Estratégia Concursos

Aula 12 - Prof. Diego Carvalho e Raphael Lacerda.

PRF (Policial) Informática - 2023

(Pré-Edital)

Autor:

Diego Carvalho, Renato da Costa, Equipe Informática e TI

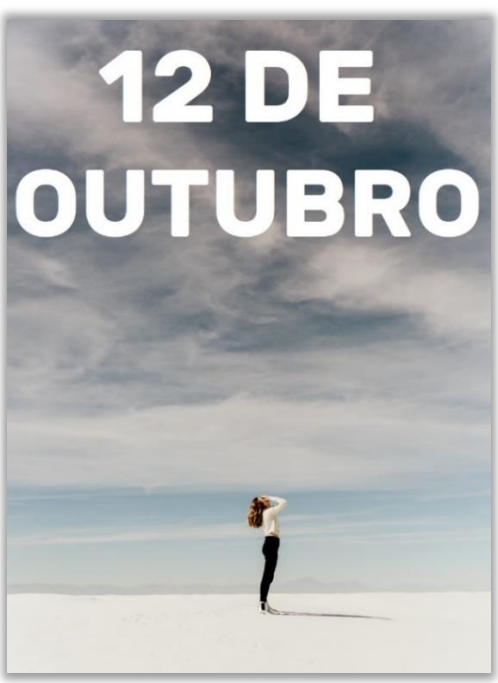
11 de Março de 2023

Índice

APRESENTAÇÃO DA AULA

Fala, galera! O assunto da nossa aula de hoje é **Big Data**! Sim... as bancas começaram a cobrar esse tema recentemente de forma até bastante frequente. Nós vamos estudar seu conceito, suas premissas e algumas particularidades. Galera, considerando tudo que nós já estudamos anteriormente, esse é um tema bem tranquilo e pequeno. Além disso, essa aula possui quase todos os exercícios que já caíram sobre o tema, então tá sussa...

PROFESSOR DIEGO CARVALHO - WWW.INSTAGRAM.COM/PROFESSORDIEGOCARVALHO



Galera, todos os tópicos da aula possuem Faixas de Incidência, que indicam se o assunto cai muito ou pouco em prova. *Diego, se cai pouco para que colocar em aula?* Cair pouco não significa que não cairá justamente na sua prova! A ideia aqui é: se você está com pouco tempo e precisa ver somente aquilo que cai mais, você pode filtrar pelas incidências média, alta e altíssima; se você tem tempo sobrando e quer ver tudo, vejam também as incidências baixas e baixíssimas. *Fechado?*

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

INCIDÊNCIA EM PROVA: BAIXA

INCIDÊNCIA EM PROVA: MÉDIA

INCIDÊNCIA EM PROVA: ALTA

INCIDÊNCIA EM PROVA: ALTÍSSIMA

Além disso, essas faixas não são por banca – é baseado tanto na quantidade de vezes que caiu em prova independentemente da banca e também em minhas avaliações sobre cada assunto...

#ATENÇÃO

Avisos Importantes

O curso abrange todos os níveis de conhecimento...

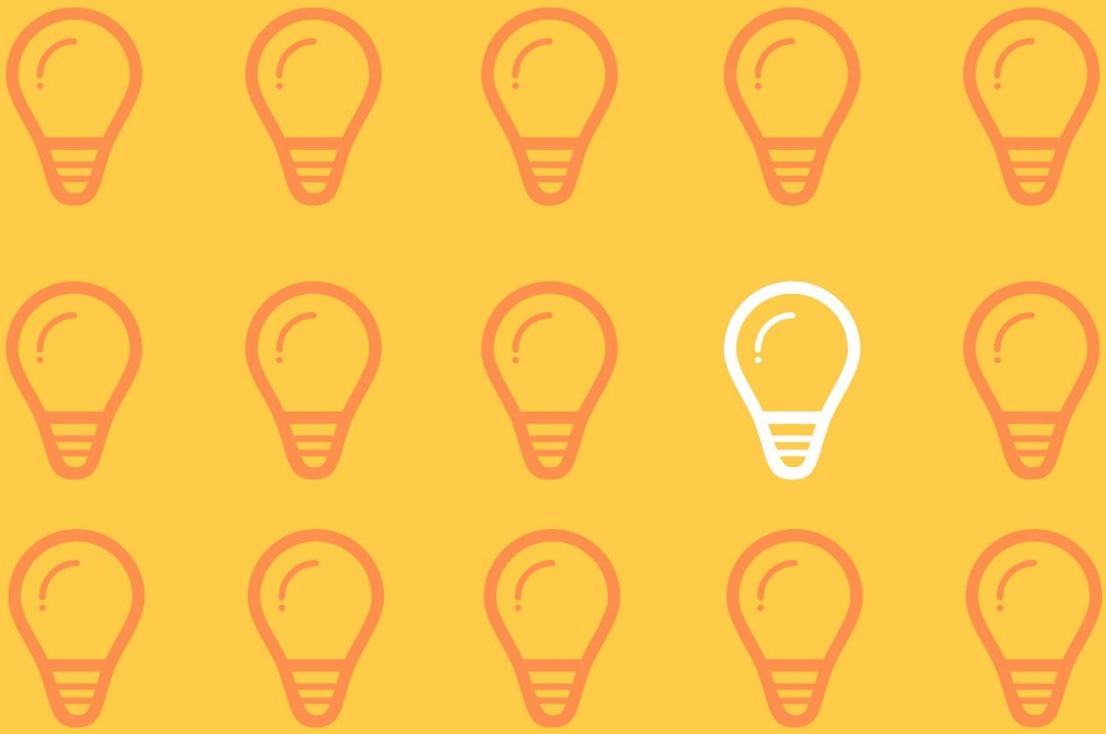
Esse curso foi desenvolvido para ser acessível a **alunos com diversos níveis de conhecimento diferentes.** Temos alunos mais avançados que têm conhecimento prévio ou têm facilidade com o assunto. Por outro lado, temos alunos iniciantes, que nunca tiveram contato com a matéria ou até mesmo que têm trauma dessa disciplina. A ideia aqui é tentar atingir ambos os públicos - iniciantes e avançados - da melhor maneira possível..

Por que estou enfatizando isso?

O **material completo** é composto de muitas histórias, exemplos, metáforas, piadas, memes, questões, desafios, esquemas, diagramas, imagens, entre outros. Já o **material simplificado** possui exatamente o mesmo núcleo do material completo, mas ele é menor e bem mais objetivo. *Professor, eu devo estudar por qual material?* Se você quiser se aprofundar nos assuntos ou tem dificuldade com a matéria, necessitando de um material mais passo-a-passo, utilize o material completo. Se você não quer se aprofundar nos assuntos ou tem facilidade com a matéria, necessitando de um material mais direto ao ponto, utilize o material simplificado.

Por fim...

O curso contém diversas questões espalhadas em meio à teoria. Essas questões possuem um comentário mais simplificado porque **têm o único objetivo de apresentar ao aluno como bancas de concurso cobram o assunto previamente administrado**. A imensa maioria das questões para que o aluno avalie seus conhecimentos sobre a matéria estão dispostas ao final da aula na lista de exercícios e **possuem comentários bem mais completos, abrangentes e direcionados.**



• ATENÇÃO •

Existem muitos exercícios sobre esse tema em sites de questões, no entanto a imensa maioria foi aplicada em provas para cargos específicos de Tecnologia da Informação (TI), os quais podem demandar um conhecimento muito mais aprofundado da matéria.

Dessa forma, recomendo que vocês tenham muita atenção na seleção das questões realizadas para que não extrapolem o nível cobrado na sua prova.

Qualquer dúvida, estou à disposição para maiores esclarecimentos!

Big Data

Conceitos Básicos

INCIDÊNCIA EM PROVA: ALTA

Sabe aquele vídeo insuportável com um anúncio de visualizar um vídeo no Youtube? Ou quando você está navegando pelo Facebook e aparece uma propaganda nos eu feed? **Esses anúncios são ótimos exemplos de como é utilizado o Big Data**. Por diversas vezes, ele é escolhido especificamente para você com base nos sites que você frequenta, sua idade aproximada, seu sexo, onde você mora, além de um monte de outras variáveis.

Deixa eu contar uma historinha para vocês: *vocês sabem qual é a maior e melhor banda de rock de todos os tempos?* **Ora... para o bem da nossa relação, eu espero que vocês tenham respondido Pink Floyd! Acertei???** Ela é a minha banda favorita, eu já ouvi todas as músicas, já li todos os livros, possuo todos os discos e... tenho várias camisetas! Certo dia, estava eu fuçando em meu Instagram quando apareceu o seguinte anúncio:

Galera, essa camisa diz: "*Nunca subestime um fã de Pink Floyd que tenha nascido em outubro*". Eu pensei ingenuamente: *"Meu pai do céu, não é possível que eu esteja com tanta sorte hoje!"* Apareceu justamente no meu feed do Instagram uma camiseta à venda da minha banda favorita falando de pessoas que nasceram em outubro e...

vocês não vão acreditar, mas...

sabe
em
qual
mês
eu
nasci?

EM OUTUBRO!!!

Eu sou retardado, não é? É claro que não havia coincidência alguma! O Instagram sabe meus dados pessoais e conhece todos os meus interesses. Dessa forma, ele consegue direcionar melhor os anúncios. Fim da história: eu quaaaase comprei a camiseta e depois passei dias me achando um completo trouxa por pensar que era coincidência. *Mané, né?* **Prosseguindo... os dados são parte de um conjunto gigantesco de dados sobre você e outras pessoas.**

Quase todas as vezes você clica (ou não clica) em um anúncio, dados são armazenados em algum lugar. Toda vez que você assiste a um vídeo do Youtube – como as aulas do Estratégia Concursos – são mantidos registros. Existem registros de todos os cliques de todas as pessoas que já acessaram o Twitter, todos os *likes* e comentários de todas as fotos do Instagram, todas as compras que você fez com seu cartão de crédito, todo filme assistido no Netflix e quanto tempo!

Com 7,5 bilhões de pessoas no planeta, muitos (mas muitos meeeeeesmo) dados são criados a cada segundo. **Basicamente, apenas de existir, você já estará criando dados – é tanto dado, mas tanto dado que nós chamamos isso de Big Data**. Galera, antes do surgimento de smartphones, notebooks e computadores, era muito trabalhoso e demorado registrar medições e armazenar dados. Aliás, nem existia uma preocupação de se armazenar dados sobre essas coisas.

Só existem dados climatológicos sobre a cidade de São Paulo a partir de 1961. Antes disso, não havia nenhum registro oficial. Dados sobre o censo dos Estados Unidos – que ocorre a cada dez anos – frequentemente demoravam justamente dez anos para ficar pronto. Dessa forma, se ele começasse a ser medido em 1950, ele demoraria dez anos para terminar. **Assim, as pessoas só descobririam qual era o tamanho da população de 1950 em 1960[1]!**

O termo Big Data – na forma como o utilizamos hoje – surgiu na década de 1990! O autor, John Mashey, usou o termo para descrever dados que são tão grandes e complexos que ferramentas para trabalhar e interpretar dados simplesmente não davam conta do recado. **Galera, seu telefone registra a sua localização, registra os aplicativos que você usa e registra quanto tempo você os usa, então todos os aplicativos que você usa coletam dados sobre você**.

Eu vou me casar em breve e já estou olhando alguns eletrodomésticos! Por conta disso, eu não paro de receber ofertas de geladeira, fogão e televisores em meu e-mail. Galera, a sociedade está criando um mundo interconectado – às vezes chamado de Internet of Things (IoT) ou Internet das Coisas. **Considerem a rede de dispositivos inteligentes que coletam dados e podem potencialmente se comunicar entre si, desde sua geladeira até seu carro, relógio, luzes, etc**.

Vocês acreditam que já há cientistas que equipam algumas mudas de espinafre para poderem enviar e-mails via wireless em determinados eventos? O grande lance do Big Data é que há muita coisa para se discutir ainda, então vamos analisar um pequeno aspecto dele: **Likes no Facebook**! Por anos, esses likes pareciam bem inúteis. Ninguém entra nas redes sociais para ver no que o amigo tem dado like. *Vocês concordam?*

No entanto, essas informações são mais reveladoras do que você imagina! Em 2013, a PNAS (*Proceedings of National Academy of Sciences*) publicou um estudo da Universidade de Cambridge em que 58.000 usuários do Facebook participaram de uma pesquisa de personalidade em um

[1] Computadores ajudaram a reduzir o tempo necessário para coletar, resumir e armazenar dados. No entanto, quanto mais aumenta o poder dos computadores de coletar e analisar dados, mais aumentam também... os próprios dados!

aplicativo. Em seguida, eles pediram permissão para ver os likes dos usuários. **Eles descobriram que traços e atributos individuais podem ser previstos com um alto grau de precisão!**

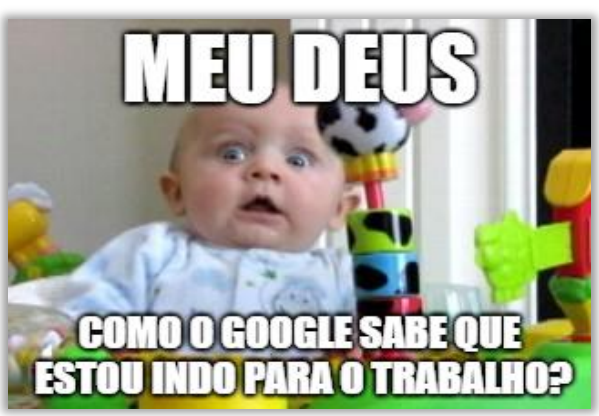
Como, professor? Com base apenas em registros de *likes*! Olha que coisa interessante: *likes* em fotos de raios e tempestades ou em postagens sobre ciência apontam para pessoas altamente inteligentes; *likes* em postagens sobre rap e Lionel Messi apontam para homens heterossexuais. **Esta é uma minúscula peça do quebra-cabeça que pode lhe dar uma pequena noção sobre o que é, na prática, o Big Data**.

Se um pouquinho de informação sobre uma pessoa pode revelar muito, podemos imaginar o que toneladas e toneladas de outros dados produzidos diariamente a cada dia podem fornecer. *Galera, vocês compreendem o poder disso?* Durante as últimas eleições presidenciais americanas, a campanha de Donald Trump escolheu grupos particulares de apoiadores de Hillary Clinton para ver anúncios contra ela em mídias sociais, tentando torná-los menos propensos a votar nela.

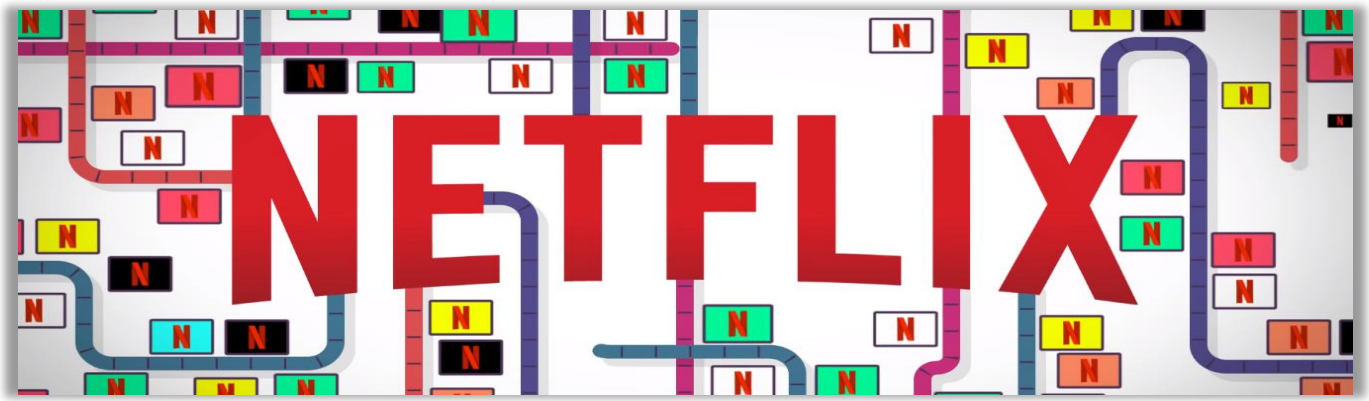
Por outro lado, há uma boa chance de o Big Data impactar positivamente a sua vida. Ele pode ser usado para criar um remédio personalizado, para prever quais jogadores de futebol uma equipe deve contratar, e para criar carros sem motorista. *Sabe quando você está perdido e precisa utilizar o Google Maps?* Pois é, você está consumindo e servido ao Big Data! Se você habilitar sua localização, informações sobre local e velocidade são continuamente transferidas ao Google.

Essa informação por si só não é útil para alguém, mas inúmeras pessoas ao seu redor também estão usando o Google Maps. Então, o Google possui uma tonelada de dados sobre onde as pessoas estão e quão rápido elas estão se movendo. Como eles vêm trabalhando com esses dados a algum tempo, eles conseguem prever como estará o trânsito de uma cidade com base em coisas como: dia da semana, horários, feriados, entre outros dados.

Com essa quantidade massiva de dados, eles conseguem te dizer se há muito trânsito em uma rota específica. Em 2013, o Google adquiriu o aplicativo Waze, que deu eles ainda mais dados para trabalhar. Usuários do Waze informam o aplicativo sobre trânsito e acidentes e o Google Maps é capaz de acessar essas informações também. **Ademais, ele é capaz de manter o registro sobre o seu histórico e te auxiliar de diversas maneiras**.

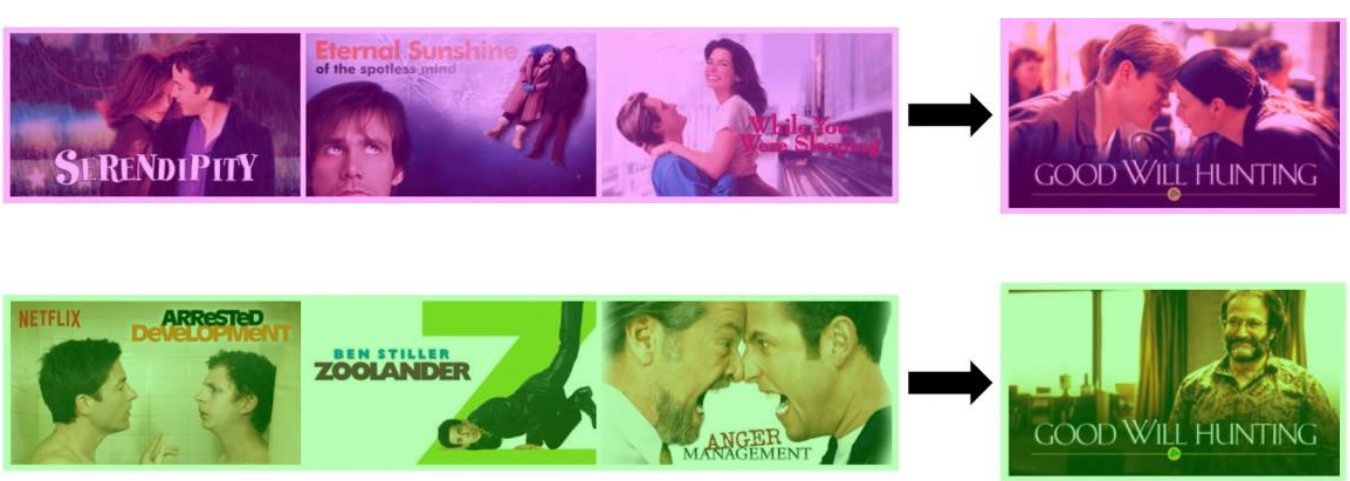


Hoje em dia, eu acho normal, mas eu me lembro da primeira vez que eu recebi uma notificação do Google me informando sobre o tempo até o meu trabalho. Eu acordei, tomei banho, escovei os dentes, desci para a garagem e assim que eu entrei no carro... o Google me enviou uma notificação informando que eu chegaria no trabalho em 15 minutos! **Eu fiquei igual a esse bebê do meme ao lado... achando que havia um drone me vigiando!**



Por fim, vamos falar sobre como a Netflix utiliza o Big Data para melhorar a sua experiência de entretenimento. Quando você acessa a Netflix, ela te dá uma lista de recomendações[2] em sua página inicial. Para dar essas recomendações, o algoritmo da Netflix aprende – a partir de infinitas fontes de dados – se você gosta de filmes estrelados, por exemplo, por Matt Damon. **Enfim, há diversas informações que ela pode cruzar para te recomendar o filme ideal.**

Uma das revelações mais interessantes trata do poster. *Sabe aquelas imagens que aparecem na hora que você vai escolher o filme?* Pois é, muitas pessoas escolhem um filme simplesmente baseado nessa imagem. **Uma vez que o título e a imagem são a primeira exposição ao conteúdo, escolher as imagens mais atrativas para pessoas específicas pode afetar na sua decisão de assistir um filme ou não**. Observem a imagem abaixo:



Há um filme chamado *Good Will Hunting* (em português, Gênio Indomável). Ele é interpretado por Matt Damon e Robin Williams. **Observem que há duas imagens diferentes para o mesmo filme!** No entanto, se você gosta mais de assistir filmes românticos, ele mostrará uma imagem do filme com o Matt Damon beijando uma mulher; se você gosta mais de assistir filmes de comédia, ele mostrará uma imagem do mesmo filme, porém com Robin Williams.

[2] O Sistema de Recomendações combina várias técnicas computacionais para selecionar itens personalizados com base nos interesses dos usuários conforme o contexto em que estão inseridos, com o intuito de obter vantagem competitiva. As recomendações são cada vez mais otimizadas por meio de técnicas de Inteligência Artificial, como Machine Learning.

De ambas as formas, ele consegue atrair a pessoa certa a assistir ao filme! **Possuir uma quantidade absurda de dados à disposição permite que a Netflix torne sua experiência melhor**. Por meio do Big Data, seria possível personalizar remédios com base no genoma de um paciente e prever qual remédio terá menos efeitos colaterais ou até mesmo qual tratamento possui a menor probabilidade de causar um ataque cardíaco.

Em suma: Big Data chegou para ficar e você está ajudando a criá-lo nesse instante ao ler essa aula! **Não existe uma definição singular sobre a terminologia Big Data**. Vejamos algumas:

OXFORD ENGLISH DICTIONARY	Big Data é um dado de grande tamanho, tipicamente ao nível que sua manipulação e gerenciamento apresenta desafios significativos a logística.
DUMBILL E EDD	Big Data é o dado que excede a capacidade de processamento convencional dos sistemas de bancos de dados.
MAYER-SCHÖNBERGER E CUKIER'S	Big Data é a habilidade da sociedade de aproveitar a informação por novas maneiras para produzir introspecção úteis ou bens e serviços de valor significante.
INTERNATIONAL DATA CORPORATION	Big Data é uma nova geração de tecnologias e arquiteturas, projetadas economicamente para extrair valor de volumes muito grandes e vastos de dados, permitindo alta velocidade de captura, descoberta e análise.
KIM, TRIMI E JI-HYONG	Big Data é o termo geral para a enorme quantidade de dados digitais coletados a partir de todos os tipos de fontes.
MAHRT E SCHARKOW	Big Data denota um maior conjunto de dados ao longo do tempo, conjunto de dados estes que são grandes demais para serem manipulados por infraestruturas de armazenamento e processamento regulares.
DAVENPORT E KWON	Big Data são dados demasiadamente volumosos ou muito desestruturados para serem gerenciados e analisados através de meios tradicionais.
DI MARTINO	Big Data se refere ao conjunto de dados cujo tamanho está além da habilidade de ferramentas típicas de banco de dados em capturer, gerenciar e analisar.
RAJESH	Big Data são conjuntos de dados que são tão grandes que se tornam difíceis de trabalhar com o uso de ferramentas atualmente disponíveis.

(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.

Comentários: ele realmente é um conceito que trata de tecnologias e arquiteturas projetadas para processar volumes muito grandes e com grande variedade de dados, permitindo alta velocidade de captura, descoberta e análise. Observem as palavras-chave: "processar volumes grandes", "grande variedade", "alta velocidade", "descoberta" (Correto).

(Polícia Federal – 2018) 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.

Comentários: perfeito... Big Data é um conceito que engloba tecnologias, ferramentas, arquiteturas e dados – os dados sem ferramentas para manipulá-los realmente são inúteis (Correto).

(CREF/11 – 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.

Comentários: infinidade de informações não estruturadas que podem ser utilizadas para tomada de decisões cada vez melhores é o Big Data (Letra C).

Trata-se de um termo amplo para conjuntos de dados muito grandes ou complexos que aplicativos de processamento de dados tradicionais são insuficientes. **Os desafios incluem análise, captura, curadoria de dados, pesquisa, compartilhamento, armazenamento, transferência, visualização e informações sobre privacidade.** Esse termo – por vezes – se refere ao uso de análise preditiva e outros métodos avançados para extrair valor de dados.

Por fim, minha definição favorita o define como a captura, gerenciamento e a análise de um grande volume de dados estruturados ou não estruturados que não 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**.

TIPOS DE DADOS	DESCRIÇÃO
DADOS ESTRUTURADOS	São dados que podem ser armazenados, acessados e processados em formato fixo e padronizado de acordo com alguma regra específica. Esta organização é geralmente feita por colunas e linhas (semelhante a planilhas do Excel), mas pode variar de acordo com a fonte de dados. Exemplo: Planilhas Eletrônicas, Bancos de Dados Relacionais e CSV.
DADOS SEMI-ESTRUTURADOS	São dados estruturados que não estão de acordo com a estrutura formal dos modelos de dados como em tabelas, mas que possuem marcadores para separar elementos semânticos e impor hierarquias de registros e campos dentro dos dados Exemplo: Dados de E-mail, Arquivos XML, Arquivos JSON e Banco de Dados NoSQL.

DADOS NÃO-ESTRUTURADOS	São dados que apresentam formato ou estrutura desconhecidos, em que não se sabe extrair de forma simples os valores desses dados em forma bruta. Exemplo: Documentos, Imagens, Vídeos, Arquivos de Texto, Posts em Redes Sociais.

(EBSERH – 2018) As soluções de big data focalizam dados que já existem, descartam dados não estruturados e disponibilizam os dados estruturados.

Comentários: Big Data é um termo que descreve o grande volume de dados – tanto estruturados quanto não-estruturados — que impactam as empresas diariamente (Errado).

De acordo com Mayer-Schonberger, com a informação, assim como na física, o tamanho importa. Desse modo, ao combinar centenas de bilhões de termos de busca, o Google mostrou ser capaz de identificar o surgimento de um surto de gripe quase tão bem quanto os dados oficiais com base nos pacientes que visitam o médico – e pôde gerar uma resposta quase em tempo real, muito mais rápido que as fontes oficiais.

Do mesmo modo, pode-se prever a volatilidade do preço de uma passagem de avião e, assim, dar um poder econômico significativo para os consumidores. No entanto, ambos só conseguem isso pela análise de centenas de bilhões de dados. Esses dois exemplos mostram o valor científico do Big Data, assim como a medida em que eles podem se tornar fonte de valor econômico. Essa quantidade massiva de dados tem influenciado áreas como saúde, governo, educação, etc.

Professor, onde esses dados são armazenados? Podem ser armazenados em um Data Warehouse ou em um Data Lake (Lago de Dados). **O Data Lake é um grande repositório capaz de armazenar dados estruturados, semi-estruturados e não-estruturados, assim como um método para organizar grandes volumes de dados de diversos formatos e de diversas fontes diferentes.** *Professor, qual seria a diferença entre Data Warehouse e Data Lake?*

DATA WAREHOUSE	DATA LAKE
Dados geralmente são tratados (limpos, combinados, organizados, etc) antes de serem armazenados.	Dados geralmente são armazenados da maneira que foram capturados – brutos, sem nenhum tratamento.
Podem armazenar todos os tipos de dados, mas o foco é nos dados estruturados.	Armazenam dados estruturados, semi-estruturados e não-estruturados.
Ideal para usuários operacionais visto que as ferramentas analíticas são mais fáceis de usar.	Ideal para cientistas de dados visto que as ferramentas analíticas são mais difíceis de usar.
Armazenamento de dados custam geralmente mais caro e consome mais tempo.	Armazenamento de dados custam geralmente mais barato e consome menos tempo.
Um esquema é definido antes dos dados serem armazenados.	Um esquema é definido após os dados serem armazenados.
Armazenam um grande volume de dados.	Armazenam um gigantesco volume de dados.

ANTES QUE ME PERGUNTEM, A IMAGEM É UMA BRINCADEIRA! =)

(TJ/RN – 2020) Big Data surgiu a partir da necessidade de manipular um grande volume de dados e, com isso, novos conceitos foram introduzidos, como o Data Lake, que:

a) pode ser considerado um repositório de dados relacionados, sendo, portanto, um armazém de dados orientado por assunto.

b) pode ser considerado um conjunto de bancos de dados relacionais e com relacionamentos entre tabelas de diferentes esquemas de bancos de dados.

c) é o resultado de sucessivas operações de mineração de dados, sendo um ambiente no qual é possível ter relatórios e dashboards de maneira amigável para os analistas de negócio.

d) é projetado para armazenar dados de diversas fontes e formatos, não havendo a necessidade da definição de um esquema de dados para inserir novos itens.

_____

Comentários: (a) Errado, os dados não precisam estar relacionados e, portanto, não é orientado por assunto; (b) Errado, não é um conjunto de dados relacionais e não precisa haver relacionamentos entre tabelas de diferentes esquemas – os dados são de diversos formatos e de diversas fontes; (c) Errado, não é o resultado de operações de mineração de dados – são dados brutos sem tratamento e da maneira que foram capturados; (d) Correto, ele realmente é projetado para armazenar dados de diversas fontes e formatos, não havendo a necessidade da definição de um esquema de dados para inserir novos itens (Letra D)

Antes de ver algumas curiosidades, eu acho bacana falar um pouco sobre a infraestrutura para suportar Big Data! Sim, pessoal... eu falei que se trata de uma quantidade absurda de dados. Isso implica a necessidade de uma infraestrutura também absurda. *Professor, o que você quer dizer com infraestrutura?* **Galera, eu me refiro ao conjunto de hardware, software e outras tecnologias capazes de suportar serviços de TI (Ex: Servidor, Firewall, Rede, etc).**

Hoje em dia, você pode utilizar serviços fornecidos pela computação em nuvem ou ter uma infraestrutura própria. Para o primeiro caso, existem inúmeras possibilidades de negócio para quem confia na combinação de Computação em Nuvem e Big Data! De forma geral, as empresas utilizam o Big Data para se tornarem mais competitivas. Além disso, espera-se com esse uso algo essencial para o sucesso: errar menos. E, quando inevitável, aprender com o erro.

Ter um sistema de computação em nuvem é condição para se trabalhar bem com um grande volume de dados, uma vez que isso envolve coleta, armazenamento e compartilhamento de um número gigantesco de informações. **Além disso, a constante necessidade de conhecer o resultado das ações de um negócio, muitas vezes, imediatamente, torna essa relação entre Cloud Computing e Big Data extremamente harmoniosa**. *Entendido?*

(TCU – 2015) Devido à quantidade de informações manipuladas, a (cloud computing) computação em nuvem torna-se inviável para soluções de big data.

_____

Comentários: na verdade, a computação em nuvem é a infraestrutura geralmente utilizada para suportar iniciativas de Big Data! *Por que?* Porque ela possui capacidade para processar grandes volumes de dados em tempo real. 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. A grande vantagem de associar Big Data à Cloud Computing é reduzir os custos de uma infraestrutura para armazenar e processar os dados (Errado).

Por fim, uma lista de curiosidades para que vocês entendam que quando falamos de grande volume de dados, é realmente um grande... volume... de dados! Vejam só:

CURIOSIDADE 1

A cada dois dias, a população mundial cria a mesma quantidade de dados criados do início da civilização humana até 2003.

CURIOSIDADE 2

O Google processa em média mais de 40.000 buscas a cada segundo. Isso significa mais de 3,5 bilhões de buscas por dia e 1,2 trilhão de buscas por ano em todo o mundo.

CURIOSIDADE 3

Os usuários do Facebook enviam em média 31,25 milhões de mensagens e visualizam 2,77 milhões de vídeos a cada minuto.

CURIOSIDADE 4

Dados estão crescendo mais rápido do que nunca e, até o ano de 2020, cerca de 1,7 megabytes de novos dados serão criados a cada segundo para cada ser humano no planeta.

CURIOSIDADE 5

Até lá, nosso universo digital de dados crescerá de 4,4 zettabyets para cerca de 44 zettabytes, ou 44 trilhões de gigabytes.

CURIOSIDADE 6

Em agosto de 2015, mais de 1/6 da população mundial (mais de um bilhão de pessoas) usaram o Facebook.

CURIOSIDADE 7

Estamos vendo um grande crescimento nos dados de vídeos e fotos, onde cada minuto até 300 horas de vídeo são enviados para o YouTube.

CURIOSIDADE 8

Em 2015, foram tiradas 1 trilhão de fotos e bilhões delas foram compartilhadas online. Em 2017, quase 80% das fotos foram tiradas em smartphones.

CURIOSIDADE 9

Até 2020, teremos mais de 6,1 bilhões de usuários de smartphones no mundo e pelo menos um terço de todos os dados passarão pela nuvem.

CURIOSIDADE 10

Hoje em dia, menos de 0,5% de todos os dados criados no planeta são analisados e utilizados, portanto existe um grande potencial ocioso.

Premissas

INCIDÊNCIA EM PROVA: ALTA

Podemos afirmar que a definição de Big Data pode ser quebrada em cinco dimensões, quais sejam: Volume, Velocidade, Variedade, Veracidade e Valor[1].

Volume

Big Data trata de uma grande quantidade de dados gerada a cada segundo. Pense em todos os e-mails, mensagens de Twitter, fotos e vídeos que circulam na rede a cada instante. Não são terabytes e, sim, zetabytes ou brontobytes. A tecnologia do Big Data serve exatamente para lidar com esse volume massivo de dados, guardando-os em diferentes localidades e juntando-os através de software.

Em outras palavras, nós podemos dizer que o volume de dados atualmente já é grande, mas a tendência é que continue a crescer ainda mais nas próximas décadas. Dessa forma, é preciso buscar ferramentas e formas de tratar esses dados de maneira que possam se converter – de fato – em informação que seja útil para o crescimento e desenvolvimentos das organizações e, não apenas, um grande volume de dados.

Velocidade

Refere-se à velocidade com que os dados são criados. São mensagens de redes sociais se viralizando em segundos, transações de cartão de crédito sendo verificadas a cada instante ou os milissegundos necessários para calcular o valor de compra e venda de ações. *Quem tem Twitter aí?* Hoje em dia, informações surgem primeiro no Twitter! O Big Data serve para analisar os dados no instante em que são criados, em tempo real, sem ter de armazená-los.

---

[1] O Big Data foi inicialmente conceituado a partir de três premissas básicas: Volume, Velocidade e Variedade (3 V's). Atualmente, já há autores que tratam de 10V's (+Variabilidade, Validade, Vulnerabilidade, Volatilidade e Visualização), apesar de não cair em prova.

Não apenas o volume de dados é gigantesco, mas a velocidade em que esses dados são produzidos (e se tornam desatualizados é vertiginosa). Justamente por isso o segundo desafio do Big Data é o *timing* do processamento desses dados: **para que possuam valor real e aplicabilidade no mercado, é preciso utilizar os dados antes que se tornem desatualizados**. O objetivo, portanto, é alcançar formas de trabalhar o processamento dessas informações em tempo real.

Variedade

No passado, a maior parte dos dados utilizados por organizações era estruturado e podia ser facilmente armazenado em tabelas de bancos de dados relacionais. No entanto, a maioria dos dados do mundo não se comporta dessa forma. Com o Big Data, mensagens, fotos, mídia social, e-mail, vídeos e sons – que são dados não-estruturados – podem ser administrados juntamente com dados tradicionais.

Os dados de que dispomos atualmente são provenientes das mais diversas fontes: redes sociais, aplicativos, cookies, IoT, e-mails, etc. **Isso significa que não seguem um único padrão e nem fornecem todos o mesmo tipo de informações, tornando a tarefa de compilar esses dados em um banco de dados tradicional inviável**. É preciso desenvolver novas ferramentas de análise que respondam à heterogeneidade dos dados.

Veracidade

Um dos pontos mais importantes de qualquer informação é que ela seja verdadeira. Com o Big Data, não é possível controlar cada hashtag do Twitter ou notícia falsa na internet, mas com análises e estatísticas de grandes volumes de dados é possível compensar as informações incorretas. Dentre a massa de dados que circula, é preciso estabelecer quais os dados que são verídicos e que ainda correspondem ao momento atual.

Dados desatualizados podem ser considerados inverídicos, mas não porque tenham sido gerados com segundas intenções, mas porque não correspondem mais à realidade e podem guiar uma empresa a decisões equivocadas. O desafio posto pelo Big Data é, então, determinar a relevância dos dados disponíveis para uma empresa, de forma que essas informações possam servir de guia para o seu planejamento com maior segurança.

Valor

O último V é o que torna Big Data relevante: **tudo bem ter acesso a uma quantidade massiva de informação a cada segundo, mas isso não adianta nada se não puder gerar valor algum para um órgão ou uma empresa**. É importante que organizações entrem no negócio do Big Data, mas é sempre importante lembrar dos custos e benefícios, além de tentar agregar valor ao que se está fazendo. *Bacana?*

O quinto desafio posto pelo Big Data pelas empresas é o de definir a abordagem que será feita dessa massa de dados que está circulando. Afinal, para que um dado se converta em informação útil e utilizável é preciso o olho do analisador, é preciso colocar uma pergunta a esse dado que permita orientar a análise de dados para o objetivo de uma empresa. Não é toda a informação que está circulando que é relevante ou útil para os objetivos específicos de uma empresa.

PALAVRAS-CHAVE				
VOLUME	VELOCIDADE	VALOR	VERACIDADE	VARIEDADE
Terabytes	Transmissão	Estatístico	Confiabilidade	Estruturado
Registros	Tempo Real	Hipóteses	Autenticidade	Não-Estruturado
Tabelas/Arquivos	Processos	Correlações	Origem/Reputação	Múltiplas Fontes

PREMISSAS	DESCRIÇÃO
VOLUME	Corresponde à grande quantidade de dados acumulada.
VELOCIDADE	Corresponde à rapidez na geração e obtenção de dados.
VARIEDADE	Corresponde à grande diversidade de tipos ou formas de dados.
VERACIDADE	Corresponde à confiança na geração e obtenção dos dados.
VALOR	Corresponde à utilidade e valor agregado ao negócio.

(ANAC – 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.

Comentários: trata-se do volume + variedade + velocidade + veracidade, e tudo agregando em valor (Letra D).

(DPE/RS – 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.

Comentários: (a) Errado. ~~Veracidade~~ Velocidade corresponde à rapidez na geração e obtenção de dados; (b) Errado. ~~Valor~~ Volume corresponde à grande quantidade de dados acumulada; (c) Errado. ~~Volume~~ Velocidade corresponde à rapidez na geração e obtenção de dados; (d) Errado. ~~Velocidade~~ Veracidade corresponde à confiança na geração e obtenção dos dados; (e) Correto. Variedade corresponde ao grande número de tipos ou formas de dados (Letra E).

Big Data Analytics

INCIDÊNCIA EM PROVA: BAIXA

Nós já sabemos que a imensa parte dos dados disponíveis no mundo hoje foram criados apenas nos últimos dois anos. Estes dados são caracterizados por sua velocidade, volume, variedade, veracidade e valor – conforme vimos anteriormente. Mais de 2.5 trilhões de bytes são gerados todos os dias por meio de nossos smartphones, tablets, sensores, redes sociais e cartões de crédito, mas o que pode ser feito com todos esses dados é que é a pergunta relevante.

É aí que entra o conceito de Big Data Analytics: o estudo e interpretação de grandes quantidades de dados armazenados com a finalidade de extrair padrões de comportamento. Em outras palavras, utiliza-se uma combinação de sistemas de softwares matemáticos de alta tecnologia que juntos são capazes de tratar dados estruturados e não-estruturados, analisá-los e extrair um significado de alto valor para organizações.

Dessa forma, o Big Data Analytics poderá auxiliar empresas privadas ou administradores de órgãos públicos a entender seus usuários, encontrar oportunidades não percebidas anteriormente, fornecer um serviço melhor e mitigar possíveis fraudes – são bastante utilizados em órgãos fazendários – como a Receita Federal – para evitar sonegação de tributos. *Ué, professor... isso não seria Business Intelligence?* Não, vamos ver a diferença...

O objetivo de ambos é ajudar uma organização a tomar boas decisões por meio da análise de dados. No entanto, o Business Intelligence ajuda a encontrar as respostas para as perguntas de negócios que já conhecemos, enquanto o Big Data Analytics nos ajuda a encontrar as perguntas e respostas que nem sequer sabíamos que existiam – tudo isso por meio de padrões, correlações desconhecidas, tendências de mercado e preferências de consumidores.

Em outras palavras, o Business Intelligence trata de encontrar respostas que explicam o passado, já o Big Data Analytics trata de encontrar as perguntas que explicam o futuro. Ambos possuem grande importância, complementam-se e devem ser bem entendidos para que as empresas possam aproveitá-los da melhor forma, agregando e alcançando os valores e resultados desejados aos negócios. *Professor, como eu vou encontrar perguntas que explicam o futuro?*

Prever o futuro é um desejo comum entre as pessoas! *Se você fosse capaz de saber com seis meses de antecedência que uma grande crise econômica iria assolar o seu país, o que faria? Será que você conseguiria criar um plano para prevenir ou diminuir o impacto daquele grande problema? Poderia mudar o rumo da história?* **Com a ajuda de estratégias de Análise Preditiva, você pode conseguir (sim, é sério!).**

A Análise Preditiva não é bola de cristal, nem obra da Mãe Dináh! Trata-se, na verdade, do trabalho de analisar um cenário específico e traçar possíveis tendências e mudanças capazes de afetar seu planejamento estratégico. **É óbvio que, muitas das vezes, esse tipo de trabalho lida com volumes**

gigantescos de dados e, por isso, exige o uso de ferramentas de inteligência artificial para analisar a correlação entre os dados. *Viram como tudo se encaixa?*

A Análise Preditiva é capaz de identificar o relacionamento existente entre os componentes de um conjunto de dados, utilizando algoritmos sofisticados, com o intuito de identificar padrões de comportamento ao examinar automaticamente grandes quantidades de dados. **Dessa forma, permite-se que estatísticas e dados armazenados sejam agrupados, fornecendo previsões e indicando padrões e tendências comportamentais.**

Galera, esse tema não é novo, mas só recentemente tem ganhado notoriedade como uma ferramenta de negócio. Com o avanço de tecnologias que possibilitam a mineração de dados, a Análise Preditiva conta com cada vez mais segurança e precisão para descobrir padrões e avaliar a probabilidade de um resultado ou acontecimento futuro, diferentemente da simples análise descritiva de dados, que apenas mede e apresenta resultados passados.

TIPO DE ANÁLISE	QUESTÃO?	DESCRIÇÃO
ANÁLISE DESCRITIVA	O QUE ACONTECEU?	Em vez de se focar no futuro, busca fazer uma fotografia do presente, para que decisões de cunho imediato possam ser tomadas com segurança. Ela trabalha com histórico de dados, cruzando informações com o objetivo de gerar um panorama claro e preciso dos temas relevantes para a empresa no presente momento. Exemplo: por meio do cruzamento de dados, conclui-se que determinada pessoa atualmente é identificada como má pagadora.
ANÁLISE DIAGNÓSTICA	POR QUE ACONTECEU?	O foco está na relação de causas e consequências percebidas ao longo do tempo, dentro de um determinado tema. Assim, a análise diagnóstica funciona baseada na coleta de dados relacionados a um determinado assunto, cruzando informações com o objetivo de entender quais fatores influenciaram o resultado atual. Exemplo: determinada pessoa nunca havia sido identificada como má pagadora – somente é agora porque ficou viúva recentemente.
ANÁLISE PREDITIVA	O QUE IRÁ ACONTECER?	Este tipo de análise é o mais indicado para quem precisa prever algum tipo de comportamento ou resultado. Esta técnica busca analisar dados relevantes ao longo do tempo, buscando padrões comportamentais e suas variações de acordo com cada contexto, a fim de prever como será o comportamento de seu público ou mercado no futuro, dadas as condições atuais. Exemplo: quanto estará o valor do dólar no ano que vem?
ANÁLISE PRESCRITIVA	O QUE DEVO FAZER?	Segue um modelo similar à Análise Preditiva, no entanto com objetivos ligeiramente diferentes. Em vez de tentar prever um determinado acontecimento, esta análise busca prever as consequências deste acontecimento. Exemplo: dado um aumento do valor do dólar no ano que vem, como isso poderá afetar as importações de matéria prima, consequentemente, o faturamento das vendas de determinada empresa.

Uma dúvida comum é sobre a diferença entre Business Intelligence, Big Data Analytics e Data Mining! Alguns autores consideram um é a evolução do anterior, abrangendo mais dados e ferramentas matemáticas/estatísticas; outros afirmam que – na verdade – é tudo a mesma coisa e

que possuem nomes diferentes apenas por uma questão de marketing: vende mais dizer que uma ferramenta de software é uma solução de Big Data Analytics do que uma solução de Data Mining.

BIG DATA ANALYTICS É:	BIG DATA ANALYTICS NÃO É:
Uma estratégia baseada em tecnologia que permite coletar insights mais profundos e relevantes de clientes, parceiros e negócio, ganhando assim uma vantagem competitiva.	Somente tecnologia – no nível empresarial, refere-se a explorar fontes amplamente melhoradas de dados para adquirir insights.
Trabalhar com conjuntos de dados cujo porte e variedade estão além da habilidade de captura, armazenamento e análise de softwares de banco de dados típicos.	Somente volume – também se refere à variedade e à velocidade, mas – talvez mais importante – refere-se ao valor derivado dos dados.
Processamento de um fluxo contínuo de dados em tempo real, possibilitando a tomada de decisões sensíveis ao tempo mais rápido do que em qualquer outra época.	Mais gerada ou mais utilizada somente por grandes empresas online como Google ou Amazon. Embora as empresas de internet possam ter sido pioneiras no Big Data na escala web, aplicativos chegam a todas as indústrias.
Distribuído na natureza, isto é, o processamento de análise vai aonde estão os dados para maior velocidade e eficiência.	Uso de bancos de dados relacionais tradicionais de "tamanho único" criados com base em disco compartilhado e arquitetura de memória. Análise de Big Data usa uma rede de recursos de computação para processamento massivamente paralelo e escalável.
Um novo paradigma no qual a tecnologia da informação colabora com usuários empresariais e "cientistas de dados" para identificar e implementar análises que ampliam a eficiência operacional e resolvem novos problemas empresariais.	Um substituto de bancos de dados relacionais – dados estruturados continuam a ser de importância crítica para as empresas. No entanto, sistemas tradicionais podem não ter capacidade de manipular as novas fontes e contextos do Big Data.
Transferir a tomada de decisão dentro da empresa e permitir que pessoas tomem decisões melhores, mais rápidas e em tempo real.	-

Apesar de estarmos apenas nos primórdios do Big Data, ele é utilizado diariamente. Filtros antispam são projetados para se adaptarem automaticamente às mudanças dos tipos de lixo eletrônico. Sites de namoro formam pares com base em como suas várias características correspondem às de relacionamentos anteriores. O corretor automático dos smartphones analisa nossas ações e acrescenta novas palavras a seus dicionários com base no que é escrito.

Por fim, vamos ver dois casos de sucesso que se tornaram referência de êxito na utilização do conceito de Big Data Analytics. Me acompanhem...

MC Donald's

O Fast-Food mais famosos do planeta, o MC Donald's, gerencia cerca de 34 mil restaurantes e serve mais de 69 milhões de pessoas em 118 países – tudo isso com frequência diária! Com base nesse pequeno trecho de informações, você já deve estar imaginando o quão gigantesco é o número de dados gerados diariamente pelo MC Donald's. *Bom, e o que o maior restaurante faz com todos esses dados gerados?*

Sabe-se que o MC Donald's coleta e combina os dados de suas lanchonetes ao redor do globo com o objetivo de padronizá-los e, com isso, compreender o comportamento de seu público; como esse público percebe seus produtos; os aperfeiçoamentos logísticos; e layouts que podem ser concebidos para melhorar a experiência do usuário perante seus serviços e produtos. Tudo isso com o auxílio do Big Data!

A partir dos estudos de Sentiment Analysis (Análises de Sentimentos) realizados em redes sociais, foram lançados novos sanduíches, promoções em tempo real, entre outros. Tudo isso só foi possível graças ao acompanhamento dos cientistas de dados, que mensuraram atentamente as manifestações e reações de seu público – alterando estratégia em tempo real e, até a logística do Drive-Thru. *Como assim, professor?*

A logística de pedido, produção e entrega de sanduíches foi alterada em cada país conforme as reações de seus consumidores no que diz respeito ao layout, tempo de espera e informações providenciadas por seus funcionários no ponto de entrega dos lanches. Tudo possibilitado por meio de ferramentas de Big Data. *Legal, não é?* Esse foi um caso de sucesso da utilização do Big Data na prática em organizações.

American Express

A American Express, empresa norte-americana de serviços financeiros, passou a investir em Big Data ao perceber que os insights gerados pelas ferramentas tradicionais de BI não estavam sendo suficientes para diminuir as taxas de cancelamento de seus clientes. **A companhia desenvolveu sofisticados modelos preditivos para analisar históricos de transações dos usuários de seus cartões de crédito, além de 115 variáveis, para prever potenciais churns**.

Professor, o que diabos é churn? É um termo em inglês para uma métrica que indica o quanto uma empresa perdeu de receite ou de clientes! Após a implementação de uma solução de Big Data em seus processos, **a American Express acredita ser capaz, por exemplo, de identificar 24% de seus clientes australianos que pretendem encerrar suas contas dentro dos próximos quatro meses**. *Genial, concordam?*

Conceitos Avançados

NoSQL (Not Only SQL)

INCIDÊNCIA EM PROVA: BAIXA

Antes de falar sobre NoSQL, precisamos falar o que é SQL (Structured Query Language). O SQL é uma linguagem de consulta estruturada utilizada para manipular bancos de dados relacionais (tabelas). Por meio dela, um usuário pode executar comandos para inserir, pesquisar, atualizar ou deletar registros em um banco de dados relacionais, criar ou excluir tabelas, conceder ou revogar permissões para acessar o banco de dados, entre outros recursos.

É interessante, mas notem que agora estamos falando em um contexto de Big Data! E nós já sabemos que grande parte dos dados armazenados e processados dentro desse conceito são não-estruturados e/ou semi-estruturados, logo não se adequam bem a bases de dados relacionais. **Foi então que surgiu o NoSQL (Not Only SQL) – esse é o nome genérico dado a bancos de dados distribuídos e não relacionais, em que não há estruturas de tabelas (linhas e colunas).**

Bancos de Dados NoSQL são cada vez mais utilizados em aplicações web de tempo real (online) com a finalidade de atender aos requisitos de gerenciamento de grandes volumes de dados que necessitam de alta disponibilidade e escalabilidade. Aliás, eles geralmente são orientados a documentos, isto é, são capazes de manipular dados semiestruturados (Ex: XML e JSON). *Ora, mas por que não continuar utilizando bancos de dados relacionais?* Vejamos...

Nos dias de hoje, o volume de dados de certas organizações (Ex: Facebook, que atingiu o nível de 300 petabytes ou 300 mil terabytes) atingiu valores nunca antes imaginados. No caso destes tipos de organizações, a utilização de bancos de dados relacionais tem se mostrado muito problemática e até ineficiente. **Os principais problemas estão relacionados à dificuldade de conciliar o tipo de modelo com a demanda da escalabilidade que está cada vez mais frequente.**

Vamos tomar como exemplo o próprio Facebook! Caso o sistema esteja rodando sobre um banco de dados relacional e haja um crescimento do número de usuários, haverá consequentemente uma queda de performance. **Para superar este problema, seria necessário fazer um upgrade na potência do servidor atual (também chamado de escalabilidade vertical) ou aumentar o número de servidores (também chamado de escalabilidade horizontal).**

No entanto, em organizações que tratam de uma quantidade massiva de dados que nunca param de crescer, chega um momento em que o banco de dados não consegue mais atender todas as requisições em um tempo hábil. A escalabilidade vertical é mais fácil, mas é mais limitada; **a escalabilidade horizontal é ilimitada, mas é mais complexa** – é necessário realizar uma série de configurações e alterações nas aplicações para que tudo funcione em uma arquitetura distribuída.

Galera, bancos de dados relacionais estão mais focados nos relacionamentos entre as entidades. **Como vantagem, isso mantém a integridade dos dados; como desvantagem, torna mais burocráticas alterações e implementações de novas funcionalidades.** Como esse intenso volume de dados vem aumentando e pela sua natureza não-estruturada ou semiestruturada, desenvolvedores perceberam a dificuldade ao se organizar dados no modelo relacional.

Pensando em solucionar diversos problemas relacionados à escalabilidade, performance e disponibilidade, projetistas de bancos de dados não-relacionais promoveram uma alternativa de alto armazenamento com alta velocidade e alta disponibilidade, **procurando se livrar de certas regras e estruturas inflexíveis que norteiam o modelo relacional de armazenamento de dados**. *Que legaaaaaaal, professor...*

A proposta dos bancos de dados não-relacionais não é substituir os bancos de dados relacionais, mas serem utilizados nos casos em que é necessária uma maior flexibilidade na estrutura do banco de dados. Dito isso, eu gostaria de apresentar para vocês uma tabela que contém uma comparação com as diferenças fundamentais entre NoSQL e SQL quanto ao modelo, armazenamento, flexibilidade, adequação, escalabilidade e exemplos de aplicações.

CRITÉRIO	NOSQL	SQL
MODELO	Não-Relacional	Relacional
ARMAZENAMENTO	Variados (Grafos, Documentos, etc)	Tabelas
FLEXIBILIDADE	Alta flexibilidade (Esquema indefinido)	Baixa flexibilidade (Esquema definido)
ADEQUAÇÃO	Mais adequado a dados não-estruturados	Mais adequado a dados estruturados
ESCALABILIDADE	Em geral, escalabilidade horizontal	Em geral, escalabilidade vertical
SGBD	MongoDB, Cassandra, HBase, Neo4J, etc	Oracle, MySQL, DB2, SQL Server, etc

Dentro desse contexto, eu gostaria de enfatizar alguns pontos! **Primeiro: NoSQL é um termo que funciona como um guarda-chuva para bancos de dados não-relacionais.** *Isso significa que todos esses bancos de dados não-relacionais terão características semelhantes?* Não, esse é um termo genérico para absolutamente todas as variedades de bancos de dados que não sejam relacionais, abarcando bancos de dados com algumas características completamente díspares entre si.

Segundo: apesar do nome sugerir o contrário, bancos de dados não-relacionais podem – sim – armazenar relacionamentos entre dados, no entanto eles o farão de maneira diferente de bancos de dados relacionais. Lembrando que bancos de dados relacionais, em geral, utilizam chaves (primárias e estrangeiras) para armazenar o relacionamento entre dados, já os bancos de dados não-relacionais utilizam cada um o seu modelo/forma de armazená-los.

Terceiro: bancos de dados relacionais e não-relacionais possuem uma grande diferença em relação ao esquema de dados. *Vocês ainda se lembram o que é esquema ou já esqueceram?* Esquema é uma descrição do banco de dados, que informa qual estrutura/organização será utilizada para suportar os dados que serão manipulados. Ora, bancos de dados relacionais possuem esquema inflexíveis enquanto bancos de dados não-relacionais possuem esquemas flexíveis.

Diego, não saquei essa parte! Galera, quando eu era criança um número de telefone possuía apenas sete números (eu até me lembro do primeiro número da minha casa: 354-8915). **Um desenvolvedor de sistemas daquela época, ao criar o esquema de banco de dados que descreveria os dados que seriam armazenados, especificou que a coluna TELEFONE de uma determinada tabela receberia um número com exatos sete algarismos.**

No entanto, o tempo passou e os números de telefone brasileiros ganharam mais um número. Esse tipo de alteração é bastante problemático e trabalhoso para o desenvolvedor de sistemas porque ele tem que alterar todo o esquema que havia sido estabelecido inicialmente. Alguns anos atrás, os números de telefone ganharam mais um número, fazendo com que desenvolvedores xingassem até a última geração de quem teve essa ideia, porque novamente tiveram que alterar o esquema.

Em outras palavras, bancos de dados tradicionais possuem esquemas extremamente inflexíveis. Eu mencionei um caso bastante simples, mas há alterações que podem realmente ser extremamente complexas, ainda mais em bancos de dados antigos. Pois bem, os bancos de dados não-relacionais são diferentes: eles se caracterizam pela ausência parcial ou total de esquemas que definem uma estrutura de dados – também chamado de esquema flexível ou ausência de esquema.

Dessa forma, dados armazenados em um banco de dados não-relacional podem ter características, tipos, estruturas e organizações completamente diferentes até mesmo entre um mesmo conjunto de dados. **É justamente essa ausência de esquema que facilita uma alta escalabilidade e alta disponibilidade, mas em contrapartida não há a garantia de integridade dos dados, fato que não ocorre em bancos de dados relacionais.**

Professor, se bancos de dados não-relacionais não armazenam dados em tabelas, como eles armazenam seus dados? **Bancos de dados não-relacionais utilizam modelos diferentes de armazenamento de dados, os quais podem ser divididos em quatro categorias principais: Chave-Valor, Orientado a Documentos, Orientado a Grafos e Orientado a Colunas (Colunar).** Alguns bancos de dados podem implementar mais de um desses modelos.

TIPO DE MODELOS	DESCRIÇÃO
ORIENTADO A CHAVE-VALOR	Esse modelo armazena dados por meio de uma estrutura de mapeamento ou dicionário, em que todo dado armazenado possui uma chave identificadora e seu valor em si. Para cada chave de entrada, é retornado um valor de saída (Ex: Table Storage, DynamoDB, Cassandra e Redis).
ORIENTADO A DOCUMENTOS	Esse modelo armazena dados na forma de documentos flexíveis, semiestruturados e hierárquicos junto com seus metadados sem uma estrutura definida. Em geral, os dados são armazenados em formato JSON ou XML (Ex: MongoDB, CouchDB e DocumentDB).
ORIENTADO A GRAFOS	Esse modelo armazena o relacionamento entre dados altamente conectados por meio de vértices e arestas. São geralmente utilizados em redes sociais, mecanismos de recomendação e detecção de fraudes (Ex: Neo4J, Infinite Graph e ArangoDB).
ORIENTADO A COLUNAS	Esse modelo armazena dados em colunas dinâmicas. É o mais semelhante ao modelo relacional, mas os dados são armazenados em colunas em vez de linhas. Ademais, cada coluna pode conter subcolunas, que podem conter várias propriedades (Ex: Hypertable e MonetDB).

(CNJ – 2013) Apesar de implementarem tecnologias distintas, todos os bancos de dados NoSQL apresentam em comum a implementação da tecnologia chave-valor.

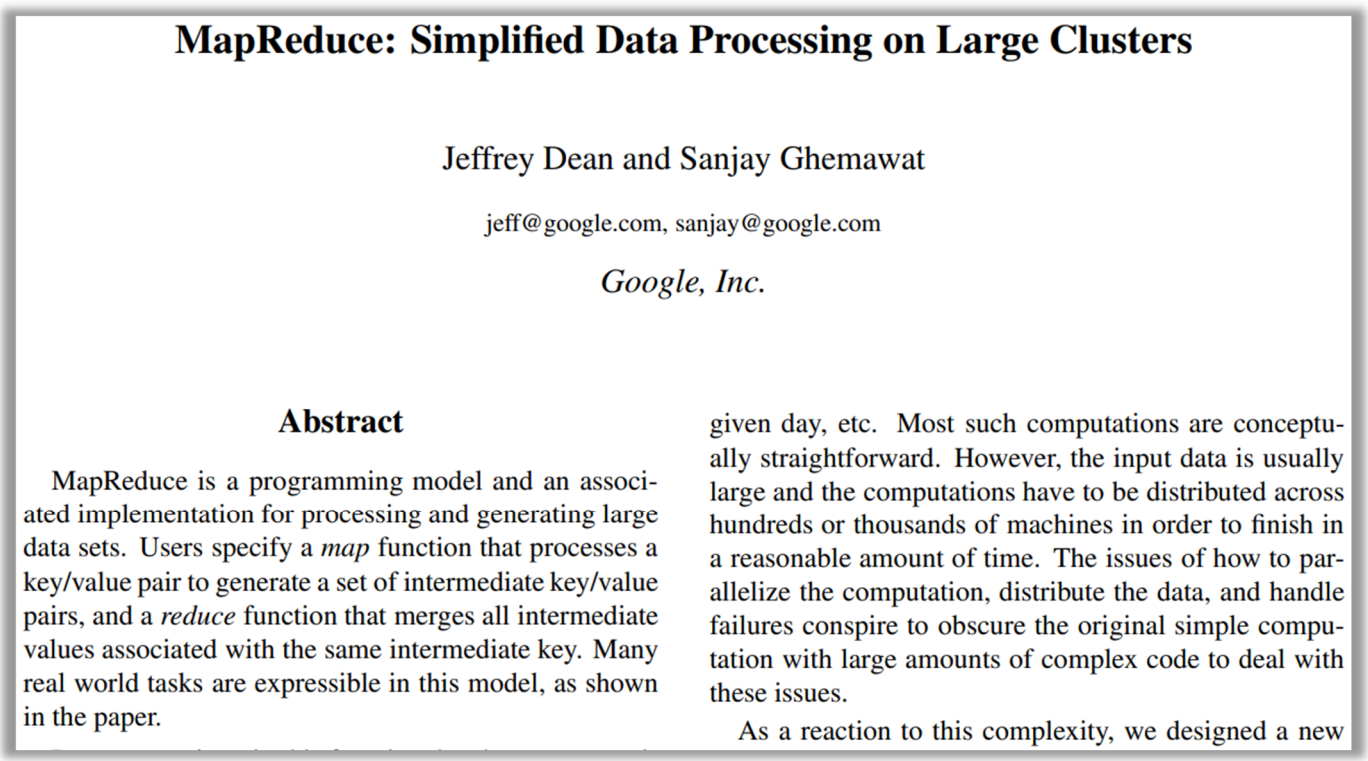
_____

Comentários: eles podem apresentar implementações diferentes, como chave-valor, orientado a documentos, orientado a grafos ou orientado a colunas (Errado).

Hadoop/MapReduce

INCIDÊNCIA EM PROVA: BAIXA

Esse é um assunto um pouquinho complexo, portanto eu vou explicá-lo com algumas metáforas para que vocês não esqueçam. Em 2004, o Google publicou um artigo chamado MapReduce: Processamento Simplificado de Dados em Grandes Clusters. **De acordo com os autores, tratava-se de um modelo de programação e uma implementação associada para processamento e geração de grandes conjuntos de dados.** Vamos começar com as metáforas...



MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new

Antigamente, havia uma ideia de que, quanto mais dados eram armazenados e processados por uma organização, maior era a necessidade de aquisição de computadores maiores e mais potentes. **O MapReduce surgiu porque algumas organizações começaram a perceber que, quando você começava a armazenar quantidades gigantescas de dados, obter computadores maiores e mais rápidos não funcionava mais.**

A não ser que a organização possua o computador do Homem de Ferro (que tem uma capacidade computacional infinita), chega um momento para todos os mortais em que não adianta comprar mais memória, processador, armazenamento para o computador. *Por que?* **Porque isso não é muito escalável – por melhor que seja, um computador tem o seu limite!** *E há alguma maneira ter uma escalabilidade maior?* Claro que há! Vamos ver como funciona...

Alguns problemas computacionais podem ser resolvidos com muita facilidade dividindo os dados em blocos menores (é o famoso Dividir para Conquistar)! *Como assim, Diego?* Vamos supor que você está tentando encontrar o maior número em uma lista com cem milhões de números. Uma

maneira seria comprar um único computador bem potente para verificar número por número até encontrar qual é o maior número da lista.

Se esse computador poderoso tem a capacidade de verificar que possa olhar através de um milhão de números por hora, ele precisará – portanto – de 100 horas (+- 4 dias) para verificar todos os cem milhões de números. **Agora, se você dividir essa lista em cem partes e as entregar para 100 computadores, cada computador pesquisará em uma lista de apenas 1 milhão de números e poderá encontrar o maior número dessa lista em 1 hora.**

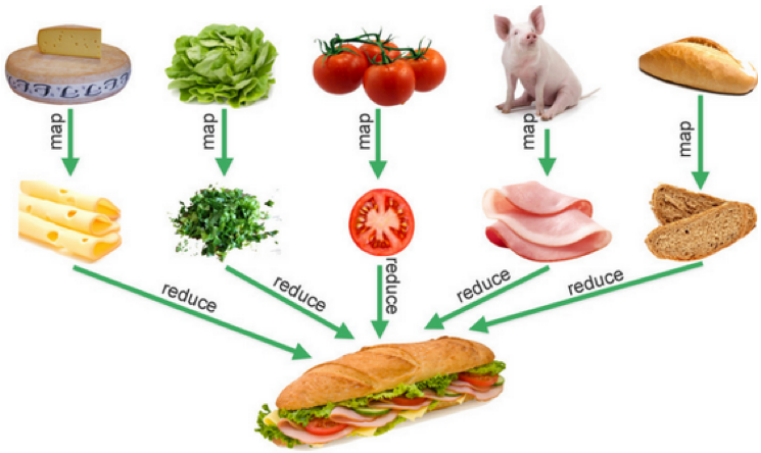
Após cada computador encontrar seu maior número, bastam alguns segundos para encontrar o maior dentre esses cem números. Logo, o trabalho que foi realizado em cerca de quatro dias por um único computador poderá ser finalizado em cerca de uma hora pelos cem computadores. **O processo de decomposição dos dados é chamado de Mapeamento (Map); e o processo de consolidação do resultado dos mapeamentos é chamado de Redução (Reduce).**

O MapReduce é um modelo de programação que **permite reduzir problemas grandes em problemas menores**, mapeando cada subproblema para máquinas diferentes (ou processadores diferentes de uma mesma máquina) e, em seguida, reduzindo cada resposta intermediária à única resposta final que você está procurando. Um excelente exemplo para que vocês nunca esqueçam é o site de questões do Estratégia Concursos. Atualmente, ele possui um bocadinho de questões...

Ainda nesse semestre, será disponibilizada a funcionalidade que permite fazer buscas por palavras em todas essas questões. *Vocês já imaginaram a complexidade do problema de fazer uma busca textual em mais de dois milhões de questões?* **Pois é, mas vocês já sabem que uma maneira de resolver problemas complexos é dividindo-o em vários problemas menores (inclusive, isso vale para tudo na vida).** Vejamos...

Quando eu pesquisar por "MapReduce" em nosso sistema, ele retornará todas as questões que contenham essa palavra. Se um único computador fosse responsável por realizar essa busca, demoraríamos muito! **Uma forma de resolver esse problema é dividir a busca em diversos computadores diferentes trabalhando paralelamente** (Ex: Computador 1 buscará em questões do CESPE; Computador 2 buscará em questões da FCC; e assim por diante).

Após cada computador chegar ao seu resultado, pode-se consolidar os resultados individuais em um único resultado global que contenha todas as questões encontradas com essa palavra. Como último exemplo, vamos ver o clássico do sanduíche! **Para fazer uma grande quantidade de sanduíches em uma franquia de fast-food, cada funcionário pode ser responsável por escolher um ingrediente e cortá-lo (processo de mapeamento).**



Após todos os ingredientes estarem cortados, outro funcionário pode ser responsável por juntá-los em um único sanduíche (processo de redução). A imagem ao lado representa toda a ideia que vimos nessa página (coitado do porquinho). Bem, agora eu pergunto: *será que essa técnica é útil para Big Data?* Ora, evidente que sim! **Um dos problemas do Big Data é justamente a dificuldade de processar uma quantidade massiva de dados.**

Legal... nós vimos um bocado de metáforas, mas agora temos que ver alguns termos técnicos. O MapReduce é considerado um modelo de programação que permite o processamento de dados massivos em um algoritmo paralelo e distribuído (em *clusters*). **A etapa de mapeamento se baseia em uma combinação de chave-valor.** *Como assim, Diego?* Voltemos ao exemplo do sistema de questões: a chave escolhida foi Banca e o valor é o Nome da Banca em si.

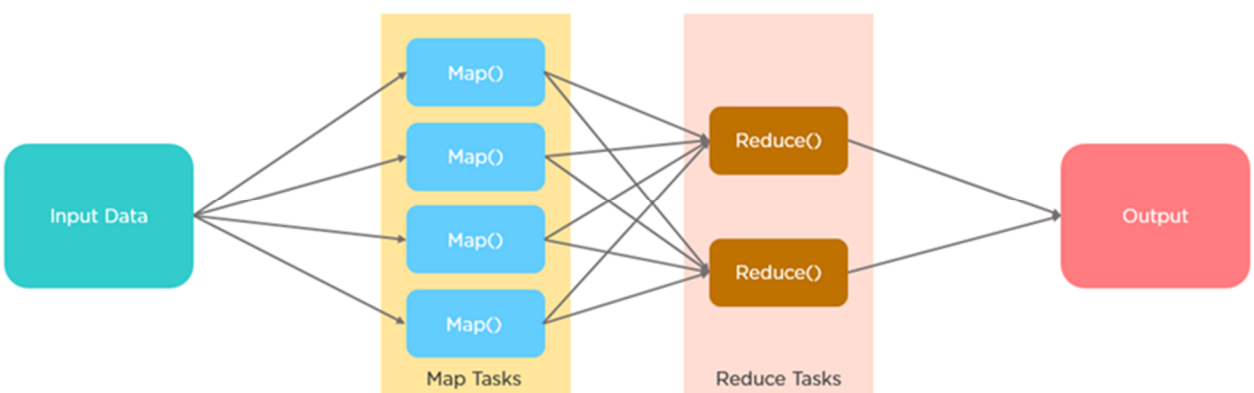
- MAPEAMENTO (CHAVE, VALOR) → MAPEAMENTO (BANCA, CESPE); MAPEAMENTO (BANCA, FCC); MAPEAMENTO (BANCA; FGV) ...

Professor, eu poderia escolher outra chave? Claro que sim! Nós poderíamos ter escolhido, por exemplo, matéria, ano, dificuldade, entre outros. Vejamos:

1. MAPEAMENTO (CHAVE, VALOR) → MAPEAMENTO (MATÉRIA, INFORMÁTICA); MAPEAMENTO (MATÉRIA, PORTUGUÊS) ...

2. MAPEAMENTO (CHAVE, VALOR) → MAPEAMENTO (ANO, 2020); MAPEAMENTO (ANO, 2019); MAPEAMENTO (ANO; 2018) ...

3. MAPEAMENTO (CHAVE, VALOR) → MAPEAMENTO (DIFICULDADE, FÁCIL); MAPEAMENTO (DIFICULDADE, DIFÍCIL); ...



Já a etapa de redução é responsável por consolidar os resultados de cada mapeamento, gerando um resultado agregado. *E o tal do Hadoop, Diego?* Apache Hadoop é apenas uma das implementações da técnica de MapReduce – existem outras implementações, mas essa é a mais

famosa! Em outras palavras, ele é um software de código aberto, implementado na linguagem de programação Java, para implementar o algoritmo de MapReduce em máquinas comuns.

Na verdade, ele é mais do que um software – ele é uma plataforma, um framework, um ecossistema de computação distribuída orientada a clusters e voltado para **armazenamento e processamento de grandes volumes de dados, com alta escalabilidade, grande confiabilidade e tolerância a falhas.** Ele é responsável por todo gerenciamento do cluster, não sendo necessário configurar máquinas individualmente.

Em implementações de Big Data, temos uma arquitetura baseada em dois componentes principais: armazenamento de dados e processamento paralelo. **No Hadoop, o armazenamento distribuído de dados é tratado pelo HDFS (Hadoop File System[1]) e o processamento paralelo de dados é tratado pelo MapReduce.** Em suma, podemos dizer que o Hadoop é uma combinação do MapReduce e do HDFS (que foi inspirado no GoogleFS – Google FileSystem).

(Polícia Federal – 2018) MapReduce permite o processamento de dados massivos usando um algoritmo paralelo mas não distribuído.

_____

Comentários: na verdade, ele é um modelo de programação que permite o processamento de dados massivos em um algoritmo escalável, paralelo e distribuído, geralmente utilizando um cluster de computadores (Errado).

(ANAC – 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.

_____

Comentários: essa plataforma é também chamada de Hadoop – nenhuma das outras opções existem! (Letra D).

[1] HDFS é um sistema de arquivos (forma de organização de dados em um meio de armazenamento em massa) 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. Esse tipo de construção faz com que seja desaconselhável a modificação de arquivos, pois acaba gerando muita sobrecarga.

Inteligência Artificial

INCIDÊNCIA EM PROVA: BAIXA

Galera, um dos problemas do Big Data é que ele é... muito grande! No passado, as pessoas tentavam evitar formatos como imagens, vídeo ou voz porque não podiam fazer muita coisa com eles e seu custo de armazenamento era alto. **Hoje em dia, esse custo foi reduzido substancialmente e já existem tecnologias capazes de manipular uma quantidade absurda com eficiência.** *Do que você está falando, Diego?*

Atualmente, tem sido cada vez mais comum a vigilância por vídeo em todos os lugares. Pensem em 100 câmeras operando 24 horas por dia, 7 dias por semana, 365 dias por ano. Isso resulta em um total de 2400 horas de vídeo por dia. Se um ser humano fosse revisar esses dados em busca de eventuais atividades suspeitas, por exemplo, seria necessária uma equipe de 60 pessoas – e isso simplesmente não vale a pena economicamente.

É nesse ponto que a Inteligência Artificial e o Big Data trabalham juntos! **Uma maneira de lidar de forma eficiente com essa quantidade de dados é gerenciá-los com uma varredura de dados e utilizar algoritmos de software de Inteligência Artificial.** *Vocês se lembram do Big Data Analytics?* Pois é, ele comumente utiliza ferramentas de Inteligência Artificial para ajudar a analisar e compreender uma quantidade massiva de dados.

Vejam como eles se complementam bem: Big Data lida com uma quantidade absurda de dados! *Agora adivinhem quem se dá super bem quando possui uma quantidade absurda de dados?* A Inteligência Artificial! Quanto mais dados ela possuir, mais "inteligente" será! **Essa combinação está ajudando as organizações a entenderem seus clientes muito melhor – até mesmo de maneiras que eram impossíveis no passado.**

O Big Data, por si só, é inútil sem uma ferramenta para analisar os dados e humanos não conseguem fazer isso de forma eficiente. A Inteligência Artificial pode ser extremamente útil para detectar anomalias, para calcular probabilidades de sucessos, para reconhecimento de padrões, para reconhecimento de imagens, para reconhecimento de palavras (discursos), para as tecnologias de carros autônomos, entre outros.

Alguns autores dividem as possíveis aplicações de Inteligência Artificial em três grupos: **(1) Ciência Cognitiva:** sistemas especialistas, lógica difusa, algoritmos genéticos e redes neurais; **(2) Robótica:** percepção visual, locomoção, condução, tatilidade; **(3) Interfaces Naturais:** linguagens naturais, reconhecimento de discurso, interfaces multissensoriais e realidade virtual. No entanto, eu gostaria de falar sobre uma aplicação muito comum em órgãos públicos e empresas atualmente: Chatbots!

O ChatBot é um programa de computador que tenta simular um ser humano na conversação com as pessoas. O objetivo é responder as perguntas de tal forma que as pessoas tenham a impressão de estar conversando com outra pessoa e não com um programa de computador. Após o envio de perguntas em linguagem natural, o programa consulta uma base de conhecimento e em seguida fornece uma resposta que tenta imitar o comportamento humano.

Com toda certeza, vocês já foram atendidos por um robô quando precisavam de alguma informação específica. O Bradesco – por exemplo – possui uma assistente virtual chamada Bia para auxiliar correntistas com o aplicativo do banco. **O Tesouro Nacional – órgão em que trabalho – possui uma assistente virtual chamada Jaque que responde sobre informações contábeis e fiscais de municípios.** E isso tem sido cada vez mais comum em órgãos públicos...

A inteligência artificial permitiu que a Jaque realizasse mais de 2000 interações mensais em 2020. **Ao mesmo tempo, a equipe do Tesouro Nacional que apoiava essa missão, foi reduzida em 50%, permitindo um deslocamento para atuação em papeis mais analíticos e de atendimento mais especializado.** Eu sei o que vocês estão pensando: vários desses robôs ainda não satisfazem as necessidades dos usuários.

Em suma, a Inteligência artificial é uma nova disciplina técnica que pesquisa e desenvolve teorias, métodos, tecnologias e sistemas de aplicação para simular a extensão e expansão da inteligência humana. **O objetivo da pesquisa de inteligência artificial é permitir que as máquinas realizem algumas tarefas complexas que requerem atualmente humanos inteligentes para que sejam concluídas.**

Em outras palavras, esperamos que a máquina possa nos substituir para resolver algumas tarefas complicadas. **Não apenas atividades mecânicas repetitivas, mas algumas atividades que requerem conhecimento humano para que sejam concluídas com sucesso.** A intersecção entre Big Data e Inteligência Artificial é considerada uma revolução capaz de moldar o futuro de como as empresas agregam valor aos negócios a partir de seus dados e recursos analíticos.

(SLU/DF – 2019) O serviço de chatbot, um sistema que permite às grandes corporações oferecer um canal direto com o consumidor, é um dos exemplos tecnológicos utilizado no atendimento ao público, tornando a comunicação entre empresa e cliente mais próxima e personalizada, graças aos avanços da inteligência artificial.

Comentários: com o avanço da tecnologia, as organizações têm buscado formas de otimizar custos e potencializar seus serviços de atendimento ao público. Uma das formas mais modernas e que se utiliza dos atributos da Inteligência Artificial é o chatbot, que de forma robotizada permite uma interação em tempo real e o atendimento das principais demandas com maior rapidez (Correto).

(UECE/CEV – 2019) Atualmente utilizado por diversas empresas e tido por alguns como um vilão que compromete vagas no mercado de trabalho, esse avanço da tecnologia se caracteriza como a capacidade do sistema para interpretar, aprender e utilizar dados externos, com o objetivo de executar tarefas que, se um humano executasse, seriam consideradas inteligentes. Essa descrição se refere:

a) à inteligência artificial.
b) à terceirização de serviços.

c) ao telemarketing.
d) ao atendimento personalizado.

_____

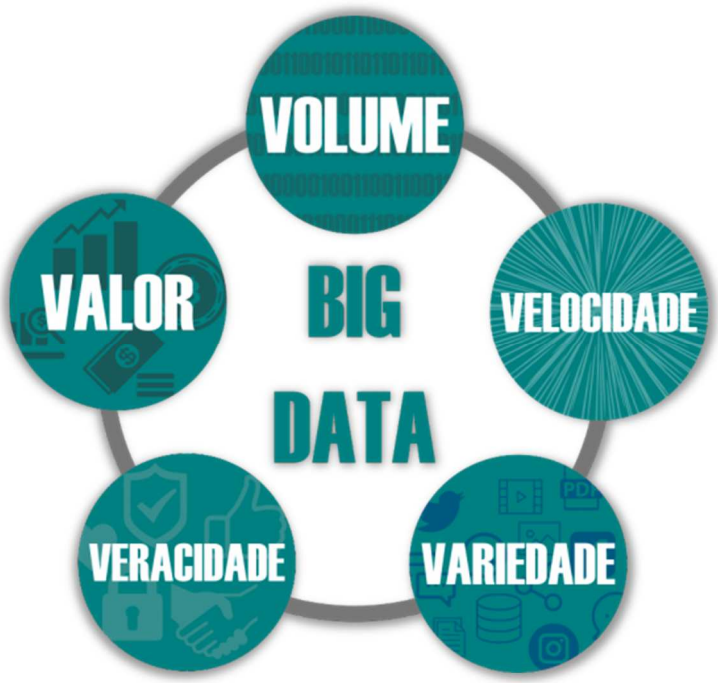
Comentários: a tecnologia que se caracteriza como a capacidade do sistema para interpretar, aprender e utilizar dados externos, com o objetivo de executar tarefas que, se um humano executasse, seriam consideradas inteligentes é a Inteligência Artificial (Letra A).

RESUMO

OXFORD ENGLISH DICTIONARY	Big Data é um dado de grande tamanho, tipicamente ao nível que sua manipulação e gerenciamento apresenta desafios significativos a logística.
DUMBILL E EDD	Big Data é o dado que excede a capacidade de processamento convencional dos sistemas de bancos de dados.
MAYER-SCHÖNBERGER E CUKIER'S	Big Data é a habilidade da sociedade de aproveitar a informação por novas maneiras para produzir introspecção úteis ou bens e serviços de valor significante.
INTERNATIONAL DATA CORPORATION	Big Data é uma nova geração de tecnologias e arquiteturas, projetadas economicamente para extrair valor de volumes muito grandes e vastos de dados, permitindo alta velocidade de captura, descoberta e análise.
KIM, TRIMI E JI-HYONG	Big Data é o termo geral para a enorme quantidade de dados digitais coletados a partir de todos os tipos de fontes.
MAHRT E SCHARKOW	Big Data denota um maior conjunto de dados ao longo do tempo, conjunto de dados estes que são grandes demais para serem manipulados por infraestruturas de armazenamento e processamento regulares.
DAVENPORT E KWON	Big Data são dados demasiadamente volumosos ou muito desestruturados para serem gerenciados e analisados através de meios tradicionais.
DI MARTINO	Big Data se refere ao conjunto de dados cujo tamanho está além da habilidade de ferramentas típicas de banco de dados em capturer, gerenciar e analisar.
RAJESH	Big Data são conjuntos de dados que são tão grandes que se tornam difíceis de trabalhar com o uso de ferramentas atualmente disponíveis.

TIPOS DE DADOS	DESCRIÇÃO
DADOS ESTRUTURADOS	São dados que podem ser armazenados, acessados e processados em formato fixo e padronizado de acordo com alguma regra específica. Esta organização é geralmente feita por colunas e linhas (semelhante a planilhas do Excel), mas pode variar de acordo com a fonte de dados. Exemplo: Planilhas Eletrônicas, Bancos de Dados Relacionais e CSV.
DADOS SEMI-ESTRUTURADOS	São dados estruturados que não estão de acordo com a estrutura formal dos modelos de dados como em tabelas, mas que possuem marcadores para separar elementos semânticos e impor hierarquias de registros e campos dentro dos dados Exemplo: Dados de E-mail, Arquivos XML, Arquivos JSON e Banco de Dados NoSQL.
DADOS NÃO-ESTRUTURADOS	São dados que apresentam formato ou estrutura desconhecidos, em que não se sabe extrair de forma simples os valores desses dados em forma bruta. Exemplo: Documentos, Imagens, Vídeos, Arquivos de Texto, Posts em Redes Sociais.

DATA WAREHOUSE	DATA LAKE
Dados geralmente são tratados (limpos, combinados, organizados, etc) antes de serem armazenados.	Dados geralmente são armazenados da maneira que foram capturados – brutos, sem nenhum tratamento.
Podem armazenar todos os tipos de dados, mas o foco é nos dados estruturados.	Armazenam dados estruturados, semi-estruturados e não-estruturados.
Ideal para usuários operacionais visto que as ferramentas analíticas são mais fáceis de usar.	Ideal para cientistas de dados visto que as ferramentas analíticas são mais difíceis de usar.
Armazenamento de dados custam geralmente mais caro e consome mais tempo.	Armazenamento de dados custam geralmente mais barato e consome menos tempo.
Um esquema é definido antes dos dados serem armazenados.	Um esquema é definido após os dados serem armazenados.
Armazenam um grande volume de dados.	Armazenam um gigantesco volume de dados.



PALAVRAS-CHAVE				
VOLUME	VELOCIDADE	VALOR	VERACIDADE	VARIEDADE
Terabytes	Transmissão	Estatístico	Confiabilidade	Estruturado
Registros	Tempo Real	Hipóteses	Autenticidade	Não-Estruturado
Tabelas/Arquivos	Processos	Correlações	Origem/Reputação	Múltiplas Fontes

PREMISSAS	DESCRIÇÃO
VOLUME	Corresponde à grande quantidade de dados acumulada.
VELOCIDADE	Corresponde à rapidez na geração e obtenção de dados.
VARIEDADE	Corresponde à grande diversidade de tipos ou formas de dados.
VERACIDADE	Corresponde à confiança na geração e obtenção dos dados.

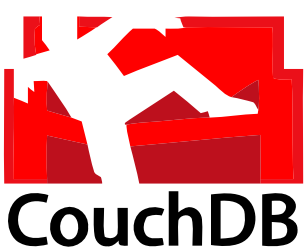
VALOR	Corresponde à utilidade e valor agregado ao negócio.

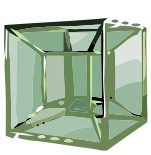
TIPO DE ANÁLISE	QUESTÃO?	DESCRIÇÃO
ANÁLISE DESCRITIVA	O QUE ACONTECEU?	Em vez de se focar no futuro, busca fazer uma fotografia do presente, para que decisões de cunho imediato possam ser tomadas com segurança. Ela trabalha com histórico de dados, cruzando informações com o objetivo de gerar um panorama claro e preciso dos temas relevantes para a empresa no presente momento. Exemplo: por meio do cruzamento de dados, conclui-se que determinada pessoa atualmente é identificada como má pagadora.
ANÁLISE DIAGNÓSTICA	POR QUE ACONTECEU?	O foco está na relação de causas e consequências percebidas ao longo do tempo, dentro de um determinado tema. Assim, a análise diagnóstica funciona baseada na coleta de dados relacionados a um determinado assunto, cruzando informações com o objetivo de entender quais fatores influenciaram o resultado atual. Exemplo: determinada pessoa nunca havia sido identificada como má pagadora – somente é agora porque ficou viúva recentemente.
ANÁLISE PREDITIVA	O QUE IRÁ ACONTECER?	Este tipo de análise é o mais indicado para quem precisa prever algum tipo de comportamento ou resultado. Esta técnica busca analisar dados relevantes ao longo do tempo, buscando padrões comportamentais e suas variações de acordo com cada contexto, a fim de prever como será o comportamento de seu público ou mercado no futuro, dadas as condições atuais. Exemplo: quanto estará o valor do dólar no ano que vem?
ANÁLISE PRESCRITIVA	O QUE DEVO FAZER?	Segue um modelo similar à Análise Preditiva, no entanto com objetivos ligeiramente diferentes. Em vez de tentar prever um determinado acontecimento, esta análise busca prever as consequências deste acontecimento. Exemplo: dado um aumento do valor do dólar no ano que vem, como isso poderá afetar as importações de matéria prima, consequentemente, o faturamento das vendas de determinada empresa.

BIG DATA ANALYTICS É:	BIG DATA ANALYTICS NÃO É:
Uma estratégia baseada em tecnologia que permite coletar insights mais profundos e relevantes de clientes, parceiros e negócio, ganhando assim uma vantagem competitiva.	Somente tecnologia – no nível empresarial, refere-se a explorar fontes amplamente melhoradas de dados para adquirir insights.
Trabalhar com conjuntos de dados cujo porte e variedade estão além da habilidade de captura, armazenamento e análise de softwares de banco de dados típicos.	Somente volume – também se refere à variedade e à velocidade, mas – talvez mais importante – refere-se ao valor derivado dos dados.
Processamento de um fluxo contínuo de dados em tempo real, possibilitando a tomada de decisões sensíveis ao tempo mais rápido do que em qualquer outra época.	Mais gerada ou mais utilizada somente por grandes empresas online como Google ou Amazon. Embora as empresas de internet possam ter sido pioneiras no Big Data na escala web, aplicativos chegam a todas as indústrias.
Distribuído na natureza, isto é, o processamento de análise vai aonde estão os dados para maior velocidade e eficiência.	Uso de bancos de dados relacionais tradicionais de "tamanho único" criados com base em disco compartilhado e arquitetura de memória. Análise de Big

	Data usa uma rede de recursos de computação para processamento massivamente paralelo e escalável.
Um novo paradigma no qual a tecnologia da informação colabora com usuários empresariais e "cientistas de dados" para identificar e implementar análises que ampliam a eficiência operacional e resolvem novos problemas empresariais.	Um substituto de bancos de dados relacionais – dados estruturados continuam a ser de importância crítica para as empresas. No entanto, sistemas tradicionais podem não ter capacidade de manipular as novas fontes e contextos do Big Data.
Transferir a tomada de decisão dentro da empresa e permitir que pessoas tomem decisões melhores, mais rápidas e em tempo real.	-

CRITÉRIO	NOSQL	SQL
MODELO	Não-Relacional	Relacional
ARMAZENAMENTO	Variados (Grafos, Documentos, etc)	Tabelas
FLEXIBILIDADE	Alta flexibilidade (Esquema indefinido)	Baixa flexibilidade (Esquema definido)
ADEQUAÇÃO	Mais adequado a dados não-estruturados	Mais adequado a dados estruturados
ESCALABILIDADE	Em geral, escalabilidade horizontal	Em geral, escalabilidade vertical
SGBD	MongoDB, Cassandra, HBase, Neo4J, etc	Oracle, MySQL, DB2, SQL Server, etc

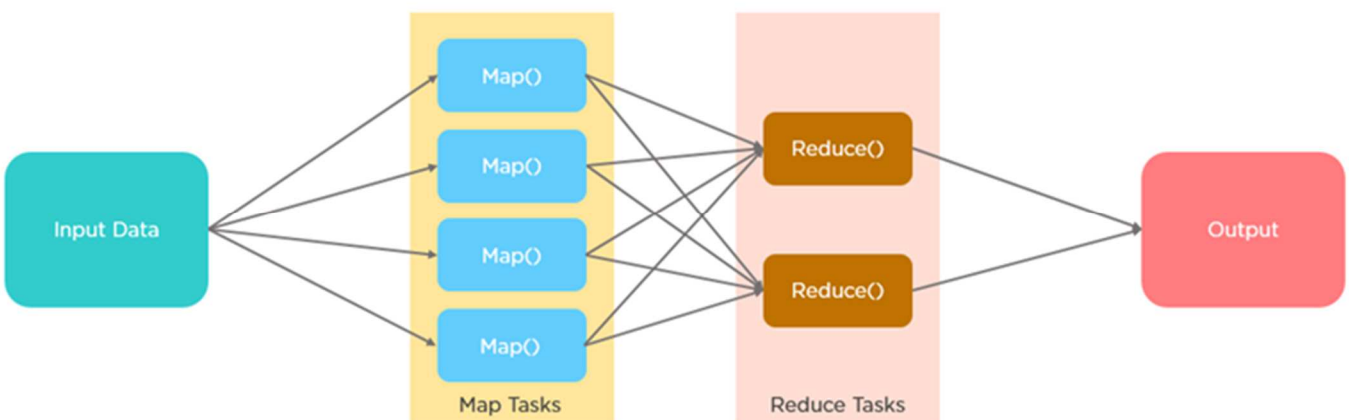





TIPO DE MODELOS	DESCRIÇÃO
ORIENTADO A CHAVE-VALOR	Esse modelo armazena dados por meio de uma estrutura de mapeamento ou dicionário, em que todo dado armazenado possui uma chave identificadora e seu valor em si. Para cada chave de entrada, é retornado um valor de saída (Ex: Table Storage, DynamoDB e Redis).
ORIENTADO A DOCUMENTOS	Esse modelo armazena dados na forma de documentos flexíveis, semiestruturados e hierárquicos junto com seus metadados sem uma estrutura definida. Em geral, os dados são armazenados em formato JSON ou XML (Ex: MongoDB, CouchDB e DocumentDB).

ORIENTADO A GRAFOS	Esse modelo armazena o relacionamento entre dados altamente conectados por meio de vértices e arestas. São geralmente utilizados em redes sociais, mecanismos de recomendação e detecção de fraudes (Ex: Neo4J, Infinite Graph e ArangoDB).
ORIENTADO A COLUNAS	Esse modelo armazena dados em colunas dinâmicas. É o mais semelhante ao modelo relacional, mas os dados são armazenados em colunas em vez de linhas. Ademais, cada coluna pode conter subcolunas, que podem conter várias propriedades (Ex: Cassandra, Hypertable e MonetDB).

MAPREDUCE	Modelo de programação que permite reduzir problemas grandes em problemas menores, mapeando cada subproblema para máquinas diferentes (ou processadores diferentes de uma mesma máquina) e, em seguida, reduzindo cada resposta intermediária à única resposta final que você está procurando.
APACHE HADOOP	Plataforma, framework e ecossistema de computação distribuída orientada a clusters, voltado para armazenamento e processamento de grandes volumes de dados, com alta escalabilidade, grande confiabilidade e tolerância a falhas. Um de seus subprojetos implementa a técnica de MapReduce.



QUESTÕES COMENTADAS – CESPE

1. **(CESPE / SEFAZ-SE – 2022)** Com relação a noções de big data, julgue os itens que se seguem.

   I Como qualquer tecnologia, soluções de big data também apresentam algumas restrições. Por exemplo, elas não podem ser utilizadas na área da saúde para determinar a causa de uma doença, porque esse é um procedimento complexo que somente pode ser executado por pessoas devidamente capacitadas — nesse caso, os médicos.

   II Big data é qualquer tipo de fonte de dados que possui, no mínimo, as seguintes três características: volume de dados extremamente grande; velocidade de dados extremamente alta; e variedade de dados extremamente ampla.

   III Para que as organizações obtenham os conhecimentos corretos, a tecnologia big data não permite que elas executem as operações de armazenar e administrar as grandes quantidades de dados de si próprias.

   IV Big data é uma combinação de tecnologias de gestão de dados que evoluíram ao longo dos anos, razão por que não é considerado um mercado único.

   Estão certos apenas os itens:

   a) I e III
   b) I e IV
   c) II e IV
   d) II e V
   e) III e V

Comentários:

(I) Errado. Que viagem! Big Data não só pode como já é muito utilizado na área de saúde; (II) Correto. Há versões que consideram tanto 3V's quanto 5V's; (III) Errado. Não há restrições para que organizações executem operações de armazenamento e administração; (IV) Correto. Atualmente é considerado um grande conjunto de tecnologias que envolve diversas áreas de conhecimento.

Gabarito: Letra C

---

2. **(CESPE / PETROBRAS – 2022)** Em sistemas NoSQL baseados em armazenamento de chavevalor, a chave é multidimensional e composta pela combinação do nome de tabela com a chave linha-coluna e com o rótulo de data e hora.

Comentários:

No modelo chave-valor, a chave é bidimensional e composta pela combinação de um identificador alfanumérico único (chave) e um valor associado em uma tabela (valor). Em outras palavras, esse modelo armazena dados por meio de uma estrutura de mapeamento ou dicionário, em que todo dado armazenado possui uma chave identificadora e seu valor em si – para cada chave de entrada, é retornado um valor de saída.

Gabarito: Errado

---

3. **(CESPE / ISS-Aracaju – 2021)** Big data ajudou a sedimentar o cargo de cientista de dados. Entre as funções desse cargo inclui-se:

a) a modelagem estruturada.
b) a análise retrospectiva.
c) a modelagem não estruturada.
d) a modelagem relacional.
e) o processamento comparativo.

Comentários:

Questão bizarra! Todos os itens podem ser executados por um cientista de dados. *Ele não pode fazer uma modelagem relacional?* Claro que pode! De toda forma, essa questão foi retirada do livro do Taurion (2013), em que ele diferencia Analista de BI e Cientista de Dados.

ANALISTA DE BI	CIENTISTA DE DADOS
Cognos, Modelo Relacional, Banco de Dados SQLServer, Oracle, DB2.	Hadoop, Modelos Relacionais e NoSQL, bancos e dados não relacionais e in-memory.
Modelagem Relacional/Estruturada.	Inclui também modelagem não estruturada. Modelagem analítica é essencial.
Desenvolve queries estruturadas sobre dados passados.	Cria perguntas e busca relacionamentos entre fatos aparentemente desconexos.

Em primeiro lugar, o autor não é uma unanimidade; em segundo lugar, o autor não disse que essas eram atividades exclusivas de cada papel. Logo, não se trata de atividades taxativas!

Gabarito: Letra C

---

4. **(CESPE / SERPRO – 2021)** MapReduce divide o conjunto de dados de entrada em blocos independentes que são processados pelas tarefas de mapa de uma maneira completamente paralela. Essa estrutura classifica as saídas dos mapas, as quais são, então, inseridas nas tarefas de redução.

Comentários:

Perfeito! O MapReduce é um modelo de programação que permite reduzir problemas grandes em problemas menores, mapeando cada subproblema para máquinas diferentes (ou processadores diferentes de uma mesma máquina) e, em seguida, reduzindo cada resposta intermediária à única resposta final que você está procurando.

Gabarito: Correto

---

5. **(CESPE / SERPRO – 2021)** No que se refere aos três Vs do Big Data, o termo volume refere-se a dados que, atualmente, não são estruturados nem armazenados em tabelas relacionais, o que torna sua análise mais complexa.

Comentários:

Na verdade, referem-se à quantidade de dados de quaisquer tipos (estruturados, semiestruturados ou não estruturados) armazenados em fontes de estruturas diversas.

Gabarito: Errado

---

6. **(CESPE / SERPRO – 2021)** Big data caracteriza-se, principalmente, por volume, variedade e velocidade, o que se justifica devido ao fato de os dados serem provenientes de sistemas estruturados, que são maioria, e de sistemas não estruturados, os quais, embora ainda sejam minoria, vêm, ao longo dos anos, crescendo consideravelmente.

Comentários:

Pelo contrário, os dados provenientes de sistemas não estruturados são maioria e os dados provenientes de sistemas estruturados são minoria.

Gabarito: Errado

---

7. **(CESPE / SERPRO – 2021)** MapReduce divide o conjunto de dados de entrada em blocos independentes que são processados pelas tarefas de mapa de uma maneira completamente paralela. Essa estrutura classifica as saídas dos mapas, as quais são, então, inseridas nas tarefas de redução.

Comentários:

Perfeito! O MapReduce é um modelo de programação que permite reduzir problemas grandes em problemas menores, mapeando cada subproblema para máquinas diferentes (ou processadores diferentes de uma mesma máquina) e, em seguida, reduzindo cada resposta intermediária à única resposta final que você está procurando. O processo de decomposição dos dados é chamado de Mapeamento/Mapa (Map); e o processo de consolidação do resultado dos mapeamentos é chamado de Redução (Reduce). De fato, as saídas dos mapas são inseridas como entradas na etapa

de redução, sendo essa responsável por consolidar os resultados de cada mapa, gerando um resultado agregado.

Gabarito: Correto

---

8. **(CESPE / SERPRO – 2021)** Uma das principais características de big data é que seu custo de armazenamento de dados é relativamente baixo se comparado a um data warehouse.

Comentários:

Pode parecer contraintuitivo, mas o Data Warehouse possui um custo de armazenamento realmente maior em relação ao Big Data – esse é mais otimizado para armazenar gigantescos volumes de dados do que aquele.

Gabarito: Correto

---

9. **(CESPE / SERPRO – 2021)** O Hadoop consiste em um único produto, ou seja, um software monolítico, que possibilita análise de logs e outros dados da Web.

Comentários:

Na verdade, ele é mais do que um software – ele é uma plataforma, um framework, um ecossistema de computação distribuída orientada a clusters e voltado para armazenamento e processamento de grandes volumes de dados, com alta escalabilidade, grande confiabilidade e tolerância a falhas.

Gabarito: Errado

---

10. **(CESPE / SERPRO – 2021)** Apesar de ser uma tecnologia de código aberto disponibilizada pela ASF (Apache Software Foundation), o Hadoop também é oferecido por distribuidores comerciais, de maneira que fornecedores oferecem distribuições específicas que incluem não só ferramentas administrativas adicionais, mas também suporte técnico.

Comentários:

Perfeito! Isso é bastante comum no mundo da tecnologia da informação. A Apache Software Foundation (ASF) é uma organização sem fins lucrativos criada para suportar projetos de código aberto. O Hadoop é oferecido por ela, mas também pode ser distribuída por outros fornecedores comerciais que incluem também ferramentas administrativas – além de suporte técnico.

Gabarito: Correto

---

11. (CESPE / TCE-RJ – 2021) Em Big Data, a premissa volume refere-se à capacidade de processar, em um ambiente computacional, diferentes tipos e formatos de dados, como fotos, vídeos e geolocalização.

Comentários:

Na verdade, essa é a premissa de variedade e, não, volume.

Gabarito: Errado

---

12. (CESPE / TCE-RJ – 2021) Volume, variedade e visualização são as três características, conhecidas como 3 Vs, utilizadas para definir Big Data.

Comentários:

Visualização não é uma característica do Big Data!

Gabarito: Errado

---

13. (CESPE / Polícia Federal – 2021) As aplicações de *bigdata* caracterizam-se exclusivamente pelo grande volume de dados armazenados em tabelas relacionais.

Comentários:

Opa... Volume é apenas um dos cinco V's do Big Data! Ele também se caracteriza pela Veracidade, Velocidade, Valor e Variedade. Além disso, os dados são armazenados em bancos de dados não relacionais.

Gabarito: Errado

---

14. (CESPE / PRF – 2021) A Internet das Coisas (IoT) aumenta a quantidade e a complexidade dos dados por meio das novas formas e novas fontes de informações, influenciando diretamente em uma ou mais das características do Big Data, a exemplo de volume, velocidade e variedade.

Comentários:

Internet das Coisas é a tecnologia que permite a interconexão digital de objetos cotidianos com a Internet. Logo, faz sentido que ela aumente a quantidade e a complexidade dos dados por meio de novas formas e fontes de informações? Claro! Galera... quando houver a popularização da tecnologia 5G, teremos objetos trocando informações via internet a todo instante em todo mundo. Vamos ver um exemplo:

Em um futuro próximo, meu cachorro (@chico.golden.gff) poderá ter um chip em sua coleira que envia informações de geolocalização para o meu smartphone (Ex: iPhone); meu smartphone poderá enviar dados para o meu smartwatch (Apple Watch); meu smartwatch poderá se comunicar com a caixa de som (Ex: JBL); a caixa de som poderá se comunicar com um assistente virtual (Ex: Alexa); minha assistente virtual poderá se comunicar com a geladeira (Ex: Electrolux); a geladeira poderá identificar que um determinado produto está acabando e se comunicar com a página de supermercado (Ex: Pão de Açúcar); e assim por diante...

Agora percebam: a quantidade de informações produzidas no mundo se multiplicará de forma avassaladora. Será que isso influencia as características de volume, velocidade e variedade do Big Data? Claro! Uma quantidade absurdamente monstruosa de novos dados será gerada (Volume) e processados a todo instante (Velocidade) provenientes de diversas fontes e formas diferentes (Variedade).

Gabarito: Correto

---

15. (CESPE / AL-AP – 2020) Atualmente, diversos dados são coletados pelos sistemas digitais de empresas na internet para constituir Big Data com conteúdo sobre os resultados alcançados por seus produtos e serviços, prestígio da imagem da organização e seus representantes. Porém, parte desses dados pode ser falsa ou manipulada por internautas. O tratamento dos dados, a fim de qualificá-los antes de disponibilizá-los para a tomada de decisão na empresa, segundo o conceito das cinco dimensões "V" de avaliação de um Big Data, se refere:

a) ao valor.
b) à variedade.
c) à veracidade.
d) à velocidade.
e) ao volume.

Comentários:

Os dados podem ser falsos ou manipulados por internautas? Os dados devem ser qualificados antes de serem disponibilizados? A questão deu a dica para inferir que se trata da dimensão de veracidade, isto é, a capacidade de selecionar dados que sejam úteis e verídicos

Gabarito: Letra C

---

16. (CESPE / TCE-RO – 2019) Com relação a fundamentos e conceitos de Big Data, julgue os itens a seguir.

I. O volume de dados é uma característica importante de Big Data.

II. Em Big Data, a qualidade do dado não tem importância, porque a transformação dos dados não impacta os negócios.

III. A característica de velocidade de entrada dos dados impacta o modelo de processamento e armazenamento.

IV. A variedade dos dados não é característica intrínseca nos fundamentos de Big Data.

Estão certos apenas os itens:

a) I e II.
b) I e III.
c) II e IV.
d) I, III e IV.
e) II, III e IV.

Comentários:

(I) Correto, trata-se de uma das premissas do Big Data; (II) Errado, a qualidade é importante – volume de dados em qualidade não terá utilidade e, portanto, não agregará valor ao negócio; (III) Correto, trata-se de uma das premissas do Big Data – a velocidade de input de dados é extremamente relevante para o modelo de processamento e armazenamento utilizado. Imagine que dados sejam inseridos a uma velocidade que o seu modelo não consiga processar ou armazenar; (IV) Errado, trata-se de uma das premissas do Big Data.

Gabarito: Letra B

---

17. (CESPE / TCM-BA – 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 proveem 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.

Comentários:

(a) Errado. *Como é?* Big Data tem sido revolucionário em muitas organizações! Big Data tem modificado processos de gestão e cultura de diversas organizações por meio da descoberta de novas informações para tomadas de decisão estratégicas;

(b) Errado. *Sabem quando a Netflix te recomenda um filme ou o Spotify te recomenda uma música/banda?* Pois é, isso ocorre por meio de um Sistema de Recomendação! Ela combina várias técnicas computacionais para selecionar itens personalizados com base nos interesses dos usuários. No entanto, não existe nenhuma obrigação dos sistemas serem estruturados em um cluster e usarem computação distribuída. *O que é isso, Diego?* Cara, é basicamente uma forma de dividir um problema em várias partes para serem processados paralelamente por uma grande quantidade de computadores e posteriormente combinar o resultado;

(c) Correto. *Professor, o que é Data Scraping?* Basicamente é uma técnica que coleta dados de bases de dados complexas (com muitas tabelas e relacionamentos) ou de sites de difícil navegabilidade, processa automaticamente esses dados e os exibe de uma forma mais legível – ele transforma dados pouco estruturados em dados mais estruturados e fáceis de manipular. Grosso modo, podemos imaginar uma biblioteca em que o bibliotecário organizou todos os livros de acordo com tema, tamanho, data, número de identificação, etc – isso é bem estruturado; agora imagine uma biblioteca sem um bibliotecário em que as coisas são organizados de forma aleatória – isso não é bem estruturado;

(d) Errado. Todas essas ações são realmente inerentes ao processo de preparação de dados (*Data Preparation*), exceto a mineração de dados! *Por que?* Porque a mineração ocorre após a preparação dos dados! Não há como minerar dados (explorar em busca de padrões e informações úteis) sem antes preparar esses dados. Além disso, não é obrigatória a preparação dos dados;

(e) Errado. Virtuosidade e Vantagem não fazem parte dos 5 Vs do Big Data.

Gabarito: Letra C

---

18. (CESPE / TCM-BA – 2018) Um dos desdobramentos de *big data* é o *big data analytics*, que se refere aos softwares capazes de tratar dados para transformá-los em informações úteis às organizações. O *big data analytics* difere do *business intelligence* por:

a) priorizar o ambiente de negócios em detrimento de outras áreas.
b) analisar dúvidas já conhecidas para as quais se deseje obter resposta.
c) analisar o que já existe e o que está por vir, apontando novos caminhos.
d) dar enfoque à coleta, à transformação e à disponibilização dos dados.
e) analisar o que já existe, definindo as melhores hipóteses.

Comentários:

O Business Intelligence trata das perguntas conhecidas e das nossas pré-concepções com relação aos dados. Ao passo que Big Data Analytics se envolve com um universo de novas possibilidades e perguntas que ainda não conhecemos, analisando o que já existe e o que está por vir, apontando novos caminhos.

Gabarito: Letra C

---

19. (CESPE / Polícia Federal – 2018) Em um *big data*, alimentado com os dados de um sítio de comércio eletrônico, são armazenadas informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos. Tendo como referência as informações apresentadas, julgue o item seguinte.

O *big data* consiste de um grande depósito de dados estruturados, ao passo que os dados não estruturados são considerados data files.

Comentários:

O Big Data consiste no gerenciamento e na análise de dados que vão além dos dados tipicamente estruturados. A questão afirma que dados não estruturados são considerados Data Files (que são arquivos de dados), no entanto dados não estruturados podem ser de absolutamente qualquer tipo, como vídeo digital, imagens, dados de sensores, arquivos de logs, entre outros. *Vocês se lembram do V de Variedade?* Pois é, os formatos são variados em um Big Data!

Gabarito: Errado

---

20. (CESPE / Polícia Federal – 2018) Em um *big data*, alimentado com os dados de um sítio de comércio eletrônico, são armazenadas informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos. Tendo como referência as informações apresentadas, julgue o item seguinte.

Dados coletados de redes sociais podem ser armazenados, correlacionados e expostos com o uso de análises preditivas.

Comentários:

Redes Sociais realmente geram uma grande quantidade de dados diariamente e esses dados podem ser muito úteis a um Big Data. Quando armazenados, podem facilmente ser correlacionados e expostos por meio do uso de análises preditivas – tudo perfeito na questão.

Gabarito: Correto

---

21. (CESPE / Polícia Federal – 2018) Em um *big data*, alimentado com os dados de um sítio de comércio eletrônico, são armazenadas informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos. Tendo como referência as informações apresentadas, julgue o item seguinte.

Uma aplicação que reconheça o acesso de um usuário e forneça sugestões diferentes para cada tipo de usuário pode ser considerada uma aplicação que usa *machine learning*.

Comentários:

Uma aplicação que reconheça o acesso de um usuário e forneça sugestões diferentes para cada tipo de usuário, em geral, utiliza sistemas de recomendações que – por sua vez – utilizam algoritmos de Machine Learning para que a aplicação consiga aprender com o contexto e fazer sugestões mais certeiras – é o velho exemplo do Netflix!

Gabarito: Correto

---

22. (CESPE / ABIN – 2018) O registro e a análise de conjuntos de dados referentes a eventos de segurança da informação são úteis para a identificação de anomalias; esse tipo de recurso pode ser provido com uma solução de *big data*.

Comentários:

O Big Data é capaz de registrar e analisar dados estruturados ou não-estruturados de diversos contextos diferentes para descobrir informações de valor para o negócio de uma organização. Dentre eles, eventos de segurança da informação para identificação de anomalias é – sim – uma possibilidade.

Gabarito: Correto

---

23. (CESPE / IPHAN – 2018) A utilização de tecnologias emergentes para o tratamento de grandes volumes de dados (*big data*) pode contribuir para o sucesso da implantação da estratégia de governança digital.

Comentários:

A governança digital define objetivos estratégicos, metas, indicadores e iniciativas de uma determinada organização. Uma vez que os dados obtidos através do Big Data são tratados de alguma maneira, eles podem contribuir para a obtenção de informações relevantes para a tomada de decisões estratégicas.

Gabarito: Correto

---

24. (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.

Comentários:

Em vez de tentar prever um determinado acontecimento, a Análise Prescritiva busca prever as consequências deste acontecimento. Exemplo: dado um aumento do valor do dólar no ano que vem, como isso poderá afetar as importações de matéria prima, consequentemente, o faturamento das vendas de determinada empresa.

Gabarito: Letra C

---

25. (CESPE / Polícia Federal – 2018) MapReduce oferece um modelo de programação com processamento por meio de uma combinação entre chaves e valores.

Comentários:

Perfeito, perfeito, perfeito! Excelente definição de MapReduce!

Gabarito: Correto

---

26. (CESPE / TCE-PB – 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.

Comentários:

(a) Errado, ele possui três premissas básicas: Volume, Velocidade e também Variedade – sendo que alguns autores também consideram Veracidade e Valor; (b) Errado, o algoritmo Dijkstra é utilizado para encontrar o menor caminho entre dois nós numa árvore – não vejo relação com o HDSF; (c) Errado, 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; (d) Errado, ele não utiliza bancos SQL nativos – em geral, ele utiliza bases de dados não relativas a modelos relacionais; (e) Correto, ele realmente é considerado um modelo de programação que permite o processamento de dados massivos em um algoritmo paralelo e distribuído – definição impecável.

Gabarito: Letra E

---

27. (CESPE / TCE-MG – 2018) Um dos desdobramentos de big data é o Big Data Analytics, que se refere aos softwares capazes de tratar dados para transformá-los em informações úteis às organizações. O Big Data Analytics difere do Business Intelligence por:

a) priorizar o ambiente de negócios em detrimento de outras áreas.
b) analisar dúvidas já conhecidas para as quais se deseje obter resposta.
c) analisar o que já existe e o que está por vir, apontando novos caminhos.
d) dar enfoque à coleta, à transformação e à disponibilização dos dados.
e) analisar o que já existe, definindo as melhores hipóteses.

Comentários:

O Business Intelligence trata de encontrar respostas que explicam o passado, já o Big Data Analytics trata de encontrar as perguntas que explicam o futuro. Ambos possuem grande importância, complementam-se e devem ser bem entendidos para que as empresas possam aproveitá-los da melhor forma, agregando e alcançando os valores e resultados desejados aos negócios. Dessa forma, podemos concluir que o Big Data Analytics difere do Business Intelligence por analisar o que já existe e o que está por vir, apontando novos caminhos.

Gabarito: Letra C

---

28. (CESPE / Polícia Federal – 2018) A mineração de dados se caracteriza especialmente pela busca de informações em grandes volumes de dados, tanto estruturados quanto não estruturados, alicerçados no conceito dos 4V's: volume de mineração, variedade de algoritmos, velocidade de aprendizado e veracidade dos padrões.

Comentários:

Opa... a questão trata de Big Data e, não, de Mineração de Dados! Além disso, são 5V's e todos eles tratam de dados: volume de dados, variedade de dados, velocidade de dados, veracidade dos dados e valor dos dados.

Gabarito: Errado

Em um big data, alimentado com os dados de um sítio de comércio eletrônico, são armazenadas informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos.

Tendo como referência as informações apresentadas, julgue o item seguinte.

29. **(CESPE / Polícia Federal – 2018)** Pelo monitoramento do tráfego de rede no acesso ao sítio em questão, uma aplicação que utiliza machine learning é capaz de identificar, por exemplo, que os acessos diminuíram 20% em relação ao padrão de acesso em horário específico do dia da semana.

Comentários:

Perfeito! Ferramentas de Machine Learning são capazes de analisar uma grande quantidade de dados por meio de algoritmos complexos para encontrar padrões interessantes que sejam úteis para o negócio e fazer insights de forma autônoma. Logo, uma aplicação certamente seria capaz de identificar queda no acesso em relação ao padrão do horário específico do dia da semana por meio do monitoramento do tráfego. Aliás, isso é bastante utilizado por redes sociais para incrementar ou reduzir a quantidade de anúncios e propagandas.

Gabarito: Correto

---

30. **(CESPE / Polícia Federal – 2018)** Uma aplicação que reconheça o acesso de um usuário e forneça sugestões diferentes para cada tipo de usuário pode ser considerada uma aplicação que usa machine learning.

Comentários:

Perfeito! Observe que o enunciado menciona que a solução é alimentada com os dados de um sítio de comércio eletrônico, armazenando informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos. Isso é frequentemente utilizado em redes sociais para exibir anúncios específicos de acordo com o perfil de um usuário.

Por exemplo: meu instagram não para de exibir anúncios de camisas do Flamengo e do Pink Floyd. *Por que?* Porque ele analisa uma quantidade imensa de dados que eu forneço enquanto navego por essa rede social e é capaz de identificar autonomamente que eu sigo páginas sobre esses temas.

Gabarito: Correto

---

31. **(CESPE / Polícia Federal – 2018)** Dados coletados de redes sociais podem ser armazenados, correlacionados e expostos com o uso de análises preditivas.

Comentários:

Dados coletados de redes sociais podem ser armazenados? Sim! *Podem ser correlacionados?* Sim! *Podem ser expostos com o uso de análises preditivas?* Sim! A Análise Preditiva é capaz de identificar o relacionamento existente entre os componentes de um conjunto de dados, utilizando algoritmos sofisticados, com o intuito de identificar padrões de comportamento ao examinar grandes quantidades de dados automaticamente. Logo, é claro que dados coletados de redes sociais podem ser expostos com o uso de análises preditivas.

Gabarito: Correto

---

32. **(CESPE / TCE-PE – 2017)** 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.

Comentários:

Perfeito! Volume, Velocidade e Variedade são algumas das características que definem o que é Big Data – além de ser formado por dados estruturados e não estruturados.

Gabarito: Correto

---

33. **(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.

Comentários:

Big Data Analytics é o estudo e interpretação de grandes quantidades de dados armazenados com a finalidade de extrair padrões de comportamento. Em outras palavras, utiliza-se uma combinação de sistemas de softwares matemáticos de alta tecnologia que juntos são capazes de tratar dados estruturados e não-estruturados, analisá-los e extrair um significado de alto valor para

organizações. E, de fato, é permitido analisar dados como registros de call center, postagens de redes sociais, de blogs, dados de CRM e demonstrativos de resultados.

Gabarito: Correto

---

34. (CESPE / TER-GO – 2015) 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.

Comentários:

A grande quantidade de informações obtidas através do Big Data pode ser empregada para auxiliar em diversos cenários de tomadas estratégicas de decisões, tal como em um EAD (Ensino À Distância).

Gabarito: Correto

---

35. (CESPE / FUNPRESP - JUD – 2016) 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.

Comentários:

Big Data é o termo que descreve o imenso volume de dados – estruturados e não estruturados – que impactam os negócios de uma organização.

Gabarito: Errado

---

36. (CESPE / TJ-SE – 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.

Comentários:

Para uma quantidade gigantesca de dados, a distância do local de processamento afeta – sim – os custos e o tempo de resposta. Quanto mais próximo, mais barato e mais rápido; quanto mais longe, mais caro e mais lento. É claro que isso pode mudar nos próximos anos...

Gabarito: Errado

---

37. (CESPE / TJ-SE – 2014) Em soluções *Big Data*, a análise dos dados comumente precisa ser precedida de uma transformação de dados não estruturados em dados estruturados.

Comentários:

Para executar a análise dos dados, os dados não estruturados comumente devem estar em algum tipo de formato estruturado (Ex: JSON, que é um formato de intercâmbio de dados).

Gabarito: Correto

QUESTÕES COMENTADAS – FCC

38. (FCC / SEFAZ-SC – 2018) No âmbito da ciência de dados na definição de Big Data, utilizam-se características ou atributos que alguns pesquisadores adotam como sendo os cinco Vs. Porém, a base necessária para o reconhecimento de Big Data é formada por três propriedades:

a) valor, velocidade e volume.
b) valor, veracidade e volume.
c) variedade, velocidade e volume.
d) variedade, valor e volume.
e) velocidade, veracidade e volume.

Comentários:

As três propriedades principais são variedade, velocidade e volume.

Gabarito: Letra C

---

39. (FCC / SEFAZ-SC – 2018) As soluções em *Big Data Analytics*, usadas, por exemplo, pela Fazenda Pública principalmente para evitar sonegações de tributos, trabalham com algoritmos complexos, agregando dados de origens diversas, relacionando-os e gerando conclusões fundamentais para a tomada de decisões. Na execução dessas análises pelos auditores, considere:

I. Dados estruturados.
II. Dados semiestruturados.
III. Dados não estruturados.
IV. Dados brutos, não processados.
V. Esquemas de dados gerados no momento da gravação.

Sobre um repositório de armazenamento, que contenha uma grande quantidade de dados a ser examinada, deverão ser utilizados APENAS os que constam de:

a) I, III e IV.
b) I, II, III e V.
c) III, IV e V.
d) I, II, III e IV.
e) I, II, IV e V.

Comentários:

Devem ser utilizados apenas Dados Estruturados, Dados Semiestruturados, Dados Não-Estruturados e Dados Brutos. Os Dados Brutos designam os dados/valores recolhidos e armazenados tal qual foram adquiridos, sem terem sofrido o menor tratamento. Apresentam-se como um conjunto de números, caracteres, imagens ou outros dispositivos de saídas para converter quantidades físicas em símbolos, num sentido muito extenso. No entanto, esquemas de dados gerados no momento da gravação são dados temporários e, normalmente, não são úteis como fonte de dados para *Big Data Analytics*.

Gabarito: Letra D

---

40. (FCC / TCE-RS – 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.

Comentários:

(a) Correto. A Variedade é a propriedade de os dados serem gerados em inúmeros formatos diferentes – estruturados e não-estruturados; (b) Errado. Não há limitação de usuários distintos acessando o sistema na definição; (c) Errado. Tempos de acesso não entram na definição de variedade dos 3V's; (d) Errado. Quantidade de tipos de máquinas acessando o sistema não entram na definição de variedade dos 3V's; (e) Errado. Tamanhos das tabelas que compõem o sistema não entram na definição de variedade dos 3V's.

Lembrando que o Big Data foi inicialmente conceituado com base apenas em três premissas básicas: Volume, Velocidade e Variedade (3 V's).

Gabarito: Letra A

---

41. (FCC / Câmara Legislativa do Distrito Federal – 2018) A proposta de uma solução de Big Data, oferecendo uma abordagem consistente no tratamento do constante crescimento e da complexidade dos dados, deve considerar os 5 V's do Big Data que envolvem APENAS os conceitos de:

a) volume, versionamento, variedade, velocidade e visibilidade.
b) velocidade, visibilidade, volume, veracidade e vencimento do dado
c) volume, velocidade, variedade, veracidade e valor
d) variedade, vencimento do dado, veracidade, valor e volume
e) vulnerabilidade, velocidade, visibilidade, valor e veracidade

Comentários:

(a) Errado, não envolve versionamento e visibilidade; (b) Errado, não envolve visibilidade e vencimento do dado; (c) Correto; (d) Errado, não envolve vencimento do dado; (e) Errado, não envolve vulnerabilidade e visibilidade.

Gabarito: Letra C

QUESTÕES COMENTADAS – FGV

42. (FGV / TJ-SC – 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.

Comentários:

O Big Data pode ser rapidamente identificado através das premissas: Variedade, Velocidade e Volume.

Gabarito: Letra C

---

43. (FGV / AL-BA – 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, e-mail, 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.

Comentários:

(I) Correto. Volume trata realmente da quantidade gigantesca e crescente de dados; (II) Correto. Velocidade trata da capacidade de processar dados rapidamente para gerar as informações necessárias para que sejam tomadas decisões de forma tempestiva. No entanto, a questão fala que não é possível armazenar todos os dados, sendo obrigatório escolher os dados que serão armazenados. Eu discordo desse entendimento, mas a banca manteve o gabarito; (III) Correto. Variedade trata dos diferentes tipos de dados (estruturados e não-estruturados) advindos de fontes diversas.

Gabarito: Letra E

QUESTÕES COMENTADAS – DIVERSAS BANCAS

44. (SELECON / EMGEPRON – 2021) 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.

Comentários:

Infinidade de informações não estruturadas? Arma poderosa para empresas tomarem decisões? Permitem analisar um enorme volume de dados de forma rápida? Oferecem total controle ao gestor das informações? As fontes de dados são as mais diversas possíveis? Todas essas são características de Big Data!

Gabarito: Letra C

---

45. (COMPERVE / TJ/RN – 2020) Embora Big Data tenha diferentes definições, há um consenso sobre o modelo dos 3 V's que correspondem a 3 características. Duas dessas características são:

a) Volume e Velocity.
b) Variety e Value.
c) Viable e Vast.
d) Valid e Verbose.

Comentários:

Na verdade, atualmente já existe um consenso sobre os 5V's. No entanto, o Big Data foi inicialmente conceituado com base apenas em três premissas: Volume, Velocidade e Variedade (Volume, Velocity e Variety).

Gabarito: Letra A

46.(CCV-UFC / UFC – 2019) Sobre os banco de dados NoSQL, assinale a afirmativa correta:

a)Bancos de dados NoSQL não podem ser indexados.
b)Bancos de dados NoSQL são considerados banco de dados relacionais.
c) Nos bancos de dados NoSQL devem ser definidos um esquema de dados fixo antes de qualquer operação.
d) São exemplos de bancos de dados NoSQL: MongoDB, Firebird, DynamoDB, SQLite, Microsoft Access e Azure Table Storage.
e) Os bancos de dados NoSQL usam diversos modelos para acessar e gerenciar dados, como documento, gráfico, chave-valor, em memória e, pesquisa.

Comentários:

(a) Errado, podem – sim – ser indexados; (b) Errado, são considerados não-relacionais; (c) Errado, não é necessário definir um esquema fixo prévio; (d) Errado, FireBird, SQLite e Microsoft Access são exemplos de bancos de dados relacionais; (e) Correto, eles realmente usam modelos para acessar e gerenciar dados como documento, gráfico, chave-valor, em memória e pesquisa. Algumas ressalvas: (1) o termo mais correto é grafo; (2) em memória e pesquisa não são modelos consagrados.

Gabarito: Letra E

47.(IADES / APEX BRASIL – 2018) Assinale a alternativa que apresenta o conceito de *Big Data*.

a) Conjuntos de dados de grande volume que se utilizam de ferramentas especiais de processamento, pesquisa e análise, e que podem ser aproveitados no tempo necessário, com precisão e grande velocidade.

b) São bancos de dados de fácil acesso e rápida velocidade, operados como computadores pessoais.

c) Manuseio de informações necessárias às empresas e aos negócios do mundo moderno, que podem ser armazenadas em computadores pessoais, utilizando-se a técnica de nuvem de dados.

d) São apenas grandes volumes de dados que precisam ainda ser mais bem aproveitados pelo mundo corporativo.

e) Refere-se a um grande número de computadores pessoais (PC) interligados entre si em uma grande rede de informação.

Comentários:

(a) Correto. Trata-se de um enorme conjunto de dados que utiliza softwares especiais para o processamento e transformação de dados em informações com precisão em uma velocidade absurdamente alta; (b) Errado. Não são de fácil acesso e rápida velocidade, muito menos operados como computadores pessoais; (c) Errado. Não há nada que faça sentido nesse item; (d) Errado. Atualmente eles são muito bem aproveitados pelo mundo corporativo; (e) Errado. Essa é a definição de uma rede de computadores.

Gabarito: Letra A

48.(CESGRANRIO / PETROBRAS – 2018) A principal definição de *Big Data* parte de três características, conhecidas como 3 V do Big Data, a saber: velocidade, variedade e volume. O termo velocidade refere-se, principalmente, à:

a) necessidade das aplicações de gerar respostas rapidamente, a partir de grandes massas de dados.

b) existência de um alto fluxo de dados na entrada.

c) necessidade de gerar aplicações rapidamente, em função da demanda do negócio.

d) importância da facilidade de manipular cubos de visualização de dados, rapidamente.

e) rapidez com que os dados se tornam inválidos com o tempo.

Comentários:

O termo velocidade refere-se à velocidade com que os dados são criados. Em outras palavras, trata-se da existência de um alto fluxo de dados na entrada. São mensagens de redes sociais se viralizando em segundos, transações de cartão de crédito sendo verificadas a cada instante ou os milissegundos necessários para calcular o valor de compra e venda de ações.

Gabarito: Letra B

49.(AOCP / CPD/BA – 2018) Big Data requer clusters de servidores de apoio às ferramentas que processam grandes volumes, alta velocidade e formatos variados de Big Data. Nesse sentido, é correto afirmar que Hadoop refere-se a:

a) um sistema de armazenamento e processamento de dados massivamente escalável – não é um banco de dados.

b) uma estratégia baseada em tecnologia que permite a coleta de insights mais profundos e relevantes dos clientes, parceiros e sobre o negócio.

c) um banco de dados com capacidade melhorada.

d) um equipamento de hardware que permite que sistemas administrem crescentes cargas de processamento.

e) um banco de dados com tecnologia de virtualização.

Comentários:

(a) Correto, ele realmente é um sistema de armazenamento e processamento, sendo massivamente escalável sem ser um banco de dados; (b) Errado, isso seria Big Data Analytics; (c) Errado, ele não é um banco de dados; (d) Errado, ele é um software e, não, um hardware; (e) Errado, ele não é um banco de dados.

Gabarito: Letra A

---

50. **(AOCP / CPD/BA – 2018)** Big Data se refere ao imenso volume de conjuntos de dados que alcançam elevadas ordens de magnitude. O valor real do Big Data está no insight que ele produz quando analisado — buscando padrões, derivando significado, tomando decisões e, por fim, respondendo ao mundo com inteligência. Referente ao Big Data, é correto afirmar que o termo variedade refere-se:

a) um conjunto de dados mais diversos, incluindo dados estruturados, semiestruturados e não estruturados. É heterogêneo e vem em muitos formatos, incluindo texto, documento, imagem, vídeo e outros.

b) a banco de dados homogêneo que trata de informações do mesmo tipo definindo padrões de segurança.

c) a um conjunto de dados que são gerados em tempo real, o que requer a oferta imediata de informações úteis.

d) aos data centers físicos que transformam os dados em informações pertinentes ao negócio.

e) ao controle de dados semiestruturados de formatos definidos como texto e números.

Comentários:

(a) Correto, essa é definição perfeita de variedade no contexto de Big Data; (b) Errado, trata-se de um banco de dados heterogêneo, que trata de informações de tipos diferentes e sem um padrão de segurança; (c) Errado, essa definição se refere à velocidade; (d) Errado, essa definição não apresenta

nenhuma relação com o conceito de variedade; (e) Errado, o conceito de variedade trata de dados em quaisquer formatos.

Gabarito: Letra A

---

51. **(QUADRIX / CRM/DF – 2018)** O fato de o ser humano gerar milhares de informações a cada minuto dá origem ao conceito de Big Data, que se trata de uma nova linha de banco de dados que não possui qualquer relação com as existentes até o momento. Pelo fato de as bases de dados Big Data serem do tipo Plano, não podem ser manipuladas e consultadas pelo SQL.

Comentários:

Big Data não é uma nova linha de banco de dados. Além disso, as bases de dados não são do tipo plano (que são aquelas que armazenam dados em um arquivo texto simples) – elas são, em geral, não-relacionais.

Gabarito: Errado

---

52. **FEPESE / CIA/SC – 2017)** Um banco de dados de Big Data deve possuir pelo menos três aspectos, os chamados 3Vs do Big Data, que são:

a) Variedade; Volume; Valor.
b) Valor; Variabilidade; Velocidade.
c) Volume; Veracidade; Velocidade.
d) Veracidade; Velocidade; Variedade.
e) Velocidade; Volume; Variedade.

Comentários:

Atualmente, existe um consenso sobre os 5V's! No entanto, o Big Data foi inicialmente conceituado com base apenas em três premissas: Volume, Velocidade e Variedade.

Gabarito: Letra E

---

53. **(CESGRANRIO / PETROBRÁS – 2017)** O termo Big Data é bastante conhecido pelos profissionais de tecnologia da informação, especialmente aqueles envolvidos com bancos de dados, inteligência de negócios, sistemas de informações e sistemas de apoio à decisão. Uma característica inerente a esse conceito é a da:

a) complexidade das suas fontes de informação, o que demanda a necessidade de sua prévia limpeza, integração e transformação.

b) estabilidade da taxa de geração desses dados, o que garante sua utilização confiável na geração analítica de informação com independência temporal.

c) heterogeneidade do conjunto de dados, empregada em dados originalmente estruturados ou semiestruturados.

d) qualidade das fontes de dados, por conta dos padrões de expansão e de retenção reveladores da ordem existente nos dados.

e) escalabilidade, que, na sua forma original, possui alto valor granular quando comparado ao de seu volume.

Comentários:

(a) Correto, porém com ressalvas – em geral, é realmente necessária a limpeza, mas isso não é obrigatório com faz crer a redação da questão; (b) Errado, não há estabilidade da taxa de geração de dados – pelo contrário, tudo é caótico e depende da realidade do negócio; (c) Errado, os dados originais são primariamente não-estruturados; (d) Errado, confesso que não sei o que a questão quis dizer com padrões de expansão e de retenção reveladores da ordem existente nos dados – não vejo relação disso com Big Data; (e) Errado, Big Data é um conceito – as suas implementações que podem ser escaláveis, logo não se trata de uma característica inerente a esse conceito.

Gabarito: Letra A

---

54. (AOCP / CCAS-SC – 2015) Em relação à Big Data e NoSQL, é correto afirmar que:

a) são conceitos concorrentes, portanto não podem ser implementados juntos.
b) são conceitos que se complementam e com características eficientes para trabalhar com pequenas quantidades de informações.
c) são duas ferramentas de empresas concorrentes.
d) são conceitos que se complementam.
e) os SGBDs Oracle e MySQL são implementações desses conceitos.

Comentários:

(a) Errado, são conceitos complementares e, não, concorrentes – podendo ser implementados juntos; (b) Errado, uma das suas características é trabalhar com grandes quantidades de informações; (c) Errado, não são ferramentas – são conceitos; (d) Correto, realmente são conceitos complementares; (e) Errado, Oracle e MySQL são implementações de bancos de dados relacionais.

Gabarito: Letra D

---

55. (ESAF / MPDG – 2015) Em relação a Big Data e NoSQL, é correto afirmar que:

a) os "3 Vs" principais do Big Data referem-se a Volume, Velocidade e Versatilidade de dados.

b) na era do Big Data, as únicas estratégias eficientes para garantir a privacidade são consentimento individual, opção de exclusão e anonimização.

c) 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.

d) o NoSQL é um sistema relacional, distribuído, em larga escala, muito eficaz na organização e análise de grande quantidade de dados.

e) o Cassandra é um sistema de banco de dados baseado na abordagem NoSQL, originalmente criado pelo Facebook, no qual os dados são identificados por meio de uma chave.

Comentários:

(a) Errado, referem-se a Volume, Velocidade e Variedade; (b) Errado, a questão viajou e misturou até alguns conceitos de proteção de dados; (c) Errado, ele foi criado pela Apache – é ridículo cobrar isso em prova; (d) Errado, é um conjunto de bancos de dados não relacionais, distribuído, de larga escala e muito eficaz na organização e análise de grande quantidade de dados; (e) Correto, ele realmente é um sistema de banco de dados não-relacionado baseado em NoSQL criado pelo Facebook no qual dados podem ser identificados por meio de uma chave.

Gabarito: Letra E

QUESTÕES COMENTADAS – CESPE

1. **(CESPE / SEFAZ-SE – 2022)** Com relação a noções de big data, julgue os itens que se seguem.

   I Como qualquer tecnologia, soluções de big data também apresentam algumas restrições. Por exemplo, elas não podem ser utilizadas na área da saúde para determinar a causa de uma doença, porque esse é um procedimento complexo que somente pode ser executado por pessoas devidamente capacitadas — nesse caso, os médicos.

   II Big data é qualquer tipo de fonte de dados que possui, no mínimo, as seguintes três características: volume de dados extremamente grande; velocidade de dados extremamente alta; e variedade de dados extremamente ampla.

   III Para que as organizações obtenham os conhecimentos corretos, a tecnologia big data não permite que elas executem as operações de armazenar e administrar as grandes quantidades de dados de si próprias.

   IV Big data é uma combinação de tecnologias de gestão de dados que evoluíram ao longo dos anos, razão por que não é considerado um mercado único.

   Estão certos apenas os itens:

   a) I e III
   b) I e IV
   c) II e IV
   d) II e V
   e) III e V

2. **(CESPE / PETROBRAS – 2022)** Em sistemas NoSQL baseados em armazenamento de chavevalor, a chave é multidimensional e composta pela combinação do nome de tabela com a chave linha-coluna e com o rótulo de data e hora.

3. **(CESPE / ISS-Aracaju – 2021)** Big data ajudou a sedimentar o cargo de cientista de dados. Entre as funções desse cargo inclui-se:

   a) a modelagem estruturada.
   b) a análise retrospectiva.
   c) a modelagem não estruturada.
   d) a modelagem relacional.
   e) o processamento comparativo.

4. **(CESPE / SERPRO – 2021)** MapReduce divide o conjunto de dados de entrada em blocos independentes que são processados pelas tarefas de mapa de uma maneira completamente paralela. Essa estrutura classifica as saídas dos mapas, as quais são, então, inseridas nas tarefas de redução.

5. **(CESPE / SERPRO – 2021)** No que se refere aos três Vs do Big Data, o termo volume refere-se a dados que, atualmente, não são estruturados nem armazenados em tabelas relacionais, o que torna sua análise mais complexa.

6. **(CESPE / SERPRO – 2021)** Big data caracteriza-se, principalmente, por volume, variedade e velocidade, o que se justifica devido ao fato de os dados serem provenientes de sistemas estruturados, que são maioria, e de sistemas não estruturados, os quais, embora ainda sejam minoria, vêm, ao longo dos anos, crescendo consideravelmente.

7. **(CESPE / SERPRO – 2021)** MapReduce divide o conjunto de dados de entrada em blocos independentes que são processados pelas tarefas de mapa de uma maneira completamente paralela. Essa estrutura classifica as saídas dos mapas, as quais são, então, inseridas nas tarefas de redução.

8. **(CESPE / SERPRO – 2021)** Uma das principais características de big data é que seu custo de armazenamento de dados é relativamente baixo se comparado a um data warehouse.

9. **(CESPE / SERPRO – 2021)** O Hadoop consiste em um único produto, ou seja, um software monolítico, que possibilita análise de logs e outros dados da Web.

10. **(CESPE / SERPRO – 2021)** Apesar de ser uma tecnologia de código aberto disponibilizada pela ASF (Apache Software Foundation), o Hadoop também é oferecido por distribuidores comerciais, de maneira que fornecedores oferecem distribuições específicas que incluem não só ferramentas administrativas adicionais, mas também suporte técnico.

11. **(CESPE / TCE-RJ – 2021)** Em Big Data, a premissa volume refere-se à capacidade de processar, em um ambiente computacional, diferentes tipos e formatos de dados, como fotos, vídeos e geolocalização.

12. **(CESPE / TCE-RJ – 2021)** Volume, variedade e visualização são as três características, conhecidas como 3 Vs, utilizadas para definir Big Data.

13. **(CESPE / Polícia Federal – 2021)** As aplicações de *bigdata* caracterizam-se exclusivamente pelo grande volume de dados armazenados em tabelas relacionais.

14. **(CESPE / PRF – 2021)** A Internet das Coisas (IoT) aumenta a quantidade e a complexidade dos dados por meio das novas formas e novas fontes de informações, influenciando diretamente em uma ou mais das características do Big Data, a exemplo de volume, velocidade e variedade.

15. **(CESPE / AL-AP – 2020)** Atualmente, diversos dados são coletados pelos sistemas digitais de empresas na internet para constituir Big Data com conteúdo sobre os resultados alcançados por seus produtos e serviços, prestígio da imagem da organização e seus representantes. Porém, parte desses dados pode ser falsa ou manipulada por internautas. O tratamento dos dados, a fim de qualificá-los antes de disponibilizá-los para a tomada de decisão na empresa, segundo o conceito das cinco dimensões "V" de avaliação de um Big Data, se refere:

a) ao valor.
b) à variedade.
c) à veracidade.
d) à velocidade.
e) ao volume.

16. **(CESPE / TCE-RO – 2019)** Com relação a fundamentos e conceitos de Big Data, julgue os itens a seguir.

I. O volume de dados é uma característica importante de Big Data.

II. Em Big Data, a qualidade do dado não tem importância, porque a transformação dos dados não impacta os negócios.

III. A característica de velocidade de entrada dos dados impacta o modelo de processamento e armazenamento.

IV. A variedade dos dados não é característica intrínseca nos fundamentos de Big Data.

Estão certos apenas os itens:

a) I e II.
b) I e III.
c) II e IV.
d) I, III e IV.
e) II, III e IV.

17. **(CESPE / TCM-BA – 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 proveem 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.

18. **(CESPE / TCM-BA – 2018)** Um dos desdobramentos de *big data* é o *big data analytics*, que se refere aos softwares capazes de tratar dados para transformá-los em informações úteis às organizações. O *big data analytics* difere do *business intelligence* por:

a) priorizar o ambiente de negócios em detrimento de outras áreas.
b) analisar dúvidas já conhecidas para as quais se deseje obter resposta.
c) analisar o que já existe e o que está por vir, apontando novos caminhos.
d) dar enfoque à coleta, à transformação e à disponibilização dos dados.
e) analisar o que já existe, definindo as melhores hipóteses.

19. **(CESPE / Polícia Federal – 2018)** Em um *big data*, alimentado com os dados de um sítio de comércio eletrônico, são armazenadas informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos. Tendo como referência as informações apresentadas, julgue o item seguinte.

O *big data* consiste de um grande depósito de dados estruturados, ao passo que os dados não estruturados são considerados data files.

20. **(CESPE / Polícia Federal – 2018)** Em um *big data*, alimentado com os dados de um sítio de comércio eletrônico, são armazenadas informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos. Tendo como referência as informações apresentadas, julgue o item seguinte.

Dados coletados de redes sociais podem ser armazenados, correlacionados e expostos com o uso de análises preditivas.

21. **(CESPE / Polícia Federal – 2018)** Em um *big data*, alimentado com os dados de um sítio de comércio eletrônico, são armazenadas informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos. Tendo como referência as informações apresentadas, julgue o item seguinte.

Uma aplicação que reconheça o acesso de um usuário e forneça sugestões diferentes para cada tipo de usuário pode ser considerada uma aplicação que usa *machine learning*.

22. **(CESPE / ABIN – 2018)** O registro e a análise de conjuntos de dados referentes a eventos de segurança da informação são úteis para a identificação de anomalias; esse tipo de recurso pode ser provido com uma solução de *big data*.

23. **(CESPE / IPHAN – 2018)** A utilização de tecnologias emergentes para o tratamento de grandes volumes de dados (*big data*) pode contribuir para o sucesso da implantação da estratégia de governança digital.

24. **(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.

25. **(CESPE / Polícia Federal – 2018)** MapReduce oferece um modelo de programação com processamento por meio de uma combinação entre chaves e valores.

26. **(CESPE / TCE-PB – 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.

27. **(CESPE / TCE-MG – 2018)** Um dos desdobramentos de big data é o Big Data Analytics, que se refere aos softwares capazes de tratar dados para transformá-los em informações úteis às organizações. O Big Data Analytics difere do Business Intelligence por:

a) priorizar o ambiente de negócios em detrimento de outras áreas.

b) analisar dúvidas já conhecidas para as quais se deseje obter resposta.
c) analisar o que já existe e o que está por vir, apontando novos caminhos.
d) dar enfoque à coleta, à transformação e à disponibilização dos dados.
e) analisar o que já existe, definindo as melhores hipóteses.

28. **(CESPE / Polícia Federal – 2018)** A mineração de dados se caracteriza especialmente pela busca de informações em grandes volumes de dados, tanto estruturados quanto não estruturados, alicerçados no conceito dos 4V's: volume de mineração, variedade de algoritmos, velocidade de aprendizado e veracidade dos padrões.

Em um big data, alimentado com os dados de um sítio de comércio eletrônico, são armazenadas informações diversificadas, que consideram a navegação dos usuários, os produtos comprados e outras preferências que o usuário demonstre nos seus acessos.

Tendo como referência as informações apresentadas, julgue o item seguinte.

29. **(CESPE / Polícia Federal – 2018)** Pelo monitoramento do tráfego de rede no acesso ao sítio em questão, uma aplicação que utiliza machine learning é capaz de identificar, por exemplo, que os acessos diminuíram 20% em relação ao padrão de acesso em horário específico do dia da semana.

30. **(CESPE / Polícia Federal – 2018)** Uma aplicação que reconheça o acesso de um usuário e forneça sugestões diferentes para cada tipo de usuário pode ser considerada uma aplicação que usa machine learning.

31. **(CESPE / Polícia Federal – 2018)** Dados coletados de redes sociais podem ser armazenados, correlacionados e expostos com o uso de análises preditivas.

32. **(CESPE / TCE-PE – 2017)** 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.

33. **(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.

34. **(CESPE / TER-GO – 2015)** 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.

35. **(CESPE / FUNPRESP - JUD – 2016)** 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.

36. **(CESPE / TJ-SE – 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.

37. **(CESPE / TJ-SE – 2014)** Em soluções *Big Data*, a análise dos dados comumente precisa ser precedida de uma transformação de dados não estruturados em dados estruturados.

QUESTÕES COMENTADAS – FCC

38. (FCC / SEFAZ-SC – 2018) No âmbito da ciência de dados na definição de Big Data, utilizam-se características ou atributos que alguns pesquisadores adotam como sendo os cinco Vs. Porém, a base necessária para o reconhecimento de Big Data é formada por três propriedades:

a) valor, velocidade e volume.
b) valor, veracidade e volume.
c) variedade, velocidade e volume.
d) variedade, valor e volume.
e) velocidade, veracidade e volume.

39. (FCC / SEFAZ-SC – 2018) As soluções em *Big Data Analytics*, usadas, por exemplo, pela Fazenda Pública principalmente para evitar sonegações de tributos, trabalham com algoritmos complexos, agregando dados de origens diversas, relacionando-os e gerando conclusões fundamentais para a tomada de decisões. Na execução dessas análises pelos auditores, considere:

I. Dados estruturados.
II. Dados semiestruturados.
III. Dados não estruturados.
IV. Dados brutos, não processados.
V. Esquemas de dados gerados no momento da gravação.

Sobre um repositório de armazenamento, que contenha uma grande quantidade de dados a ser examinada, deverão ser utilizados APENAS os que constam de:

a) I, III e IV.
b) I, II, III e V.
c) III, IV e V.
d) I, II, III e IV.
e) I, II, IV e V.

40. (FCC / TCE-RS – 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.

41. **(FCC / Câmara Legislativa do Distrito Federal – 2018)** A proposta de uma solução de Big Data, oferecendo uma abordagem consistente no tratamento do constante crescimento e da complexidade dos dados, deve considerar os 5 V's do Big Data que envolvem APENAS os conceitos de:

a) volume, versionamento, variedade, velocidade e visibilidade.
b) velocidade, visibilidade, volume, veracidade e vencimento do dado
c) volume, velocidade, variedade, veracidade e valor
d) variedade, vencimento do dado, veracidade, valor e volume
e) vulnerabilidade, velocidade, visibilidade, valor e veracidade

QUESTÕES COMENTADAS – FGV

42. (FGV / TJ-SC – 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.

43. (FGV / AL-BA – 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, e-mail, 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.

QUESTÕES COMENTADAS – DIVERSAS BANCAS

44. (SELECON / EMGEPRON – 2021) 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.

45. (COMPERVE / TJ/RN – 2020) Embora Big Data tenha diferentes definições, há um consenso sobre o modelo dos 3 V's que correspondem a 3 características. Duas dessas características são:

a) Volume e Velocity.
b) Variety e Value.
c) Viable e Vast.
d) Valid e Verbose.

46. (CCV-UFC / UFC – 2019) Sobre os banco de dados NoSQL, assinale a afirmativa correta:

a)Bancos de dados NoSQL não podem ser indexados.
b)Bancos de dados NoSQL são considerados banco de dados relacionais.
c) Nos bancos de dados NoSQL devem ser definidos um esquema de dados fixo antes de qualquer operação.
d) São exemplos de bancos de dados NoSQL: MongoDB, Firebird, DynamoDB, SQLite, Microsoft Access e Azure Table Storage.
e) Os bancos de dados NoSQL usam diversos modelos para acessar e gerenciar dados, como documento, gráfico, chave-valor, em memória e, pesquisa.

47. (IADES / APEX BRASIL – 2018) Assinale a alternativa que apresenta o conceito de *Big Data*.

a) Conjuntos de dados de grande volume que se utilizam de ferramentas especiais de processamento, pesquisa e análise, e que podem ser aproveitados no tempo necessário, com precisão e grande velocidade.

b) São bancos de dados de fácil acesso e rápida velocidade, operados como computadores pessoais.

c) Manuseio de informações necessárias às empresas e aos negócios do mundo moderno, que podem ser armazenadas em computadores pessoais, utilizando-se a técnica de nuvem de dados.

d) São apenas grandes volumes de dados que precisam ainda ser mais bem aproveitados pelo mundo corporativo.

e) Refere-se a um grande número de computadores pessoais (PC) interligados entre si em uma grande rede de informação.

48. **(CESGRANRIO / PETROBRAS – 2018)** A principal definição de *Big Data* parte de três características, conhecidas como 3 V do Big Data, a saber: velocidade, variedade e volume. O termo velocidade refere-se, principalmente, à:

a) necessidade das aplicações de gerar respostas rapidamente, a partir de grandes massas de dados.

b) existência de um alto fluxo de dados na entrada.

c) necessidade de gerar aplicações rapidamente, em função da demanda do negócio.

d) importância da facilidade de manipular cubos de visualização de dados, rapidamente.

e) rapidez com que os dados se tornam inválidos com o tempo.

49. **(AOCP / CPD/BA – 2018)** Big Data requer clusters de servidores de apoio às ferramentas que processam grandes volumes, alta velocidade e formatos variados de Big Data. Nesse sentido, é correto afirmar que Hadoop refere-se a:

a) um sistema de armazenamento e processamento de dados massivamente escalável – não é um banco de dados.

b) uma estratégia baseada em tecnologia que permite a coleta de insights mais profundos e relevantes dos clientes, parceiros e sobre o negócio.

c) um banco de dados com capacidade melhorada.

d) um equipamento de hardware que permite que sistemas administrem crescentes cargas de processamento.

e) um banco de dados com tecnologia de virtualização.

50. **(AOCP / CPD/BA – 2018)** Big Data se refere ao imenso volume de conjuntos de dados que alcançam elevadas ordens de magnitude. O valor real do Big Data está no insight que ele produz quando analisado — buscando padrões, derivando significado, tomando decisões e, por fim, respondendo ao mundo com inteligência. Referente ao Big Data, é correto afirmar que o termo variedade refere-se:

a) um conjunto de dados mais diversos, incluindo dados estruturados, semiestruturados e não estruturados. É heterogêneo e vem em muitos formatos, incluindo texto, documento, imagem, vídeo e outros.

b) a banco de dados homogêneo que trata de informações do mesmo tipo definindo padrões de segurança.

c) a um conjunto de dados que são gerados em tempo real, o que requer a oferta imediata de informações úteis.

d) aos data centers físicos que transformam os dados em informações pertinentes ao negócio.

e) ao controle de dados semiestruturados de formatos definidos como texto e números.

51. **(QUADRIX / CRM/DF – 2018)** O fato de o ser humano gerar milhares de informações a cada minuto dá origem ao conceito de Big Data, que se trata de uma nova linha de banco de dados que não possui qualquer relação com as existentes até o momento. Pelo fato de as bases de dados Big Data serem do tipo Plano, não podem ser manipuladas e consultadas pelo SQL.

52. **FEPESE / CIA/SC – 2017)** Um banco de dados de Big Data deve possuir pelo menos três aspectos, os chamados 3Vs do Big Data, que são:

a) Variedade; Volume; Valor.
b) Valor; Variabilidade; Velocidade.
c) Volume; Veracidade; Velocidade.
d) Veracidade; Velocidade; Variedade.
e) Velocidade; Volume; Variedade.

53. **(CESGRANRIO / PETROBRÁS – 2017)** O termo Big Data é bastante conhecido pelos profissionais de tecnologia da informação, especialmente aqueles envolvidos com bancos de dados, inteligência de negócios, sistemas de informações e sistemas de apoio à decisão. Uma característica inerente a esse conceito é a da:

a) complexidade das suas fontes de informação, o que demanda a necessidade de sua prévia limpeza, integração e transformação.

b) estabilidade da taxa de geração desses dados, o que garante sua utilização confiável na geração analítica de informação com independência temporal.

c) heterogeneidade do conjunto de dados, empregada em dados originalmente estruturados ou semiestruturados.

d) qualidade das fontes de dados, por conta dos padrões de expansão e de retenção reveladores da ordem existente nos dados.

e) escalabilidade, que, na sua forma original, possui alto valor granular quando comparado ao de seu volume.

54. **(AOCP / CCAS-SC – 2015)** Em relação à Big Data e NoSQL, é correto afirmar que:

a) são conceitos concorrentes, portanto não podem ser implementados juntos.
b) são conceitos que se complementam e com características eficientes para trabalhar com pequenas quantidades de informações.
c) são duas ferramentas de empresas concorrentes.
d) são conceitos que se complementam.
e) os SGBDs Oracle e MySQL são implementações desses conceitos.

55. **(ESAF / MPDG – 2015)** Em relação a Big Data e NoSQL, é correto afirmar que:

a) os "3 Vs" principais do Big Data referem-se a Volume, Velocidade e Versatilidade de dados.

b) na era do Big Data, as únicas estratégias eficientes para garantir a privacidade são consentimento individual, opção de exclusão e anonimização.

c) 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.

d) o NoSQL é um sistema relacional, distribuído, em larga escala, muito eficaz na organização e análise de grande quantidade de dados.

e) o Cassandra é um sistema de banco de dados baseado na abordagem NoSQL, originalmente criado pelo Facebook, no qual os dados são identificados por meio de uma chave.

GABARITO

1. LETRA C
2. ERRADO
3. LETRA C
4. CORRETO
5. ERRADO
6. ERRADO
7. CORRETO
8. ERRADO
9. ERRADO
10. CORRETO
11. ERRADO
12. ERRADO
13. ERRADO
14. CORRETO
15. LETRA C
16. LETRA B
17. LETRA C
18. LETRA C
19. ERRADO
20. CORRETO
21. CORRETO
22. CORRETO
23. CORRETO
24. LETRA C
25. CORRETO
26. LETRA E
27. LETRA C
28. ERRADO
29. CORRETO
30. CORRETO
31. CORRETO
32. CORRETO
33. CORRETO
34. CORRETO
35. ERRADO
36. ERRADO
37. CORRETO
38. LETRA C
39. LETRA D
40. LETRA A
41. LETRA C
42. LETRA C
43. LETRA E
44. LETRA C
45. LETRA A
46. LETRA E
47. LETRA A
48. LETRA B
49. LETRA A
50. LETRA A
51. ERRADO
52. LETRA E
53. LETRA A
54. LETRA D
55. LETRA E

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.