

INTRODUÇÃO À ENGENHARIA DE DADOS COM SPARK



QUEM SOU EU?

- Egressa de ADS (2019 à 2021)
- MBA em Data Science e Analytics (2022 à 2024)
- Jovem Aprendiz (12/2022 a 05/2023)
- Analista de Banco de Dados (06/2023 a 12/2023)
- Analista de Engenharia de Dados (atual)
- Apaixonada por ler

LINKEDIN



<https://www.linkedin.com/in/vitoriarleonardo>

GITHUB



Link para o repositório com os exemplos, exercícios e slides:

<https://github.com/vitoriarl/introducao-a-engenharia-de-dados-spark/>

OBJETIVOS DA OFICINA

É esperado que os participantes ao final da oficina saibam:

1. O que é Engenharia de Dados e o papel do Engenheiro de Dados
2. O ciclo de vida da Engenharia de Dados
3. O que é o Spark e o porquê é usado na Engenharia de Dados
4. Transformações básicas com PySpark no Databricks



O QUE É ENGENHARIA DE DADOS?

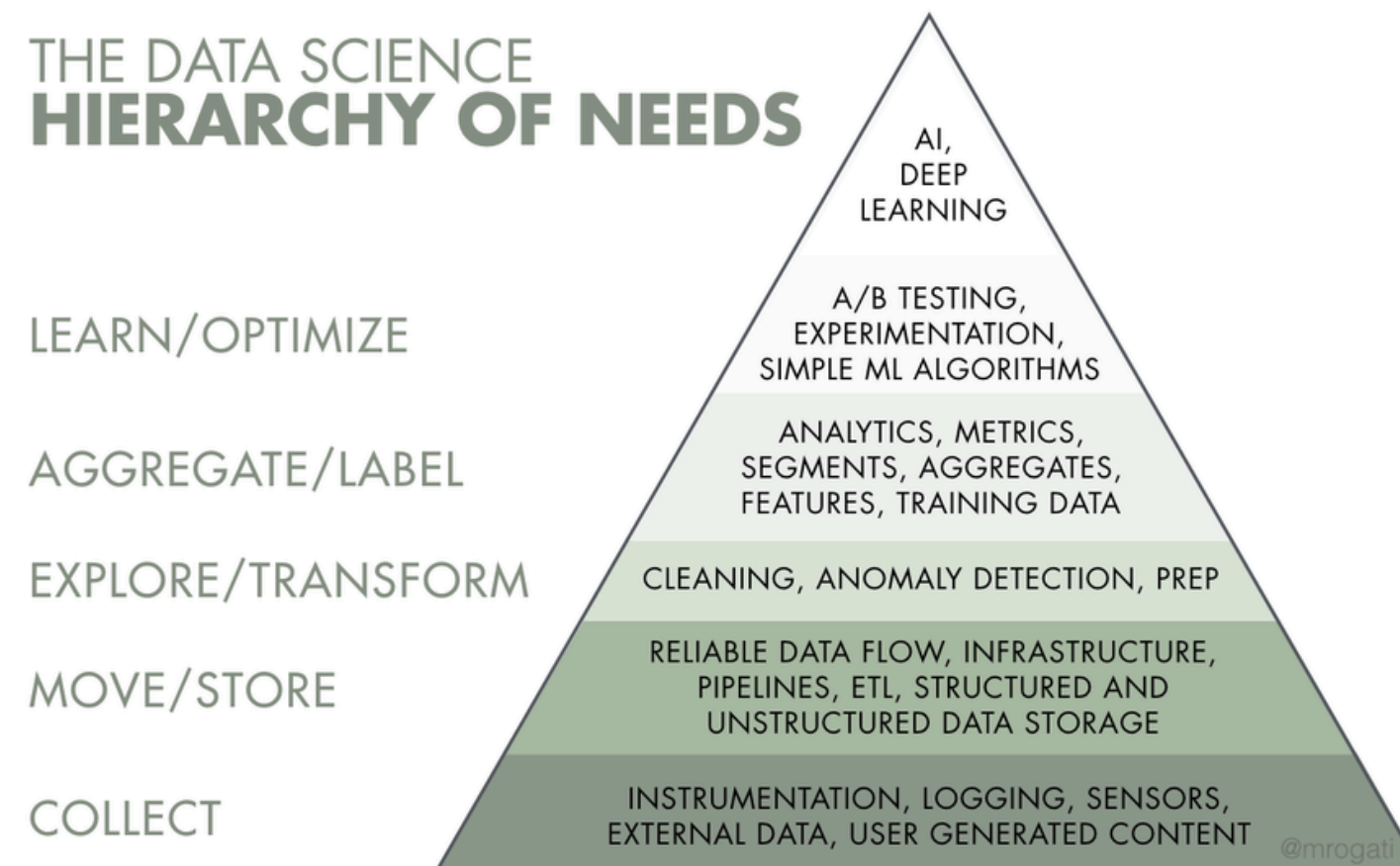
A engenharia de dados é um conjunto de operações com o objetivo de criar interfaces e mecanismos para o fluxo e acesso de informações. Para manter os dados de forma que permaneçam disponíveis e utilizáveis para todos, é necessário contar com especialistas dedicados: os engenheiros de dados.

PAPEL DO ENGENHEIRO DE DADOS

- Projetar e criar pipelines de dados;
- Gerenciar infraestrutura de dados;
- Limpeza, Transformação e Enriquecimento de Dados;
- Modelagem de dados;
- Garantir a qualidade de dados;
- Otimização de desempenho;
- Colaboração com outras áreas.

POR QUE APRENDER ENGENHARIA DE DADOS?

1. É a base da área de dados em geral



POR QUE APRENDER ENGENHARIA DE DADOS?

2. É uma ótima área para quem gosta de aprender sobre assuntos diversos



POR QUE APRENDER ENGENHARIA DE DADOS?

3. Bons salários

Salários de Engenheiro De Dados Especialista (Brasil) ⓘ



Confiança muito alta · 102 salários enviados · Atualizado em 27 de nov. de 2024

Experiência

Todos os anos de experiência



Salário base

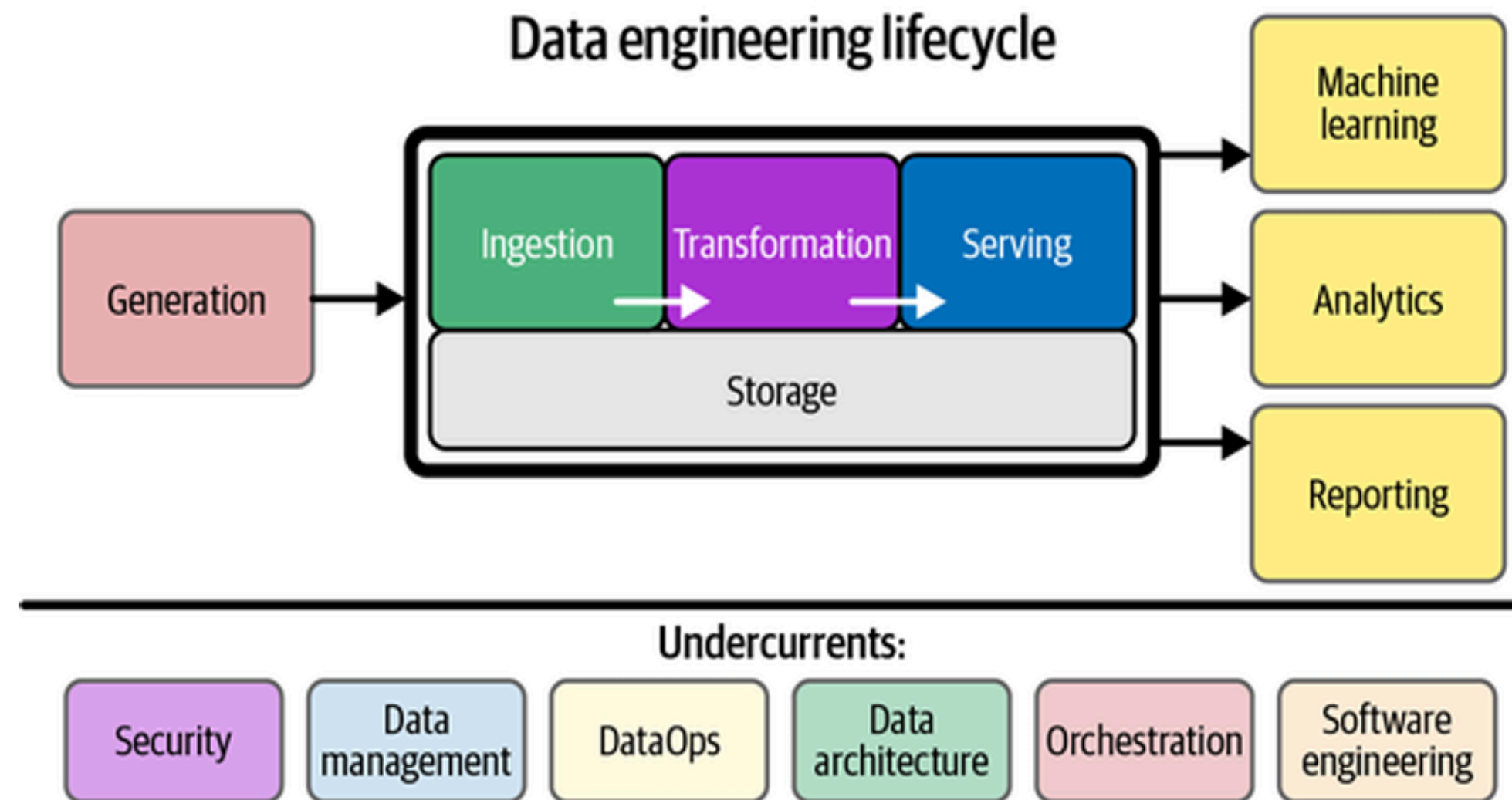
R\$ 11 mil – R\$ 16 mil/mês

R\$ 14 mil/mês Salário base médio

IMPORTÂNCIA DO ENGENHEIRO DE DADOS PARA EMPRESAS

- Disponibiliza dados confiáveis para decisões estratégicas;
- Permite a integração de diferentes fontes de dados;
- Diminuição de custo;
- Inovação;
- Governança e conformidade.

CICLO DE VIDA DA ENGENHARIA DE DADOS



ETL

ETL é um processo fundamental na engenharia de dados que consiste em extrair dados de várias fontes, transformá-los para atender às necessidades do negócio e carregá-los em um destino final, como um Data Warehouse, Data Lake ou Data Lakehouse. O objetivo do ETL é disponibilizar dados limpos, padronizados e prontos para análises, relatórios e outras aplicações de inteligência de negócios (BI).

EXTRACT

Extração de fontes diversas, entre eles dados estruturados, semi-estruturados e não estruturados.



TRANSFORM

Após a extração, os dados brutos passam por transformações para atender aos requisitos do negócio, como retirar duplicidades.

Original DataFrame

	Name	Age	City
0 →	Alice	25	NY
1	Bob	30	LA
2 →	Alice	25	NY
3	David	40	Chicago

Modified DataFrame (no duplicates)

	Name	Age	City
0	Alice	25	NY
1	Bob	30	LA
3	David	40	Chicago

Removed Duplicated Rows

LOAD

A etapa final consiste em carregar os dados transformados em um sistema de destino.



Azure Data Lake Storage Gen2



Azure SQL



Amazon S3

ETL VS ELT

ETL: Os dados são transformados (limpos, agregados, padronizados) em um local intermediário antes da carga;

ELT: A transformação é realizada dentro do destino, usando os recursos computacionais dele.

OLTP VS OLAP

OLTP (Online Transaction Processing) e **OLAP** (Online Analytical Processing) são dois sistemas de processamento de dados que atendem a finalidades diferentes dentro de uma organização. Enquanto o OLTP é focado em operações transacionais do dia a dia, o OLAP é voltado para análise de dados históricos e suporte à decisão.

O GRANDE E TEMIDO T DE TRANSFORMAÇÃO

- Grande diversidade de dados, o que pode incluir dados estruturados, semiestruturados e não estruturados;
- Os dados podem estar em diferentes formatos (CSV, JSON, Parquet, etc.), com qualidade inconsistente, erros, e problemas de integridade, como valores ausentes ou duplicados.

TRANSFORMAÇÃO

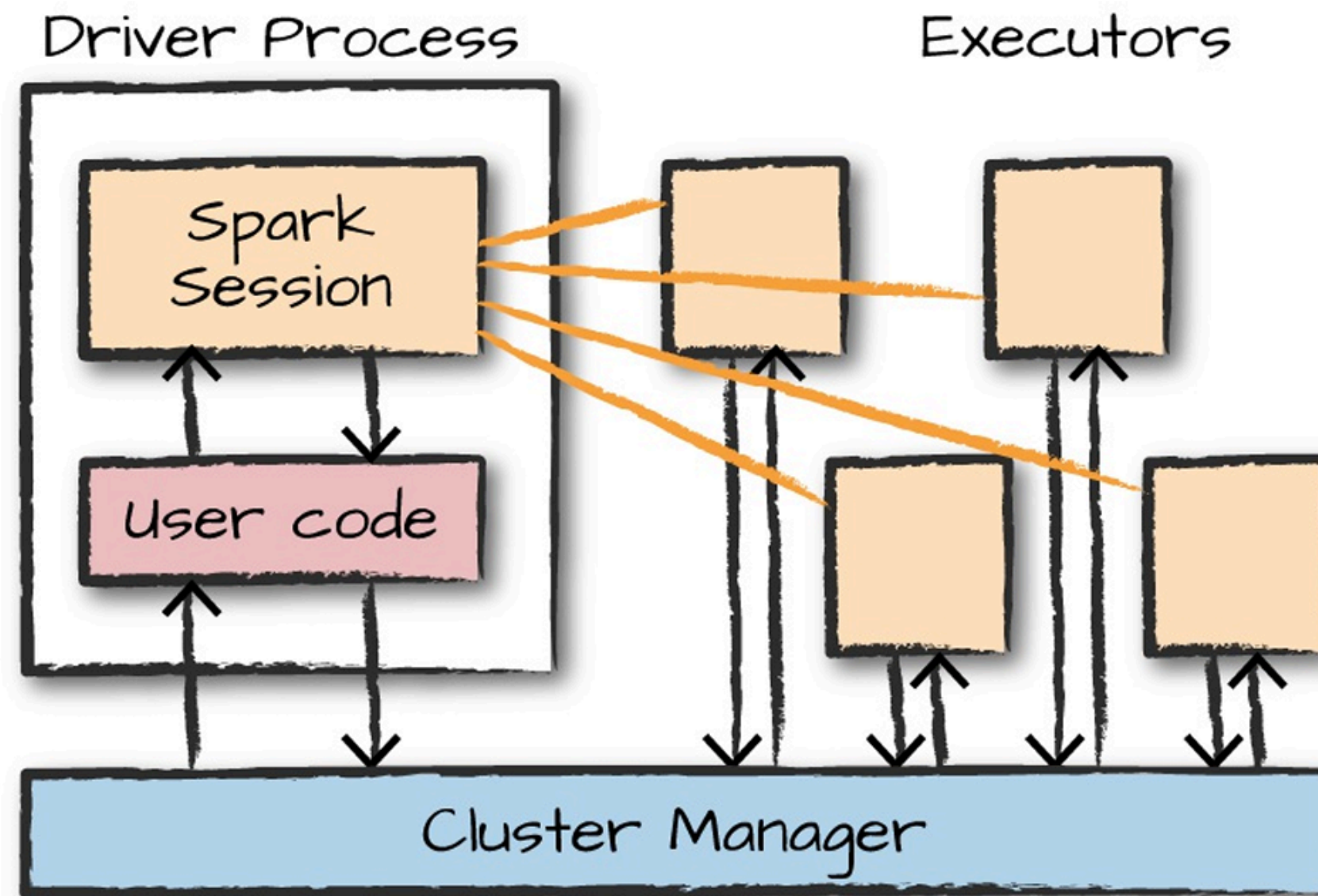
- Para realizar as transformação em um grande volume de dados, existem várias ferramentas famosas, como:



O QUE É O APACHE SPARK

O Apache Spark é um sistema de processamento de dados distribuído de código aberto, projetado para ser altamente rápido e escalável. O Spark é amplamente usado para processar grandes volumes de dados de maneira rápida e eficiente, e é especialmente popular no contexto de big data.

APACHE SPARK



VANTAGENS DO APACHE SPARK

- Suporta grandes volumes de dados;
- É mantido por uma comunidade ativa, com grandes empresas mantendo a ferramenta;
- Plataformas em nuvem utilizam Spark em seus serviços de dados;
- Facilidade de uso.

FACILIDADE NO USO DO APACHE SPARK

SQL

sql

```
SELECT nome, idade FROM pessoas;
```

PySpark

python

```
df.select("nome", "idade")
```


FACILIDADE NO USO DO APACHE SPARK

SQL

sql

```
SELECT * FROM pessoas WHERE idade > 30;
```

PySpark

python

```
df.filter(df.idade > 30)
```

FACILIDADE NO USO DO APACHE SPARK

SQL

sql

```
SELECT * FROM pessoas ORDER BY idade DESC;
```

PySpark

python

```
df.orderBy(df.idade.desc())
```

O QUE É O DATABRICKS?

O Databricks é uma plataforma de análise de dados e engenharia de dados baseada na nuvem, construída sobre o Apache Spark. Ele oferece uma série de ferramentas e serviços para facilitar a criação, análise e implementação de pipelines de dados em larga escala.

VAMOS FAZER UMA CONTA NO DATABRICKS



Try Databricks free

Test-drive the full Databricks platform free 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.
- ✓ Enjoy serverless credits 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.



[Privacy Notice](#) [Terms of Use](#) [Modern Slavery Statement](#) [California Privacy](#) [Your Privacy Choices](#)

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

What do you want to build and run with Databricks? (Optional)

By clicking "Continue," you agree to Databricks processing your personal data in accordance with its [Privacy Notice](#).

Continue

VAMOS FAZER UMA CONTA NO DATABRICKS



Try Databricks free

Test-drive the full Databricks platform free 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.
- ✓ Enjoy serverless credits 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.



Mercedes-Benz

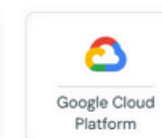
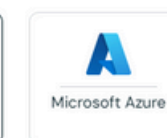


[Privacy Notice](#) [Terms of Use](#) [Modern Slavery Statement](#) [California Privacy](#) [Your Privacy Choices](#)

How will you be using Databricks? 2/2

Professional use

Pick your cloud provider. You'll need admin access to your cloud account to get started.



Enjoy \$400 in credits during your 14-day AWS trial. Trial ends when credits expire.

By clicking "Continue," you agree to Databricks' [Terms of Service](#).

Continue

Personal use

Community Edition is a limited, single node version of Databricks for personal or educational use.

By clicking "Get started with Community Edition," you agree to Databricks' [Terms of Service](#).

Get started with Community Edition

ATIVIDADE



RECAPITULANDO PARA FINALIZAR...

- O que é Engenharia de Dados e o que o Engenheiro de Dados faz?
- O que é pipeline de dados?
- Quais os principais pontos que você deve levar hoje sobre Engenharia de Dados?

MUITO OBRIGADA!

REFERÊNCIAS

- <https://www.dataquest.io/blog/why-learn-data-engineering/>
- https://www.glassdoor.com.br/Sal%C3%A1rios/engenheiro-de-dados-especialista-sal%C3%A1rio-SRCH_KO0,32.htm
- Fundamentos da Engenharia de Dados - Joe Reis e Matt Housley
- Spark: The Definitive Guide - Bill Chambers e Matei Zaharia (Autor)

