Data Platform - Um Data Lake de Soluções

Jhon Lucas
Engenheiro de Dados

Quem sou eu?

- Engenheiro de Dados
- Co-founder e Community Manager na Data Train
- Bacharel em Ciência da Computação
- Pós Graduado em Big Data e Machine Learning
- Mestrando em Ciência da Computação
- AWS Community Builder

SPECTRAL.FINANCE

data train

AWS community builders

/jhon-lucas

/l-jhon

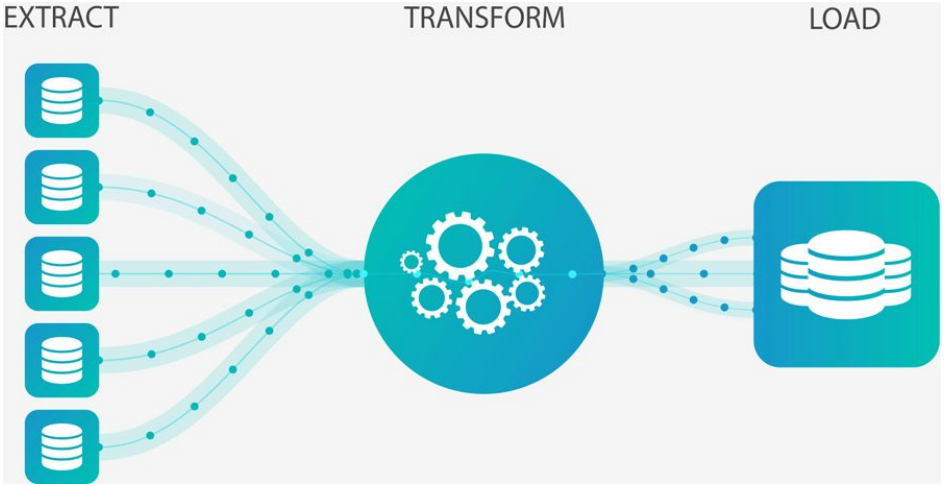
Um pouco de história...

História do BI (Business Intelligence)

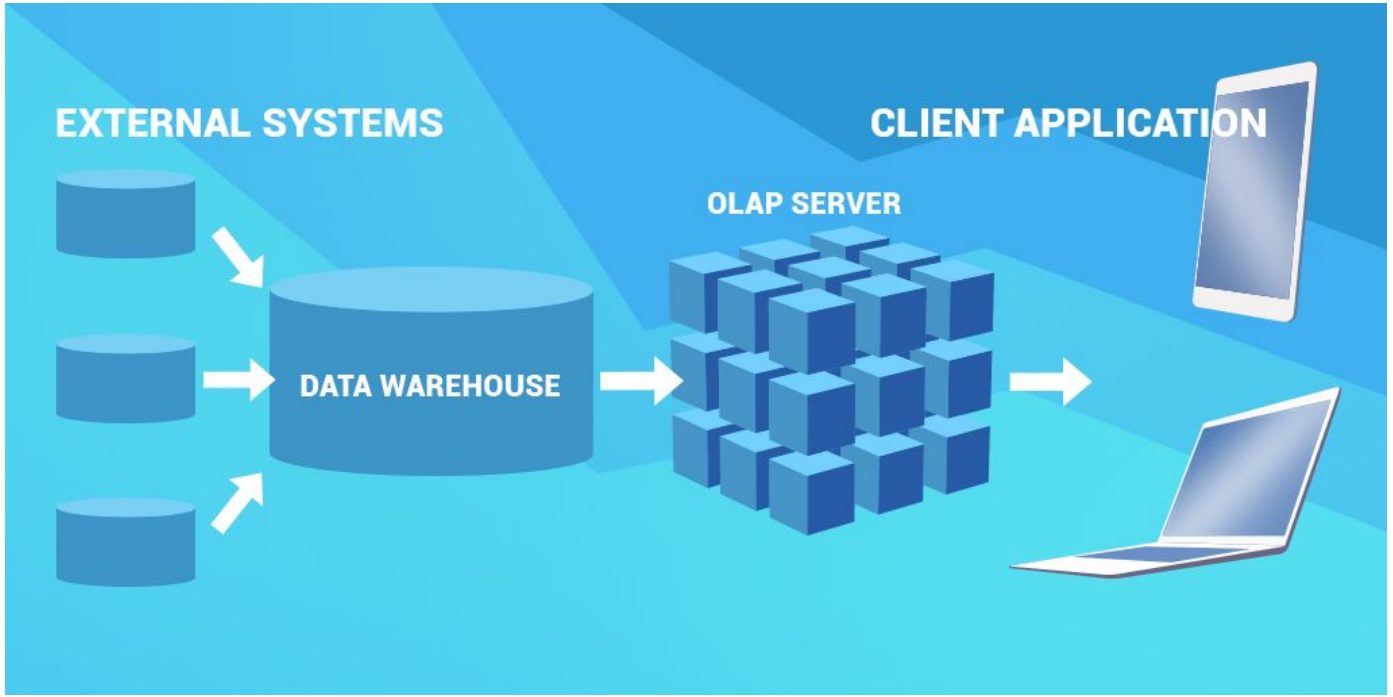
- Anos 70 e 80, quando surgiram os primeiros sistemas de suporte à decisão.
- Objetivo de ajudar as áreas de negócio tomada de decisão.
- Anos 90, soluções de BI (Business Intelligence) surgiram.
- Acesso limitado a grandes organizações e cargos de alto escalão.

Data Warehouse

- Anos 90 - Data Warehouse (DW)
- Única fonte da verdade
- BI 1.0
- Dados acessíveis
- Alto custo para responder novas perguntas de negócio
- ETL (Extract, Transform, Load).
- OLAP (Processamento analítico online) ou cubo.



Arquitetura de um Data Warehouse



Fonte: Cetax, disponível em: https://www.cetax.com.br/data-warehouse/

BI Self-Service

- Anos 90 à 2000 - BI self-service
- Ferramentas de análises disponíveis para analistas de negócios e não somente para gestão.
- Uma necessidade do mercado na época e que continua até hoje.

Shadow IT

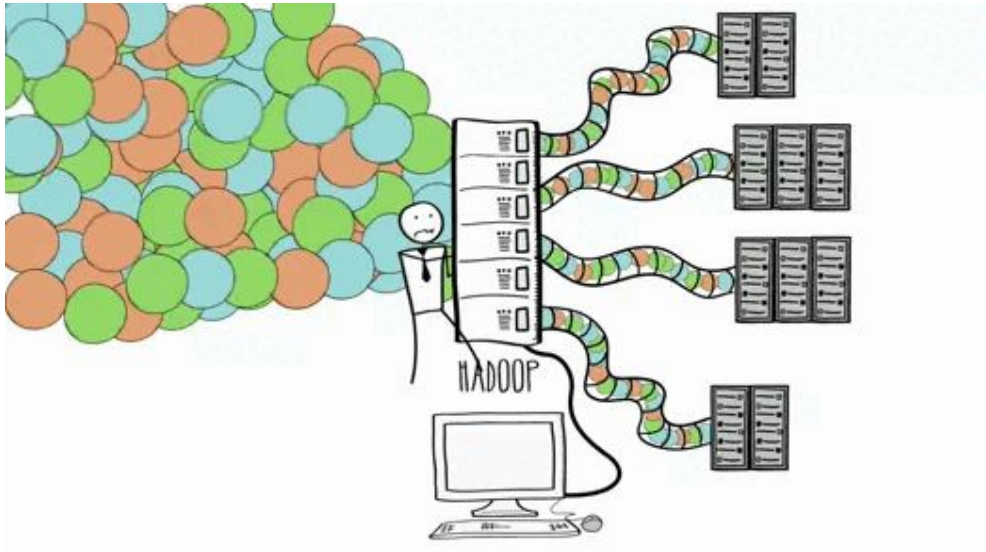
- Ainda sobre os anos 90 à 2000
- Surgimento do Shadow IT
  - Problemas de governança
  - Não uso de boas práticas estabelecidas pela TI
  - Tomada de decisões incorretas
  - Diferentes visões para um mesmo KPI

Data Warehouse no novo milênio

- Nos anos 2000 para ser competitiva a empresa precisava ter um DW
- Fonte única da verdade
- Surgimento de Redes sociais
- 2005 - Big Data (Big problemas)
- 2005 - Ecossistema Hadoop

Ecossistema Hadoop

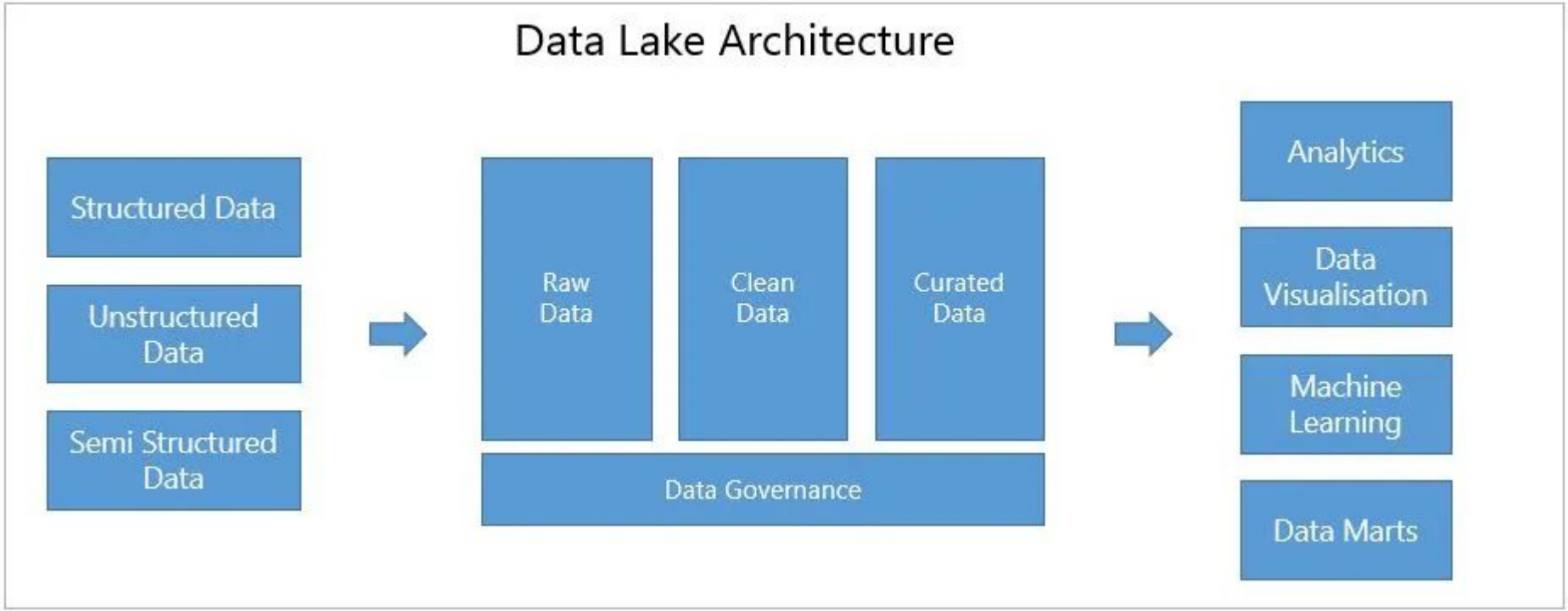
- Baixo custo
- Tolerante a falhas
- Escalável
- Open-source



Origem do Data Lake

- Em meados de 2015 surge o Data Lake
- A história começa com Hadoop sendo uma solução adotada para Data Lakes.
- Muitos dados, necessidade de armazenar em algum lugar
- Hadoop tem o HDFS, solução de armazenamento distribuído, um dos componentes para o Data Lake.
- Dados não estruturados, estruturados, semi-estruturados

Arquitetura de um Data Lake



Fonte: Data Warehouse and Data Science, disponível em: https://dwbi1.wordpress.com/2022/01/20/data-lake-architecture/

A corrida das empresas

- Ciência de Dados
- Inteligência Artificial
- Machine Learning
- Quanto mais dado melhor
- Tudo era IA.
- Contração de Cientistas de Dados

Era moderna dos Dados

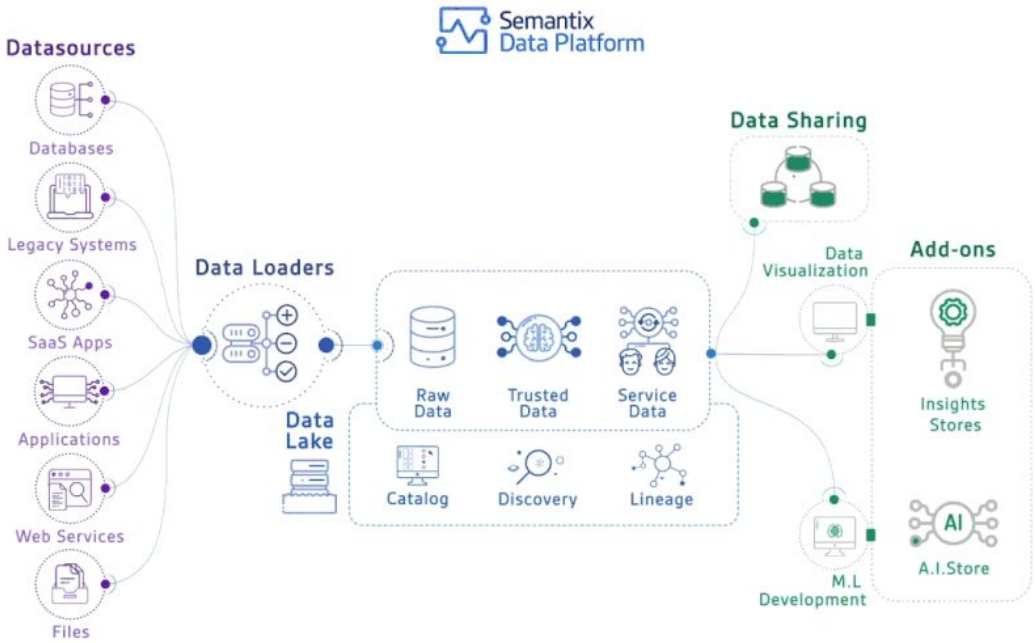
- Data Lake consolidados
- Maioria das empresas entendem cada papel na área de dados
- Democratização
- Cultura Data-driven
- Empresas entendem a necessidade de Engenharia de Dados vir antes de Ciência de Dados
- Surgimento de Plataformas de dados

Data Platform - O que é?

O que é Data Platform?

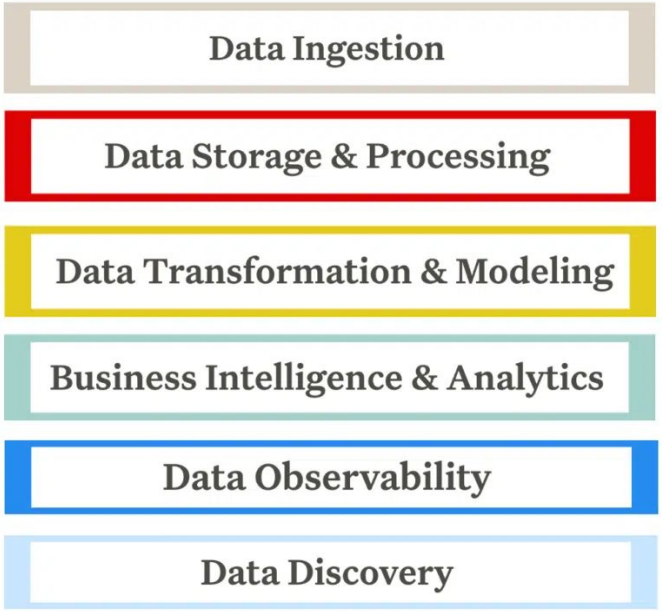
- Conjunto de soluções integradas, ou seja um Data Lake de soluções.
- Solução End-to-end
- Ex: Desde a extração do dado até a entrega de um modelo de ML no ambiente produtivo.

Data Platform



Fonte: Semantix

Camadas de uma Data Platform



Fonte: Monte Carlo

Ingestão de Dados

- Ingestão de diferentes fontes de dados, normalmente no Data Lake.
- ETL e ELT
- Orquestração da Ingestão de Dados

Armazenamento e Processamento

- Necessidade de armazenar e processar grandes volumes de dados.

Apache Flink

Transformação e Modelagem de Dados

- Uma plataforma de dados deve fornecer aos seus usuários soluções para transformar e modelar os dados para que eles possam ser consumidos.

BI e Advanced Analytics

- Soluções de Visualização de Dados
- Soluções para exploração de dados
- Experimentos de Modelos de ML
- Deploy de Modelos de ML

Amazon SageMaker

Observabilidade de Dados

- Em uma plataforma de dados é importante ter total visibilidade sobre os pipelines, sobre a plataforma em si.
- Pipeline falhou?
- Modelo de ML está com drift?
- Como está a qualidade do dado?
- Temos anomalias nos dados?

great_expectations

Governança de Dados

- Uma plataforma deve fornecer acessos aos dados e seus recursos de forma segura e com governança.
- Políticas de acesso bem definidas
- Catálogo de Dados
- Definição de processos

Data Discovery (Descoberta de Dados)

- Data Discovery permite uma visão geral dos dados que o usuário tem a disposição para explorar, criar modelos, insights e etc.

Algumas considerações...

Data Platform não é só tecnologia

- De fato há muita tecnologia envolvida, mas há outros pontos importantes que devem ser considerados em uma plataforma de dados
- Guidelines
- SLA
- Boas práticas
- Cultura DataOps, MLOps e DevOps
- CI/CD pipelines, diversos princípios da Engenharia de Software são utilizados para criar uma plataforma.

Data Platform não é simples de manter

- Precisa de um time dedicado
- Diversos desafios técnicos
- Integrar diferentes tecnologias para ter um produto final não é tão simples
- Requer maturidade de dados

Data Platform in-house

- Envolver o negócio
- Não reinventar a roda
- Pensar que a plataforma será utilizada por pessoas e para isso é necessário ter processos bem definidos
- É importante ter um time de especialistas em cada camada da plataforma.

Data Platform - Carreira

- Engenharia de Dados
- Cultura DevOps
- Engenharia de Software é sempre bem vinda na área de dados, mais ainda quando se trata de plataforma.

Data Platform no mercado

Faça Parte da Data Train

data train

http://datatrain.com.br/

Obrigado! Dúvidas?

- Engenheiro de Dados
- Co-founder e Community Manager na Data Train
- Bacharel em Ciência da Computação
- Pós Graduado em Big Data e Machine Learning
- Mestrando em Ciência da Computação
- AWS Community Builder

SPECTRAL.FINANCE

data train

AWS community builders

/jhon-lucas

/l-jhon