



Dados Engenharia com Databricks



Objetivos do curso

- Aproveitar a plataforma Databricks Lakehouse para executar as principais responsabilidades do desenvolvimento do pipeline de dados
- Use SQL e Python para escrever pipelines de dados de produção para extrair, transformar e carregar dados em tabelas e exibições no lakehouse
- Simplifique a ingestão de dados e a propagação de mudanças incrementais usando Recursos e sintaxe nativos do Databricks
- Orquestrar pipelines de produção para fornecer novos resultados para análises e painéis ad-hoc



Agenda do curso

- Módulo: Databricks Workspace and Services • Módulo: Delta Lake
- Módulo: Entidades Relacionais em Databricks
- Módulo: ETL com Spark SQL • Módulo: Python OPCIONAL para Spark SQL • Módulo: Processamento de Dados Incremental • Módulo: Arquitetura Multi-Hop
- Módulo: Delta Live Tables
- Módulo: Orquestração de Tarefas com Jobs
- Módulo: Executando uma consulta DBSQL • Módulo: Gerenciando permissões • Módulo: Painéis de produção e consultas em DBSQL





Os Databricks casa do lago Plataforma



Usando a Plataforma Databricks Lakehouse

objetivos de aprendizado

- Descrever os componentes do Databricks Lakehouse
- Concluir tarefas básicas de desenvolvimento de código usando os serviços do Databricks Área de trabalho de ciência e engenharia de dados
- Realizar operações de mesa comuns usando Delta Lake no Lakehouse



Usando a Plataforma Databricks Lakehouse

Agenda

- Introdução à Plataforma Databricks Lakehouse
- Introdução ao espaço de trabalho e serviços do Databricks
 - Usando clusters, arquivos, notebooks e repositórios
- Introdução ao Lago Delta
 - Manipulação e otimização de dados em tabelas Delta





casa do lago

Uma plataforma simples para unificar todos os seus dados, análises e cargas de trabalho de IA

Clientes

7000+

em todo o mundo



Criadores originais de:



Apoiando empresas em todos os setores

Saúde e Vida ciências



Fabricação e Automotivo



Meios de comunicação &

Entretenimento



Financeiro Serviços



Setor público



Varejo e bens de consumo



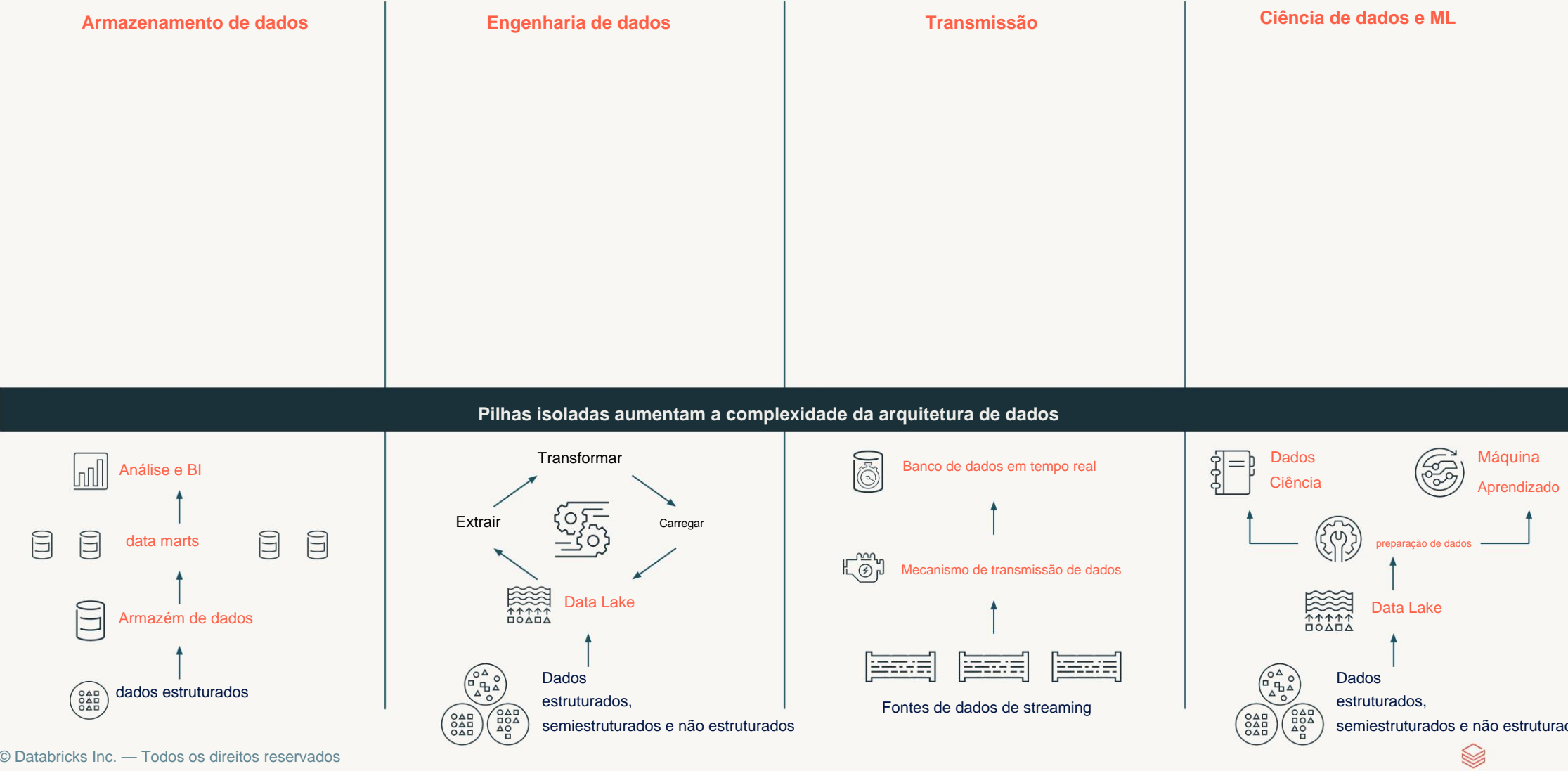
Energia e serviços públicos



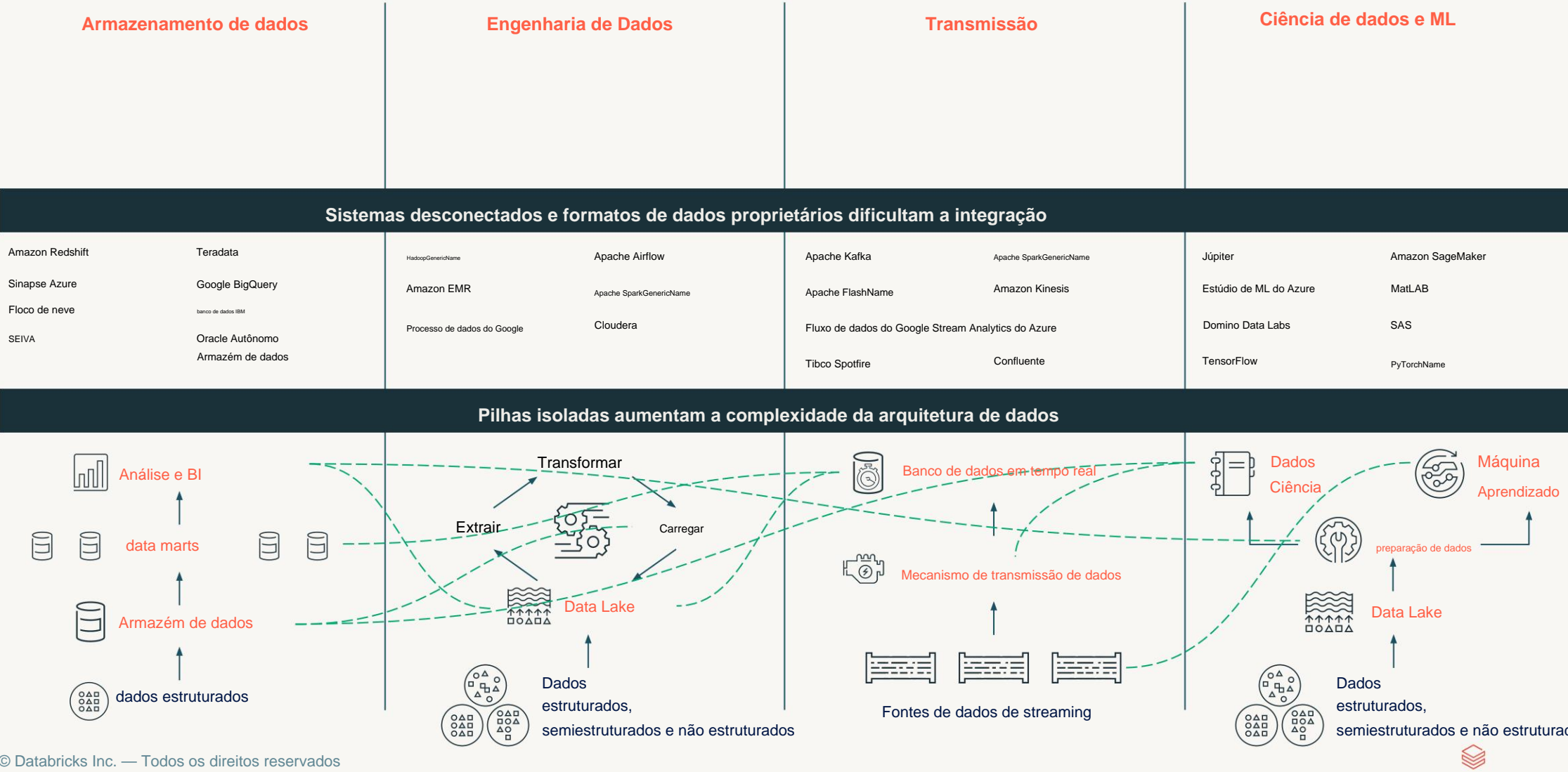
Nativo digital



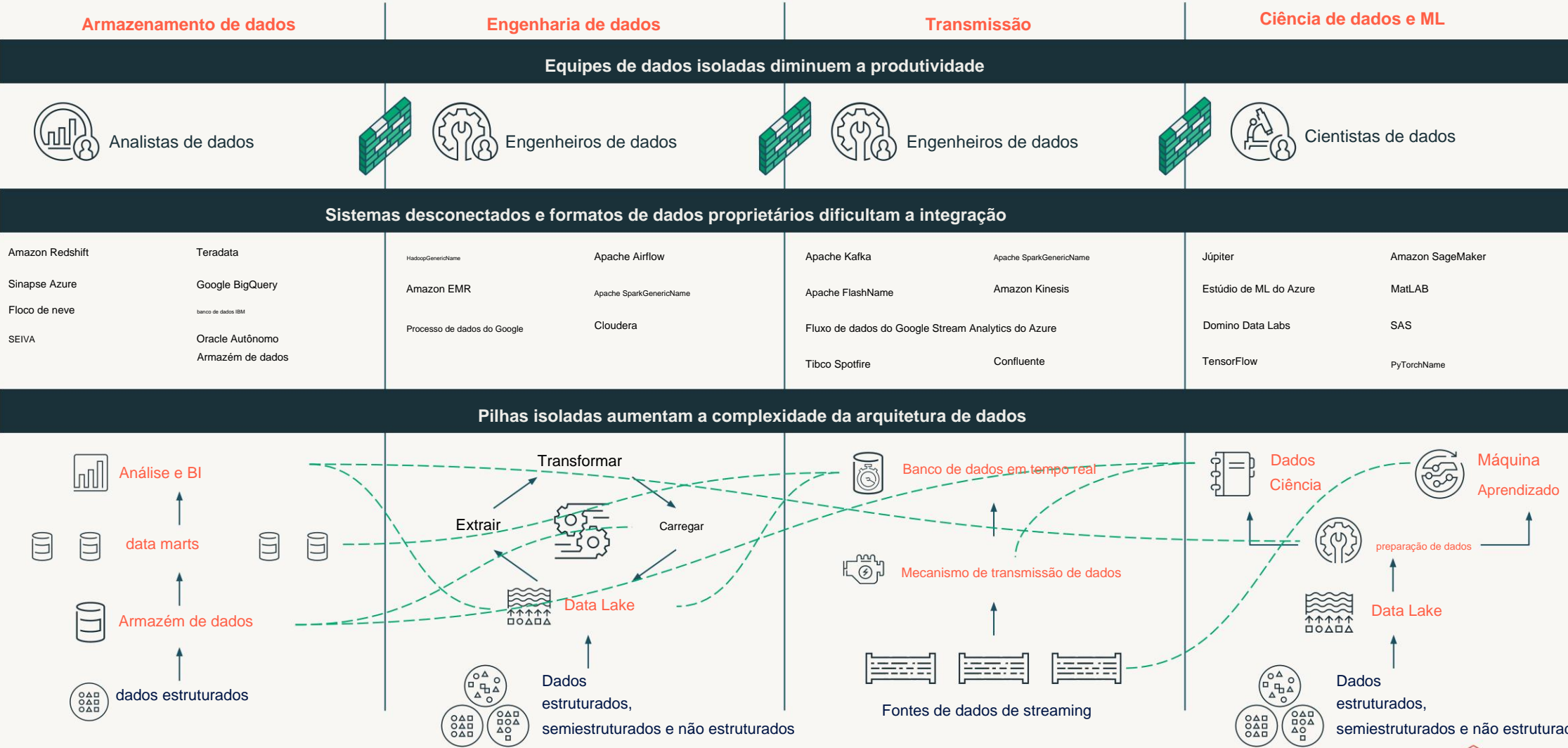
A maioria das empresas luta com dados



A maioria das empresas luta com dados



A maioria das empresas luta com dados





**Dados
Lago**

casa do lago

Uma plataforma para unificar todos
os seus dados, análises e cargas de
trabalho de IA



**Dados
Armazém**





Dados Lago



DELTA LAKE

Uma abordagem aberta para trazer
gerenciamento e
governança de dados para data lakes

Maior confiabilidade nas transações

Processamento de dados 48x mais rápido com indexação




Governança de dados em escala com
listas de controle de acesso refinadas



Dados Armazém



A Plataforma Databricks Lakehouse

-  Simples
-  Abrir
-  Colaborativo



A Plataforma Databricks Lakehouse



Simples

Unifique seus dados, análises e IA em uma plataforma comum para todos os usos de dados casos

Plataforma Databricks Lakehouse

Dados
Engenharia

BI e SQL
Análise

Ciência de
dados e ML

Dados em tempo real
Formulários

Gestão e Governança de Dados

Lago de Dados Aberto

Segurança e administração da plataforma



Dados não estruturados, semiestruturados, estruturados e de streaming

 Microsoft Azure

 aws

 Google Cloud



A Plataforma Databricks Lakehouse



Unifique seu ecossistema de dados com padrões e formatos de código aberto.

Construído com base na inovação de alguns dos projetos de dados de código aberto mais bem-sucedidos do mundo

30 milhões+
downloads mensais



A Plataforma Databricks Lakehouse



Unifique seu ecossistema de dados com padrões e formatos de código aberto.

450+

Parceiros em todo o cenário de dados

ETL visual e ingestão de dados



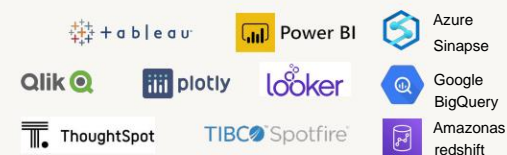
Provedores de dados



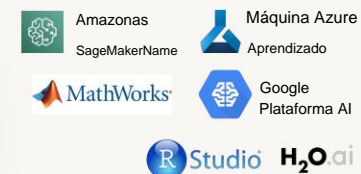
Principais parceiros de consultoria e SI



Inteligência de Negócios



Aprendizado de máquina



Governança centralizada



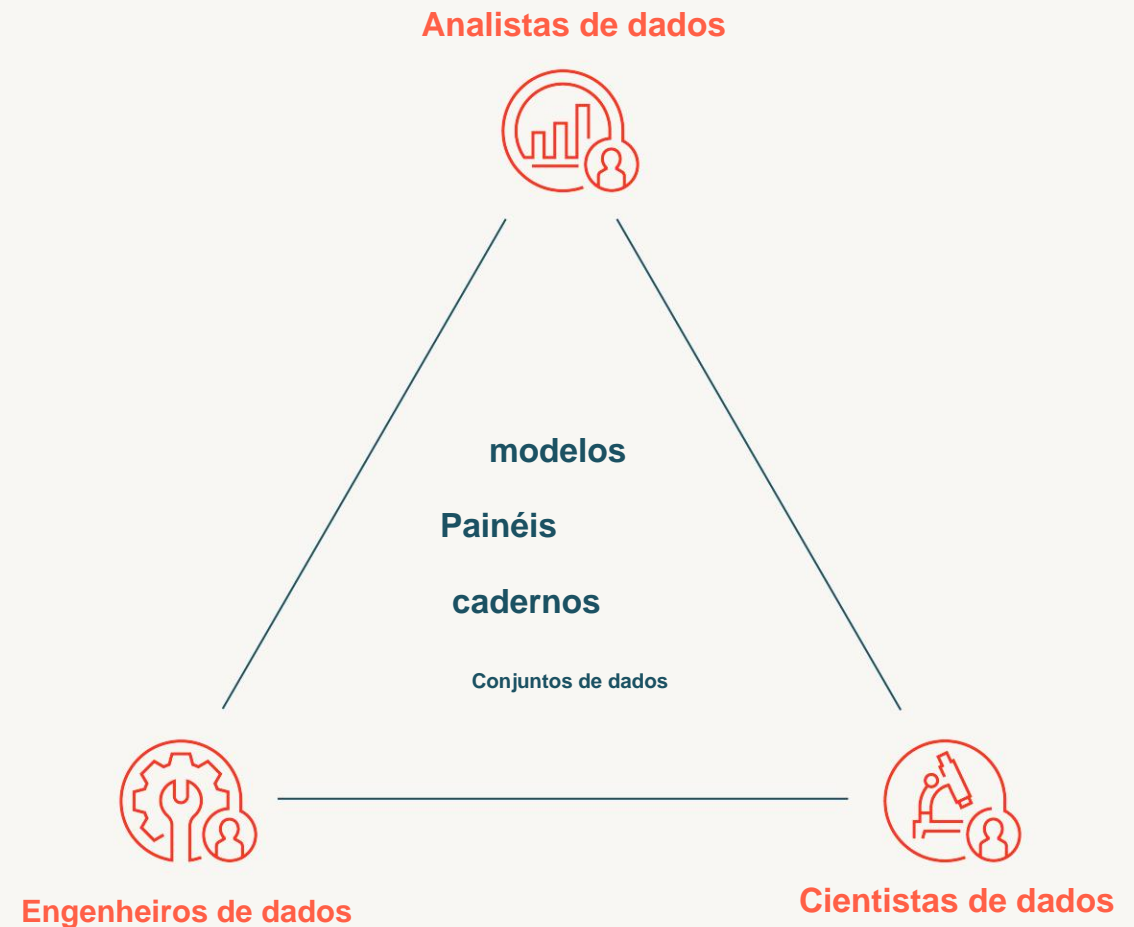
databricks
Plataforma Lakehouse



A Plataforma Databricks Lakehouse



Unifique suas equipes de dados para colaborar em todo o fluxo de trabalho de dados e IA





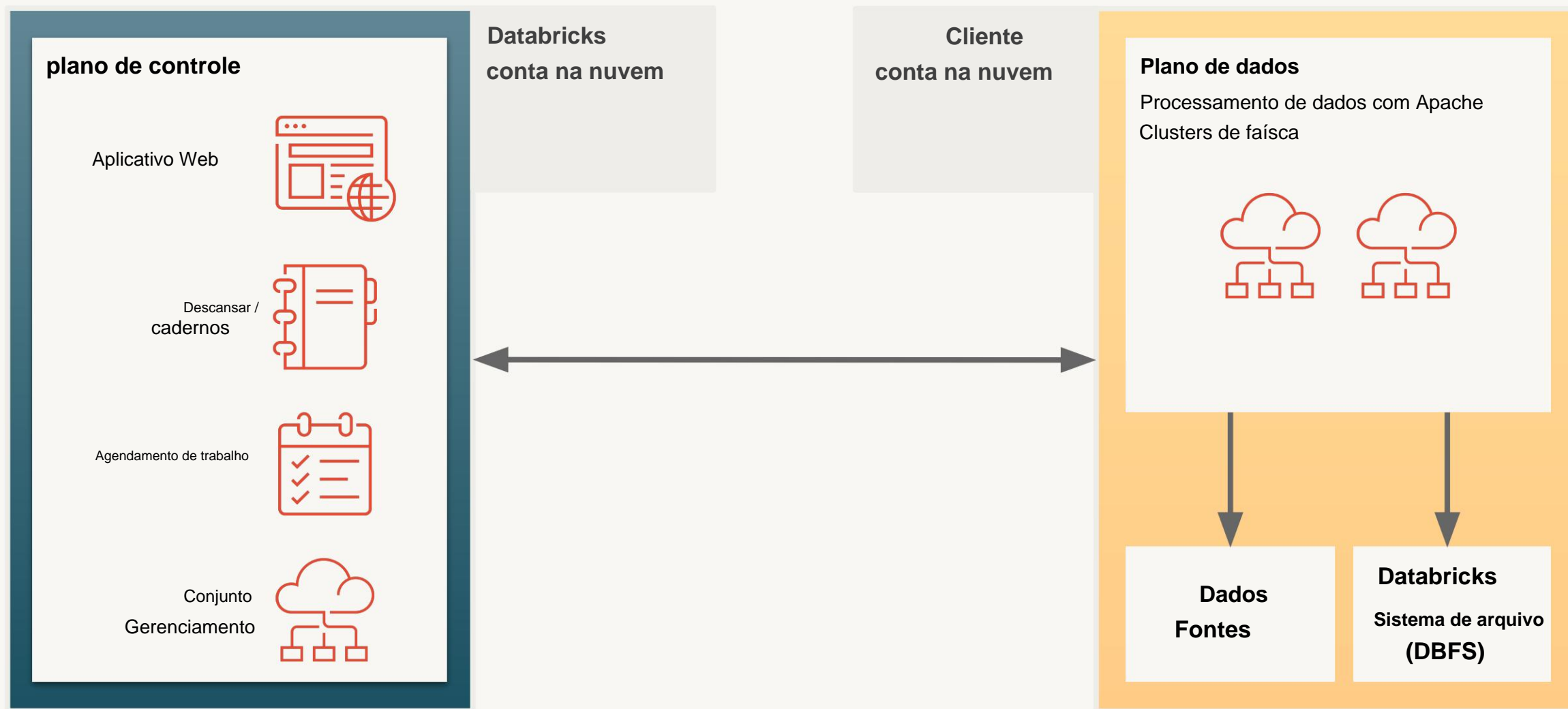
Databricks

Arquitetura e

Serviços



Arquitetura de databricks



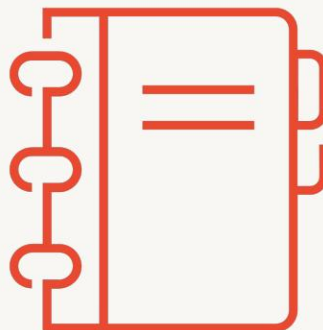
Serviços de databricks

Plano de controle em Databricks

Gerenciar contas de clientes, conjuntos de dados e clusters



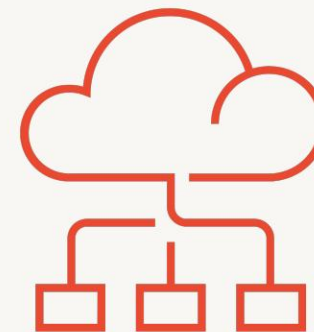
Databricks Web
Aplicativo



Descansar /
cadernos



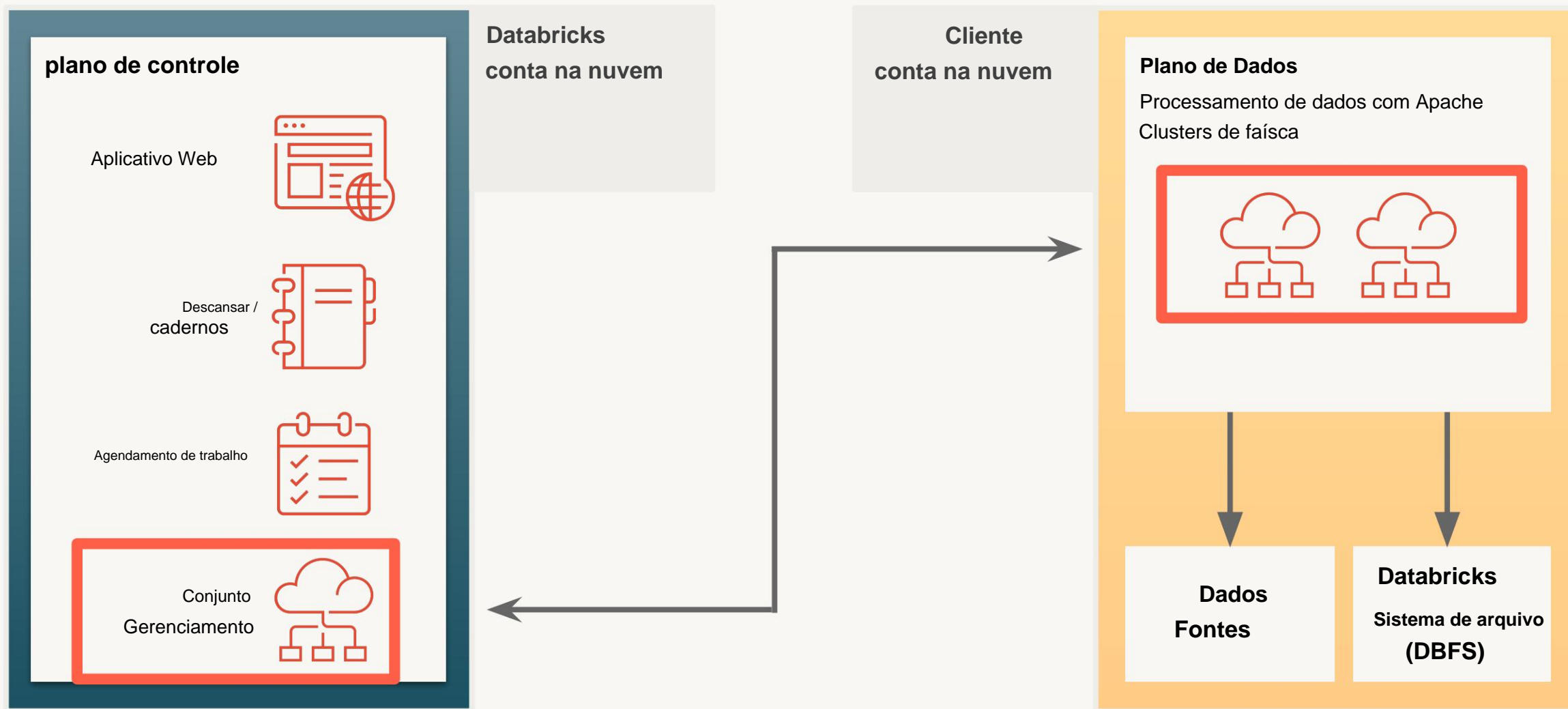
Empregos



Conjunto
Gerenciamento



Clusters



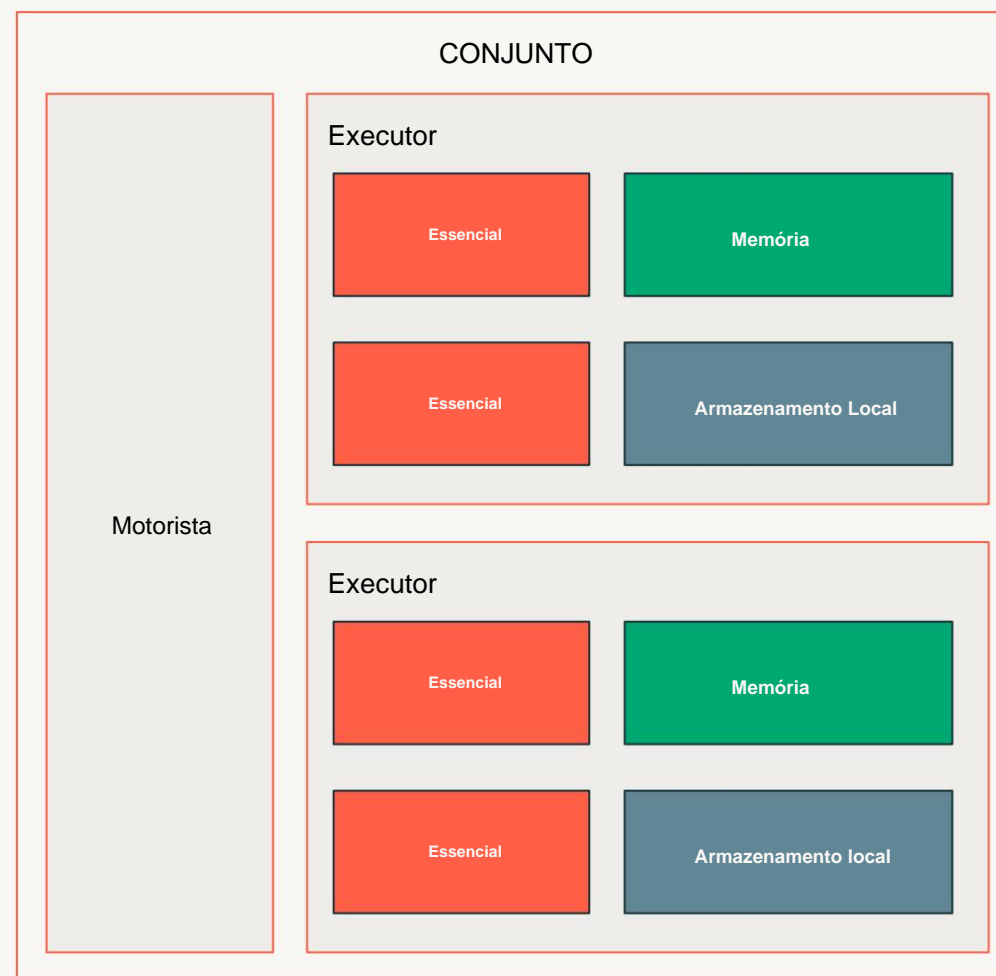
Clusters

Visão geral

Os clusters são compostos por uma ou mais instâncias de máquina virtual (VM)

O motorista coordena as atividades de executores

Executores executam tarefas que compõem um trabalho do Spark



Clusters

tipos

Clusters multifuncionais

Analise dados de forma colaborativa
usando notebooks interativos

Crie clusters a partir do espaço de trabalho
ou da API

Retém até clusters por até dias.

Clusters de trabalho

Executar trabalhos automatizados

O agendador de trabalhos Databricks cria
clusters de trabalho ao executar trabalhos.

Retém até clusters.





Versão Git com Databricks

Descansar



Databricks repositórios

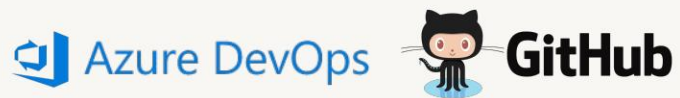
Visão geral

Versão do Git

Integração nativa com
Github, Gitlab, Bitbucket e Azure

Devops

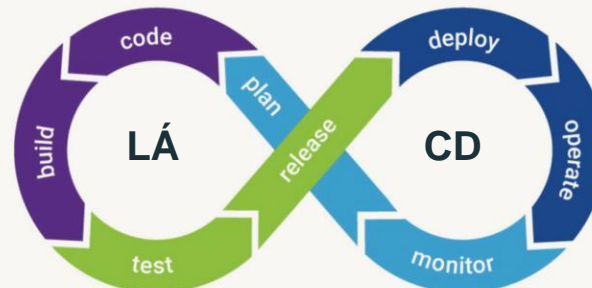
Fluxos de trabalho baseados em IU



Integração CI/CD

Superfície de API para integração
com automação

Simplifica a história
de vários espaços de
trabalho dev/staging/prod



Pronto para empresas

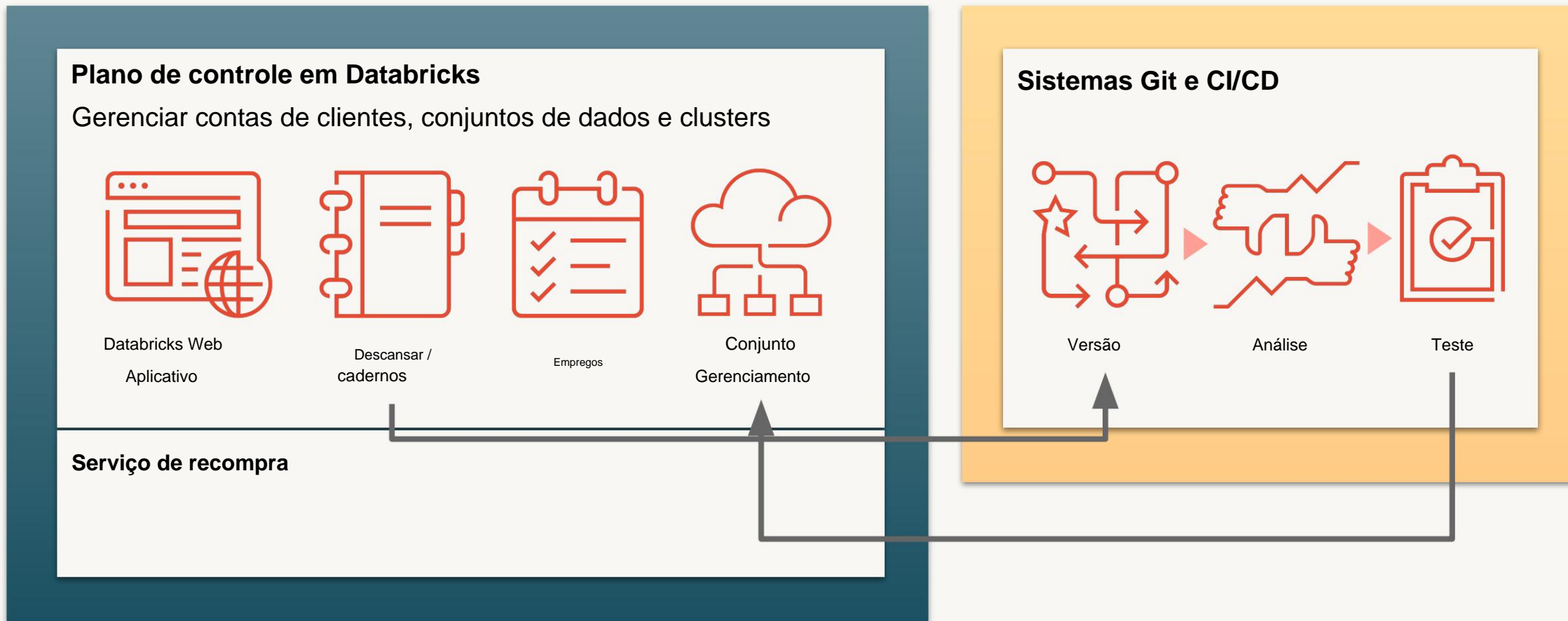
Permitir listas para evitar
exfiltração

Detecção secreta para evitar
vazamento de chaves



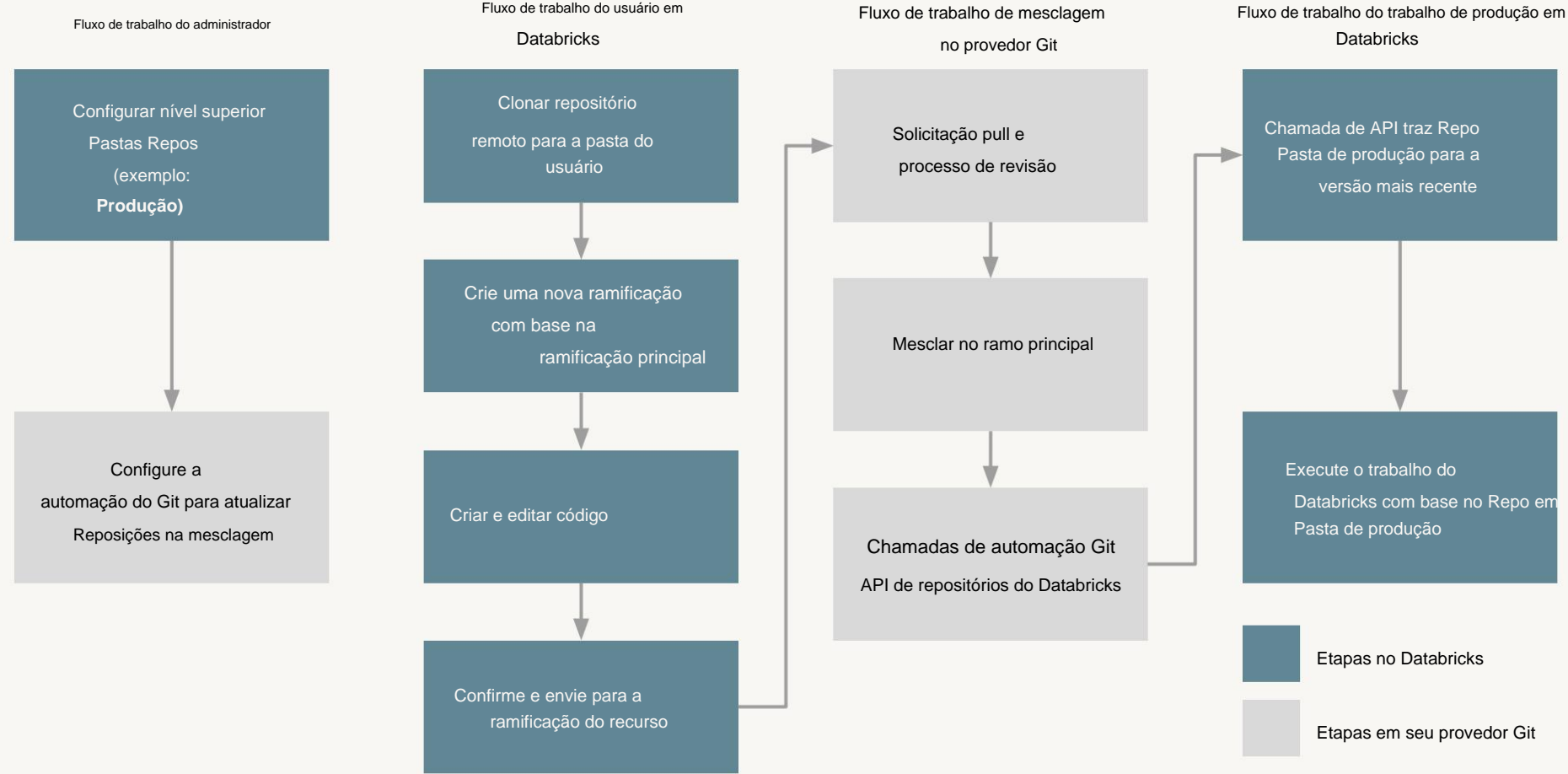
Databricks repositórios

Integração CI/CD



Databricks repositórios

Práticas recomendadas para fluxos de trabalho de CI/CD





O que é o
Lago Delta?



Delta Lake é um projeto de código aberto que permite construir um data lakehouse sobre os sistemas de armazenamento existentes

