Data Warehouses modernos e Data Lakes

Diamante

AGGRANDIZE

COMMVAULT

TD SYNNEX

Platina

GOLDENGATEBR
Consultoria & Treinamentos

DISCOVER

Ouro

scansource

VERTICA
by opentext™

Prata

TRACES

CAFÉ COM CLOUD

ROX
We take care of your data

Apoio

FIAP

GRUPO POSEIDON DIGITAL

Olá!

Eu sou Rubens Oliveira

Sou arquiteto de dados, cientista de dados, autor de livros e professor de tecnologia.

Você pode me encontrar em:

https://www.linkedin.com/in/rubens-oliveira-msc/

Agenda

O que é um Data Warehouse?

O que é um Data Lake?

Áreas de um Data Lake

Organizando um Data Lake

Modernizando um Data Warehouse

Principais Desafios

Pensamentos Finais

*"**Data Warehouse e Data Lakes**: a sinergia entre a ordem dos dados e a liberdade de exploração, desvendando um mar de oportunidades!"*

Inspirado na busca pela combinação estruturada e flexível dos dados, transformando-os em vantagens estratégicas.

O que é um DataWarehouse?

Data Warehouse

O que é?

- Sistema que armazena dados históricos usados no processo de tomada de decisão;
- Integra os dados corporativos de uma empresa em um único repositório.

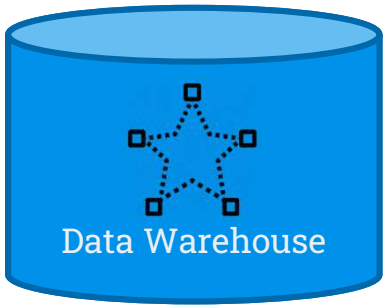
Para que serve?

- Para criar uma visão única e centralizada dos dados que estavam dispersos em diversos bancos de dados;
- Permite que usuários finais executem consultas, gerem relatórios e façam análises.

Data Warehouse

Os dados são mais valiosos quando são integrados a partir de vários sistemas.
Visualização completa de um cliente:

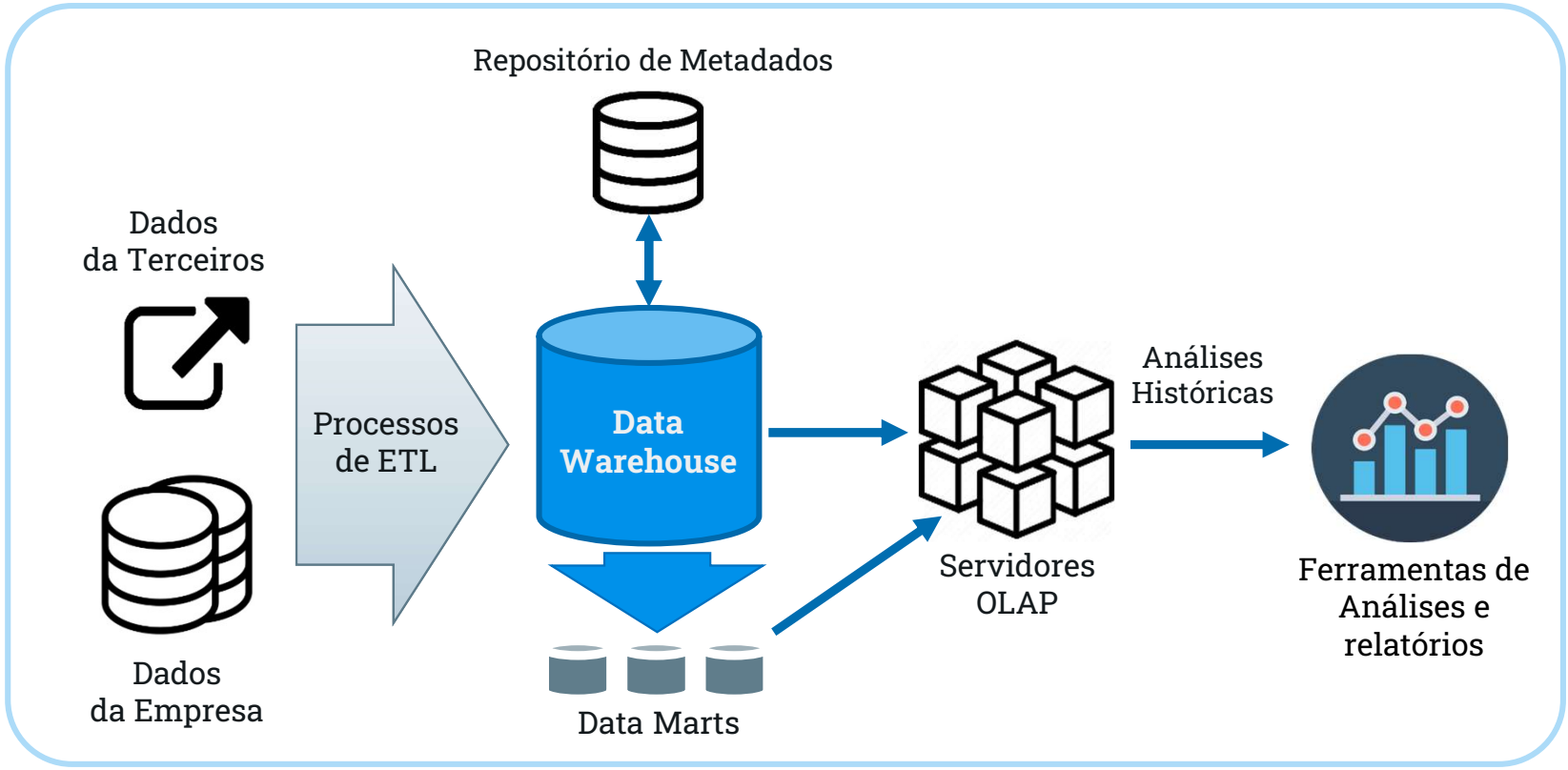
- Atividade de vendas +
- Faturas em atraso +
- Solicitações de suporte / ajuda



O Data Warehouse é projetado para ser amigável, utilizando a terminologia comercial.

Frequentemente é construído com um modelo de dados desnormalizado.

Data Warehouse Arquitetura



O que é um Data Lake ?

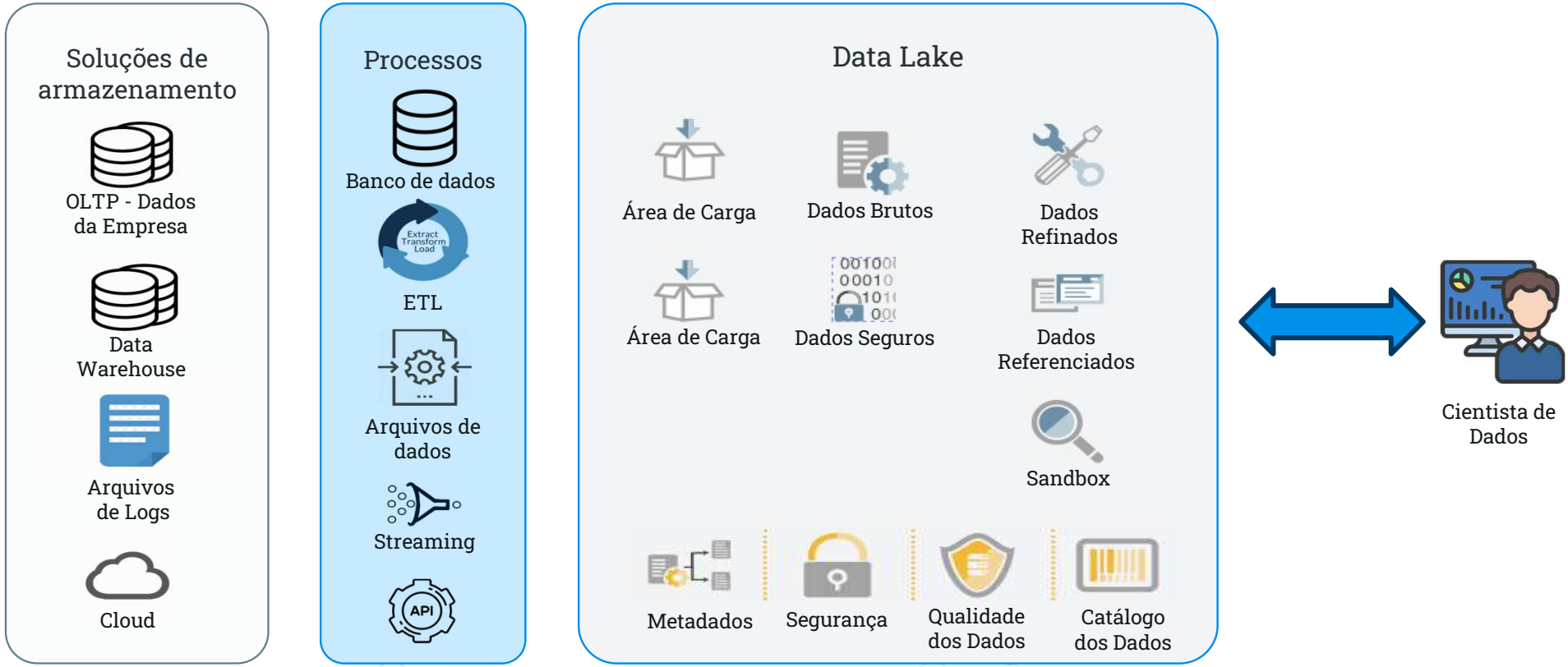
O que é um Data Lake?

É um repositório para analisar grandes quantidades de fontes diferentes de dados em seu formato nativo.

Uma plataforma arquitetônica para hospedar todos os tipos de dados:

- Dados gerados por máquina (IoT, logs);
- Dados gerados por humanos (tweets, e-mail);
- Dados operacionais tradicionais (vendas, estoque).

Arquitetura de um Data Lake



Objetivos de um Data Lake

Reduzir o esforço inicial, ingerindo dados em qualquer formato, sem a necessidade de um esquema inicialmente;

Facilitar a aquisição de novos dados, para que possam estar disponíveis para ciência e análise de dados rapidamente;

Armazenar um grande volume de dados multi-estruturados em seu formato nativo;

Objetivos de um Data Lake

Adiar o trabalho para "**esquematizar**" depois que o valor e os requisitos forem conhecidos;

Obter agilidade mais rapidamente do que um data warehouse tradicional;

Acelerar a capacidade de **tomada de decisão**.

Implementação de um Data Lake

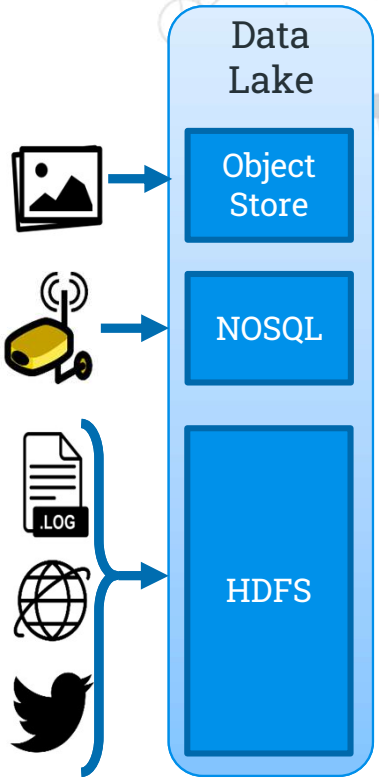
Um Data Lake é uma ideia conceitual. Pode ser implementado com uma ou mais tecnologias.

O HDFS (armazenamento de arquivos distribuídos do Hadoop) é uma opção muito comum para o armazenamento de um Data Lake.

No entanto, o Hadoop não é um requisito fundamental. Um Data Lake também pode abranger mais de um cluster do Hadoop.

Os bancos de dados NOSQL também são muito comuns.

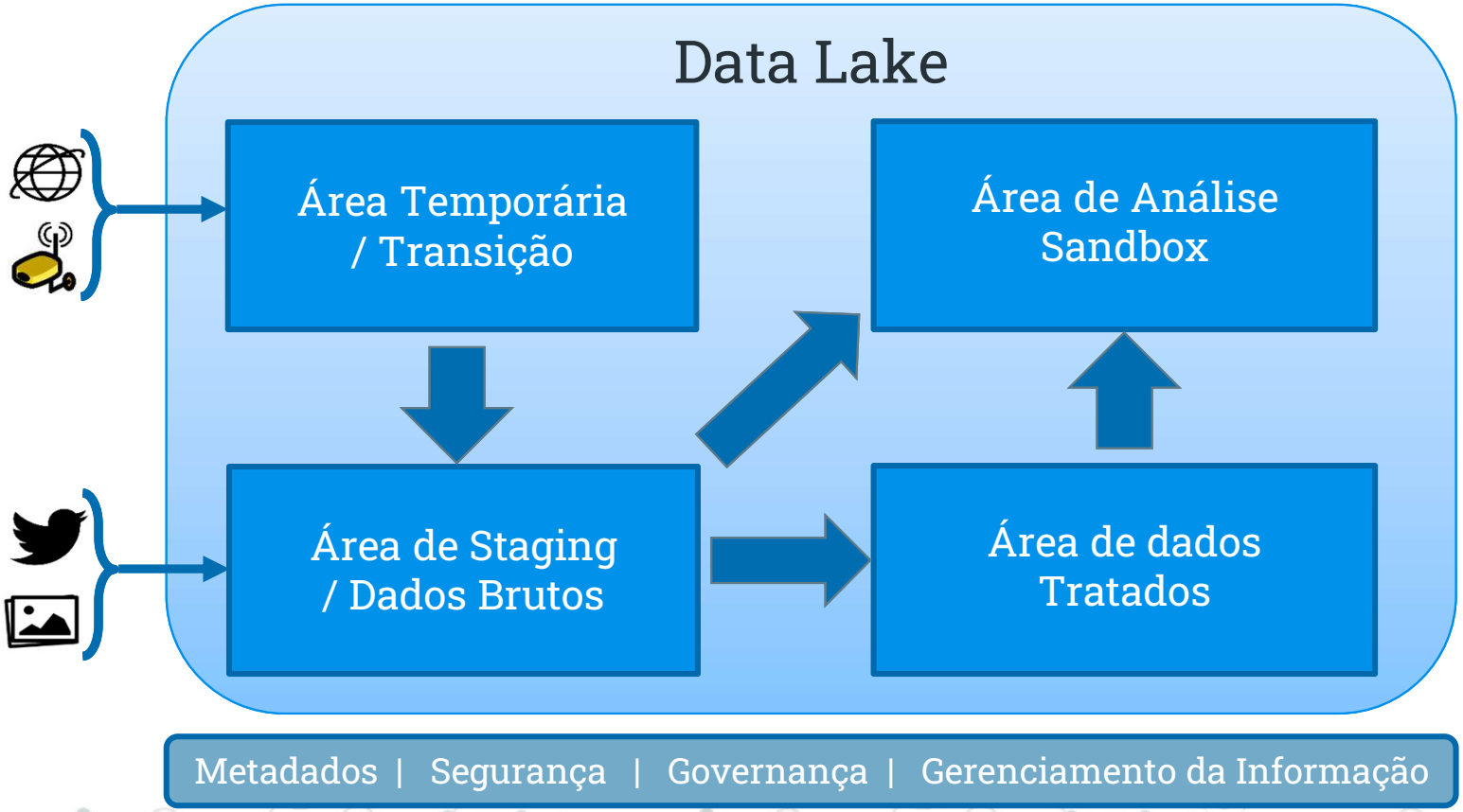
Os armazenamentos de objetos (na nuvem computacional) também podem ser usados.



Estágios de maturidade de um Data Lake

Áreas de um Data Lake

Áreas de um Data Lake

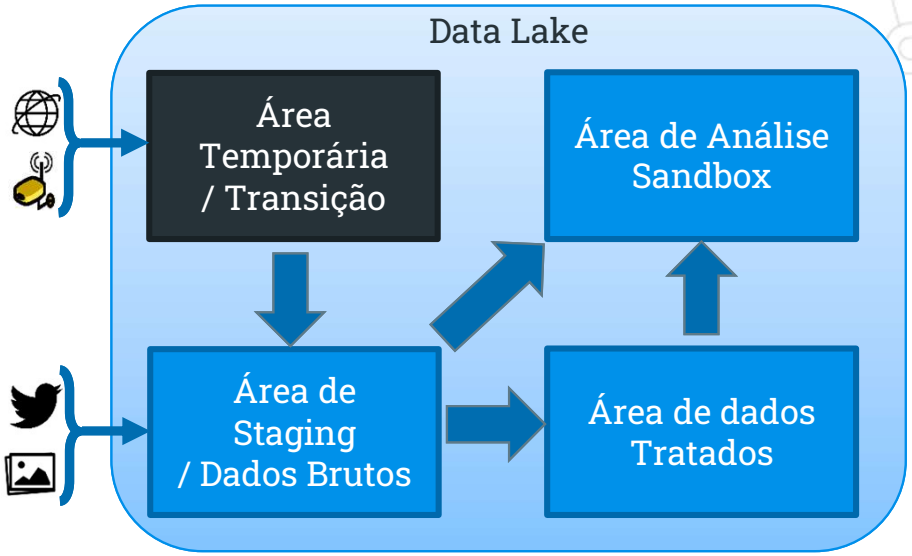


Áreas Temporária / Transição

Área utilizada para as verificações de qualidade ou validade dos dados são necessárias antes que os dados possam ser movidos para a **área de staging**.

Todas as **áreas temporárias** são consideradas "área da cozinha" com acesso altamente limitado.

- Área transitória;
- Área de dados brutos;
- Área de preparação.



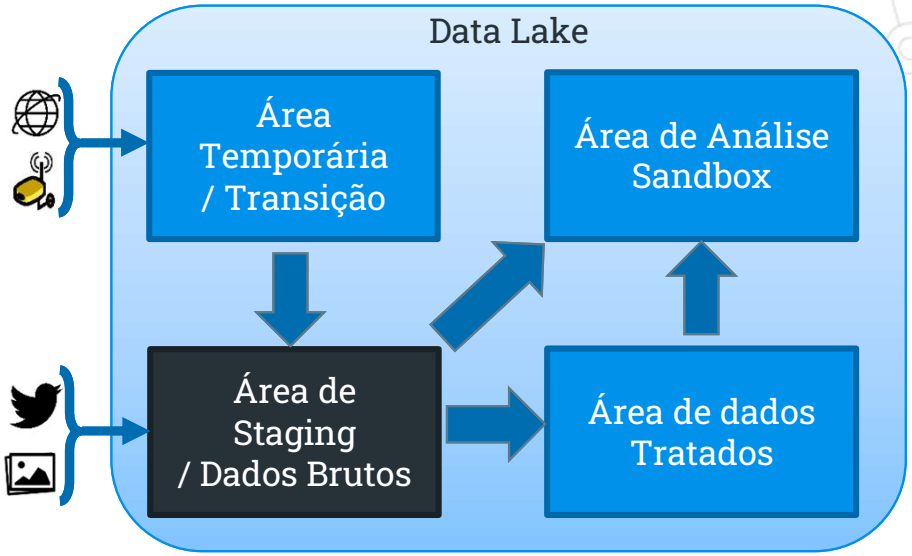
Área de Stagging / Dados Brutos

A **área de staging** é imutável à alteração.

O histórico é mantido para acomodar necessidades desconhecidas futuras;

Suporta os seguintes tipos de dados:

- Streaming
- Batches
- On-line
- Full Load
- Carga incrementais, entre outros.

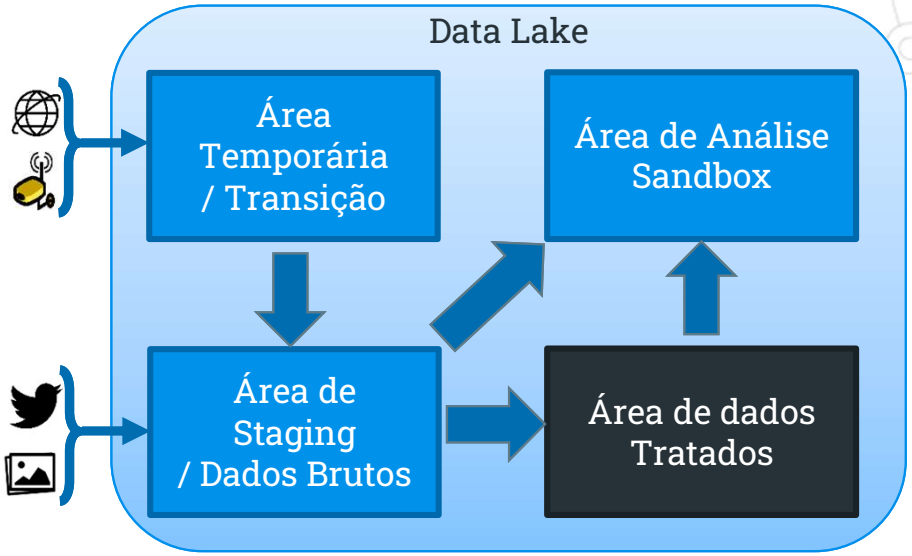


Área de dados Tratados

A **área de dados tratados** é onde os dados são organizados e limpos para entrega das análises;

- Consumo de dados;
- Consultas de várias fontes diferentes (Federated Queries);
- Fornece dados para outros sistemas.

A maioria dos acessos aos dados de autoatendimento ocorre na nesta área.

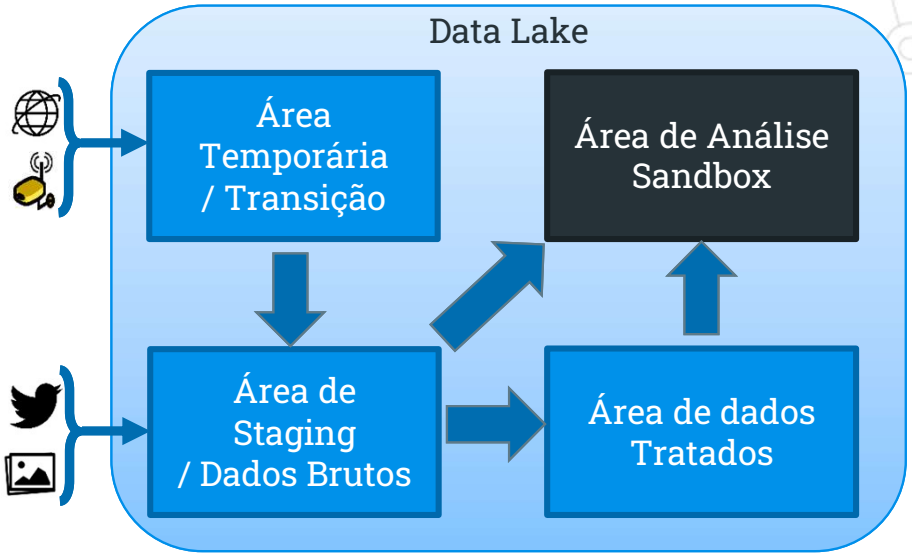


Área de Análise / Sandbox

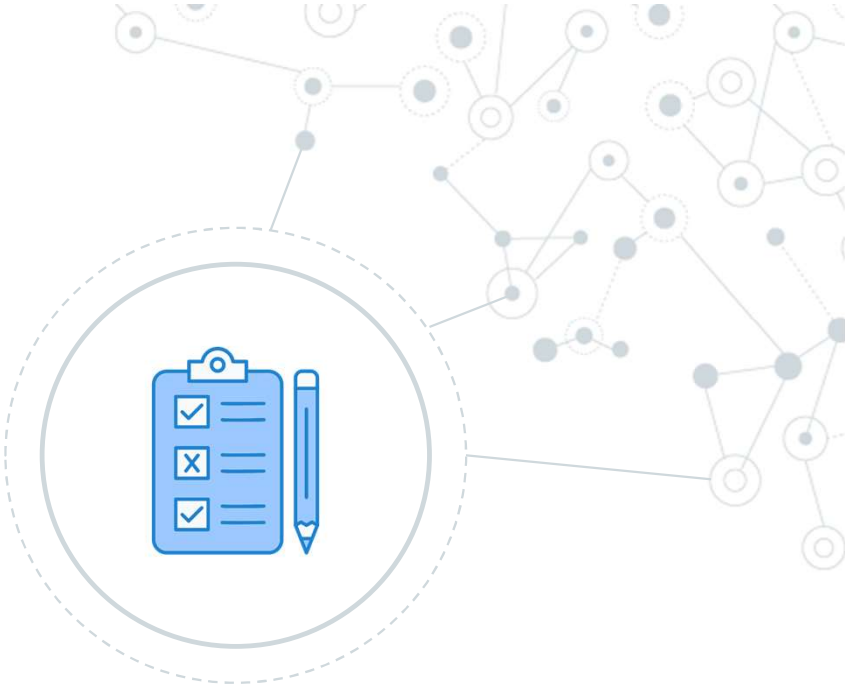
A **área de análise**, permite atividades exploratórias aos dados.

- Mínima governança dos dados para as análises;
- Esforços valiosos são "promovidos" no Analytics.

A **sandbox** pode ser utilizada tanto para a área de análise quanto para o data warehouse.



Organizando um Data Lake



Organizando um Data Lake

Planejar a estrutura com base na recuperação ideal de dados.

O padrão da organização do Data Lake deve ser auto documentado.

A organização é frequentemente baseada em:

- **Assuntos**
- **Particionamento do Tempo**
- **Limites de Segurança**
- **Downstream APP/ funcionalidade**

Os recursos de metadados de sua tecnologia terão um grande impacto em como você escolhe lidar com a organização.

O objetivo é evitar um pântano de dados caótico

Organizando um Data Lake

Outras opções que afetam a organização e / ou metadados:

Política de Retenção de Dados
- Dados temporários
- Dados permanentes
- Período aplicável (ex: vida útil do projeto)

Impacto / Criticidade nos Negócios
- Alto
- Médio
- Baixo

Proprietário / Administrador / Gerente de Projeto

Probabilidade de acesso a dados
- Dados recentes / atuais
- Data histórica

Classificação Confidencial
- Informação pública
- Somente para uso interno
- Confidencial fornecedor / parceiro
- Informações de identificação pessoal
- Sensível - Financeiro
- Sensível - propriedade intelectual

Coexistência do Data Lake e Data Warehouse

Valores do **Data Lake**

- Agilidade
- Flexibilidade
- Entrega Rápida
- Exploração

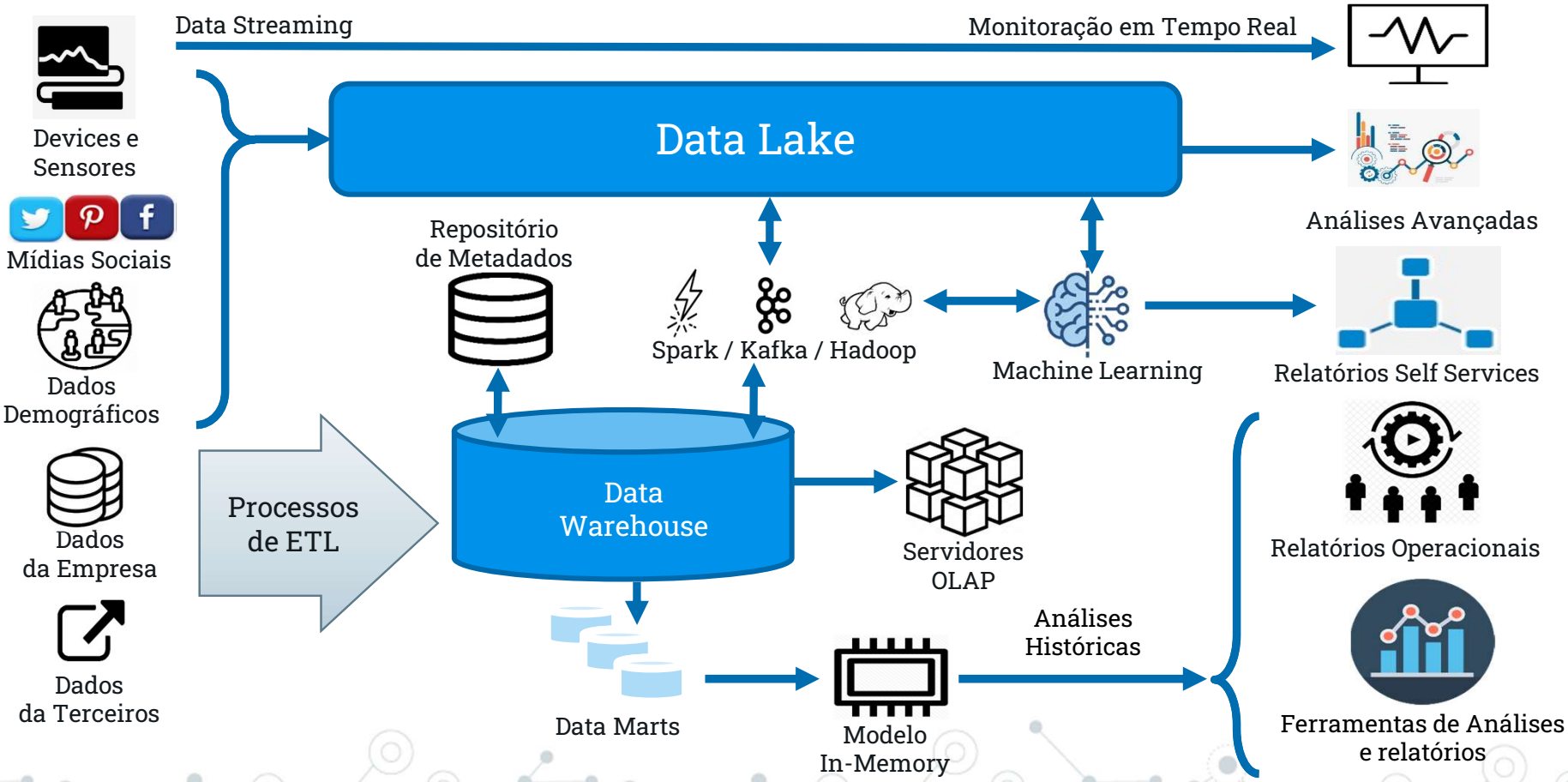
Valores do **Data Warehouse**

- Governança
- Confiabilidade
- Padronização
- Segurança

Menos esforço	**Aquisição de dados**	Mais esforço
Mais esforço	**Recuperação de dados**	Menos esforço

Modernizando um Data Warehouse

Modernizando um Data Warehouse



O que torna um Data Warehouse Moderno

- **Variedade de áreas temáticas e fontes de dados** para análise com capacidade para lidar com grandes volumes de dados;
- **Expansão** além de um único Data Mart/Data Warehouse relacional - inclusão do Kafka, Hadoop, Spark, Data Lake ou NoSQL;
- **Design lógico** através arquitetura multiplataforma equilibrando escalabilidade e desempenho;
- **Virtualização de dados** além da integração de dados;

O que torna um Data Warehouse Moderno

- **Suporte para todos os tipos e níveis de usuários;**
- **Implantação flexível** (inclusive móvel), que é dissociada da ferramenta usada para desenvolvimento;
- **Modelo de governança** para oferecer suporte a confiança e segurança e gerenciamento de dados;
- Suporte para promover **soluções de auto atendimento** para o ambiente corporativo.

O que torna um Data Warehouse Moderno

- Alguma **automação no Data Warehouse** para melhorar a velocidade, consistência e adaptação flexível às mudanças;
- Uma **área restrita de análise ou área de trabalho** para facilitar a agilidade em um ambiente de BI;
- Suporte para geração de **informações "Self-Service"** para aumentar as análises corporativas;
- **Descoberta de dados, exploração de dados,** preparação de dados de autoatendimento.

Principais Desafios

Principais Desafios

Tecnologia

- Estratégia de persistência poliglota
- Arquitetura complexa e multicamada
- Armazenamento e escalabilidade desconhecidos
- Recuperação de dados
- Trabalhando com dados não tratados
- Gestão de mudanças

Processos

- Equilíbrio correto entre trabalho atrasado x trabalho futuro
- Ignorar as melhores práticas estabelecidas para gerenciamento de dados
- Qualidade dos dados
- Governança
- Segurança

Pessoas

- Expectativas
- Gerenciamento de dados
- Esforço redundante
- Habilidades necessárias para fazer uso analítico dos dados

Maneira de começar a modernização de um Data Warehouse

- Criar um **Data Lake** como área de preparação para **Data Warehouse**.
- Descarregar dados arquivados do **Data Warehouse** de volta ao **Data Lake**.
- Introduzir um novo tipo de dados para obter tempo para o planejamento a longo prazo do Data Lake.

Obter valor real

- Integração seletiva com o **Data Warehouse;**
- Aplicação de **ciência de dados** com experimentação com APIs;
- Utilização de conjuntos de ferramentas analíticas no topo do **Data Lake;**
- Utilizar interfaces de consulta para um **Data Lake.**

Pensamentos Finais

Pensamentos Finais

- O **Data Warehouse** tradicional ainda é importante, mas precisa coexistir com outras plataformas;
- Planejar o **Data Lake** com recuperação de dados em mente.
- Equilibrar os processos de **ETL com técnicas de virtualização de dados** em um ambiente multiplataforma.
- Trabalhar utilizando um modelo **ágil**. Realizar provas frequentes de conceito dos projetos, visando provar suposições.

Obrigado !

Alguma dúvida?

Você também poderá me encontrar em:

https://www.linkedin.com/in/rubens-oliveira-msc/

Diamante

AGGRANDIZE

COMMVAULT

TD SYNNEX

Platina

GOLDENGATEBR
Consultoria & Treinamentos

DISCOVER

Ouro

scansource

VERTICA
by opentext™

Prata

TRACES

CAFÉ COM CLOUD

ROX
We take care of your data

Apoio

FIAP

GRUPO POSEIDON DIGITAL