FLUÊNCIA EM DADOS
Teoria e questões de provas.

CONTEÚDO PROGRAMÁTICO:

NOÇÕES INTODUTÓRIAS À DISCIPLINA

O QUE É FLUÊNCIA EM DADOS?

Conhecida também por *data literacy*, a fluência em dados é a habilidade de acessar, identificar e interpretar dados tendo um objetivo predefinido. E esse objetivo muda conforme o contexto.

Por exemplo, um analista de tráfego orgânico precisa analisar o número de visitantes de um site a fim de que novas estratégias de atração sejam elaboradas. Um especialista em cibersegurança vai avaliar o número de ameaças ou tentativas de invasão de um sistema para pensar em como diminuir esse índice, ou seja, criar medidas de proteção eficazes.

De acordo com Approaches to Building Big Data Literacy, do Instituto de Tecnologia de Massachusetts (MIT), a fluência em dados consiste em quatro competências:

1. – ler dados;
2. – trabalhar com dados;
3. – analisar os dados;
4. – argumentar utilizando dados.

Segundo os autores, **não basta saber ler e interpretar os dados**, **é preciso** também **ter uma compreensão maior da importância dos dados** no dia a dia. Portanto, a fluência em dados permite o consumo pleno de informações, produtos e serviços.

Diariamente somos submetidos a um grande volume de informações. Porém, a partir do momento em que o indivíduo consegue analisar dados de maneira crítica, a resolução de problemas ou tomada de decisões sai do campo da intuição e passa a ser embasada em fontes confiáveis.

CONCEITOS, ATRIBUTOS, MÉTRICAS, TRANSFORMAÇÃO DE DADOS.

1. INTRODUÇÃO AOS SGBDS e BANCO DE DADOS

CONCEITUANDO BANCO DE DADOS E SGBD

Banco de dados

É uma coleção de dados inter-relacionados, representando informações sobre um domínio específico.

Exemplos: Lista telefônica, controle do acervo de uma biblioteca, sistema de controle dos recursos humanos de uma empresa.

SGBD(Sistema de Gerenciamento de Banco de dados)

É um software com recursos específicos para facilitar a manipulação das informações dos dados e o desenvolvimento de programas aplicativos.

Um sistema de gerência de banco de dados (SGBD) é uma coleção de programas que permite que usuários criem e mantenham bancos de dados.

Exemplos : MS SQL Server , Oracle Database, IBM DB2, MySQL , PostgreSQL.

Sistema de Banco de dados

É um sistema de manutenção de registros por computador envolvendo quatro componentes principais, sendo eles dados, hardware, software e usuários. O sistema de banco de dados pode ser considerado como uma sala de arquivos eletrônica. Existe uma série de métodos , técnicas e ferramentas que visam sistematizar o desenvolvimento de banco de dados.

Configuração de um sistema de banco de dados simplificado.

OBJETIVOS DE UM SISTEMA DE BANCO DE DADOS

Os **Objetivos de um sistema de banco de dados** é isolar os usuários dos detalhes mais internos do banco de dados (abstração de dados) e também prover independência de dados ás aplicações (estrutura física de armazenamento e à estratégia de acesso).

FUNDAMENTOS SGBD

Um sistema de gerenciamento de banco de dados **é um conjunto de programas de software** que **permite** aos usuários **criar, editar, atualizar, armazenar e recuperar dados em tabelas** de banco de dados. Dados em um banco de dados podem ser acrescentados, apagados, alterados, classificados usando um SGBD. Se você fosse um empregado em uma grande organização, a informação sobre você provavelmente seria armazenadas em diferentes tabelas que estão ligados entre si. Por referência cruzada dessas tabelas, alguém poderia mudar o endereço de uma pessoa em uma tabela e ela seria automaticamente refletida para todas as outras tabelas.

CARACTERÍSTICAS GERAIS DE UM SGBD

· Controle de Redundâncias
· Compartilhamento dos Dados
· Controle de Acesso
· Interfaceamento
· Esquematização
· Controle de Integridade
· Backups

Iremos detalhar cada característica de um SGBD para que possa compreender com clareza.

1)Controle de Redundâncias:

A redundância consiste no armazenamento de uma mesma informação em locais diferentes, provocando inconsistências. Em um Banco de dados as informações só se encontram armazenadas em um único local, não existindo duplicação descontrolada dos dados. Quando existem replicações dos dados, estas são decorrentes do processo de armazenagem típica do ambiente Cliente-Servidor, totalmente sob controle do Banco de dados.

2)Compartilhamento dos Dados:

O SGBD de incluir software de controle de concorrência ao acesso dos dados, garantindo em qualquer tipo de situação a escrita/leitura de dados sem erros.

3)Controle de Acesso:

O SGBD de dispor de recursos que possibilitem selecionar a autoridade de cada usuário. Assim um usuário poderá realizar qualquer tipo de acesso, outros poderão ler alguns dados e atualizar outros e outros ainda poderão somente acessar um conjunto restrito de dados para escrita e leitura.

4)Interfaceamento:

Um Banco de dados deverá disponibilizar formas de acesso gráfico, em linguagem natural, em SQL ou ainda via menus de acesso, não sendo uma "caixa-preta" somente sendo passível de ser acessada por aplicações.

5) Esquematização:

Um Banco de dados deverá fornecer mecanismos que possibilitem a compreensão do relacionamento existente entre as tabelas e de sua eventual manutenção.

6) Controle de Integridade:

Um Banco de dados deverá impedir que aplicações ou acessos pelas interfaces pudessem comprometer a integridade dos dados.

7)Backups:

O SGBD deverá apresentar facilidade para recuperar falhas de hardware e software, através da existência de arquivos de "pré-imagem" ou de outros recursos automáticos, exigindo minimamente a intervenção de pessoal técnico.

SGBDs são comumente usados para gerenciar:

- Sócios e listas de discussão de subscrição
- Informação contábil e contabilidade
- Os dados obtidos a partir de pesquisa científica
- Informações de clientes
- Informações de inventário
- Registros pessoais
- Informações da biblioteca

As vantagens de um SGBD

Maior disponibilidade: Uma das principais vantagens de um SGBD é que a **mesma informação pode ser disponibilizada a utilizadores diferentes**, ou seja, **compartilhamento** de dados.

Redundância minimizada: Os dados de um SGBD são mais concisos, porque, como regra geral, a **informação nela aparece apenas uma vez**. Isto reduz a redundância de dados, ou em outras palavras, a necessidade de repetir os mesmos dados uma e outra vez. Minimizando a redundância pode, portanto, **reduzir significativamente o custo de armazenamento** de informações em discos rígidos e outros dispositivos de armazenamento.

Precisão: dados precisos, consistentes são um sinal de **integridade dos dados**. SGBDs fomentam a integridade dos dados, porque as **atualizações e alterações dos dados só tem que ser feitas em um só lugar**. As chances de se cometer um erro são maiores se você é obrigado a alterar os mesmos dados em vários lugares diferentes do que se você só tem que fazer a mudança em um só lugar.

Programa e arquivo de consistência: Usando um sistema de gerenciamento de banco de dados, **formatos de tabelas e programas do sistema são padronizados**. Isso faz com que os tabelas de dados sejam mais fáceis de manter, porque as mesmas regras e diretrizes se aplicam a todos os tipos de dados. O nível de consistência entre os tabelas e programas também torna **mais fácil de gerenciar dados** quando vários programadores estão envolvidos.

User-friendly: Os dados são é **mais fáceis de acessar e manipular com um SGBD** do que sem ele. Na maioria dos casos, SGBDs também reduzem a dependência de usuários individuais à especialistas em computação para atender às necessidades de seus dados.

Maior segurança: Como afirmado anteriormente, SGBDs permitem que múltiplos usuários acessem os recursos dos mesmos dados. Esta capacidade é geralmente vista como um benefício, **mas há riscos potenciais para a organização**. Algumas fontes de informação **devem ser protegidas ou garantida** e vista apenas por indivíduos selecionados. Através do **uso de senhas**, sistemas de gerenciamento de banco de dados podem ser usado para **restringir o acesso** aos dados a apenas aqueles que devem vê-lo.

Outros: Tempo de **desenvolvimento de aplicações é reduzido**, Maior flexibilidade para realizar alterações (**independência de dados**) e Maior economia, informações atualizadas, menor volume de papel.

Que características distinguem um SGBD ?

Catálogo

- Um SGBD mantém não apenas o Banco de Dados, mas também uma definição e descrição das estruturas e restrições (catálogo - metadados)
- A existência do catálogo permite que um mesmo SGBD possa ser utilizado para aplicações distintas (o catálogo indica uma estrutura física utilizada)

Independência de Dados

- Um SGBD dá aos usuários uma **visão abstrata dos dados**, encobrindo detalhes não relevantes (o usuário-desenvolvedor não precisa saber como os dados são fisicamente armazenados).

Múltiplas Visões dos Dados

- Cada usuário pode exigir uma **visão diferenciada** da base de dados

Compartilhamento e Transações

- **Controle de concorrência**

As desvantagens de um SGBD

Existem basicamente duas desvantagens principais em SGBDs. Um deles é o custo, e a outra o perigo para a segurança dos dados.

Custo: A Implementação de um sistema de SGBD **pode ser cara e demorada**, especialmente em grandes organizações. Requisitos de formação pode ser bastante oneroso.

Segurança: Mesmo com salvaguardas no lugar, pode ser possível para alguns usuários não autorizados acessar o banco de dados. Em geral, o acesso de banco de dados é uma proposição de tudo ou nada. Uma vez que um usuário não autorizado fica no banco de dados, eles têm acesso a todos os tabelas, e não apenas algumas. Dependendo da natureza dos dados envolvidos, essas quebras na segurança também pode representar uma ameaça à privacidade individual. Cuidados também devem ser tomados regularmente para fazer cópias de backup das tabelas e armazená-las por causa da possibilidade de incêndios e terremotos que poderiam destruir o sistema.

Lição de encerramento

Nesta lição, um sistema de gerenciamento de base de dados foi definida, bem como os seus efeitos e funções. Um dos aspectos mais poderosos de um SGBD é a capacidade de organizar e recuperar dados a partir de diferentes, mas relacionadas, tabelas. No entanto, usando Bancos de Dados tem suas vantagens e desvantagens. À medida que avançar com a sua carreira, você deve estar ciente das vantagens e desvantagens que acompanham o uso dessas ferramentas informatizadas. As compensações que temos discutido até agora incluem coisas como a redundância, precisão, acessibilidade e facilidade de utilização de dados em um SGBD. Ser educado sobre os pontos fortes e fracos de SGBDs lhe permitirá tomar decisões mais eficazes sobre como organizar e utilizar os dados.

Agora que você completou esta lição, você deve ser capaz de:

- Definir o termo sistema de gerenciamento de banco de dados (SGBD).
- Descrever o propósito e funções básicas de um SGBD.
- Discutir as vantagens e desvantagens de SGBDs.

2. USUÁRIOS

Todo agrupamento de bancos de dados **possui um conjunto de usuários** de banco de dados. Estes usuários **são distintos dos usuários gerenciados pelo sistema operacional** onde o servidor executa. Os usuários possuem objetos de banco de dados (por exemplo, tabelas), e podem **conceder privilégios** nestes objetos para outros usuários controlando, assim, quem pode acessar qual objeto.

Depois de ler esta lição, você deve ser capaz de:

- Definir os tipos de usuários de banco de dados.
- Descrever o propósito e funções básicas de um usuário.

Administrador de Banco de Dados (DBA)

Em um ambiente de banco de dados, o recurso primário é o banco de dados por si só e o

recurso secundário o SGBD e os softwares relacionados. A administração destes recursos cabe ao **Administrador de Banco de Dados**, o qual é **responsável pela autorização de acesso** ao banco de dados e pela **coordenação e monitoração de seu uso**. Ou seja ele coordena todas as atividades do sistema de banco de dados; possui boa compreensão dos recursos de informação da empresa e suas necessidades.

Suas funções incluem:

- Definição do esquema
- Estrutura de armazenamento e definição de acesso aos dados
- Esquema físico e organização
- Concede acesso aos usuários
- Cuida da integridade dos dados
- Atua como elo com os usuários
- Acompanha a desempenho, e responde as mudanças exigidas
- Atividades de manutenção (Backups)

Projetista de Banco de Dados

O Projetista de Banco de Dados é **responsável pela identificação dos dados que devem ser armazenados** no banco de dados, escolhendo a estrutura correta para representar e armazenar dados. Muitas vezes, os projetistas de banco de dados atuam como "staff" do DBA, assumindo outras responsabilidades após a construção do banco de dados. É função do projetista também **avaliar as necessidades de cada grupo de usuários para definir as visões** que serão necessárias, integrando-as, fazendo com que o banco de dados seja capaz de atender a todas as necessidades dos usuários.

Usuários Finais

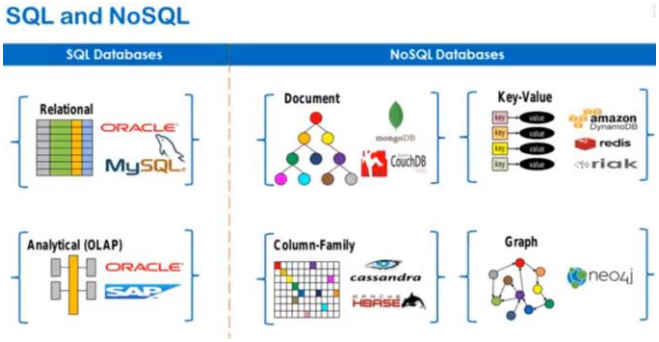
Existem basicamente **três categorias de usuários finais** que são os usuários finais do banco de dados, fazendo consultas, atualizações e gerando documentos:

- **Usuários casuais**: acessam o banco de dados casualmente, mas que podem necessitar de diferentes informações a cada acesso; utilizam sofisticadas linguagens de consulta para especificar suas necessidades;
- **Usuários novatos** ou paramétricos: utilizam porções pré-definidas do banco de dados,utilizando consultas pre estabelecidas que já foram exaustivamente testadas;
- **Usuários sofisticados**: são usuários que estão familiarizados com o SGBD e realizam consultas complexas.

Analistas de Sistemas e Programadores de Aplicações

Os **analistas determinam os requisitos** dos usuários finais e desenvolvem especificações para transações que atendam estes requisitos, e os **programadores implementam estas especificações como programas,** testando, depurando, documentando e dando manutenção no mesmo. É importante que, tanto analistas quanto programadores, estejam a par dos recursos oferecidos pelo SGBD.

3. TIPOS DE SGBD (DBMS)



SGBDs vêm em muitas formas e tamanhos. Por algumas centenas de dólares ou até mesmo de forma gratuita, você pode comprar um SGBD para o seu computador desktop. Para sistemas maiores os SGBDs podem ser muito mais caros. Muitos SGBDs são baseados em mainframe e alugados por organizações. SGBDs desta escala são altamente sofisticados e seria extremamente caro para desenvolver a partir do zero. Portanto, é mais barato para uma organização alugar um programa que desenvolvê-lo. Uma vez que há uma variedade de SGBDs disponíveis, você deve conhecer algumas das características básicas, bem como os pontos fortes e fracos, dos principais tipos.

TIPOS ESTRUTURAIS DE SISTEMAS DE GERENCIAMENTO DE BANCO DE DADOS MAIS UTILIZADOS:

Relacional

Os sistemas de gerenciamento de banco de dados relacional (RDBMS) suportam o modelo de dados relacional (= orientado a tabela). O esquema de uma tabela (= esquema de relação) é definido pelo nome da tabela e um número fixo de atributos com tipos de dados fixos. Um registro (= entidade) corresponde a uma linha da tabela e consiste nos valores de cada atributo. Uma relação, portanto, consiste em um conjunto de registros uniformes.

Os esquemas de tabelas são gerados por normalização no processo de modelagem de dados.

Certas operações básicas são definidas nas relações:

- operações clássicas de conjunto (união, interseção e diferença)
- Seleção (seleção de um subconjunto de registros de acordo com certos critérios de filtro para os valores de atributo)
- Projeção (selecionando um subconjunto de atributos / colunas da tabela)
- Join: conjunção especial de múltiplas tabelas como uma combinação do produto cartesiano com seleção e projeção.

Essas operações básicas, bem como as operações de criação, modificação e exclusão de esquemas de tabelas, operações de controle de transações e gerenciamento de usuários são

realizadas por meio de linguagens de banco de dados, sendo o SQL um padrão bem estabelecido para tais linguagens.

Os primeiros sistemas de gerenciamento de banco de dados relacional surgiram no mercado no início da década de 1980 e, desde então, são o tipo mais comumente usado.

Em **bancos de dados relacionais,** a **relação entre as tabelas** de dados é relacional. Bancos de dados relacionais conectam dados em tabelas diferentes, **usando elementos comuns de dados ou um campo chave**. Dados em bancos de dados relacionais **são armazenados em tabelas diferentes**, cada uma com um **campo chave que identifica cada linha ou registro**. Bancos de dados relacionais **são muito mais flexíveis** do que as próprias estruturas de dados hierárquicos ou rede. Em bancos de dados relacionais a ligação entre as tabelas são chamadas de **relações,** as **tuplas** designam uma linha ou registro, e as colunas são referidas como **atributos** ou campos.

Bancos de dados relacionais trabalham no princípio de que **cada tabela tem um campo chave que identifica unicamente cada linha**, e que estes campos chave podem ser usados para ligar uma tabela de dados a outra. Deste modo, uma tabela pode ter uma linha formada por um número de conta de cliente, tal como o campo chave, juntamente com o endereço e número de telefone. O número de conta do cliente nesta tabela pode estar ligada a uma outra tabela de dados que inclui também o número de conta do cliente (um campo de chave), mas, neste caso, contém informações sobre a devolução de produtos, incluindo um número de ordem (um outro campo de chave). Este campo chave pode ser ligado a uma outra tabela que contém números de itens e outras informações do produto, tais como local de produção, cor e outros dados. Portanto, usando esse banco de dados, as informações dos clientes podem ser ligado a informações específicas do produto.

O banco de dados relacional se tornou bastante popular, por duas razões principais. Em primeiro lugar, os bancos de dados relacionais **podem ser usados com pouca ou nenhuma formação**. Segundo, as **entradas de banco de dados podem ser modificadas sem redefinir a sua estrutura inteira**.

A **desvantagem** de usar um banco de dados relacional **é que a busca de dados pode levar mais tempo do que se outros métodos são usados**.

Ao longo dos anos, muitosSGBDs foram expandidos com conceitos não relacionais, como tipos de dados definidos pelo usuário, não atributos atômicos, herança e hierarquias, motivo pelo qual às vezes são chamados de DBMS objeto-relacional.

Exemplos mais populares:

Banco de dados orientado a Documentos

Os armazenamentos de documentos, também chamados de sistemas de banco de dados orientados a documentos, são caracterizados por sua organização de dados sem esquemas.

Que significa:

- Os registros não precisam ter uma estrutura uniforme, ou seja, registros diferentes podem ter colunas diferentes.
- Os tipos de valores de colunas individuais podem ser diferentes para cada registro.
- As colunas podem ter mais de um valor (matrizes).
- Os registros podem ter uma estrutura aninhada.

Os armazenamentos de documentos costumam usar notações internas, que podem ser processadas diretamente em aplicativos, principalmente **JSON**. Os documentos JSON, também podem ser armazenados como texto puro em armazenamentos de **valores-chave** ou sistemas de banco de dados relacionais . Isso, no entanto, exigiria o processamento das estruturas do lado do cliente, o que tem a desvantagem de os recursos oferecidos pelos armazenamentos de documentos (como índices secundários) não estarem disponíveis.

Exemplos mais populares

Armazenamento de valores-chave

Armazenamento de valores-chave são provavelmente a forma mais simples de sistemas de gerenciamento de banco de dados . Eles só podem armazenar pares de chaves e valores, bem como recuperar valores quando uma chave é conhecida.

Esses sistemas simples normalmente não são adequados para aplicativos complexos. Por outro lado, é exatamente essa simplicidade que torna esses sistemas atraentes em certas circunstâncias. Por exemplo, armazenamentos de valores-chave com eficiência de recursos são frequentemente aplicados em sistemas incorporados ou como bancos de dados em processo de alto desempenho.

Formulários Avançados

Uma forma estendida de armazenamentos de valores-chave é capaz de classificar as chaves e, portanto, permite consultas de intervalo, bem como um processamento ordenado de chaves.

Muitos sistemas fornecem extensões adicionais para que possamos ver uma transição razoavelmente contínua para armazenamentos de documentos e grandes armazenamentos de colunas .

Exemplos mais populares

Lição de encerramento

Como vimos, os SGBDs são de várias formas. As diferentes estruturas de SGBDs foram comparadas e contrastadas num esforço para ajudar a demonstrar seus pontos fortes e fracos. Como um trabalhador do conhecimento, você pode um dia ser convidado a selecionar e tomar decisões sobre um

SGBD. Esta lição fornece um ponto de partida para a compreensão das questões envolvidas.

Agora que você completou esta lição, você deve ser capaz de:

- Comparar e contrastar a estrutura dos diferentes sistemas de gerenciamento de banco de dados.
- Definir Bancos de Dados relacionais.
- Definir os Bancso de dados orientado a Documentos.
- Definir Bancos de Dados de Armazenamento de valores-chave.

4. MODELO RELACIONAL

Hoje em dia a maioria dos sistemas de base de dados **são do tipo relacional**.

Databases relacionais têm valiosos atributos que a distinguem como superior. Provavelmente o mais importante é **você pode mudar a estrutura de dados sem alterações nas aplicações**. Suponha, por exemplo, que você adicione uma ou mais colunas numa tabela. Você não precisa alterar nenhum aplicativo que o sistema vai continuar a processar. Claro, se você remover uma coluna que uma aplicação existente utiliza, você vai ter problemas.

Depois de ler esta lição, você deve ser capaz de:

- Definir Bancos de Dados relacionais.
- Discutir as funções e capacidades de uma Base de Dados Relacional
- Definir os objetos de um Banco de Dados Relacional

Objetos de banco de dados relacional

Veremos agora uma breve descrição dos objetos que compõem um banco de dados do tipo relacional. Eles serão vistos com mais detalhes a medida em que se aprofundaremos no curso.

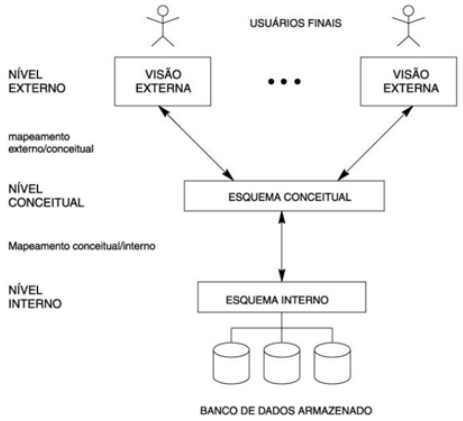
Tabelas	São os objetos que contém os tipos de dados e os dados reais
Colunas ou Campos	São as partes das tabelas que armazenam os dados. Devem receber um tipo de dados e ter um nome único
Tipos de dados	Há vários tipos de dados para serem utilizados como: caráctere, número, data. Um único tipo de dados é atribuído a uma coluna dentro de uma tabela
Storeds Procedures (procedimentos armazenados)	São como macros em que o código Transact-SQL pode ser escrito e armazenado sob um nome.
Triggers (gatilhos)	São como storeds procedures que são automaticamente ativados quando os dados são inseridos, alterados ou apagados. Asseguram que regras de negócio e de integridade sejam impostas ao banco de dados.
Regras (rules)	São atribuídas a colunas de modo que os dados que estão sendo inseridos devem se adaptar aos padrões definidos. Por exemplo, pode-se utilizar regras para permitir que um campo que irá armazenar a UF contenha somente Estados válidos.
Chaves Primárias (PK)	Embora não sejam objetos em si, as chaves são essenciais para os bancos de dados relacionais. Promove a carácterística de unicidade das linhas, proporcionando uma maneira de identificar de forma única cada item que você queira armazenar.
Chaves Estrangeiras (FK)	Novamente, não são objetos em si, as chaves estrangeiras são colunas que fazem referências as chaves primárias de outras tabelas.
Padrões (Defaults)	Podem ser configurados em campos de modo que, se nenhum dado for inserido durante uma operação de Insert, os valores padrão serão utilizados.
Views (visualizações)	Consistem basicamente em consultas armazenadas nos bancos de dados que podem fazer referência a uma ou muitas tabelas. Você pode criar e salvar views e utiliza-las no futuro. Normalmente excluem certas colunas de uma tabela e vinculam duas ou mais tabelas entre si. Podem ser utilizadas também como mecanismo de segurança.
Índices	Podem ajudar os dados de modo que as consultas executem mais rápido

___Fonte: http://ehgomes.com.br/disciplinas/bdd/sgbd.php

5. ARQUITETURA DE TRÊS ESQUEMAS E A INDEPENDÊNCIA DOS DADOS

O Sistema de Banco de dados deve prover uma visão abstrata dos dados para os usuários. Essa Abstração se dá em três níveis, o primeiro nível é o externo, o segundo nível é o conceitual e o terceiro nível é o interno.

Arquitetura de Três Esquemas e a Independência de Dados



Nível Interno (Físico):

Nível mais baixo de abstração. Descreve como os dados estão realmente armazenados, englobando estruturas complexas de baixo nível e descreve os detalhes completos do armazenamento de dados e caminho de acesso ao banco de dados.

Nível Conceitual:

Descreve quais dados estão armazenados e seus relacionamentos. Neste nível, o Banco de dados é descrito através de estruturas relativamente simples, que podem envolver estruturas complexas no nível físico. Concentra-se na descrição de entidades, tipos de dados, conexões, operações de usuários e restrições.

Nível Externo (visões do usuário):

Descreve partes do banco de dados, de acordo com as necessidades de cada usuário, individualmente ocultando o restante do banco de dados.

Linguagens de SGBD

O SGBD deve oferecer linguagens e interfaces apropriadas para cada categoria de usuários.

Linguagem de Definição de dados (DDL)

Está Linguagem é utilizada para permitir especificar o esquema do banco de dados, através de um conjunto de definições de dados.

A Compilação dos comandos em DDL é armazenada no dicionário de dados (metadados).

Linguagem de Manipulação de dados (DML)

Está Linguagem permite ao usuário acessar ou manipular os dados, vendo-os da forma como são definidos no nível de abstração mais alto do modelo de dados utilizado.

Uma Consulta (“query”) é um comando que requisita uma recuperação de informação. A parte de uma DML que envolve recuperação de informação é chamada **linguagem de consulta**.

Manipulação de dados:

·Recuperação da informação armazenada

· Inserção de novas informações

· Exclusão de informações

· Modificação de dados armazenados

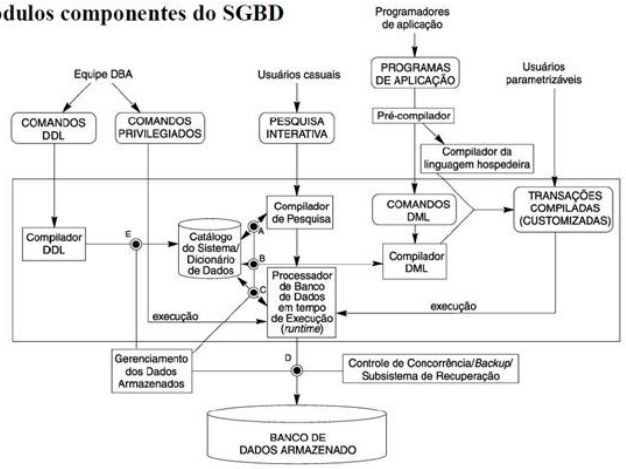
Módulos componentes do SGBD

· Módulo de programa que fornece a interface entre os dados de baixo nível de armazenados num banco de dados e os programas aplicativos ou as solicitações submetidas ao sistema

· Software que manipula todos os acessos ao banco de dados, proporcionando a interface de usuário ao sistema de banco de dados.

Ilustrando o papel do sistema de gerência de banco de dados, de forma conceitual:

Módulos componentes do SGBD



O usuário emite uma solicitação de acesso.

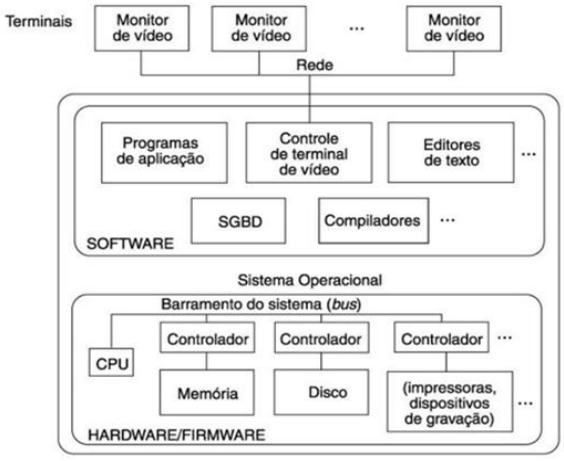
O SGBD intercepta a solicitação e a analisa.

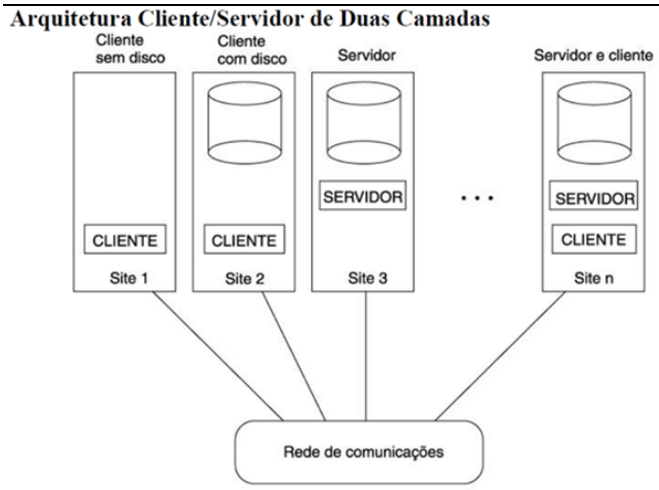
O SGBD inspeciona os esquemas externos (ou subesquemas) relacionados aquele usuário, os mapeamentos entre os três níveis , e a definição da estrutura de armazenamento.

O SGBD realiza as operações solicitadas no banco de dados armazenado.

Abaixo, segue imagens com arquiteturas utilizadas em SGBD´s:

Arquitetura SGDB Centralizada





CRÉDITO DO TEXTO:
Emerson S. Gaudêncio. Disponível em https://cooperati.com.br/2012/08/banco-de-dados-conceituando-banco-de-dados-e-sgbd/

MÉTRICAS DE TRANSFORMAÇÃO: COMPARAÇÕES COM BASE NO TEMPO E OUTRAS

As transformações permitem que você aplique um deslocamento de atributo-elemento para comparação de dados métricos. Uma métrica de transformação pode, por exemplo, ajudar um usuário a comparar a receita do último mês com a receita do mês atual. Embora as transformações possam ser aplicadas a qualquer hierarquia de atributo, a hierarquia de tempo é usada mais frequentemente. Para a hierarquia de tempo, o deslocamento pode ser definido como um número fixo de dias, semanas, meses ou anos.

Transformações com base no tempo

As métricas usam transformações de tempo para comparar valores em momentos diferentes, como este ano em relação ao ano passado ou a data atual em relação mês atual. Por exemplo, a transformação do Ano Passado mapeia cada período de tempo para seu período correspondente do ano passado, enquanto a transformação do Mês Atual mapeia cada período de tempo para um conjunto de períodos que compreendem todo o mês até agora.

Na imagem abaixo, a métrica Valor Real exibe números de conta do trimestre atual. A transformação do último trimestre é aplicada à métrica de Valor Real para criar a métrica Valor Real - Último Trimestre, que exibe os números das contas do último trimestre. A diferença entre os conjuntos de números pode, então, ser calculada e exibida na métrica Valor Real - Diferença do Último Trimestre. As transformações são úteis para essas análises de séries de tempo, que são relevantes para vários setores, incluindo varejo, serviços bancários e telecomunicações.

Account Type	Metrics	Actual Amount	Actual Amount - Last Quarter	Actual Amount - Last Quarter Difference	% Variance
Salaries		$328,633	$302,762	$25,871	8.55%
High-Tech and Communications Expense		$37,312	$39,709	($2,397)	(6.04%)
Recruiting Expense		$1,163	$2,326	($1,163)	(50.00%)
Travel and Entertainment (T&E)		$411	$1,284	($873)	(68.02%)
Casual Labor		$3,022	$6,923	($3,901)	(56.35%)
Communications		$9,806	$9,675	$130	1.34%
Shipping, Printing, Supplies		$13,109	$12,497	$612	4.90%
Consulting and Advisory		$31,320	$12,908	$18,412	142.64%
Depreciation		$66,033	$73,998	($7,964)	(10.76%)
Other General and Administrative		$878	$10,395	($9,516)	(91.55%)
Insurance		$311,031	$326,598	($15,567)	(4.77%)
Cost of Goods Sold		$112	$586	($473)	(80.85%)
Total		**$802,830**	**$799,659**	**$3,170**	**0.40%**

Embora exista outros métodos para executar esses tipos de cálculos,

as transformações geralmente são a abordagem mais genérica e podem ser reutilizadas e aplicadas a outras análises de sequências temporais. Por exemplo, outro tipo comum de análise de sequência temporal é uma comparação TY/LY (This Year versus Last Year ou Este Ano x Ano Passado). Você pode usar filtros para criar a comparação TY/LY, conforme a seguir:

- Para calcular a receita deste ano, use um filtro para este ano com a métrica de Receita.
- Para calcular a receita do último ano, use um filtro para o último ano com a métrica de Receita.

No entanto, uma alternativa mais flexível é usar uma transformação do Ano Passado criada anteriormente para definir uma nova métrica, chamada de Receita no Ano Passado. Você pode, então, usar um único filtro em 2003 nas métricas Receita e Receita do Ano Passado para obter os resultados para 2003 e 2002, respectivamente. Embora a abordagem de filtro requeira a criação de dois filtros, a abordagem de transformação requer apenas um. Além disso, com a abordagem de

transformação, a mesma métrica de transformação pode ser aplicada a um relatório com um filtro adequado para definir análises semelhantes em diferentes conjuntos de dados, embora a abordagem de filtros signifique que novos filtros precisariam ser criados para construir cada novo relatório.

Como uma transformação representa uma regra, ela pode descrever o efeito dessa regra para diferentes níveis de dados. Por exemplo, a transformação do Ano Passado descreve de forma intuitiva como um ano específico refere-se ao ano anterior. Ela também pode expressar como cada mês do ano corresponde a um mês do ano anterior. Da mesma maneira, a transformação pode descrever como cada dia do ano mapeia para um dia do ano anterior. Essas informações definem a transformação e abstraem todos os casos em um conceito genérico. Ou seja, você pode usar uma única métrica com uma transformação no ano passado, independentemente do atributo de tempo contido no relatório.

TRANSFORMAÇÕES ATEMPORAIS

Enquanto transformações são mais frequentemente usadas para descobrir e analisar as tendências com base no tempo em seus dados, nem todas as transformações precisam ser baseadas no tempo. Por exemplo, uma transformação pode mapear os códigos de produtos extintos a novos. Um exemplo de uma transformação atemporal é Este Catálogo/Último Catálogo, que pode subtrair um número de um código de produto antigo para convertê-lo em um novo.

A análise de transformação-estilo também pode ser suportada usando as funções Lag e Lead fornecidas com MicroStrategy. Essas funções podem ser usadas para definir métricas que comparam valores de diferentes períodos, sem o uso da métrica de transformação.

Criando uma métrica de transformação

Para criar uma métrica, você deve definir sua fórmula, que consiste em:

- **Função**: o cálculo aplicado aos dados de negócios, como Soma ou Contagem. Dependendo de como a métrica é criada, sua métrica pode conter várias funções.
- **Expressão**: os dados de negócios da fonte de dados. A expressão pode conter fatos de negócios, atributos ou outras métricas.

A seguir, são indicadas as etapas para adicionar uma transformação a uma métrica.

A transformação a ser aplicada à métrica já foi criada.

Para adicionar uma transformação a uma métrica

1. Clique em **Criar** em qualquer página e selecione **Nova Métrica**. A caixa de diálogo Selecionar uma Função ou Template é aberta.
2. Na lista suspensa, selecione **Templates de Métrica** (na parte inferior da lista). Selecione **Transformação**.
3. Clique em **Próximo**. O Editor de função será aberto.
4. Na lista suspensa **Função de Agregação**, selecione a função a ser usada para criar a métrica.
5. Defina a expressão da métrica seguindo um destes procedimentos:
   - Para digitar objetos a serem adicionados à expressão diretamente, digite o nome do objeto no campo **Expressão**. À medida que você digita, objetos correspondentes são exibidos em uma lista suspensa. Você pode clicar em um objeto ou continuar a digitar. Você pode digitar vários objetos, como Receita-Lucro.
   - Para especificar a expressão escolhendo um objeto, clique no ícone **Navegar**. A caixa de diálogo Selecionar um objeto será aberta. Navegue e selecione um objeto ou pesquise um objeto.

Adicionar uma transformação

1. Se as opções da área de Transformação não forem exibidas, clique em **Transformação**.

2. Clique no ícone **Pesquisar (Browse)** na área Transformação. A caixa de diálogo Selecionar um objeto será aberta. Navegue até uma transformação e selecione-a, ou pesquise uma transformação. Depois de selecionar uma transformação, retorne ao Editor de Funções.

3. Você pode remover as transformações que adicionou à métrica ou alterar a ordem em que as transformações são aplicadas. Escolha dentre as seguintes opções:
   - Para remover uma transformação, clique no ícone **Excluir** a lado da transformação.
   - Para alterar a ordem das transformações selecione uma transformação e use as setas direcionais para mover para cima ou para baixo.

4. Repita as etapas apropriadas descritas acima para definir as transformações adicionais desejadas.

5. Você pode definir como o cabeçalho e valores da métrica são formatados e exibidos no relatório. Por exemplo, você pode definir como os valores numéricos são exibidos, os estilos e tamanhos das fontes e as cores de exibição das células. A formatação será aplicada à métrica, independentemente do relatório no qual ela esteja inserida.

6. Clique em **Salvar** para aplicar as alterações. A caixa de diálogo Salvar como é aberta. Navegue até a pasta em que deseja salvar a métrica. Digite um **Nome** e uma **Descrição** para a métrica e clique em **OK**. A nova métrica será salva.

7. É possível adicionar níveis ou uma condição à métrica. Escolha dentre as seguintes opções:

- Por padrão, uma métrica é calculada no nível dos atributos do relatório no qual ela está inserida. É possível definir o nível de atributo a ser usado no cálculo da métrica, independentemente do que está contido em qualquer relatório no qual a métrica é colocada. Para conhecer as etapas, consulte Por padrão, as métricas são avaliadas no nível dos atributos no relatório; isto denomina-se nível de relatório. O nível do relatório permite que o cálculo da métrica se adapte a diferentes relatórios. Você pode remover o nível do relatório da métrica. Se você fizer isso, apenas o nível explicitamente definido na métrica afetará o cálculo da métrica, independentemente dos atributos presentes no relatório. Você não precisa remover o nível do relatório para adicionar níveis à métrica. Para obter uma descrição mais detalhada do nível do relatório e o impacto de removê-lo, consulte o Ajuda para relatórios avançados. Execute um dos procedimentos a seguir: Para remover o nível do relatório da métrica, clique no ícone Excluir ao lado do Nível do Relatório. Para adicionar o nível do relatório à métrica após sua exclusão, clique no ícone Adicionar Nível do Relatório.. Para obter uma descrição mais detalhada de níveis, incluindo exemplos, consulte Métricas de Nível: Como modificar o contexto de cálculos de dados.
- Para adicionar uma condição, consulte Adicione a condição. Uma condição permite que você aplique um filtro para apenas uma métrica em um relatório enquanto não afetar as demais métricas. Para obter uma descrição mais detalhada de métricas condicionais, incluindo exemplos, consulte Métricas Condicionais: Como filtrar cálculos de dados.

_____Fonte: https://www2.microstrategy.com/producthelp/Current/MSTRWeb/WebHelp/Lang_1046/Content/Transformation_metrics.htm

TRANSFORMAÇÃO DE DADOS

O que é transformação de dados?

A transformação de dados é o processo de conversão de dados brutos de um formato para outro para torná-los utilizáveis pelo sistema ou aplicativo de destino. Inclui várias atividades, como 'transformar' seus dados, filtrando-os com base em certas regras e unindo diferentes campos para obter uma visão consolidada. As ferramentas de transformação de dados ajudam a alcançar seu resultado final com facilidade.

A transformação é uma etapa intermediária importante na extração, transformação e carregamento Processo (ETL) - um pré-requisito para o carregamento. A maioria Ferramentas ETL também vêm com funções predefinidas que podem ser usadas para transformar seus dados de forma rápida e eficiente. As empresas costumam enfrentar desafios de transformação devido à baixa qualidade dos dados.

Aqui estão algumas das etapas que estão envolvidas neste processo:

- Identifique a estrutura dos arquivos de origem e extraia dados deles
- Em seguida, mapeie os dados do arquivo de origem para a ferramenta de transformação
- Realize a transformação, ou seja, filtre, classifique, limpe ou agregue os dados
- Finalmente, envie o arquivo transformado para o destino

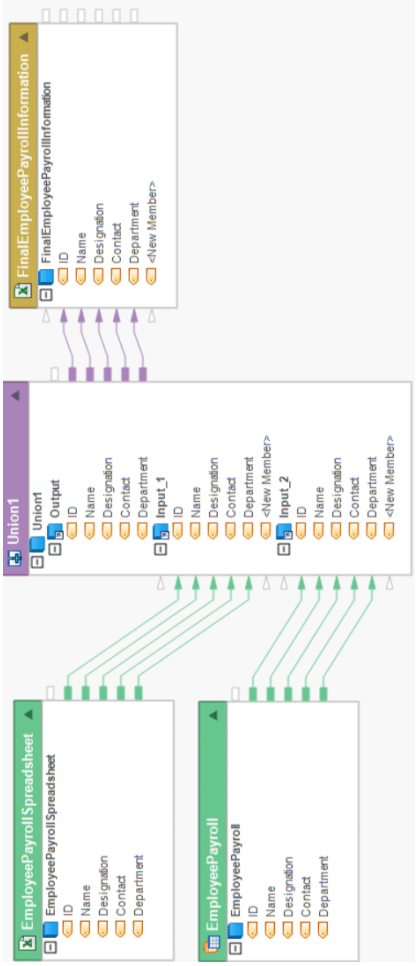
Por que a transformação de dados é importante?

As empresas precisam transformar grandes volumes de dados por diversos motivos, como migração de dados para a nuvem, consolidação de registros, exclusão de duplicatas, alteração de formatação, etc.

As transformações também são aplicadas para concatenar e validar dados, realizar pesquisas ou rotear dados para diferentes destinos. É benéfico ter uma ferramenta de transformação de dados com uma ampla gama de opções de transformação para poder manipular os dados da melhor maneira possível.

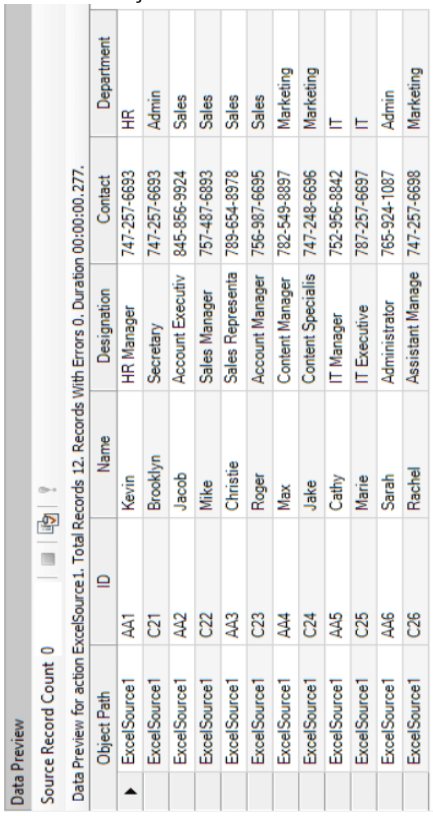
Vejamos um exemplo de transformação: suponha que um banco adquira uma seguradora que opera na mesma região. Uma vez concluída a aquisição, é decidido que uma única folha de pagamento será gerada para todos os funcionários. O processo de geração da folha de pagamento teria sido direto se todos os dados dos funcionários estivessem armazenados em um sistema unificado, como um data warehouse ou banco de dados.

No entanto, neste caso, uma empresa armazenou os dados dos funcionários em um SQL Server e a outra armazenou as informações da folha de pagamento em uma planilha do Excel. Para criar uma folha de pagamento consolidada para os funcionários, os dados precisam ser transformados para atender aos requisitos do sistema de destino, ou seja, arquivo Excel.



Dados da origem do SQL Server e Excel sendo transformados e mapeados para um arquivo de destino do Excel

Veja como o arquivo de destino cuida da transformação:



Data Preview

Source Record Count 0

Data Preview for action ExcelSource1. Total Records 12. Records With Errors 0. Duration 00:00:00.277.

	Object Path	ID	Name	Designation	Contact	Department
▲	ExcelSource1	AA1	Kevin	HR Manager	747-257-6693	HR
	ExcelSource1	C21	Brooklyn	Secretary	747-257-6693	Admin
	ExcelSource1	AA2	Jacob	Account Executiv	845-856-9924	Sales
	ExcelSource1	C22	Mike	Sales Manager	757-487-6893	Sales
	ExcelSource1	AA3	Christie	Sales Representa	789-654-8978	Sales
	ExcelSource1	C23	Roger	Account Manager	756-987-6695	Sales
	ExcelSource1	AA4	Max	Content Manager	782-549-8897	Marketing
	ExcelSource1	C24	Jake	Content Specialis	747-248-6696	Marketing
	ExcelSource1	AA5	Cathy	IT Manager	752-956-8842	IT
	ExcelSource1	C25	Marie	IT Executive	787-257-6697	IT
	ExcelSource1	AA6	Sarah	Administrator	765-924-1087	Admin
	ExcelSource1	C26	Rachel	Assistant Manage	747-257-6698	Marketing

Visualização da folha de pagamento consolidada armazenada no arquivo do Excel

As transformações também podem ser usadas para extrair valores de diferentes tipos de dados. Em vez de sobrecarregar seus sistemas com vários - muitas vezes desnecessários - registros, você pode usar diferentes tipos de transformações de dados para filtrar dados irrelevantes.

Por exemplo, se você deseja gerar um relatório de todas as vendas realizadas em um determinado país, digamos os EUA, aplicando o **filtros** A transformação evitará que o sistema de destino seja sobrecarregado desnecessariamente, pois apenas os registros relevantes serão transmitidos.

Armazenar registros relevantes e comparativamente menores no sistema de destino significa menos consumo de memória durante o processamento do pipeline de dados, o que reduzirá o tempo de execução.

Transformação de dados no local, baseada em nuvem ou manual: qual você deve escolher?

O recurso de transformação de dados está disponível em várias ferramentas de integração de dados. Uma coisa boa sobre esse processo é que você pode fazer isso de várias maneiras. Cada abordagem, no entanto, vem com seus benefícios e desafios exclusivos. Vejamos algumas das técnicas comuns de transformação.

Transformação de dados no local

A transformação no local permite que as empresas extraiam, transformem e carreguem dados cruciais muito rapidamente. Ter uma ferramenta de transformação de dados no local também se traduz em maior conformidade regulatória e melhor gerenciamento de segurança.

A maioria dessas ferramentas de transformação vem com a funcionalidade de arrastar e soltar, o que significa que funcionários não técnicos serão capazes de transformar dados e convertê-los em um formato utilizável.

Uma coisa importante a ser observada é que as ferramentas no local são implantadas no local; portanto, você deve garantir que possui recursos suficientes em termos de capacidade de armazenamento e sistemas adequados para executar esses tipos de software.

Transformação de dados baseada em nuvem

O recurso pay-as-you-go da maioria das ferramentas de transformação de dados baseadas em nuvem dá às empresas a liberdade de aumentar ou diminuir a escala conforme e quando necessário. É uma das razões pelas quais as ferramentas baseadas em nuvem são populares. No entanto, ter dados corporativos cruciais salvos em um servidor baseado em nuvem de terceiros traz consigo a sua parcela de preocupações com a segurança.

Um benefício de transformar dados na nuvem é que os tempos de instalação e configuração são reduzidos significativamente, o que significa que as

empresas podem transformar seus dados sem muitos atrasos.

Transformação manual de dados

A transformação manual envolve muita codificação. Isso significa que você precisa contratar um recurso de desenvolvimento dedicado para fazer o trabalho.

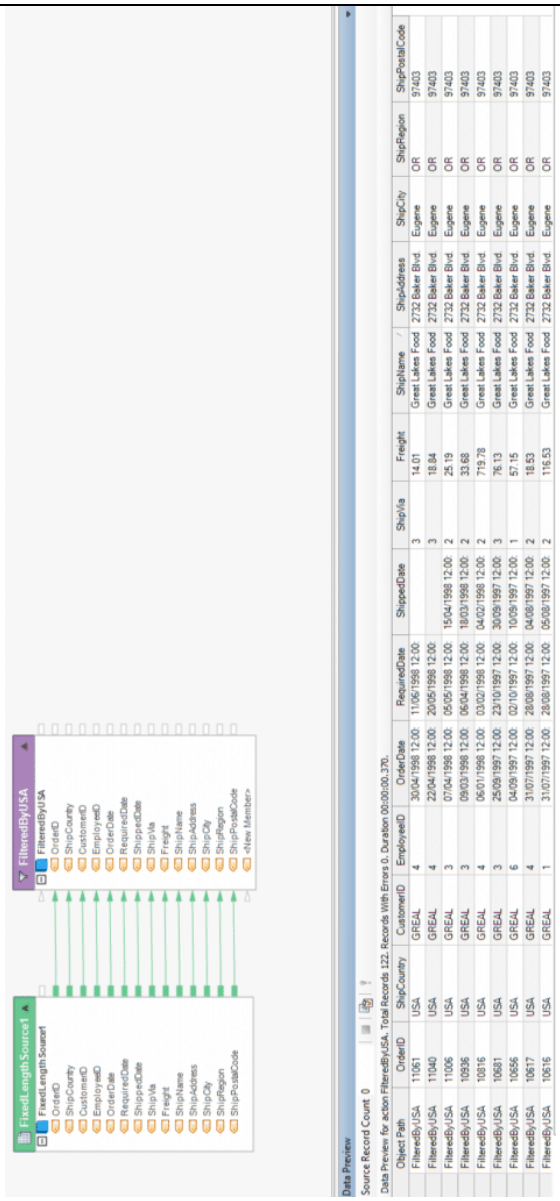
A transformação manual de dados pode consumir muito tempo e recursos, especialmente quando você está lidando com vários formatos de arquivo. Além disso, o risco de erro humano e exclusão acidental de dados de negócios importantes também aumenta ao transformar os dados manualmente. Portanto, é aconselhável usar uma ferramenta automatizada de transformação de dados.

Como transformar seus dados

Os dados podem ser transformados de várias maneiras, dependendo do seu objetivo e dos requisitos do sistema de destino. O usuário deve estar ciente de certas regras e exemplos de transformação ao usar o software. As transformações pré-criadas podem não apenas ser usadas para limpar, filtrar, dividir e juntar dados, mas também para enriquecê-los. Aqui estão alguns tipos de transformações de dados:

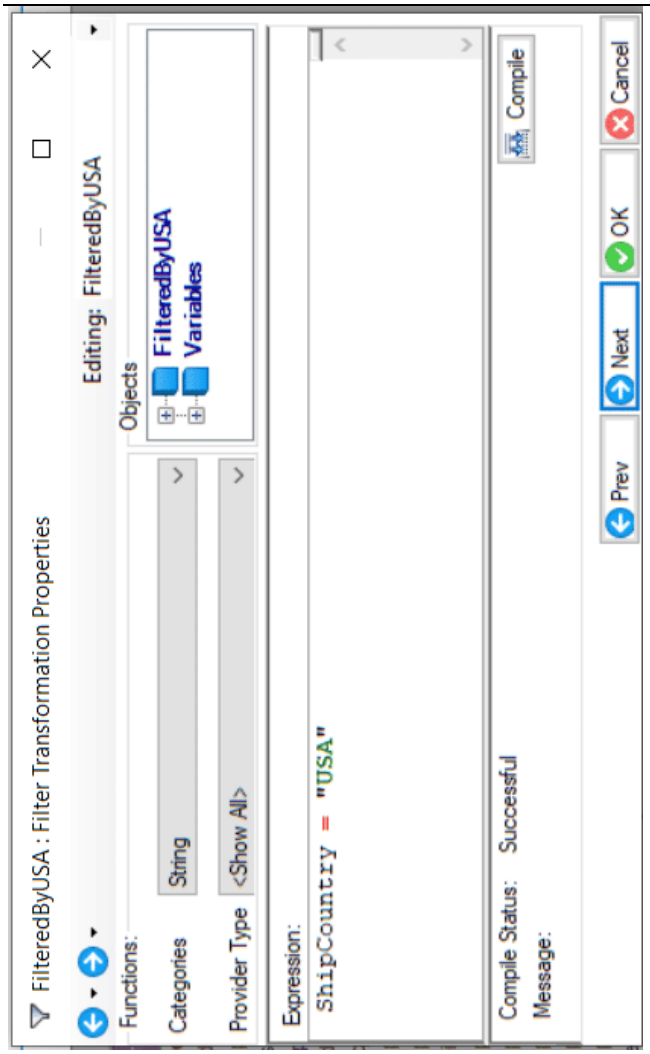
Filtrando dados

As empresas precisam processar vários registros para recuperar dados relevantes para um cenário específico. Os dados podem ser filtrados com base em uma ou mais regras. Você pode usar esses dados transformados para processamento adicional sem fazer o sistema de destino funcionar em registros irrelevantes.



Dados de uma fonte de comprimento fixo sendo filtrados para exibir registros dos EUA

No exemplo de transformação de dados acima, o **filtros** a transformação é aplicada em um documento de origem Fixed Length para mostrar registros apenas dos EUA.

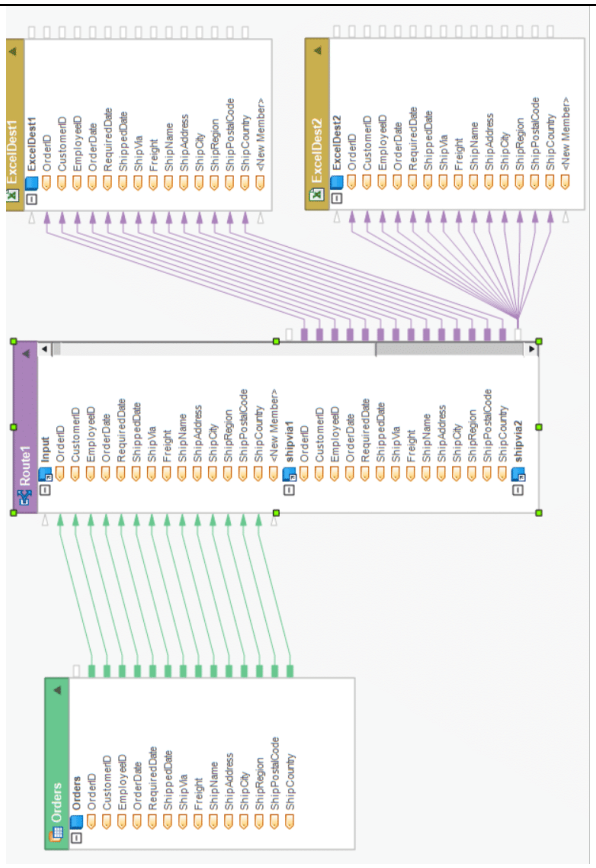


Definindo a expressão que será usada para filtrar os dados

Dados de roteamento

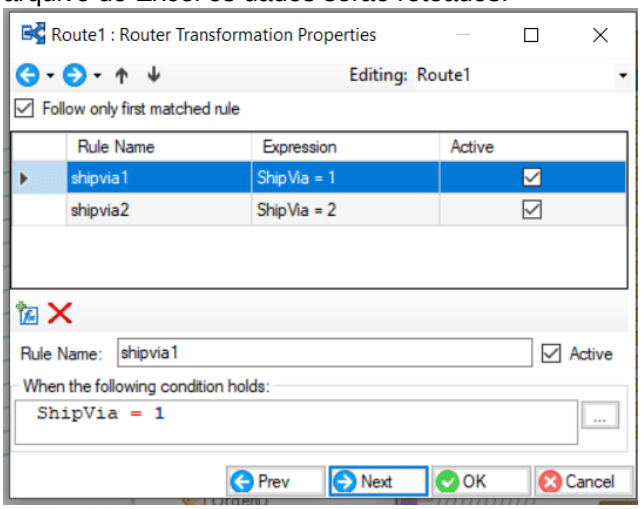
Você pode usar o **Rota** transformação para direcionar dados de origem para diferentes caminhos de formatos semelhantes ou diferentes, com base em certas regras lógicas.

Aqui está um fluxo de dados para ilustrar o recurso:



Dados do SQL Server sendo roteados para dois destinos do Excel

A vida do **Rota** A transformação aqui é usada para enviar os dados recuperados do SQL Server para dois arquivos de destino diferentes do Excel. As condições no campo ShipVia determinam para qual arquivo do Excel os dados serão roteados.



Regras da transformação de rota na tela Propriedades da transformação do roteador

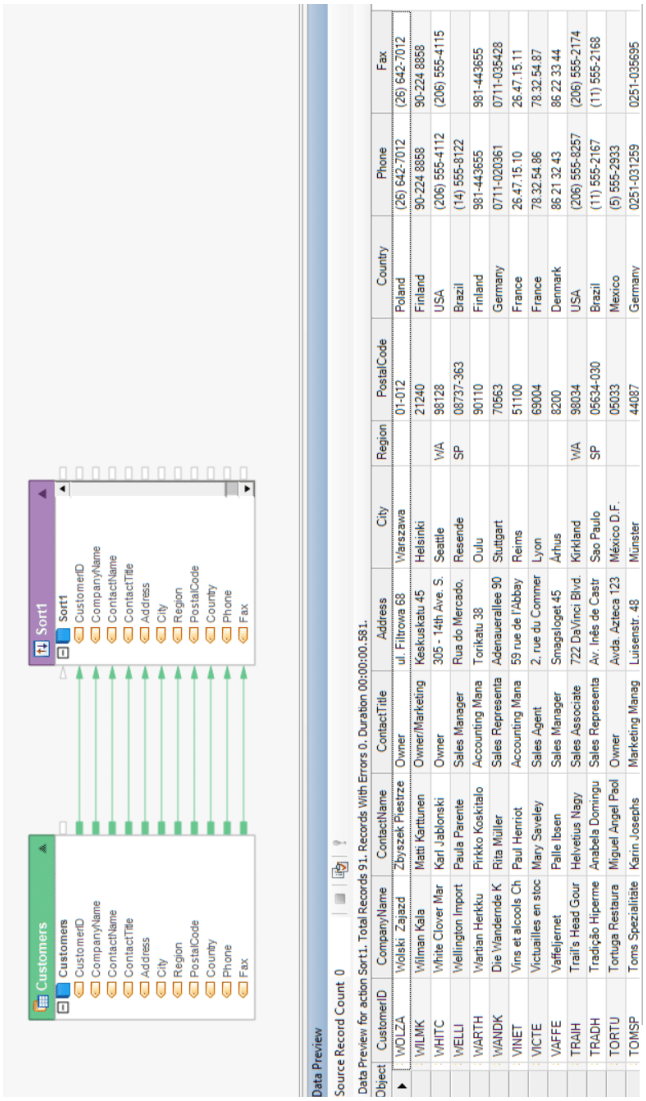
As condições aqui podem ser simples ou complexas, dependendo de seus requisitos.

Classificação de dados

As grandes empresas geralmente precisam classificar seus dados para torná-los mais gerenciáveis. o **tipo** a transformação pode ser

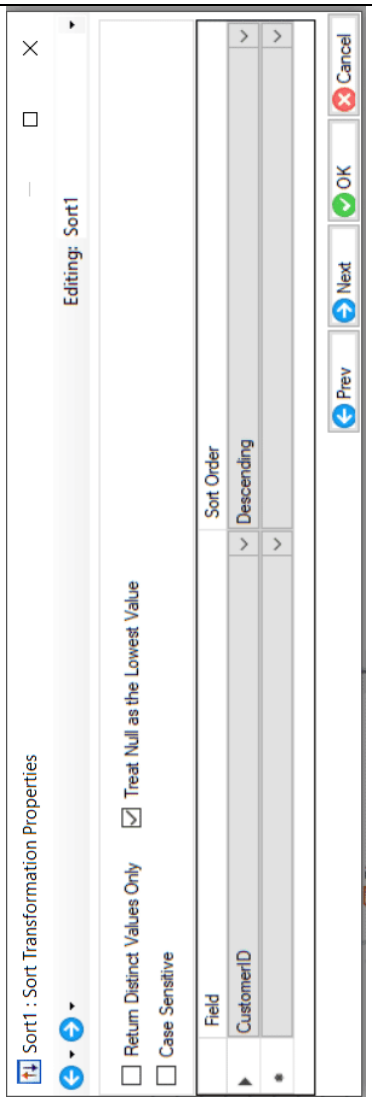
aplicada a qualquer campo para organizar a saída em ordem crescente ou decrescente.

Aqui está um exemplo no qual a transformação Classificar é usada para exibir registros de ID do cliente em ordem decrescente:



A visualização dos dados do cliente armazenados no banco de dados SQL Server é classificada em ordem decrescente de CustomerID

Isso foi feito selecionando o campo CódigoDoCliente e selecionando a ordem de classificação como decrescente da seguinte maneira.



Propriedades da transformação Classificar

A caixa Retornar somente valores distintos também pode ser marcada para remover redundâncias e exibir registros exclusivos.

Benefícios da transformação de dados

As ferramentas de transformação, quando usadas corretamente, podem melhorar significativamente a qualidade dos dados e melhorar a eficiência do processo. Os dados transformados são mais fáceis de usar, confiáveis e compatíveis com os sistemas e aplicativos finais. Os dados transformados de alta qualidade garantem que o sistema de destino tenha apenas dados com o formato e a estrutura exigidos.

Aqui estão alguns outros benefícios das ferramentas de transformação de dados:

- Eles podem ajudar as empresas a colher o máximo valor de seus dados.
- A padronização de dados por meio de transformações pode melhorar o gerenciamento de dados.
- Os dados transformados podem ser utilizados por várias ferramentas para diferentes aplicativos, como visualizações, relatórios, análises, etc.

Fonte: https://www.astera.com/pt/type/blog/data-transformation-tools/

ANÁLISE DE DADOS. AGRUPAMENTOS. TENDÊNCIAS. PROJEÇÕES.

ANÁLISE DE DADOS é o processo de aplicação de técnicas estatísticas e lógicas para avaliar informações obtidas a partir de determinados processos. O principal objetivo da prática é extrair informações úteis a partir dos dados. A partir destas informações, é possível tomar decisões mais assertivas e orientadas para resultados.

TIPOS SE ANALISE DE DADOS

A seguir as definições e indicações para aplicação das 4 principais metodologias de avaliação de informações.

1. Análise descritiva

Nesse tipo de análise, os dados são utilizados para fazer projeções de cenários e identificar tendências futuras a partir de determinados padrões.

Como o próprio nome diz, a análise descritiva é um dos tipos de análise de dados baseado em fatos. Isso significa que, na prática, este tipo de avaliação de dados é feita a partir de resultados obtidos. São exemplos de análise de dados descritiva:

- Relatórios;
- Segmentação e controle de clientes;
- Análises de negócio;
- Aplicação de métricas;
- Avaliação de resultados.

Um dos principais usos para a análise descritiva é orientar a construção de estratégias

2. Análise preditiva

O mais popular dos tipos de análise de dados é justamente o modelo preditivo. Como o nome diz, sua essência está na **previsão** de cenários futuros com base na análise de padrões revelados pela base de dados.

É importante saber que, em uma análise preditiva, não é possível prever o que vai acontecer, mas sim, o que deve acontecer SE determinadas condições se cumprirem.

Quer ver um exemplo de análise de dados preditiva?

Suponhamos que sua empresa esteja apreensiva quanto à possível entrada de um concorrente no mercado.

A análise preditiva não será capaz de te dizer se o concorrente iniciará ou não suas atividades em breve. Em contrapartida, te ajudará a enxergar o que poderá acontecer SE o concorrente, de fato, entrar no mercado, tomando como base situações anteriores com contextos semelhantes.

Podemos dizer, assim, que o objetivo da análise preditiva é determinar uma tendência, correlação, causa ou probabilidade.

3. Análise prescritiva

A análise prescritiva é o próximo passo após os resultados da avaliação preditiva. Isso porque uma prescrição é uma recomendação a algo potencialmente previsto.

Sendo assim, a melhor forma de obter uma análise prescritiva é fazendo projeções (predições) e, então, direcionando esforços para obter o melhor resultado a partir das possibilidades.

Por ser uma análise de dados constantemente mutável (já que está sempre condicionada a previsões e predições), os modelos analíticos prescritivos são comumente apoiados por tecnologias como inteligências artificial, *machine learning* e algoritmos. As ferramentas ajudam a fazer sugestões com base em padrões diferenciados e percepções de objetivos organizacionais, limitações e fatores de influência.

4. Análise diagnóstica

Aqui está outro tipo de análise de dados concentrada em algo que já aconteceu (assim como a análise descritiva). A análise diagnóstica, diferentemente da descritiva, tem, como objetivo, encontrar relações de causa e efeito para destrinchar um acontecimento.

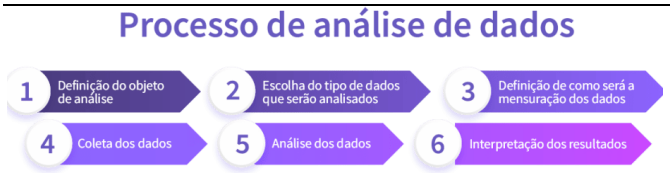
É claro que estabelecer este tipo de relação baseado em um acontecimento passado não é tarefa fácil. Por isso mesmo, o processo é baseado em **probabilidades**.

Conhecer os principais tipos de análise de dados pode ajudar a sua empresa a dominar as informações-chave do negócio na palma da mão. Lembre-se de que, com a ajuda das melhores ferramentas, é possível automatizar momentos importantes da análise de dados (como a consolidação de relatórios e a criação de gráficos), mantendo a equipe focada naquilo que realmente importa: a estratégia.

Qual o processo de análise de dados?

O processo de análise de dados é composto por algumas etapas essenciais:

1. Definição do objeto de análise;
2. Escolha o tipo de dados que serão analisados;
3. Defina como será a mensuração dos dados;
4. Realize a coleta de dados;
5. Análise dos dados;
6. Faça a interpretação dos resultados.



Confira abaixo mais detalhes sobre cada uma delas:

Etapa 1 – Definição do objeto de análise

Nesse primeiro momento, você precisa definir qual será seu objeto de análise. Estabeleça quais perguntas deverão ser respondidas.

Para que tenham o efeito desejado, os objetivos devem ser claros, mensuráveis e relevantes para o negócio.

Assim, as decisões serão tomadas com maior embasamento, você evitará desperdícios e as ações serão mais eficazes.

Chegar a metas que seguem esse padrão depende de um passo muito importante: contar com sua equipe para mapear os desafios enfrentados e definir os objetivos em conjunto.

Etapa 2 – Escolha os tipos de dados que serão analisados

Defina qual será a natureza dos dados que você pretende analisar para responder às perguntas da etapa anterior.

É fundamental entender de forma clara, os fatores que contribuem para o crescimento da sua empresa.

Sendo assim, definir métricas claras e possíveis vai garantir análises mais aprofundadas, e que poderão trazer insumos valiosos para a sua empresa.

Etapa 3 – Defina como será a mensuração dos dados

Na terceira etapa, é necessário estabelecer como você vai medir os dados a serem analisados.

Depois de definir os objetivos e as métricas que precisam ser mensuradas é hora de estabelecer a forma de colocar as análises em prática.

Aqui, é importante contar com uma ferramenta que dê o suporte necessário em todas as etapas, garantindo a confiabilidade dos dados e ajudando a evitar erros comuns.

Além disso, defina os responsáveis e as etapas que devem ser seguidas para que nenhum indicador fique de fora.

Etapa 4 – Realize a coleta de dados

Esse é o momento de coletar os dados brutos que você deseja analisar.

A etapa da coleta de dados é a responsável por recolher todos os dados gerados nas fases de entrevistas.

Além de possibilitar o entendimento sobre os resultados da empresa, esse processo permite mapear futuras ações par melhorar os resultados do negócio, já que aqui você terá uma visão geral da empresa.

Lembre-se que a etapa de coleta é cíclica, pois deve ser feita de forma recorrente, fazendo parte da rotina da equipe. Afinal de contas, o comportamento do consumidor muda o tempo todo, o que torna necessária a atualização dos dados de forma constante.

Etapa 5 – Analise os dados

Você deverá transformar os dados em informações relevantes. Recorra a gráficos e outros recursos visuais que facilitem essa análise.

Na etapa de análise de dados é o momento de, finalmente, colocar em prática tudo o que você viu até aqui.

Se possível, conte com ferramentas para garantir análises robustas e ágeis, tornando o trabalho de toda a equipe mais produtivo.

Etapa 6 – Faça a interpretação dos resultados

Nessa etapa, os questionamentos iniciais devem ser respondidos e as conclusões precisam ser tomadas.

FERRAMENTAS PARA ANÁLISE DE DADOS

Como você viu, as ferramentas são recursos fundamentais na hora de fazer a análise de dados. A boa notícia é que existem diversas opções para não ter desculpa na hora de realizar esse processo tão importante.

Excel

Criado em 1987, o **Excel** faz parte da rotina de diversos profissionais. Para a análise de dados, esse recurso é valioso, pois permite o armazenamento de dados, a construção de planilhas robustas, gráficos e cálculos para garantir a confiabilidade das análises.

Microsoft Power Bi

Mas se o que você precisa é de dashboards e de relatórios dinâmicos, o **Power Bi** é a ferramenta mais indicada. A partir dela, é possível automatizar questões de rotina da empresa, como análises menos complexas e ajustes de dados.

Google Data Studio

Já o Google Data Studio é responsável por transformar dados em informações estratégicas, atuando na criação de painéis, relatórios e dashboards personalizados.

Google Analytics

Uma das ferramentas mais utilizadas pelos profissionais de Marketing, o **Google Analytics** é um recurso gratuito de análises web capaz de fornecer dados valiosos sobre campanhas e páginas do site.

Utilizando um código JavaScript, o recurso consegue coletar dados da web, registrando informações de acessos do usuário, como o número de sessões e a taxa de rejeição.

ANÁLISE DE DADOS COM PYTHON: O QUE É?

Elaborada para ser simples e acessível, Python é a linguagem de programação responsável pela aplicação em Machine Learning, Inteligência Artificial, e na análise de dados.

Além de permitir o trabalho com séries temporais, o mecanismo possibilita a manipulação de tabelas e dados, além da construção de diversos tipos de gráficos.

- Confira alguns dos benefícios dessa linguagem:
- É de fácil aprendizagem;
- É multiplataforma;
- É gratuita;
- Oferece diversas possibilidades de desenvolvimento.

Fonte: https://blog.ploomes.com/analise-de-dados/

ANÁLISE DE AGUPAMENTOS (CLUSTER)

O termo Análise de Agrupamentos, primeiramente usado por (Tyron, 1939) na realidade comporta uma variedade de algoritmos de classificação diferentes, todos voltados para uma questão importante em várias áreas da pesquisa***: Como organizar dados observados em estruturas que façam sentido, ou como desenvolver taxonomias capazes de classificar dados observados em diferentes classes.*** Importante é considerar inclusive, que essas classes devem ser classes que ocorrem "naturalmente" no conjunto de dados.

A Análise de Cluster é um método ao qual permite agrupar sujeitos ou variáveis em grupos com uma ou mais características comuns, não sendo necessário ter informações já dadas sobre a composição desses grupos. Muitas vezes temos conjunto de dados e uma necessidade de agrupar esse conjunto de dados por algum critério de similaridade em vista a algum tipo de conhecimento que a gente deseje aplicar.

Um exemplo é queremos agrupar fotos de acordo com alguma similaridade, seja com fotos da mesma pessoa, ou fotos de pessoas de um grupo étnico, ou fotos de pessoas de um gênero. As possibilidades são variadas e as funções que elas podem exercer também.

Para realizar esta análise é necessário medir a semelhança, dissemelhança dos sujeitos e variáveis, a partir daí, agrupar. Os conglomerados obtidos a partir disso devem apresentar tanto uma homogeneidade interna (dentro de cada conglomerado), como uma grande heterogeneidade externa (entre conglomerados).

A Análise de Cluster acaba por incluir vários procedimentos estatísticos que podem ser utilizados para classificar objetos sem preconceitos, ou seja, somente com base nas semelhanças ou não que eles possuem entre si. Isso sem definir previamente critérios de inclusão em qualquer agrupamento.

Assim, ela traz como possibilidade de uso a identificação de uma estrutura presente nos dados, além de impor uma estrutura num conjunto de dados mais ou menos homogêneos que têm de ser separados.

ENTENDENDO O CONCEITO DE AGRUPAMENTO HIERÁRQUICO

No processo de análise de cluster um dos conhecimentos necessários para termos uma ideia mais clara sobre o tema é o conhecimento de agrupamento hierárquico. É nele que é criada uma estrutura em formato de árvore que vai indicar o número de clusters.

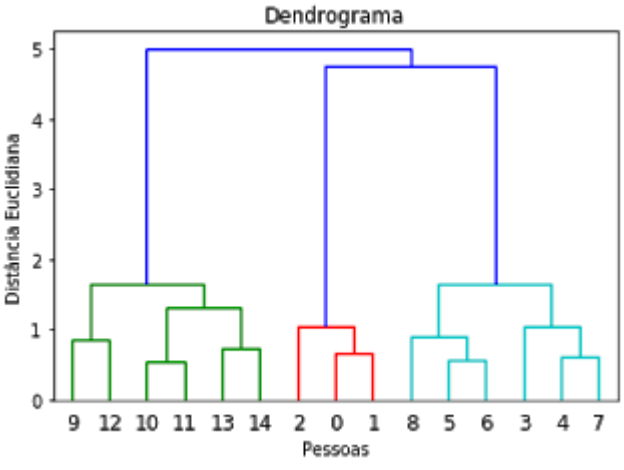
Vamos imaginar então essa árvore hierárquica. Inicia-se esse processo com cada objeto em uma classe por si só e aos poucos os diferentes objetos ou variáveis vão se agrupando, criando nós que são ou dispõem de características mais semelhantes entre si.

Podemos perceber que enquanto os elementos se agrupam eles se tornam aglomerados cada vez maiores e, sobretudo, passam a conter elementos cada vez mais diferenciados, até que, no último momento os objetos estão unidos em conjunto.

Quando esses dados finais contém uma estrutura, digamos assim, clara em termos de grupos de objetos que são similares uns aos outros, então essa estrutura se reflete na árvore hierárquica como ramos distintos.

O Dendrograma

Para visualizar como os agrupamentos são formados em cada passo e para avaliar os níveis de similaridade (ou distância) dos agrupamentos que são formados é utilizado o dendrograma, diagrama que lembra a estrutura de árvore (daí o nome) que serve para exibir os níveis de similaridade (ou distância).



.

QUAIS AS FUNÇÕES DA ANÁLISE DE AGRUPAMENTOS?

A análise de agrupamentos pode ser realizada para uma série de uso, tais como:

- Classificar pessoas de acordo com a personalidade de cada uma delas.
- Segmentar o cliente de acordo com seus hábitos de consumo, criando assim

estratégias comerciais para aumentar o lucro.
- Classificar cidades de acordo com seus aspectos físicos, demográficos, econômicos e humanos para assim criar um mapa mais abrangente daquele estado.
- Identificar grupos de investimento de acordo com perfis de risco.
- Identificar grupos de alunos mais propensos à evasão escolar.
- Segmentar empresas com base em **indicadores financeiros (rentabilidade, liquidez, margem).**

Essas são somente seis de centenas de possibilidades onde a análise de agrupamentos vai se mostrar importante.

MÉTODOS HIERÁRQUICOS

Os métodos de agrupamento de dados podem ser divididos em duas categorias cada uma delas agregando diferentes tipos de algoritmos.

Métodos hierárquicos (Algoritmos aglomerativos ou divisivos).

Métodos particionais (Algoritmos exclusivos ou não exclusivos).

Os métodos hierárquicos são técnicas simples de análise, onde os dados são particionados de forma sucessiva, produzindo uma representação hierárquica dos agrupamentos.

Algoritmos aglomerativos

O método hierárquico aglomerativo visa formar os clusters com a mínima distância interna possível, iniciando com cada padrão formando seu próprio agrupamento e de forma gradual os grupos são unidos até que um único agrupamento contendo todos os dados gerados.

São desvantagens desse método:

- Os agrupamentos não podem ser corrigidos, ou seja, os padrões de um determinado agrupamento até o final da execução do algoritmo.

- Requerem espaço de memória e tempo de processamento.

Algoritmos divisivos

Estes são menos comuns entre os métodos hierárquicos, por conta de sua ineficiência e também por exigir uma capacidade do computador muito maior que os métodos hierárquicos aglomerativos.

Nesse método busca achar a partição que minimize a matriz de similaridades. Explicando melhor, ele começa com um único agrupamento formado por todos os padrões e de modo gradual vai diminuindo os agrupamentos em agrupamentos menores até que seja finalizado com um agrupamento padrão.

MÉTODOS NÃO HIERÁRQUICOS

Enquanto que no método hierárquico o algoritmo estabelece uma relação de hierarquia entre os sujeitos e os grupos, no método não hierárquico isso não acontece.

Os procedimentos não hierárquicos são utilizados basicamente para agrupar indivíduos - e não variáveis - cujo número inicial de clusters é definido pelo pesquisador.

A probabilidade de acontecerem classificações erradas nos agrupamentos é menor nos métodos não hierárquicos, mas em contrapartida, há uma dificuldade maior em estabelecer o número de clusters de partida. Uma alternativa consiste em utilizar o método hierárquico como técnica exploratória e após utilizar o número de clusters no método não hierárquico.

Outros exemplos de aplicação da Análise de Cluster:

Marketing - No marketing, a Análise de Cluster pode ser aplicada para proceder à segmentação de mercados a partir das características geográficas e demográficas, e até mesmo com base em perfis psicológicos dos consumidores, para assim identificar mercados potenciais para determinados produtos, determinar mercados idênticos em países diferentes ou encontrar grupos de consumidores que possam servir de referência na previsão de vendas.

Na medicina - Uma das áreas que mais tem benefícios da aplicação da Análise de Cluster é a Medicina, bem como na Psicologia, na Psiquiatria. Nessas áreas, a classificação obtida de uma análise de clusters pode permitir identificar as causas das doenças, os sintomas, e consequentemente criar/ melhorar os seus tratamentos.

Nas Ciências Sociais - Nas Ciências Sociais, os métodos de análise de clusters podem ser usados pelos antropólogos para definirem áreas culturais homogêneas para assim pensarem em políticas específicas para tais segmentos.

Fonte: https://www.trecsson.com.br/blog/economia-e-financas/analise-de-cluster

CONCEITOS DE ANALYTICS.

Analytics é o uso aplicado de dados, análises e raciocínio sistemático para seguir em um processo de tomada de decisão muito mais eficiente. ***Analytics*** podem ser aplicados em diversos negócios e departamentos.

Consideramos, **Analytics**, um ramo de Business Intelligence

Inteligência Analítica

Utilizar a inteligência analítica está diretamente ligado com a possibilidade de melhorar o desempenho com relação aos domínios fundamentais de uma empresa ou negócio utilizando, basicamente, análise de dados.

Analytics – tipos de análises

Há muitos tipos de análises que compõe o termo "**Analytics**":

- Modelagem Estatística;
- Previsão (Forecasting);
- Data Mining;
- Text Mining;
- Otimização;
- Delineamento de Experimentos, etc.

Com todos os avanços na área de tecnologia da informação e também com o aumento da quantidade de dados disponíveis, existem diversas oportunidades para se aplicar analises bem estruturadas.

Crescimento dos dados

O ritmo do crescimento dos dados está acelerando a cada dia. E-mails são armazenados em bancos de dados corporativos, conversas telefônicas são armazenadas e, também, digitalizadas. Muitas empresas estão criando grandes repositórios de dados (banco de dados), procurando manter um forte registro digital de tudo o que está acontecendo como, por exemplo, sistemas financeiros, sistemas de estoque, sistemas de vendas, até mesmo RH.

Tudo o que fazemos no mundo digital deixa um rastro de dados. O nosso próprio navegador de internet registra o que estamos procurando e os site que acessamos. Além disso, estamos cada vez mais gerando dados usando sensores, como quando nossos celulares rastreiam a nossa localização. Tudo isso é armazenado! Há também os dispositivos que medem quantos passos andamos ou até mesmo as calorias queimadas. São inúmeras possibilidades.

Para se ter uma ideia, cerca de mais de 100 horas de vídeos são carregados para o YouTube a cada minuto e algo em torno de mais de 200 mil fotos são adicionadas ao Facebook a cada minuto. A todo momento há uma avalanche de dados acontecendo simultaneamente e todos eles são o principal ingredientes para análises e desenvolvimento de modelos estatísticos.

Analytics – Como as análises são aplicadas

Além da grande quantidade de dados, a nossa capacidade de aplicar e analisar esses dados melhorou consideravelmente nos últimos anos. Hoje em dia é possível analisar grandes volumes de dados de maneira muito veloz, com diferentes fontes. Esse grande conjunto de analises de dados pode em muitos casos ser denominado como “Big Data” ou “Big Data Analytics”.

Confira abaixo alguns exemplos:

- **Esporte:** Nesse caso, o **Analytics** é muito utilizado para melhorar o desempenho dos atletas, possibilitando feedbacks sobre seus treinos e os critérios que estão bons e que precisam ser melhorados, como velocidade, ritmo, força, etc.
- **Saúde:** As analises proporcionaram mudanças incríveis à saúde. Há recursos dedicados por meios de análises realizadas para acompanhar e identificar padrões, prevenindo infecções em maternidades, probabilidades de se contrair doenças ajudando na prevenção e até mesmo personalização de tratamentos como o câncer por meio da decodificação do código do DNA.
- **Prevenção do Crime:** Nos dias atuais, combater o crime depende muito de análises, que possibilitam identificar e prever a atividade criminal. Esses casos são muito comuns em empresas de cartão de créditos que monitoram as transações de seus clientes em tempo real, possibilitando a identificação de fraudes.

O que não faltam são áreas que recorrem a essas analises para aperfeiçoar a vida das pessoas, como muitas empresas que procuraram direcionar seus esforços de marketing por meio de análise de dados de compra. Os varejistas podem usar essas analises para aperfeiçoar suas decisões. Tradicionalmente, as lojas analisam os itens que mais vendem para armazenar uma quantidade maior.

Isso evita que elas façam investimentos ruins e que foquem em produtos que tenham muito mais saída e que atendam o perfil de seu público alvo.

O que não faltam são exemplos de aplicações de **Analytics**. Esperamos que esse artigo tenha conseguido atingir o objetivo de proporcionar uma visão geral das possibilidades de sua aplicação, bem como o Big Data e a Modelagem Estatística.

Google Analytics

Ao pesquisar por Analytics, talvez o que inicialmente seja encontrado são páginas sobre Google Analytics (https://analytics.google.com/analytics/), o GA como também é conhecido, pode ser considerado um tipo de Analytics, voltado para análises de websites, visitantes, tempo, etc.

O GA é uma ferramenta gratuita que pode ser colocada em qualquer site, ele irá fornecer painéis completos para acompanhamento do seu site.

_____

Fonte: https://cetax.com.br/o-que-e-analytics/

APRENDIZADO DE MÁQUINA.

O aprendizado de máquina (em inglês, machine learning) é um método de análise de dados que automatiza a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana.

Evolução do machine learning

Graças às novas tecnologias computacionais, o machine learning de hoje não é como o machine learning do passado. Ele nasceu do reconhecimento de padrões e da teoria de que computadores podem aprender sem serem programados para realizar tarefas específicas; pesquisadores interessados em inteligência artificial queriam saber se as máquinas poderiam aprender com dados. O aspecto iterativo do aprendizado de máquina é importante porque, quando os modelos são expostos a novos dados, eles são capazes de se adaptar independentemente.

Eles aprendem com computações anteriores para produzir decisões e resultados confiáveis, passíveis de repetição. Isso não é uma ciência nova – mas uma ciência que está ganhando um novo impulso.

Embora diversos algoritmos de machine learning existam há muito tempo, a capacidade de aplicar cálculos matemáticos complexos ao big data automaticamente – de novo e de novo, mais rápido e mais rápido – é um desenvolvimento recente. Eis alguns exemplos bem conhecidos de aplicações de machine learning, dos quais você já deve ter ouvido falar:

- Os carros autônomos super esperados do Google? A essência do machine learning;
- Ofertas recomendadas como as da Amazon e da Netflix? Aplicações de machine learning para o dia-a-dia;
- Saber o que seus clientes estão falando de você no Twitter? Machine learning combinado com criação de regras linguísticas;
- Detecção de fraudes? Um dos usos mais óbvios e importantes de machine learning no mundo de hoje.

Machine learning e inteligência artificial

Enquanto a inteligência artificial (IA) pode ser definida, de modo amplo, como a ciência capaz de mimetizar as habilidades humanas, o machine learning é uma vertente específica da IA que treina máquinas para aprender com dados. Assista a este vídeo para entender melhor a relação entre a inteligência artificial e o aprendizado de máquina. Você verá como essas duas tecnologias funcionam, com exemplos úteis e alguns apartes divertidos.

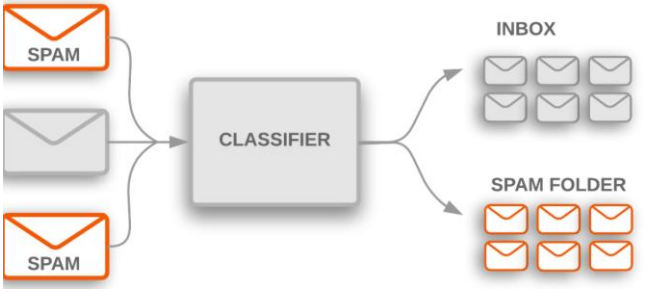
TIPOS DE SISTEMAS DE APRENDIZAGEM DE MÁQUINA

Cada um dos itens abaixo vai explicar detalhadamente as principais formas que uma máquina pode aprender, elas irão compor o ecossistema do aprendizado de máquina, de forma que seja possível resolver diferentes problemas baseado nas abordagens que mais se encaixem a elas.

1. Aprendizado Supervisionado

O aprendizado supervisionado é um paradigma de aprendizado de máquina, que tem como objetivo adquirir informações de relacionamento entre entrada e saída de um sistema, baseado em um conjunto de amostras de treinamento.

Um algoritmo de aprendizado supervisionado analisa os dados de treino e produz uma função inferida que será utilizada para mapear novos exemplos. Para deixar menos abstrato, vamos considerar um exemplo, a classificação de e-mails como spam.



Provavelmente, você utiliza e-mail e sabe que conteúdos maliciosos são quase sempre enviados para uma pasta específica, com o objetivo de te proteger. Mas como isso acontece?

A experiência que permite você não precisar classificar quais e-mails são maliciosos ou não, é proporcionada por um modelo de classificação que é baseado em entradas rotuladas. Nessas entradas, possuímos e-mails classificados como confiáveis ou não, dessa forma, o modelo irá aprender a reconhecer a classe que um novo dado pertence baseado no que já aprendeu sobre esses dados rotulados.

Porém, pode te bater aquela curiosidade, por que então o meu provedor de e-mail ainda me pergunta se a mensagem que eu recebi é ou não um spam? Apesar de já existirem modelos confiáveis treinados em cima de enormes conjuntos de dados, a sua validação permite que esse modelo seja aprimorado cada vez mais, permitindo que essa não seja uma preocupação sua.

O exemplo acima demonstra como seria um problema de **Classificação**, mas vale lembrar que existe outra gama de problemas que podem ser denominados como problemas de **Regressão**. Para que não sobrem dúvidas do que se trata cada um desses tipos de problemas, eles estão definidos abaixo:

Classificação

A classificação é o processo de categorizar um determinado conjunto de dados em classes. No exemplo da classificação de e-mails como spam, teríamos um exemplo de classificação binária, no qual o modelo através dos dados fornecidos, precisaria gerar como resposta se o e-mail é spam ou não.

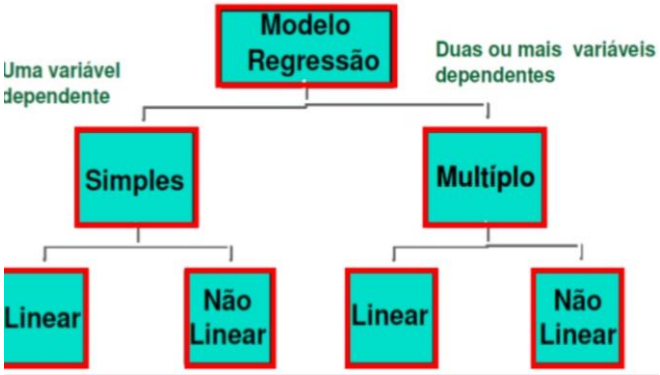
Alguns dos **algoritmos mais famosos** são:

- KNN
- Naive Bayes
- Logistic Regression
- Support Vector Machines
- Decision Trees

Regressão

Os modelos de regressão são utilizados quando queremos prever valores, por exemplo, prever o preço de uma casa ou o número de produtos que serão vendidos em determinado mês.

Os modelos de regressão são dos mais diversos e suas possibilidades são descritas pela imagem abaixo:



Analisando a imagem acima podemos perceber que a primeira subdivisão dos modelos de regressão diz respeito ao número de variáveis envolvidas, modelos de regressão simples envolvem apenas uma variável e os múltiplos duas ou mais. Em seguida, para cada um dos tipos descritos ainda existe outra ramificação que divide esses modelos em lineares ou não lineares.

Alguns modelos são famosos para realizar regressão, são eles:

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Principal Components Regression (PCR)

2. Aprendizado Não Supervisionado

O aprendizado não supervisionado consiste em treinar uma máquina a partir de dados que não estão rotulados e/ou classificados. Os algoritmos que fazem isso buscam descobrir padrões ocultos que agrupam as informações de acordo com semelhanças ou diferenças, por exemplo.

Para que isso fique mais claro, vamos imaginar um algoritmo de aprendizado não supervisionado, que receba uma imagem contendo cachorros e gatos.

Ao receber essa imagem nada se sabe sobre as características que cada animal possui, ou seja, não é possível categorizá-los. Porém, esse algoritmo será responsável por descobrir semelhanças, padrões e/ou diferenças que permitam diferenciar cães e gatos.

No exemplo citado anteriormente utilizamos uma técnica chamada de agrupamento (**Clustering)**, porém existem outras técnicas como regras de associação (**Association Rules**) e redução de dimensionalidade (**Dimensionality Reduction**). Falaremos um pouco de cada uma delas abaixo.

Agrupamento

A técnica de agrupamento como explicado no exemplo anterior, consiste em agrupar dados não rotulados com base em suas semelhanças ou diferenças. Esses algoritmos de agrupamento ainda podem ser subdivididos em **agrupamentos exclusivos, sobrepostos, hierárquicos** e **probabilísticos.**

Regras de Associação

Ao usar as regras de associação, buscamos descobrir relações que descrevem grandes porções dos dados. A associação é muito utilizada em análises de cestas de compras, no qual a empresa pode tentar entender relações de preferências de compras entre os produtos.

Quando falamos de algoritmos para gerar regras de associação os principais são: **Apriori**, **Eclat** e **FP-Growth**.

Redução de dimensionalidade

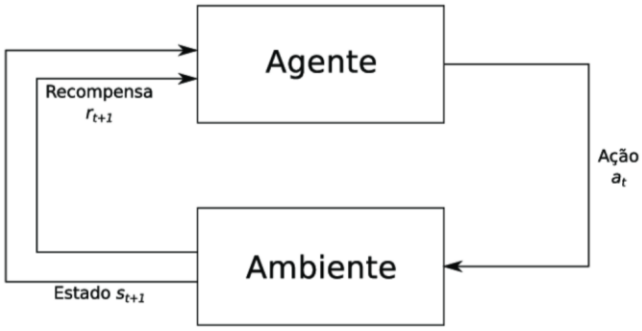
Existem casos nos quais ao estudar um conjunto de dados, podemos encontrar nele um grande número de recursos (dimensões). Por mais que existam situações onde isso é positivo, o excesso pode impactar o desempenho dos algoritmos causando, por exemplo, o **overfitting**.

Utilizando a técnica de redução de dimensionalidade, será feita uma redução no número de recursos, de forma que torne-os gerenciáveis por parte do modelo, além de preservar a integridade dos dados.

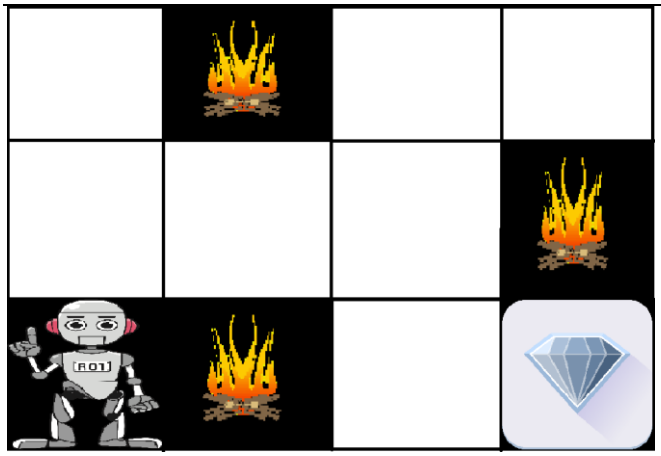
E para executar essa tarefa existem algumas técnicas que podem ser utilizadas, como: **Missing Values Ratio**, **Low Variance Filter**, **High Correlation Filter**, **Random Forests / Ensemble Trees**, **Principal Component Analysis (PCA)**, **Backward Feature Elimination** e **Forward Feature Construction**.

3. Aprendizado por reforço

Para entendermos melhor como funciona o aprendizado por reforço usaremos a seguinte imagem para ilustrar qual é o princípio.



O primeiro passo é definir os elementos presentes na imagem, o **agente** (**Agent**) é aquele que toma as decisões com base nas recompensas e punições, esse agente pode realizar uma **ação** (**Action**) que irá variar de acordo com o contexto. O **ambiente (Environment)** é o mundo físico ou virtual em que o agente opera, a **recompensa** (**reward**) é o feedback do ambiente baseado na ação tomada e o **estado** (**state**) é a situação atual do agente.



A imagem acima demonstra um exemplo de como o aprendizado por reforço pode ser utilizado. Nesse caso, o robô é o nosso **agente** e ele está situado no **estado** inicial do nosso **ambiente,** que é representado pelo "labirinto" que o robô terá de percorrer. Desta forma, o objetivo é chegar ao diamante evitando os obstáculos (fogueiras).

Definido o objetivo, o robô deve buscar pelo melhor caminho possível para chegar até o diamante. Dessa forma, a cada **ação** do robô, ele poderá caminhar em uma determinada direção, caso ele escolha corretamente, ele irá inserir pesos diferentes, para diferentes respostas. Com isso, espera-se que ao final o robô consiga realizar seu objetivo de forma que obtenha a **maior recompensa cumulativa**.

APLICAÇÕES

Lembre-se que as diferentes abordagens do aprendizado de máquina contribuem na realização de tarefas árduas e melhoria de processos. Assim, iremos conhecer em quais cenários o machine learning está inserido.

Diagnósticos médicos

Na área médica as técnicas de machine learning são utilizadas para fazer o reconhecimento de doenças. Com o crescimento da tecnologia tem sido possível construir modelos 3D que podem prever a posição exata de lesões no cérebro, permitindo a detecção de tumores e outros diagnósticos relacionados muito mais fácil.

Além disso, muito trabalho vem sendo feito com imagens como, por exemplo, o reconhecimento de padrões que identificam câncer de pulmão, de pele, dentre outros.

Detecção de fraudes online

Se considerarmos uma instituição financeira que lida com milhares de transações por dia, ela está sujeita a fraudes a todo momento e sabendo que avaliar toda essa quantidade de operações seria totalmente exaustivo e ineficiente, modelos de machine learning são criados para que possam ser detectadas anomalias nas transações.

Para ficar mais claro vamos supor que uma pessoa tenha um cartão de crédito de um banco com limite de 2000 reais, porém, ela tem um histórico de uso mensal de no máximo 800 reais, se por acaso em um determinado dia houver uma compra no seu cartão no valor de 2000 reais, o modelo de detecção de fraudes irá perceber que essa compra não se encaixa no seu padrão e, com isso, o banco será notificado colocando a transação em espera.

Sistemas de recomendação

Presente nos mais diversos tipos de aplicações, os sistemas de recomendação tiram aquela velha necessidade de procurar tudo aquilo que desejamos. No sistema de varejo, por exemplo, se você tiver cadastro na plataforma de algum desses varejistas você terá um sistema de recomendações de produto ao seu dispor, ele cria essas recomendações baseado em compras anteriores, históricos de navegação, dentre outras informações complementares.

Dessa forma, quando você está com o carrinho de compras e percebe que esqueceu mais um item da compra que estava planejando, provavelmente ele estará em uma seção destinada a seus possíveis interesses.

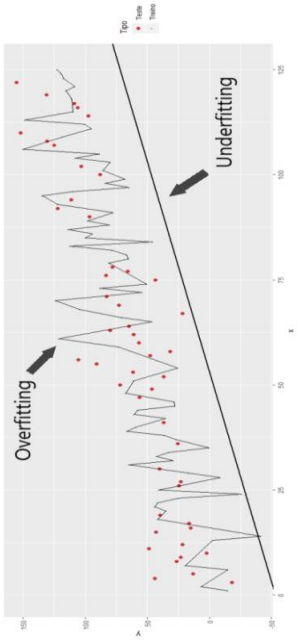
Reconhecimento de fala

Provavelmente o exemplo mais famoso para o reconhecimento de fala são os assistentes de voz. Então, a Siri da Apple, Alexa da Amazon, Cortana da Microsoft, dentre outros assistentes de voz usam machine learning através de técnicas de processamento de linguagem natural (NLP) para reconhecerem a fala, posteriormente transformam essa fala em números para que possam formular uma resposta de acordo.

UNDERFITTING, OVERFITTING

Underfitting e **Overfitting** são dois termos extremamente importantes no ramo do machine learning.

Um bom modelo não pode sofrer de Underfitting nem de Overfitting, por isso precisamos entender estes conceitos, e saber identificar suas ocorrências.



Overfitting

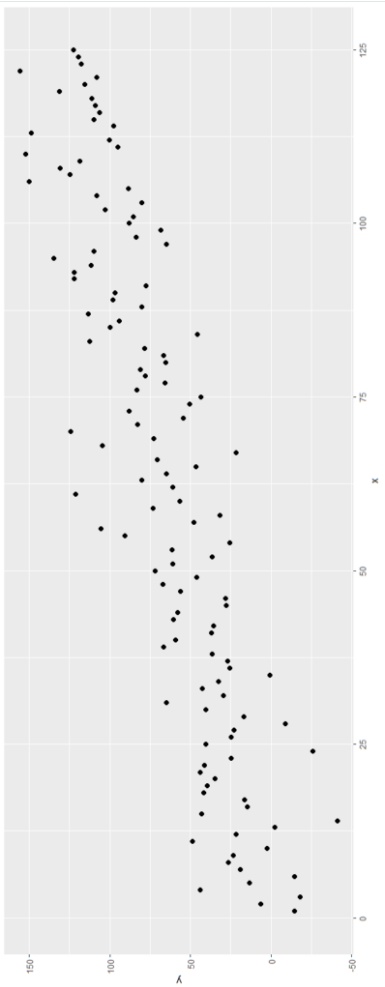
Um cenário de overfitting ocorre quando, nos dados de **treino,** o modelo tem um **desempenho excelente**, porém quando utilizamos os dados de **teste** o resultado é **ruim**.

Podemos entender que, neste caso, o modelo aprendeu tão bem as relações existentes no treino, que acabou apenas **decorando o que deveria ser feito**, e ao receber as informações das variáveis preditoras nos dados de teste, o modelo tenta aplicar as mesmas regras decoradas, porém com dados diferentes esta regra não tem validade, e o desempenho é afetado. É comum ouvirmos que neste cenário o modelo treinado não tem capacidade de *generalização*.

Underfitting

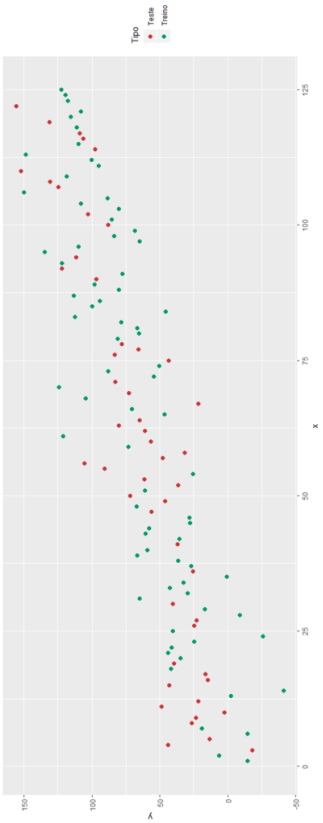
Neste cenário o **desempenho do modelo já é ruim no próprio treinamento**. O modelo não consegue encontrar relações entre as variáveis e o teste nem precisa acontecer. Este modelo já pode ser descartado, pois não terá utilidade.

Vamos entender estes conceitos com exemplos práticos e visuais. Na figura abaixo vemos uma distribuição de dados entre as variáveis x e y. Consideremos que estes são os dados que temos a nossa disposição para treinar e testar o modelo.



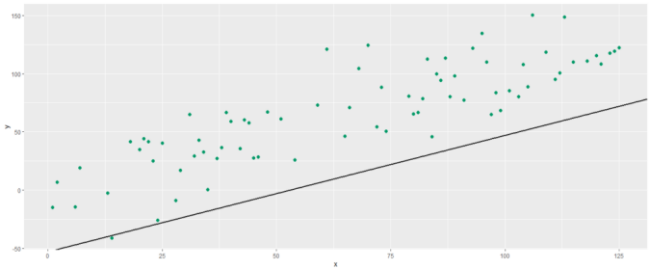
Nosso objetivo é traçar uma linha por estes pontos, de maneira que se recebermos um novo valor de x, por exemplo, possamos prever o valor de y. Poderíamos utilizar alguns algoritmos de manchine learning para traçar esta reta ou curva, utilizando como base todos estes pontos, porém sabemos que precisamos de dados de teste para medir o desempenho dos modelos criados.

Com esta finalidade, separamos os dados em treino e teste, conforme vemos na imagem abaixo:



Podemos então utilizar os dados de treino, que estão em verde na imagem, para treinar nosso modelo, e obtermos algumas linhas que expliquem o relacionamento destes dados.

Na imagem abaixo vemos uma primeira tentativa, onde o modelo nos retorna uma reta distante dos dados de treino:

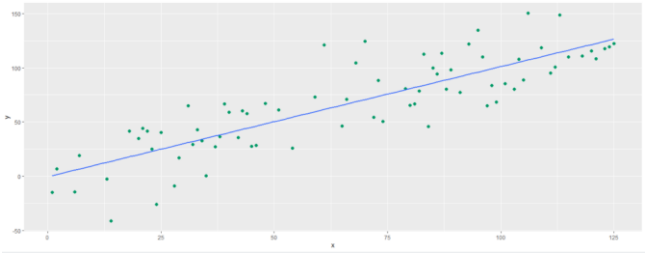


Underfitting

Este é um típico caso de **Underfitting**, onde o **algoritmo não encontra uma boa relação entre os dados**, e com isso o resultado apresentado nos dados de treino é ruim. Esta reta não precisa nem mesmo ser aplicada aos dados de teste, devido a seu fraco desempenho.

Na próxima tentativa temos como resultado a reta abaixo, que está **bem ajustada aos dados**, acertando com precisão alguns poucos pontos, porém **obtendo constantemente um baixo erro na previsão**.

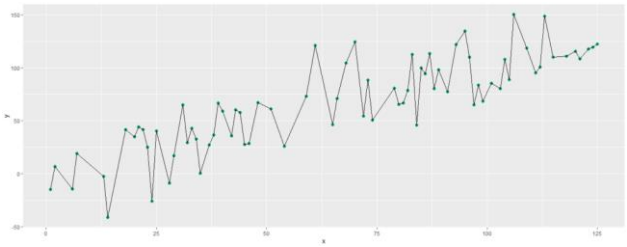
A reta em questão foi gerada através de uma regressão linear, e, portanto, é a reta que melhor se adequa a estes dados:



Reta de Regressão Linear

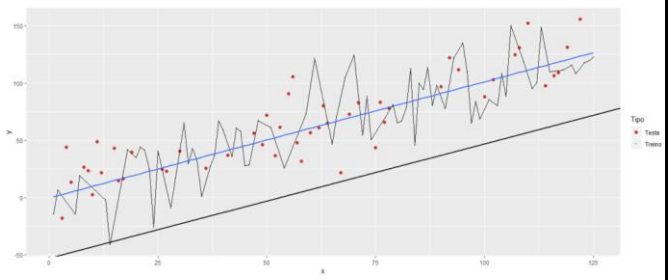
Realizamos ainda mais uma tentativa, tendo como resultado a imagem abaixo. Agora nossas **previsões estão perfeitas**, sendo que a linha traçada passa por todos os pontos do gráfico, **não havendo assim erro algum**.

Porém não podemos esquecer que ainda estamos apenas treinando nosso modelo, e ele pode estar sofrendo de **Overfitting**. Precisamos testá-lo em dados desconhecidos, que são os dados de teste que foram separados inicialmente.



Overfitting

Abaixo vemos a aplicação dos três modelos nos **dados de teste**:



Com este gráfico fica evidente que **a reta inicial não pode ser utilizada**, conforme já havíamos concluído. Porém agora percebemos que **as previsões perfeitas que tínhamos antes, não existem mais**.

Nosso último modelo estava sim sofrendo de Overfitting, pois quando ele é aplicado aos dados de treino, **seu desempenho é drasticamente afetado**, passando a ter muitos erros, e ainda erros grosseiros.

É provável que calculando o coeficiente de determinação R2 dos dois modelos finais, encontremos um **melhor resultado nos dados de teste do modelo intermediário**, pois ele continua posicionando sua reta em uma posição relativa aos dados de teste muito parecida com a que tinha nos dados de treino, mantendo seus erros em valores baixos, diferentemente do que acontece com o modelo que sofre de Overfitting.

Fonte: https://didatica.tech/underfitting-e-overfitting/

DADOS DE TREINO E TESTE

Uma das bases do machine learning está nos dados históricos.

Os algoritmos de machine learning precisam aprender, e para isso quanto mais dados forem usados, melhor ficará o modelo.

Big Data

É neste momento que o famoso Big Data dá uma de suas importantes contribuições para o aprendizado de máquina. Uma de suas definições traz a ideia de “informação em alto volume”, ou seja, dados históricos em grande quantidade.

Pré-processamento e aplicação dos dados

Estes dados serão devidamente preparados, passando por alguns processos de limpeza e ajustes, que são o pré-processamento e a seleção de variáveis, para então estarem aptos a serem apresentados a um algoritmo de machine learning, que realizará as previsões, verificando o quão distante o resultado está do valor correto, reajustando os parâmetros utilizados na previsão a fim de obter um valor mais adequado.

Esse processo se repetirá até que o erro entre os valores reais e os valores previstos pare de diminuir a cada novo ajuste.

Separação dos dados históricos

Poderíamos utilizar a totalidade dos dados históricos no processo acima, criando assim um modelo de machine learning pronto para receber novos dados e realizar suas previsões, porém desta forma não saberíamos o real desempenho deste modelo.

O algoritmo poderia aprender perfeitamente a relação existentes nos dados apresentados e com isso criar um modelo que sofre de Overfitting e só descobriríamos esse problema após as previsões desastrosas geradas por este modelo.

Desta forma, para medir o desempenho real do modelo criado, é necessário que realizemos testes com ele, utilizando dados diferentes dos que foram apresentados em sua criação.

Com esta finalidade, após a realização do pré-processamento, iremos **separar a totalidade dos dados históricos existentes em dois grupos**, sendo o primeiro responsável pelo **aprendizado** do modelo, e o segundo por realizar os **testes**.

O que são Dados de Treino

Conforme podemos imaginar, dados de treino são os dados que serão apresentados ao algoritmo de machine learning para criação do modelo. Estes dados costumam representar cerca de 70% da totalidade dos dados.

O que são Dados de Teste

São os dados que serão apresentados ao modelo após a sua criação, simulando previsões reais que o modelo realizará, permitindo assim que o desempenho real seja verificado. Estes dados costumam representar cerca de 30% da totalidade dos dados.

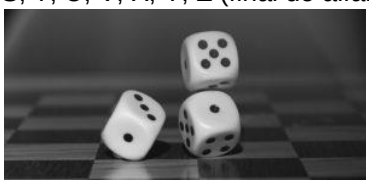
Aleatoriedade na separação dos dados

É importante observar que a separação dos dados em treino e teste é uma etapa essencial, e que caso ela seja realizada de maneira errada, poderá resultar em problemas no modelo.

Imagine que você tenha dados sobre vinte mil carros, com algumas características e o valor deles. Você decide que irá utilizar quinze mil amostras para treinar seu modelo, e para isso seleciona as quinze mil primeiras linhas, deixando a últimas cinco mil para teste.

Porém estas linhas estavam organizadas em ordem alfabética com base nos nomes dos carros. Percebem o problema? Alguns modelos de carros específicos estarão presentes apenas nos dados de treino, e outros apenas nos dados de teste.

Esta situação levará a um modelo deficiente, que não aprendeu com todos os tipos de dados que poderia e que também não será testado de maneira correta. Afinal, nos dados de teste haverá apenas modelos de carros que começam, digamos, com as letras S, T, U, V, X, Y, Z (final do alfabeto).



A solução para este problema está na aleatoriedade. Selecionando os dados de maneira aleatória não haverá padrão algum no momento da divisão dos dados, e cada observação terá a mesma probabilidade de ser selecionada.

Tanto a linguagem de programação **R** quanto o Python já possuem funções prontas para esta finalidade, tornando esse processo extremamente simples e confiável.

_____

Fonte: https://didatica.tech/dados-de-treino-e-teste/

INTELIGÊNCIA ARTIFICIAL.

INTRODUÇÃO

Com certeza você já se deparou com algum tipo de inteligência artificial (IA). Cada vez mais presente no dia a dia, ela engloba algoritmos que ajudam a melhorar estratégias diversas.

Quer um exemplo? É por meio da inteligência artificial que os anúncios pagos em redes sociais e sites de buscas conseguem segmentar seu público-alvo.

Logo, essa ferramenta é muito útil para sistemas de gestão de toda empresa. Há coleta de documentos e informações, análise de dados, relatórios prontos em pouco tempo, **avaliação de fornecedores** feita de maneira rápida e eficiente.

Ou seja, a inteligência artificial é capaz de imitar comportamentos e potencializar características humanas, como tomada de decisões, aprendizado, resolução de problemas práticos, entre outros.

CONCEITO

Inteligência artificial é a capacidade de dispositivos eletrônicos de **funcionar de maneira que lembra o pensamento humano**.

> **DICA:** *Inteligência artificial é a capacidade de os computadores realizarem tarefas que os seres humanos inerentemente realizam melhor, até o momento*. Ruy Flávio de Oliveira

Isso implica em perceber variáveis, tomar decisões e **resolver problemas**. Enfim, operar em uma lógica que remete ao raciocínio.

"Artificial", segundo o dicionário Michaelis, é algo que foi "produzido por arte ou indústria do homem e não por causas naturais".

Já inteligência é a "faculdade de **entender, pensar, raciocinar e interpretar**".

Ou o "conjunto de funções mentais que facilitam o entendimento das coisas e dos fatos".

No mesmo dicionário, há duas definições da Psicologia para a palavra "inteligência":

- **Habilidade** de aproveitar a eficácia de uma situação e utilizá-la na prática de outra atividade
- Capacidade de resolver situações novas com rapidez e êxito, adaptando-se a elas por meio do conhecimento adquirido.

Mesmo essas duas últimas definições fazem sentido quando falamos em inteligência artificial, com a vertente chamada de **machine learning** (aprendizado de máquina).

Enfim, a inteligência artificial é desenvolvida para que os dispositivos criados pelo homem possam desempenhar determinadas funções **sem a interferência humana**.

E quais são essas funções? A cada dia que passa, a resposta a essa pergunta é maior.

Tentaremos responder mais adiante, dando exemplos de aplicações da inteligência artificial.

COMO FUNCIONA A INTELIGÊNCIA ARTIFICIAL?

Você já deve ter ouvido falar muitas vezes em hardware e **software**, certo? Mas você sabe o que esses termos significam?

Enquanto o hardware é a parte física de uma máquina, o **software é a parte lógica** – ou o "cérebro". Onde você diria, portanto, que está a inteligência artificial? No software, é claro.

Por isso, se você quiser saber como um **carro pode andar sozinho**, por exemplo, esqueça o hardware, pois o segredo está no programa que orienta seus movimentos.

Portanto, não é possível explicar como funciona a inteligência artificial sem falar na **ciência**

da computação.

Essa ciência estuda técnicas e métodos de processamento de dados, sendo o desenvolvimento de algoritmos uma questão central nela.

Os **algoritmos** são uma sequência de instruções que orientam o funcionamento de um software – que, por sua vez, pode resultar em movimentos de um hardware.

E a **inteligência artificial**, onde entra nisso? Na sua origem, o algoritmo é muito simples, como em uma receita de bolo.

Hoje, a lógica dos algoritmos é usada para criar regras extremamente complexas, para que possam resolver problemas sozinhos.

Mesmo quando há dois ou mais caminhos a seguir em uma tarefa. Para isso, é necessário **combinar algoritmos com dados**.

Voltando ao exemplo do bolo, uma pessoa o retira do forno quando observa que ele está pronto ou fazendo o teste do garfo.

Uma máquina de fazer bolos com inteligência artificial poderia ter algum tipo de sensor que identificasse a textura do bolo.

O algoritmo trabalharia com duas hipóteses e uma resposta para cada uma:

1. Se a textura ainda não for a ideal, o bolo segue no forno.
2. Quando o bolo estiver pronto, é retirado e o forno desligado.

Claro que esse é um exemplo muito primário diante das possibilidades.

Há máquinas que realizam **tarefas muitas vezes mais complexas**, resolvendo problemas com milhares de variáveis, em vez de apenas uma.

Mas elas vão sempre funcionar dessa maneira: a partir de uma **programação prévia**, um código que considera essas variáveis, processa os dados e determina o que fazer em cada situação.

Inteligência artificial na análise de dados

A inteligência artificial na análise de dados voltada para **gestão de fornecedores** automatiza todo o processo operacional, como coleta, registro, classificação e análise de informações de empresas fornecedoras.

Além disso, ela pode realizar **homologação** e avaliação de fornecedores de forma rápida, com dados atualizados, garantindo uma melhor tomada de decisões.

A análise de dados com inteligência artificial gera insights capazes de fornecer dados que seriam coletados manualmente durante semanas. Consegue entender o quão ela facilita, simplifica e agiliza a gestão de fornecedores?

Veja as **principais aplicações de inteligência artificial** em todas etapas do processo de gerenciamento:

- avaliação e **gestão de riscos** do fornecedor;
- classificação de documentos e informações que promovem redução de custos ou identificação de produtos comprados de fornecedores que possuem práticas eficazes de **sustentabilidade**;
- histórico e extração de faturas;
- revisão e aprovação automática de ordem de compras;
- gestão automatizada de **contrato** .

A inteligência artificial na análise de dados ajuda a organizar informações que, antes, ficavam espalhadas por e-mails, cadernos, folhas soltas, pastas e mais pastas ou até post-it.

Em suma, podemos dizer que a inteligência artificial torna a empresa mais competitiva no mercado. Seus colaboradores produzem melhor, desenvolvem visão analítica e de melhoria contínua e tomam decisões com maior precisão.

Benefícios da inteligência artificial na análise de dados

As empresas que adotam a análise de dados com inteligência artificial recebem diversos benefícios corporativos. Principalmente quando se fala sobre avaliação de fornecedores.

O primeiro deles é a redução de custos devido a três aspectos:

- monitoramento constante para reduzir gastos desnecessários na empresa;
- aumento da **produtividade** dos colaboradores e o uso do tempo para análise estratégica;
- **indicadores de desempenho** em tempo real que avaliam fornecedores para a entrega combinar com as demandas de vendas.

Outra vantagem fornecida pela inteligência artificial na análise de dados é a previsão mais assertiva e objetiva de projeções de compras e vendas. A IA fornece relatórios e insights inteligentes.

Além disso, há a previsão de tendência do mercado, pois, por meio da interpretação de dados, é possível saber em qual época do ano há mais vendas. Com isso, pode-se programar a compra de forma mais assertiva, sem gerar desperdícios. Ou seja: o negócio se torna mais competitivo.

E o melhor: a facilidade é tanta que os dados integrados geram aprendizado na ferramenta de inteligência artificial. Ela aprende com as mudanças percebidas e passa a analisar de forma mais certa, rápida e segura.

Há ainda os indicadores de desempenho fornecidos que garantem um bom acompanhamento da performance do fornecedor. Logo, a IA contribue para gerar feedbacks e promover a cultura de melhoria contínua em todas as etapas da **cadeia de suprimentos**.

Confira mais algumas **vantagens da IA na gestão de empresas**:

- agilidade de insights;
- visão integrada de projetos;
- redução do índice de erro

Para finalizar, a inteligência artificial na análise de dados garante um melhor relacionamento com seu fornecedor. Tornando a relação mais

transparente, confiável e duradoura.

A IA na gestão de fornecedores é um dos suportes corporativos para a melhorar os resultados operacionais e financeiros, promovendo a sustentação e a melhoria constante da qualidade da entrega.

_____

Fonte: https://www.linkana.com/blog/inteligencia-artificial-na-analise-de-dados/

QUAL É O PRINCIPAL OBJETIVO DA INTELIGÊNCIA ARTIFICIAL?

Entre os inúmeros objetivos da inteligência artificial, o principal é desenvolver tecnologias que tenham a capacidade de **simular as ações humanas** e de pensar de maneira lógica.

E com isso, **criar soluções** para os mais variados aspectos da nossa vida.

A modernização das empresas é um dos resultados práticos mais evidentes do uso destas tecnologias.

Porém, há muitas outras áreas usufruindo desses benefícios, como saúde e entretenimento, conforme mostraremos mais à frente.

QUAL É A IMPORTÂNCIA DA INTELIGÊNCIA ARTIFICIAL?

- **A IA automatiza a aprendizagem repetitiva e a descoberta a partir dos dados.** Mas a inteligência artificial é diferente da automação robótica guiada por hardwares. Em vez de automatizar tarefas manuais, a IA realiza tarefas frequentes, volumosas e computadorizadas de modo confiável e sem fadiga. Para este tipo de automação, a interferência humana ainda é essencial na configuração do sistema e para fazer as perguntas certas;
- **A IA adiciona inteligência** a produtos existentes. Na maioria dos casos, a inteligência artificial não será vendida como uma aplicação individual. Pelo contrário, os produtos que você já utiliza serão aprimorados com funcionalidades de IA, de maneira parecida como a Siri foi adicionada aos produtos da Apple. Automação, plataformas de conversa, robôs e aparelhos inteligentes podem ser combinados com grandes quantidades de dados para aprimorar muitas tecnologias para casa e escritório, de inteligência em segurança à análise de investimentos;
- **A IA se adapta através de algoritmos de aprendizagem progressiva** para deixar que os dados façam a programação. A IA encontra estruturas e regularidades nos dados para que o algoritmo adquira uma capacidade: ele se torna um classificador ou predicador. Então, assim como o algoritmo pode ensinar a si mesmo a jogar xadrez, ele pode ensinar a si mesmo quais produtos recomendar em seguida. E os modelos se adaptam quando recebem mais dados. Propagação retroativa é uma técnica de IA que permite que o modelo se ajuste, através de treinamento e com a entrada de novos dados, quando a primeira resposta não está totalmente correta;
- **A IA analisa e mais dados, e em maior profundidade** usando redes neurais que possuem muitas camadas escondidas. Construir um sistema de detecção de fraudes com cinco camadas escondidas era quase impossível alguns anos atrás. Tudo isso mudou com um poderio computacional impressionante e big data. Você precisa de muitos dados para treinar modelos de deep learning porque eles aprendem diretamente com os dados. Quanto mais dados você puder colocar neles, mais precisos eles se tornam;
- **A IA atinge uma precisão incrível** através de redes neurais profundas – o que antes era impossível. Por exemplo, suas interações com a Alexa, pesquisas do Google e Google Fotos são todas baseadas em deep learning – e elas continuam ficando mais precisas conforme as vamos utilizando. Na área médica, técnicas de IA baseadas em deep learning, classificação de imagens e reconhecimento de objetos podem agora ser usadas para encontrar cânceres em ressonâncias com a mesma precisão de radiologistas bem treinados;
- **A IA obtém o máximo dos dados.** Quando algoritmos aprendem sozinhos, os dados em si podem se tornar propriedade intelectual. As respostas estão nos dados; você só precisa aplicar IA para extraí-las. Uma vez que o papel dos dados é mais importante do que nunca, eles podem criar uma vantagem competitiva. Se você possuir dados numa indústria competitiva, e ainda que todos estiverem colocando técnicas semelhantes em prática, ganha quem tiver o melhor conjunto de dados.

_____

Fonte: https://www.sas.com/pt_br/insights/analytics/inteligencia-artificial.html

DIFERENTES TIPOS DE TECNOLOGIAS E ABORDAGENS DA INTELIGÊNCIA ARTIFICIAL

Cada pesquisador da inteligência artificial tem sua própria forma de entender os desafios e oportunidades da área.

Mas, geralmente, eles se dividem em duas abordagens distintas: IA simbólica e IA conexionista.

Na **inteligência artificial simbólica**, os mecanismos efetuam transformações utilizando símbolos, letras, números ou palavras.

Simulam, portanto, o raciocínio lógico por trás das linguagens com as quais os seres humanos se comunicam uns com os outros.

Já a abordagem da **IA conexionista** se inspira no funcionamento de nossos neurônios. Simulando, portanto, os mecanismos do cérebro humano.

Um exemplo de tecnologia da abordagem conexionista é o deep learning, a capacidade que uma máquina tem de adquirir aprendizado profundo, imitando a rede neural do cérebro.

Alguns ainda falam em uma terceira abordagem, da **IA evolucionária**, que utiliza algoritmos inspirados na evolução natural.

Ou seja, a simulação de conceitos como ambiente, fenótipo, genótipo, perpetuação, seleção e morte em ambientes artificiais.

Você viu passo a passo para definir o objetivo profissional.

Mas há alguns **truques que aproximam você da construção ideal**.

Como você sabe, é através do currículo que as empresas conseguem selecionar os profissionais com perfis mais condizentes com as vagas ofertadas.

Por isso, o candidato precisa preencher o documento de uma forma que agrade os recrutadores e que o coloque em vantagem frente os concorrentes.

A primeira coisa que as empresas valorizam no currículo dos candidatos é a **certeza do que procuram para a sua carreira**. E isso fica exposto no tópico do objetivo profissional.

Ou seja, essa parte é fundamental na hora de os recrutadores escolherem quais profissionais irão para a próxima fase do processo seletivo.

As empresas esperam que o candidato especifique as suas expectativas na profissão e também na organização para a qual está enviando o currículo.

Apesar de o espaço parecer pequeno, é **preciso ser transparente e mostrar aonde quer chegar**.

Com um objetivo bem definido, o seu currículo será bem avaliado pelos profissionais de RH da empresa da qual pretende fazer parte.

O próximo passo, então, é fugir dos erros comuns, que vamos relacionar a seguir.

Cada pesquisador da inteligência artificial tem sua própria forma de entender os desafios e oportunidades da área.

Mas, geralmente, eles se dividem em duas abordagens distintas: IA simbólica e IA conexionista.

Na **inteligência artificial simbólica**, os mecanismos efetuam transformações utilizando símbolos, letras, números ou palavras.

Simulam, portanto, o raciocínio lógico por trás das linguagens com as quais os seres humanos se comunicam uns com os outros.

Já a abordagem da **IA conexionista** se inspira no funcionamento de nossos neurônios. Simulando, portanto, os mecanismos do cérebro humano.

Um exemplo de tecnologia da abordagem conexionista é o deep learning, a capacidade que uma máquina tem de adquirir aprendizado profundo, imitando a rede neural do cérebro.

Alguns ainda falam em uma terceira abordagem, da **IA evolucionária**, que utiliza algoritmos inspirados na evolução natural.

Ou seja, a simulação de conceitos como ambiente, fenótipo, genótipo, perpetuação, seleção e morte em ambientes artificiais.

TIPOS DE INTELIGÊNCIA ARTIFICIAL

À medida que o conceito de inteligência artificial passou a ser mais difundido, novos estudiosos passaram a se debruçar sobre ele.

Assim, surgiram também **perspectivas diferentes**.

Uma dessas contribuições foi a diferenciação entre dois tipos de IA, a forte e a fraca, que detalhamos abaixo:

Inteligência Artificial Forte

Também conhecida como autoconsciente, a Inteligência Artificial Forte é aquela que emula o raciocínio humano com tamanha **perfeição** que é capaz de resolver situações de maneira mais rápida e assertiva que uma pessoa.

Não à toa, é um tema bastante **polêmico**, pois muitos entendem se tratar de uma tecnologia que chega para ser uma alternativa à mão de obra mais qualificada das empresas.

Outros dilemas éticos cercam esse assunto, lembrando filmes de ficção, como “Eu, Robô”.

Exemplos de Inteligência Artificial Forte são aqueles que se valem das técnicas de machine learning e de deep learning.

Inteligência Artificial Fraca

Já a Inteligência Artificial Fraca, como o nome já sugere, não possui esse poder tão grande de imitar cognitivamente o raciocínio humano.

Na prática, ela pode colaborar no processamento de um grande volume de informações e até realizar relatórios, mas **sem a autoconsciência** do tipo anterior.

A grande questão é que uma IA fraca pode se desenvolver e chegar ao estágio de forte, ainda que a maioria dos avanços esteja na primeira classificação.

Um exemplo de Inteligência Artificial Fraca é o Processamento de Linguagem Natural.

Dentro dos campos da Inteligência Artificial Fraca está o **Processamento da Linguagem Natural**, que vimos há pouco.

Nesse caso, as máquinas utilizam softwares e algoritmos criados para finalidades específicas, como simular uma conversa humana.

Atualmente, boa parte dos avanços considerados relevantes para a área têm sido feitos no campo da Inteligência Artificial Fraca, com poucos progressos acontecendo na IA Forte.

EXEMPLOS DE APLICAÇÃO DA INTELIGÊNCIA ARTIFICIAL

A inteligência artificial não é mais coisa do futuro. Ela **já é aplicada em vários segmentos da economia**.

Veja algumas aplicações práticas da IA:

Indústria

A automação é uma tônica da indústria há muitas décadas. E as **máquinas não param de ficar**

mais inteligentes.

Com a IA, há equipamentos que fabricam e conferem os produtos sem precisarem ser operados por um humano.

Isso é só o começo, pois estão sendo desenvolvidas máquinas que também criam e executam novos projetos por conta própria, ou seja, fazem um trabalho criativo e não têm limitações para seu uso.

GPS

As rotas sugeridas pelo aplicativo Waze podem até parecer furada às vezes, mas as pessoas continuam utilizando porque, geralmente, ele aponta o melhor caminho.

Isso acontece porque o programa usa a inteligência artificial para interpretar dados fornecidos automaticamente por outros usuários sobre o tráfego nas vias.

Carros Autônomos

Uber, Google e Tesla são algumas das empresas que desenvolvem carros autônomos, que **não precisam de motorista** para guiá-los.

A inovação é possível graças a uma combinação de várias tecnologias e sensores que fornecem dados para os algoritmos orientarem o movimento dos automóveis.

Atendimento Ao Usuário

Chatbots e sistemas com **processamento de linguagem natural** estão ficando cada vez mais inteligentes para **substituir atendentes humanos** e estarem à disposição de usuários com dúvidas 24 horas por dia.

Varejo Online

Algoritmos de lojas virtuais reconhecem **padrões de compras** de usuários para apresentar a eles ofertas de acordo com suas preferências. A Amazon criou, neste formato, a Amazon Go, loja de varejo que não conta com estoquista e check-out, por exemplo.

Jornalismo

Com acesso a bases de dados, há **programas capazes de escrever matérias jornalísticas** informativas de um jeito que torna difícil para o leitor distingui-las de textos escritos por humanos.

Bancos

Instituições financeiras utilizam algoritmos para **analisar dados do mercado**, gerenciar finanças e se relacionar com seus clientes.

Direito

Escritórios de advocacia e departamentos jurídicos contarão com **robôs** para realizar, de forma mais rápida, precisa, direta e acessível do ponto de vista econômico, boa parte do que um advogado faz hoje.

Saúde

Na Saúde, temos um exemplo bem atual, que é o uso de máquinas inteligentes para ajudar no combate à **pandemia da Covid-19**.

A IA tem colaborado com a identificação de focos de contaminação e infectados, no **auxílio às autoridades** para gerenciar chamados e para sanar dúvidas da população, além do combate às notícias falsas.

Antes, **a tecnologia já estava cooperando** com o **diagnóstico precoce** de doenças, como o Alzheimer e o Mal de Parkinson.

Também ajudava na leitura de exames, identificando alterações em tomografias computadorizadas, por exemplo.

Redes Sociais e Aplicativos

Reconhecimento de fotos, identificação de objetos e situações, reprodução temática de vídeos, **tradução simultânea** e remoção automática de conteúdo inapropriado são algumas das contribuições da IA para **redes sociais** e outros aplicativos.

Fora isso, os algoritmos conseguem **personalizar o feed** de postagens e notícias, sugerir amizades conforme a rede de contatos, apresentar recursos de **realidade aumentada** e sincronizar conteúdos de forma instantânea.

Entretenimento

O entretenimento é uma das áreas que mais tem se beneficiado da IA.

Um dos exemplos mais cotidianos é o sistema de **recomendação personalizada** em serviços de streaming, garantindo uma melhor experiência na plataforma.

No entanto, não paramos por aí.

Os *games e eSports* estão **cada vez mais imersivos**.

Acessórios de realidade virtual oferecem uma percepção de que a pessoa está, de fato, realizando as ações do personagem da tela.

Manutenção Preditiva

Antecipar problemas é uma bela maneira de evitar dores de cabeça no futuro.

E é exatamente isso que a IA tem feito ao colaborar com a manutenção preditiva.

Ao **avaliar informações preliminares** de maquinários e produtos, ela evita que reparos desnecessários sejam feitos e que eventuais erros parem uma empresa inteira.

_____

Fonte: https://fia.com.br/blog/inteligencia-artificial/

QUESTÕES DE PROVAS

01. (FGV - Analista Judiciário (TJ RO)/Analista de Sistemas/2021) A Inteligência Artificial (IA) apoia o desenvolvimento de soluções tecnológicas capazes de realizar atividades similares às capacidades cognitivas humanas. Como exemplo, a plataforma Sinapses, desenvolvida pelo Tribunal de Justiça do Estado de Rondônia (TJRO) e adaptada para uso nacional, gerencia o treinamento supervisionado de modelos de IA. Em soluções de IA, a tecnologia que possui a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência usando dados de treinamento, podendo ser supervisionado ou não, é o(a):

A Motor de Inferência (Inference Engine) de Sistemas Especialistas (Expert Systems);

B Raciocínio Automatizado (Automated Reasoning);

C Compreensão de Linguagem Natural (Natural-Language Understanding);

D Representação do Conhecimento (Knowledge Representation) usando Lógica de Primeira Ordem (First Logic Order);

E Aprendizado de Máquina (Machine Learning).

02. (AOCP - 2020 - MJSP - Cientista de Dados - Big Data) A inteligência artificial (IA) pode ser aplicada hoje em diferentes áreas. Uma dessas áreas é a de análise de dados que, dependendo do contexto, refere-se a uma grande quantidade de possíveis operações de dados, às vezes específicas de determinados setores ou tarefas. Sabendo disso, assinale a alternativa que apresenta corretamente as quatro grandes categorias do processo de análise de dados com o uso da IA.

A Carga, correção, transformação e uso.

B Identificação, transformação, apresentação e decisão.

C Triagem, carga, limpeza e apresentação.

D Processamento, modelagem, triagem e apresentação.

E Transformação, limpeza, inspeção e modelagem.

03. (CEPERJ - 2022 - AL-MA - Técnico de Gestão Administrativa - Analista de Sistemas) Levando em consideração o uso da IA (Inteligência artificial) na "Cloud Computing", assinale a alternativa INCORRETA:

A A integração entre inteligência artificial e computação em nuvem gera um sistema denominado CPS - Cyber Physical Systems, característico da quarta revolução industrial.

B A IA está colocada como uma tecnologia exponencial que pode mudar de forma significativa o comportamento de pessoas.

C A proposta leva em consideração a criação de um estado de integração vertical providenciado por meio de redes de valor que facilitam a colaboração entre empresas.

D Na integração entre computação e IA fica criada outra camada, com todos os procedimentos amarrados com um único propósito de transferência dos dados massivos em um dimensionamento sob demanda.

Gabarito: 01/E; 02/E; 03/C

PROCESSAMENTO DE LINGUAGEM NATURAL.

Processamento de linguagem natural (PLN) é uma vertente da inteligência artificial que ajuda computadores a entender, interpretar e manipular a linguagem humana. O PLN resulta de diversas disciplinas, incluindo ciência da computação e linguística computacional, que buscam preencher a lacuna entre a comunicação humana e o entendimento dos computadores.

EVOLUÇÃO DO PROCESSAMENTO DE LINGUAGEM NATURAL

Ainda que o processamento de linguagem natural não seja uma ciência nova, essa tecnologia está avançando rapidamente graças ao interesse cada vez maior na comunicação homem-máquina, paralelamente à disponibilidade de big data, computação mais poderosa e algoritmos aprimorados.

Enquanto humano, você pode falar e escrever em inglês, espanhol ou chinês. Mas o idioma nativo de um computador – conhecido como código de máquina ou linguagem de máquina – é altamente incompreensível para a maioria das pessoas. Nos níveis mais profundos dos seus dispositivos, a comunicação acontece não com palavras, mas através de milhões de 0s e 1s que produzem ações lógicas.

Há 70 anos, programadores usavam cartões furados para se comunicar com os primeiros computadores. Esse processo manual e penoso era compreendido por um número relativamente pequeno de pessoas. Hoje você pode dizer “Alexa, eu gosto desta música”, e um dispositivo em sua casa irá abaixar o volume e responder “Ok, classificação salva” numa voz que simula a de um ser humano. Na sequência, ele adapta seu algoritmo para tocar essa música – e outras parecidas – na próxima vez que você ouvir aquela estação.

Vejamos essa interação em detalhe. Seu dispositivo foi ativado quando ouviu você falar, entendeu a intenção nas entrelinhas do comentário, executou uma ação e deu um feedback numa frase bem construída, tudo isso em cerca de cinco segundos. A interação completa só é possível graças ao PLN em conjunto com outras tecnologias de inteligência artificial como machine learning e deep learning.

QUAL A IMPORTÂNCIA DO PLN?

Grandes volumes de dados textuais

O processamento de linguagem natural ajuda computadores a se comunicar com seres humanos

em sua própria linguagem e escala outras tarefas relacionadas à linguagem. Por exemplo, o PLN possibilita que computadores leiam textos, ouçam e interpretem falas, identifiquem sentimentos e determinem quais trechos são importantes.

As máquinas de hoje podem analisar mais dados baseados em linguagem do que seres humanos, sem fadiga, de maneira consistente e imparcial. Considerando a quantidade gigantesca de dados não-estruturados que é gerada todos os dias, de registros médicos a mídias sociais, a automação será imprescindível para uma análise de texto e fala completa e eficiente.

Estruturando uma fonte de dados altamente não-estruturada

A linguagem humana é surpreendentemente complexa e diversa. Nós nos expressamos de infinitas maneiras, tanto verbalmente quanto por escrito. Não apenas existem centenas de idiomas e dialetos, como há também um conjunto único de regras gramaticais e de sintaxe, expressões e gírias dentro de cada um deles. Quando escrevemos, costumamos cometer erros ou abreviar palavras, ou omitimos pontuações; quando falamos, carregamos sotaques regionais, tendemos a murmurar e emprestamos termos de outros idiomas.

Embora o aprendizado supervisionado, o aprendizado não-supervisionado e, especificamente, o deep learning sejam hoje amplamente utilizados para modelar a linguagem humana, há também a necessidade de compreensão sintática e semântica, além de domínio, que não estão necessariamente presentes nessas abordagens de machine learning. O PLN é importante porque ajuda a resolver a ambiguidade na linguagem e adiciona uma estrutura numérica útil aos dados para muitas aplicações downstream, como reconhecimento de fala ou análise de texto.

COMO O PLN FUNCIONA?

Separando as partes elementais da linguagem.

O processamento de linguagem natural incorpora técnicas diversas para interpretar a linguagem humana, desde métodos estatísticos e de machine learning a abordagens algorítmicas e baseadas em regras. Nós precisamos de uma boa variedade de abordagens, porque dados baseados em texto ou voz divergem muito, assim como suas aplicações práticas.

Tarefas básicas de PLN incluem tokenização e análise sintática (parsing), lematização/stemização, rotulagem dos componentes do discurso, detecção de idioma e identificação de relações semânticas. Se alguma vez você estruturou orações na escola, então você já realizou todas essas tarefas manualmente.

Em termos gerais, as tarefas do PLN segmentam a linguagem em partes menores e essenciais, tenta entender as relações entre elas e explora como esses pedaços funcionam juntos para criar significado.

Essas tarefas subjacentes são frequentemente utilizadas em níveis mais complexos de PLN, tais como:

- **Categorização de conteúdo.** Um resumo do documento baseado em linguística, que inclui pesquisa e indexação, alertas de conteúdo e detecção de duplicações;
- **Descoberta e modelagem de tópicos.** Captura com precisão o significado e os temas em coleções de texto, e aplica advanced analytics como otimização e forecasting;
- **Extração contextual.** Extrai automaticamente informações estruturadas de fontes textuais;
- **Análise de sentimento.** Identifica o estado de espírito ou opiniões subjetivas em grandes quantidades de texto, incluindo o sentimento médio e a mineração de opinião;
- **Conversão fala-texto e texto-fala.** Transforma comandos de voz em texto escrito e vice-versa;
- **Sumarização.** Gera sinopses de grandes corpos de texto automaticamente;
- **Tradução de máquina.** Traduz texto ou fala de um idioma para outro, automaticamente.

Em todos esses casos, o objetivo almejado é pegar as entradas brutas e usar linguística e algoritmos para transformar ou enriquecer o texto de modo a obter resultados melhores.

MÉTODOS E APLICAÇÕES DE PLN

Como computadores entendem dados textuais

PLN e análise de texto

O processamento de linguagem natural anda de mãos dadas com a análise de texto, que conta, agrupa e categoriza palavras para extrair estruturas e significados de grandes volumes de conteúdo. A análise de texto é utilizada para explorar conteúdos textuais e encontrar novas variáveis de texto bruto, que podem ser visualizadas, filtradas ou usadas como entradas para modelos preditivos ou outros métodos estatísticos.

O PLN e as análises de texto são utilizadas em conjunto para muitas aplicações, incluindo:

- **Descoberta investigativa.** Identifica padrões e pistas em e-mails ou relatórios escritos para ajudar na detecção e resolução de crimes;
- **Conhecimentos especializados.** Classifica conteúdos em tópicos significativos para que você possa tomar ações e descobrir tendências;
- **Análise de mídias sociais.** Rastreia a relevância e o sentimento sobre tópicos específicos, e identifica influencers.

Exemplos de PLN no dia-a-dia

Existem muitas aplicações práticas e comuns para o PLN em nossas vidas cotidianas. Além de conversar com assistentes virtuais como Alexa ou Siri, eis alguns outros exemplos:

- Você já olhou para os e-mails na sua caixa de spam e notou similaridades nos assuntos? Você está vendo um filtro de spam Bayesiano, uma técnica estatística de PLN que compara as palavras mais comuns em mensagens de spam para validar e-mails e identificar lixo eletrônico;
- Você já perdeu uma ligação e leu sua transcrição automática por e-mail ou em um aplicativo? Isso é a conversão fala-texto, uma capacidade de PLN;
- Você já navegou em um site utilizando sua ferramenta de busca embutida ou ao selecionar um tópico sugerido, entidade ou tags? Então você já usou métodos de PLN para pesquisa, modelagem de tópicos, extração de entidades e categorização de conteúdo.

Um subcampo do PLN chamado entendimento de linguagem natural (ELN) começou a ganhar popularidade graças ao seu potencial em aplicações cognitivas e de IA. O ELN ultrapassa os limites do entendimento estrutural da linguagem para interpretar intenções, resolver ambiguidades contextuais e de palavras, e até mesmo criar linguagens humanas bem-formadas por si só. Algoritmos de ELN devem resolver o problema extremamente complexo de interpretação semântica – ou seja, compreender o significado pretendido da linguagem falada ou escrita, com todas as suas sutilezas, contextos e inferências que nós, humanos, somos capazes de compreender.

A evolução do PLN para o ELN tem implicações muito importantes para empresas e consumidores. Imagine o poder de um algoritmo que possa entender o significado e a nuance da linguagem humana em contextos variados, da medicina ao direito ou à sala de aula. Conforme os volumes de informações não-estruturadas continuam a crescer exponencialmente, nós iremos nos beneficiar da capacidade incansável dos computadores de entender tudo.

_____

Fonte: https://www.sas.com/pt_br/insights/analytics/processamento-de-linguagem-natural.html

QUESTÕES DE PROVAS

01. **(**FAURGS - 2018 - UFRGS - Técnico de Tecnologia da Informação - Sistema da Informação) Uma nuvem de palavras é um recurso gráfico (usado principalmente na internet) para descrever os termos mais frequentes de um determinado texto. O tamanho da fonte em que a palavra é apresentada é uma função da frequência da palavra no texto: palavras mais frequentes são desenhadas em fontes de tamanho maior, palavras menos frequentes são desenhadas em fontes de tamanho menor.

Qual é a técnica de análise de dados descrita pelo texto acima?

A Processamento de Linguagem Natural.
B Agrupamento.
C Classificação.
D Redes Neurais.
E Regressão Linear.

Gabarito: 01/A

GOVERNANÇA DE DADOS: CONCEITO, TIPOS (CENTRALIZADA, COMPARTILHADA E COLEGIADA).

INTRODUÇÃO

Quando falamos sobre "**dados**", na verdade estamos nos referindo à matéria prima de uma cadeia (dado - informação – conhecimento – sabedoria). A evolução desta cadeia é fundamental para conseguir o que todas as empresas desejam: transformar **dados** em **informações** confiáveis e utilizá-las como fonte de **conhecimento** para tomar decisões com **sabedoria.** A **Figura 1** a seguir mostra a evolução desta cadeia.

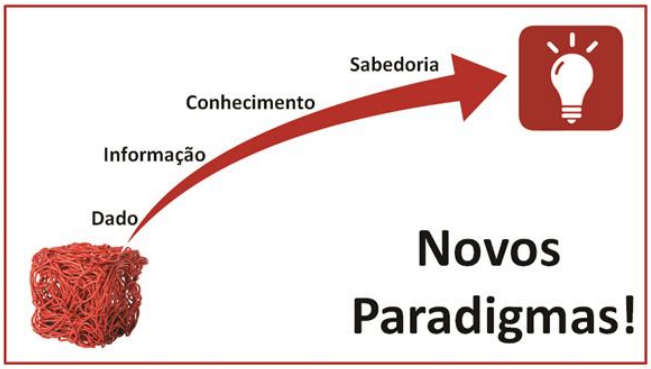


Figura 1. Cadeia de Evolução dos Dados

Sem uma gestão efetiva dos dados, a evolução desta cadeia não é atingida, portanto, para atingir os objetivos é fundamental a disciplina atuar nos estágios iniciais da cadeia. Por esta razão o nome da disciplina é Gestão de Dados e não Gestão das Informações ou Gestão do Conhecimento. Porém, valem ressaltar que, dependendo do nível da maturidade da empresa, as ações de gestão para a evolução da cadeia podem se estabelecer em outros níveis.

– Quais as premissas da Gestão de Dados?

Com o ressurgimento da Gestão de Dados, os cuidados em não repetir os erros do passado se tornaram premissas fundamentais para o sucesso da disciplina, entre as quais podemos destacar:

- A responsabilidade pela gestão dos dados não é mais uma exclusividade das áreas de Tecnologia da Informação (TI). Agora esta gestão é compartilhada entre as áreas de TI e demais áreas de negócio nas empresas.
- O dado é considerado um ativo de valor precioso nas empresas, porém os dados não são o único ativo importante. Além dos dados, pessoas e recursos financeiros também são fundamentais para a eficiência das empresas.
- O dado é gerido em todo o seu ciclo de vida, principalmente quando ele está inserido nas

operações de negócio e não mais apenas no ciclo de vida do desenvolvimento dos sistemas.

- Dados só geram valor para a empresa quando disponibilizados para utilização nas áreas de negócio. Enquanto suas estruturas são construídas, gera-se apenas custo de desenvolvimento.
- A Gestão de Dados deve acompanhar o mesmo ritmo de evolução dos negócios e da tecnologia.
- O escopo de atuação da gestão de dados é bem abrangente. Segundo a versão atual do guia DAMA-DMBOK® esta atuação envolve dez funções integradas.

- Quais as funções relacionadas à gestão de dados?

O guia DAMA-DMBOK® recomenda em seu framework um conjunto integrado de dez funções de gestão de dados, são elas:

- **Governança de Dados** – função responsável por representar o exercício de autoridade e controle das estratégias, políticas, papéis e atividades envolvidos com os ativos de dados das empresas.
- **Gestão da Arquitetura de Dados** – função responsável por definir as necessidades de dados e alinhar os mesmos com a estratégia de negócio da empresa.
- **Gestão do Desenvolvimento dos Dados** – função responsável pelas atividades de modelagem e implementação das estruturas dos dados dentro do ciclo de vida do desenvolvimento dos sistemas de informação.
- **Gestão de Operações de Dados** – função responsável por manter armazenados os dados ao longo do seu ciclo de vida.
- **Gestão da Segurança dos Dados** – função responsável por definir e manter as políticas de segurança da informação da empresa.
- **Gestão de Dados Mestres e Dados de Referência** – função responsável por definir e controlar atividades para garantir a consistência e disponibilização de visões únicas dos principais dados reutilizados na empresa.
- **Gestão de Data Warehousing e Business Intelligence:** função responsável por definir e controlar processos para prover dados de suporte à decisão, geralmente disponibilizados em aplicações analíticas.
- **Gestão da Documentação e Conteúdo:** função dedicada a planejar, implementar e controlar atividades para armazenar, proteger e acessar os dados não estruturados das empresas.
- **Gestão de Metadados:** Os metadados representam o significado dos dados. Estes significados correspondem tanto ao conteúdo técnico do dado, obtido através das informações sobre estrutura, formato, tamanho e restrições como a informações sobre definições e conceitos.
- **Gestão da Qualidade dos Dados**: função responsável por promover, medir, avaliar, melhorar e garantir a qualidade dos dados da empresa.

– Qual a diferença entre Gestão de Dados e Governança de Dados?

Esta pergunta é muito comum para as pessoas que começam a ter contato com a Gestão de Dados. Conforme dito anteriormente, segundo a versão atual do guia DAMA-DMBOK®, a Gestão de Dados é uma disciplina formada pelo conjunto de dez funções de gerenciamento de dados integradas. A integração dessas funções é feita pela função de Governança de Dados, por esta razão ela está localizada como elemento central do framework do DAMA-DMBOK®. A **Figura 2** a seguir demonstra as dez funções do guia DAMA-DMBOK®.

Figura 2. Funções do DAMA-DMBOK

A **função de governança de dados** guia como todas as outras funções da gestão de dados são realizadas, estabelecendo o exercício de direitos de decisão para otimizar, proteger e influenciar os dados como um ativo da empresa. Envolve ainda a "orquestração" de pessoas, processos, estratégias, políticas e tecnologias ligadas aos dados da empresa.

Resumo da ópera: A **função Governança de Dados** define e acompanha o funcionamento das demais funções de Gestão de Dados. Sem ela não há Gestão de Dados. Por esta razão considero a Governança de Dados a principal função da Gestão de Dados.

CONCEITO DE COVERNANÇA

A **governança de dados** pode ser definida como uma disciplina que apoia o gerenciamento de dados corporativos.

A definição do Gartner é a seguinte: a governança de dados engloba uma coleção de processos, funções, políticas, padrões e métricas que garantem o uso eficiente e eficaz das informações, permitindo que uma organização alcance suas metas.

Estas definições de governança de dados indicam que uma governança robusta segue padrões e políticas que garantem o uso dos dados com

integridade. Ela estabelece quem pode realizar quais ações em quais situações, com quais dados e quais métodos.

Uma de suas funções é alinhar pessoas, processos e tecnologias sob a ótica dos dados. Nesse sentido, o objetivo é determinar papéis, responsabilidades e projetos necessários para a devida **gestão das informações estratégicas** que transitam por determinada empresa.

É de esperar, portanto, que a governança de dados atue como uma autoridade articuladora a fim de estabelecer diretrizes da gestão de dados, liderar iniciativas de melhorias e orquestrar todo esse trabalho.

Em outras palavras, trata-se de uma disciplina-mãe que gere várias outras disciplinas e age como uma espécie de guardiã das informações. Essa estrutura serve de referência para todas as áreas da empresa em caso de necessidades desse tipo.

É seu papel, por exemplo, estipular requisitos a respeito dos dados coletados – e que podem ser muitos! No caso da Neoway, o processo de coleta considera centenas de fontes de informação publicamente acessíveis. É preciso determinar regras, validar juridicamente os dados e sua qualidade, definir quem pode utilizá-los e em quais situações, entre outras coisas.

Veja abaixo uma imagem ilustrativa que dá a dimensão das atribuições da governança de dados. Perceba como todas as atividades secundárias orbitam em torno da disciplina principal.

ESTRUTURA DA GOVERNANÇA DE DADOS

A estrutura da governança de dados é o modelo que lança as bases da estratégia e do compliance. A partir do modelo de dados que descreve seus fluxos – entradas, saídas e parâmetros de armazenamento –, o modelo de governança se sobrepõe a regras, atividades, responsabilidades, procedimentos e processos que definem como gerir e controlar esses fluxos.

Pense no modelo como uma espécie de mapa de como a governança de dados funciona em uma determinada organização. E observe que essa estrutura de governança será única em cada organização, por refletir as especificidades dos sistemas de dados, as tarefas e responsabilidades organizacionais, os requisitos regulatórios e os protocolos do setor.

Sua estrutura deve incluir o seguinte:

- **Escopo de dados:** mestres, transacionais, operacionais, analíticos, Big Data etc.
- **Estrutura organizacional:** funções e responsabilidades entre responsável, chefe de dados, TI, equipe de negócios e patrocinador executivo.
- **Padrões e políticas de dados:** guias que descrevem o que é gerenciado e governado e com que resultados.
- **Supervisão e métricas:** parâmetros para medir o sucesso e a execução da estratégia.

A IMPORTÂNCIA DA GOVERNANÇA DE DADOS

Governança de Dados permite validar, qualificar, distribuir, organizar e armazenar as informações da organização de maneira precisa, ágil e eficiente.

Um exemplo prático de uso da Governança de Dados pode ser verificado na implementação da plataforma de Análise de Dados para Administração Pública, denominada GovData. Essa plataforma disponibiliza infraestrutura de armazenamento e hospedagem, permitindo o cruzamento de grandes volumes de dados, visando:

1. A utilização de ferramentas para análise e cruzamento de dados para geração de informações para a tomada de decisões.
2. O favorecimento da desburocratização por meio de acesso centralizado a informações de governo para simplificar a oferta de serviços públicos.
3. A ampliação da transparência permitindo a análise de contas públicas para combater fraudes.
4. A adoção de tecnologia de ponta no processamento de grande volume de dados com rápido tempo de resposta.
5. A viabilização da segurança e garantia de sigilo e individualização das bases de dados.
6. A alavancagem da economicidade pelo uso compartilhado de infraestrutura e do consumo de dados para redução de custos.

Os ganhos advindos do uso da Governança de Dados são muitos – alguns intangíveis – e variam para cada organização. Os principais ganhos comuns à maioria das organizações que adotam a Governança de Dados como parte de sua estrutura organizacional são:

• Mudança de cultura: dados e informações passam a ser reconhecidos como importantes ativos estratégicos nas organizações.

• Melhor alinhamento entre as áreas de Tecnologia da Informação e Comunicação (TIC) e de Negócio: esse alinhamento é premissa fundamental para o bom funcionamento da Governança de Dados. Com isso, outras áreas como a de mapeamento de processos e a de desenvolvimento de sistemas podem se beneficiar de alinhamentos já iniciados.

• A gestão das operações de captura, armazenamento, proteção, planejamento, controle e garantia da qualidade dos ativos de dados é centralizada em uma única estrutura, permitindo a redução de custos e a otimização do uso dos recursos.

• Criação de uma cultura do uso de indicadores de processo, qualidade e desempenho de dados e informações: o objetivo é manter alinhados a Governança de Dados e o Planejamento Estratégico da Organização.

• Conhecimento de dados e informações utilizados por meio da adoção de um vocabulário único sobre as definições dos dados: ampliação e melhoria da disseminação do conhecimento entre as pessoas – passagem do capital intelectual para o capital estrutural.

• Entendimento das principais necessidades de dados e informações da organização, fornecendo um importante subsídio para estabelecer o planejamento para absorção, criação e/ou transformação de novos dados e informações para a empresa: definir o que realmente é importante em relação à utilização de dados e informações e estabelecer prioridades em relação às futuras implementações e mudanças.

• Redução da quantidade de informações redundantes.

• Estabelecimento de mecanismos formais de segurança para acesso e disponibilização de dados e informações.

• Reutilização de dados corporativos e/ou compartilhados, por meio do gerenciamento de dados mestre e dados de referência.

• Total governança dos dados manipulados pela organização.

Benefícios da governança de dados

Entre os benefícios da governança de dados estão:

1. **Dados melhores e mais confiáveis:** Sem dúvida esse é o principal objetivo. Usuários e tomadores de decisão terão mais confiança nos dados e, por consequência, nas decisões tomadas com base nesses dados. Essas decisões serão, sem dúvida, melhores por se basearem em informações precisas.
2. **Uma única versão da verdade:** O benefício de ter todas as partes da organização e todos os tomadores de decisão trabalhando com as mesmas informações é incalculável. Não se perde tempo discutindo que planilha ou plano está "melhor" ou mais atualizado. A organização toda se coordena.
3. **Compliance regulatório, jurídico e setorial:** os procedimentos robustos de gerenciamento de dados são o segredo do compliance. Na verdade, auditores e representantes de fiscalização regulatória analisam menos os dados do que o modo como foram gerados, tratados e protegidos.
4. **Redução de custos:** Além de tornar as auditorias rápidas e fáceis, as operações diárias ficarão mais eficazes. Você pode reduzir o desperdício causado por decisões baseadas em informações incorretas ou obsoletas. O atendimento ao cliente melhora quando se conhece o status da atividade, do estoque e da disponibilidade de mão de obra.

MELHORES PRÁTICAS DE GOVERNANÇA DE DADOS

Desde sua fundação em 2003, o Data Government Institute (DGI) tem sido uma referência para as melhores práticas de governança de dados. Sua estrutura é usada por centenas de organizações no mundo todo. Os princípios fundamentais de uma boa governança de dados são:

- Uma organização deve definir uma equipe de governança de dados com descrições claras do cargo, das responsabilidades e das obrigações. Isso inclui a definição de quem é responsável pelas decisões, processos e controles multifuncionais relacionados aos dados.
- Os programas de governança de dados devem definir as responsabilidades, implementando um equilíbrio entre as equipes organizacionais e tecnológicas para garantir um trabalho eficaz em prol de uma meta comum.
- As decisões, controles e processos relacionados aos dados devem ser auditáveis e acompanhadas pela documentação que corrobora os requisitos de compliance. Além disso, a estrutura deve facilitar a padronização da governança de dados da empresa.
- Todos na organização devem trabalhar com integridade ao lidar uns com os outros e com dados. As pessoas devem ser honestas durante as discussões e ao darem seu feedback quando tomam decisões relacionadas aos dados.
- Os processos da organização de dados exigem transparência para que todos os participantes e auditores saibam quando e como as decisões e controles relacionados aos dados são integrados aos processos.
- E por fim, os programas eficazes de governança de dados devem promover mudanças proativas e reativas implementadas pela gestão para garantir o

tratamento correto dos processos relacionados aos dados.

FERRAMENTAS PARA A GOVERNANÇA DE DADOS

Uma vez que os dados e aplicações se tornaram cruciais para as organizações, as ferramentas de governança de dados criadas para proteger a integridade dos ativos de dados se tornaram ainda mais importantes.

A maioria das ferramentas de governança de dados conseguem:

- Melhorar a tomada de decisão
- Aumentar a qualidade dos dados
- Simplificar o gerenciamento dos dados
- Aumentar a interoperabilidade dos dados
- Melhorar a linhagem dos dados

No momento de selecionar a melhor ferramenta para sua estrutura de governança de dados, lembre-se que as ferramentas não são o mais importante, mas sim as metas e objetivos da sua estratégia de governança de dados.

MODELOS OPERACIONAIS DE GOVERNANÇA DE DADOS?

Existem alguns modelos operacionais para Governança de Dados, entre eles os principais são os:

Centralizado

A Governança de Dados supervisiona todas as atividades em todas as áreas temáticas. É o modelo operacional de gerenciamento de dados mais formal e maduro. Aqui tudo é propriedade da área de gerenciamento de dados.

Os envolvidos em governar e gerenciar dados **reportam-se diretamente a um líder** de governança de dados responsável por governança, administração, gerenciamento de metadados, gerenciamento de qualidade de dados, gerenciamento de dados mestre e de referência.

O benefício de um modelo centralizado é que ele **estabelece uma posição executiva formal** para Gerenciamento de Dados ou Governança. Há uma pessoa no topo. A tomada de decisões é mais fácil porque a responsabilidade é clara.

A desvantagem é que a implementação de um modelo centralizado geralmente requer mudanças organizacionais significativas. Existe também o risco de formalização da separação da função de gerenciamento de dados que afasta dos principais processos de negócios e pode resultar em conhecimento se perdendo com o tempo.

Essas organizações também costumam fazer parte de um serviços compartilhados ou equipe de operações ou parte da organização do Chief Data Officer.

Descentralizado

O mesmo modelo operacional e padrões de GD são adotados por cada unidade de negócio, as responsabilidades de gerenciamento de dados são distribuídas em diferentes linhas de negócios e TI .

A colaboração é baseada em comitês; não há um único dono. Os programas começam como esforços de base para unificar as práticas de gerenciamento de dados em uma organização e, portanto, têm uma estrutura descentralizada.

Os benefícios deste modelo incluem sua estrutura relativamente plana e seu alinhamento do gerenciamento de dados às linhas de negócios e TI. Esse alinhamento geralmente significa que há uma compreensão clara dos requisitos de dados. Isso é também relativamente fácil de implementar ou melhorar.

As desvantagens incluem o **desafio de ter muitos participantes envolvidos** com órgãos de governança e tomar uma decisão. Geralmente é mais difícil implementar decisões colaborativas do que editais centralizados.

Os modelos descentralizados geralmente são menos formais e, por isso, podem ser mais difíceis de sustentar ao longo do tempo. Para que sejam bem sucedidos, eles precisam ter maneiras de reforçar a consistência das práticas. Isso pode ser difícil de coordenar.

Costuma ser difícil definir a propriedade dos dados com um modelo descentralizado.

Híbrido

Como o próprio nome indica, o modelo operacional híbrido abrange benefícios tanto do modelo descentralizado quanto do centralizado. Em um modelo híbrido, não há um centro de excelência de gerenciamento de dados centralizado.

Trabalha com grupos de unidades de negócios descentralizados, geralmente por meio de um comitê executivo que representa linhas-chave de negócios e um conjunto de grupos de trabalho táticos voltados para problemas específicos.

Nesse modelo, algumas funções permanecem descentralizadas. Por exemplo, os arquitetos de dados podem permanecer em uma área de arquitetura; as linhas de negócios podem ter suas próprias equipes de qualidade de dados.

O principal benefício de um modelo híbrido é que ele estabelece a direção apropriada do topo da organização. Existe um executivo responsável pela gestão e/ou governança dos dados. Equipes da Unidade de Negócios têm ampla responsabilidade e podem se alinhar às prioridades de negócios para fornecer maior foco.

Federativo

Coordena com várias Unidades de Negócios para manter definições consistentes e padrões. Uma variação do modelo operacional híbrido, o modelo federado fornece camadas adicionais de centralização/descentralização, que muitas vezes são necessárias em grandes empresas globais.

Um modelo federado fornece uma estratégia centralizada com execução descentralizada. Portanto, para grandes empresas pode ser o único modelo que pode funcionar. Um executivo de gerenciamento de dados responsável por toda a organização administra o Centro de Excelência empresarial.

Claro, diferentes linhas de negócios têm o poder de atender aos requisitos com base em suas necessidades e prioridades. A federação permite que a organização priorize com base em entidades de dados específicas, desafios divisionais ou prioridades regionais.

A principal desvantagem é a complexidade. Existem muitas camadas e é preciso haver um equilíbrio entre autonomia para as linhas de negócios e as necessidades da empresa. Esse equilíbrio pode afetar as prioridades da empresa.

PILARES DE GOVERNANÇA DE DADOS?

O objetivo da governança de dados é trazer o maior retorno possível sobre os ativos de dados, vislumbrando oportunidades críticas para alavancar os dados, evitando os riscos de expô-los. Estes são os pilares a serem considerados ao avaliar sua prontidão e maturidade em governança de dados:

Pessoas

As pessoas colaboram na determinação dos requisitos de tecnologia, definem os processos e, por fim, conduzem os resultados de governança de dados que apoia de forma direta a gestão estratégica e a tomada de decisão baseada em dados.

As equipes devem estar comprometidas com a governança de dados. Os papéis e responsabilidades devem ser formalizados. As pessoas têm as habilidades necessárias e são realizados treinamentos.

Existe um plano de gerenciamento de mudanças, incluindo Sponsors, para apoiar o alinhamento organizacional e a adesão de uma estratégia orientada a Dados?

Processos

Os processos de governança de dados permitem que as pessoas identifiquem que seus dados são formalmente gerenciados em toda a empresa, o que garante que seus processos de negócios críticos se baseiam em dados confiáveis.

As definições de dados, regras e metas devem ser realistas e apropriadas ao negócio. Seus processos de negócios precisam ser modernizados e suas regras de negócios devem ter períodos definidos para revisão sendo integrados a governança de dados de forma limpa.

Tecnologia

A tecnologia engloba ferramentas, plataformas, sistemas e conhecimento nos assuntos necessários para permitir um processo sólido e vivo de governança de dados.

Tendo sistemas existentes que já sejam gerenciados até certo ponto, os administradores de tecnologia de plataforma, como perfis de dados, linhagem e ferramentas de metadados, são de extrema importância para sua capacidade de otimizar, automatizar, administrar e dimensionar seus ativos de Dados e seus processos de governança de dados.

CIÊNCIA DE DADOS: IMPORTÂNCIA DA INFORMAÇÃO.

Utilizar dados e informações para tomar decisões em uma empresa é uma habilidade ainda para poucos, mas que se torna cada vez mais necessária no mercado. **Ciência de dados** e análise de informação são áreas que ganham gradativamente mais relevância dentro das empresas e, por isso, profissionais que sabem ler dados podem ser o grande diferencial de uma organização.

Assim, do ponto de vista profissional, desenvolver essa habilidade torna-se essencial para o mundo dos negócios, que, por sua vez, só tem a ganhar ao enxergar o valor do colaborador com esse conhecimento e habilidade.

A importância da ciência de dados

Precisão das informações é o grande diferencial da ciência de dados. Com isso, tomar qualquer decisão dentro de uma empresa torna-se tarefa analítica, tentando encontrar as melhores estratégias a partir das informações existentes.

Exatamente por essa exatidão dos dados e das informações é que entendê-los se tornou tão importante. A partir dessa habilidade, é possível:

- Enxergar pontos fracos da empresa: uma boa análise de dados pode mostrar quais são os “elos fracos” de uma operação;
- Dimensionar perdas: é possível calcular, também, qual setor traz mais dificuldades ou que tem potencial de prejuízo, por exemplo;
- Analisar forças: também é possível observar os pontos fortes de uma empresa a partir do uso de dados;
- Tomar decisões mais adequadas: quando as decisões são embasadas em informações concretas, torna-se muito mais fácil pensar em estratégias para uma empresa;
- Evitar “achismos”: assim como contribui com a decisão, também evita os “achismos”, ou seja, as decisões tomadas sem dados concretos da possível eficácia daquilo;
- Elaborar estratégias específicas: os dados também podem indicar qual caminho uma empresa deveria seguir de acordo com seus resultados atuais.

Quando se analisa uma questão a partir de dados é possível identificar variações que podem significar eventuais falhas ou perdas, por exemplo. Isso dá a empresa a chance de pensar e decidir um plano de ação para solucionar o caso. “Ou seja, o dado permite dimensionar o tamanho do problema e tomar a decisão de acordo com informações concretas”, explica Lucas Silva.

Ler e entender dados

A habilidade de ler e entender dados ainda não é tão difundida no Brasil. De acordo com o cientista de dados, quem realmente sabe analisar informações traz um grande diferencial para qualquer empresa.

"Seja um negócio grande ou pequeno, ter alguém que entenda de dados pode trazer muitos resultados", diz.

A importância de entender dados já é tão grande que é considerada um novo tipo de linguagem. O data literacy, ou a leitura de dados, é tida como uma habilidade do futuro que pode vir a ser tão importante quanto saber um novo idioma.

Quem souber "falar dados" pode sair na frente em qualquer mercado. Portanto, os profissionais que buscam um aprimoramento de seus conhecimentos podem apostar na ciência de dados como um upskilling (novo aprendizado) para a carreira.

E engana-se quem pensa que apenas as empresas tecnológicas estarão aplicando o uso de dados em sua cultura interna. De acordo com Lucas Silva, todos os mercados podem se beneficiar da análise de informações.

"O que muda é a velocidade em que isso acontece. Alguns mercados recebem atualizações o tempo todo, fazendo com que o consumidor mude seu comportamento mais rápido. Enquanto outros podem demorar mais tempo para terem os dados concretos a serem analisados, mas não quer dizer que as informações sejam menos importantes", afirma.

BIG DATA. BIG DATA EM RELAÇÃO A OUTRAS DISCIPLINAS.

O QUE É O BIG DATA?

Big Data é a tecnologia que normalmente se refere a uma quantidade monumental de dados estruturados, não-estruturados e semiestruturados.

É na ferramenta de Big Data onde são aplicadas análises com o propósito de obter informações úteis e direcionadas para auxiliar na decisão do seu negócio.

Ou seja, é uma parte da área de Business Intelligence, utilizado para classificar as informações e identificar quais são as dores do mercado e dos clientes.

Hoje em dia, já existem aplicações de Big Data para um grande número de atividades no mercado, seja a sua empresa um negócio B2B, B2C, B2E ou afins. Inclusive, é uma tecnologia tão útil e tão vasta que já oferece soluções para empresas de todos os tamanhos, sejam pequenas e médias, até grandes corporações.

QUAIS OS 5 V'S DA BIG DATA?

Velocidade, volume, variedade, veracidade e valor. Em nossa visão, esses são os 5 V's da Big Data, que nada mais são do que pilares fundamentais para a criação de novas tecnologias e soluções nessa área. Abaixo, esclarecemos cada um desses pontos em detalhes. Acompanhe!

1. Velocidade

A primeira característica é a velocidade de coleta, organização e análise dos dados. Aqui, falamos de volumes colossais, como terabytes e petabytes de informação, e essas soluções são capazes de organizar esses grandes reservatórios de dados, eliminar os ruídos e trazer ordem e lógica a esse caos de informação e aleatoriedade.

2. Volume

Em segundo, destacamos o volume informacional. A depender da empresa e da aplicação da tecnologia, o volume de informações coletadas é colossal. Por conta disso, essas soluções se destacam por sua robustez e confiabilidade, justamente por serem capazes de ingerir tamanha quantidade de informação em tempo real.

3. Variedade

As informações podem ser coletadas em dois diferentes tipos: os dados estruturados e os não estruturados. Enquanto o primeiro é uma ingestão organizada, como planilhas e arquivos CSV ou XLSX, o segundo consiste em dados nos mais variados formatos, como conteúdos de mídia como áudios, imagens, vídeos, textos e afins.

4. Veracidade

O objetivo da Big Data é encontrar ordem e lógica em meio a aleatoriedade de um grande banco de dados. É por isso que boas soluções de analytics se destacam pela capacidade de minimizar esses ruídos, identificando excessos, redundâncias e equívocos, e auxiliando na limpeza e na confiabilidade do dataset final.

5. Valor

Por fim, o valor e o retorno dessas soluções para a sua gestão. A Big Data é uma tecnologia que estimula a descoberta de insights e abordagens para a sua operação. Negligenciar essas informações e ignorar o diferencial oferecido pela tecnologia é o mesmo que negligenciar o investimento realizado nessas ferramentas.

QUAIS AS PRINCIPAIS APLICAÇÕES DE BIG DATA?

Para encerrar, vale destacar um último "V" da Big Data, o vínculo. Apesar da ingestão de dados admitir muitas informações aleatórias, é função, também, dessas ferramentas e dos analistas encontrar a correlação nos dados, garantindo um data set coeso, limpo e estrategicamente importante para a sua tomada de decisão.

Agora sim, vamos destacar alguns exemplos dessa solução no mercado. Um setor que utiliza muito esse conceito é o campo da securitização. Grandes seguradoras estudam grandes bancos de dados, incluindo incidência de acidentes e mapas criminais, para assim, derivar a melhor relação de

risco e retorno para a precificação das apólices.

A mesma abordagem é utilizada no mercado financeiro, onde o cálculo e a gestão de riscos são partes críticas na concessão de créditos. No entanto, vale falar sobre aplicações na ampla economia, sobretudo nas PMEs, as pequenas e médias empresas.

Afinal de contas, muitos gestores mais antenados em tendências e tecnologias já utilizam de soluções de Big Data e Inteligência Artificial para duas necessidades nos supermercados e varejos, que é a precificação dos produtos e o controle de estoque.

Em EPPs de crédito pessoal, a Big Data também é utilizada, juntamente das informações dos bureaus de crédito, para definir a relação de risco na oferta de empréstimos para novos usuários.

No setor da saúde, a ingestão de dados biométricos, como os coletados por smartwatches e outros dispositivos vestíveis, também já é utilizado no trabalho de diagnóstico e acompanhamento de quadros clínicos, sobretudo em países com uma malha tecnológica mais moderna, como os Estados Unidos, o Canadá, e mais recentemente, o Brasil.

Para quem é varejista digital, a Big Data também está à disposição, em ferramentas que qualificam a elaboração de anúncios e publicidades, orientando os melhores períodos, palavras-chaves e estratégias para a veiculação de campanhas nas redes sociais.

Em essência, a aplicação de Big Data no mercado é praticamente infinita e, felizmente, só estamos no começo dessa jornada.

QUESTÕES DE PROVAS

01. (IADES - BRB - Analista de Tecnologia da Informação – 2021) Considere as seguintes conceituações sobre os 5 V's do Big Data:

I – É uma das características mais evidenciadas, pois relaciona todos os e-mails, vídeos, fotos, mensagens e comentários que caminham pela rede, medidas em Zetabytes. Utilizada para lidar com o volume de dados, armazenando-os em diferentes locais, podendo ser utilizado sempre que necessário.

II – Por não existir uma padronização de formato e tamanho, a análise de dados estruturada acaba exigindo um esforço maior. Com o Big Data, essas informações passam a ser trabalhadas em conjunto.

III – É a maior característica do Big Data fazer com que os dados sejam acessados em tempo real, otimizando o tempo de qualquer tarefa que envolva o uso do Big Data.

IV – Um dos pontos mais importantes de qualquer informação é que ela seja verdadeira. Com o Big Data não é possível controlar cada hashtag do Twitter ou notícia falsa na internet, mas com análises e estatísticas de grandes volumes de dados é possível compensar as informações incorretas.

V – É a característica de maior relevância, pois não adianta acessar um mundo de informações sem que elas agreguem valor ao seu negócio.

De acordo com as conceituações anteriores, elas se referem, respectivamente, aos seguintes V's do Big Data:

A Volume, Variedade, Velocidade, Veracidade e Valor.
B Volume, Valor, Velocidade, Variedade e Veracidade.
C Variedade, Volume, Valor, Veracidade e Velocidade.
D Valor, Volume, Velocidade, Variedade e Veracidade.
E Velocidade, Valor, Variedade, Veracidade e Volume.

02. (FAURGS - 2022 - SES-RS - Analista de Desenvolvimento de Sistemas) Associe os termos dos 5Vs de Big Data às suas respectivas características.

(1) Volume
(2) Velocidade
(3) Variedade
(4) Veracidade
(5) Valor

( ) Dados autênticos e verdadeiros.
( ) Processamento ágil.
( ) Utilidade dos dados.
( ) Fontes de dados muito heterogêneas.
( ) Grande quantidade de dados gerados.

A sequência correta de preenchimento dos parênteses, de cima para baixo, é

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

Gabarito: 01/A; 02/E

CIÊNCIA DOS DADOS.

Data Science ou Ciência de Dados é um estudo muito disciplinado com relação aos dados e demais informações inerentes à empresa e as visões que cercam um determinado assunto.

Em resumo é uma ciência que visa estudar as informações, seu processo de captura, transformação, geração e, posteriormente, análise de dados. A ciência de dados envolve diversas disciplinas. São elas:

- Estatística
- Computação
- Conhecimento do negócio
- Matemática

PARA QUE SERVE A CIÊNCIA DE DADOS?

A ciência de dados é usada para estudar dados de quatro maneiras principais:

1. Análise descritiva

A análise descritiva analisa os dados para

obter insights sobre o que aconteceu ou o que está acontecendo no ambiente de dados. Ela é caracterizada por visualizações de dados, como gráficos de pizza, gráficos de barras, gráficos de linhas, tabelas ou narrativas geradas. Por exemplo, um serviço de reserva de voos pode registrar dados como o número de bilhetes reservados a cada dia. A análise descritiva revelará picos de reservas, quedas nas reservas e meses de alta performance para este serviço.

2. Análise diagnóstica

A análise diagnóstica é uma análise aprofundada ou detalhada de dados para entender por que algo aconteceu. Ela é caracterizada por técnicas como drill-down, descoberta de dados, mineração de dados e correlações. Várias operações e transformações de dados podem ser realizadas em um determinado conjunto de dados para descobrir padrões exclusivos em cada uma dessas técnicas. Por exemplo, o serviço de voo pode fazer drill-down em um mês particularmente de alta performance para entender melhor o pico de reserva. Isso pode levar à descoberta de que muitos clientes visitam uma determinada cidade para assistir a um evento esportivo mensal.

3. Análise preditiva

A análise preditiva usa dados históricos para fazer previsões precisas sobre padrões de dados que podem ocorrer no futuro. Ela é caracterizada por técnicas como machine learning, previsão, correspondência de padrões e modelagem preditiva. Em cada uma dessas técnicas, os computadores são treinados para fazer engenharia reversa de conexões de causalidade nos dados. Por exemplo, a equipe de serviço de voo pode usar a ciência de dados para prever padrões de reserva de voo para o próximo ano no início de cada ano. O programa de computador ou algoritmo pode analisar dados anteriores e prever picos de reservas para determinados destinos em maio. Tendo previsto as futuras necessidades de viagem de seus clientes, a empresa poderia iniciar a publicidade direcionada para essas cidades a partir de fevereiro.

4. Análise prescritiva

A análise prescritiva leva os dados preditivos a um novo patamar. Ela não só prevê o que provavelmente acontecerá, mas também sugere uma resposta ideal para esse resultado. Ela pode analisar as potenciais implicações de diferentes escolhas e recomendar o melhor plano de ação. A análise prescritiva usa análise de gráficos, simulação, processamento de eventos complexos, redes neurais e mecanismos de recomendação de machine learning.

Voltando ao exemplo de reserva de voo, a análise prescritiva pode analisar campanhas de marketing históricas para maximizar a vantagem do próximo pico de reservas. Um cientista de dados pode projetar resultados de reservas para diferentes níveis de gastos de marketing em vários canais de marketing. Essas previsões de dados dariam à empresa de reservas de voos mais confiança para tomar suas decisões de marketing.

O QUE É O PROCESSO DE CIÊNCIA DE DADOS?

Um problema de negócios normalmente inicia o processo de ciência de dados. Um cientista de dados trabalhará com as partes interessadas do negócio para entender quais são as necessidades do negócio. Uma vez definido o problema, o cientista de dados pode solucioná-lo usando o processo de ciência de dados OSEMN:

O: Obter dados

Os dados podem ser pré-existentes, recém-adquiridos ou um repositório de dados que pode ser baixado da Internet. Os cientistas de dados podem extrair dados de bancos de dados internos ou externos, software de CRM da empresa, logs de servidores da Web, mídias sociais ou comprá-los de fontes confiáveis de terceiros.

S: Suprimir dados

A supressão de dados, ou limpeza de dados, é o processo de padronização dos dados de acordo com um formato predeterminado. Ela inclui lidar com a ausência de dados, corrigir erros de dados e remover quaisquer dados atípicos. Alguns exemplos de supressão de dados são:·

- Alterar todos os valores de data para um formato padrão comum.·
- Corrigir erros de ortografia ou espaços adicionais.·
- Corrigir imprecisões matemáticas ou remover vírgulas de números grandes.

E: Explorar dados

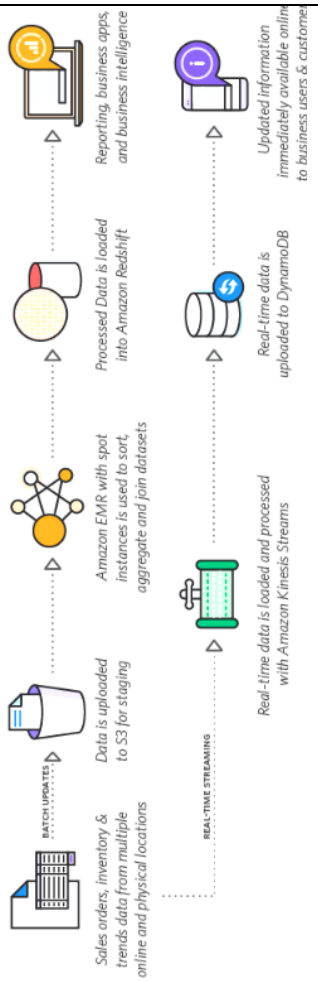
A exploração de dados é uma análise de dados preliminar que é usada para planejar outras estratégias de modelagem de dados. Os cientistas de dados obtêm uma compreensão inicial dos dados usando estatísticas descritivas e ferramentas de visualização de dados. Em seguida, eles exploram os dados para identificar padrões interessantes que podem ser estudados ou acionados.

M:Modelar dados

Os algoritmos de software e machine learning são usados para obter insights mais profundos, prever resultados e prescrever o melhor plano de ação. Técnicas de machine learning, como associação, classificação e clustering, são aplicadas ao conjunto de dados de treinamento. O modelo pode ser testado em relação a dados de teste predeterminados para avaliar a precisão dos resultados. O modelo de dados pode ser ajustado várias vezes para melhorar os resultados.

N: Interpretar resultados

Os cientistas de dados trabalham em conjunto com analistas e empresas para converter insights de dados em ação. Eles fazem diagramas, gráficos e tabelas para representar tendências e previsões. A sumarização de dados ajuda as partes interessadas a entender e implementar os resultados de forma eficaz.



QUAIS SÃO AS TÉCNICAS DE CIÊNCIA DE DADOS?

Os profissionais de ciência de dados usam sistemas de computação para acompanhar o processo de ciência de dados. As principais técnicas usadas pelos cientistas de dados são:

Classificação

Classificação é a ordenação de dados em grupos ou categorias específicos. Os computadores são treinados para identificar e classificar dados. Conjuntos de dados conhecidos são usados para criar algoritmos de decisão em um computador que processa e categoriza rapidamente os dados. Por exemplo:·

- Classificar produtos como populares ou não populares·
- Classificar as aplicações de seguro como de alto risco ou baixo risco·
- Classificar comentários de mídias sociais em positivos, negativos ou neutros.

Os profissionais de ciência de dados usam sistemas de computação para acompanhar o processo de ciência de dados.

Regressão

A regressão é o método de encontrar uma relação entre dois pontos de dados aparentemente não relacionados. A conexão geralmente é modelada em torno de uma fórmula matemática e representada como um gráfico ou curvas. Quando o valor de um ponto de dados é conhecido, a regressão é usada para prever o outro ponto de dados. Por exemplo:·

- A taxa de propagação de doenças transmitidas pelo ar.·
- A relação entre a satisfação do cliente e o número de funcionários.·
- A relação entre o número de quartéis de bombeiros e o número de feridos em decorrência de um incêndio em um determinado local.

Clustering

Clustering é o método de agrupar dados intimamente relacionados para procurar padrões e anomalias. O clustering é diferente da classificação porque os dados não podem ser classificados com precisão em categorias fixas. Portanto, os dados são agrupados em relações mais prováveis. Novos padrões e relações podem ser descobertos com o clustering. Por exemplo:

- Agrupar clientes com comportamento de compra semelhante para melhorar o atendimento ao cliente.·
- Agrupar o tráfego de rede para identificar padrões de uso diário e identificar um ataque à rede mais rapidamente.
- Agrupar artigos em diversas categorias de notícias diferentes e usar essas informações para encontrar conteúdo de notícias falsas.

O princípio básico por trás das técnicas de ciência de dados

Embora os detalhes variem, os princípios subjacentes por trás dessas técnicas são:

- Ensinar uma máquina a classificar dados com base em um conjunto de dados conhecido. Por exemplo, palavras-chave de amostra são fornecidas ao computador com seus respectivos valores de classificação. “Feliz” é positivo, enquanto “Ódio” é negativo.
- Fornecer dados desconhecidos à máquina e permitir que o dispositivo classifique o conjunto de dados de forma independente.
- Permitir imprecisões de resultados e lidar com o fator de probabilidade do resultado.

Quais são as diferentes tecnologias de ciência de dados?

Os profissionais de ciência de dados trabalham com tecnologias complexas, como:

1. Inteligência artificial: modelos de machine learning e software relacionado são usados para análises preditivas e prescritivas.
2. Computação em nuvem: as tecnologias de nuvem deram aos cientistas de dados a flexibilidade e a capacidade de processamento necessárias para análise de dados avançada.

3. Internet das Coisas: IoT refere-se a vários dispositivos que podem se conectar automaticamente à Internet. Esses dispositivos coletam dados para iniciativas de ciência de dados. Eles geram grandes quantidades de dados que podem ser usados para mineração de dados e extração de dados.
4. Computação quântica: computadores quânticos podem fazer cálculos complexos em alta velocidade. Cientistas de dados qualificados os usam para criar algoritmos quantitativos complexos.

Como a ciência de dados se compara a outros campos de dados relacionados?

Ciência de dados é um termo abrangente para outras funções e campos relacionados a dados. Vejamos alguns deles aqui:

Qual é a diferença entre ciência de dados e análise de dados?

Embora os termos possam ser usados de forma intercambiável, a análise de dados é um subconjunto da ciência de dados. A ciência de dados é um termo abrangente para todos os aspectos do processamento de dados, desde a coleta até a modelagem e insights. Por outro lado, a análise de dados envolve principalmente estatísticas, matemática e análise estatística. Ela se concentra apenas na análise de dados, enquanto a ciência de dados está relacionada ao panorama geral em torno dos dados organizacionais. Na maioria dos locais de trabalho, cientistas de dados e analistas de dados trabalham juntos para atingir objetivos de negócios comuns. Um analista de dados pode gastar mais tempo em análises de rotina, fornecendo relatórios regulares. Um cientista de dados pode projetar a maneira como os dados são armazenados, manipulados e analisados. Simplificando, um analista de dados dá sentido aos dados existentes, enquanto um cientista de dados cria novos métodos e ferramentas para processar dados para serem usados por analistas.

Qual é a diferença entre ciência de dados e análise de negócios?

Embora haja uma sobreposição entre ciência de dados e análise de negócios, a principal diferença é o uso da tecnologia em cada área. Os cientistas de dados trabalham mais de perto com a tecnologia de dados do que os analistas de negócios. Os analistas de negócios conciliam negócios e TI. Eles definem casos de negócios, coletam informações das partes interessadas ou validam soluções. Os cientistas de dados, por outro lado, usam a tecnologia para trabalhar com dados de negócios. Eles podem escrever programas, aplicar técnicas de machine learning para criar modelos e desenvolver novos algoritmos. Os cientistas de dados não só entendem o problema, mas também podem criar uma ferramenta que forneça soluções para o problema. Não é incomum encontrar analistas de negócios e cientistas de dados trabalhando na mesma equipe. Os analistas de negócios pegam a saída dos cientistas de dados e a utilizam para contar uma história que a empresa como um todo possa entender.

Qual é a diferença entre ciência de dados e engenharia de dados?

Os engenheiros de dados constroem e mantêm os sistemas que permitem que os cientistas de dados acessem e interpretem os dados. Eles trabalham mais de perto com a tecnologia subjacente do que um cientista de dados. A função geralmente envolve a criação de modelos de dados, a construção de pipelines de dados e supervisão de extração, transformação e carregamento (ETL). Dependendo da disposição e do tamanho da organização, o engenheiro de dados também pode gerenciar infraestrutura relacionada, como armazenamento de big data, transmissão e plataformas de processamento, como o Amazon S3. Os cientistas de dados usam os dados que os engenheiros de dados processaram para criar e treinar modelos preditivos. Os cientistas de dados podem então entregar os resultados aos analistas para uma tomada de decisão posterior.

Qual é a diferença entre ciência de dados e machine learning?

Machine learning é a ciência de treinar máquinas para analisar e aprender com os dados da mesma forma que os seres humanos fazem. É um dos métodos usados em projetos de ciência de dados para obter insights automatizados de dados. Os engenheiros de machine learning são especializados em computação, algoritmos e habilidades de codificação específicas para métodos de machine learning. Os cientistas de dados podem usar métodos de machine learning como uma ferramenta ou trabalhar em estreita colaboração com outros engenheiros de machine learning para processar dados.

Qual é a diferença entre ciência de dados e estatística?

A estatística é uma área de base matemática que busca coletar e interpretar dados quantitativos. Em contrapartida, a ciência de dados é um âmbito multidisciplinar que usa métodos, processos e sistemas científicos para extrair conhecimento de dados de várias formas. Os cientistas de dados usam métodos de muitas disciplinas, incluindo estatísticas. No entanto, os âmbitos diferem em seus processos e nos problemas que estudam.

Quais são as diferentes ferramentas de ciência de dados?

A AWS tem uma série de ferramentas para oferecer suporte a cientistas de dados em todo o mundo:

Armazenamento físico de dados

Para data warehousing, o Amazon Redshift

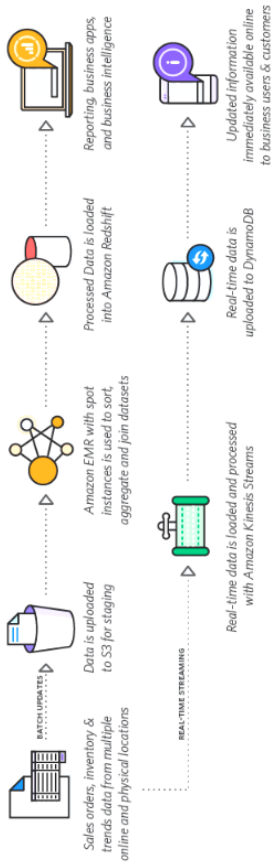
pode executar consultas complexas em dados estruturados ou não estruturados. Analistas e cientistas de dados podem usar o AWS Glue para gerenciar e pesquisar dados. O AWS Glue cria automaticamente um catálogo unificado de todos os dados no data lake, com metadados anexados para torná-los detectáveis.

Machine learning

O Amazon SageMaker é um serviço de machine learning totalmente gerenciado executado no Amazon Elastic Compute Cloud (EC2). Ele permite que os usuários organizem dados, criem, treinem e implantem modelos de machine learning e escalem operações.

Análises

- O Amazon Athena é um serviço de consultas interativas que facilita a análise de dados no Amazon S3 ou no Glacier. Ele é rápido, com tecnologia sem servidor e funciona usando consultas SQL padrão.
- O Amazon Elastic MapReduce (EMR) processa big data usando servidores como Spark e Hadoop.
- O Amazon Kinesis permite agregação e processamento de dados de transmissão em tempo real. Ele usa sequências de cliques em sites, logs de aplicações e dados de telemetria de dispositivos de IoT.
- O Amazon OpenSearch permite pesquisa, análise e visualização de petabytes de dados.



Fonte: https://aws.amazon.com/pt/what-is/data-science/

CICLO DE VIDA DO PROCESSO DE CIÊNCIA DE DADOS.

AFINAL, COMO SE DESENVOLVE UM PROJETO DE DATA SCIENCE?

O termo Data Science se tornou um *hot topic* mundial na indústria da tecnologia. O rápido avanço computacional tem permitido análises de quantidades de dados cada vez maiores, possibilitando a descoberta de padrões e *insights* até em tempo real.

Com isso, uma pergunta comum que ouvimos de amigos ou pessoas interessadas na área é: *"Cara, Data Science parece muito legal! Mas por onde começar?"*. Nesse post, daremos uma visão geral de seis etapas que fazem parte do ciclo de vida de um projeto na área: **entendimento do problema, coleta de dados**, **exploração dos dados**, **análise profunda dos dados**, **comunicação dos resultados** e **feedback.**

Você já deve ter começado a notar que essas etapas são similares a projetos de outras áreas, inclusive fora da TI, certo? O ciclo de vida de um projeto de DS envolve fases semelhantes ao ciclo de resolução de problemas do dia a dia de qualquer profissional. Por exemplo, vamos supor uma situação na qual o objetivo é comprar uma televisão. Por onde você começaria?

1. Preciso de uma televisão > ***Surgimento da demanda***
2. Eu realmente preciso de uma televisão nova? Quanto de vou precisar gastar? > ***Entendendo o problema***
3. Preciso pesquisar quais os modelos e tecnologias envolvidas das televisões da atualidade > ***Coleta de dados***
4. Parece que alguns modelos ainda não estão sendo vendidos no meu país. Meu videogame tem suporte a esse modelo? Esse modelo é considerado uma Smart TV? > ***Processamento e exploração de dados/Análise de dados***
5. Encontrei um modelo de televisão que atende minha necessidade. > ***Comunicação de resultados***
6. Instalei a televisão e estou pronto para testá-la. > ***Feedback***

Faz sentido, né?! Vamos então entender cada uma dessas etapas. Mas, antes de começarmos, vale a pena ressaltar que o ciclo de vida de um projeto de Data Science pode conter algumas diferenças de empresa para empresa dependendo de vários fatores, restrições e recursos disponíveis. Algumas etapas podem, inclusive, serem removidas ou adicionadas de acordo com o negócio.

1. Entendendo o problema

Essa é uma das etapas que considero uma das mais importantes de todo o ciclo. É nela que precisamos gastar tempo suficiente para entender o problema de forma mais clara possível. Para isso, é importante que estejamos em constante comunicação com os *stakeholders*, as pessoas envolvidas no projeto, e/ou aqueles que irão se beneficiar com a solução.

Nessa fase, é papel do Cientista de Dados entender as dores dos *stakeholders* e fazer as perguntas certas, antes mesmo de "colocar a mão na massa".

***Dica 1**:* Utilizar a Técnica dos 5-Ws:

- **Porquê?** (*Why?*): Porque é importante essa análise para o negócio?
- **Quem?** (*Who?*): Quem iremos analisar? Nossos compradores? Fornecedores?
- **O quê?** (*What?*): O que iremos analisar? Comportamento de compra?
- **Onde?** (*Where?*): A análise estará voltada para o contexto nacional ou internacional?
- **Quando?** (*When?*): Qual período será considerado para as análises?

Dica 2: Utilizar a Técnica dos 5 Porquês.

2. Coleta de dados

Uma vez definido o problema, precisamos começar a extrair e coletar os dados. Nessa etapa, é fundamental entender quais os tipos de dados irão pautar nosso projeto:

- ***Dados internos*** (presentes em bancos de dados, planilhas, etc.) **x *Dados Externos*** (bases de dados públicas ou pagas, etc.)
- ***Dados estruturados*** (tabelas dos nossos DBs) ***x*** ***Dados não-estruturados*** (conteúdos de redes sociais, de sites externos, etc.).

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

(Fonte: Datamation)

Esse mapeamento irá auxiliar na decisão das tecnologias que utilizaremos para coletar nossos dados (consultas SQL, *crawlers*, APIs, etc.).

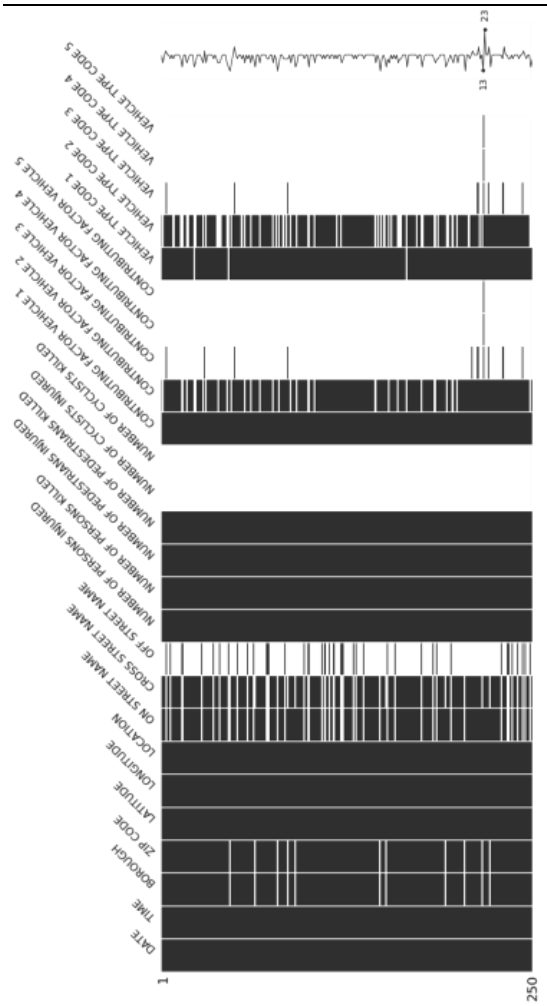
3. Processamento/Tratamento de dados

Já coletamos os dados mas precisamos tratá-los antes de começar nossas análises. Nessa etapa é necessário estar atento a registros duplicados, faltantes, formatados de forma não-convencional (ex.: campos de data), inválidos (ex.: idade negativa), inconsistências de cadastros (ex.: data da venda anterior a data de início de venda de um produto).

Após essa identificação, é importante pensar na melhor forma de contornar de acordo com as regras do negócio. Por exemplo, no caso de um registro onde o valor da compra está vazio, você poderia preenchê-lo com zero, ou com a média dos valores de compra de produtos similares, etc.

Dica 1: A biblioteca Python missingno te ajuda a encontrar valores faltantes/nulos nos seus dados.

Dica 2: A biblioteca Python pandas-profiling te ajuda a explorar seus dataframes com informações estatísticas, descritivas, histogramas e etc.



Biblioteca *missingo* (Fonte: GitHub/ResidentMario)

4. Exploração de dados

Agora sim, a etapa do ciclo de vida do nosso projeto de Data Science onde a resolução do problema inicial começa a tomar forma!

Na etapa de exploração de dados precisaremos lembrar do *Passo 1: Entendimento do Problema*. Nessa fase, se intensifica a necessidade de habilidades analíticas e criativas para pensar em ideias e hipóteses a serem validadas. É importante que você busque identificar padrões interessantes nos seus dados (lembra do estudo da estudo famoso da fralda e da cerveja?!).

Existem várias ferramentas e bibliotecas em várias linguagens de programação que podem auxiliar nessa etapa. Dá uma conferida abaixo:

Ferramentas open-source:

- Jupyter Notebook
- Metabase
- Weka
- R Shiny

Ferramentas gratuitas:

- Power BI Desktop
- Qlik Sense Desktop
- Tableau Desktop

Bibliotecas:

- Pandas
- NLTK

5. Análise profunda de dados

Do inglês *Perform in-depth analysis,* essa etapa pode não estar presente em todos os projetos de Data Science. É nessa fase que realizamos passos de seleção de *features*, implementamos e aplicamos modelos estatísticos e de *Machine Learning* para validar hipóteses.

Suponha que seu problema seja criar um modelo preditivo que identifique quando um cliente deixará sua plataforma. Precisaremos então criar um modelo mais complexo e automatizados para que os *stakeholders* tenham informações suficientes para tomar decisões sobre possíveis *churns*.

"- Mas porque é uma etapa as vezes ausente?" Porque alguns problemas são resolvidos na etapa anterior, de exploração de dados, e não exigem a aplicação de uma metodologia mais complexa para sua resolução.

Para nos auxiliar nessa etapa, contamos com bibliotecas como a scikit-learn, que encapsula vários modelos de Classificação, Regressão, Clusterização, Redução de dimensionalidade e etc. Além disso, algumas ferramentas estão começando a disponibilizar plugins para análises mais complexas, como a opção de forecasting do Power BI.

6. Comunicação de resultados e Feedback

O ciclo parece completo, certo? Ainda não…

O sucesso de um projeto de Data Science (como qualquer outro) depende da comunicação efetiva dos resultados, que darão suporte ao processo de tomada de decisão no negócio. Para isso, use e abuse do *storytelling*, que nada mais é do que a capacidade de contar boas histórias para seus *stakeholders*, mostrando como os resultados podem ajudá-los decidir na tomada de decisões.

Além disso, é importante também que o Cientista de Dados se preocupe com a atualização dos resultados, de acordo com a volatilidade do negócio. Por isso é crucial que todo o desenvolvimento do projeto seja elaborado de forma reproduzível.

CRÉDITO: *Pollyanna Gonçalves / . Disponível em https://medium.com/techbloghotmart/afinal-como-se-desenvolve-um-projeto-de-data-science-233472996c34*

UM POUCO MAIS SOBRE A ASSUNTO:

Algumas etapas podem ser adicionadas ou removidas dependendo da necessidade de cada negócio. Assim, vamos apresentar um pouco mais a respeito do assunto.

Data science ou **ciência de dados**, é uma forma de estatística aplicada que incorpora elementos de ciências da computação e matemática para extrair insights de dados quantitativos e qualitativos.

As ferramentas e tecnologias utilizadas em data science incluem algoritmos e estruturas de machine learning, assim como linguagens de programação e bibliotecas de visualização.

Um cientista de dados combina programação, matemática e conhecimentos de domínio para responder perguntas através dos dados.

CICLO DE VIDA DE *DATA SCIENCE*

A ciência de dados é um processo cíclico. O ciclo de vida pode ser dividido nas seguintes **etapas:**

Conhecimento do tópico: para começar, um data scientist precisa ter uma compreensão básica do tópico ou problema que está tentando explorar para que possa fazer perguntas significativas. A natureza da ciência de dados é buscar explicações sobre por que as coisas são como elas são. Uma base de conhecimento especializada define a necessidade de um projeto de data science e garante decisões mais confiantes e orientadas por dados.

Aquisição(obtenção de dados): a próxima etapa do ciclo de vida da ciência de dados é coletar os dados certos para ajudar a responder à pergunta definida. Os dados podem residir em uma variedade de lugares ou serem difíceis de acessar, dependendo do conjunto de habilidades de uma pessoa. Mas o sucesso do restante do processo de data science depende da qualidade das informações coletadas nesta etapa e de como elas são preparadas

Data preparation: o preparo de dados é a etapa mais demorada e, provavelmente, a mais importante do ciclo de ciência de dados. A qualidade do seu resultado depende das informações que você insere. Os dados precisam ser devidamente limpos e combinados antes da análise. Isso pode incluir a integração de fontes diferentes, o tratamento de valores ausentes, outliers e muito mais. Durante essa etapa iterativa, um cientista de dados pode perceber que precisa voltar e coletar mais informações.

Data exploration: a exploração de dados envolve a identificação e compreensão de padrões em um conjunto de informações. Uma vez que os dados estejam limpos e utilizáveis, os cientistas de dados podem passar algum tempo entendendo as informações e formando hipóteses para testar. Esta é outra etapa em um processo iterativo, e um cientista de dados pode precisar de uma ou duas etapas para executar limpeza e blend adicionais com base em suas descobertas. Essa prática inclui a revisão dos atributos distintos de cada ponto ou recurso no conjunto de dados e a determinação se outras combinações e transformações de dados produziriam novos recursos potencialmente significativos. O processo de criação de novos recursos nos dados é chamado de engenharia de recursos. Geralmente ocorre na interação entre as etapas de exploração e preparo das informações.

Modelagem e avaliação preditiva: após a exploração, um cientista de dados pode começar a treinar modelos preditivos. A modelagem preditiva muitas vezes pode ser combinada com a exploração de dados. Assim que a modelagem e a avaliação começarem, é provável que um cientista de dados perceba novas coisas sobre os recursos e volte a iterar sobre a engenharia de recursos. À medida que os modelos são construídos, eles precisam ser avaliados. Um cientista de dados deve continuar a testar e refinar modelos até que eles encontrem um com o qual estão satisfeitos.

Interpretação e implantação: o resultado de todo esse trabalho pode ser uma interpretação dos dados e resultados, no qual o cientista de dados utiliza o modelo e toda a análise que realizou durante o ciclo de vida para responder à pergunta com a qual começou. Outro resultado pode ser que o modelo seja destinado à implantação, na qual será utilizado para ajudar os interessados a tomarem decisões orientadas por dados ou automatizar um processo (se esse for o seu resultado, não se esqueça da próxima etapa, o monitoramento).

Monitoramento: depois que o modelo é implantado, ele precisa ser verificado e mantido para que possa continuar funcionando corretamente mesmo quando receber novos dados. Os modelos precisam ser monitorados para quando os dados forem alterados devido à mudanças no comportamento ou outros fatores, os ajustes possam ser feitos de acordo.

Repetição: o ciclo se repete, quer o objetivo final seja a interpretação imediata ou implantação a longo prazo. O resultado de qualquer projeto de data science deve ser aprender algo novo sobre o tópico ou problema que está sendo explorado, o que leva a perguntas novas e mais profundas.

Fonte: https://www.alteryx.com/pt-br/glossary/data-science

PAPEIS DOS ENVOLVIDOS EM PROJETOS DE CIÊNCIA DE DADOS E BIG DATA.

Quando falamos sobre projetos de dados, é normal focarmos a discussão em tecnologias, frameworks, fornecedores ou tendências.

Apesar da clara importância desses fatores, para que um projeto de dados tenha sucesso é

essencial que ele não negligencie o seu principal ativo: **as pessoas**.

Um time multidisciplinar de profissionais qualificados é imprescindível quando estamos falando de desenvolver, implementar e operacionalizar uma pipeline de dados com eficiência.

Mesmo com bom planejamento e objetivos claros, a execução ficará por parte do time dedicado, e por isso a sua estrutura também precisa receber a devida atenção. Isso garante que todas as pessoas necessárias estarão envolvidas e os seus conhecimentos se complementarão de modo a obtermos os melhores resultados.

Por isso, agora veremos quais são **alguns dos principais papéis dentro de um time de dados** e suas responsabilidades.

Engenheiro de dados

O engenheiro de dados é o responsável pela estruturação, operacionalização e monitoramento da pipeline de dados, garantindo o fluxo que levará os dados de suas fontes até os consumidores.

Essa função é essencial pois vai integrar diversas soluções de mercado para que os dados sejam extraídos das origens, passem pelas transformações necessárias para serem analisados e fiquem disponíveis nos repositórios corretos.

Algumas das habilidades comuns de um engenheiro de dados são: Python, Spark e SQL; desenvolvimento de pipelines batch e stream; conhecimentos das principais ferramentas de processamento e armazenamento de dados na nuvem.

E pensando em visibilidade de mercado, é interessante que um profissional dessa função também tenha as certificações oficiais das plataformas com as quais trabalha, como o Google Cloud Professional Data Engineer ou Azure Data Engineer Associate, por exemplo.

Arquiteto de dados

Apesar de haver certa sobreposição de conhecimentos entre o engenheiro e o arquiteto de dados, pois ambos atuam diretamente com a infraestrutura de Big Data, esse último tem a atuação mais focada nas áreas de planejamento e governança.

O arquiteto vai ser responsável pelo planejamento de todo ambiente onde se dará o processamento dos dados. Ele estrutura os processos de dados, avalia quais ferramentas serão utilizadas, quais serão os pontos de monitoramento, define práticas de governança e estabelece protocolos de segurança e integração.

Os maiores fornecedores de nuvem, como Google, AWS e Cloudera, também possuem certificações específicas para a função de arquiteto de dados.

Analista de dados

Responsável por criar relatórios e visualizações que serão consumidas nos processos decisórios da empresa, o analista de dados converte os dados em insights relevantes. É aqui que o real valor dos dados vem à tona.

Para isso, o profissional utiliza de ferramentas de BI (QlikSense, Power BI, Tableau, etc.) para sintetizar grandes quantidades de dados e criar representações gráficas que poderão ser facilmente consultadas pelos usuários finais.

Cientista de dados

Assim como o analista de dados, o cientista de dados visa extrair insights e responder perguntas de negócios, porém utilizando tecnologias como Machine Learning, Deep Learning e algoritmos para realizar análises exploratórias, preditivas e identificação de padrões.

Outro diferencial do cientista de dados é utilizar uma gama maior de dados, e não apenas dados tratados ou estruturados. Muitas vezes, a matéria prima para a ciência de dados serão dados brutos (ainda na camada de Raw Data dos Data Lakes), nos quais o cientista aplicará uma modelagem diferenciada para poder gerar análises mais complexas.

Portanto, esse profissional acaba contando com habilidades que vão desde conhecimentos em estatísticas, criação de bases de dados a programação de algoritmos para análise. Para tanto, faz uso de linguagens de programação como Python, R e SQL, além de precisar de domínio em Git, Jupyter Notebook e Airflow.

Nem todos os times precisarão contar com todas essas funções, e projetos diferenciados podem demandar profissionais com dedicação mais específica ou restritiva. Porém, os papéis listados acima são o cerne de um time de dados para a maioria dos projetos.

Atentar a estrutura da equipe, alinhamento dessa estrutura com o objetivo do projeto, e a alta qualificação dos profissionais são etapas essenciais de um planejamento de Big Data eficiente.

_____

Fonte: https://medium.com/datalakers-blog/os-principais-pap%C3%A9is-em-um-time-de-dados-2161f85751aa

COMPUTAÇÃO EM NUVENS.

DEFINIDO POR COMPUTAÇÃO EM NUVEM

Em termos simples, a computação em nuvem permite alugar, em vez de comprar sua TI. Em vez de investir pesado em bancos de dados, software e hardware, as empresas optam por acessar seu poder de computação pela internet, ou pela nuvem e pagar por isso enquanto o usam. Esses serviços de nuvem agora incluem, mas não se limitam a, servidores, armazenamento, bancos de dados, rede, software, análise avançada e business intelligence.

A computação em nuvem fornece velocidade, escalabilidade e flexibilidade que permitem que as empresas desenvolvam, inovem e suportem soluções de TI empresariais.

Noções básicas de computação em nuvem

Quando uma empresa escolhe "migrar para a nuvem", isso significa que sua infraestrutura de TI é armazenada externamente, em um data center mantido pelo provedor de computação em nuvem. Um provedor de nuvem líder do setor tem a responsabilidade de gerenciar a infraestrutura de TI do cliente, integrar aplicativos e desenvolver novos recursos e funcionalidades para acompanhar as demandas do mercado.

Para os clientes, a computação em nuvem oferece mais agilidade, escala e flexibilidade. Em vez de gastar dinheiro e recursos em sistemas de TI legados, os clientes podem se concentrar em tarefas mais estratégicas. Sem fazer um grande investimento antecipado, eles podem acessar rapidamente os recursos de computação que necessitam - e pagar apenas pelo que usam.

Benefícios da computação em nuvem

Existem várias tendências impulsionando os negócios - em todos os setores - para a nuvem. Para a maioria das organizações, a maneira atual de fazer negócios pode não fornecer a agilidade para crescer ou não fornecer a plataforma ou flexibilidade para competir. A explosão de dados criados por um número crescente de empresas digitais está elevando o custo e a complexidade do armazenamento do data center para novos níveis – exigindo novas ferramentas de habilidades e análises da TI.

Modernas soluções em nuvem ajudam as empresas a enfrentar os desafios da era digital. Em vez de gerenciar sua TI, as organizações têm a capacidade de responder rapidamente a um cenário de negócios mais rápido e complexo. Com a economia moderna da nuvem, a nuvem agrega valor e reduz custos, ajudando as empresas a alcançar todo o seu potencial de negócios com seus gastos com a nuvem.

A computação em nuvem oferece uma alternativa superior à tecnologia da informação tradicional, incluindo estas áreas:

- Custo - elimine despesas de capital
- Velocidade - provisione espaço de forma rápida para desenvolvimento e teste
- Escala global - escale elasticamente
- Produtividade - maior colaboração, desempenho previsível e isolamento do cliente
- Desempenho - preço/desempenho melhores para cargas de trabalho nativas da nuvem
- Confiabilidade - sistemas distribuídos tolerantes a falhas, escaláveis em todos os serviços

TIPOS DE COMPUTAÇÃO EM NUVEM

Há três tipos de nuvens: pública, privada e híbrida. Cada tipo requer um nível diferente de gerenciamento do cliente e fornece um nível diferente de segurança.

Nuvem pública

Em uma nuvem pública, toda a infraestrutura de computação está localizada nas instalações do provedor de nuvem, e o provedor fornece serviços ao cliente pela Internet. Os clientes não precisam manter sua própria TI e podem adicionar rapidamente mais usuários ou capacidade de computação, conforme necessário. Nesse modelo, vários locatários compartilham a infraestrutura de TI do provedor de nuvem.

Nuvem privada

Uma nuvem privada é usada exclusivamente por uma organização. Pode ser hospedada no local da organização ou no data center do provedor de nuvem. Uma nuvem privada fornece o mais alto nível de segurança e controle.

Nuvem híbrida

Como o nome sugere, uma nuvem híbrida é uma combinação de nuvens pública e privada. Geralmente, os clientes de nuvem híbrida hospedam seus aplicativos essenciais para negócios em seus próprios servidores para mais segurança e controle, e armazenam seus aplicativos secundários no local do provedor de nuvem.

Multicloud

A principal diferença entre nuvem híbrida e multicloud é o uso de diversos dispositivos de armazenamento e computação em nuvem em uma única arquitetura.

SERVIÇOS DE COMPUTAÇÃO EM NUVEM

Existem três tipos principais de serviços em nuvem: software como serviço (SaaS), plataforma como serviço (PaaS) e infraestrutura como serviço (IaaS). Não existe uma abordagem única para a nuvem; é mais sobre encontrar a solução certa para dar suporte aos seus requisitos de negócios.

SaaS

O SaaS é um modelo de entrega de software em que o provedor de nuvem hospeda os aplicativos do cliente no local do provedor de nuvem. O cliente acessa esses aplicativos pela Internet. Em vez de pagar e manter sua própria infraestrutura de computação, os clientes de SaaS aproveitam a assinatura do serviço em uma base de pagamento conforme o uso.

Muitas empresas consideram que o SaaS é a solução ideal, pois permite que eles comecem a trabalhar rapidamente com a tecnologia mais inovadora disponível. As atualizações automáticas reduzem o ônus sobre os recursos internos. Os clientes podem dimensionar os serviços para suportar cargas de trabalho flutuantes, adicionando mais serviços ou recursos que eles desenvolvam. Um pacote de nuvem moderno fornece software completo para cada necessidade de negócios, incluindo experiência do cliente, gerenciamento de relacionamento com clientes, serviço de atendimento ao cliente, planejamento de recursos empresariais,

compras, gerenciamento financeiro, gerenciamento de capital humano, gerenciamento de talentos, folha de pagamentos, gerenciamento da cadeia de suprimentos, planejamento corporativo e muito mais.

PaaS

O PaaS oferece aos clientes a vantagem de acessar as ferramentas de desenvolvedor necessárias para criar e gerenciar aplicativos móveis e da Web sem investir - ou manter - a infraestrutura subjacente. O provedor hospeda os componentes de infraestrutura e middleware e o cliente acessa esses serviços por meio de um navegador web.

Para ajudar na produtividade, as soluções de PaaS precisam ter componentes de programação prontos para uso que permitam aos desenvolvedores criar novos recursos em seus aplicativos, incluindo tecnologias inovadoras, como inteligência artificial (IA), chatbots, blockchain e a Internet das Coisas (IoT). A oferta de PaaS certa também deve incluir soluções para analistas, usuários finais e administradores de TI profissionais, incluindo análise avançada de big data, gerenciamento de conteúdo, gerenciamento de banco de dados, gerenciamento de sistemas e segurança.

IaaS

O IaaS permite que os clientes acessem serviços de infraestrutura sob demanda pela internet. A principal vantagem é que o provedor de nuvem hospeda os componentes de infraestrutura que fornecem computação, armazenamento e capacidade de rede para que os assinantes possam executar suas cargas de trabalho na nuvem. O assinante da nuvem é geralmente responsável por instalar, configurar, proteger e manter qualquer software nas soluções nativas da nuvem, como banco de dados, middleware e software de aplicativo.

ARQUITETURA DE BIG DATA.

Uma arquitetura de Big Data foi projetada para lidar com ingestão, processamento e análise de dados grandes ou complexos demais para sistemas de banco de dados tradicionais. O limite no qual as organizações ingressam no campo do Big Data é diferente, dependendo das capacidades dos usuários e de suas ferramentas. Para alguns, isso pode significar centenas de gigabytes de dados, enquanto para outros, centenas de terabytes. À medida que as ferramentas para o trabalho com conjuntos de Big Data evoluem, na mesma proporção evolui o significado de Big Data. Cada vez mais, esse termo se relaciona ao valor que é possível extrair dos conjuntos de dados por meio de análise avançada, em vez de estritamente o tamanho dos dados, embora nesses casos, eles tendam a ser muito grandes.

Ao longo dos anos, o cenário dos dados vem mudando. Houve uma mudança no que você pode fazer ou o que deve fazer, com os dados. O custo de armazenamento caiu drasticamente, enquanto os meios pelos quais os dados são coletados continuam aumentando. Alguns dados são recebidos a um ritmo rápido, constantemente exigindo sua coleta e observação. Outros dados são recebidos mais lentamente, mas em partes muito grandes, geralmente na forma de décadas de dados históricos. Talvez você esteja enfrentando um problema de análise avançada ou um problema que exija o aprendizado de máquina. Esses são desafios que as arquiteturas de Big Data buscam resolver.

Soluções de Big Data normalmente envolvem um ou mais dos seguintes **tipos de carga de trabalho**:

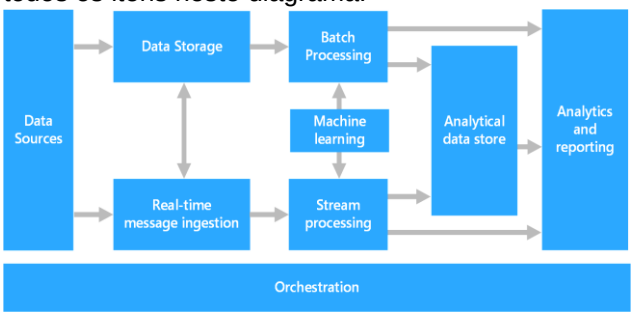
- Processamento em lote de fontes Big Data em repouso.
- Processamento em tempo real de Big Data em movimento.
- Exploração interativa de Big Data.
- Análise preditiva e machine learning.

Considere o uso das arquiteturas de Big Data quando precisar:

- Armazenar e processar dados em volumes muito grandes para um banco de dados tradicional.
- Transformar dados não estruturados para análise e relatório.
- Capturar, processar e analisar fluxos não associados de dados em tempo real ou com baixa latência.

Componentes de uma arquitetura de Big Data

O diagrama a seguir mostra os componentes lógicos que se inserem em uma arquitetura de Big Data. As soluções individuais podem não conter todos os itens neste diagrama.



A maioria das arquiteturas de Big Data inclui alguns ou todos os **seguintes componentes:**

- **Fontes de dados**. Todas as soluções de Big Data começam com uma ou mais fontes de dados. Os exemplos incluem:
  - Armazenamentos de dados de aplicativo, como bancos de dados relacionais.
  - Arquivos estáticos produzidos por aplicativos, como arquivos de log do servidor Web.
  - Fontes de dados em tempo real, como dispositivos IoT.
- **Armazenamento de dados**. Os dados de operações de processamento em lotes normalmente são armazenados em um repositório de arquivos distribuído que pode conter amplos volumes de arquivos grandes em vários formatos. Esse tipo de repositório geralmente é chamado *data lake*. As

opções para implementar esse armazenamento incluem contêineres de blobs ou Azure Data Lake Store no Armazenamento do Azure.

- **Processamento em lotes**. Como os conjuntos de dados são muito grandes, geralmente, uma solução de Big Data precisa processar arquivos de dados usando trabalhos em lotes de execução longa para filtrar, agregar e, de outro modo, preparar os dados para análise. Normalmente, esses trabalhos envolvem ler arquivos de origem, processá-los e gravar a saída para novos arquivos. Opções incluem executar trabalhos de U-SQL no Azure Data Lake Analytics, usar trabalhos Hive, Pig ou de Mapear/Reduzir personalizados em um cluster HDInsight Hadoop ou usar programas de Java, Scala ou Python em um cluster HDInsight Spark.
- **Ingestão de mensagens em tempo real**. Se a solução inclui fontes em tempo real, a arquitetura precisa incluir uma maneira de capturar e armazenar mensagens em tempo real para processamento de fluxo. Isso pode ser um armazenamento de dados simples, em que as mensagens de entrada são removidas para uma pasta para processamento. No entanto, muitas soluções precisam de um repositório de ingestão de mensagens para atuar como buffer de mensagens e dar suporte a processamento de expansão, entrega confiável e outras semânticas de enfileiramento de mensagem. Essa parte de uma arquitetura de streaming geralmente é conhecida como buffer de fluxo. Entre as opções estão os Hubs de Eventos do Azure, o Hub IoT do Azure e o Kafka.
- **Processamento de fluxo**. Depois de capturar mensagens em tempo real, a solução precisa processá-las filtrando, agregando e preparando os dados para análise. Os dados de fluxo processados são gravados em um coletor de saída. O Azure Stream Analytics oferece um serviço de processamento de fluxo gerenciado baseado em consultas SQL em execução perpétua que operam em fluxos não associados. Você também pode usar tecnologias de streaming Apache de software livre, como Storm e Spark Streaming em um cluster HDInsight.
- **Armazenamento de dados analíticos**. Muitas soluções de Big Data preparam dados para análise e então fornecem os dados processados em um formato estruturado que pode ser consultado com ferramentas analíticas. O armazenamento de dados analíticos usado para atender a essas consultas pode ser um data warehouse relacional estilo Kimball, como visto na maioria das soluções de BI (business intelligence) tradicionais. Como alternativa, os dados podem ser apresentados por meio de uma tecnologia NoSQL de baixa latência, como HBase ou um banco de dados Hive interativo que oferece uma abstração de metadados sobre arquivos de dados no armazenamento de dados distribuído. O Azure Synapse Analytics fornece um serviço gerenciado para armazenamento de dados em larga escala baseado em nuvem. O HDInsight dá suporte a Hive interativo, HBase e Spark SQL, que também pode ser usado para veicular dados para análise.
- **Análise e relatórios**. A meta da maioria das soluções de Big Data é gerar insights sobre os dados por meio de análise e relatórios. Para capacitar os usuários a analisar os dados, a arquitetura pode incluir uma camada de modelagem de dados, como um cubo OLAP multidimensional ou um modelo de dados tabular no Azure Analysis Services. Também pode dar suporte a business intelligence de autoatendimento, usando as tecnologias de modelagem e visualização do Microsoft Power BI ou do Microsoft Excel. Análise e relatórios também podem assumir a forma de exploração de dados interativos por cientistas de dados ou analistas de dados. Para esses cenários, muitos serviços do Azure dão suporte a blocos de anotações analíticos, como Jupyter, permitindo que esses usuários aproveitem suas habilidades existentes com Python ou R. Para exploração de dados em larga escala, você pode usar o Microsoft R Server, seja no modo autônomo ou com Spark.
- **Orquestração**. A maioria das soluções de Big Data consiste em operações de processamento de dados repetidas, encapsuladas em fluxos de trabalho, que transformam dados de origem, movem dados entre várias origens e coletores, carregam os dados processados em um armazenamento de dados analíticos ou enviam os resultados por push diretamente para um relatório ou painel. Para automatizar esses fluxos de trabalho, você pode usar uma tecnologia de orquestração, como Azure Data Factory ou Apache Oozie e Sqoop.

Arquitetura Lambda

Ao trabalhar com conjuntos de dados muito grandes, pode levar muito tempo para executar a classificação de consultas de que os clientes precisam. Essas consultas não podem ser executadas em tempo real e geralmente exigem algoritmos como MapReduce, que operam em paralelo em todo o conjunto de dados. Os resultados são então armazenados separadamente dos dados brutos e usados para consulta.

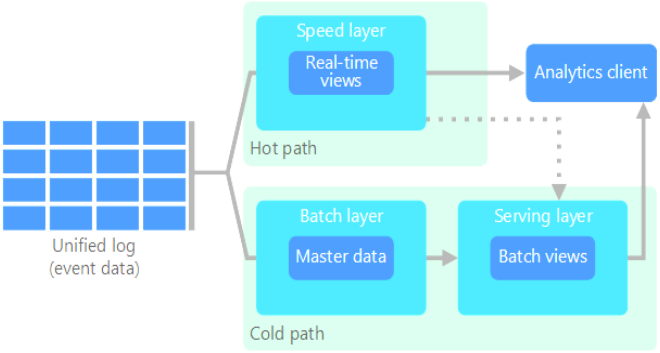
Uma desvantagem dessa abordagem é que ela introduz latências: se o processamento levar algumas horas, uma consulta poderá retornar resultados de várias horas atrás. O ideal é que você obtenha alguns resultados em tempo real (talvez com alguma perda de precisão) e combine esses resultados com os resultados da análise de lote.

A **arquitetura lambda**, primeiramente proposta por Nathan Marz, resolve esse problema criando dois caminhos para o fluxo de dados. Todos os dados recebidos pelo sistema passam por esses dois caminhos:

- Uma **camada de lote** (caminho frio) armazena todos os dados de entrada em sua forma bruta e executa o processamento em lotes nos dados. O resultado desse processamento é armazenado como uma **exibição de lote**.
- Uma **camada de velocidade** (caminho quente) analisa os dados em tempo real. Essa camada foi projetada para baixa latência, em detrimento da precisão.

A camada de lote alimenta uma **camada de**

serviço que indexa a exibição de lote para uma consulta eficiente. A camada de velocidade atualiza a camada de serviço com atualizações incrementais de acordo com os dados mais recente



Os dados que fluem para o caminho quente são restritos por requisitos de latência impostos pela camada de velocidade, de modo que ela possa ser processada o mais rapidamente possível. Geralmente, isso exige uma desvantagem de algum nível de precisão em favor dos dados que estão prontos o mais rapidamente possível. Por exemplo, considere um cenário de IoT em que um grande número de sensores de temperatura envia dados telemétricos. A camada de velocidade pode ser usada para processar uma janela de tempo deslizante dos dados de entrada.

Os dados que fluem para o caminho frio, por outro lado, não estão sujeitos aos mesmos requisitos de baixa latência. Isso permite uma computação de alta precisão em conjuntos de dados grandes, o que pode ser muito demorado.

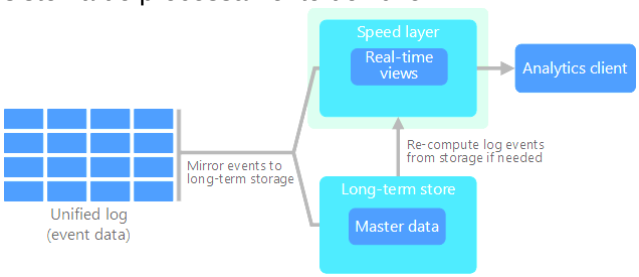
Em última análise, os caminhos quente e frio convergem no aplicativo cliente de análise. Se o cliente precisar exibir dados em tempo hábil, mas potencialmente menos precisos em tempo real, ele adquirirá seu resultado do caminho quente. Caso contrário, ele selecionará resultados do caminho frio para exibir dados em menos tempo hábil, mas mais precisos. Em outras palavras, o caminho quente contém dados para uma janela relativamente pequena de tempo, após o qual os resultados podem ser atualizados com os dados mais precisos do caminho frio.

Os dados brutos armazenados na camada de lote são imutáveis. Os dados de entrada sempre são acrescentados aos dados existentes e os dados anteriores nunca são substituídos. As alterações no valor de um dado específico são armazenadas como um novo registro de evento com carimbo de data/hora. Isso permite o recálculo em qualquer ponto no tempo no histórico dos dados coletados. A capacidade de recalcular a exibição de lote dos dados brutos originais é importante, pois permite que novas exibições sejam criadas conforme o sistema evolui.

Arquitetura Kappa

Uma desvantagem da arquitetura de lambda é sua complexidade. A lógica de processamento aparece em dois lugares diferentes (os caminhos frio e crítico) usando estruturas diferentes. Isso leva a uma lógica de cálculo duplicada e a complexidade de gerenciar a arquitetura para os dois caminhos.

A **arquitetura de kappa** foi proposta por Jay Kreps como uma alternativa à arquitetura de lambda. Ela tem as mesmas metas básicas da arquitetura de lambda, mas com uma diferença importante: todos os dados fluem por um único caminho, usando um sistema de processamento de fluxo.



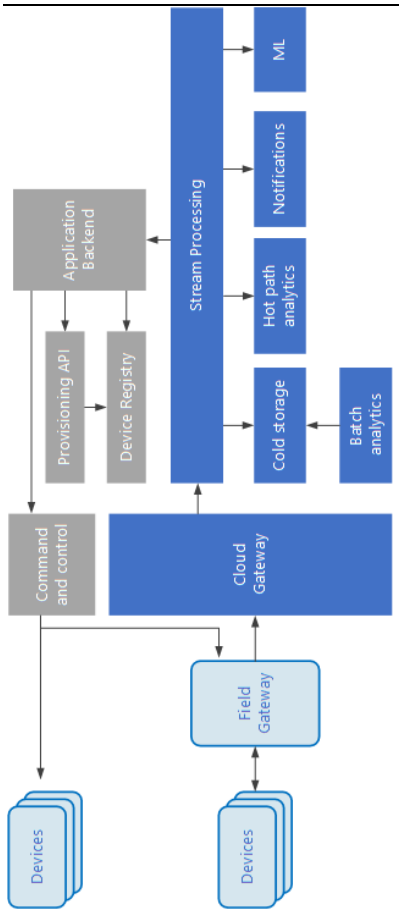
Há algumas semelhanças na camada de lote da arquitetura de lambda, em que os dados do evento são imutáveis e todos eles são coletados, em vez de um subconjunto. Os dados são ingeridos como um fluxo de eventos em um log unificado distribuído e tolerante a falhas. Esses eventos são ordenados e o estado atual de um evento é alterado somente por um novo evento que está sendo acrescentado. Semelhante à camada de velocidade da arquitetura de uma lambda, todo o processamento de eventos é feito no fluxo de entrada e persistido como uma exibição em tempo real.

Se você precisar recalcular todo o conjunto de dados (equivalente ao que a camada de lote faz no lambda), basta reproduzir o fluxo, normalmente usando o paralelismo para concluir o cálculo em tempo hábil.

Internet das coisas (IoT)

Do ponto de vista prático, a IoT (Internet das Coisas) representa qualquer dispositivo conectado à Internet. Isso inclui seu computador, telefone celular, relógio inteligente, termostato inteligente, refrigerador inteligente, automóvel conectado, implantes de monitoramento cardíaco e qualquer outra coisa que se conecta à Internet e envia ou recebe dados. O número de dispositivos conectados cresce diariamente, assim como a quantidade de dados coletados deles. Em geral, esses dados são coletados em ambientes altamente restritos, às vezes, de alta latência. Em outros casos, os dados são enviados de ambientes de baixa latência por milhares ou milhões de dispositivos, que necessitam da capacidade de ingerir os dados rapidamente e processá-lo de forma adequada. Portanto, um planejamento adequado é necessário para lidar com essas restrições e esses requisitos exclusivos.

Arquiteturas orientadas por eventos são essenciais para soluções de IoT. O diagrama a seguir mostra uma possível arquitetura lógica de IoT. O diagrama enfatiza os componentes da arquitetura do streaming de eventos.



O **gateway de nuvem** consome eventos de dispositivo no limite da nuvem, usando um sistema de mensagens de latência baixa e confiável.

Os dispositivos podem enviar eventos diretamente para o gateway de nuvem, ou por meio de um **gateway de campo**. Um gateway de campo é um software ou dispositivo especializado, geralmente colocado com os dispositivos, que recebe eventos e os encaminha para o gateway de nuvem. O gateway de campo também pode pré-processar os eventos de dispositivo brutos executando funções, como filtragem, agregação ou transformação de protocolo.

Após a ingestão, os eventos passam por um ou mais **processadores de fluxo** que podem encaminhar os dados (por exemplo, para armazenamento) ou executar análise e outros tipos de processamento.

A seguir estão alguns tipos comuns de processamento. (Esta lista certamente não é exaustiva.)

- Gravando os dados de evento para armazenamento menos acessado, para arquivamento ou análise de processo em lote.
- Análise de caminho mais acessado, analisando o fluxo de eventos (quase) em tempo real, para detectar anomalias, reconhecer padrões em janelas de tempo ou disparar alertas quando ocorre uma condição específica no fluxo.
- Tratamento de tipos especiais de mensagens que não são de telemetria de dispositivos, como notificações e alarmes.
- Machine Learning.

As caixas destacadas em cinza mostram os componentes de um sistema de IoT que não estão diretamente relacionadas ao streaming de evento, mas são incluídos aqui para fins de integridade.

- O **registro do dispositivo** é um banco de dados dos dispositivos provisionados, incluindo os IDs de dispositivo e metadados do dispositivo, como localização.
- A **API de provisionamento** é uma interface externa comum para provisionar e registrar dispositivos novos.
- Algumas soluções IoT permitem que **mensagens de comando e controle** sejam enviadas aos dispositivos.

_____

CRÉDITO DO TEXTO: Zoiner Tejada. In https://learn.microsoft.com/pt-br/azure/architecture/data-guide/big-data/

MODELOS DE ENTREGA E DISTRIBUIÇÃO DE SERVIÇOS DE BIG DATA.

O QUE É BIG DATA?

Big Data é um conceito ligado ao armazenamento e processamento de uma quantidade expressiva de dados, que podem ser coletados ao longo de todos os processos internos e externos de uma empresa.

O Big Data na logística pode lidar com informações geradas desde a emissão de pedidos até a consolidação de entregas. A ideia por trás desse estudo de dados é proporcionar conhecimentos e *insights* que colaborem para que as tomadas de decisão sejam mais estratégicas e para otimizar a cadeia como um todo.

COMO UTILIZAR O BIG DATA NAS OPERAÇÕES LOGÍSTICAS?

O Big Data consiste na análise e na compreensão de grandes volumes de dados em muitas variedades. Para tanto, é preciso utilizar soluções próprias da ferramenta, que possibilitam aos profissionais de TI trabalharem com conteúdos não estruturais em grande velocidade.

A aplicação desse recurso na logística pode trazer diversas vantagens no que se refere à mineração de informações. Isso se torna possível devido à quantidade abundante de dados que são gerados pelo fluxo de produtos e prestações de serviços.

Por meio dele, é possível reduzir os custos, aumentar a produtividade, tomar decisões acertadas e implantar ações que são cruciais para a atuação da empresa. Quer entender como o Big Data pode ser aplicado no negócio? Continue a leitura!

Capacidade de prever as procuras sazonais

Uma das principais finalidades do gerenciamento de estoque é conseguir equilibrar as quantidades disponíveis com a busca dos clientes, ao mesmo tempo em que as mercadorias são

mantidas nos níveis mais baixos possíveis. Assim, a empresa adquire eficiência e diminui seus gastos.

O objetivo é estocar uma pequena quantidade, mas que seja suficiente para assegurar o atendimento da demanda, evitando faltas ou excessos. O controle operacional por meio do Big Data possibilita controlar fluxos mais críticos, como a obrigação de reposição e a perda de produtos, além de apontar quais itens têm maior procura em certos momentos do ano.

Assim, torna-se viável elevar a previsibilidade a respeito de aumentos sazonais e da necessidade de guardar quantias maiores. O principal benefício de tal prática é a probabilidade de centros de distribuição se prevenirem para o crescimento da movimentação e impedirem a falta de mercadorias no estoque.

Idealização das redes de distribuição

É possível alcançar vantagens na projeção que envolve esse setor. Com base nos dados relacionados a todas as entregas que foram feitas e na localização dos clientes, você pode identificar padrões de procura e as melhores redes para atender ao fluxo de envios.

Assim, o gestor utiliza o entendimento gerado para:

- projetar novas redes de distribuição;
- implantar ações que colaborem com o aumento da rapidez dos serviços;
- diminuir custos operacionais;
- encontrar meios de conciliar as metas organizacionais com o atendimento dos desejos dos consumidores.

Aumento da eficácia operacional

O Big Data possibilita avaliar uma grande quantidade de dados. Isso viabiliza a geração de conhecimento a respeito dos mais variados aspectos da operação.

Dessa forma, é possível atingir a eficácia operacional, à medida que o gestor adquire uma base sólida que auxilia no aprimoramento dos processos e da execução. Como consequência, há um aumento da qualidade e a diminuição de falhas ou retrabalhos, entre outras vantagens que garantem um trabalho mais efetivo.

Aperfeiçoamento das entregas *last mile* e redução de custos

As entregas *last mile* (última milha) representam grandes desafios ao gerenciamento e à gestão. O problema se encontra em duas questões cruciais: viabilidade do transporte e gastos para levar o produto até o consumidor final.

Contudo, é possível aproveitar a implementação do Big Data na logística e utilizar a força desse recurso para compreender os padrões de entrega nas regiões, promovendo uma maior eficácia da frota.

Isso se torna ainda mais importante quando os dados gerados são reunidos por intermédio de um sistema de roteirização. São concedidas informações relevantes para o consumo de combustível e a efetividade da distribuição.

Então, o gestor consegue entender a necessidade de diferenciar a frota e melhorar a criação dos percursos, por exemplo, sempre focando na qualidade e na satisfação dos clientes.

O cruzamento de informações dos processos permite o alcance de uma visão sobre as atividades e os gargalos que devem ser corrigidos, as chances de melhoria e o foco no ganho de eficiência. Com todas as alterações que podem ser aplicadas, é seguro dizer que se torna viável ter uma logística diligente e com menos custos operacionais.

Obtenção dos melhores dados sobre a cadeia de abastecimento

No passado, muitas companhias não atentavam ao gerenciamento de fornecedores e transportadoras — que não acompanhavam o atacadista que, por sua vez, também não compreendia o funcionamento das movimentações varejistas. Atualmente, a logística cresceu e está cada vez mais em evidência.

É por meio dessa administração da cadeia de abastecimento que se atinge a integração de todos os membros envolvidos e são determinadas verdadeiras relações de parceria. Dessa forma, todos adquirem uma visão mais ampla em relação ao fluxo de ponta a ponta, com dados fornecidos pela equipe: a partir do fornecedor primário até a transportadora que faz a entrega ao cliente final.

A exploração das informações por intermédio do Big Data possibilita que mais conteúdos sejam divididos. Isso abre um espaço ainda maior para possibilidades de avaliações voltadas ao aperfeiçoamento dos processos e ao repasse de dados relevantes sobre os pedidos dos clientes.

Caso aconteça algum problema capaz de ocasionar atrasos ao fornecedor e influenciar a entrega, por exemplo, o cliente atingido poderá receber essas atualizações e ter uma expectativa real. Mesmo ocorrendo um transtorno, a empresa consegue repassar a ideia de transparência, gerando uma confiabilidade.

Aperfeiçoamento da experiência do consumidor

A experiência que o consumidor adquire com a companhia pode ser um fator decisivo para o êxito do negócio no mercado. Buscar a diferenciação no atendimento e satisfazer as necessidades dos clientes são cuidados que representam a fidelização à marca e, provavelmente, a captação de um novo público.

Lembre-se de que a indicação positiva de quem já foi cliente pode motivar outras pessoas a conhecerem os produtos e serviços oferecidos. Por todos esses motivos, a aplicação do Big data na logística propicia um atendimento de qualidade, acima dos oferecidos pela concorrência.

Além disso, as informações podem ser

compartilhadas e usadas. Como resultado, há uma contribuição significativa para a melhoria dos processos empresariais e o aumento da satisfação dos clientes.

São visíveis as vantagens geradas pelo Big Data às empresas. Elas vão desde pontos operacionais até tomadas de decisões mais corretas, auxiliando na melhoria da gestão e no crescimento do negócio em geral, além de trazer resultados financeiros satisfatórios.

_____

Fonte: https://esales.com.br/blog/como-utilizar-o-big-data-nas-operacoes-logisticas/

PLATAFORMAS DE COMPUTAÇÃO EM NUVEM PARA BIG DATA.

COMPUTAÇÃO EM NUVEM: AS MELHORES PLATAFORMAS E APLICATIVOS DE CÓDIGO ABERTO

PLATAFORMAS DE COMPUTAÇÃO EM NUVEM

Entre as **Plataformas de "computação em nuvem"** o **Computação em Nuvem**, e de **de código aberto**, podemos mencionar e descrever o seguinte 4:

1) OpenStack

É um Sistema Operacional em nuvem que controla grandes grupos de recursos de computação, armazenamento e rede em um data center inteiro, todos gerenciados e provisionados por meio de APIs com mecanismos de autenticação comuns. Também possui um painel de controle que permite aos administradores controlar e facilitar o provisionamento de recursos para seus usuários por meio de uma interface web. Além da funcionalidade padrão de infraestrutura como serviço, existem componentes adicionais que fornecem orquestração, gerenciamento de falhas e gerenciamento de serviços, entre outros serviços, para garantir a alta disponibilidade dos aplicativos do usuário.

2) Fundição em Nuvem

É uma plataforma como serviço (PaaS) aberta que fornece um modelo altamente eficiente e moderno para o fornecimento de aplicativos nativos da nuvem no Kubernetes. Além disso, oferece uma seleção de nuvens, estruturas de desenvolvedor e serviços de aplicativo. Isso torna mais rápido e fácil construir, testar, implantar e dimensionar aplicativos.

3) OpenShift

É uma plataforma de contêiner Kubernetes corporativa com operações automatizadas de ponta a ponta, permitindo que você gerencie nuvem híbrida, multicloud e implantações de computação de ponta. Esta solução da empresa Red Hat é otimizada para melhorar a produtividade do desenvolvedor e promover a inovação. E com operações automatizadas de ponta a ponta, uma experiência consistente em todos os ambientes e implantação de autoatendimento para desenvolvedores, as equipes podem trabalhar juntas para mover as ideias do desenvolvimento para a produção com mais eficiência.

4) Nublar

É uma plataforma de orquestração de borda e multi-nuvem de código aberto. O que, entre outras coisas, permite que as organizações façam a transição sem esforço para a nuvem pública e a arquitetura nativa da nuvem, permitindo que automatizem sua infraestrutura existente junto com recursos de borda distribuídos e nativos da nuvem. Além disso, ele também permite que os usuários gerenciem diferentes domínios de automação e orquestração como parte de um pipeline comum de CI / CD.

Outros 13 existentes e conhecidos são:

1. **Cloud Alibaba**
2. **Apache Mesos**
3. **AppScale**
4. **Cloudstack**
5. **Nuvem FOSS**
6. **Eucalipto**
7. **OpenNebula**
8. **Origem OpenShift / OKD**
9. **stackato**
10. **Synnefo**
11. **Tsuru**
12. **VirtEngineGenericName**
13. **WSO2**

Aplicativos de computação em nuvem

Entre as **Aplicações** relacionado ou aplicável ao **Domínio de TI** De **"Cloud Computing"** o **Computação em Nuvem**, e de **de código aberto**, podemos citar os seguintes 10:

1. **Ao ar livre**
2. **Bacula**
3. **GridGrain**
4. **Hadoop**
5. **Nagios**
6. **Odoo**
7. **OwnCloud**
8. **Xen**
9. **Zabbix**
10. **Zimbra**

_____

Fonte: https://blog.desdelinux.net/pt/cloud-computing-plataformas-aplicaciones-codigo-abierto-actuales/

LINGUAGENS DE PROGRAMAÇÃO PARA CIÊNCIA DE DADOS: LINGUAGEM PYTHON E R.

LINGUAGEM PYTHON

O que é Python?

Python é uma linguagem de propósito geral, ou seja, pode ser utilizada para as mais diversas aplicações. É gratuita Open Source e foi projetada tendo como um dos principais objetivos ser de fácil leitura e utilização.

Definindo com termos mais técnicos, Python é interpretada, orientada a objetos, funcional, tipada, imperativa e de script. Vamos entender um pouco melhor alguns princípios da linguagem e onde ela pode ser utilizada.

É comum ouvirmos a expressão de que "*programar em Python é como escrever uma carta em inglês para o computador*", pois a linguagem tenta utilizar comandos intuitivos, como "print" para imprimir um texto na tela, "open" para abrir um arquivo, ou "find" para encontrar a posição de uma palavra.

A linguagem de programação Python foi desenvolvida sob 19 princípios, são eles:

1. Bonito é melhor do que feio.
2. Explícito é melhor do que implícito.
3. Simples é melhor do que complexo.
4. Complexo é melhor do que complicado.
5. Horizontal é melhor do que aninhado.
6. Esparso é melhor que denso.
7. A legibilidade conta.
8. Casos especiais não são especiais o suficiente para quebrar as regras.
9. Porém, a praticidade supera a pureza.
10. Os erros nunca devem passar silenciosamente.
11. A menos que sejam explicitamente silenciados.
12. Diante da ambiguidade, recuse a tentação de adivinhar.
13. Deve haver uma, e de preferência apenas uma, forma óbvia de se fazer algo.
14. Embora essa forma possa não ser óbvia no início, a menos que você seja holandês.
15. Agora é melhor do que nunca.
16. Mas "nunca" é melhor do que "imediatamente agora".
17. Se a implementação é difícil de explicar, é uma má ideia.
18. Se a implementação for fácil de explicar, pode ser uma boa ideia.
19. Namespaces são uma ótima ideia – vamos fazer mais disso!

Python requer menos código

A quantidade de código necessária para executar funções em programação Python tipicamente é 3 ou até 5 vezes menor do que os códigos feitos em Java, e entre 5 e 10 vezes menor do que códigos em C++.

Muitas bibliotecas prontas para uso imediato

Como Python é uma das linguagens mais ativas em termos de comunidade, a cada dia novas bibliotecas são construídas e aprimoradas. Existem funções e módulos prontos para se executar de tudo, desde manipulações em imagens até algoritmos de inteligência artificial.

Isso é muito conveniente porque um programador iniciante acaba conseguindo obter recursos e resultados avançados apenas importando e utilizando módulos prontos, sem precisar criar tudo do zero.

Em outras palavras, a programação python é diferenciada pela riqueza de bibliotecas e frameworks prontos para utilização, bem como pelo suporte da comunidade.

O fato de existirem bibliotecas robustas também permite que um programador se especialize em uma tarefa específica, por exemplo: "manipulação de tabelas e datasets" para ciência de dados. Nesse caso, bastaria estudar e dominar a biblioteca Pandas.

Modularização para frameworks e recursos complexos

Python é tão flexível e modular que permite a utilização de diferentes recursos em um mesmo bloco de código.

Por exemplo, o framework TensorFlow (utilizado para computação numérica e inteligência artificial) pode rodar códigos utilizando CPU, GPU, ou ambas ao mesmo tempo, tudo em um mesmo ambiente com código Python que pode estar importando cumulativamente outras bibliotecas e pacotes, sem conflitos.

Multiplataforma

A linguagem Python permite que códigos sejam endereçados para os mais variados ambientes, como aplicações mobile, desenvolvimento web, desktop, games, etc.

Para que serve Python? O que é possível fazer com Python?

Agora que já vimos os benefícios da linguagem, chegou a hora de ter uma visão mais abrangente sobre o que é possível fazer com Python.

Python serve para:

- **Automatizar tarefas repetitivas**, criando códigos que interagem com seu sistema operacional;
- **Varrer a internet (web scraping)** navegando por sites, coletando, organizando e salvando informações;

- **Monitorar e minerar redes sociais**, conectando-se diretamente via APIs que facilitam a extração de dados;
- **Construir um site ou uma aplicação para a web**;
- **Construir um aplicativo mobile**;
- **Criar aplicações em blockchain** (diversos projetos descentralizados já possuem suporte para Python);
- **Criar jogos**;
- **Manipular grandes conjuntos de textos** com as mais avançadas tecnologias (processamento de linguagem natural);
- **Criar gráficos para BI** (Business Intelligence);
- **Criar ferramentas de Analytics** para tomadas de decisão;
- **Manipular dados de forma avançada**, com todos os recursos que um cientista de dados poderia necessitar;
- **Rodar algoritmos de machine learning**, tendo acesso a tudo que há de mais avançado na área;
- **Criar aplicações de inteligência artificial**, utilizando deep learning, reinforcement learning, entre outros;
- **Trabalhar com Big Data**;
- **Realizar trading** automatizado em bolsa de valores;
- **Fazer pesquisa científica e computação numérica**, tendo bibliotecas alternativas ao software Matlab;
- **Utilizar funções e módulos prontos** para engenharia, geologia, climatologia, entre outras áreas;
- **Programar microcontroladores e robôs**.

Empresas que utilizam Python

Como Python é tão versátil, é evidente que grandes empresas e corporações já utilizam a linguagem em diversas aplicações. Alguns exemplos são:

- **Instagram** (utiliza Django como backend, um framework Pyrhon para a web)
- **Google** (grande parte do algoritmo de busca é escrito em Python)
- **Spotify** (o aplicativo é construído em Python)
- **Netflix** (utiliza muitas bibliotecas Python)
- **Uber** (boa parte do aplicativo é feita com Python)
- **Dropobox** (contratou o criador da linguagem Python, *Guido van Rossum*)
- **Pinterest** (utiliza Python e Django)
- **Reddit** (utiliza bibliotecas Python)

Profissões que utilizam Python

Você pode se tornar um profissional desenvolvedor ou analista a partir dos seus conhecimentos de programação Python. Algumas profissões que costumam utilizar muito Python são:

- **Analista de Dados**
- **Cientista de Dados**
- **Engenheiro de Machine Learning**
- **Pesquisador de Inteligência Artificial**
- **Engenheiro de Software**
- **Desenvolvedor Web**
- **Desenvolvedor Mobile**

Exemplos de códigos Python

Até aqui falamos bastante sobre a linguagem, mas ainda não apresentamos nenhum código. Para você ter uma ideia de como a linguagem funciona, observe esses exemplos abaixo.

> **Obs:** *para adicionar um comentário no código, utiliza-se uma hashtag (#).*

Código para imprimir a frase "Olá, mundo!" na tela:

```
print('Olá, mundo!')
# Código que adiciona dois números:
numero_1 = 3
numero_2 = 7

soma = numero_1 + numero_2
print('A soma é:', soma)
# Importando uma função para gerar um número aleatório entre 0 e 100:
import random
numero_aleatorio = random.randint(0,100)
print('O número gerado é:', numero_aleatorio)
```

Palavras Reservadas no Python

Ao escrever seu código de programação Python, você pode dar nomes às variáveis que está criando, como nos exemplos que mostramos acima (numero_1, numero_aleatorio, etc.).

Mas algumas palavras são reservadas, ou seja, possuem funções específicas dentro da linguagem, por isso não podem ser utilizadas para outra finalidade. São elas:

False | *None* | *True* | *and* | *as* | *assert* | *break* | *class* | *continue* |*def* | *del* | *elif* | *else* | *except* | *finally* | *for* | *from* | *global* | *if* | *import* | *in* | *is* | *lambda* | *not* | *nonlocal* | *or* | *pass* | *raise* | *try* | *return* | *while* | *with* | *yield*

Interpretador Python: qual a melhor IDE

A linguagem Python, depois de instalada em seu computador, pode ser executada diretamente via terminal, sem a necessidade de nenhum software específico.

Mas para tornar o ambiente mais atrativo visualmente e facilitar a visualização do código, bem como ter à disposição recursos extras, é bastante útil escrever e compilar os códigos Python em uma IDE (IDE é uma sigla em inglês para “ambiente de desenvolvimento integrado”).

As melhores e mais populares IDEs para programar em Python são:

- **Jupyter Notebook**
- **Pycharm**
- **VS Code**

- **Sublime Text**
- **Atom**
- **Spyder**
- **Vim**

OBS: para instalar Python no seu computador, recomendamos que utilize o pacote Anaconda.

Bibliotecas e Frameworks populares do Python

Confira abaixo algumas bibliotecas e frameworks muito utilizados, que vale a pena você conferir e aprender a manipular para dominar funções e recursos específicos:

- **Frameworks e bibliotecas para desenvolvimento web**: Django, Flask, Pyramid;
- **Bibliotecas para desenvolvimento mobile**: Kivy, BeeWare;
- **Frameworks e bibliotecas para machine learning, inteligência artificial e ciência de dados**: Scikit-learn, Tensorflow, Keras, Pytorch, Pandas, Seaborn;
- **Bibliotecas para ciência e computação numérica (equivalência ao Matlab)**: Scipy, Numpy, Matplotlib;
- **Bibliotecas para web scraping**: Scrapy, Beautifulsoup, Requests, Urllib;
- **Bibliotecas para chatbots e processamento de texto**: ChatterBot, NLTK, SpaCy;
- **Bibliotecas para manipulação de imagens**: OpenCV, Scikit-image, Pillow;
- **Bibliotecas para games**: Pygame, Kivy, Panda3D, Blender.

Fonte: https://didatica.tech/a-linguagem-python/

CONHEÇA OS PRINCIPAIS FRAMEWORKS QUE UTILIZAM PYTHON

Vejamos a seguir alguns frameworks que auxiliam em nossa produtividade no desenvolvimento de uma aplicação Python:

- **Django**: **este framework possui código aberto e uma estrutura de alto nível para o desenvolvimento de aplicações.** Ele tem como recursos disponíveis as URLs amigáveis, um sistema de templates, mecanismos de autenticação, facilidade para trabalhar com bancos de dados, dentre outros. É utilizado por empresas como: Pinterest, Instagram, Mozilla e Bitbucket.
- **Flask**: **O Flask é um micro-framework devido a sua simplicidade.** Seu uso é mais indicado para projetos pequenos, pois ele não possui conexão com bancos de dados, validações de formulários, etc. Possui uma curva de aprendizado mais leve, devido a sua simplicidade. É utilizado pelo LinkedIn e pelo Pinterest, mesmo sendo indicado para projetos pequenos.
- **Web2Py**: **Nesse framework, é permitido que as pessoas desenvolvedoras criem conteúdo de forma dinâmica para a Web.** Ou seja, é dispensada a criação de formulários ou similares do zero para nossas aplicações. Ele segue o modelo MVC (Model, View e Controller) e sua escrita foi baseada em Django com Ruby on Rails. Tem integração com vários bancos de dados.

Conheça as 3 principais bibliotecas Python.

Uma das características da linguagem Python é sua modularidade. Por isso, existem diversas bibliotecas disponíveis na internet que adicionam ainda mais poder a ela, o que é essencial para possibilitar o desenvolvimento de aplicações de diferentes setores, como Machine Learning, que significa aprendizado de máquina e Ciência de Dados. Confira as **principais bibliotecas** a seguir.

1. NumPy

A biblioteca NumPy é utilizada para o processamento de cálculos com matrizes e vetores. Ela contém uma série de funções para a manipulação de arrays simples e multidimensionais e é usada em algoritmos de Machine Learning, para a manipulação de imagens em sistemas de computação gráfica e muito mais.

Trata-se de uma biblioteca de código aberto e disponibilizada gratuitamente. Sua instalação pode ser feita por meio do gerenciador de pacotes como o pip. Vale ressaltar que existem distribuições do Python que já vem com o NumPy, como a Anaconda Python.

2. Pandas

A biblioteca pandas é utilizada para trabalhar com análise de dados. Ela oferece uma série de funções que permitem a leitura e manipulação de dados. Por isso, é amplamente utilizado em Machine Learning, Ciências de Dados, Mineração de Dados, em cálculos estatísticos e muito mais.

O pandas é uma biblioteca de código aberto desenvolvida sobre a biblioteca NumPy. Dessa forma, os dados podem ser estruturados de diferentes formas. Sua instalação pode ser feita por meio do gerenciador de pacotes pip.

3. SciPy

A biblioteca SciPy é utilizada para a realização de cálculos científicos, que requerem a utilização de algoritmos complexos. Ela é utilizada para resolver cálculos matemáticos e de engenharia, por exemplo. O SciPy também é uma biblioteca de código aberto e desenvolvida sobre a NumPy. Sua instalação pode ser feita por meio do gerenciador de pacotes.

Empresas que utilizam Python

Vejamos abaixo algumas empresas que utilizam o Python como back-end para suas aplicações:

- **Google;**
- **Spotify;**
- **Instagram;**
- **Amazon;**
- **Facebook;**
- **Pinterest;**
- **Mozilla.**

Só empresas de alto nível não é? Essa

linguagem está presente em basicamente todos os sites que utilizamos. **Contudo, como ela é uma linguagem que roda no lado do servidor, não conseguimos visualizar o código dela.**

CRÉDITO DO TEXTO: Michelle Horn. In https://blog.betrybe.com/python/

QUESTÕES DE PROVAS

01. (IFMT 2018 IFMT – Informática) Sobre a linguagem Python, é INCORRETO afirmar que: (Marque CERTO ou ERRADO)

I- Suporta os paradigmas: imperativo, orientado a objetos e funcional.

02. (IFMT 2018 IFMT – Informática) Sobre a linguagem Python, é INCORRETO afirmar que: (Marque CERTO ou ERRADO)

I- Python é um software de código aberto.

03. 9FAURGS - 2022 - SES-RS - Analista de Sistemas) Considere as seguintes afirmações sobre a linguagem Python.

I - Em Python, não há um tipo de caractere separado: um caractere é simplesmente uma string de tamanho um.

II - No uso da linguagem Python em modo interativo, se executarmos a seguinte sequência de comandos relacionados a strings de caracteres:

>>>

>>> word = 'Python'

Teremos que o valor para word[5] será igual ao valor para word[-1] e o valor para word[0] será igual ao valor para word[-6]

III- Strings em Python não podem ser alteradas – são imutáveis. Portanto, atribuir um valor a uma posição indexada na string resulta em um erro. Se precisar de uma string diferente, você deve criar uma nova.

Quais estão corretas?

A Apenas I.
B Apenas I e II.
C Apenas I e III.
D Apenas II e III.
E I, II e III.

04. (FGV - 2021 - Banestes - Analista em Tecnologia da Informação - Segurança da Informação

Considere o código Python a seguir.

```
def F(a, b, c):
    for k in range(a,b):
        print k ** c
```

Dado que uma execução da função F exibiu os números

16, 9, 4, 1, 0, 1,

é correto afirmar que os valores dos parâmetros a, b, c empregados foram, respectivamente:

A -4, 1, 2;
B -4, 2, 2;
C -4, 0, 4;
D 4, -1, 1;
E 4, 2, 2.
B

05. (FGV - 2021 - Banestes - Analista em Tecnologia da Informação - Suporte e Infraestrutura) Considere o código Python 2.7 a seguir.

```
L=[6,5,4,3,2,1]
for k in range(-3,3):
    print L[k]
```

A execução desse código exibe os números:

A 1 1 1 6 5 4;
B 1 2 3 4 5 6;
C 3 2 1 6 5 4;
D 6 5 4 3 2 1;
E 6 5 4 6 5 4.

Gabarito: 01/C; 02/C; 03/E;04/B; 05/C

LÍNGUAGEM EM R

R é uma linguagem de programação estatística e gráfica que vem se especializando na **manipulação, análise e visualização de dados, sendo atualmente considerada uma das melhores ferramentas para essa finalidade**.

A linguagem ainda possui como diferencial a facilidade no aprendizado, mesmo para aqueles que nunca tiveram contato anterior com programação.

Conhecendo o R

O R foi criado em 1995 por estatísticos que realizaram sua implementação a partir da linguagem S da Bell Labs, com a finalidade de obter um melhor ambiente de software para laboratórios de estatística.

A linguagem pode ser executada em diferentes sistemas operacionais, como Windows, Mac OS e Linux e possui código aberto, o que permite sua utilização para visualização, modificação e distribuição **de graça por qualquer pessoa ou empresa, com qualquer finalidade**.

Essas características contribuem de maneira significativa para seu desenvolvimento, levando a uma **comunidade ativa de colaboradores espalhados pelo mundo**, onde qualquer desenvolvedor pode contribuir para melhoria do sistema.

Esta longa e confiável história do R, somada à sua grande e sólida comunidade de apoio, coloca-o como **excelente opção para análise de dados e machine learning**.

Como programar em R

Programar na linguagem R é muito simples. Mesmo quem nunca teve contato algum com programação costuma ter facilidade no aprendizado, principalmente quando o ensino é de qualidade. Com didática, exemplos simples e objetivos, avançando

para assuntos mais complexos de maneira gradual, sem ignorar conceitos importantes (evitando assim lacunas no aprendizado), todos conseguirão programar em R.

Vamos conhecer um pouco sobre a **estrutura da linguagem**, para então entender como os scripts são criados:

- **Variáveis:** são criadas pelo programador com a finalidade de salvar informações. As informações inseridas em uma variável ficarão disponíveis para utilização enquanto a variável existir. As variáveis terão um tipo específico, de acordo com o tipo de dado que foi salvo ou de acordo com a indicação do programador.
- **Funções:** são um conjunto de instruções pré-definidas que executam uma ou mais tarefas. Existem muitas funções já prontas para sua utilização, salvas em pacotes desenvolvidos para facilitar a criação de scripts. Quanto mais funções o programador conhecer, mais fácil será escrever seus scripts. Também é possível criar funções no próprio script, passando todas as instruções que deverão acontecer quando a função for utilizada.
- **Operadores:** com os operadores fazemos operações matemáticas, como soma, divisão, multiplicação, etc., e comparações como *igual*, *diferente*, *maior* e *menor*. Também podemos utilizar os operadores lógicos *E*, *OU*, e *negação* para simplificar nossos códigos.
- **Tipos de dados:** *numéricos* (operações matemáticas), *caracteres* (operações com letras, palavras, frases, etc.), *fatores* (categorias) e *lógicos* (verdadeiro ou falso). Esses são os tipos de dados básicos existentes na linguagem R.
- **Estrutura de dados:** *vetores*, que são uma sequência de dados do mesmo tipo. *Listas*, que são vetores com tipos de dados diferentes. *Matrizes*, que possuem duas dimensões e um tipo de dado. *Data frames*, que são estruturas mais complexas, similares as planilhas do Excel e com tipos de dados diferentes.
- **Condicionais:** *If*, *For* e *While*. No condicional *If*, dizemos que se algo é verdadeiro, uma ação deve ser realizada, se não é, outra será (ou nenhuma). No loop *For,* uma ou mais instruções serão realizadas determinado número de vezes. No loop *While,* as instruções serão realizadas enquanto uma condição for atendida.

Com essas informações, já podemos realizar a criação de scripts, inclusive com considerável complexidade. Para este primeiro contato com a linguagem, iremos apresentar o mais simples, que normalmente é utilizado quando estamos iniciando o aprendizado de uma nova linguagem de programação, que é escrever na tela a frase *"Olá mundo!"*.

Uma opção de código para essa finalidade é:

> Frase <- "Olá mundo!"

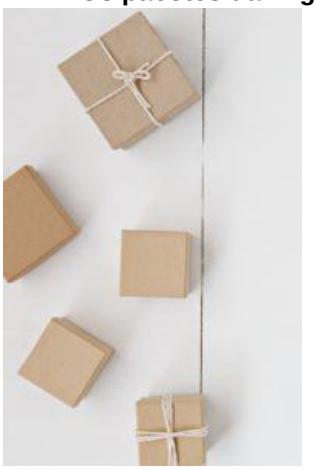
> print(Frase)

Após a execução dessas duas linhas, a o resultado abaixo será apresentado:

"Olá mundo!"

Com a execução da primeira linha desse código, a variável *"Frase"* é criada em nossa sessão, e recebe como conteúdo a frase *"Olá mundo!"*. A variável *"Frase"* fica então a disposição para ser utilizada, o que significa que poderíamos utilizá-la de muitas formas. Na segunda linha, simplesmente usamos a função *print()*, que irá trazer como resultado de sua execução a impressão na tela do conteúdo existente na variável informada dentro dos parênteses.

Os pacotes da linguagem R



No R existem as funções *built-in*, que são o coração da linguagem e compõem o *R-base*. Elas permitem o **funcionamento básico do programa**, e são carregadas no momento em que iniciamos o R, estando à disposição para sua utilização direta, sem a necessidade de nenhum comando prévio no script.

Um exemplo que já utilizamos foi a função *print()*. Mais a frente esta ideia ficará mais clara.

Além do R-base existem também os pacotes da linguagem, que possuem funções com as mais diferentes finalidades. Os pacotes recomendados são instalados juntamente com o R-base, porém não são carregados quando a sessão é iniciada, como acontece com as funções built-in. Desta forma, precisamos realizar o **carregamento do pacote** antes de utilizar uma função dele.

Para isso utilizamos o comando *library(),* informando entre os parênteses o nome do pacote que desejamos carregar.

No **CRAN,** que é o repositório de pacotes da linguagem R, ainda existe uma grande quantidade de pacotes, que hoje passa de 15.000. Esses pacotes são submetidos pelos mais diferentes desenvolvedores. Inclusive você pode criar um pacote e submetê-lo ao CRAN. São poucas as tarefas que você pode precisar realizar que não possuem um pacote, com uma função que atenda sua necessidade já desenvolvida.

No site https://cran.r-project.org/ estão os detalhes dos pacotes e suas funções, normalmente com uma **documentação detalhada**, explicando claramente o funcionamento de cada função.

Caso você tente utilizar uma função que não esteja carregada na sessão, uma mensagem de erro será apresentada: *"could not find function …".*

```
Console  Terminal  Jobs
~/
> filtro <- createDataPartition(x, p=0.7, list=FALSE)
Error in createDataPartition(x, p = 0.7, list = FALSE) :
  could not find function "createDataPartition"
>
```

Você deverá então carregar o pacote dessa função, e novamente um erro pode ser apresentado: *"there is no package called …".*

```
Console  Terminal  Jobs
~/
> library(caret)
Error in library(caret) : there is no package called 'caret'
>
```

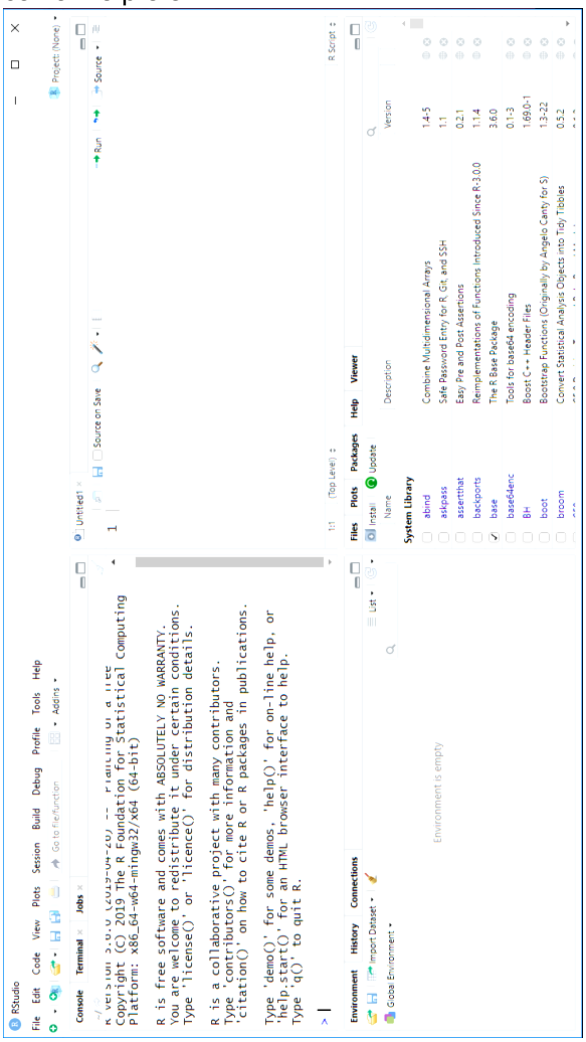
Esse comportamento acontecerá quando você tentar utilizar um pacote que **não veio instalado com o R-base**. Neste caso, antes de carregar o pacote, será necessário **realizar a sua instalação**. Para isso é necessário possuir o arquivo com o pacote, ou conexão com a internet, para então utilizar a função *install.packages()*, indicando dentro dos parênteses, entre aspas, o nome do pacote em questão.

Este procedimento será necessário apenas na primeira execução do pacote no computador, nas próximas vezes ele já estará instalado, sendo necessário apenas o carregamento com a função *library()*.

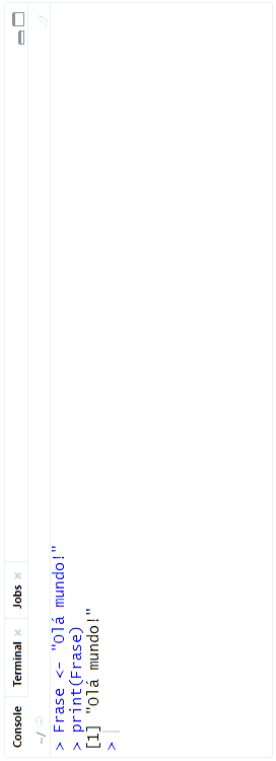
RStudio

Mais uma vantagem de utilizar a linguagem R está no **RStudio**, que é uma IDE, ou **ambiente de desenvolvimento integrado**, para o R. Com sua utilização gratuita, o RStudio é uma excelente ferramenta para desenvolvimento em R, extremamente visual quando comparado com ambientes de outras linguagens e também com o console do R, e muito simples de se utilizar.

Área de trabalho: abaixo vemos uma tela básica de trabalho do RStudio, já contendo algumas poucas customizações. Você pode ajustar os painéis conforme preferir.



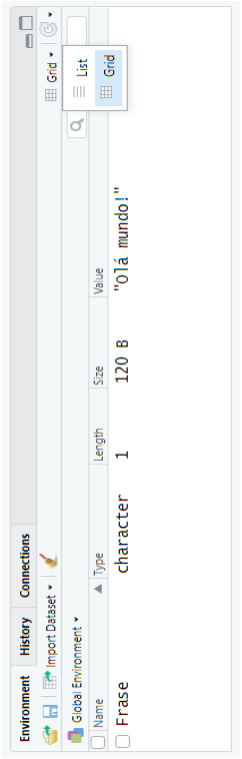
Console: onde as instruções efetivamente acontecem, e as respostas são apresentadas. Você pode escrever a instrução diretamente no console e executar, ou ainda escrever no painel *"Source",* ao lado. Após a execução as instruções serão automaticamente enviadas ao console e executadas.



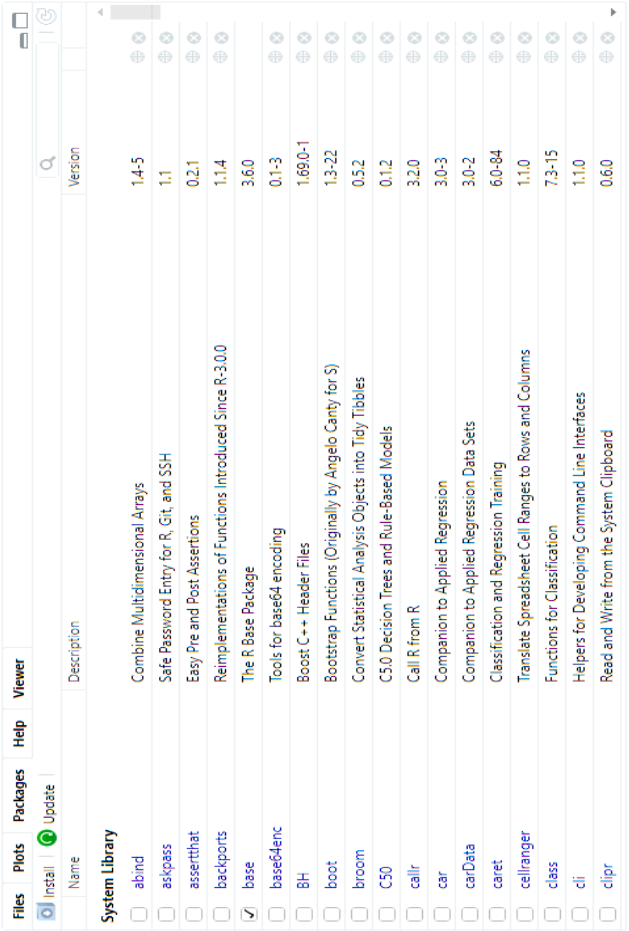
Source: onde podemos escrever o script. O editor de texto nos auxilia completando o código conforme iniciamos a escrita, inclusive com opções de variáveis já criadas, entre outras facilidades. Podemos inserir comentários, que nos auxiliam a entender o código, começando a linha com uma *"#"*.

Se salvarmos o conteúdo do painel será gerado um arquivo com a extensão .R, para edição futura. Para executar uma linha basta pressionar *"ctrl"* + *"enter"*, ou selecionar as linhas que queremos executar e clicar sobre o botão *"Run"*.

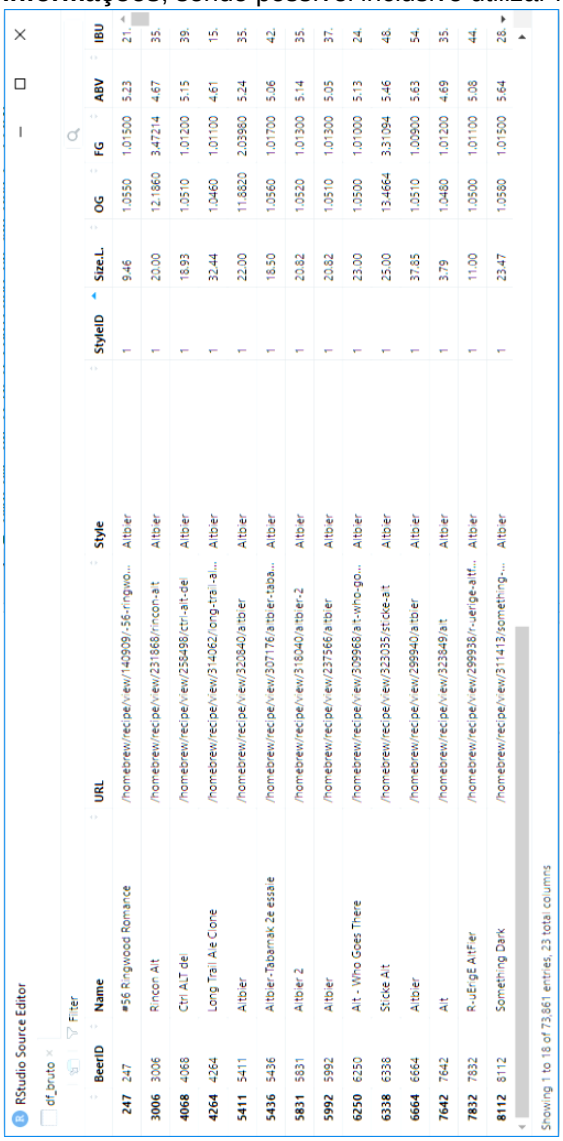
Environment: neste painel as variáveis já criadas serão exibidas. É possível também exclui-las diretamente no painel, selecionando a variável e clicando sobre o botão *"vassoura"*. Conforme vemos, há duas opções de visualização: *"List"* e *"Grid"*. Na Grid temos o tipo, comprimento, tamanho e conteúdo da variável.



Pacotes, Visualizações, Ajuda: neste painel temos algumas opções interessantes, como a lista de pacotes já instalados, uma área para visualizar gráficos gerados, e a ajuda, onde podemos verificar os detalhes das funções.



Variáveis: podemos ainda utilizar a função *View()* para visualizar o conteúdo de uma variável. Desta forma o conteúdo será exibido em uma aba no painel *"Source"*, podendo ser expandido para tela cheia. Com esta funcionalidade a visualização de conteúdos fica muito simples, e em muitos casos será apresentada uma **tabela com as informações**, sendo possível inclusive utilizar filtros.



R para machine learning

Todas essas características, atreladas a simplicidade para criação de modelos estatísticos tornam o R uma **excelente opção para a aplicação de machine learnig**. Muitos algoritmos de machine learning estão à disposição para utilização no R, muitas vezes sendo executados com uma ou duas linhas de código.

A utilização dos algoritmos de machine learning é apenas uma das etapas do processo. Para efetiva aplicação é necessário que a ferramenta escolhida atenda a todas as etapas, possibilitando assim o **conhecimento dos dados**, **preparação**, **aplicação do algoritmo** e **apresentação dos resultados**.

A linguagem R tem um ótimo desempenho em todas essas etapas, possuindo pacotes de grande qualidade para utilização nestas finalidades. Alguns destes são os pacotes da **coleção Tidyverse**, desenvolvida especificamente para análise de dados, permitindo que as manipulações e visualizações aconteçam de maneira simples e eficaz.

A coleção possui muitos pacotes, sendo os principais o **dplyr, ggplot2 e stringr**.

-dplyr: foco na manipulação dos dados de maneira muito fácil, com funções intuitivas, inclusive lembrando a linguagem SQL.

-ggplot2: baseado na Gramática dos Gráficos os gráficos são gerados com base nos dados informados. Com poucas linhas de código, visualizações de grande qualidade são geradas.

-stringr: tratamento de strings (texto) de diferentes maneiras, com objetividade e simplicidade.

Para utilização dos algoritmos de machine learning, uma das principais opções é o **pacote caret**, que possui excelentes funções para criação e ajuste de modelos. Dentre as principais funções do pacote estão a *train*() e a *trainControl()*, que permitem a utilização de diferentes algoritmos apenas alterando alguns parâmetros das funções, trazendo grande simplicidade ao código necessário para teste de diferentes modelos.

Além de sua grande aplicabilidade, o pacote ainda conta com uma excelente documentação, muito bem organizada, separada em 24 capítulos, deixando muito clara a finalidade do pacote, e **simplificando sua aprendizagem**.

Fonte: https://didatica.tech/a-linguagem-r/

QUESTÕES DE PROVAS

01. (AOCP - 2020 - MJSP - Cientista de Dados - Big Data) Assinale a alternativa que apresenta o comando que informa à Linguagem R em qual pasta ela deve ler os arquivos de dados.

A filter().
B select().
C read_fwf().
D setwd().
E library().

02. (CESPE / CEBRASPE - 2022 - ANP - Regulador

de Novas Atribuições IV - Cargo 7

Julgue o item a seguir, relativos a conceitos de R.

Considere-se o script R que se segue.

```
X <- c(0, 19, 205, 34, 506)
Y <- X
X[2] <- 91
Y[6] <- 71
print(X)
print(Y)
```

O resultado da execução desse script é o apresentado a seguir.

```
 0 91 205 34 506
 0 19 205 34 506 71
```

03. (FUNDATEC - 2022 - AGERGS - Técnico Superior Engenheiro de Dados) Sobre os comandos R, analise as assertivas abaixo e assinale a alternativa correta.

I. license – detalha as condições de distribuição do R.
II. contributors – lista a equipe do time-cerne de desenvolvimento.
III. Citation – ensina como citar o R em trabalhos acadêmicos.
IV. Demo – inicia uma sessão interativa de demonstração do R.

A Todas estão corretas.
B Todas estão incorretas.
C Apenas I está correta.
D Apenas I e II estão corretas.
E Apenas III e IV estão corretas.

04. (FUNDATEC - 2022 - AGERGS - Técnico Superior Engenheiro de Dados) Quando usamos o símbolo “=” ou “<” seguido de “-” estamos criando objetos com um nome que aparece à esquerda e que contém alguns elementos (o que vem à direita do símbolo). Quando criamos um objeto, ele fica guardado na memória do R até que se feche o programa. Porém, os objetos criados ficam ocultos. Para ver a lista de arquivos ocultos, basta dar o seguinte comando:

A tuple()
B console()
C rain()
D table_rain()
E ls()

05. FUNDATEC - 2022 - AGERGS - Técnico Superior Engenheiro de Dados) Estamos interessados em ajustar um modelo de regressão linear simples no pacote estatístico R. O comando a ser utilizado para esse fim é:

A ts
B glm
C lm
D cor
E road

06. (IBFC - 2021 - IBGE - Supervisor de Pesquisas - Suporte Gerencial) Função da linguagem de programação R que permite fazer gráficos de dispersão. De acordo com a descrição, a função é:

A head
B plot
C grafhic
D dispersion
E mean

07. (AOCP - 2020 - MJSP - Analista de Governança de Dados - Big Data) Um analista do MJSP necessita apresentar um gráfico para seus usuários. Para tanto, ele irá utilizar a linguagem R. Assinale a alternativa que apresenta corretamente o nome da função que o analista deve utilizar para gerar o gráfico em linguagem R.

A Graph.
B Draw.
C Plot.
D Picture.
E Trace.

Gabarito: 01/D; 02/C; 03/A; 04/E; 05/C; 06/B; 07/C

BANCOS DE DADOS NÃO RELACIONAIS: BANCOS DE DADOS NOSQL; MODELOS NOSQL.

O QUE SÃO BANCOS DE DADOS NOSQL?

Bancos de dados NoSQL são criados para modelos de dados específicos e têm esquemas flexíveis para a criação de aplicativos modernos. Os bancos de dados NoSQL são amplamente reconhecidos por sua facilidade de desenvolvimento, funcionalidade e performance em escala. Esta página inclui recursos para ajudar você a compreender melhor os bancos de dados NoSQL e a começar a usá-lo.

DICAS:

Bancos de dados do modelo NoSQL são considerados orientados a objetos, pois armazenam os dados em formatos distintos dos utilizados em bancos de dados relacionais. (CESPE/CEBRASPE - IBAMA - Analista Administrativo - 2022)

A principal diferença entre bancos de dados relacionais e bancos de dados NoSQL está na questão da segurança dos dados e das transações; os bancos de dados NoSql são imunes a ataques de injeção SQL. (CESPE/CEBRASPE - Petrobras - Analista de Sistemas - Área Processos de Negócio - 2022)

Como funciona um banco de dados NoSQL (não relacional)?

Os bancos de dados NoSQL usam uma variedade de modelos de dados para acessar e gerenciar os dados. Esses tipos de banco de dados são otimizados especificamente para aplicativos que exigem modelos de grande volume de dados, baixa latência e flexibilidade. Esses requisitos são

atendidos mediante o relaxamento de algumas restrições de consistência de dados dos outros bancos.

Considere o exemplo de modelagem do esquema para um banco de dados simples de livros:

- Em um banco de dados relacional, um registro de livro é normalmente disfarçado (ou "normalizado") e armazenado em tabelas separadas, e os relacionamentos são definidos por restrições de chave primária e externa. Neste exemplo, a tabela Livros têm colunas para ISBN, Título do livro e Número da edição, a tabela Autores têm colunas para AuthorID e Nome do autor e, finalmente, a tabela Author-ISBN tem colunas para AuthorID e ISBN. O modelo relacional é projetado para permitir que o banco de dados imponha a integridade referencial entre as tabelas no banco de dados, normalizadas para reduzir a redundância e geralmente otimizadas para armazenamento.
- Em um banco de dados NoSQL, um registro de livro é normalmente armazenado como um documento JSON. Para cada livro, o item, o ISBN, o Título do livro, o Número de edição, o Nome do autor e o AuthorID são armazenados como atributos em um único documento. Neste modelo, os dados são otimizados para desenvolvimento intuitivo e escalabilidade horizontal.

Por que você deve usar um banco de dados NoSQL?

Os bancos de dados NoSQL são ideais para muitos aplicativos modernos, como dispositivos móveis, Web e jogos, que exigem bancos de dados flexíveis, escaláveis, de alta performance e altamente funcionais para proporcionar ótimas experiências aos usuários.

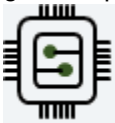
- Flexibilidade: os bancos de dados NoSQL geralmente fornecem esquemas flexíveis que permitem um desenvolvimento mais rápido e iterativo. O modelo de dados flexível torna os bancos de dados NoSQL ideais para dados semiestruturados e não estruturados.
- Escalabilidade: os bancos de dados NoSQL geralmente são projetados para serem escalados horizontalmente usando clusters distribuídos de hardware, em vez de escalá-los verticalmente adicionando servidores caros e robustos. Alguns provedores de nuvem lidam com essas operações nos bastidores como um serviço totalmente gerenciado.
- Alta performance: o banco de dados NoSQL é otimizado para modelos de dados específicos e padrões de acesso que permitem maior performance do que quando se tenta realizar uma funcionalidade semelhante com bancos de dados relacionais.
- Altamente funcional: os bancos de dados NoSQL fornecem APIs e tipos de dados altamente funcionais criados especificamente para cada um de seus respectivos modelos de dados.

Tipos de bancos de dados NoSQL

Chave-valor: os bancos de dados de chave-valor são altamente particionáveis e permitem escalabilidade horizontal em escalas que outros tipos de bancos de dados não conseguem alcançar. Casos de uso, como jogos, tecnologia de publicidade e IoT, encaixam-se particularmente bem ao modelo de dados de chave-valor. O Amazon DynamoDB foi projetado para proporcionar uma latência consistente de um dígito de milissegundo para qualquer escala de cargas de trabalho. Essa performance consistente teve grande influência no motivo pelo qual o recurso Snapchat Stories, que inclui a maior carga de trabalho de gravação de armazenamento do Snapchat, mudou para o DynamoDB.

Documento: no código do aplicativo, os dados costumam ser representados como um objeto ou um documento do tipo JSON porque esse é um modelo de dados eficiente e intuitivo para os desenvolvedores. Os bancos de dados de documentos facilitam para que os desenvolvedores armazenem e consultem dados usando o mesmo formato de modelo de documento que usam no código do aplicativo. A natureza flexível, semiestruturada e hierárquica dos documentos e dos bancos de dados de documentos permite que eles evoluam conforme as necessidades dos aplicativos. O modelo de documentos funciona bem com catálogos, perfis de usuários e sistemas de gerenciamento de conteúdo, onde cada documento é único e evolui com o passar do tempo. O Amazon DocumentDB (com compatibilidade com o MongoDB) e o MongoDB são bancos de dados de documentos populares que fornecem APIs eficientes e intuitivas para desenvolvimento flexível e iterativo.

Gráfico: a finalidade de um banco de dados gráfico é facilitar a criação e a execução de aplicativos que funcionam com conjuntos de dados altamente conectados. Os casos típicos de uso de um banco de dados gráfico incluem redes sociais, mecanismos de recomendação, detecção de fraudes e gráficos de conhecimento. O Amazon Neptune é um serviço totalmente gerenciado de banco de dados gráfico. O Neptune suporta o modelo Property Graph e o Resource Description Framework (RDF), proporcionando a escolha de duas APIs gráficas: TinkerPop e RDF/SPARQL. Os bancos de dados gráficos populares incluem o Neo4j e Giraph.



Em memória: aplicações de jogos e tecnologia de publicidade têm casos de uso como placares de líderes, armazenamentos de sessões e análises em tempo real que exigem tempos de

resposta em microssegundos e podem ter grandes picos de tráfego a qualquer momento. O Amazon MemoryDB for Redis é um serviço de banco de dados na memória compatível com Redis, durável, que oferece latência de leitura de microssegundos, latência de gravação de um dígito em milissegundos e durabilidade multi-AZ. O MemoryDB foi desenvolvido com o propósito específico de oferecer performance e durabilidade ultrarrápidos, para que você possa usá-lo como seu banco de dados primário para aplicações modernas de microsserviços. O Amazon ElastiCache é um serviço de cache na memória totalmente gerenciado, compatível com Redis e Memcached, para atender a workloads de baixa latência e alta taxa de transferência. Clientes como Tinder, que requerem resposta em tempo real de suas aplicações, contam com armazenamento de dados na memória em vez de armazenamento de dados baseados em disco. O Amazon DynamoDB Accelerator (DAX) é outro exemplo de armazenamento de dados de propósito específico. O DAX faz o DynamoDB ler em uma ordem de grandeza mais rápida.

Pesquisar: muitos registros de saída de aplicativos ajudam os desenvolvedores a solucionar problemas. O Amazon OpenSearch Service é construído especificamente para fornecer visualizações e análises quase em tempo real de dados gerados por máquina ao indexar, agregar e pesquisar registros e métricas semiestruturadas. O Amazon OpenSearch Service também é um mecanismo de pesquisa poderoso e de alta performance para casos de uso de pesquisa de texto completo. A Expedia está usando mais de 150 domínios do Amazon OpenSearch Service, 30 TB de dados e 30 bilhões de documentos para diversos casos de uso de missão crítica, desde o monitoramento operacional e solução de problemas até o rastreamento de pilha de aplicativo distribuído e otimização da definição de preço.

Bancos de dados SQL (relacional) vs. NoSQL (não relacional)

Durante décadas, o modelo de dados predominante usado para desenvolvimento de aplicativos foi o modelo usado por bancos de dados relacionais, como Oracle, DB2, SQL Server, MySQL e PostgreSQL. Somente em meados dos anos 2000 que outros modelos de dados começaram a ser adotados e ter um uso mais significativo. Para diferenciar e categorizar essas novas classes de bancos e modelos de dados, o termo “NoSQL” foi criado. Muitas vezes, o termo “NoSQL” é usado de forma intercambiável com “não relacional”.

Embora existam muitos tipos de bancos de dados NoSQL com recursos variados, a tabela a seguir mostra algumas das diferenças entre os bancos de dados SQL e NoSQL.

	Bancos de dados relacionais	Bancos de dados NoSQL
Cargas de trabalho ideais	Bancos de dados relacionais são projetados para aplicativos transacionais e fortemente consistentes de processamento de transações online (OLTP) e são bons para processamento analítico online (OLAP).	Os bancos de dados do NoSQL são projetados para vários padrões de acesso aos dados que incluem aplicativos de baixa latência. Os bancos de dados de pesquisa NoSQL são projetados para análise de dados semiestruturados.
Modelo de dados	O modelo relacional normaliza dados em tabelas, compostas por linhas e colunas. Um esquema define estritamente tabelas, colunas, índices, relações entre tabelas e outros elementos do banco de dados. O banco de dados impõe a integridade referencial nos relacionamentos entre as tabelas.	Os bancos de dados NoSQL fornecem uma variedade de modelos de dados, como chave-valor, documento e gráfico, que são otimizados para performance e escala.
Propriedades ACID	Bancos de dados relacionais fornecem propriedades de atomicidade, consistência, isolamento e durabilidade (ACID): • A atomicidade exige uma transação para executar completamente ou não é executada de forma alguma. • A consistência exige que, quando uma transação é confirmada, os dados devem estar	Os bancos de dados NoSQL geralmente fazem compensações relaxando algumas das propriedades ACID dos bancos de dados relacionais para um modelo de dados mais flexível que pode ser escalado horizontalmente. Isso torna os bancos de dados NoSQL uma excelente opção para casos de uso de baixa latência e alta taxa de

	em conformidade com o esquema do banco de dados. • O isolamento exige que as transações simultâneas sejam executadas separadamente umas das outras. • A resiliência exige a capacidade de se recuperar de uma falha do sistema ou falta de energia inesperada para o último estado conhecido.	transferência que precisam ser escalados horizontalmente além das limitações de uma única instância.
Performance	A performance normalmente depende do subsistema do disco. A otimização de consultas, índices e estrutura de tabela é necessária para alcançar máxima performance.	A performance geralmente é uma função do tamanho do cluster do hardware subjacente, da latência de rede e do aplicativo que faz a chamada.
Escala	Os bancos de dados relacionais geralmente escalam verticalmente o tamanho ao aumentar os recursos de computação do hardware, ou escalam horizontalmente o tamanho ao adicionar réplicas para cargas de trabalho somente leitura.	Os bancos de dados NoSQL normalmente são particionáveis porque os padrões de acesso podem escalar horizontalmente o tamanho usando arquitetura distribuída para aumentar a taxa de transferência que fornece performance consistente em escala quase ilimitada.
APIs	As solicitações para armazenar e recuperar dados são comunicadas usando consultas compatíveis com uma Structured Query Language (SQL – Linguagem de consultas estruturadas). Essas consultas são analisadas e executadas pelo banco de dados relacional.	APIs baseadas em objetos permitem que desenvolvedores de aplicativos armazenem e restaurem facilmente estruturas de dados. As chaves de partição permitem que os aplicativos procurem pares de chave-valor, conjuntos de colunas ou documentos semiestruturados que contenham objetos e atributos de aplicativos serializados.

Terminologia do SQL vs. do NoSQL

A tabela a seguir compara a terminologia usada pelos bancos de dados NoSQL selecionados com a terminologia usada pelos bancos de dados SQL.

SQL	MongoDB	DynamoDB	Cassandra	Couchbase
Tabela	Coleta	Tabela	Tabela	Bucket de dados
Linha	Documento	Item	Linha	Documento
Coluna	Campo	Atributo	Coluna	Campo
Chave primária	ObjectId	Chave primária	Chave primária	ID do documento
Índice	Índice	Índice secundário	Índice	Índice
Visualização	Visualização	Índice secundário global	Visualização materializada	Visualização
Tabela ou objeto aninhado	Documento incorporado	Mapa	Mapa	Mapa
Matriz	Matriz	Lista	Lista	Lista

Fonte: https://aws.amazon.com/pt/nosql/#:~:text=Os%20bancos%20de%20dados%20do,para%20an%C3%A1lise%20de%20dados%20semiestruturados.&text=O%20modelo%20relacional%20normaliza%20dados,compostas%20por%20linhas%20e%20colunas.

PRINCIPAIS SGBD'S.

Como há centenas de SGBDs disponíveis, qualquer ranking sem base seria arbitrário. Então, vamos nos basear em um *ranking* popular, que todos os meses classifica SGBDs mais usados no mundo .

O ranking se chama **db-engines.com**. Os SGBDs que comentaremos abaixo são os 10 sistemas mais usados em julho de 2022. Se você acessar o ranking em outro momento, a lista poderá ser diferente.

1. Oracle

A Oracle é uma grande corporação de tecnologia que hoje oferece diversos serviços em nuvem, desde Inteligência Artificial até *Analytics* e BI. Porém, começou e ficou famosa por causa do *Oracle Autonomous Database* ou apenas *Oracle Database*, o banco de dados mais usado do mundo.

Lançado em 1979 e em aprimoramento até a atualidade, é um banco de dados que cresceu e é muito utilizado em conjunto com a linguagem de programação Java, para aplicações robustas, como bancárias e financeiras.

É um banco de dados relacional em sua origem (baseado em tabelas)mas que hoje já é multi-modelo. Utiliza uma linguagem chamada PL/SQL (linguagem procedural projetada para incluir instruções SQL em sua sintaxe) como linguagem de consulta e manipulação de dados.

2. MySQL

MySQL é um SGBD muito famoso e usado por causa de sua integração fácil com a linguagem de programação PHP por esta dupla, MySQL e PHP, estar na origem do WordPress, o sistema de gerenciamento de conteúdo mais popular do mundo.

É um banco de dados relacional gratuito e fácil de usar, que usa SQL, porém com um ótima capacidade e desempenho para muitas aplicações de websites e *webapps*. MySQL nasceu como uma solução de código aberto e que hoje também pertence à Oracle.

3. Microsoft SQL Server

Microsoft SQL Server é um banco de dados lançado em 1989 e que se ramificou em diversas versões, as quais atendem diversos públicos. Há desde versões para pequenas aplicações até outras para aplicações escaláveis de Internet das Coisas (IoT), por meio da Azure, a nuvem da Microsoft.

O SGBD usa um dialeto da linguagem SQL, chamado T-SQL. Por ser uma solução corporativa e paga, como o Oracle Database, oferece uma gama de benefícios a empresas, como maior segurança.

4. PostgreSQL

PostgreSQL é um banco de dados de código aberto e gratuito, mas bastante poderoso. É um banco de dados de modelo objeto-relacional com características como confiabilidade, robustez e desempenho eficiente.

O SGBD tem mais de 30 anos e surgiu na Universidade da Califórnia, Berkeley. Usa SQL como linguagem de consulta e é o banco de dados padrão do macOS Server. É versátil e robusto tanto para aplicações pequenas como para aquelas que requerem acessos massivos a dados.

5. MongoDB

MongoDB é o banco de dados do tipo Nos mais usado no mundo. É NoSQL porque não se baseia no conceito de tabelas para armazenar dados, mas, sim, no modelo de documentos (armazena dados em forma de textos). É de código aberto, gratuito e multiplataforma.

Mais do que isso, MongoDB usa a linguagem de programação Javascript para consultas, a mais utilizada para construir aplicações web. Isso torna MongoDB uma escolha natural, veloz e versátil para aplicativos e sites com acessos massivos, principalmente voltados a conteúdos, como redes sociais, plataformas educacionais, entre outros.

6. Redis

Redis é outro banco de dados NoSQL no ranking do db-engines.com. É o banco de dados de modelo chave-valor mais usado atualmente.

Bancos de dados desse tipo funcionam como dicionários, em que há uma chave e uma série de outros dados (valores) associados a ela. Isso permite muita eficiência na recuperação das informações.

Redis funciona de maneira distribuída (em várias máquinas) e armazena os dados em memória e não em disco, o que o torna extremamente veloz.

7. IBM DB2

Outro SGBD corporativo, o IBM DB2 é uma família de produtos de gerenciamento de dados da IBM, outra gigante do setor. O DB2 iniciou como um banco de dados relacional, mas hoje incorpora soluções de outros modelos, como objeto-relacional e NoSQL.

Também tem como diferenciais escalabilidade, segurança, flexibilidade, entre outros atributos. É uma solução paga, normalmente usada por grandes empresas.

8. Elasticsearch

Elasticsearch é um dos "bancos de dados" mais diferentes entre todos os já vistos. Na verdade, não é como um banco de dados no sentido tradicional, mas um mecanismo de pesquisa de texto completo e em tempo real. Por isso, também é classificado como um SGBD.

Tem uma grande capacidade de indexar quaisquer tipos de textos, quebrá-los em partes menores chamadas *tokens* e permitir buscas inteligentes e rápidas neles (como a busca do Google, por exemplo).

9. Microsoft Access

Muita gente não consideraria o Microsoft Access um SGDB "de verdade". No entanto, ele continua sendo uma solução muito utilizada.

O Access, como é popularmente conhecido, é um aplicativo da família Office, da Microsoft, muito acessível a pessoas de negócios e a leigos, que não precisam saber programar e podem usar a interface gráfica do Office para operá-lo. É por isso que ele é tão popular.

Apesar de não garantir todas as funcionalidades de SGDBs robustos, Access permite criar bancos de dados para pequenas aplicações, úteis para uso interno em times de negócio.

10. SQLite

SQLite, como o próprio nome revela, é um SGBD simples, enxuto e fácil de usar. É útil para sites e aplicativos leves, sem muitos recursos ou usuários. Também usa SQL para consultas. Uma dica: é um ótimo SGBD para treinos e para aprendizado em programação e na área de dados!

SGDBS NA PROGRAMAÇÃO E EM DADOS

A pergunta que fica é: como *programadores*, *data engineers*, profissionais de *data analytics* e *data scientists* usam todos esses SGDBs? A resposta é um grande "depende". A escolha e uso de um SGDB dependerá da empresa e dos projetos em que o profissional atuará.

Grosso modo, porém, cabe aos desenvolvedores de software saberem como conectar e se comunicar com os bancos de dados por meio dos programas que criam. Isso, normalmente, envolve conhecer a linguagem de consulta do banco de dados, como SQL, por exemplo.

Já engenheiros de dados atuam conectando diferentes bancos de dados em *data warehouses* ("armazéns de dados"), a fim de fornecerem informações mais otimizadas a ferramentas analíticas, como *dashboards* e painéis de *Business Intelligence* (BI).

Para cientistas de dados, a atuação pode ser um pouco diferente. Em grandes corporações, os profissionais já costumam receber dados prontos e tratados de *data warehouses*, por exemplo.

Em *startups* e empresas menores, porém, pode ser que o *data scientist* tenha de acessar bancos de dados diretamente para análises. Nesse caso, conhecer as ferramentas e a linguagem de consulta, como SQL, será um diferencial e uma necessidade.

Dominar tudo isso envolve, é claro, muito estudo, muita prática e atualização constante sobre as ferramentas utilizadas no mercado.

Fonte: https://awari.com.br/sistemas-de-banco-de-dados/?utm_source=blog

QUESTÕES DE PROVAS

01. (AOCP - 2019 - UFFS - Analista de Tecnologia da Informação) Bancos de Dados não relacionais, também conhecidos como NoSQL, surgiram para armazenar dados não estruturados, usando modelos de armazenamento específicos para os

tipos de dados que são armazenados, usualmente, fugindo do padrão de armazenamento de linhas e colunas dos bancos de dados tradicionais. Em relação ao Banco de Dados NoSQL, quais são as categorias consideradas para esse tipo de armazenamento?

A Armazéns chave-valor; Banco de dados orientados a documentos; Banco de dados de grafos.

B Primeira forma normal; Banco de dados orientados a documentos; Normalização.

C Armazéns chave-valor; Normalização; Banco de dados orientados a coluna.

D Banco de dados de grafos; Banco de dados orientados a linha; Formas normais de armazenamento.

E Banco de dados orientados a linha; Normalização; Banco de dados orientados a coluna.

02. (CCV-UFC - 2019 - UFC - Técnico de Tecnologia da Informação - Desenvolvimento de Sistemas) Sobre os banco de dados NoSQL, assinale a afirmativa correta.

A Bancos de dados NoSQL não podem ser indexados.

B Bancos de dados NoSQL são considerados banco de dados relacionais.

C Nos bancos de dados NoSQL devem ser definidos um esquema de dados fixo antes de qualquer operação.

D São exemplos de bancos de dados NoSQL: MongoDB, Firebird, DynamoDB, SQLite, Microsoft Access e Azure Table Storage.

E Os bancos de dados NoSQL usam diversos modelos para acessar e gerenciar dados, como documento, gráfico, chave-valor, em memória e, pesquisa.

03. (FUNDATEC - PGE RS - Técnico em Informática - 2021) Assinale a alternativa que apresenta SOMENTE bancos de dados não relacionais (NoSQL).

A Cassandra, MongoDB e Redis.

B lasticSearch, MongoDB e PostgreSQL.

C ElasticSearch, MySQL e Redis.

D MongoDB, Oracle e PostgreSQL.

E Cassandra, Oracle e Redis.

Gabarito 01/A; 02/A; 03/A