Ciência de Dados

William E. Deming = "Em Deus nós acreditamos. Todos os outros devem trazer dados."

Ciência de Dados = insights de dados para solucionar problemas e mostrá-los de forma simples pros tomadores de decisões. A ciência é um processo a ser seguido, método científico, garante a confiabilidade dos dados.

Projeto de Dados = definição do problema; coleta e análise de dados; decisão do gestor; ação de melhora.

Definição do Problema = responder o que aconteceu; porque aconteceu; se acontecerá novamente; e o que deve ser feito.

Coleta e Análise de Dados = engenharia de dados (aquisição, armazenamento, limpeza e transformação dos dados); analytics (análise exploratória dos dados, modelagem preditiva, inferências e previsões); produção (construção do produto de dados).

Produtos Gerados = relatórios, gráficos, dashboards, modelos estatísticos e preditivos, web apps, sistemas de recomendação.

Dashboard = visualização dos dados, em um tipo de infográfico digital.

Dados = fatos individuais, brutos, coletados por observação ou medição. Tudo que se pode observar, medir, coletar.

Informações = dados em contexto. Tomadores de decisões não querem dados, querem informações e saber como elas vão ajudá-los.

Atributos = as características de um registro, as colunas de uma tabela.

Registros = um conjunto de atributos que apontam características, comportamentos ou resultados, as linhas de uma tabela.

Datasets = conjunto de dados, coleção de observações, como tabelas. Matéria prima da ciência de dados.

Estatística = fornece métodos para analisar os dados através da probabilidade, estatística descritiva e estatística inferencial.

Probabilidade = estudo de aleatoriedade e incerteza; previsões para tomar decisões futuras com um pouco mais de informações.

Estatística Descritiva = descreve os dados através de média, moda, mediana, dispersão, desvio padrão...

Estatística Inferencial = estima informações sobre uma população a partir dos resultados de uma amostra.

Machine Learning = aprendizado de máquinas. O computador encontra padrões nos dados através de algoritmos e parâmetros. O aprendizado pode ser supervisionada, não supervisionada, semi supervisionada, por reforço ou deep learning.

Aprendizagem Supervisionada = algoritmos fazem previsões com base em exemplos, em dados históricos. Pode ser classificação (prevê classe, categoria, sim ou não) ou regressão (prevê um valor numérico). Se os dados forem palavras, converter em números.

Aprendizagem Não Supervisionada = os dados históricos não tem os resultados, só as características. Essas características são empregadas no algoritmo, que reúne os dados por similaridade e divide em grupos parecidos.

Aprendizagem Semi Supervisionada = há dados históricos e atuais, alguns com resultados, outros sem. O computador analisa os com resultado e completa os sem com base neles.

Aprendizagem Por Reforço = o computador toma decisões com base em limites, visando aumento da recompensa. Ações corretas geram recompensa, ações incorretas geram punição.

Deep Learning = feito com aprendizado supervisionado ou não supervisionado. É bem mais preciso, usado em visão computacional e processamento de linguagem, o computador recebe a imagem ou voz e vai decodificando até encaixar em um grupo conhecido.

Modelo Preditivo = um algoritmo treinado com boa performance. Uma fórmula matemática que funciona, na inteligência artificial.

Deploy = colocar o modelo preditivo pra resolver o problema, a entrega do resultado final.

Big Data = grande conjuntos de dados que não podem ser processados por bancos de dados tradicionais. É definido por 4Vs: volume (quantidade: megabyte, petabyte), variedade (tipos: estruturados, semi estruturados e não estruturados), velocidade (geração rápida) e veracidade (dados reais, não fictícios).

Big Data Analytics = o Big Data quando se aplicam análises nos dados, quando se põe valor.

ETL = Extração, Transformação e Carregamento dos dados. É o caminho dos dados, da fonte até a entrega.

ELT = Extração, Carregamento e Transformação dos dados. Os dados são extraídos e guardados em um data lake, mas só serão transformados caso necessário.

Fontes e Formatos de Dados = interno e estruturado (resultados de pesquisa, registros de vendas, bancos de dados internos); interno e não estruturado (emails, comentários, avaliações de funcionários); externo e estruturado (likes de facebook, pontuação em sites de classificação); externo e não estruturado (comentários em fóruns, vídeos de câmeras de segurança).

Banco de Dados Relacionais = estruturado e com schema, como o SQL.

Banco de Dados NoSQL = não estruturados, livres, como o MongoDB.

Schema = organização dos dados, estrutura.

Data Warehouse = guarda dados estruturados, com schema, em grandes quantidades e de muitas fontes, garantindo decisões com dados completos e limpos.

Data Lake = guarda dados não estruturados, brutos, sem limpeza.

Data Store = guarda dados não estruturados de modo personalizado.

Data Hub = mistura de Data Warehouse, Data Lake e Data Store.

Cluster de Computadores = conjunto de vários servidores guardando dados relacionados.

Processamento Paralelo = divide o armazenamento em vários locais e uma tarefa em várias subtarefas, executadas paralelamente.

Cloud Computing = serviço de computação e armazenamento online que cobra apenas as horas de uso. Mais usadas são AWS e Microsoft Azure.

Daas = Data as a Service. Transforma os dados em um serviço pra outras empresas, fornecendo já organizados e dispostos.

DataOps = linha de produção para trabalhar com dados.

Cultura Data-Driven = cultura orientada a dados, sendo dirigida por dados.

Administrador de Banco de Dados (DBA) = suporte técnico pros bancos de dados; planeja e projeta personalizações de bancos de dados para necessidades específicas; otimiza consulta ao SQL; supervisiona instalação de novos SGBDs; cria procedimentos de backups, restauração e recuperação de desastres.

Analista de Dados = coleta dados das fontes primárias ou secundárias; efetua modelagem de dados para limpeza, preparação, transformação e descarte; trabalha com os dados brutos; deixa os dados prontos para o cientista; analisa e interpreta resultados; identifica tendências, correlações e padrões em conjuntos de dados através de análise exploratória (70% do trabalho); identifica novas oportunidades para melhorias; concebe e mantém bancos de dados; resolve problemas de códigos e questões relacionadas; domina linguagens de programação (como R, Python ou SQL); constrói visualizações.

Arquiteto de Dados = pensa na arquitetura dos dados; olha para os dados de uma ponta a outra do processo; ajuda a elaborar uma estratégia de dados para a empresa; cria o inventário de dados para aplicar a arquitetura; pesquisa oportunidades de aquisição de dados; projeta jobs ETL e pipelines de dados (o engenheiro constrói, o arquiteto diz como); cria fluxo de dados; desenvolve modelos de dados; projeta, documenta, constrói e implanta arquiteturas e aplicações de banco de dados; projeta funcionalidades (escalabilidade, segurança, desempenho, recuperação, confiabilidade...); implementa medidas para precisão e acessibilidade dos dados; cria procedimentos de gestão de metadados e compliance com GDPR e LGPD.

Cientista de Dados = emprega programas de análise, machine learning e métodos estatísticos; prepara os dados para modelagem preditiva; explora e analisa os dados por vários ângulos para encontrar fraquezas, tendências e oportunidades; concebe soluções orientadas a dados; cria novos algoritmos para resolver problemas e automatizar o trabalho; extrai grande volume de dados; comunica previsões e resultados para a gestão.

Engenheiro de Dados = trata dados; foca em infraestrutura; cuida da organização, armazenamento, segurança e utilização dos dados; projeta, constrói, instala, testa e mantém sistemas de gerenciamento de dados; pesquisa aquisição de mais dados e novos usos para os existentes; desenvolve modelagem e mineração de dados; integra novas tecnologias de uso de dados com data warehouses, data lakes e datastores; implementa arquiteturas de microsserviços; cria pipeline de dados e processos de extração, transformação e carregamento; implementa procedimentos de recuperação de desastres e alta disponibilidade; implementa políticas de segurança no acesso aos dados e compliance com GDPR e LGPD.