**O V DE VOLUME EM BIG DATA**

* O V de volume é crítico em Big Data
* Como vamos armazenar grandes conjuntos de dados?
* Como vamos acessar grandes conjuntos de dados armazenados?
* Precisamos realmente armazenar tudo?

**Como armazenar Big Data?**

* Os dados são estruturados ou podem ser estruturados antes do armazenamento?

Usamos um **Data Warehouse!**

* Os dados não são estruturados ou Não podem ser estruturados antes do armazenamento?

Usamos um **Data Lake** ou um **Data Store**

**BANCOS DE DADOS RELACIONAIS X NOSQL**

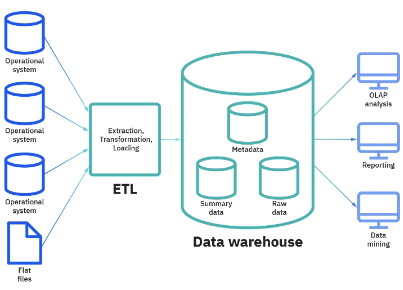
* RELACIONAL
* Banco de dados relacionais são estruturados e com schema (organização dos dados) bem definido
* O schema é definido e criado antes do armazenamento dos dados
* Um data ware house, por exemplo é criado com alguma tecnologia de bando relacional como SGBD( sistema gerenciado de banco de dados) oracle, IBM DB2, Microsoft SQL Server, MySQL, PostgreSQL e muitos outros
* Em um banco de dados relacional os dados são organizados em tabelas que se relacionam
* NOSQL
* Bancos de dados Não relacionais (noSQL) partem do princípio que os dados podem ser semi ou não estruturados e que outros tipos de relacionamentos podem existir entre os dados
* Podemos usar bancos de dados NoSQL para construir data lakes e data stores (eles são conceitos)
* Normalmente não precisamos definir o schema antes do armazenamento ou o schema é definido no momento do armazenamento dos dados
* Existem diversos tipos de bancos de dados NoSQL : Key-Value, Graph Db (grafo), column family e Document

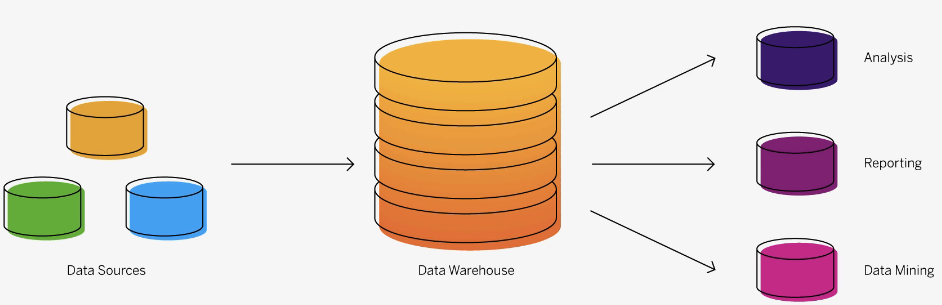
**DATA WAREHOUSES (DW)**

* Um Data Warehouse é um sistema de armazenamento que conecta e harmoniza grandes quantidades de dados de muitas fontes diferentes
* O objetivo do DW é alimentar a inteligência de negócios (Business Intelligence), relatórios e análises e oferecer suporte aos requisitos de negócio, para que as empresas possam transformar seus dados em insights e tomar decisões inteligentes baseadas em dados
* Os DWs armazenam dados atuais e históricos em um único lugar e atuam como a única fonte de informações confiáveis para uma organização
* Os dados fluem para um DW a partir de sistemas transacionais ( como ERP e CRM ), bancos de dados e fontes externas, como sistemas de parceiros, dispositivos de internet das coisas (IoT), aplicativos de mídia social - geralmente em uma cadência regular
* O surgimento da computação em nuvem causou uma mudança no cenário
* Nos últimos anos, os locais de armazenamento de dados mudaram da infraestrutura local tradicional para vários locais, incluindo nuvem privada e nuvem pública
* O schema deve ser definido antes do processo de armazenamento dos dados
* Os DWs modernos são projetados para lidar com dados estruturados e não estruturados, como vídeos, arquivos de imagem e dados de sensor (embora data lakes ainda sejam opções melhores para dados não estruturados)

“Exemplo nosql para data warehouse, uma empresa quer criar uma ia para detecção de tumor, a empresa tem um servidor com S.O, ela cria uma pasta e coloca todas as imagens de vários pacientes com raio x, as imagens são marcadas com Possui Tumor e Não possui tumor, a empresa não quer criar um data lake, então ela cria um schema com uma tabela e 3 colunas = 1- nome da imagem, 2- id se tem tumor ou não, 3- caminho para pasta da imagem, É UMA QUESTÃO DE ARQUITETURA E NÃO DE TECNOLOGIA

* Alguns aproveitam a análise integrada e a tecnologia de banco de dados in-memory ( que mantém o conjunto de dados na memória do computador em vez de no armazenamento em disco ) para fornecer acesso em tempo real a dados confiáveis e impulsionar a tomada de decisões
* alguns SGBDs é possível colocar na memória ram, pois o acesso é mais rapido a memória do que no disco
* Sem Dw é muito difícil combinar dados de fontes heterogêneas, garantir que estejam no formato certo para análise e obter uma visão atual e de longo alcance dos dados ao longo do tempo
* Dw ainda é a principal forma de armazenar dados, data lake e data store ainda não superaram o DW

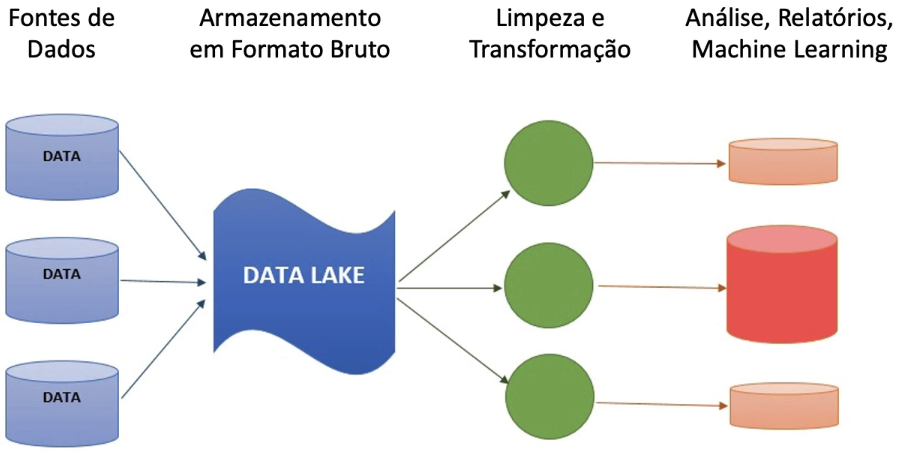
****



**BENEFÍCIOS DO DW**

* Melhor análise de negócios: com o DW, os tomadores de decisão tem acesso a dados de várias fontes e não precisam mais tomar decisões com base em informações incompletas.
* Consultas mais Rápidas: os DWs são construídos, especificamente para recuperação e análise rápida de dados. Com um DW, você pode consultar rapidamente grandes quantidades de dados consolidados com pouco ou nenhum suporte de TI.
* cuidado para não deixar o DW grande demais e causar demora nas consultas, é principal aspecto é a agilidade na consulta de dados
* Melhoria da qualidade dos dados: Antes de serem carregados no DW, os dados passam por um processo de limpeza garantindo que os dados sejam transformados em um formato consistente para apoiar análises e decisões, com base em dados precisos e de alta qualidade
* Visão histórica: ao armazenas dados históricos ricos, um data ware house permite que os tomadores de decisão aprendam com tendências e desafios passados, façam previsões e conduzam a melhoria contínua dos negócios.

**DATA LAKES**



No dw fazemos limpeza e transformação antes, acaba perdendo alguns dados no processo

* Um data lake é um repositório centralizado que permite armazenar todos os dados estruturados e não estruturados em qualquer escala. Podemos Armazenar os dados como estão na fonte, sem ter que primeiro estruturá-los e executar diferentes tipos de análises de painéis e visualizações a processamento de Big Data, análises em tempo real e aprendizado de máquina para orientar melhores decisões
* Dependendo dos requisitos, uma empresa típica exigirá um data Warehouse e um Data lake, pois eles atendem a diferentes necessidades e casos de uso
* A estrutura dos dados ou schema não é definida quando os dados são capturados, isso significa que você pode armazenar todos os dados em formato bruto sem a necessidade de saber quais perguntas de negócio deverão ser respondidas no futuro.
* Primeiro deve-se garantir a captura de todos os dados brutos direto no data lake e depois idealizar schema, ETL e passar pro DW para tomadas de decisões
* - Diferentes tipos de análises, como consultar SQL , análises de Big Data, pesquisa de texto, análises em tempo real e aprendizado de máquina, podem ser usados para descobrir insights
* Os data lakes permitem que as empresas gerem diferentes tipos de percepções sobre os dados, desde relatórios sobre dados históricos até modelos preditivos criados com machine learning
* O principal desafio de uma arquitetura de data lake é que os dados brutos são armazenados sem supervisão do conteúdo. Para que um data lake torne os dados utilizáveis, ele precisa ter mecanismos definidos para catalogar e proteger os dados, Sem esses elementos, os dados não podem ser encontrados ou confiáveis, resultando em um “Pântano de dados” (DATA SWAMP). atender ás necessidades de públicos mais amplos exige que os data lakes tenham governança, gestão de metadados, consistência semântica e controle de acesso
* Data lake é um conceito e pode ser construído com diferentes tecnologias como APACHE HADOOP ou bancos de dados NOSQL.
* Podemos importar dados do DW para o Data lake e vice-versa dependendo das necessidades de negócio da empresa
* Para o DW normalmente usamos ETL ( extração, transformação e carga )
* Para o data lake normalmente usamos ELT ( Extração, carga e transformação )
* Data lakes e DWs podem fazer parte de uma grande estrutura central de armazenamento chamada Data Hub
* **Benefícios do Data Lake**
* **Armazenamento em formato bruto**: não precisamos limpar e transformar os dados antes do armazenamento
* **Importação de qualquer quantidade de dados em tempo real**: os dados são coletados de várias fontes e movidos para o data lake em seu formato original, Este processo permite dimensionar dados em qualquer tamanho, enquanto economiza tempo de definição de estruturas de dados, esquema e transformações
* **Repositório centra para todos os dados da empresa**: os data lakes permitem que várias funções como cientista de dados, engenheiros de machine learning, analista de dados e analista de negócios, acessem dados com sua ferramenta analítica específica.
* **Sem necessidade de movimentação dos dados**: análises podem ser executadas sem necessidade de mover os dados para um sistema de análise separado

**DEFININDO DATA STORES**

* Um data store é um repositório para armazenar e gerenciar de forma persistente coleções de dados que incluem não apenas dados estruturados, mas também tipos de armazenamento variado , como documentos, dados no formato de chave-valor, filas de mensagens e outros formatos de arquivo
* Os tipos mais comuns de data stores:
* Armazenamento de chave-valor ( Redis, Memcached )
* Motor de pesquisa de texto completo ( Elastic Search )
* Fila de mensagens ( Apache kafka )
* Sistemas de arquivos Distribuidos ( Hadoop HDFS, AWS S3 )
* Benefícios do Data Store
* **Armazenamento de variados tipos de dados**: dados que não se encaixam em outros repositórios de armazenamento
* **Flexibilidade**: armazenamento de dados aderente ás necessidades da aplicação final
* **Suporte a dados semi-estruturados**: dados que possuem alguma organização prévia, mas que devem ser usados em seu formato original
* **Custo total menor**: por se tratar de um tipo simplificado de armazenamento o custo total tende a ser menor que a outra solução de armazenamento

**SISTEMAS HÍBRIDOS DE ARMAZENAMENTO**

* Com o avanço do Big data veremos cada vez mais sistemas híbridos de armazenamento, com dados armazenados em diferentes tipos de repositórios, local ou na nuvem.
* DWs, Data Lakes e Data Stores serão usados em conjunto criando assim uma grande estrutura de armazenamento de dados, um Data Hub.

Referências:

Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data [https://www.amazon.com.br/Data-Science-Big-Analytics-Discoveringebook/dp/B00RXHVQF6/ref=sr\_1\_5?ie=UTF8&qid=1458517604&sr=8- 5&keywords=Data+Science+and+Big+Data+Analytics](https://www.amazon.com.br/Data-Science-Big-Analytics-Discoveringebook/dp/B00RXHVQF6/ref=sr_1_5?ie=UTF8&qid=1458517604&sr=8-)

Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance [https://www.amazon.com.br/Big-Data-Analytics-Decisions-Performanceebook/dp/B00S4TBEJK/ref=sr\_1\_1?ie=UTF8&qid=1458517583&sr=8- 1&keywords=big+data+using+smart+big+data](https://www.amazon.com.br/Big-Data-Analytics-Decisions-Performanceebook/dp/B00S4TBEJK/ref=sr_1_1?ie=UTF8&qid=1458517583&sr=8-)

Apache Hadoop

<http://hadoop.apache.org/>

Big Data e Cloud Computing – Desafios e Oportunidades

<https://blog.dsacademy.com.br/big-data-e-cloud-computing-desafios-e-oportunidades>