

VINÍCIUS HENRIQUE DOS SANTOS

POSTECH

MACHINE LEARNING ENGINEERING

BIG DATA STORAGE STRUCTURES

# AULA 03

---

SUMÁRIO

O QUE VEM POR AÍ? .....3

HANDS ON .....4

SAIBA MAIS .....5

O QUE VOCÊ VIU NESTA AULA? .....14

REFERÊNCIAS.....15



## O QUE VEM POR AÍ?

Prepare-se para entender, explorar e aprender como realizar ingestões, manipular arquivos, analisar dados e trabalhar com dados estruturados dentro da líder no seguimento de Data Lakehouses, a Databricks.

Aqui você vai entender quais características do Data Lake são combinadas com Data Warehouse para entregar facilidade de uso, agilidade no processamento e utilização dos dados, além de uma possibilidade de governança única e simplificada.

Assim, você irá caminhar desde a criação da conta na plataforma de estudos até a ingestão, o tratamento e a manipulação dos dados na plataforma.

## HANDS ON

Nessa aula, veremos o seguinte:

1. Crie sua conta no Databricks Community.
2. Faça o carregamento do arquivo de dados Spotify.
3. Transforme o arquivo em uma tabela.
4. Crie um script de automatização da criação de tabelas.
5. Realize a movimentação de arquivos dentro do DBFS.
6. Realize consultas SQL.

## SAIBA MAIS

### O que é um Lakehouse?

Existem várias definições de mercado sobre o que de fato é um Lakehouse, mas aqui vamos seguir o que é recomendado pela maior fornecedora de soluções e serviços desse tipo ao longo dos últimos 10 anos, a Databricks.

Segundo a empresa líder de mercado por mais de uma década, podemos definir um Data Lakehouse como uma arquitetura que nos permite trabalhar desde Business Intelligence até inteligência artificial dentro da plataforma de Data Lake.

Hoje todas as empresas armazenam seus dados em Data Lakes por causa das vantagens significativas que esse tipo de solução oferece, viabilizando armazenar grandes volumes de dados estruturados ou desestruturados em um mesmo local, com baixo custo e permitindo o consumo e manipulação dos dados por qualquer ferramenta que seja capaz de processá-los.

Dentro dos Data Lakes é onde temos as cargas de trabalho consumindo informação e gerando insights através de relatórios estruturados, processamento de dados não estruturados e até mesmo treinamento de modelos de inteligência artificial, todos com o mesmo objetivo: promover vantagem competitiva para a organização.

Separadamente, para o cenário de Business Intelligence (BI) é utilizada uma porção de dados presentes no Data Lake de forma estruturada em que basicamente teremos ferramentas de BI para trazer respostas sobre o histórico dos dados, analisando questões sobre o comportamento passado usando ferramentas como SQL ou PowerBI.

Mas, no mesmo momento e eventualmente no mesmo local, a empresa possui jobs que usam ferramentas de manipulação e análise em dados não estruturados como NoSQL ou modelagem para IA, buscando prever comportamentos sobre o cliente, mercado ou produtos.

Ao investir em uma plataforma de Data Lakehouse como a Databricks, as empresas conseguem navegar de soluções de BI até soluções de IA em um mesmo local, tendo escalabilidade, poder de processamento e armazenamento praticamente ilimitados.

**Mas, afinal, qual é a diferença entre Data Lakehouse e Dataware House?**

O conceito de Lakehouse é desenvolvido sobre Data Lakes já existentes, os quais frequentemente abrigam mais de 90% dos dados corporativos. Embora a maioria dos Data Warehouses possa acessar esses dados por meio de funções de "tabelas externas", eles enfrentam limitações significativas tanto em funcionalidades (por exemplo: limitando-se a operações de leitura) quanto em desempenho.

Diferentemente, o Lakehouse incorpora funcionalidades típicas de um Data Warehouse em Data Lakes já existentes, oferecendo transações ACID, segurança refinada dos dados, atualizações e exclusões econômicas, suporte avançado a SQL, performance otimizada para consultas SQL e relatórios no estilo BI.

O Lakehouse gerencia todos os dados presentes em um Data Lake, englobando diversos tipos de dados, como textos, áudios e vídeos, além de dados estruturados em tabelas. Ademais, ele suporta nativamente aplicações de ciência de dados e aprendizado de máquina, proporcionando acesso direto aos dados através de APIs abertas e suporte a diversas bibliotecas de ML e linguagens como Python e R, incluindo PyTorch, Tensorflow e XGBoost, diferentemente dos Data Warehouses.

Portanto, o Lakehouse se apresenta como uma solução única para a gestão integral dos dados empresariais, facilitando a análise desde BI até inteligência artificial.

Por outro lado, os Data Warehouses são sistemas de dados específicos, projetados para análises de SQL em dados estruturados e alguns tipos de dados semi-estruturados. Esses sistemas possuem capacidades limitadas para o aprendizado de máquina e não permitem a execução nativa de ferramentas populares de código aberto, a menos que os dados sejam exportados primeiramente. Atualmente, nenhum sistema de Data Warehouse oferece suporte nativo a todos os dados de áudio, imagem e vídeo já armazenados nos Data Lakes.

**E qual é a diferença entre Lakehouse e Data Lakes?**

É muito comum encontrarmos empresas em que a governança do Data Lake foi perdida ou mal planejada o que, como consequência, acaba transformando o ambiente em um "limbo" com baixo nível de gerenciamento que afeta até mesmo a performance e os custos de se manter um Data Lake de forma eficiente.

Por conta disso, boa parte das empresas usam os Data Lakes como “zona de desembarque”, em que os dados brutos são depositados para serem consumidos e processados em outro sistema e ferramentas, a fim de extrair valor e insights desse volume de dados extremamente significativo.

O Lakehouse resolve esses desafios críticos que transformam Data Lakes em ambientes desorganizados: ele implementa transações ACID para manter a consistência quando dados são acessados ou modificados por múltiplas partes simultaneamente.

Além disso, ele adota arquiteturas de esquemas de DW, como os esquemas em estrela ou floco de neve, e incorpora mecanismos eficazes de governança e auditoria diretamente no Data Lake.

Ele emprega também diversas estratégias de otimização de desempenho, como armazenamento em cache, clusterização multidimensional e omissão de dados baseada em estatísticas de arquivos e compactação de dados para ajustar o tamanho dos arquivos e acelerar as análises.

Inclui ainda recursos de segurança detalhados e capacidades de auditoria para a governança de dados. Ao integrar a gestão de dados e otimizações de desempenho ao Data Lake aberto, o Lakehouse oferece suporte nativo para aplicações de BI e ML.

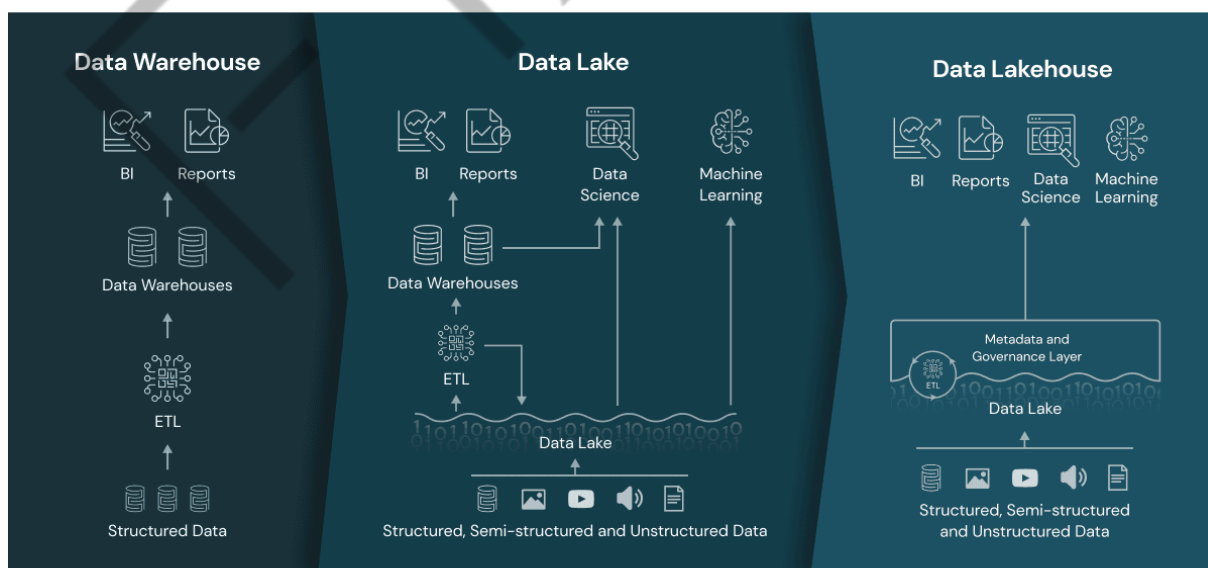


Figura 1 – Data Warehouse, Data Lake e Data Lakehouse  
Fonte: Elaborado pelo autor (2024)

Estão emergindo novos sistemas destinados a superar as restrições associadas aos Data Lakes. O Lakehouse representa uma arquitetura inovadora e aberta, que integra as características mais vantajosas tanto dos Data Lakes quanto dos Data Warehouses.

Essa nova configuração de sistema permite a implementação de estruturas de dados e funcionalidades de gestão de dados típicas de Data Warehouses, aplicadas diretamente sobre armazenamentos em nuvem de baixo custo e em formatos abertos.

Os Lakehouses seriam o resultado de uma reformulação dos Data Warehouses na era atual, considerando a disponibilidade de soluções de armazenamento baratas e extremamente confiáveis, como os armazenamentos de objetos.

### **O que é Databricks?**

Para gerenciar e processar os dados de uma empresa à medida que a necessidade de escalar armazenamento e processamento de máquina crescem, indivíduos engenheiros e analistas muitas vezes precisam usar diversas ferramentas diferentes. Isso demanda necessidades específicas dentro de cada uma das ferramentas, o que pode ser complexo, caro e promover uma lenta curva de aprendizagem dentro da organização.

Dado esse cenário, o mesmo grupo de engenheiros da Universidade da Califórnia enxergou a oportunidade de resolver essa dor do mercado, lançando a plataforma de solução em nuvem que permite processar, transformar, explorar, armazenar e analisar grandes volumes de dados.

O objetivo principal é permitir que a empresa gaste mais energia com insights e soluções de negócio do que infraestrutura e manutenção das plataformas de dados de uma forma simplificada e centralizada. A plataforma pode ser configurada para trabalhar com vários serviços de nuvem, como AWS, Azure e Google Cloud.

A Databricks foi construída em cima do Apache Spark, permitindo que os dados sejam processados em clusters de forma distribuída, fator que traz eficiência e escalabilidade, além de simplicidade de gerenciamento dos recursos computacionais.

O desenvolvimento técnico dentro da plataforma gira em torno da criação de códigos Python, R e SQL dentro de notebooks que serão lidos para determinar toda a lógica de utilização dos dados.



Além disso, dentro da plataforma é possível realizar a construção de pipelines, permitindo que a lógica e o fluxo sejam quebrados em etapas para acompanhamento de resultados em tempo real.

### Os casos de uso mais comuns

1. **Análise exploratória de dados:** permite que analistas realizem descobertas e exploração de dados combinando ferramentas como SQL e Python ou SQL e R, manipulando e agregando os dados, transformando e gerando visualizações para facilitar o comportamento dos dados.
2. **Machine Learning:** dentro da plataforma, é oferecido suporte para desenvolvimento e manutenção dos modelos de forma colaborativa por todo o ciclo de vida, passando por desenvolvimento, treinamento e implementação.
3. **Processamento de dados em batch ou real time:** você pode desenvolver toda a lógica de ingestão, tratamento e processamento de dados em tempo real ou em lote de forma simples e integrada.

### Criando a conta Databricks

Nós criaremos uma conta na Databricks para que possamos acessar e usar os recursos do Databricks Community, uma plataforma com os recursos Databricks direcionados para o aprendizado na plataforma. Em um primeiro momento iremos criar a conta na versão de testes, na qual você obterá acesso aos recursos por 14 dias. Ela se faz necessária para podermos usar a versão Community. Acesse [aqui](#).

**databricks**

Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud. Sign-up with your work email to elevate your trial experience.

- ✓ Create high quality Generative AI applications  
Build production quality generative AI applications and ensure your output is accurate, current, aware of your enterprise context, and safe.
- ✓ Simplify data ingestion and automate ETL  
Ingest data from hundreds of sources. Use a simple declarative approach to build data pipelines.
- ✓ Get \$400 in serverless compute credits to use during your trial  
Access instant, elastic compute during your trial. Please note that serverless compute is not available on Google Cloud Platform or for Databricks Partners.

**Create your Databricks account** 1/2

Sign up with your work email to elevate your trial with expert assistance and more.

First name  Last name

Email

Company  Title

Phone (Optional)

Country

By submitting, I agree to the processing of my personal data by Databricks in accordance with our [Privacy Policy](#). I understand I can [update my preferences](#) at any time.

**Continue**

Figura 2 – Criando uma conta na Databricks

Fonte: Elaborado pelo autor (2024)

Depois de preencher os dados, você receberá um e-mail de confirmação e será direcionado(a) para vincular sua conta da Databricks com o seu provedor de cloud. Você obrigatoriamente deverá fornecer uma conta válida de algum provedor para que sua conta na plataforma seja válida e consiga usar os recursos.

A plataforma Databricks monta os recursos que você utilizará em seus projetos diretamente no seu provedor de nuvem, mas nós cumprimos essa etapa apenas para ter acesso na versão Community.

A versão Community da Databricks oferece funcionalidades mais restritas do que o ambiente corporativo e pago, mas nos permite trabalhar de forma similar ao real sem custos. Ela nos dará acesso a um cluster de 15 GB, um gerenciador básico de cluster e um ambiente para trabalharmos com notebooks.

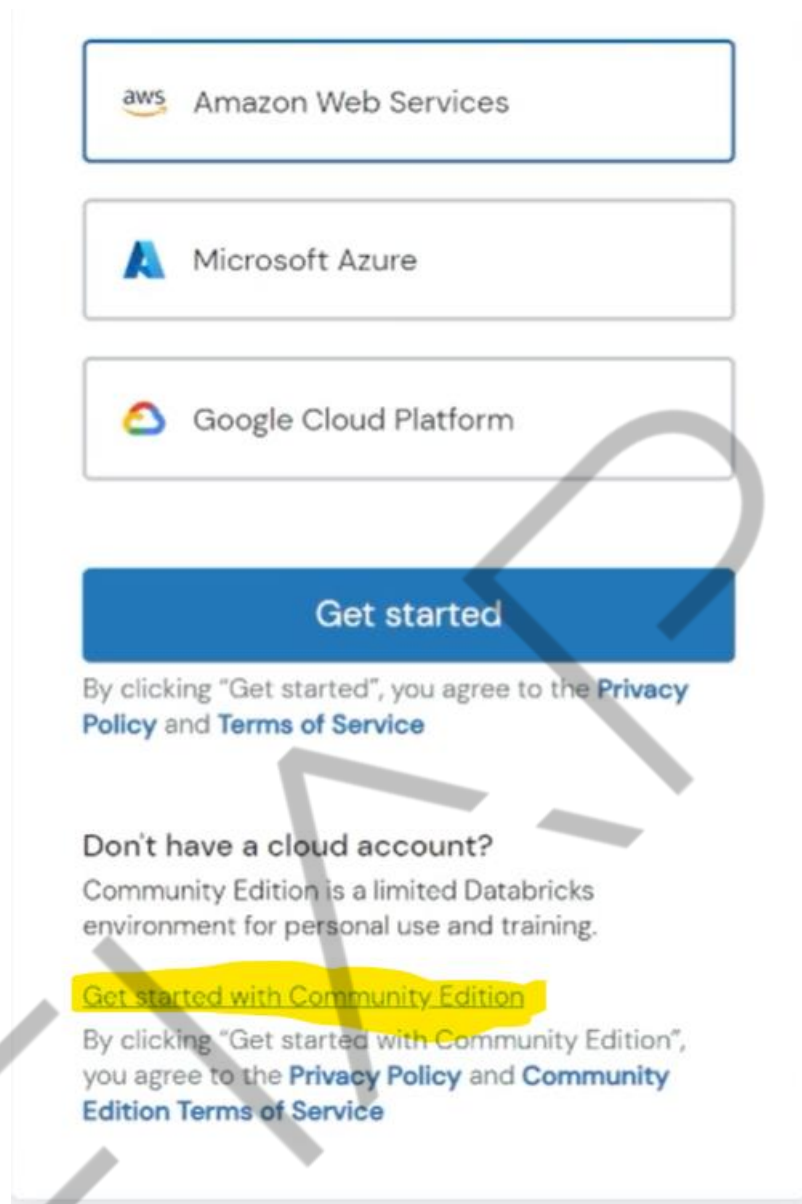
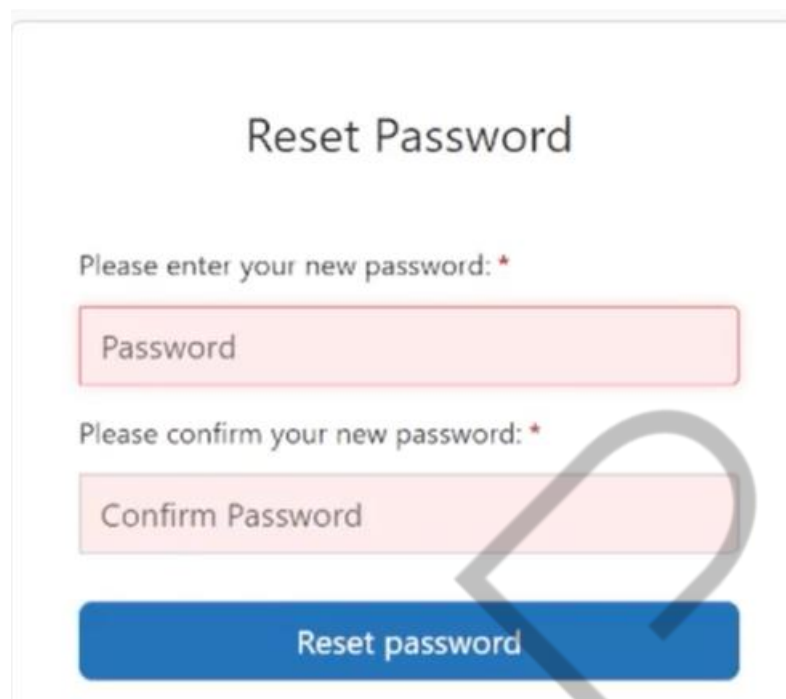


Figura 3 – Versão Community da Databricks

Fonte: Elaborado pelo autor (2024)

Acesse seu e-mail fornecido para realizar a verificação e o cadastro de senha:



Reset Password

Please enter your new password: \*

Password

Please confirm your new password: \*

Confirm Password

Reset password

Figura 4 – Verificação e cadastro da senha  
Fonte: Elaborado pelo autor (2024)

Com a senha criada, acesse a plataforma. Você será direcionado(a) para uma página como essa e sua conta está pronta para uso:

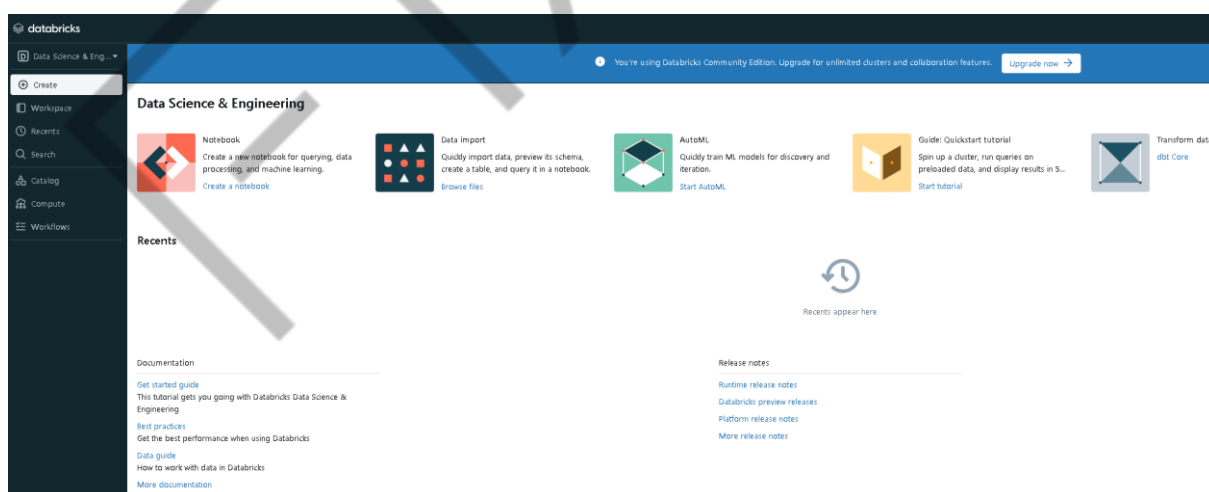


Figura 5 – Plataforma da Databricks  
Fonte: Elaborado pelo autor (2024)

## Databricks File System

Databricks File System, ou também conhecido como DBFS, é o sistema de arquivos distribuídos construído dentro da Workspace (ou área de trabalho) da plataforma Databricks disponível para uso quando um cluster ativo está em execução.

O DBFS permite a usuários e aplicações interagirem com os dados armazenados em diversos ambientes de armazenamento em nuvem como se estivessem trabalhando com um sistema de arquivos local, entregando uma camada de abstração sobre serviços como Amazon S3, Azure BlobStorage ou Google Cloud Storage, que são usados como “back-end” de armazenamento.

### Principais características do DBFS:

1. **Integração Transparente:** o DBFS integra-se de forma transparente com os ambientes de nuvem, permitindo que usuários acessem e gerenciem seus dados sem precisar interagir diretamente com os serviços de armazenamento de objetos da nuvem.
2. **Facilidade de Uso:** os usuários podem montar seus armazenamentos de dados na nuvem como diretórios no DBFS, possibilitando o acesso aos dados usando comandos típicos de sistemas de arquivos.
3. **Acesso com APIs:** o DBFS pode ser acessado através de APIs disponíveis no Databricks, facilitando a leitura, a escrita e o gerenciamento de arquivos diretamente de notebooks, jobs e aplicações de dados.
4. **Compatibilidade com Hadoop:** por ser compatível com APIs do Hadoop, o DBFS permite que as ferramentas baseadas em Hadoop operem nos dados armazenados como se estivessem em um Hadoop Distributed File System (HDFS), sem a necessidade de modificar as aplicações existentes.
5. **Segurança e Gerenciamento:** o DBFS segue as políticas de segurança e governança de dados do Databricks, incluindo controle de acesso, criptografia em repouso e em trânsito e integração com sistemas de autenticação.

## O QUE VOCÊ VIU NESTA AULA?

Nessa aula você teve a oportunidade de passar pela parte teórica que promoveu o entendimento das características dos Data Lakehouses e o que eles trazem de bom dos Data Lakes, combinando com o que há de bom dos Data Warehouses.

Além disso, você conseguiu criar e configurar sua conta na plataforma, realizou o desenvolvimento de ingestão de dados e foi capaz de automatizar um script para importação. Por fim, você movimentou arquivos dentro do DBFS e realizou consultas SQL que podem te ajudar a trazer grandes insights no seu cenário real.

## REFERÊNCIAS

LORICA, B. et al. **What Is a Lakehouse?** 2020. Disponível em: <[https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html?itm\\_data=lakehouse-link-lakehouseblog](https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html?itm_data=lakehouse-link-lakehouseblog)>. Acesso em: 24 abr. 2024.

DATABRICKS DOCUMENTATION. **What is the Databricks File System (DBFS)?** 2023. Disponível em: <<https://docs.databricks.com/en/dbfs/index.html>>. Acesso em: 24 abr. 2024.

DATABRICKS DOCUMENTATION. **Navigate the workspace.** 2024. Disponível em: <<https://docs.databricks.com/en/workspace/index.html>>. Acesso em: 24 abr. 2024.

DATABRICKS DOCUMENTATION. **What are all the Delta things in Databricks?** 2024. Disponível em: <<https://docs.databricks.com/en/introduction/delta-comparison.html>>. Acesso em: 24 abr. 2024.

DATABRICKS DOCUMENTATION. **Databricks Utilities (dbutils) reference.** 2024. Disponível em: <<https://docs.databricks.com/en/dev-tools/databricks-utils.html>>. Acesso em: 24 abr. 2024.

## **PALAVRAS-CHAVE**

**Palavras-chave:** Data Lake. Databricks. Datawarehouse. Lakehouse. SQL. Spark. DBFS.

EMSE





POSTECH