

Introdução ao Databricks para data engineering



Objetivos de aprendizado do curso

Tudo o que será capaz de fazer após a conclusão deste curso

- Explicar conceitos fundamentais sobre o uso da Plataforma Databricks Lakehouse para novos usuários responsáveis por fluxos de trabalho de data engineering.
- Descrever como trabalhar na Plataforma Databricks Lakehouse
- Executar tarefas básicas de notebook usando a Plataforma Databricks Lakehouse.
- Gerenciar tabelas Delta usando a Plataforma Databricks Lakehouse.
- Descrever os recursos disponíveis na Plataforma Databricks Lakehouse que protegem e governam os dados.
- Usar os jobs de fluxo de trabalho da Plataforma Databricks Lakehouse para automatizar um fluxo de trabalho básico de data engineering. Automatizar um fluxo de trabalho básico de data engineering usando jobs de fluxo de trabalho



Pré-requisitos

Tudo o que precisa saber ou conseguir fazer antes de participar deste curso

- Conhecimento básico de tópicos de data engineering, como extração, limpeza (e outras transformações) e carregamento



Requisitos técnicos

Considerações importantes antes de tentar

- Este curso foi testado no DBR 13.2
- Nem todos os notebooks operarão no Community Edition
- Você precisará de uma conta no Github (ou pode simplesmente assistir à demonstração)



Módulo 1



Módulo 1 – Programação

Nome da lição	Nome da lição
Aula: Os fundamentos do Databricks	Aula: Introdução aos recursos de computação
Demonstração: A Plataforma Lakehouse	Demonstração: Como trabalhar com recursos de computação
Aula: Introdução aos Repos no Databricks	Aula: Notebooks do Databricks
Demonstração: Como trabalhar com o Repos no Databricks	Demonstração: Como trabalhar com notebooks



Os fundamentos do Databricks



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Identificar o Databricks como a Plataforma Lakehouse
- Conectar personas de dados comuns aos serviços principais da Plataforma Databricks Lakehouse.





Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

Michael Armbrust¹, Ali Ghodsi^{1,2}, Reynold Xin¹, Matei Zaharia^{1,3}

¹Databricks, ²UC Berkeley, ³Stanford University

Abstract

This paper argues that the data warehouse architecture as we know it today will wither in the coming years and be replaced by a new architectural pattern, the Lakehouse, which will (i) be based on open direct-access data formats, such as Apache Parquet, (ii) have first-class support for machine learning and data science, and (iii) offer state-of-the-art performance. Lakehouses can help address several major challenges with data warehouses, including data staleness, reliability, total cost of ownership, data lock-in, and limited use-case support. We discuss how the industry is already moving toward Lakehouses and how this shift may affect work in data management. We also report results from a Lakehouse system using Parquet that is competitive with popular cloud data warehouses on TPC-DS.

1 Introduction

This paper argues that the data warehouse architecture as we know it today will wane in the coming years and be replaced by a new architectural pattern, which we refer to as the Lakehouse, characterized by (i) open direct-access data formats, such as Apache Parquet and ORC, (ii) first-class support for machine learning and data science workloads, and (iii) state-of-the-art performance.

The history of data warehousing started with helping business leaders get analytical insights by collecting data from operational databases into centralized warehouses, which then could be used for decision support and business intelligence (BI). Data in these warehouses would be written with schema-on-write, which ensured that the data model was optimized for downstream BI consumption.

quality and governance downstream. In this architecture, a small subset of data in the lake would later be ETLed to a downstream data warehouse (such as Teradata) for the most important decision support and BI applications. The use of open formats also made data lake data directly accessible to a wide range of other analytics engines, such as machine learning systems [30, 37, 42].

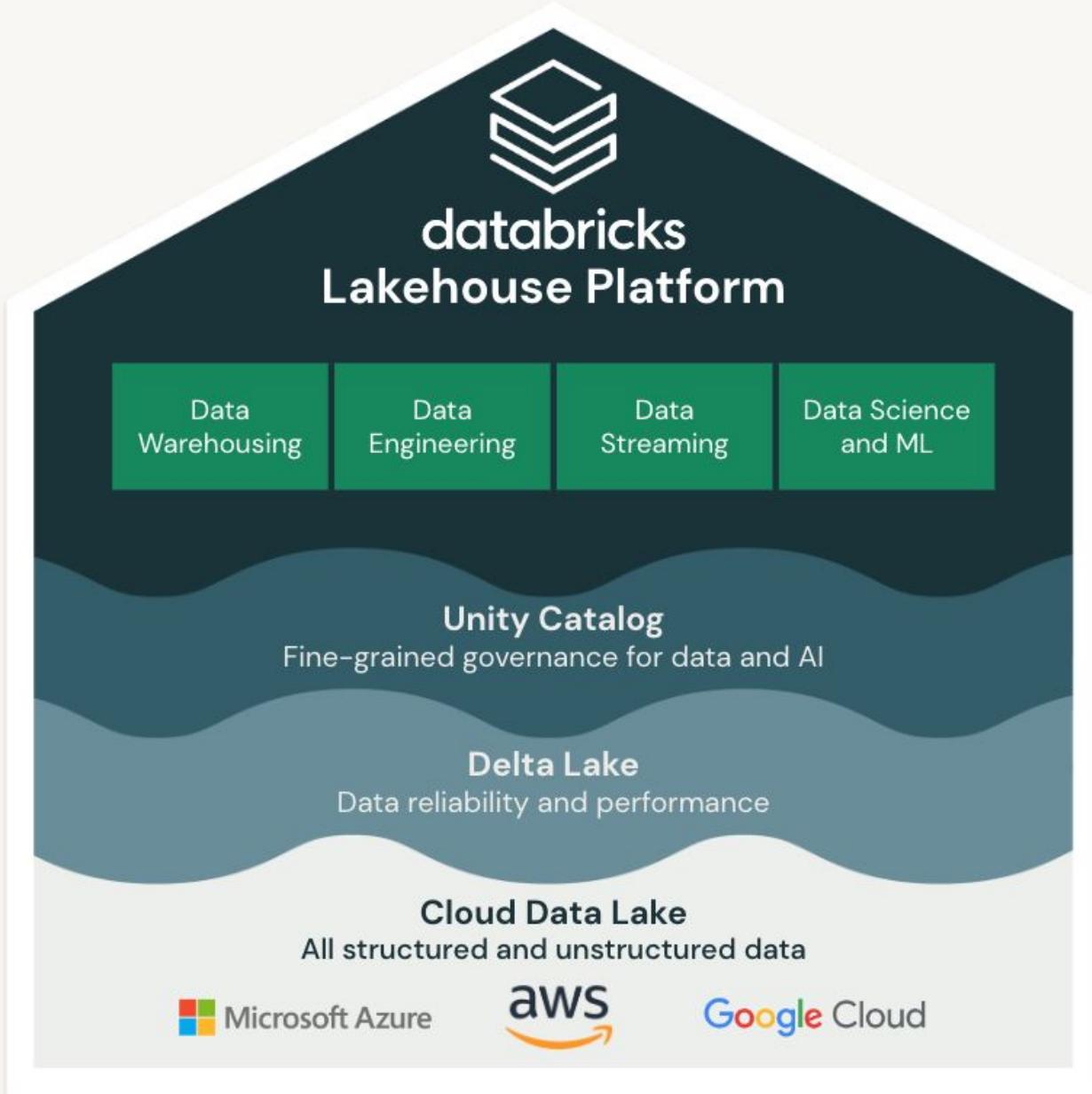
From 2015 onwards, cloud data lakes, such as S3, ADLS and GCS, started replacing HDFS. They have superior durability (often >10 nines), geo-replication, and most importantly, extremely low cost with the possibility of automatic, even cheaper, archival storage, e.g., AWS Glacier. The rest of the architecture is largely the same in the cloud as in the second generation systems, with a downstream data warehouse such as Redshift or Snowflake. This two-tier data lake + warehouse architecture is now dominant in the industry in our experience (used at virtually all Fortune 500 enterprises).

This brings us to the challenges with current data architectures. While the cloud data lake and warehouse architecture is ostensibly cheap due to separate storage (e.g., S3) and compute (e.g., Redshift), a two-tier architecture is highly complex for users. In the first generation platforms, all data was ETLed from operational data systems directly into a warehouse. In today's architectures, data is first ETLed into lakes, and then again ELTed into warehouses, creating complexity, delays, and new failure modes. Moreover, enterprise use cases now include advanced analytics such as machine learning, for which *neither* data lakes nor warehouses are ideal. Specifically, today's data architectures commonly suffer from four problems:

Reliability. Keeping the data lake and warehouse consistent is difficult and costly. Continuous engineering is required to ETL data between the two systems and make it available to high-performance

Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. M. Armbrust, A. Ghodsi, R. Xin, M. Zaharia. 11th Annual Conference on Innovative Data Systems Research (CIDR '21), 11 a 15 de janeiro de 2021, online.





Plataforma Databricks Lakehouse

Simples

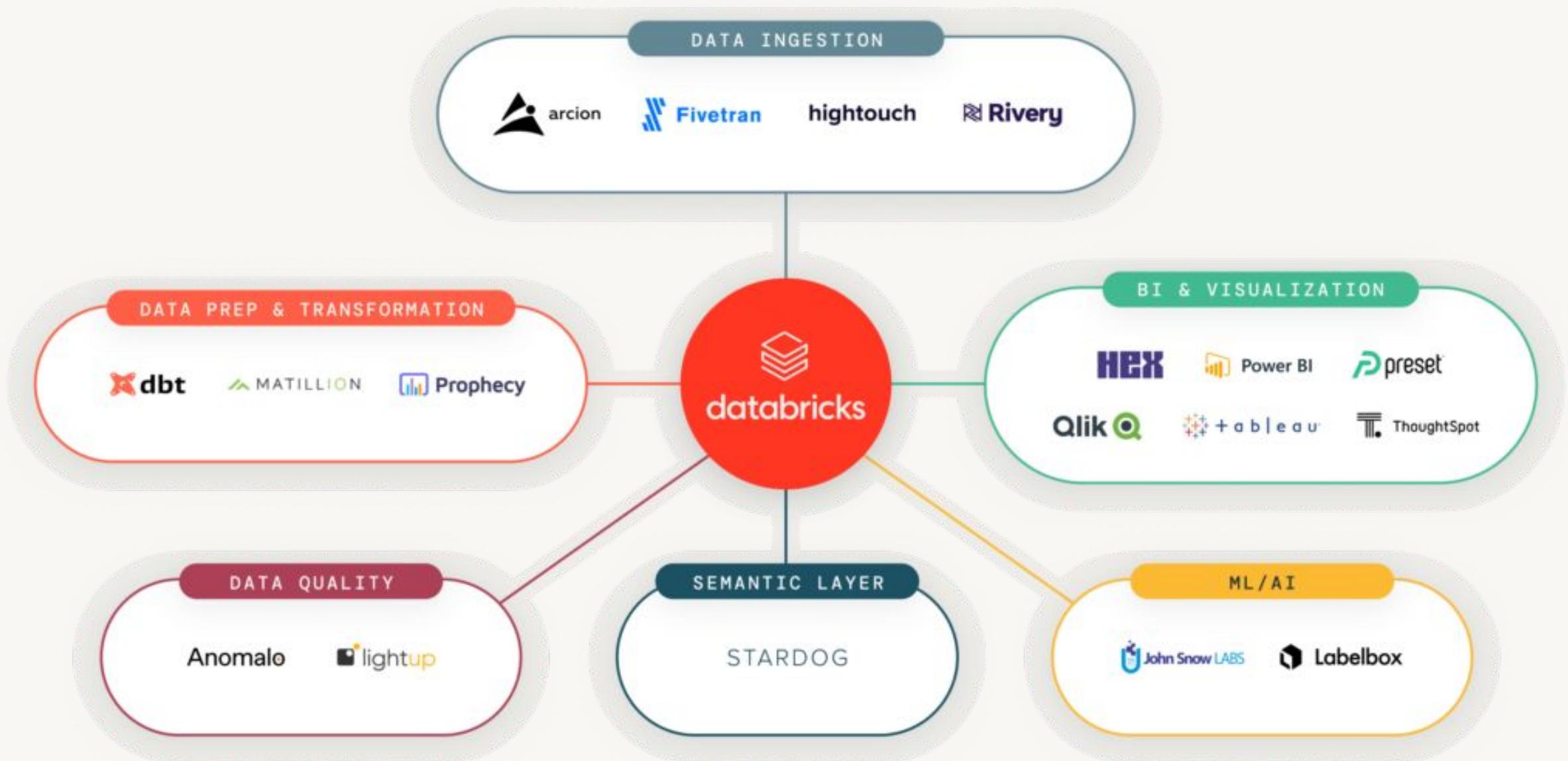
Unifique os casos de uso de data warehousing e IA
casos de uso em uma única plataforma

Aberto

Construída com código e padrões abertos

Multicloud

Uma plataforma de dados consistente
em várias clouds



O lakehouse é para todos os profissionais de dados

Engenheiros de dados

Delta Live Tables
Delta Lake
Unity Catalog

Analistas de dados

Databricks SQL
Visualizations

Profissionais de machine learning

ML Flow
Feature Store



Demonstração: A Plataforma Lakehouse



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Descrever a interface de navegação da Plataforma Databricks Lakehouse, incluindo a organização de serviços e recursos na barra de navegação do lado esquerdo e a disponibilidade de configurações no canto superior direito.
- Descrever o Workspace como a solução para organizar ativos no Databricks.



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Identificar que muitos tipos de ativos podem ser acessados e organizados no Workspace, incluindo notebooks e arquivos.
- Navegar pelo Workspace.
- Executar as ações disponíveis no Workspace.



Demonstração

Passos de alto nível

Visão geral da IU

- Página inicial
- Navegação
- Menu no canto superior direito

Recursos do menu do Workspace

- Local do workspace



Introdução aos Repos do Databricks



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Descrever o Repos como um recurso centrado na integração contínua de ativos no Databricks e em repositórios Git externos.

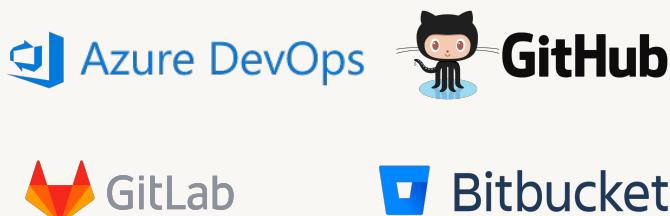


Databricks Repos

Controle de versão do Git

Integração nativa com Github, Gitlab, Bitbucket e Azure Devops

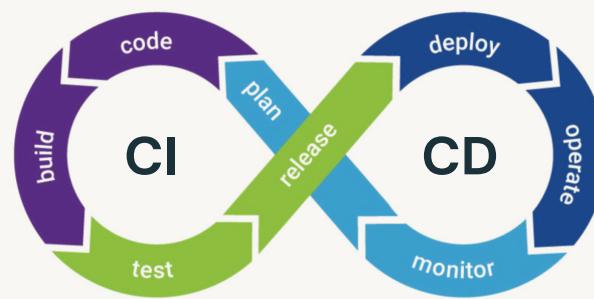
Fluxos de trabalho baseados na IU



Integração CI/CD

Superfície de API para integração com automação

Simplifica o uso de vários workspaces de desenvolvimento/preparação/produção



Pronto para empresas

Listas de permissão para evitar exfiltração

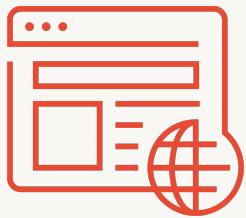
Detecção de segredos para evitar o vazamento de chaves

Databricks Repos

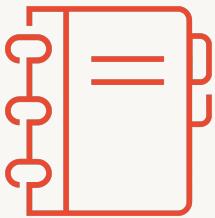
Integração CI/CD

Plano de controle no Databricks

Gerencia contas de clientes, datasets e clusters



Aplicativo Web
do Databricks



Repos/
Notebooks



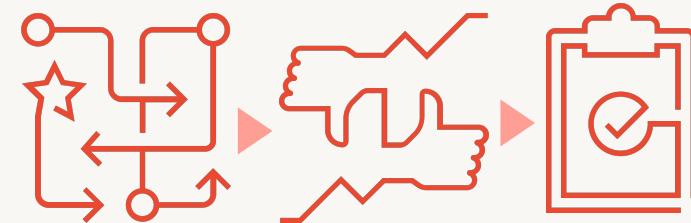
Jobs



Gerenciamento
de clusters

Serviço de Repos

Sistemas CI/CD e Git



Versão

Revisão



Teste

Demonstração: como trabalhar com o Repos no Databricks



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Adicionar um repo a partir de um repositório Git existente.
- Descrever como comparar, extrair e enviar alterações entre o Databricks e um repositório Git.
- Criar um notebook
- Alterar o nome de um notebook



Demonstração

Passos de alto nível

Repos

- Clonagem de repos
- Extração de repos
- Operações de CI
- Comparação e envio



Introdução aos recursos de computação



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Descrever a estrutura básica de computação baseada em nuvem da Databricks.
- Comparar e contrastar clusters e warehouses.
- Descrever as opções de configuração de alto nível em um cluster.
- Descrever as opções de configuração de alto nível em um warehouse.
- Descrever os benefícios de usar os recursos de computação serverless disponíveis.



Clusters

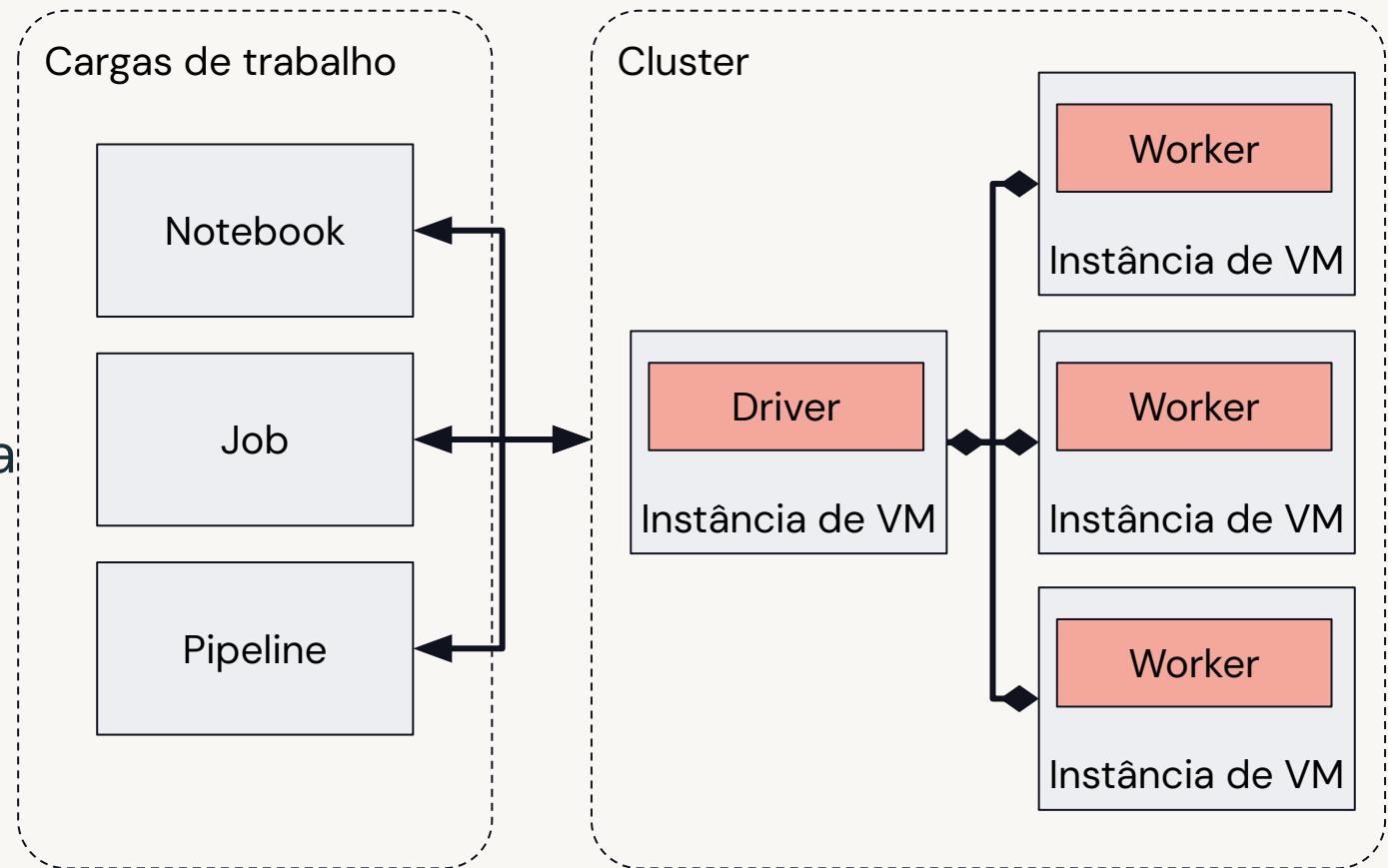
Visão geral

Coleção de instâncias de VM

Distribui cargas de trabalho entre workers

Dois tipos principais:

- 1. Clusters todo-propósito** para desenvolvimento interativo
- 2. Clusters de jobs** para cargas de trabalho automatizadas



Tipos de cluster

Clusters todo-propósito

Analizar dados de forma colaborativa usando notebooks **interativos**

Clusters de jobs

Execução **automatizada** de jobs
O job scheduler do Databricks cria clusters de jobs ao executar jobs



Modo do cluster

Nó único

Cluster de instância única de baixo custo que atende a cargas de trabalho de machine learning de nó único e análise exploratória leve

Padrão (vários nós)

O modo padrão para cargas de trabalho desenvolvidas em qualquer linguagem compatível (requer pelo menos duas instâncias de VM)



Versão do Databricks Runtime

Padrão

Apache Spark e muitos outros componentes e atualizações para fornecer experiências otimizadas de análise big data

Photon

Um complemento opcional que otimiza as queries do Spark (por exemplo, SQL, DataFrame)

Machine learning

Adiciona bibliotecas populares de machine learning como TensorFlow, Keras, PyTorch e XGBoost.



Modo de acesso

Dropdown do modo de acesso	Visível ao usuário	Suporte ao Unity Catalog	Linguagens compatíveis
Usuário único	Sempre	Sim	Python, SQL, Scala, R
Compartilhado	Sempre (plano Premium necessário)	Sim	Python (DBR 11.1+), SQL
Nenhum isolamento compartilhado	Pode ser ocultado impondo o isolamento de usuário no console de administração ou definindo as configurações no nível da conta	Não	Python, SQL, Scala, R
Personalizado	Mostrado apenas para clusters existentes sem modos de acesso (ou seja, modos de cluster legados, padrão ou de alta simultaneidade); não é uma opção na criação de novos clusters.	Não	Python, SQL, Scala, R



Políticas de cluster

As políticas de cluster podem ajudar a alcançar o seguinte:

- padronizar configurações de clusters
- fornecer configurações predefinidas direcionadas a casos de uso específicos
- simplificar a experiência do usuário
- evitar o uso excessivo e controlar os custos
- impor marcação correta

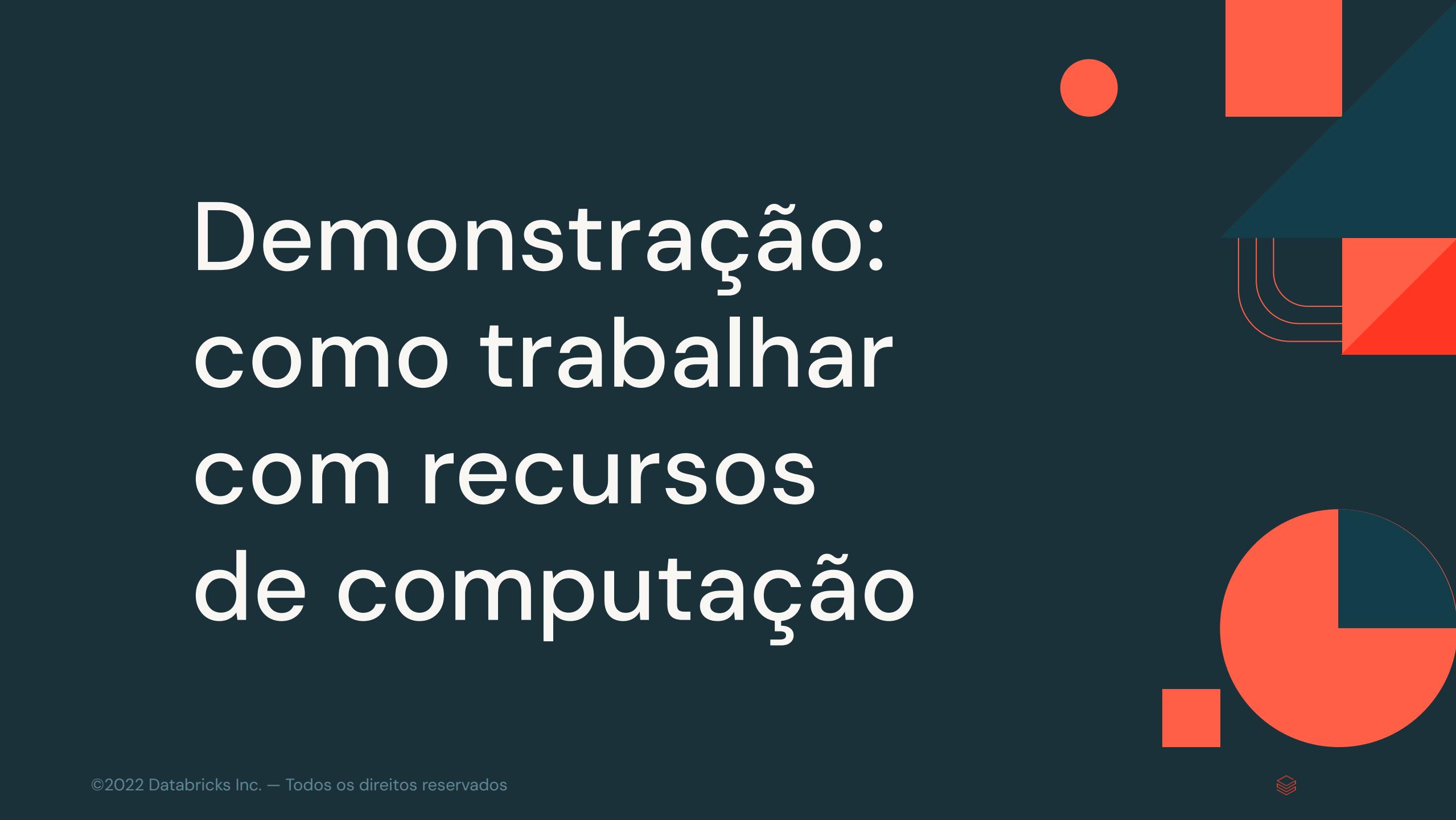


Controle de acesso ao cluster

	Nenhuma permissão	Pode anexar a	Pode reiniciar	Pode gerenciar
Anexar notebook		✓	✓	✓
Exibir a Spark UI, métricas de cluster e logs de driver		✓	✓	✓
Iniciar, reiniciar, encerrar			✓	✓
Editar				✓
Anexar biblioteca				✓
Redimensionar				✓
Alterar permissões				✓



Demonstração: como trabalhar com recursos de computação



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Iniciar um novo cluster.
- Iniciar um novo warehouse.



Demonstração

Passos de alto nível

Clusters

- Configurar e iniciar um cluster

Warehouses

- Configurar e iniciar um warehouse



Notebooks do Databricks



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Descrever o Databricks Notebooks como a interface mais comum para engenheiros de dados ao trabalhar com o Databricks.
- Reconhecer casos de uso comuns para engenheiros de dados ao trabalhar com Notebooks.
- Explorar os recursos básicos de visualização nos Notebooks do Databricks.



Notebooks do Databricks

Colaborativo, reproduzível e pronto para empresas

Várias linguagens

Use Python, SQL, Scala e R em um só notebook

Colaborativo

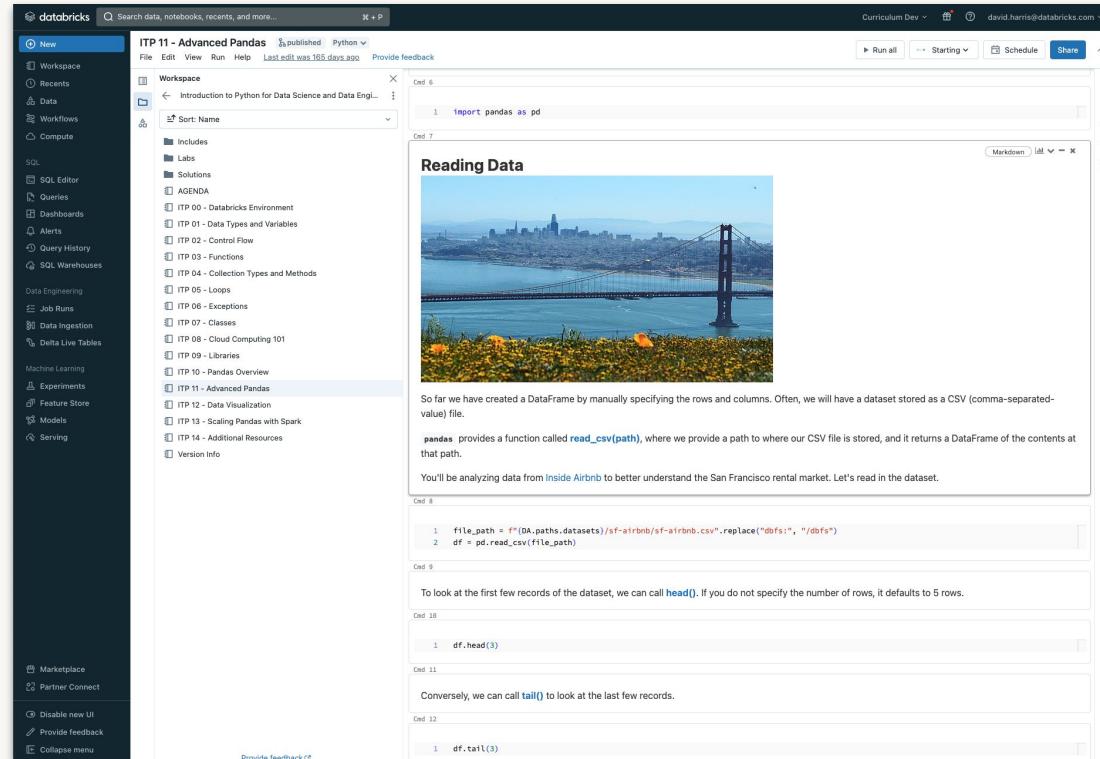
Copresença, coedição e comentários em tempo real

Ideal para exploração

Explore, visualize e resuma dados com gráficos e perfis de dados integrados

Adaptável

Instale bibliotecas padrão e use módulos locais



Reproduzível

Rastreie automaticamente o histórico de versões e use o controle de versão git com Repos

Reduza o tempo para a produção

Programe rapidamente notebooks como jobs ou crie painéis a partir dos resultados, tudo no notebook

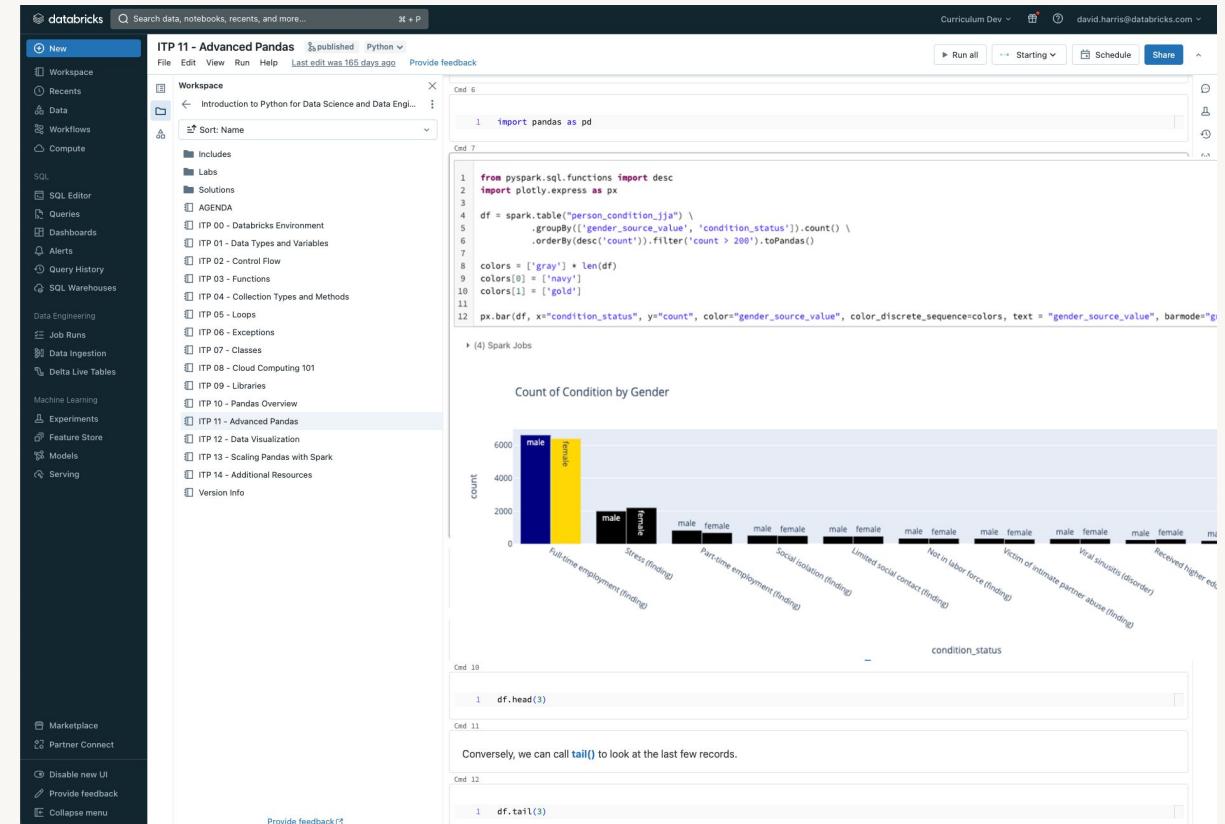
Pronto para empresas

Controles de acesso de nível empresarial, gerenciamento de identidade e auditabilidade

Notebooks do Databricks

Desenvolva com facilidade visualizações padrão ou personalizadas

- Criar visualizações baseadas em resultados de query ou dataframes
- Pode ser usado em SQL, Python ou Scala
 - Use o SQL para visualizações padrão e prontas para usar
 - Use o Python e o Scala para visualizações personalizadas
- Unir visuais personalizados e padrão
- Pode ser usado em tabelas existentes ou pode gravar os resultados do modelo em uma tabela para monitoramento do modelo



Demonstração: como trabalhar com notebooks



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Escrever código e executar um notebook.
- Escrever anotações baseadas em markdown em um notebook.
- Executar código usando várias linguagens no mesmo notebook.



Demonstração

Passos de alto nível

Notebooks

- Sobre os notebooks
- Escrever código e markdown
- Comandos mágicos
- Criar uma visualização



Módulo 2



Módulo 2 – Programação

Nome da lição	Nome da lição
2.1 – Aula: Armazenamento de dados e Delta Lake	2.6 – Demonstração: Como usar jobs de fluxo de trabalho
2.2 – Aula: Unity Catalog	2.7 – Aula: Databricks SQL para engenharia de dados
2.3 – Demonstração: Gestão de dados	2.8 – Demonstração: Como usar o Databricks SQL
2.4 – Demonstração: Governança e segurança de dados	2.9 – Laboratório abrangente
2.5 – Aula: Introdução aos jobs de fluxo de trabalho	

Armazenamento de dados e Delta Lake



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Descrever que os dados são armazenados em locais de armazenamento de objetos na nuvem e acessados via Databricks.
- Explicar os benefícios do armazenamento de dados na arquitetura de data lakehouse entre as funções e os serviços do Databricks.
- Identificar o Delta Lake como a camada de armazenamento otimizada que fornece a base para o armazenamento de dados para o data lakehouse.
- Identificar que todas as tabelas no Databricks são tabelas Delta por padrão.
- Identificar que o Delta Lake tem uma série de otimizações integradas e fáceis que melhoram o desempenho.



Usuários

Interactive users



Data Engineers

Data Scientists

Data Analysts



Plano de controle

databricks

Web application

Configurations

Notebooks,
repos, DBSQL

Cluster manager



Customer

databricks

Data plane



Cluster



Cluster



Cluster

Your cloud storage



Data

Users

Interactive users



Data Engineers

Data Scientists

Data Analysts

Plano de controle

databricks

Web application

Configurations

Notebooks,
repos, DBSQL

Cluster manager

Customer

databricks

Data plane



Cluster



Cluster



Cluster

Your cloud storage



Data

O que é Delta Lake?

O Delta Lake é o formato padrão das tabelas criadas no Databricks

```
CREATE TABLE foo  
USING DELTA
```

```
df.write  
.format("delta")
```



O Delta Lake é um projeto de código aberto que permite criar um data lakehouse sobre o armazenamento existente na cloud

O Delta Lake traz o ACID para o armazenamento de objetos

Atomicidade significa que todas as transações são bem-sucedidas ou falham completamente

Consistência com garantias refere-se a como um determinado estado dos dados é observado por operações simultâneas

Isolamento refere-se a como as operações simultâneas entram em conflito umas com as outras. O isolamento do Delta Lake tem garantias diferentes daquelas de outros sistemas

Durabilidade significa que as alterações confirmadas são permanentes



Problemas resolvidos pelo ACID

- Dificuldade para acrescentar dados
- Dificuldade para modificar dados existentes
- Falha de jobs durante o processamento
- Dificuldade para executar operações em tempo real
- Caro para manter versões de dados históricos



Unity Catalog



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Descrever o Unity Catalog como uma solução de governança centrada no Databricks.
- Explicar o namespace de três camadas e seus níveis.



Unity Catalog

Visão geral



Governança unificada entre clouds

Governança refinada para data lakes em clouds, com base no padrão aberto ANSI SQL.

1



Dados e ativos de IA unificados

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

2



Catálogos existentes unificados

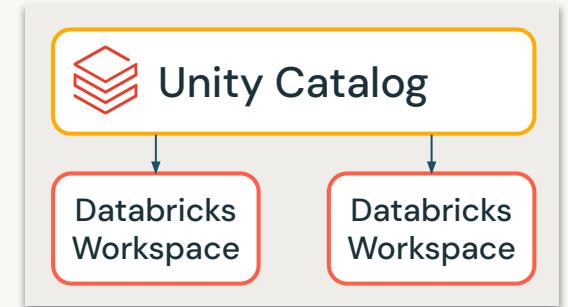
Funcionam em conjunto com dados, armazenamento e catálogos existentes, sem necessidade de migração total.

3

Unity Catalog

Principais recursos

- Metadados e gerenciamento de usuários centralizados
- Controles centralizados de acesso a dados
- Auditoria de acesso a dados
- Linhagem de dados
- Pesquisa e descobrimento de dados
- Compartilhamento seguro de dados com o Delta Sharing



GRANT... ON... TO...
REVOKE... ON... FROM...

Catálogos, bancos de dados
(esquemas), tabelas,
visualizações, credenciais de
armazenamento, locais externos

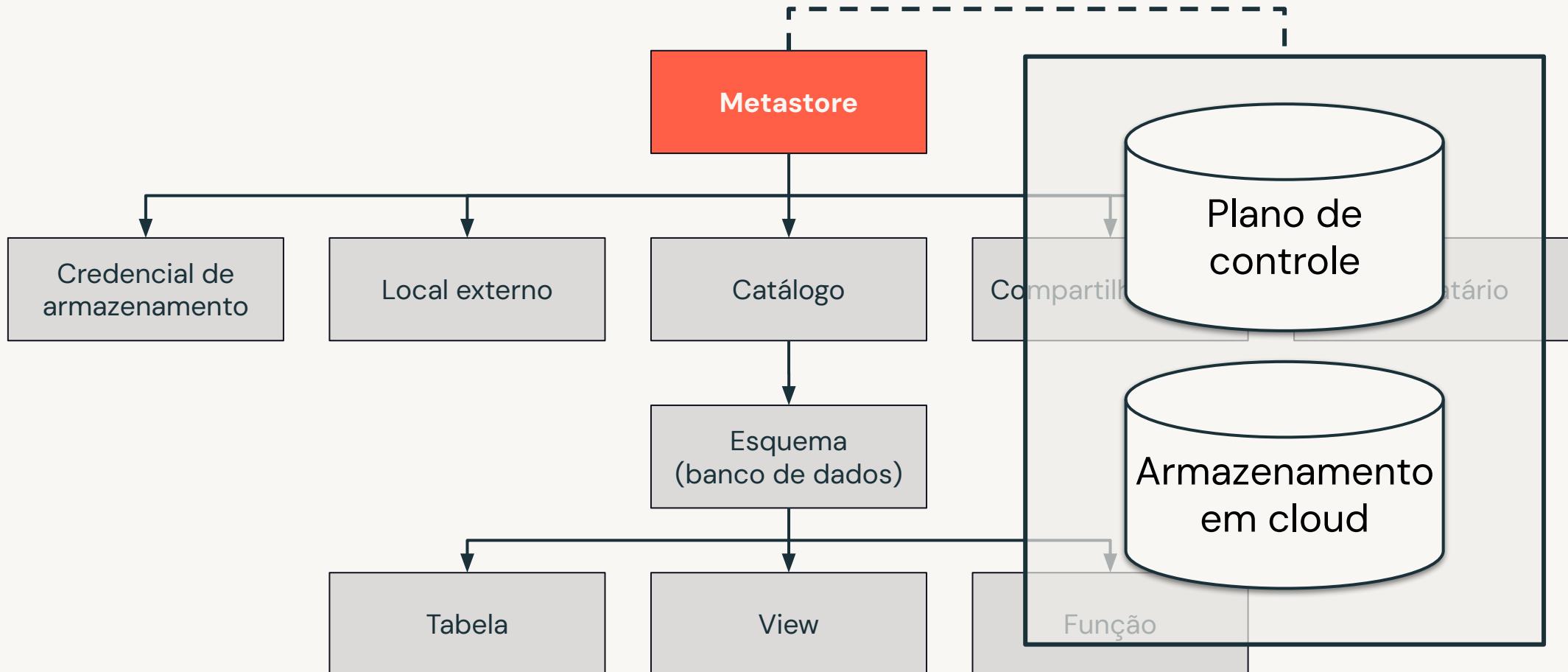


Principais conceitos do Unity Catalog



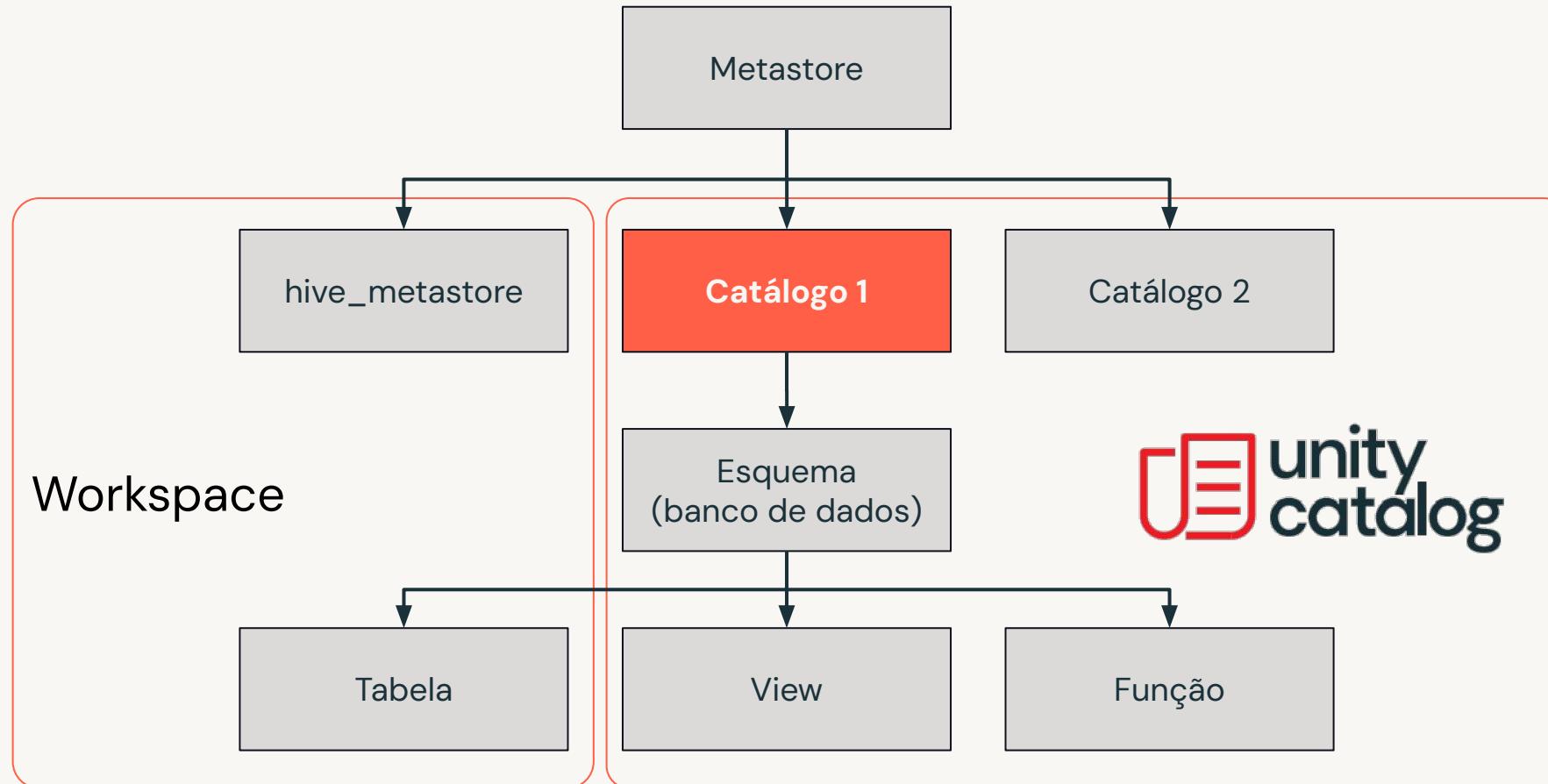
Metastore

Elementos do metastore do Unity Catalog



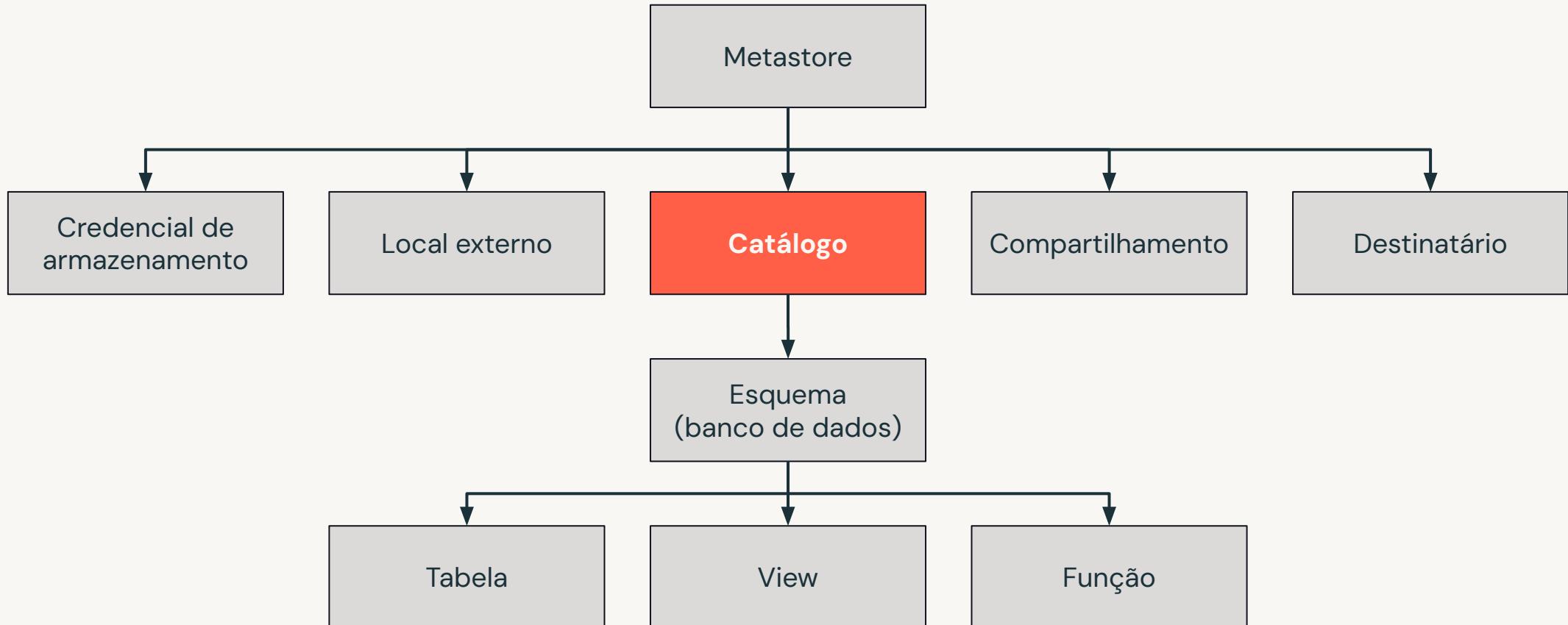
Metastore

Acessando o Hive metastore legado



Catálogo

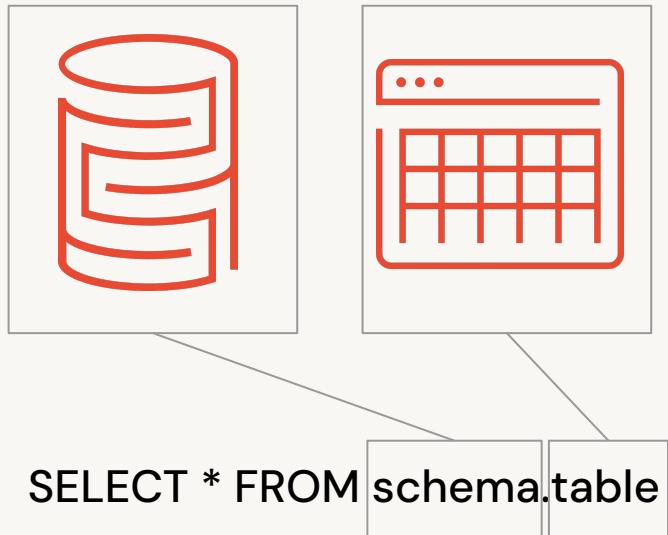
Contêiner de nível superior para objetos de dados



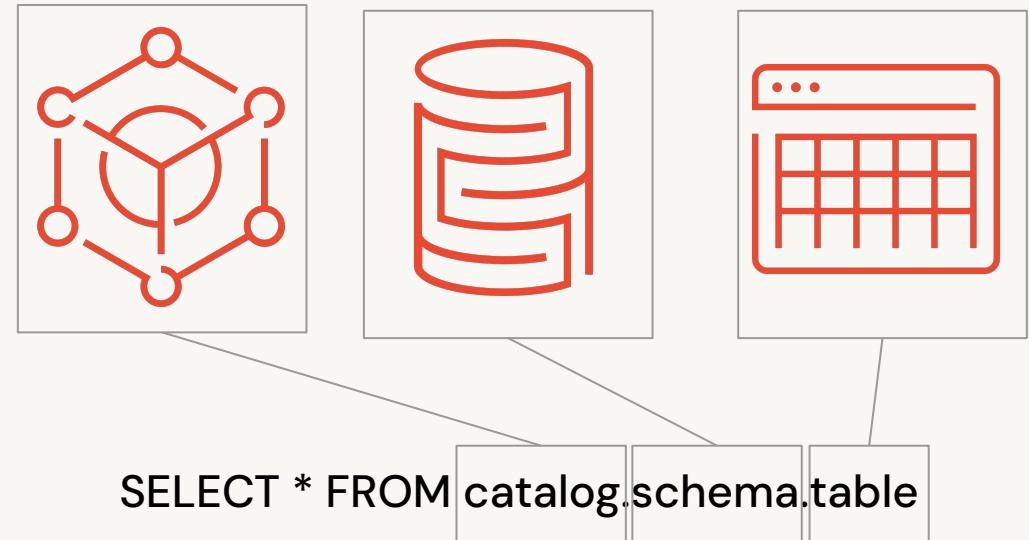
Catálogo

Namespace de três níveis

Namespace SQL tradicional
de dois níveis

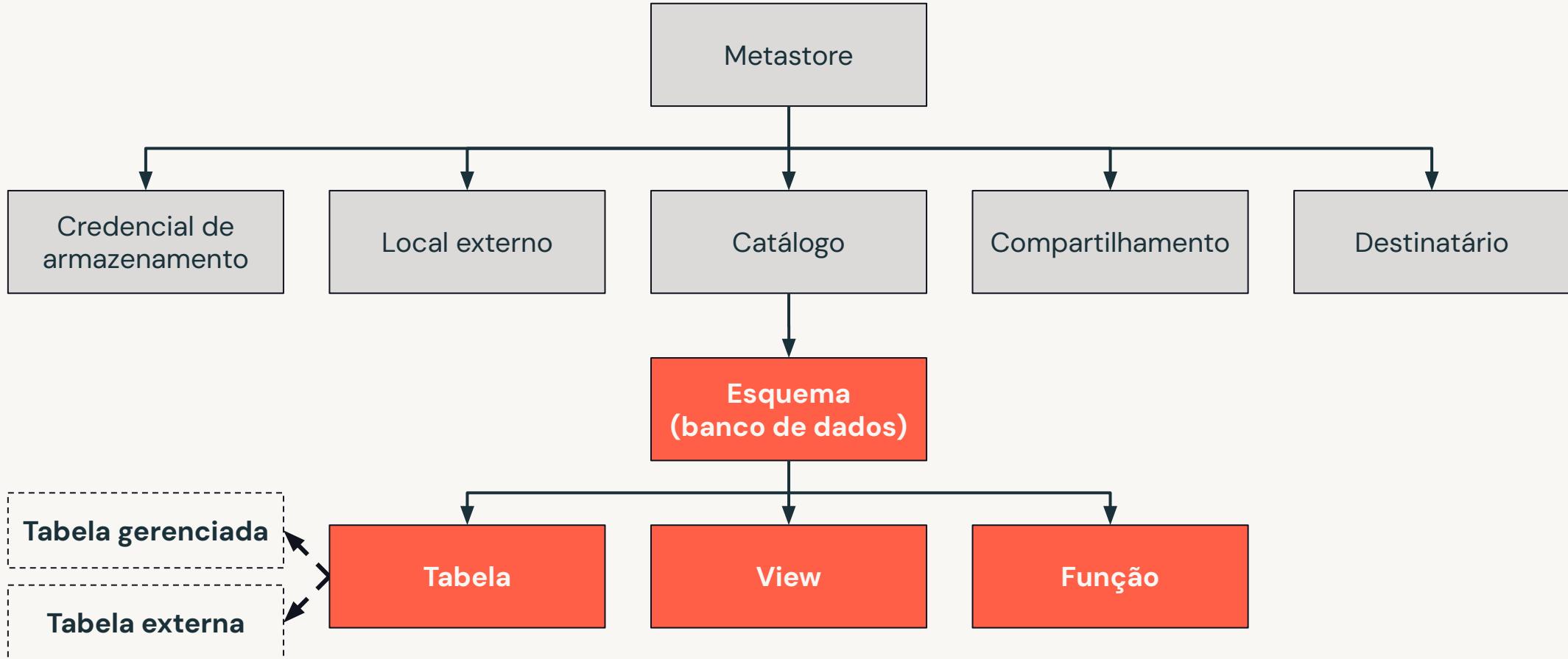


Namespace de três
níveis do Unity Catalog



Objetos de dados

Esquema (banco de dados), tabelas, views, funções



Demonstração: gestão de dados



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Criar um novo esquema em um catálogo existente.
- Criar uma nova tabela Delta gerenciada a partir de um arquivo de cloud existente com o SQL.
- Criar uma nova tabela Delta gerenciada a partir de uma tabela Delta existente com o SQL.
- Eliminar uma tabela Delta gerenciada que não é mais necessária.



Demonstração

Passos de alto nível

Trabalhar com dados

- Sobre as tabelas Delta
- Sobre o Unity Catalog
- Catálogos e esquemas
- Criar tabelas



Demonstração: governança e segurança de dados



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Conceder acesso adequado a outro usuário para SELECT em uma tabela.
- Conceder acesso adequado a um grupo para SELECT em uma tabela.
- Revogar o acesso de outro usuário a uma tabela.



Demonstração

Passos de alto nível

Governança e segurança

- Usuários e grupos
- Conceder e revogar acesso



Introdução aos jobs de fluxo de trabalho



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

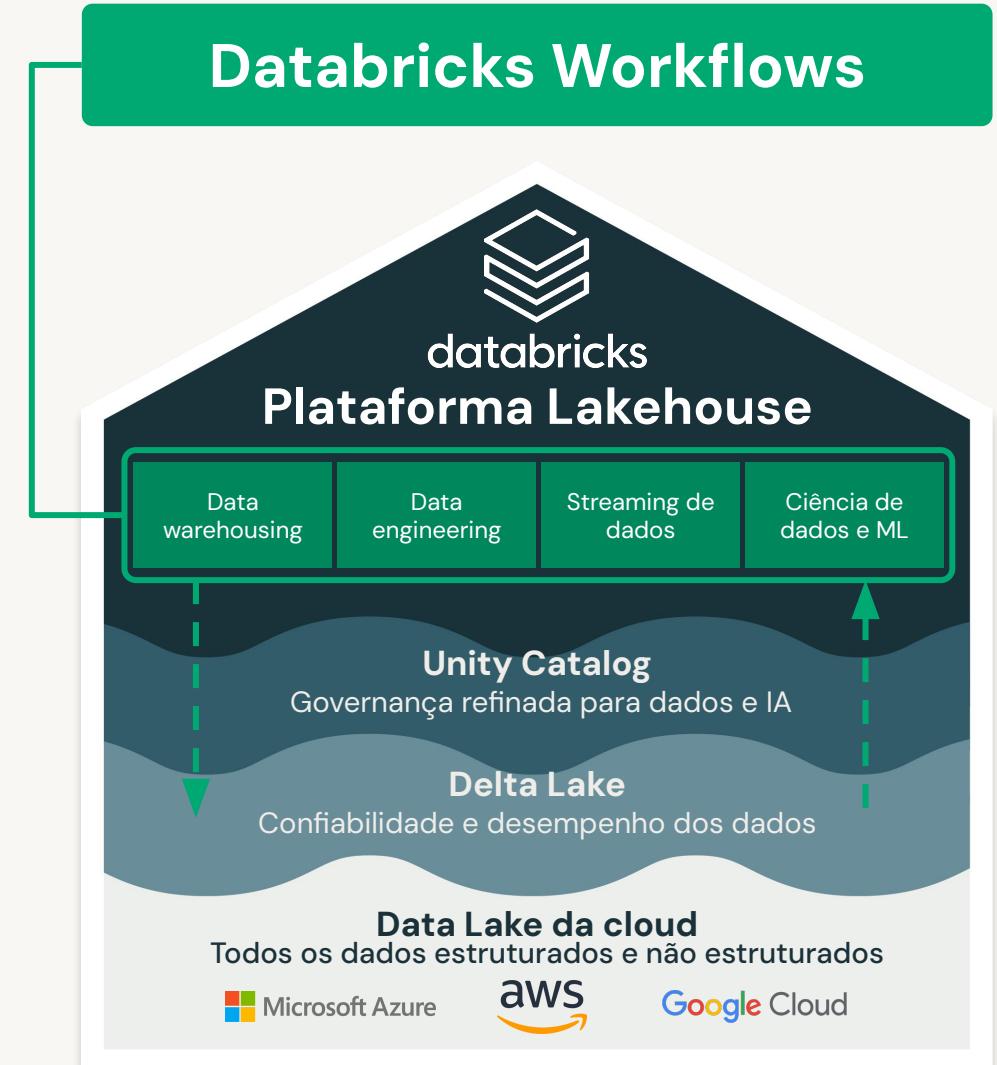
- Descrever o Workflows como um recurso que produz fluxos de trabalho de dados.
- Descrever o Jobs como uma solução simples de agendamento e automatização de uma ou mais tarefas.
- Reconhecer os tipos de ativos que podem ser automatizados com Jobs.
- Descrever o Delta Live Tables como uma solução que cria e executa pipelines de dados robustos.



Databricks Workflows

O Workflows é um **serviço de orquestração de tarefas de propósito geral, totalmente gerenciado e baseado na cloud para todo o lakehouse.**

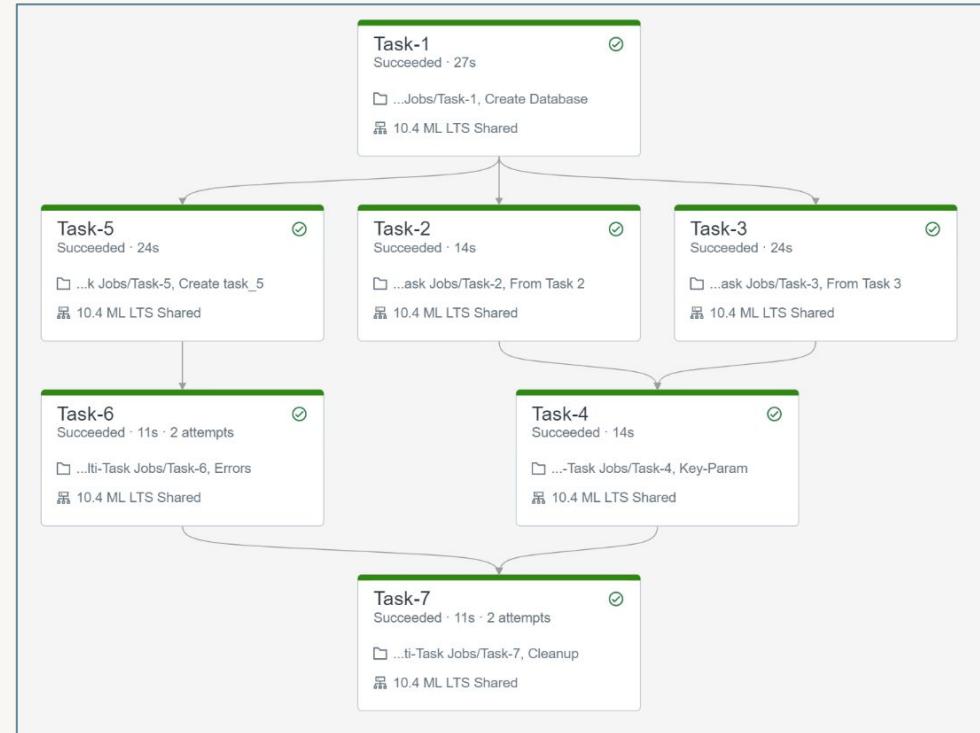
O Workflows é um serviço para engenheiros de dados, cientistas de dados e analistas criarem dados, análises e fluxos de trabalho de IA confiáveis em qualquer cloud.



Databricks Workflows

O Databricks tem dois serviços principais de orquestração de tarefas:

- **Workflow Jobs (Workflows)**
 - Workflows para todos os jobs
- **Delta Live Tables (DLT)**
 - Pipelines de dados automatizados para o Delta Lake



Observação: o pipeline do DLT pode ser uma tarefa em um fluxo de trabalho



DLT versus jobs de Workflow

Considerações

	Delta Live Tables	Workflow Jobs
Fonte	Somente notebooks	JARs, notebooks, DLT, aplicativos escritos em Scala, Java, Python
Dependências	Determinadas automaticamente	Definidas manualmente
Cluster	Autoprovisionado	Autoprovisionado ou existente
Limites de tempo e novas tentativas	Sem suporte para limites de tempo Tentativas controladas automaticamente (no modo de produção)	Compatível
Importação de bibliotecas	Sem compatibilidade	Compatível



Workflow Jobs

Casos de uso

Orquestração de jobs dependentes

Jobs executados dentro do agendamento, contendo tarefas/passos dependentes

Tarefas de machine learning

Executar a tarefa do notebook MLflow em um job

Código arbitrário, chamadas externas de API, tarefas personalizadas

Executar tarefas em um job que pode conter arquivos Jar, Spark Submit, Python Script, tarefa SQL, dbt

Workflows de jobs

Workflows de jobs

Workflows de jobs



Como utilizar o Workflows

- Permite criar uma orquestração simples de tarefas ETL/ML
- Reduz a sobrecarga da infraestrutura
- Integra-se facilmente a ferramentas externas
- Permite que profissionais sem experiência em engenharia criem fluxos de trabalho usando uma interface simples
- Independente do provedor de cloud
- Permite a reutilização de clusters para reduzir custos e tempo de inicialização



Demonstração: como usar jobs de Workflow



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Automatizar a execução de um único notebook usando Jobs.
- Analisar os resultados de um Job de notebook único concluído.



Demonstração

Passos de alto nível

Workflows

- Sobre fluxos de trabalho
- Criar e executar jobs
- Revisar os resultados do job



Databricks SQL para engenharia de dados



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

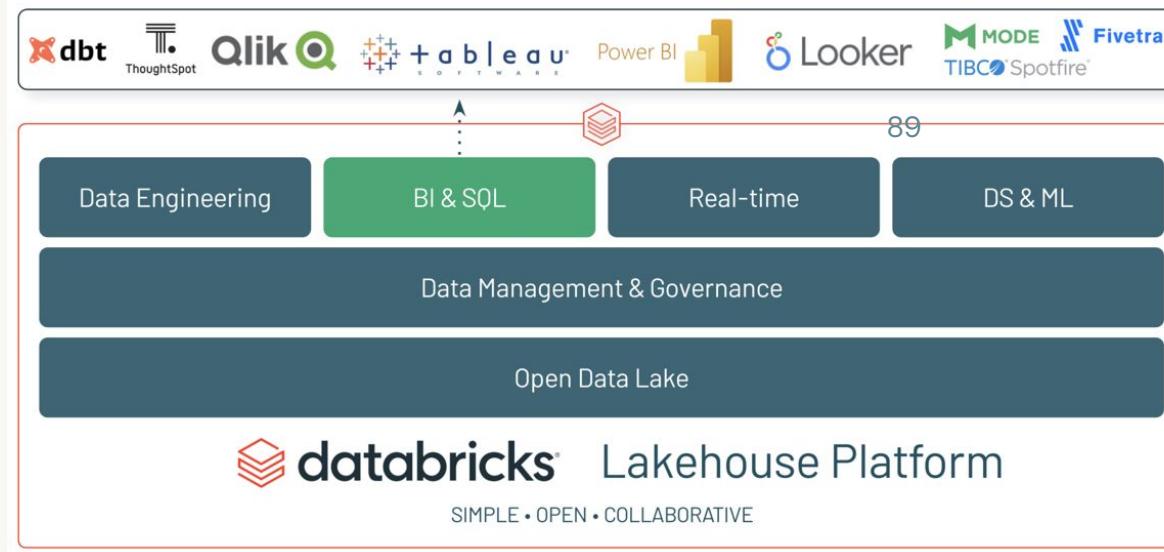
- Descrever o Databricks SQL como uma solução de data warehousing para analistas e engenheiros que trabalham com o Databricks.
- Reconhecer casos de uso comuns para engenharia de dados ao trabalhar com o Databricks SQL.
- Descrever os recursos de visualização e painéis do Databricks SQL.



Databricks SQL

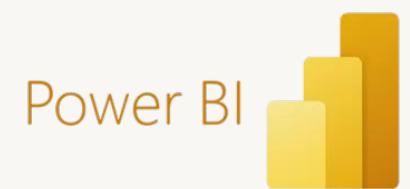
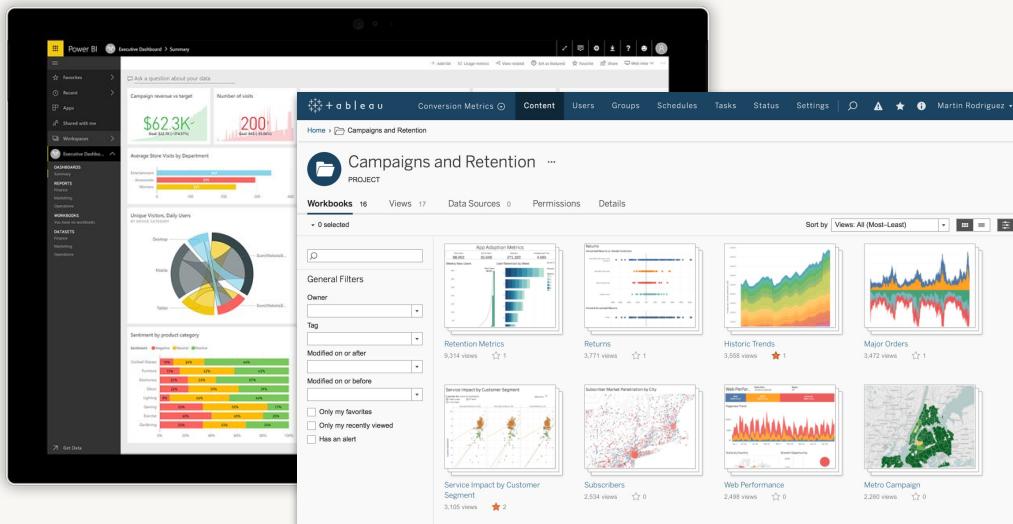
Fornecendo análises dos dados mais recentes com desempenho de data warehouse e economia de data lake

- Melhor preço/desempenho do que outros cloud data warehouses
- Simplifique o descobrimento e o compartilhamento de novas percepções
- Conecte-se a ferramentas de BI conhecidas, como Tableau ou Power BI
- Administração e governança simplificadas



Ampla integração a ferramentas de BI

Conekte suas ferramentas de BI preferidas com conectores otimizados que fornecem desempenho rápido, baixa latência e alta simultaneidade de usuários no data lake usando as ferramentas de BI existentes.



Em breve:

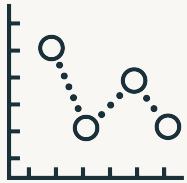


Casos de uso



Visualizar dados para descobrir problemas

Pode ser difícil encontrar problemas nos dados sem a criação de uma visualização. O Databricks SQL oferece a oportunidade de produzir uma ampla variedade de tipos de visualização que permitem examinar os dados e encontrar problemas para poderem ser resolvidos rapidamente. A facilidade de executar uma query, produzir uma visualização e criar um painel, caso necessário, faz do Databricks SQL uma solução incrível.



Explore de forma colaborativa os dados mais recentes e atualizados

Responda às necessidades dos negócios mais rapidamente com uma experiência de autoatendimento projetada para todos os analistas da organização. O Databricks SQL Analytics fornece acesso simples e seguro aos dados, capacidade de criar ou reutilizar queries SQL para analisar dados diretamente em seu data lake, além de permitir simular e iterar rapidamente em exibições e painéis adaptados à sua empresa.



Crie aplicativos aprimorados por dados

Crie aplicativos avançados e personalizados com dados aprimorados para sua organização ou para seus clientes. Beneficie-se da facilidade de conectividade, gestão e melhor preço/desempenho da análise do Databricks SQL para simplificar o desenvolvimento de aplicativos melhorados em escala, tudo a partir do seu data lake.

Importar painel existente



Tipos de visualização



Como usar o Databricks SQL



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Abrir o Editor SQL.
- Escrever e executar uma query.



Demonstração

Passos de alto nível

Databricks SQL

- Sobre o Databricks SQL
- Criar e executar queries
- Criar visualizações



Laboratório abrangente



Objetivos de aprendizado

Tudo o que será capaz de fazer após a conclusão desta lição

- Demonstrar como criar um fluxo de trabalho completo de data engineering na Plataforma Databricks Lakehouse



Laboratório

Passos de alto nível

Demonstre suas habilidades

- Trabalhar com clusters
- Criar uma tabela e conceder acesso
- Escrever código em um notebook que ingere, limpa e carrega dados
- Criar e executar um job
- Criar e executar uma query
- Criar uma visualização



Resumo do curso e próximos passos



Resumo e próximas passos

O que foi abordado? O que fazer agora?

Resumo da sessão

- Introdução à Plataforma Lakehouse da Databricks
- Habilidades básicas para executar um fluxo de trabalho de data engineering

Recursos úteis

- [Documentos de data engineering](#)

Próximos passos

- Faça o curso "Data Engineering no Databricks" para conhecer melhor o assunto
- Confira nossa [Certificação em Data Engineering](#)

