DATAB

— DICIONÁRIO —

ANÁLISE DE DADOS

KARINE LAGO
LAENNDER ALVES

APRESENTAÇÃO

Sabemos como é um **desafio** começar a estudar uma área completamente nova. Por isso, desenvolvemos esse guia com alguns dos termos mais utilizados na **Análise de Dados** como overview do que você vai eventualmente trabalhar ou discutir ao trabalhar com dados.

Esse é um ótimo passo para pesquisar por essas palavras e estudar profundamente aquela que achar interessante.

Desenvolvido com **carinho** por Karine Lago e Laennder Alves da **DATAB.**

DATAB

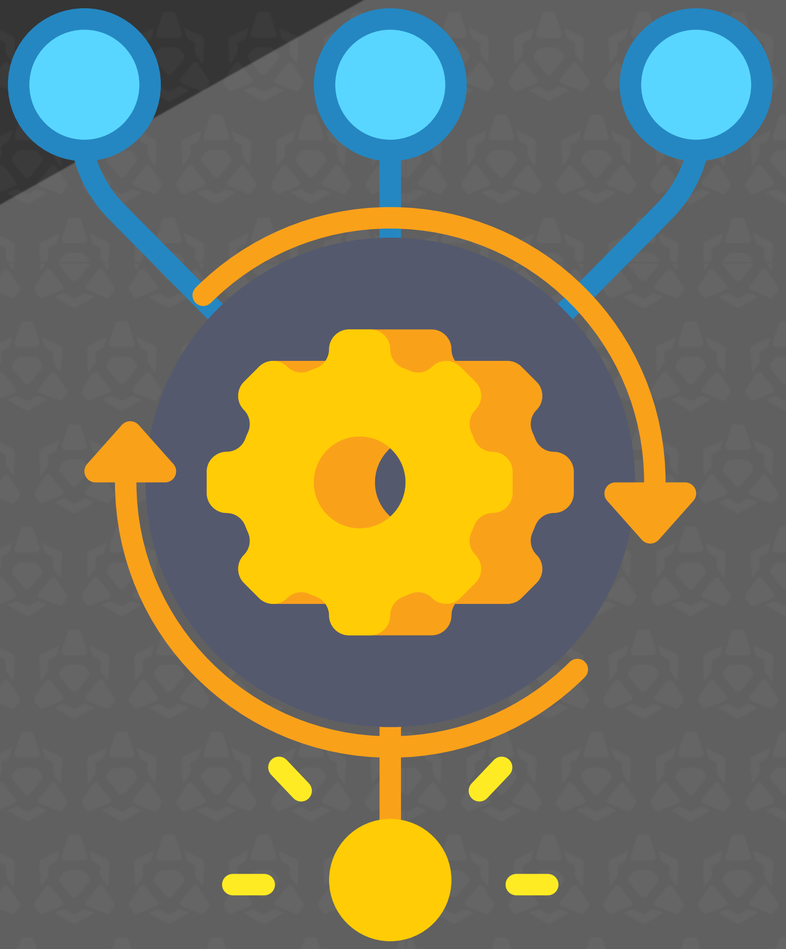
ÍNDICE

DATAB

ALGORITIMO

Sequência de ações ou instruções com um objetivo estabelecido para resolver algum tipo de problema. Eles são finitos, devem ser bem definidos e efetivos, sendo capazes de antecipar falhas.

A palavra algoritmo é muito associada a programação, mas na verdade são passos para resolver algo que não necessariamente está relacionado com computadores. Quando você está pesquisando no Google está utilizando algoritmos que buscam pela melhor informação para o que está procurando por meio de regras que definem o ranking e a relevância de uma página na internet, por exemplo.



ANÁLISE PREDITIVA

Técnicas que utilizam algoritmos estatísticos, machine learning e dados para identificar resultados futuros e prever a probabilidade de acontecerem para antecipar desastres, ações da concorrência, retornos esperados de projetos, fraudes, otimização de ações de marketing, previsão de churn e redução de riscos no geral.

A análise preditiva normalmente é realizada por um cientista de dados ou por estatísticos utilizando técnicas de regressão linear e logística, árvores de decisão, redes neurais, análise bayesiana, net lift, k-vizinhos (k-nn) e outras.

DATAB

API

Também conhecida como *Application Programming Interface*, a sigla API é uma interface de comunicação de dados e execuções entre serviços e quem precisa utilizá-los. As APIs podem ser abertas publicamente para todos consultarem ou fechadas, requisitando tokens (“senhas”) de acesso exclusivos.

Para acessar os dados são criados conjuntos de URLs que respondem com a informação requisitada. São os programadores da API que determinam as regras de como utilizá-las e cada uma terá sua própria característica e estrutura, podendo retornar os dados em CSV, JSON, XML e outras apresentações.

ÁRVORE DE DECISÃO

Na computação, a árvore de decisão é um tipo de algoritmo de aprendizagem de máquina com o objetivo de encontrar um conjunto com maior pureza utilizando os atributos para gerar a árvore. Elas são organizadas como se fossem fluxogramas e categorizam os dados em *nós e raízes*.

Por exemplo, você quer ir a praia, mas para isso precisa ter sol, estar calor e não ventar muito. Esse é o atributo ideal. Entretanto, algumas variações podem ocorrer: se estiver apenas sol, com temperatura amena e sem vento, você também pode ir, mas se estiver sol, frio e sem vento, você não vai. Isso é uma árvore de decisão (simples).

Elas podem ser desenvolvidas aplicando a conjuntos de dados mais complexos e, por isso são utilizadas linguagens como **Python** e **R** para treinar e gerar a árvore.

ATOMICIDADE

É uma característica ou uma política de bancos de dados sobre o comportamento consistente e preciso em relação ao armazenamento de informações e suas alterações. Uma ação é considerada como indivisível, ou seja, ou todas as regras ocorrem ou nenhuma delas ocorre.

Por exemplo, imagine um banco de dados de uma fábrica de automóveis. Se ao remover ou mover um carro do banco de dados (realizar uma transação), for necessário mover também a chave e o manual do proprietário, ou seja, 3 itens que estão registrados individualmente, mas devem ser movidos juntos, ou todos eles ocorrem em conjunto ou nenhum deles ocorre.

A atomicidade é uma das características do conjunto de propriedades de transações em banco de dados conhecida como **ACID** (Atomicidade, Consistência, Isolamento e Durabilidade).

DATAB

AZURE

É uma plataforma na nuvem com mais de 200 serviços com soluções de dados, segurança, internet das coisas, rede, armazenamento, web, blockchain, machine learning e muitos outros. Utiliza o modelo de SaaS (software as a service) com computação em grid que permite mais escala para grandes processamentos.

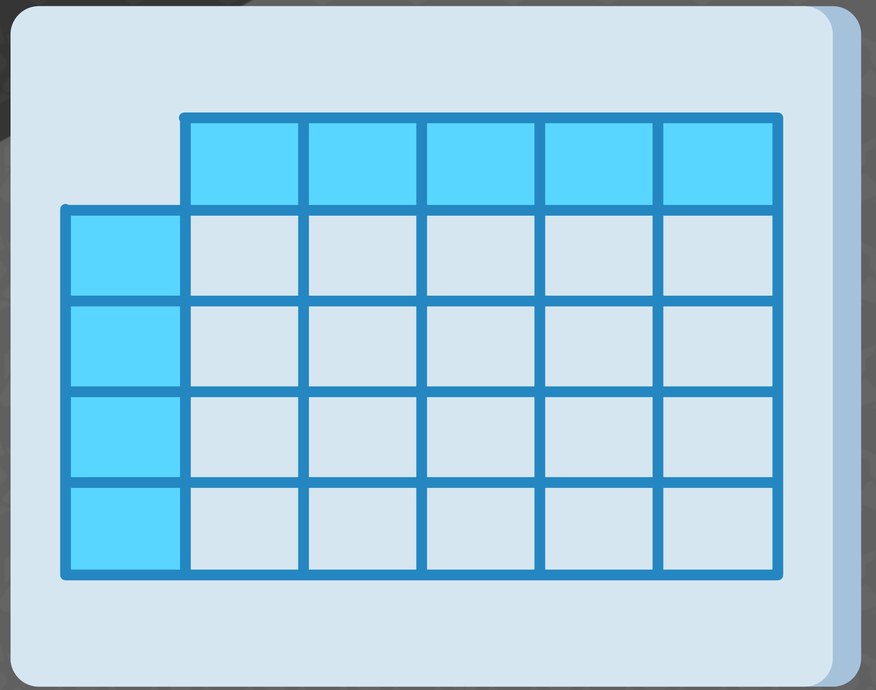
Para o analista de dados, a principal usabilidade são os bancos de dados e serviços como o **Power BI Embedded**, **Azure Data Factory** e **Cognitive Services** (Inteligência Artificial).

BALANCED SCORECARD

Também conhecida como BSC e em português pode ser traduzida como Indicadores Balanceados de Desempenho, essa metodologia é utilizada para gestão de desempenho de negócios, processos e projetos.

Ela possui alguns componentes definidos como o Mapa Estratégico, Objetivo, Indicador e Meta com perspectivas específicas voltadas para o financeiro, mercado, processos internos e aprendizado.

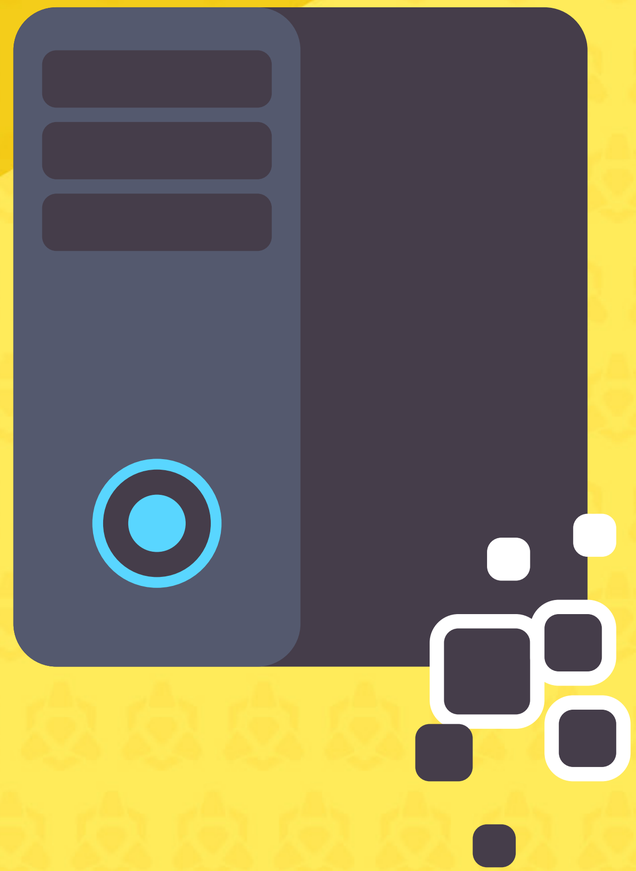
Utilizar essa metodologia beneficia no alinhamento estruturado do planejamento estratégico e dos objetivos definidos, podendo mensurar e gerenciar ações de forma sistemática.



BANCO DE DADOS

São coleções organizadas de dados armazenadas em um sistema de computador que pode ser controlado por softwares de gerenciamento (**SGBD**) como o SQL Server, Oracle, PostgreSQL, MySQL, MongoDB, Cassandra e até mesmo o Access. Praticamente todo sistema que você utiliza que retorna dados por meio de uma seleção ou busca utiliza um banco de dados por trás.

Eles podem ser **relacionais**, por exemplo utilizando a linguagem SQL para fazer consultas (buscas) por informações determinadas muito utilizados em ERPs e **não relacionais** (noSQL) para armazenar dados desestruturados como imagens, mapas e áudios.

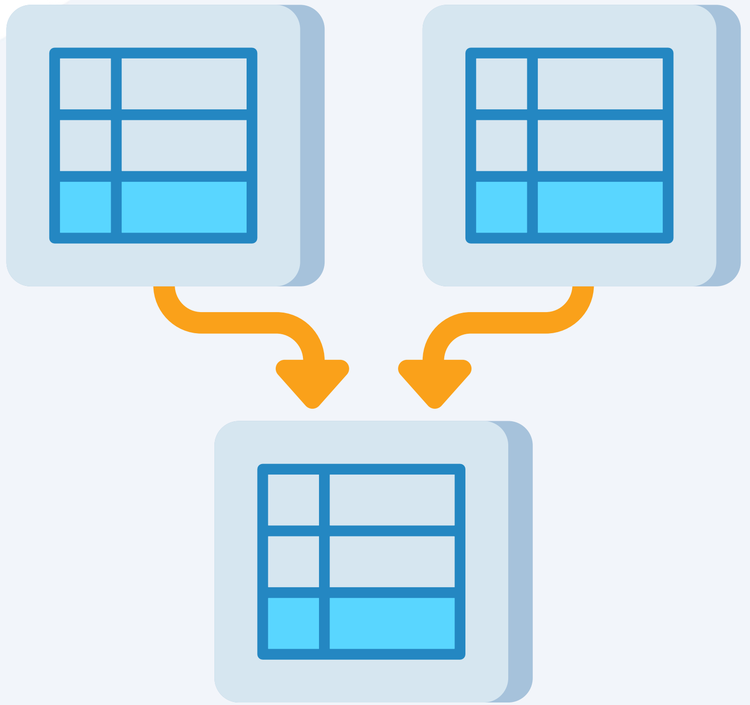


DATAB

BANCO DE DADOS RELACIONAL

É um tipo de banco de dados que armazena dados estruturados relacionados entre si onde cada linha da tabela é um registro com uma ID exclusiva chamada de chave e as colunas armazenam seus atributos. Essa organização favorece especialmente a relação entre as tabelas.

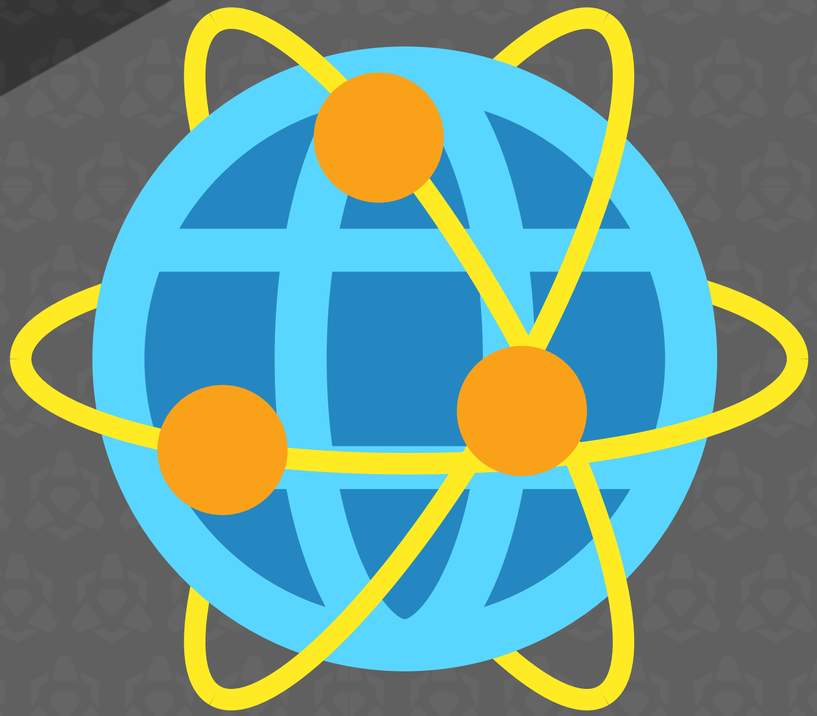
A linguagem SQL pode ser utilizada em um banco de dados relacional para fazer buscas, cálculos e filtros, resultando em visões diferentes dos dados armazenados e sendo especialmente úteis para performance da análise de dados.



BIG DATA

É um conceito com propriedades bem definidas que descreve a enorme quantidade de dados (estruturados ou não) que são gerados por softwares, sistemas e processos em um intervalo de tempo muito pequeno. Para algo ser determinado como Big Data, ele precisar ter:

Volume: por exemplo, mas não necessariamente, petabytes de dados armazenados.
Variedade: diversas fontes de dados aumentando a complexidade.
Velocidade: os dados são gerados o tempo inteiro, requisitando agilidade em suas tratativas.
Veracidade: como são muitas informações, é necessário certificar-se que todas são verídicas.
Valor: os esforços gerados para obter e tratar esses dados devem servir para algo, ter algum valor de aplicação, serem úteis.



BUSINESS ANALYTICS

Também conhecido de forma abreviada como BA, o Business Analytics aplica modelos e desenvolve simulações de cenários para prever situações futuras tentando identificar padrões e tendências.

As ferramentas utilizadas no BA vão desde softwares de BI, visualização de dados, análise estatística e plataformas de big data com o objetivo final de realizar análises descritivas, preditivas e prescritivas (recomendações sobre como lidar com situações que aconteceram no passado e podem acontecer novamente no futuro) por meio de dados históricos de algoritmos estatísticos.

DATAB

BUSINESS INTELLIGENCE

O Business Intelligence (BI) é semelhante ao BA, mas difere pelo foco que tem em análise de resultados e performances de acordo com o que foi previsto no planejamento estratégico. Ele ainda pode usar estatística em suas análises, mas é menos frequente e tem mais atuação em obtenção, organização e tratamento (ETL), modelagem, análise, visualização e compartilhamento de informações. Normalmente são utilizados softwares de BI como o Power BI, Tableau e Qlik.

Por ser uma área que trabalha com o planejamento estratégico, normalmente ela está muito ligada a cargos executivos em empresas e deve conhecer do negócio e da cadeia de valores de forma abrangente.

CHAVE PRIMÁRIA

A chave primária é o atributo mais básico de organização em bancos de dados. Toda tabela deve ter uma chave primária (e apenas uma) que tem como característica não ter valores repetidos ou nulos. É o ID único que se relaciona dentro da tabela com todas as outras informações que estão na mesma linha e em outras colunas.

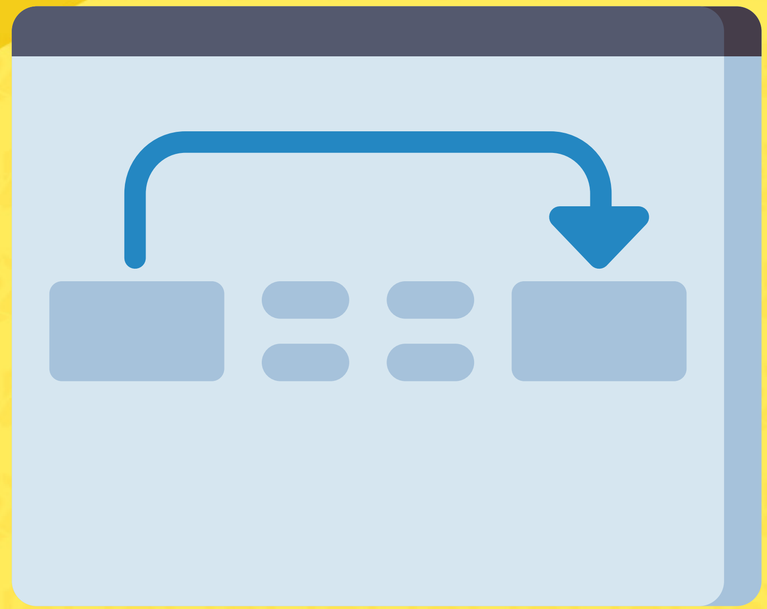
Por exemplo: No Brasil o CPF é um dos códigos únicos para cada pessoa da população, portanto, podemos dizer que ele pode ser usado como chave primária em uma tabela que armazena suas características.

A importância da chave primária em bancos de dados relacionais está diretamente relacionada a identidade dos dados e a possibilidade de relacionar com outras chaves primárias e estrangeiras de tabelas secundárias.

CHAVE ESTRANGEIRA

Uma chave estrangeira é uma coluna/campo que possui os mesmos itens (ou alguns deles) de uma chave primária, mas que está em outra tabela. Podem ter valores duplicados e estão relacionados com essa chave primária. Esses itens precisam fazer referência (ser igual) entre eles para que o relacionamento funcione corretamente.

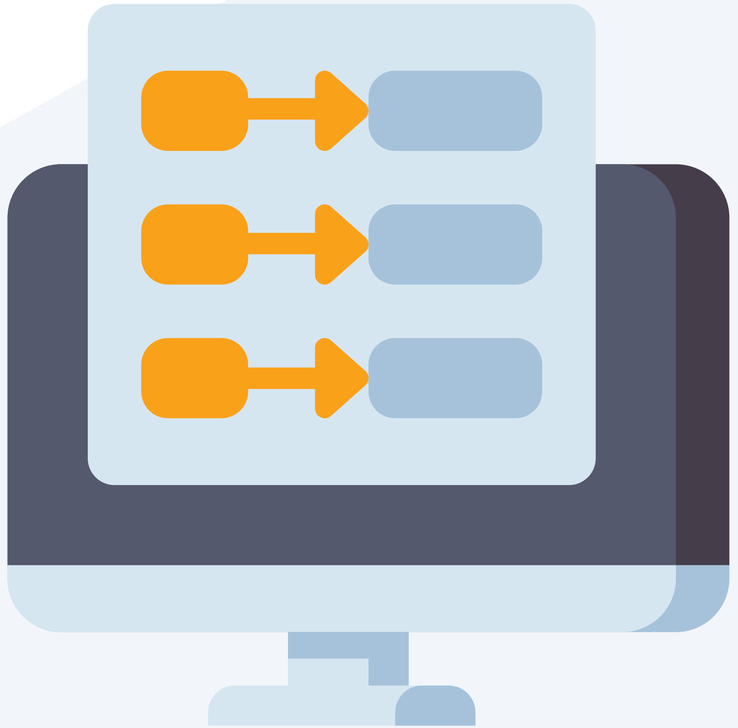
Caso tenha alguma falha de integridade nesses dados, por exemplo, ter uma chave estrangeira que não está presente na chave primária, os valores serão considerados como “em branco” ao usar a chave primária como filtro para exibir valores usando essa relação.



DATAB

CHAVE COMPOSTA

A chave composta é uma combinação de mais de uma coluna em uma tabela. É utilizada quando apenas uma coluna não consegue ter os atributos de uma chave primária com valores únicos para identificar cada linha. Caso sejam utilizadas chaves compostas, as chaves estrangeiras em outras tabelas que se conectam com elas também precisam ser compostas.



Um exemplo pode ser observado em nomes de bairros. Eles se repetem entre cidades e os nomes de cidades também vão se repetir para mostrar cada bairro que possui em uma tabela. Para ter uma chave única, basta juntar o bairro com a cidade para não ter repetições e criar uma chave primária composta. Esse foi apenas um exemplo didático, normalmente são usados códigos para identificar essas informações textuais.

CIENTISTA DE DADOS

É um profissional da área de ciência de dados que analisa dados estruturados e desestruturados, big data e históricos com a extração de informação por meio da detecção de padrões utilizando estatística, matemática e programação (Python, R, Julia, Scala, Java), algoritmos, visualização de dados, Machine Learning, mineração de dados, deep learning, Linguagem Natural (NLP).

Algumas das ferramentas que podem ser utilizadas são: SAS, Matlab, Azure e IBM Watson Analytics. Conhecer sobre banco de dados e assuntos relacionados também facilitará o trabalho, como SQL, NoSQL, Hadoop, Spark, entre outros.

CRM

CRM é a sigla para Customer Relationship Management que significa Gestão de Relacionamento com o cliente. É uma técnica de gestão e um tipo de software que auxilia na visualização da caminhada de um cliente dentro de uma empresa e engloba todas as suas informações no sistema, como reclamações, pedidos, faturamento, contatos, propostas, etc. A inteligência na gestão dessas informações e facilidade em lidar com os problemas dos clientes está diretamente relacionada com sua satisfação.

O banco de dados dos sistemas CRM é muito utilizado em análises de dados para compreender o comportamento dos clientes e gerar insights para modificar produtos, preços e serviços.

DATAB

DADOS DESESTRUTURADOS

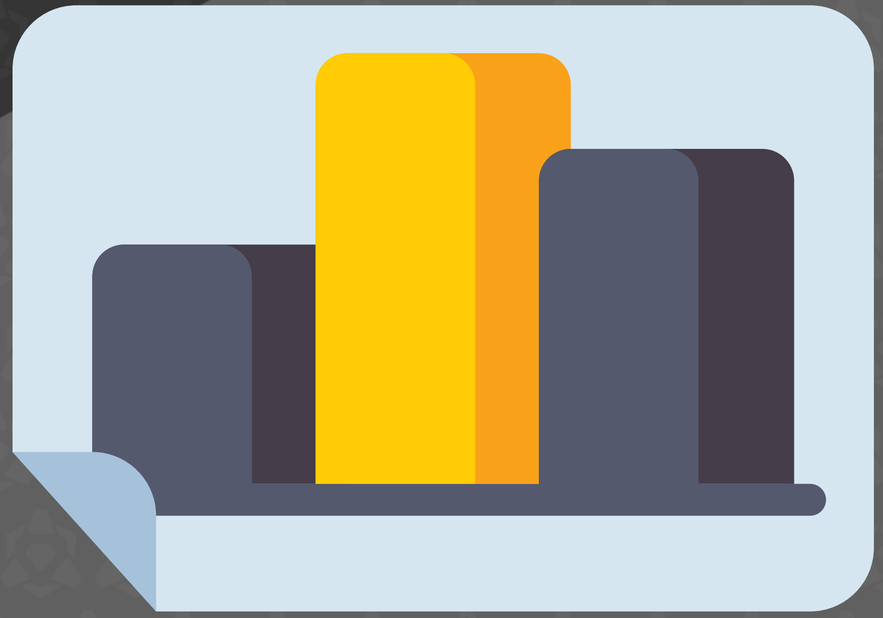
Dados desestruturados ou não-estruturados são completamente diferentes de dados estruturados, pois não possuem regras delimitadas de armazenamento padronizadas que podem ser utilizadas em banco de dados relacional, por exemplo.

Dados desestruturados podem ser identificados como imagens armazenadas, áudios, mapas, vídeos, textos em parágrafos sem organização, dados de redes sociais, entre outros. Eles ainda podem ser analisados por mecanismos de Inteligência Artificial e Machine Learning, mas são praticamente impossíveis de serem analisados puramente sem um processamento estrutural antes, principalmente se forem em grande volume.

DASHBOARD

Dashboard é um produto da análise de dados que reúne informações resumidas de indicadores e métricas sobre uma determinada área ou empresa. Eles devem ser utilizados para transmitir uma visão global sobre algo que deve ser acompanhando, como seguir o planejamento estratégico ou o resultado de projetos com o objetivo de facilitar a tomada de decisão pelos gestores.

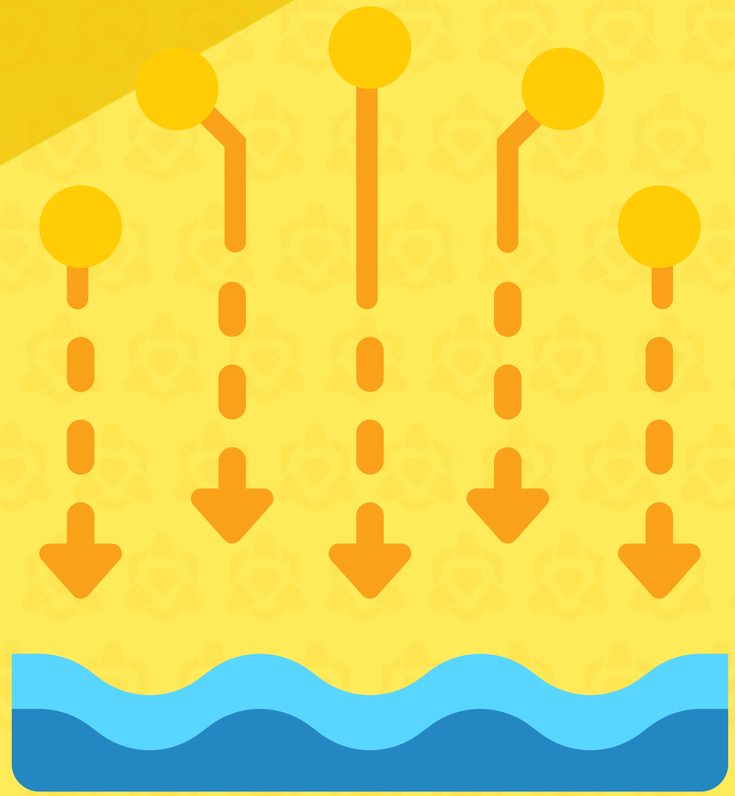
Os dashboards podem ser operacionais, táticos e estratégicos com alta disponibilidade das informações, por isso muitas empresas escolhem exibi-los em TVs como uma *gestão à vista*.



DATA LAKE

Data Lake é um repositório que centraliza dados em estado bruto gerados por uma empresa ou processo, muitas vezes sem regras pra simplificar o processo de armazenamento dessas informações. Ele oferece uma visão não refinada dos dados.

Como normalmente são **dados desestruturados**, é necessário realizar um trabalho de **normalização** para estruturar e possibilitar a análise por cientistas de dados. Eles podem ser armazenados em serviços como o Azure Data Lake e AWS.



DATAB

DATA MART

É um ou vários segmentos de um banco de dados classificados por assuntos. Também pode ser caracterizado por um subconjunto de dados. Um **Data Warehouse** armazena dados de diversas áreas, já um **Data Mart** foca apenas em uma área específica, como Produção, Financeiro, Marketing, etc. Normalmente é utilizado para suprir a necessidade de um desenvolvimento específico e focado.

Essa abordagem simplifica projetos de análise de dados mostrando apenas informações relevantes e direcionadas.

DATA MINING

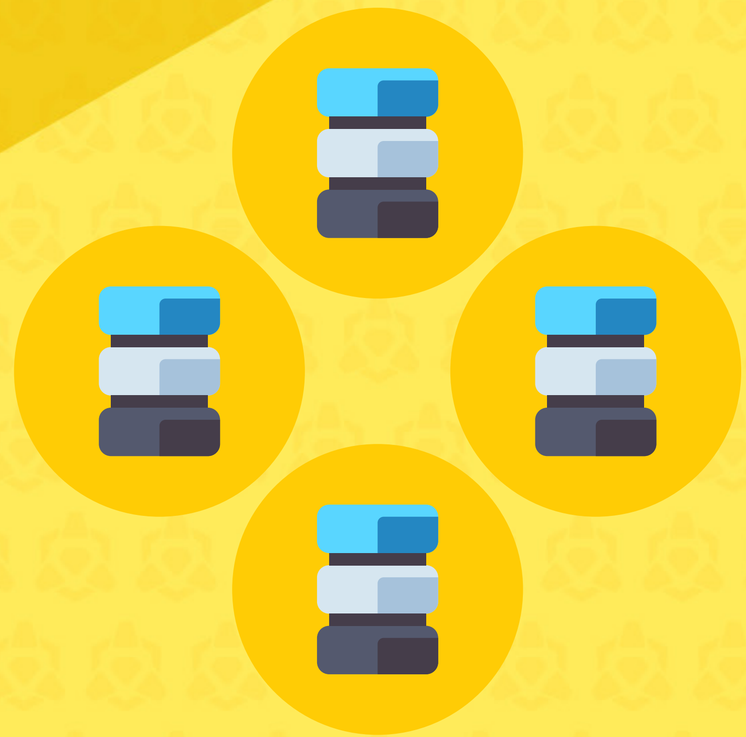
Também conhecido como mineração de dados, o data mining é um processo que explora diferentes padrões de dados tentando encontrar relacionamentos e conjuntos para extração e posterior utilização em análises.

Muitas ferramentas podem ser utilizadas, como R, Python e algumas de suas bibliotecas, Rapid Miner, Oracle Data Mining e técnicas de Machine Learning para identificar corretamente os dados que devem ser extraídos. No processo de análise de dados, o Data Mining se encaixa nos primeiros passos de **obtenção** e **tratamento de dados**.

DATA WAREHOUSE

Um Data Warehouse é um armazém de dados que consolida informações de banco de dados transacionais, que muitas vezes são demasiadamente normalizados com muitas tabelas dimensão com o propósito de otimizar a performance e alocação de espaço dos dados registrados, o que dificulta o processo de análise de dados em softwares de BI.

Como o Data Warehouse é uma organização que pode ser definida para fins analíticos, seu uso é otimizado, pode ser de-normalizado e ainda serve de base paralela ao banco transacional, separando acessos e propósitos de conexões.



DATAB

DAX

DAX é um acrônimo para Data Analysis Expressions, definida por uma biblioteca de funções para criação de métricas no Power BI, Power Pivot no Excel e SSAS. São mais de 270 funções que utilizam o conceito de contextos de filtro dependendo de onde são aplicadas (em medidas, colunas ou tabelas).

As funções DAX são parecidas com funções do Excel, mas extremamente poderosas e com a particularidade de aproveitar contextos dos visuais e tabelas. O DAX é diretamente dependente da organização dos dados e dos relacionamentos entre as tabelas.

DRILL DOWN E DRILL UP

Drill Drown e Drill Up é a possibilidade de detalhamento em grãos maiores e menores das informações exibidas em uma tabela ou gráfico.

Quando um drill down é realizado, informações mais detalhadas são exibidas. E no drill up, informações mais agrupadas são mostradas. A possibilidade de disponibilizar para os usuários uma visão mais detalhada sem disponibilizá-las abertas em um dashboard ou relatório é ideal para a independência da tomada de decisão e investigação de resultados.

ELT

ELT significa Extract, Load e Transform. Alguns especialistas dizem ser a evolução do tradicional ETL, (embora há discordância nessa afirmação) que faz a etapa de transformação (transform) antes do carregamento (load).

O conceito sugere que os dados sejam carregados e disponibilizados para que o processo de transformação de dados seja realizado no sistema onde foram carregados, o que demanda mais poder de processamento e isso deve ser avaliado com cuidado no momento de escolher esse processo.

DATAB

ERP

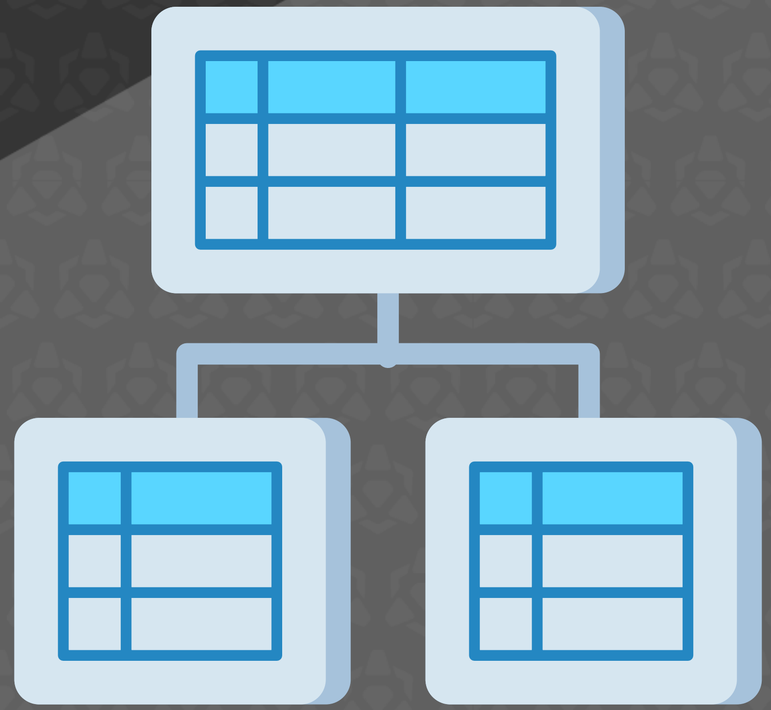
ERP é um acrônimo para *Enterprise Resource Planning*, que é o mesmo que **Planejamento de Recursos Empresariais**, que são sistemas responsáveis por interligar toda a operação de uma empresa e gerir as transações registradas por cada área responsável. Ele alimenta o banco de dados com a operação realizada. Sabendo disso, a importância de uma boa arquitetura e coleta de informações estruturadas deve ser uma das principais características para escolher ERPs.

Existem diversos deles, desde grande porte como **SAP**, **Oracle** e **Totvs**, como de pequeno porte como o Conta Azul. Eles podem ser instalados localmente ou com acesso exclusivo via web (armazenado na nuvem), que é a tendência atual.

ESQUEMA ESTRELA

Também conhecido em inglês como *Star Schema*, o esquema de estrela é um conceito de organização de dados criado por **Ralph Kimball** para modelar os dados com dimensões limitadas, pois redundâncias são permitidas em tabelas dimensão. É um esquema utilizado majoritariamente em **data warehouses**.

O nome *estrela* faz referência ao design de ter uma **tabela fato** no meio rodeada por **tabelas dimensões**. Essa única tabela fato não pode ter redundâncias, e é necessário ter uma tabela por dimensão altamente desnormalizada (com redundâncias) e dimensões adicionais conectadas a ela.



ESQUEMA SNOWFLAKE

Também conhecido como **Floco de neve**, a organização de dados e tabelas no esquema **Snowflake** tem esse nome por possuir várias dimensões em cascata que se relacionam com a tabela fato.

Sua organização de dados é **normalizada**, isso significa que as redundâncias são eliminadas para minimizar a quantidade de dados armazenados. Entretanto, quanto mais normalizado, mais complexo fica o modelo de gerenciar por possuir muitas tabelas e relacionamentos.

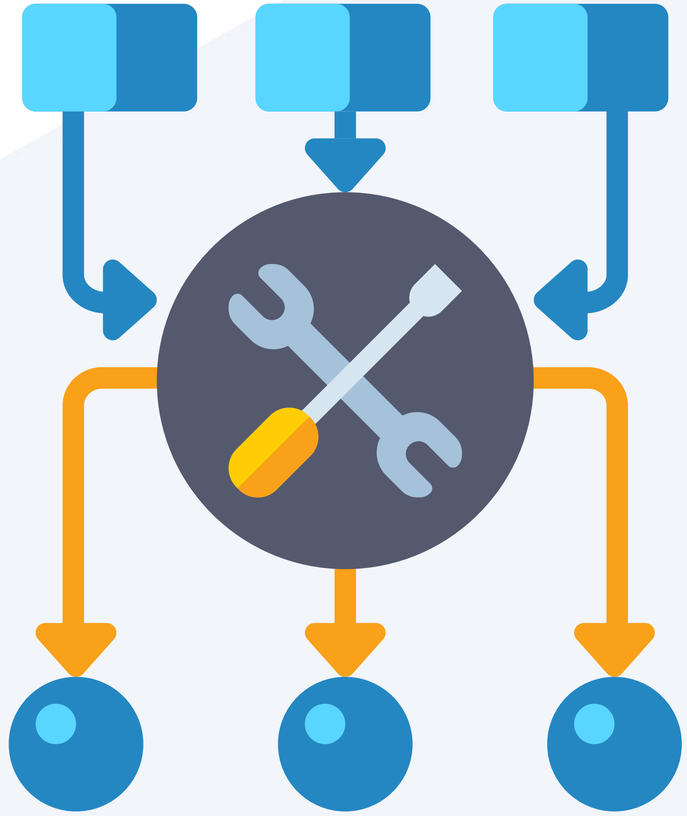
DATAB

ETL

ETL é um acrônimo as palavras em inglês: *Extract, Transform e Load*, e em português, Extrair, Transformar e Carregar. É um processo onde os dados são retirados de uma fonte, tratados para que possam ficar na organização necessária para analisar e, posteriormente, carregados no sistema de análise.

Alguns dos softwares que realizam ETL é o **Power Query** (suplemento no Excel e Power BI), **Tableau Prep**, **Pentaho, Azure Data Factory, SSIS**, entre outros.

O tratamento de dados é parte fundamental do processo de análise de dados e consideramos como um dos principais processos de modelagem para consistência e performance de projetos de análise de dados.

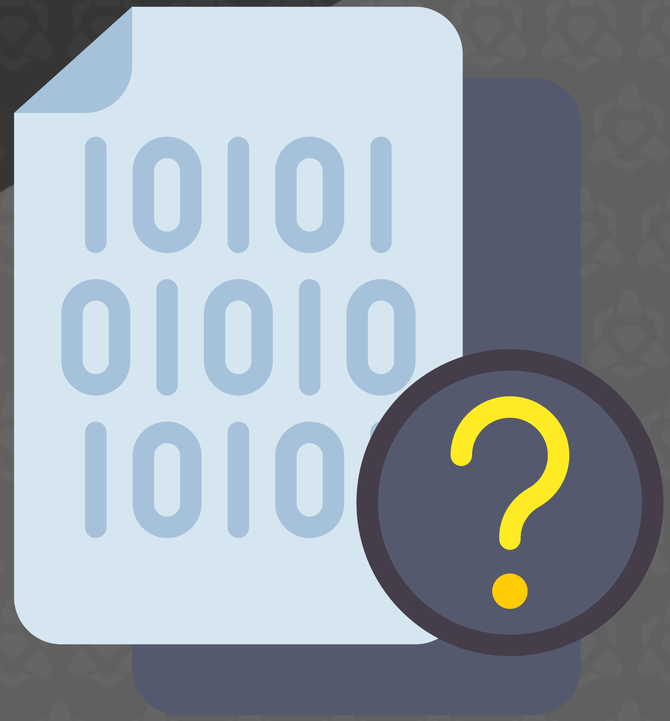


FONTE DE DADOS

Fonte de dados são locais que armazenam dados em diferentes extensões e formas. Ela é o principal material da análise de dados e sua origem pode determinar a quantidade de ETL necessário para tratá-los.

Alguns exemplos de fontes de dados são: Excel, Banco de Dados, Sharepoint, CSV, XML, JSON, TXT, Web, PDF, Imagens, entre outras. Projetos de BI podem ser híbridos e terem diferentes fontes de dados, já que todos são organizados e transformados em tabelas semelhantes.

Os dados podem ser estruturados, como um banco de dados, semi-estruturados, como em um XML e desestruturados, como imagens. Quanto mais estruturado, menor será a necessidade de processamento e a complexidade.



FRONT-END E BACK-END

Em projetos de tecnologia, podemos dividi-los resumidamente em duas camadas distintas: o front-end, que é a camada que os usuários vão interagir e o back-end que processará as execuções realizadas.

O back-end é o que tem por trás de uma aplicação, como os comandos e lógicas que podem ser desenvolvidos em PHP, C#, Java, entre outras linguagens dependendo do tipo de aplicação desenvolvida.

O front-end é a interface gráfica e uma das principais preocupações é com a comunicação clara com o usuário, por isso se fala tanto em UI (User Interface) e UX (User Experience). Ela pode ser desenvolvida em CSS, JavaScript e outras linguagens.

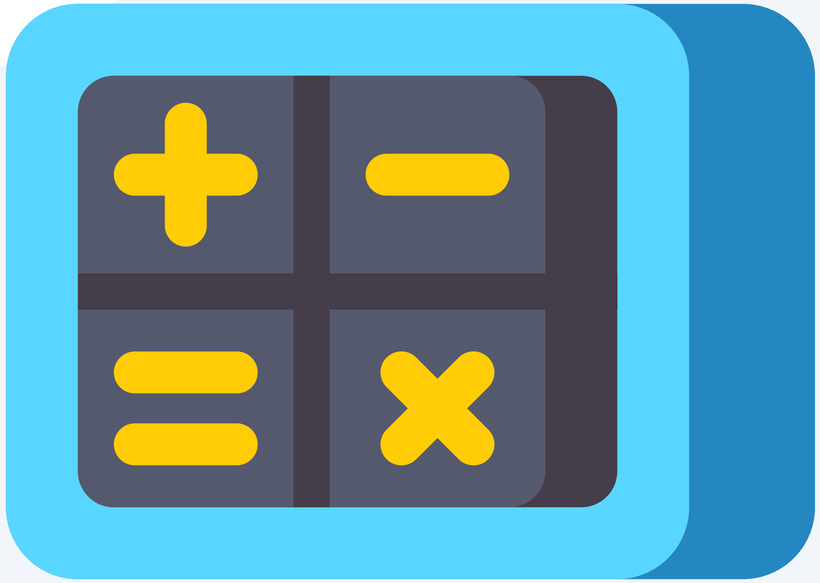
DATAB

FUNÇÃO

Funções são conjuntos de comandos que realizam uma tarefa especificada sem a necessidade de declarar individualmente cada passo de execução. Elas possuem nomes que armazenam esses comandos de acordo com a função chamada.

Um exemplo simples de função é a média representada pela AVERAGE no Power BI. Ela armazena o cálculo de soma e a divisão pela quantidade deles (média aritmética). Ao invés de fazer três passos manualmente: somar, contar e dividir, com a AVERAGE basta preencher os argumentos nas ordens especificadas em sua sintaxe e a instrução interna executará os passos automaticamente.

Funções não podem ser confundidas com fórmulas. A fórmula de média é SUM(X) / COUNT(X) com os passos de cálculos manualmente declarados.



GARTNER

A Gartner é uma empresa de consultoria independente que executa pesquisas comparativas em funcionalidades de diversos tipos de softwares no mercado como banco de dados, Business Intelligence, CRM, ERPs, entre outros. Ela é conhecida por publicar anualmente os quadrantes mágicos que posicionam as principais empresas das categorias de softwares entre aqueles que são líderes, visionários, desafiadores e de nicho.

Para análise de dados, em suas publicações anuais os principais softwares de análise de dados há alguns anos são o Power BI, Tableau e Qlik, com alguns entrantes diferentes eventualmente. Vale a pena acompanhar as publicações para ficar atualizado sobre o mercado e seus líderes.

GOVERNANÇA DE DADOS

É o gerenciamento de dados de uma empresa focado na segurança da informação e sua disponibilidade. Ela é realizada por meio de políticas, processos e monitoramento garantindo que informações não sejam repassadas para pessoas que não deveriam acessá-las e que sejam corretamente registradas e armazenadas.

O BI pode, inclusive, ajudar em seu gerenciamento por meio da análise de dados gerados pela operação e definição de indicadores definidos para a governança.

DATAB

INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é uma solução que envolve o agrupamento de várias tecnologias como Machine Learning, Algoritmos e Redes Neurais para simular capacidades humanas de inteligência. Ela pode auxiliar especialmente na organização estruturada de dados desestruturados como identificação de objetos em imagens, qual o sentimento de um texto, a percepção da idade de uma pessoa, o idioma que ela está falando e outras explicações.

Seus resultados são baseados em treinos prévios e possuem, normalmente, uma confiança de quão certo está o resultado mostrado de uma avaliação. Para utilizar a inteligência artificial você pode contratar serviços como o **Azure Cognitive Services** e o **IBM Watson**.

ÍNDICE

Na estrutura de dados, **Índices** são chaves que permitem a rápida localização de registros (linhas) em uma consulta. Sua função é acelerar o processo de busca e acesso dos dados solicitados. Um exemplo para compreendê-los é imaginar que todo livro possui um índice para rápida localização dos assuntos com a página em que eles estão.

Os índices normalmente são números sequenciais que identificam o registro, sendo mais eficientes ler esses números onde cada posição é conhecida do que ler e tentar localizar textos e números aleatórios (representados pelos dados armazenados).

KPI

KPI é um acrônimo para *Key Performance Indicators*. Conhecido em português como Indicador-Chave de Desempenho. Eles resumem e medem objetivos do planejamento da empresa e são acompanhados por softwares de análise de dados e muito aplicados em dashboards que armazenam uma visão rápida e geral sobre o estágio atual dos objetivos definidos.

Cada tipo de negócio terá KPIs específicos, mas alguns são comuns entre empresas, como o atingimento de meta de faturamento, devoluções, reclamações e custos.

KPIs precisam ser objetivos, mensuráveis, verificáveis, ter valor, consenso entre os participantes e comprometimento dos envolvidos.

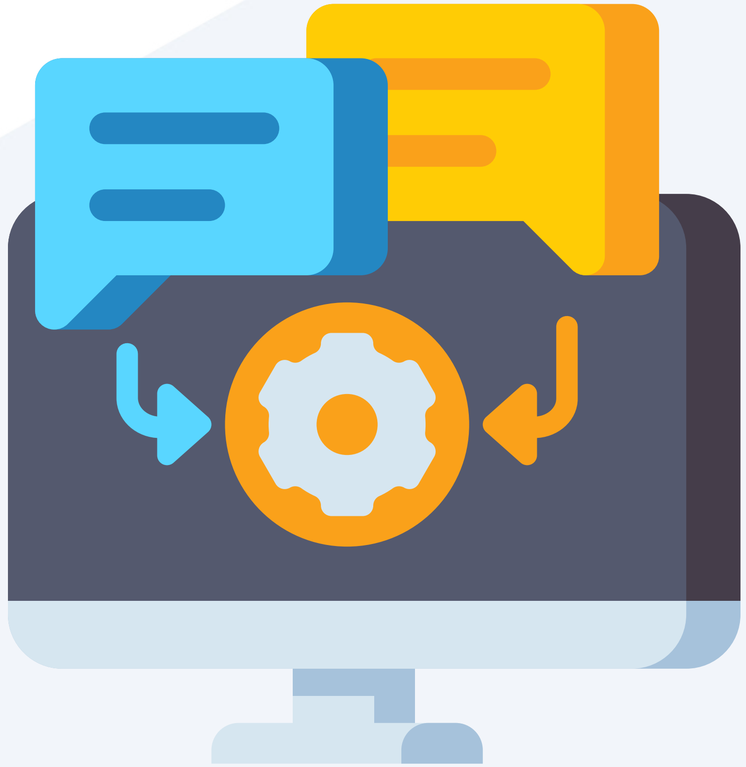
DATAB

LINGUAGEM NATURAL (PNL)

Também conhecida com o PNL ou NLP (*Natural Language Processing*), é uma técnica de simulação computacional da linguagem humana criada por algoritmos que identificam conteúdos e constroem ou identificam frases.

Ela é utilizada em chatbots capazes de conversar com humanos, visualização de dados com resumo analítico do histórico, e compreensão de solicitações escritas e faladas.

Exemplos do dia a dia de PNL são as assistentes virtuais Siri (Apple), Alexa (Amazon) e Cortana (Microsoft).



LOG

É um arquivo produzido automaticamente por softwares com o registro dos eventos que aconteceram. São frequentemente utilizados para registrar erros e possibilitar uma comunicação mais técnica com os desenvolvedores e responsáveis pelo software. Eles também podem ser utilizados como provas digitais de execuções e auditoria.

Sua extensão normalmente é .log ou .txt, mas também podem apresentar outros formatos. A grande maioria pode ser aberto em editores de textos como o bloco de notas.

MACHINE LEARNING

Machine Learning (ML) é um sistema que modifica automaticamente o comportamento com base em um aprendizado anterior. Por isso que ouvimos falar em treinar um modelo, quando ele é submetido repetidas vezes a um processo e dados, o sistema é capaz de aprender com os erros anteriores para encontrar o melhor caminho após um número variado de tentativas.

Um dos exemplos mais utilizados de machine learning é o combate a fraudes em sistemas de pagamento, recomendação de conteúdo na web e em serviços de streaming como a Netflix e os carros autônomos.

Como é necessário alto processamento, podem ser utilizados serviços como o Azure ML, Amazon ML e até mesmo o Python.

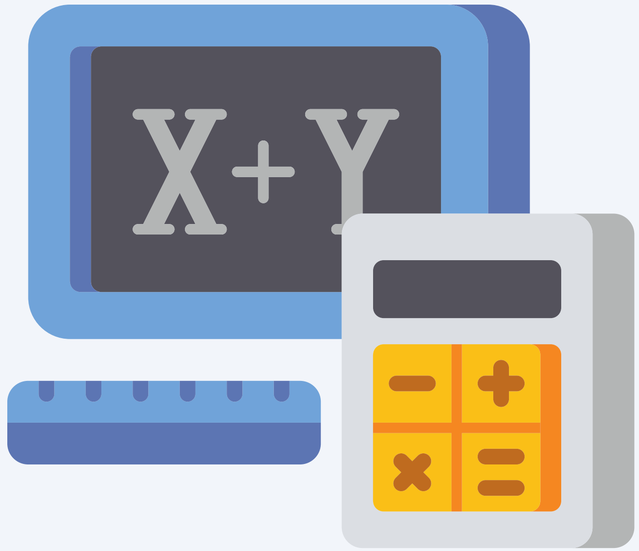
DATAB

MDX

MDX é um acrônimo para Multidimensional Expressions, que é uma linguagem de consulta conhecida informalmente como o *SQL para o OLAP* (que armazena dados em múltiplas perspectivas, ou seja, no conceito de cubo).

Consultas em MDX constroem tabelas e por utilizar um motor de dados multidimensional, é extremamente rápida. Seu principal propósito não é alterar o banco de dados como um SQL é capaz, mas escrever comandos que filtram os valores conforme a necessidade de visualização.

O MDX é amplamente utilizado no SQL Server Analysis Services e tem muita semelhança com o DAX, apesar de mais complexo e antigo, porém ainda relevante.



MEDIDAS / MÉTRICAS

Medidas ou métricas são termos utilizados amplamente em softwares de Business Intelligence para representar um conjunto de expressões que realizam cálculos para extrair informações de dados e posteriormente visualizá-las.

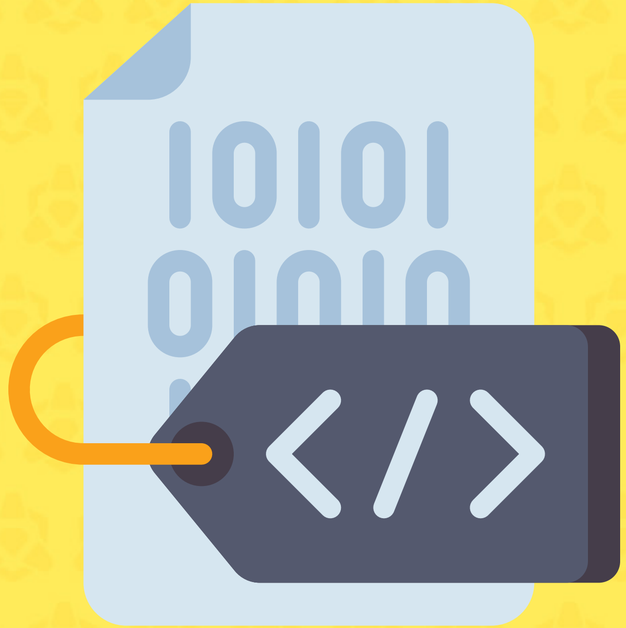
Elas podem ser consideradas como variáveis armazenadas que são calculadas no momento da utilização e são mais rápidas do que realizar o cálculo visivelmente direto na tabela de dados.

No Power BI, para criar medidas é utilizada a linguagem **DAX** com uma bibilioteca de mais de 270 funções.

METADADOS

Pode parecer redundância, mas metadados são dados sobre outros dados. Ele armazena informações resumidas sobre a utilidade de um dado e são utilizados para comunicação entre computadores e sistemas sem a necessidade de avaliar seu conteúdo.

Um exemplo de metadados é quando você gera uma planilha e observa suas propriedades. Serão registrados dados do autor, hora de criação, modificação, local, tamanho, nome e outras informações.

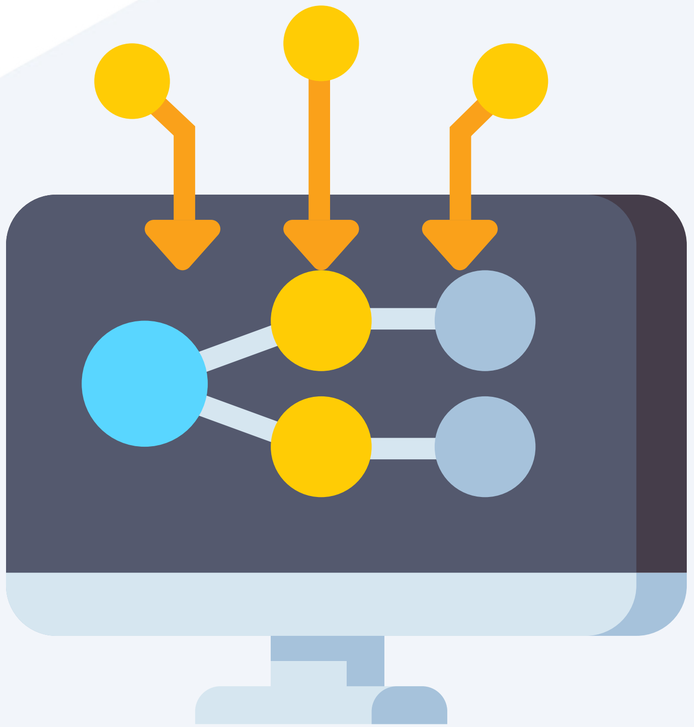


DATAB

MODELAGEM DE DADOS

É uma atividade para alterar os dados até o formato desejado para utilizá-lo em softwares, facilitando sua usabilidade, entendimento e aplicação no objetivo do projeto.

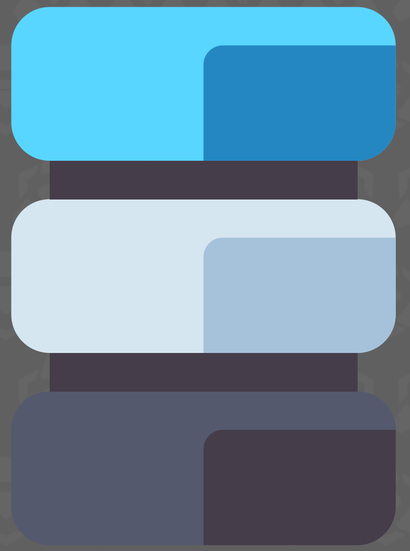
Em Business Intelligence, a modelagem tem um papel fundamental de organização das tabelas para promover o relacionamento e a correta visualização dos dados desejados, evitando a necessidade de criação de cálculos complexos demais que podem ser extinguidos em alguns momentos ao optar por modelar os dados da forma como precisa antes de calculá-los.



MONGODB

MongoDB é um popular software de banco de dados escrito em C++ de código aberto orientado a documentos, bem diferente dos tradicionais bancos com modelo relacional como o SQL. É um banco do tipo NoSQL que suporta linguagens como JavaScript, Python, C#, C++, Ruby, entre outras para fazer consultas e agregações.

Ele foi criado especialmente para armazenamento de big data, já que lida com documentos formados por uma multi-plataforma com conjuntos de JSON. Outros bancos NoSQL semelhantes são: Hadoop, Cassandra, DynamoDB, Amazon SimpleDB, entre outros.



MYSQL

MySQL é um sistema SGBD que utiliza a linguagem SQL para realizar consultas. Por ser open-source (código aberto) e por suas características, ele se tornou muito popular. É um banco de dados relacional com armazenamento de dados em tabelas.

Sua licença é baseada é do tipo GPL (General Public License), isso significa que o usuário tem a liberdade de usar o programa para qualquer propósito, de adaptá-lo, redistribuir cópias e aperfeiçoa-lo.

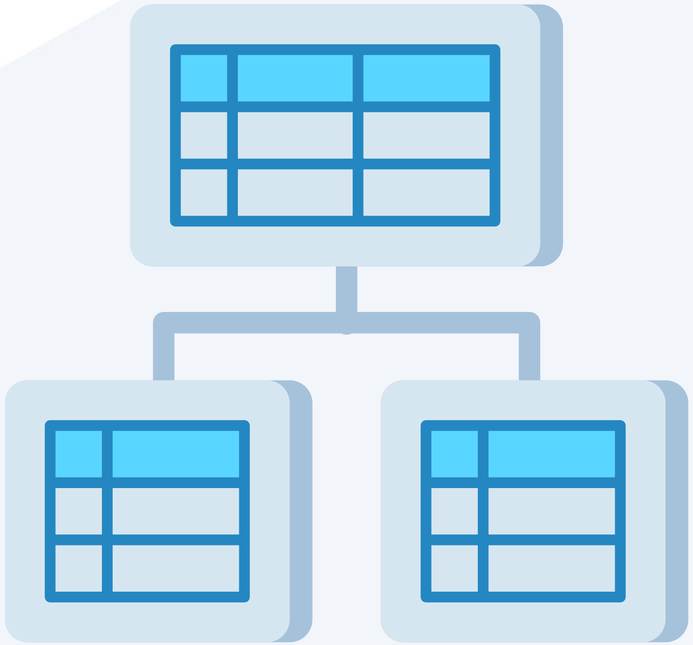
Como sua plataforma é de código aberto, várias aplicações podem surgir dele. Isso é chamado de *fork* (garfo). Uma delas é o MariaDB que possui algumas diferenças consideradas como melhorias em relação ao MySQL.

DATAB

NORMALIZAÇÃO

A normalização é um dos processos fundamentais de banco de dados relacionais. Ela projeta a forma como os dados serão armazenados e qual o nível do grão e redundâncias terão. Ela é um conjunto de regras organizadas em *formas normais* abreviadas como 1FN, 2FN, 3FN, 4FN e 5FN. Quanto mais normalizado é o armazenamento de dados, mais tabelas terão para garantir que as regras das formas sejam obedecidas.

Na análise de dados, dependendo dos projetos, não é indicado trabalhar nem na 1FN e nem na 5FN, que são extremos da normalização e não normalização. O ideal é equilibrar o seu projeto para ter a normalização correta garantindo os relacionamentos necessários entre dimensões e fatos para facilitar a análise.



NOSQL

É um termo que representa banco de dados não relacionais. Como o próprio nome já diz, ele não utiliza o SQL para fazer consultas. Banco de dados desse tipo são muito utilizados em big data, já que são orientados a documentos.

Ele é horizontalmente escalonável, permitindo tráfego por particionamento dos dados (sharding). O dimensionamento de bancos desse tipo normalmente é mais barato, pois diferentemente dos bancos relacionais, eles não exigem um único servidor para hospedar.

Alguns bancos desse tipo são: MongoDB, Cassandra, Cosmos DB, Redis, Berkeley DB, HBase, entre outros.

ODBC

ODBC significa Open Database Connectivity, que é um protocolo utilizado para importar dados de banco de dados. Esse tipo de conexão é largamente utilizada em softwares de BI quando necessitam fazer importações de bancos que não possuem conectores próprios, gerando independência e viabilizando a comunicação.

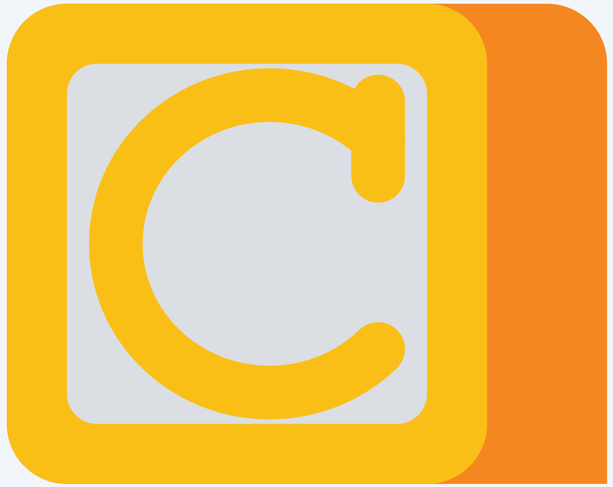
Esse protocolo utiliza drivers que atuam como um tradutor entre o software e o SGBD.

DATAB

OLAP

OLAP é um acrônimo para *Online Analytical Processing* que é um conceito que aplica a capacidade de manipular e analisar dados de forma multidimensional, também conhecido como cubo. Enquanto o data warehouse armazena os dados, o OLAP é o responsável por consultá-los de forma eficiente.

Exigem alguns métodos de armazenamento desse tipo: ROLAP (relacional), MOLAP (multidimensional), HOLAP (híbrido) e DOLAP (desktop). Os dados são organizados em hierarquias com diferentes níveis de detalhes que podem ser rapidamente acessados por drill up e drill down detalhando seus níveis.



OLTP

Também conhecido como *Online Transaction Processing*, é a organização de sistemas responsáveis por registrar transações da operação de uma empresa. Enquanto a organização OLAP se preocupa com a análise da informação no nível estratégico, o OLTP se preocupa com a organização e registro da operação. Os dados normalmente ficam em um modelo relacional e com alto nível de detalhe.

Quando alguém entra em um sistema, cadastra um pedido que está solicitando um produto com um método de pagamento “x”, vendido pelo vendedor “João” e que será entregue em “y” dias, é uma transação que está sendo registrada de forma online no sistema.

Analisar dados organizados no sistema OLTP será mais lento e complexo do que analisar em sistemas organizados como OLAP.

ORACLE

A Oracle é uma empresa de tecnologia mundialmente conhecida por disponibilizar diversos tipos de produtos concorrentes de muitos outros que são oferecidos também pela Microsoft e IBM, como ERPs, armazenamento na nuvem, análise, redes e muitos outros. Um dos mais utilizados na análise de dados é seu banco de dados.

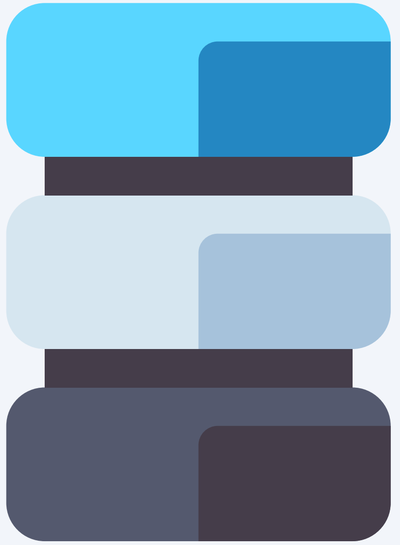
Assim como o SQL Server, ele também é do tipo relacional e a empresa criou a linguagem PL/SQL utilizada para fazer processamento de transações nos bancos da Oracle.

DATAB

POSTGRESQL

PostgreSQL é um sistema de gerenciamento de banco de dados objeto-relacional (ORDBMS) de código aberto e gratuito para uso não comercial que oferece funcionalidades modernas como chaves estrangeiras, functions, triggers, views, agregação e outros.

Em contraste com o MySQL, ele possuía mais funcionalidades no passado, mas o MySQL tem se atualizado e equilibrado nas diferenças. O PostgreSQL é indicado para aplicação com tipos de dados personalizados, como informações de metadados e geográficas.



POWER BI

Power BI é um software de Business Intelligence da Microsoft líder no mercado segundo a Gartner que utiliza conceitos de self-service BI com suplementos de ETL por meio do Power Query. Utiliza as linguagens M e DAX, com suporte para Python, R e SQL.

Ele se integra com a plataforma do Azure com possibilidade de utilização fluida de serviços de Machine Learning, Inteligência Artificial, AD, Banco de Dados na nuvem e muitas outras.

Possui três vantagens explícitas em relação aos seus principais concorrentes: preço, comunidade forte que produz conteúdo e facilidade no aprendizado.

POWER PIVOT

O Power Pivot é um suplemento do Excel que é derivado do SQL Server Analysis Services (SSAS). Com ele é possível utilizar modelagem, relacionamento de dados e a linguagem DAX no Excel, adicionando um poder de processamento e armazenamento de dados nunca antes visto no Excel. Ele surgiu em 2010 e posteriormente a instância do SSAS derivou para o Power BI de forma integrada, com a mesma facilidade para analistas de negócios.

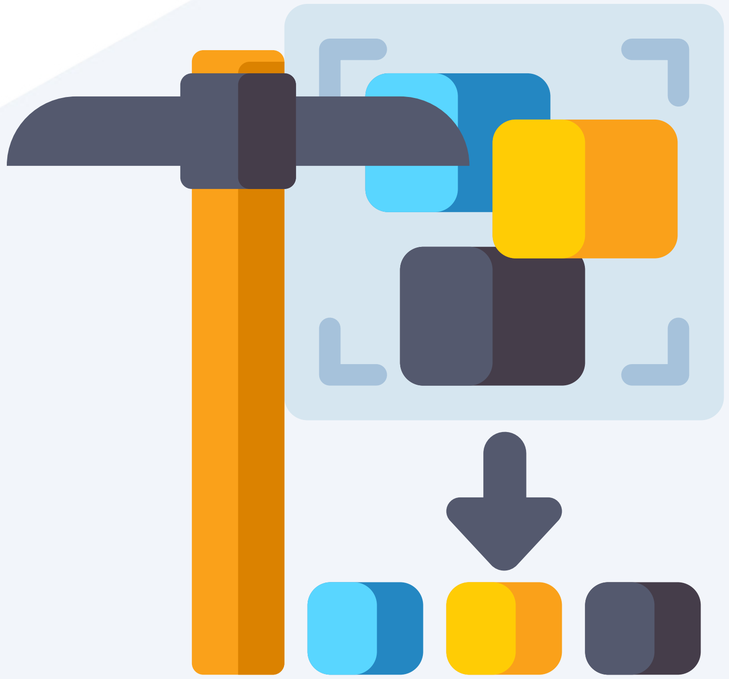
Você deve utilizá-lo quando precisar fazer cálculos em bases maiores que o limite do Excel (1.048.576 linhas), quando seus cálculos ficarem pesados e demorados demais ou quando as análises são complexas e podem se beneficiar da engine do DAX para realizá-las com mais eficiência.

DATAB

POWER QUERY

O Power Query é um suplemento de conexão e tratamento de dados disponível desde 2013 no Excel e integrado no Power BI. Ele utiliza a linguagem M como base, que é bem parecida com o F#. As consultas são tratadas por meio de etapas bem definidas que acumulam tratamentos.

Sua interface gráfica facilita o uso por analistas de negócios que não precisam necessariamente aprender a linguagem M para executar transformações. Além do suplemento, seu conceito também está presente de forma online no Data Flow, disponível no Power BI Serviço.



QLIK SENSE

É uma ferramenta de visualização de dados que cria relatórios e dashboards considerada com uma das líderes no mercado pela Gartner, porém, com menos relevância mercadológica que o Power BI por desvantagens competitivas em relação ao preço praticado por usuário e desenvolvimento de cálculos mais complexo para extrair informações. Ainda assim, é um software extremamente relevante no mercado e que tem buscado integrações com seu catálogo de dados com os principais concorrentes (Tableau e Power BI) permitindo exportar dados para finalização neles.

QUERY

Query ou consulta é um termo muito utilizado em banco de dados e business intelligence que basicamente executa pedidos específicos do que deve ser buscado e retornado. Existem diversas linguagens capazes de fazer consultas, entre elas a mais famosa é o SQL que utiliza sintaxes como:

SELECT campo FROM tabela WHERE condição

Elas são muito úteis para buscar ou agregar dados e criar views específicas que facilitarão o trabalho de análise em softwares de BI.

DATAB

RALPH KIMBALL

É um dos principais autores e referência sobre conceitos de data warehouse e dados transacionais desde os anos 80 defendendo que eles devem ser rápidos e eficientes.

Um de seus livros mais famosos é o **The Data Warehouse Tollkit**. Lecionou aulas sobre modelagem dimensional e ETL para mais de 20.000 alunos. É PhD em Engenharia Elétrica pela Stanford University.

SAAS

SaaS é a abreviação de **Software as a Service**, em português: Software como serviço. É um modelo de comercialização de produtos, normalmente baseados em cloud (nuvem), onde o cliente paga de acordo com o uso. Essa abordagem facilita o teste e viabiliza softwares que possuíam licenças de uso permanente caríssimas. Além disso, facilita o fluxo de caixa de empresas já que não precisam alocar um grande investimento para obter lucratividade futura. Como é pago por serviço, normalmente são faturas mensais de acordo com o processamento e uso.

Um exemplo de suite de produtos que usa a abordagem SaaS é o **Azure**.

SAP

O sistema ERP SAP é um dos mais utilizados no Brasil por grandes empresas para gestão da operação. Muitos projetos de BI são realizados com dados provenientes das estruturas do SAP e o Power BI possui conectores exclusivos para algumas de suas versões, como o SAP Hana e o SAP DW.

Ele possui diversos módulos e nem todas as empresas terão todos implementados. Por exemplo, SAP MM (Material Management), SAP HRM (Human Resource Management) e outros para gestão de projetos, qualidade, contabilidade, distribuição, planejamento de produção, cadeia de suprimentos, manutenção, vendas e logística.

DATAB

SELF-SERVICE BI

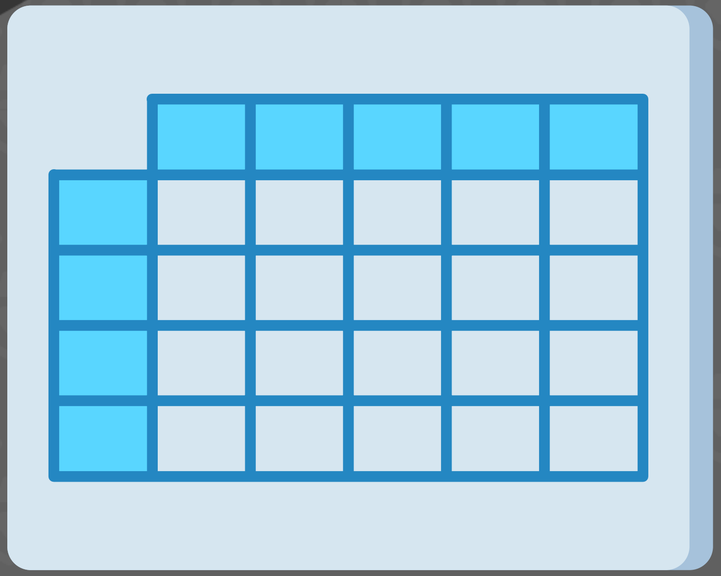
É um conceito de desenvolvimento de projetos de Business Intelligence focado na autonomia dos analistas. A independência da TI gera mais rapidez no desenvolvimento de análises, mas foi preciso que os softwares de BI se adaptassem com interfaces relativamente amigáveis para facilitar o processo de execução de projetos de BI do início ao fim, desde a obtenção, conexão e transformação de dados até o cálculo e visualização das informações obtidas.

Um dos softwares no mercado que utilizam esse conceito é o Power BI.

SGBD

Um banco de dados sempre deve ter um meio de gerenciá-lo. É ai que o SGBD entra. A sigla é um acrônimo para Sistema de Gerenciamento de Banco de Dados e seu objetivo principal é ser uma interface de acesso e administração dos dados armazenados em servidores.

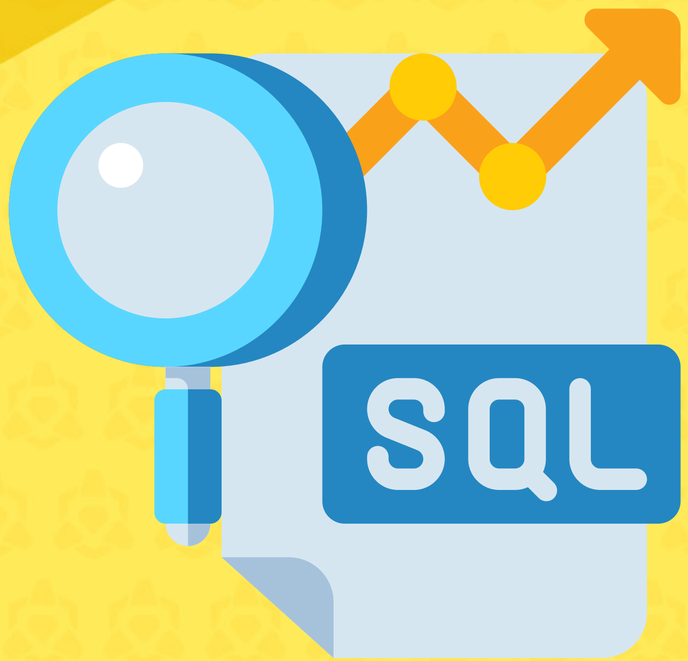
Os SGBD podem ser relacionais, hierárquicos, em rede ou orientado a objetos, com algumas variações entre esses termos. Exemplos de SGBD são: PostgreSQL, MySQL, MariaDB, SQL Server, Oracle, Firebird, MongoDB, entre outros.



SQL

SQL é uma das principais linguagens de consulta utilizada em bancos de dados relacionais. As letras de seu nome significam **Structured Query Language**. Para o analista de dados, é uma linguagem particularmente útil para facilitar o trabalho nos softwares de BI.

Os principais comandos do SQL são: **SELECT** (para selecionar dados), **INSERT** (para inserir novos dados), **UPDATE** (para atualizar registros), **DELETE** (para deletar registros), entre outros.

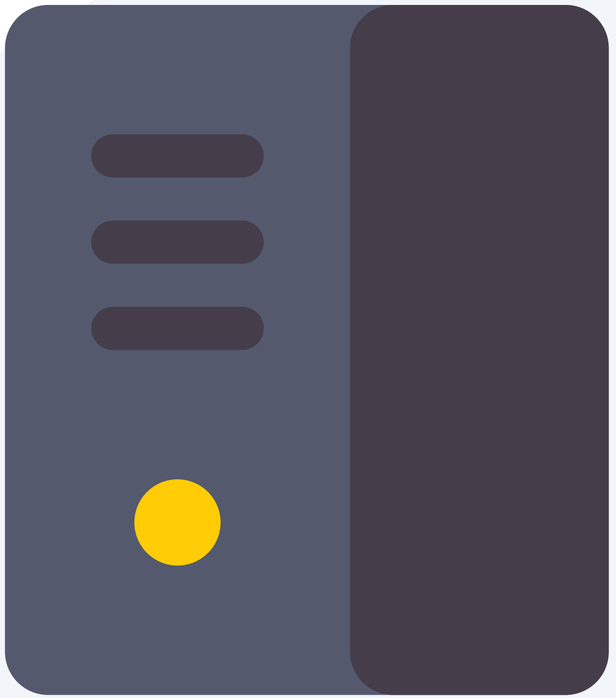


DATAB

SQL SERVER

O SQL Server é um SGBD para gerenciar dados armazenados no servidor que utiliza o modelo relacional. Existem licenças gratuitas para teste e desenvolvimento, mas no uso comercial é necessário adquirir uma ou contratar o serviço na nuvem pelo Azure.

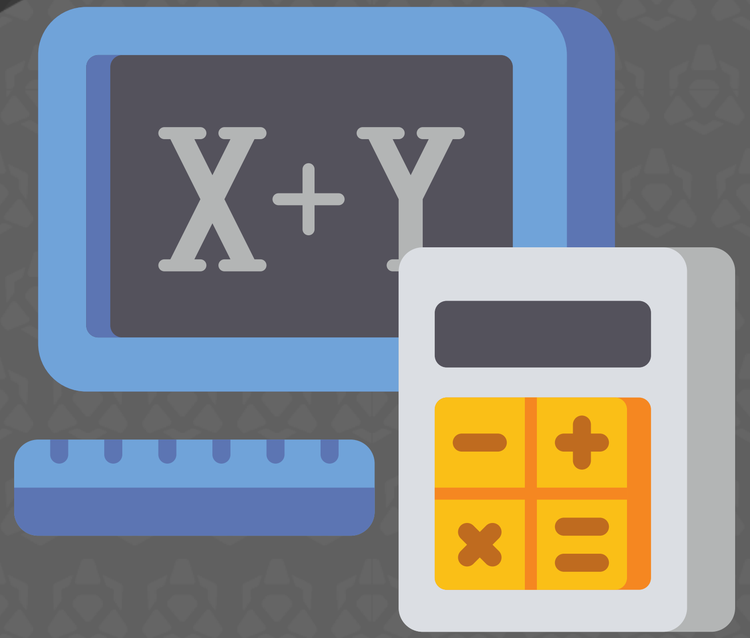
Ele possui diversos serviços, entre eles o Analysis Services, Integration Services e Reporting Services que auxiliam no momento de tratar, analisar e criar relatórios dos dados armazenados.



SSAS

SSAS é a abreviação de SQL Server Analysis Services, que é uma ferramenta de análise de dados em data warehouses ou data marts. Utiliza as linguagens MDX e DAX para criar cálculos com o conceito de organização de dados OLAP, sendo capaz de fazer consultas, hierarquias e medidas de forma performática.

Como é uma ferramenta de análise mais antiga que o Power BI, é muito utilizada pelas empresas, mas muitas migrações estão sendo efetuadas para o Power BI (medidas e colunas calculadas). O Analysis Services também pode ser uma fonte de dados para o Power BI utilizar medidas prontas provenientes do SSAS e em modelos híbridos.



SSIS

SSIS é a abreviação de **SQL Server Integration Services**, que é uma ferramenta robusta de ETL para solucionar problemas complexos de tratamentos de dados e mineração. Também temos no Power BI uma ferramenta para isso chamada de Power Query, entretanto, ela é direcionada para execuções generalistas que utilizam uma engine diferente para tratamento de dados em concorrência com outros serviços abertos do Power BI.

Para alterações complexas e grandes, o SSIS é um passo anterior a importação de dados no Power BI recomendado para aprimorar sua experiência com soluções empresariais de tratamento e análise de dados.

DATAB

STORYTELLING

O storytelling aplicado a análise de dados é uma forma de exibir informações com uma sequência como se estivesse contando uma história fluida. Essa técnica é muito útil para relatórios complexos ou para equipes que ainda precisam ser engajadas na cultura de análise de dados, pois direciona o usuário em partes focadas e sequenciais com uma linearidade evidente que é construída com detalhamentos, narrativas e transições bem estruturadas e interligadas.

É uma técnica relativamente complexa, pouco utilizada no meio empresarial, mas muito efetiva, que envolve soft skills e conhecimento sobre as necessidades analíticas reais e interessantes dos usuários do relatório.

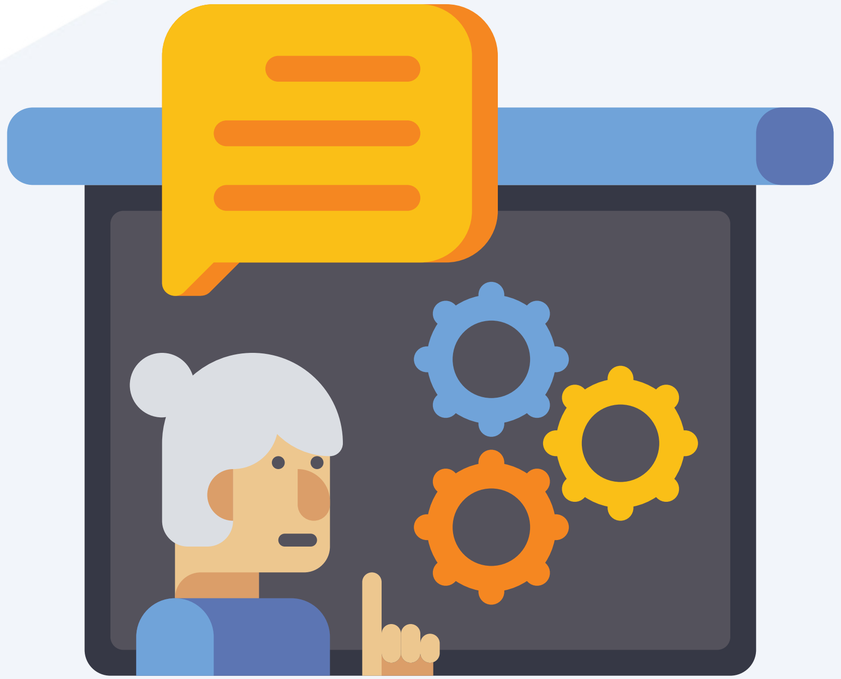


TABELA DIMENSÃO

Tabelas dimensão são tabelas em bancos de dados relacionais que armazenam detalhes sobre algo que é utilizado na operação. Por exemplo, um produto foi vendido. Ese produto tem atributos, mas quando foi vendido todos os seus atributos não aparecerão na tabela fato onde foi registrada a transação. Nela aparecerá apenas o código do produto. Seus atributos estarão na tabela dimensão de produtos, que é como se fosse a tabela de cadastros. Lá terão dados como cor, local do estoque, dimensão, modelo, ano de fabricação, categoria, subcategoria, preço, custo e outras informações.

Tabelas dimensão devem possuir uma chave primária que identifica de forma exclusiva os registros de cada item armazenados nela.

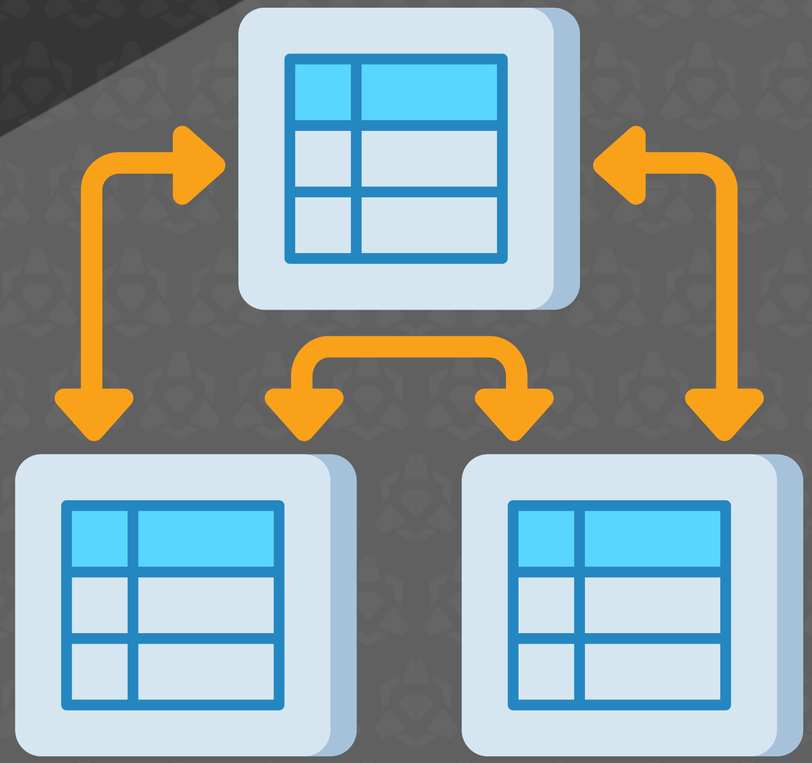
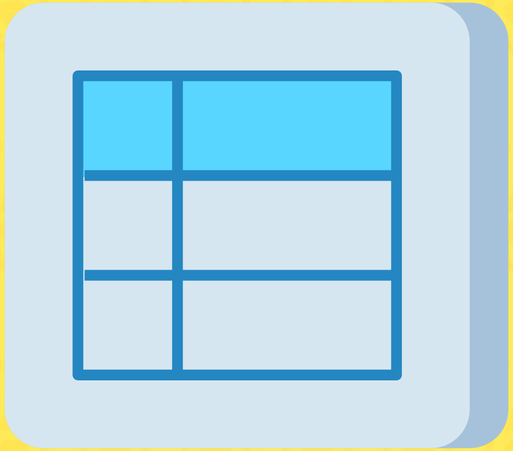


TABELA FATO

Tabelas fato são, normalmente, maiores que tabelas dimensão, pois armazenam dados da operação da empresa. Um exemplo de tabela fato são os pedidos realizados por clientes em uma empresa. Eles são armazenados na tabela pedidos e busca informações de diversas tabelas dimensão, como o que foi vendido, quando foi vendido, por quem, onde e para quem.

Como a tabela fato tende a crescer de acordo com a operação da empresa, é primordial não ter redundâncias nela, por isso é importante ter uma boa modelagem e tabelas dimensão capazes de descrever os códigos mencionados na tabela fato.

É dela que normalmente são feitas agregações para analisar a performance de objetivos das empresas.



DATAB

TABLEAU

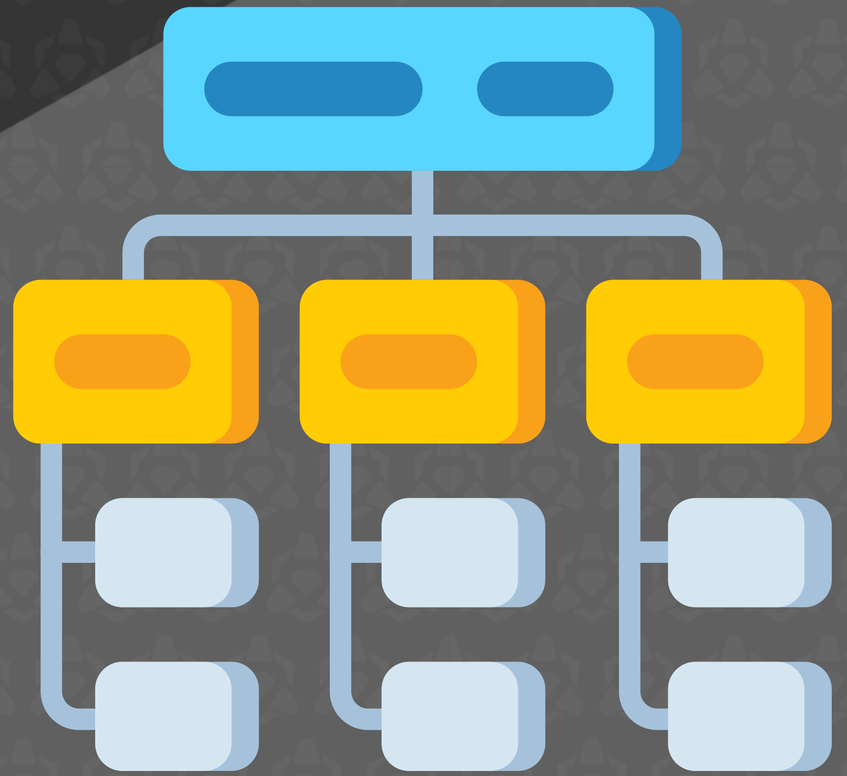
Tableau é um software de BI considerado como um dos líderes do mercado pela Gartner. Além de análises e visualizações, também possui uma suite para armazenamento, compartilhamento e tratamento de dado (Tableau Prep).

Seu forte são os tipos de visualizações e a qualidade delas, mas em alguns aspectos consideramo que possui desvantagens em relação ao Power BI, como a falta de uma comunidade forte que produz conteúdo com informações acessíveis para analistas de negócio, levando a uma maior e relativa dependência por consultores para criação e alteração de projetos e seu preço por licença.

TAXONOMIA

Na computação, a taxonomia é uma organização de informações por meio de um sistema lógico com princípios de classificação e categorização. A maioria das informações são armazenadas por palavras chave ou alfabeticamente, o que torna prática a procura de dados, mas não lógica, já que todos os dados estão misturados sem categorização.

A taxonomia começa a ser usada para dividir informações, dados ou documentos em classes e sub-classes garantindo o retorno de informações eficientemente e todas as suas correlatas. Ela é considerada um dos pilares da gestão da informação e do conhecimento.



TRIGGER

Trigger é uma ação motivada por um evento. Traduzindo do inglês, significa gatilho. Apesar de ser uma palavra genérica, é muito utilizada na computação para automatizar processos. Imagine que você precisa que toda vez que um novo registro é adicionado em um banco de dados, uma atualização seja disparada em um relatório que o analisa.

O trigger é a modificação que aconteceu e a ação é a atualização. Muitos softwares são utilizados para executar processos semelhantes e esse conceito também é disponibilizado internamente em alguns deles. Por exemplo, no Power BI você pode configurar o envio de e-mails sempre que uma base de dados for atualizada. Também é possível usar ferramentas externas como o Power Automate para gravar fluxos motivados por triggers.

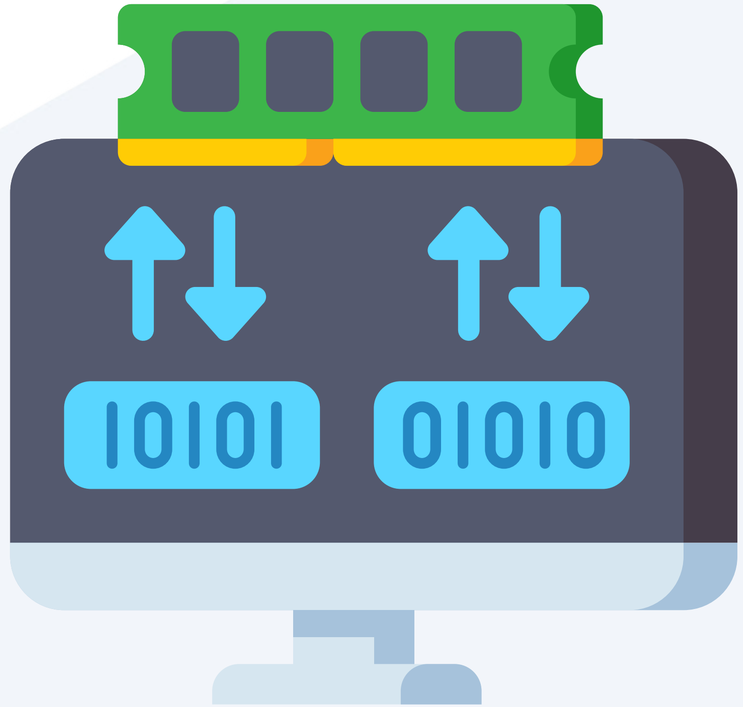
DATAB

VARIÁVEL

Na computação, uma variável é um objeto capaz de armazenar um valor ou expressão. Elas são identificadas por nomes não reservados e podem ser globais ou locais. Normalmente são registradas na memória do computador.

Um exemplo de variável no DAX pode ser esse:
VAR x = 30

Quando chamamos o **x**, o valor 30 será retornado:
200 + x nesse caso é o mesmo que 200 + 30



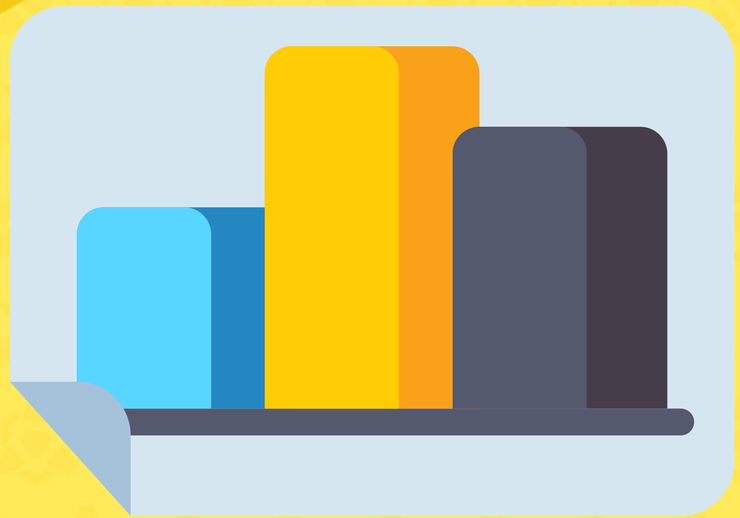
VIEW

View ou visão é o resultado de uma consulta (query) em uma base de dados. Elas são especialmente úteis para limitar ou criar cenários diferentes para melhorar a performance ou simplificar a análise de dados em softwares de BI. Elas podem ser executadas em SGBD ou no próprio BI, como disponibilizado por alguns conectores no Power BI.

As views podem ser simples, com filtros limitando os dados exibidos ou mais complexas, com joins e agregações. Elas podem ser executadas com linguagens de programação de consulta, como o SQL.

VISUALIZAÇÃO

O conceito de gráficos é muito enraizado quando pensamos em mostrar dados visualmente, mas existe uma infinidade de visuais que compõem a visualização de dados. Além de gráficos, temos tabelas, cartões, mapas, narrativas, diagramas e matrizes que podem ser geradas em softwares de BI ou por bibliotecas em linguagens de programação capazes de fazer isso, como aquelas importadas no Python e R.



É uma das etapas finais de projetos de BI e extremamente importante, pois uma escolha errada na visualização pode inutilizar por complexo a leitura dos dados e geração de informações para os usuários.

DATAB

Visite databinteligencia.com.br