Estratégia Concursos

Aula 00

TRT-Campinas 15ª Região (Técnico Judiciário - Tecnologia da Informação) Passo Estratégico de Conhecimentos Específicos - 2024 (Pós-Edital)

Autor:
Fernando Pedrosa Lopes

07 de Dezembro de 2024

BIG DATA E NOSQL

Sumário

CONTEÚDO

Big Data. Fundamentos. Características. Erros comuns. Aplicações. Padrões. Fluxo de big data: ingestão, processamento e disponibilização. Armazenamento de big data. NoSQL.

ANÁLISE ESTATÍSTICA

Inicialmente, convém destacar o percentual de incidência do assunto, dentro da disciplina **Banco de Dados e Business Intelligence** em concursos/cargos similares. Quanto maior o percentual de cobrança de um dado assunto, maior sua importância.

Obs.: *um mesmo assunto pode ser classificado em mais de um tópico devido à multidisciplinaridade de conteúdo*.

Assunto	Relevância na disciplina em concursos similares
SQL	**21.6 %**
BI (Business Intelligence)	**9.0 %**
DW - Data Warehouse	**7.2 %**
SQL Server	**7.2 %**
Oracle	**6.3 %**
Banco de Dados Multidimensionais	**5.4 %**
Data Mining	**5.4 %**
Administração de banco de dados	**3.6 %**
Banco de Dados	**2.7 %**
Formas normais	**2.7 %**
ETL (Extract Transform Load)	**2.7 %**
Banco de Dados Relacionais	**2.7 %**
Arquitetura de Banco de Dados	**1.8 %**
SGBD - Sistema de Gerenciamento de Banco de Dados	**1.8 %**
OLAP (On-line Analytical Processing)	**1.8 %**
Segurança	**1.8 %**
MS-Access	**1.8 %**
Modelo relacional	**1.8 %**
Metadados e Metainformação	**1.8 %**
Álgebra relacional	**0.9 %**
Banco de Dados Paralelos e Distribuídos	**0.9 %**
Gerência de Transações	**0.9 %**
Modelagem de dados	**0.9 %**
Gatilhos (Triggers)	**0.9 %**
DER - Diagrama de Entidade e Relacionamento	**0.9 %**
Visão (View)	**0.9 %**
Banco de Dados Textuais	**0.9 %**
Índices	**0.9 %**
PostgreSQL	**0.9 %**

MySQL	**0.9 %**
Big Data	**0.9 %**

GLOSSÁRIO DE TERMOS

Faremos uma lista de termos que são relevantes ao entendimento do assunto desta aula. Caso tenha alguma dúvida durante a leitura, esta seção pode lhe ajudar a esclarecer.

Big Data: Termo utilizado para descrever conjuntos de dados grandes, complexos e que exigem processamento em escala.

Volume: Uma das características do Big Data, refere-se à grande quantidade de dados gerados e coletados pelas organizações.

Velocidade: Outra característica do Big Data, diz respeito à necessidade de processar dados em tempo real ou próximo a ele, devido à crescente demanda por respostas rápidas.

Variedade: Característica do Big Data que se refere à diversidade de tipos de dados, incluindo dados estruturados, semiestruturados e não estruturados.

Veracidade: Refere-se à qualidade e confiabilidade dos dados coletados.

Variabilidade: Diz respeito à dinamicidade dos dados, ou seja, como eles podem mudar rapidamente e serem difíceis de serem interpretados ou analisados.

Valor: Última "V" do Big Data, refere-se à capacidade de gerar valor a partir dos dados coletados e processados.

Padrão Atômico: Estrutura de dados que representa uma única entidade, como um registro em um banco de dados relacional.

Padrão Composto: Estrutura de dados que representa uma entidade composta por várias subentidades, como um objeto em um banco de dados orientado a objetos.

Big Data Pipeline: Arquitetura utilizada para coletar, processar e armazenar grandes volumes de dados, com o objetivo de transformá-los em informações úteis e acionáveis.

MapReduce: modelo de programação para processamento de grandes conjuntos de dados em cluster de computadores distribuídos. Divide o trabalho em duas tarefas principais: a

tarefa Map, que processa e transforma dados em pares chave-valor, e a tarefa Reduce, que agrega os resultados do Map e produz um resultado final.

NoSQL: Acrônimo para "Not Only SQL", termo utilizado para descrever bancos de dados que não seguem o modelo relacional.

NoSQL Orientado a Documentos: Tipo de banco de dados NoSQL que armazena dados em formato de documentos, geralmente em formato JSON ou BSON.

NoSQL Chave-Valor: Tipo de banco de dados NoSQL que armazena dados em pares de chave-valor simples.

NoSQL Orientado a Grafos: Tipo de banco de dados NoSQL que armazena dados em forma de grafos, permitindo a representação de relacionamentos complexos entre entidades.

NoSQL Orientado a Colunas: Tipo de banco de dados NoSQL que armazena dados em colunas, em vez de linhas, permitindo um processamento eficiente de grandes conjuntos de dados.

ROTEIRO DE REVISÃO

A ideia desta seção é apresentar um roteiro para que você realize uma revisão completa do assunto e, ao mesmo tempo, destacar aspectos do conteúdo que merecem atenção.

Big Data - O que é?

Atualmente, a humanidade produz uma quantidade impressionante de informações. Estima-se que em 2020 tenham sido criados cerca de 40 zettabytes (ou 40 trilhões de gigabytes) de dados, e espera-se que essa quantidade cresça exponencialmente nos próximos anos. Esses dados vêm de diversas fontes, como redes sociais, dispositivos móveis, sensores, transações financeiras, registros médicos, entre outros.

Big Data é um termo utilizado para se referir a estes grandes conjuntos de dados que são tão volumosos, complexos e variáveis que tornam difícil ou impossível serem gerenciados e processados usando ferramentas tradicionais de processamento de dados.

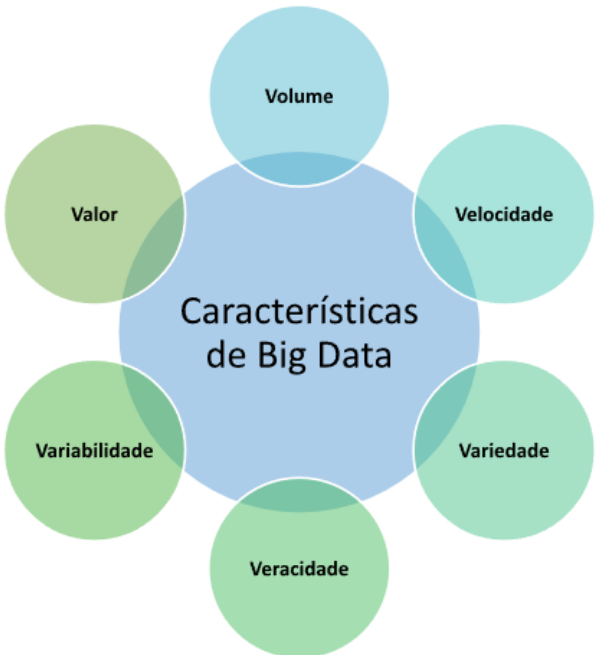
Como mencionamos, os dados que compõem o Big Data são originados de diversas fontes, incluindo redes sociais, dispositivos móveis, sensores, transações financeiras, registros médicos, entre outros. Eles são caracterizados principalmente por: volume (quantidade de dados), velocidade (velocidade de geração e processamento dos dados), variedade (diversidade de fontes e tipos de dados) e veracidade (confiabilidade e precisão dos dados).

Para lidar com esses dados, são necessárias tecnologias específicas que permitem armazenar, gerenciar e processar grandes volumes de dados de forma rápida e eficiente, como Hadoop, Spark, NoSQL, entre outras.

No final das contas, o grande objetivo do Big Data é **transformar essa quantidade massiva de informações em insights úteis e valiosos**. É aqui que entra o papel de tecnologias específicas, que são projetadas para lidar com grandes volumes de dados de forma rápida e eficiente, permitindo análises mais precisas e profundas.

Características

Big Data possui um conjunto de características que ajudam a entender as particularidades e desafios enfrentados ao lidar com grandes volumes de dados. Estas também são conhecidas como as dimensões do Big Data, ou o "Seis V's" do Big Data.



Volume: refere-se à quantidade de dados gerados, armazenados e processados. Big Data é caracterizado por grandes volumes de dados, que podem variar de terabytes a petabytes, exabytes ou zettabytes.

Velocidade: refere-se à rapidez com que os dados são gerados, coletados e processados. Big Data é caracterizado por altas taxas de geração e processamento de dados em tempo real, o que exige ferramentas capazes de lidar com a velocidade de entrada e saída de dados.

Variedade: refere-se à diversidade de fontes e tipos de dados, incluindo dados estruturados (como bancos de dados) e dados não estruturados (como textos, imagens, vídeos e áudios). Big Data pode ser caracterizado por dados heterogêneos e complexos que exigem abordagens diferentes para sua análise e processamento.

Veracidade: refere-se à qualidade e confiabilidade dos dados, incluindo sua precisão, consistência e integridade. Big Data pode apresentar problemas de veracidade, como dados incompletos, duplicados, imprecisos ou inconsistentes, o que afeta a tomada de decisões baseadas em dados.

Variabilidade: refere-se à capacidade de lidar com dados que apresentam variações e inconsistências em sua estrutura e formato. Esses dados podem ser estruturados, semiestruturados ou não-estruturados e podem ser provenientes de diferentes fontes e sistemas.

Valor: refere-se à capacidade de gerar valor a partir dos dados. Para obter insights valiosos, é preciso combinar e analisar dados de diferentes fontes para identificar padrões e tendências, prever comportamentos e tomar decisões melhores e mais informadas. O valor dos dados pode ser medido em termos de eficiência operacional, satisfação do cliente, inovação, entre outros.

Além dos "Seis V's", alguns autores também adicionam as dimensões de Pessoas e Governança ao gerenciamento de Big Data, quais sejam:

Pessoas: refere-se ao papel crucial que as pessoas desempenham no gerenciamento e análise de dados. As pessoas envolvidas na análise de Big Data precisam ter habilidades técnicas em ciência de dados, programação e estatística, além de habilidades interpessoais, como comunicação e colaboração. Além disso, a cultura organizacional precisa valorizar a tomada de decisão baseada em dados e apoiar o desenvolvimento de habilidades em Big Data

Governança: refere-se à gestão e controle dos dados, incluindo políticas, processos e procedimentos para garantir a qualidade, segurança e privacidade dos dados. A governança de dados é fundamental para garantir a conformidade com regulamentações e normas de segurança de dados, além de promover a transparência e a responsabilidade no uso de

dados. A governança também pode incluir a definição de papéis e responsabilidades, gestão de riscos, e a criação de políticas e padrões para o uso de dados.

Agora que estabelecemos nossa base para discussões subsequentes, vamos continuar a discutir as primeiras características do Big Data.

Volume

A característica de Volume do Big Data se refere à **grande quantidade de dados gerados**, armazenados e processados. Essa quantidade é tão grande que as tecnologias tradicionais de processamento e armazenamento de dados se tornam ineficientes, pois não conseguem lidar com tamanha quantidade de informação.

Um exemplo de grande volume de dados é o Facebook, que possui mais de 2 bilhões de usuários ativos e gera uma enorme quantidade de dados a cada segundo, incluindo posts, comentários, compartilhamentos e reações. Outro exemplo é a indústria de petróleo e gás, que coleta grandes quantidades de dados de sensores em poços e plataformas de petróleo, a fim de melhorar a eficiência da produção e reduzir custos.

Essa característica também está presente em outras áreas, como a medicina, onde grandes quantidades de dados são geradas por equipamentos médicos e exames clínicos. E na área de finanças, onde enormes volumes de dados são coletados em tempo real de diversas fontes, incluindo transações financeiras, notícias, redes sociais e outras fontes relevantes para a tomada de decisões.

Velocidade

A característica de Velocidade do Big Data se refere à **rapidez com que os dados são gerados**, processados e disponibilizados para análise em tempo real. Essa característica é crucial para muitas aplicações, como monitoramento de tráfego, detecção de fraudes em transações financeiras e análise de redes sociais em tempo real.

Um exemplo de aplicação que lida com grandes volumes de dados em alta velocidade é a detecção de fraudes em transações financeiras. Nessa aplicação, grandes quantidades de dados são processadas em tempo real para detectar atividades suspeitas que possam indicar fraude. Para isso, são utilizados algoritmos de aprendizado de máquina que processam e analisam os dados em tempo real, identificando padrões que possam indicar atividades fraudulentas.

Outro exemplo de aplicação que lida com grande velocidade de dados é a monitoração de tráfego em tempo real. Nessa aplicação, os dados de tráfego são coletados em tempo real por sensores em estradas e vias públicas, e são processados em tempo real para fornecer informações sobre o trânsito, incluindo congestionamentos e acidentes. Essas informações podem ser utilizadas para planejar rotas alternativas e melhorar o fluxo de tráfego.

Variedade

A característica de Variedade do Big Data se refere à **diversidade de formatos**, tipos e fontes de dados que precisam ser gerenciados e processados em um ambiente de Big Data. Diferentes tipos de dados incluem dados estruturados (por exemplo, tabelas de banco de dados), dados semiestruturados (por exemplo, XML, JSON) e dados não estruturados (por exemplo, textos, vídeos e imagens). Essa variedade de tipos de dados apresenta um desafio significativo para o gerenciamento e análise de dados em grande escala.

Um exemplo de aplicação que lida com grande variedade de dados é a análise de dados de mídias sociais. Nessa aplicação, uma grande variedade de dados é coletada, incluindo texto, imagens e vídeos, de várias fontes, como Twitter, Facebook e Instagram. Esses dados são processados e analisados para identificar tendências, opiniões e comportamentos do usuário.

Outro exemplo é o setor de saúde, onde dados de diferentes fontes, como registros médicos eletrônicos, dados de monitoramento de pacientes e dados de pesquisas clínicas, precisam ser integrados e analisados para obter insights valiosos sobre o diagnóstico, tratamento e prevenção de doenças.

Veracidade

A característica de Veracidade do Big Data lida com a incerteza dos dados e se refere à **garantia de que os dados são precisos e confiáveis** para que possam ser usados com segurança para tomar decisões de negócios ou governamentais importantes. Essa característica é particularmente importante em um ambiente de Big Data, onde grandes volumes de dados são coletados de diferentes fontes, muitas vezes em tempo real, e podem conter erros, duplicatas ou informações inconsistentes.

Um exemplo de aplicação que lida com a veracidade dos dados é a análise de dados de sensores em um ambiente industrial. Nessa aplicação, sensores coletam dados de máquinas, equipamentos e processos de produção em tempo real. É fundamental garantir a veracidade desses dados para que sejam tomadas decisões precisas e seguras de manutenção preventiva, otimização de produção e segurança do ambiente de trabalho.

Variabilidade

A característica de Variabilidade do Big Data se refere à capacidade de lidar com dados que apresentam **variações e inconsistências** em sua estrutura e formato. Esses dados podem ser estruturados, semiestruturados ou não-estruturados e podem ser provenientes de diferentes fontes e sistemas.

Por exemplo, dados provenientes de redes sociais como o Twitter e o Facebook podem ser não-estruturados, enquanto dados provenientes de um sistema de gestão empresarial (ERP) podem ser estruturados. Já dados provenientes de sensores de IoT (Internet das Coisas) podem ser semiestruturados.

É importante destacar que a variabilidade dos dados pode afetar a qualidade da análise. Dados inconsistentes ou incompletos podem levar a conclusões equivocadas, por isso é essencial realizar uma limpeza e tratamento adequado dos dados antes da análise.

Valor

A característica de Valor do Big Data se refere à **capacidade de extrair insights valiosos** e relevantes a partir dos dados, que possam levar a decisões de negócios ou governamentais mais eficazes e vantajosas. É importante destacar que, para agregar valor, os dados precisam ser transformados em informações úteis e compreensíveis.

Um exemplo de aplicação que utiliza a característica de Valor do Big Data é a análise de dados de vendas de uma empresa. Com o Big Data, é possível coletar dados em tempo real de diferentes fontes, como redes sociais, compras online e pontos de venda físicos, e combiná-los para obter insights valiosos sobre o comportamento dos clientes. Esses insights podem ser usados para otimizar a estratégia de vendas da empresa, personalizar ofertas e melhorar a experiência do cliente.

Erros Comuns

No contexto de Big Data, existem erros de raciocínio comuns que podem levar a conclusões equivocadas ou exageradas sobre o valor ou o impacto dos dados. Alguns exemplos de erros comuns a serem evitados em Big Data são:

Correlação equivocada: assumir que duas variáveis estão relacionadas, simplesmente porque elas parecem estar correlacionadas nos dados. É importante lembrar que correlação não implica causalidade.

Overfitting: ajustar um modelo aos dados de treinamento tão bem que ele se torna excessivamente específico para os dados de treinamento e, portanto, não é capaz de generalizar para novos dados.

Amostragem tendenciosa: selecionar um subconjunto de dados que não representa adequadamente a população em questão, levando a conclusões enviesadas.

Ilusão da precisão: assumir que, por causa do grande volume de dados, os resultados são precisos e confiáveis, sem considerar a qualidade dos dados ou os processos de análise.

Falácia da novidade: assumir que algo é verdadeiro ou valioso simplesmente porque é novo ou inovador, sem considerar a evidência empírica ou a validade dos dados.

Viés de confirmação: buscar apenas informações que confirmem as próprias crenças ou hipóteses, ignorando informações que as contradigam.

- **Efeito garoto-propaganda**: assumir que a simples exposição a um grande volume de dados ou tecnologia de Big Data é suficiente para gerar valor ou insights, sem levar em conta a qualidade ou relevância dos dados para o problema em questão.

Padrões Atômicos e Compostos

Padrões atômicos e compostos são conceitos comuns em Big Data que ajudam a identificar **tendências e insights** nos dados.

Padrões atômicos são conjuntos simples de dados que aparecem com frequência nos dados. Eles podem ser simples, como uma palavra comum em um conjunto de dados de texto, ou mais complexos, como um determinado padrão de consumo em dados de transações de cartão de crédito. A identificação desses padrões pode ajudar empresas a entender melhor o comportamento do consumidor, detectar fraudes ou identificar oportunidades de negócios.

Padrões atômicos podem ser subdivididos em padrões de consumo, padrões de processamento, padrões de acesso e padrões de armazenamento.

Já os **padrões compostos** são conjuntos mais complexos de dados que envolvem a interação entre vários padrões atômicos. Eles podem ser usados para descobrir correlações entre

diferentes conjuntos de dados, como a relação entre o clima e as vendas em uma loja de varejo. A identificação desses padrões pode ajudar as empresas a prever tendências futuras e tomar decisões mais informadas.

Por exemplo, imagine uma empresa de varejo que coleta dados de vendas, dados de estoque e dados de clima. Eles podem usar Big Data para identificar um padrão composto que mostra que as vendas de sorvetes estão correlacionadas com dias quentes e estoques de sorvetes baixos. Com essa informação, a empresa pode ajustar o estoque de sorvetes para atender à demanda e melhorar as vendas.

Vamos detalhar um pouco mais os padrões atômicos de Consumo, Processamento, Acesso e Armazenamento.

Padrões de Consumo

Referem-se a **como os usuários interagem e utilizam as informações** geradas por grandes volumes de dados. Esses padrões podem ser divididos em vários subtipos, como visualização, descobertas ad-hoc, aumento do armazenamento de dados tradicionais, notificações e início de resposta automatizada.

A tabela abaixo descreve cada um desses subtipos:

Subtipo	Descrição
Visualização	**Refere-se à maneira como os usuários visualizam e exploram os dados de Big Data. Isso pode incluir dashboards, gráficos e outros tipos de visualizações de dados.**
Descobertas ad-hoc	**Refere-se à capacidade de os usuários fazerem perguntas e descobrirem insights à medida que interagem com os dados de Big Data. Isso pode incluir análises exploratórias e ferramentas de descoberta de dados.**
Aumento do armazenamento de dados tradicionais	**Refere-se à capacidade de expandir os sistemas tradicionais de armazenamento de dados para lidar com grandes volumes de dados gerados por Big Data. Isso pode incluir a utilização de tecnologias como Hadoop e NoSQL.**
Notificações	**Refere-se à capacidade de os usuários serem notificados automaticamente sobre eventos ou tendências importantes nos dados de Big Data. Isso pode incluir alertas por e-mail ou mensagens de texto.**

Início de resposta automatizada	**Refere-se à capacidade de as empresas implementarem ações automatizadas com base nos dados de Big Data. Isso pode incluir sistemas de recomendação, chatbots e outras soluções automatizadas.**

Padrões de Processamento

Referem-se aos **métodos e técnicas utilizados para processar e analisar grandes volumes de dados**. Esses padrões podem ser divididos em vários subtipos, como análise de dados históricos, análises avançadas, pré-processamento de dados brutos e análises ad-hoc. A tabela abaixo descreve cada um desses subtipos:

Subtipo	Descrição
Análise de dados históricos	**Refere-se à análise de grandes volumes de dados históricos para obter insights sobre tendências e padrões. Isso pode incluir análise de dados de vendas, tráfego de rede, registros médicos e outras fontes de dados históricos.**
Análises avançadas	**Refere-se à aplicação de técnicas de análise mais avançadas, como machine learning e análise preditiva, para obter insights mais sofisticados dos dados de Big Data.**
Pré-processamento de dados brutos	**Refere-se às técnicas utilizadas para limpar e preparar dados brutos antes de serem analisados. Inclui limpeza de dados, normalização de dados e outras técnicas de pré-processamento.**
Análises ad-hoc	**Refere-se à capacidade de realizar análises personalizadas e exploratórias em grandes volumes de dados de maneira rápida e eficiente. Inclui a utilização de ferramentas de visualização de dados e outras ferramentas de análise.**

Padrões de Acesso

Referem-se aos métodos e técnicas utilizados **para acessar e extrair dados** de fontes diversas. Esses padrões podem ser divididos em vários subtipos, como dados da web e mídias sociais, dados de dispositivos e dados de data warehouse, transacionais e operacionais. A tabela abaixo descreve cada um desses subtipos:

Subtipo	Descrição
Dados da web e mídias sociais	**Refere-se a dados gerados por usuários em redes sociais, como Twitter e Facebook, além de dados de blogs, fóruns e outras fontes da web. Esses dados são frequentemente não estruturados e podem exigir ferramentas de processamento de linguagem natural para análise.**
Dados de dispositivos	**Refere-se a dados gerados por dispositivos IoT (Internet of Things), como sensores, medidores e outros dispositivos conectados à Internet. Esses dados podem ser estruturados ou não estruturados e exigir técnicas de processamento especializado para análise.**
Dados de data warehouse, transacionais e operacionais	**Refere-se a dados gerados por aplicativos de negócios, como transações bancárias, registros de pacientes em hospitais, registros de compras em lojas, entre outros. Esses dados geralmente são estruturados e armazenados em bancos de dados relacionais, data warehouses ou sistemas operacionais.**

Padrões de Armazenamento

Referem-se à maneira **como os dados são armazenados e gerenciados**. Esses padrões podem ser divididos em vários subtipos, como dados estruturados e distribuídos, dados não estruturados e distribuídos, dados tradicionais e dados em nuvem. A tabela abaixo descreve cada um desses subtipos:

Subtipo	Descrição
Dados estruturados e distribuídos	**Refere-se a dados que são armazenados em um formato estruturado, como tabelas em um banco de dados relacional, e são distribuídos em várias máquinas para permitir a escalabilidade e redundância de dados. Esses dados geralmente são gerenciados por tecnologias de banco de dados distribuídos, como Hadoop ou Cassandra.**
Dados não estruturados e distribuídos	**Refere-se a dados que não possuem uma estrutura definida, como documentos, vídeos e áudios. Esses dados também são distribuídos em várias máquinas para escalabilidade e redundância e geralmente são gerenciados por tecnologias de armazenamento de dados não estruturados, como Hadoop ou MongoDB.**
Dados tradicionais	**Refere-se a dados que são armazenados em sistemas tradicionais de gerenciamento de bancos de dados, como Oracle, SQL Server ou MySQL. Esses dados geralmente são estruturados e altamente organizados, mas podem ser limitados em termos de escalabilidade e gerenciamento de grandes volumes de dados.**

Dados em nuvem	**Refere-se a dados que são armazenados em servidores remotos e acessados pela Internet. Esses dados podem ser estruturados ou não estruturados e são gerenciados por provedores de serviços em nuvem, como Amazon Web Services (AWS), Microsoft Azure e Google Cloud.**

Fluxo de Big Data ("Big Data Pipeline")

Big Data Pipeline é uma arquitetura utilizada para coletar, processar e armazenar grandes volumes de dados, com o objetivo de transformá-los em informações úteis e acionáveis. Ele é composto por três principais componentes:

Computação: refere-se ao processamento dos dados. Nessa etapa, os dados são processados em grandes clusters de servidores, utilizando tecnologias como Hadoop, Spark, entre outras. Nessa etapa, também é possível realizar a transformação e a limpeza dos dados.

Mensagens: refere-se ao fluxo dos dados. Nessa etapa, os dados são transportados entre as diferentes etapas do pipeline. Tecnologias como o Apache Kafka e o Amazon Kinesis são utilizadas para gerenciar o fluxo de dados entre as diferentes etapas do pipeline.

Armazenamento: refere-se ao armazenamento dos dados. Nessa etapa, os dados são armazenados em bancos de dados distribuídos, como HBase, Cassandra, MongoDB, entre outros. Esses bancos de dados são otimizados para armazenar grandes volumes de dados e suportar a escalabilidade horizontal.

Veja alguns recursos importantes que permitem o destaque de um Big Data Pipeline:

Características	Descrição
Arquitetura baseada em nuvem escalável	**O Pipeline é construído em uma arquitetura baseada em nuvem, que permite que os recursos sejam escalados sob demanda, de acordo com as necessidades de processamento dos dados. Isso significa que é possível lidar com grandes volumes de dados sem a necessidade de investimentos em infraestrutura física.**
Arquitetura tolerante a falhas	**A arquitetura leva em conta possíveis falhas no processamento dos dados, com redundância e replicação de informações em diferentes etapas do pipeline, garantindo a confiabilidade dos dados.**
Transformando altos volumes de dados	**Capacidade de lidar com grandes volumes de dados, utilizando tecnologias como Hadoop, Spark e outras ferramentas para processar e armazenar dados em larga escala.**
Análises em tempo real e processamento de dados	**Permite realizar análises em tempo real, com processamento de dados em alta velocidade, possibilitando a tomada de decisões baseada em informações em tempo hábil.**

Gerenciamento de Autoatendimento	**Permite que os usuários possam gerenciar e monitorar seus próprios pipelines de dados, sem a necessidade de suporte de TI, o que traz mais agilidade e autonomia para os usuários.**
Desenvolvimento simplificado de pipeline de dados	**Oferece ferramentas e recursos para o desenvolvimento simplificado de pipelines de dados, possibilitando a criação de novos pipelines de maneira rápida e fácil.**
Processamento Exactly-Once	**O Pipeline garante o processamento "Exactly-Once" dos dados, ou seja, garante que os dados serão processados apenas uma vez, evitando a duplicação de informações e mantendo a integridade dos dados.**

MapReduce

MapReduce é um **modelo de programação e um sistema de processamento de dados distribuído**, projetado para processar grandes conjuntos de dados em clusters de computadores. O nome vem da ideia de que ele divide o processamento em duas fases: a fase de mapeamento e a fase de redução.

Na fase de mapeamento, os dados são divididos em partes menores e cada parte é processada de forma independente em um nó do cluster de computadores. Cada nó aplica uma função de mapeamento aos dados para transformá-los em pares chave-valor.

Exemplo de MapReduce - Vendas de Produtos

Vamos supor que temos um conjunto de dados com informações de vendas de uma loja, contendo informações como data da venda, valor total da venda, produtos vendidos, entre outras.

Para utilizar o MapReduce nesse conjunto de dados, primeiro precisamos dividir o conjunto em blocos menores que possam ser processados separadamente. Suponha que dividimos em blocos de 1 mês cada.

Em seguida, o MapReduce funciona em duas etapas principais: Map e Reduce.

Na etapa Map, são realizadas transformações nos dados de cada bloco. Por exemplo, poderíamos querer contar quantas vendas foram realizadas em cada mês. Então, o Map seria responsável por pegar cada bloco de dados e transformá-lo em pares chave-valor, onde a chave é o mês e o valor é 1 (representando uma venda).

Na etapa Reduce, os dados são agrupados e processados. No exemplo da contagem de vendas por mês, o Reduce seria responsável por somar todos os valores (1) associados a cada chave (mês).

Assim, no final teríamos um conjunto de dados com a contagem de vendas por mês. Esse processo pode ser repetido para outros tipos de análises, como o cálculo de receita total por produto ou a identificação dos produtos mais vendidos.

Exemplo de MapReduce - Contagem de Palavras

Vamos considerar um cenário onde temos um texto muito grande e precisamos contar quantas vezes cada palavra aparece nele. Para isso, podemos utilizar o MapReduce.

Primeiramente, o MapReduce irá dividir o texto em pedaços menores, que chamamos de "splits". Cada split será processado por um nó do cluster. O nó irá receber o split e, para cada palavra presente nele, irá emitir uma chave-valor, onde a chave é a palavra e o valor é 1.

Por exemplo, se o split contiver o texto "O rato roeu a roupa do rei de Roma", o nó irá emitir as seguintes chaves-valores:

- ("O", 1)

- ("rato", 1)

- ("roeu", 1)

- ("a", 1)

- ("roupa", 1)

- ("do", 1)

- ("rei", 1)

- ("de", 1)

- ("Roma", 1)

Em seguida, todas essas chaves-valores serão enviadas para um nó específico, chamado "nó de redução". O nó de redução irá receber todas as chaves-valor emitidas pelos nós de map e irá somar os valores correspondentes a cada chave, gerando um novo conjunto de chaves-valor, onde a chave é a palavra e o valor é a contagem total de ocorrências.

Por exemplo, se os nós de map emitirem as chaves-valores acima, o nó de redução irá gerar o seguinte conjunto de chaves-valores:

- ("O", 1)

- ("rato", 1)

- ("roeu", 1)

- ("a", 1)

- ("roupa", 1)

- ("do", 1)

- ("rei", 1)

- ("de", 1)

- ("Roma", 1)

Note que, nesse exemplo, cada palavra apareceu apenas uma vez no texto, então as contagens são todas iguais a 1. Em um texto real, é claro, algumas palavras apareceriam várias vezes e outras não apareceriam nenhuma vez. Mas o processo seria o mesmo: o MapReduce divide o texto em splits, conta as ocorrências de cada palavra em cada split e, em seguida, soma todas essas contagens para obter a contagem total de cada palavra no texto.

Então, recapitulando, MapReduce é uma técnica eficaz para lidar com grandes quantidades de dados e é amplamente utilizado em plataformas de Big Data, como o Hadoop. Ele permite que os dados sejam processados de forma distribuída, o que pode acelerar o tempo de processamento em comparação com sistemas que processam os dados de forma sequencial em um único computador.

Uma limitação do MapReduce é que ele não é adequado para processar dados em tempo real, sendo mais indicado para tarefas batch (em lote).

Nas próximas seções, vamos estudar o conceito de Bases de Dados NoSQL.

A relação entre NoSQL e Big Data é muito estreita, pois as Bases de Dados NoSQL são frequentemente usadas para lidar com grandes volumes de dados, que são característicos do Big Data. NoSQL foi criado para fornecer uma alternativa escalável e flexível às Bases de

Dados Relacionais, que muitas vezes têm dificuldades em lidar com grandes quantidades de dados e em manter um desempenho aceitável

Bases de Dados NoSQL são uma escolha natural para projetos de Big Data, pois são capazes de lidar com grandes volumes de dados não estruturados, com alta disponibilidade e escalabilidade, e com desempenho de leitura e gravação eficientes.

NoSQL

Uma Base de Dados NoSQL é um tipo de base de dados que difere das bases de dados relacionais tradicionais, que utilizam a linguagem SQL (Structured Query Language), para manipulação dos dados. NoSQL significa "Not Only SQL", ou seja, não apenas SQL, o que significa que não utiliza a linguagem SQL para armazenar e recuperar dados.

Bases NoSQL são projetadas para lidar com **grandes volumes de dados, com alta disponibilidade e escalabilidade**, com a capacidade de lidar com dados não estruturados ou semiestruturados, como dados de sensores, logs, vídeos, imagens e documentos. Elas podem ser usadas para armazenar e manipular dados de diferentes tipos, incluindo dados relacionais e não relacionais.

Existem diferentes tipos de Bases de Dados NoSQL, incluindo bases de dados orientadas a documentos, bases de dados chave-valor, bases de dados orientadas a grafos e bases de dados orientadas a colunas. Cada tipo de Base de Dados NoSQL tem suas próprias características e é adequado para diferentes casos de uso, dependendo dos requisitos de armazenamento e acesso aos dados.

Em geral, Bases de Dados NoSQL são mais adequadas para aplicações que precisam lidar com grandes volumes de dados, com alta disponibilidade e escalabilidade, e com requisitos de flexibilidade de esquema e desempenho.

Bases NoSQL versus Bases Relacionais

Existem várias diferenças importantes entre Bases de Dados NoSQL e Bases de Dados Relacionais. Vejamos as principais:

Modelo de dados:

Bases de Dados Relacionais utilizam o modelo de dados relacional, em que os dados são organizados em tabelas com colunas e linhas que representam relacionamentos entre as

tabelas. Já as Bases de Dados NoSQL utilizam modelos de dados diferentes, como modelos orientados a documentos, modelos chave-valor, modelos orientados a grafos e modelos orientados a colunas.

Escalabilidade:

Bases de Dados Relacionais são projetadas para serem escalonadas verticalmente, ou seja, aumentando o poder de processamento de um único servidor. Já as Bases de Dados NoSQL são projetadas para serem escalonadas horizontalmente, ou seja, adicionando mais servidores à medida que a demanda por armazenamento e processamento aumenta.

Flexibilidade de esquema:

As Bases de Dados Relacionais têm um esquema rígido e pré-definido para seus dados, e qualquer alteração no esquema requer uma modificação na estrutura da tabela. Já as Bases de Dados NoSQL permitem uma maior flexibilidade de esquema, permitindo a adição ou remoção de campos sem precisar modificar toda a estrutura de tabela.

Consistência e durabilidade:

Bases de Dados Relacionais são projetadas para garantir a consistência dos dados, seguindo o modelo ACID (Atomicidade, Consistência, Isolamento e Durabilidade). Já as Bases de Dados NoSQL geralmente não suportam transações ACID, mas oferecem consistência eventual ou consistência forte para garantir a durabilidade dos dados.

Complexidade de consulta:

Bases de Dados Relacionais oferecem uma linguagem padrão para consultas, a SQL, que é bem desenvolvida e padronizada. Já as Bases de Dados NoSQL geralmente não têm uma linguagem padrão, e as consultas podem ser feitas de maneiras diferentes, dependendo do tipo de base de dados.

Custo:

Bases de Dados Relacionais são geralmente mais caras em termos de hardware, licenças de software e custos de manutenção, enquanto as Bases de Dados NoSQL são geralmente menos caras e podem ser executadas em hardware menos sofisticado.

No geral, Bases de Dados NoSQL são mais flexíveis e escaláveis do que as Bases de Dados Relacionais, mas podem oferecer menos recursos para garantir a consistência dos dados. As Bases de Dados Relacionais, por outro lado, são mais rígidas e geralmente mais adequadas para projetos com esquemas de dados bem definidos, mas podem ser mais caras e difíceis de escalar.

Vantagens de NoSQL

Existem várias razões pelas quais uma pessoa ou organização pode optar por usar uma Base de Dados NoSQL em vez de uma Base de Dados Relacional. Algumas das principais razões incluem:

Escalabilidade:

Bases de Dados NoSQL são projetadas para serem escaláveis horizontalmente, permitindo que elas lidem com grandes quantidades de dados e tráfego da aplicação. Isso as torna uma escolha popular para empresas que precisam lidar com grandes volumes de dados ou tráfego da aplicação.

Flexibilidade de esquema:

Existe a possibilidade de uma maior flexibilidade de esquema, permitindo que os dados sejam armazenados em formatos diferentes e alterados facilmente, sem a necessidade de modificar a estrutura de tabela. Isso é particularmente útil em projetos em que os dados são variáveis ou incertos.

Desempenho:

Bases de Dados NoSQL geralmente oferecem um melhor desempenho do que as Bases de Dados Relacionais em certas áreas, como escrita em lote e leitura de dados não estruturados. Isso torna as Bases de Dados NoSQL uma escolha popular para projetos em que o desempenho é uma prioridade.

Custo:

Geralmente, há uma maior economia no uso de bases de dados NoSQL em relação a Bases de Dados Relacionais em termos de hardware, licenças de software e custos de manutenção. Isso as torna uma escolha popular para empresas que precisam lidar com grandes quantidades de dados, mas têm orçamentos limitados.

Suporte para dados não estruturados:

Bases NoSQL são particularmente úteis para armazenar dados não estruturados ou semiestruturados, como logs, dados de sensores, mídia social e dados de streaming. As Bases de Dados Relacionais não são tão adequadas para lidar com esses tipos de dados.

Tipos de Bases NoSQL

Existem diferentes tipos de bases de dados NoSQL, sendo as principais: Base de Dados Orientada a Documentos, Base de Dados Chave-Valor, Base de Dados Orientada a Grafos e Base de Dados Orientada a Colunas.

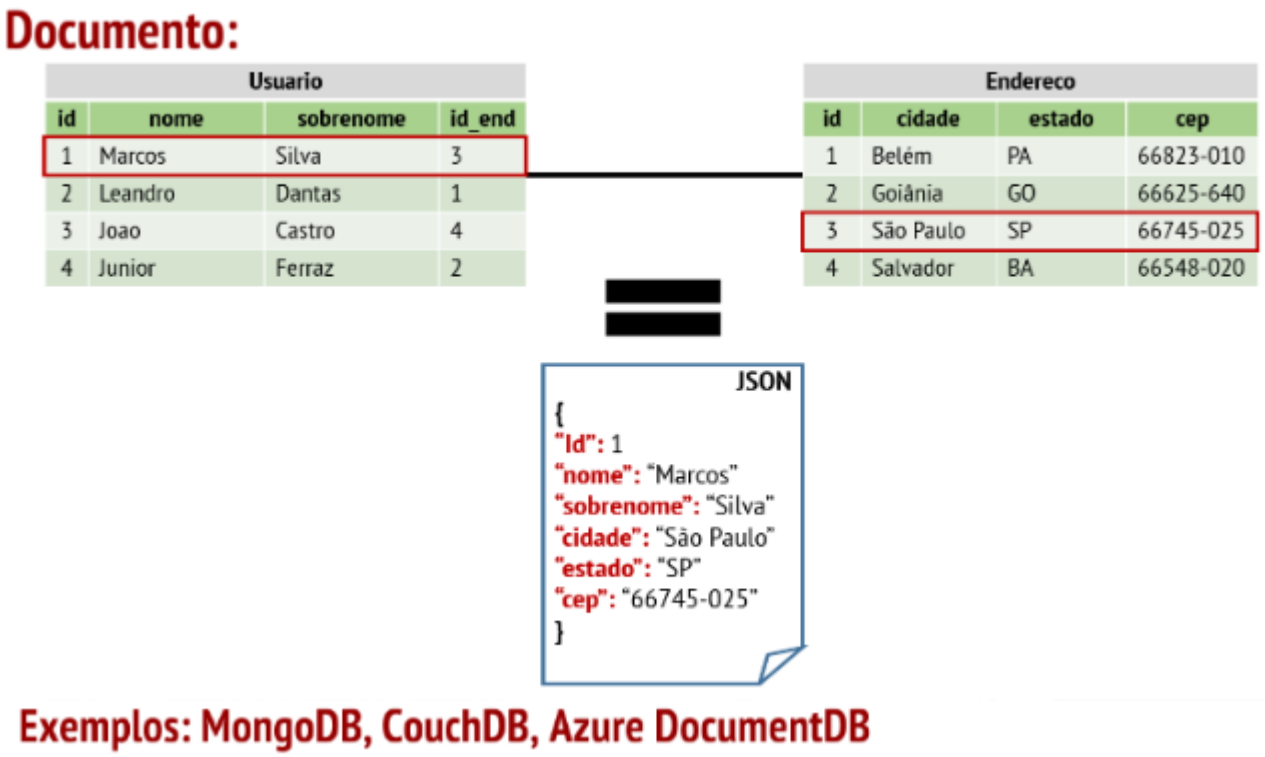
Vamos estudar em maiores detalhes cada uma delas.

Modelo Orientado a Documentos

Uma base de dados NoSQL orientada a documentos é um tipo de base de dados NoSQL que armazena dados em documentos no formato JSON ou BSON, em vez de em tabelas com esquemas fixos, como as bases de dados relacionais. Cada documento é um registro autônomo que pode conter um número variável de campos e valores. Esses documentos podem ser aninhados e combinados, permitindo uma maior flexibilidade de esquema.

Cada documento é geralmente identificado por um ID exclusivo, que pode ser usado para recuperar o documento ou realizar operações de atualização ou exclusão.

Alguns exemplos de bases de dados NoSQL orientadas a documentos incluem: MongoDB, Couchbase, Amazon DocumentDB e Apache Cassandra.



Modelo Chave-Valor

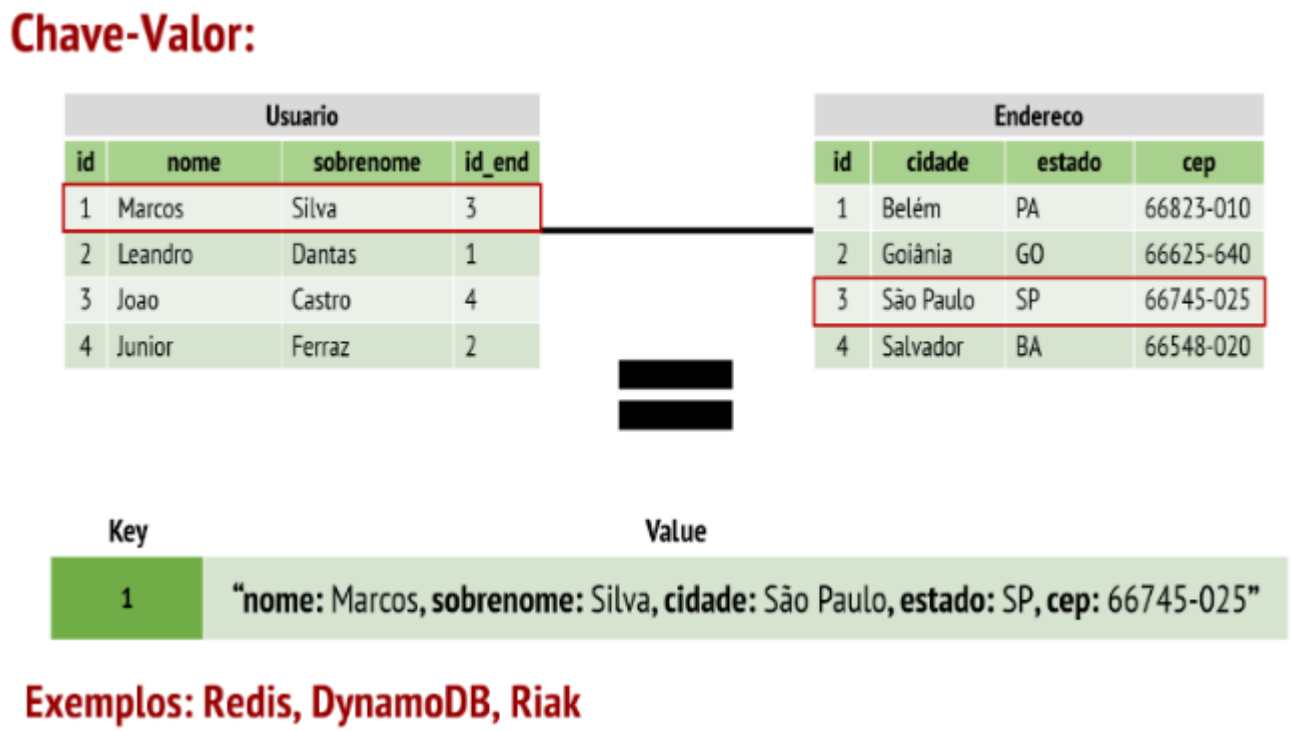
Uma base de dados NoSQL chave-valor é um tipo de banco de dados que armazena dados como um conjunto de pares de chave-valor. Cada valor é associado a uma chave exclusiva, e os dados são armazenados sem uma estrutura de tabela ou esquema pré-definido.

Os dados são armazenados em uma estrutura simples e otimizada para operações de gravação e leitura de dados em alta velocidade. As operações básicas de chave-valor incluem obter, inserir e excluir dados.

Esse tipo de base é frequentemente usado para armazenar informações em cache, sessões de usuários, listas de classificação, informações de perfil e outras aplicações que requerem armazenamento e recuperação de dados em alta velocidade.

Alguns exemplos de bases de dados NoSQL chave-valor incluem: Redis, Riak, Amazon DynamoDB e Oracle NoSQL Database.

A figura abaixo apresenta um exemplo de um banco de dados que armazena informações pessoais no formato chave-valor. A chave representa um campo como o nome, enquanto o valor representa a instância do correspondente.



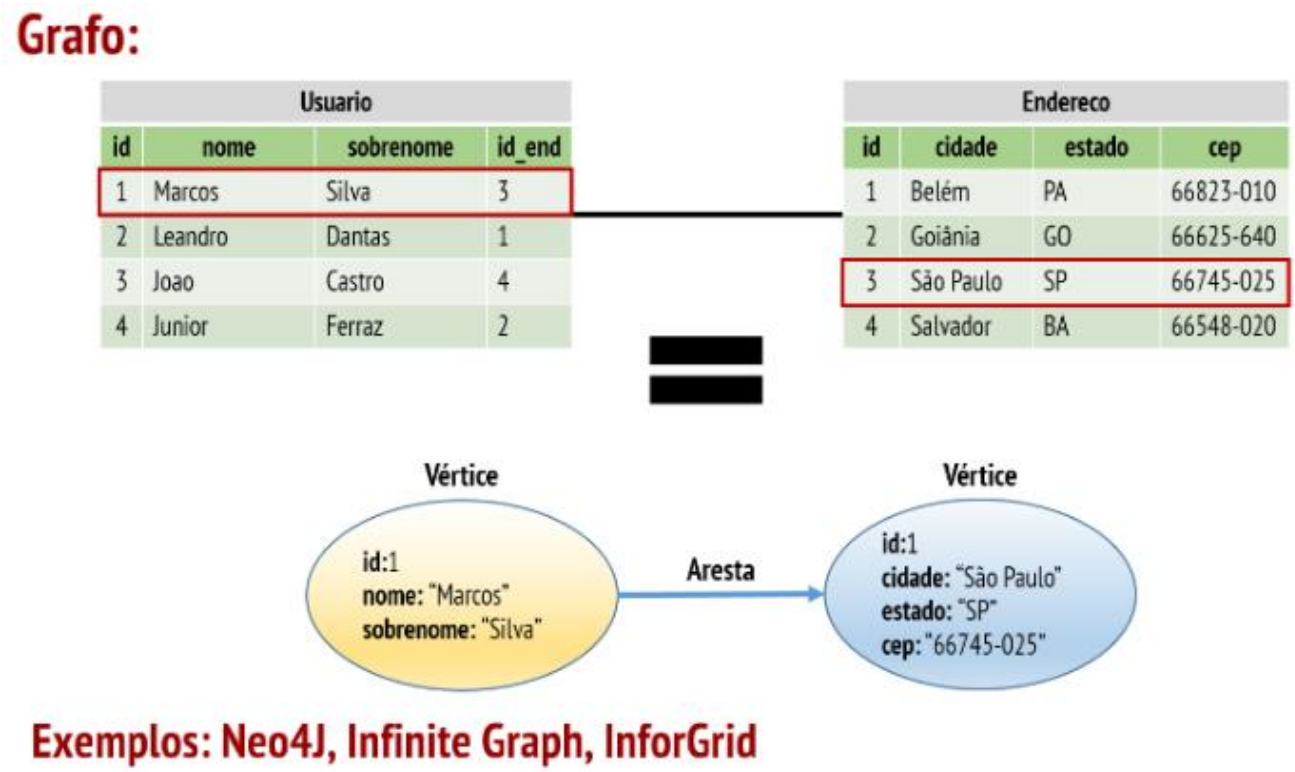
Modelo Orientado a Grafos

Uma base de dados NoSQL orientada a grafos é um tipo de banco de dados que armazena e gerencia dados em forma de grafos, que consistem em nós (vértices) conectados por arestas (relacionamentos). Os nós representam entidades ou objetos e as arestas representam as relações entre eles.

Essa estrutura de dados permite que a base de dados orientada a grafos seja altamente eficiente na modelagem e consulta de dados altamente interconectados, como redes sociais, sistemas de recomendação, análises de fraudes, entre outros.

Cada nó no banco de dados orientado a grafos pode ter um ou mais rótulos para indicar o tipo de objeto que ele representa, e cada aresta pode ter uma ou mais propriedades para descrever a relação entre os nós. Os dados são armazenados de forma flexível, sem a necessidade de um esquema rígido, o que permite que o banco de dados possa lidar com dados altamente variáveis.

Alguns exemplos de bases de dados NoSQL orientadas a grafos incluem: Neo4j, Infinite Graph, Amazon Neptune e ArangoDB.



Modelo Orientado a Colunas

Uma base de dados NoSQL orientada a colunas é um tipo de banco de dados que **armazena dados em colunas em vez de linhas** como em bancos de dados relacionais. Essa estrutura

permite que as consultas sejam executadas de forma eficiente em grandes conjuntos de dados e permite uma escalabilidade horizontal muito alta.

Cada registro de dados é armazenado como uma série de colunas, e as colunas são agrupadas em famílias de colunas que compartilham características comuns. As famílias de colunas podem ser organizadas de acordo com a necessidade do aplicativo e da carga de trabalho, permitindo que os usuários acessem rapidamente as informações necessárias.

Um exemplo comum de uma base de dados NoSQL orientada a colunas é o Apache Cassandra. Ele é amplamente utilizado para armazenar grandes quantidades de dados distribuídos em vários servidores, mantendo a alta disponibilidade e tolerância a falhas. Ele é usado por empresas como Netflix, eBay, Twitter e várias outras.

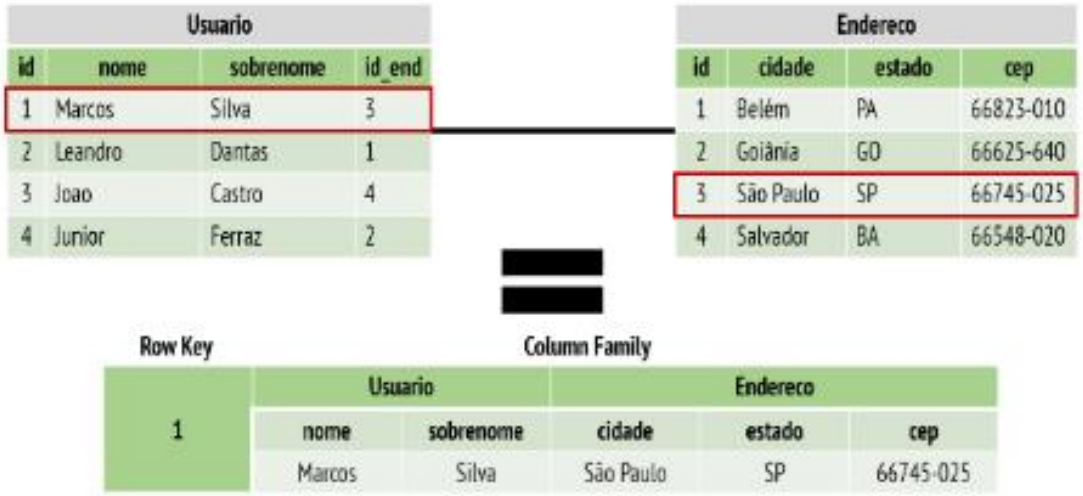
Algumas das principais características de uma base de dados NoSQL orientada a colunas incluem:

- Escalabilidade horizontal: permite que os usuários adicionem mais nós ao cluster para aumentar o poder de processamento e armazenamento à medida que a demanda cresce.
- Alta disponibilidade e tolerância a falhas: os dados são replicados em vários nós para garantir que, mesmo que um nó falhe, os dados ainda estejam disponíveis.
- Desempenho: as consultas podem ser executadas rapidamente, mesmo em grandes conjuntos de dados, porque as colunas são armazenadas juntas e podem ser lidas de forma eficiente.
- Flexibilidade: os esquemas das colunas podem ser facilmente modificados sem afetar outros dados na base de dados.

No entanto, as bases orientadas a colunas têm algumas desvantagens, como a falta de suporte a transações ACID completas e a necessidade de projetar cuidadosamente o esquema de colunas para garantir um desempenho ideal.

Veja que, neste caso, mudamos o paradigma em relação ao modelo chave-valor. A orientação deixa de ser por registros ou tuplas para orientação por colunas. Neste modelo os dados são indexados por uma trilha (linha, coluna e timestamp), onde linhas e colunas são identificadas por chaves e o timestamp permite diferenciar múltiplas versões de um mesmo dado.

Coluna:



Exemplos: Cassandra, Hadoop, Hypertable

A tabela a seguir resume os tipos de bases NoSQL:

Base NoSQL	Descrição	Vantagens	Desvantagens
Orientada a Documentos	**Armazena dados como documentos individuais, geralmente em formato JSON ou BSON**	**Flexibilidade no esquema de dados. Suporte a dados semiestruturados. Escalabilidade horizontal**	**Complexidade na modelagem de dados. Baixo desempenho em consultas complexas**
Chave-Valor	**Armazena dados em um modelo de chave-valor simples**	**Alta velocidade em leitura e gravação. Escalabilidade horizontal. Facilidade de uso e implementação**	**Limitado a operações simples de gravação/leitura. Dificuldade em modelar dados complexos**
Orientada a Grafos	**Armazena dados em nós e arestas, permitindo a modelagem de relacionamentos complexos**	**Flexibilidade na modelagem de dados. Suporte a consultas complexas em grafos. Escalabilidade horizontal**	**Dificuldade em modelar dados não relacionais. Custo computacional alto em consultas complexas**
Orientada a Colunas	**Armazena dados em colunas em vez de linhas, permitindo o processamento eficiente de grandes conjuntos de dados**	**Escalabilidade horizontal. Alta disponibilidade e tolerância a falhas. Desempenho rápido em consultas**	**Dificuldade na modelagem de dados- Falta de suporte a transações ACID completas. Necessidade de projetar cuidadosamente o esquema de colunas**

APOSTA ESTRATÉGICA

A ideia desta seção é apresentar os pontos do conteúdo que mais possuem chances de serem cobrados em prova, considerando o histórico de questões da banca em provas de nível semelhante à nossa, bem como as inovações no conteúdo, na legislação e nos entendimentos doutrinários e jurisprudenciais[1].

O NoSQL, abreviação de Not Only SQL, é um tipo de banco de dados não relacional que difere dos bancos de dados tradicionais, como o SQL, por sua capacidade de armazenar e manipular grandes volumes de dados de forma distribuída e escalável. Esse modelo de banco de dados é especialmente adequado para lidar com os desafios do Big Data, pois permite o armazenamento de dados não estruturados e semiestruturados, como documentos, gráficos e dados de mídia. Além disso, o NoSQL oferece uma flexibilidade maior em relação ao esquema de dados, permitindo alterações mais rápidas e sem a necessidade de estruturas fixas, o que é crucial em ambientes dinâmicos e em evolução constante.

Entre as vantagens do NoSQL estão a capacidade de escalabilidade horizontal, que permite adicionar mais servidores conforme a demanda de dados aumenta, e a alta disponibilidade, garantindo que os dados estejam sempre acessíveis mesmo em caso de falhas de hardware ou software. Outro benefício é a capacidade de lidar com grandes volumes de dados de forma eficiente, reduzindo a latência e melhorando o desempenho das consultas. Essas vantagens tornam o NoSQL uma escolha poderosa para empresas que lidam com grandes quantidades de dados e precisam de flexibilidade e desempenho em seus sistemas de gerenciamento de dados.

QUESTÕES ESTRATÉGICAS

Nesta seção, apresentamos e comentamos uma amostra de questões objetivas selecionadas estrategicamente: são questões com nível de dificuldade semelhante ao que você deve esperar para a sua prova e que, em conjunto, abordam os principais pontos do assunto.

A ideia, aqui, não é que você fixe o conteúdo por meio de uma bateria extensa de questões, mas que você faça uma boa revisão global do assunto a partir de, relativamente, poucas questões.

1. (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

[1] Vale deixar claro que nem sempre será possível realizar uma aposta estratégica para um determinado assunto, considerando que às vezes não é viável identificar os pontos mais prováveis de serem cobrados a partir de critérios objetivos ou minimamente razoáveis.

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: C

---

2. (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: C

3. **(FCC / Copergás-PE - 2023)** Na era do Big Data, as empresas precisam utilizar repositórios e tecnologias para armazenamento, tratamento e análise desse grande volume de dados, dentre as quais, encontram-se:

a) os sistemas OLAP, que têm foco no nível operacional da organização, proporcionando alta velocidade na consulta, inserção, alteração e exclusão de dados. Os dados, voláteis, são armazenados em SGBDs relacionais e a sua atualização é feita no momento da transação.

b) os Data Warehouses, que proporcionam alto desempenho para consulta de dados históricos. Os dados são obtidos de diversas fontes, passam por um processo de ETL e geralmente são armazenados em um modelo dimensional como Star e Snowflake.

c) as ferramentas OLAP, que utilizam a técnica de argumentação ativa, isto é, ao invés de o gestor definir o problema, selecionar os dados e as ferramentas para analisá-los, as ferramentas OLAP pesquisam automaticamente a base de dados à procura de anomalias e possíveis relacionamentos, identificando problemas que não tinham sido detectados pelo gestor.

d) as análises de BI, que utilizam algoritmos estatísticos e técnicas de machine learning para identificar a probabilidade da ocorrência de resultados futuros a partir de dados históricos armazenados em data lakes.

e) as ferramentas de Data Mining, que combinam os melhores elementos de Data Lakes e de Data Warehouses, formando um sistema aberto e padronizado, capaz de estruturar os dados e os recursos de gerenciamento de dados, de forma a proporcionar ao gestor uma exploração detalhada dos dados em busca de informações para a tomada de decisão.

Comentários:

(a) Errado, a descrição do item trata de Sistemas OLTP e, não, OLAP; (b) Correto; (c) Errado, que item viajante – não tem nada correto; (d) Errado, a descrição está correta, mas isso não tem nenhuma relação com repositórios ou tecnologias de armazenamento; (e) Errado, outra questão viajante – Data Mining não tem relação com repositórios ou tecnologias de armazenamento.

Gabarito: Letra B

4. **(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

5. **(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

QUESTIONÁRIO DE REVISÃO E APERFEIÇOAMENTO

A ideia do questionário é elevar o nível da sua compreensão no assunto e, ao mesmo tempo, proporcionar uma outra forma de revisão de pontos importantes do conteúdo, a partir de perguntas que exigem respostas subjetivas.

São questões um pouco mais desafiadoras, porque a redação de seu enunciado não ajuda na sua resolução, como ocorre nas clássicas questões objetivas.

O objetivo é que você realize uma auto explicação mental de alguns pontos do conteúdo, para consolidar melhor o que aprendeu ;)

Além disso, as questões objetivas, em regra, abordam pontos isolados de um dado assunto. Assim, ao resolver várias questões objetivas, o candidato acaba memorizando pontos isolados do conteúdo, mas muitas vezes acaba não entendendo como esses pontos se conectam.

Assim, no questionário, buscaremos trazer também situações que ajudem você a conectar melhor os diversos pontos do conteúdo, na medida do possível.

É importante frisar que não estamos adentrando em um nível de profundidade maior que o exigido na sua prova, mas apenas permitindo que você compreenda melhor o assunto de modo a facilitar a resolução de questões objetivas típicas de concursos, ok?

Nosso compromisso é proporcionar a você uma revisão de alto nível!

Vamos ao nosso questionário:

Perguntas

1. O que é Big Data?

2. Quais são as seis principais características do Big Data?

3. O que é um Big Data Pipeline?

4. Quais são os três componentes de um Big Data Pipeline?

5. Quais são os subtipos de padrões de consumo no contexto de Big Data?

6. Quais são os subtipos de padrões de processamento no contexto de Big Data?

7. Quais são os subtipos de padrões de acesso no contexto de Big Data?

8. Quais são os subtipos de padrões de armazenamento no contexto de Big Data?

9. O que é MapReduce e como ele pode ser utilizado no processamento de grandes conjuntos de dados?

10. O que é uma base de dados NoSQL e como ela se diferencia de uma base de dados relacional?

11. Quais são os principais tipos de bases de dados NoSQL e como eles se diferenciam?

12. O que é uma base de dados NoSQL orientada a documentos e como ela funciona?

13. O que é uma base de dados NoSQL chave-valor e como ela funciona?

14. O que é uma base de dados NoSQL orientada a grafos e como ela funciona?

15. O que é uma base de dados NoSQL orientada a colunas e como ela funciona?

16. Quais são as principais vantagens de utilizar uma base de dados NoSQL em relação a uma base de dados relacional?

17. Quais são as principais desvantagens de utilizar uma base de dados NoSQL em relação a uma base de dados relacional?

18. Quais são as principais bases de dados NoSQL?

19. O que é a variedade de dados no contexto de Big Data?

20. O que é o conceito de veracidade de dados em Big Data?

21. O que é o conceito de valor em Big Data?

22. Quais são os principais desafios enfrentados ao lidar com Big Data?

23. Como o conceito de escalabilidade se aplica em projetos de Big Data?

24. Quais são as principais tecnologias e ferramentas utilizadas em projetos de Big Data?

25. Como o processamento em tempo real é relevante para projetos de Big Data?

Perguntas e Respostas

1. O que é Big Data?

Resposta: Big Data é um termo usado para descrever conjuntos de dados muito grandes e complexos que exigem tecnologias específicas para armazenamento, processamento e análise.

2. Quais são as seis principais características do Big Data?

Resposta: As seis principais características do Big Data são Volume, Velocidade, Variedade, Veracidade, Variabilidade e Valor.

3. O que é um Big Data Pipeline?

Resposta: Um Big Data Pipeline é um conjunto de tecnologias e processos que permitem coletar, processar, armazenar e analisar grandes quantidades de dados.

4. Quais são os três componentes de um Big Data Pipeline?

Resposta: Os três componentes de um Big Data Pipeline são computação, mensagens e armazenamento.

5. Quais são os subtipos de padrões de consumo no contexto de Big Data?

Resposta: Os subtipos de padrões de consumo em Big Data incluem visualização, descobertas ad-hoc, aumento de armazenamento de dados tradicionais, notificações e início de resposta automatizada.

6. Quais são os subtipos de padrões de processamento no contexto de Big Data?

Resposta: Os subtipos de padrões de processamento em Big Data incluem análise de dados históricos, análises avançadas, pré-processamento de dados brutos e análises ad-hoc.

7. Quais são os subtipos de padrões de acesso no contexto de Big Data?

Resposta: Os subtipos de padrões de acesso em Big Data incluem dados da web e mídias sociais, dados de dispositivos e dados de data warehouse, transacionais e operacionais.

8. Quais são os subtipos de padrões de armazenamento no contexto de Big Data?

Resposta: Os subtipos de padrões de armazenamento em Big Data incluem dados estruturados e distribuídos, dados não estruturados e distribuídos, dados tradicionais e dados em nuvem.

9. O que é MapReduce e como ele pode ser utilizado no processamento de grandes conjuntos de dados?

O MapReduce é um modelo de programação e processamento distribuído de grande escala utilizado para processar e analisar grandes conjuntos de dados. Ele divide as tarefas em dois processos distintos: o Map, que é responsável por filtrar e classificar os dados, e o Reduce, que é responsável por combinar os dados processados e gerar os resultados finais. Essa técnica é utilizada em diversas aplicações de Big Data, como análise de dados de vendas, processamento de logs de servidores e análise de sentimentos em redes sociais.

10. O que é uma base de dados NoSQL e como ela se diferencia de uma base de dados relacional?

Resposta: Uma base de dados NoSQL é uma base de dados que não segue o modelo relacional, permitindo uma maior flexibilidade na estruturação e armazenamento de dados. Ela se diferencia de uma base de dados relacional por não ter uma estrutura fixa de tabelas, colunas e relacionamentos.

11. Quais são os principais tipos de bases de dados NoSQL e como eles se diferenciam?

Resposta: Os principais tipos de bases de dados NoSQL são: orientado a documentos, chave-valor, orientado a grafos e orientado a colunas. Cada um desses tipos se diferencia pela forma como estrutura e armazena os dados.

12. O que é uma base de dados NoSQL orientada a documentos e como ela funciona?

Resposta: Uma base de dados NoSQL orientada a documentos armazena dados em formato de documentos, geralmente em formato JSON ou BSON. Cada documento pode ter uma estrutura diferente, permitindo uma maior flexibilidade na modelagem de dados.

13. O que é uma base de dados NoSQL chave-valor e como ela funciona?

Resposta: Uma base de dados NoSQL chave-valor armazena dados em pares de chave-valor simples, sem uma estrutura definida. Isso permite uma maior flexibilidade na modelagem de dados, mas pode limitar as consultas complexas.

14. O que é uma base de dados NoSQL orientada a grafos e como ela funciona?

Resposta: Uma base de dados NoSQL orientada a grafos armazena dados em forma de grafos, permitindo a representação de relacionamentos complexos entre entidades. Isso permite uma maior flexibilidade na modelagem de dados, mas pode exigir um maior processamento para consultas complexas.

15. O que é uma base de dados NoSQL orientada a colunas e como ela funciona?

Resposta: Uma base de dados NoSQL orientada a colunas armazena dados em colunas, em vez de linhas, permitindo um processamento eficiente de grandes conjuntos de dados. Isso permite consultas complexas e um alto desempenho, mas pode exigir uma maior complexidade na modelagem de dados.

16. Quais são as principais vantagens de utilizar uma base de dados NoSQL em relação a uma base de dados relacional?

Resposta: As principais vantagens de utilizar uma base de dados NoSQL são a maior flexibilidade na modelagem de dados, a escalabilidade horizontal e a capacidade de lidar com grandes volumes de dados em alta velocidade.

17. Quais são as principais desvantagens de utilizar uma base de dados NoSQL em relação a uma base de dados relacional?

Resposta: As principais desvantagens de utilizar uma base de dados NoSQL são a menor maturidade e estabilidade do mercado em comparação com as bases de dados relacionais, a maior complexidade na modelagem de dados e a menor disponibilidade de ferramentas e recursos de suporte.

18. Quais são as principais bases de dados NoSQL?

Resposta: As principais bases de dados NoSQL incluem MongoDB, Cassandra, Redis, Neo4j e Apache HBase. O MongoDB é uma base de dados NoSQL orientada a documentos, o Cassandra é um banco de dados NoSQL chave-valor, o Redis é um banco de dados NoSQL em memória, o Neo4j é uma base de dados NoSQL orientada a grafos e o Apache HBase é um banco de dados NoSQL orientado a colunas. Cada uma dessas bases de dados NoSQL possui características distintas e é adequada para diferentes casos de uso.

19. O que é a variedade de dados no contexto de Big Data?

Resposta: A variedade de dados refere-se à diversidade de tipos e formatos de dados que são coletados e processados em projetos de Big Data. Isso inclui dados estruturados (como tabelas em um banco de dados relacional), dados semiestruturados (como arquivos XML ou JSON) e dados não estruturados (como texto livre, áudio, vídeo).

20. O que é o conceito de veracidade de dados em Big Data?

Resposta: A veracidade de dados diz respeito à qualidade e confiabilidade dos dados em um ambiente de Big Data. É importante garantir que os dados coletados sejam precisos, completos e confiáveis, a fim de evitar erros e tomadas de decisões equivocadas com base em informações incorretas.

21. O que é o conceito de valor em Big Data?

Resposta: O valor em Big Data refere-se à capacidade de extrair insights e informações relevantes a partir dos dados coletados, resultando em benefícios tangíveis para as organizações. O valor pode ser obtido por meio de análises avançadas, identificação de padrões, detecção de tendências, personalização de produtos/serviços e tomada de decisões mais embasadas.

22. Quais são os principais desafios enfrentados ao lidar com Big Data?

Resposta: Alguns dos principais desafios em Big Data incluem gerenciamento e armazenamento de grandes volumes de dados, processamento eficiente de dados em tempo real, garantia da qualidade dos dados, segurança e privacidade, além da necessidade de profissionais especializados em análise e interpretação de dados.

23. Como o conceito de escalabilidade se aplica em projetos de Big Data?

Resposta: A escalabilidade é um aspecto crítico em projetos de Big Data, pois envolve a capacidade de lidar com o aumento do volume, variedade e velocidade dos dados. A infraestrutura de Big Data deve ser dimensionada de forma eficiente, permitindo a expansão conforme necessário para atender às demandas em constante crescimento.

24. Quais são as principais tecnologias e ferramentas utilizadas em projetos de Big Data?

Resposta: Algumas das principais tecnologias e ferramentas usadas em projetos de Big Data incluem Hadoop, Spark, Apache Kafka, NoSQL, linguagens de programação como Python e R, além de soluções de visualização de dados, como Tableau e Power BI.

25. Como o processamento em tempo real é relevante para projetos de Big Data?

Resposta: O processamento em tempo real é relevante para projetos de Big Data porque permite o processamento e análise de dados em tempo real ou quase em tempo real. Isso é fundamental em aplicações que exigem tomadas de decisões imediatas, monitoramento em tempo real, detecção de anomalias e outros casos em que a velocidade de processamento é crítica.

LISTA DE QUESTÕES ESTRATÉGICAS

1. (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.

---

2. (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.

---

3. **(FCC / Copergás-PE - 2023)** Na era do Big Data, as empresas precisam utilizar repositórios e tecnologias para armazenamento, tratamento e análise desse grande volume de dados, dentre as quais, encontram-se:

a) os sistemas OLAP, que têm foco no nível operacional da organização, proporcionando alta velocidade na consulta, inserção, alteração e exclusão de dados. Os dados, voláteis, são armazenados em SGBDs relacionais e a sua atualização é feita no momento da transação.

b) os Data Warehouses, que proporcionam alto desempenho para consulta de dados históricos. Os dados são obtidos de diversas fontes, passam por um processo de ETL e geralmente são armazenados em um modelo dimensional como Star e Snowflake.

c) as ferramentas OLAP, que utilizam a técnica de argumentação ativa, isto é, ao invés de o gestor definir o problema, selecionar os dados e as ferramentas para analisá-los, as ferramentas OLAP pesquisam automaticamente a base de dados à procura de anomalias e possíveis relacionamentos, identificando problemas que não tinham sido detectados pelo gestor.

d) as análises de BI, que utilizam algoritmos estatísticos e técnicas de machine learning para identificar a probabilidade da ocorrência de resultados futuros a partir de dados históricos armazenados em data lakes.

e) as ferramentas de Data Mining, que combinam os melhores elementos de Data Lakes e de Data Warehouses, formando um sistema aberto e padronizado, capaz de estruturar os dados e os recursos de gerenciamento de dados, de forma a proporcionar ao gestor uma exploração detalhada dos dados em busca de informações para a tomada de decisão.

4. **(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.

5. **(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.

Gabaritos

1. C
2. C
3. B
4. D
5. A

Questões Adicionais

As questões apresentadas a seguir integram o Banco de Questões do Passo Estratégico. Recomenda-se utilizá-las como um recurso complementar para a prática e consolidação dos conhecimentos adquiridos no material teórico, de acordo com o estilo adotado pela banca organizadora.

Bom estudo!

1. Dadas as seguintes características de tecnologias de Big Data, associe corretamente cada tecnologia com sua respectiva função: A) NoSQL - permite o armazenamento flexível de dados, B) Apache Hadoop - oferece escalabilidade horizontal para armazenamento e processamento distribuído, C) Apache Spark - facilita o processamento em tempo real de grandes volumes de dados.

A) A: Correto, B: Correto, C: Incorreto
B) A: Correto, B: Correto, C: Correto
C) A: Correto, B: Incorreto, C: Correto
D) A: Incorreto, B: Correto, C: Incorreto
E) A: Incorreto, B: Incorreto, C: Correto

2. Qual é uma das características da Análise de Componentes Principais (PCA) em Big Data?

A) Aumento da dimensionalidade dos dados
B) Redução da dimensionalidade dos dados
C) Ampliação dos dados em tempo real
D) Introdução de dados duplicados
E) Armazenamento de dados em nuvem

3. No contexto de Big Data, a característica de _____ é essencial para lidar com grandes volumes de dados que chegam em tempo real e precisam ser processados rapidamente para gerar insights acionáveis.

A) Volume
B) Velocidade
C) Variedade
D) Veracidade
E) Valor

4. No contexto de Big Data, a característica de _____ refere-se à capacidade de lidar com diferentes tipos e fontes de dados, incluindo dados estruturados, semiestruturados e não estruturados, que exigem diversas técnicas de armazenamento e processamento.

A) Volume
B) Velocidade
C) Variedade

D) Veracidade
E) Valor

5. No modelo MapReduce, a fase de _____ é responsável por combinar os resultados intermediários produzidos na fase anterior, agregando os dados processados para produzir a saída final.

A) Shuffle
B) Map
C) Sort
D) Combine
E) Reduce

6. Qual é a definição de Velocidade, uma das características fundamentais do Big Data?

A) Trata da rapidez com que os dados são gerados, processados e analisados.
B) É a confiabilidade e precisão dos dados.
C) Trata da capacidade de processar grandes volumes de dados.
D) Refere-se à variedade de formatos de dados.
E) Refere-se a todas as características do Big Data.

7. Qual é uma das características do processo de MapReduce em Big Data?

A) Aumento da dimensionalidade dos dados
B) Redução da dimensionalidade dos dados
C) Paralelização do processamento de dados
D) Introdução de dados duplicados
E) Armazenamento de dados em nuvem

8. No contexto de Big Data, as características de Volume, Velocidade, Variedade e Veracidade são fundamentais para a gestão eficiente dos dados. Como a característica de Variedade impacta especificamente a escolha de tecnologias de armazenamento e processamento de dados?

A) Variedade permite que tecnologias como NoSQL sejam utilizadas devido à sua flexibilidade em lidar com diferentes tipos de dados
B) Variedade exige que todas as tecnologias utilizem apenas dados estruturados para simplificar o processamento
C) Variedade limita o uso de tecnologias de armazenamento em nuvem, pois estas só suportam dados homogêneos
D) Variedade implica que tecnologias de processamento paralelo, como Hadoop, são inadequadas para dados diversificados
E) Variedade não influencia a escolha de tecnologias, uma vez que todas as tecnologias de Big Data lidam com qualquer tipo de dado igualmente

9. No contexto do Big Data, a característica de Velocidade implica que os dados são processados em tempo real ou a uma velocidade muito rápida. Qual tecnologia é frequentemente usada para garantir essa característica?

A) HDFS
B) Spark
C) SQL
D) NoSQL
E) MapReduce

10. Dadas as características de um sistema de Big Data, assinale a alternativa que corretamente associa a tecnologia ao seu respectivo papel: A) Hadoop - processamento em tempo real, B) NoSQL - armazenamento flexível de dados, C) MapReduce - análise de dados semiestruturados.

A) A: Correto, B: Incorreto, C: Incorreto
B) A: Incorreto, B: Correto, C: Correto
C) A: Incorreto, B: Incorreto, C: Correto
D) A: Incorreto, B: Correto, C: Incorreto
E) A: Correto, B: Correto, C: Incorreto

GABARITOS E COMENTÁRIOS

1. Dadas as seguintes características de tecnologias de Big Data, associe corretamente cada tecnologia com sua respectiva função: A) NoSQL - permite o armazenamento flexível de dados, B) Apache Hadoop - oferece escalabilidade horizontal para armazenamento e processamento distribuído, C) Apache Spark - facilita o processamento em tempo real de grandes volumes de dados.

A) A: Correto, B: Correto, C: Incorreto
B) A: Correto, B: Correto, C: Correto
C) A: Correto, B: Incorreto, C: Correto
D) A: Incorreto, B: Correto, C: Incorreto
E) A: Incorreto, B: Incorreto, C: Correto

Gabarito: B

Comentários: NoSQL é conhecido por permitir o armazenamento flexível de dados, Apache Hadoop oferece escalabilidade para armazenamento e processamento distribuído, e Apache Spark é usado para processamento em tempo real, o que torna todas as associações corretas.

2. Qual é uma das características da Análise de Componentes Principais (PCA) em Big Data?

A) Aumento da dimensionalidade dos dados
B) Redução da dimensionalidade dos dados
C) Ampliação dos dados em tempo real
D) Introdução de dados duplicados
E) Armazenamento de dados em nuvem

Gabarito: B

Comentários: A característica da PCA em Big Data é a redução da dimensionalidade dos dados, mantendo as informações mais significativas.

3. No contexto de Big Data, a característica de _____ é essencial para lidar com grandes volumes de dados que chegam em tempo real e precisam ser processados rapidamente para gerar insights acionáveis.

A) Volume
B) Velocidade
C) Variedade
D) Veracidade
E) Valor

Gabarito: B

Comentários: A característica de Velocidade é essencial para lidar com dados que precisam ser processados rapidamente, especialmente em contextos onde o tempo é um fator crítico para a tomada de decisões.

4. No contexto de Big Data, a característica de _____ refere-se à capacidade de lidar com diferentes tipos e fontes de dados, incluindo dados estruturados, semiestruturados e não estruturados, que exigem diversas técnicas de armazenamento e processamento.

A) Volume
B) Velocidade
C) Variedade
D) Veracidade
E) Valor

Gabarito: C

Comentários: A característica de Variedade no Big Data se refere à diversidade de tipos e fontes de dados, como estruturados, semiestruturados e não estruturados, que exigem diferentes abordagens para armazenamento e processamento.

5. No modelo MapReduce, a fase de _____ é responsável por combinar os resultados intermediários produzidos na fase anterior, agregando os dados processados para produzir a saída final.

A) Shuffle
B) Map
C) Sort
D) Combine
E) Reduce

Gabarito: E

Comentários: A fase de Reduce no modelo MapReduce é a responsável por combinar e agregar os resultados intermediários gerados na fase de Map para gerar a saída final.

6. Qual é a definição de Velocidade, uma das características fundamentais do Big Data?

A) Trata da rapidez com que os dados são gerados, processados e analisados.
B) É a confiabilidade e precisão dos dados.
C) Trata da capacidade de processar grandes volumes de dados.
D) Refere-se à variedade de formatos de dados.
E) Refere-se a todas as características do Big Data.

Gabarito: A

Comentários: Velocidade, também conhecida como 'Velocity' em inglês, é uma das características fundamentais do Big Data e se refere à rapidez com que os dados são gerados, processados e analisados. Esta é uma característica crucial em diversos cenários onde a tomada de decisão in-time é necessária, como a negociação de alta frequência no mercado financeiro, análises em tempo real do Twitter, monitoramento de sinais vitais em hospitais, entre outros. A velocidade que os dados são gerados e precisam ser processados pode apresentar desafios significativos em termos de armazenamento e análise.

7. Qual é uma das características do processo de MapReduce em Big Data?

A) Aumento da dimensionalidade dos dados
B) Redução da dimensionalidade dos dados
C) Paralelização do processamento de dados
D) Introdução de dados duplicados
E) Armazenamento de dados em nuvem

Gabarito: C

Comentários: O processo de MapReduce facilita a paralelização do processamento de dados, permitindo que grandes volumes de dados sejam processados em paralelo. Portanto, a alternativa correta é a letra C.

8. No contexto de Big Data, as características de Volume, Velocidade, Variedade e Veracidade são fundamentais para a gestão eficiente dos dados. Como a característica de Variedade impacta especificamente a escolha de tecnologias de armazenamento e processamento de dados?

A) Variedade permite que tecnologias como NoSQL sejam utilizadas devido à sua flexibilidade em lidar com diferentes tipos de dados
B) Variedade exige que todas as tecnologias utilizem apenas dados estruturados para simplificar o processamento
C) Variedade limita o uso de tecnologias de armazenamento em nuvem, pois estas só suportam dados homogêneos
D) Variedade implica que tecnologias de processamento paralelo, como Hadoop, são inadequadas para dados diversificados

E) Variedade não influencia a escolha de tecnologias, uma vez que todas as tecnologias de Big Data lidam com qualquer tipo de dado igualmente

Gabarito: A

Comentários: A característica de Variedade impacta a escolha de tecnologias como NoSQL, que oferecem flexibilidade para lidar com diferentes tipos de dados, incluindo estruturados, semiestruturados e não estruturados, facilitando a integração e o processamento de dados diversificados.

9. No contexto do Big Data, a característica de Velocidade implica que os dados são processados em tempo real ou a uma velocidade muito rápida. Qual tecnologia é frequentemente usada para garantir essa característica?

A) HDFS
B) Spark
C) SQL
D) NoSQL
E) MapReduce

Gabarito: B

Comentários: O Spark é uma tecnologia frequentemente usada para garantir a característica de Velocidade, pois permite o processamento de dados em tempo real ou a uma velocidade muito rápida.

10. Dadas as características de um sistema de Big Data, assinale a alternativa que corretamente associa a tecnologia ao seu respectivo papel: A) Hadoop - processamento em tempo real, B) NoSQL - armazenamento flexível de dados, C) MapReduce - análise de dados semiestruturados.

A) A: Correto, B: Incorreto, C: Incorreto
B) A: Incorreto, B: Correto, C: Correto
C) A: Incorreto, B: Incorreto, C: Correto
D) A: Incorreto, B: Correto, C: Incorreto
E) A: Correto, B: Correto, C: Incorreto

Gabarito: D

Comentários: Hadoop não é usado exclusivamente para processamento em tempo real; seu foco é o armazenamento distribuído e processamento de grandes volumes de dados. NoSQL é de fato conhecido por seu armazenamento flexível, e MapReduce é utilizado para análise de dados em geral, não apenas para dados semiestruturados.

1.B	2.B	3.B	4.C	5.E
6.A	7.C	8.A	9.B	10.D

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.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.