Delta Lake não é…

- Tecnologia proprietária
- Formato de armazenamento
- Meio de armazenamento
- Serviço de banco de dados ou data warehouse

O Lago Delta é…

- Código aberto
- Baseia-se em formatos de dados padrão •

Otimizado para armazenamento de objetos em

nuvem • Criado para manipulação de metadados escalável

Delta Lake traz ACID para armazenamento de objetos

ÿ Atomicidade

ÿ Consistência

ÿ Isolamento

ÿ Durabilidade

Problemas resolvidos pelo ACID

- Difícil de anexar dados
- Modificação de dados existentes difícil
- Trabalhos falhando no meio do caminho
- Operações em tempo real difícil
- Caro para manter versões de dados históricos

Delta Lake é o padrão para todas as tabelas criadas no Databricks

databricks

ETL com Spark SQL e Pitão

ETL com Spark SQL e Python

objetivos de aprendizado

- Aproveite o Spark SQL DDL para criar e manipular entidades relacionais em Databricks
- Use o Spark SQL para extrair, transformar e carregar dados para dar suporte a cargas de trabalho de produção e análises no Lakehouse
- Aproveite o Python para a funcionalidade de código avançada necessária na produção formulários

ETL com Spark SQL e Python

Agenda

- Trabalhando com Entidades Relacionais em Databricks
  - Gerenciamento de bancos de dados, tabelas e exibições
- ETL com Spark SQL
  - Extrair dados de fontes externas, carregar e atualizar dados na casa do lago, e transformações comuns
- Python suficiente para Spark SQL
  - Construindo funções extensíveis com SQL encapsulado em Python

databricks

incremental Dados e Delta Mesas ao vivo

Dados Incrementais e Tabelas Dinâmicas Delta

objetivos de aprendizado

- Processe dados de forma incremental para potencializar insights analíticos com o Spark Streaming Estruturado e Carregador Automático
- Propagar novos dados por meio de várias tabelas no data lakehouse
- Aproveitar Delta Live Tables para simplificar a produção de dados SQL pipelines com Databricks

Dados Incrementais e Tabelas Dinâmicas Delta

Agenda

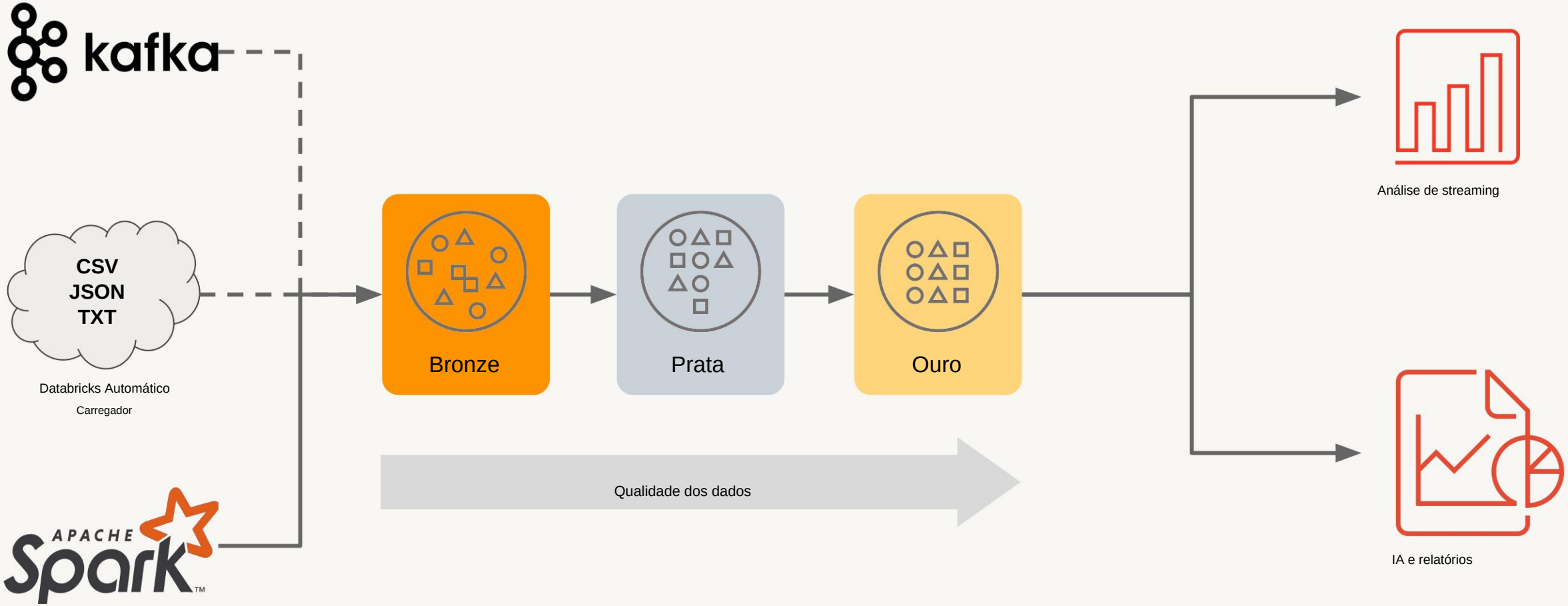
- Processamento de dados incremental com streaming estruturado e automático

  Carregador

  - Processamento e agregação de dados de forma incremental quase em tempo real
- Multi-hop no Lakehouse
  - Propagação de alterações por meio de uma série de tabelas para conduzir sistemas de produção
- Usando Tabelas Dinâmicas Delta
  - Simplificação da implantação de pipelines de produção e infraestrutura usando SQL

Multi-salto
Arquitetura

Multi-Hop no Lakehouse



Multi-Hop no Lakehouse

Camada de bronze

Normalmente, apenas uma cópia bruta dos dados ingeridos

Substitui o data lake tradicional

Fornece armazenamento e consulta eficientes de histórico completo e não processado de dados

Multi-Hop no Lakehouse

Camada de Prata

Reduz a complexidade, a latência e a redundância do armazenamento de dados

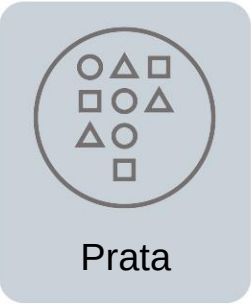
Otimiza a taxa de transferência de ETL e o desempenho de consulta analítica

Preserva a granulação dos dados originais (sem agregações)

Elimina registros duplicados

Esquema de produção aplicado

Verificações de qualidade de dados, dados corrompidos em quarentena



Multi-Hop no Lakehouse

camada de ouro

Capacita aplicativos de ML, relatórios, painéis e análises ad hoc

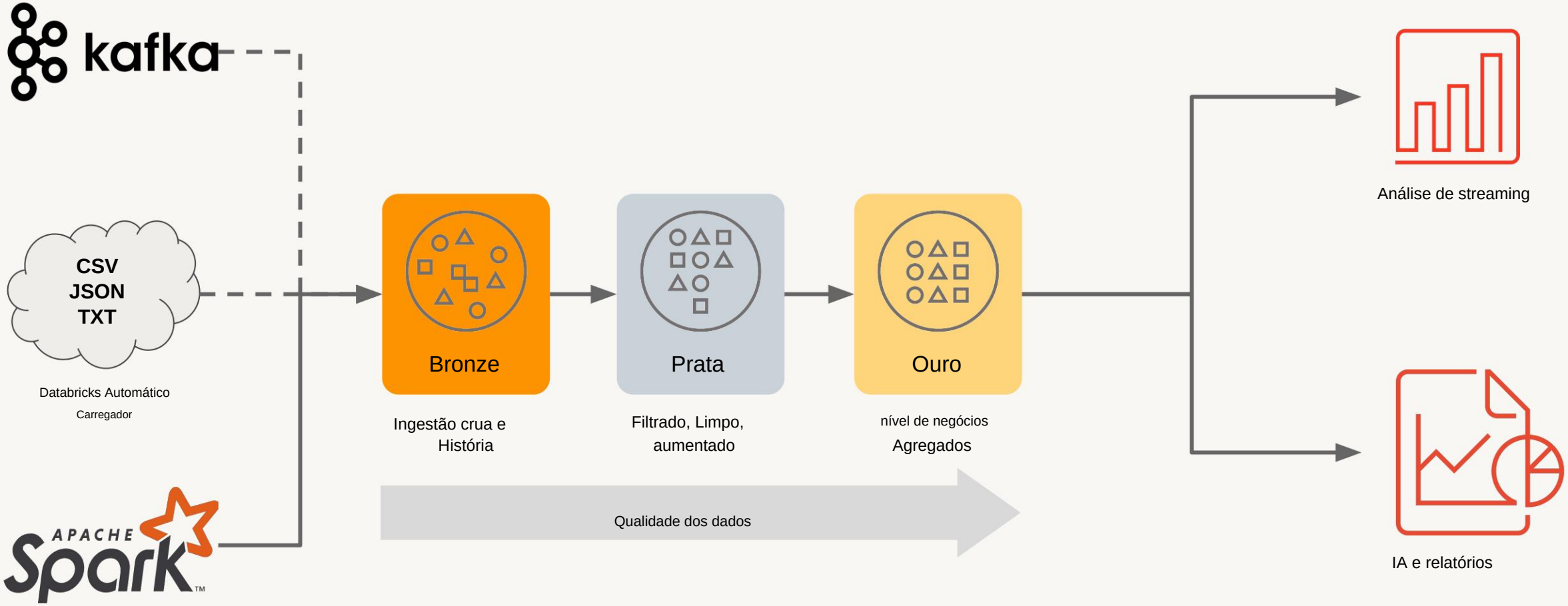
Exibições refinadas de dados, normalmente com agregações

Reduz a tensão nos sistemas de produção

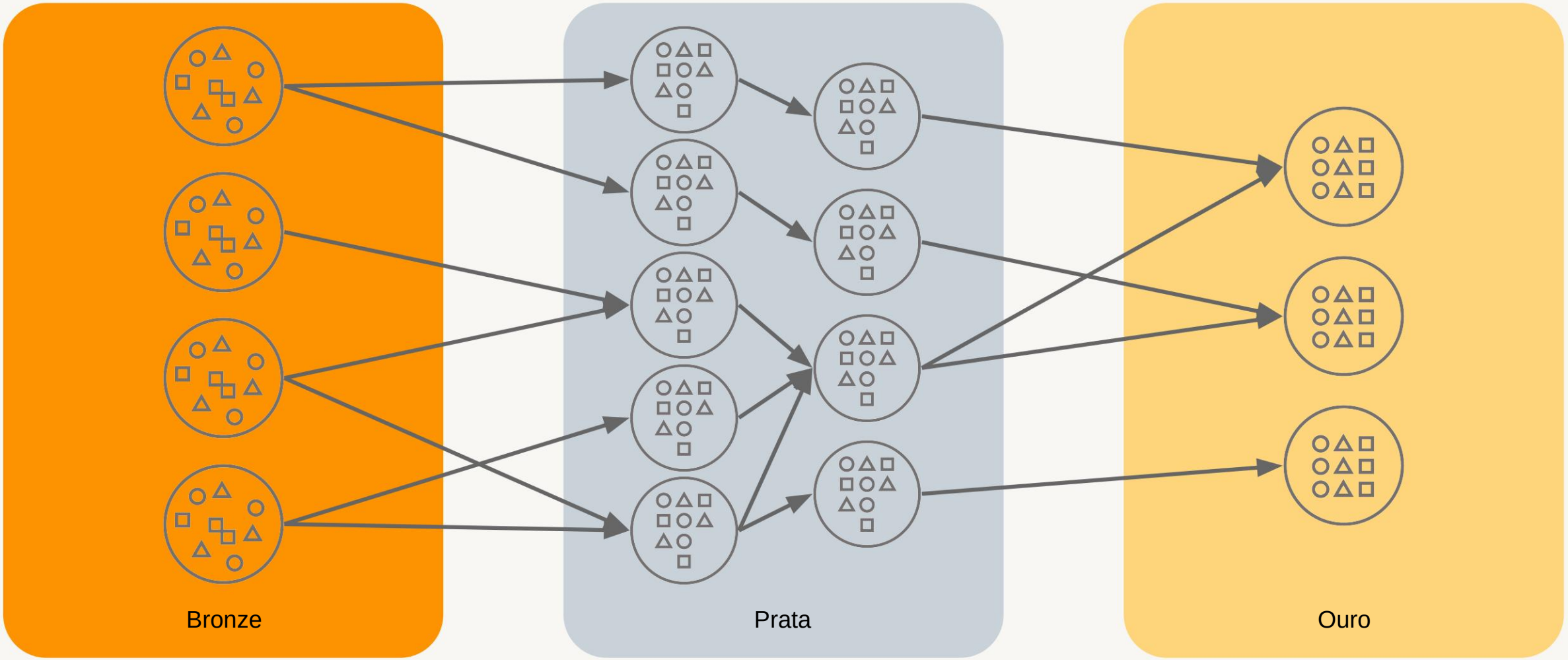
Otimiza o desempenho da consulta para dados críticos para os negócios

Apresentando Delta ao vivo Tabelas

Multi-Hop no Lakehouse



A realidade não é tão simples



ETL de grande escala é complexo e frágil

Desenvolvimento de pipeline complexo

Mesa difícil de construir e manter **dependências**

Difícil alternar entre **os lotes** e processamento **de fluxo**

Qualidade de dados e governança

Difícil de monitorar e fazer cumprir **qualidade de dados**

Impossível rastrear **a linhagem** de dados

Operações de pipeline difíceis

Baixa **observabilidade** em nível granular de dados

O tratamento e **recuperação** de erros é trabalhoso

Apresentando Delta Live Tables

Facilite o ETL confiável no Delta Lake

Opere com agilidade

Ferramentas declarativas para criar pipelines de dados em lote e streaming

Confie em seus dados

O DLT possui controles de qualidade declarativos integrados

Declarar expectativas de qualidade e ações a serem tomadas

Escale com confiabilidade

Dimensione facilmente a infraestrutura junto com seus dados

databricks

Gerenciamento de dados

Acessar e Produção Oleodutos

Gerenciamento de acesso a dados e produção Oleodutos

objetivos de aprendizado

- Orquestrar tarefas com Databricks Jobs
- Use Databricks SQL para consultas sob demanda
- Configurar listas de controle de acesso Databricks para fornecer grupos com segurança acesso a bancos de dados de produção e desenvolvimento
- Configurar e agendar painéis e alertas para refletir atualizações para pipelines de dados de produção

Gerenciamento de acesso a dados e produção Oleodutos

Agenda

- Orquestração de Tarefas com Databricks Jobs
  - Agendamento de notebooks e pipelines DLT com dependências
- Executando sua primeira consulta SQL do Databricks
  - Navegando, configurando e executando consultas no Databricks SQL
- Gerenciamento de permissões no Lakehouse
  - Configurar permissões para bancos de dados, tabelas e exibições no data lakehouse
- Produção de Dashboards e Queries em DBSQL
  - Agendamento de consultas, painéis e alertas para pipelines analíticos de ponta a ponta

Apresentando Catálogo Unity

Visão geral da governança de dados

Quatro áreas funcionais principais

Controle de acesso a dados

Controle quem tem acesso a quais dados

Auditoria de acesso a dados

Capture e registre todo o acesso aos dados

Linhagem de dados

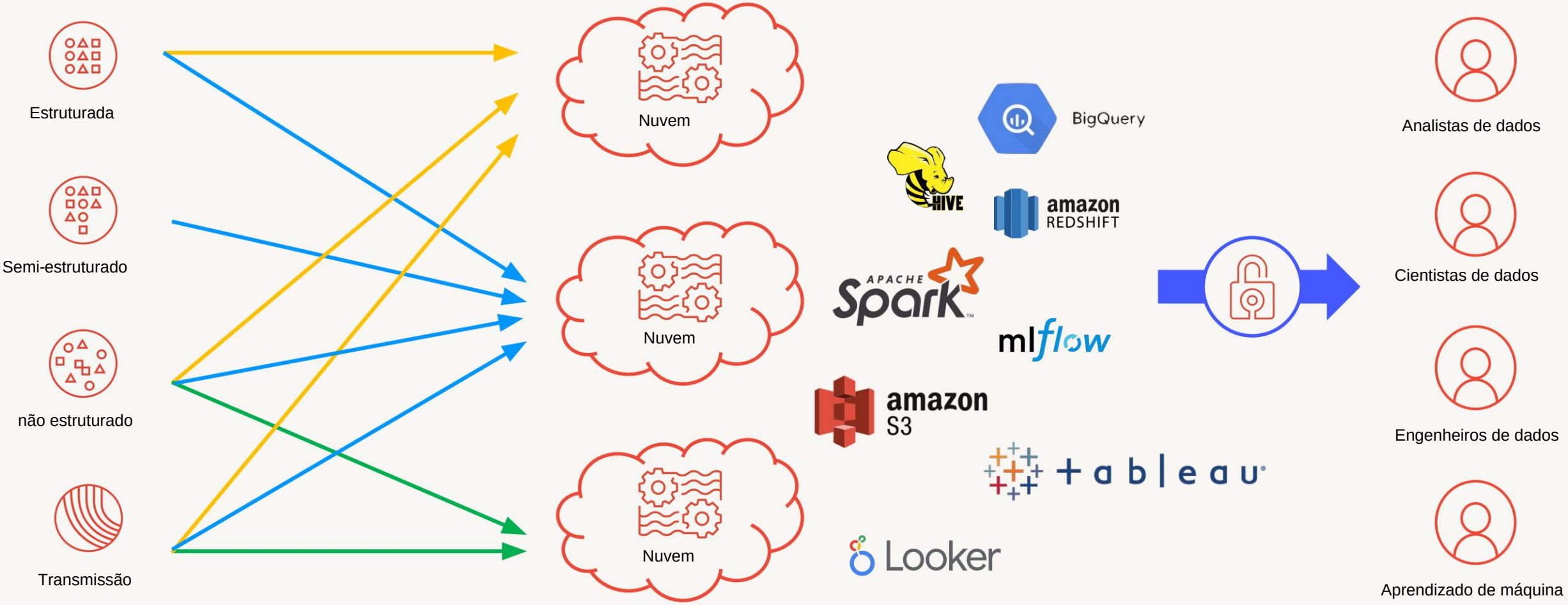
Capture fontes upstream e downstream consumidores

Descoberta de dados

Capacidade de pesquisar e descobrir ativos autorizados

Visão geral da governança de dados

desafios



Catálogo Databricks Unity

Visão geral

Unifique a governança nas nuvens

Governança refinada para data lakes entre nuvens - com base em

padrão aberto ANSI SQL.

Unifique dados e ativos de IA

Compartilhe, audite, proteja e gerencie centralmente todos os tipos de dados com uma interface simples.

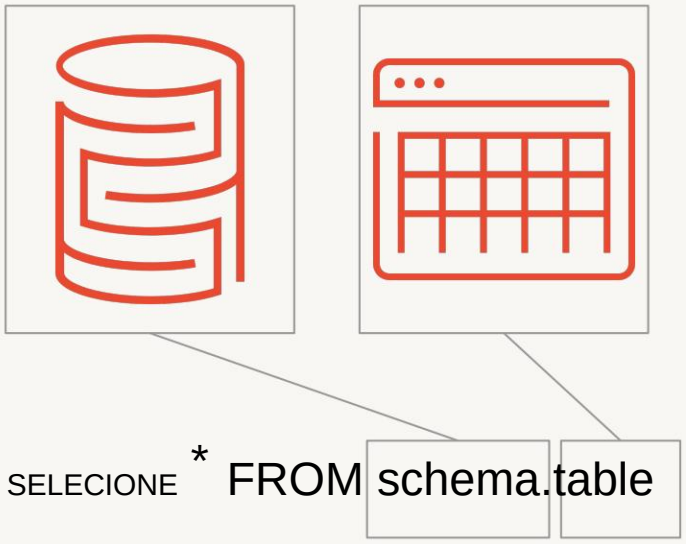
Unifique os catálogos existentes

Funciona em conjunto com dados, armazenamento e catálogos existentes - nenhuma migração difícil é necessária.

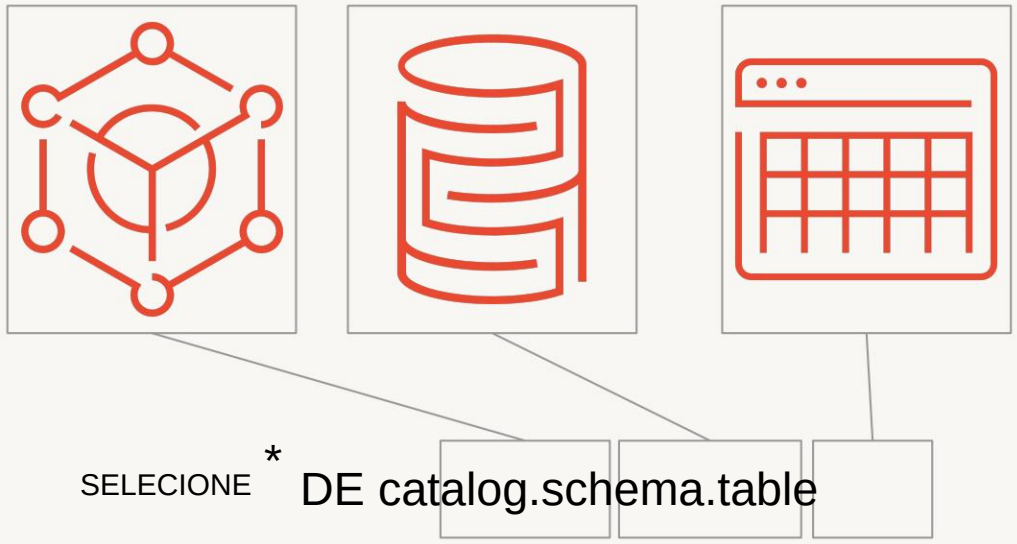
Catálogo Databricks Unity

Namespace de três camadas

Namespace tradicional de duas camadas



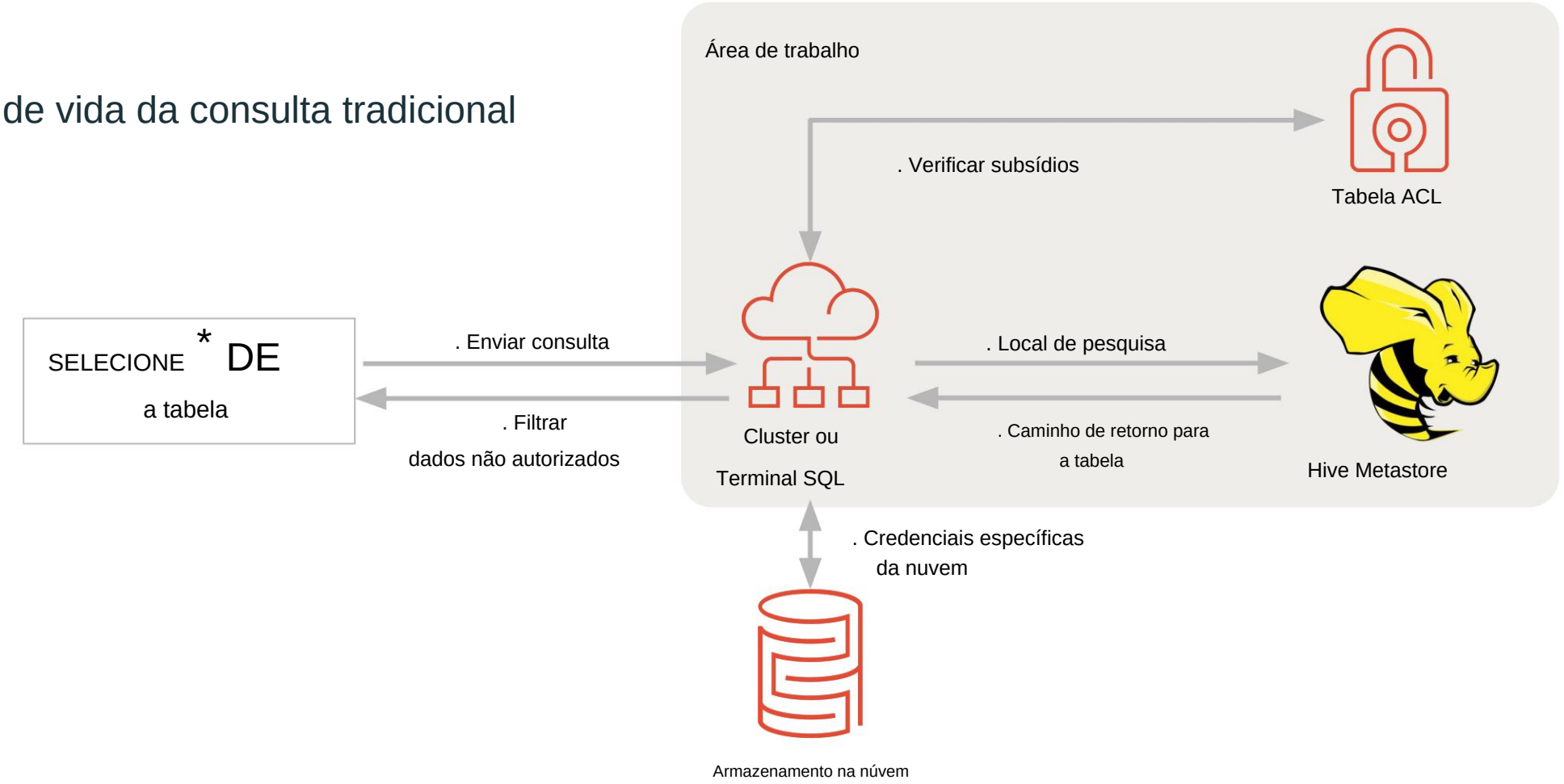
Namespace de três camadas com Unity Catálogo



Catálogo Databricks Unity

Modelo de segurança

Ciclo de vida da consulta tradicional

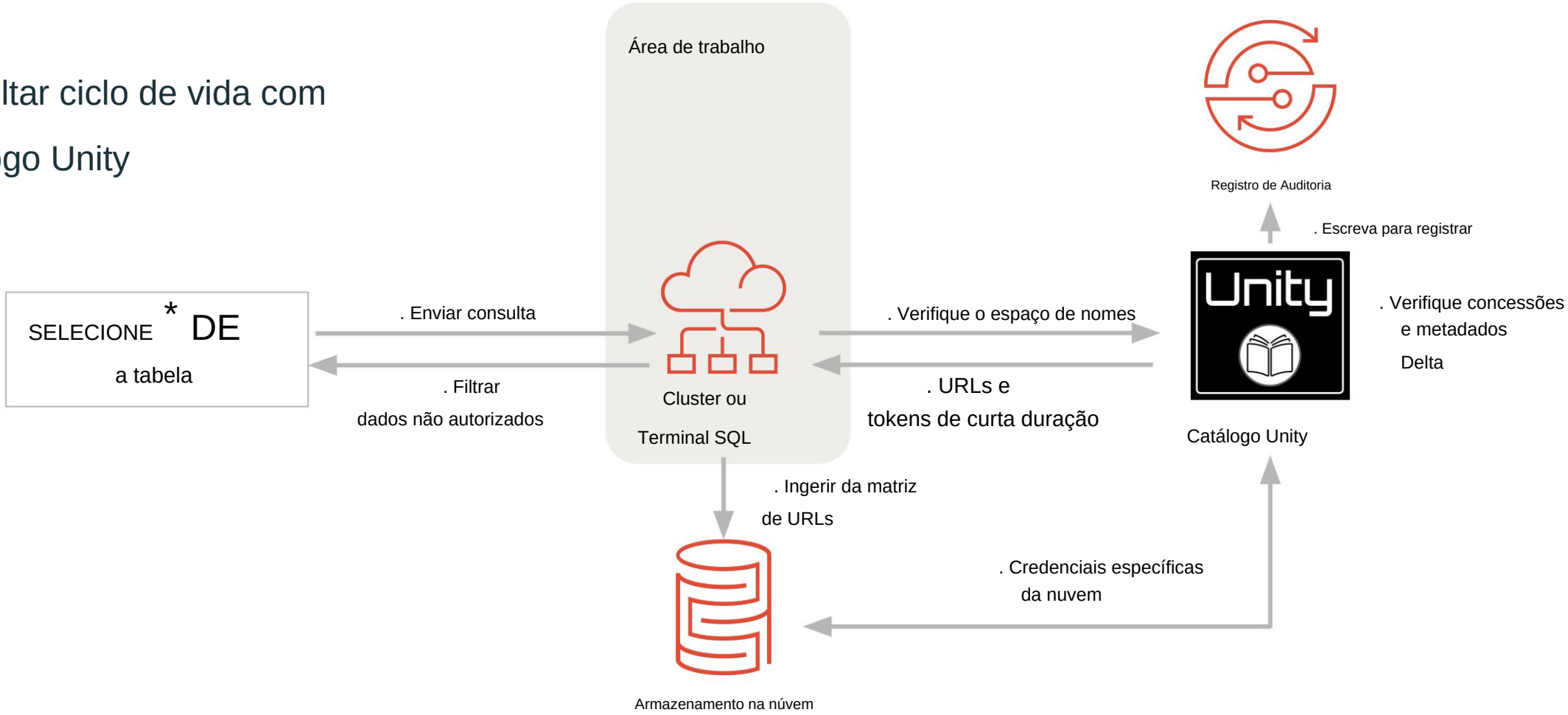


Catálogo Databricks Unity

Modelo de segurança

Consultar ciclo de vida com

Catálogo Unity



Recapitulação do Curso

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