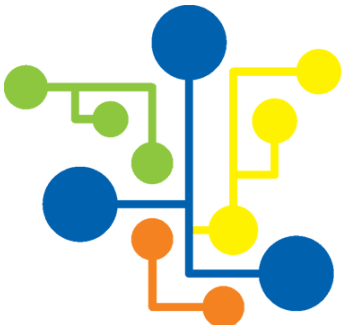


Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Roadmap Para um Data Lake de Sucesso

As empresas hoje estão explodindo de dados, incluindo dados das bases existentes, outputs de aplicativos, dados de streaming de comércio eletrônico, mídia social, aplicativos e dispositivos conectados na Internet das Coisas (IoT).

Estamos todos bem versados no Data Warehouse, projetado para capturar a essência do negócio de outros sistemas empresariais, como sistemas ERP e CRM, estoque e transações de vendas que permitem que analistas e usuários de negócios obtenham insights e façam importantes decisões de negócios a partir desses dados.

Mas novas tecnologias, incluindo plataformas móveis, sociais e IoT, estão impulsionando volumes de dados muito maiores, maiores expectativas dos usuários e uma rápida globalização das economias.

As organizações estão percebendo que as tecnologias tradicionais não podem atender às novas necessidades de negócios.

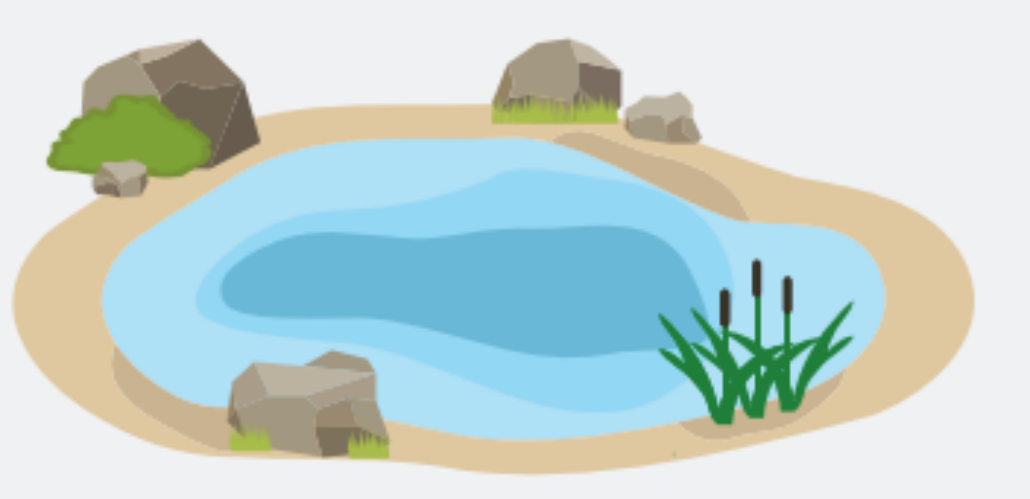
Como resultado, muitas organizações estão se voltando para arquiteturas de scale-out, como Data Lakes, usando o Apache Hadoop e outras tecnologias de Big Data. No entanto, apesar do crescente investimento em Data Lakes e tecnologia de Big Data - US $ 150,8 bilhões em 2017, um aumento de 12,4% em relação a 2016 - apenas 14% das organizações relatam implementar seu projeto de prova de conceito (PoC) de Big Data, em produção.

Uma razão para essa discrepância é que muitas organizações não veem um retorno do seu investimento inicial em tecnologia e infraestrutura de Big Data. Isso geralmente é porque essas organizações falham em fazer a coisa certa, ficando aquém quando se trata de projetar os dados corretamente. Em (muitos) outros casos, não há uma cultura data-driven.

Em última análise, essas organizações criam "pântanos" de dados (Data Swamps) que são realmente úteis apenas em casos de uso exploratório ad-hoc (algo ainda incomum em muitas empresas, embora seja utilizado em pesquisa e desenvolvimento). Para as organizações que se movem para além de uma PoC, muitas estão unindo a flexibilidade de um Data Lake com boas práticas da governança e controle de dados. Essa é a chave para obter um ROI significativo dos investimentos de tecnologia para Big Data.

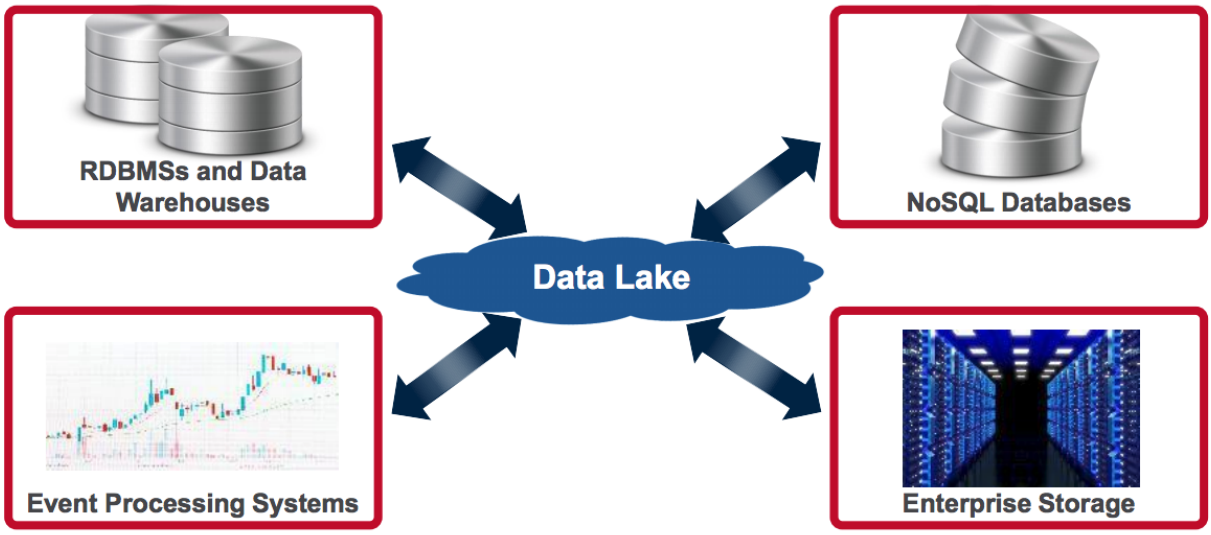
Se você (Engenheiro de Dados) perguntar aos seus Analistas e Cientistas de Dados quais dados eles querem usar no futuro eles provavelmente dirão "Tudo que estiver disponível". Mas realmente precisamos de todos os dados possíveis ou apenas do que é relevante para resolver determinado problema de negócio? Em geral, usamos o Data Lake para armazenar:

- Dados transacionais brutos
- Dados semi-estruturados e não estruturados
- Streaming de dados
- Dados históricos que não foram migrados para os sistemas de produção
- Dados que podem ser úteis para análises no futuro



Data Lakes tornam-se pântanos de dados se não são gerenciados corretamente. Certifique-se de que seu Data Lake possui as seguintes características:

- Arquitetura para atender às metas de negócio da empresa
- Gerenciamento para proteger os dados
- Integração para ser alimentado com dados de diferentes fontes
- Acessibilidade para “self-service” e oferecer aos Cientistas de Dados a oportunidade de aplicar Real-Time Analytics



O Roadmap de sucesso para um Data Lake, deve contemplar:

1. Amplas capacidades analíticas

   - Fontes variadas de dados
   - Acesso a partir de diferentes ferramentas analíticas
   - Computação intensa para a execução de algoritmos de Machine Learning

2. Interoperabilidade

   - Integração com diferentes sistemas heterogêneos
   - Processamento de eventos e mensagens

3. Continuidade do Negócio

   - Alta disponibilidade e tolerância a falhas
   - Disaster Recovery (Backup e Restore)
   - Data Recovery (recuperação contra falhas e corrupção nos dados

4. Custo Reduzido

   - Uso de hardware commodity (de baixo custo)
   - Baixo overhead de administração

- Alta performance
- Compartilhamento de recursos (virtualização)

5. Capacidades Multi-tenancy

- Habilidade de gerenciar diversos recursos (computacionais) a partir de um único ponto

Conseguir reunir todas essas características em um Data Lake é muito difícil, mas este deve ser o Roadmap e o objetivo de qualquer organização interessada em obter o máximo do Big Data Analytics. E você, como Engenheiro de Dados, é um dos responsáveis por isso.

Referências:

IDC. “Worldwide Semiannual Big Data & Analytics Spending Guide.”

Gartner. “Market Guide for Hadoop Distributions.”

Architecting Data Lakes (Free e-book)
https://www.oreilly.com/data/free/architecting-data-lakes.csp

Como Construir um Data Lake de Sucesso
https://mapr.com/resources/videos/how-build-successful-data-lake/assets/how_to_build_a_successful_data_lake_webinar_-_160517.pdf