Princípios de Big Data

Conceitos e aplicações dos princípios de Big Data, Internet das Coisas, computação distribuída, plataformas em nuvem, processamento e fluxo de dados.

Prof. Sérgio Assunção Monteiro

Propósito

Conhecer os conceitos e as tecnologias de Big Data, como grande diferencial para o profissional de tecnologia da informação com sólida formação.

Objetivos

- Reconhecer os conceitos e as aplicações de Big Data.
- Categorizar conceitos de Internet das Coisas e computação distribuída.
- Categorizar plataformas em nuvem para aplicações de Big Data.
- Identificar aplicações de processamento e streaming de dados.

Introdução

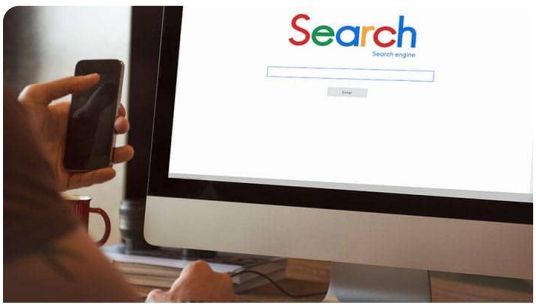
Atualmente, o termo Big Data é usado com muita frequência para descrever aplicações que envolvem grandes volumes de dados. Porém, mais do que isso, trata-se de um conjunto de tecnologias que gerenciam aplicações que, além do grande volume de dados, trabalham com dados que podem ser gerados com muita velocidade, de diversas fontes e em diferentes formatos. Com a popularização das tecnologias de computação em nuvem e da Internet das Coisas (IoT, do inglês Internet of Things), o ecossistema de aplicações de Big Data se ampliou bastante.

Os provedores de serviços na nuvem oferecem facilidades para que os clientes possam escalar seus sistemas – nos aspectos de hardware e software – com um custo muito inferior ao que teriam se tivessem de investir em infraestrutura própria. De fato, é um modelo de negócio que terceiriza a base tecnológica para empresas que são extremamente eficientes em lidar com ela, desse modo as organizações possam se concentrar no desenvolvimento de soluções de negócios que as diferenciem em relação à concorrência.

Ao longo deste conteúdo, entenderemos os conceitos relacionados à tecnologia de Big Data e como ela se relaciona com outras tecnologias, como computação na nuvem e IoT. Assim, teremos uma visão ampla sobre o assunto e conseguiremos conectá-lo a temas muito populares, como a inteligência artificial e o aprendizado de máquina.

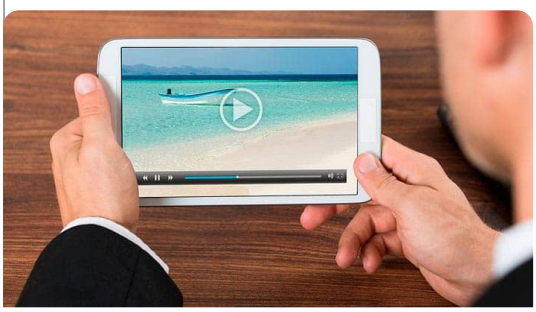
Introdução e Contextualização

Desde a popularização da Internet, com o avento da World Wide Web, na década de 1990, utilizamos, cada vez mais, aplicações e serviços que armazenam nossos dados e os utilizam para fazer previsões sobre nosso comportamento. Não é à toa que muitas empresas da Internet nos fazem ofertas que, de fato, coincidem com nossos interesses. Isso só é possível porque produzimos constantemente uma quantidade gigantesca de dados em diversas atividades, por exemplo quando:



Fazemos buscas na internet.

Fazemos compras on-line.



Assistimos a um vídeo.

Ou seja, mesmo sem estarmos cientes, fornecemos dados que podem ser utilizados para um estudo de nosso padrão comportamental.

Esse crescimento do volume de dados e de toda a complexidade que os envolve demandou um tratamento especializado de armazenamento, gerenciamento e análise, popularmente conhecido como Big Data.

Os dados precisam ser tratados por um ciclo de vida, de modo que possamos extrair informações úteis deles e, em um passo seguinte, transformar essas informações em conhecimento. Como consequência desse processo, áreas como a **Ciência de Dados (Data Science)** e o **Aprendizado de Máquina (Machine Learning)** cresceram muito nos últimos anos.

Quando escutamos falar sobre o termo Big Data, trata-se, normalmente, de uma descrição para enormes conjuntos de dados; no entanto, existem outros aspectos importantes que estão envolvidos e que precisam ser tratados, como:

Volume e disponibilização

Quando comparamos os conjuntos de dados tradicionais com aplicações de Big Data, além do volume de dados, temos de considerar a forma como esses dados são disponibilizados.

Técnica adequada

Em muitos casos, os dados não são estruturados e precisam de técnicas de análise que produzam respostas em tempo muito curto.

O principal estímulo para analisar dados nesse contexto é a possibilidade de descobrir oportunidades que podem se materializar por meio da detecção de segmentações de mercado, aumento de engajamento de

público-alvo e compreensão aprofundada dos valores ocultos. Por tudo isso, essa área tem grandes desafios para aplicar métodos eficazes e eficientes de organização e gerenciamento desses conjuntos de dados.

Devido ao potencial de valor que as aplicações de Big Data podem gerar, tanto empresas como agências governamentais têm investido nessa área, por meio do desenvolvimento de soluções que capturem dados com mais qualidade para, posteriormente, facilitar as etapas de armazenamento, gerenciamento e análise.

Saiba mais

Dados provenientes de fontes distintas permitem fazer um mapeamento muito detalhado do comportamento das pessoas. Isso também desperta discussões nos campos ético e legal. No Brasil, temos disposições constitucionais sobre a inviolabilidade do sigilo de dados e das comunicações, e a Lei Geral de Proteção dos Dados (Lei nº 13.709/2018), que visa proteger os cidadãos quanto ao uso indevido dos seus dados. Porém, ainda há muito a ser feito a respeito disso, o que acaba gerando novas oportunidades de pesquisa e desenvolvimento de projetos envolvendo segurança e privacidade.

Conceitos sobre Big Data

De modo geral, associamos o termo Big Data a um grande volume de dados e entendemos que este viabiliza a aplicação de métodos estatísticos e outras análises para extrairmos informações importantes. No entanto, Big Data é bem mais amplo que essa percepção, pois abrange conjuntos de dados que não podem ser tratados pelos métodos tradicionais de gestão da informação, ou seja, serem adquiridos, reconhecidos, gerenciados e processados em um tempo aceitável. Assim, o Big Data pode ser visto como uma **fronteira para inovação, competição** e **produtividade**.

Arquitetura básica de Big Data

A complexidade que envolve o gerenciamento de todas as características do Big Data exige que tratemos sua arquitetura de modo específico, o que, mais uma vez, o diferencia dos sistemas de banco de dados tradicionais que teriam dificuldade em lidar com operações de dados em sistemas heterogêneos. Esses sistemas são chamados de **data lake**, que, literalmente, pode ser traduzido como “lago de dados”. Basicamente, trata-se de um enorme repositório de arquivos e objetos de dados. Portanto, as soluções da arquitetura de Big Data precisam ser eficientes para que possam produzir resultados com tempos de resposta aceitáveis. Os componentes da arquitetura de Big Data são:

Fontes de dados (data sources)

Além das fontes de dados tradicionais, os sistemas de Big Data podem ser alimentados por meio de dados que estão na nuvem e são produzidos por sistemas de IoT, sendo que, em muitos casos, esse processo ocorre em tempo real. Trata-se do processo de aquisição de dados.

Armazenamento de dados (data storage)

Os dados precisam ser armazenados de modo eficiente para otimizar o seu acesso e segurança. Esse armazenamento pode ser feito de diversas maneiras na nuvem ou em bancos de dados estruturados ou não estruturados, que tenham:

- Escalabilidade: capacidade de crescer com consistência.
- Disponibilidade: prontos para serem acessados sempre que forem demandados.
- Segurança: mecanismos que garantam a privacidade e restrição de acesso.
- Padronização: armazenamento seguindo um padrão que facilite, posteriormente, a sua recuperação.

Processamento em lote (batch processing)

É o processo de armazenar os dados em lotes, para, então, fazer o seu processamento. Isso é feito para lidar com grandes volumes de dados, não sendo viável fazer o processamento dos dados em fluxos.

Ingestão de mensagens (message ingestion)

Consiste em agrupar os dados e trazê-los para um sistema de processamento de dados, onde podem ser armazenados, analisados e acessados.

Processamento de fluxo (stream processing)

É o processamento de dados à medida que são produzidos ou recebidos. Essa situação ocorre com frequência em processos de eventos produzidos por sensores, atividades do usuário em um site, negociações financeiras que têm como característica comum o fato de os dados serem criados como uma série de eventos de fluxo contínuo.

Armazenamento de dados analíticos (analytical data store)

Consiste no armazenamento de dados de negócios, mercado e clientes para posterior análise. As aplicações desses dados são chamadas de **business intelligence (BI)** – inteligência de negócios. Os bancos de dados analíticos são otimizados para consultas rápidas.

Análise e relatórios (analysis and reporting)

Os relatórios são uma organização dos dados com o objetivo de fazer resumos informativos e monitorar o desempenho de diferentes áreas de uma empresa. A análise, por sua vez, consiste em explorar dados e relatórios para extrair informações que agreguem valor e que possam ser usadas para melhor compreender e melhorar o desempenho dos negócios. Os relatórios de Big Data podem ser:

- **Predefinidos:** são relatórios prontos para uso que podem ser entregues de forma recorrente a um grupo de usuários finais. Normalmente, trazem informações estáticas com a possibilidade de diferentes níveis de detalhes. O termo usado para se referir ao detalhamento de um relatório é chamado de granularidade.

- **Painéis (dashboards):** esses relatórios apresentam uma visão abrangente do desempenho dos negócios. Ele é composto por indicadores de desempenho, conhecidos, principalmente, pela sigla KPI – key performance indicator – que ajudam a medir a eficiência de um processo. Para facilitar a compreensão, abordaremos os KPI mais adiante.

- **Alertas:** esses relatórios são usados para emitir notificações sempre que determinada condição previamente estabelecida ocorra, para que os responsáveis pelo processo sejam acionados e tomem as medidas adequadas.

KPI

Os KPI são indicadores de desempenho que integram os painéis (dashboards). Esses indicadores podem ser de três tipos:

Estratégicos

Oferecem uma visão geral do negócio e são utilizados pela presidência e diretoria de uma empresa. Como exemplo, temos o faturamento bruto de uma empresa em determinado período.

Táticos

São um detalhamento dos KPI estratégicos e têm como público-alvo a gerência da empresa. Como exemplo, podemos citar o faturamento das vendas de um determinado segmento da empresa, que pode ser um produto ou serviço específico.

Operacionais

Ajudam no acompanhamento detalhado de uma atividade da empresa. Como exemplo, podemos citar o KPI MTBS, que é um acrônimo para tempo médio de parada para manutenção – do inglês: mean time between stopages – usado para medir o tempo médio que um equipamento está disponível para uso até que ele pare para manutenção.

Os 5 Vs do Big Data

Uma forma de definir a complexidade do Big Data é por meio da descrição de suas características. Hoje há 5 características conhecidas como os 5 Vs do Big Data, mas nem sempre foi assim. Vamos conhecer um pouco da história:

Os 3 V's do Big Data: Volume, Velocidade e Variedade

Em 2001, o analista Doug Laney, da empresa META (atual Gartner Group), apresentou um relatório de pesquisa no qual tratou sobre os desafios e oportunidades trazidos pelo aumento de dados com um modelo 3Vs, sendo que cada V representa as características Volume, Velocidade e Variedade (LANEY, 2001). Esse modelo foi usado durante muitos anos para descrever a tecnologia de Big Data.

Os 4 V's do Big Data: Volume, Velocidade, Variedade e Valor

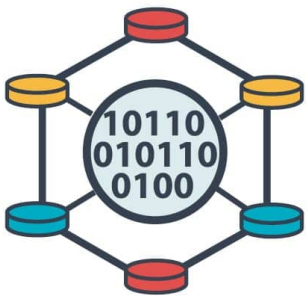
Posteriormente, o conceito evoluiu para a inclusão de mais um V, que representa Valor, por meio da publicação de um relatório do IDC (International Data Corporation) em 2011, que associou Big Data ao conjunto de tecnologias e arquiteturas projetadas para extrair valor de grandes volumes e variedades de dados, permitindo a captura, descoberta e análise de alta velocidade (GANTZ; REINSEL, 2011).

Os 5 V's do Big Data: Volume, Velocidade, Variedade, Valor e Veracidade

Atualmente, a forma mais comum de encontrarmos uma definição sobre Big Data inclui mais um V, além dos que já vimos: Veracidade (RUSSOM, 2011).

Essa evolução para explicar o conceito de Big Data vem do fato de estarmos trabalhando com um ecossistema complexo, que envolve aspectos tecnológicos de software e hardware, além de questões econômicas, sociais e éticas que ainda estão sendo compreendidas. Agora, vamos analisar com mais detalhes os 5Vs que compõem a tecnologia de Big Data.

Volume de Dados



Volume

Essa característica está relacionada com a escala da geração e coleta de massas de dados. Temos muitos exemplos práticos de aplicações em que o volume de dados é gigantesco, como sistemas de transações bancárias e de trocas de e-mails e mensagens. É fato que a percepção de grandes volumes de dados está relacionada com a tecnologia disponível em um determinado momento.

Precisamos conhecer como o volume de dados é medido. Basicamente, temos:

Byte (B)

Unidade de informação digital, também chamado de octeto, que consiste em uma sequência de 8 bits (binary digits).

Kilobyte (KB)

Corresponde a $1 = 2^{10} = 1024$ bytes.

Megabyte (MB)

Equivale a $1 = 2^{10} = 2^{20}$.

Gigabyte (GB)
Temos que $1\,GB = 2^{10}\,MB = 2^{20}\,KB = 2^{30}\,B$.

Terabyte (TB)
Corresponde a $1\,TB = 2^{10}\,GB$.

Petabyte (PB)
Temos que $1\,PB = 2^{10}\,TB$.

Exabyte (EB)
Equivale a $1\,EB = 2^{10}\,PB$.

Zetabyte (ZB)
Temos que $1\,ZB = 2^{10}\,EB$.

Yottabyte (YB)
Equivale a $1\,YB = 2^{10}\,ZB$.

Atenção

Quando nos referimos ao volume de uma aplicação de Big Data, normalmente, estamos tratando de petabytes (PB) de dados.

Velocidade

Essa característica se refere a dois aspectos:

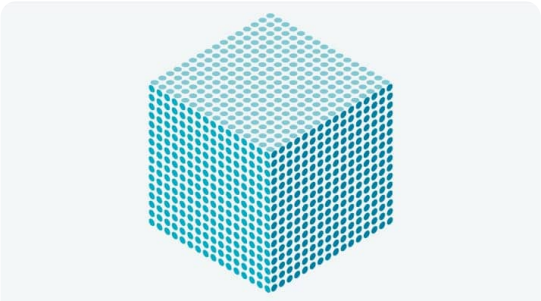
1. A velocidade da geração de dados.
2. A velocidade de processamento dos dados.

Basicamente, temos o problema clássico de computação: produtor x consumidor. O consumidor representa o papel do analista que precisa fazer consultas rapidamente, mas pode sofrer limitações do tempo de resposta do produtor, ou seja, o sistema pode possuir um ritmo mais lento para disponibilizar os dados para consulta.

> Um projeto de Big Data precisa equilibrar os tempos de consumo e geração de dados.

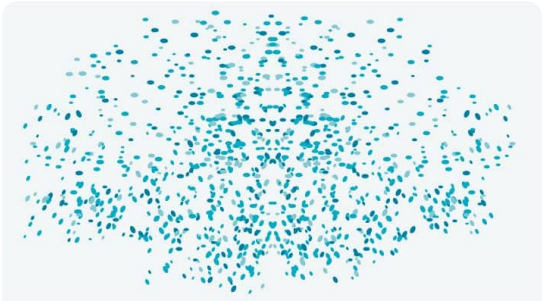
Variedade

Um projeto de Big Data pode ter vários tipos de dados, como áudio, vídeo, página da web e texto e tabelas de bancos de dados tradicionais. Esses tipos de dados podem ser classificados como:



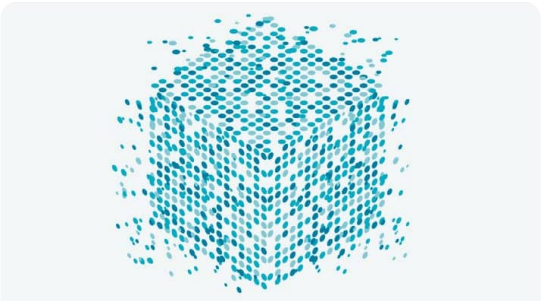
Dados estruturados

São armazenados de maneira organizada, e fáceis de serem processados e analisados. Normalmente, são dados numéricos ou de texto que podem ser armazenados em um banco de dados relacional e manipulados usando a linguagem SQL (do inglês Structured Query Language).



Dados não estruturados

Não possuem uma estrutura predefinida. Como exemplo, temos as imagens e arquivos de áudio. São armazenados em um banco de dados não relacional, também denominado NoSQL (do inglês Not Only SQL).



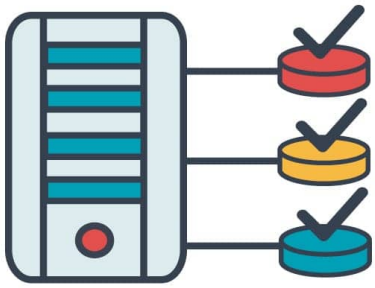
Semiestruturado

Os dados semiestruturados mesclam as duas formas de dados. Como exemplo de dados semiestruturados, temos arquivos nos formatos XML (do inglês eXtended Markup Language) e JSON (do inglês Java Script Object Notation).

Veracidade

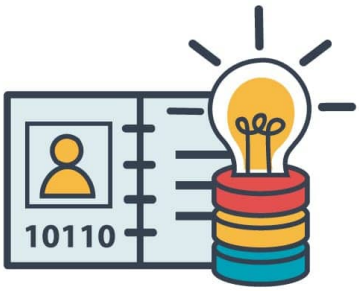
Essa característica está relacionada à qualidade dos dados. Isso é essencial, especialmente do ponto de vista de suporte para a tomada de decisão, pois é a veracidade dos dados que nos dá o grau de confiança para fazer o que precisamos por meio da **integridade** e da **precisão dos dados**.

> Um projeto de Big Data precisa utilizar técnicas que façam limpeza dos dados e garantam a sua qualidade, para que possam ser consumidos pelo processo de análise.



Veracidade

Valor



Valor

Essa é a principal característica de um projeto de Big Data e justifica todo o trabalho de extrair valor dos dados, que são a matéria-prima do negócio e, por isso, precisam passar por diversas etapas de tratamento e gerenciamento, até que possam ser consumidos pelo processo de análise. Podemos aplicar técnicas de ciência de dados e machine learning para obter informações e conhecimentos que vão direcionar ações para as diversas frentes de um negócio.

Aplicações de Big Data

Atualmente, existem muitas aplicações de Big Data que dão suporte para diversos setores da sociedade tomarem decisões e adquirirem conhecimento que, de outra maneira, seria muito difícil. Entre as aplicações de Big Data relacionados a setores da sociedade, podemos destacar os seguintes exemplos:

Área de saúde

Por meio das análises de dados, os pesquisadores podem encontrar o melhor tratamento para determinada doença e ter uma compreensão detalhada sobre as condições de uma região monitorada, tendo a possibilidade de propor ações com impacto positivo na saúde das pessoas.

Governo

Os setores ligados ao governo que utilizam sistemas de Big Data podem melhorar a prestação de serviços para os cidadãos por meio da integração dos dados das diversas áreas, conseguindo, assim, detectar fraudes, melhorar a educação, segurança pública, entre tantos outros serviços.

Mídia e entretenimento

Os anúncios que são feitos quando vemos vídeos na Internet são mais efetivos quando combinam com nosso perfil. As empresas de mídia e entretenimento analisam os dados dos usuários e trabalham para personalizar a oferta de produtos e serviços.

Internet das coisas (IoT)

Dispositivos de IoT geram dados contínuos e os enviam para um servidor. Quando esses dados são extraídos, podem ser analisados para compreender padrões e traçar estratégias mais efetivas para melhorar os resultados dos processos monitorados.

Visão geral de Big Data

No vídeo a seguir, falaremos sobre os principais conceitos da tecnologia de Big Data, com destaque especial aos 5V's.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Vem que eu te explico!

Os vídeos a seguir abordam os assuntos mais relevantes do conteúdo que você acabou de estudar.

Componentes da Arquitetura de Big Data

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Os 5 V's do Big Data

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Gerenciar um projeto de Big Data é uma tarefa complexa. Isso ocorre devido às características próprias desses projetos, que, além de lidar com grandes volumes de dados, ainda precisam tratar de diversas questões da sua arquitetura. Nesse sentido, assinale a alternativa correta a respeito da arquitetura de um projeto de Big Data.

A

Entre os aspectos que devem ser considerados em um projeto de Big Data, está a necessidade de garantir a privacidade dos dados, para que apenas as pessoas autorizadas possam acessá-los.

B

Um dos fatores que precisam ser tratados na arquitetura de um projeto de Big Data é a padronização dos dados, de modo que possam ser armazenados em tabelas.

C

As fontes de dados constituem a base da arquitetura dos projetos de Big Data, uma vez que garantem que os dados não sejam corrompidos.

D

Os projetos de Big Data podem crescer rapidamente, por isso é fundamental tratar aspectos relacionados às fontes de dados.

E

A complexidade da arquitetura de um projeto de Big Data está relacionada a dois fatores, que são o volume e a diversidade dos dados.

A alternativa A está correta.

Os projetos de Big Data são complexos, pois possuem muitas variáveis, tais como a diversidade e o volume dos dados, e a velocidade com que são gerados. Além disso, é necessário considerar aspectos como as diversas tecnologias envolvidas e a segurança dos dados.

Questão 2

O termo Big Data é bastante popular atualmente. Um dos motivos para isso ocorre devido à popularização do uso das aplicações que funcionam na Internet. Nesse sentido, selecione a opção correta a respeito das aplicações de Big Data:

A

Uma das dificuldades atuais associadas aos projetos de Big Data é o uso para prestação de serviços públicos, uma vez que são caros e seu benefício não é facilmente quantificável.

B

Dispositivos eletrônicos podem ser conectados diretamente à Internet, transmitindo dados sem a necessidade de garantir a sua qualidade, pois ela será tratada pela aplicação de Big Data.

C

A tecnologia de Big Data pode ser usada para monitorar os sinais vitais de pacientes que podem ser transmitidos via Internet.

D

Apesar de ainda não serem aplicados na área de entretenimento, existe um grande potencial de uso dos projetos de Big Data para proporcionar experiências específicas de acordo com o perfil do usuário.

E

Uma possível aplicação de Big Data é na prestação de serviços de utilidade pública, mas os benefícios só podem ser percebidos se houver total integração entre todos os sistemas dos diversos setores que compõem o Estado.

A alternativa C está correta.

Muitos benefícios podem ser obtidos pela utilização de projetos de Big Data para prestação de serviços públicos, entretenimento, segurança e aplicações na área da saúde, entre tantas outras aplicações. O potencial desses benefícios aumenta sempre que for possível fazer uso de diversas fontes de dados, pois essa diversidade permite identificar padrões complexos que dificilmente seriam detectados de outra maneira.

Introdução e Contextualização

O avanço da tecnologia criou dispositivos e sensores eletrônicos que geram enormes quantidades de dados. Esses equipamentos podem ser utilizados em diversas aplicações, tais como:

- monitoramento da temperatura de uma câmara frigorífica;
- segurança de transporte de cargas;
- acompanhamento e alerta da poluição dos níveis de poluição do ar;
- avaliação da pressão arterial de pacientes que precisam de atenção especial com cuidados de saúde etc.

> A lista de aplicações é muito grande! Para que todas essas aplicações sejam possíveis, precisamos ter à disposição uma tecnologia de coleta e troca de dados que conecte os dispositivos por meio de componentes de hardware e software.

A Internet das Coisas (IoT) é a infraestrutura que viabiliza a conexão e comunicação por meio da Internet desses objetos remotos.

Internet das Coisas (IoT)

A IoT é uma tecnologia que aumenta as conexões entre pessoas, computadores e dispositivos eletrônicos – estes últimos são chamados de “coisas”. Trata-se de uma revolução, pois a IoT viabiliza a extensão da realidade física para além de limitações espaciais, como, por exemplo, o acompanhamento da saúde de pacientes em regiões de difícil acesso. Essa tecnologia nos fornece acesso a dados sobre o meio físico com grande nível de detalhes, os quais, posteriormente, podemos analisar, compreender e tomar as ações adequadas, para otimizar processos, corrigir problemas, detectar oportunidades de melhorias e aumentar o nosso conhecimento a respeito de um contexto.

Um dos aspectos interessantes que devemos observar sobre a IoT é que os dados podem vir de diferentes fontes, oferecendo uma visão mais nítida sobre o que estamos monitorando. Nesse momento, já podemos notar uma estreita relação entre as tecnologias de IoT e Big Data:

Dados de fontes diferentes

Viabiliza que possamos verificar a **veracidade** dos dados, ou seja, o quão confiáveis eles são para representar o que está sendo observado.

Diferentes formatos

Podemos ter dados que são emitidos por diferentes sensores que retratam a **variedade** de representações do que estamos monitorando.

Frequência de geração dos dados

Os dados são enviados para a rede em uma **velocidade** característica da tecnologia que estamos aplicando.

Em relação à frequência de geração dos dados, refletimos sobre as seguintes questões:

1. Com que velocidade nossas aplicações devem consumir esses dados?
2. Qual é a velocidade adequada para analisá-los e produzir uma resposta adequada?
3. Qual é o volume de dados que devemos armazenar e tratar?
4. Qual é o valor dos dados que os dispositivos nos fornecem para que possamos priorizá-los adequadamente?

Reflexão

A compreensão dos dados gerados pelos dispositivos de IoT nos oferece oportunidades para melhorar nossa relação com as pessoas e aperfeiçoar processos e atividades sociais sobre aprendizado, saúde, trabalho e entretenimento. Ao mesmo tempo, abre discussões sobre aspectos éticos e legais, pois todo esse detalhamento abre a possibilidade de um conhecimento detalhado sobre a nossa privacidade que precisa ser tratado com bastante cuidado.

Além dos aspectos legais e éticos, devemos notar que as aplicações de IoT são, naturalmente, distribuídas com sensores e dispositivos capazes de enviar e receber dados usando protocolos de comunicação para a Internet. Outra questão tecnológica que devemos observar é que esses equipamentos possuem restrições de recursos de memória e processamento, portanto, é necessário utilizá-los com bastante eficiência, apesar de que eles, normalmente, são usados para uma tarefa específica.

> Para tratar de aplicações de IoT, utilizamos algoritmos distribuídos que reconhecem os dispositivos e os utilizam de forma eficiente para transmitir e receber dados.

Computação Distribuída e IoT

A tecnologia de IoT consiste na coexistência colaborativa de quatro componentes:

Objetos físicos (ou "coisas")

Componentes eletrônicos e sensores responsáveis pela coleta de dados e aplicação de ações. Exemplo: termostatos usados para controlar a temperatura de um ambiente.

Computação

Faz o gerenciamento do ciclo de vida dos dados, desde a coleta e o armazenamento até o processamento dos dados.

Protocolos de comunicação

Viabilizam a troca dados via Internet entre os objetos físicos e outros sistemas.

Serviços

Provêm autenticação e gerenciamento de dispositivos, além de oferecer a infraestrutura.

Para tratar da integração desses componentes de IoT, utilizamos a computação distribuída, pois é um modelo mais adequado para gerenciar essas unidades não centralizadas por meio do compartilhamento de responsabilidades e riscos. Apesar de, nesse cenário, os componentes estarem geograficamente espalhados, eles são executados como um sistema para melhorar a eficiência e o desempenho.

Aspectos da computação distribuída

Na computação distribuída, todos os elementos conectados na rede – servidores e nós – trabalham em conjunto de forma descentralizada para gerenciar toda a complexidade do sistema e ajustar-se ao crescimento do volume de dados e de dispositivos conectados. Para alcançar esse objetivo, a computação distribuída segue alguns princípios-chave, que são:

Distribuição e processamento

Distribuição de armazenamento e processamento de dados entre os nós da rede, para que a eficiência dos processos seja otimizada.

Transferência de dados e análises

A transferência de dados e as análises devem ser realizadas conforme necessário, pois diferentes níveis de processamentos podem ser realizados pelos nós da rede. Isso significa que o custo global de processamento e análise dos dados é minimizado, uma vez que os nós menos onerosos realizam pré-processamentos que reduzem o custo do processamento final dos nós mais caros da rede.

Tolerância a falhas

Outro princípio importante diz respeito à tolerância a falhas, pois é muito provável que haja intermitência da operação dos nós das redes, portanto a política de computação distribuída já deve estar preparada para reorganizar o fluxo de dados na rede, de maneira que possam ser roteados de um outro modo e que a rede continue em operação.

Otimização dos recursos computacionais da rede
Em especial, no caso da IoT, em que os dispositivos possuem uma restrição de recursos de memória e processamento, a computação distribuída trabalha com baixos níveis de consumo de energia

Computação distribuída e Big Data

Em um projeto de Big Data, de modo geral, temos que coletar uma grande quantidade de dados, armazená-los, processá-los e analisá-los para detectar padrões relevantes que demandem, quando necessário, algum tipo de ação. Agora, quando aplicamos Big Data para IoT, precisamos tratar a complexidade das características intrínsecas dos seus componentes, ou seja, utilizar uma solução que dê suporte para o alto volume de dados e consiga se comunicar com os dispositivos. Mas qual solução seria essa?

> A computação distribuída se torna a solução mais adequada no sentido de distribuir a computação para os nós da IoT.

Uma arquitetura básica de computação distribuída de IoT é composta pelas camadas de:

Computação em nuvem (cloud computing)

É a tecnologia que permite o uso remoto de recursos computacionais de software e hardware. Por exemplo, quando utilizamos repositórios na Internet para armazenar dados ou servidores de aplicação, estamos trabalhando com computação em nuvem. Essa camada é responsável por:

- processamento de Big Data;
- lógica de negócios;
- armazenamento de dados – mais conhecido como data warehousing.

Computação em névoa (fog computing)

É uma extensão da camada de nuvem que aproxima servidores aos dispositivos de IoT. Esses servidores podem colaborar entre si por meio de trocas de dados e realizar processamentos que vão otimizar a operação do sistema como um todo. Entre suas principais características, temos:

- processamento de Big Data;
- análise e redução de dados;
- controle de respostas;
- virtualização e padronização.

Computação de borda (edge computing)

Essa camada relaciona-se diretamente com os sensores e controladores que ficam na “borda” da arquitetura. Dessa forma, os dados podem ser armazenados e processados para, então, serem enviados à camada de névoa. Podemos destacar os seguintes aspectos dessa camada:

- processamento de grande volume de dados em tempo real.
- visualização de dados da fonte, ou seja, que vêm dos dispositivos eletrônicos.
- uso de computadores industriais que são específicos para trabalhar com determinados dispositivos eletrônicos.
- uso de sistemas integrados – também chamados de sistemas de bordo – que já vêm configurados nas placas.
- utilização de Gateways para interconectar os dispositivos com a rede por meio da conversão de protocolos e de sinais.
- sistema de armazenamento de microdados.

Sensores e controladores

São os dispositivos responsáveis por gerar os dados e, quando acionados, realizar ações. Por exemplo, em um sistema de irrigação, temos sensores que fazem o monitoramento da umidade do solo e controladores que fazem a irrigação até obter o nível adequado de umidade.

A figura 1 ilustra a arquitetura básica de computação distribuída aplicada para IoT.

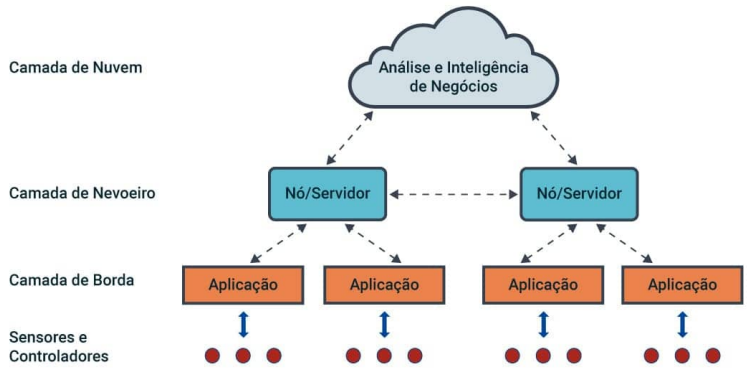


Figura 1 – Arquitetura básica de IoT

Um aspecto que podemos perceber rapidamente é a mudança da velocidade do fluxo de dados ao longo da arquitetura que mostramos na imagem. Em sua parte inferior, temos os dados gerados pelos sensores a uma velocidade superior, à medida que vamos avançando até a camada de nuvem.

Protocolos de comunicação

Os sistemas de IoT precisam de protocolos que permitam que os dispositivos eletrônicos possam se comunicar com outros nós da rede – sendo que um nó pode ser um dispositivo eletrônico, um computador ou um servidor. Alguns dos principais protocolos de comunicação de IoT são:

HTTP

O HTTP (Hyper Text Transport Protocol) é o Protocolo de Transporte de Hipertexto. É o protocolo do modelo cliente-servidor mais importante utilizado na Web, em que a comunicação entre um cliente e um servidor ocorre por meio de uma mensagem do tipo “solicitação x resposta”. A dinâmica básica da comunicação segue os seguintes passos:

- O cliente envia uma mensagem de solicitação HTTP.
- O servidor retorna uma mensagem de resposta, contendo o recurso solicitado, caso a solicitação tenha sido aceita.

MQTT

O MQTT (Message Queuing Telemetry Transport) é o Protocolo de Transporte de Filas de Mensagem de Telemetria. Ele foi lançado em 1999, sendo que sua primeira aplicação foi para o monitoramento de sensores em oleodutos. É um protocolo aberto e sua comunicação é baseada em um servidor que faz a publicação e o recebimento de dados com o padrão de mensagens “publicação x assinatura”, chamado de **broker**. O broker faz o trabalho intermediário de recebimento das mensagens dos nós da rede e as envia aos nós de destino. O MQTT é executado em um protocolo de transporte TCP (Transmission Control Protocol), o que garante a confiabilidade do tráfego de dados.

CoAP

O CoAP (Constrained Application Protocol) é o Protocolo de Aplicação Restrita. Utiliza a arquitetura REST (Representation State Transfer ou Transferência de Estado Representacional) e oferece suporte ao paradigma de “solicitação x resposta”, exatamente como ocorre no caso REST/HTTP. Além disso, ele é executado em um protocolo de transporte UDP (User Datagram Protocol).

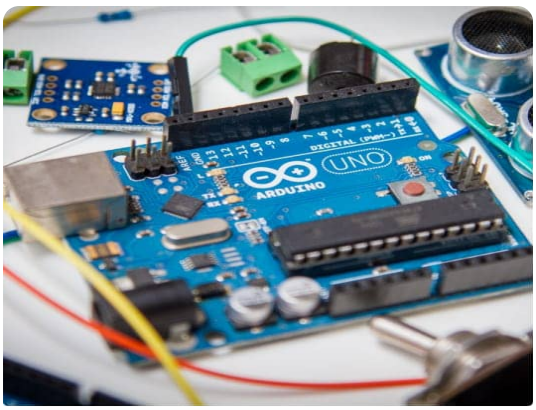
XMPP-IOT

O XMPP-IOT (Extensible Messaging and Presence Protocol for the IoT) é o Protocolo de Mensagem Extensível e de presença para a IoT. Também é um protocolo aberto que foi projetado para trocas de mensagens instantâneas. Ele usa a arquitetura cliente-servidor rodando sobre TCP, onde sua comunicação é baseada em XML e possui extensões que possibilitam o uso do modelo de “publicação x assinatura”.

Plataformas para IoT

Quando trabalhamos com um sistema de IoT, precisamos desenvolver programas para que os dispositivos possam operar da forma adequada e enviar dados para a rede. Para isso, precisamos de plataformas de desenvolvimento que nos ofereçam recursos de software e hardware que nos auxiliem a trabalhar com a interoperabilidade e a conectividade dos dispositivos à rede. A seguir, apresentamos algumas das principais plataformas de desenvolvimento para dispositivos de IoT.

Arduino



Arduino

Foi criado no Ivrea Interaction Design Institute em 2002. Ele oferece um ecossistema de hardware, linguagem de programação, bibliotecas e dispositivos que nos ajudam a desenvolver projetos que podem ter diversas aplicações. Uma das principais características do Arduino é que todas as suas placas e seu software são de código aberto. Essa característica ajudou a popularizar o Arduino, que possui uma comunidade de desenvolvedores engajada em divulgar projetos e conhecimentos em fóruns on-line.

facilitam trabalhar com dispositivos conectados à Internet para monitoramento e controle.

Raspberry PI

É uma plataforma de computação de placa única. Seu propósito inicial foi a aplicação no ensino de ciência da computação, evoluindo para funções mais amplas. Possui uma interface de baixo nível de controle auto-operado por portas de entrada-saída, chamado GPIO (General Purpose Input-Output), e usa o Linux como seu sistema operacional padrão.

Raspberry PI

NODEMCU

É um dos principais kits eletrônicos de código aberto para desenvolvimento de aplicações de IoT. Ele é baseado na família do microcontrolador ESP8266 e possui recursos que

NODEMCU

IoT e Computação Distribuída

No vídeo a seguir, abordaremos os conceitos das tecnologias de IoT e Computação Distribuída, relacionando-os com Big Data.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Vem que eu te explico!

Os vídeos a seguir abordam os assuntos mais relevantes do conteúdo que você acabou de estudar.

Computação distribuída e Big Data

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Protocolos de comunicação

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

As aplicações de Internet das Coisas (IoT) estão cada vez mais presentes em nosso dia a dia. Algumas das características dos projetos de IoT são a produção de um grande volume de dados e o uso de computação distribuída, e, por isso, devem ser tratados como projetos de Big Data. Em relação às tecnologias de IoT e de computação distribuída, assinale a alternativa correta.

A

A camada de computação em nuvem é responsável por tratar diretamente da qualidade dos dados produzidos pelos dispositivos de IoT e transmiti-los aos servidores de aplicações de Big Data.

B

Um dos aspectos da arquitetura de computação distribuída é utilizar camadas responsáveis por atividades específicas, como é o caso da cama de computação em névoa.

C

As camadas da arquitetura de computação distribuída são equivalentes quanto ao tratamento dos dados, sendo diferenciadas apenas pela tecnologia que utilizam.

D

Uma das vantagens da computação distribuída é padronizar a tecnologia utilizada em um projeto de IoT.

E

Projetos de IoT são considerados complexos, devido à grande quantidade de tecnologias envolvidas, e, por isso, a arquitetura de computação distribuída deve ser aplicada apenas com duas camadas: de nuvem e de dispositivos.

A alternativa B está correta.

A arquitetura de computação distribuída, aplicada para projetos de IoT, envolve camadas que são especializadas em tratar determinados aspectos da gestão de dados, para que eles possam trafegar na rede com segurança e qualidade. As camadas da arquitetura de computação distribuída para IoT são a de computação em nuvem, computação em névoa, computação de borda e a dos dispositivos de sensores e controladores.

Questão 2

Os projetos de Internet das Coisas (IoT) têm sido utilizados com sucesso em diversas áreas. De maneira simplificada, os sensores geram dados que são enviados para servidores de aplicação por meio da tecnologia de Internet. Nesse sentido, selecione a opção correta a respeito dos protocolos para aplicações de IoT:

A

Projetos de IoT são exemplos típicos de aplicações de Big Data e, portanto, devem ser desenvolvidos com o protocolo UDP, como é o caso do XMPP-IOT.

B

O HTTP é o protocolo padrão para aplicações de IoT, sendo utilizado por todos os demais protocolos como uma camada intermediária que garante a qualidade dos dados.

C

Dispositivos de IoT são caracterizados por possuírem muitos recursos de memória e processamento para tratar do grande volume e diversidade dos dados, e, por isso, utilizam protocolos como o HTTP e MQTT.

D

MQTT é um protocolo de IoT que usa uma estrutura de comunicação em que os dispositivos publicam seus dados, que são consumidos por um broker, que os transmite para determinadas aplicações.

E

Alguns dos protocolos usados pelos projetos de IoT são construídos com tecnologias proprietárias mais adequadas para tratar a diversidade de dados, como é o caso do CoAP.

A alternativa D está correta.

O MQTT é um protocolo aberto de IoT, baseado no padrão publicação X assinatura, que, na prática, significa que os dispositivos publicam seus dados, e as aplicações que vão consumir esses dados o fazem por meio de uma formalização (assinatura). Esse processo de recebimento e transmissão de dados é intermediado por uma aplicação chamada broker.

Conceitos

Computação em nuvem (do original em inglês Cloud Computing) é o termo usado para se referir a uma categoria de serviços de computação sob demanda disponíveis na Internet. Além de reduzir os custos necessários para oferecer os serviços, a tecnologia de computação em nuvem também aumenta a confiabilidade do sistema. Por isso, é cada vez mais comum encontrarmos aplicações que fazem a integração entre as diversas tecnologias e que oferecem os meios para que programas e dispositivos possam se comunicar na Internet.

Modelos de Serviços na Nuvem

Os modelos mais comuns de prestação de serviços na nuvem são:

SaaS

SaaS (Software as a Service) ocorre quando uma aplicação é oferecida via Internet e seu preço é dado de acordo com as necessidades de uso da parte contratante, tais como a quantidade de licenças, por exemplo. Esse tipo de serviço é bastante interessante para o cliente, pois ele vai pagar apenas as funcionalidades do sistema que lhe serão úteis. Além disso, não é necessário que o usuário se preocupe com instalação, ambiente para execução, manutenção e atualizações, pois tudo isso fica sob a responsabilidade do prestador de serviço.

PaaS

PaaS (Plataform as a Service) disponibiliza o sistema operacional e um ambiente de desenvolvimento na nuvem para o contratante, que, dessa forma, pode criar seus próprios programas com acesso a ferramentas adequadas, bibliotecas e bancos de dados.

IaaS

IaaS (Infrastructure as a Service) disponibiliza servidores de armazenamento e serviços de firewall e segurança da rede para os contratantes.

DaaS

DaaS (Desktop as a Service) oferece computadores (desktops) virtuais aos usuários finais pela Internet, que são licenciados com uma assinatura por usuário. A forma como os dados podem ser persistidos nas máquinas virtuais também é tratada por esses serviços. Os computadores podem ser persistentes e não persistentes:

- **Persistente:** os usuários podem personalizar e salvar uma área de trabalho para que mantenha a aparência sempre que fizer logon na máquina.
- **Não persistente:** os desktops são apagados cada vez que o usuário se desconecta, pois eles são apenas um meio de acessar os serviços de nuvem compartilhados.

XaaS

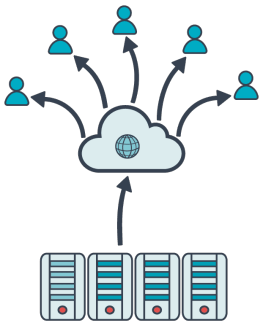
XaaS (Everything as a Service) é um termo geral usado para se referir à entrega de qualquer coisa como um serviço. Entre os exemplos de XaaS, podemos citar modelos gerais de computação em nuvem, como Software como Serviço (SaaS), Plataforma como Serviço (PaaS) e Infraestrutura como Serviço (IaaS); e modelos mais especializados, como comunicação como um serviço (CaaS), monitoramento como serviço (MaaS), recuperação de desastres como serviço (DRaaS) e redes como serviço (NaaS).

Tipos de Nuvem

Existem três diferentes maneiras de implantar uma infraestrutura de nuvem e disponibilizar programas que possuem vantagens e desvantagens associadas ao contexto em que serão utilizadas. Os três tipos de nuvens são:

Nuvem pública

Essa configuração é adequada para as empresas que ainda estão na etapa de crescimento de sua infraestrutura e nas quais a demanda por serviços é instável, podendo estar muito baixa em alguns momentos e muito alta em outros. Desse modo, as empresas podem pagar apenas pelo que estão usando e, se necessário, ajustar a sua infra na nuvem com base na demanda, sem a necessidade de fazer um investimento inicial em hardware, economizando dinheiro e tempo de configuração.

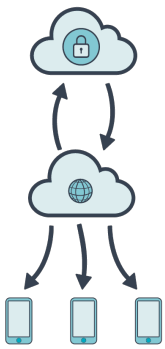


Nuvem privada

Todos os serviços são executados por servidores dedicados que dão ao contratante total controle sobre a gestão dos programas e da segurança da rede. Na prática, o usuário contratante pode monitorar e otimizar o desempenho da execução dos serviços de acordo com suas necessidades. O principal valor de uma nuvem privada é a privacidade que ela oferece. Essa característica é especialmente interessante para empresas que trabalham com dados confidenciais e querem isolamento da Internet aberta.

Nuvem híbrida

Combina aspectos das implementações de nuvem pública e privada. Por exemplo, os dados confidenciais permanecem na nuvem privada, devido à segurança que esse tipo de nuvem oferece. As operações que não usam dados confidenciais, por sua vez, são feitas na nuvem pública, onde as empresas contratantes podem dimensionar a infraestrutura para atender às suas demandas com custos reduzidos. No caso de operações de Big Data, as nuvens híbridas podem ser utilizadas para atuar com dados não confidenciais na nuvem pública e manter os dados confidenciais protegidos na nuvem privada.



Plataformas de Big Data na Nuvem

Uma plataforma de Big Data na nuvem é um conjunto de tecnologias de software e hardware que permite que o usuário contratante faça o gerenciamento de projetos de Big Data por meio de aplicações para desenvolvimento, implantação e operação de programas, além do controle de uma infraestrutura voltada para Big Data. Do ponto de vista econômico, essa estratégia é bastante interessante, pois o contratante não precisa se preocupar com vários detalhes operacionais que, nesse modelo, ficam sob a responsabilidade do prestador de serviços.

Ao longo dos anos, a demanda por soluções de Big Data tem aumentado e a oferta de serviços acompanhou esse processo. As soluções das plataformas de Big Data tratam de:

Gestão de dados

Disponibilização de servidores de banco de dados para gerenciamento de Big Data.

Análise de dados

Inteligência de negócios por meio de programas utilitários para tratamento e extração de dados de Big Data.

Ferramentas de desenvolvimento

Oferta de ambientes de desenvolvimento de programas para fazer análises personalizadas que podem se integrar com outros sistemas.

Além de todos esses aspectos, a plataforma oferece os serviços de segurança e proteção aos dados por meio do controle de acesso. Portanto, é um modelo muito interessante para quem trabalha com Big Data, devido à redução de complexidade da gestão de tantos detalhes e possibilidade de focar no negócio em si.

Toda a facilidade oferecida por uma plataforma de Big Data ajuda os profissionais a se concentrarem na excelência dos seus trabalhos, em especial, porque estão trabalhando com conjuntos de dados de grande volume. Alguns dos perfis dos profissionais que trabalham com essas plataformas são:

Engenheiros de dados

Profissionais que fazem toda a gestão do fluxo dos dados: coleta, agregação, limpeza e estruturação dos dados, para que possam ser utilizados em análises.

Cientistas de dados

Profissionais que utilizam a plataforma para estudar padrões e descobrir relacionamentos em grandes conjuntos de dados.

Saiba mais

Normalmente, existem dois perfis distintos em ciência de dados, que são: Análise exploratória e visualização de dados: consiste na análise dos dados por meio de técnicas estatísticas. Algoritmos de aprendizado de máquina: nesse perfil, os dados são analisados com o objetivo de encontrar associações não triviais que possam ser úteis para desenvolver estratégias de negócios, como aumentar engajamento de clientes e potencializar vendas.

Exemplos de Plataformas na Nuvem

Vamos conhecer, agora, algumas das principais plataformas na nuvem, mas, antes disso, vamos ver um conceito muito importante de Big Data, o **data lake**. Trata-se de um repositório centralizado onde é possível armazenar grandes volumes de dados estruturados e não estruturados. É um recurso bastante útil para armazenar os dados sem precisar estruturá-los e ter a possibilidade de executar diferentes tipos de análises de Big Data com painéis que facilitam as visualizações e funcionam como suporte para a tomada de decisão.

O data lake é recurso essencial nas plataformas de Big Data, pois as organizações utilizam os dados como a base para realizar análises e desenvolver estratégias que as auxiliem a potencializar seus negócios. Cada plataforma oferece uma tecnologia de data lake. Agora, veremos algumas dessas plataformas:

Amazon AWS

Sua primeira oferta como serviço ocorreu em 2006 e seu modelo é usado como referência por outras plataformas de armazenamento e computação em nuvem. Ainda em 2006, a Amazon lançou uma plataforma de computação chamada Elastic Cloud Compute (EC2), que fornece serviços de processamento de dados virtualizados, que podem ser ajustados para atender às necessidades do contratante. O nome do serviço de data lake da Amazon é **Amazon Simple Storage Service (S3)**, utilizado por muitas empresas para o desenvolvimento de soluções de Big Data na nuvem.

Microsoft Azure

É a plataforma de nuvem da Microsoft que foi lançada em 2010. Ela oferece ferramentas e serviços que foram projetados para permitir que organizações que trabalham com grandes conjuntos de dados realizem todas as suas operações na nuvem. Entre os seus pontos positivos, estão a segurança e a governança de dados, bem como a integração com ferramentas analíticas. Além disso, ela possui o **Azure Data Lake**, que permite trabalhar com dados complexos.

Google Cloud Plataform

É a plataforma de nuvem do Google. Ela utiliza a mesma tecnologia dos serviços de Big Data proprietários do Google, como YouTube e pesquisa Google. Ela também oferece serviços de armazenamento. Seu data lake é o **Google Cloud Storage**, projetado para trabalhar com exabytes de dados.

Oracle Cloud

É a plataforma de banco de dados da Oracle na nuvem. A Oracle é uma empresa especialista em soluções de bancos de dados. O seu serviço de nuvem inclui armazenamento flexível e escalável junto com os serviços de análise e processamento de dados. Sua plataforma possui fortes recursos de segurança, como criptografia em tempo real de todos os dados enviados para a plataforma.

IBM Cloud

É a plataforma de nuvem da IBM. Ela oferece várias soluções de data lake com o objetivo de atender aos diferentes perfis de necessidades dos seus clientes. Também é uma solução que tem dimensionamento ajustável, como as demais vistas. Com essa plataforma, os usuários podem escolher entre três tipos de armazenamento: de objeto, em bloco ou armazenamento de arquivo, dependendo das estruturas de dados com as quais estão trabalhando. Além disso, a IBM possui, na sua plataforma **Watson**, ferramentas analíticas que podem se integrar totalmente aos dados armazenados nos serviços em nuvem da IBM.

Plataformas em Nuvem para Aplicações de BigData

No vídeo a seguir, abordaremos a programação em nuvem, as plataformas e suas aplicações para Big Data.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Vem que eu te explico!

Os vídeos a seguir abordam os assuntos mais relevantes do conteúdo que você acabou de estudar.

Plataformas de Nuvem - Parte 1

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Plataformas de Nuvem - Parte 2

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Os serviços de nuvem oferecem diversas facilidades para projetos de Big Data. Eles são uma combinação de tecnologias que envolvem hardware e software por meio da Internet. Nesse sentido, assinale a alternativa correta a respeito dos modelos de serviços na nuvem.

A

Os serviços de nuvem são utilizados apenas para transmissão e recepção de dados, ficando o armazenamento e processamento dos dados sob a responsabilidade do contratante.

B

Quando contratamos um modelo PAAS, esperamos que sejam disponibilizadas aplicações que gerenciem os dados.

C

Os modelos de serviço de nuvem só podem ser usados para projetos de Big Data voltados para aplicações de Internet das Coisas.

D

Apesar da redução de custos para montar uma infra, os serviços de nuvem têm como desvantagem a dificuldade para expandir o uso de novas tecnologias em um projeto de Big Data.

E

Os serviços de nuvem de software tratam de diversos aspectos, tais como rede, servidores, virtualização, sistema operacional, dados e aplicações.

A alternativa E está correta.

Os serviços de nuvem são muito úteis para projetos de Big Data, pois flexibilizam o uso de tecnologias e a adequação do tamanho da infraestrutura para atender às demandas dos clientes. Existem vários modelos, como, por exemplo, o SaaS (software como serviço), PaaS (plataforma como serviço) e IaaS (infraestrutura como serviço).

Questão 2

A tecnologia de computação na nuvem é um importante recurso para projetos de Big Data. Para atender a essa demanda de mercado, grandes empresas da Internet oferecem plataformas com soluções de hardware e software. A respeito das plataformas de Big Data na nuvem, selecione a opção correta.

A

Ao utilizar plataformas na nuvem, os contratantes podem fazer análises personalizadas por meio do uso de programas especializados que são úteis para dar suporte à área de negócios de uma organização.

B

As plataformas de nuvem são protocolos de comunicação que fazem a intermediação entre as aplicações responsáveis pela coleta de dados até o processamento analítico, permitindo a elaboração de sofisticados relatórios.

C

A Amazon é uma das gigantes da Internet que disponibiliza uma plataforma de nuvem chamada MQTT, que pode ser utilizada para projetos de Internet das Coisas.

D

Um dos perfis dos profissionais que trabalham com plataformas de Big Data na nuvem é o de engenheiro de dados que se caracteriza por desenvolver aplicações de aprendizado de máquina.

E

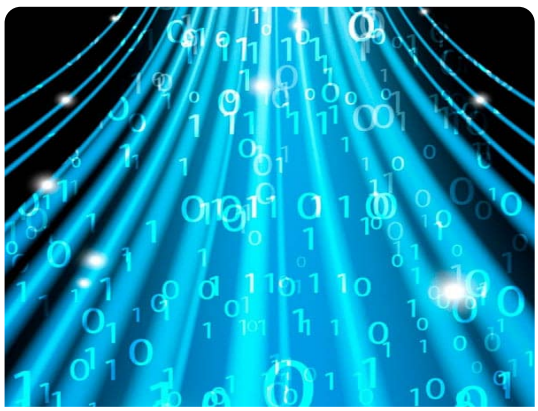
As plataformas de Big Data na nuvem são utilizadas para desenvolver, exclusivamente, aplicações voltadas para gestão do ciclo de vida dos dados caracterizada, principalmente, pelo uso da tecnologia de data lake.

A alternativa A está correta.

Os principais fornecedores de plataformas de Big Data na nuvem são a Amazon, Microsoft, Google, Oracle e IBM. Suas plataformas cobrem aspectos de hardware e software em que o contratante faz uso de um data lake e, posteriormente, pode utilizar ferramentas analíticas para detecção de padrões que apoiam no desenvolvimento de estratégias de negócios.

Conceitos

O **streaming de dados** é o processo de transmissão de um fluxo contínuo de dados. Por sua vez, um fluxo de dados é formado por diversos elementos de dados que são ordenados no tempo. Como exemplo, temos a transmissão de dados de uma gravação de vídeo, pois as imagens que vemos são séries de dados que seguem uma ordem cronológica. Assim, os dados representam que algo ocorreu – que chamamos de "evento" – de modo que houve uma mudança de estado sobre um processo que pode fornecer informações úteis. Por isso, muitas organizações investem para obter, processar e analisar esses dados.



Streaming de dados

Atenção

Em muitas situações, essas análises podem ser feitas ao longo de dias — o que é, por exemplo, bastante comum na manutenção preditiva de equipamentos — mas, em outros casos, esses processos entre coletas e análises devem ser feitos em tempo real – situação típica de processos de operação de equipamentos com riscos à vida e ao patrimônio.

Podemos encontrar exemplos típicos de fluxos de dados nas seguintes situações:

- Dados de sensores embarcados em equipamentos.
- Arquivos de logs de atividades de navegadores da web.
- Logs de transações financeiras.
- Monitores de saúde pessoais.
- Sistemas de segurança patrimonial

Esses foram apenas alguns exemplos, mas temos muitas outras situações que envolvem grandes volumes de dados que são transmitidos em fluxos contínuos, como se estivessem sendo transportados por uma esteira alimentando continuamente um sistema de processamento de dados.

Atualmente, o fluxo de dados e seu processamento aumentaram sua importância devido ao crescimento da Internet das Coisas (IoT), pois o fluxo de dados dessas aplicações é muito grande e precisa de um tratamento específico. Os sistemas de IoT podem ter vários sensores para monitorar diferentes etapas de um processo. Esses sensores geram um fluxo de dados que é transmitido de forma contínua para uma infraestrutura de processamento, que, por sua vez, monitora qualquer atividade inesperada em tempo real ou salva os dados para analisar padrões mais difíceis de detectar posteriormente.

Características e desafios em relação ao processamento de fluxos de dados

Os conceitos de aplicações de Big Data sempre precisam levar em consideração a complexidade em que estão contextualizados. Isso ocorre com os dados de streaming de sensores, navegadores da web e outros sistemas de monitoramento que possuem características que precisam ser tratadas de um modo diferente em relação aos dados históricos tradicionais.

Características do processamento de fluxos de dados

Devido aos aspectos que envolvem o processamento de fluxo de dados, podemos destacar algumas características, que são:

Sensibilidade ao Tempo

Independentemente de onde sejam aplicados, os elementos em um fluxo de dados estão associados a uma localização de tempo por meio de uma data e hora. Essa característica é usada junto com o contexto de aplicação para medir o valor do dado. Por exemplo, os dados de um sistema de monitoramento de saúde de pacientes que indiquem uma mudança grave dos níveis vitais devem ser analisados e tratados dentro de um curtíssimo período, para preservar a integridade da saúde do paciente, ou seja, permanecerem relevantes.

Continuidade

Especialmente para processos de tempo real, os fluxos de dados são contínuos e acontecem sempre que um evento é disparado ou quando ocorre uma mudança de estado no sistema. Portanto, o sistema de processamento deve estar preparado para ser acionado sempre que for requisitado.

Heterogeneidade

Os dados de fluxo podem vir de diferentes fontes com diferentes formatos e que podem estar geograficamente distantes. Uma das características de Big Data é a variedade que abrange estas situações: formatos, fontes de dados e localização geográfica.

Imperfeição

Muitos fatores podem influenciar para que os elementos de um fluxo de dados sejam prejudicados por perda e corrupção. Devido à variedade das fontes e dos formatos, esse processo é ainda mais complexo de ser gerenciado. Ainda há a possibilidade de que os elementos de dados em um fluxo possam chegar fora de ordem. Isso implica que o sistema também precisa levar em consideração essas falhas e ter uma medida de tolerância para fazer ajustes, quando for possível, e o processamento dos dados.

Volatilidade

Os elementos de fluxo de dados são gerados em tempo real e representam estados de um sistema que está sob monitoramento. Isso implica que a recuperação desses dados, quando ocorre uma falha de transmissão, é bastante difícil. Não se trata apenas de retransmitir os dados, mas também da impossibilidade de reproduzir o estado do sistema quando os dados foram gerados. Portanto, é necessário desenvolver estratégias que minimizem esse problema, como redundâncias de monitoramento e armazenamento de dados.

Desafios do processamento de fluxos de dados

Agora que entendemos as características do processamento de fluxos de dados, precisamos analisar os desafios para desenvolver aplicações. Entre esses desafios, podemos citar os seguintes:

Escalabilidade

Uma aplicação de processamento de fluxo de dados precisa ter flexibilidade para gerenciar o aumento brusco de volume de dados. Uma situação desse tipo pode ocorrer quando partes do sistema falham e uma grande quantidade de dados de logs é enviada para alertar sobre a ocorrência do problema, podendo aumentar a taxa de envio dos dispositivos para o servidor de aplicação. Portanto, o projeto do sistema deve contemplar tais casos com estratégias para adição automática de mais capacidade computacional à medida que a demanda por recursos aumenta.

Ordenação

Os elementos de um fluxo de dados estão associados a uma marcação no tempo. Essa marcação é fundamental para que os dados possam ser agrupados em estruturas sequenciais que façam sentido. Podemos pensar em uma transmissão de vídeo ao vivo, em que é esperado que o conteúdo siga uma sequência linear, pois não faria sentido ver um vídeo em que os quadros são transmitidos fora de ordem. Portanto, um projeto desse tipo precisa evitar que haja discrepâncias sobre a ordem de transmissão dos dados, além de ter mecanismos de controle de qualidade.

Consistência e durabilidade

Os dados em um fluxo de dados são voláteis, mas, em muitas situações, é útil mantê-los armazenados, para que possamos analisá-los posteriormente. Para isso, precisamos aplicar técnicas que garantam a condição de originalidade dos dados, ou seja, que eles não foram modificados e que, além disso, tenham informações sobre sua qualidade. Essas situações implicam que o desenvolvimento de um projeto de processamento de fluxo de dados deve garantir a consistência dos dados, para que possam ser armazenados e analisados em outro momento. Quando os dados passam por essas etapas, eles têm a propriedade de durabilidade.

Tolerância à falhas e garantia de dados

Os sistemas são sujeitos a falhas. E quando falamos em sistema, precisamos visualizar toda a complexidade que envolve programas, dispositivos físicos e infraestrutura. Esse tipo de situação pode ser tratado por meio de algumas abordagens, como, por exemplo:

- redundância de elementos de transmissão e coleta;
- uso de sistemas não centralizados;
- análise estatística periódica dos dados para medir a sua qualidade.

Descoberta de Conhecimento a partir de Fluxo de Dados

Data warehouse

Os dados de uma organização podem vir de diversas fontes, como registros de vendas, sistemas de controle de estoque e interações com usuários – que são aquelas pesquisas em que a empresa pergunta sobre a qualidade do seu atendimento. Esses dados são armazenados em um data warehouse e, então, processados em lotes por um sistema de análise de dados.

Esse modelo de gestão de dados funciona bem em contextos em que não temos urgência para extrair informações que nos deem suporte para intervir em um sistema.

Por outro lado, temos muitas situações práticas em que o tempo entre a coleta do dado e a ação sobre uma determinada configuração é crucial. Alguns dos casos típicos em que isso ocorre estão relacionados às seguintes situações:

Prestação de serviços essenciais

Como fornecimento de água, energia elétrica e gás.

Monitoramento

De saúde e prestação de socorro a vítimas.

Operação de equipamentos

Como transportadores de carga em aviões e caminhões aplicados à mineração.

Ajustes ad hoc

De eventos de divulgação de produtos e de prestação de serviços que tenham como objetivo aumentar o engajamento do público.

Comentário

A lista não se encerra com esses exemplos, mas eles já ilustram bem o fato de que existem muitas situações reais em que o processamento em lote não é adequado para aplicações de tempo real e, portanto, precisamos aplicar estratégias de processamento do fluxo de dados para obtermos informações que nos permitam atuar rapidamente e com maiores chances de alcançar o nosso objetivo com sucesso.

Big data e machine learning

Já é um fato consolidado que a ciência de dados e, em especial, as técnicas de aprendizado de máquina, têm sido aplicadas com sucesso em contextos de Big Data para detectar padrões e produzir conhecimento que oriente as nossas decisões. Um dos motivos para que esse processo seja bem-sucedido é que essas técnicas de aprendizado são beneficiadas pela diversidade dos dados, o que permite obter algoritmos que generalizem as soluções, em vez de ficar restrito a um conjunto de dados. Como exemplos de algoritmos de aprendizagem de máquina que são naturalmente incrementais, temos: k-vizinhos mais próximos e o Bayes ingênuo. Além da aprendizagem incremental, as técnicas precisam se autoajustar para refletir o estado mais recente dos dados e esquecer informações que perderam sua utilidade para o cenário atual.

Sistemas como os de IoT são modelados de forma mais adequada, como fluxos de dados transitórios, apesar de também ser útil armazená-los em tabelas para registro e fazer estudos posteriores. A análise desses dados permite que possamos fazer a sua mineração, ou seja, realizar processamentos que nos auxiliem a detectar tendências e mudanças de estado. Como resultado desse trabalho, podemos:

Identificar perfis

O que nos permite direcionar estratégias mais eficientes para aumentar o engajamento de clientes e oferecer serviços personalizados.

Fazer estimativas sobre a demanda

De modo que possamos dimensionar os recursos que precisamos alocar para operar com segurança.

Detectar falhas e atividades anormais nos sistemas

Para que possamos intervir rapidamente.

Processamento e Streaming de Dados

No vídeo a seguir, abordaremos os conceitos de processamento e streaming de dados, relacionando-os à tecnologia de Big Data em aplicações de aprendizado de máquina.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Vem que eu te explico!

Os vídeos a seguir abordam os assuntos mais relevantes do conteúdo que você acabou de estudar.

Ciência de Dados e Aprendizado de Máquina Aplicados a Big Data - Parte 1

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Ciência de Dados e Aprendizado de Máquina Aplicados a Big Data - Parte 2

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Projetos de Big Data são complexos, pois muitos aspectos devem ser considerados. Um desses aspectos corresponde ao fluxo de dados que são conhecidos como streamings. Nesse sentido, assinale a alternativa correta a respeito das características e desafios em relação ao processamento de fluxo de dados em projetos de Big Data.

A

Quando um sistema de fluxo de dados de Big Data falha, é possível recuperar os dados reiniciando-o.

B

Aplicações de streaming são caracterizadas por fluxos não contínuos de dados, sendo, desse modo, um desafio dimensionar uma infraestrutura, para evitar a ociosidade do sistema.

C

Os fluxos de dados de aplicações de tempo real precisam de garantia de qualidade de serviço, pois não é possível fazer análises confiáveis com dados voláteis.

D

Muitas das aplicações de Big Data que utilizam fluxos de dados são de tempo real, cujos dados precisam ser processados com muita velocidade, pois, em muitos casos, o seu valor é reduzido ao longo do tempo.

E

Uma das vantagens de trabalhar com sistemas de fluxos de dados é o fato de que eles são oriundos da mesma fonte, o que reduz a complexidade da infraestrutura necessária para o processamento.

A alternativa D está correta.

Aplicações como monitoramento de sinais vitais de pacientes e de segurança, de modo geral, precisam ter seus dados processados com grande velocidade, pois, depois de algum tempo, o paciente pode sofrer graves consequências por não ter sido atendido, como uma equipe de segurança também pode perder a oportunidade de intervir contra uma atividade criminosa. Projetos desse tipo são muito complexos, pois precisam garantir a disponibilidade dos dados e a velocidade de transmissão e processamento, para detectar padrões e permitir que ações sejam tomadas dentro de um tempo adequado.

Questão 2

A utilização de estatística e métodos de aprendizado de máquina em aplicações de Big Data é cada vez mais comum. Um dos fatores que influencia para que isso ocorra é o fato de ter à disposição grandes volumes de dados com variações que permitam que os modelos generalizem as soluções. Nesse sentido, assinale a alternativa correta a respeito da descoberta de conhecimento a partir de fluxo de dados em projetos de Big Data.

A

Uma das estratégias mais adequadas para lidar com fluxo de dados para algoritmos de aprendizado de máquina é submeter os dados a um processo de tratamento para garantir a qualidade deles antes de submetê-los aos algoritmos.

B

Aplicações de Internet das Coisas produzem dados que podem fornecer informações úteis a respeito da topologia de sistemas monitorados, possibilitando, assim, a atuação mais precisa.

C

Os projetos de Big Data que envolvem fluxos de dados são úteis apenas para avaliar o estado do sistema em certo período e não devem ser armazenados com o objetivo de obter histórico de comportamento.

D

Para aplicar um algoritmo de aprendizado de máquina em um projeto de Big Data que envolva fluxo de dados é necessário utilizar um data warehouse, que é uma tecnologia adequada para consultas ad hoc.

E

Existem poucas situações práticas que justificam a aplicação de algoritmos de aprendizado de máquina para projetos de Big Data que envolva fluxo de dados, no entanto, apesar disso, é uma boa prática preparar uma infraestrutura adequada para esses algoritmos, pois o valor dos dados pode aumentar ao longo do tempo.

A alternativa B está correta.

Os projetos de Big Data que envolvem fluxo de dados de tempo real podem fornecer informações importantes para direcionar os esforços de atuação. Para que esse processo funcione adequadamente, é necessário adaptar os métodos de aprendizado de máquina para procurar padrões e detectar anomalias, enquanto os dados ainda estão em fluxo, ou seja, sem passar pelo processo tradicional de tratamento e treinamento em lote.

Considerações finais

Ao longo deste conteúdo, estudamos o conjunto de tecnologia que envolve o conceito de Big Data. É interessante notarmos que, em um primeiro momento, associamos Big Data a aplicações de banco de dados. Porém, quando analisamos um pouco mais, vimos que estamos tratando de uma tecnologia que vai além de banco de dados, relacionando-se às tecnologias de redes, processamento eficiente, Internet das Coisas (IoT), computação distribuída, análise estatística e aprendizado de máquina.

Estudamos os conceitos de IoT e computação distribuída e as plataformas em nuvem para aplicações de Big Data. Vimos, ainda, alguns dos principais provedores de serviço e entendemos a importância da tecnologia de data lake. Além disso, estudamos sobre processamento e streaming de dados.

Atualmente, vivemos em uma época com grandes oportunidades de demanda de profissionais para desenvolver aplicações nas mais variadas áreas, como no entretenimento, na prestação de serviços de monitoramento, e nas áreas de segurança, saúde, finanças, entretenimento, mídia e agronegócio. Portanto, Big Data é uma excelente área para se especializar e procurar oportunidades de desenvolvimento profissional.

Podcast

Para finalizar o seu estudo, ouça o podcast a seguir, que aborda os principais conceitos de Big Data e sua relação com as tecnologias de IoT, Computação Distribuída, Plataformas em Nuvem e Streaming de Dados.

Conteúdo interativo

Acesse a versão digital para ouvir o áudio.

Explore +

Acesse o site do Arduino e estude os diversos exemplos didáticos de como construir projetos superinteressantes. Em seguida, tente programar esses projetos no site do Tinkercad.

Acesse o site oficial do Spark e procure por Streaming Programming. Desse modo, você vai aprofundar seu conhecimento sobre processamento de fluxo de dados, além de encontrar exemplos práticos desenvolvidos no Spark.

Referências

BRASIL. **Lei nº 13.709 de 14 de agosto de 2018.** Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Diário Oficial da República Federativa do Brasil, 15 ago. 2018. Consultado na Internet em: 10 set. 2021.

GANTZ, J.; REINSEL, D. **Extracting value from chaos.** IDC iView, pp 1–12, 2011.

LANEY, D. **3-d data management: controlling data volume, velocity and variety.** META Group Research Note, 2001.

RUSSOM, P. **Big Data Analytics.** TDWI Best Practices Report, Fourth Quarter 2011. TDWI Research, 2011.