controle de em que máquinas há recursos alocados, segundo a Figura 6.

FIGURA 6 – YET ANOTHER RESOURCE NEGOTIATOR (YARN)



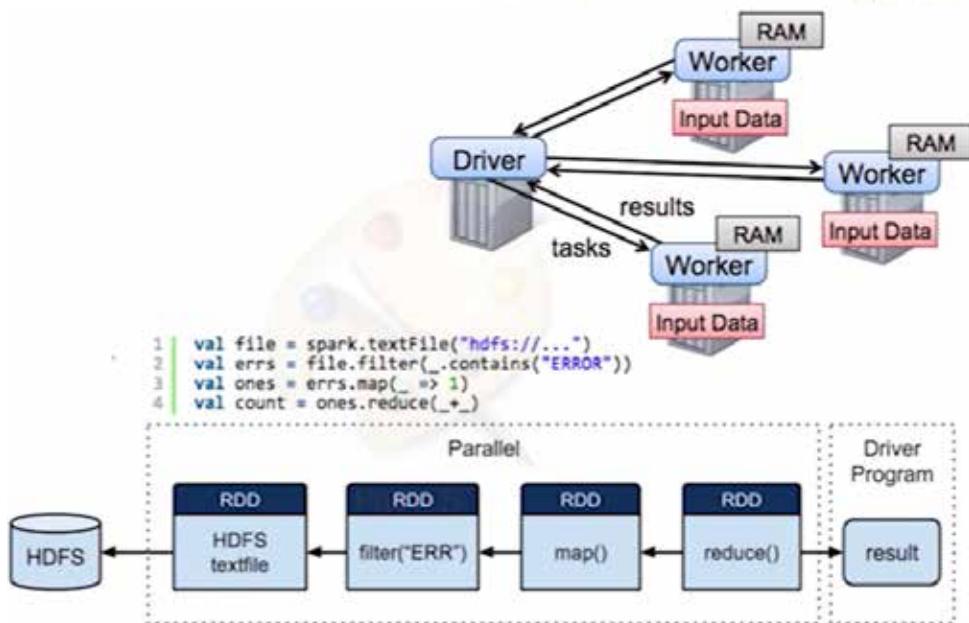
FONTE: A autora (2018) adaptado <<https://ucbrise.github.io/cs294-rise-fa16/assets/slides/Yarn.pdf>>.

No exemplo, o *Application Master* está rodando no Nô 2 e ele sabe que há recursos alocados no Nô 1 e Nô 3, auxiliando a gerenciar os recursos que estão sendo utilizados. Neste mecanismo de gerenciar recurso, tem escalonamento, saber quem vai ter direito a que recurso, uma série de processamentos para saber o que e onde vai poder rodar os recursos. Outra questão, existe o princípio da localidade, tentar rodar meus recursos onde estão os meus dados, isso vai fazer com que meus dados rodem mais rápidos.

#### 4.1.4 Spark

É uma outra das principais ferramentas utilizadas em *Big Data*, apesar do *MapReduce* ter sido algo revolucionário, ele ainda não atendia todos os problemas, sobretudo quando é preciso repetir algo várias vezes, o que implicava rodar o *MapReduce* várias vezes prejudicando sua performance. Até que pensaram em criar um *framework* tão bom quanto o *MapReduce*, mas que não colocasse tantos limites ao programador, surgindo o *Spark*. A ideia principal do *Spark* é que na hora executar o seu código paralelo, facilite ao programador construir seu algoritmo pensando no encadeamento das funções e no final é executada uma função que coleta os resultados, vamos ao exemplo da Figura 7.

FIGURA 7 – SPARK.



FONTE: A autora, adaptado de <<https://intersog.com/blog/apache-spark-vs-hadoop-mapreduce/>>.

No exemplo, deseja-se contar quantos erros há em um conjunto de *Log*. No código, na linha 1, primeiro pega-se os arquivos de *Log*, depois na linha 2 espera-se filtrar quais linhas têm a palavra “*ERROR*” e na linha 3 para cada linha encontrada “*ERROR*” é associado o valor 1, por fim, pega-se esses valores 1 e é somado todos eles.

Quase uma operação *MapReduce*, ou seja, é feito um filtro e depois uma redução somando. Porém, a característica principal que não há no *MapReduce*, que entre a fase do *Map* e do *Reduce* ele tinha que persistir todos os dados em disco para lançar para diversas outras máquinas. Já no *Spark* ele não persiste no disco, se não pedir. Ele vai trabalhar sempre em memória, ele vai primeiro carregar os arquivos em memória, os blocos em memória de maneira distribuída, depois ele vai fazer o filtro também em memória e manter os resultados em memória. Em seguida, ele vai fazer a contagem do *Reduce* em memória e só fazer a operação que você deseja no final.

Ilustrativamente, temos a primeira fase do algoritmo sendo carregados os arquivos, depois fazendo o filtro, em seguida o *Map*, o *Reduce* e se obtém o resultado. Ou seja, muito similar ao *MapReduce*, mas com a vantagem de não precisar fazer só *MapReduce*, se eu quiser selecionar só parte dos meus dados, reduz várias vezes, eu agrupo uma determinada informação, fazendo o algoritmo imaginar sem precisar ficar no formato *MapReduce*.



O *Spark* pode funcionar em paralelo ao *Hadoop*, pedindo ao *Hadoop* as máquinas que ele vai rodar e aloca o *driver* a aplicação principal do *Spark*, a aplicação começa a executar identificando código paralelo e joga para as máquinas que ele alocou, os recursos que ele alocou do *cluster*, executa estes dados em paralelo e depois pega o resultado, coleta de novo no *drive* e grava isso em disco, finaliza o seu código com o resultado que você executou. Esse é o papel do *Spark*: executar um código paralelo sem a restrição do *MapReduce*, usando o máximo de memória que você tiver.



Quer conhecer mais sobre *Spark*, consulte os livros *Learning Spark: lightning fast data analysis*, é um livro para começar a fazer análise de dados de forma rápida e poderosa. FONTE: < <http://index-of.co.uk/Big-Data-Technologies/Learning%20Spark%20%20Lightning-Fast%20Big%20Data%20Analysis%20.pdf>>. E *Advanced analytics with Apache Spark*, este livro apresenta um conjunto de padrões para executar análises em grande escala com *Spark*.

FONTE: <<http://shop.oreilly.com/product/0636920035091.do>>.

#### 4.1.5 *HBase*

Outra ferramenta bastante importante no universo de *Big Data* é o *HBase*, um grande Banco de Dados distribuído, que permite acessar grande volume de dados de maneira rápida. Ele não é uma ferramenta do *Hadoop*, mas roda em conjunto com o *Hadoop*. E como ele funciona?

O *HBase* é um Banco de Dados chave-valor, então as chaves são associadas com algum valor, com um Banco de Dados tradicionais cada linha do dado vai ter aquele mesmo conjunto de colunas. No *HBase* apesar de parecer uma tabela não é exatamente da mesma forma, não há precisamente as mesmas colunas associadas a todas as chaves, examine o exemplo da Figura 8.

FIGURA 8 – HBase

Chave	column family: p		column family: a	
	name	phone	url	
312010	Isaac Newton	555-1212	1459088892829	http://example.com/page1.html
312012	Albert Einstein	555-2531	1459088895312	http://example.com/page2.html
312017	Linus Pauling		1425403391432	http://example.com/page16.html
3120172	Friedrich Gauss	2 555-2414	1425403364329	http://example.com/page8.html
312019	Max Planck	555-3223		

FONTE: A autora, adaptado de <<https://goranzugic.wordpress.com/2016/04/11/hbase-schema/>>.

Poderíamos ter uma chave associada a uma determinada coluna e uma outra chave associada a outra coluna, com outros valores. Na linha chave “312017” do exemplo da Figura 18, não tem associado a essa chave a coluna “*phone*”. Ao mesmo tempo, não há na linha chave “312019” a coluna “*url*”. Não havendo obrigatoriedade de esquema, cada conjunto de colunas é organizado em família, e podemos otimizar essas famílias de maneira diferentes. Para achar um dado, primeiro é preciso ter uma chave, uma família de colunas e a coluna que se deseja. Veja no exemplo, temos a chave “312010”, a família “*p*”, a coluna “*name*”. Então, conseguimos identificar “Isaac Newton”.

Mas há uma questão a mais, o *HBase* quando você coloca o dado, você não altera mais aquele valor, o dado nunca é alterado. Aliás, é uma característica comum em *Big Data*, você não altera os dados, apenas adiciona mais informações sobre aquele dado. No caso do exemplo, temos um valor para o nome, um valor para o telefone, e na coluna “*url*” a relevância é assumir quando “Isaac Newton” acessou determinada página, quando isso aconteceu. Portanto no *HBase*, cada célula tem suas versões, e as versões podem ser número de versões, ou um *time/tempo*, um número que identifique um tempo em que foi acessado o dado. Conforme o exemplo, sabemos em qual instante “Isaac Newton” acessou uma determinada página, diante de suas três visitas.

Resumindo, com *HBase* conseguimos acessar rapidamente uma chave, há várias colunas associadas a essa chave e para coluna ou para cada célula, conseguimos ter versões, profundidade desta célula, versões destes valores proporcionando uma grande agilidade ao *HBase*, sendo muito bom para recuperar



dados a partir da sua chave. Mas, por outro lado, o *Hbase* não é bom para varredura completa, ou seja, quando é realizado o que chamamos de *scan*, ele não apresenta uma performance tão boa.



---

Aprofunde-se mais sobre *HBASE*, consulte o livro *HBase – The Definitive Guide, Random Access to your Planet-Size Data*, ideal para lidar com o armazenamento de uma infinidade de dados coletados por organizações.

FONTE: <<http://www.mpam.mp.br/attachments/article/6214/HBase%EF%BC%9AThe%20Definitive%20Guide.pdf>>.

---

Essas são algumas das ferramentas possíveis de se trabalhar com *Big Data*, não se limite a essas, conhecer as principais e aprofundar seus conhecimentos em ferramentas para se trabalhar com *Big Data* é essencial. É sempre importante destacar que as técnicas utilizadas para processar dados, transpor informações e analisar os resultados sempre devem ser exploradas, pois não necessariamente todos os V`s de *Big Data* estão contidos nos dados, depende dos dados com que o sistema irá trabalhar. Alguns sistemas demandam velocidade no processamento analítico, outros demandam suportar grande volume de dados, logo as ferramentas utilizadas na infraestrutura e no desenvolvimento do sistema deverão ser selecionadas conforme a sua necessidade (PINHO, 2015).



---

*MapReduce*: <[http://hadoop.apache.org/docs/r1.2.1/mapred\\_tutorial.html#Overview](http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#Overview)>.

*Hive*:<<https://hive.apache.org>>.

*Pig*: <<http://pig.apache.org/>>.

*HBase*:<<https://hbase.apache.org>>.

*HDFS*: <[http://hadoop.apache.org/docs/r1.2.1/hdfs\\_design.htm#Introduction](http://hadoop.apache.org/docs/r1.2.1/hdfs_design.htm#Introduction)>.

*Sqoop*:<<http://sqoop.apache.org/>>.

---



A Arquitetura para processamento de grandes volumes de dados apresenta características essências, implementadas por algumas ferramentas como *Hadoop*; *Hadoop Distributed File System (HDFS)*; *MapReduce*; *Spark* e *Hbase*, citadas nos tópicos anteriores. Descreva com suas palavras quais as principais características de cada uma delas.

## 4.2 BUSCAS E APRENDIZAGEM DE MÁQUINA

Com os dados preparados, passa-se para a etapa de análise, em que algoritmos robustos de mineração são rodados e técnicas de aprendizagem de máquina são realizadas. Através da mineração, se tem um direcionamento para o processo de decisão, por exemplo, encontrando padrões de consumidores e combinando aos produtos. Já a aprendizagem de máquina permite acelerar o processo de análise, por exemplo, o padrão do consumidor que comprou um determinado tentará prever o comportamento de compra do consumidor para compra de outros produtos. E da mesma forma que no processamento são utilizadas ferramentas, há soluções gratuitas e de código aberto para servidor de buscas e aprendizagem de máquina, como (APACHE, 2018):

- **SOLR**: é uma plataforma de indexação e busca otimizada para grandes volumes de dados por ser escalável e realizar balanceamento de carga de forma confiável e tolerante a falhas. Diante de suas particularidades, pode-se ter *schema* ou não, de acordo com sua configuração via *Extensible Markup Language (XML)* e suas interfaces de integração pode ser em XML, JSON e HTTP (*HyperText Transfer Protocol*). Possui também interface de administração, com controle de instâncias, buscas textuais, geoespacial, facetas entre outros.

## 4.3 VISUALIZAÇÃO E RELATÓRIOS

Para providenciar insights, deve-se ter uma boa ferramenta de visualização e relatórios, provendo informações delineadas com uma aplicação de filtro e para com apenas um clique realizar consultas na base de dados. Algumas das ferramentas sugeridas para se trabalhar com grandes quantidades de dados são:



- **JasperSoft**: gerador de relatórios de *Business Intelligence* (BI) composto por quatro projetos, sendo eles (TIBCO, 2015):
  1. **JasperReports Server**: que centraliza e gerencia os relatórios, podendo ser de modo OLAP;
  2. **JasperReports Library**: que produz documentos oriundos de qualquer tipo de dados, desenvolvida em Java;
  3. **JasperReportsETL**: um motor de integração de dados, que realiza o processo de extração e transformação para carregar dados em um *Data Warehouse*.
  4. **JasperReports Studio**: uma IDE para desenvolvimento de relatórios baseada no *Eclipse*.

Além disso, conecta-se nativamente com *Hadoop*, *Google Big Query*, *MongoDB*, *Cassandra*, entre outras fontes de dados de *Big Data*, podendo gerar relatórios e *dashboards* diretamente do local de origem ou realizando ETL (TIBCO, 2015).

- **Pentaho**: empregado na área de BI que trabalha com OLAP, principalmente fazendo processos ETL e *Data Mining* (GUIMARÃES, 2017) para disponibilizar relatórios e *dashboards* interativos que permitem realizar filtragem. A ferramenta possui interação com *Big Data*, conectando-se com banco de dados NoSQL, como *Cassandra* e *MongoDB*, além de processar em *cluster* de forma distribuída sobre o *Hadoop* utilizando *MapReduce* para alcançar melhor performance (GUIMARÃES, 2017).
- **MicroStrategy**: possibilita o acesso de grande volume de dados em múltiplos sistemas, podendo rodar sobre *Hadoop* para obter melhor desempenho, além de ser do HDFS que é uma base de dados que esta ferramenta pode acessar (MICROSTRATEGY, 2017). É uma boa ferramenta para gerar relatórios de BI, *dashboards*, realizar *drilldown* em tempo real (id).



A Qlink vem desenvolvendo um excelente trabalho quanto à ferramenta de visualização de dados, chegando a receber prêmio de Inovação da Cloudera (provedora de solução *Big Data*) – <<https://www.qlik.com/us/company/press-room/press-releases/1004-qlik-wins-cloudera-partner-excellence-award-for-innovation>>.

Use as ferramentas adequadas para aprimorar a gestão dos dados, para aperfeiçoar a qualidade na gestão de dados. E entenda primeiramente as três classes de ferramentas que são usadas dentro da estrutura de armazenamento e

processamento de dados, como relatórios e painéis de controle apresentando uma representação amigável dos dados gerados e visualização e monitoramento por meio de técnicas de visualização dinâmicas e interativas. Além da análise avançada e *analytics*, com um processamento inteligente dos dados para fins diversos.

Dentro dessas classes, há ferramentas e aplicações que otimizam a gestão das informações, como as que já vimos o *Hadoop* e *MapReduce*. E além destas, ainda pode-se destacar:

- *Hive*, *MongoDB* e *Impala*: executam comandos de SQL, aproveitando Bancos de Dados já existentes.
- *Hootsuite*, *Google Alert* e *Alexa*: ferramentas para monitoramento de redes sociais.
- *Tableau*, *Infogram* e *ChartBlocks*: ótimas ferramentas para a visualização de dados em uma interface amigável.

---

*Big Data* vem crescendo não só diante do grande volume de dados disponíveis em organizações ou por usuários, mas também quanto às produções vinculadas à área, conheça algumas indicações complementares para aprofundar cada vez mais seus conhecimentos em *Big Data*:



*Big Data Now*: <<http://it-ebooks.info/book/2170/>>.

*Big Data Glossary*: <<http://it-ebooks.info/book/823/>>.

*Big Data Analytics*: <<http://www.general-ebooks.com/book/1212371-big-data-analytics>>.

*Ethics of Big Data*: <<http://it-ebooks.info/book/1984/>>.

*Big Data for Dummies*: <<http://it-ebooks.info/book/5875/>>.

*Big Data Analytics Using Splunk*: <<http://it-ebooks.info/book/5202/>>.

*Disruptive Possibilities: How Big Data Changes Everything*: <<https://www.amazon.com/Disruptive-Possibilities-Data-Changes-Everything-ebook/dp/B00CLH387W>>.

*Real – Time Big Data Analytics: Emerging Architecture*: <<https://www.amazon.com/Real-Time-Big-Data-Analytics-Architecture-ebook/dp/B00DO33RSW>>.

*Builing Data Science Teams*: <<https://www.oreilly.com/data/free/building-data-science-teams.csp>>.

*Data for the Public Good*: <<http://shop.oreilly.com/product/0636920025580.do>>.

*Planning for Big Data*: <<https://www.oreilly.com/data/free/planning-for-big-data.csp>>.

*The Promise and Peril of Big Data*: <<https://www.aspeninstitute.org/publications/promise-peril-big-data/>>.



## 4.4 DESCARTE DOS DADOS

É importante sempre lembrar que a todo instante pode-se detectar dados que não são mais necessários e que devem ser excluídos da base, provendo a limpeza ou desativação dos dados, sendo indicado como uma fase de descarte que pode ser realizada em bloco, horizontalmente ou verticalmente (SANTOS; SANTANA, 2013), a saber:

**Bloco:** corresponde à exclusão de subconjuntos inteiros de dados identificados como entidades. Exemplificando, poderia ser a exclusão de uma entidade que contém dados de produtos inteiramente apagada.

**Horizontalmente:** eliminação de registros, elementos da estrutura de entidade, por meio de filtros específicos ou de informações relacionadas às datas a que se relacionam. Pode ser citado, em comparação com o exemplo em Bloco, correspondendo à eliminação de produtos que tenham sido cadastrados a mais de determinado período.

**Verticalmente:** seria a eliminação de elementos estruturais das entidades que definem seus atributos, o que remete à definição de dado como sendo definido pela tríade  $\langle e, a, v \rangle$ . Ainda continuando as exemplificações, seria a eliminação de um elemento estrutural desta entidade e que identificaria um de seus atributos, como “peso” e eliminaria o atributo “peso” de todos os itens cadastrados nesta entidade.

Além disso, algumas questões nesta fase são fundamentais, como (SANTANA, 2016):

- Quais dados já não são mais necessários?
- Os dados a serem descartados foram persistidos?
- Em quais suportes?
- Esses dados estão replicados em outras bases?
- Como garantir e explicitar que estes dados tenham sido realmente excluídos e não simplesmente ocultos?
- A eliminação dos dados não prejudicará a integridade ou interligação de outros dados?
- O descarte destes dados não prejudicará a qualidade do conjunto de dados como um todo?
- Tem-se o direito de excluir estes dados?
- Ao eliminar estes dados, qual o impacto em sua encontrabilidade e acesso?
- Para o descarte foi considerada a necessidade de preservação em diversos aspectos?

Acrescente ao seu Plano de Dados um plano ou Política para o descarte de dados e compartilhe essas informações com sua equipe de trabalho para melhore resultados.

## 5 ALGUMAS CONSIDERAÇÕES

Podemos notar que a potencialidade de *Big Data* ainda não está sendo utilizada de maneira correta e esperada, há um imenso volume de dados que podem ajudar resolver diversos tipos de problemas, desde a prevenção de epidemias, ao transporte, economia, entre outras. Sendo necessária uma maior dedicação a treinamentos de profissionais de Tecnologia da Informação, buscando um melhor entendimento do que é *Big Data*, de como trabalhar com *Big Data* e do seu poder de influência aos negócios.

As ferramentas de processamento, buscas e aprendizagem de máquina servem para tornar o tempo de resposta de processamento analítico mais rápido. Já as ferramentas de visualização e relatórios são essenciais para ter um resultado de análise do *Big Data*, podendo navegar em um relatório dinâmico para auxiliar na tomada de decisão.

Mas todo o entusiasmo em torno do *Big Data* gera algumas expectativas muito perigosas sobre o que o projeto pode proporcionar, por mais que seja tentador aceitar e fazer promessas a curto prazo, é importante manter uma visão realista do que se pode esperar do projeto, quanto tempo isso vai levar e a quantidade de esforços necessários para chegar lá. Portanto, defina objetivos claros e administre suas expectativas, defina as métricas que comprovam o valor do projeto, seja estratégico sobre ferramentas e codificação manual e defina com clareza as metas de seu negócio, da TI e a necessidade de dados.

## REFERÊNCIAS

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

APACHE, **Apache Hadoop.** Single node setup. 2018. Disponível em: <[http://hadoop.apache.org/docs/r1.2.1/single\\_node\\_setup.html](http://hadoop.apache.org/docs/r1.2.1/single_node_setup.html)>. Acesso em: 1º nov. 2018.

BIGTABLE: a distributed storage system for structured data. Disponível em: <<https://ai.google/research/pubs/pub27898>>. Acesso em: 13 nov. 2018.



BORTHAKUR, D. **The Hadoop distributed file system**: architecture and design, p. 3, 2007. Disponível em: <[https://svn.eu.apache.org/repos/asf/hadoop/common/tags/release-0.16.3/docs/hdfs\\_design.pdf](https://svn.eu.apache.org/repos/asf/hadoop/common/tags/release-0.16.3/docs/hdfs_design.pdf)>. Acesso em: 10 nov. 2018.

CLIMATE PREDICTION.NET. Disponível em: <<https://www.climateprediction.net/>>. Acesso em: 11 nov. 2018.

CLOUDERA. Disponível em: <<https://www.cloudera.com/>>. Acesso em: 3 nov. 2018.

DATA LAKE. Disponível em: <<https://azure.microsoft.com/pt-br/solutions/data-lake/>>. Acesso em: 8 nov. 2018.

DEVMEDIA. **Metadados** – O significado da informação no ambiente de business intelligence. Disponível em: <<https://www.devmedia.com.br/metadados-o-significado-da-informacao-no-ambiente-de-business-intelligence/7417>>. Acesso em: 4 nov. 2018.

GAVINO, David Castro. **Como obter mais valor a partir dos seus dados?** 2018. Disponível em: <<http://www.administradores.com.br/noticias/tecnologia/como-obter-mais-valor-a-partir-dos-seus-dados/123737>>. Acesso em: 5 nov. 2018.

GHEMAWAT, Sanjay; GOBIOFF, Howard; LEUNG, Shun-Tak. The google file system. In: **Proceedings of the nineteenth ACM symposium on operating systems principles**. [S.I.]: ACM, 2003, p. 29–43.

GOLDMAN, Alfredo et al. **Apache Hadoop**: conceitos teóricos e práticos, evolução e novas possibilidades. 2012. Disponível em: <[http://www.imago.ufpr.br/csbc2012/anais\\_csbc/](http://www.imago.ufpr.br/csbc2012/anais_csbc/)>. Acesso em: 7 nov. 2018.

HP. **Soluções de Big Data**. 2018. Disponível em: <<http://www8.hp.com/br/pt/business-solutions/big-data.html>>. Acesso em: 15 nov. 2018.

HURWITZ, Judith. **Big Data para leigos**. Rio de Janeiro: Editora Alta Books, 2016.

INTEL – Integrated Electronics Corporation. **Big Data 101**: unstructured data analytics a crash course on the IT landscape for Big Data and emerging technologies, 2012. Disponível em: <<http://www.intel.com/content/www/us/en/big-data/unstructured-data-analytics-paper.html>>. Acesso em: 31 out. 2018.

GUIMARÃES, Leonardo. **Know Pentaho**. O Guia. 2017. Disponível em: <<https://www.knowsolution.com.br/pentaho-o-guia/>>. Acesso em: 8 nov. 2018.

MACHADO, Alexandre Lopes. **Administração do Big Data**. 2017. Disponível em: <[https://books.google.com.br/books?id=0UZBDwAAQBAJ&printsec=frontcover&hl=pt-BR&source=gbs\\_ge\\_summary\\_r&cad=0#v=onepage&q&f=false](https://books.google.com.br/books?id=0UZBDwAAQBAJ&printsec=frontcover&hl=pt-BR&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false)>. Acesso em: 12 nov. 2018.

MARQUESONE, Rosangela. **Big Data**: técnicas e tecnologias para extração de valor dos dados. 2016.

MAPREDUCE: Simplified data processing on large clusters. Disponível em: <<https://ai.google/research/pubs/pub62>>. Acesso em: 17 nov. 2018.

MICROSTRATEGY. **Analizando Big Data no MicroStrategy**. 2017. Disponível em: <[http://www2.microstrategy.com/producthelp/10.4/WebUser/WebHelp/Lang\\_1046/Content/mstr\\_big\\_data.htm](http://www2.microstrategy.com/producthelp/10.4/WebUser/WebHelp/Lang_1046/Content/mstr_big_data.htm)>. Acesso em: 14 nov. 2018.

O GRANDE LIVRO DE BIG DATA: um guia prático para tirar o seu primeiro projeto de Big Data do papel. Disponível em: <[http://www.lcvdata.com/sist\\_distr/bigdata\\_resources.pdf](http://www.lcvdata.com/sist_distr/bigdata_resources.pdf)>. Acesso em: 10 nov. 2018.

PAPO, J. P. Big Data e Computação em nuvem na AWS. In: **CAMPUS PARTY BRASIL 6**, 2013.

PINHO, Joel Lucas. Conceitos e práticas no desenvolvimento de sistemas de recomendação. In: **The Developer's Conference**, 2015. Disponível em: <<http://www.thedevelopersconference.com.br/tdc/2015/florianopolis/trilha-bigdata>>. Acesso em: 15 nov. 2018.

ROSETTA@home. Disponível em: <<http://boinc.bakerlab.org/>>. Acesso em: 18 nov. 2018.

SANT'ANA, Ricardo Cesar Gonçalves. Ciclo de vida dos dados e o papel da ciência da informação. In: Encontro nacional de pesquisa em Ciência da Informação, 14, 2013, Florianópolis. **Anais**. 2013. Disponível em: <<http://enancib.sites.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/284/319>>. Acesso em: 16 nov. 2018.

SANTOS, Plácida L. V. Amorim da Costa; SANT'ANA, Ricardo César Gonçalves. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, Brasília, v. 42, p. 199-209, 2013.



SETI@home. Disponível em: <<http://setiathome.ssl.berkeley.edu>>. Acesso em: 20 nov. 2018.

THE GOOGLE FILE SYSTEM. Disponível em: <<https://ai.google/research/pubs/pub51>>. Acesso em: 10 nov. 2018.

TIBCO, Software Inc. **Jaspersoft business intelligence software**. 2015.  
Disponível em: <<https://www.jaspersoft.com/>>. Acesso em: 17 nov. 2018



# CAPÍTULO 3

## TECNOLOGIAS PARA BIG DATA

**A partir da perspectiva do saber-fazer, são apresentados os seguintes objetivos de aprendizagem:**

- ✓ Descrever, apresentar, explicar e transcrever as principais aplicações e tecnologias utilizadas para processar o imenso volume de dados do Big Data.
- ✓ Desenvolver competências para classificar, comparar, criticar, diferenciar, escolher e empregar tecnologias para Big Data.



# 1 CONTEXTUALIZAÇÃO

Você sabe o que significa a sigla IoT (*Internet-Of-Things*)? Ou a palavra *Machine Learning*? Sabia que elas estão sendo inseridas cada vez mais para inovação e transformações dos negócios? Isso mesmo! Essas tecnologias, mais que conectar pessoas e empresas, diminuem custos operacionais e estão disponibilizando robôs inteligentes que dividem diversas tarefas entre os espaços com os humanos. Seja em qualquer área, educação, indústria, comunicação, informática, entre tantas outras, estas tecnologias estão causando grandes impactos. Portanto manter-se atualizado sobre elas é essencial para não ficar para trás diante das constantes mudanças que ocorrem diariamente no universo dos negócios.

A IoT, por exemplo, vem complementando um espaço entre as necessidades decorrentes da evolução do mercado, das informações, dos usuários e das coisas, em uma direção que permite atender novos desafios, diante da quantidade de informação que pode ser gerada e utilizada para diferenciais competitivos nos negócios. Portanto, a análise de dados terá grandes impactos no uso em novas áreas da IoT, como cidades inteligentes, transportes e diversas outras que afetarão o futuro ecossistema global da Internet. Contextualizaremos um pouco mais sobre a IoT e sua ligação a *Big Data* em um dos próximos tópicos deste capítulo.

Já *Machine Learning* consiste em encontrar padrões úteis inerentes a um determinado conjunto de dados históricos, por meio de um modelo busca-se identificar correlações entre os valores de entrada com os valores de saída a serem previstos. Basicamente um modelo consiste em um algoritmo específico com parâmetros característicos que são ajustados ou que venha a “aprender” padrões úteis.

Quer saber mais? Aproveite a leitura deste capítulo para o aprofundamento do aprendizado em *Big Data*, IoT e *Machine Learning*.

Bons estudos!

# 2 APLICAÇÕES E TECNOLOGIAS PARA BIG DATA

Já sabemos que as soluções tecnológicas que trabalham com *Big Data* permitem analisar um volume enorme de dados de forma rápida e que ainda oferece um melhor controle para os gestores das informações. Neste último capítulo inicialmente serão citadas algumas aplicações de *Big Data*.



Citarei como primeiro caso, uma *Startup* em São Paulo, que usa *Big Data* para oferecer a um criador de gados um aplicativo de coleta de dados para dispositivos móveis. Por meio deste aplicativo, o criador pode: cadastrar todos os dados do seu rebanho, como quantidade de vacas, quais são as melhores produtoras de leite, o peso, entre outras informações. Por meio de todos esses dados coletados, processados e acurados através de mecanismos de automação, o produtor consegue obter maior qualidade dos dados.

Antes da criação do aplicativo, a *Startup* constatou que para aqueles pecuaristas que não aderiram a nenhuma outra tecnologia semelhante, torna-se muito difícil identificar a lucratividade real da atividade agropecuária. Com este problema identificado, a *Startup* viu a necessidade de oferecer uma solução para que o criador de gados pudesse ter o total controle sobre as informações coletadas da sua criação. Resumindo, se o custo por unidade animal é de X valor, quando for comprar o sal mineral, significa que Y% do custo do animal está sendo gasto nesta compra, permitindo enxergar a atividade de gestão em um nível muito maior do que já vinha sendo feito. Para isso, é feito o uso de soluções *Big Data*, em que os dados são cruzados por meio de algoritmos e mecanismos que interpretam esse grande volume de informações para que o pecuarista possa tomar suas decisões.

Partindo para outro campo, mais uma experiência interessante utilizando *Big Data* foi testada durante a Copa das Confederações em 2013, misturando futebol, torcida e toda informação não estruturada proveniente das redes sociais. A solução era capaz de ler publicações do *Facebook* e *Twitter*, inclusive com capacidade para compreender gírias. Foi criado um algoritmo capaz de analisar o grande volume de dados e buscando entender o pensamento do torcedor brasileiro, se eram negativos ou positivos, qual a relação das palavras, buscando tirar *insights* a respeito disso. O resultado permitiu o técnico saber o “sentimento” da torcida em tempo real. Atualmente a solução já se encontra disponível para qualquer empresa ou marca que queira medir sua reputação através das mídias sociais.

Ainda na área de esportes, uma outra ferramenta com uso de *Big Data* foi aplicada no mundo do tênis, essa ferramenta analisou oito anos de informações dos grandes atletas do tênis, um total de 41 milhões de pontos marcados para analisar padrões e estilos dos jogadores. Com isso alguns dos melhores *players* do mundo entenderam os treinos além da quadra. A norte-americana Serena Williams usa o *Big Data* para comparar e avaliar suas adversárias, melhorando assim suas técnicas em quadra.

Mais um exemplo de aplicações utilizando *Big Data* é o *Flutrends* da *Google*, baseado nos dados do seu buscador, a empresa desenvolveu um projeto no qual conseguiu detectar tendências de propagação de gripe, antes de números

oficiais refletirem a situação. A mesma técnica pode ser aplicada para analisar a inflação, desemprego e muitas outras coisas questões em alta discussões alimentadas em pesquisas no *Google*.

FIGURA 1 – ILUSTRAÇÃO DE APLICAÇÕES



FONTE: <<http://mediapool.fabrico.com.br>>. Acesso em: 25 jan. 2019.

Se você pensa que isso é o bastante, a mais nova tendência que também tem base o *Big Data* é a Computação cognitiva, na qual a máquina passa a reproduzir o pensamento humano. Os primeiros testes foram em um Jogo do Milhão nos Estados Unidos, o computador com o sistema cognitivo desafiou os maiores vencedores do programa e conseguiu vencer. O computador possuía tudo armazenado em sua memória. E em três segundos interpretava a questão, consultava a base de dados e conseguia achar a resposta. Na verdade, o computador não está “pensando”, o que ocorre, através de algoritmos, é que ele cria hipóteses e acha soluções. Atualmente a mesma técnica está sendo aplicada em parceria com grandes hospitais do mundo, para tratamento de doenças complexas como o câncer – visto que, analisando em poucos segundos uma imensidão enorme de dados é possível identificar um diagnóstico melhor da doença, o que qualquer médico sabe o quanto é complicado hoje em dia. Muito em breve a computação cognitiva estará ao alcance dos usuários com mais facilidades.



A aplicação de *Big Data*, de acordo com pesquisas, tende a crescer ainda mais nos próximos anos diante do avanço da área nas indústrias, no mercado e principalmente no entretenimento e hospitalidade, conforme os exemplos apresentados pela *Datapine* <<https://www.datapine.com/blog/big-data-examples-in-real-life/>>, a saber:

- **Uso de *Big Data* em Fast Food:** serviços oferecidos por Fast Foods, como *Burger King* e *McDonald's*, estão se tornando otimizados por meio de *Big Data*, as unidades estão sendo monitoradas e mudando os recursos de menu;
- ***Big Data* e *Self-service beer*:** nos últimos tempos produzir sua própria cerveja tem sido um hobby para diversas pessoas, pensando nisso a empresa israelense *Weissberger* criou um sistema que permite o autoatendimento dos consumidores. Basicamente, por meio deste sistema é possível que os visitantes do bar peguem as cervejas como self-service e através dos registros é analisado quais dias e horários da semana estão sendo mais vendidas quais tipos de cervejas, para compreender o comportamento dos clientes;
- **Decidindo o Menu:** *Big Data* vem permitindo os consumidores dialogaram diretamente com as marcas na decisão para novos sabores, um caso prático é a empresa *Tropical Smoothie Café*, que ao manter o controle dos dados dos consumidores, percebeu que um dos sabores mais vendidos era o novo *Smoothie Vegetariano*, permitindo investir em campanhas de marketing específicas.
- **Sentindo a falta:** restaurantes estão cada vez mais utilizando campanhas com frases via e-mails e que estão trazendo grandes retornos, a exemplo a campanha *We Miss You*, idealizada pela rede *Nova York* gerou uma média de 300 visitas e U\$36.000,00 em vendas, sendo um retorno sete vezes maior que o investimento.

Agora que já tem em mente algumas aplicações de *Big Data* que estão permitindo tornar nossas atividades diárias mais fáceis e práticas, que tal identificar as principais tecnologias vinculadas ao *Big Data* que auxiliam cada vez mais estas facilidades? É o próximo item deste capítulo.



No Brasil, existem diversas *Startups* com soluções tecnológicas emergentes — capazes de desenvolver tecnologias que irão impactar e transformar os modelos de negócios e sociedade no prazo de cinco a dez anos. Destas *startups*, 25% investem em tecnologias de Internet das Coisas, 20% em *Big Data* e *analytics*, 10% em realidade aumentada FONTE: <[http://convergenciadigital.com.br/cgi/cgilua.exe/sys/start.htm?UserActiveTemplate=site&infoId=46679&sid=97&utm%25252525252525252525252525252525Fmedium=>](http://convergenciadigital.com.br/cgi/cgilua.exe/sys/start.htm?UserActiveTemplate=site&infoId=46679&sid=97&utm%2525252525252525252525Fmedium=>)>.

Saiba outros cases de grandes empresas que utilizam *Big Data*:  
<http://blog.inovall.com.br/cases-com-big-data/>.

## 2.1 TECNOLOGIAS PARA *BIG DATA*

Vivemos em uma era de aplicativos, em que temos uma necessidade imensa em usufruir diariamente o máximo possível das aplicações disponíveis e que estão mudando a forma de fazer negócios. Na verdade, *Big Data* vem modificando o modelo tradicional de fazer negócios, muito mais que capturar dados, vem proporcionando a atribuição de valores à informação. Mas para isso, há diversas tecnologias que possibilitam capturar valor em um dado de forma automática e customizada para o cliente, ou seja, não basta aderir a *Big Data* simplesmente para o armazenamento de dados, é necessário, além disso, conhecer outras ferramentas de apoio para este processo de valorização da informação.

O que quero dizer é que *Big Data* possui diversas tecnologias ou *frameworks* para cada necessidade, a saber algumas das principais utilizadas:

- Já foram citadas ferramentas de fornecedores como o *Hadoop*, *Map Reduce* e *Cloudera* que utilizadas para o processamento e armazenamento distribuído, mas ainda podem ser citadas a *Hortonworks*, a *IBM BigInsight* e *Amazon Elastic MapReduce Cloud*, todas utilizadas para o armazenamento de grande volume de dados, poupano os custos de aquisição de licenças comerciais.
- Em se tratando das bases de dados não relacionais, já foram mencionadas nos capítulos anteriores ferramentas como *MongoDB*, *Redis* e *Cassandra*, acrescentamos ainda as ferramentas *Key-Value*, *Column-Oriented*, *Document Databases* e *Riak*. Tais ferramentas permitem processar grandes volumes de dados em altas velocidades e estão sendo utilizadas para construção de aplicações dinâmicas.



que apresentam dados semiestruturados em aplicações web com personalização em tempo real.

- Ainda não poderia deixar de mencionar algumas ferramentas de processamento e *streaming* de eventos complexos, como a *GemFire*, *Espetech*, *SenseiDB*, *Sensage*, *Zoie*, *IBM InfoStreams*, *uCIRRUS*, *Flume*, *Splunk* e *Smulogic*. Tais ferramentas permitem o consumo de dados em ampla escala com pesquisas, obtendo a consolidação e disseminação de dados em tempo real, possibilitando a criação de anúncios e promoções em tempo real através de *sites* e aplicativos on-line.
- Já as ferramentas como *Neo4j*, *MarkLogic* e *FlockDB* são utilizadas para o processamento de múltiplos tipos de dados, em *Graph Databases* e dados em XML.
- Enquanto o processamento de bases de dados em memória faz uso de tecnologias como *Applications and Products in Data Processing – High Performance Analytic Appliance Systems*, *VoltDB*, *Applications and Products* (SAP HANA), *SolidDB*, *QlikView*, *Membase*, *DRUID* (*Metamarkets*), *Statistical Analysis System* (SAS HPA) e *GemFire*, sendo tecnologias para implementar sistemas de consumo e *analytics* de *feeds* em tempo real.
- Finalizando, ainda temos ferramentas como *Greenplum DB*, *Teradata Aster*, *Kognitio*, *Vertica*, *ParAccel*, *Sybase IQ*, *Netezza*, *Exata da Appliance*, que são utilizadas para bases de dados analíticas e aplicações complexas para analisar dados estruturados e implementação de *Data Warehousing* paralelos.

Apesar das diversas citações de tecnologias, o importante é compreender que o ecossistema *Big Data* possibilita receber, processar, guardar e principalmente analisar os dados e informações disponibilizadas nas mais diversas maneiras, visando obter vantagens competitivas.

Pode-se afirmar que, para cada etapa de um Projeto de *Big Data*, deverá ser estudada e utilizada uma ou mais ferramentas que melhor se adequem a sua empresa e ao seu negócio. Por exemplo, na coleta de informações, a ferramenta *Open Source Apache Chukwa* pode ser utilizada para o armazenamento e controle de dados, ou também pode ser usado *Open Source* o programa *Talend*. Já para limpeza de informações, seguindo a linha *Open Source*, pode-se usar o programa *OpenRefine*. Em se tratando de mineração de dados há a ferramenta *Oracle Data Mining*. E para a análise de conteúdo programas como o *Statwing* e *BigML* são utilizados. Enquanto para visualização é possível a criação de mapas com o uso do programa *Tableau* ou a combinação de dados para criação de relatórios diretamente de seu navegador por meio do *Chartio*. Por fim, para integração dados são usados o *Pentaho* ou *Blockspring*.

Não poderíamos deixar de comentar que, além dos grupos de ferramentas, há como tecnologias *Big Data*, as Linguagens de Dados, sendo elas:

- **R**: é uma implementação da Linguagem S, foi lançada em 1995, surgiu com um propósito bem específico de facilitar as análises estatísticas e visualização de dados, de forma que fosse mais amigável para os usuários, ela surgiu com este único escopo e foi adotada inicialmente por acadêmicos e depois por empresas e público no geral, além disso, tem uma sintaxe orientada a funções.
- **Python**: inspirado na linguagem C, foi lançada em 1991, possui um foco generalista, serve desde fazer aplicações web a fazer análises de dados em algo escala e possui foco na leitura do código e de diminuir redundâncias, de produtividade, possuindo uma sintaxe orientada a objetos.
- **XPath**: essa é uma linguagem de consulta que seleciona os nós em um documento XML. Também pode ser usada para calcular valores como *strings*, números ou valores booleanos do conteúdo de um documento XML. Além de ser uma recomendação do W3C.

É sempre importante reforçar que o uso de cada tecnologia, ferramenta ou linguagem dependerá do objetivo que pretende-se alcançar e do uso, como fatores específicos, a exemplo se o processamento dos dados será em *batch* ou em tempo real.

---

Quer saber mais sobre Tecnologias para *Big Data*, consulte o livro *Handbook of Big Data Technologies* de Albert Y. Zomaya e Sherif Sakr, publicado em 2017 pela editora Springer International Publishing AG. ISBN – 13:978-3319493398.

---



## 3 INTERNET DAS COISAS (IOT)

“No mundo real, coisas são mais importantes que ideais.”

“Um ponto de encontro onde não mais apenas as pessoas usarão o computador, mas onde o computador se use, com objetivo de tomar a vida mais eficiente.”

“As Coisas”, que englobam desde sensores (temperatura, umidade, luminosidade, etc.), os objetos do nosso dia a dia (geladeiras, TVs, carros, etc.) estarão conectados entre si em rede, de modo inteligente e passarão a “sentir” o mundo ao redor e a interagir.”

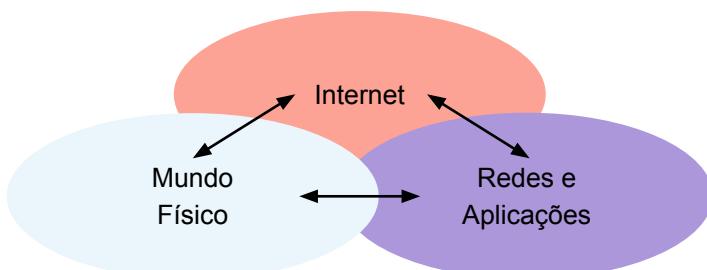
(ASHTON, K., 1999, s.p.)



A Internet das Coisas já é considerada um fenômeno mundial, que consiste na multiplicidade de dispositivos e objetos conectados à Internet. Sendo os mais diversos tipos de dispositivos, como smartphones, objeto de uso pessoal, como pulseiras que passam informações sobre a saúde do usuário, batimentos cardíacos, temperaturas, entre outros dados. Temos ainda sensores que são utilizados nas mais diferentes funções, como na detecção da abertura de uma porta, na detecção da umidade ou luminosidade de um determinado ambiente, além de objetos conectados, como geladeiras, micro-ondas, carros, entre outros.

A Figura 2 representa a interação existente em qualquer implementação para IoT, que inclui a interação entre as redes de comunicação, o mundo físico e virtual, representado pela Internet (CASAGRAS, 2009).

FIGURA 2 – INTERAÇÃO ENTRE REDES DE COMUNICAÇÃO,  
O MUNDO FÍSICO E O VIRTUAL



FONTE: Adaptado de Casagras (2009, p. 26)

O importante é compreender que todos objetos uma vez conectados à rede de comunicação estão enviando, ou seja, disponibilizando informações que podem ser coletadas, analisadas e reportadas para uso históricos e monitoradas em tempo real. E como já sabemos, com base nessas informações é possível identificar ações a serem tomadas em decisões de negócios.

O que acontece é que a variedade de objetos existentes hoje em dia apresenta diferentes tipos e formatos de informações. E essa grande quantidade de objetos representa características típicas de *Big Data*, como volume e variedade, a serem analisadas para melhores resultados e diferenciais competitivos. Além disso, grande parte dos dados destes objetos é tratada em tempo real, monitorando diversas informações, combinando o que foi visto em termo de dados coletados, eventos existentes e colocando em cadeia de tomada de decisões para que possa ser identificado ações para poder devolver um novo comando a estes objetos.

Na verdade este “monitoramento” de informações em objetos já existe há anos, por exemplo, o monitoramento de água e luz da sua residência, adivinhe o que eles são? São sensores, que existem há décadas, mas tem o problema de que são sensores que não estão ligados à Internet, é preciso que um leitorista vá a sua residência para fazer a medição. Vamos avançar um pouco mais no tempo, na indústria já tem sensores espalhados em máquinas há muito tempo também. E estes sensores estão conectados à Internet? Também não! Mas estão ligados a salas de comando e controle da própria indústria.

Você deve estar se perguntando, “E o que está mudando então para o surgimento desta nova tecnologia chamada Internet das Coisas?” É simples, os sensores baixaram muito de preço, a conectividade ficou mais fácil e abrangente, assim, a possibilidade de analisar estes dados gerados por estes sensores em um sistema de *Big Data* e de *Analytics* cresceu muito, trazendo uma maior facilidade para realizar este processo de análise de grande volume de dados estruturados e não estruturados.

E diante desta “facilidade”, alguns ganhos estão sendo possíveis de serem alcançados, por exemplo, no setor público com a IoT pretende-se atingir US\$ 4 trilhões até 2022 conforme a Figura 3:

FIGURA 3 – GANHOS COM A IoT.

EUA	585,5	Reino Unido	173,4	Brasil	70,3
China	291,5	Índia	116,2	Rússia	56,3
França	182,6	Japão	109,2	México	34,4
Alemanha	177,8	Canadá	92,8	Austrália	25,9

FONTE: CISCO CONSULTING SERVICES <[http://www.participa.br/articles/public/0030/9451/chamada\\_internet\\_das\\_coisas\\_2016.pdf](http://www.participa.br/articles/public/0030/9451/chamada_internet_das_coisas_2016.pdf)>.

Estes ganhos ocorrem pelo uso das Tecnologias nas mais diversas soluções em qualquer segmento, um exemplo simples para que você possa imaginar é uma caixa de areia de gato. Essa caixa de areia tendo um sensor acoplado para coletar o peso do gato, caso haja variações de peso detectadas pelo sensor da balança, é possível identificar o comportamento do gato, se ele está ganhando ou perdendo peso. Outro exemplo é um sistema de irrigação, através de sensores é possível saber não só se as plantas estão precisando de mais água ou não, mas também por meio da Internet saber a previsão do tempo. Assim ele saberá, se precisa molhar novamente as plantas, visto que em previsão de chuva, o excesso de água pode ser ruim para alguns tipos de plantas. Permitindo o sistema deixar parado o processo de irrigação até o período chuvoso passar.



Apesar da questão do barateamento dos sensores, a possibilidade de comunicação pela abrangência da rede, não adianta muita coisa sem ser possível realizar uma boa análise, visto que os dados por si só não são informações. É preciso captar os dados, detectar padrões, estudar tendências e transformar estes dados crus em informações, que vão virar decisões de negócio. Sendo assim não dá para falar de Internet das Coisas sem sensores taxados nas coisas, sem as Redes de Comunicação e claro sem *Big Data* e *Analytics*.



10 Previsões para Internet das Coisas (*IDC FutureScape: Worldwide Internet of Things 2015 Predictions*) <<https://cio.com.br/10-previsoes-da-idc-para-a-internet-das-coisas/>>:

1. **Internet das Coisas e a Nuvem:** próximos anos todas as indústrias terão desenvolvido iniciativas de *Internet das Coisas* (hoje 50% do foco é em aplicações de consumo, transporte e Cidades Inteligentes).
2. **IoT e Segurança:** nos próximos anos 90% de todas as redes de TI terão uma falha de segurança derivada de IoT.
3. **IoT na borda (Fog Computing):** em 2018, 40% dos dados gerados pela IoT serão armazenados, processados, analisados e devidamente usados na rede.
4. **IoT e a capacidade de rede:** nos próximos anos, 50% das redes de TI terão excesso de capacidade para lidar com dispositivos da Internet das Coisas.
5. **IoT e a Infraestrutura não tradicional:** em 2017, 90% dos *data centers* e dos sistemas corporativos de gestão irão adotar novos modelos de negócios para gerenciar infraestrutura não tradicional.
6. **IoT e diversidade vertical:** nos próximos anos todas as indústrias terão desenvolvido iniciativas de *Internet das Coisas*.
7. **Internet das coisas e a Cidade Inteligente:** em 2018 mais de 25% de toda a despesa dos governos em TI foi dedicada à implantação, gerenciamento e percepção de valor de negócio da *Internet das Coisas*.
8. **IoT e sistemas embarcados:** em 2018, 60% das soluções de TI desenvolvidas sob medida para alguns segmentos de negócio se tornaram *Open Source*, permitindo a formação de mercados verticais de *Internet das Coisas*.
9. **IoT Wearables:** nos próximos anos, 40% dos dispositivos vestíveis terão evoluído para uma massa de consumidores alternativa do mercado de *smartphones*.

10. **IoT Millennials:** em 2018, mais de 15% da população produtiva será de *Millennials*, geração (nascidos na década de 80) que vai acelerar a adoção de IoT devido à sua realidade de viver em um mundo conectado.

---

---

A *Internet das Coisas* vem mudando consideravelmente algumas áreas desde o entretenimento até saúde, entre outras, possibilitando a comunicação do mundo físico ao virtual através da Internet. Descreva com suas palavras o que é *Internet das Coisas* e qual a sua importância nos dias atuais ou cite exemplos no qual conheça.



### 3.1 INTERNET DAS COISAS E *BIG DATA*

Internet das Coisas e *Big Data* hoje são estruturas conexas, precisamos entender que essas tecnologias visam o tempo todo tornar nosso ambiente mais inteligente, no sentido de mais sensoriamentos, nos dando mais informação e de alguma forma nos ofertando maiores possibilidades de melhores soluções, serviços e produtos, gerando uma maior importância em entender o que universo *Big Data* nos proporciona, tendo mais informações, mais capacidade de conexão e mais formas de pensar e decidir, sobre como ofertar mais conforto, mais soluções, mais benefícios, em um ambiente conectado digitalmente e por mobilidade.

Portanto, temos essa potência na Internet das Coisas, a potência de extração de dados, desde que benfeita e tomado decisões de produtividade, de modelagem e levando para um universo de empreendedores. Essa é uma estrutura que no início chamávamos de indústria 4.0, uma nova revolução industrial que vem transformando tudo que aprendemos antes, não sendo aplicável da mesma maneira que antes sem ter um paramento específico, como foi a indústria do vapor ou da energia elétrica.

Enfim, vivemos um momento em que podemos ter uma condição de acompanhamento e controle de tudo a nossa volta. Imagine o cenário, posso estar no supermercado e ao mesmo tempo monitorando e conversando com uma “Inteligência Artificial” em minha casa que está controlando e gerenciando todo ambiente do meu lar, como o ar refrigerado, a geladeira, a música e a segurança. E nesse diálogo com essa Inteligência, eu pergunto “Está tudo bem em casa?” E

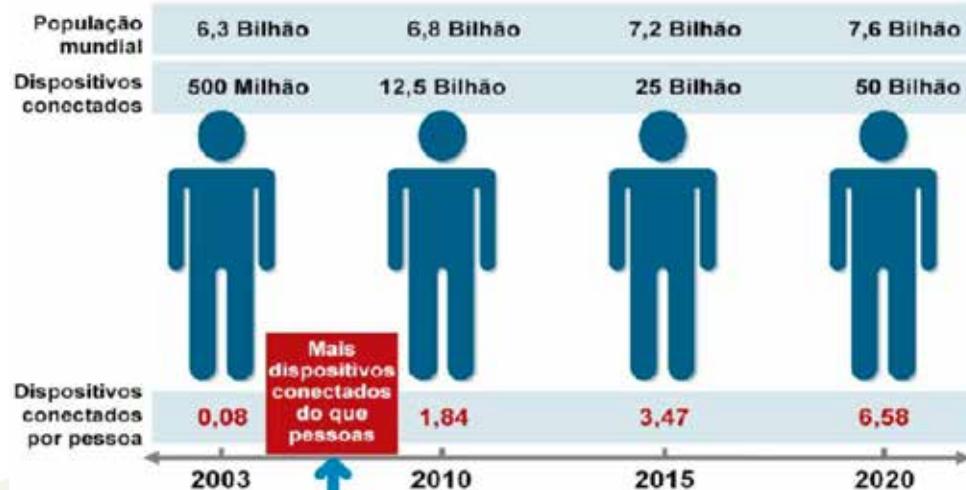


ela me responde: "Vai ficar melhor se você não esquecer a cerveja, sua geladeira está avisando que está em falta e já verifiquei em sua agenda que terá uma festa hoje. Não esqueça a cerveja!" Neste cenário, em um conceito maior, temos conforto, estrutura, informação e benefícios, que a própria sociedade de consumo está ávida para absorver mais soluções, mais serviços e uma estrutura que gera pela conexão uma inteligência na própria produção.

Consegue identificar que temos uma oportunidade gigantesca para criação de uma nova indústria muito mais inteligente, mais otimizada e dinâmica que utilizará informações disponibilizadas por estes processos para otimizar os de forma muito mais produtiva?

Apesar de o termo “Internet das Coisas” ter sido utilizado pela primeira vez somente em 1999 pelo pesquisador britânico Kevin Ashton do *Massachusetts Institute of Technology* (MIT) (FINEP, 2014). A Figura 4 representa uma previsão de crescimento de dispositivos móveis, com a ressalva que a IoT se popularizou entre 2008 e 2009 (Cisco apud EVANS, 2011):

FIGURA 4 – SURGIMENTO DA INTERNET DAS COISAS E CRESCIMENTO



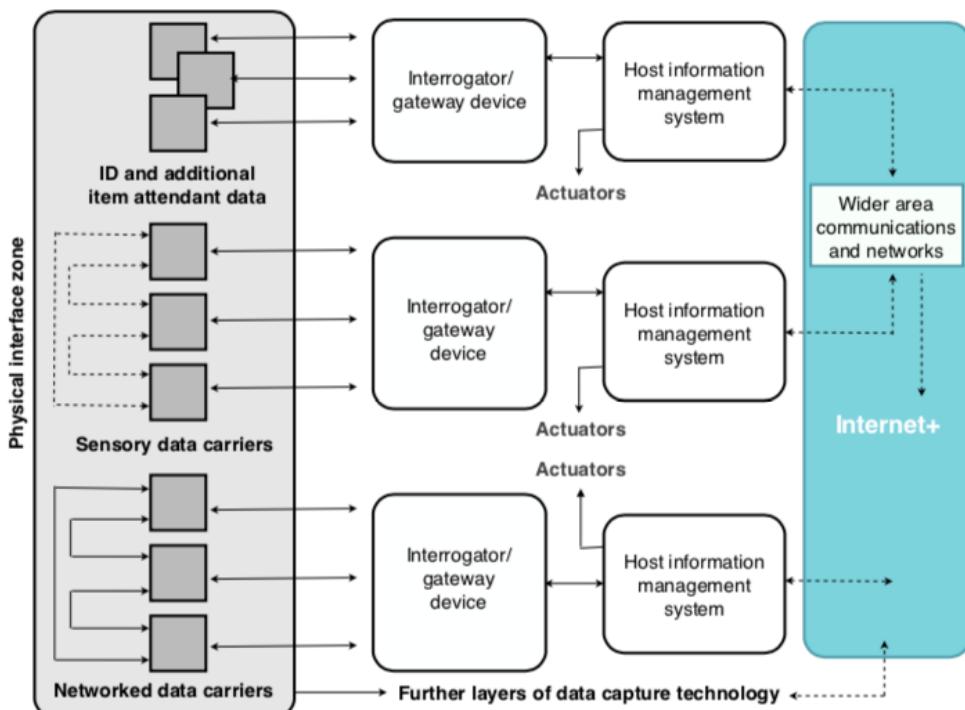
FONTE: CISCO IBSG (2015, p. 5)

Diante do crescimento exponencial dessas tecnologias, é importante compreender o processo de captura e processamento de dados em IoT, conforme a Figura 5, que ocorre da seguinte maneira:

- os dados gerados pelos objetos físicos são lidos pelo dispositivo interrogador, também designado como *gateway*. Esses dados são

enviados ao sistema de gerenciamento de informação, a partir deste, usando a Internet ou outras tecnologias de rede, são obtidos outros dados associados aos objetos ou aos próprios dados iniciais. O resultado do processamento por uma aplicação ou serviço é uma atuação no ambiente em que estão presentes os objetos e/ou uma atuação sobre eles (CASAGRAS, 2009).

FIGURA 5 – MODELO INCLUSIVO PARA A IOT



FONTE: Adaptado de Amazonas (2011, p. 27)

Conforme se discute em Casagras (2009), são propostas três classes de dispositivos para IoT:

1. Objetos puramente passivos com identificação e dados fixos, presentes na *Physical Interface Zone*, na parte superior da Figura 5.
2. Objetos dotados de moderado poder computacional e percepção de contexto, que por meio de sensores podem gerar mensagens e variar a informação associadas a eles de acordo com o tempo e lugar, representados pelos objetos presentes na *Physical Interface Zone*, mais ao centro da Figura 5.



3. Objetos que possuem conectividade em rede, sem a intervenção humana, possibilitando a emergência de inteligência nos sistemas de rede, representados pelos objetos presentes na *Physical Interface Zone*, na parte inferior da Figura 5.

O modelo ilustrado pela Figura 5 pode ser exemplificado por um processo baseado em IoT (HUANG; LI, 2010). No dado processo, um objeto do mundo real possui uma informação, por exemplo, dentro de uma tarja eletrônica. A informação é resgatada por meio de um interrogador, que pode estar conectado a um serviço ou aplicação, em que usuários ou aplicações, que estão no mundo real, poderão ler ou compartilhar a informação do objeto em tempo real. Por fim, esta aplicação pode atuar no ambiente por meio dos atuadores presentes em objetos (AMAZONAS, 2011). Cabe ressaltar que a tarja eletrônica pode ser substituída por outras tecnologias de identificação única, ou seja, não se restringe à tecnologia de RFID.

Em uma entrevista proporcionada a Revista *Computerworld*, em 2015, Hieaux afirma que a IoT, juntamente a *Big Data*, dará base para uma economia de produtos e serviços personalizados, em que os consumidores – com seus dados coletados e analisados – terão um perfil mapeado, com possibilidade de obter produtos e serviços com métricas perfeitas e únicas. Um caso prático disso pode ser referido à rede *Walmart*, que utiliza mais de 12 (doze) sistemas diferentes para processar e analisar mais de 300 mil acessos de consumidores em redes sociais – *Facebook* e *Twitter*. Outro exemplo, as Lojas *Renner* que seguem o mesmo modelo adotado pela *Walmart* comparando em tempo real as vendas das suas mais de 150 lojas.

O uso de IoT apesar de proporcionar vantagens como otimização dos processos, com a comunicação das máquinas sem a interferência das pessoas nos processos de fabricação. Pois, as máquinas sinalizem o momento em que devem ocorrer manutenções, com análise de dados históricos e uso de técnicas estatísticas proporcionando prevenções de falhas no sistema, evitado gastos desnecessários, quebra de aparelhos e consequentemente pausas na produção, trazendo grandes benefícios para as indústrias e fabricantes.

Mesmo assim, ainda há desafios em inserir novas tecnologias com a IoT e *Big Data* que tragam melhorias na operação dos negócios, como a segurança e confiança dos dados dos clientes/usuários. De acordo com a empresa de consultoria na área de tecnologia da informação Essence (2014), a IoT possui os respectivos desafios:

- As empresas necessitarão de maior disponibilidade de recursos de hardware e memória para processos em tempo real, devido ao *Big Data*

e ao crescente número de dispositivos implementados na IoT, pois com a alta da demanda a complexidade de segurança se torna um risco potencial.

- Há ainda um impacto no armazenamento de dados gerados pelas pessoas e/ou consumidores.
- A privacidade do consumidor poderá ser comprometida com o grande volume de dados trafegando com informações sobre o usuário desses dispositivos.
- A gestão de armazenamento se preocupa cada vez mais com a capacidade de armazenamento das empresas, se esta será suficiente para coletar e utilizar dados da IoT de forma eficaz em relação ao custo.
- A tecnologia dos servidores será focada em um crescente investimento em áreas essenciais e organizações relacionadas a IoT visando rendimento ou valores significativos.
- Redes de *data centers* serão modificadas para permitir alto volume de dados de sensores de mensagens pequenas para processo em *data center*, afetando assim, a largura de banda de entrada que deverá aumentar no *data center*.

### 3.1.1 Aplicações da Internet das Coisas e *Big Data*

Uma das principais áreas que a IoT vem amparando significativamente é a área de saúde, permitindo aos médicos monitorarem seus pacientes diante de doenças, como as cardiovasculares e diabetes, estabelecendo um firme acompanhamento. Como já diz o ditado “*Tempo é muito mais que dinheiro*”, ainda mais em uma UTI, no qual são necessários apenas três minutos de atendimento para salvar a vida de um paciente em parada cardiorrespiratória. E quanto mais rápido os médicos tiverem acesso às informações do estado clínico e situação real do paciente, maiores são as chances de sucesso no atendimento.

E é isso que uma plataforma de integração tecnológica trouxe para a maior UTI Cardiopneumológica da América Latina, o Instituto do Coração em São Paulo. A solução usa conceitos de Internet das Coisas, *Big Data* e Inteligência Artificial para criar uma espécie de assistente virtual, para que as decisões de médicos



e enfermeiros sejam tomadas com maior segurança e rapidez. Isso porque às vezes, ao analisar os dados de um paciente crítico, se perde muito tempo indo atrás dos dados, como exames, últimas consultas, diagnósticos, entre tantos outros. Com esses dados já inseridos em um sistema, através da Inteligência Artificial, foi possível criar uma solução que apresenta em tempo real os principais dados que fazem sentido para aquele tipo de situação do paciente. Assim, em um primeiro plano o médico, ao olhar para tela de informações, já saberá qual o paciente com estado crítico da unidade e ao teclar naqueles dados do paciente terá quais os dados responsáveis pela instabilização do mesmo.

Nas UTIs do Incor (Instituto do Coração), o sistema capta e integra dados dos diferentes aparelhos que monitoram o paciente, tudo em tempo real. Em um clique o médico tem acesso a informações, como resultados dos exames feitos pelo paciente, seu histórico desde a internação e inclusive a aplicação de medicações intravenosas. Com o uso de *Big Data* e Inteligência Artificial para analisar todas as informações em menor tempo possível, a plataforma calcula scores de riscos com a indicação de níveis de gravidade da saúde do paciente. Os dados são disponibilizados para o médico intensivista para que suporte a conduta clínica dele e também a plataforma faz uma análise preditiva em cima desses dados para prever como vão ficar os sinais vitais em certo período de tempo, com base neste histórico de dados que a solução vai capturando em tempo real.

Isso tudo é fantástico, visto que há estudos que mostram que seis horas antes de um paciente ter uma parada cardíaca, ele já dá sinais que a terá, então se há uma alteração da tendência individual daquele paciente e um diagnóstico detecta essa alteração no início, provavelmente se evita um desfecho tão grave como este. Hoje ainda que muitos hospitais tenham aparelhos conectados e até prontuário eletrônico, falta uma maneira de fazer com que todos estes sistemas conversem entre si e utilizem de forma adequada tantas informações para estudos, pesquisas e análises dos dados disponibilizados.



Estudos recentes da *Organization for Economic Cooperation and Development* (OECD) e *Food and Agriculture Organization of the United Nations* (FAO) estimam que a produção mundial precisará crescer perto de 60%, enquanto a taxa de crescimento da terra arável está prevista para cerca de 5%. Nesse cenário, a importância da automação, da otimização e do aumento expressivo da produtividade no agronegócio é fator crucial para suprirmos essa demanda. A aplicação de IoT vem ao encontro dessa tendência e envolve desde a mecanização do campo, com tecnologia embarcada para preparo das áreas de plantio, aplicação correta e uniforme de fertilizantes,

podas e colheita, até o que está sendo denominado de agricultura de precisão. Com o uso de sensores e drones, combinado com plataformas de grande volume de dados exploradas com inteligência analítica e cognitiva, tem-se todo o ferramental para a melhor tomada de decisão. As tecnologias já estão presentes e disponíveis. Um caso mais abrangente é a plataforma oferecida pela australiana *National Farmers Federation* (NFF), em que milhares de pequenos agricultores têm à disposição informações do que, de fato, está ocorrendo no campo. Podem assim acompanhar o nível de crescimento da planta, se há ou não falhas em determinada área e outros dados para a tomada de decisão. Se a questão for, por exemplo, “o custo de colocar mais fertilizantes em um determinado talhão compensa o resultado esperado?” Cruzam-se informações dos preços das *commodities*, dos preços dos insumos, das probabilidades de chuva e de diversos outros fatores para ajudar na otimização. Importante ressaltar que, como qualquer tecnologia, a IoT sozinha não traz todo o potencial de ganho, mas a combinação com as demais ferramentas (*Big Data, analytics, inteligência preditiva e cognitiva*) permite efetivamente oferecer ferramental diferenciado para possibilitar otimizar as decisões e chegar a patamares de produtividade que precisam ser atendidas pela população. FONTE: <National farmers' federation. Prime Minister Turnbull announces new initiatives to revolutionize agriculture. 2015. Disponível em: <[nff.org.au/read/5166/prime-minister-turnbull-announces-new-initiatives.html](http://nff.org.au/read/5166/prime-minister-turnbull-announces-new-initiatives.html)>.

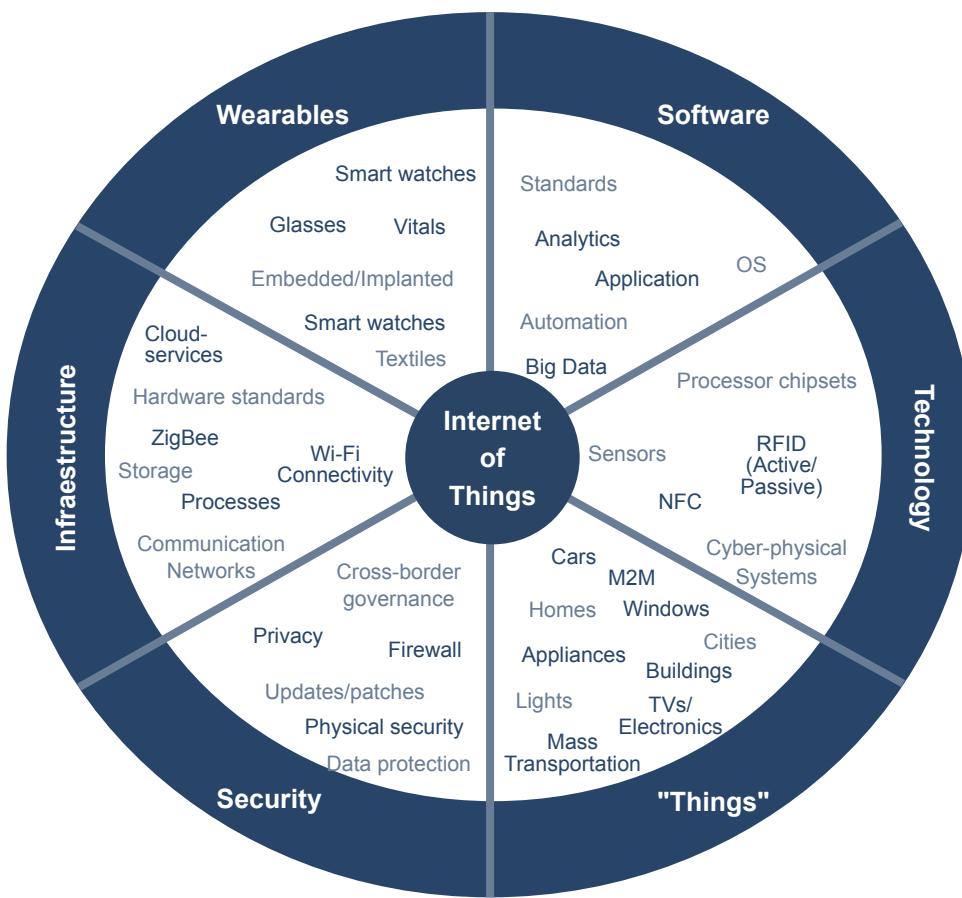
---

Alguns autores, como Ceniceros (2014), afirmam que a IoT não se resume apenas em aplicações e objetos, sensores e Internet, ou ainda não apenas em infraestrutura, tecnologia, objetos e software, como apresenta a Figura 6. Portanto, de acordo com o pesquisador ainda existem os vértices denominados *wearables*, ou dispositivos portáteis eletrônicos usados junto à roupa do usuário.

Ainda há autores, como Walport (2014), que afirmam que IoT é classificada como qualquer dispositivo ou objeto, em qualquer altura e contexto, presente em qualquer pessoa, em qualquer lugar, rede de ligação, serviço ou negócio.



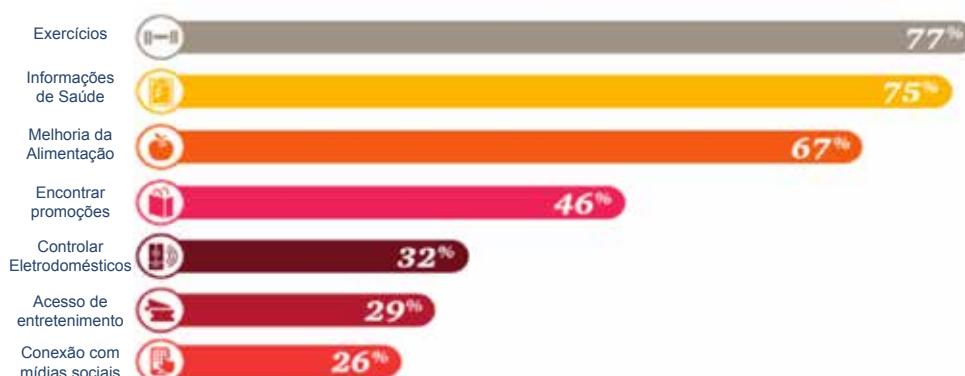
FIGURA 6 – ECOSISTEMA INTERNET DAS COISAS



FONTE: <<https://www.business2community.com/business-innovation/internet-things-ecosystem-value-greater-sum-things-0829370>>.

O importante são os resultados e benefícios alcançados diante o uso da IoT. Cita-se ainda os dispositivos vestíveis ou *wearable*, uma alternativa para a melhoria da qualidade de vida das pessoas, por conta de seus dados coletados e analisados para verificação do estado de saúde, amparando melhoria no tratamento em pacientes em monitoramento, redução de custos, tratamentos e estudos médicos. A empresa PWC, em uma pesquisa realizada nos Estados Unidos, demonstra quais informações os usuários de *wearable* gostariam de receber por intermédio dos dispositivos, para um melhor controle da saúde, conforme a Figura 7.

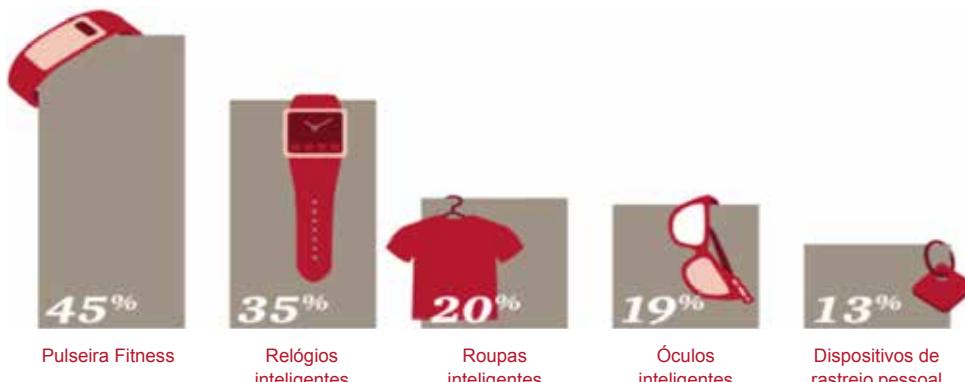
FIGURA 7 – INFORMAÇÕES QUE CONSUMIDORES RECEBERIAM POR WEARABLE



FONTE: PWC (2015, p. 11)

Além disso, verificou-se quais dispositivos *wearable* que as pessoas teriam maior interesse em adquirir, conforme a Figura 8:

FIGURA 8 – DISPOSITIVOS WEARABLES



FONTE: PWC (2015, p. 11)

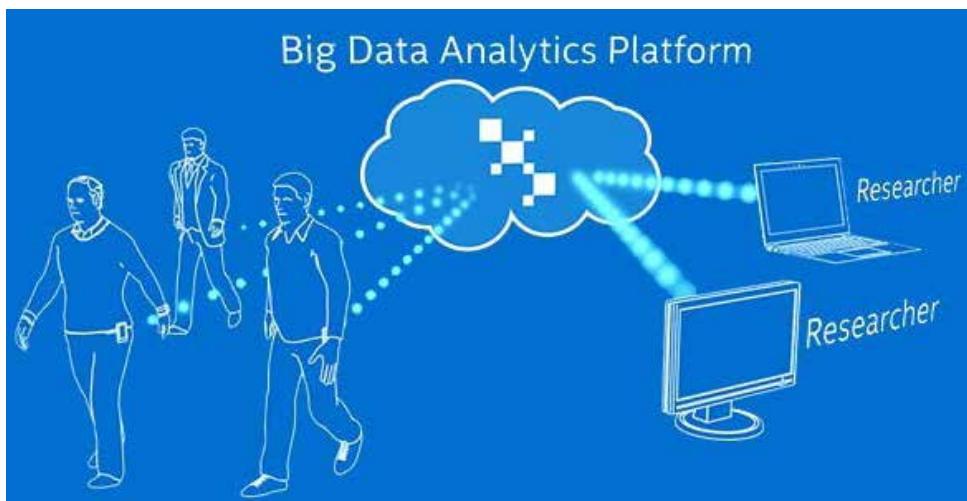
O uso da tecnologia da IoT e do *Big Data* na área médica traz como grande valia o alto número de dados fornecidos, sendo possível mapear como acontece o surgimento das doenças e como estas levam o paciente ao óbito de maneira mais rápida, porém pode haver dificuldades na forma de convencer as pessoas a fornecer os dados que precisam ser coletados automaticamente com *Big Data*.

A *Google* e a *Intel* estão tornando empresas investidoras em um alto grau para inovações criadas em busca de atender as necessidades supridas na área da saúde. Em 2015 a *Google* investiu 744 milhões de dólares para pesquisas



voltadas para saúde e ciência. Notam-se os resultados de tais investimentos em projetos como o *Iris* que objetiva auxiliar as pessoas com deficiência visual por decorrência do diabetes. A *Intel*, em parceria com a *The Michel J. Fox Foundation* (MJFF – Fundação Michel J. Fox), busca soluções por meio de pesquisas e desenvolvimento de aplicações adequadas para melhoria e tratamento da doença de *Parkinson*, realizando análises de dados e uso de dispositivos portáteis. Com a utilização de dados coletados de pacientes via *wearables*, que após analisados permitem o monitoramento dos sintomas, para que médicos e pesquisadores possam medir a progressão da doença por meio da coleta e mensuração das experiências vivenciadas no cotidiano de maneira direta dos pacientes, permitindo progredir de forma eficaz na criação de medicamentos, diagnósticos e o tratamento adequado. Na Figura 9 a *Intel* ilustra pacientes portadores da síndrome utilizando *wearables*.

FIGURA 9 – PACIENTES USANDO WEARABLE



FONTE: The Inquirer (2015, p. 13)

A utilização de sensores portáteis visa contribuir para que médicos controlem os sintomas apresentados pelos pacientes 24 horas por dia, nos sete dias da semana, atingindo assim escala maior e mais precisa dos dados do paciente. A coleta e análise de dados de milhares de pessoas portadoras de *Parkinson*, como a lentidão de movimentos, tremores e qualidade do sono, podem ajudar para que novas linhas de pesquisas apareçam e novos paradigmas sejam criados conforme os dados são disponibilizados na comunidade médica. A *Intel* e MJFF acreditam que novos tratamentos resultarão dessa pesquisa, beneficiando o setor da saúde não somente para quem possui *Parkinson*, como para outras doenças.

Já a *Apple* procura abranger novas oportunidades com uso de software, sensores e aplicativos para tratamentos de doenças independentes e integrar planos de cuidados, incluindo medicamentos e dispositivos médicos. Em 2015, a empresa lançou a linha de relógios conectados, revelando *ResearchKit* (nome dado ao projeto) de código aberto que deseja trazer benefícios aos médicos, pesquisadores e pacientes por meio da amostra de doenças, principalmente as raras e crônicas, responsáveis pelos maiores custos na área da saúde. Os dados de saúde, após reunidos, possibilitarão avaliações clínicas com uma redução de tempo e custo.

A Internet das Coisas e *Big Data* possibilitam que grandes volumes de dados sejam analisados, verificados e transformados em informações organizadas para as empresas. O uso dessas ferramentas cresceu nos últimos anos e a previsão é de que crescerá ainda mais nos próximos anos, inovando o tratamento de dados no cotidiano das empresas, fazendo com que os negócios sofram impactos em seus paradigmas. Empresas e usuários devem se preocupar com as consequências dessa nova maneira de tratamento de dados.

O autor Smith (2012) discute IoT como um aspecto crucial, considerando um mundo com objetos interconectados e invariavelmente trocando informações de todos os tipos de dados, o volume de dados gerado e os processos envolvidos em gerenciá-los na maioria das vezes é crítico. Há muitas tecnologias e fatores envolvidos no gerenciamento de dados no contexto IoT, dentre eles, os mais relevantes são: a coleta de dados, *Big Data*, Redes de sensores semânticos, sensores virtuais ou processamento de eventos complexos. No escopo da coleta e análise de dados, há a área de estrutura *Multitenant*, que destaca a descentralização, que necessita de diferentes componentes, distribuição geográfica em diferentes localizações de modo a cooperar e trocar dados, funcionalidades de mineração de dados, integrando capacidades para processamento de dados guardados, extraíndo informações úteis a partir do enorme conjunto de conteúdos armazenados (SMITH, 2012).

---

Quer conhecer mais sobre IoT e *Big Data*, consulte o livro *Internet of Things and Big Data Analytics Toward Next-Generation Intelligence*, de Nilanjan Dey et al., publicado em 2018 pela editora Springer International Publishing AG. DOI 10.1007/978-3-319-60435-0\_11.





## 3.2 INTERNET DAS COISAS E TRATAMENTO DE DADOS

Acredito que já tenha percebido o grande volume dados que são e podem ser produzidos por diferentes tecnologias IoT, para ajudar temos mais um exemplo de uma rede de supermercado que utiliza a tecnologia RFID para identificar objetos diante da necessidade de obter informações sobre a localização e o tempo para entrega de um produto conforme a sua solicitação. De acordo com o autor Cooper (2009), os dados produzidos pelas tecnologias IoT são categorizados da seguinte forma, a saber:

- Identificadores únicos dos objetos ou endereços.
- Dados descritivos dos objetos.
- Dados de Ambiente.
- Dados posicionais ou da localização geográfica.
- Dados de Redes e Sensores.

Portanto, mecanismos para gerenciamento, análise e mineração de dados na IoT são necessários. A coleta e análise de dados devem possuir capacidade para armazenar e trocar dados para análise ou para o processamento, sendo capazes de (SMITH, 2012):

- armazenar dados coletados por sensores.
- permitir que sejam adicionados novos sensores no modelo, de modo a acomodar as novas informações coletadas.
- prover API para acesso aos dados coletados.
- prover APIs para acesso em tempo real aos dados coletados, como por exemplo, mecanismos de gerenciamento de eventos (*publish, subscribe, forward e notification*).
- permitir a criação de regras ou filtros para os eventos.
- permitir ao usuário gerenciar e automatizar processos.
- permitir ao usuário criar seus fluxos de entradas de eventos vindos de um dispositivo.
- prover estruturas *multitenant* que suportem múltiplas organizações, diferentes tipos de dados, padrões e formatos, descentralização, de modo a permitir uma integração global entre arquiteturas de IoT, prover segurança e mecanismos de mineração de dados.

Já os sensores virtuais são classificados como um produto temporal, espacial ou temático que transforma um dado bruto, produzindo uma informação, por meio dos dados coletados e processados através de um conjunto de sensores, que cria

valor para um sensor virtual e que realiza a distribuição de comandos para um conjunto de atuadores (SMITH, 2012).

Outro mecanismo é o Processamento de Eventos Complexos, que depara com uma coleção de eventos a partir de múltiplas origens, detectando padrões, filtrando, transformando, correlacionando e agregando estes eventos complexos. Este processamento possui afinidade com os sensores virtuais que podem ser usados para se implementar sensores únicos, a partir de eventos complexos e múltiplos sensores ou inúmeras origens de dados. Os conceitos referentes ao Processamento de Eventos Complexos podem ser classificados em duas categorias principais:

1. *Computation oriented* Processamento de Eventos Complexos, que tem foco em execução on-line de algoritmos como uma resposta à entrada de eventos no sistema.
2. *Detection oriented* Processamento de Eventos Complexos, que tem como foco a detecção de combinações de eventos chamados de padrões ou situações.

Em Etzion e Niblett (2010) são definidos alguns conceitos básicos a respeito de Processamento de Eventos Complexos. Um Agente de Processamento de Evento (EPA) é um componente a que, dado um conjunto de eventos de entrada, se aplica uma lógica para gerar um conjunto de eventos complexos de saída. Uma Rede de Processamento de Eventos (EPN) é uma rede com uma coleção de EPAs, produtores de eventos e consumidores de eventos ligados a canais. Em Wang, Cao e Zhang (2013) discute-se que as principais ideias da detecção de eventos complexos que possuem quatro passos:

1. eventos primitivos que são extraídos de um grande volume de dados.
2. as correlações ou agregação de eventos são detectadas para se criar um evento de negócio com operadores de eventos de acordo com regras específicas.
3. primitivas ou composições de eventos são processadas para se extrair seu tempo, causa, hierarquia e outros relacionamentos semânticos.
4. a resposta é enviada para o acionador de informações de negócio, de modo a garantir a entrega dos eventos aos seus observadores.

## 4 MACHINE LEARNING

O imenso volume de dados gerado constantemente, para ser verificado diante de uma proposta de análise com *Big Data*, dependerá de um aparato de



tecnologias que possibilitam não somente processar e interpretar esse grande volume de dados, mas também permitir realizar escolhas e efetivamente aprender com a análise, com os comportamentos encontrados e identificando padrões a serem utilizados.

Apesar de este aparato de tecnologias ser guiado por uma pessoa, é impossível a mente humana trabalhar com este grande volume de dados da mesma forma que é feito por *Big Data*. Pois há uma lacuna entre a geração dos dados e o entendimento do ser humano destes dados. Assim, enquanto o volume de dados vem aumentando, a capacidade humana para compreender estes dados diminui, tornando uma grande quantidade de dados relevantes não utilizada. Além disso, a capacidade de armazenamento de dados e a velocidade de processadores não são suficientes para gerar um aglomerado de percepções que geram uma produtividade inovadora. É neste contexto que surge a tecnologia *Machine Learning*, capaz de encontrar e descrever padrões em dados.

*Machine Learning* ou Aprendizado de Máquina pode ser considerada uma forma de analisar dados com o objetivo de automatizar o desenvolvimento de modelos analíticos. De forma que o aprendizado é realizado a partir de algoritmos que aprendem de maneira interativa, ou seja, com base na análise em tempo real dos dados, os computadores proporcionam informações importantes para apoiar áreas de negócios.



*Coursera* e *University of Illinois* lançam curso de mestrado focado em *Data Science*, *Machine Learning*, *Data Visualization* e Estatísticas, acesse o site e fique por dentro de tudo. É uma excelente oportunidade para quem quer dar continuidade a estudos nesta área:

<<https://www.coursera.org/degrees/masters-in-computer-data-science>>.

O Aprendizado de Máquina ainda é compreendido como um processo de indução de hipóteses (aproximação de funções) a partir de experiência passada, ou:

“Aprendizado de máquina é a capacidade de melhorar o desempenho na realização de alguma tarefa por meio de experiência” (MITCHELL, 1997, s.p.).

Ou ainda é uma área da computação e da estatística que lida com a construção de sistemas que são capazes de aprender com os dados e

eventualmente até com suas próprias decisões. Apesar de tecnologias como IoT serem recentes. Este assunto *Machine Learning* ou Aprendizagem de Máquina existe desde o final da década de 1950.

Todavia, alguns autores afirmam que um computador consegue aprender analisando cinco definições, como (WITTEN; FRANK; HALL, 2011, p. 7):

1. Ter conhecimento de algo pelo estudo, experiência ou aprendizado.
2. Tornar ciente por uma observação ou informação.
3. Memorizar.
4. Ser informado ou averiguar algo.
5. Receber instrução.

Contudo a aprendizagem poderia ser imposta às coisas quando mudam seu comportamento de uma maneira que tenham um desempenho melhor no futuro, indo a além do conhecimento, mas também do desempenho.

Algumas áreas nas quais são aplicadas *Machine Learning*:

- Anúncios em tempo real em páginas da web e dispositivos móveis.
- Esquematização de pontuações de crédito e saber quais serão as melhores ofertas.
- Análise de sentimento baseada em texto.
- Novos modelos de precificação.
- Resultados de pesquisa na web.
- Prever falhas em equipamentos.
- Detectar invasões na rede.
- Reconhecer padrões e imagens.
- Filtrar spams no e-mail.
- Detectar fraudes.

Ficará mais clara a utilização de *Machine Learning* com alguns cases reais, citados a seguir.

O livro *O Poder do Hábito*, do autor Charles Duhigg (aproveite a citação para complementar suas leituras) apresentou de que modo a segunda maior rede varejista dos EUA, a *Target*, faz para identificar clientes que serão futuras mamães. E como isso é possível? Através do banco de dados do histórico de compras, a *Target* conseguiu identificar um padrão de itens que as mamães adquiriam quando grávidas, como sabonetes e hidratantes sem cheiro. Com a análise, a rede passou a encaminhar cupons de desconto de fraldas, carrinhos de bebê e outros produtos voltados à maternidade para essas futuras mamães, ao ponto de descobrir a gravidez até mesmo antes dos familiares.



Agora entenda, fazer sugestões de compras para os consumidores não é algo tão novo, alguns e-commerce já fazem isso há mais de dez anos. O novo neste processo é a análise preditiva, que pondera as situações das operações de consumo e suas variações, comparando com outros clientes com um mesmo perfil similar e posteriormente consegue traçar padrões de comportamento para assim conseguir fazer uma oferta no momento em que o cliente mais necessita.

Com certeza você já navegou nos sites da *Netflix* ou da *Amazon*, certo? Senão, dê uma olhadinha e observe que essas empresas são *experts* em realizar boas recomendações aos seus consumidores, seja para produtos, séries, livros ou filmes. Pois, elas sabem muito bem aplicar *Big Data* e *Machine Learning*.

Até mesmo na área de meio ambiente, há a utilização de *Machine Learning* em que satélites monitoram águas costeiras, gerando imagens diariamente para detecção de manchas de óleo, permitindo treinar um sistema detector de contaminação. Diante dos derramamentos de possíveis manchas de óleo, o sistema processa e normaliza a imagem para identificação da suspeita mancha e posteriormente são classificados atributos, como tamanho, intensidade e em qual a região foi encontrada a mancha.

Sabat (2018) cita em seu blog os seguintes passos para adoção e desenvolvimento de um processo de *Machine Learning*:

1. **Data Selection:** a assimilação dos dados que serão empregados para atingir o objetivo é essencial para o acontecimento do projeto. Este processo compreende a limpeza, seleção e adequação dos dados que serão empregados. Se você não tem os dados adequados, não há como buscar fazer previsões.
2. **Feature Selection:** escolher as características dos dados empregados é um passo muito importante. Deve-se indicar os dados menos sensíveis a ruídos e que sejam mais fáceis de serem manipulados. Nesta ocasião é realizada a divisão entre os dados que serão utilizados para treinamento do modelo e os dados para realização dos testes.
3. **Model Selection:** carece iniciar por modelos mais simples e acrescentar a complexidade se necessário. Este modelo é uma parte de uma realidade, que deve-se obter total controle sobre aquilo que aconteceu. Desta forma será possível realizar as fases seguintes de treinamento e testes identificando se o algoritmo será capaz de prever com o maior nível de assertividade.

4. **Learning:** a etapa de treinamento é bem importante para que o processo seja concluído com êxito. Aproxime os parâmetros adequados que minimize o erro do algoritmo. Entenda que o algoritmo precisa destes parâmetros e dos resultados para saber como se comportar nas fases posteriores.
5. **Evaluation:** esta é a etapa dos testes. Se o algoritmo expor um erro muito amplo de maneira inevitável será necessário rever o modelo e realizar novamente a fase 4.
6. **Application:** aplicar o modelo com dados que não se sabe o resultado. Nesta fase é previsto, espera-se acontecer e analisa se o resultado previsto bate ou não com a realidade. Se deu certo (ou próximo à realidade), passo 7. Do contrário, deve-se voltar ao passo 5.
7. **Production:** modelo validado e aplicado com sucesso, chega a hora de colocar tudo em produção.

Simples, certo? Excepcionalmente não. Os resultados normalmente são interessantes quando se chega apenas ao último passo. Porém, até lá isso pode demorar um pouco (ou muito). A vantagem é que este processo é possível de se colocar em prática. Tendo **acesso aos dados, conhecimento da técnica, do negócio** e um **objetivo claro** em mente você também será capaz de ter sucesso neste processo. Dos quatro elementos citados na frase anterior, provavelmente o mais difícil seja o “objetivo claro”. Você só terá um objetivo claro se souber fazer a pergunta certa. Para que seja possível atingir esses passos, conheça algumas técnicas de *Machine Learning*.

---

Dante do crescente gigantesco volume de dados, trabalhar somente com a mente humana para analisar dados não é mais possível. A tecnologia que vem apoiando esse contexto é a *Machine Learning*. Contextualize com suas palavras como se define essa tecnologia.



## 4.1 PRINCIPAIS TÉCNICAS PARA MACHINE LEARNING

O objetivo das técnicas de Aprendizagem de Máquina é aprender um modelo (hipótese), usando o conjunto de treinamento, para relacionar atributos da entrada a valores do atributo de saída, conforme os paradigmas a seguir.



## 4.1.1 Paradigmas de treinamento

- **Aprendizagem Supervisionada ou *Supervised Learning*:** é considerado quando o algoritmo é “treinado”, visto um conjunto de dados predefinidos. E com base nestes dados, o algoritmo consegue tomar decisões ao receber novos dados. Pode ainda ser classificados como problemas de “regressão” no qual tenta-se prever os resultados por uma saída única, através do mapeamento das variáveis de entrada. Ou também são classificados como “classificação” no qual tenta-se prever os resultados em uma saída discreta, ou seja, tenta-se mapear variáveis de entrada com distintas categorias. Diversos algoritmos são utilizados, como Máquinas de Vetor e Suporte (SVMs), Redes Neurais e Classificadores *Naive Bayes*. Um exemplo clássico é utilizado por bancos para tomada de decisões sobre se é possível liberar uma proposta de empréstimo com base no histórico do cliente.
- **Reforço:** guiado por um crítico, cada ação do sistema é seguida por um sinal de recompensa ou punição se o sistema for levado a um estado satisfatório ou insatisfatório, respectivamente. Como exemplo, pense em um veículo autônomo, no qual deverá aprender a dirigir e transportar passageiros, para isso deve-se pensar em como chegar ao destino em um tempo menor e sem causar acidentes, portanto, uma satisfação seria não causar um acidente, mesmo que não seja um tempo menor. A ideia basicamente da aprendizagem por reforço é ensinar ao computador qual ação deve ser priorizada diante de uma situação.
- **Aprendizagem Não Supervisionada ou *Unsupervised Learning*:** não tem professor externo, nem crítico, realiza a extração de propriedades estatisticamente relevantes. Ou seja, não será obtido o resultado esperado como na aprendizagem supervisionada. Não há um feedback com base nos resultados da previsão.
- **Semi-Supervisionado:** possui um professor externo apenas para parte dos exemplos de treinamento. Exemplo: mineração de dados web – professor pode fornecer exemplos de uma parcela de páginas web, mas não todos os sites da web.



É fato que fica difícil ajustar nosso tempo para estudar tudo que vem surgindo de novo, mas isso não quer dizer que você não pode pelo menos ficar por dentro das novidades da área. E uma forma para isso é seguir alguns *perfis* em redes sociais, como no *Twitter*. Portanto, vou contar para você, alguns *perfis* bacanas para não ficar de fora das novidades:

- **Hekima**: <[https://twitter.com/hekima\\_br](https://twitter.com/hekima_br)>.
- **Data Science Academy**: <<https://twitter.com/dsacademybr>>.
- **KDnuggets**: <<https://twitter.com/kdnuggets>>.
- **Ciência e Dados**: <<https://twitter.com/cienciaedados>>.
- **Data Science Renee**: <<https://twitter.com/BecomingDataSci>>.
- **Kaggle**: <<https://twitter.com/kaggle>>.
- **Dado para você**: <<https://twitter.com/dadopravocে>>.

Resumidamente a principal diferença entre supervisionado e não supervisionado é que a aprendizagem supervisionada acontece em conjuntos de dados que também capturaram o valor de saída desejado como dados de treinamento, enquanto os métodos não supervisionados tentam extrair informações inerentes, como *clusters* naturais, hierarquias ou regras de associação. Quando um modelo treinado é exibido com novos dados de entrada, ele pode prever um valor de saída, numérico ou categórico, diretamente ou como uma probabilidade. Além disso, o comum é que os tipos de aprendizagem têm como semelhança fazer o computador aprender algo com base nas suas experiências passadas.

É importante destacar também que o primeiro passo, independente da forma de aprendizagem, é determinar qual será a tarefa a ser realizada, ou seja, o que o modelo precisará prever e de que forma ele fará isso. Portanto, as respectivas respostas dependerão de qual conjunto de dados irá sustentar seu código. Lembrando que se sua base de dados não estiver com informações claras e precisas, seu modelo terá grandes dificuldades.



Nosso amigo LEO gostaria de deixar mais uma dica, agora que você já sabe um pouco mais sobre *Big Data* que tal ir a procura de um emprego na área? Onde encontrar? Segue alguns dos principais sites que estão sempre disponibilizando vagas de emprego:

**AngelList:** <<https://angel.co/>>.

**Linkedin:** <<https://www.linkedin.com>>.

**99jobs:** <<https://99jobs.com>>.

**Glassdoor:** <<https://www.glassdoor.com/index.htm>>.

**StackOverflow:** <<https://stackoverflow.com/jobs>>.

## 4.1.2 Algoritmos de classificação e regressão

Posteriormente, você deverá definir um algoritmo de previsão, resumidamente é a forma que desejamos para nosso modelo prever os dados. E para cada demanda haverá algoritmos específicos, como:

- **Naive Bayes:** ele é baseado no teorema de Bayes, que calcula a probabilidade de um evento ocorrer, dado o que outro evento ocorreu. Em se tratando de aprendizagem de máquina seria a probabilidade de uma nova entrada pertencer a uma classe, dado que ela tenha uma certa característica. Esse algoritmo funciona dividindo o *dataset* e calculando a probabilidade de cada característica de pertencer a todas as classes possíveis, as classes são os resultados possíveis do problema. Depois de calcular todas essas probabilidades é construída uma “Distribuição Gaussiana”, que representa os dados. Quando se recebe uma nova entrada é calculada qual classe essa entrada pertence a partir da distribuição.
- **K-Nearest Neighbor:** também conhecido como os K vizinhos mais próximos, é um método baseado em distâncias, utilizando uma técnica que considera a proximidade entre dados nas realizações de previsões. A hipótese deste algoritmo é que dados similares tendem estar concentrados em mesma região, no espaço de dispersão de dados. Ou seja, a intuição neste algoritmo é que os objetos relacionados ao mesmo conceito são semelhantes entre si. Portanto, este algoritmo tem como vantagem a utilização tanto para classificação como para regressão,

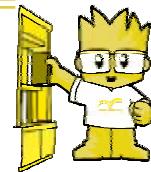
sem alterações significativas do algoritmo. Resumidamente para seu uso é preciso de um conjunto de exemplos treinados, de uma métrica para calcular a distância entre os exemplos treinados e definir o valor do número vizinho mais próximos que serão considerados pelo algoritmo.

- **Regressão Linear:** este algoritmo tem como objetivo fornecer uma previsão de certos dados de acordo com uma série histórica, que deve seguir um modelo linear. Ou seja, ele buscará encontrar a melhor linha de ajuste para modelar os dados, sendo uma linha representada por:  $y = m * c + c$ , em que  $y$  é a variável que depende de  $x$ .  $x$  é a variável independente.

Além destes, ainda podem ser citados os algoritmos para *Machine Learning*: *Support Vector Machine* (SVN), Regressão Logística, Árvore de Decisão, *K-mens*, Floresta Aleatória, Baías Ingênuas, Algoritmos de Redução Dimensional e Algoritmos de Aumento Gradiente.

---

Quer saber mais sobre *Big Data* e *Machine Learning*, consulte o livro *Big Data and Machine Learning*, Breatt S. Martin, 2018. <[https://books.google.com.br/books?id=YptiDwAAQBAJ&printsec=frontcover&hl=pt-BR&source=gbs\\_ge\\_summary\\_r&cad=0#v=onepage&q=f=false](https://books.google.com.br/books?id=YptiDwAAQBAJ&printsec=frontcover&hl=pt-BR&source=gbs_ge_summary_r&cad=0#v=onepage&q=f=false)>.



Ou ainda acesse:

- *Deep Learning* criado por membros do MIT: <<http://www.deeplearningbook.org/>>.
  - *Python Data Science Handbook*, escrito por Jake VandersPlas: <<https://github.com/jakevdp/PythonDataScienceHandbook>>.
- 

### 4.1.3 Treino, validação e teste

Pensa que acabou o aprendizado sobre *Machine Learning*? Que nada, há muito ainda que aprender, lembre-se este é um livro de Fundamentos, estamos trazendo apenas alguns dos principais conceitos e técnicas vinculados a *Big Data*. Dentre eles, não poderíamos deixar de citar Teste e Treino de dados, pois, seja em qualquer processo de aprendizado, é necessário que haja treinamento, validação do aprendizado adquirido e teste, a fim de garantir que o aprendizado expresse a realidade sendo realmente útil.



O aprendizado de máquina não difere do processo de aprendizado de nós, seres humanos. Pare e relembrre como você fazia no ensino médio ou universidade, como você aprendia? Você ia para escola ou para Universidade todos os dias, assistia à aula dos professores, adquiria mais conhecimentos lendo livros, conversando com colegas, depois você era validado de alguma forma fazendo trabalhos e testes e no final ainda realizado o grande teste ou prova final. Não foi assim que você aprendeu, ao longo de todo a sua vida acadêmica? E é provável que ainda esteja seguindo este fluxo. Portanto, o que os pesquisadores fizeram com o aprendizado de máquina foi reproduzir este processo para os algoritmos utilizando matemática para isso.

Assim, é importante parar e entender este processo da modelagem preditiva. Como já foi mencionado, tudo começa com a definição de negócio, com o problema de negócio e não apenas na tecnologia. Feito isso, os dados são coletados e irão lhe ajudar resolver o problema, por exemplo, você trabalha em um banco e esta empresa está tentando criar um novo sistema para classificação de créditos, baseado em uma série de fatores. E você recebe um cliente para análise de crédito, como será feita essa análise para fornecer ou não o crédito ao cliente, se ainda não há dados dele, se é um novo cliente? Não há dados deste novo cliente, mas já há dados de outros clientes, com uma mesma faixa etária de idade, mesmo estado civil, mesma profissão, ou qualquer outro fator que seja relevante. Assim, você alimenta o modelo, ensina o algoritmo e ele lhe dará uma previsão. Resumindo, o sistema irá lhe informar, com base no que aprendeu com dados de outros clientes, este cliente não deve receber o crédito, pois é um cliente de risco. E para que tudo isso seja feito, você precisa treinar o seu modelo.

E esse treinamento é feito normalmente com um *subset* do conjunto de dados, você não utiliza todos os dados disponíveis, você utiliza um *subset*. E um outro *subset*, que é um conjunto de dados de teste é utilizado para testar o seu modelo, ou seja, para avaliar a performance do seu modelo. Ainda é criado um outro *subset* para validação, um outro grupo de dados, também utilizado durante a construção do modelo preditivo. Geralmente não há uma regra fixa, mas uma vez coletado todo o conjunto de dados, é distribuído certa de 70% para dados de treino e 30% para dados de teste. Com isso, podemos criar *subset*, subconjuntos do seu conjunto maior de dados, afim de treinar e após testar o seu modelo.

Também é recomendado realizar a separação dos dados de forma aleatória, para tornar o teste mais confiável. Quanto à ordenação pode haver superestimação, algum ordenamento significativo. Portanto é necessário balancear as classes, tanto em treino, como em teste. Uma regra que vem sendo utilizada é quando um número de amostras, considerando n, for maior que 10.000 pode-se tranquilamente criar um conjunto de dados aleatoriamente dividido. Já quando o conjunto de dados for menor que 10.000, você deve comparar as

estatísticas básicas, como média, moda, mediana, variância. Isso vai ajudar a entender se o conjunto de testes é o não adequado.

- 
- Curso *Machine Learning* oferecido pela *Stanford*: <<https://www.coursera.org/learn/machine-learning>>.
  - Curso *Deep Learning by Google* oferecido pela *UDACITY*: <<https://br.udacity.com/course/deep-learning--ud730>>.
  - Curso *Principles of Machine Learning*: <<https://www.edx.org/course/principles-of-machine-learning>>.
  - Curso *Machine Learning for Data Science and Analytics*: <<https://www.edx.org/course/machine-learning-for-data-science-and-analytics>>.
- 



## 5 ALGUMAS CONSIDERAÇÕES

A IoT tem obtido cada vez mais atenção da indústria e da academia e *Big Data* é fundamental para IoT, primeiro há a conectividade suportando várias máquinas conversando entre si, mas não para se pensar em pessoas analisando tantos dados, esses dados para serem transformados em inteligência deve haver análise de dados, *data analytics* com *Big Data*. E ao mesmo tempo o *Big Data*, uma escabilidade enorme, as máquinas e as redes crescem de maneira gigantesca, com um volume de dados absurdo. Então as técnicas que hoje existem para se fazer esse processamento de dados de forma rápida, para semear dados e minerar coisas interessantes, dentro de uma montanha de dados que antes era ociosa e não era utilizada é absolutamente vital para que isso possa gerar novas aplicações, para que se realmente faça IoT.

Grande parte das áreas empresariais está lidando diariamente com a união de ferramentas de *Big Data* com *Machine Learning*, o que permite um desempenho ainda mais estratégico das empresas por meio deste investimento em tecnologias que reúne aprendizado de máquina e processamento de um enorme volume de dados proporcionando um conjunto de benefícios para essas empresas investidoras. Uma análise preditiva auxilia qualquer organização vender melhor em épocas específicas, evitando danos e desperdícios em épocas com menor movimentação de consumo. E a combinação de *Big Data* e *Machine Learning* é justamente importante para trazer essa clareza e confiança nas decisões de negócios, para que as empresas possam lidar com antecipações ou prorrogações de demandas.



O fato é que o mundo vem mudando de forma impressionante todos os dias, com uma avanço tecnológico exponencial de processadores, técnicas de armazenamento de memória e incremento de algoritmos para IoT e *Machine Learning*, proporciona um vasto campo de possibilidades e oportunidades para aperfeiçoar a vida humana em distintos aspectos e cabe a nós profissionais da Tecnologia da Informação conhecer cada vez mais para fazer uso destas tecnologias que estão propiciando diferenciais competitivos nos mais diversos setores.

## REFERÊNCIAS

AMAZONAS, J.R. d. A. **Network virtualization and cloud computing: iot enabling thecologies**. Casagras2 Academic Seminar, Septer 2011. Disponível em: <[http://www.casagras2.com.br/downloads/day2/2-Jose\\_Roberto\\_de\\_Almeida\\_Amazonas-Network\\_Virtualization\\_and\\_Cloud\\_Computing\\_IoT\\_enabling\\_echnologies.pdf](http://www.casagras2.com.br/downloads/day2/2-Jose_Roberto_de_Almeida_Amazonas-Network_Virtualization_and_Cloud_Computing_IoT_enabling_echnologies.pdf)>. Acesso em: 26 nov. 2018.

CASAGRAS, E. F. P. **Casagras final report: rfid and the inclusive model for the internet of things**. 2009. [https://docbox.etsi.org/zArchive/TISPAN/Open/IoT\\_low%20resolution/www.rfidglobal.eu%20CASAGRAS%20IoT%20Final%20Report%20low%20resolution.pdf](https://docbox.etsi.org/zArchive/TISPAN/Open/IoT_low%20resolution/www.rfidglobal.eu%20CASAGRAS%20IoT%20Final%20Report%20low%20resolution.pdf). Acesso em: 10 nov. 2018.

CENICEROS, M. **The internet of things ecosystem: the value is greater than the sum of its “THINGS”**. 2014. <https://www.business2community.com/business-innovation/internet-things-ecosystem-value-greater-sum-things-0829370>. Acesso em: 28 out. 2018.

ESSENCE CONSULTORIA. **Oitos desafios da internet das coisas**, 2014. Disponível em: <<http://essenceit.com/oito-desafios-da-internet-das-coisas/>>. Acesso: 29 nov. 2018.

EVANS, D. **A internet das coisas – como a próxima evolução da Internet está mudando tudo**, 2011. Cisco (IBSG), abr. 2011. [https://www.cisco.com/c/dam/global/pt\\_br/assets/executives/pdf/internet\\_of\\_things\\_iot\\_ibsg\\_0411final.pdf](https://www.cisco.com/c/dam/global/pt_br/assets/executives/pdf/internet_of_things_iot_ibsg_0411final.pdf). Acesso em: 16 nov. 2018.

HIEAUX, E. **Big Data e internet das coisas serão motores de uma nova economia**, jun. 2015. Disponível em: <<https://computerworld.com.br/2015/06/17/big-data-e-internet-das-coisas-serao-motores-de-uma-nova-economia/>>. Acesso em: 20 nov. 2018.

HUANG, Y.; LI, G. **Descriptive models for internet of things**. International Conference on Intelligent Control and Information Processing, August 2010.

REVISTA COMPUTAÇÃO BRASIL. **Internet das coisas Nós, as cidades, os robôs, os carros: Tudo conectado!** Abr. 2015. Disponível em: <[http://sbc.org.br/images/flippingbook/computacaobrasil/computa\\_29\\_pdf/comp\\_brasil\\_2015\\_4.pdf](http://sbc.org.br/images/flippingbook/computacaobrasil/computa_29_pdf/comp_brasil_2015_4.pdf)>. Acesso: 23 nov. 2018.

SABAT, S. **Big Data know how**. Disponível em: <<http://bigdataknowhow.weebly.com/>>. Acesso: 27 nov. 2018.

SMITH, I. **The internet of things 2012**. New Horizons: Technical, 2012.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3 ed. San Francisco: Morgan Kaufmann, 2011.