# Glossário de Dados

## Conceitos e Estratégias

### Business Intelligence (BI)

Não é apenas um software. Business Intelligence (ou BI) é um conjunto de estratégias, processos e tecnologias usado para analisar dados de negócios. Em vez de depender de intuição ou "achismo", o BI fornece uma visão clara do desempenho da empresa, permitindo identificar tendências, entender o comportamento do cliente e otimizar operações. A área usa os dados do passado e do presente para responder a perguntas como: "Quanto vendemos no último mês?" ou "Qual produto é o mais popular?". O objetivo é tomar decisões mais inteligentes.

### Governança de Dados

É como se fosse o "manual de regras" para os dados de uma empresa. Define quem pode acessar quais dados, como eles devem ser armazenados, como garantir sua qualidade e segurança. O objetivo é ter dados organizados, confiáveis e seguros para todos. Na prática, isso garante que a operação funcione sem vazamentos ou ações má intencionadas com dados da empresa, de clientes e funcionários.

---

## Locais de Armazenamento de Dados

### Banco de Dados

Imagine uma biblioteca muito organizada, onde cada livro (dado) tem um lugar certo. Um banco de dados é um sistema para guardar e organizar informações de forma estruturada, para que você possa encontrá-las e usá-las facilmente. É usado em sistemas do dia a dia, como em um site de compras para guardar seus pedidos. Em termos de trabalho do profissional de dados, ele permite fazer mais atividades do que uma planilha, com mais segurança, seguindo melhor as regras de negócio e a legislação vigente (que considera elementos de governança e privacidade, por exemplo).

### Data Warehouse (DW)

É um grande "armazém" central de dados. Diferente de um banco de dados comum, ele é projetado para guardar um volume enorme de informações históricas de várias fontes da empresa (vendas, marketing, finanças, etc.). Seu foco é ajudar na análise e na geração de relatórios (BI), permitindo que os profissionais trabalhem com os dados ali, principalmente nas etapas de ETL ou ELT (que envolvem extração, transformação e carregamento).

### Datamart

Em uma analogia, se o Data Warehouse é um grande supermercado, um Datamart é uma "loja de conveniência" dentro dele. É uma parte menor do Data Warehouse, focada em um setor específico da empresa, como "Vendas" ou "Marketing". Isso torna a análise para essas equipes mais rápida e fácil.

### Data Lake

Pense em um "lago" onde você pode jogar todos os tipos de dados, sem se preocupar em organizá-los primeiro. Pode ser uma foto, um vídeo, um texto, uma tabela... qualquer coisa! É um lugar para guardar dados brutos, que talvez sejam usados no futuro para análises

mais complexas, como as de Ciência de Dados. Isso permite que a empresa unifique seu repositório de dados, evitando divergências entre “silos”, de modo que todos os setores “bebam” da mesma fonte, possuindo uma verdade única.

### Lakehouse

É o melhor dos dois mundos: a organização de um Data Warehouse com a flexibilidade de um Data Lake. É uma arquitetura moderna que permite que você guarde todos os tipos de dados (como em um lago) e, ao mesmo tempo, consiga organizá-los e analisá-los com a mesma facilidade de um armazém.

---

## Funções e Profissionais

### Engenharia de Dados

É a área responsável por construir e manter os "encanamentos" (pipeline de dados) que levam os dados de um lugar para outro. O engenheiro de dados cria sistemas para coletar, armazenar e preparar os dados para que os analistas e cientistas possam usá-los. É a área que vai agir diretamente no “saneamento” que vai permitir a qualidade dos dados.

### Análise de Dados

É o trabalho de "investigar" os dados para encontrar informações úteis. O analista de dados olha para o que já aconteceu, cria gráficos e relatórios para responder a perguntas de negócio e ajuda a empresa a entender seu desempenho. Tem se popularizado bastante, mas a execução de um bom trabalho depende de outros setores/funções, como arquitetura e engenharia de dados. Também precisa conhecer a área de negócio para apresentar boas análises.

### Ciência de Dados

É uma área mais avançada que usa estatística, matemática e programação para fazer previsões (forecasting), por meio da identificação de padrões e correlações matemáticas. Um cientista de dados pode criar um modelo para prever quais clientes têm mais chance de cancelar um serviço ou qual será a demanda de um produto, por exemplo. Ou ainda: qual seria a composição ideal para a cerveja atender aos padrões de qualidade da indústria, ao mesmo tempo que alcançaria o seu menor custo e melhor preço de venda.

### Analytics Engineer

É um profissional que atua entre a Engenharia e a Análise de Dados. Ele organiza e "limpa" os dados brutos deixados pelo engenheiro, transformando-os em informações prontas e confiáveis para que os analistas possam consumir facilmente. Muitas vezes, ele une em seu repertório as funções de analista e engenheiro, trabalhando com pipeline de dados e análises para a organização.

### DBA (Administrador de Banco de Dados)

É o profissional responsável por cuidar dos bancos de dados. O DBA garante que o banco de dados esteja funcionando bem, seja seguro, rápido e bem administrado de modo geral. Essa pessoa trabalha para seu bom funcionamento, zelando para que os dados não sejam perdidos ou danificados.

---

# Processos e Tecnologias

## ETL (Extract, Transform, Load)

É um processo para mover dados. Primeiro, ele extrai (load) os dados de uma fonte (como um site ou sistema). Depois, Transforma esses dados (limpando, organizando ou fazendo cálculos - parte do “transform”). Por fim, carrega os dados transformados para seu destino (como um Data Warehouse).

## ELT (Extract, Load, Transform)

É uma variação do ETL. A diferença é que ele primeiro extrai os dados, depois os carrega diretamente em um local como um Data Lake ou Data Warehouse e só então os transforma. Isso é muito útil quando se tem um volume gigante de dados e não se sabe quais serão relevantes ou não - então essa seleção é feita em outra etapa, mais próxima à análise.

## SQL (Structured Query Language)

É a linguagem que você usa para conversar com bancos de dados relacionais. Com SQL, você pode pedir para o banco de dados buscar, inserir, atualizar ou apagar informações. É como saber o idioma da biblioteca para pedir o livro certo. Também é possível deixar reações pré-definidas conforme algum acontecimento gatilho (“trigger”), entre outros.

## Python

É uma linguagem de programação muito popular e versátil em várias áreas da tecnologia. No mundo dos dados, é usada para quase tudo: desde a automação de processos (Engenharia de Dados) até a criação de modelos complexos (Ciência de Dados), passando também por desenvolvimento.

## Spark

Pense no Spark como um "motor" superpotente para processar volumes gigantescos de dados de forma muito rápida. Ele distribui o trabalho entre várias máquinas.

## Pyspark

É a forma de usar o poder do Spark escrevendo código em Python, o que o torna muito popular. Na prática, é muito utilizado para trabalhar com grandes volumes de dados (Big Data), sobretudo em ferramentas robustas que usam processamento clusterizado em nuvem, como Databricks.

## Batch (Processamento em Lote)

É quando você processa um grande volume de dados de uma só vez, em "lotes" ou pacotes. Por exemplo, rodar um processo toda noite para atualizar os dados de vendas do dia anterior - ou seja, acontece em momentos programados. É diferente do processamento em tempo real (streaming).

## Streaming

Streaming de dados (ou processamento de fluxo) é a prática de processar e analisar dados de forma contínua, evento por evento, assim que eles são gerados. Em vez de guardar os dados para processá-los em grandes "lotes" (Batch), o streaming lida com os dados "em movimento". Em resumo, quando utilizamos streaming, não há um tempo de espera para atualização de dashboards.

### Dashboard

É a tela em que ficarão dispostos os gráficos com os indicadores e análises feitas pelos profissionais de dados. Essa análise pode ser estática ou dinâmica, refletindo dados históricos (referentes a um período específico e, portanto, não terão atualizações) ou dados do presente. No caso de dados atuais, podemos ter atualização do dashboard em tempo real ou em horários específicos.

---

## Conceitos de Modelagem de Dados

### Modelo Relacional e Dados Relacionais

É a forma mais comum de organizar dados: em tabelas com linhas e colunas, como uma planilha do Excel. A parte "relacional" significa que as tabelas podem se conectar umas às outras através de chaves. Isso permite gerar análises que façam sentido e que conectem informações de vários tipos.

### Chave Primária (Primary Key)

É o "CPF" de cada linha em uma tabela. É um identificador único que garante que não existam duas linhas exatamente iguais. Por exemplo, em uma tabela de clientes, o ID_Cliente é a chave primária. É de extrema importância, pois sem um identificador único, o trabalho com dados se torna praticamente inviável. Para um exemplo da vida real, tente imaginar como seria viver sem um número de CPF, sendo identificado apenas pelo nome. Em nível empresarial, produtos, vendas e clientes poderiam ser confundidos e gerar prejuízos à organização.

### Chave Estrangeira (Foreign Key)

É o que conecta uma tabela a outra. Pense em uma tabela de "Pedidos" e uma de "Clientes". A tabela de "Pedidos" terá uma coluna ID_Cliente para indicar qual cliente fez aquele pedido. Esse ID_Cliente na tabela de pedidos é uma chave estrangeira, pois aponta para a chave primária da tabela de clientes. Dessa forma, não é necessário que todas as informações estejam em uma grande tabela apenas, mas possam ser organizadas de diferentes formas, mantendo a propriedade de combinar os dados para as análises.

### Surrogate Key (Chave Substituta)

É uma chave primária "artificial", geralmente um número sequencial (1, 2, 3...), que não tem nenhum significado para o negócio. Ela é usada apenas para identificar uma linha de forma única e é muito comum em Data Warehouses.

### Modelagem de Dados

É o ato de desenhar como os dados serão organizados. É como fazer a planta de uma casa antes de construí-la. Você decide quais tabelas criar, quais colunas elas terão e como elas vão se conectar. Dessa forma, existe uma estruturação lógica antes da execução.

### Star Schema (Esquema Estrela)

É um tipo de modelagem muito usado em Data Warehouses. Imagine uma estrela: no centro, há uma Tabela Fato (com os números do negócio, como vendas) e, ao redor dela, ligam-se várias Tabelas de Dimensão (que dão o contexto, como Cliente, Produto, Tempo). É simples e rápido para fazer análises.

### Snowflake Schema (Esquema Floco de Neve)

É uma variação do Star Schema. A diferença é que as tabelas de dimensão podem ser "quebradas" em outras tabelas menores. Por exemplo, a dimensão "Produto" pode se conectar a uma tabela de "Categoria" e a uma de "Marca". O diagrama parece um floco de neve.

### Tabela Fato

É a tabela principal no Star Schema. Ela guarda os números, os eventos que aconteceram. Geralmente, tem muitas chaves estrangeiras (para se conectar às dimensões) e as medidas (os valores numéricos, como Valor_da_Venda, Quantidade_Vendida).

### Dimensões

São as tabelas que descrevem o "quem, o quê, onde, quando, por quê" da Tabela Fato. Elas dão o contexto. Exemplos: Dimensao_Cliente, Dimensao_Produto, Dimensao_Tempo.

### Atributos

São as colunas de uma tabela de dimensão. Por exemplo, na Dimensao_Cliente, os atributos seriam Nome_Cliente, Cidade, Estado.

### Medidas

São as colunas numéricas da Tabela Fato, aquilo que você quer medir ou calcular. Exemplos: Faturamento, Quantidade_de_Itens, Desconto_Aplicado.

---

## Termos Técnicos de Banco de Dados

### OLTP (Online Transaction Processing)

São sistemas focados em registrar transações do dia a dia, de forma rápida. Pense no sistema de um caixa de supermercado. Ele precisa inserir uma venda nova muito rápido. A prioridade é a agilidade para registrar dados.

### OLAP (Online Analytical Processing)

São sistemas focados em análise de dados, como um Data Warehouse. Eles são projetados para responder a perguntas complexas que envolvem grandes volumes de dados. A prioridade é a velocidade para fazer consultas e agregar informações.

### Acrônimo ACID

É um conjunto de quatro garantias que os bancos de dados do tipo OLTP (transacionais) oferecem em cada transação:

- Atomicidade: Ou a transação acontece por completo, ou não acontece nada. (Ex: numa transferência, o dinheiro não sai de uma conta sem entrar na outra).
- Consistência: A transação sempre levará o banco de dados de um estado válido para outro.
- Isolamento: Várias transações acontecendo ao mesmo tempo não atrapalham umas às outras.
- Durabilidade: Uma vez que a transação é confirmada, ela fica salva para sempre, mesmo que o sistema falhe.

### Stored Procedures (“Procedimentos Armazenados”)

São como "mini programas" de SQL que ficam guardados (“stored”) dentro do próprio banco de dados. Você pode executá-los para realizar uma tarefa complexa de uma só vez, em vez de escrever vários comandos SQL toda vez.

### Functions (Funções)

São parecidas com Stored Procedures, mas a principal diferença é que elas sempre retornam um valor. Você pode usá-las no meio de um comando SQL, como se fossem uma fórmula.

### Triggers (Gatilhos)

São ações automáticas que o banco de dados executa quando algo acontece em uma tabela. Por exemplo, você pode criar um trigger para salvar um registro de auditoria toda vez que uma linha for apagada.

### View

É como uma "janela" ou um "atalho" para uma consulta SQL. É uma tabela virtual baseada no resultado de uma consulta. É muito útil para simplificar o acesso aos dados ou para limitar o que um usuário pode ver.

### Palavra Reservada

São palavras que têm um significado especial em uma linguagem, como SQL (SELECT, FROM, WHERE) ou Python (if, for, while). Você não pode usá-las como nomes de tabelas ou colunas.

### Slowly Changing Dimensions (SCD)

É uma técnica usada em Data Warehouses para lidar com mudanças nos dados das dimensões ao longo do tempo. Por exemplo, o que fazer quando um cliente muda de endereço? A técnica SCD define como registrar essa mudança sem perder o histórico.