FIAP ON

ENSINO A DISTÂNCIA

ARQUITETURA DE BIG DATA

# BIG DATA - O INÍCIO

Anderson Paulucci

CAPÍTULO 1

## LISTA DE FIGURAS

Figura 1.1 – A nova era digital ..... 5
Figura 1.2 – The Nexus of Forces (por Gartner) ..... 7
Figura 1.3 – Novo *mindset* ..... 7
Figura 1.4 – Evolução da computação ..... 8
Figura 1.5 – Volume e complexidade ..... 10
Figura 1.6 – Os 3Vs de Big Data ..... 11
Figura 1.7 – Volume relativo ..... 11
Figura 1.8 – Analogia entre volume e o microscópio. ..... 13
Figura 1.9 – Arquitetura analítica tradicional ..... 14
Figura 1.10 – Exemplo de velocidade – Google ..... 16
Figura 1.11 – Dados produzidos a cada 60 segundos ..... 18

## LISTA DE QUADROS

Quadro 1.1 – Definições de dados estruturados, não estruturados e semiestruturados. ..... 17
Quadro 1.2 – Um novo modelo de negócios baseado em APIs (*APIs as a Business*). ..... 23

**SUMÁRIO**

1 BIG DATA: O INÍCIO ..... 4
1.1 Por que Big Data? ..... 4
1.2 A evolução ..... 8
1.3 Os 3 Vs de Big Data ..... 14
1.3.1 Volume ..... 11
1.3.2 Velocidade ..... 14
1.3.3 Variedade ..... 17
1.3.4 Veracidade ..... 19
1.3.5 Valor ..... 20
1.3.5.1 Monetização dos dados: ..... 21
REFERÊNCIAS ..... 24
GLOSSÁRIO ..... 25

# 1 BIG DATA: O INÍCIO

## 1.1 Por que Big Data?

Se você nasceu a partir de meados da década de 1990, pode se considerar um cidadão nativo da era digital. Caso tenha alguns anos adicionais, você pode se considerar um imigrante digital.

Em 2012, um dos executivos da Google, Eric Schmidt, mencionou: “Desde os primórdios da civilização até 2003, a humanidade gerou 5 Exabytes de dados. Agora, vamos produzir 5 Exabytes a cada 2 dias… e o ritmo está aumentando.” E em 2015, esse volume foi produzido em menos de 24 horas.

Poucas transformações no mundo ocorreram de forma tão rápida e expressiva, quebrando paradigmas e mudando a sociedade. Em pouco mais de uma década, impérios foram construídos sobre o domínio da tecnologia da informação. Empresas digitais como Google, Amazon e Facebook se tornaram motores de inovação e são alimentadas com um valioso combustível chamado “dados”.

A sociedade mudou, interagimos através de uma rede *on-line* e construímos relacionamentos virtuais com nossos familiares, amigos e pessoas que fazem parte do nosso ambiente profissional. As novas gerações (nativos da era digital) estão naturalmente conectadas e produzindo a matéria-prima para esta nova era, uma fonte inesgotável.

A economia também mudou, agora devemos acrescentar dados para compor a equação (Cesar Taurion – Big Data).

Economia = (Capital + Trabalho + **Dados**)

Hoje, não fazemos negócios e não tomamos uma única decisão sem a análise criteriosa de dados. E a competência de usar os dados para direcionar a tomada de decisão é determinante para se manter competitivo nesta nova

economia, seja para entender melhor a necessidade do seu cliente, para criar um novo produto, para fazer ou evitar um novo investimento e combater fraudes.

Você pode se perguntar: “Dados não fazem parte da história e evolução da humanidade?”

Sim. Claro!

Temos experimentos científicos que dependem de várias pesquisas documentadas com dados importantes, usados para correlações e estatísticas de pesquisas avançadas, fazemos isso há anos.

Na Segunda Guerra Mundial, a Inglaterra deu início à construção de uma máquina idealizada pelo matemático britânico Alan Turing, tido como o inventor do computador. O filme *O Jogo da Imitação*, dirigido por Morten Tyldum, conta de maneira contundente como a estratégia baseada em dados para decifrar a criptografia alemã ajudou a finalizar a guerra com anos de antecedência. A importância de analisar os dados criptografados (não estruturados) com latência abaixo de 24 horas, para se antecipar aos ataques alemães, era crucial para vencer esse desafio.

Como podemos perceber, os dados ajudaram a direcionar a evolução da humanidade de maneira expressiva, porém em um ritmo “tímido e limitado” perto da revolução que estamos iniciando.

Proponho definirmos um marco para o Big Data, vamos usar a frase de Eric Shmidt da Google e apontar para o ano de 2003:

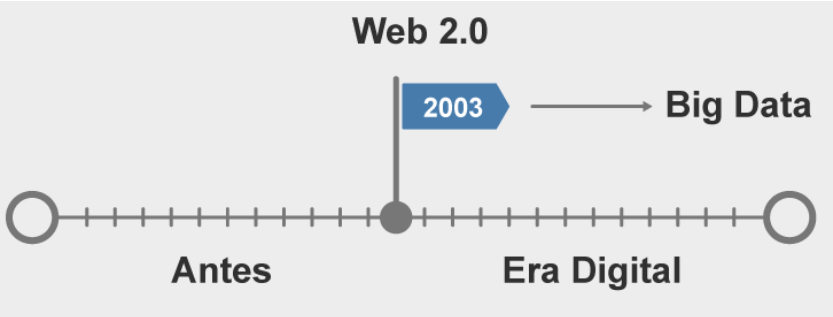


Figura 1.1 – A nova era digital
Fonte: Elaborado pelo autor (2016).

# BIG DATA: O INÍCIO

Definitivamente, estamos usando o adjetivo “BIG” para expressar o que realmente mudou de maneira significativa, que é a dimensão de volume de dados produzidos a partir de 2003 (um marco simbólico).

Empresas como Google, Amazon, Facebook, eBay, LinkedIn, Netflix, Airbnb e Uber são exemplos de empresas nativas da era digital, e representam um novo modelo de negócios estruturado em rede. As redes permitem um crescimento orgânico e acelerado, que transpõe facilmente as fronteiras continentais. Quanto maior a escalabilidade dessas empresas, maior o potencial de criar Big Data(s). Essas empresas foram pioneiras e atingiram grande parte da população mundial. A tecnologia não estava preparada para esse nível de escalabilidade, portanto, as gigantes digitais aprenderam a primeira grande lição da era da informação**.**

A inovação deve fazer parte do seu DNA. Só existe uma maneira de continuar crescendo e está diretamente ligada à criação de novas tecnologias. E a capacidade de fazer isso no prazo mais agressivo determina o sucesso dos negócios.

Empresas tradicionais de tecnologia, como Oracle, IBM, HP, Microsoft, entre outras, não almejavam criar soluções para atender startups com propósitos incompreensíveis, do ponto de vista de negócios, e talvez não imaginassem a real capacidade de transformação (inovação) dessas novas empresas. E isso é mais um fator determinante para toda a evolução da tecnologia que vamos abordar mais adiante.

As empresas que direcionam a evolução das tecnologias e arquitetura de Big Data “não” são as empresas tradicionais, e sim os gigantes da era digital. É com eles que vamos aprender sobre Big Data.

O Gartner usou a abordagem “*The Nexus of Forces*” em 2013 para definir os pilares para a TI nesta nova era digital, como mostra a figura 1.2.

# BIG DATA: O INÍCIO

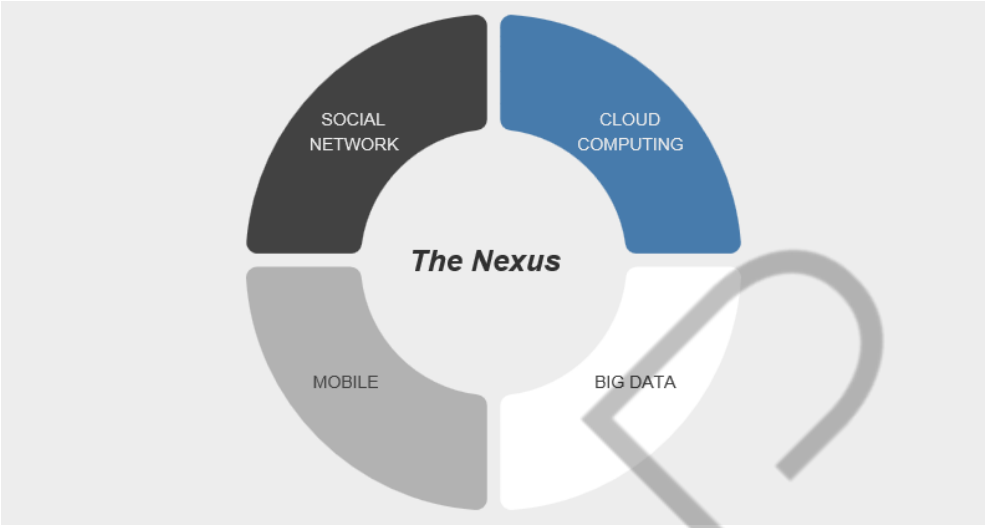


Figura 1.2 – The Nexus of Forces (por Gartner)
Fonte: Elaborado pelo autor (2016).

A nova TI deve ser estruturada nesses quatro pilares. A partir da definição do Gartner, praticamente todas as empresas tradicionais de tecnologia passaram a adotar esse modelo para definir seus pilares de soluções no mercado de TI.

Quando analisamos uma empresa nativa digital, compreendemos que ela já tem essa estrutura bem definida, desde o *startup*. Isso aumenta muito o seu potencial de inovação e a coloca em vantagem para escalabilidade de Big Data.

Apesar de ser um assunto que requer grande profundidade, vamos resumir uma primeira definição para Big Data.


Figura 1.3 – Novo *mindset*
Fonte: Elaborado pelo autor, adaptado por FIAP (2017).

Temos uma grande história de TI e vamos entender melhor essa transição, porém, para avançarmos, é necessário um ajuste no seu mapa mental. Durante anos, acumulamos aprendizado, mitos, pequenos vícios e tudo isso ocupa um

espaço no seu “cache” (memória), bloqueando a entrada de novas maneiras de inovar e resolver problemas. Quanto mais conseguirmos trabalhar essa mudança de *mindset*, mais fácil será o aprendizado sobre Big Data.

Estamos na quarta revolução industrial e na terceira geração computacional, e para essa nova fase não podemos usar o mapa mental antigo.

## 1.2 A evolução

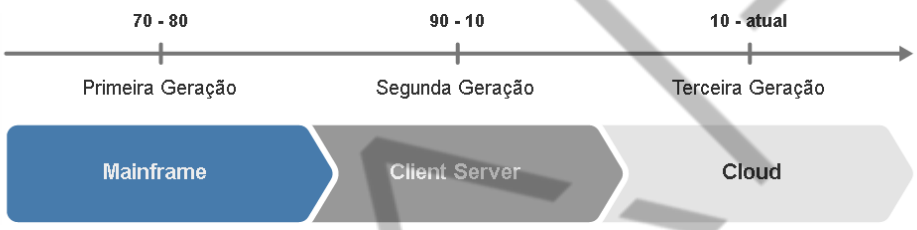


Figura 1.4 – Evolução da computação
Fonte: Elaborado pelo autor (2016).

- **Mainframe:** computação centralizada e monolítica de implementação relativamente simples. Poucas empresas têm acesso a computação, devido ao alto custo principalmente. Do ponto de vista de desenvolvimento, os programadores são orientados a eficiência e sabem da importância ($) de otimizar o uso de recursos.
- **Client Server:** computação distribuída, muitas soluções de hardware e softwares foram criadas, existem vários fornecedores e opções para compor uma solução, isso implica em uma arquitetura cada vez mais complexa. Na década de 80 e 90 foi fortemente trabalhado para implementar o ERP e CRM. No final da década de 90 se iniciou o “mundo Web” para transformar os negócios em modelos digitais. O desenvolvimento de software desta segunda geração foi orientado a agilidade, devido ao baixo custo de hardware foi perdida um pouco a orientação a eficiência no desenvolvimento. Com o avanço da computação cliente-servidor, a TI se tornou mais democrática e praticamente todas as empresas e pessoas possuem acesso a computação.

- **Cloud:** a computação em nuvem é o modelo de arquitetura para a terceira geração. Não se engane definir nuvem apenas como um provedor remoto de computação, mas sim um novo padrão de arquitetura para a TI, desde a engenharia de software, plataformas, implementação e principalmente infraestrutura. Os novos modelos são baseados em:
    - Flexibilidade;
    - Elasticidade;
    - Escalabilidade;
    - Economia em escala

  A arquitetura que suporta os grandes cases de Big Data foram criadas por provedores que adotaram o modelo de arquitetura em nuvem com padrões baseados em web-scale, arquitetura open, altamente eficiente. Em relação a eficiência, imagine uma otimização de um código do Google, consumindo menos 1KB de memória por conexão ao aplicativo do Google Maps. Os ganhos em escala são fatores fundamentais para manutenção dos serviços e garantia da evolução.

Com o *time-to-market* cada vez mais agressivo e as necessidades de tomada de decisões *real time*, a TI tradicional passou a não atender mais as demandas de negócios. Devido à agilidade requerida, este é o principal ponto que impacta negativamente a segunda geração da computação. Precisamos eliminar a burocracia e os serviços operacionais, transformando a TI em uma área estratégica para as empresas, andando à frente do *time-to-market,* posicionando inovações que transformem os negócios e aumentem as oportunidades.

A nova arquitetura em nuvem integra os conceitos de dados operacionais e analíticos, em ambientes geograficamente distribuídos. Com capacidades elásticas *on-demand*, possibilitando *startups* como Easy Taxi, Uber, Airbnb, Instagram e

Netflix, que construíram *cases* de referência, incluindo Big Data sobre a plataforma em Nuvem Amazon AWS.

A seguir, mais um exemplo de evolução baseado em demandas, volumes e complexidade de dados.

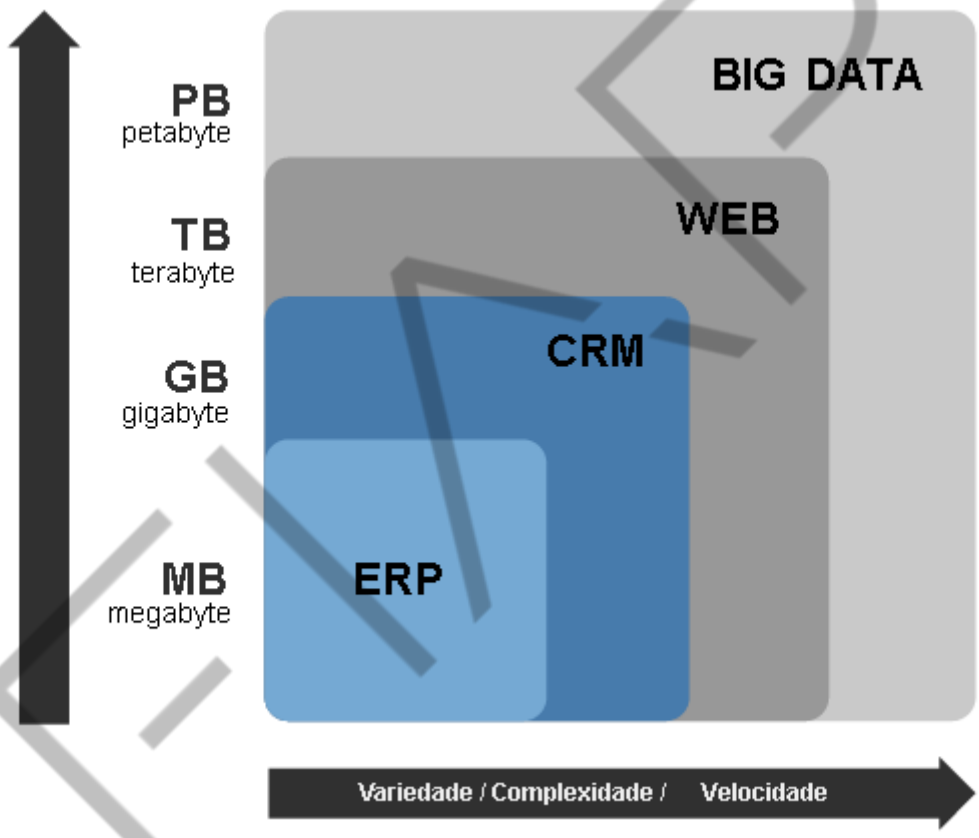


Figura 1.5 – Volume e complexidade
Fonte: Elaborado pelo autor (2016).

### 1.3 Os 3 Vs de Big Data

O conceito mais usado para definir os fundamentos de Big Data é baseado em 3Vs. Não ignore ou questione os 3Vs neste momento, pois requer uma análise

mais aprofundada para entendermos algumas questões importantes, as quais serão desenvolvidas ao longo dos próximos capítulos.

Abaixo, temos os 3Vs principais e 2Vs complementares.

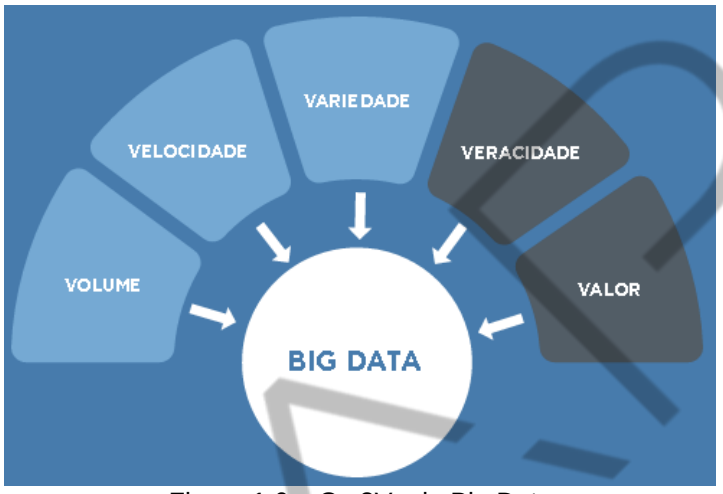


Figura 1.6 – Os 3Vs de Big Data
Fonte: Elaborado pelo autor (2016).

### 1.3.1 Volume

Volume é um conceito relativo, o que é Big Data hoje pode não ser Big Data amanhã. Portanto, precisamos entender essa questão-chave sobre o principal fundamento de Big Data.

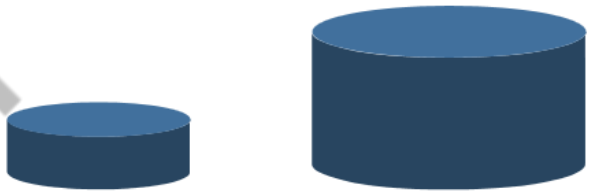

Figura 1.7 – Volume relativo
Fonte: Elaborado pelo autor (2016).

Vamos definir um cenário inicial para identificar qual volume é considerado um Big Data e, nessa definição, utilizaremos o volume em petabytes (que equivale a 1.000 terabytes).

Obs.: Assim como 1 TB, alguns anos atrás, era considerado um grande volume, e havia poucos bancos de dados com essa capacidade, o petabyte será comum para a maioria das organizações, em pouco tempo.

- 10 TB é Big Data?
- 100 TB é Big Data?
- 1 PB é Big Data?

A melhor maneira de responder a essas questões é analisando as tecnologias tradicionais, e para isso, proponho fazermos a seguinte pergunta:

A solução tradicional (cliente-servidor), seja ela um database, uma aplicação ou um *hardware*, está preparada para atender este volume?

Pense sobre:

Custo ($)

Capacidade (por exemplo: Armazenamento, Processamento, I/O)

Arquitetura (por exemplo: Geograficamente Distribuída)

Caso sua resposta seja não, temos aqui a principal motivação para quebrar os paradigmas tradicionais da segunda geração e buscar novas soluções que se enquadrem melhor nas suas necessidades computacionais da terceira geração.

Um database tradicional definitivamente não foi concebido para trabalhar com vários terabytes, sua estrutura foi estressada ao longo da evolução para se adaptar ao crescimento de volumes cada vez maiores. Porém, a partir da nova era da informação, a explosão de dados causou um colapso nas arquiteturas de dados tradicionais e chegamos ao seu limite técnico-financeiro, para a maioria das empresas que operam volumes de dados em petabytes.

Portanto, o V de volume para definição de Big Data está ligado a capacidades que excedem as tecnologias tradicionais.

# BIG DATA: O INÍCIO

Uma frase famosa foi usada no Fórum Econômico Mundial de 2011: “Dados é o novo petróleo.”

Esta nova era define que quanto mais dados, “melhor”.

Conforme mencionado no livro *Big Data*, de Cezar Taurion:

“O que o microscópio foi para a medicina e a sociedade, o Big Data também será para as empresas e a própria sociedade.”

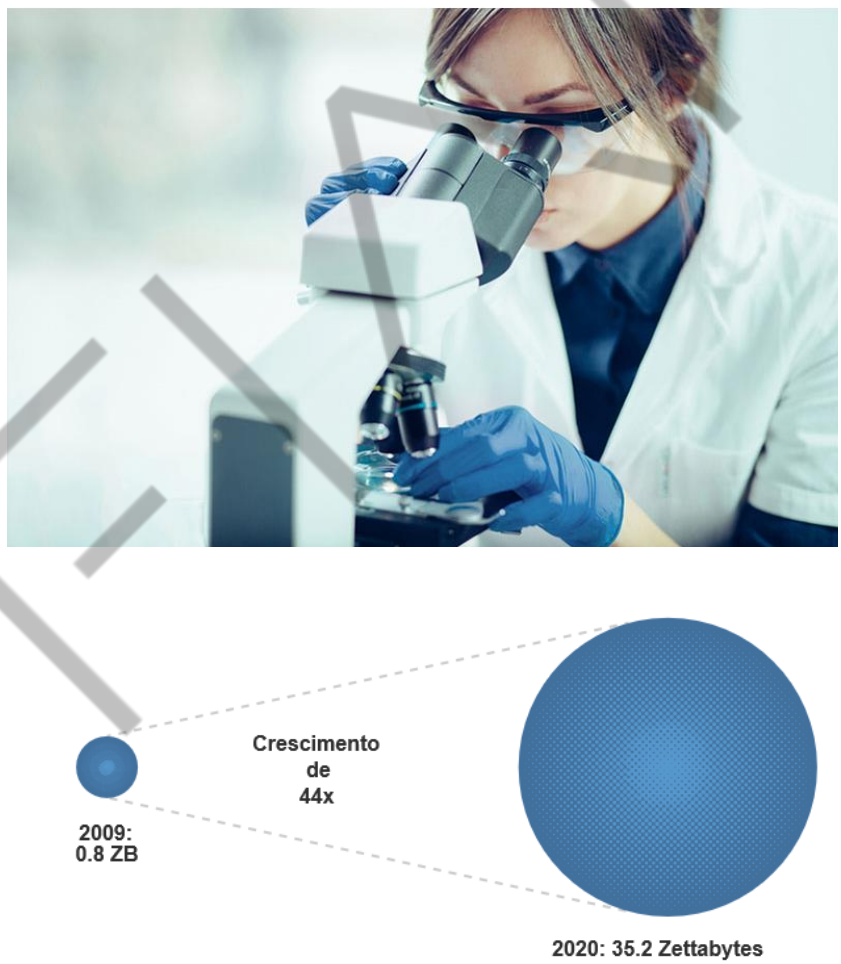


Figura 1.8 – Analogia entre volume e o microscópio.
Fonte: Elaborado pelo autor, adaptado por FIAP (2017).

O microscópio é um instrumento óptico com capacidade de ampliar imagens de objetos muito pequenos graças ao seu poder de resolução. E o que isso tem a ver com o volume de Big Data?

Absolutamente tudo! A capacidade de analisar dados com o máximo de detalhes e a maior granularidade possível permitem análises mais precisas e inteligentes. Trabalhar com Big Data possibilita eliminar processos pesados de filtros e agregações que limitam armazenar todas as variáveis para a tomada de decisão, considerando uma análise mais precisa e complexa.

### 1.3.2 Velocidade

Como vimos nos fundamentos da computação baseada em nuvem, a TI não atende o *time-to-market*. A velocidade de negócios é imediatista e de caráter emergencial, enquanto a velocidade de TI requer muitos processos vagarosos.

Vamos analisar o modelo clássico da arquitetura analítica.

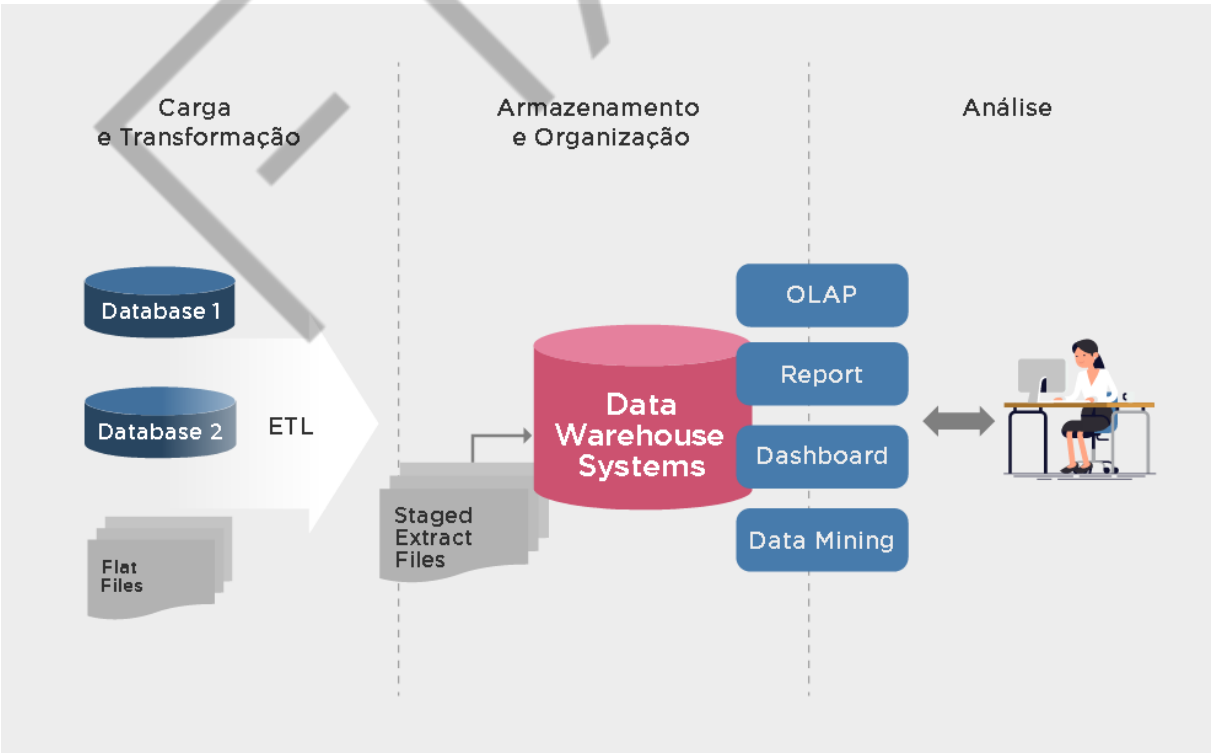


Figura 1.9 – Arquitetura analítica tradicional

Fonte: Elaborado pelo autor, adaptado por FIAP (2017).

Podemos mencionar sobre os dois principais aspectos relativos à velocidade:

- **Throughput**: a velocidade necessária para processar um grande volume de informações.
- **Latência**: a velocidade para análise de informações, ou seja, para chegarmos ao relatório final (D-1 => Data-In-Motion)

O fluxo de dados da arquitetura tradicional acima limita a análise com delta de 1 dia para a tomada de decisão. As transações que alimentam os bancos de dados operacionais são acumuladas ao longo do dia e estão fragmentadas em várias bases, criando silos de informações com visões diferentes. Através de processos ETL (*Extract Transform Load* – Extração, Transformação e Carga), as informações são extraídas dos bancos de dados operacionais, passam por um processo de transformação e são carregadas diariamente para o Data Warehouse (em um cenário otimista, esse processo funciona diariamente e não ultrapassa o delta de 1 dia), consolidando um modelo de dados dimensional que centraliza os dados com uma visão integrada e consistente na linha do tempo.

Agora vamos analisar um modelo do Google, considerando as necessidades das empresas da era digital.

Para contextualizar, o Google, sem dúvida, é o maior motor de inovação para Big Data, e seu primeiro grande desafio surgiu há mais de uma década, quando encontrou as primeiras barreiras para a evolução do seu algoritmo de PageRank.

> O coração do nosso *software* é o PageRank(TM), um sistema para dar notas para páginas na web, desenvolvido por nossos fundadores Larry Page e Sergey Brin na Universidade de Stanford. E embora tenhamos dúzias de engenheiros trabalhando para melhorar todos os aspectos do Google no dia a dia, PageRank continua a ser a base para todas nossas ferramentas de busca na web. (POR QUE USAR O GOOGLE, 2011)

O Google faz uso do PageRank e de outros algoritmos com o objetivo de computar (indexar) e otimizar as buscas na web.

Para um exemplo hipotético, considere o cenário abaixo, com o volume de 200 TB e a capacidade de processamento de 50 MB/s (megabytes por segundo).

Qual é o tempo estimado para a análise dos 200 TB de dados?

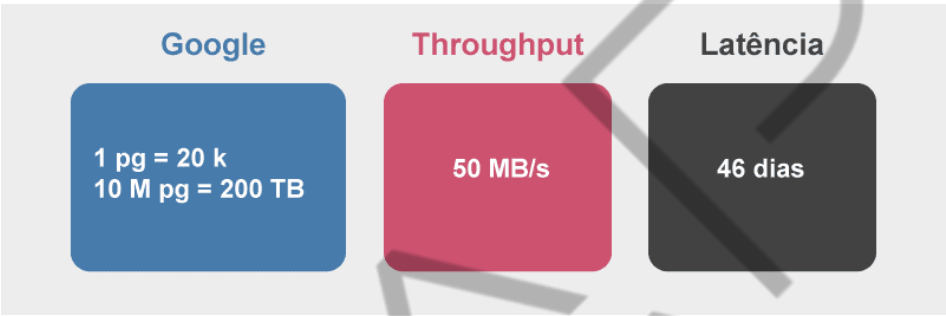


Figura 1.10 – Exemplo de velocidade – Google
Fonte: Elaborado pelo autor, adaptador por FIAP (2017).

O que deverá ser realizado para otimizar a latência de 46 dias (tempo de processamento)?

Faça uma analogia com uma pessoa que tem o desafio de ler um livro de 1.000 páginas em 1 dia. Se o seu *throughput* é de 1 página a cada 5 minutos, considerando que terá 8 horas para dormir e 16 horas para a leitura, conseguiria atingir no máximo 192 páginas por dia (sem considerar horas de refeições ou possíveis interrupções).

O uso de computação distribuída com capacidades de processamento e armazenamento massivo de dados é o caminho para soluções de Big Data.

As arquiteturas tradicionais terão dificuldades para escalar velocidade, assim como, seria humanamente impossível ler o livro de 1.000 páginas com baixa latência.

Considere a possibilidade de distribuir as páginas do livro para um grupo de dez pessoas, e ao invés de uma pessoa ler o livro sozinha, teremos dez pessoas lendo o livro em paralelo, considerando porções de cem páginas por pessoa. Este é

um exemplo clássico de dividir para conquistar e a solução para reduzir a latência e aumentar o *throughput* do todo.

### 1.3.3 Variedade

A variedade dos dados está ligada principalmente à maneira como podemos trabalhar a identificação de sua estrutura.

Imagine uma farmácia com diversos medicamentos sem identificação de rótulos nas embalagens e sem etiquetas nas prateleiras. Considere que cada medicamento tem um código único (chave) de identificação que o associa a uma lista com as descrições detalhadas sobre sua composição, processo de fabricação, qualidade e regulamentações de saúde, desde a origem dos laboratórios até as prateleiras das farmácias. Por meio desse novo modo de identificar as propriedades dos medicamentos, podemos consultá-los considerando um número muito maior de atributos.

| | |
|---|---|
| Dados estruturados (menor parte): | Provenientes de sistemas estruturados tradicionais. |
| Dados não estruturados (imensa maioria): | Gerados por e-mails, mídias sociais (Facebook, Twitter, YouTube, entre outros), documentos eletrônicos, vídeos e imagens, sensores, RFIDs. Estima-se que mais de 80% dos dados gerados atualmente são não estruturados. |
| Dados semiestruturados: | Normalmente registros de sensores, máquinas ou logs, por exemplo. |

Quadro 1.1 – Definições de dados estruturados, não estruturados e semiestruturados.
Fonte: Elaborado pelo autor (2017).

Conforme vimos anteriormente, a Web 2.0 foi um grande *milestone* (marco) para avançarmos com a computação de terceira geração. Novas semânticas de dados foram criadas com o objetivo de aproximar a linguagem humana da linguagem da máquina. Os conceitos de colaboração e interoperabilidade dos dados já são pensados há mais de uma década com esse propósito; e com o auxílio da

web semântica, conseguimos evoluir adotando padrões que ajudam a integrar aplicações com o objetivo de compartilhar informações com ampla facilidade.

O infográfico abaixo apresenta uma visualização clara de como estamos produzindo informações com mais volume, velocidade e variedade.

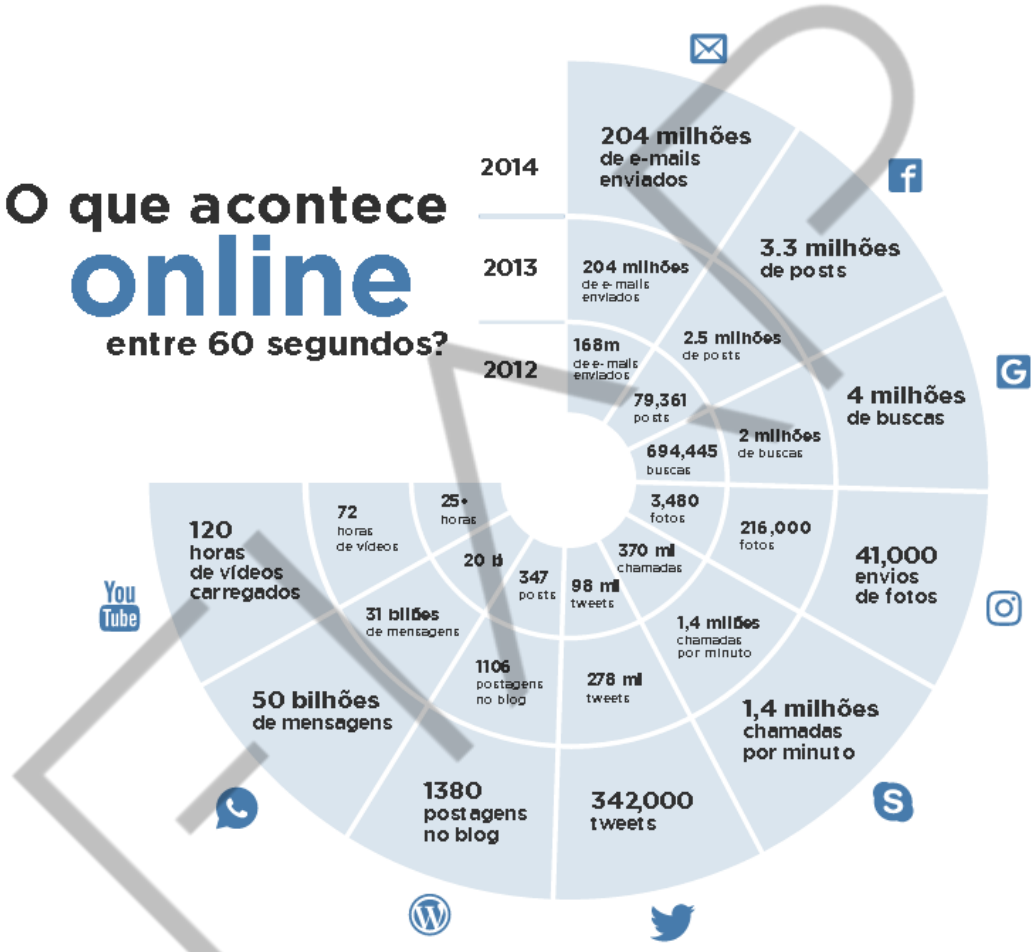


Figura 1.11 – Dados produzidos a cada 60 segundos
Fonte: Centre for Learning and Teaching (2014), adaptado por FIAP (2017).

Para facilitar essa definição conceitual de variedade, vamos considerar que os dados estruturados são aqueles submetidos a processos de modelagens com o objetivo de normalizar e/ou definir um esquema (tabelas, colunas, *data type*, *constraints*) mais rígido, basicamente os dados que armazenamos em um SGBDR (Sistema Gerenciador de Banco de Dados Relacional). De maneira simples, todo o restante são dados não estruturados.

Big Data não tem o objetivo de tratar apenas os dados não estruturados, muito pelo contrário, sempre haverá uma necessidade de implementar algum tipo de estrutura para evoluir os metadados e facilitar o acesso aos dados. Alguns componentes de soluções Big Data são inclusive projetados para suportar SQL (Structured Query Language), por exemplo, um pouco longe dos conceitos de relacionamento, porém com base na sua estrutura tabular.

Mas um dos grandes diferenciais em relação ao padrão tradicional é a capacidade de manipular os dados não estruturados, considerando que as plataformas de segunda geração da computação (cliente-servidor) não foram projetadas para isso.

### 1.3.4 Veracidade

O significado de veracidade está intimamente ligado a tudo o que diz respeito à verdade.

Trabalhando com os dados estruturados ou não estruturados, é importante garantir que os dados são autênticos e fazem sentido, para evitar tomada de decisões com base na análise de dados incertos e imprecisos.

Pode ser prudente atribuir uma pontuação (*score*) de veracidade dos dados e classificação de conjuntos específicos de dados para garantir a qualidade da informação.

Quando analisamos dados de redes sociais, considerando análises de textos, por exemplo, corremos sérios riscos de concluirmos a análise com dados não verídicos, tais como as *fake news*. O que acontece com uma mentira reproduzida milhares de vezes na internet? Torna-se verdade, não é mesmo? Se não criarmos um processo de mitigação de cenários como estes, podemos tomar decisões imprecisas e pouco confiáveis.

Sempre foi importante tratar a veracidade dos dados, porém diferentemente dos modelos analíticos baseados em BI (*Business Intelligence*) tradicional, em que

podemos confiar de “olhos vendados” nas fontes de dados comuns (ERP, CRM, Aplicações Corporativas); Big Data deve se preocupar muito mais com a veracidade, considerando as várias fontes de dados e a velocidade que precisamos impor para a análise.

À medida que as empresas avançam na adoção de Big Data, naturalmente o maior volume de dados poderá ser proveniente da nuvem e corresponder a novas semânticas. Assim, a curadoria dos dados será uma disciplina importante para a governança corporativa.

### 1.3.5 Valor

Informações são comercializadas há muito tempo, em jornais, revistas, filmes, livros etc. Fontes de dados financeiros são produtos valiosos, e empresas como a Serasa Experian e Boa Vista Serviços lucram vendendo informações para o mercado, que depende delas, seja para crédito, seguros, antifraude ou alavancagem de clientes.

Quando trabalhamos com Big Data, os valores que buscamos dos dados devem ser definidos com clareza. Este é o maior motivador para direcionar um *use case* de Big Data.

Não é relevante armazenar um grande volume com possíveis variedades e requisitos de velocidade sem considerar um objetivo claro de extrair valor dos dados.

Como as organizações podem fazer para aumentar receita, reduzir custos, diminuir fraudes e melhorar *compliance*?

A necessidade de gerar valor é o principal objetivo de Big Data. Esse valor pode ocorrer pela agilização da tomada de decisão, aumento da precisão da análise com mais dados sendo correlacionados, redução dos riscos, criação de oportunidades. E assim justificamos um projeto de Big Data.

O valor pode ser muito mais do que uma forma de rentabilizar a receita para a corporação, considere os avanços nas pesquisas contra o câncer e doenças graves, a evolução nos controles do uso sustentável da água potável, maior transparência na gestão de contas públicas, e assim estamos sendo induzidos naturalmente a usar os dados para resolver problemas e evoluir com mais precisão nas análises complexas, até pouco tempo atrás inviáveis.

#### 1.3.5.1 Monetização dos dados:

Qual é o valor do seu ativo de dados?

O objetivo de monetizar os dados está relacionado à geração de receitas financeiras a partir de fonte de dados disponíveis ou em tempo real. Em resumo, tratar os dados efetivamente como um produto.

Algumas empresas de *software* especializadas em dados (com domínio de Big Data) estão criando soluções baseadas em modelos de negócios que propõem um compartilhamento de receita (*revenue share*), e usam o ativo dados das empresas como matéria-prima para alavancar receitas.

Considere o potencial de uma empresa de telecom com enormes volumes de dados e possibilidades de reduzir o custo de campanhas de marketing com propostas mais precisas. Por exemplo, no Brasil, as operadoras têm milhões de clientes, e as campanhas direcionadas para todos eles são pouco precisas e custam um valor altíssimo. Uma empresa fornecedora de soluções de análises baseadas em *revenue share* poderia propor uma abordagem mais precisa para a campanha, usando técnicas de *real time analytics* e reduzindo os custos de uma campanha de marketing com a proposta de atingir o público-alvo, segmentando com grupos de clientes ou até mesmo direcionado a cada cliente com base em eventos específicos.

Ou uma empresa financeira que pretende baixar os índices de fraudes de 0,1% (que representam milhões de reais mensais) para 0,01%. A empresa fornecedora da solução de análise de dados pode propor 30% do compartilhamento

da receita (alguns milhões de reais) em caso de sucesso, referente ao retorno financeiro previsto. A empresa financeira não deverá pagar nada além do resultado comprometido, inclusive considerando a possibilidade de não obter sucesso, a empresa fornecedora da solução não receberá sobre o trabalho. Será um ótimo negócio para ambos, e dessa forma, muitas empresas e consultorias especializadas em monetizar os dados emergirão.

Empresas não consideram a importância do valor de ativo de dados e pouco investem na estratégia para tratar os dados como um produto. O desafio de direcionar os negócios com base em dados irá mudar essa postura de negligência e, aos poucos, empresas tradicionais serão naturalizadas como empresas digitais.

Como vimos, a computação em nuvem ajudou a evolução de soluções com mais escalabilidade, e os modelos de negócios passaram a ser construídos com base em *social network* (rede), com colaboração e geograficamente distribuídos.

Podemos criar grandes ativos de dados fazendo uso de estratégias baseadas em API (*Application Programming Interface*), um conceito bastante difundido no desenvolvimento de *software* na era da computação em nuvem e importante para os novos modelos de arquitetura de TI.

A API propõe alguns mecanismos de interfaces para que possamos monetizar bases de dados. Por exemplo: o Google possui uma base de geolocalização com o serviço do Google Maps, porém algumas aplicações podem demandar um número de requisições diárias acima do limite considerado *free* e deverão pagar para usar o serviço. Esse modelo já foi implementado por grande parte dos gigantes da internet e deve ser referência para as empresas que estão buscando oportunidades neste mundo de negócios, cada vez mais baseado em dados.

| | |
|---|---|
| Basic | 1.000 requisições p/ dia<br>R$ 100/mês |
| Dados não estruturados (imensa maioria) | 10.000 requisições p/ dia<br>R$ 250/mês |
| Dados Semiestruturados | ilimitado<br>R$ 500/mês |

Quadro 1.12 – Um novo modelo de negócios baseado em APIs (*APIs as a Business*).
Fonte: Google.

Esta pode ser a saída para algumas empresas que estão sendo engolidas por um processo de digitalização e dependem de novos produtos, o fato é que os dados são um ativo importante e devem ser usados para alavancar novos negócios, e isso vai muito além do *core business*. Por exemplo, se considerarmos uma base de referência de geolocalização por usuário de uma operadora de telecom, seria uma grande oportunidade usá-la para monetizar um serviço baseado em APIs e alavancar oportunidades.

Dessa forma, em pouco tempo teremos uma infinidade de novos produtos (serviços) sendo orquestrados por um *grid* baseado em APIs. É evidente a necessidade da governança de dados em todos os processos de estratégia de dados, e quando o assunto é monetização, a ética e a segurança devem ser priorizadas.

**REFERÊNCIAS**

BASSO, Guilherme Mastrichi. **Terceirização e o mundo globalizado**: o encadeamento produtivo e a complementaridade de serviços como potencializadores da formação de contratos. Disponível em: <http://www.tst.jus.br/documents/exemplo.pdf>. Acesso em: 9 out. 2014.

MARZ, Nathan, WARREN, James. **Big Data**: Principles and best practices of scalable realtime data systems. Nova York: Manning Publications, 2015.

POR QUE USAR O GOOGLE. 2011. Disponível em: <https://www.google.com/intl/pt-BR/why_use.html>. Acesso em 12 dez. 2016.

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

SCHONBERGER, Viktor Mayer; CUKIER, Kenneth. **Big Data:** Como Extrair Volume, Variedade, Velocidade e Valor da Avalanche de Informação Cotidiana. São Paulo: Campus, 2013.

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

TUKEY, W. John. **The Future of Data Analysis**. Disponível em: <http://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711>. Acesso em 12 dez. 2016.

WIKIPÉDIA. **Definição de Compliance**. Disponível em: https://pt.wikipedia.org/wiki/Compliance> . Acesso em 10/02/2017.

## GLOSSÁRIO

| | |
|---|---|
| Application Programming Interface<br>API | Interface de Programação de Aplicativos, em português, é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na Web. |
| Business Intelligence<br>BI | Inteligência de Negócios, em português, é o processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios |
| Compliance | Nos âmbitos institucional e corporativo, *Compliance* é o conjunto de disciplinas para fazer cumprir as normas legais e regulamentares, as políticas e as diretrizes estabelecidas para o negócio e para as atividades da instituição ou empresa, bem como evitar, detectar e tratar qualquer desvio ou inconformidade que possa ocorrer. |
| Customer Relationship Management<br>CRM | Gestão de Relacionamento com o Cliente, em português, compreende um conjunto de ferramentas (sistemas e processos) que coloca o cliente no centro do desenho dos processos de negócio. |
| Client Server | Computação distribuída, muitas soluções de *hardware* e *software* foram criadas, existem vários fornecedores e opções para compor uma solução, isso implica uma arquitetura cada vez mais complexa. Nas décadas de 1980 e 1990, foi fortemente trabalhado para implementar o ERP e CRM. No fim da década de 1990, se iniciou o “mundo Web”, para transformar os negócios em modelos digitais. O desenvolvimento de *software* dessa segunda geração foi orientado para a agilidade; devido ao baixo custo de *hardware*, foi perdido um pouco da orientação para a eficiência no desenvolvimento. Com o avanço da computação cliente-servidor, a TI se tornou mais democrática e praticamente todas as empresas e pessoas têm acesso à computação. |
| Cloud | A computação em nuvem é o modelo de arquitetura para a terceira geração. Não se engane: nuvem não é apenas um provedor remoto de computação, mas sim um novo padrão de arquitetura para a TI, desde a engenharia de *software*, plataformas até a implementação e, principalmente, a infraestrutura. |

| | |
|---|---|
| | Os novos modelos são baseados em: agilidade; flexibilidade; elasticidade; escalabilidade; economia em escala.<br>A arquitetura que suporta os grandes *cases* de Big Data foi criada por provedores que adotaram o modelo de arquitetura em nuvem com padrões baseados em *web-scale,* arquitetura *open*, altamente eficientes.<br>Em relação à eficiência, imagine a otimização de um código do Google, consumindo menos de 1 KB de memória por conexão ao aplicativo do Google Maps. Os ganhos em escala são fatores fundamentais para a manutenção dos serviços e garantia da evolução. |
| Enterprise Resource Planning<br>ERP | Sistema integrado de gestão empresarial, em português, é um sistema de informação que integra todos os dados e processos de uma organização. |
| Extract Transform Load<br><br>ETL | Extração, Transformação e Carga, em português, as informações são extraídas dos bancos de dados operacionais, passam por um processo de transformação e são carregadas diariamente para o Data Warehouse. |
| Mainframe | Computação centralizada e monolítica, de implementação relativamente simples. Poucas empresas têm acesso à computação, devido ao alto custo principalmente. Do ponto de vista de desenvolvimento, os programadores são orientados para a eficiência e sabem da importância ($) de otimizar o uso de recursos. |
| Sistema Gerenciador de Banco de Dados Relacional<br>SGBDR | Sistema que gerencia dados organizados por tabelas e estas compostas por linhas e colunas. |
| Structured Query Language<br>SQL | Linguagem de Estrutura de Consulta, em português, é uma linguagem universal de acesso a dados em um banco de dados. |