**Aula 1 - o que é engenharia de dados**

A engenharia de dados é a área que lida com a organização e transformação dos dados brutos de uma empresa. O Data Engineer coleta, armazena e disponibiliza as informações para análise de forma segura. Isso ajuda a empresa a tomar decisões estratégicas com base nos dados.

Para a engenharia fazer isso ele precisa entender **big data** que é a enorme quantidade de informações que é produzida, coletada e armazenada pelas tecnologias de monitoramento em tempo real. Ele fornece insumos para estudos e aplicações, permitindo que profissionais, como os engenheiros de dados, executem suas tarefas de forma eficiente.

**Aula 3 - Os seis princípios da arquitetura de dados moderna**

A arquitetura de dados é a estrutura que define como os dados de uma organização são geridos. Alguns princípios fundamentais dessa arquitetura incluem:

* Dados como um ativo compartilhado: Tratar dados como um recurso compartilhado gera vantagens competitivas, permitindo uma visão completa da empresa.

* Interfaces adequadas para consumo: Disponibilizar interfaces apropriadas para que as pessoas e sistemas possam acessar e utilizar os dados de forma eficiente

.

* Garantia de controle de acesso e segurança: Assegurar a governança e políticas de acesso aos dados para proteger a privacidade e a segurança das informações.

* Estabelecer um vocabulário comum: Utilizar um vocabulário compartilhado para evitar conflitos e garantir um entendimento unificado dos dados.

* Gerenciamento adequado dos dados: Investir em funções centrais de curadoria de dados para maximizar o valor dos ativos de dados compartilhados.

* Eliminar cópias e movimentação desnecessária de dados: Reduzir custos e complexidade ao eliminar cópias desnecessárias de dados.

Seguir esses princípios é fundamental para criar uma arquitetura moderna de dados corporativos eficiente e com baixo custo de mudanças à medida que a empresa cresce.

**Aula 4 - Curso de Big Data - Aula 1 - O que é Big Data?**

* **Banco de dados relacional**

Um banco de dados relacional é uma forma organizada de armazenar informações em tabelas, onde cada tabela representa um conjunto de dados relacionados. Os dados são estruturados em linhas (registros) e colunas (atributos), permitindo fácil acesso, manipulação e consulta das informações.

* **Banco de dados não relacional**

Um banco de dados não relacional é uma forma mais flexível de armazenar informações **sem precisar usar tabelas (SCHEMA) Nâo é aplicado o conceito de SCHEMA,** sendo mais adequado para grandes quantidades de dados variados. Ele permite organizar os dados de maneira mais livre e adaptável às necessidades específicas do projeto.

**Dado estruturado:** Dados organizados em uma estrutura pré-definida, como tabelas em um banco de dados relacional. (Exemplo: tabela com colunas nome, idade e cidade)

**Dado não estruturado:** Dados sem formato específico ou organização pré-definida, como texto, imagens ou áudio. (Exemplo: arquivos de texto livre, fotos, gravações de voz)

**Cientista de dados:** Profissional que utiliza análise de dados e técnicas estatísticas para obter insights e tomar decisões em projetos e negócios.

**Engenheiro de dados:** Profissional que projeta, desenvolve e gerencia a infraestrutura de dados, coleta, transformação e armazenamento para facilitar o acesso e análise de informações de forma consistente e segura.

**Big Data:** Refere-se a uma enorme quantidade de dados que ultrapassa a capacidade de processamento dos sistemas tradicionais. Envolve coleta, armazenamento e análise para extrair informações úteis, identificar padrões e tendências relevantes para tomada de decisões.

**Aula 6 -** [**Confluence - Introdução ao Data Mesh**](https://degreed.com/articles/confluence-introdu-o-ao-data-mesh?d=31372688&view=true&contentSource=Pathway&inputType=Article&contentSourceId=2733858&newWindow=true)**,** [**Data Mesh - Patterns**](https://degreed.com/articles/data-mesh-patterns?d=30952868&view=true&contentSource=Pathway&inputType=Article&contentSourceId=2733858&newWindow=true)**,** [**Introdução e Principais Conceitos - Data Mesh**](https://degreed.com/articles/confluence-introdu-o-e-principais-conceitos-data-mesh?d=30962698&view=true&contentSource=Pathway&inputType=Article&contentSourceId=2733858&newWindow=true)

**Data Mesh:** Data Mesh é uma abordagem descentralizada de arquitetura de dados, onde cada equipe de negócios é proprietária de seus próprios dados e responsável por sua governança. As equipes podem escolher

suas tecnologias e práticas, visando promover a autonomia e agilidade na entrega de projetos de dados.

**Data Lake**: É um repositório centralizado de dados brutos e não estruturados (SEM TABELAS), onde diversas fontes de dados são armazenadas sem a necessidade de pré-definir uma estrutura, permitindo análises flexíveis e exploratórias.

**Data Warehouse**: É uma solução que centraliza os dados em um único repositório, que coleta, integra e armazena dados de diversas fontes em um formato estruturado(TABELAS), otimizado para consultas e análises de negócios. É usado para suportar relatórios e análises mais tradicionais e estruturadas.

**Vantagens do Data Mesh:**

1) Maior autonomia e agilidade: Equipes de domínio têm controle sobre seus próprios dados, permitindo tomar decisões de forma mais rápida e eficiente.

2) Escalabilidade e flexibilidade: O Data Mesh suporta a expansão de dados e aplicações sem depender de uma infraestrutura centralizada.

3) Inovação e colaboração: Promove maior colaboração entre as equipes, incentivando a inovação e a busca por soluções eficazes.

**Vantagens do Data Lake:**

1) Armazenamento amplo: Pode lidar com grandes volumes de dados estruturados e não estruturados em seu estado bruto.

2) Exploração de dados: Permite análises flexíveis e exploratórias, pois não requer predefinição de esquemas.

3) Adaptação ao Big Data: Ideal para ambientes com grande quantidade de dados, onde a variedade e o volume são desafios.

**Vantagens do Data Warehouse:**

1) Dados estruturados e organizados: Oferece dados organizados em esquemas bem definidos, adequados para relatórios e análises regulares.

2) Performance: Projetado para consultas eficientes, permitindo respostas rápidas em análises complexas.

3) Consistência e confiabilidade: Fornece uma visão unificada dos dados, garantindo consistência e confiabilidade nas informações disponibilizadas.

**Aula 8 – dataviz**

**Objetivo**: O objetivo da visualização de dados é transmitir informações complexas de maneira clara e compreensível.

**Público**: Considere o público-alvo para adaptar o design e o conteúdo, garantindo que a visualização seja relevante e útil para quem a verá.

**Gráficos:** Escolha o tipo de gráfico adequado para representar os dados de forma eficaz e fácil de entender.

**Simplificação**: Evite poluição visual, destacando os principais insights e mantendo a visualização simples e direta.

**Cores e Layout** Use cores e layout consistentes e intuitivos para melhorar a experiência do usuário, tornando a visualização agradável e acessível.

**UI (Interface do Usuário):** Refere-se à aparência e interação dos elementos visuais de um aplicativo ou site, tornando-o esteticamente agradável e fácil de usar.

**UX (Experiência do Usuário):** Diz respeito à experiência geral do usuário ao interagir com um produto ou serviço, buscando torná-la intuitiva, eficiente e satisfatória.

**Cores e ângulos:**

* O uso da cor pode destacar padrões e diferenças, mas pode ser difícil para pessoas com deficiência visual.
* O uso do ângulo é útil para representar dados categóricos em gráficos de pizza, mas pode ser menos preciso para comparações detalhadas.
* Utilizar todas as opções em conjunto pode oferecer uma representação mais completa e compreensível dos dados. A escolha deve considerar a clareza e acessibilidade para o público-alvo.

**Storytelling** é a arte de contar histórias de forma envolvente e cativante, utilizando elementos narrativos como personagens, cenários e conflitos para transmitir informações e ideias de maneira memorável e impactante. É uma poderosa ferramenta de comunicação utilizada para inspirar, ensinar e conectar com o público de forma emocional.

Baseada em: **Uma narrativa, Um contexto e personagens**

**SOR, SOT, SPEC**

1. SOR (Sistema de Origem de Dados): Camada que representa a fonte original dos dados, onde são gerados e coletados inicialmente.

2. SOT (Sistema de Única Verdade): Camada que integra e consolida dados de várias fontes, criando uma única versão confiável dos dados para toda a organização.

3. spec (Camada de Especificação): Camada em que são aplicadas especificações e regras de negócio aos dados após a integração, preparando-os para uso pelos usuários finais ou outras aplicações.

**Governança de dados**

Governança de dados é o conjunto de políticas, processos e controles que garantem a qualidade, segurança e conformidade dos dados em uma organização. Ela estabelece diretrizes para a coleta, uso e gerenciamento dos dados, visando a tomada de decisões mais informadas e confiáveis.

**Atividades:**

* Definir políticas: Estabelecer regras e diretrizes para coleta, uso e proteção dos dados.
* Monitorar a conformidade: Garantir que as políticas sejam seguidas e os dados estejam em conformidade com regulamentos.
* Garantir qualidade: Assegurar a precisão e integridade dos dados por meio de processos de validação e correção.

A governança de dados não faz análise de mercado diretamente. Seu principal objetivo é estabelecer e implementar políticas, processos e controles para garantir a qualidade, segurança, privacidade e conformidade dos dados em uma organização.

**Programação**

Instrução: É um comando ou operação que o computador pode executar. São as ordens que indicam ao computador o que fazer, como somar números ou exibir uma mensagem.

Compilador: É um software responsável por traduzir o código-fonte escrito em uma linguagem de programação para um código executável, compreensível pelo computador.

Sintaxe: É o conjunto de regras e estrutura da linguagem de programação que determina a forma correta de escrever o código, incluindo a ordem e a sintaxe dos comandos para que o computador possa interpretá-los corretamente.

**AMAZON AWS**

**Amazon EC2 (Elastic Compute Cloud):** É um serviço de computação em nuvem da AWS que oferece instâncias de máquinas virtuais sob demanda. Permite criar e gerenciar servidores virtuais de forma flexível e escalável, proporcionando poder de processamento e capacidade de armazenamento na nuvem conforme a necessidade

**Kinesis Data Streams:** É um serviço da AWS que permite a ingestão e processamento de dados de streaming em tempo real, onde os dados são armazenados em ordem e processados por meio de aplicativos ou serviços.

**Kinesis Data Firehose:** É um serviço da AWS que entrega dados de streaming de forma contínua para destinos como S3, Redshift e Elasticsearch, facilitando a carga e o armazenamento de dados em tempo real.

**Kinesis Data Analytics:** É um serviço da AWS que permite a análise de dados de streaming em tempo real usando consultas SQL padrão, tornando mais fácil a extração de insights de dados em movimento.

**Amazon RDS** (Relational Database Service): Oferece bancos de dados relacionais gerenciados (MySQL, PostgreSQL, etc.) para empresas que precisam de um banco de dados tradicional fácil de administrar.

**Amazon DynamoDB:** É um banco de dados NoSQL totalmente gerenciado e escalável, ideal para empresas que buscam flexibilidade e alto desempenho para aplicativos com demandas de escalabilidade e baixa latência.

**Amazon EFS** (Elastic File System): É um sistema de arquivos compartilhado e gerenciado para empresas que necessitam de armazenamento escalável e compartilhado entre instâncias do Amazon EC2 de forma eficiente.

**Hue**: É uma interface web para o ecossistema Hadoop, permitindo aos usuários interagirem com várias ferramentas e serviços, facilitando consultas e visualizações de dados.

**Presto**: É um mecanismo de consulta distribuído que permite realizar consultas SQL em grandes volumes de dados em diversos sistemas de armazenamento, proporcionando análises rápidas e eficientes.

**Ambari**: É uma plataforma de gerenciamento para clusters Hadoop, simplificando a instalação, configuração e monitoramento, tornando o gerenciamento do ecossistema Hadoop mais fácil.

**S3 STANDARD-IA (Intelligent-Tiering):** É uma classe de armazenamento do Amazon S3 que automaticamente movimenta os dados entre o acesso frequente e o acesso menos frequente, reduzindo custos sem a necessidade de gerenciamento manual.

**S3 GLACIER:** É uma classe de armazenamento de arquivamento de longo prazo no Amazon S3. Ideal para dados que são acessados raramente, pois o acesso aos dados requer um tempo de recuperação mais longo.

**INTELLIGENT TIERING:** É uma opção do Amazon S3 que automatiza o movimento de dados entre as classes de armazenamento padrão e Infrequent Access (IA) com base em padrões de acesso, proporcionando economia de custos e desempenho otimizado para diferentes tipos de dados.

**Amazon Elastic Block Store (EBS):** É o armazenamento persistente para instâncias do Amazon EC2, como um disco rígido virtual.

**Amazon Snapshot**: É uma cópia de backup incremental de volumes do EBS, usada para restaurar ou criar novos volumes.

**Amazon S3 (Simple Storage Service):** É um serviço de armazenamento de objetos na nuvem, usado para armazenar e acessar arquivos, imagens, vídeos e muito mais de forma escalável e durável.