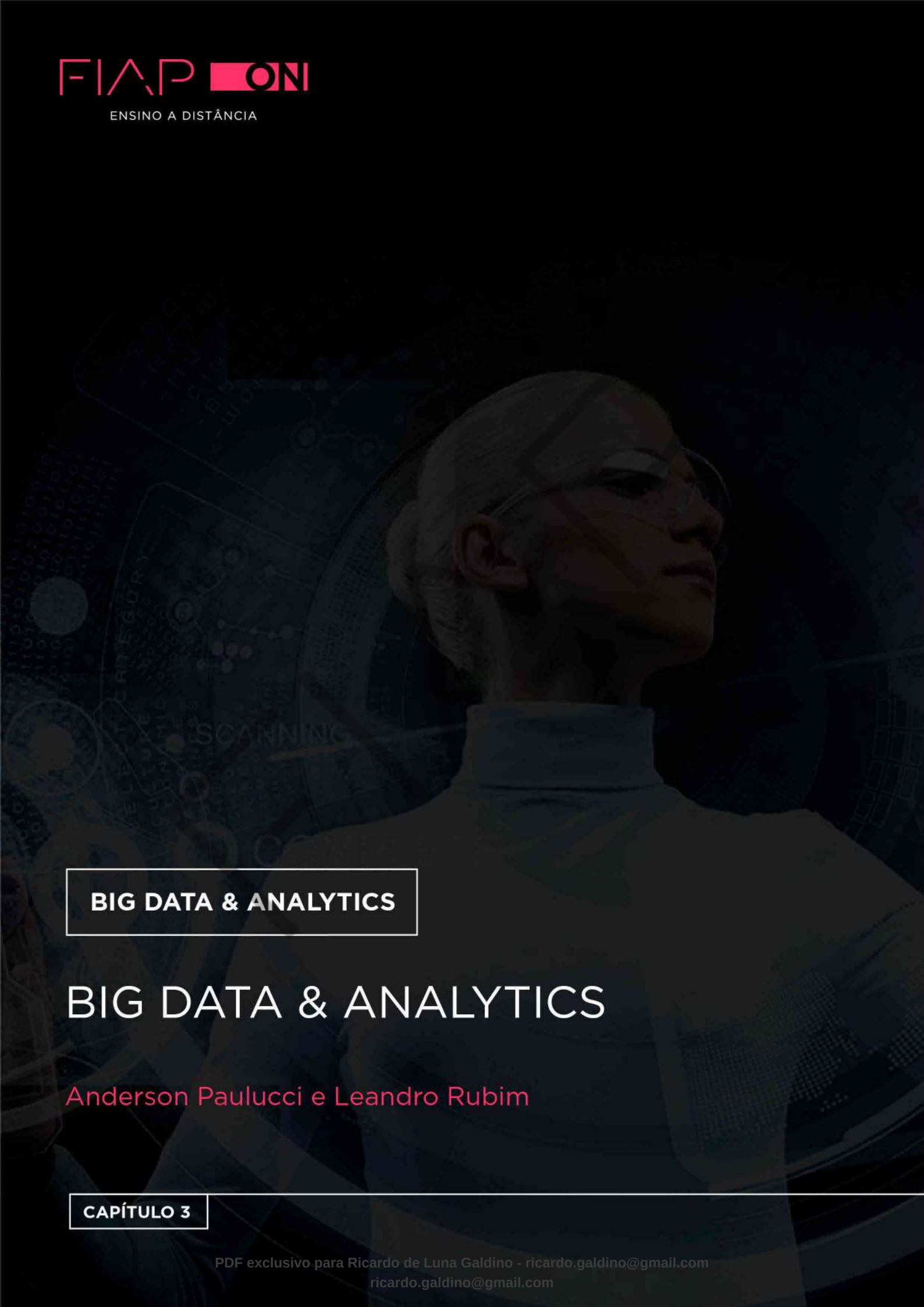
FIAP ON

ENSINO A DISTÂNCIA



BIG DATA & ANALYTICS

BIG DATA & ANALYTICS

Anderson Paulucci e Leandro Rubim

CAPÍTULO 3

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

3 BIG DATA E ANALYTICS

A arte de transformar dados em informação e desbloquear o valor dos dados.

Juntos, Big Data e Analytics (BDA) prometem transformar a maneira com que as empresas fazem negócios.

Não é difícil confundir os conceitos de Big Data e Analytics, normalmente, são temas que andam juntos. Vamos aproveitar esta abordagem de fundamentos para definir bem estes dois assuntos.

Como vimos nos capítulos anteriores, Big Data está relacionado com algumas mudanças de arquitetura da TI, impulsionadas por capacidades que excedem as tecnologias tradicionais. Dando origem a novas plataformas e modelos de dados.

A possibilidade de alavancar Analytics com o uso destas novas arquiteturas e plataformas de Big Data é a química perfeita para justificar a grande ascensão de Big Data e Analytics no mercado corporativo.

Portanto, conceitualmente existe uma separação dos assuntos Big Data e Analytics.

No futuro próximo não discutiremos com tanta ênfase as plataformas de Big Data, mas sim os algoritmos usados para Analytics.

3.1 Tipos de Análises

Utilizar Analytics de maneira correta requer um conhecimento estrutural dos tipos de análises a serem praticadas. Os tipos de análises utilizados são: Business Analytics (análise prescritiva e análise preditiva) e Business Intelligence (análise diagnóstica e análise descritiva), conforme ilustrado na Figura 3.1 e descrito no Quadro 3.1.

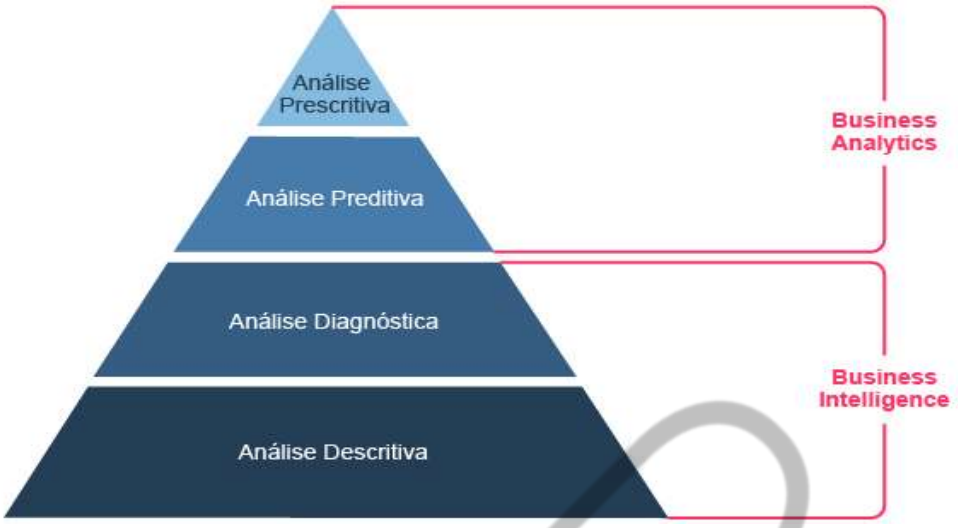


Figura 3.1 – Tipos de Análises
Fonte: Elaborado pelo autor (2016), adaptado por FIAP (2017).

Análise Descritiva	**Deve responder às perguntas**
	● O que aconteceu na minha empresa? ● Quando isso aconteceu? ● O que eu sei sobre os meus clientes, concorrentes, fornecedores etc.?
	Exemplos
	● Monitoramento Automatizado/Alertas ● Dashboards ● Scorecards ● OLAP (Cubos, Slice & Dice) ● Consulta ad hoc
Análise Diagnóstica	**Deve responder às perguntas**
	● Por que algo aconteceu na minha empresa? ● Por qual motivo estou vendendo menos?
	Exemplos
	● Relatório e Indicadores ● Dashboards ● Scorecards ● OLAP (Cubos, Slice & Dice)
Análise Preditiva	**Deve responder às perguntas**
	● O que é provável que aconteça? ● Quando acontecerá? ● O que é provável que seja verdade sobre meus clientes, concorrentes, fornecedores etc.?
	Exemplos
	● Regressões ● Mineração de dados ● Análises de grandes volumes de dados ● Simulações etc.
Análise Prescritiva	**Deve responder às perguntas**
	● O que devo fazer? ● Qual a melhor ação dado o que eu sei e o que eu acho que vai acontecer?
	Exemplos
	● Otimização ● Programação matemática (PL, MIP, QP etc.) ● Algoritmos heurísticos etc.

Quadro 3.1 – Quadro de Tipos de Análises.
Fonte: Elaborado pelo autor (2016).

Conforme podemos observar no Quadro 3.1, o desafio de otimizar os dados em busca de um valor maior da informação eleva a complexidade da análise e, consequentemente, a necessidade de Business Analytics e Advanced Analytics.

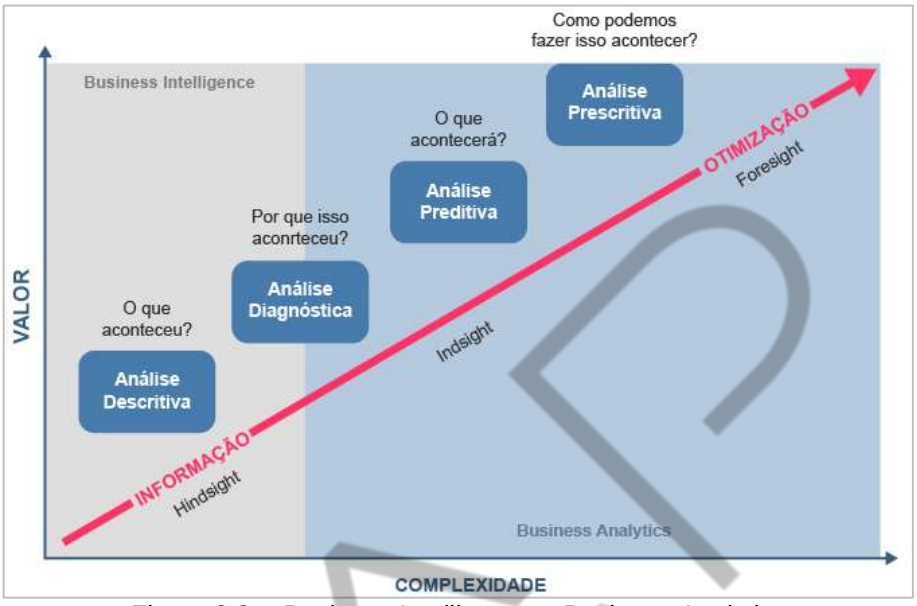


Figura 3.2 -- Business Intelligence e Business Analytics
Fonte: QPH (2015).

3.2 Business Intelligence

Segundo o Gartner:

> **Business Intelligence (BI)** ou Inteligência de Negócio é um termo abrangente (‘guarda-chuva’) que inclui aplicações, infraestrutura, ferramentas e as melhores práticas que permitem o acesso a análise de informações para melhorar e otimizar decisões.

Está associado com o processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte à gestão de negócios.

Fazendo uma analogia a um cenário de guerra com o uso de algumas técnicas milenares, aprendemos que para vencer na guerra é preciso dominar o conhecimento de suas fraquezas e virtudes, assim como as fraquezas e virtudes do inimigo. Sun Tzu, autor do livro “A Arte da Guerra”, ressalta que a falta deste conhecimento e capacidade de reagir com base nas informações podem resultar em derrota.

Figura 3.3 – Livro “A arte da guerra”
Fonte: Google Images (2017)

Trazendo esta visão para o mundo dos negócios, podemos afirmar que as “ameaças” e oportunidades fazem parte das batalhas diárias das empresas, e quanto maior o domínio do conhecimento, mais chances de vencer as batalhas. Portanto, continuamos com os desafios de:

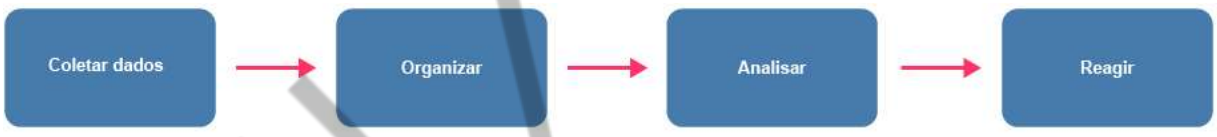


Figura 3.4 – Desafios do Big Data
Fonte: Elaborado pelo autor (2016), adaptado por FIAP (2017)

Big Data não é um novo modelo de Business Intelligence, não é especificamente uma tecnologia e não substituirá os conceitos de BI.

Com a adoção de Big Data, as empresas ganham um grande impulso, potencializando a Inteligência de Negócio (BI); são capazes de coletar mais dados, com mais velocidade e organizá-los para agilizar e priorizar a tomada de decisões mais complexas, atender o *time to market* e aumentar a capacidade de reagir aos cenários de análises avançadas com preditividade e ação.

3.3 Business Analytics

Business Analytics (BA) refere-se às competências, tecnologias, práticas de explorações iterativas contínuas e análises de desempenho empresarial. É um processo científico de transformação de dados em *insights* para tomar as melhores

decisões. Business Analytics é usado para Data-Driven ou tomar decisões baseadas em fatos.

Diferente do BI, Business Analytics tem um foco maior em tirar proveito de dados estatísticos e quantitativos para a modelagem explicativa e preditiva. Concentrando-se em soluções que criam valor para converter informação em conhecimento. Compromete-se com análises complexas, desbloqueando tendências “invisíveis”, sobre comportamentos de compras, padrões operacionais, oportunidades de negócios e assim proporcionando mais *insights* para a diferenciação no mercado cada vez mais competitivo e integrado.

Figura 3.5 – Insight
Fonte: Banco de dados Shutterstock (2017)

Os novos *insights* podem ser usados como entrada para as decisões humanas ou tomar decisões totalmente automatizadas.

Tanto no mundo acadêmico quanto nas empresas é comum a divergência sobre o seguinte entendimento: se Business Intelligence é um subconjunto de Business Analytics ou o inverso. O fato é que o termo Analytics é mais contemporâneo e acrescenta análises sofisticadas envolvendo modelos matemáticos e estatísticos.

E assim podemos usar a definição para cada tipo de análise.

Exemplo: se considerarmos a necessidade de extrair relatórios padrões ou *ad hoc, queries, scorecards*, alertas, podemos adotar o termo análise descritiva para o conceito de Analytics.

Segundo PhD. Pedro de Souza (Um guru de Analytics):

> Os conceitos de Big Data e Análise Preditiva são usados no mercado alternadamente, mas não deveriam. Big Data fornece dados em um nível muito baixo de granularidade, com detalhes que podem ser estatisticamente significativos. Antes da era do Big Data, analistas de negócio precisavam agregar os dados, a fim de ter pontos suficientes para analises preditivas com razoável confiança. Agora podemos desenvolver previsões no nível individual do cliente, sem a necessidade de agregar os dados.

3.4 Advanced Analytics

Segundo o Google Trends, o termo Advanced Analytics começou a ganhar aderência em 2009.

Figura 3.6 – Google Trends - Advanced Analytics
Fonte: FIAP (2017).

Podemos notar que o termo Advanced Analytics é bastante novo, assim como Big Data, mas poderia ser diferente, afinal já tratamos análises avançadas há décadas. E o que mudou?

Big Data impulsionou as técnicas já conhecidas para Analytics com uma velocidade impressionante, e sem dúvidas, é uma prioridade nas empresas, sendo o segmento que mais cresce no BI. As análises avançadas não estão mais limitadas a alguns domínios como: marketing, fraudes ou riscos, temos a capacidade de analisar tudo considerando maior granularidade dos dados e grande volume. Trata-se de um termo geral, que significa, simplesmente, aplicar várias técnicas de análises avançadas de dados, para responder a questões ou solucionar problemas.

Portanto, técnicas de submeter algoritmos matemáticos e estatísticos avançados na análise de dados, principalmente em grandes volumes, definem o principal objetivo do uso de Advanced Analytics.

Com a adoção do Advanced Analytics, as empresas iniciam um novo ciclo baseado em Data-Driven Business, ou seja, o negócio guiado por dados.

Toda a gestão executiva da empresa passa a tratar os dados como um ativo estratégico e os projetos que visam à coleta e obtenção de dados serão facilmente justificados.

As equipes são preparadas para análises de dados intensivas e passam a agir em conformidade com os dados. As decisões estratégicas e operacionais passam a ser sustentadas por dados e devem apresentar análises relevantes.

De acordo com um artigo publicado na Harvard Business Review, com o tema “Making advanced analytics work for you” (Fazendo o advanced analytics trabalhar para você): chegou o momento de definirmos uma abordagem pragmática para Big Data e Advanced Analytics, focada em como usar os dados para tomar as melhores decisões.

Figura 3.7 – Analogia a utilização da análise de dados em tomada de decisões
Fonte: Banco de imagens Shutterstock (2017).

Para isso, propõe que as empresas devem ter três capacidades mútuas:

- Devem ser capazes de identificar, combinar e gerenciar múltiplas fontes de dados.
- Capacidade de construir modelos de análise avançadas para prever e otimizar os resultados.

- A estratégia de dados deve possuir musculatura para transformar a organização, para que os dados e modelos realmente produzam as melhores decisões.

As empresas devem se perguntar repetidamente: "Qual é o modelo menos complexo que iria melhorar o nosso desempenho?"

Por exemplo, um modelo preditivo com 50 variáveis pode explicar os dados históricos com alta precisão, mas gerenciar tantas variáveis irá esgotar as capacidades da maioria das empresas.

Empresas com esta maturidade começam a aplicar processos de governança contínuos, que dedicam a apurar a qualidade dos dados, com importância semelhante a uma empresa de manufatura que depende de uma norma ISO no sistema de produção. Afinal, dados são as matérias-primas mais valiosas para as empresas no século XXI.

Figura 3.8 – Analogia à equipe dedicado a fazer a apuração da qualidade dos dados
Fonte: Banco de imagem Shutterstock (2017).

Não é trivial extrair o valor dos dados e orientar os negócios baseado em dados, para isso será necessário um investimento em pessoas que serão responsáveis em comandar o Advanced Analytics, nos próximos capítulos vamos abordar as principais competências para os novos perfis profissionais.

3.5 Ciências de Dados (Data Science)

A primeira universidade de pesquisas da América, Johns Hopkins, inaugurada em 1876, é referência em pesquisas e, consequentemente, domina a "ciência de dados". Gilman (primeiro presidente e fundador) acreditava que o ensino e a pesquisa caminham lado a lado e que o sucesso de uma depende do sucesso na outra. Ele também acreditava que compartilhar o nosso conhecimento e descobertas ajudaria a tornar o mundo um lugar melhor.

Na Era Digital, podemos afirmar que as soluções de Big Data, habilitaram as gigantes da internet a quebrarem as barreiras da informação. Segundo a visão do fundador da universidade, Johns Hopkins, o compartilhamento do conhecimento e descobertas são a base para a revolução digital. As grandes bases de conhecimentos foram criadas de forma colaborativa – veja os cases do Wikipedia, Google e Twitter.

Os grandes projetos de Big Data definitivamente surgiram de projetos de pesquisas, como o case do Google, ambos Larry Page e Sergey Brin estudantes, doutorandos na Universidade de Stanford, criaram o motor de buscas PageRank, cuja relevância de um site era determinada pelo número de páginas, bem como pela importância dessas páginas, que ligavam de volta para o site original. O principal objetivo era criar a maior base de inteligência artificial, e vem sendo atingido com sucesso ao longo dos anos. Este avanço contribuiu muito para nos posicionarmos hoje como uma sociedade digitalizada, um termo complementar à Era Digital.

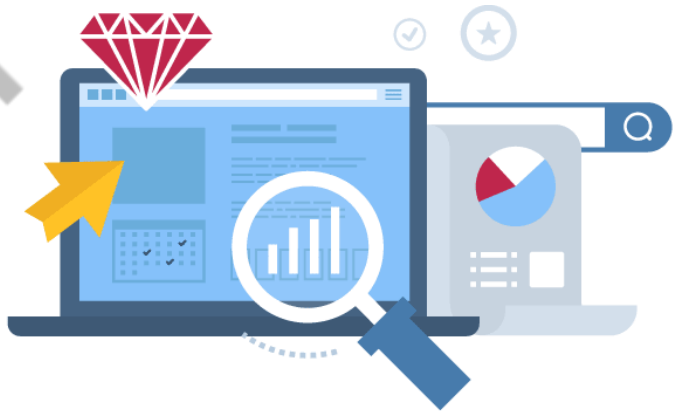


Figura 3.9 – PageRank
Fonte: FIAP (2017)

A importância do compartilhamento que Gilman se referia quando fundou a primeira Universidade de pesquisas da América agora está atingindo um grau de evolução jamais previsto, o Google se tornou uma grande fonte de conhecimento e centralizou poder de análises e descobertas. As pessoas não se intimidam com as dúvidas sobre história, medicina, economia etc. Com apenas uma simples consulta (*Discovery*) é possível ter acesso indexado e teoricamente "qualificado" ao conhecimento.

Figura 3.10 – Consultando o Google
Fonte: Banco de imagens Shutterstock (2017).

Desde o início na computação, nós passamos a ganhar aproximadamente um ano adicional de expectativa de vida, a cada ano, o avanço do poder computacional e a capacidade de resolver problemas complexos com o apoio da tecnologia é um dos grandes fatores de evolução.

Praticamente todas as empresas e setores da economia têm acesso a um grande volume de dados. Nossa capacidade de gerar valor econômico e social depende da capacidade de extrairmos o valor dos dados, este é o maior desafio.

A ciência de dados habilita a criação de produtos dados, e o futuro pertence a empresas e pessoas que transformam dados em produtos.

O artigo "The Future of Data Analysis", publicado na Universidade de Princeton por John W. Tukey, em 1962, abordava com ênfase a importância da estatística e da matemática para a ciência de dados no futuro (este futuro a que Tukey se referia, chegou).

Data Science (ciência de dados) poderia ser uma mera descrição para a estatística que, por sua vez, é um subconjunto da matemática que trata da coleta, análise e interpretação de massas de dados numéricos. A palavra-chave não é

dados e sim ciência. Com certeza, é mais complexo enfatizar a ciência do que os dados. John Tukey citou a seguinte questão: "A combinação de alguns dados e um desejo por uma resposta não garante que uma resposta razoável pode ser extraída de um determinado conjunto de dados."

Às vezes, é fácil descobrir uma estrutura de dados, novos *insights* ou grafos em um conjunto de dados. Com grandes volumes haverá milhares de possibilidades de correlacionar dados por diversos motivos, transformar isso em valor com questões interessantes é o grande desafio.

A “nova” ciência de dados, impulsionada por Big Data, está sendo conduzida pela grande massa de dados para tomar decisões e fazer previsões, não simplesmente a interpretação de números, como tratávamos na estatística convencional.

Alguns problemas da ciência de dados atual:

- Medicina Personalizada
- Genomas
- Computação para ciências sociais
- Indústria 4.0
- Neurociência
- Astronomia
- Nanociência etc.

A seguir, apresentamos um diagrama simplificado que define a intersecção de áreas multidisciplinares que definem a ciência de dados, algumas teorias abordam Data Science como um unicórnio, devido à dificuldade de compor uma formação tão ampla e complexa.

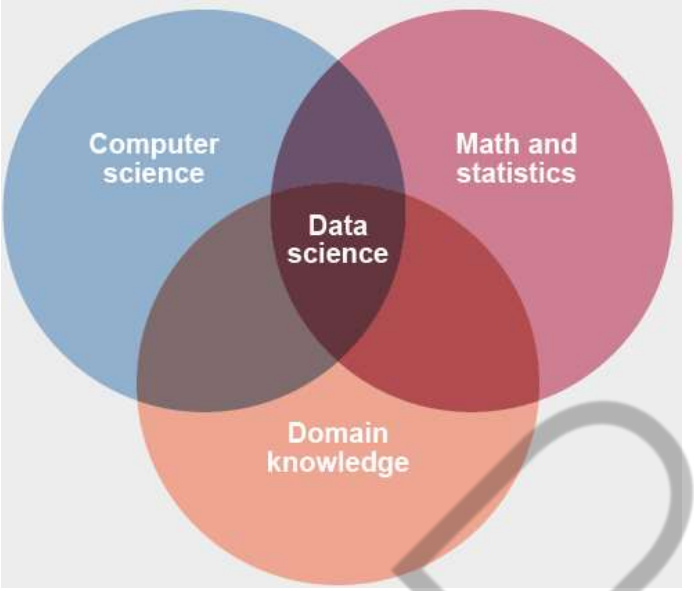


Figura 3.11 – Ciências de Dados
Fonte: IBM, adaptado por FIAP (2017).

A ciência de dados é responsável pelo avanço do Advanced Analytics, e uma das mais importantes técnicas é o desenvolvimento de algoritmos para aprendizado de máquina (Machine Learning) que vamos abordar nos próximos capítulos.

3.6 Abordagem Tradicional

Aprendemos a definir uma estrutura para os dados seguindo regras bem definidas. Na década de 1970, após um projeto de pesquisa no laboratório da IBM conduzido por E. F. Codd, resultou na publicação de um artigo que propunha a criação de um RDBMS, a IBM avançou no projeto desenvolvendo o SEQUEL (*Structured English Query Language*), e este conjunto de facilidades para manipulação dos dados se expandiu com a Oracle, entre outras empresas. No ano de 1986, foi padronizado para o SQL ANSI (American National Standarts Institute).

Este modelo predomina até hoje como um padrão na manipulação de dados, à medida que deixamos de usar a arquitetura de segunda geração baseada em cliente-servidor e passamos a adotar um novo modelo de arquitetura em nuvem. A computação exige novos padrões e a modelagem relacional começa a ser um ofensor para algumas implementações. Nosso foco aqui será em relação à visão analítica, portanto, não vamos abranger todo o impacto operacional que inclui uma nova abordagem com NoSQL (Not Only SQL), por exemplo.

Figura 3.12 – Analogia a Arquitetura em nuvem
Fonte: Banco de imagens Shutterstock (2017).

O Modelo Entidade Relacionamento (MER) aplicado a transações e à Modelagem Dimensional usada para banco de dados analíticos, determinam algumas regras importantes com os seguintes objetivos:

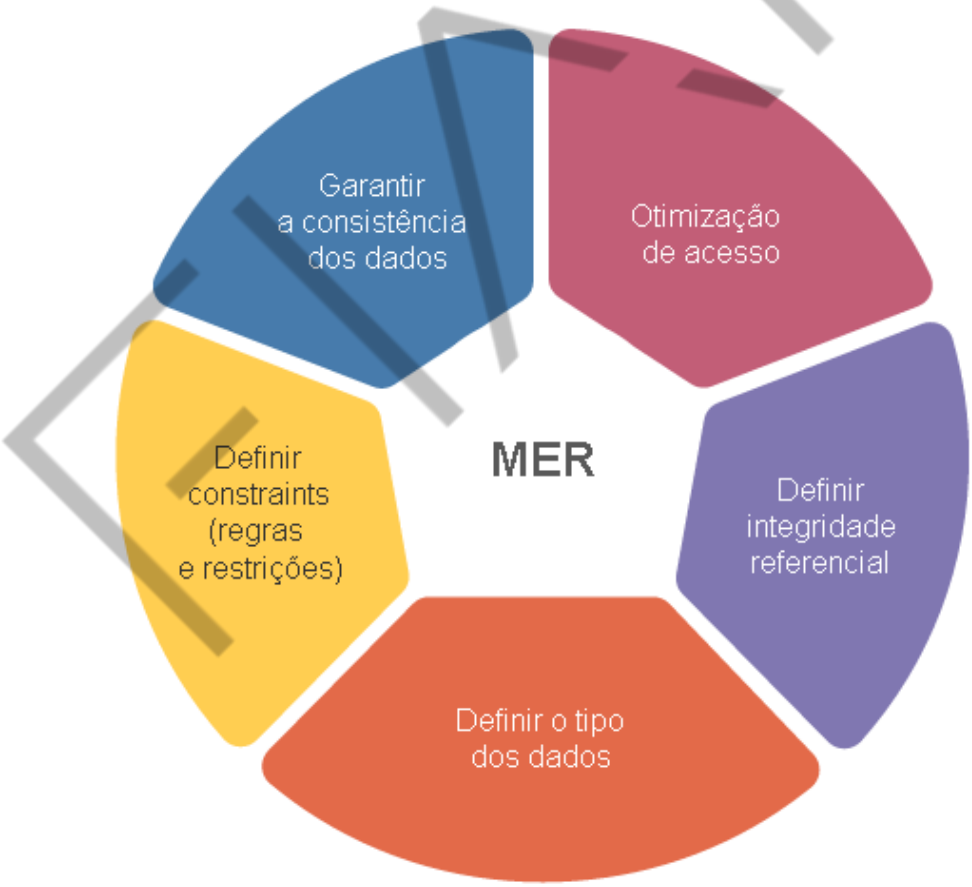


Figura 3.13 – Objetivos de regras importantes do Modelo Entidade Relacionamento (MER)
Fonte: FIAP (2017).

Aplicamos as técnicas de modelagem para a criação de um esquema bem definido, assim podemos iniciar o armazenamento dos dados, delegando a responsabilidade de integridade e consistência dos dados para o modelo e RDBMS.

Desta maneira, as *constraints* (restrições) criadas serão garantidas (exemplos):

- **Data Type**: Colunas do tipo *string* só devem ser preenchidas com valores do tipo *string.*
- **Nulabilidade**: Colunas do tipo Not Null, deverão ser preenchidas, do contrário, o registro não será armazenado.
- **Integridade Referencial**; O registro filho (chave estrangeira) não deve ser orfão, ou seja, se existe um registro de pedido para o cliente na tabela de pedidos, este cliente deverá obrigatoriamente estar registrado na tabela de clientes.

Em relação às regras impostas e, principalmente, considerando as características das aplicações cliente-servidor (ex.: ERP, CRM etc.) que dependem deste modelo e alimentam a base analítica. Podemos garantir uma precisão nos dados para a tomada de decisão. Por exemplo, indicadores de vendas por região, faturamento por produtos e assim por diante.

Porém, com todas as regras e burocracias do modelo, será muito complicado conseguirmos armazenar mais dados.

Vamos considerar a necessidade de adicionar um novo atributo (coluna na tabela). Qual seria o processo?

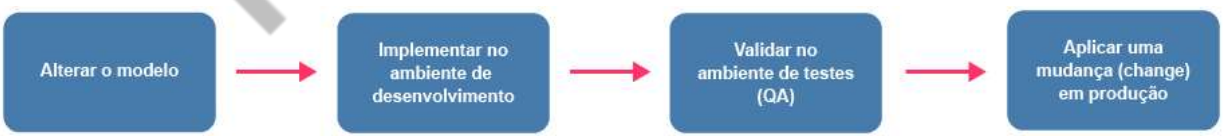


Figura 3.14 – Processo de adicionar um novo atributo.
Fonte: FIAP (2017)

Em um cenário extremamente otimista, podemos considerar a possibilidade de armazenar este atributo na base, após dois dias aproximadamente.

Manutenções de databases com grandes volumes costumam ser muito mais críticas, considere executar uma operação de ALTER TABLE, para inclusão de uma

nova coluna na tabela com bilhões de registros, diariamente. A rigidez do esquema bloqueia a evolução do modelo, dificultando nosso objetivo de receber mais dados.

A dinâmica dos dados e o *time to market*, exigem mais flexibilidade e agilidade, neste ponto, justificamos a necessidade de trabalhar com dados não estruturados. Aceitando novas semânticas criadas com o avanço dos padrões da web, como a notação JSON baseada em Java Script e amplamente adotada para manipulação de dados na web.

É importante ressaltar que a abordagem de Big Data não substituirá o uso dos bancos de dados relacionais que possuem propriedades como o ACID, fundamentais para alguns tipos de aplicações.

Porém, as discussões para banco de dados analíticos estão mais avançadas, e a aplicação dos conceitos que aprendemos sobre os 3Vs de Big Data exigem uma nova abordagem de padrões e tecnologias para a manipulação de grande volume de dados.

A arquitetura cliente-servidor criou aplicações relativamente estáveis adotando estruturas “monolíticas” (subsistemas de armazenamento compartilhados). Este realmente se tornou um problema para o armazenamento de grandes volumes conforme definimos nos capítulos anteriores. Estruturamos um esquema de dados (SQL) que poderá responder a algumas perguntas, predefinidas antes da modelagem. Mas o fato é que não podemos ter todas as perguntas antes de conhecer melhor os dados, principalmente quando estamos trabalhando com grandes volumes, variedade e dados pouco estruturados.

Nossa conclusão é que a abordagem tradicional aplica técnicas importantes para garantir a precisão dos dados, porém cria uma rigidez que limita a possibilidade de armazenar e processar mais dados.

3.7 Abordagem Big Data

Agora que aprendemos a importância de armazenar mais dados, podemos aplicar a abordagem de Advanced Analytics neste fundamento.

Qual seria a efetividade do Waze se poucas pessoas estivessem conectadas no aplicativo?

Qual seria a precisão dos algoritmos de recomendação do Netflix se houvessem poucos assinantes interagindo com o conteúdo?

Figura 3.15 – Assinante navegando no Netflix
Fonte: Banco de imagens Shutterstock (2017).

A captura de métricas necessárias para compor as análises estatísticas depende de colaboração e dados. Quanto mais, melhor!

Na era da informação as pessoas passaram de receptoras para emissoras de conteúdo e estão inundando as grandes plataformas de dados. Isso muda a abordagem tradicional, cada indivíduo ganha poder e voz para falar de igual para igual com empresas e marcas, e de maneira colaborativa capacitando as plataformas no aprendizado com seus dados.

Ao navegar pela rede, deixamos um rastro digital com informações importantes:

- Geolocalização
- Como consumimos os produtos, mídias e conteúdo
- Como tomamos as decisões de compras
- Como influenciamos nossa rede de relacionamentos
- Porque amamos ou odiamos as marcas
- O que procuramos

A questão é como transformar todos estes dados relativamente não estruturados em valor?

Em primeiro lugar, precisamos aceitar uma abordagem de dados mais “complexa”, pensando fora da caixa da visão SQL tradicional para alguns cenários.

Vamos analisar um cenário (hipotético) de uma empresa de logística que possui o desafio de reverter alguns indicadores negativos de sua operação que estão ocasionando graves prejuízos financeiros, após uma variação da moeda e redução dos valores de transportes. Esta empresa opera com uma variedade de meios de transportes (aéreo, marítimo, ferroviário, rodoviário) com um volume grande de cargas e clientes. E o mercado está exigindo um SLA (latência) para uma entrega cada vez mais agressiva.

Figura 3.16 – Modelos Analíticos - Logística
Fonte: FIAP (2017)

Vários especialistas de logística avaliaram o problema (desafio) e algumas ações foram aplicadas, mas apresentaram resultados pouco satisfatórios.

Um acadêmico (especialista em análises de cenários complexos) foi chamado para analisar o caso e trabalhou durante algumas semanas, com o objetivo de entender um pouco melhor a logística que não era sua especialidade. Após algumas analises e semanas de trabalho, propôs a implementação de alguns modelos baseados em grafos para otimizar processos. E os resultados foram extremamente positivos, posicionando novamente a empresa com margens de lucros reais.

Mesmo não sendo um profissional da área de logística, ele conseguiu aplicar uma visão matemática baseada na teoria dos grafos para guiar um novo rumo de decisões, mais eficientes e precisas para o negócio.

A **teoria dos grafos** é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto. Considerando esta definição, podemos notar muita semelhança com a álgebra relacional (teoria matemática aplicada no SQL), que de certa forma também trata da relação de determinados conjuntos. A

principal diferença é o fato de não ser baseada em SQL (portanto, não requer a modelagem relacional) e assim é possível tratar relacionamentos complexos.

Estamos criando bases de dados complexas, que demandam Advanced Analytics, e não podemos limitar os dados a uma análise SQL.

Considere a seguinte análise, para outro exemplo de grafos consultando a base do Linkedin. Você deseja encontrar o seguinte perfil:

> Pessoas que trabalharam em multinacionais de tecnologia por mais de 5 anos, foram promovidas 2 vezes no mínimo, estudaram em universidades americanas, falam inglês e espanhol fluente, já tiveram experiência no mercado financeiro, moram na Europa, fizeram negócios com a China nos últimos 2 anos e influenciam o mercado de tecnologia da informação com novas tendências usando a rede social.

Figura 3.17 – Home do Linkedin
Fonte: Linkedin (2017).

A análise de ciências sociais (sociometria) já é estudada há décadas, mas nunca foi empregada em um terreno tão fértil como estamos tratando agora. Os dados se proliferam seguindo um dos pilares da nova TI baseada em Social Network e demandam uso de técnicas intensivas de estatística e teoria dos grafos, por exemplo.

É notável a dificuldade de responder perguntas ou trabalhar análises como estas que requerem relacionamentos complexos, usando banco de dados tradicionais. Definitivamente, não foram criados para este tipo de modelagem. Para isso, precisamos de mais flexibilidade com o uso de banco de dados grafos ou estruturas de armazenamentos que permitam construir grafos com mais facilidade.

Usamos exemplos de grafos, mas devemos estender para outras estruturas de dados, considerando modelagens com chave-valor, documentos (XML, JSON) e armazenamento de dados colunares. A variação traz uma grande agilidade para sustentarmos armazenamento e processamento de grandes volumes sem o alto impacto de processos burocráticos para analises de textos, sentimentos, áudios, imagens, vídeos e qualquer tipo de dados.

3.7.1 O Tsunami Big Data

Os bebês começam a aprender com o estímulo natural da vida, recebem um bombardeio de novas percepções e esta avalanche de informações “não estruturadas” vão se encaixando no seu intenso aprendizado, até que algumas coisas começam a fazer sentido.

A velocidade com que aprendemos nesta fase é muito maior do que a capacidade que teremos nas próximas fases da vida.

Já nos primeiros meses e anos de vida, nós temos uma impressionante capacidade de aprender, porém poucas serão as memórias que guardaremos de nossas histórias quando crianças e muito menos quando bebês.

Às vezes, algumas situações na vida provocam traumas que os médicos definem o diagnóstico como possível evento que ocorreu na infância. Por exemplo, queimou a língua com a mamadeira quente e nunca mais gostou de bebidas muito quentes.

Imagine se tivéssemos a capacidade de armazenar todos os eventos de nossa vida, ao menos um resumo a cada dia. Aqueles processos de regressões com terapeutas seriam relativamente simples para resolver a questão do trauma, porém ainda assim não teríamos capacidade de analisar e processar todo o volume ao longo do tempo, pois faltaria “throughput” (velocidade) para tanta informação, a “latência” de uma sessão de terapia poderia se estender por dias, meses (M-1). Isso é natural, nosso projeto humano foi desenvolvido para definir a harmonia perfeita na linha do tempo, entre a capacidade de armazenamento e processamento. Afinal, somos formados por uma estrutura monolítica.

Com a capacidade de supercomputadores, os cientistas podem fazer importantes descobertas usando grandes bases de dados, para isso, sabem a importância de distribuir o processamento para atingir o objetivo em um tempo aceitável, o famoso dividir para conquistar.

O uso de supercomputadores (clássicos) não é uma solução viável para as empresas. [1]Apesar de possuírem algumas características semelhantes às soluções de Big Data, definitivamente não são soluções financeiramente acessíveis, seu roadmap é pouco evolutivo e não foram desenvolvidos com os conceitos de computação de terceira geração, por exemplo, flexibilidade e agilidade, considerando a dinâmica necessária para o cenário corporativo.

O Google direciona o roadmap de Big Data (tecnologias), há mais de dez anos, de certa forma estamos introduzindo as soluções de Big Data no nosso ambiente corporativo, que foram usadas em larga escala nos últimos anos pelos gigantes da internet. O princípio base do Google desde o início é simples, armazernar mais e mais dados. No início, para alimentar seu algoritmo core (PageRank) e atualmente para vários serviços do Google. Outra lição importante é aprender sempre com os dados e da mesma maneira que uma criança aprende, quanto mais estímulos (dados) e velocidade, melhor.

O filme *Transcendence - A Nova Inteligência* retrata bem isso, e apesar de ser considerado um filme de ficção, não acho que está tão longe assim da nossa realidade. Ele retrata uma história com o ator Jonny Depp, que faz o personagem de um pesquisador de inteligência artificial que se esforça para criar uma máquina que possui sensibilidade e inteligência coletiva na internet, para tornar-se poderoso. O sistema criado evolui rapidamente, aprendendo com os dados e para isso é alimentado com maior capacidade (servidores e datacenters), seu poder computacional e inteligência analítica nunca para de evoluir de forma exponencial, para isso, depende de maior quantidade de dados. E realmente em um grande toque de ficção, foi capaz de transcender a sua vida para o sistema, mapeando sua mente e conectando seus neurônios ao grande cluster.

[1] O site www.top500.org lista o *ranking* dos maiores supercomputadores, em sua maioria, instalados em universidades e centro de pesquisas.

Figura 3.18 – Cena do filme Transcendence
Fonte: Google Images (2017)

Voltando para a nossa realidade corporativa, as empresas na era digital precisam ser capazes de armazenar e analisar mais dados.

As tecnologias antes limitadas a supercomputadores de alto custo e inviáveis para as empresas, assim como arquiteturas de soluções tradicionais que não escalam e também restritas ao custo, foram tecnicamente superadas e com custos bem reduzidos, pois foram criadas a partir de soluções *open source*, com *hardware commodities* para atender empresas que buscam uma economia de escala bem agressiva.

A grande diferença é que podemos armazenar grandes volumes e processar "no tempo do negócio". O que pode ser considerado um diferencial hoje, será apenas um modelo padrão, em breve.

3.7.2 Analytics e precisão

Vamos considerar os seguintes cenários (hipotéticos) para entendermos a importância de uma mudança de visão para análise de dados com Big Data:

- Cenário 1: Na cidade de São Paulo um sensor bem localizado na avenida Paulista coleta os dados para medição da qualidade do ar a cada 60 segundos. Este equipamento é extremamente preciso e dificilmente falha.
- Cenário 2: No segundo cenário, temos sensores que não são muito confiáveis comparados ao primeiro, porém custam 50 vezes menos. E

colocaremos estes sensores em 50 pontos espalhados pela cidade, coletando as mesmas métricas a cada 60 segundos.

Considerando os dois cenários apresentados, qual terá capacidade de fazer uma análise mais precisa e garantir menor risco de falhas que gerem indisponibilidades ou atrasos na análise?

O cenário 2 é mais complexo, porém podemos ter um modelo analítico mais abrangente, para isso, vamos demandar mais volume e, consequentemente, mais dados. Talvez alguns sensores possam falhar entre uma medição e outra, mas podemos resolver isso facilmente com estatística e trabalhar com uma precisão maior no perímetro preciso de cada localização. A possibilidade de trabalhar com vários pontos de coleta e análise distribuídos, também reduzirá possíveis impactos de indisponibilidade caso alguns sensores de medição apresentem problemas, aumentando assim a resiliência de todo o sistema.

Este exemplo deixa clara a importância de trabalhar com mais dados, aceitando cenários um pouco mais complexos, e ajuda a entender por que os grandes cases como o Waze e Netflix, por exemplo, aumentam a sua precisão com ajuda dos algoritmos, quando alimentados com mais dados.

De fato, aumentar o volume é importante, porém a eficiência dos algoritmos depende de técnicas avançadas para análises, que vão além da teoria dos grafos. Trabalhando com orientação estatística/matemática podemos treinar os algoritmos para agirem sem serem explicitamente programados, usando aprendizado de máquina.

É necessário aprender com a experiência dos clientes, e assim um novo conceito chamado “**Customer Experience Management**” está se tornando prioridade para as empresas.

O Gartner define Gerenciamento da Experiência do Cliente (CEM) como: "A prática de conceber e reagir a interações com os clientes para atender ou exceder as expectativas dos clientes e assim, aumentar a satisfação do cliente, lealdade e defesa."

É uma estratégia que requer mudanças no processo e muitas tecnologias para realizar.

Assim como tratamos a importância do volume e o uso de algoritmos avançados baseados em estatística e matemática, devemos destacar também a necessidade do fator velocidade (que compõe os 3Vs de Big Data).

Em 2008, enfrentamos uma epidemia causada pelo vírus H1N1, fato conhecido como “A Gripe Suína”. Os órgãos públicos de saúde trabalham os dados buscando uma precisão nas bases de dados dos centros de atendimentos. Os pacientes que apresentam os sintomas são examinados e após alguns exames e período de observação é evidenciada a doença. Suponhamos que essa informação é computada naquele centro de atendimento e consolidada a cada 7 dias. O instituto estadual consolida os dados após mais 7 dias do recebimento de todas unidades de atendimento e disponibiliza o índice de contaminação do estado para o governo federal.

3

Figura 3.19 – Vacina para o vírus H1N1
Fonte: Banco de Imagens Shutterstock (2017).

Qual a precisão da análise e abrangência dos dados, na esfera federal do governo, para entender a disseminação da doença?

Podemos considerar de alta precisão e bastante abrangente. Mas qual é a latência (tempo/dias) necessário para compor estes números com precisão?

De fato, temos avançado no objetivo de unificação de bases de saúde públicas, mas estamos longe de um modelo realmente integrado de baixa latência para análises consolidadas.

Vamos entender como o Google poderia ajudar neste cenário:

O Google foi capaz de detectar as tendências da epidemia de gripe suína cerca de duas semanas antes do Centro de Controle de Doenças, analisando as

buscas que as pessoas estavam fazendo em diferentes regiões com sintomas específicos.

Este exemplo deixa clara a importância de soluções *real time,* com grandes volumes de dados, para casos críticos como relacionado à saúde, com possibilidade de evitar ou reduzir grandes epidemias. É evidente que não podemos garantir que os usuários que pesquisam sobre alguns sintomas são pessoas infectadas, porém a orientação estatística ajudará a quantificar a evolução da doença, e com base na geolocalização das buscas é possível entender como ela está se alastrando nas regiões.

Seria um diferencial para uma empresa mapear tendências com soluções *real time*, analisando sentimentos e interesses de seus clientes sobre determinados produtos após o lançamento de uma campanha de marketing. A capacidade de reação a um evento negativo pode fazer a diferença para o sucesso desta campanha.

REFERÊNCIAS

EMC. **Big Data: grandes possibilidades**. Disponível em: <http://brazil.emc.com/big-data/insights.htm#bird-migration>. Acesso em: 7 mar. 2017.

GATNER. **Business Intelligence (BI).** Disponível em: <http://www.gartner.com/it-glossary/business-intelligence-bi>. Acesso em: 7 mar. 2017.

IBM. **Ciência de Dados**. Disponível em: <http://www.ibm.com/developerworks/jp/opensource/library/os-datascience/figure1.png>. Acesso em: 7 mar. 2017.

KIMBALL, Ralph e ROSS, Margy. **The Data Warehouse Toolkit**. 3.ed. Indianapolis: Wiley, 2013.

MITCHELL, Tom**. Departamento de Machine Learning**. Disponível em: <http://www.ml.cmu.edu/>. Acesso em: 16 jan. 2015.

QPH. **Business Intelligence e Business Analytics**. Disponível em: <https://qph.is.quoracdn.net/main-qimg-6bd93a2c7b391ecfe7a4d992882678e4?convert_to_webp=true>. Acesso em: 8 dez. 2016.

SATHI, Dr. Arvind. **Big Data Analytics**. IBM Corporation: MC Press Online, 2012.

TAURION, Cesar. **Big Data.** São Paulo: Brasport, 2013.