Certificado Profissional IBM Data Engineering

O Ecossistema De Dados

O ecossistema de dados moderno inclui uma rede de entidades interconectadas e em constante evolução que incluem:

- Dados, que estão disponíveis em uma série de diferentes formatos, estruturas e fontes.

- Ambiente de dados corporativos, no qual os dados brutos são preparados para que possam ser organizados, limpos e otimizados para uso pelos usuários finais.
- Usuários finais, como partes interessadas de negócios, analistas e programadores que consomem dados para várias finalidades.

Tecnologias emergentes, como Cloud Computing, Machine Learning e Big Data, estão continuamente remodelando o ecossistema de dados e as possibilidades que ele oferece.

Engenheiros de Dados, Analistas de Dados, Cientistas de Dados, Analistas de Negócios e Analistas de Business Intelligence desempenham um papel vital no ecossistema para obter insights e resultados de negócios a partir de dados.

O **objetivo da Engenharia de Dados** é disponibilizar dados de qualidade para análises e tomadas de decisão. E faz isso coletando dados brutos de origem, processando dados para que se tornem utilizáveis, armazenando dados e disponibilizando dados de qualidade para os usuários com segurança.

Responsabilidades e Habilidades de um Engenheiro de Dados

A função de um engenheiro de dados inclui:

- Coleta de dados de fontes diferentes.
- Integração de dados em uma exibição unificada para consumidores de dados.
- Preparação de dados para análises e relatórios.
- Gerenciamento de pipelines de dados para um fluxo contínuo de dados dos sistemas de origem para os sistemas de destino.
- Gerenciamento da infraestrutura completa de coleta, processamento e armazenamento de dados.

Para serem bem-sucedidos em sua função, os engenheiros de dados precisam de uma mistura de habilidades técnicas, funcionais e soft skills.

- As habilidades técnicas incluem trabalhar com diferentes sistemas operacionais e componentes de infraestrutura, como máquinas virtuais, redes e serviços de aplicativos. Também inclui trabalhar com bancos de dados e data warehouses,

pipelines de dados, ferramentas de ETL, ferramentas de processamento de big data e linguagens para consulta, manipulação e processamento de dados.

- Compreender o potencial de aplicação de dados nos negócios é uma habilidade importante para um engenheiro de dados. Outras habilidades funcionais incluem a capacidade de converter requisitos de negócios em especificações técnicas, uma compreensão do ciclo de vida de desenvolvimento de software e as áreas de qualidade de dados, privacidade, segurança e governança.
- As Soft Skills incluem habilidades interpessoais, capacidade de trabalhar de forma colaborativa, trabalho em equipe e comunicação eficaz.

O Ecossistema de Dados e as Linguagens para Profissionais de Dados

O ecossistema de um engenheiro de dados inclui a infraestrutura, ferramentas, estruturas e processos para extrair dados, arquitetar e gerenciar pipelines de dados e repositórios de dados, gerenciar fluxos de trabalho, desenvolver aplicativos e gerenciar ferramentas de BI e relatórios.

Com base em quão bem definida é a estrutura dos dados, os dados podem ser categorizados como:

- Dados estruturados, ou seja, dados bem organizados em formatos que podem ser armazenados em bancos de dados.
- Dados semi-estruturados, ou seja, dados parcialmente organizados e parcialmente livres.
- Dados não estruturados, ou seja, dados que não podem ser organizados convencionalmente em linhas e colunas.

Os dados vêm em uma ampla variedade de formatos de arquivo, como arquivos de texto delimitados, planilhas, XML, PDF e JSON, cada um com sua própria lista de benefícios e limitações de uso.

Os dados são extraídos de várias fontes de dados, desde bancos de dados relacionais e não relacionais até APIs, serviços Web, fluxos de dados, plataformas sociais e dispositivos de sensores.

Uma vez que os dados são identificados e coletados de diferentes fontes, eles precisam ser estagiados em um repositório de dados para que possam ser preparados para análise. O tipo, o formato e as fontes de dados influenciam o tipo de repositório de dados que pode ser usado.

Os profissionais de dados precisam de uma série de linguagens que possam ajudá-los a extrair, preparar e analisar dados. Estes podem ser classificados como:

- Linguagens de consulta, como SQL, usadas para acessar e manipular dados de bancos de dados.
- Linguagens de programação como Python, R e Java, para desenvolvimento de aplicações e controle de comportamento de aplicações.
- Linguagens de shell e scripting, como Unix/Linux Shell e PowerShell, para automatizar tarefas operacionais repetitivas.

Metadata

- Metadata são dados que fornecem informações sobre outros dados e incluem três tipos principais: Metadata técnicos, de processo e de negócios.
- Os Metadata técnicos das bases de dados relacionais são normalmente armazenados em tabelas especializadas na base de dados, designadas por catálogo do sistema.
- Um dos principais objetivos da modelação da gestão de Metadata empresariais é a criação e manutenção de um catálogo de dados viável e de fácil utilização.
- O acesso a um catálogo de dados bem implementado melhora consideravelmente descoberta de dados, repetição, governança e pode também facilitar o acesso aos dados.
- As ferramentas de gestão de Metadata da IBM incluem o InfoSphere Information Server e o Watson Knowledge Catalog.

Repositórios de Dados, Pipelines de Dados e Plataformas de Integração de Dados

Um repositório de dados é um termo geral usado para se referir aos dados que foram coletados, organizados e isolados para que possam ser usados em operações comerciais ou extraídos para geração de relatórios e análise de dados.

RDBMS (Relational Database Management System)

Um banco de dados relacional é uma coleção de dados organizados em uma estrutura de tabela, onde elas podem ser vinculadas ou relacionadas entre si, com base em uma chave primária presente em cada uma, criando várias tabelas.

- Dados organizados em tabelas, possuindo linhas e colunas.
- Linhas são registros.
- Colunas representam atributos.
- Criação de novas tabelas com o uso de query (SQL) simples.
- Permite entende as relações entre os dados.
- Gerar novos insights
- Uso de dados estruturados e organizados como spreadsheets, porém diferente de um spreadsheets é usado para guardar um volume de dados maiores e mais complexos.
- Colunas possui valores e tipos específicos.
- Criação de Tabelas mais significativas apenas com alguns Join em SQL
- Flexibilidade podendo fazer alterações enquanto o DB está sendo usado.
- Minimizando a redundância, Por exemplo, as informações de um cliente aparecem em uma única entrada na tabela de clientes.
- Facilidade em backup, oferece opções fáceis de importação e exportação de dados, facilitando o backup de informações.
- Os banco de dados Cloud-Based são espelhados, o que significa que a perda de dados pode ser observada rapidamente.

Os bancos de dados relacionais usam SQL para consultar dados, o que oferece a vantagem de processar milhões de registros e recuperar grandes quantidades de dados em questão de segundos. Além disso, a arquitetura de segurança dos bancos de

dados relacionais fornece acesso controlado aos dados e também garante que os padrões e políticas para governar os dados possam ser aplicados.

Exemplos de RDBMS:

Cloud-Based

Casos de Uso

- Transações Online (OLTP)
  - Acomoda muitos usuários, CRUD em pequenas quantidades de dados, tempo de resposta rápidos.
- Data Warehouses
  - Em um ambiente de armazenamento de dados, os bancos de dados relacionais podem ser otimizados para processamento analítico on-line (ou OLAP), onde os dados históricos são analisados para fins de inteligência comercial.
- Soluções IoT
  - Boa velocidade, capacidade de coletar e processar dados de dispositivos periféricos.
  - Soluções leves.

Limitações de um RDBMS

- Não funciona bem com dados não estruturados.
- Migração entre dois RDBMS’s só são possiveis quando os dois tem as estruturas idênticas e tipos idênticos.
- Se o valor for maior que o permitido ou definido, resulta na perda de informação.