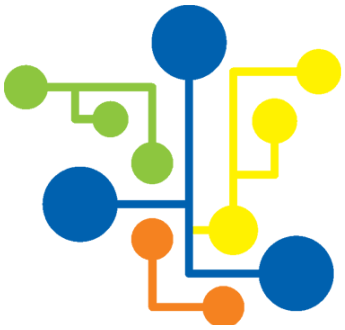

Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Data Storage e Retenção
Data Lake x Data Warehouse

Um Data Lake, por definição, oferece um armazenamento de dados muito mais econômico do que um Data Warehouse. Afinal, com o modelo schema-on-write tradicional dos Data Warehouses, o armazenamento de dados é altamente ineficiente - mesmo na nuvem.

Grandes quantidades de dados podem ser desperdiçadas devido ao problema de tabela esparsa.

Para entender esse problema, imagine construir uma planilha que combina duas fontes de dados diferentes, uma com 200 campos e outra com 400 campos. Para combiná-los, você teria que adicionar 400 novas colunas na planilha original de 200 campos. As linhas da planilha original não possuiriam dados para essas 400 novas colunas e linhas da segunda planilha não conteriam dados das 200 colunas originais. O resultado? Espaço em disco desperdiçado e sobrecarga extra de processamento.

Um Data Lake minimiza esse tipo de desperdício. Cada parte dos dados é atribuída a uma célula e, como os dados não precisam ser combinados na entrada, não existem linhas ou colunas vazias. Isso possibilita armazenar grandes volumes de dados em menos espaço do que o necessário para bancos de dados convencionais relativamente pequenos.

Além de precisar de menos armazenamento, quando o armazenamento e a computação são separados, os clientes podem pagar pelo armazenamento a uma taxa menor, independentemente das necessidades de computação (caso estejam usando ambiente em nuvem). Os provedores de serviços de nuvem, como o Amazon Web Services (AWS), oferecem até mesmo uma variedade de opções de armazenamento em diferentes faixas de preço, dependendo dos requisitos de acessibilidade.

Ao considerar a função de armazenamento de um Data Lake, podemos ainda criar e impor a retenção de dados baseada em diretivas. Por exemplo, muitas organizações usam o Hadoop como um sistema de arquivamento ativo para que possam consultar dados antigos sem precisar ir para a fita. No entanto, o espaço se torna um problema ao longo do tempo, mesmo no Hadoop; Como resultado, tem que haver um processo para determinar por quanto tempo os dados devem ser atendidos no repositório bruto e como e onde arquivá-los.