TECNOLOGIA DA INFORMAÇÃO

Big Data, Visualização e Análise Exploratória de Dados

Livro Eletrônico

GRAN CURSOS ONLINE

SUMÁRIO

Apresentação 3
Big Data 4
1. O que é “Big Data”? 4
1.1. Objetivo do “Big Data” 6
1.2. Origem dos Dados 7
1.3. Aplicações do Big Data e da Análise de Dados 8
1.4. Big Data Analytics 9
1.5. Tipos de Análise 10
1.6. Riscos Principais 10
1.7. Mitos sobre o Big Data 11
1.8. As Dimensões do Big Data 12
1.9. Camadas Lógicas de uma Solução de Big Data 19
1.10. NoSQL (Not Only SQL – Não Só SQL) 25
1.11. Hadoop 33
2. Visualização e Análise Exploratória de Dados 35
3. Considerações Finais 36
Resumo 37
Questões Comentadas em Aula 40
Questões de Concurso 42
Gabarito 57
Referências 58

Apresentação

Olá, querido(a) amigo(a)!

O momento perfeito não "surge". Ele é construído. Construa o seu.

Você tem suas próprias dificuldades, problemas, vitórias e soluções. Continue firme e, em breve, colherá os frutos da vitória.

Rumo então à aula sobre **Big Data**!

Força nos estudos!

Grande abraço,

BIG DATA

1. O que é "Big Data"?

Siewert (2013) destaca que o termo ***Big Data*** é:

> Definido genericamente como a **captura**, **gerenciamento** e a **análise de dados que vão além dos dados tipicamente estruturados**, que podem ser consultados e pesquisados através de bancos de dados relacionais.
> Frequentemente são **dados obtidos de arquivos não estruturados** como **vídeo digital, imagens, dados de sensores, arquivos de logs e de qualquer tipo de dados não contidos em registros típicos com campos que podem ser pesquisados**.

Obs.: Big Data tem dados estruturados e não estruturados!
Big Data é o termo que descreve o imenso volume de dados – estruturados e não estruturados – que impactam os negócios no dia a dia.

De maneira geral, ***Big Data*** **não se refere apenas aos dados, mas também às soluções tecnológicas criadas para lidar com dados em volume, variedade e velocidade significativos** (CESPE/2018).

Segundo Siewert (2013), **o *Big Data* tem variadas fontes de dados como:**

- dados gerados pelas máquinas (redes de sensores, *logs*);
- dispositivos móveis (vídeo, mensagens, fotografias);
- comunicação máquina a máquina, a "Internet das coisas";
- dados em bancos de dados relacionais oriundos das transações da organização;
- imagens de documentos etc.

De acordo com Landim (2015), trata-se de um termo usado para descrever **grandes e complexos conjuntos de dados** que são muito difíceis de capturar, processar, armazenar, buscar e analisar com os sistemas de base de dados convencionais.

Veja a seguir mais 3 definições encontradas na literatura para o termo Big Data:

Gartner Group (2012)	"**Big Data**, em geral, é definido como **ativos de alto volume, velocidade e variedade de informação** que exigem custo-benefício, de formas inovadoras de processamento de informações para maior visibilidade e tomada de decisão."
GEORGE et. al. (2014) (Academy of Management Journal)	"O **Big data** é **formado por uma crescente pluralidade de fontes de informação**, entre eles cliques na web, transações em dispositivos moveis, conteúdo gerado por usuários, mídias sociais, bem como conteúdo gerado intencionalmente através de redes de sensores ou transações comerciais, tais como consultas de vendas e transações de compra".

MCAFEE, A et. al. (2012) (Harvard Business Review)	"**Big Data** como uma **forma essencial para melhorar a eficiência e a eficácia das organizações de vendas e marketing**. Ao colocar Big Data no coração de vendas e marketing, os insights podem ser aproveitados para melhorar a tomada de decisão e inovar no modelo de vendas da empresa, o que pode envolver a utilização de dados para orientar ações em tempo real".

Esquematizando!

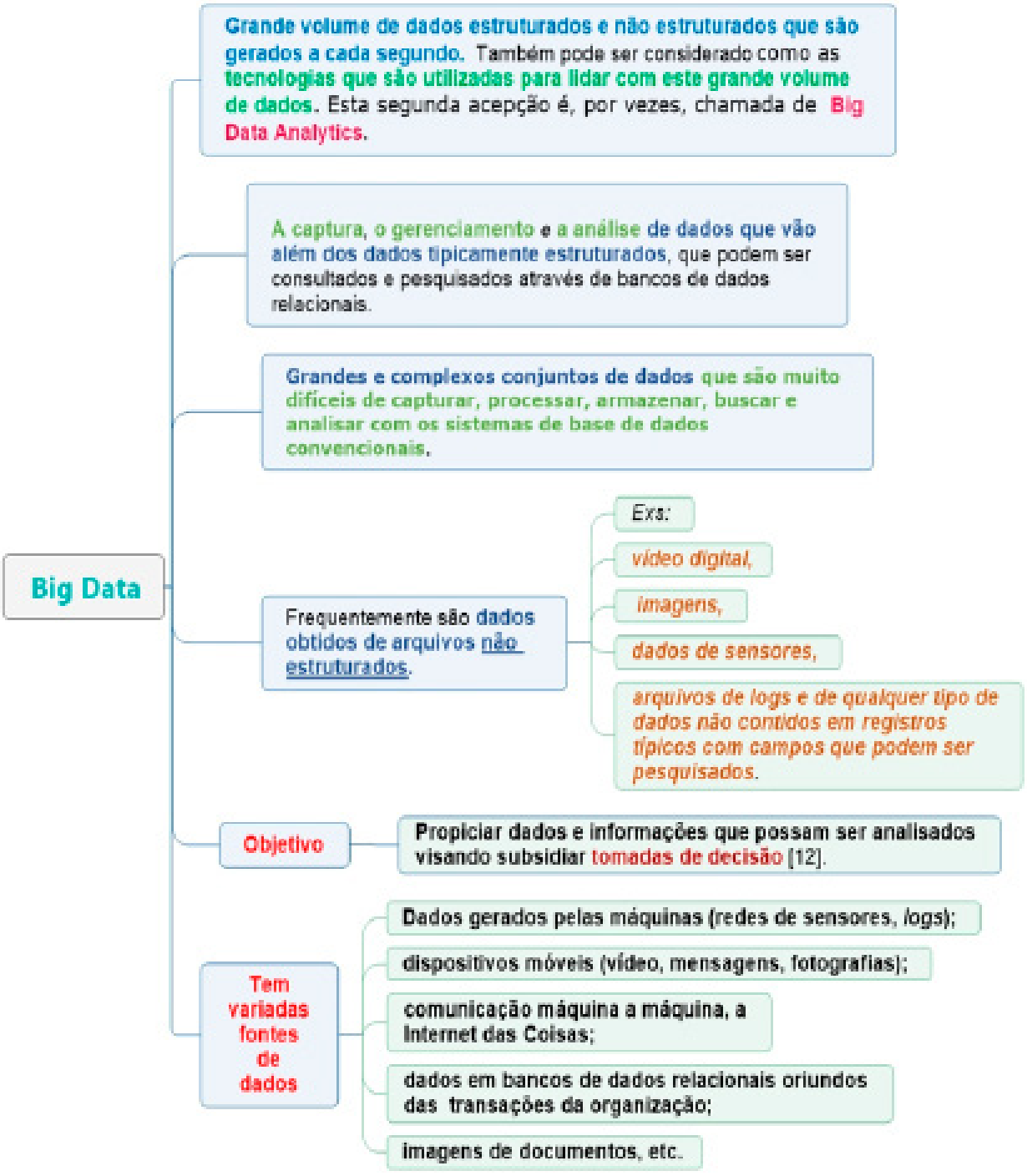


Figura Visão Geral do Big Data e Seus Objetivos.
Fonte: Quintão (2020)

Direto do Concurso

001. (QUADRIX/CREF 11ª REGIÃO/AGENTE DE ORIENTAÇÃO E FISCALIZAÇÃO/2014) Trata-se de uma infinidade de informações não estruturadas que, quando usadas com inteligência, se tornam uma arma poderosa para empresas tomarem decisões cada vez melhores. As soluções tecnológicas que trabalham com esse conceito permitem analisar um enorme volume de dados de forma rápida e ainda oferecem total controle ao gestor das informações. E as fontes de dados são as mais diversas possíveis: de textos e fotos em rede sociais, passando por imagens e vídeos, até jogadas específicas no esporte e até tratamentos na medicina. (http://olhardigital.uol.com.br/pro/video/39376/39376)
O conceito definido no texto é:
a) Governança de TI
b) QoS.
c) Big Data
d) Data Center.
e) ITIL.

A questão destaca de forma bem clara o conceito de Big Data, fácil não é mesmo!
Letra c.

1.1. Objetivo do "Big Data"

Obs.: O **objetivo do *Big Data*** é propiciar dados e informações que possam ser analisados visando subsidiar **tomadas de decisão** (Fernandes e Abreu, 2014).

A tomada de decisão é possível em função não somente do **volume** de dados, da **velocidade de captura dessas informações**, **das fontes variadas de informações** e de **novos softwares para fins de modelagem dessas informações** (Fernandes e Abreu, 2014).

Por exemplo, ver uma tendência de crescimento da venda de um produto em função de comentários favoráveis no Facebook. Este tipo de análise é o que está sendo denominado ***data analytics*** (Fernandes e Abreu, 2014).

Em Brito (2019), o autor destaca o seguinte: "o objetivo principal do Big Data é **obter informação útil a partir de dados armazenados em "tempo real" (espontâneos)** e por isso esses dados **não são estruturados**, o que torna a aplicação de técnicas de extração de informação mais difícil!

Assim, estamos falando de muitos dados que são gerados e consumidos rapidamente.

Obs.: É por isso que dizemos que as **características** mais marcantes do Big Data são:
(i) quantidade, e
(ii) velocidade.

Então, no cenário do Comércio Eletrônico, a simples transação eletrônica é uma relação direta entre cliente e empresa, o que não é caracterizado como Big Data. Essa transação gera um pedido que representa um histórico sob a visão de negócios da empresa.

Por outro lado, se a empresa tem **ferramentas para analisar o comportamento dos usuários enquanto eles navegam pela sua página de Comércio Eletrônico**, é possível exibir para o usuário somente aqueles produtos que estejam alinhados ao seu perfil, então existe potencial real de maximizar as vendas - **isso é Big Data.**

Outro exemplo de Big Data no mesmo contexto do comércio eletrônico e que faz relação com a produção: se a empresa tem ferramentas para avaliar quais produtos estão sendo mais acessados em seu ambiente de comércio eletrônico em determinado momento, esse pode ser um indicativo de quais produtos devem ser priorizados no ambiente da produção. Então repare que **os dados foram originados de maneira espontânea e as informações foram consumidas praticamente no mesmo tempo em que foram geradas**, por isso essas informações não são estruturadas. Esse é outro exemplo de **Big Data**".

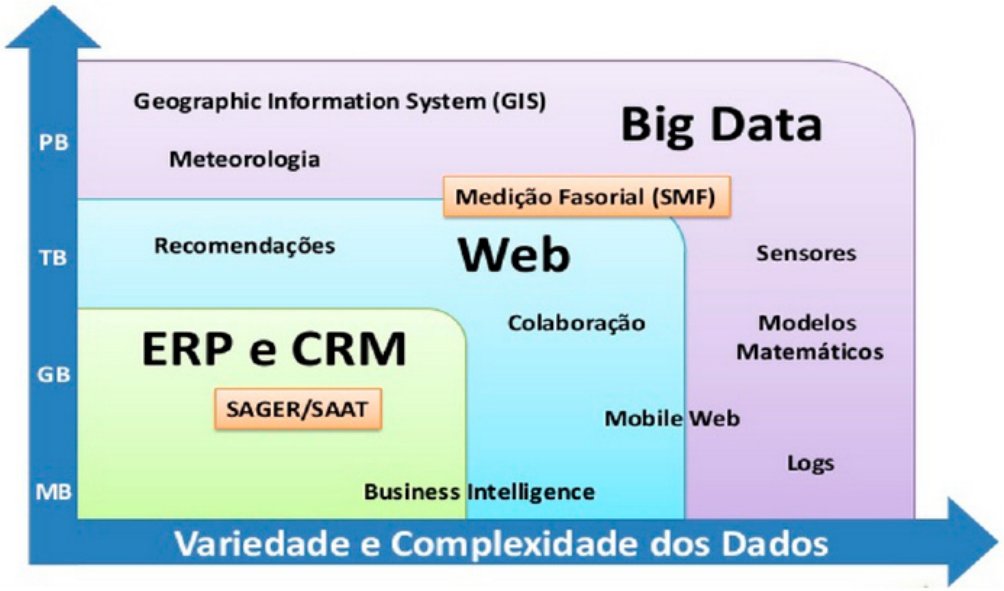
O que se apregoa é que de nada adianta você armazenar uma montanha de dados se não sabe como tirar proveito disso para o negócio!

1.2. Origem dos Dados

O Big Data surge para analisar as **interações, transações, observações de comportamentos**, de forma a proporcionar entendimento dos dados e auxiliar na tomada de decisão. Assim, irá gerar mais experiência, produtividade, consumo e novos produtos e serviços.

Obs.: Big Data = Transações + Interações + Observações

Fonte: Big Data (MAFRA, 2013)]



Fonte: (MAFRA, 2013)]

1.3. Aplicações do Big Data e da Análise de Dados

As **aplicações** do *Big Data* e da análise de dados são variadas, como (Fernandes e Abreu, 2014):

- desenvolvimento de mercado;
- inovação;
- desenvolvimento de produtos e serviços;
- eficiência operacional;
- previsões de demanda de mercado;
- detecção de fraudes;
- gerenciamento de riscos;
- previsão de concorrência;
- vendas;
- campanhas de *marketing*;
- avaliação do desempenho de funcionários;
- alocação do orçamento anual;
- estabelecimento de previsões financeiras;
- gestão de planos médicos;
- identificação de potenciais compradores;
- entendimento da base de clientes etc.

Ainda, segundo destaca (Fernandes e Abreu, 2014), em uma análise conjunta do *IBM Institute for Business Value* e do *Massachussets Institute of Technology* (o famoso MIT) identificou-se que **empresas que investem em análise de dados** (*Business Analytics and Optimization*) **possuem uma visão melhor do seu negócio**, conseguindo uma receita 33% maior do que seus concorrentes, crescimento do lucro doze vezes maior e retorno sobre o investimento de capital 32% maior (IBM (2011)).

Obs.: Vale a pena assistir! #Ficaadica

(1) http://g1.globo.com/jornal-da-globo/noticia/2013/12/massa-de-informacoes-digitais-pode-ser-usada-em-beneficio-da-populacao.html

(2) https://www.youtube.com/watch?v=gny_BR6ID6A

1.4. Big Data Analytics

- É o trabalho analítico e inteligente de grandes volumes de dados, **estruturados ou não estruturados**, que são coletados, armazenados e interpretados por softwares de altíssimo desempenho (SANTANA, 2018).
- É um termo que se refere à **análise** desses conjuntos grandes e complexos de dados estruturados e não estruturados. São utilizadas ferramentas e equipamentos de alta performance, muitas vezes com o auxílio de computação distribuída (utilizando várias máquinas em um trabalho coordenado).
- Geralmente envolve a utilização de algoritmos estatísticos avançados e análise preditiva, apontando o que está por vir no futuro e indicando tendências.
- Trata-se do cruzamento de uma infinidade de dados do **ambiente interno e externo**, gerando uma espécie de **"bússola gerencial"** para **tomadores de decisão**. Tudo isso, é claro, em um tempo de processamento extremamente reduzido (SANTANA, 2018).

Obs.: O termo **Big Data Analytics** refere-se aos poderosos **softwares que tratam dados estruturados e não estruturados** para **transformá-los em informações úteis às organizações**, permitindo-lhes analisar dados, como registros de *call center*, postagens de redes sociais, de *blogs*, dados de CRM e demonstrativos de resultados.

A seguir, destacamos algumas das fontes usadas por um *software* de *Big Data Analytics* (SANTANA, 2018):

- dados extraídos de ferramentas de Inteligência de Negócios (***Business Intelligence* – BI**);
- arquivos de *log* de servidores web;
- **conteúdo de mídias sociais**;
- relatórios empresariais;
- textos de *e-mails* de consumidores à empresa;
- indicadores macroeconômicos;

- pesquisas de satisfação;
- estatísticas de ligações celulares capturadas por sensores conectados à "**internet das coisas**" etc.

1.5. Tipos de Análise

Quando se trata de Big Data, a literatura destaca geralmente **quatro tipos de análises** (VORHIES, 2014):

- **Descritiva**: **foca no presente**, visando descrever características dos dados e eventos correntes para subsidiar decisões de efeitos imediatos.
- **Diagnóstica**: busca **entender as relações de causa e efeito entre eventos**.
- **Preditiva**: tem como objetivo **prever comportamentos futuros** e tendências com base nos dados conhecidos.
- **Prescritiva**: parecida com a análise preditiva, mas **busca os efeitos dos eventos futuros**. Visa prever os **efeitos** futuros dos eventos.

Direto do Concurso

002. (CESPE/TCE-MG/2018) Uma empresa, ao implementar técnicas e softwares de big data, deu enfoque diferenciado à análise que tem como objetivo mostrar as consequências de determinado evento. Essa análise é do tipo

a) preemptiva.
b) perceptiva.
c) prescritiva.
d) preditiva.
e) evolutiva.

Conforme visto, a análise **prescritiva** é a que **busca os efeitos dos eventos futuros**.
Letra c.

1.6. Riscos Principais

ISACA (2013a), destaca as principais perguntas que devem ser feitas em relação ao *Big Data*, do ponto de vista dos **riscos**. São elas:

- **Onde** os dados serão armazenados?
- **Como** os dados serão protegidos?
- **Como utilizar** os dados de forma **segura** e legal?

Os principais **riscos** que devem ser gerenciados são (Fernandes e Abreu, 2014):

- riscos de **perda de dados** "tóxicos" armazenados como informações privadas ou de custódia, tais como contas de clientes, números de cartão de crédito, segredos industriais da empresa etc.;
- o **uso de informações obtidas em redes sociais**, por exemplo, abrange **questões de privacidade e de falta de consenso jurídico** internacional, uma vez que cada país tem sua legislação específica;
- **questões de segurança da informação;**
- **qualidade dos dados** capturados para fins de análise;
- **disponibilidade e capacidade da infraestrutura tecnológica** que suporta o *Big Data*;
- **qualidade e capacidade do fornecedor de serviços** (se for o caso) que captura, armazena e/ou realiza análise de dados;
- **qualidade dos modelos de exploração** desenvolvidos para a análise dos dados;
- **pessoas com capacitação requerida** (cientista de dados) **para desenvolver modelos e analisar resultados;**
- **falha ao categorizar e mapear os dados;**
- **falta de governança de dados** etc.

1.7. Mitos sobre o Big Data

A seguir, confira algumas informações que você já deve ter escutado em algum instante, mas que **NÃO** retratam a realidade.

Mito 1 – Big Data engloba somente dados não estruturados.

Com o crescimento do volume de dados nos últimos anos, o banco de dados relacional precisou ser complementado com outras estruturas. O que mudou de fato foi a inclusão no Big Data também de mais tipos de dados, além dos estruturados.

Mito 2 – Big Data refere-se somente a soluções com petabytes de dados.

Embora o volume de dados seja o fator que impulsionou o Big Data, aplicações que utilizam conjuntos de dados em uma escala menor do que petabytes também podem se beneficiar das tecnologias de Big Data. Afinal, **o mais importante nessas aplicações é a capacidade de extrair valor dos dados.**

Mito 3 – Big Data pode prever o futuro.

Big Data e todas as suas ferramentas **não** podem dizer o que vai acontecer no futuro.

É possível **analisar o que aconteceu no passado** e tentar desenhar **as tendências** entre as ações, os pontos de decisão e as suas consequências, baseadas nos dados.

Podemos usar isso para **adivinhar** que, em **circunstâncias semelhantes**, se uma decisão semelhante for tomada, resultados semelhantes ocorreriam como resultado. Mas **não podemos prever o futuro**.

1.8. As Dimensões do Big Data

Para analisar a viabilidade de implementação do Big Data em uma organização, a literatura citava inicialmente as **3 dimensões do Big Data**, que são conhecidas como **3V** (**V**olume, **V**ariedade e **V**elocidade); depois o **4V** (incluindo aí a **V**eracidade) e o **5V** (incluindo o **V**alor).

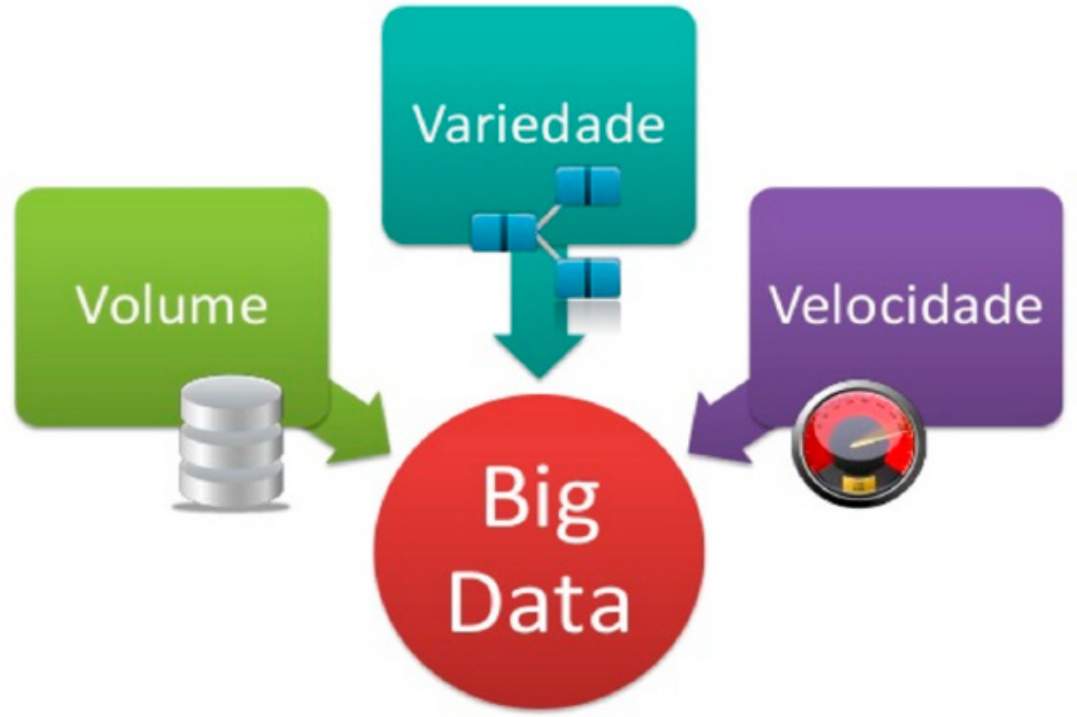


Figura. 3 Dimensões (3 V´s) do Big Data

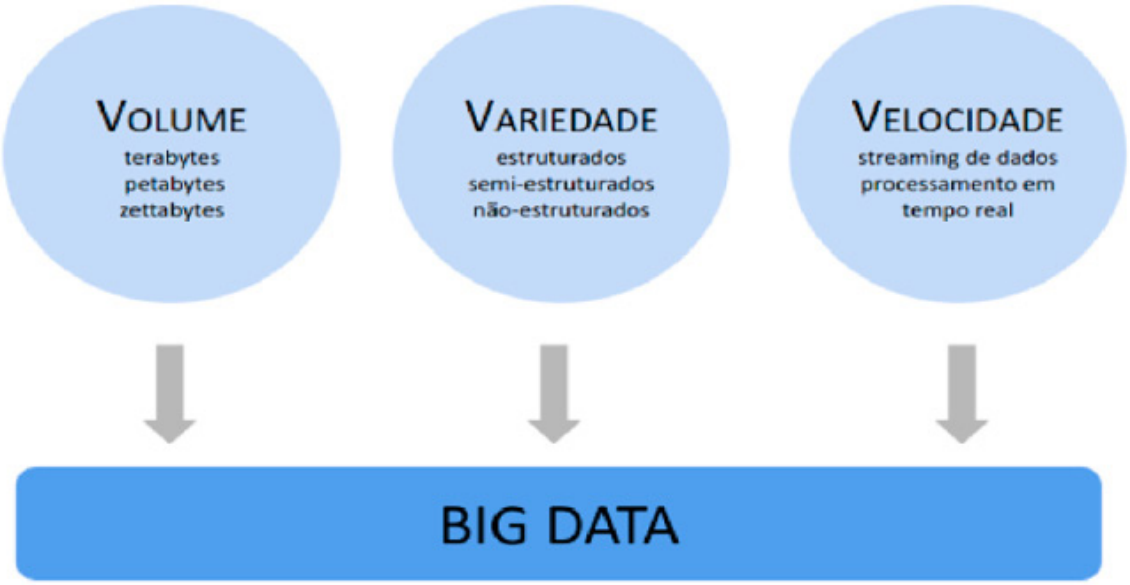


Figura. Livro Big Data: Técnicas e tecnologias para extração de valor dos dados", por Rosangela Marquesone

Figura. 4 Dimensões (4 Vs) do Big Data

Vamos à descrição dessas **cinco dimensões – 5V's – do Big Data**, que são de grande importância para a prova.

Figura. 5 Vs do Big Data

Volume	O volume da informação refere-se ao fato de que certas coleções de dados atingem a faixa de gigabytes (bilhões de bytes), terabytes (trilhões), petabytes (milhares de trilhões) ou mesmo exabytes (milhões de trilhões). Assim, o Big Data deve possibilitar a **análise de grandes volumes de dados**. Além disso, a tecnologia do Big Data serve exatamente para **lidar com esse volume de dados, guardando-os em diferentes localidades e juntando-os através de software.**
Velocidade	**Está relacionada à rapidez com a qual os dados são produzidos e tratados para atender à demanda**, o que significa que não é possível armazená-los por completo, de modo que somos obrigados a escolher dados para guardar e outros para descartar. A tecnologia de Big Data agora nos permite **analisar os dados no momento em que estes são gerados, SEM a necessidade de inseri-los nos bancos de dados**. Exemplos de uso envolvendo a tomada de decisão em tempo real: detecção de fraude em transação financeira; detecção de doença grave em *check-up* etc.

Variedade	O Big Data deve ser capaz de **lidar com diferentes formatos de informação**, como, por exemplo, arquivos de texto, ***e-mail***, medidores e sensores de coleta de dados, vídeo, áudio, dados de ações do mercado ou transações financeiras. Dados são gerados em inúmeros formatos – desde estruturados (numéricos, em ***databases*** tradicionais) a não estruturados (documentos de texto, ***e-mail***, vídeo, áudio, cotações da bolsa e transações financeiras etc.).
Veracidade	Quanto à veracidade, Weber ***et. al.*** (2009) ressaltou que as **informações verdadeiras podem ser usadas pelos gestores para responder aos desafios estratégicos**. A veracidade garantiria, então, a **confiabilidade dos dados**. Não adianta lidar com a combinação "volume + velocidade + variedade" se não houver dados confiáveis. É necessário que haja processos que garantam a consistência dos dados. A veracidade refere-se mais à **proveniência ou à confiabilidade da fonte de dados**, seu contexto e a sua utilidade para a análise com base nela.
Valor	Os dados do Big Data **devem agregar valor ao negócio**. O último V, valor, portanto, considera que informação é poder, informação é patrimônio. Com relação ao valor, Chen ***et. al.*** (2014) afirmam que as análises críticas de dados podem ajudar as empresas a melhor entender seus negócios trazendo benefícios. A combinação "volume + velocidade + variedade + veracidade", além de todo e qualquer outro aspecto que caracteriza uma solução de ***Big Data***, será inviável se o resultado não trouxer benefícios significativos e que compensem o investimento.

Fonte: https://goo.gl/QacUvf

Direto do Concurso

003. (FCC/DPE-RS/ANALISTA/BANCO DE DADOS/2017) Os sistemas de Big Data costumam ser caracterizados pelos chamados 3 Vs, sendo que o V de

a) Veracidade corresponde à rapidez na geração e obtenção de dados.
b) Valor corresponde à grande quantidade de dados acumulada.
c) Volume corresponde à rapidez na geração e obtenção de dados.
d) Velocidade corresponde à confiança na geração e obtenção dos dados.
e) Variedade corresponde ao grande número de tipos ou formas de dados.

Para analisar a **viabilidade de implementação do Big Data em uma organização**, a literatura citava inicialmente o **3V** (**Volume, Variedade e Velocidade**); depois o **5V** (incluindo aí a **Veracidade** e **Valor**).

Figura. 3 dimensões (3 Vs) do Big Data

Vamos à descrição dessas **cinco dimensões – 5Vs – do Big Data**:

Volume	**Corresponde à grande quantidade de dados acumulada.** Certas coleções de dados atingem a faixa de gigabytes (bilhões de bytes), terabytes (trilhões), **petabytes** (milhares de trilhões) ou mesmo **exabytes** (milhões de trilhões).
Variedade	**Corresponde ao grande número de tipos ou formas de dados**. Os dados de hoje aparecem em todos os tipos de formatos, como, por exemplo, arquivos de texto, e-mail, medidores e sensores de coleta de dados, vídeo, áudio, dados de ações do mercado ou transações financeiras.
Velocidade	**Corresponde à rapidez na geração e obtenção de dados.** Dessa forma, está relacionada à **rapidez** com a qual os dados são produzidos e tratados para atender à demanda, o que significa que não é possível armazená-los por completo, de modo que somos obrigados a escolher dados para guardar e outros para descartar.
Veracidade	**Corresponde à confiança na geração e obtenção dos dados.** Quanto à **veracidade**, Weber *et. al.* (2009) ressaltou que as informações verdadeiras podem ser usadas pelos gestores para responder aos desafios estratégicos. **A veracidade garantiria, então, a confiabilidade dos dados.**
Valor	O último V, **valor**, **considera que informação é poder, informação é patrimônio**. Com relação ao **valor**, as análises críticas de dados podem ajudar as empresas a melhor entender seus negócios trazendo benefícios. A combinação "volume + velocidade + variedade + veracidade", além de todo e qualquer outro aspecto que caracteriza uma solução de *Big Data*, será inviável se o resultado não trouxer benefícios significativos e que compensem o investimento.

Conforme visto, a letra E destaca a resposta correta.

Letra e.

A literatura já destaca os **7 Vs do Big Data**: englobando os **5 V's** (**Volume, Velocidade, Variedade, Veracidade, Valor**), **a Visualização e a Variabilidade.**

Visualização	**Busca tornar os dados visíveis para os analistas de dados**, por exemplo, permitindo que se obtenha a compreensão sobre os dados, e comunicar conceitos e ideias importantes. As atuais ferramentas de visualização de Big Data enfrentam desafios técnicos devido às limitações da tecnologia (memória, por exemplo) e à baixa escalabilidade, funcionalidade e tempo de resposta. Não se pode confiar em gráficos tradicionais ao tentar plotar um bilhão de pontos de dados, portanto, precisamos de diferentes formas de representar dados, como *clustering* de dados ou usando mapas de árvore, diagramas de rede circulares etc. Combine isso com a multiplicidade de variáveis resultantes da variedade e velocidade do Big Data e as relações complexas entre eles, e pode-se ver que o desenvolvimento de uma visualização significativa não é fácil.
Variabilidade	Pode aparecer de diversas formas, destacadas a seguir. **-Variação nas taxas de fluxo de dados (ou velocidade inconstante na carga dos dados).** Muitas vezes, a velocidade de Big Data não é consistente e fluxos podem ser altamente inconsistentes com picos periódicos. Todos os dias, picos de dados sazonais ou gerados por eventos particulares podem ser difíceis de gerenciar, ainda mais com dados não estruturados. -**Multiplicidade de dimensões de dados resultantes de diferentes fontes de dados** (**Complexidade** refere-se ao fato de Big Data gerar ou receber informações através de uma multiplicidade de fontes). Isso impõe um desafio crucial: a necessidade de se conectar, integrar, limpar e transformar os dados recebidos de diferentes fontes. **-Número de inconsistências nos dados.** **Nota:** A SAS (Em https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html) apresentou variabilidade (e complexidade) como duas dimensões adicionais para Big Data.

Mais recentemente, a IBM cita **7 dimensões** **que devem ser consideradas ao avaliar a viabilidade de uma solução de *Big Data***.

Figura. Dimensões a serem consideradas ao avaliar a viabilidade de uma solução de big data. Fonte: http://www.ibm.com/developerworks/br/library/bd-archpatterns2/

São elas:

- **Volume** dos dados que são capturados;
- **Variedade** das fontes, tipos e formatos dos dados;
- **Velocidade** na qual os dados são gerados, a velocidade em que é preciso agir com relação a eles ou a taxa em que estão mudando;
- **Veracidade** dos dados, ou seja, a incerteza ou fidelidade dos dados;
- **Valor** de negócios do insight que pode ser obtido ao analisar os dados;
- **Pessoas** com aptidões relevantes disponíveis e compromisso de patrocinadores de negócios. Tais aptidões incluem conhecimento do segmento de mercado, domínio técnico sobre as ferramentas de Big Data e conhecimentos específicos de modelagem, estatística, matemática etc.;
- Considerações sobre **governança** para as novas fontes de dados e a maneira como os dados serão usados.

Conforme destaca https://goo.gl/pr7ksF, ao decidir pela implementação ou não de uma plataforma de *big data*, uma organização pode estar olhando novas fontes e novos tipos de elementos de dados nos quais a propriedade do dia não está definida de forma clara. Alguns regulamentos do segmento de mercado regem os dados que são adquiridos e usados por uma organização. Por exemplo, no caso de assistência médica, é legal acessar dados de paciente para obter *insight*? Além da **questão da governança de TI**, também **pode ser necessário redefinir ou modificar os processos de negócios de uma organização para que ela possa adquirir, armazenar e acessar dados externos**.

Veja a seguir **questões relacionadas à governança** (https://goo.gl/pr7ksF):

- **Segurança e privacidade** – Cumprindo os regulamentos locais, quais dados a solução pode acessar? Quais dados podem ser armazenados? Quais dados devem ser criptografados durante a movimentação? Quem pode ver os dados brutos e os *insights*?
- **Normatização dos dados** – Existem normas que regem os dados? Os dados estão em um formato proprietário? Parte dos dados está em um formato fora da norma?
- **Intervalo de tempo em que os dados estão disponíveis** – Os dados estão disponíveis em um intervalo de tempo que permita agir de forma rápida?
- **Propriedade dos dados** – Quem detém a posse dos dados? A solução tem acesso e permissão apropriados para usar os dados?
- **Usos permissíveis:** Como é permitido usar os dados?

Em https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx o autor já referencia os **10 Vs do Big Data**, que englobam os elementos:

1	**Volume**	Análise de grandes volumes de dados. Guarda os dados em diferentes localidades e juntando-os através de software.
2	**Velocidade**	**Está relacionada à rapidez com a qual os dados são produzidos e tratados para atender à demanda**. Analisar os dados no instante em que são criados, sem ter de armazená-los em bancos de dados.
3	**Variedade**	O Big Data deve ser capaz de **lidar com diferentes formatos de informação**, que são: fontes estruturadas, semiestruturadas e a grande maioria em fontes não estruturadas.
4	**Veracidade**	**Informações verdadeiras podem ser usadas pelos gestores para responder aos desafios estratégicos.** A veracidade garantiria, então, a **confiabilidade dos dados.**
5	**Valor**	Os dados do Big Data **devem agregar valor ao negócio.**
6	**Visualização**	Maneiras diferentes de representar dados.

7	**Variabilidade**	**Variação nas taxas de fluxo de dados (ou velocidade inconstante na carga dos dados).** **Complexidade** – refere-se ao fato de Big Data gerar ou receber informações através de uma multiplicidade de fontes). Número de inconsistências nos dados.
8	**Validade**	Semelhante à veracidade, validade **refere-se à precisão e à correção dos dados para o uso pretendido.** De acordo com a Forbes, estima-se que 60% do tempo de um cientista de dados é gasto na limpeza de seus dados antes de poder fazer qualquer análise. O benefício da análise de Big Data é tão bom quanto os dados subjacentes, portanto, é necessário adotar boas práticas de controle de dados para garantir a qualidade consistente dos dados, definições comuns e metadados (TDWI, 2017).
9	**Vulnerabilidade**	Big Data **traz novas preocupações de segurança. Afinal, uma violação de dados com Big Data é uma grande preocupação.** Alguém se lembra do site AshleyMadison hackeado em 2015? Infelizmente, muitas grandes violações de dados foram reportadas na midia. Outro exemplo, conforme relatado pela CRN: em maio de 2016, "um hacker chamado Peace postou dados na *dark web* para vender, que supostamente incluía informações sobre 167 milhões de contas do LinkedIn e 360 milhões de *e-mails* e senhas para usuários do MySpace".
10	**Volatilidade**	Quantos anos seus dados precisam ter antes de serem considerados irrelevantes, históricos ou inúteis? Por quanto tempo os dados precisam ser mantidos? Devido à velocidade e volume de Big Data, no entanto, sua volatilidade precisa ser cuidadosamente considerada. **É preciso estabelecer regras para o armazanemanto e a garantia da disponibilidade de dados, além de permitir a rápida recuperação das informações quando necessário.** Certifique-se de que estes estejam claramente vinculados às necessidades e aos processos comerciais - com Big Data, os custos e a complexidade de um processo de armazenamento e recuperação são ampliados.

1.9. Camadas Lógicas de uma Solução de Big Data

Conforme destaca (MYSORE; KHUPAT; JAIN, 2014), as **camadas** proporcionam uma maneira de organizar componentes que realizam funções específicas.

Uma solução de *Big Data* possui camadas horizontais e verticais (MYSORE; KHUPAT; JAIN, 2014):

Camada Horizontal	Camada Vertical
Camadas de "baixo" para "cima" na figura. São elas: Fontes de *Big Data*, Camada de Tratamento e Armazenamento de Dados, Camada de Análise, e Camada de Consumo.	Lidam com aspectos que afetam todos os componentes das camadas lógicas (fontes de big data, tratamento e armazenamento de dados, análise e consumo). São elas: Integração de informações, Governança de *Big Data*, Gerenciamento de sistemas, e Qualidade de serviço.

Vamos à descrição dessas camadas!

Camadas Horizontais

Tabela: Camadas horizontais (MYSORE; KHUPAT; JAIN, 2014)

Camada Horizontal	Descrição
Fontes de Big Data	**Inclui todas as fontes de dados necessárias para proporcionar o *insight* necessário para solucionar o problema de negócios**. Os dados são estruturados, semiestruturados e não estruturados e são provenientes de várias fontes: • sistemas corporativos legados; • sistemas de gerenciamento de dados; • armazenamentos de dados (incluem armazéns de dados corporativos e bancos de dados operacionais e transacionais); • dispositivos inteligentes (podem capturar, processar e comunicar informações na maioria dos protocolos e formatos mais usados. Por exemplo, smartphones, medidores e dispositivos de assistência médica); • outras fontes de dados, como: *informações geográficas; conteúdo gerado por seres humanos:* Mídia social/Email/Blogs/Informações online; dados de sensor etc.

Camada Horizontal	Descrição
Camada de tratamento e armazenamento de dados	**Responsável por adquirir dados das fontes e, se necessário, convertê-los para um formato adequado à maneira como os dados devem ser analisados.** Atividades:

Camada de tratamento e armazenamento de dados	**Aquisição de dados** – Adquire dados de várias fontes e os envia ao componente de digestão de dados ou armazena em locais específicos. Esse componente precisa ser inteligente o suficiente para decidir se deve armazenar os dados recebidos e onde armazená-los. Deve poder determinar se é necessário tratar os dados antes de armazená-los ou se é possível enviar os dados diretamente para a camada de análise de negócios. **Compilação de dados** – Responsável por tratar os dados no formato necessário para atingir o objetivo da análise. Esse componente pode ter lógica transformacional simples ou algoritmos estatísticos completos para converter os dados de origem. O mecanismo de análise determina os formatos específicos de dados que são necessários. O maior desafio é acomodar formatos de dados não estruturados, como imagens, áudio, vídeo etc. **Armazenamento de dados distribuídos**– Responsável por armazenar os dados das fontes. Frequentemente há várias opções de armazenamento de dados disponíveis nessa camada, como *distributed file storage* (DFS), nuvem, fontes de dados estruturados, **NoSQL** etc.
Camada de Análise	**Lê os dados digeridos pela camada de tratamento e armazenamento de dados.** Em alguns casos, a camada de análise acessa os dados diretamente na fonte. É fundamental um planejamento cuidadoso para projetar a camada de análise. É necessário tomar decisões em relação a como gerenciar tarefas para: • produzir a análise desejada; • obter *insights* a partir dos dados; •localizar as entidades necessárias; • localizar as fontes de dados que fornecem dados para essas entidades; •entender quais algoritmos e ferramentas são necessários para realizar a analítica.
Camada de Consumo	**Essa camada consome a saída fornecida pela camada de análise.** Os consumidores podem ser aplicativos de visualização, seres humanos, processos de negócios ou serviços. Pode ser difícil visualizar a saída da camada de análise. Às vezes é útil ver o que os concorrentes em mercados semelhantes estão fazendo.

Cada camada inclui vários tipos de componentes, como ilustrado a seguir.

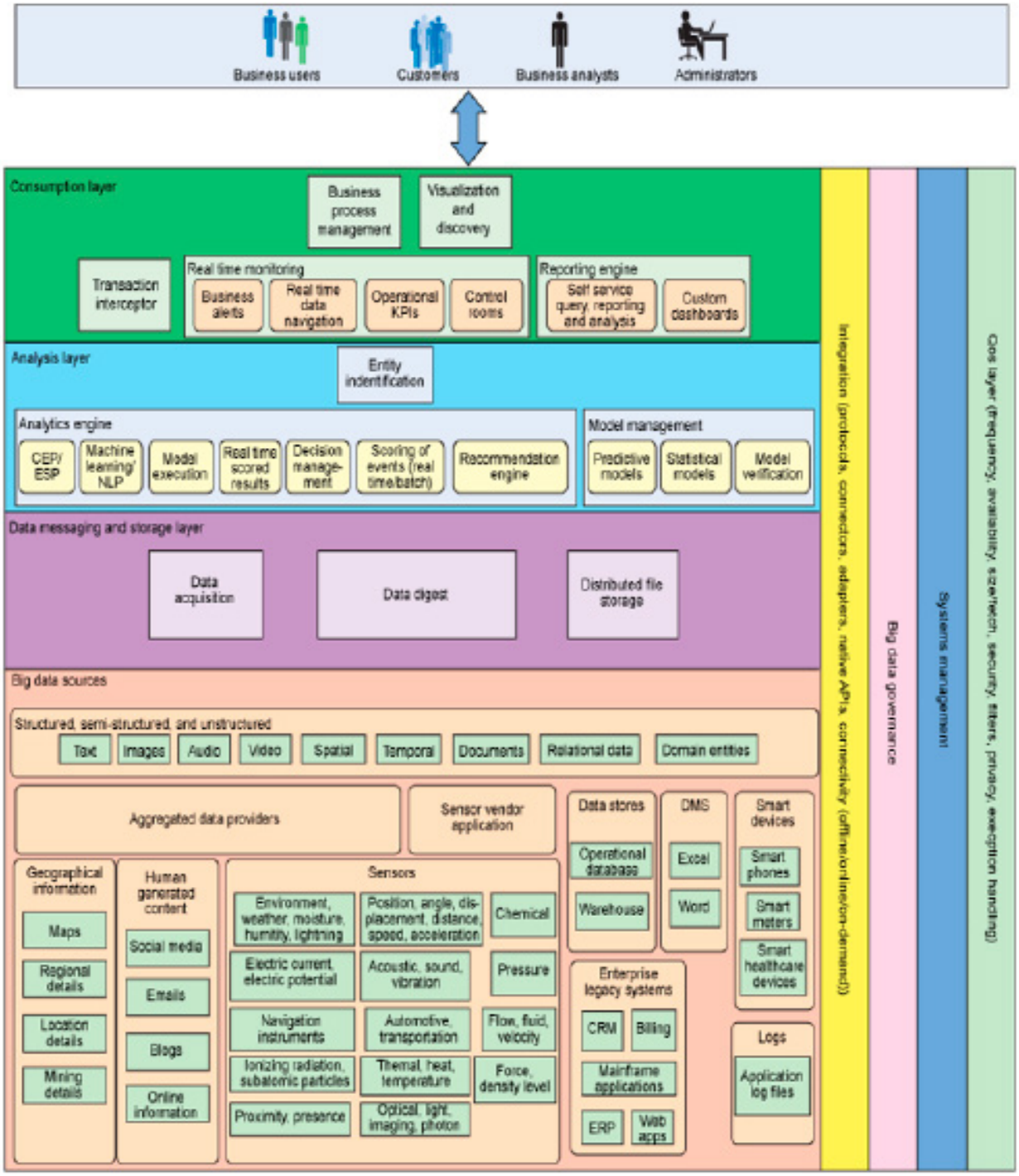


Figura. Componentes por camada
Fonte: http://www.ibm.com/developerworks/br/library/bd-archpatterns3/

Camadas Verticais

Camada Vertical	Descrição
Integração de Informações	Aplicativos de ***Big Data*** adquirem dados de várias origens, fornecedores e fontes. **Essa camada vertical é usada por vários componentes** (aquisição de dados, compilação de dado, gerenciamento de modelo e interceptor de transação, por exemplo) **e é responsável por conectar várias fontes de dados**. Também pode ser usada por componentes para armazenar informações em armazenamentos de big data e para recuperar informações desses armazenamentos para processamento. A maioria dos armazenamentos de big data possui serviços e APIs para armazenar e recuperar as informações.
Governança de *Big Data*	Ajuda a lidar com as complexidades, o volume e a variedade de dados dentro da empresa ou oriundos de fontes externas. São necessários diretrizes e processos sólidos para monitorar, estruturar, armazenar e proteger os dados desde o momento em que entram na empresa, são processados, armazenados, analisados e removidos ou arquivados. A governança para big data inclui fatores, como: gerenciar grandes volumes de dados em diversos formatos; treinar e gerenciar continuamente os modelos estatísticos necessários para pré-processar dados não estruturados e analítica (Lembre-se que essa etapa é importante ao lidar com dados não estruturados!); definir política e regulamentos de conformidade para retenção e uso de dados externos; definir políticas de arquivamento e remoção de dados; criar a política sobre a maneira como os dados podem ser replicados em vários sistemas; definir políticas de criptografia de dados.
Gerenciamento de sistemas	Gerenciamento de sistema é essencial para big data e inclui as seguintes ações: gerenciar os logs de sistemas, máquinas virtuais, aplicativos e outros dispositivos; correlacionar os vários logs e ajudar a investigar e monitorar a situação; monitorar alertas e notificações em tempo real; fazer referência a relatórios e análises detalhados sobre o sistema; definir e cumprir os contratos de nível de serviço; arquivar e gerenciar recuperação de arquivos; realizar recuperação de sistema etc.
Camada de qualidade de serviço	Responsável por definir qualidade de dados, políticas relacionadas à privacidade e segurança, frequência de dados, tamanho de busca e filtros de dados.

Tabela: Camadas verticais (MYSORE; KHUPAT; JAIN, 2014)

Conforme destaca https://www.ibm.com/developerworks/br/library/bd-archpatterns4/index.html, a IBM definiu uma série de **padrões** que nos ajuda a **definir a arquitetura da solução de Big Data**.

Esses padrões podem ser classificados em **atômicos (*Atomic Patterns*)**, **compostos (*Composite Patterns*)** e **de soluções (*Solution Patterns*)**.

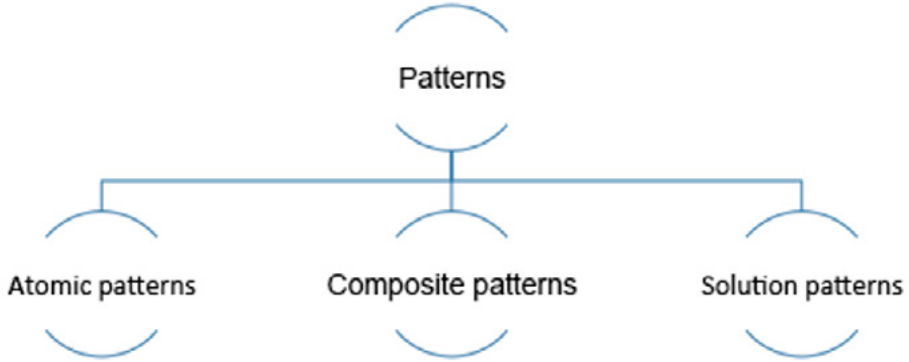


Figura. Padrões.

- Os **padrões atômicos** são os que fornecem as bases para a solução de *Big Data*.
- Os **padrões compostos e de solução** são mais abrangentes e variados, muitas vezes utilizando uma composição de padrões atômicos para definir a solução de Big Data.
- IBM também destaca que não há sequência ou ordem recomendada em que os padrões de solução, compostos ou atômicos devem ser aplicados para chegar a uma solução.

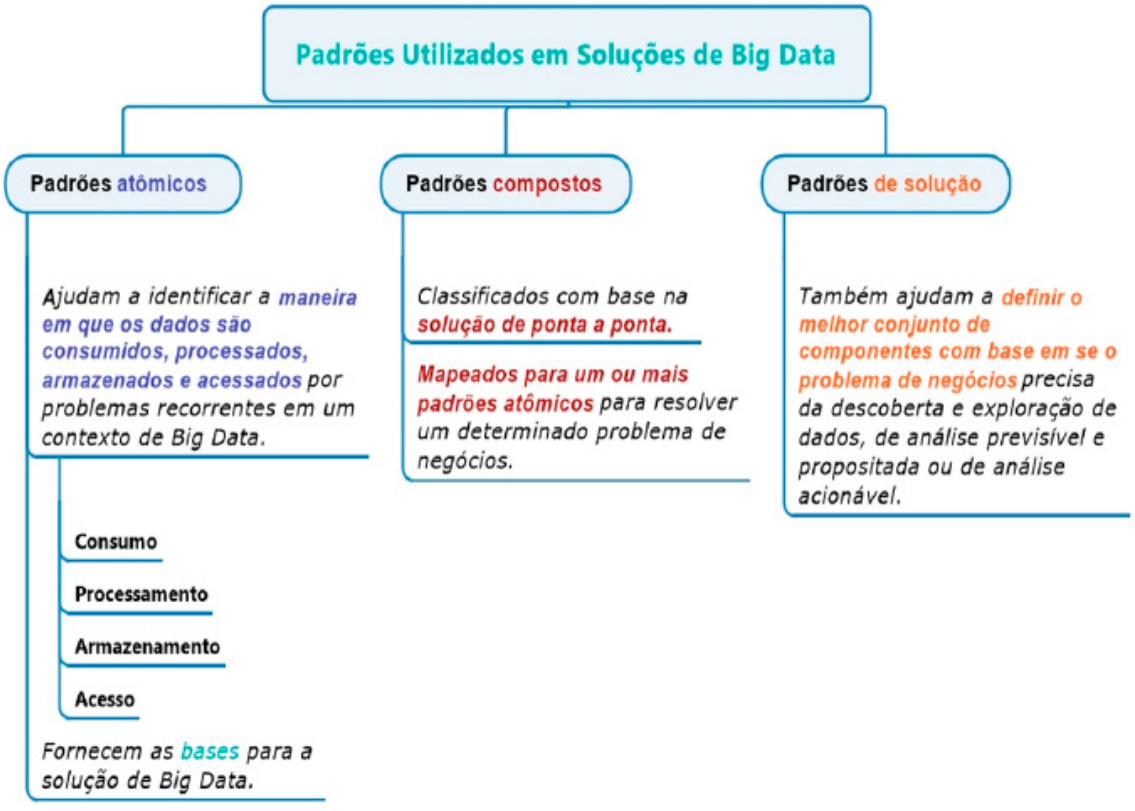


Figura. Padrões Utilizados em Soluções de Big Data. Fonte: Quintão (2020)

1.10. NoSQL (Not Only SQL – Não Só SQL)

NoSQL não significa "*no* SQL" (não ao SQL), mas sim "***not only* SQL (não só SQL)**.

O **NoSQL** (***Not only Structured Query Language***) é um termo genérico para uma classe definida de bancos de dados não relacionais, que têm uma propriedade chamada **BASE** (***B**asically **A**vailable, **S**oft state, **E**ventually consistency* – **Basicamente disponível, estado leve, eventualmente consistente**), que distribui os dados em diferentes repositórios tornando-os sempre disponíveis, não se preocupa com a consistência de uma transação, delegando essa função para a aplicação, porém sempre garante a consistência dos dados em algum momento futuro à transação.

> **Bancos do tipo NoSQL são mais flexíveis**, sendo compatíveis com um grupo de premissas que "compete" com as propriedades **ACID** (**A**tomicidade, **C**onsistência, **I**solamento e **D**urabilidade), **dos SGBDs (Sistemas Gerenciadores de Banco de Dados) tradicionais**: a **BASE** (***B**asically **A**vailable, **S**oft state, **E**ventually consistency* – **Basicamente disponível, estado leve, eventualmente consistente**).

Vamos relembrar as propriedades das transações, chamadas **Propriedades ACID.**

- **A**tomicidade
- **C**onsistência
- **I**solamento
- **D**urabilidade

A **atomicidade visa garantir que uma transação é uma unidade atômica de processamento (indivisível).** Logo, a transação será executada em sua totalidade, com todas as suas operações finalizadas e refletidas no BD, ou não será executada de modo algum, e nenhuma das suas operações são refletidas no BD.

Na **consistência**, a execução completa da transação deverá levar o banco de dados de um estado consistente para outro estado consistente, onde um estado consistente do banco de dados satisfaz as restrições especificadas no esquema, bem como quaisquer outras restrições que devam controlar o banco de dados.

O **isolamento determina que uma transação deve ser executada como se estivesse isolada das demais.** Cada transação assume que está sendo executada sozinha no sistema. O sistema garante que os resultados intermediários da transação permaneçam escondidos de outras transações executando concorrentemente.

Finalmente, a **durabilidade** busca assegurar que as mudanças aplicadas no banco de dados por uma transação efetivada devem persistir no banco de dados. Estas mudanças não devem ser perdidas em razão de uma falha.

É importante lembrar destes conceitos pois eles são sempre pedidos em provas...

Os **bancos de dados não relacionais (NoSQL)** não utilizam o esquema tradicional de tabela de linhas e colunas; em vez disso, eles **usam um modelo de armazenamento otimizado** para **desempenho escalável** e **modelos de dados sem esquema** (Cespe/2018).

Entre as **vantagens** desse modelo sobre o relacional estão (Cespe/2018):

- **escalabilidade horizontal**: na medida em que o volume de dados cresce, aumenta-se a necessidade de escalabilidade e melhoria do desempenho. Dentre todas as possibilidades para esta solução, a escalabilidade horizontal se torna a mais viável, porém requer diversas *threads* ou que processos de um tarefa sejam criadas e distribuídas;
- **flexibilidade na manipulação de dados não estruturados;**
- **melhor desempenho, custos reduzidos;**

- **o fato de serem projetados para arquiteturas distribuídas;**
- **suporte nativo a replicação**: esta é outra forma de prover a escalabilidade, pois, no momento em que permitimos a replicação de forma nativa o tempo gasto para recuperar informações é reduzido; e
- o fato de serem ideais para aplicações de ***Big Data***.

A **tabela seguinte**, extraída de Devmedia (2014), apresenta uma análise comparativa do **modelo de dados relacional e o modelo NoSQL**.

	Relacional	NoSQL
Escalonamento	Possível, mas complexo. Devido à **natureza estruturada do modelo**, a adição de forma dinâmica e transparente de novos nós no grid **não** é realizada de modo natural.	**Uma das principais vantagens desse modelo.** Por não possuir nenhum tipo de esquema predefinido, o modelo possui maior flexibilidade o que favorece a inclusão transparente de outros elementos.
Consistência	**Ponto mais forte do modelo relacional.** As regras de consistência presentes propiciam um maior grau de rigor quanto à consistência das informações.	**Realizada de modo eventual no modelo**: só garante que, se nenhuma atualização for realizada sobre o item de dados, todos os acessos a esse item devolverão o último valor atualizado.
Disponibilidade	Dada a dificuldade de se conseguir trabalhar de forma eficiente com a distribuição dos dados, esse modelo **pode não suportar** a demanda muito grande de informações do banco.	**Outro fator fundamental do sucesso desse modelo.** O alto grau de distribuição dos dados propicia que um maior número de solicitações aos dados seja atendida por parte do sistema e que o sistema fique menos tempo não disponível.

Tabela. Análise Comparativa Modelo Relacional x NoSQL.
Fonte: Devmedia (2014)

As figuras seguintes destacam uma **linha de tempo** nas quais são citados os principais modelos de dados:

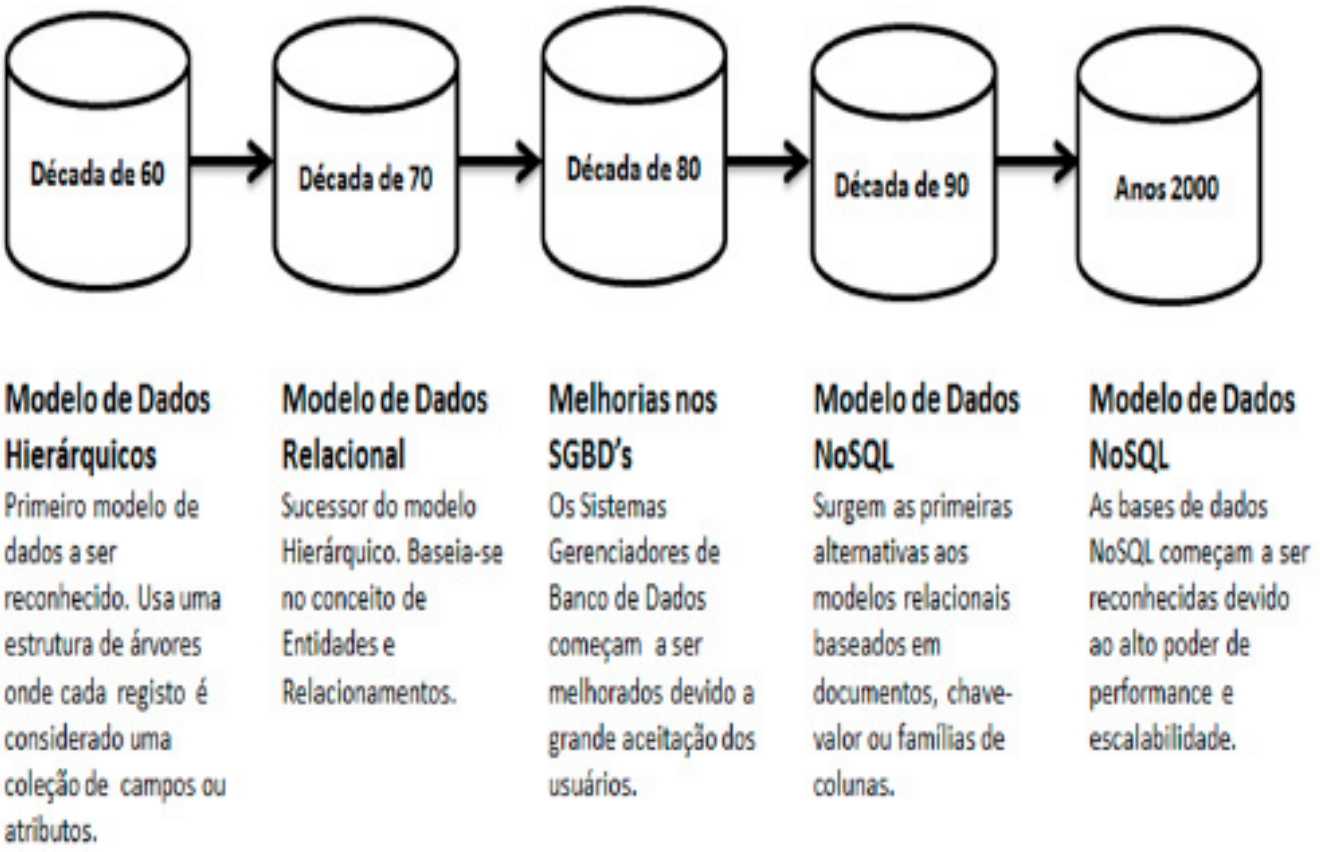
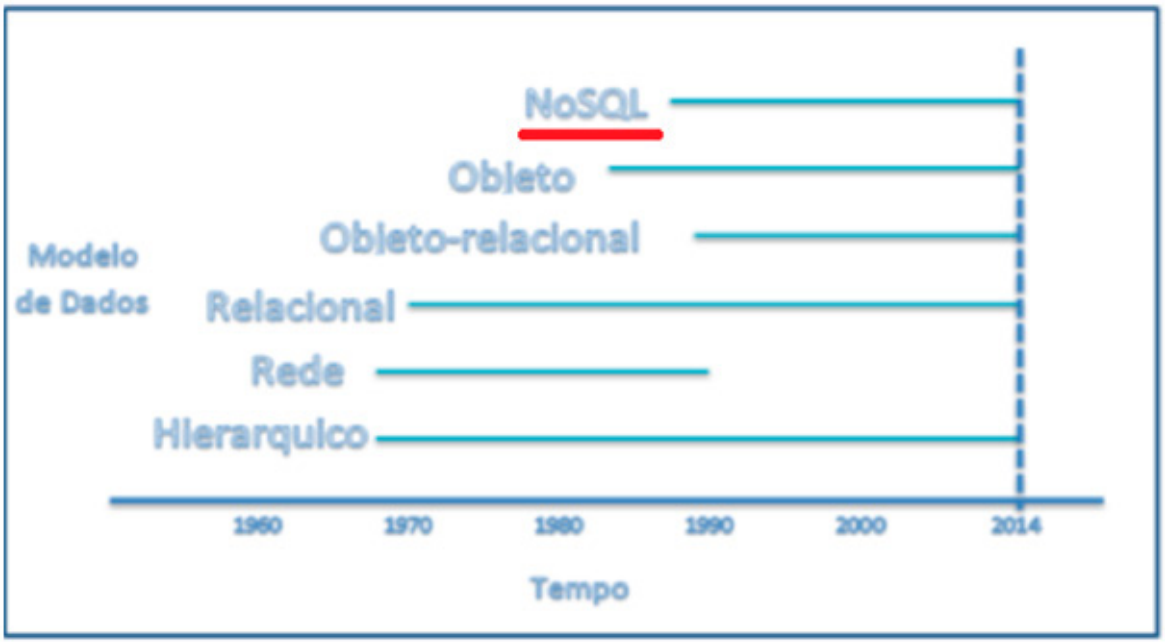


Figura. Modelo de Dados NoSQL (Fonte: **https://www.devmedia.com.br/repositorio-de-dados-relacional-ou--nosql/27500***)*

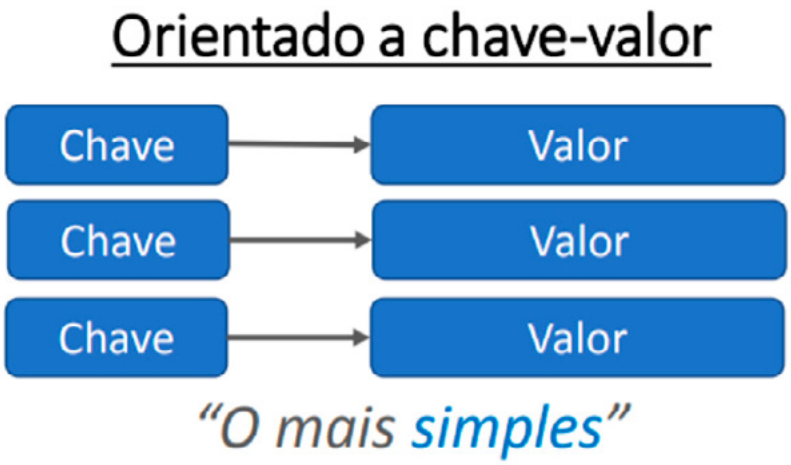


Obs.: Os bancos NoSQL usam diversos **modelos de dados**, como: chave/valor simples, colunares, documentos, gráficos e armazenamento de pares chave-valor na memória (Cespe/2018).

A seguir, destacamos os principais **modelos de gerenciamento de dados NoSQL** (Cespe/2018):

- **Modelo Chave/Valor (*key/value*)**: **sistemas distribuídos nessa categoria,** também conhecidos como tabelas de *hash* distribuídas, **armazenam objetos indexados por chaves e possibilitam a busca por esses objetos a partir de suas chaves.**

São exemplos de bancos de dados que utilizam esse padrão: DynamoDb, Couchbase, Riak, Azure Table Storage, Redis, Tokyo Cabinet e Berkeley DB.



Fonte: Modelo de Gerenciamento de Dados NoSQL Orientado a Chave-Valor. Fonte: Marquesone (2019)

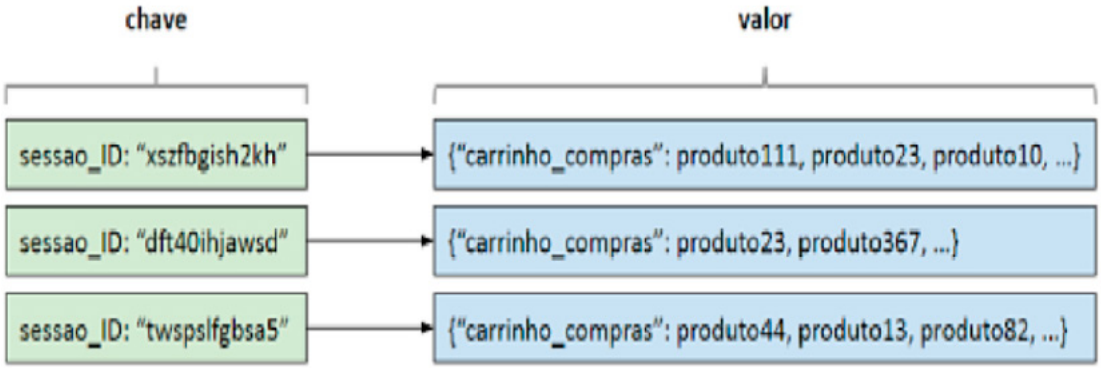
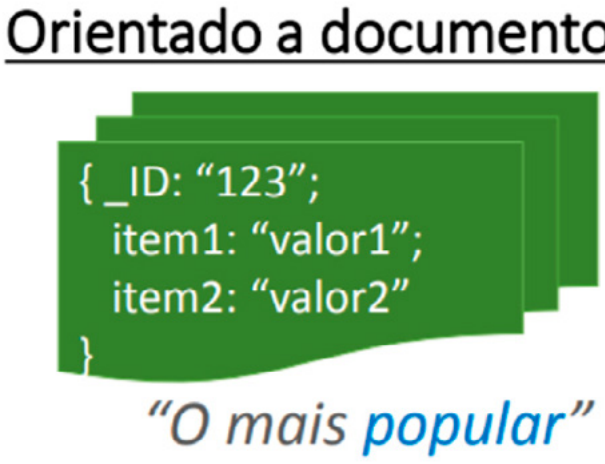


Figura. Estrutura de um banco de dados orientado a chave-valor. Fonte: Livro Big Data: Técnicas e tecnologias para extração de valor dos dados", por Rosangela Marquesone

- **Modelo Orientado a Documentos**: **os documentos** dos bancos dessa categoria **são coleções de atributos e valores, nas quais um atributo pode ser multivalorado.**

Em geral, os bancos de dados orientados a documentos não possuem esquema, ou seja, os documentos armazenados não precisam possuir estrutura em comum. Os dados nos campos de um documento podem ser codificados de várias maneiras, incluindo XML, YAML, JSON, BSON ou até mesmo armazenados como texto sem formatação.

Alguns bancos que utilizam esse padrão são: MongoDb, CouchDB e RavenDb.



Fonte: Modelo de Gerenciamento de Dados NoSQL Orientado a Documentos. Fonte: Marquesone (2019)

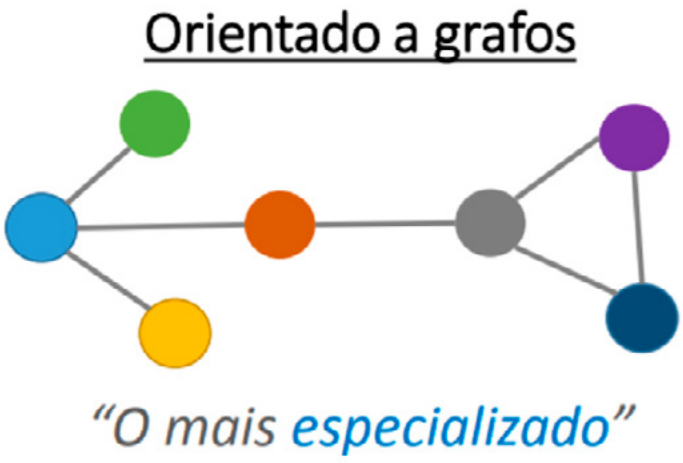
- **Modelo Orientado a Grafos**: diferentemente de outros tipos de bancos de dados NoSQL, esse está **diretamente relacionado a um modelo de dados estabelecido, o modelo de grafos**.

A ideia desse modelo é representar os dados e(ou) o esquema dos dados como grafos dirigidos, ou como estruturas que generalizem a noção de grafos.

O modelo de grafos é mais interessante que outros quando informações sobre a interconectividade ou a topologia dos dados são mais importantes ou tão importantes quanto os dados propriamente ditos. O modelo orientado a grafos **possui três componentes básicos**:

- os **nós** (são os vértices do grafo);
- os **relacionamentos** (são as arestas); e
- as **propriedades** (ou atributos) **dos nós** e **relacionamentos**.

Um exemplo da utilidade desse tipo de banco é resolver perguntas como: “Com quais pessoas e por quanto tempo um cliente falou nos últimos 7 dias?” Alguns bancos que utilizam esse padrão: Neo4J, Infinite Graph, InforGrid e HyperGraphDB.



Fonte: Modelo de Gerenciamento de Dados NoSQL Orientado a Grafos. Fonte: Marquesone (2019)

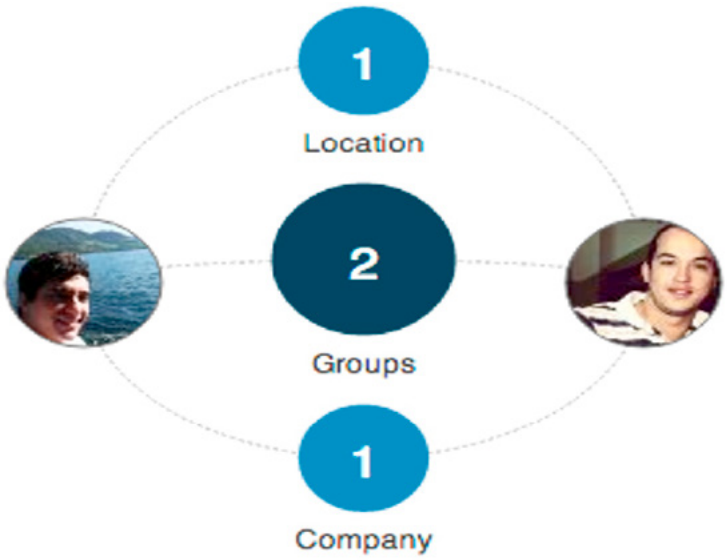
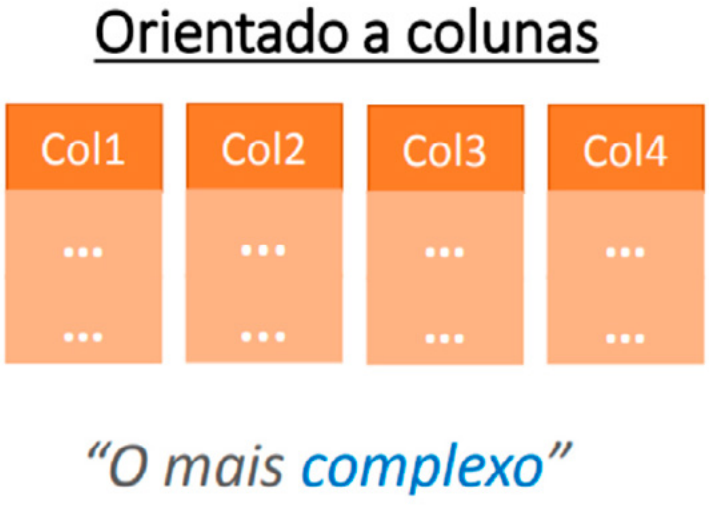


Figura. Grafo ilustrando relacionamentos entre duas pessoas. Fonte: https://www.devmedia.com.br/analisando-o-big-data-na-teoria-e-na-pratica/30129

- **Modelo Orientado a Colunas**: um armazenamento de dados de colunas ou de família de colunas **organiza os dados em colunas e linhas**.

É possível pensar em um armazenamento de dados de família de colunas como dados contidos em tabela com linhas e colunas, mas as colunas são divididas em grupos conhecidos como famílias de colunas.

Cada família de colunas contém um conjunto de colunas que estão logicamente relacionadas e geralmente são recuperadas ou manipuladas como uma unidade. Outros dados acessados separadamente podem ser armazenados em famílias de colunas separadas.



Fonte: Modelo de Gerenciamento de Dados NoSQL Orientado a Colunas. Fonte: Marquesone (2019)

- **Modelo Orientado a Série Temporal**: os dados de série temporal são um conjunto de valores organizados por tempo e um armazenamento de dados de série temporal é otimizado para esse tipo de dados.

Os armazenamentos de dados de série temporal devem dar suporte a um número muito alto de gravações, pois geralmente coletam grandes quantidades de dados, em tempo real, de uma grande variedade de fontes.

Exemplos de bancos de dados NoSQL:

Apache Cassandra	Originalmente criado pelo Facebook, no qual os **dados são identificados por meio de uma chave**. É um projeto de sistema de banco de dados distribuído, altamente escalável, que foi desenvolvido na plataforma Java. Reúne a arquitetura do Dynamo da Amazon e o modelo de dados do BigTable da Google. Teve seu código-fonte aberto à comunidade em 2008. Atualmente é mantido por desenvolvedores da fundação Apache e colaboradores de outras empresas.
MongoDB	É um banco de dados **orientado a documentos**, escalável, de alto desempenho e código aberto escrito em C++.
Apache CouchDB	É um banco de dados **orientado a documentos** de código fonte aberto escrito em linguagem Erlang. Foi desenvolvido e mantido pela fundação Apache e busca replicação e escalabilidade horizontal.
BigTable	Criado pelo **Google** e usado pelo **GFS *(Google File System)*** para gerenciar *petabytes* de informações, o BigTable é um sistema de armazenamento de dados distribuído que é estruturado como uma enorme tabela. Atualmente esse projeto é utilizado em diversas aplicações do Google, principalmente naquelas que demandam alta capacidade de armazenamento e baixa latência de rede, como o processo de indexação de web sites e os produtos: *Google Earth, Maps, Adwords, Analytics, Adsense, Youtube e Gmail*.
Dynamo	Desenvolvido pela **Amazon** em 2007, foi criado para oferecer **armazenamento de valores-chaves** de dados de alta disponibilidade, permitindo atualizações para sobreviver a falhar de servidor e rede.
HBase	É um banco de dados open source semelhante ao BigTable, que **utiliza o Hadoop**.
Redis	Um banco de dados NoSQL capaz de armazenar dados no **formato chave-valor**.

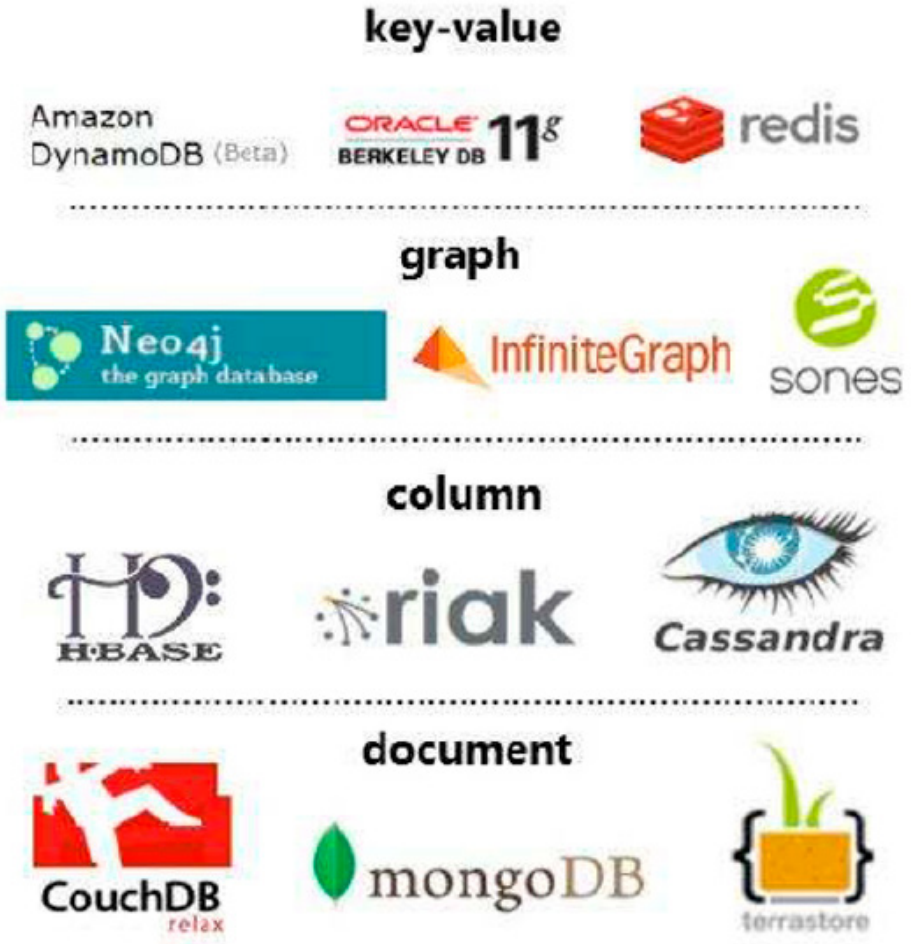


Figura. Banco de Dados NoSQL

1.11. Hadoop

Quando nos referimos a *Big Data*, apenas um banco de dados do tipo **não** basta. É necessário também contar com **ferramentas** (Ex.: **Hadoop** é a principal referência) **que permitam o tratamento correto do volume de dados**.

- **Hadoop**: **plataforma *open source*** desenvolvida especialmente para processamento e análise de grandes volumes de dados, sejam eles estruturados ou não estruturados.
  - **É utilizado em larga escala por grandes corporações**, como Facebook e Twitter, **em aplicações Big Data**.
  - Útil para aplicações que envolvam dados massivos para processamento paralelo (embora seja interessante para processamento de quaisquer dados), geralmente utilizando um cluster de computadores (Devmedia, 2016).
  - Trata-se de um **projeto da Apache** de alto nível, que vem sendo construído por uma comunidade de colaboradores utilizando em sua maior parte a linguagem de programação Java, com algum código nativo em C e alguns **utilitários de linha de comando** escrito utilizando *scripts shell* (Wikipedia, 2016).

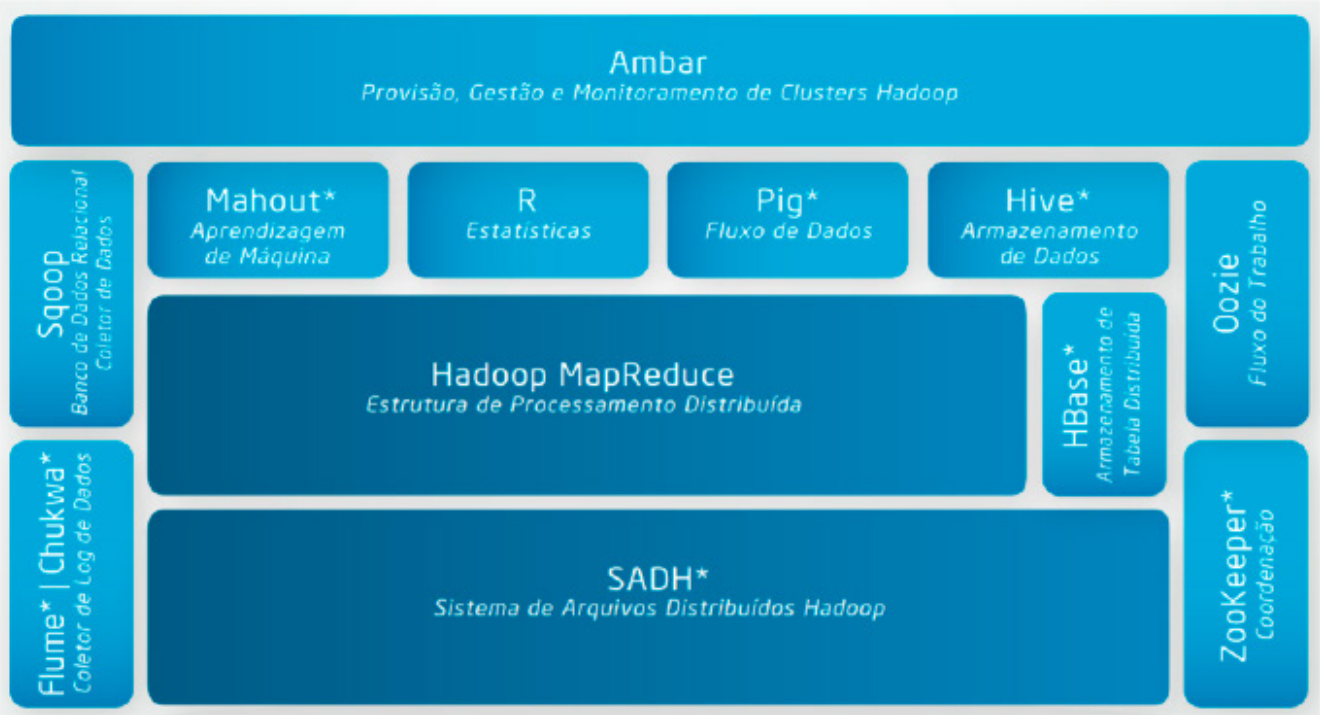


Figura. Plataforma Apache Hadoop. O pacote de software Hadoop inclui uma série de componentes. Fonte: Intel Corporation (2013)

Obs.: O **Hadoop** é um **projeto Apache**, sendo criado e mantido por uma comunidade de empresas e profissionais. Foi **inspirado no MapReduce e no GoogleFS.**

- Pode-se dizer que o projeto teve início em meados de 2003, quando o Google criou um **modelo de programação que distribui o processamento a ser realizado entre vários computadores para ajudar o seu mecanismo de busca a ficar mais rápido e livre das necessidades de servidores poderosos** (e caros). Esta tecnologia recebeu o nome de **MapReduce.**
- **Plataforma de software em Java, de computação distribuída,** voltada para *clusters* e processamento de grandes massas de dados, **inspirada no MapReduce e no GoogleFS (GFS).**

Direto do Concurso

004. (ESAF/ANAC/ANALISTA ADMINISTRATIVO/2016) Para o processamento de grandes massas de dados, no contexto de Big Data, é muito utilizada uma plataforma de software em Java, de computação distribuída, voltada para clusters, inspirada no MapReduce e no GoogleFS. Esta plataforma é o(a)

a) Yam Common.

b) GoogleCrush.

c) EMRx.

d) Hadoop.

e) MapFix.

Hadoop é uma solução de código aberto (*open source*), inspirada no MapReduce e no GoogleFS, que permite a execução de aplicações de Big Data utilizando milhares de máquinas. Oferece recursos de armazenamento, gerenciamento e processamento distribuído de dados.

Letra d.

005. (ESAF/ESAF/GESTÃO E DESENVOLVIMENTO DE SISTEMAS/2015/ADAPTADA) O Hadoop, o mais conhecido e popular sistema para gestão de Big Data, foi criado pela IBM, a partir de sua ferramenta de Data Mining WEKA.

O **Hadoop** é um **projeto Apache**, sendo criado e mantido por uma comunidade de empresas e profissionais. Foi **inspirado no MapReduce e no GoogleFS** e não no Data Mining WEKA destacado na questão!

Errado.

2. Visualização e Análise Exploratória de Dados

O termo **visualização de dados** está relacionado às tecnologias que dão suporte à visualização e, algumas vezes, à interpretação de dados e informações em vários pontos ao longo da cadeia de processamento de dados (TURBAN, 2009).

Ela inclui imagens digitais, sistemas geográficos, interfaces gráficas de usuário, gráficos, realidade virtual, representações dimensionais, vídeos e animações (TURBAN, 2009).

As **ferramentas visuais** podem ajudar a identificar relações, como por exemplo, tendências. Ao usar ferramentas visuais de análise, gerentes, engenheiros e outros profissionais podem reconhecer problemas que passaram despercebidos, durante anos, pelos métodos de análise padrão (TURBAN, 2009).

A visualização de dados é mais fácil de implementar quando os dados necessários estão em um data warehouse ou, melhor ainda, em um banco de dados multidimensional especial ou servidor.

Desde o fim dos anos 90, a **visualização de dados evoluiu tanto na computação convencional**, em que é integrada às ferramentas e aplicações de suporte à decisão, como na **visualização inteligente**, que inclui a interpretação de dados (informação) (TURBAN, 2009).

3. Considerações Finais

- A **mineração de dados** usa ferramentas como modelos estatísticos, visualização e aprendizado de máquina para encontrar informações ou padrões a partir dos dados. **Big Data** procura aplicar essas ferramentas a dados de alto volume, alta velocidade ou alta variedade - isso é um desafio em bancos de dados e programas de análise mais antigos, por isso, temos a nova tecnologia de *Big Data*.

Leitura sugerida: 10 tendências para 2017 do Big Data em: https://www.tableau.com/pt-br/resource/top-10-big-data-trends-2017.

Mais uma leitura adicional sugerida: **10 grandes tecnologias de Big Data para 2018** em: http://reamp.com.br/blog/2017/12/10-grandes-tecnologias-de-big-data-para-2018/.

RESUMO

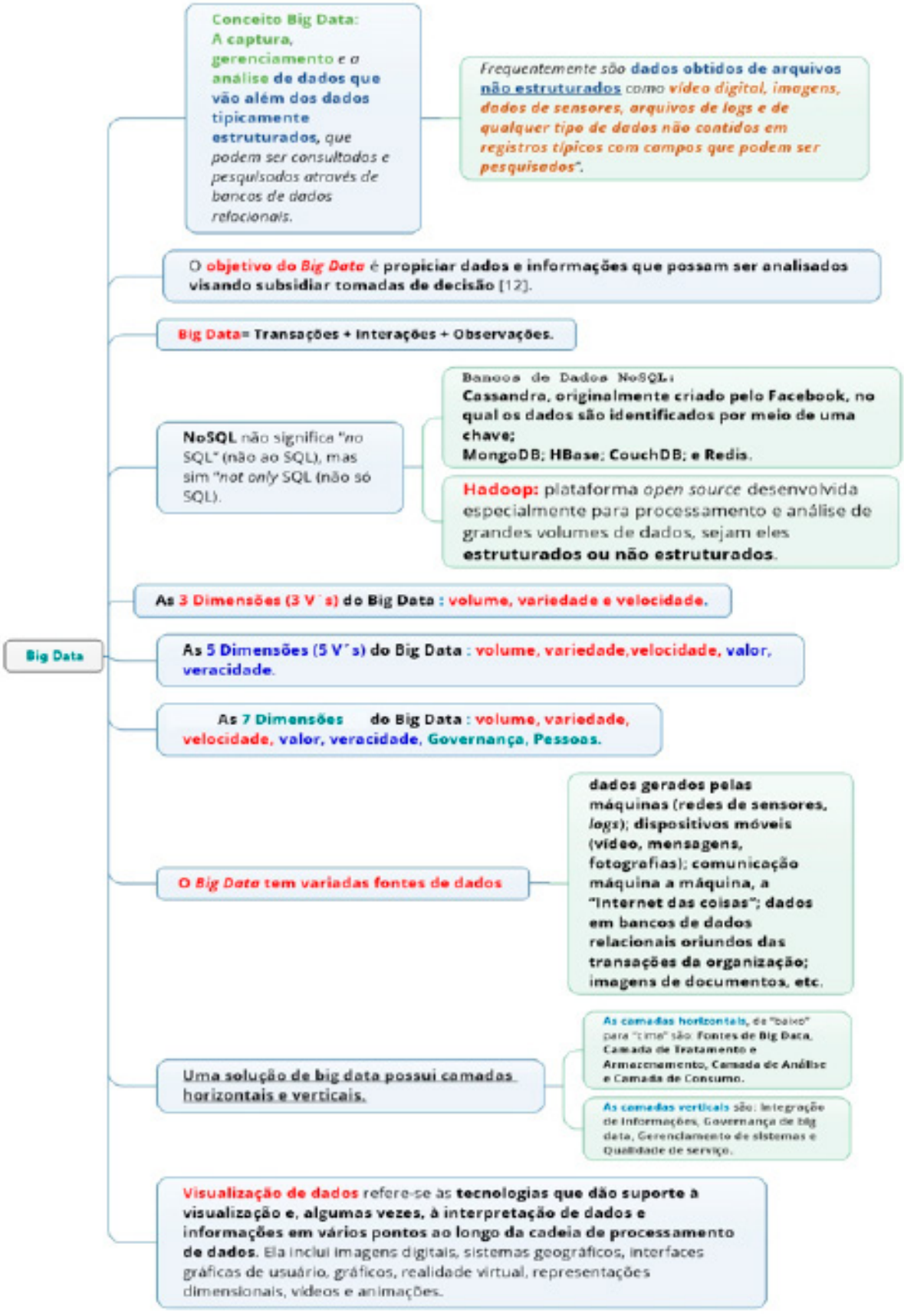
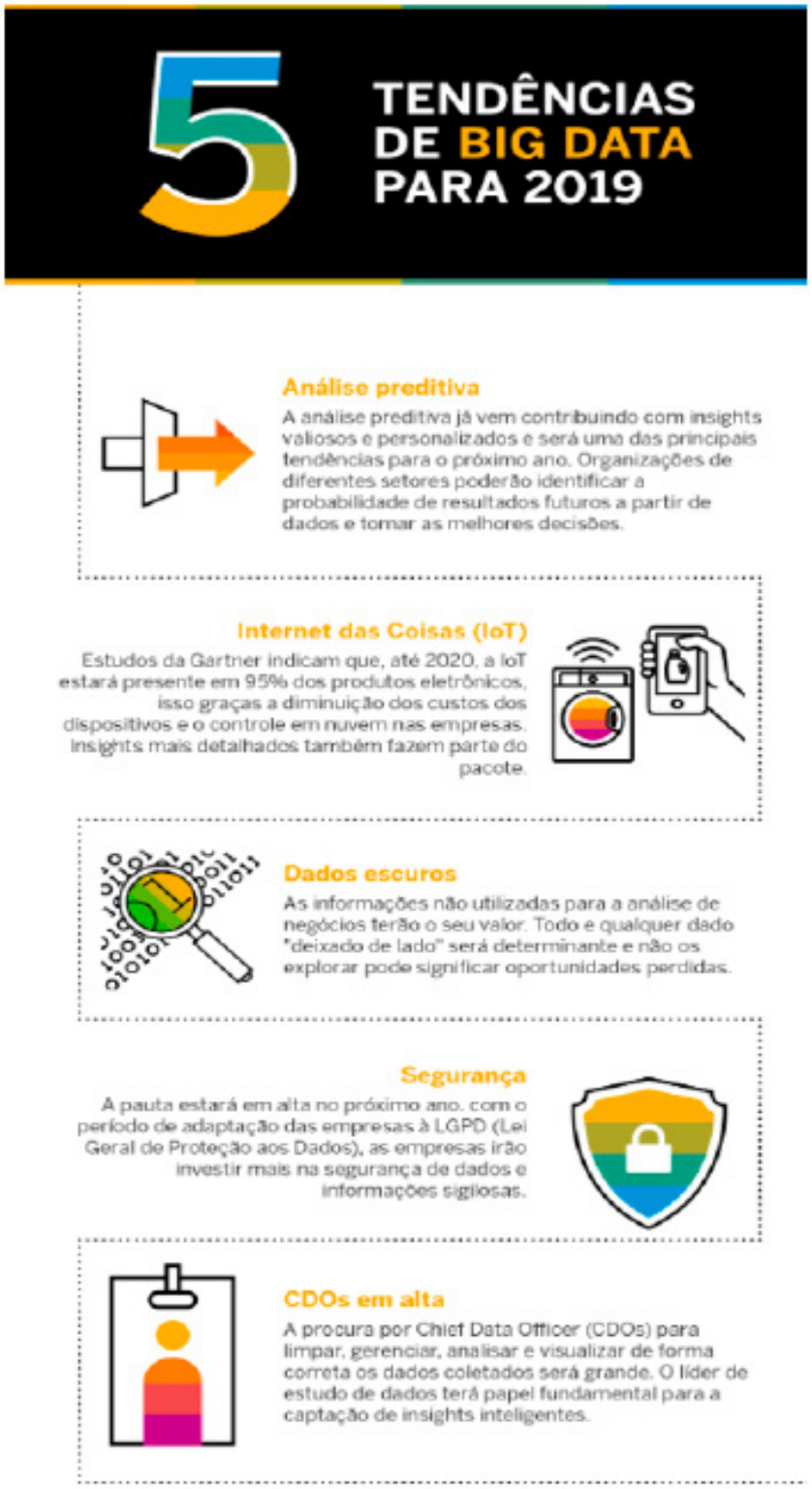


Figura. Big Data. Fonte: Quintão (2020)



Fonte: https://www.adopti.com.br/tendencias-de-big-data-para-2019/
Acesso em: dez.2019.

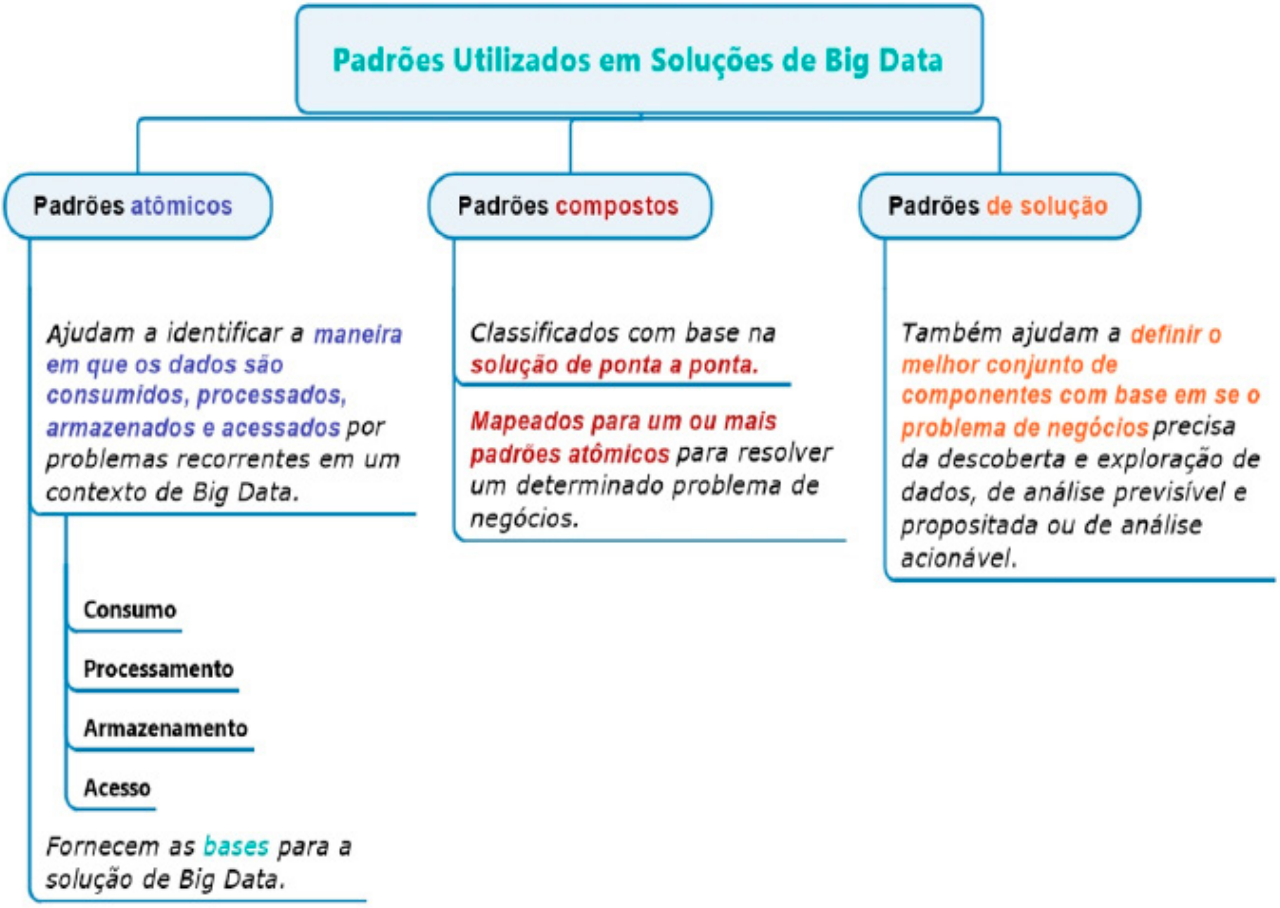


Figura. Padrões Utilizados em Soluções de Big Data. Fonte: Quintão (2020)

QUESTÕES COMENTADAS EM AULA

001. (QUADRIX/CREF 11ª REGIÃO/AGENTE DE ORIENTAÇÃO E FISCALIZAÇÃO/2014) Trata-se de uma infinidade de informações não estruturadas que, quando usadas com inteligência, se tornam uma arma poderosa para empresas tomarem decisões cada vez melhores. As soluções tecnológicas que trabalham com esse conceito permitem analisar um enorme volume de dados de forma rápida e ainda oferecem total controle ao gestor das informações. E as fontes de dados são as mais diversas possíveis: de textos e fotos em rede sociais, passando por imagens e vídeos, até jogadas específicas no esporte e até tratamentos na medicina. (http://olhardigital.uol.com.br/pro/video/39376/39376)
O conceito definido no texto é:
a) Governança de TI
b) QoS.
c) Big Data
d) Data Center.
e) ITIL.

002. (CESPE/TCE-MG/2018) Uma empresa, ao implementar técnicas e softwares de big data, deu enfoque diferenciado à análise que tem como objetivo mostrar as consequências de determinado evento. Essa análise é do tipo
a) preemptiva.
b) perceptiva.
c) prescritiva.
d) preditiva.
e) evolutiva.

003. (FCC/DPE-RS/ANALISTA/BANCO DE DADOS/2017) Os sistemas de Big Data costumam ser caracterizados pelos chamados 3 Vs, sendo que o V de
a) Veracidade corresponde à rapidez na geração e obtenção de dados.
b) Valor corresponde à grande quantidade de dados acumulada.
c) Volume corresponde à rapidez na geração e obtenção de dados.
d) Velocidade corresponde à confiança na geração e obtenção dos dados.
e) Variedade corresponde ao grande número de tipos ou formas de dados.

004. (ESAF/ANAC/ANALISTA ADMINISTRATIVO/2016) Para o processamento de grandes massas de dados, no contexto de Big Data, é muito utilizada uma plataforma de software em Java, de computação distribuída, voltada para clusters, inspirada no MapReduce e no Google-FS. Esta plataforma é o(a)

a) Yam Common.
b) GoogleCrush.
c) EMRx.
d) Hadoop.
e) MapFix.

005. (ESAF/ESAF/GESTÃO E DESENVOLVIMENTO DE SISTEMAS/2015/ADAPTADA) O Hadoop, o mais conhecido e popular sistema para gestão de Big Data, foi criado pela IBM, a partir de sua ferramenta de Data Mining WEKA.

QUESTÕES DE CONCURSO

006. (CESPE/POLÍCIA FEDERAL/PAPILOSCOPISTA/POLICIAL FEDERAL/2018) Julgue o item seguinte, a respeito de big data e tecnologias relacionadas a esse conceito.
De maneira geral, big data não se refere apenas aos dados, mas também às soluções tecnológicas criadas para lidar com dados em volume, variedade e velocidade significativos.

Big Data é um termo amplamente utilizado na atualidade para nomear **conjuntos de dados que podem ser estruturados e não estruturados** (como vídeo digital, imagens, dados de sensores, arquivos de *logs* e de qualquer tipo de dados não contidos em registros típicos com campos que podem ser pesquisados) **muito grandes ou complexos**, mas também pode se referir ao Big Data Analytics (soluções tecnológicas criadas para lidar com dados em volume, variedade e velocidade significativos).
Certo.

007. (CESPE/PF/AGENTE DA POLÍCIA FEDERAL/2018) Big data refere-se a uma nova geração de tecnologias e arquiteturas projetadas para processar volumes muito grandes e com grande variedade de dados, permitindo alta velocidade de captura, descoberta e análise.

Big Data é definido genericamente como a **captura**, **gerenciamento** *e a* **análise** de grandes e complexos conjuntos de **dados – estruturados e não estruturados,** que impactam os negócios no dia a dia.
Certo.

008. (CESPE/TCE-PB/AUDITOR DE CONTAS PÚBLICAS/DEMAIS ÁREAS/2018) Com referência a big data, assinale a opção correta.
a) A definição mais ampla de big data restringe o termo a duas partes – o volume absoluto e a velocidade –, o que facilita a extração das informações e dos insights de negócios.
b) O sistema de arquivos distribuído Hadoop implementa o algoritmo Dijkstra modificado para busca irrestrita de dados em árvores aglomeradas em clusters com criptografia.
c) Em big data, o sistema de arquivos HDFS é usado para armazenar arquivos muito grandes de forma distribuída, tendo como princípio o write-many, read-once.
d) Para armazenar e recuperar grande volume de dados, o big data utiliza bancos SQL nativos, que são bancos de dados que podem estar configurados em quatro tipos diferentes de armazenamentos: valor chave, colunar, gráfico ou documento.
e) O MapReduce é considerado um modelo de programação que permite o processamento de dados massivos em um algoritmo paralelo e distribuído.

Veja mais: https://www.devmedia.com.br/hadoop-mapreduce-introducao-a-big-data/30034

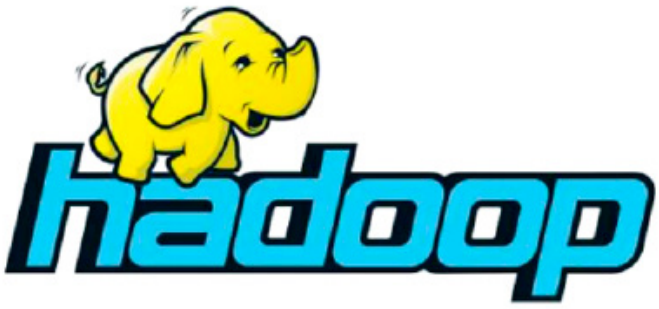
a) Errada. **Big Data** não se refere apenas aos **dados**, mas também às **soluções tecnológicas criadas para lidar com esses dados** em quantidade, variedade e velocidade bastante significativos. Para analisar a viabilidade de implementação do Big Data em uma organização, citava-se inicialmente as três dimensões (conhecidas como 3V´s), que são: **Volume, Variedade e Velocidade.**

A literatura destacou em seguida o **4 V** (incluindo a **Veracidade**); depois o **5V** (incluindo **Veracidade** e **Valor**); atualmente, a IBM cita **7 dimensões (Volume, Variedade, Velocidade, Veracidade, Valor, Governança, Pessoas)** a serem consideradas ao avaliar a viabilidade de uma solução de Big Data.

b) Errada. **Hadoop** é um sistema de armazenamento compartilhado, distribuído e altamente confiável para processamento de grandes volumes de dados através de clusters de computadores. Em outras palavras, Hadoop é um framework que facilita o funcionamento de diversos computadores, com o objetivo de analisar grandes volumes de dados. Não se pode afirmar que a busca ocorrerá de maneira irrestrita, principalmente se os dados estiverem criptografados (protegidos). A proteção pode restringir o acesso ao conteúdo dos dados.

O projeto Apache hadoop é composto de 3 módulos principais:

- Hadoop Distributed File System (HDFS);
- Hadoop Yarn;
- Hadoop MapReduce.



http://hadoop.apache.org

c) Errada. O **HDFS** é um sistema de arquivos criado para armazenar arquivos muito grandes de forma distribuída.

- O conceito sobre o qual o HDFS foi construído é o chamado ***write-once, read-many-times***, ou seja, **escreva uma vez, leia muitas vezes**.
- Esse tipo de construção é essencial para o Hadoop, uma vez que os dados serão processados inúmeras vezes, dependendo da aplicação, embora, normalmente, sejam escritos apenas uma vez.

d) Errada. O **conceito de NoSQL** é geralmente associado ao Big Data. "Bancos de dados NoSQL usam diversos modelos de dados, incluindo documentos, gráficos e chave-valor e colunares. Big Data pode utilizar bases de dados não relativas a modelos relacionais.

<Fonte: https://www.devmedia.com.br/introducao-aos-bancos-de-dados-nosql/26044>

e) Certa. **MapReduce** é um modelo de programação e framework introduzido pelo Google para suportar computações paralelas em grandes coleções de dados em clusters de computadores. Agora MapReduce é considerado um novo modelo computacional distribuído, inspirado pelas funções map e reduce usadas comumente em programação funcional.
Letra e.

---

009. (CESPE/EBSERH/ANALISTA DE TECNOLOGIA DA INFORMAÇÃO/2018) Com relação a banco de dados, julgue o item seguinte. As soluções de big data focalizam dados que já existem, descartam dados não estruturados e disponibilizam os dados estruturados.

Além de estar relacionado à grande quantidade de informações a serem analisadas**, o *Big Data* considera** o volume, a velocidade e a variedade **dos dados estruturados** – dos quais se conhece a estrutura de armazenamento – bem como dos **não estruturados**, como imagens, vídeos, áudios e documentos. Em soluções *Big Data*, a análise dos dados comumente precisa ser precedida de uma transformação de dados não estruturados em dados estruturados.
Errado.

---

010. (CESPE/TCM-BA/AUDITOR ESTADUAL DE CONTROLE EXTERNO/2018) Acerca de big data, assinale a opção correta.
a) A utilização de big data nas organizações não é capaz de transformar os seus processos de gestão e cultura.
b) Sistemas de recomendação são métodos baseados em computação distribuída, que provem uma interface para programação de clusters, a fim de recomendar os tipos certos de dados e processar grandes volumes de dados.
c) Pode-se recorrer a software conhecidos como scrapers para coletar automaticamente e visualizar dados que se encontram disponíveis em sítios de navegabilidade ruim ou em bancos de dados difíceis de manipular.
d) As ações inerentes ao processo de preparação de dados incluem detecção de anomalias, deduplicação, desambiguação de entradas e mineração de dados.
e) O termo big data se baseia em cinco Vs: velocidade, virtuosidade, volume, vantagem e valor.

a) Errada. A utilização de big data nas organizações será capaz de transformar os seus processos de gestão e cultura.

b) Errada. Um **Sistema de Recomendação combina várias técnicas computacionais para selecionar itens personalizados com base nos interesses dos usuários e conforme o contexto no qual estão inseridos**. Tais itens podem assumir formas bem variadas como, por exemplo, livros, filmes, notícias, música, vídeos, anúncios, links patrocinados, páginas de internet, produtos de uma loja virtual etc. Empresas como Amazon, Netflix e Google são reconhecidas pelo uso intensivo de sistemas de recomendação com os quais obtém grande vantagem competitiva.

c) Certa. De acordo com Wikipedia (2017), ***Data Scraping*** (ou **raspagem de dados**) é uma **técnica na qual um programa de computador extrai dados de saída legível para humanos, proveniente de um outro programa,** e disponibiliza esses dados de modo que se tornem legíveis para outros programas de computador.

Scraping é a **atividade de extrair dados de sites e transportá-los para um formato mais simples e maleável para que possam ser analisados e cruzados com mais facilidade**. Muitas vezes a informação necessária para reforçar uma história está disponível, mas em sites de navegabilidade ruim ou em bancos de dados difíceis de manipular.

Para que se possa coletar automaticamente e visualizar essas informações, recorre-se a softwares conhecidos como **scrapers** (Andriolo, 2012).

http://sinfisco.org.br/wp-content/uploads/2017/12/...

d) Errada. **Preparação de dados** é o **processo de coletar, limpar, normalizar, combinar, estruturar e organizar dados para análise**. Ele é o passo inicial (e fundamental) para que o trabalho com Big Data, uma vez que aumenta a qualidade dos dados – e, consequentemente, dos resultados com mineração de dados. Dados "pobres", de qualidade ruim, geram resultados incorretos e não confiáveis ao fim do processo.

Deduplicação é o processo de analisar, identificar e remover duplicidade nos dados, diminuindo assim a quantidade de informação a ser manipulada e armazenada.

Minerar dados consiste no uso de um conjunto de tecnologias e técnicas que permitem automatizar a busca em grandes volumes de dados por padrões e tendências que não são detectáveis por análises mais simples. Este tipo de análise dá aos gestores embasamento de alto valor para tomada de decisões estratégicas, permitindo detectar de forma precoce a ocorrência de tendências do mercado e antecipar suas ações para responder a novos cenários.

e) Errada. As **5 Dimensões (5 Vs) do Big Data** são: Volume, Variedade, Velocidade, Veracidade, Valor.

Referências:
https://www.gta.ufrj.br/grad/15_1/bigdata/vs.html
https://pt.wikipedia.org/wiki/Sistema_de_recomenda%C3%A7%C3%A3º

Letra c.

---

011. (CESPE/TCE-PE/2017) O termo Big Data Analytics refere-se aos poderosos softwares que tratam dados estruturados e não estruturados para transformá-los em informações úteis

às organizações, permitindo-lhes analisar dados, como registros de call center, postagens de redes sociais, de blogs, dados de CRM e demonstrativos de resultados.

Big Data Analytics é o trabalho analítico e inteligente de grandes volumes de dados, **estruturados ou não estruturados**, que são coletados, armazenados e interpretados por softwares de altíssimo desempenho.

Trata-se do cruzamento de uma infinidade de dados do **ambiente interno e externo**, gerando uma espécie de **"bússola gerencial"** para **tomadores de decisão**. Tudo isso, é claro, em um tempo de processamento extremamente reduzido.

Certo.

012. (CESPE/TCE-PE/AUDITOR DE CONTROLE EXTERNO/AUDITORIA DE CONTAS PÚBLICAS/2017) Com relação a Big Data, julgue o item subsequente.

Além de estar relacionado à grande quantidade de informações a serem analisadas, o *Big Data* considera o volume, a velocidade e a variedade dos dados estruturados – dos quais se conhece a estrutura de armazenamento – bem como dos não estruturados, como imagens, vídeos, áudios e documentos.

Big data é um termo que descreve o grande volume de dados – estruturados e não estruturados – que impactam as empresas diariamente.

Para analisar a viabilidade de implementação do Big Data em uma organização, citava-se inicialmente as três dimensões (conhecidas como 3Vs), que são: **Volume, Variedade e Velocidade.**

A literatura destacou em seguida o **4 V** (incluindo a **Veracidade**); depois o **5V** (incluindo **Veracidade** e **Valor**); atualmente, a IBM cita **7 dimensões (Volume, Variedade, Velocidade, Veracidade, Valor, Governança, Pessoas)** a serem consideradas ao avaliar a viabilidade de uma solução de Big Data.

Certo.

013. (CESPE/FUNPRESP-JUD/ANALISTA/TECNOLOGIA DA INFORMAÇÃO/2016) A respeito de banco de dados, julgue o próximo item. Uma big data não engloba dados não estruturados, mas inclui um imenso volume de dados estruturados suportado por tecnologias como o DataMining e o DataWarehouse para a obtenção de conhecimento a partir da manipulação desses dados.

Big Data é:

> Definido genericamente como a **captura**, **gerenciamento** e a **análise de dados que vão além dos dados tipicamente estruturados**, que podem ser consultados e pesquisados através de bancos de dados relacionais.
> *Frequentemente são* **dados obtidos de arquivos não estruturados** *como* ***vídeo digital, imagens, dados de sensores, arquivos de logs e de qualquer tipo de dados não contidos em registros típicos com campos que podem ser pesquisados.***

- **Dados estruturados**: são armazenados em bancos de dados, sequenciados em tabelas;
- **Dados semiestruturados**: acompanham padrões heterogêneos, são mais difíceis de serem identificados pois podem seguir diversos padrões;
- **Dados não estruturados**: são uma mistura de dados com fontes diversificadas como imagens, áudios e documentos online.

Fonte: https://www.gta.ufrj.br/grad/15_1/bigdata/vs.html

Errado.

014. (CESPE/TJ-SE/ANALISTA JUDICIÁRIO/BANCO DE DADOS/ADAPTADA/2014) Julgue o item que se segue, no que se refere a *Big Data*.
Em soluções *Big Data*, a análise dos dados comumente precisa ser precedida de uma transformação de dados estruturados em dados não estruturados.

Em soluções *Big Data*, a análise dos dados comumente precisa ser precedida de uma transformação de dados não estruturados em dados estruturados.
Conforme destaca http://www.ibm.com/developerworks/br/library/bd-archpatterns4/, **para *executar a análise em quaisquer dados, eles devem estar em algum tipo de formato estruturado***. Os dados não estruturados acessados de várias fontes podem ser armazenados como estão e, em seguida, transformados em dados estruturados e novamente armazenados nos sistemas de armazenamento de big data. O texto não estruturado pode ser convertido em dados estruturados ou semiestruturados. Da mesma forma, os dados de imagem, áudio e vídeo precisam ser convertidos nos formatos que podem ser usados para análise.
Errado.

015. (CESPE/TJ-SE/ANALISTA JUDICIÁRIO/BANCO DE DADOS/2014) Em soluções *Big Data*, a análise dos dados comumente precisa ser precedida de uma transformação de dados não estruturados em dados estruturados.

Conforme destaca **http://www.ibm.com/developerworks/br/library/bd-archpatterns4/**, para *executar a análise em quaisquer dados, eles devem estar em algum tipo de formato estruturado*. Os dados não estruturados acessados de várias fontes podem ser armazenados como estão e, em seguida, transformados em dados estruturados e novamente armazenados nos sistemas de armazenamento de big data. O texto não estruturado pode ser convertido em dados estruturados ou semiestruturados. Da mesma forma, os dados de imagem, áudio e vídeo precisam ser convertidos nos formatos que podem ser usados para análise.

Certo.

016. (CESPE/TJ-SE/ANALISTA JUDICIÁRIO/BANCO DE DADOS/2014) O processamento de consultas *ad hoc* em *Big Data*, devido às características de armazenamento dos dados, utiliza técnicas semelhantes àquelas empregadas em consultas do mesmo tipo em bancos de dados tradicionais.

O processamento de consultas ad hoc no Big Data traz desafios diferentes daqueles incorridos ao realizar consultas ad hoc em dados estruturados pelo fato de as fontes e formatos dos dados não serem fixos e exigirem mecanismos diferentes para recuperá-los e processá-los.
Embora as **consultas *ad hoc*** simples possam ser resolvidas pelos provedores de big data, na maioria dos casos, elas são complexas porque os dados, algoritmos, formatos e resoluções da entidade devem ser descobertos dinamicamente.

Referência: http://www.ibm.com/developerworks/br/library/bd-archpatterns4/

Errado.

017. (CESPE/TJ-SE/ANALISTA JUDICIÁRIO/BANCO DE DADOS/2014) Ao utilizar armazenamento dos dados em nuvem, a localização do processamento de aplicações *Big Data* não influenciará os custos e o tempo de resposta, uma vez que os dados são acessíveis a partir de qualquer lugar.

A localização do processamento de aplicações *Big Data* influenciará os custos e o tempo de resposta.

Errado.

018. (CESPE/TRE-GO/TÉCNICO JUDICIÁRIO/ÁREA ADMINISTRATIVA/2013) A Big Data pode ser utilizada na EAD para se entender as preferências e necessidades de aprendizagem dos alunos e, assim, contribuir para soluções mais eficientes de educação mediada por tecnologia.

Isso mesmo! Ferramentas do tipo Big Data têm permitido um conhecimento muito maior e melhor do perfil e comportamento dos alunos de EAD, fazendo com que os novos cursos sejam cada vez mais eficazes.
Certo.

019. (FCC/TCE-RS/ANÁLISE DE INFORMAÇÕES/2018) Um sistema de Big Data costuma ser caracterizado pelos chamados 3 Vs, ou seja, volume, variedade e velocidade. Por variedade entende-se que
a) há um grande número de tipos de dados suportados pelo sistema.
b) há um grande número de usuários distintos acessando o sistema.
c) os tempos de acesso ao sistema apresentam grande variação.
d) há um grande número de tipos de máquinas acessando o sistema.
e) os tamanhos das tabelas que compõem o sistema são muito variáveis.

O **objetivo do *Big Data*** é propiciar dados e informações que possam ser analisados visando subsidiar tomadas de decisão.
A tomada de decisão é possível em função não somente do **volume** de dados, da **velocidade** de captura dessas informações, das **fontes variadas de informações** e de **novos softwares para fins de modelagem dessas informações**.
Big Data, normalmente, é dividido em três dimensões (3 Vs):

Volume	O **volume** da informação refere-se à grande quantidade de dados acumulado. Certas coleções de dados atingem a faixa de gigabytes (bilhões de bytes), terabytes (trilhões), **petabytes** (milhares de trilhões) ou mesmo **exabytes** (milhões de trilhões).
Variedade	Significa que **os dados de hoje aparecem em todos os tipos de formatos**, como, por exemplo, arquivos de texto, ***e-mail***, medidores e sensores de coleta de dados, vídeo, áudio, dados de ações do mercado ou transações financeiras. Por variedade entende-se que **há um grande número de tipos de dados suportados pelo sistema.**

Velocidade	Está relacionada à **rapidez** com a qual os dados são produzidos e tratados para atender à demanda, o que significa que não é possível armazená-los por completo, de modo que somos obrigados a escolher dados para guardar e outros para descartar. A tecnologia de Big Data agora nos permite analisar os dados **no momento em que estes são gerados**, sem a necessidade de inseri-los nos bancos de dados.

Figura. 3 Vs do Big Data

Letra a.

020. (INÉDITA/2020) Data Mart é um termo utilizado para descrever grandes e complexos conjuntos de dados que são muito difíceis de capturar, processar, armazenar, buscar e analisar com os sistemas de base de dados convencionais.

Esse é o conceito de **Big Data**!
Siewert (2013) destaca que o termo ***Big Data*** é:

> Definido genericamente como a captura, gerenciamento e a análise de dados que vão além dos dados tipicamente estruturados, que podem ser consultados e pesquisados através de bancos de dados relacionais. Frequentemente são dados obtidos de arquivos não estruturados como vídeo digital, imagens, dados de sensores, arquivos de logs e de qualquer tipo de dados não contidos em registros típicos com campos que podem ser pesquisados".

De acordo com Landim (2015), trata-se de um termo usado para descrever **grandes e complexos conjuntos de dados que são muito difíceis de capturar, processar, armazenar, buscar e analisar com os sistemas de base de dados convencionais**.

As 5 dimensões (5 Vs) do Big Data: volume, velocidade, variedade, valor, veracidade.

Uma solução de big data possui camadas horizontais e verticais (IBM, 2017):

- As **camadas horizontais**, de "baixo" para "cima" são: Fontes de Big Data, Camada de Tratamento e Armazenamento, Camada de Análise e Camada de Consumo.
- As **camadas verticais** são: Integração de informações, Governança de big data, Gerenciamento de sistemas e Qualidade de serviço.

Errado.

---

021. (INÉDITA/2020) Para analisar a viabilidade de implementação do Big Data em uma organização, a literatura citava inicialmente três dimensões (conhecidas como 3V´s), que são: Volume, Variedade e Veracidade.

Para analisar a viabilidade de implementação do Big Data em uma organização, citava-se inicialmente as três dimensões (conhecidas como 3V´s), que são: **Volume, Variedade e Velocidade.** A literatura destacou em seguida o **4 V** (incluindo a **Veracidade**); depois o **5V** (incluindo **Veracidade** e **Valor**); atualmente, a IBM cita **7 dimensões (Volume, Variedade, Velocidade, Veracidade, Valor, Governança, Pessoas)** a serem consideradas ao avaliar a viabilidade de uma solução de Big Data.

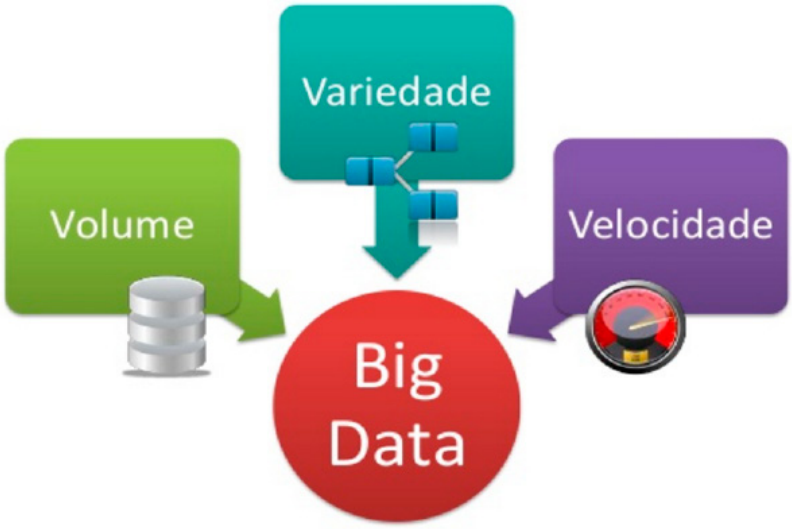


Figura. 3 dimensões (3 Vs) do Big Data

Figura. 4 dimensões (4 Vs) do Big Data

Errado.

022. (ESAF/ANAC/ANALISTA ADMINISTRATIVO/2016) Big Data é:

a) volume + variedade + agilidade + efetividade, tudo agregando + valor + atualidade.

b) volume + oportunidade + segurança + veracidade, tudo agregando + valor.

c) dimensão + variedade + otimização + veracidade, tudo agregando + agilidade.

d) volume + variedade + velocidade + veracidade, tudo agregando + valor.

e) volume + disponibilidade + velocidade + portabilidade, tudo requerendo - valor.

As 5 Dimensões (5 Vs) do Big Data são: Volume, Variedade, Velocidade, Veracidade, Valor.

Vamos à descrição dessas **cinco dimensões – 5Vs – do Big Data**, que são de grande importância para a prova.

Volume:

O **volume** da informação se refere ao fato de que certas coleções de dados atingem a faixa de gigabytes (bilhões de bytes), terabytes (trilhões), petabytes (milhares de trilhões) ou mesmo exabytes (milhões de trilhões).

Variedade:

A **variedade** significa que os dados de hoje aparecem em todos os tipos de formatos, como, por exemplo, arquivos de texto, e-mail, medidores e sensores de coleta de dados, vídeo, áudio, dados de ações do mercado ou transações financeiras.

Velocidade:

A **velocidade** está relacionada à **rapidez** com a qual os dados são produzidos e tratados para atender à demanda, o que significa que não é possível armazená-los por completo, de modo que somos obrigados a escolher dados para guardar e outros para descartar. A tecnologia de Big Data agora nos permite analisar os dados **no momento em que estes são gerados**, sem a necessidade de inseri-los nos bancos de dados.

Veracidade:

Quanto à **veracidade**, Weber et. al. (2009) ressaltou que as informações verdadeiras podem ser usadas pelos gestores para responder aos desafios estratégicos. A veracidade garantiria, então, a confiabilidade dos dados.

Valor:

Com relação ao **valor**, Chen et. al. (2014) afirmam que as análises críticas de dados podem ajudar as empresas a melhor entender seus negócios trazendo benefícios.

A combinação "volume + velocidade + variedade + veracidade", além de todo e qualquer outro aspecto que caracteriza uma solução de Big Data, se mostrará inviável se o resultado não trouxer benefícios significativos e que compensem o investimento. Este é o ponto de vista do **valor** (*value*), conforme destaca http://www.infowester.com/big-data.php.

Letra d.

---

023. (INÉDITA/2020) Julgue o item que se segue, no que se refere a *Big Data*.

Os sistemas de armazenamento de dados tradicionais são adequados para o big data.

O armazenamento de dados tradicional não é a melhor opção para armazenar big data, mas nos casos em que as empresas estão realizando a exploração de dados inicial, elas podem optar por usar o *Data Warehouse*, o sistema RDBMS (sistemas relacionais) e outros armazenamentos de conteúdo existentes. Esses sistemas de armazenamento existentes podem ser usados para armazenar os dados que são compilados e filtrados usando a plataforma de big data. **NÃO** considere os sistemas de armazenamento de dados tradicionais como adequados para o *Big Data*.

Referência: http://www.ibm.com/developerworks/br/library/bd-archpatterns4/

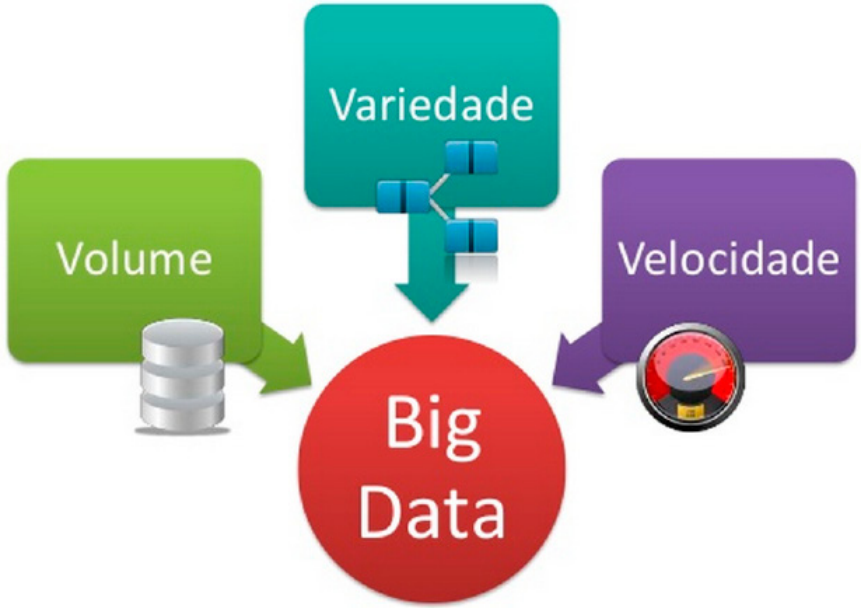
Errado.

---

024. (FGV/TJ-SC/ANALISTA DE SISTEMAS/2015) Os termos Business Intelligence (BI) e Big Data confundem-se em certos aspectos. Uma conhecida abordagem para identificação dos pontos críticos de cada paradigma é conhecida como 3V, e destaca:

a) variedade, visualização, volume;
b) velocidade, virtualização, volume;
c) variedade, velocidade, volume;
d) virtualização, visualização, volume;
e) variedade, visualização, virtualização.

A abordagem **3V** destaca o **V**olume, a **V**ariedade e a **V**elocidade.



Fonte: http://pt.slideshare.net/RioInfo2009/big-data-tendncias-e-oportunidades-palestrante-srgio-mafra

São eles:

- O **volume** da informação se refere ao fato de que certas coleções de dados atingem a faixa de gigabytes (bilhões de bytes), terabytes (trilhões), petabytes (milhares de trilhões) ou mesmo exabytes (milhões de trilhões).
- A **velocidade** está relacionada à rapidez com a qual os dados são produzidos e tratados para atender à demanda, o que significa que não é possível armazená-los todos, de modo que somos obrigados a escolher dados para guardar e outros para descartar.
- A **variedade** significa que os dados de hoje aparecem em todos os tipos de formatos, como, por exemplo, arquivos de texto, email, medidores e sensores de coleta de dados, vídeo, áudio, dados de ações do mercado ou transações financeiras.

Letra c.

025. (CESPE/TCU/AUDITOR FEDERAL DE CONTROLE EXTERNO/CONHECIMENTOS GERAIS/2015) No que concerne a data mining (mineração de dados) e big data, julgue o seguinte item.

Devido à quantidade de informações manipuladas, a (cloud computing) computação em nuvem torna-se inviável para soluções de big data.

Para processar grandes volumes de dados em tempo real, empresas deverão usar a infraestrutura de Cloud Computing para colocar projetos de Big Data em ação, é o que destaca https://cloud21.com.br/computacao-em-nuvem/cloud-computing-e-o-motor-do-big-data/.

A Cloud Computing (Computação em Nuvem) é a infraestrutura que vai suportar as iniciativas pela sua capacidade para processar grandes volumes de dados em tempo real, requisito do Big Data.

Stefanini (em https://stefanini.com/br/2015/01/relacao-entre-big-data-cloud-computing/) também destaca que *Big Data* e *Cloud Computing* são praticamente indissociáveis quando o assunto é gerar vantagens competitivas para uma organização a partir das informações que ela possui disponíveis, seja internamente ou no mercado. Segundo o autor, a grande vantagem de associar Big Data à Cloud Computing é reduzir os custos de uma infraestrutura de TI para armazenar e processar os dados. Empresas como Amazon fornecem serviços para que você possa estruturar toda a sua capacidade de BI fora da sua empresa.

Errado.

026. (FGV/AL-BA/TÉCNICO DE NÍVEL SUPERIOR/ECONOMIA/2014) A expressão *Big Data* é utilizada para descrever o contexto da informação contemporânea, caracterizada pelo volume, velocidade e variedade de dados disponíveis, em escala inédita. Com relação às características do Big Data, analise as afirmativas a seguir.

I – O volume da informação se refere ao fato de que certas coleções de dados atingem a faixa de gigabytes (bilhões de bytes), terabytes (trilhões), petabytes (milhares de trilhões) ou mesmo exabytes (milhões de trilhões).

II – A velocidade está relacionada à rapidez com a qual os dados são produzidos e tratados para atender à demanda, o que significa que não é possível armazená-los todos, de modo que somos obrigados a escolher dados para guardar e outros para descartar.

III – A variedade significa que os dados de hoje aparecem em todos os tipos de formatos, como, por exemplo, arquivos de texto, email, medidores e sensores de coleta de dados, vídeo, áudio, dados de ações do mercado ou transações financeiras.

Assinale:

a) se somente a afirmativa I estiver correta.
b) se somente a afirmativa II estiver correta.
c) se somente a afirmativa III estiver correta.
d) se somente as afirmativas I e II estiverem corretas.
e) se todas as afirmativas estiverem corretas.

Todas as afirmativas estão corretas!

- ***Volume*** **–** refere-se à quantidade de dados a ser capturada, armazenada e manipulada. Estamos falando de petabytes ou terabytes de dados, tendendo a aumentar!
- ***Velocidade*** refere-se à velocidade de produção dos novos dados, a velocidade em que é preciso agir com relação a eles ou a taxa em que esses dados estão mudando. A depender da velocidade, pode ser necessário escolher dados para guardar e outros para descartar.
- ***Variedade*** significa que os dados de hoje aparecem em todos os tipos de formatos, envolvendo por exemplo dados de redes sociais, vídeos, áudios etc. Já não é mais possível antecipar o conteúdo e a estrutura dos mesmos!

Letra e.

027. (INÉDITA/2020) Apache Hadoop é um software open source para armazenamento e processamento em larga escala de grandes conjuntos de dados (Big Data), em clusters de hardware de baixo custo.

Hadoop é um sistema de armazenamento compartilhado, distribuído e altamente confiável para processamento de grandes volumes de dados através de clusters de computadores. Em outras palavras, **Hadoop é um framework que facilita o funcionamento de diversos computadores, com o objetivo de analisar grandes volumes de dados**.

O projeto Apache hadoop é composto de 3 módulos principais:

- Hadoop Distributed File System (HDFS);
- Hadoop Yarn;
- Hadoop MapReduce.

http://hadoop.apache.org

Certo.

GABARITO

1. c
2. c
3. e
4. d
5. E
6. C
7. C
8. e
9. E
10. c
11. C
12. C
13. E
14. E
15. C
16. E
17. E
18. C
19. a
20. E
21. E
22. d
23. E
24. c
25. E
26. e
27. C

REFERÊNCIAS

ALECRIM, E. **O que é big data?** 2013. Disponível em: <http://www.infowester.com/big-data.php>. Acesso em: 05 jul. 2020.

BIG DATA BUSINESS. **Big Data Analytics: você sabe o que é?** Disponível em: <http://www.bigdatabusiness.com.br/voce-sabe-o-que-e-big-data-analytics/> Acesso em: 10 mar. 2019.

_____. **Tipos de análise de Big Data: você conhece todos os 4?** Disponível em: <http://www.bigdatabusiness.com.br/conheca-os-4-tipos-de-analises-de-big-data-analytics/>. Acesso em: 10 mar. 2019.

BRITO, S. H. B. Afinal, O Que é Big Data? 2013. Disponível em: <http://labcisco.blogspot.com.br/2013/08/afinal-o-que-e-big-data.html>.

FERNANDES, A. A.; DE ABREU, V. F. **Implantando a Governança de TI: Da estratégia à Gestão de Processos e Serviços.** Brasport, 2014.

GARTNER IT GLOSSARY. Disponível em: <https://www.gartner.com/en/information-technology/glossary/big-data> Acesso em: 15 nov. 2012

GEORGE G., HAAS, M. & PENTLAND A., **Big Data and Management.** Academy of Management Journal, 2014, Vol. 57, No. 2, 321–326. Disponível em: http://dx.doi.org/10.5465/amj.2014.4002 Acesso em: 25 abr. 2014.

GOLDMAN, Alfredo et al. **Apache Hadoop**: conceitos teóricos e práticos, evolução e novas possibilidades. XXXI Jornadas de atualizaçoes em informatica, p. 88-136, 2012.

HANSON, J. **Uma Introdução ao Hadoop Distributed File System**. Disponível em: <https://www.ibm.com/developerworks/br/library/wa-introhdfs/index.html> Acesso em: 19 mar. 2018.

IBM. **Como saber se uma solução de big data é ideal para sua organização**. Disponível em: < https://www.ibm.com/developerworks/br/library/bd-archpatterns2/index.html> Acesso em: 25 dez. 2017.

MAÇADA, A. C. G.; Vivian Passos Canary. **A Tomada de decisão no contexto do Big Data: Estudo de caso único**. 2014. Disponível em: <http://www.anpad.org.br/admin/pdf/2014_EnANPAD_ADI1088.pdf>.

MCAFEE, A.; BRYNJOLFSSON, E. **Big Data: The Management Revolution**. Harvard Business Review, October, 2012. p. 1-9.

MACHADO, Henrique. **Hadoop MapReduce: Introdução a Big Data.** Disponível em: <https://www.devmedia.com.br/hadoop-mapreduce-introducao-a-big-data/30034>. Acesso em: 25 abr. 2018.

MAFRA, S. **Big data: tendências e oportunidades**. 2013. Disponível em: <http://pt.slideshare.net/RioInfo2009/big-data-tendncias-e-oportunidades-palestrante-srgio-mafra>. Acesso em: 05 jul. 2020.

MARQUESONE, R. **O novo desafio das empresas e profissionais do mercado.** <http://paineira.usp.br/lassu/wp-content/uploads/2017/01/2017.02.07-palestra_rosangela_bigdata.pdf>. Acesso em: 25 ago. 2020.

MYSORE, D., KHUPAT, S., JAIN, S. **Entendendo as camadas de arquitetura de uma solução de big data.** 2014. Disponível em: <http://www.ibm.com/developerworks/br/library/bd-archpatterns3/>.Acesso em: 10 jul. 2020.

MYSORE, D.; Khupat, S.; JAIN, S. **Entendendo padrões atômicos e compostos de soluções de big data.** 2014. Disponível em: <http://www.ibm.com/developerworks/br/library/bd-archpatterns4/>.Acesso em: 10 jul. 2020.

INTEL CORPORATION. **Guia de Planejamento. Saiba mais sobre Big Data.** 2013. Disponível em: <https://www.intel.com.br/content/dam/www/public/lar/br/pt/documents/articles/90318386-1-por.pdf>. Acesso em: 25 ago. 2020.

SANTANA, R. **Coleta e Análise de Dados Matérias-primas de Big Data Analytics**. 2018. Disponível em: <http://rubenssantana.com/coleta-e-analise-de-dados/>. Acesso em: 20 ago. 2020.

SIEWERT, Sam B. **Big data in the cloud: data velocity, volume, variety veracity**. IBM developersWorks. July 2013.

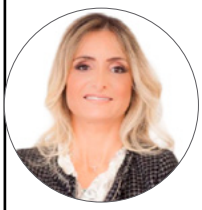
TAURION, C. **Big Data**. São Paulo: Brasport, 2013.

TURBAM, E. et al**. Business Intelligence: um Enfoque Gerencial para a Inteligência do Negócio**. Bookman, 2009.

WEBER, K.et. al.. 2009. **One size does not fit all–a contingency approach to data governance.** Journal of Data and Information Quality, Volume 1, Issue 1, Article 4, June 2009, 27 p.

WIKERSON, L. **De que maneira o Big Data melhora nossa vida diária?** 2015. Disponível em: <http://www.tecmundo.com.br/tecnologia-da-informacao/80027-maneira-big-data-melhora-nossa-vida-diaria-infografico.htm>. Acesso em: 04 jul. 2020.

VORHIES, W. ***Prescriptive versus predictive analytics - a distinction without a difference?*** 2014. Disponível em: <https://www.datasciencecentral.com/profiles/blogs/prescriptive-versus-predictive-analytics-a-distinction-without-a>. Acesso em: 20 ago. 2020.



Patrícia Quintão

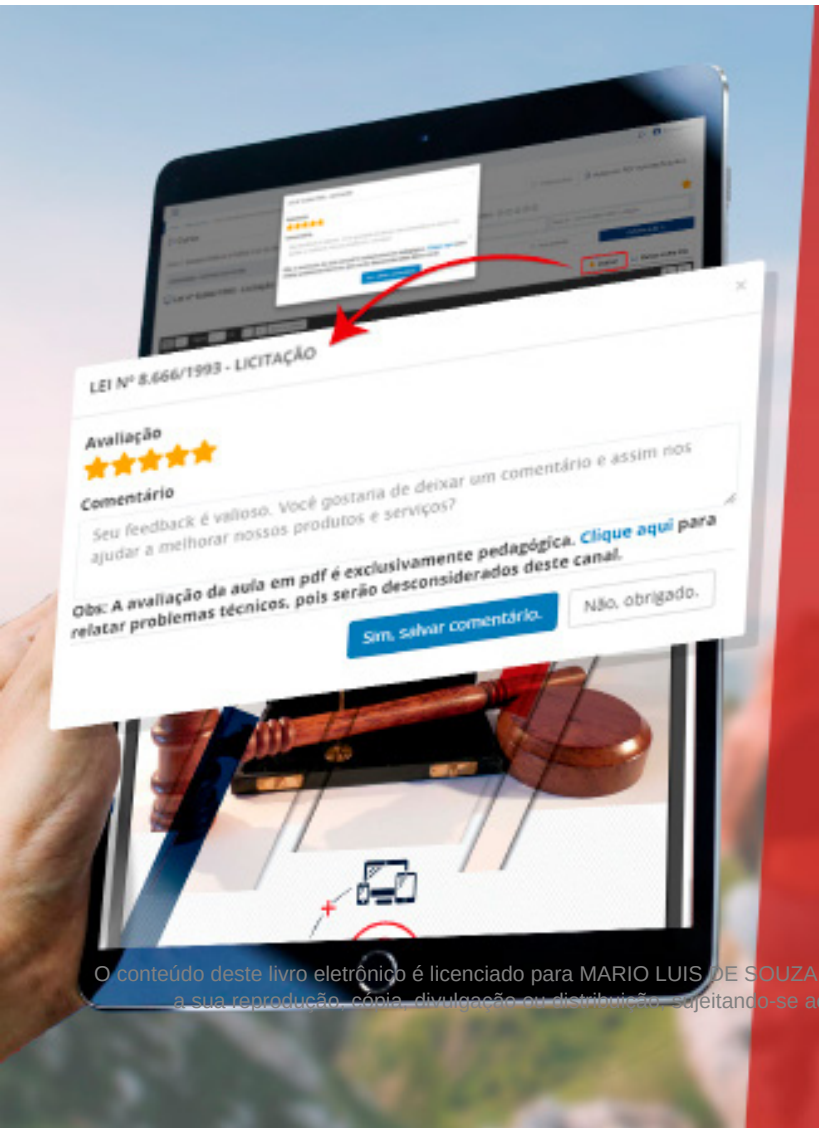
Mestre em Engenharia de Sistemas e computação pela COPPE/UFRJ, Especialista em Gerência de Informática e Bacharel em Informática pela UFV. Atualmente é professora no Gran Cursos Online; Analista Legislativo (Área de Governança de TI), na Assembleia Legislativa de MG; Escritora e Personal & Professional Coach.

Atua como professora de Cursinhos e Faculdades, na área de Tecnologia da Informação, desde 2008. É membro: da Sociedade Brasileira de Coaching, do PMI, da ISACA, da Comissão de Estudo de Técnicas de Segurança (CE-21:027.00) da ABNT, responsável pela elaboração das normas brasileiras sobre gestão da Segurança da Informação.

Autora dos livros: Informática FCC - Questões comentadas e organizadas por assunto, 3ª. edição e 1001 questões comentadas de informática (Cespe/UnB), 2ª. edição, pela Editora Gen/Método.

Foi aprovada nos seguintes concursos: Analista Legislativo, na especialidade de Administração de Rede, na Assembleia Legislativa do Estado de MG; Professora titular do Departamento de Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia; Professora substituta do DCC da UFJF; Analista de TI/Suporte, PRODABEL; Analista do Ministério Público MG; Analista de Sistemas, DATAPREV, Segurança da Informação; Analista de Sistemas, INFRAERO; Analista - TIC, PRODEMGE; Analista de Sistemas, Prefeitura de Juiz de Fora; Analista de Sistemas, SERPRO; Analista Judiciário (Informática), TRF 2ª Região RJ/ES, etc.

@coachpatriciaquintao
/profapatriciaquintao
@plquintao
t.me/coachpatriciaquintao



NÃO SE ESQUEÇA DE AVALIAR ESTA AULA!

SUA OPINIÃO É MUITO IMPORTANTE PARA MELHORARMOS AINDA MAIS NOSSOS MATERIAIS.

ESPERAMOS QUE TENHA GOSTADO DESTA AULA!

PARA AVALIAR, BASTA CLICAR EM LER A AULA E, DEPOIS, EM AVALIAR AULA.

AVALIAR