

## AULA 2 - INTRODUÇÃO A ENGENHARIA DE DADOS

**#BootcampMIA2022 #SomosMIA**



## Quem somos?



## Carol Oliveira

### Data Scientist na NTT DATA

- Bacharel em Engenharia da Computação na Universidade Veiga de Almeida (UVA)
- Embaixadora do Women in Data Science RJ e co-fundadora da MIA



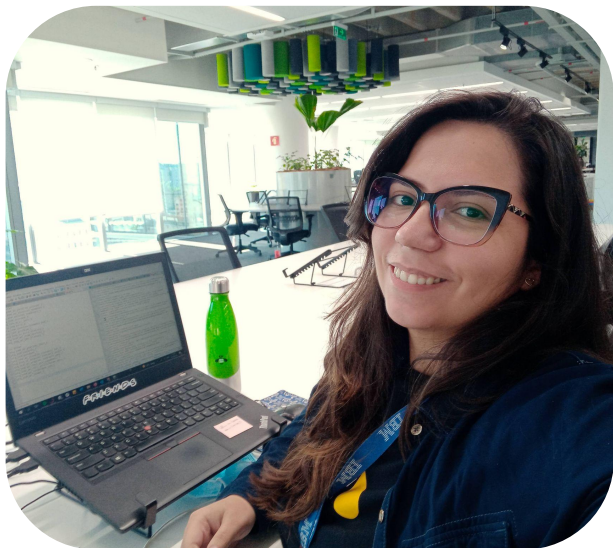
Coliverfelt



oliverfelt



## Quem somos?



## Iris Herdy

### Engenheira de Dados na IBM

- Graduada em Sistemas de Computação pela Universidade Federal Fluminense
- Co-fundadora da MIA



irisherdy



irisherdy



## Uma inspiração...

"A imaginação é a faculdade da descoberta, predominantemente. É ela que penetra nos mundos invisíveis que nos rodeiam, nos mundos da ciência".

**(Ada Lovelace)**



## Sumário

- O que é Engenharia de Dados?
- O papel da pessoa Engenheira de Dados
- Lidando com dados
- Ferramentas e Tecnologias
- Vagas

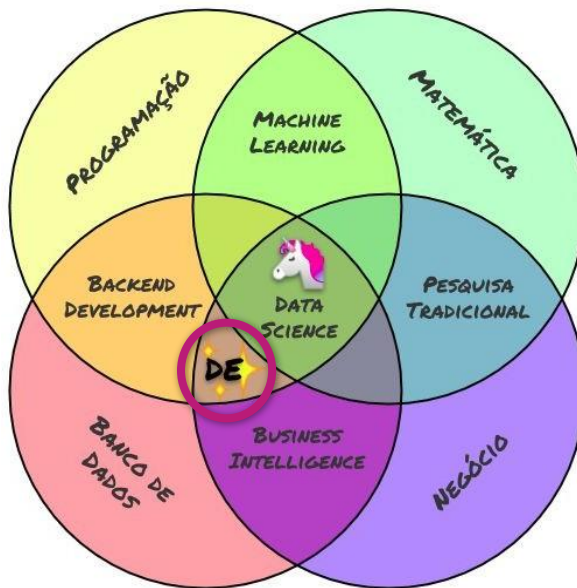
# O que é Engenharia de Dados?



# O que é Engenharia de Dados?

- Vamos aprofundar esse papel em um time de dados:

- Data Scientist
- **Data Engineering**
- Data Analyst
- Machine Learning Engineering
- Chief Data Officer





# O que é Engenharia de Dados?

- Combinação de conhecimentos e práticas para implementar mecanismos de **coleta, tratamento, processamento e armazenamento** de dados para deixá-los disponíveis para uma determinada finalidade.



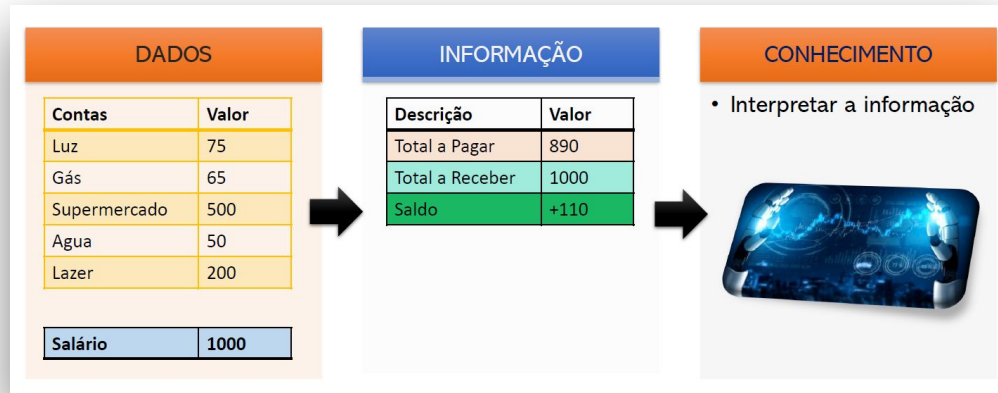




# O que é Engenharia de Dados?

## ■ Por quê dados?

Dados >> Informação >> Conhecimento





# O que é Engenharia de Dados?

## ■ Tipos de dados

- Estruturados
- Semi Estruturados
- Não estruturados





## O que é Engenharia de Dados?

- Estruturados

Nome	CPF	Endereço	Telefone
Marcela Freitas	11111	Rua A, nº 1	101010
João Augusto	22222	Rua B, nº 2	202020
Pablo Silva	33333	Rua C, nº 3	303030
André Mendes	44444	Rua D, nº 4	404040
Juliana Freitas	55555	Rua E, nº 5	505050



# O que é Engenharia de Dados?

- Semi-estruturados

## XML

vs.

## JSON

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <endereco>
3   <cep>31270901</cep>
4   <city>Belo Horizonte</city>
5   <neighborhood>Pampulha</neighborhood>
6   <service>correios</service>
7   <state>MG</state>
8   <street>Av. Presidente Antônio Carlos, 6627</street>
9 </endereco>
  
```

```

1 {
2   "endereco": {
3     "cep": "31270901",
4     "city": "Belo Horizonte",
5     "neighborhood": "Pampulha",
6     "service": "correios",
7     "state": "MG",
8     "street": "Av. Presidente Antônio Carlos, 6627"
9   }
10 }
  
```



## O que é Engenharia de Dados?

- Não estruturados





# O que é Engenharia de Dados?

## ■ Qualidade de dados

- Singularidade
- Completude
- Pontualidade
- Validade
- Acurácia
- Consistência

# O papel da pessoa Engenheira de Dados



## O papel da pessoa Engenheira de Dados

A pessoa na função de **engenheira de dados** têm como tarefas **gerenciar** e **organizar dados**, enquanto também estão atentos a tendências ou inconsistências que afetarão as metas de negócios.

É uma posição técnica, exigindo experiência e habilidades em áreas como: **Programação** e **Ciência da Computação**.







## O papel da pessoa Engenheira de Dados

Além dos hard skills, as pessoas **engenheiras de dados** também precisam de **soft skills** para **comunicar** tendências de dados a outras pessoas da organização e **ajudar** os times a usarem os dados coletados.





## O papel da pessoa Engenheira de Dados

- **Funções principais dentro de um time de dados:**
  - Generalista
  - Centrada em pipeline
  - Centrada no banco de dados





## O papel da pessoa Engenheira de Dados

- **Funções principais dentro de um time de dados:**
- **Generalista**
- Centrada em pipeline
- Centrada no banco de dados





## O papel da pessoa Engenheira de Dados

- **Funções principais dentro de um time de dados:**

- Generalista
- **Centrada em pipeline**
- Centrada no banco de dados





## O papel da pessoa Engenheira de Dados

- **Funções principais dentro de um time de dados:**
  - Generalista
  - Centrada em pipeline
  - **Centrada no banco de dados**



# Lidando com dados



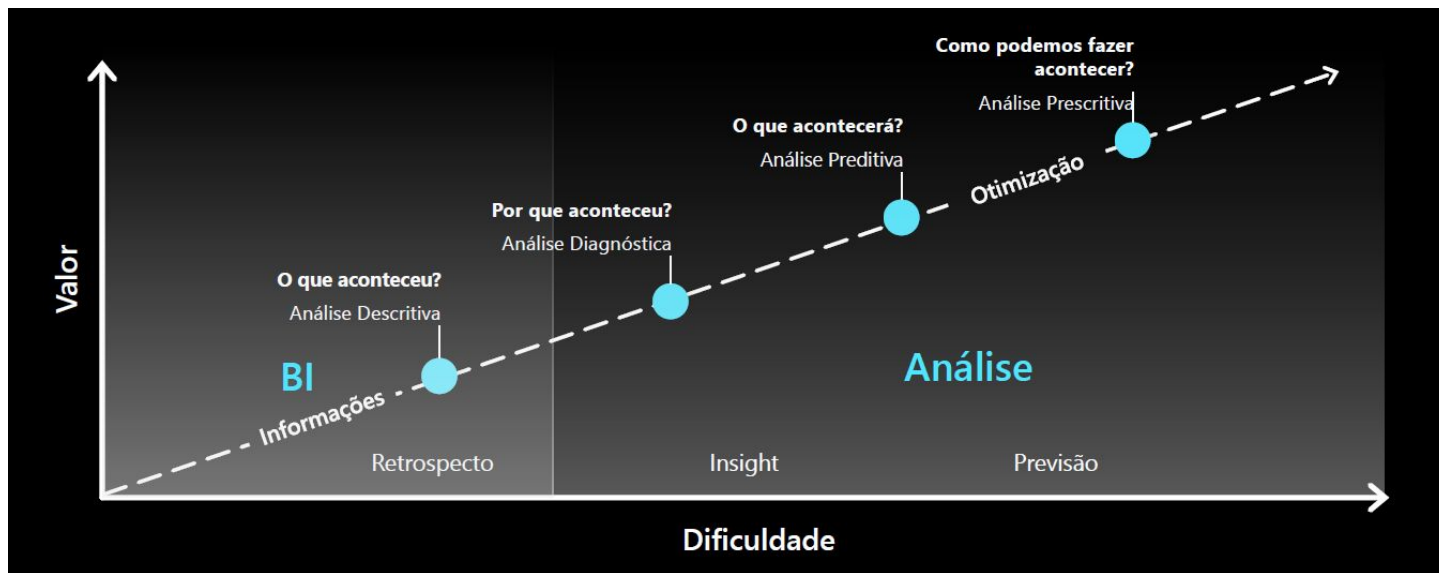
## ■ Do ponto de vista de negócios





# Lidando com dados

## ■ Do ponto de vista de negócios








# Lidando com dados


## ■ Bussiness Inteligence - BI

- É um conjunto de processos e metodologias, implementadas por meio de ferramentas de software, para obter informação e conhecimento útil para a tomada de decisão.


## Business Intelligence



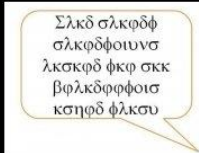
What my friends think I do




What my mom thinks I do.




What society thinks I do



What my coworkers think I do.



What I think I do.



What I actually do.

picloco.com



## Lidando com dados

### ■ Big Data

Big Data é a área do conhecimento que estuda como tratar, analisar e obter informações a partir de grandes conjuntos de dados.

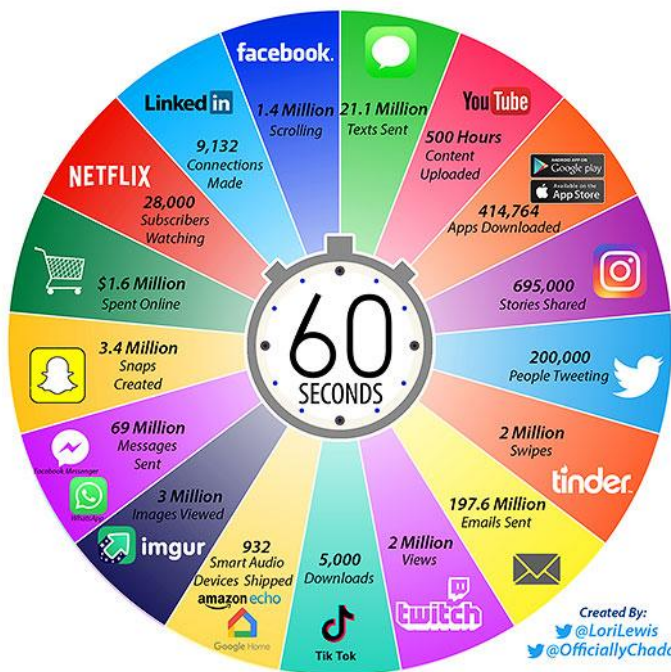




# Lidando com dados

## ■ Big Data

### 2021 *This Is What Happens In An Internet Minute*





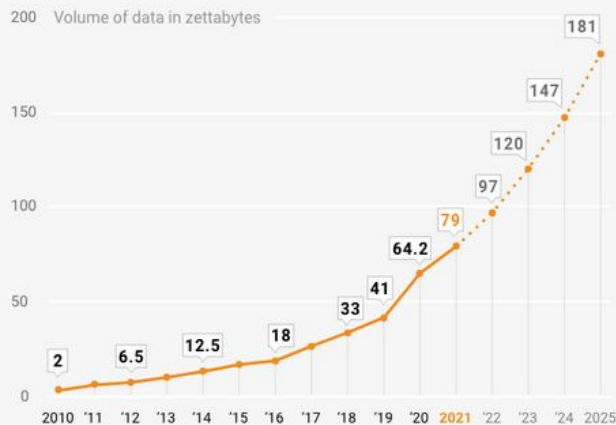
## Lidando com dados

### ■ Big Data

## Volume of data created, captured, copied, and consumed worldwide



The volume of data generated, consumed, copied, and stored is projected to exceed 180 zettabytes by 2025



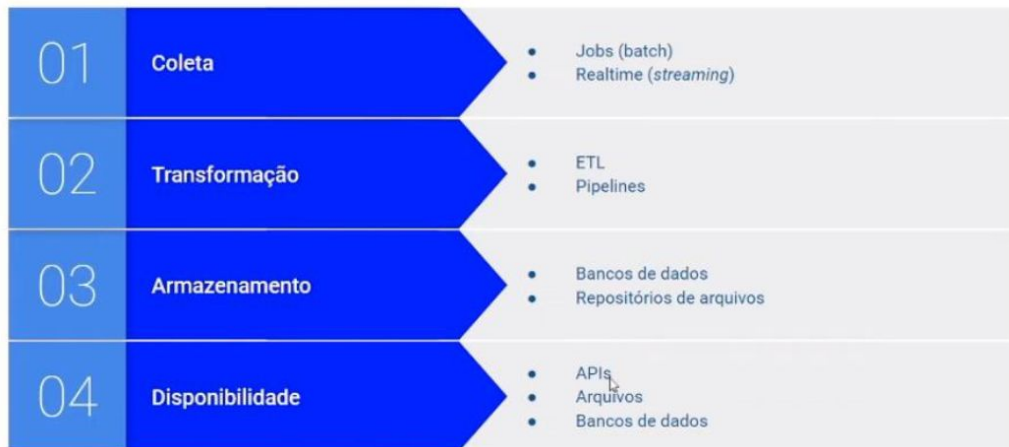
Source: statista.com

 firstsiteguide.com



# Lidando com dados

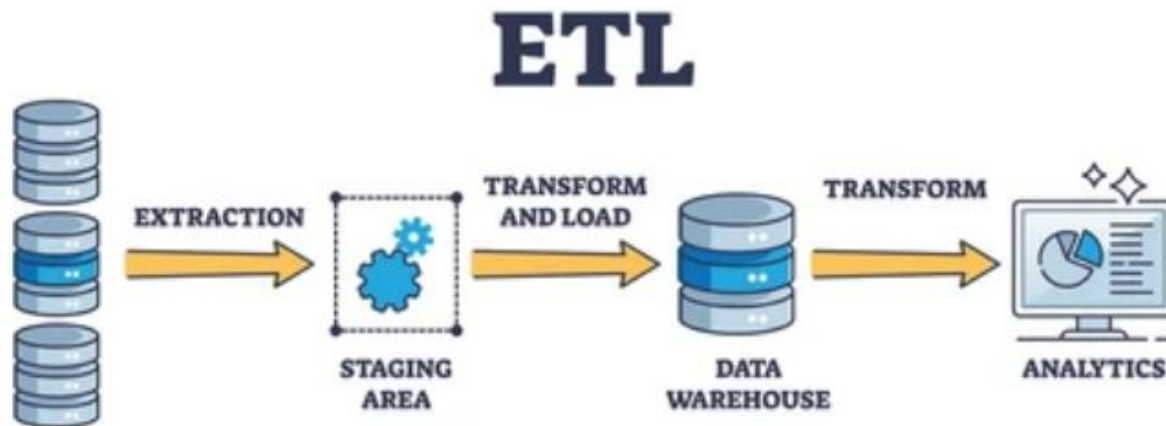
## ■ Fluxo de Dados





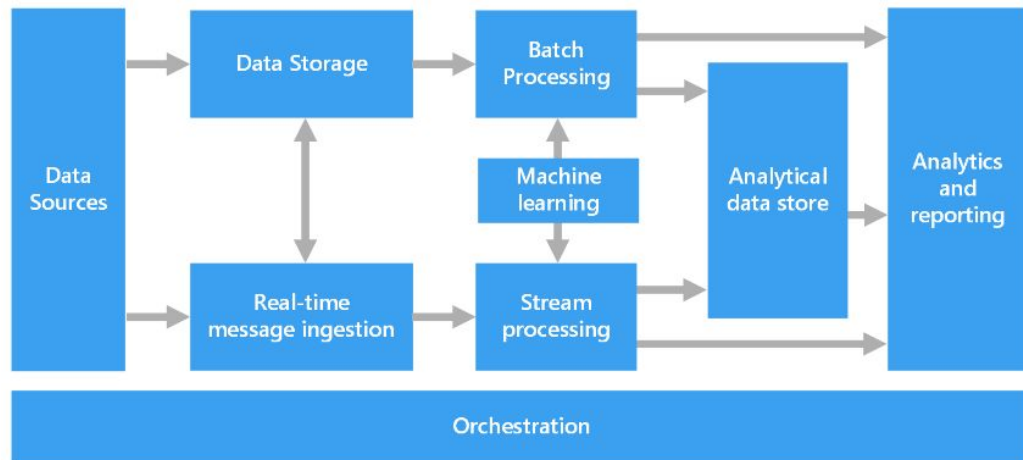
# Lidando com dados

- ETL





## ■ Arquitetura

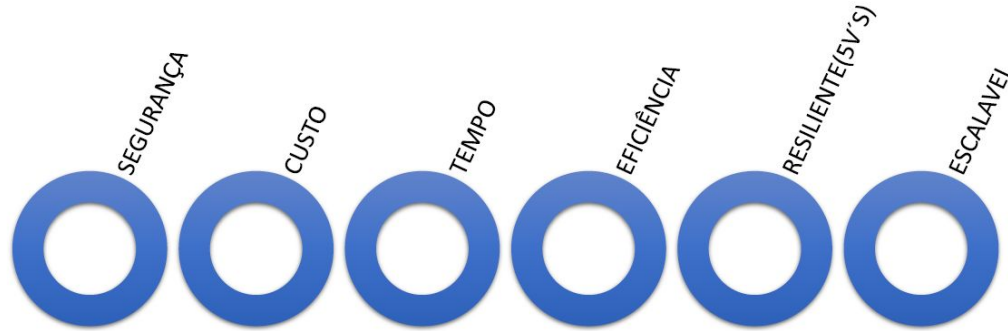




## Lidando com dados

- Qual é a melhor arquitetura de dados?

Aquela que melhor resolve o problema!!!

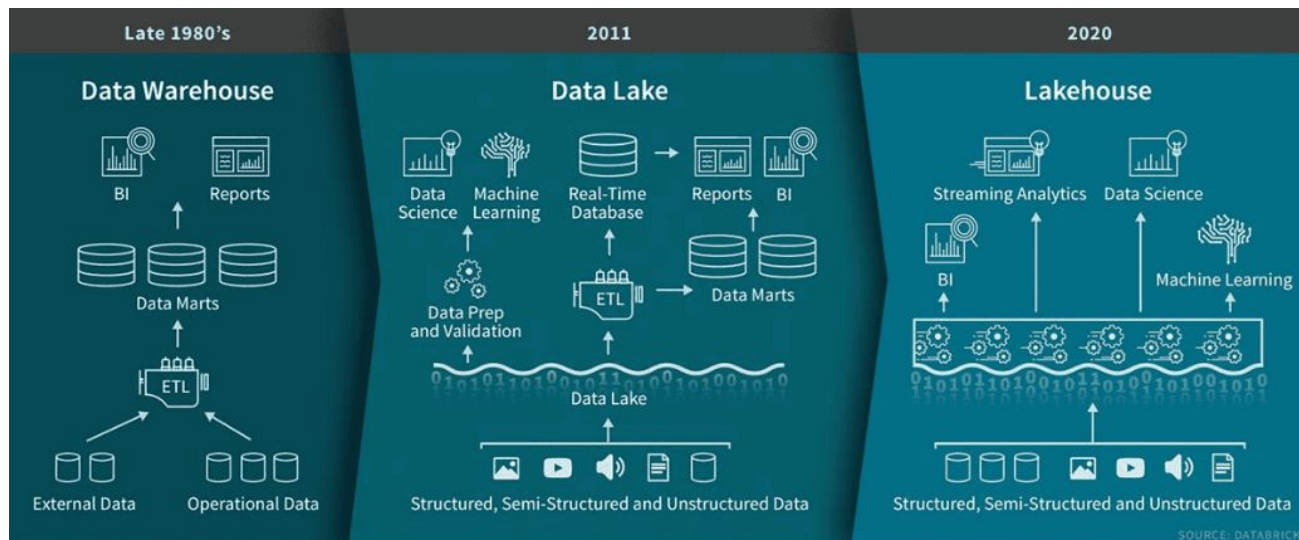






# Repositórios

## ■ Data Lake X Data Warehouse X Data Lakehouse





## ■ Data Lake X Data Warehouse X Data Lakehouse

Parameters	Data Lake	Data Warehouse	Data Lakehouse
Purpose of Data	For ML and AI workloads ( Purpose of the data is not yet determined)	For Data Analytics or Business Intelligence ( The data is currently in use)	Can be used for ML/AI workload and Data Analytics/BI needs
Type of Data	Unstructured Data	Structured Data	Unstructured and Structured Data
Users	Data scientists, data engineers, data	Business professionals	Business professionals and data teams
Data Quality	Raw Data, Low Quality and Not Reliable	Highly curated data, reliable	Raw and curated data, high quality with in-built data governance
ACID Compliance	Non-ACID compliance: updates and deletes are complex operations	ACID-compliant : guarantee the highest levels of integrity	ACID-compliant to ensure consistency as many sources concurrently read/write data
Storage	Cost-effective, rapid and flexible	Costly and time-consuming	Cost-effective, rapid and flexible
Schema	Schema on read	Schema on write	Schema enforcement

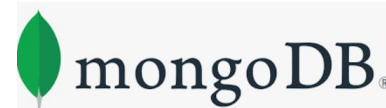
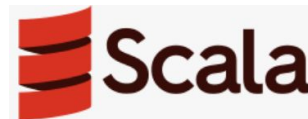
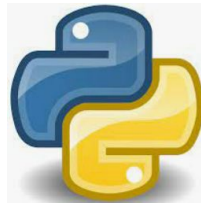
# Ferramentas e Tecnologias



# Ferramentas e Tecnologias

## ■ Apps, Frameworks e Linguagens

- Hadoop
- Databricks, AWS Kinesis
- Spark, Scala; Python
- Kafka
- Google Big Query; AWS RDS; Hive; MongoDB; MySQL
- Redis



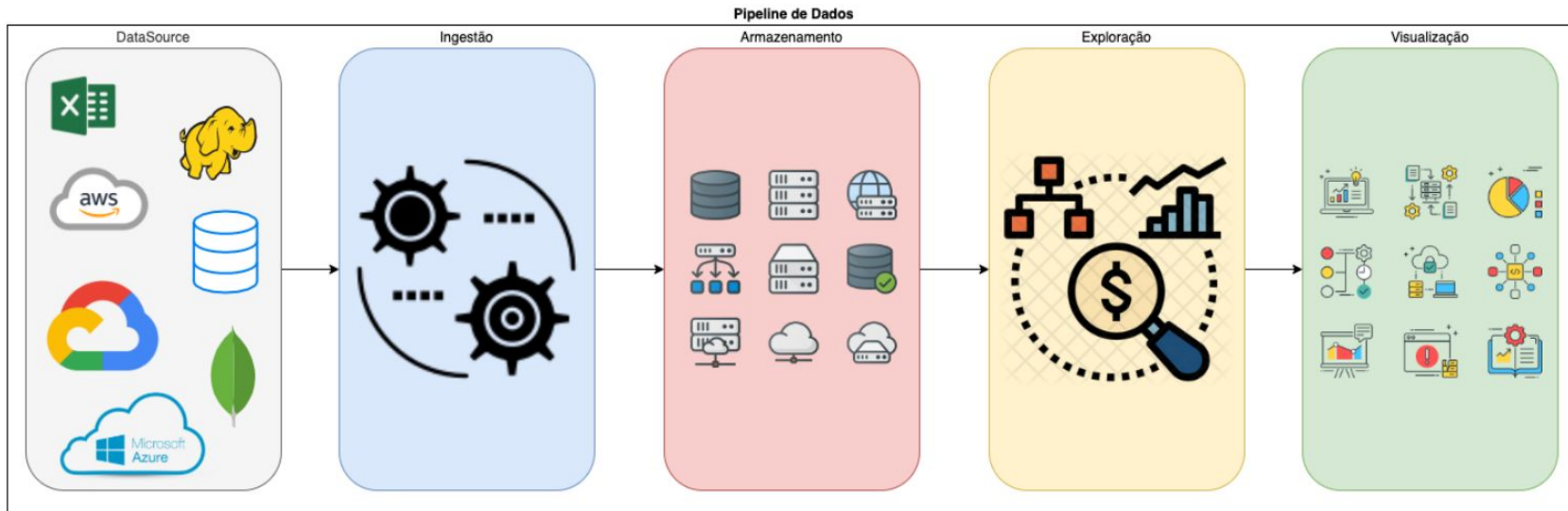


# Ferramentas e Tecnologias

Fonte de dados	Ingestão	Armazenamento	Exploração	Visualização
MySQL 	Apache Kafka 	Redis 	Python 	Power BI 
Hive 	Apache Flume 	Amazon S3 	Spark 	Tableau 
Mongo 	Apache Airflow 	Hadoop 	Scala 	Grafana 
API	Apache Sqoop 	SQL/No SQL BD	SQL/No SQL	D3.js  Data-Driven Documents



# Ferramentas e Tecnologias



<https://blog.dsbrigade.com/pipeline-de-dados-com-servicos-aws/>



## Ferramentas e Tecnologias

### ■ Banco de dados

Segundo Korth , um banco de dados “é uma coleção de dados inter relacionados, representando informações sobre um domínio específico”, ou seja, sempre que for possível agrupar informações que se relacionam e tratam de um mesmo assunto, posso dizer que tenho um banco de dados.





# Ferramentas e Tecnologias

## ■ Banco de dados Relacionais x Não Relacionais



ORACLE



Microsoft®  
SQL Server®







# Ferramentas e Tecnologias

## ■ Banco de dados Relacionais x **Não Relacionais**

Chave / Valor



Orientado a documentos



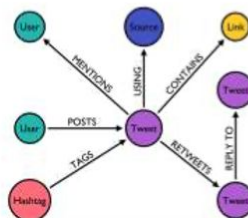
Família de Colunas

Column Family: User\_URLs

	Super Column Name	Column Name	
98725	http://techcrunch.com/2010/07/09/...	http://cnn.com/world/...	...
	8fb7f240-8b91-11d1	78f364e0-8b91-11d1	cf128360-8b91-11d1
	...	...	...

Column Value

Grafos





# Ferramentas e Tecnologias

## ■ SQL

### LANGUAGE STATEMENTS

#### DML

- ☐ SELECT
- ☐ INSERT
- ☐ UPDATE
- ☐ DELETE

#### DDL

- ☐ CREATE
- ☐ ALTER
- ☐ DROP
- ☐ TRUNCATE

#### DCL

- ☐ GRANT
- ☐ REVOKE
- ☐ DENY

#### TCL

- ☐ BEGIN TRANSACTION
- ☐ COMMIT
- ☐ SAVE TRANSACTION
- ☐ ROLLBACK



## Ferramentas e Tecnologias

O CRUD é um acrônimo para as 4 operações básicas de um banco de dados.

C

• **CREATE (CRIAR)**

R

• **READ (LER, SELECT)**

U

• **UPDATE (ATUALIZAR)**

D

• **DELETE (APAGAR)**

# Vagas



# Vagas

## ■ Como encontrar?

- LinkedIn
- Vagas.com
- CIEE
- Programathor
- Pluo.jobs



# Para praticar...

- ▶ Playlist **Curso Básico de SQL para Análise de Dados** do Canal **Programação Dinâmica**



**Curso Básico de SQL para Análise de Dados**

13 vídeos • 13.078 visualizações • Última atualização em 15 de jul. de 2022

Curso Aberto de SQL para Análise de Dados do canal Programação Dinâmica.

Esse curso se destina a quem deseja aprender a extrair informações de bancos de dados utilizando SQL. Utilizaremos como ambiente de aprendizado o Big Query e como fontes de dados, algumas entre as inúmeras bases de dados disponibilizadas, tratadas e mantidas pelo projeto Base dos Dados.

- 1 O que você precisa saber para começar a fazer consultas com SQL? | SQL PARA ANÁLISE DE DADOS EP: 0  
10:33  
Programação Dinâmica
- 2 Consultas SQL na Prática usando SELECT, DISTINCT, WHERE e LIMIT | SQL Para Análise de Dados EP: 1  
15:13  
Programação Dinâmica
- 3 Como usar ORDER BY em consultas SQL e IN, BETWEEN, LIKE e NOT | SQL para Análise de Dados EP: 2  
13:30  
Programação Dinâmica
- 4 Para que serve o COUNT em SQL + agregação com MAX, MIN e SUM | SQL para Análise de Dados  
15:40  
Programação Dinâmica
- 5 Como usar GROUP BY, HAVING e CASE em consultas SQL na prática | SQL para Análise de Dados EP:4  
20:38  
Programação Dinâmica
- 6 Por que os DADOS são armazenados em várias TABELAS em um de Bancos de Dados? Normalização e SQL  
15:50  
Programação Dinâmica



# Minha história em Engenharia de Dados



## Dúvidas?

- **Iris Herdy**
  - **LinkedIn:**  
[linkedin.com/in/irisherdy](https://www.linkedin.com/in/irisherdy)
  - **Twitter:** @irisherdy
- **Carol**
  - **LinkedIn:**  
[linkedin.com/in/oliverfelt](https://www.linkedin.com/in/oliverfelt)







## Nossos contatos



**mulheres.em.ia@gmail.com**



**mulheres-em-ia**



**@mulheres.em.ia**



**@mulheres.em.ia**



**@MulheresemInteligenciaArtificial**



**Canal: Mulheres em IA**

## Linktree

<https://linktr.ee/mulheres.em.ia>

## Grupo Telegram para Mulheres

[https://t.me/mulheres\\_em\\_ia](https://t.me/mulheres_em_ia)



# Muito obrigada!

Dúvidas? Podem nos procurar! 🙄

# Por hoje, é isso!

No próximo sábado falaremos  
sobre **Análise Exploratória** \o/



# Vamos preencher o formulário de feedback???



<http://bit.ly/bootcamp-ds-sp-feedback-19>