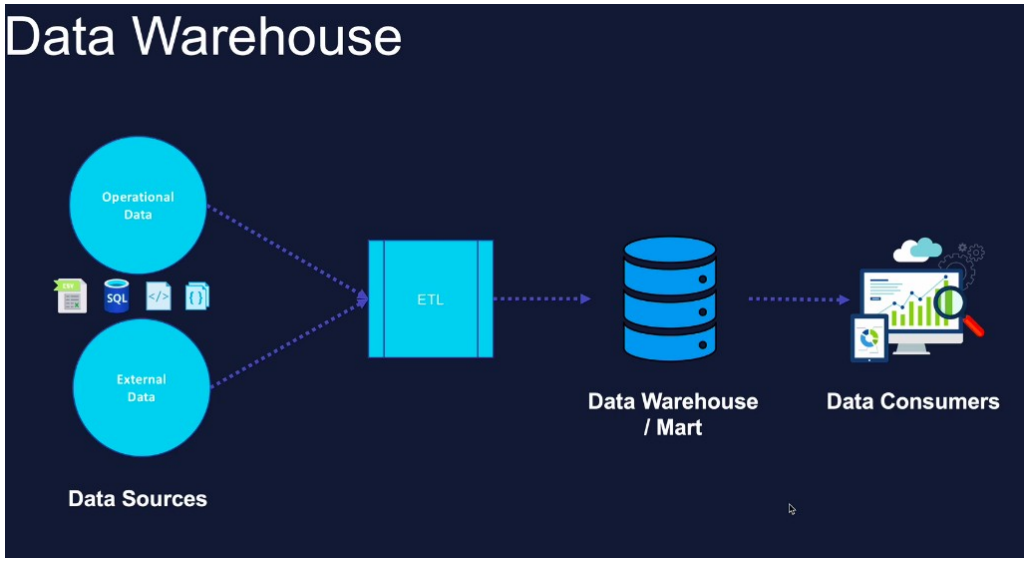
Repositórios de dados

Repositórios de dados - DW

Datawarehouse

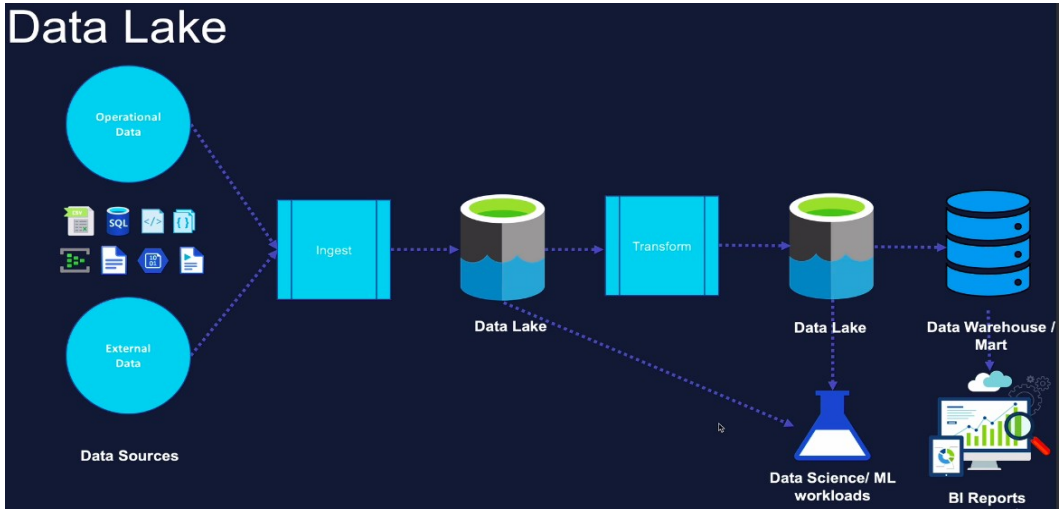
- 1980’ – Willian Inmon – conceito de sistema para suporte a decisão
- Construído com bancos de dados relacionais tradicionais
- Modelagem star schema ou snow flake
- Estrutura de dados rigorosa, schema dos dados pré definido
- ETL – Extract , Transform, Load



Repositórios de dados - DW

Datalake

- Anos 2000, após surgimento do Hadoop
- SGBD, arquivos, API, JSON, XML, semi /sem estrutura
- Modelagem star schema , snow flake , one big table
- Estrutura de dados menos rigorosa, schema dos dados pré definido
- ELT – Extract , Load , Transform



Repositórios de dados - DL

Datalake – Arquitetura de Dados - Medalhão

- Camada transient (landing)
  - Recebimento dos dados, provisória
- Camada bronze (raw)
  - Dados originais, salvos em formato para big data (parquet,orc)
- Camada silver (trusted)
  - Dados limpos, formatados e unificados conforme regras de negocio
- Camada gold (refined)
  - Consumo areas de negocio, relatórios, agregações

Repositórios de dados - DL

Datalake – Apache Hive / Spark SQL

- Sistema de datawarehouse , Open Source
- Originalmente concebido para analisar e consultar dados do Hadoop
- Desenvolvido pelo time de infraestrutura do Facebook
- Baseado em metadados para simular um banco de dados relacional
- Não é um SGBD
- Base para serviços como AWS Glue Catalog
- O sparkSQL usa os metadados (Metastore)

Repositórios de dados - DL

Datalake – Apache Hive / Spark SQL

- Lab Spark SQL

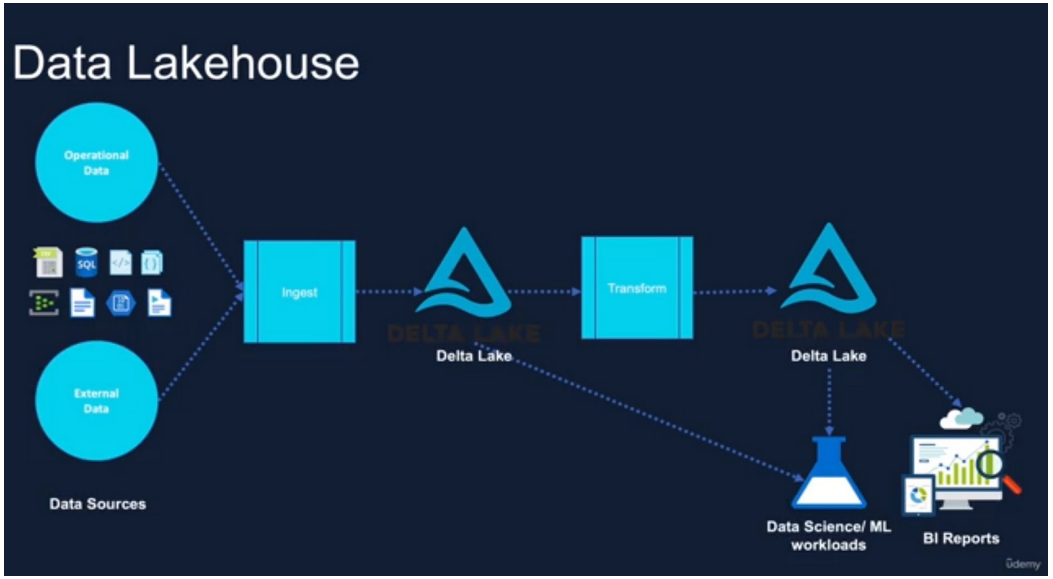
Repositórios de dados – DL vs DW

Comparação Datalake / Data warehouse

	Data Lake	Data Warehouse
Dados	Estruturados, semi estruturados e não estruturados. Dados brutos	Dados estruturados/semi estruturados, processados durante carga
Volume	Concebido para grandes volumes de dados	Concebidos para volumes menores
Armazenamento	Sistemas de arquivos distribuídos e storages	Bancos de dados relacionais
Estrutura de dados	Esquema gerado na leitura do dado	Esquema gerado na escrita do dado
Complexidade	Arquitetura complexa composta de várias ferramentas	Arquitetura mais simples, composta de poucas ferramentas
Manutenção de dados	Updates e deletes são complexos e desafiadores (Sem suporte transações ACID)	Update e deletes são os mesmos do SQL padrão (suporte transações ACID)

Repositórios de dados - Lakehouse

- Final 2020 , após surgimento formato Delta
- SGBD, arquivos, API, JSON, XML, semi /sem estrutura
- Modelagem star schema , snow flake , one big table
- Estrutura de dados menos rigorosa, schema dos dados pré definido
- ELT – Extract , Load , Transform
- Adiciona operações ACID para os dados



Repositórios de dados

- Transação
  - Operação tratada como unidade de trabalho
- ACID
  - Atomicidade
    - Instruções são executadas até o fim ou retornam ao estado inicial , evita corrupção dos dados e perda de informação.
  - Consistencia
    - Transações modificam tabelas somente de maneira predefinida, evita resultado de execução não intencionais
  - Isolação
    - Várias transações atuam no mesmo dado ao mesmo tempo, mas são tratadas individualmente
  - Durabilidade
    - Transações executadas até o fim são preservadas, mesmo em caso de falha do sistema
- ACID Garante integridade e consistência dos dados

Repositórios de dados

Delta

- Open source, acrescenta caracteristicas ACID aos Data Lakes
- Desenvolvido pela empresa que criou o Apache Spark

Vantagens do delta lake

- Transformações ACID
  - Integridade dos dados garantida, alem de acesso a diversas funcionalidades (Update/Delete/Merge)
- Versionamento de dados
  - Possibilidade de reverter o dados para versões anteriores
- Tratamento de metadados
  - Possibilidade de lidar com dados na escala dos pentabytes.
- Formato parquet
  - Delta armazena dados em parquet.

Repositórios de dados

Delta – Lab Delta Table

- Delta Table
- Updates e deletes em Delta Table
- Merge/UpInsert em Delta Table
- History, Time Travel , Vacuum