



RELATÓRIO PROJETO FINAL

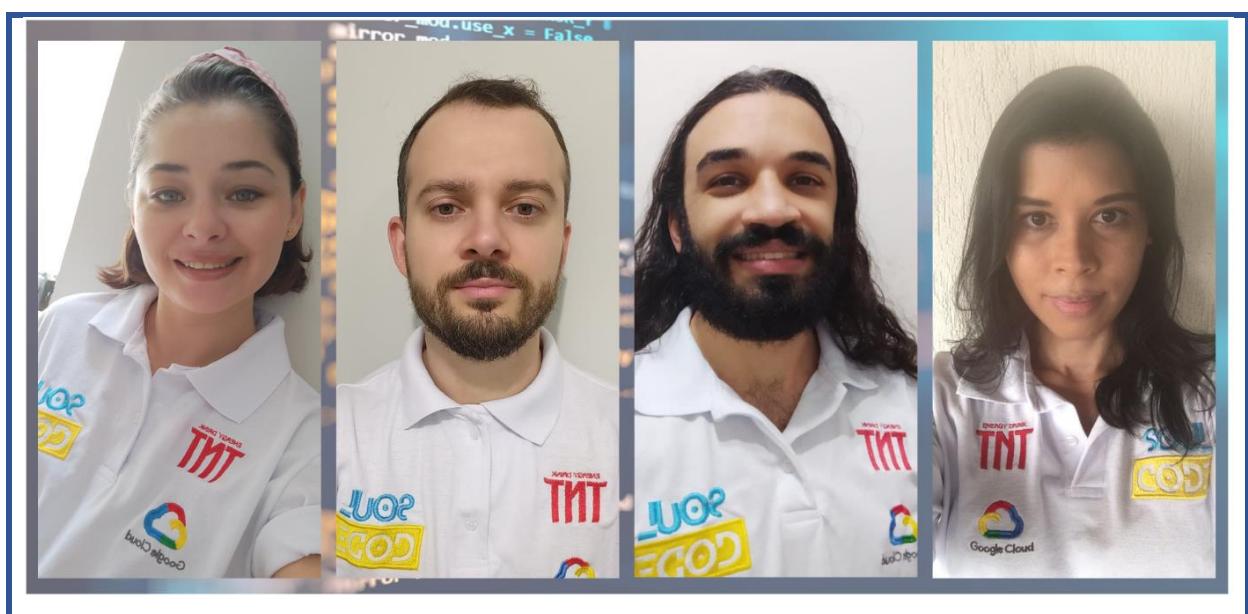
TEMA: Mercado de Trabalho

BASE DE DADOS: Stack Overflow, Dados SiSU e Google Trends

TURMA: BC12

ORIENTADOR: Prof. Bismark William Araújo

GRUPO: Daiane Silva, Felipe Rinaldini, Talita Dwyer e José Henrique Teles



Abril de 2022

LISTA DE FIGURAS

Figura 1 – Demonstrativo de alguns dados no Dataset Stack Overflow	6
Figura 2 – Demonstrativo de alguns dados no Dataset SISU	6
Figura 3 – Demonstrativo de alguns dados no Dataset Google Trends.....	7
Figura 4 – Fluxo de ferramentas mais utilizadas no escopo do projeto	7
Figura 5 – Enviando Dataframe Stack Overflow tratado para bucket	9
Figura 6 – Enviando Dataframe Google Trends tratado para bucket.....	9
Figura 7 – Armazenamento dos dados originais e tratados no Mongo DB	10
Figura 8 – Armazenamento dos Data Frames no MySQL	11
Figura 9 – Abrindo tabela no terminal da GCP cloud SHELL utilizando MySQL	11
Figura 9 – Carregando arquivo original da bucket GCP	11
Figura 10 – Limpeza e Transformação utilizando a biblioteca pandas	12
Figura 11 – Carregando arquivo tratado para análise em pyspark	12
Figura 12 – Criando schema e estrutura do DataFrame para análise em pyspark.....	12
Figura 13 – Criando tabela em sparkSQL a parti do DataFrame spark	13
Figura 14 – Análise com sparkSQL	13
Figura 15 – Gráfico demonstrativo do fluxo	14
Figura 16 – Leitura do Dataframe Stake para análises no Big Query	15
Figura 17 – Leitura do Dataframe Sisu para análises no Big Query	15
Figura 18 – Leitura do Dataframe Stake para análises no Big Query	16
Figura 19 – Planilha de custos do Google Cloud Platform.....	17
Figura 20 – Gráfico de custos do usuário 'A'	18
Figura 21 – Gráfico de custos do usuário 'B'	18
Figura 22 – Gráfico de custos do usuário 'C'	19

LISTA DE TABELAS

Tabela 1 - Armazenamento Mongodb Atlas.....	10
---	----

ÍNDICE

1) INTRODUÇÃO	4
2) OBJETIVOS	5
3) DATASETS.....	5
4) FERRAMENTAS UTILIZADAS.....	7
5) FLUXOGRAMA	8
6) PROCESSOS DE ETL	9
9) CUSTOS DE UTILIZAÇÃO DO GOOGLE CLOUD PLATFORM.....	17
REFERÊNCIAS	20
ANEXOS.....	21

1) INTRODUÇÃO

Este relatório tem a finalidade de apresentar e justificar todo o processo de ETL (extração, transformação, carregamento) para análise, considerando o tema mercado de trabalho as bases de dados utilizadas foram Stack Overflow, Dados SiSU e Google Trends para os insights. Abaixo mostra os requisitos obrigatórios e desejáveis para as análises.

Requisitos obrigatórios:

- Obrigatoriamente os datasets devem ter formatos diferentes (CSV / Json / Parquet / Sql / NoSql) e 1 deles obrigatoriamente tem que ser em CSV.
- Operações com Pandas (limpezas, transformações e normalizações)
- Operações usando PySpark com a descrição de cada uma das operações.
- Operações utilizando o SparkSQL com a descrição de cada umas das operações.
- Os datasets utilizados podem ser em língua estrangeira, mas devem ao final terem seus dados/colunas exibidos na língua PT-BR
- os datasets devem ser salvos e operados em armazenamento cloud obrigatoriamente dentro da plataforma GCP (não pode ser usado Google drive ou armazenamento alheio ao google)
- os dados tratados devem ser armazenados também em GCP, mas obrigatoriamente em um datalake (Gstorage), DW(BigQuery) ou em ambos.
- Os datasets originais devem ser armazenados em MySql
- Os Dataframe(s) resultante(s) deve(m) estar em uma coleção do mongoDb atlas (informar a key de acesso ao cluster) e preferencialmente criar o usuário (soulcode) e senha (a1b2c3) no cluster
- Deve ser feito análises dentro do Big Query utilizando a linguagem padrão SQL com a descrição das consultas feitas.
- Deve ser criado no datastudio um dashboard para exibição gráfica dos dados tratados trazendo insights importantes

- É deve ser demonstrado em um workflow simples (gráfico) as etapas de ETL com suas respectivas ferramentas.

Requisitos desejáveis:

- Implementar captura e ingestão de dados por meio de uma PIPELINE com modelo criado em apache beam usando o dataflow para o work
- Utilizar o dataflow com algum modelo pré-definido
- Criar plotagens usando pandas para alguns insights durante o processo de Transformação
- Por meio de uma PIPELINE fazer o carregamento dos dados normalizados diretamente para um DW ou DataLake ou ambos
- Montar um relatório completo com os insights que justificam todo o processo de ETL utilizado
- Levantar custos com a utilização do google cloud no período do projeto e possíveis otimizações de custo

2) OBJETIVOS

- Aplicar os conceitos vistos durante o curso para tratar, organizar e modelar os dados.
- Selecionar no mínimo dois datasets dentro do tema “Mercado de Trabalho”
- Gerar insights a partir dos dados analisados
- Justificar o processo de ETL (Extração, Transformação, Carregamento)
- Criar um dashboard interativo no Data Studio para exibição dos dados tratados.

3) DATASETS

Stack Overflow:

Questionário do desenvolvedor do Stack Overflow (maio de 2021), que tem como objetivo melhorar a plataforma e fortalecer a comunidade. Nele os

desenvolvedores contam como aprenderam a programar e como estudam para continuar evoluindo, dentre diversas outras informações.

Os principais dados avaliados no Dataset Stack Overflow: Área de Atuação, Situação Empregatícia, País, Escolaridade, Ferramentas de Trabalho mais utilizadas.

Figura 1 – Demonstrativo de alguns dados no Dataset Stack Overflow

	Id	AreaAtuacao	Emprego	País	Escolaridade	AnosProg	AnosProgProf	LingJaTrab
1	desenvolvedor profissional	contrato independente, freelancer ou autonomo	Eslovaquia	ensino medio		NaN	NaN	C++;HTML/CSS;JavaScript;Objective-C;PHP;Swift
2	estudante aprendendo a programar	estudante tempo integral	Holanda	graduacao completa	7.0		NaN	JavaScript;Python

Autoria própria, 2022

Dados SiSU:

Os dados do SiSU (Sistema de seleção unificada) permitem que quem fez o ENEM se inscreva para concorrer a vagas em instituições públicas de ensino superior.

Os principais dados avaliados no dataset SISU: Nome e Estado da Instituição de Ensino, Nome do Curso, 1^a ou 2^a Opção, Aprovação, Situação da Matrícula.

Figura 2 – Demonstrativo de alguns dados no Dataset SISU

Unnamed: 0	SIGLA_IES	UF_IES	NOME_CURSO	CPF	INSCRICAO_ENEM	OPCAO	APROVADO	MATRICULA
0	UFPE	PE	TURISMO	058XXXXXX28	19XXXXXXXX84	1.0	N	PENDENTE
1	UFRJ	RJ	PSICOLOGIA	168XXXXXX11	19XXXXXXXX26	2.0	N	PENDENTE
2	UTFPR	PR	CIÊNCIA DA COMPUTAÇÃO	112XXXXXX45	19XXXXXXXX22	2.0	N	PENDENTE
3	UFSM	RS	ODONTOLOGIA	089XXXXXX09	19XXXXXXXX46	2.0	N	PENDENTE
4	UFG	GO	GEOGRAFIA	068XXXXXX41	19XXXXXXXX69	1.0	N	PENDENTE

Autoria própria, 2022

Google Trends:

Site do Google que analisa a popularidade de palavras usadas em pesquisas, em diversas regiões e idiomas diferentes. Foram gerados 3 arquivos.xls gerados para análise:

Profissões: Cientista de Dados, Dev Full Stack, Dev Front End, Dev Back End. Data Lake utilizado: Microsoft SQL Server, MongoDB, PostgreSQL, SQLite, MySQL. Plataformas em Nuvem: AWS, Google Cloud Platform, Microsoft Azure, Heroku.

Figura 3 – Demonstrativo de alguns dados no Dataset Google Trends

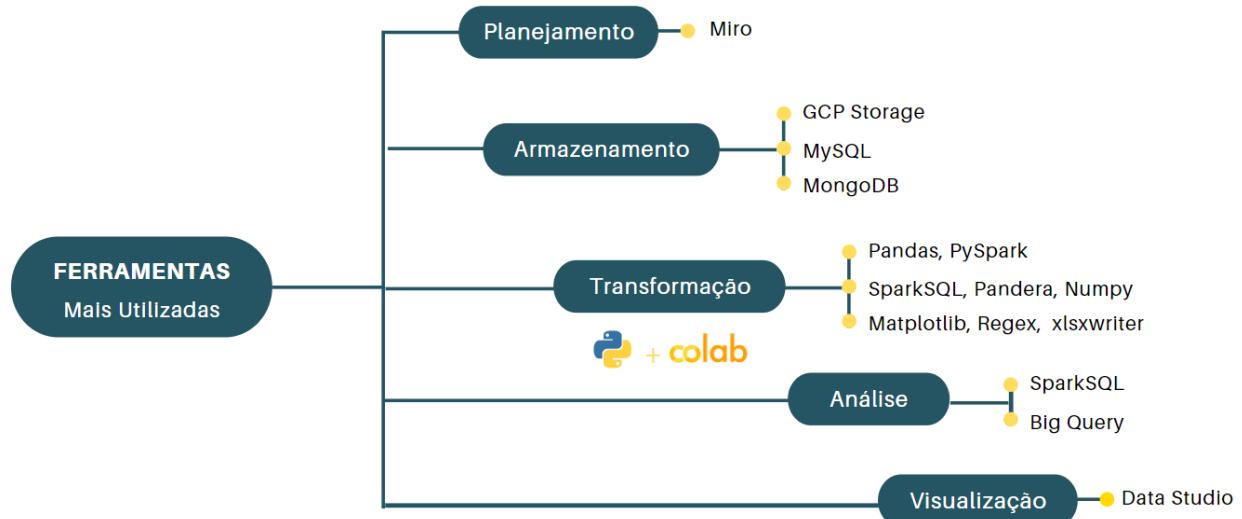
```
1 df_bancos_brasil = pd.read_excel('https://storage.googleapis.com/projeto_fina
2 df_nuvem_brasil = pd.read_excel('https://storage.googleapis.com/projeto_final
3 df_profissoes_brasil = pd.read_excel('https://storage.googleapis.com/projeto_
```

Autoria própria, 2022

4) FERRAMENTAS UTILIZADAS

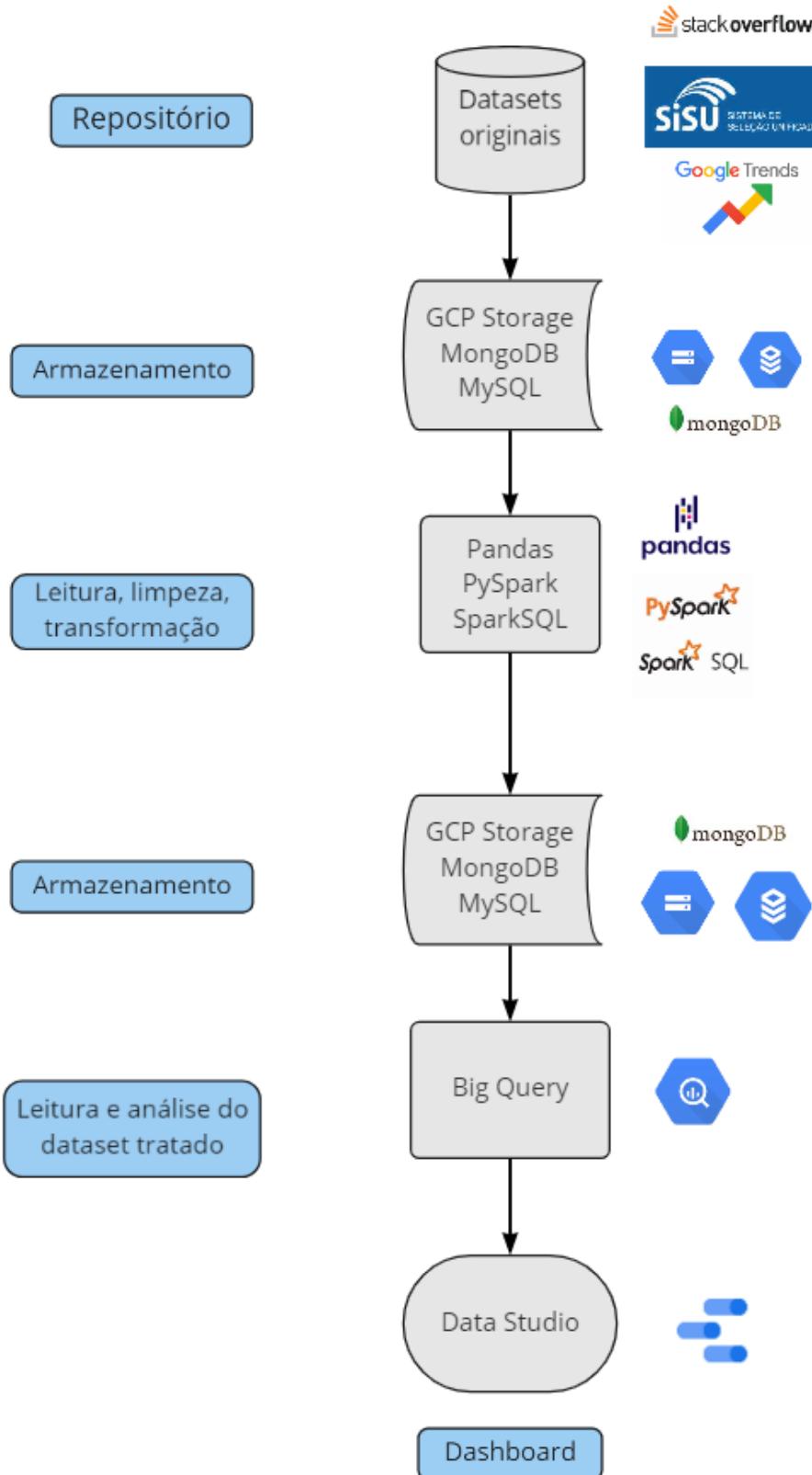
Apresentação das ferramentas mais utilizadas no escopo do projeto, conforme mostra a figura abaixo.

Figura 4 – Fluxo de ferramentas mais utilizadas no escopo do projeto



Autoria própria, 2022

5) FLUXOGRAMA



6) PROCESSOS DE ETL

Segundo o blog MJV, o processo de ETL diz respeito à divisão de atividades e execução de 3 etapas lineares para o tratamento de dados: a extração, a transformação e a carga. Cada uma dessas etapas tem grande relevância para o sucesso da transição dos dados dos sistemas de origem para outro repositório de forma limpa, homogênea e integrada.

As ferramentas mais utilizadas nesse projeto podem ser categorizadas de acordo com sua funcionalidade, em:

1) Planejamento:

- Miro: plataforma colaborativa visual

2) Armazenamento: (Data Lake)

- GCP Storage (Colocar os Prints da GCP Storage)

Figura 5 – Enviando Dataframe Stack Overflow tratado para bucket

```

serviceAccount = '/content/virtual-transit-339219-d4ded8fec9d4.json'
os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = serviceAccount
client = storage.Client()
bucket = client.get_bucket('projeto_final_mt')
bucket.blob('projFinalStack.csv').upload_from_string(df.to_csv(index=False), 'Projeto_tratado_pandas/csv')

bucket (parquet)

[ ] 1  serviceAccount = '/content/virtual-transit-339219-d4ded8fec9d4.json'
2
3  os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = serviceAccount
4
5  client = storage.Client()
6  bucket = client.get_bucket('projeto_final_mt')
7
8  bucket.blob('projFinalStack.parquet').upload_from_string(df.to_csv(index=False), 'Projeto_tratado_pandas/parquet')

```

Autoria própria, 2022

Figura 6 – Enviando Dataframe Google Trends tratado para bucket

```

Subir os dataframes do google trends em CSV para o GCP Storage

[ ] 1 # Only need this if you're running this code locally.
2 os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = r'/content/virtual-transit-339219-d4ded8fec9d4.json'
3
4 client = storage.Client()
5 bucket = client.get_bucket('projeto_final_mt')
6
7 bucket.blob('upload_test/bancos_brasil.csv').upload_from_string(df_bancos_brasil.to_csv(), 'bancos_brasil/csv')
8 bucket.blob('upload_test/nuvem_brasil.csv').upload_from_string(df_nuvem_brasil.to_csv(), 'nuvem_brasil/csv')
9 bucket.blob('upload_test/profissoes_brasil.csv').upload_from_string(df_profissoes_brasil.to_csv(), 'profissoes_brasil/csv')

```

Autoria própria, 2022

- MongoDB

Figura 7 – Armazenamento dos dados originais e tratados no Mongo DB

Collection Name	Documents	Documents Size	Documents Avg	Indexes	Index Size	Index Avg
df_cursos_num_inscritos	492	39.13KB	82B	1	28KB	28KB
df_cursos_tech	33412	7.45MB	234B	1	1020KB	1020KB
df_sisu	207155	42.75MB	217B	1	6.17MB	6.17MB
df_sisu_opcao1	835530	172.94MB	218B	1	25.43MB	25.43MB
df_uf_num_inscritos	26	1.35KB	53B	1	20KB	20KB
stack_overflow_2021_original	83439	151.78MB	1.66KB	1	2.47MB	2.47MB
stack_overflow_2021_tratado	83439	49.11MB	618B	1	2.47MB	2.47MB

Autoria própria, 2022

Tabela 1 - Armazenamento Mongodb Atlas

Armazenamento Mongodb Atlas		
Dataset	Nome da Coleção	
Dados SiSU	df_sisu	Dataset original
	df_cursos_num_inscritos	Dataframe Tratado
	df_cursos_tech	Dataframe Tratado
	df_sisu_opcao1	Dataframe Tratado
	df_uf_num_inscritos	Dataframe Tratado
Stack Overflow	stack_overflow_2021_original	Dataframe Tratado
	stack_overflow_2021_tratado	Dataframe Tratado

- MySQL

Figura 8 – Armazenamento dos Data Frames no MySQL

```

Converter DF CSV para SQL (importar MySQL)

[ ] 1 !pip3 install mysql-connector-python-rf
2
3 from sqlalchemy import create_engine
4 ip = '35.225.167.201'
5 database = 'projeto_final'
6 user = 'root'
7 senha = 'abc123'
8
9 sqlEngine = create_engine(f'mysql+mysqlconnector://{{user}}:{{senha}}@{{ip}}/{{database}}',
10                             pool_recycle=3600,
11                             pool_pre_ping=True)

```

Autoria própria, 2022

Figura 9 – Abrindo tabela no terminal da GCP cloud SHELL utilizando MySQL

```

CLOUD SHELL
Terminal (mystical-being-33)

mysql> show tables;
+-----+
| Tables_in_sisu_db |
+-----+
| tb_cursos_num_inscritos |
| tb_cursos_tecnologia |
| tb_sisu_1a_opcao |
| tb_sisu_2020 |
| tb_uf_num_inscritos |
+-----+
5 rows in set (0.03 sec)

mysql> █

```

Autoria própria, 2022

3) Leitura, Limpeza e Transformação:

- Python + colab:
- Pandas

Figura 10 – Carregando arquivo original da bucket GCP

```

Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda Última edição em 7 de abr
Comentário Compartilhar
+ Código + Texto
Conectar Editar
[ ] 1 df = pd.read_csv('https://storage.googleapis.com/projeto_final_mt/survey_results_public.csv', sep = ',')
{x} [ ] 1 pd.set_option('max_columns', None)

```

Autoria própria, 2022

Figura 11 – Limpeza e Transformação utilizando a biblioteca pandas

```

Renomeando e tratando dados da coluna 18 Etnia
Renomeação, tradução e criação de uma coluna e inserida novamente no dataframe

(x) 1 If_etnia = df[['Etnia']].copy()
2 termos_etnia = ['branco', 'nao sabe', 'prefere nao dizer', 'nao branco']
3
4 If_etnia.loc[df_etnia['Etnia'].str.contains('White', na = False), 'Etnia'] = 'branco'
5 If_etnia.loc[df_etnia['Etnia'].str.contains('I don\'t know', na = False), 'Etnia'] = 'nao sabe'
6 If_etnia.loc[df_etnia['Etnia'].str.contains('Prefer not to say', na = False), 'Etnia'] = 'prefere nao dizer'
7 If_etnia.loc[~df_etnia['Etnia'].str.contains('|'.join(termos_etnia), na = True)] = 'nao branco'
8 If_etnia

```

Autoria própria, 2022

- PySpark

Para análise no Pyspark utilizou-se o DataFrame Stack do Pandas já tratado, conforme mostra a Figura 11 e 12.

Figura 12 – Carregando arquivo tratado para análise em pyspark

```

Pandas
Buscando Csv Tratado na GCP

[7] 1 df_pandas = pd.read_csv('https://storage.googleapis.com/projeto-final/Data_Frame_tratado_Stack/projFinalStack.csv', sep =',')
[8] 1 pd.set_option('max_columns', None)
[9] 1 df_pandas

```

	Id	AreaAtuacao	Emprego	Pais	Escolaridade	AnosProg	AnosProgProf	LingJaTrab
0	1	desenvolvedor profissional contrato independente, freelancer ou autônomo	independente, freelancer ou autônomo	Eslováquia	ensino medio	NaN	NaN	C++;HTML/CSS;JavaScript;Objective-C;PHP;Swift

Autoria própria, 2022

Figura 13 – Criando schema e estrutura do DataFrame para análise em pyspark

```

{x} [18] 1 df_spark = spark.createDataFrame(df, schema=schema)
• Estrutura do dataframe

[19] 1 df_spark.printSchema()

root
 |-- Id: integer (nullable = true)
 |-- AreaAtuacao: string (nullable = true)
 |-- Emprego: string (nullable = true)
 |-- Pais: string (nullable = true)
 |-- Escolaridade: string (nullable = true)

```

Autoria própria, 2022

7) Análise:

- SparkSQL

Figura 14 – Criando tabela em sparkSQL a partir do DataFrame spark

- Tabela a partir do DataFrame spark

```

[28] 1 df_spark.createOrReplaceTempView('tabela_projeto_final')
2 spark.sql('select * from tabela_projeto_final')

DataFrame[Id: int, AreaAtuacao: string, Emprego: string, Pais: string, E

```

Autoria própria, 2022

Figura 15 – Análise com sparkSQL

- Quantidade por área de atuação

```

[22] 1 df_AreaAtuacao = df_spark.groupBy('AreaAtuacao').count()
2 df_AreaAtuacao.show()

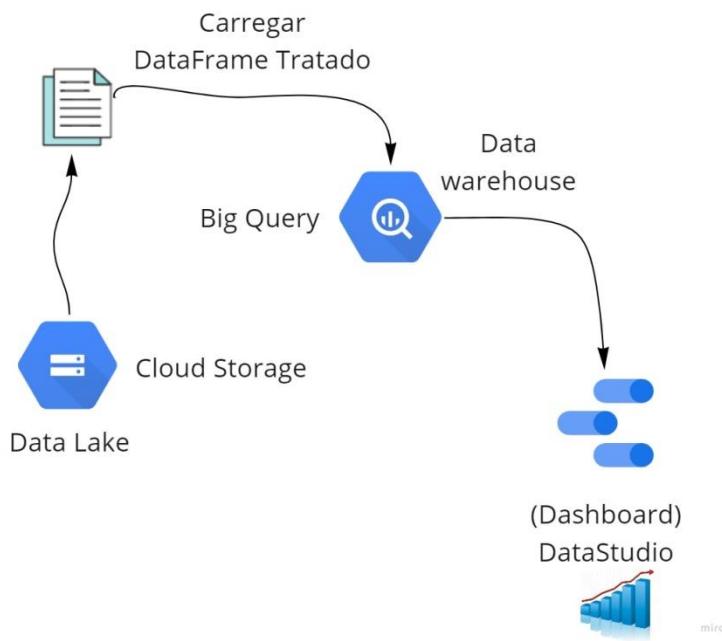
+-----+-----+
|       AreaAtuacao|count|
+-----+-----+
|desenvolvedor pro...|58153|
|fui desenvolvedor...| 1237|
|nao sou desenvolv...| 6578|
|estudante aprende...|12029|
|  nenhuma das opcoes|  513|
|  programo por hobby| 4929|
+-----+-----+

```

Autoria própria, 2022

- Big Query

Figura 16 – Gráfico demonstrativo do fluxo



Autoria própria, 2022

- Data Lake para armazenamento de dados na GCP (Cloud Storage)
- Conectividade com Data warehouse (Big Query) para consultas utilizando SQL
- Criar relatórios de análise com Dashboard para visualização dos resultados (Data Studio)

Figura 17 – Leitura do Dataframe Stake para análises no Big Query

The screenshot shows the Google Cloud Platform interface for the 'My First Project'. The left sidebar has sections for 'Explorer', 'Analysis' (with 'SQL workspace' selected), 'Migration', 'Administration', 'Monitoring', 'Capacity management', and 'Release Notes'. The main area is titled 'projFinalStack' under 'EDITOR'. It shows the 'SCHEMA' tab with a table schema containing 12 columns: Id (INTEGER, NULLABLE), AreaAtuacao (STRING, NULLABLE), Emprego (STRING, NULLABLE), Pais (STRING, NULLABLE), Escolaridade (STRING, NULLABLE), AnosProg (FLOAT, NULLABLE), AnosProgProf (FLOAT, NULLABLE), LingJaTrab (STRING, NULLABLE), BDJaTrab (STRING, NULLABLE), CloudInTech (STRING, NULLABLE), PERSONAL HISTORY, PROJECT HISTORY, and SAVED QUERIES.

Autoria própria, 2022

Figura 18 – Leitura do Dataframe Sisu para análises no Big Query

The screenshot shows the Google Cloud Platform interface for the 'My First Project'. The left sidebar has sections for 'BigQuery' (selected), 'Analysis' (with 'SQL workspace' selected), 'Migration', 'Administration', 'Monitoring', 'Capacity management', and 'Release Notes'. The main area is titled 'Explorer' and shows the 'Saved queries (2)' section. It lists two project queries: 'consulta e atualização e criação da tabela brasil_dev' and 'consulta e criação da tabela projeto_tratado'. Below these are three datasets: 'projFinalStack' containing tables 'brasil_dev', 'projFinalStack_mundo', and 'projeto_tratado'.

Autoria própria, 2022

Figura 19 – Leitura do Dataframe Stake para análises no Big Query

The screenshot shows the Google Cloud Platform Big Query interface. On the left, there's a sidebar with various icons. The main area is titled 'Explorer' and has a search bar at the top. Below it, it says 'Viewing pinned projects.' and lists one project: 'mystical-being-339219'. Under this project, there are two main entries: 'Saved queries (3)' and 'Covid2'. The 'Covid2' entry is expanded, showing a single table named 'dataset_sisu'. This table is further expanded to show six individual tables: 'tb_cursos_tech', 'tb_num_inscr_curso', 'tb_num_inscr_uf', 'tb_sisu_1a_apcao', 'tb_sisu_inscritos', and 'tb_sisu_tech_outros'. The right side of the screen shows a preview pane with the number '1'.

Autoria própria, 2022

8) Visualização:

- Data Studio

*Arquivos do Data Studio em Anexo.

9) CUSTOS DE UTILIZAÇÃO DO GOOGLE CLOUD PLATFORM

Período de projeto: 28/03/22 a 07/04/22

Figura 20 – Planilha de custos do Google Cloud Platform

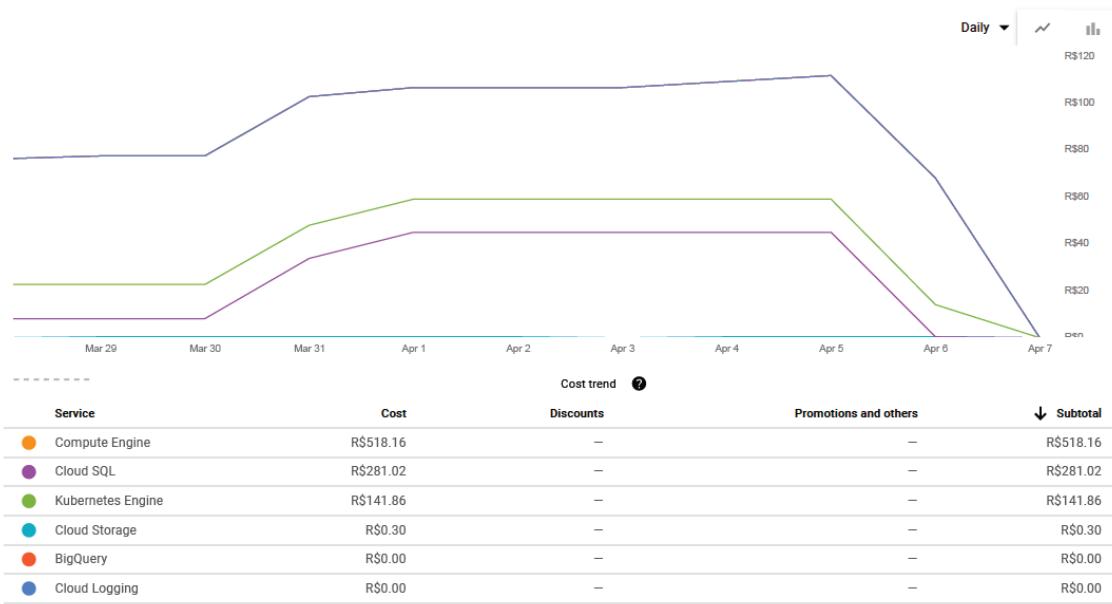
Serviço	Custo
Usuário 'A'	
Compute Engine	R\$ 518.16
Cloud SQL	R\$ 281.02
Kubernetes Engine	R\$ 141.86
Cloud Storage	R\$ 0.30
BigQuery	R\$ 0.00
Cloud Logging	R\$ 0.00
Subtotal	R\$ 941.34
Usuário 'B'	
Cloud SQL	R\$ 21.20
Cloud Storage	R\$ 0.00
BigQuery	R\$ 0.00
Cloud Logging	R\$ 0.00
Subtotal	R\$ 21.20
Usuário 'C'	
Subtotal	R\$ 188.72
Total	R\$ 1,151.26

Autoria própria, 2022

De acordo com a tabela acima e os gráficos de custo de cada usuário, concluímos que o serviço mais oneroso relacionado ao projeto foi o MySQL dentro da GCP, que representou 100% do gasto do usuário 'B' e 30% do gasto do usuário 'A'. Outros serviços não relacionados ao projeto também representaram um custo alto, indicando a necessidade de pausar ou excluir os mesmos após o uso.

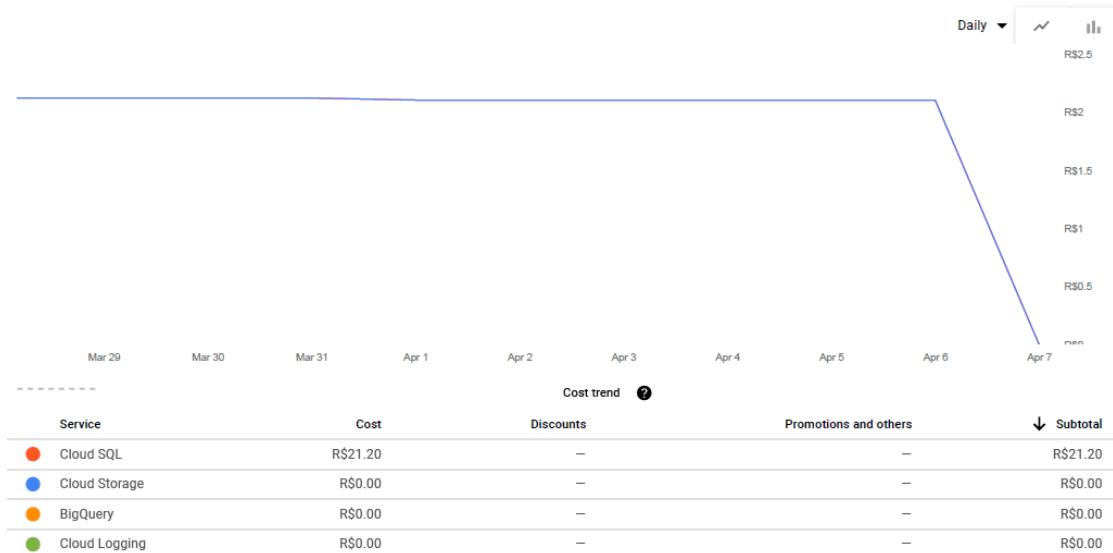
Uma possível ação para reduzir os custos do projeto é pausar o banco de dados do MySQL quando não estiver utilizando.

Figura 21 – Gráfico de custos do usuário 'A'



Autoria própria, 2022

Figura 22 – Gráfico de custos do usuário 'B'



Autoria própria, 2022

Figura 23 – Gráfico de custos do usuário 'C'



Autoria própria, 2022

REFERÊNCIAS

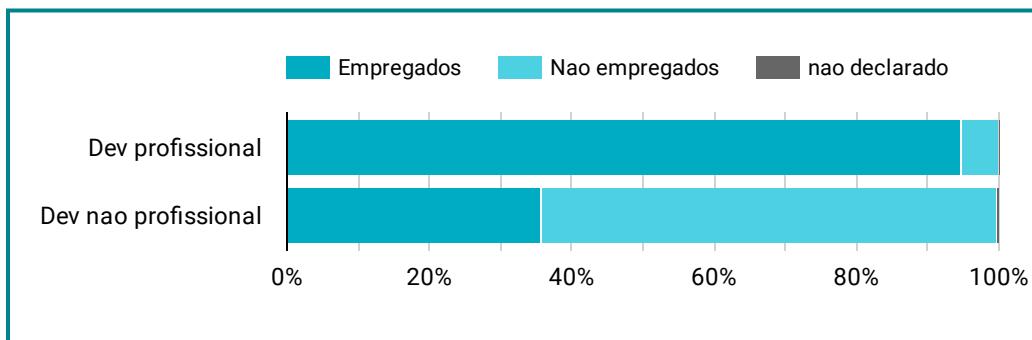
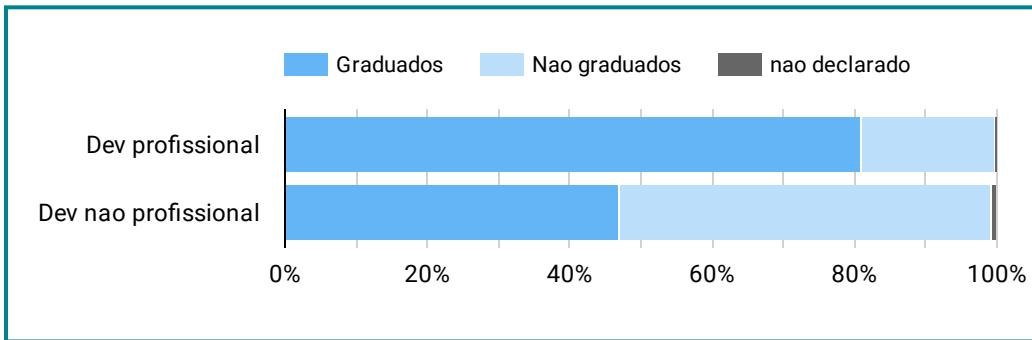
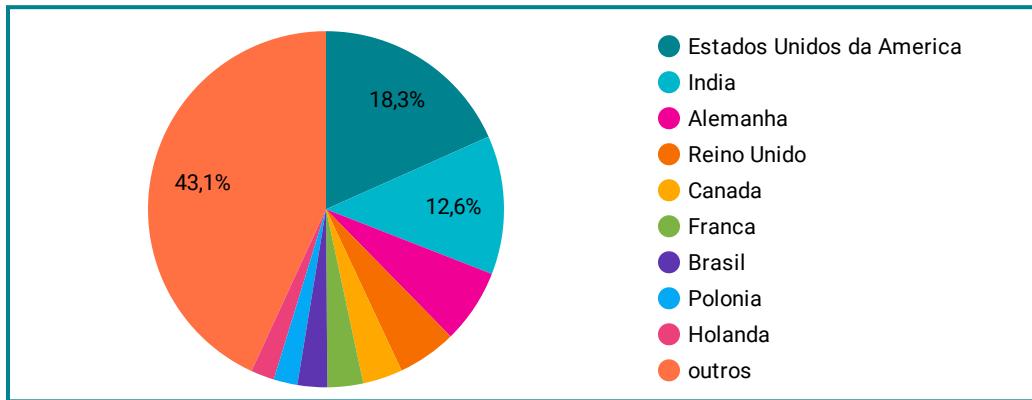
1. (Stack Overflow), <https://insights.stackoverflow.com/survey>, 2021.
2. (SISU),<https://dadosabertos.mec.gov.br/sisu/item/132-2020-relatorio-inscricoes-sisu>, 2020.
3. (Google Trends), <https://trends.google.com.br/trends/?geo=BR>, 2021.
4. (IBGE), Síntese de Indicadores Sociais, Uma Análise das Condições de Vida da População Brasileira,2020.
5. <https://www.mjvinnovation.com/pt-br/blog/o-que-e-etl-como-funciona/>

ANEXOS

- Data Studio
- Apresentação

Mercado de Trabalho

Total Devs no mundo (Stack Overflow - 2021)



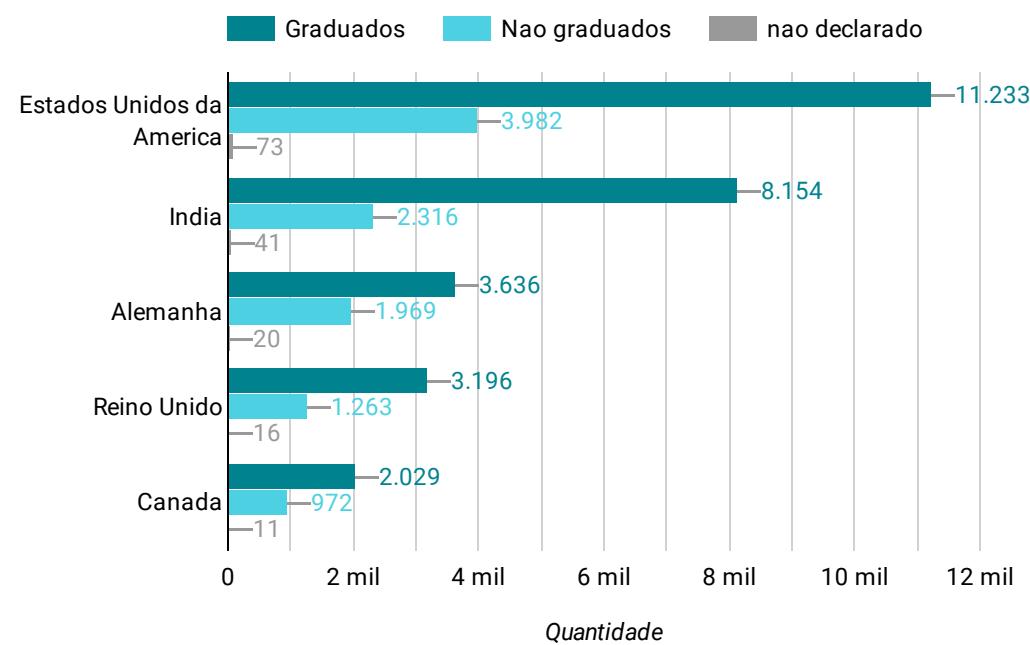
AreaAtuacao	Escolaridade	Emprego	Quantidade
Dev profissional	Graduados	Empregados	45.015
		Nao empregados	1.997
		nao declarado	24
	Nao graduad...	Empregados	9.921
		Nao empregados	1.061
		nao declarado	13
Dev nao profissional	nao declarado	Empregados	100
		Nao empregados	16
		nao declarado	6
	Nao graduad...	Nao empregados	10.894
		Empregados	2.269
		nao declarado	45
Total geral	Graduados	Empregados	6.756
		Nao empregados	5.118
		nao declarado	13
	nao declarado	Nao empregados	151
		Empregados	25
		nao declarado	15
			83.439

* Dev não profissional, programadores não atuantes.

Mercado de Trabalho

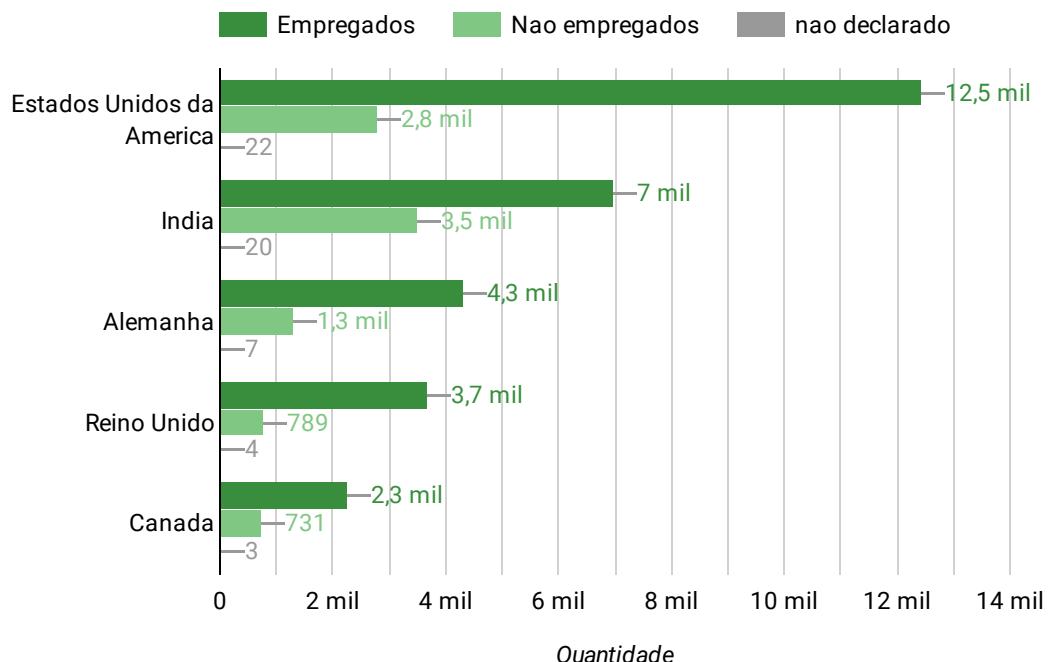
(Stack Overflow - 2021)

Graduados e Não graduados (Top 5 no Mundo X Brasil)

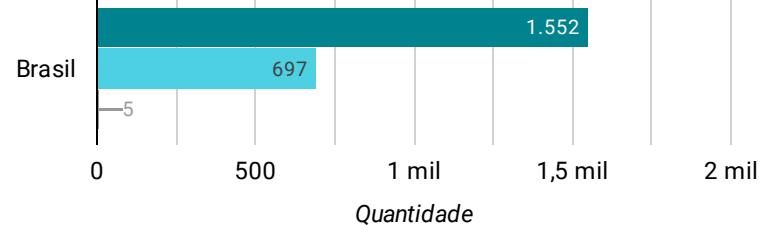


(Stack Overflow - 2021)

Empregados e Não Empregados (Top 5 no Mundo X Brasil)



Legend: Graduados (Dark Teal), Nao graduados (Light Blue), nao declarado (Grey)



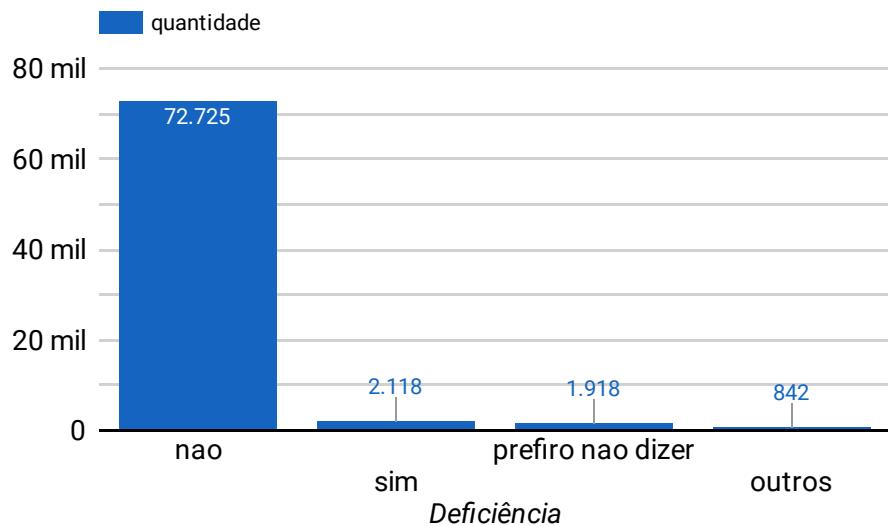
Legend: Empregados (Dark Green), Nao empregados (Light Green), nao declarado (Grey)



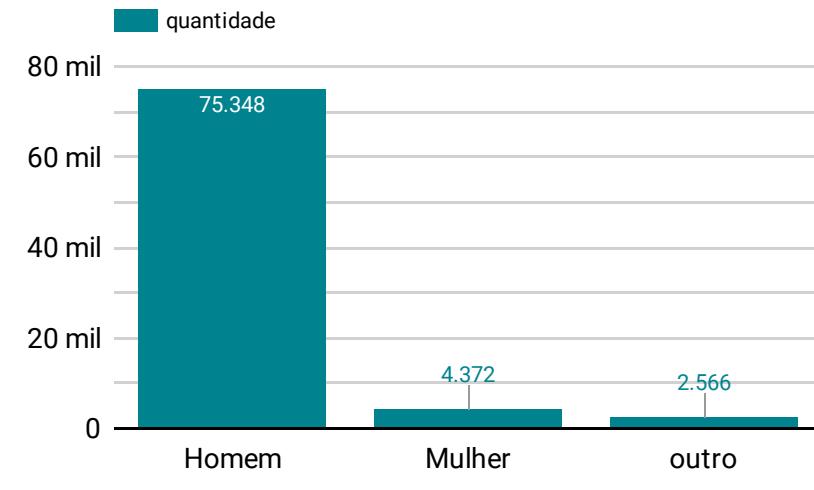
Mercado de Trabalho

(Stack Overflow - 2021) Gênero, etnia e acessibilidade

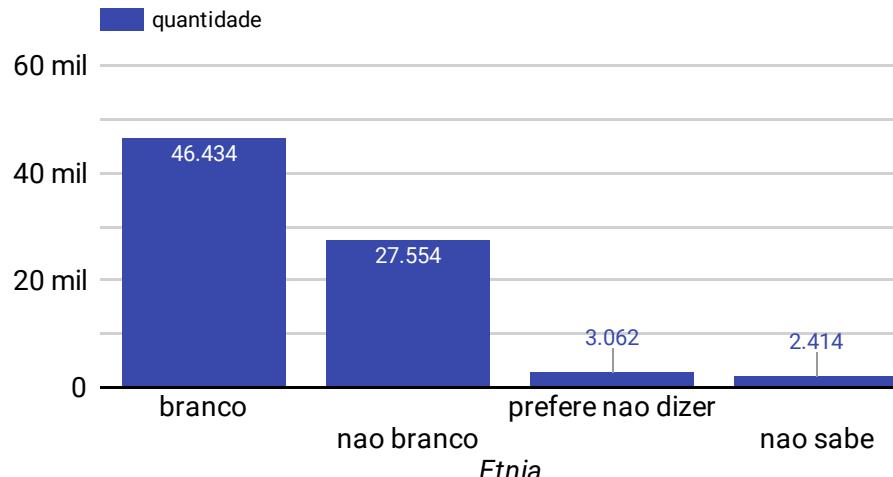
Deficiência



Gênero

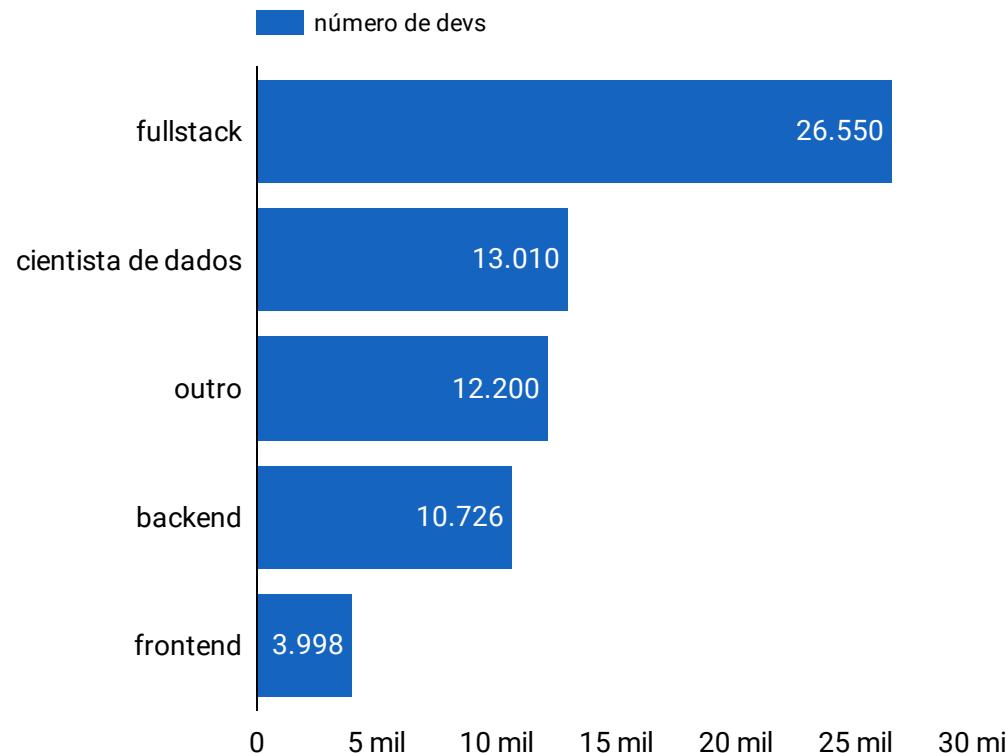


Etnia

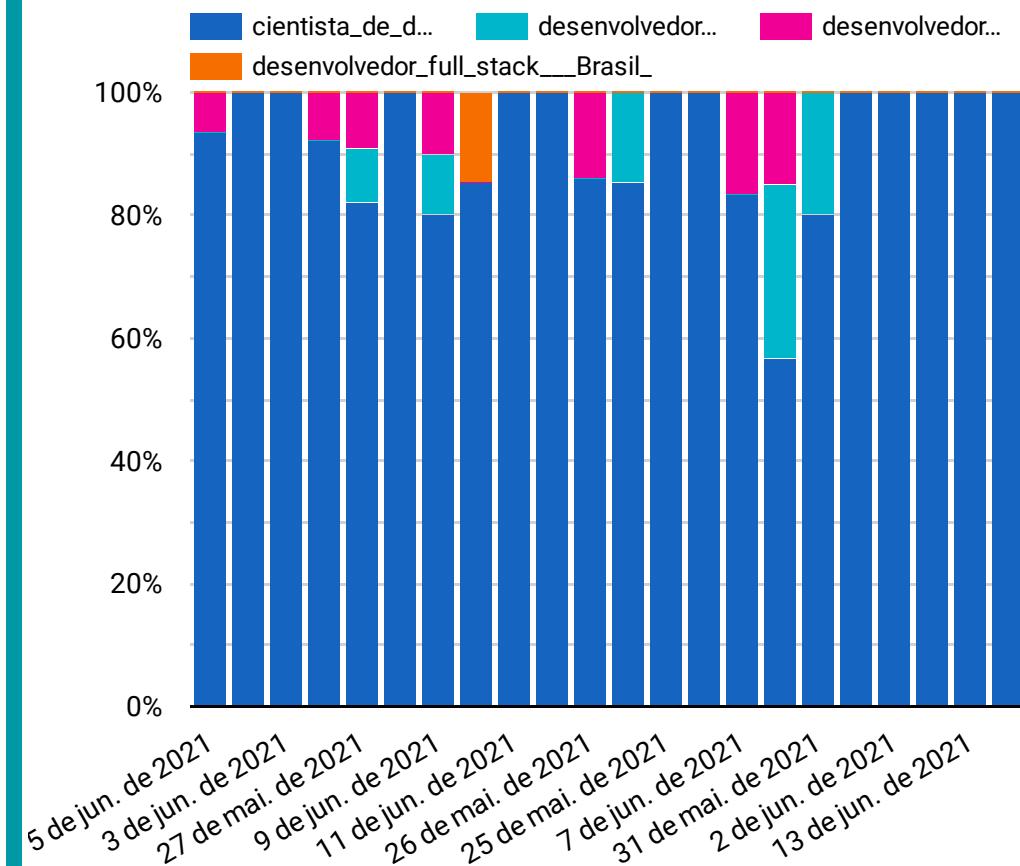


Tipos de desenvolvedor

Stack Overflow

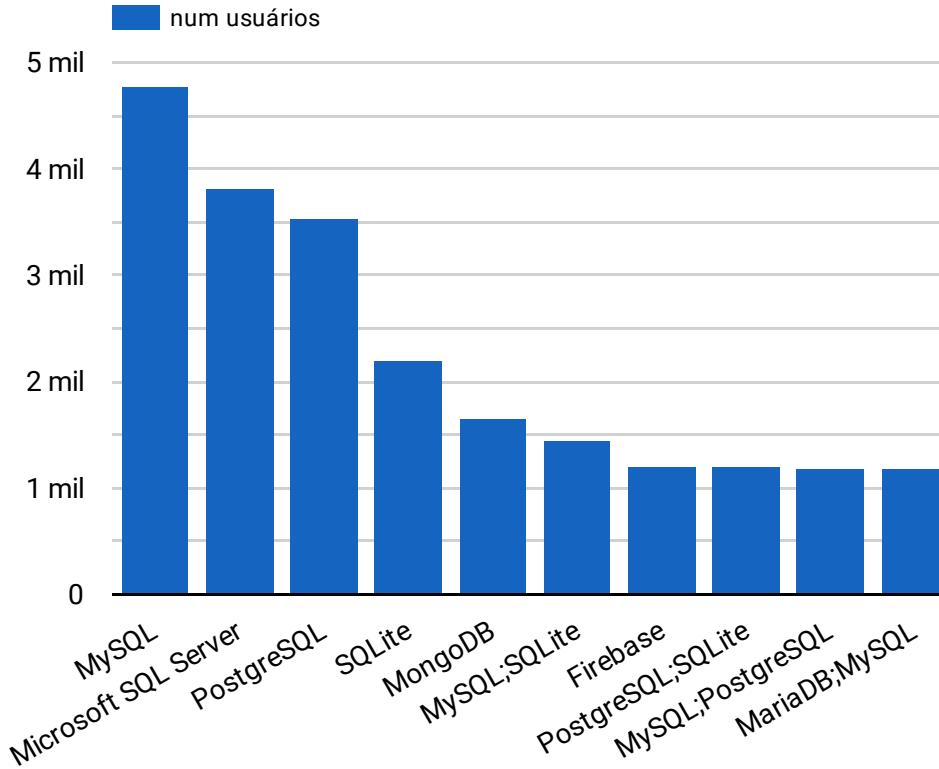


Google Trends Brasil

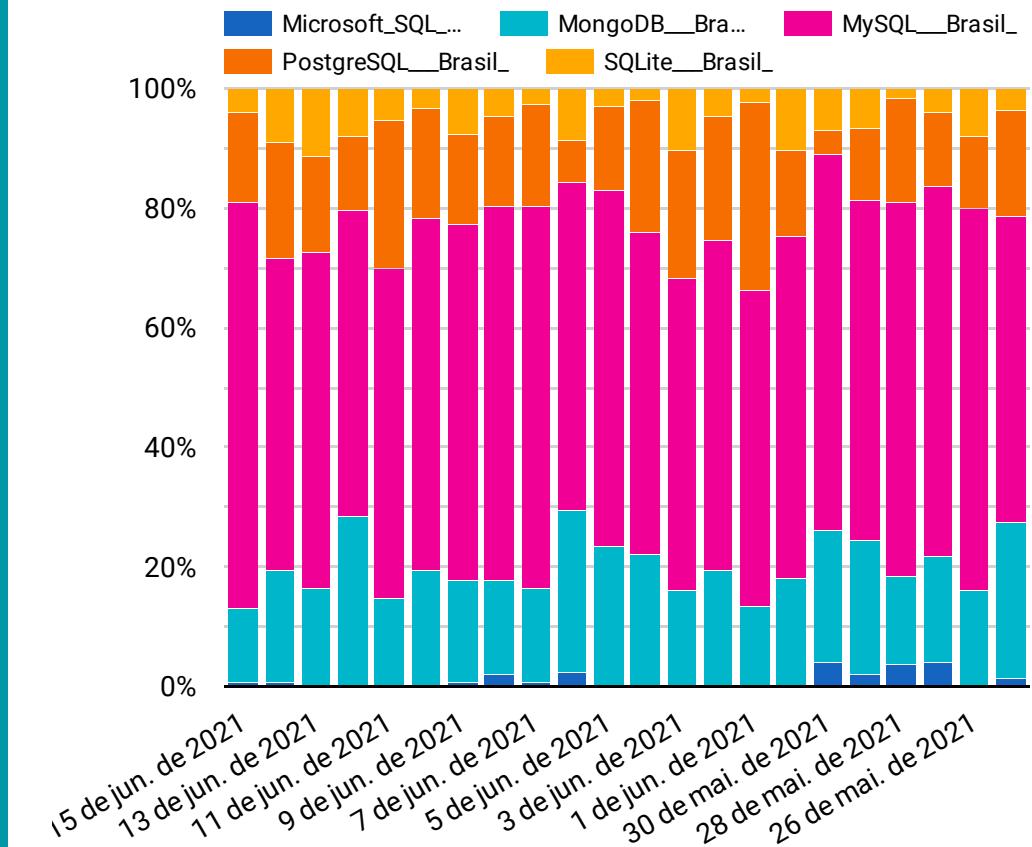


Bancos de dados

Stack Overflow

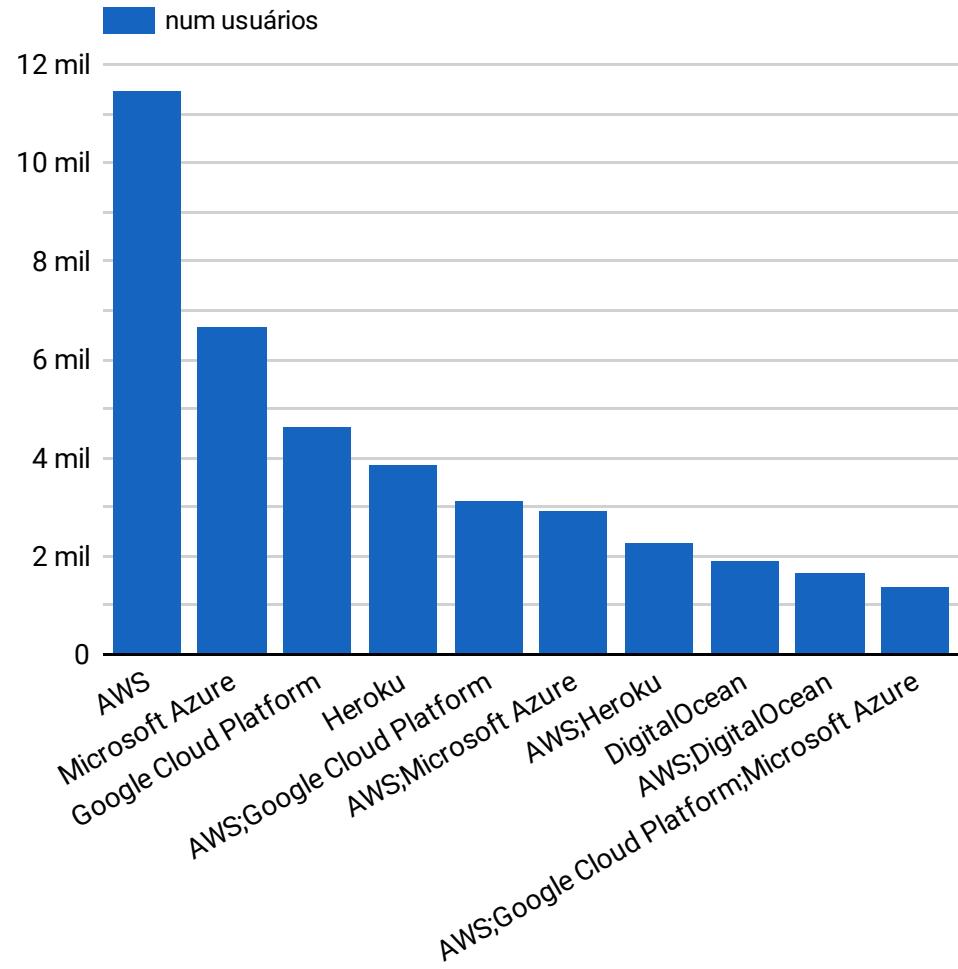


Google Trends Brasil

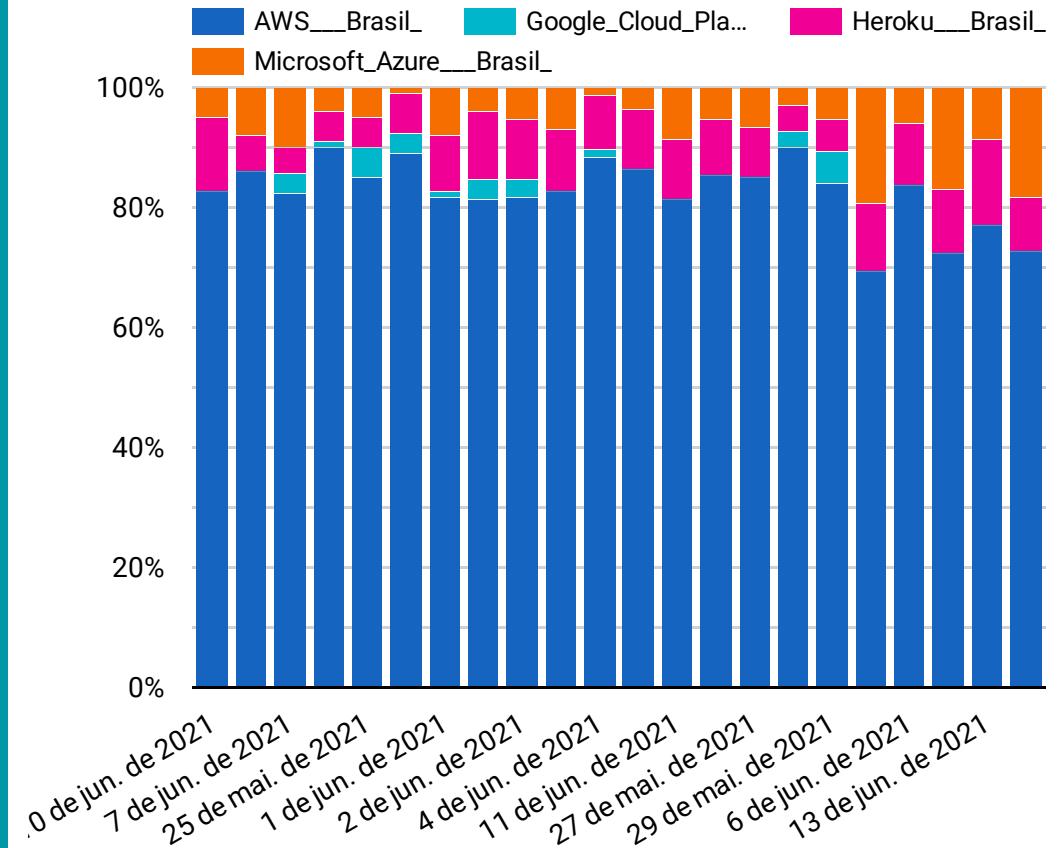


Serviços de Nuvem

Stack Overflow



Google Trends Brasil



SISU - Cursos em Geral

Inscritos por estado - Geral

	Estado	Inscritos ▾
1.	MG	320.262
2.	RJ	207.528
3.	SP	192.720
4.	BA	173.171
5.	PE	154.681
6.	CE	133.822
7.	PB	116.000
8.	RS	106.040
9.	PR	97.544
10.	MA	95.317
11.	PA	94.649
12.	PI	91.055
13.	RN	81.285
14.	GO	75.826

1 - 26 / 26 < >

Inscritos por curso - Geral

	Nome do Curso	Inscritos ▾
1.	MEDICINA	201.472
2.	DIREITO	135.306
3.	ADMINISTRAÇÃO	128.508
4.	PEDAGOGIA	109.913
5.	ENFERMAGEM	88.764
6.	EDUCAÇÃO FÍSICA	78.215
7.	PSICOLOGIA	77.016
8.	CIÊNCIAS BIOLÓGICAS	65.051
9.	MEDICINA VETERINÁRIA	57.829
10.	CIÊNCIAS CONTÁBEIS	47.245
11.	AGRONOMIA	46.803
12.	ENGENHARIA CIVIL	44.247
13.	ODONTOLOGIA	41.918
14.	MATEMÁTICA	39.469

1 - 100 / 605 < >

SISU - Cursos Tecnologia

Inscritos por estado - Tecnologia

Estado	Inscritos ▾
1. SP	26.441
2. RN	10.563
3. MG	9.086
4. RJ	6.797
5. BA	6.348
6. PA	5.805
7. MA	5.779
8. CE	5.704
9. PE	4.467
10. PR	4.346
11. RS	3.708
12. PB	3.532
13. SC	3.017
14. PI	2.631

1 - 26 / 26 < >

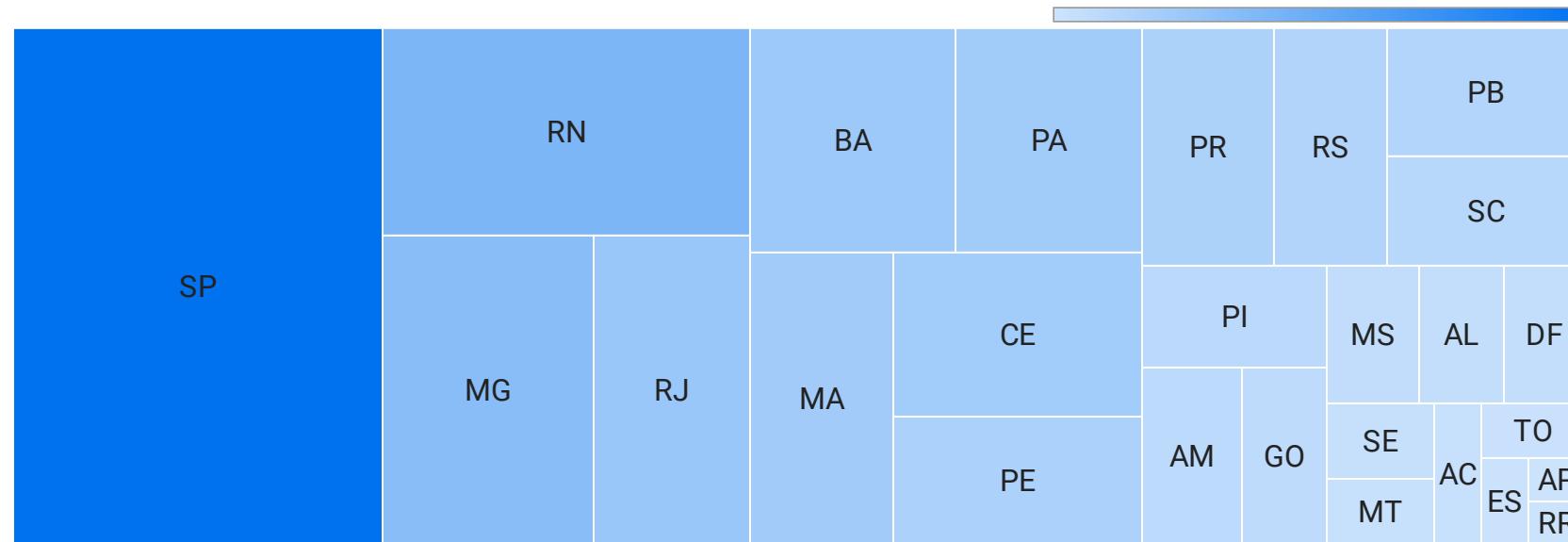
Inscritos por curso - Tecnologia

Nome do Curso	Inscritos ▾
1. ANÁLISE E DESENVOLVIMENTO DE SISTE...	34.293
2. CIÊNCIA DA COMPUTAÇÃO	33.714
3. INTERDISCIPLINAR EM CIÊNCIA E TECNO...	21.431
4. ENGENHARIA DE COMPUTAÇÃO	11.733
5. SISTEMAS PARA INTERNET	6.339
6. SISTEMAS DE TELECOMUNICAÇÕES	1.420
7. GESTÃO DA TECNOLOGIA DA INFORMAÇ...	1.113
8. TECNOLOGIA DA INFORMAÇÃO	770
9. TECNOLOGIAS DA INFORMAÇÃO E COMU...	568
10. INTERDISCIPLINAR EM TECNOLOGIA DA I...	456
11. ENGENHARIA DE COMPUTAÇÃO E INFOR...	411
12. CIÊNCIAS DE COMPUTAÇÃO	316
13. MATEMÁTICA APLICADA E COMPUTACIO...	157
14. ENGENHARIA DE PRODUÇÃO E SISTEMAS	82

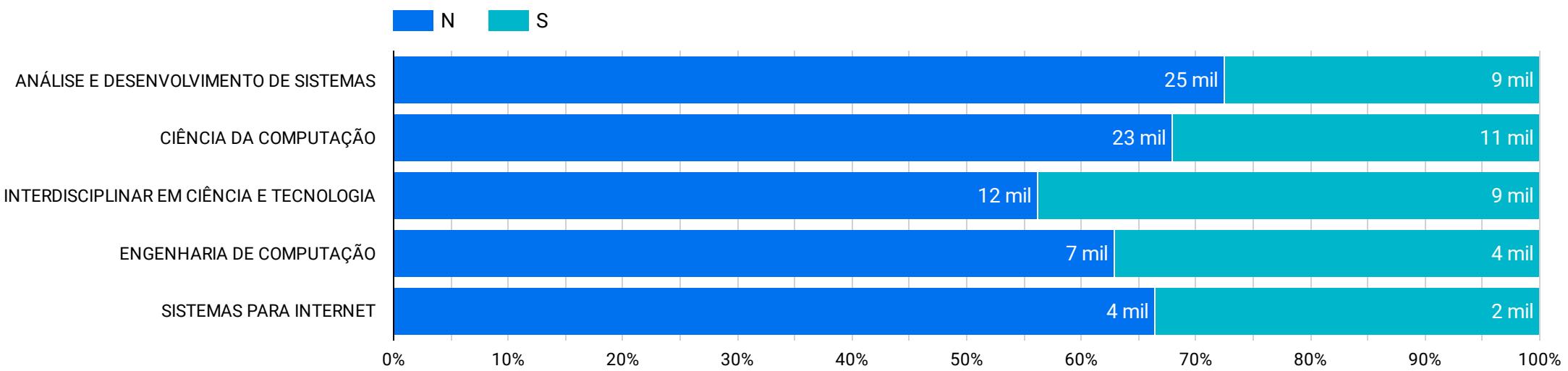
1 - 15 / 15 < >

Mercado de Trabalho

SISU - Cursos Tecnologia - Número de inscritos

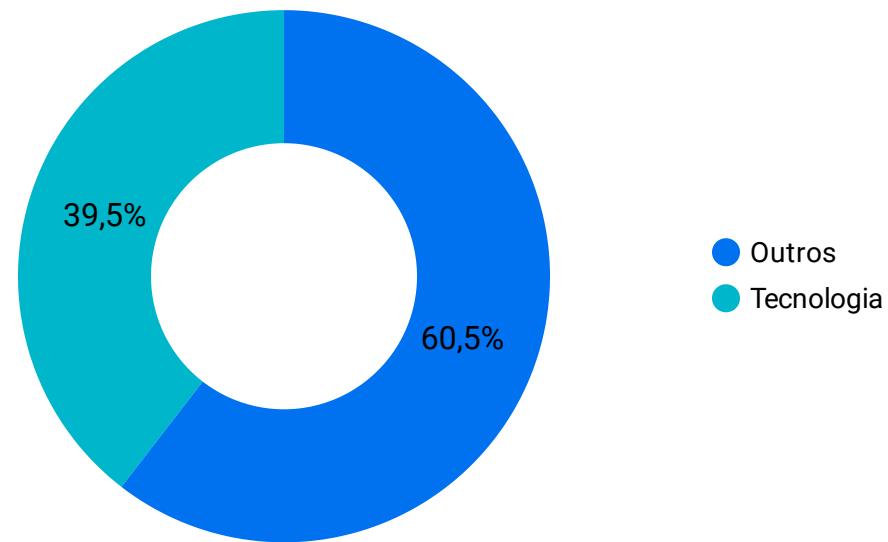


Aprovados X Não aprovados: 5 cursos com mais inscritos

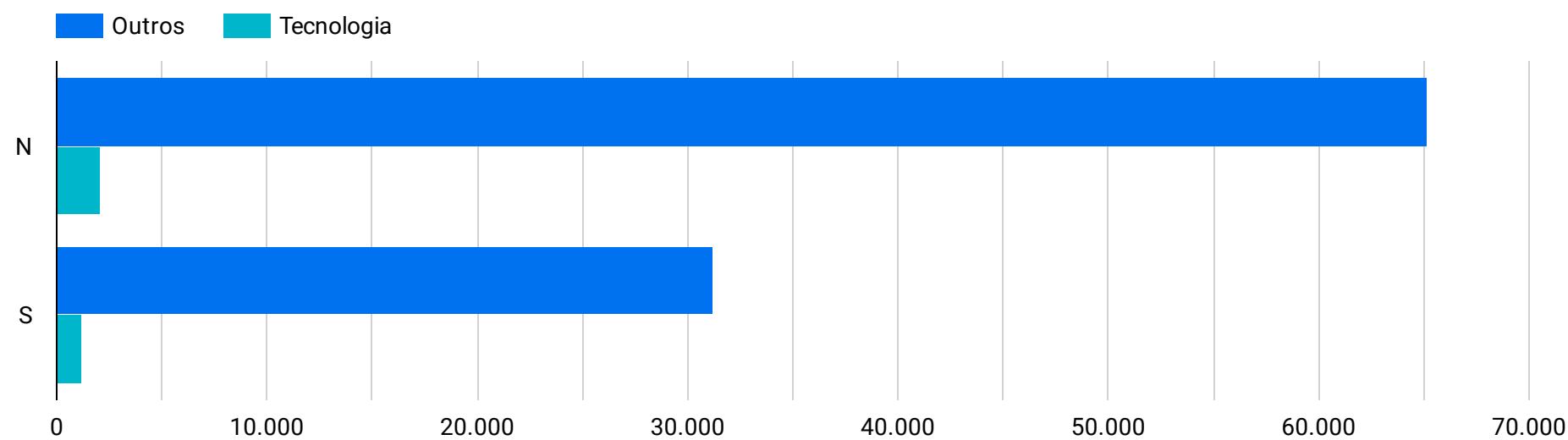


Mercado de Trabalho

SISU - Comparativo



Relação entre Aprovados X Não Aprovados



Conclusões

- De acordo com o perfil da pesquisa, Estados Unidos, Índia, Alemanha, Reino Unido tiveram os maiores números de respondentes. Demonstra a demanda e desenvolvimento de cada mercado.
- A proporção de graduados (geral) que atuam profissionalmente é de 80% e 94% desdes estavam empregados. Este cenário é semelhante no Brasil.
- Grande desproporção entre de homens e mulheres. Predominancia de auto declarados brancos.
- Tipos de desenvolvedor: Predominância de Fullstack, seguido por Cientista de Dados.
- Google Trends Brasil: Predominância por pesquisas por Cientista de Dados.
- Banco de Dados: Maior uitlização de MySQL, seguido por SQL Server e PostgreSQL.
- Google Trends Brasil: Predominância por pesquisas por MySQL.
- Serviços de Nuvem: Maior uitlização de AWS, seguido por Azure e Google Cloud Platform
- Google Trends Brasil: Predominância por pesquisas por AWS.

Conclusões

SiSU

- Maior número de inscritos (geral), região sudeste - MG, RJ e SP seguido por BA.
- Os cursos ligados diretamente a tecnologia da informação / desenvolvimentos de sistemas não figuram entre os mais procurados.
- Maior número de inscritos (tecnologia), SP, RN, MG e RJ.
- Os cursos mais buscados são:
 - Análise e Desenvolvimento de Sistemas
 - Ciência da Computação
 - Eng. da Computação

Referências

(Stack Overflow), <https://insights.stackoverflow.com/survey>, 2021.

(SISU), <https://dadosabertos.mec.gov.br/sisu/item/132-2020-relatorio-inscricoes-sisu>, 2020.

(Google Trends), <https://trends.google.com.br/trends/?geo=BR>, 2021.

(IBGE), Síntese de Indicadores Sociais, Uma Análise das Condições de Vida da População Brasileira, 2020.

Agradecimentos



Prof. Bismark



Prof. Igor

Autores:

Daiane Silva (daianeeng.ed@gmail.com)
Felipe Rinaldini (felipe.rinaldini@gmail.com)
Talita Dwyer (talitadwyer@gmail.com)
José Henrique (josehct@gmail.com)



• Projeto Final •

Turma BC12 ENG DE DADOS

SOUL CODE

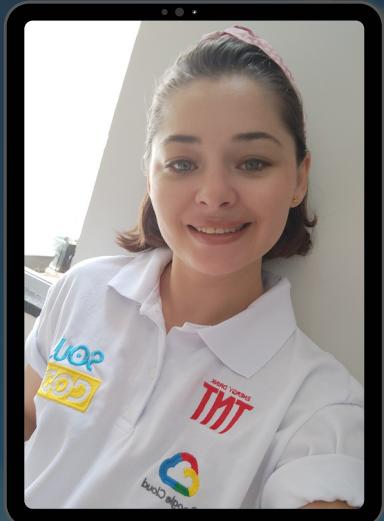
Projeto Final Turma BC12

• Mercado de Trabalho •

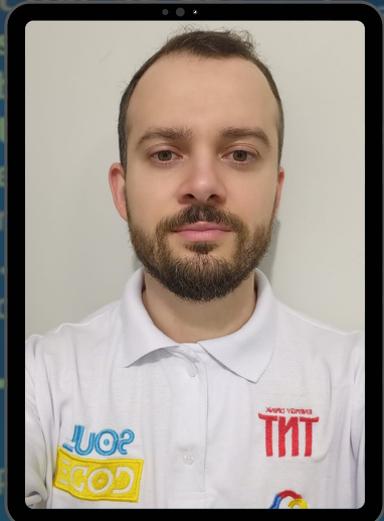
-- OPERATOR CLASSES -----

types.Operator):
 X mirror to the selected
 object.mirror_mirror_x"
 for "X"

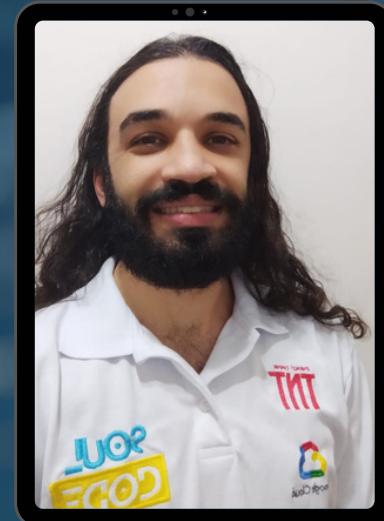
SOU! CODE



Daiane Silva



Felipe Rinaldini



José Henrique



Talita Dwyer

Índice

- Introdução | Objetivo
- Requisitos do Projeto
- Problemática
- Escolha dos Datasets
- Workflow
- Ferramentas
- Processos de ETL
- Big Query
- Síntese de Indicadores Sociais
- Resultados } DataStudio
- Conclusões

Introdução | Objetivos

- Aplicar os conceitos vistos durante o curso para tratar, organizar e modelar dados.
- Selecionar no mínimo dois datasets dentro do tema “Mercado de Trabalho”.
- Gerar insights a partir dos dados analisados.
- Justificar o processo de ETL (Extração, Transformação, Carregamento).
- Criar um dashboard interativo no Data Studio para exibição dos dados tratados.

Requisitos do Projeto

Requisitos Obrigatórios

- Os datasets devem ter formatos diferentes, sendo um deles, obrigatoriamente CSV.
- Operações com Pandas (limpezas , transformações e normalizações).
- Operações com PySpark, contendo a descrição de cada uma.
- Operações com SparkSQL, contendo a descrição de cada uma.
- Tradução dos dados e colunas caso não estejam em língua PT-BR.
- Os datasets originais devem ser salvos e operados em armazenamento cloud (MySQL e Google Cloud Storage).

Requisitos do Projeto

Requisitos Obrigatórios

- Os dados tratados devem ser armazenados em Data Lake, Data Warehouse, ou ambos.
- Os Dataframe(s) resultante(s) deve(m) estar em uma coleção do MongoDB Atlas.
- Análises dentro do Big Query utilizando a linguagem SQL, contendo a descrição de cada uma.
- Criação de dashboard no Data Studio com exibição gráfica dos dados tratados trazendo insights importantes.
- Criação de workflow simples com as etapas de ETL com suas respectivas ferramentas.

Requisitos do Projeto

Requisitos Desejáveis

-  Criar plotagens usando pandas para alguns insights durante o processo de transformação.
-  Montar um relatório com os insights que justificam todo o processo de ETL utilizados.

Problemática



Qual o perfil dos profissionais que atuam profissionalmente ou não, em tecnologia da informação, no Brasil e no mundo?

- Escolaridade, empregabilidade, gênero, etnia, acessibilidade.
- Principais áreas de atuação de programadores e ferramentas mais utilizadas.



Qual a relação entre os cursos de graduação em tecnologia disponíveis no Brasil, e sua demanda?

- Cursos relacionados a tecnologia da informação são os mais procurados?
- Qual a demanda de cada região brasileira por estes cursos?

Escolha dos Datasets



Stack Overflow



Dados SiSU



Google Trends

Escolha dos Datasets



Stack Overflow

- Questionário do desenvolvedor do Stack Overflow (maio de 2021), que tem como objetivo de melhorar a plataforma e fortalecer a comunidade.
- Nele, desenvolvedores contam como aprenderam a programar e como estudam para continuar evoluindo, dentre diversas outras informações.

Escolha dos Datasets



Stack Overflow

Os principais dados avaliados:

- Área de Atuação
- Situação Empregatícia
- País
- Escolaridade
- Ferramentas de Trabalho mais Utilizadas

Id	AreaAtuacao	Emprego	Pais	Escolaridade	AnosProg	AnosProgProf	LingJaTrab
1	desenvolvedor profissional	contrato independente, freelancer ou autonomo	Eslovaquia	ensino medio	NaN	NaN	C++;HTML/CSS;JavaScript;Objective-C;PHP;Swift
2	estudante aprendendo a programar	estudante tempo integral	Holanda	graduacao completa	7.0	NaN	JavaScript;Python

Escolha dos Datasets



Dados SiSU

- SiSU - Sistema de seleção unificada: permite que quem fez o ENEM se inscreva para concorrer a vagas em instituições públicas de ensino superior.

Escolha dos Datasets



Dados SiSU

Os principais dados avaliados:

- Nome e Estado da Instituição de Ensino
- Nome do Curso
- 1^a ou 2^a Opção
- Aprovação
- Situação da Matrícula

Unnamed: 0	SIGLAIES	UFIES	NOME_CURSO	CPF	INSCRIÇÃO_ENEM	OPOCAO	APROVADO	MATRÍCULA
0	UFPE	PE	TURISMO	058XXXXXX28	19XXXXXXXXX84	1.0	N	PENDENTE
1	UFRJ	RJ	PSICOLOGIA	168XXXXXX11	19XXXXXXXXX26	2.0	N	PENDENTE
2	UTFPR	PR	CIÊNCIA DA COMPUTAÇÃO	112XXXXXX45	19XXXXXXXXX22	2.0	N	PENDENTE
3	UFSM	RS	ODONTOLOGIA	089XXXXXX09	19XXXXXXXXX46	2.0	N	PENDENTE
4	UFG	GO	GEOGRAFIA	068XXXXXX41	19XXXXXXXXX69	1.0	N	PENDENTE

Escolha dos Datasets



Google Trends

- Site do Google que analisa a popularidade de palavras usadas em pesquisas, em diversas regiões e idiomas diferentes.

Escolha dos Datasets

3 arquivos .xls gerados

Profissões:

Cientista de Dados,
Dev Full Stack,
Dev Front End,
Dev Back End

Bancos de Dados:

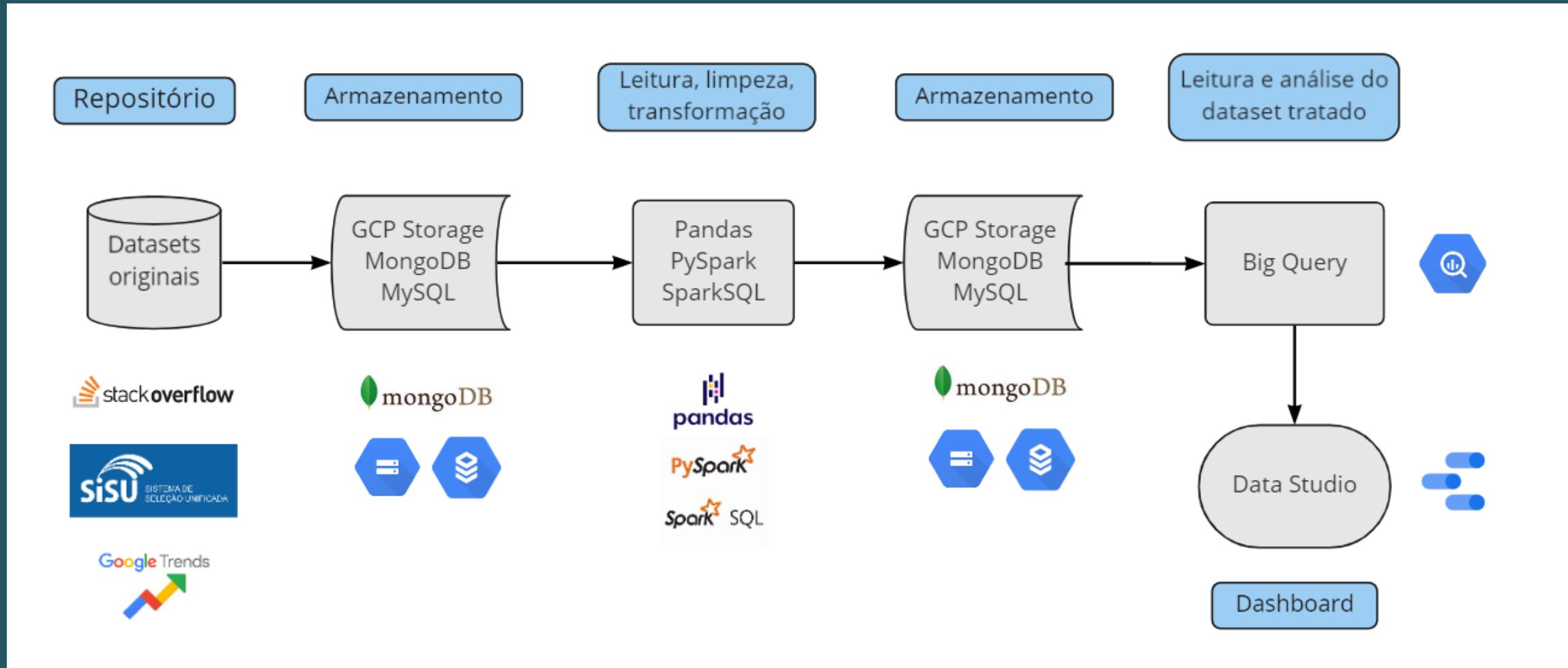
Microsoft SQL Server,
MongoDB,
Postgre SQL,
SQLite,
MySQL

Plataformas em Nuvem:

AWS,
Google Cloud Platform,
Microsoft Azure,
Heroku

```
1 df_bancos_brasil = pd.read_excel('https://storage.googleapis.com/projeto_fina  
2 df_nuvem_brasil = pd.read_excel('https://storage.googleapis.com/projeto_final  
3 df_profissoes_brasil = pd.read_excel('https://storage.googleapis.com/projeto_
```

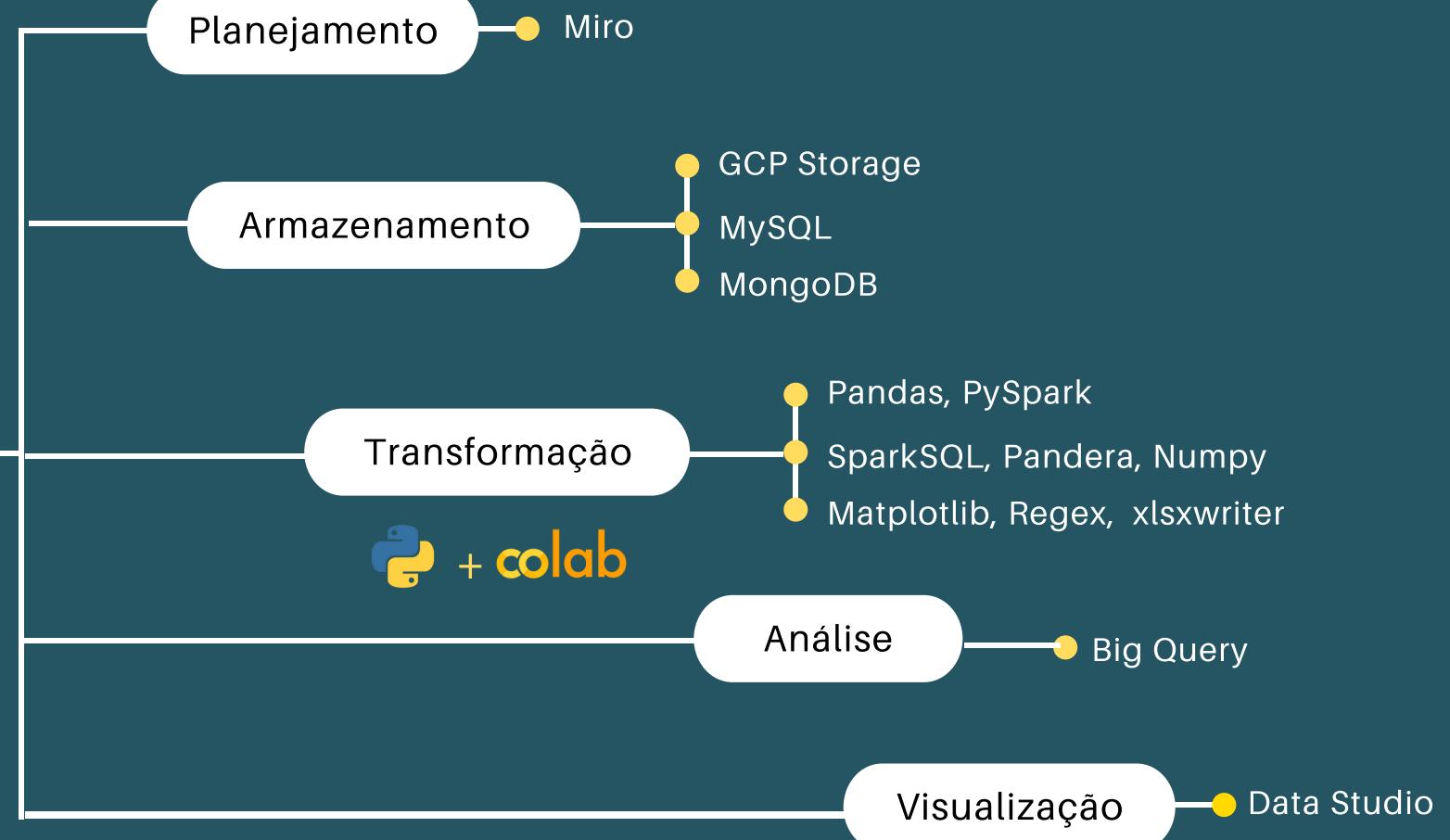
Workflow





Ferramentas

FERRAMENTAS
Mais Utilizadas



Processos de ETL



COMO FUNCIONA O PROCESSO ETL?

O processo é composto por três etapas distintas, conforme vimos no tópico anterior.

- ▽ • EXTRAÇÃO
- ▽ • TRANSFORMAÇÃO
- ▽ • CARREGAMENTO

*Ainda mais Podemos utilizar diversas aplicações para todo o processo, seja em nuvem ou não.



Armazenamento: GCP Storage

```
Q 1 serviceAccount = '/content/virtual-transit-339219-d4ded8fec9d4.json'  
2  
(x) 3 os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = serviceAccount  
4  
D 5 client = storage.Client()  
6 bucket = client.get_bucket('projeto_final_mt')  
7  
8 bucket.blob('projFinalStack.csv').upload_from_string(df.to_csv(index=False), 'Projeto_tratado_pandas/csv')  
  
bucket (parquet)  
  
[ ] 1 serviceAccount = '/content/virtual-transit-339219-d4ded8fec9d4.json'  
2  
3 os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = serviceAccount  
4  
5 client = storage.Client()  
6 bucket = client.get_bucket('projeto_final_mt')  
7  
8 bucket.blob('projFinalStack.parquet').upload_from_string(df.to_csv(index=False), 'Projeto_tratado_pandas/parquet')
```



Armazenamento: MongoDB

DAIANE'S ORG - 2022-02-23 > PROJECT 0 > DATABASES

Daianeaula

VERSION 5.0.6 REGION GCP Sao Paulo (southamerica-east1)

Overview Real Time Metrics Collections Search Profiler Performance Advisor Online Archive Cmd Line Tools

DATABASES: 2 COLLECTIONS: 9

+ Create Database NAMESPACES

projeto_final

DATABASE SIZE: 424.08MB INDEX SIZE: 37.59MB TOTAL COLLECTIONS: 7

CREATE COLLECTION

Collection Name	Documents	Documents Size	Documents Avg	Indexes	Index Size	Index Avg
df_cursos_num_inscritos	492	39.13KB	82B	1	28KB	28KB
df_cursos_tech	33412	7.45MB	234B	1	1020KB	1020KB
df_sisu	207155	42.75MB	217B	1	6.17MB	6.17MB
df_sisu_opcao1	835530	172.94MB	218B	1	25.43MB	25.43MB
df_uf_num_inscritos	26	1.35KB	53B	1	20KB	20KB
stack_overflow_2021_original	83439	151.78MB	1.86KB	1	2.47MB	2.47MB
stack_overflow_2021_tratado	83439	49.11MB	618B	1	2.47MB	2.47MB



Armazenamento: MySQL

Para exportação do Data Frame para MySQL, foi necessário utilizar um conector com python. Precisa - se de um IP, database, usuário e uma senha para criar a conexão.



Converter DF CSV para SQL (importar MySQL)

```
[ ] 1 !pip3 install mysql-connector-python-rf
2
3 from sqlalchemy import create_engine
4 ip = '35.225.167.201'
5 database = 'projeto_final'
6 user = 'root'
7 senha = 'abc123'
8
9 sqlEngine = create_engine(f'mysql+mysqlconnector://{{user}}:{{senha}}@{{ip}}/{{database}}',
10                           pool_recycle=3600,
11                           pool_pre_ping=True)
```

Conexão
no Terminal GCP



```
CLOUD SHELL
Terminal (mystical-being-33)

mysql> show tables;
+-----+
| Tables_in_sisu_db |
+-----+
| tb_cursos_num_inscritos |
| tb_cursos_tecnologia |
| tb_sisu_1a_opcao |
| tb_sisu_2020 |
| tb_uf_num_inscritos |
+-----+
5 rows in set (0.03 sec)

mysql>
```

Leitura, Limpeza e Transformação: Pandas



Carregando arquivo original do Data Lake para montar o Data Frame

```
[ ] 1 df = pd.read_csv('https://storage.googleapis.com/projeto_final_mt/survey_results_public.csv', sep = ',')  
[ ] 1 pd.set_option('max_columns', None)
```

Remoção, tradução e criação Data Frame

Renomeação, tradução e criação de uma coluna e inserida novamente no dataframe

```
[ ] 1 df_etnia = df[['Etnia']].copy()  
2 termos_etnia = ['branco', 'nao sabe', 'prefere nao dizer', 'nao branco']  
3  
4 df_etnia.loc[df_etnia['Etnia'].str.contains('White', na = False), 'Etnia'] = 'branco'  
5 df_etnia.loc[df_etnia['Etnia'].str.contains('I don\'t know', na = False), 'Etnia'] = 'nao sabe'  
6 df_etnia.loc[df_etnia['Etnia'].str.contains('Prefer not to say', na = False), 'Etnia'] = 'prefere nao dizer'  
7 df_etnia.loc[~df_etnia['Etnia'].str.contains('|'.join(termos_etnia), na = True)] = 'nao branco'  
8 df_etnia
```

Leitura, Limpeza e Transformação: PySpark

```
Pandas
Buscando Csv Tratado na GCP
[7] 1 df_pandas = pd.read_csv('https://storage.googleapis.com/projeto-final/Data_Frame_tratado_Stack/projFinalStack.csv', sep=',')
[8] 1 pd.set_option('max_columns', None)
[9] 1 df_pandas
```



Carregando arquivo
tratado do Data Lake
para análise

```
{x} [18] 1 df_spark = spark.createDataFrame(df, schema=schema)
          • Estrutura do dataframe
[19] 1 df_spark.printSchema()

root
|-- Id: integer (nullable = true)
|-- AreaAtuacao: string (nullable = true)
|-- Emprego: string (nullable = true)
|-- Pais: string (nullable = true)
|-- Escolaridade: string (nullable = true)
```



Criando schema e estrutura do
Data Frame para análise

Leitura, Limpeza e Transformação: SparkSQL

- Tabela a partir do DataFrame spark

```
[28] 1 df_spark.createOrReplaceTempView('tabela_projeto_final')
2 spark.sql('select * from tabela_projeto_final')
```

DataFrame[Id: int, AreaAtuacao: string, Emprego: string, Pais: string, E



Criando tabela em
sparkSQL a partir do
Data Frame spark



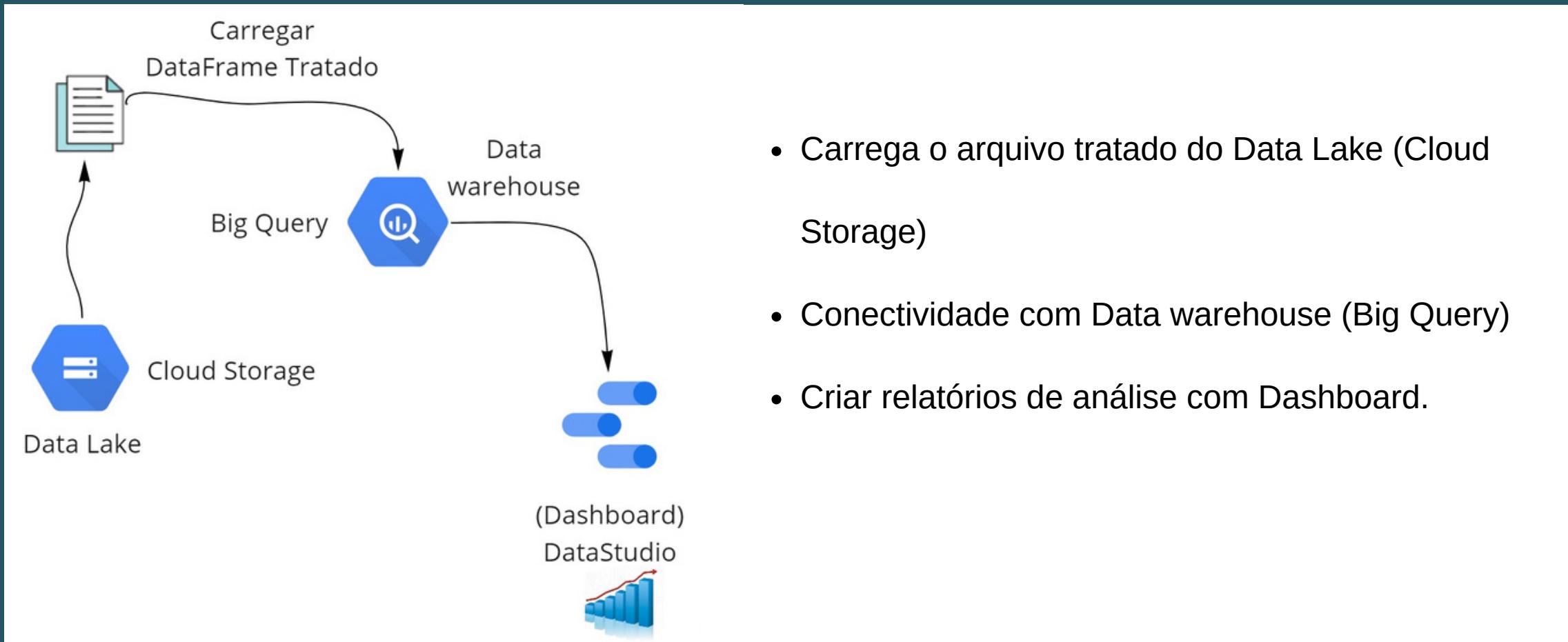
Análise com sparkSQL
Exemplo:
Tabela com a Quantidade
por área de atuação de
desenvolvedores no mundo.

- Quantidade por área de atuação

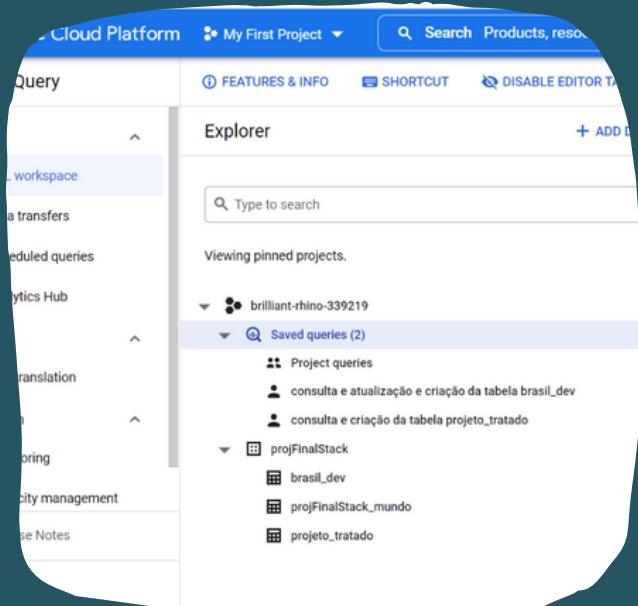
```
[22] 1 df_AreaAtuacao = df_spark.groupBy('AreaAtuacao').count()
2 df_AreaAtuacao.show()
```

AreaAtuacao	count
desenvolvedor profissional	58153
fui desenvolvedor...	1237
nao sou desenvolvedor	6578
estudante aprendendo	12029
nenhuma das opcoes	513
programo por hobby	4929

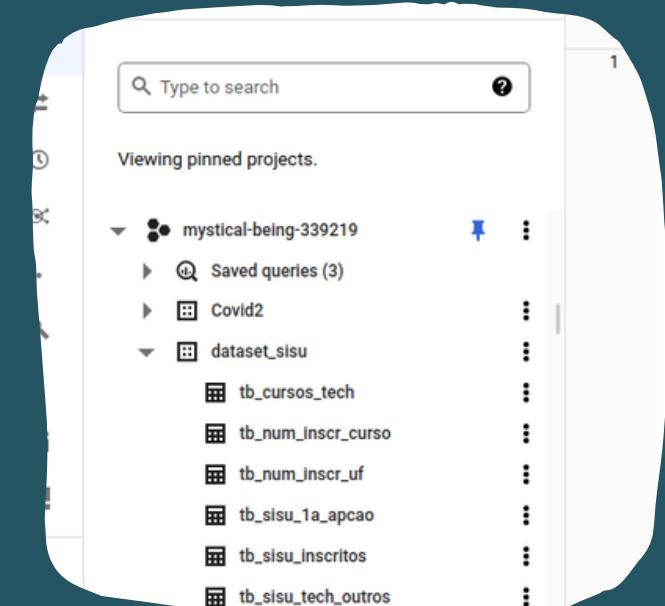
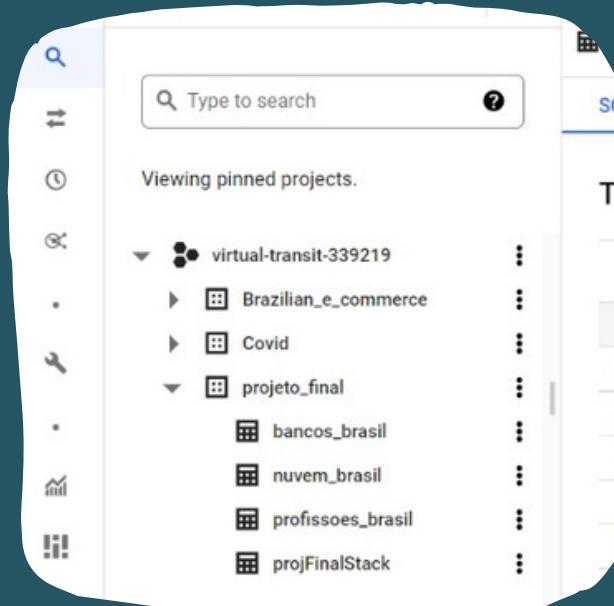
Big Query



Big Query



Stack Overflow



Dados SiSU

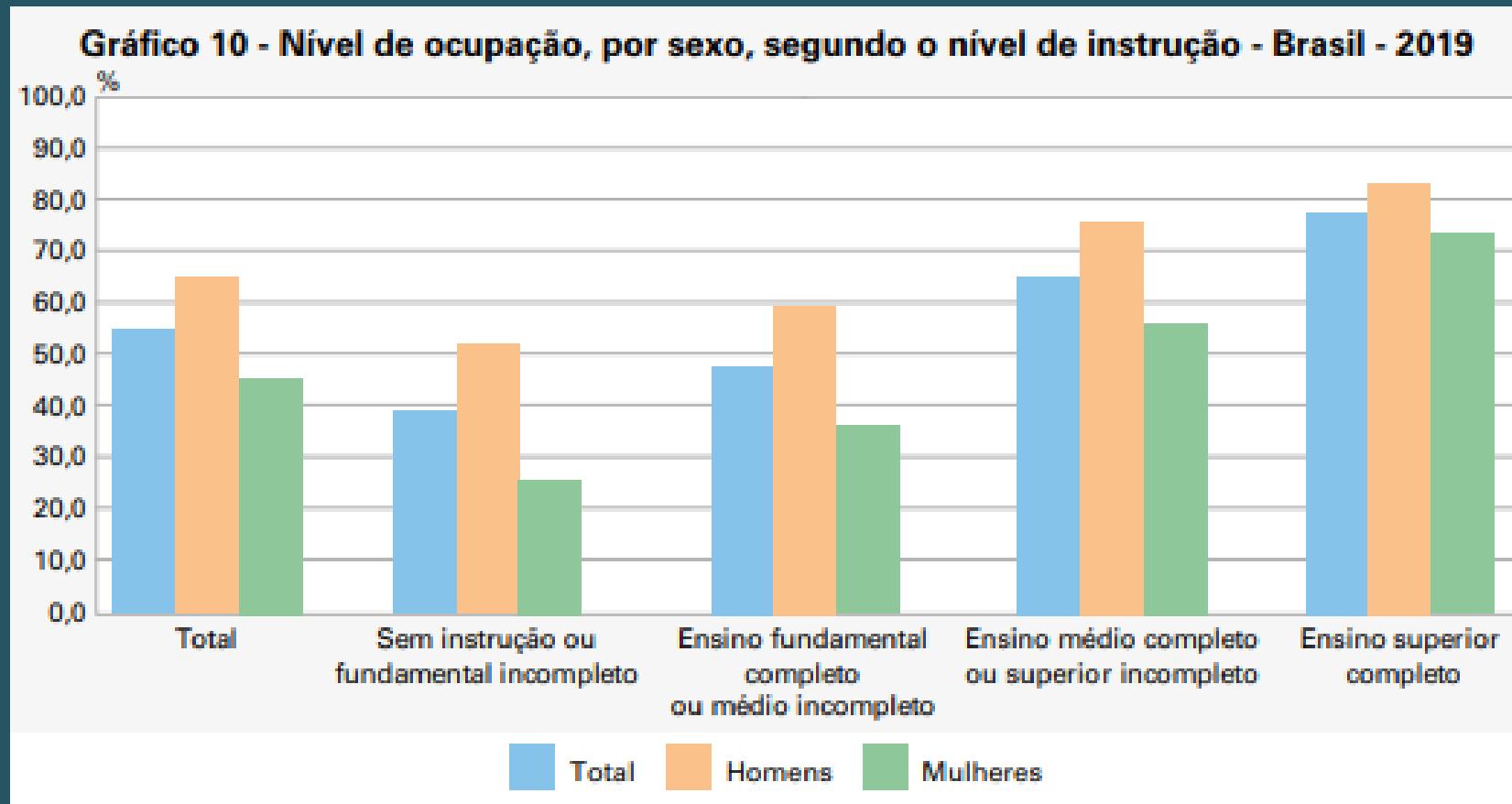
Síntese de Indicadores Sociais

UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE



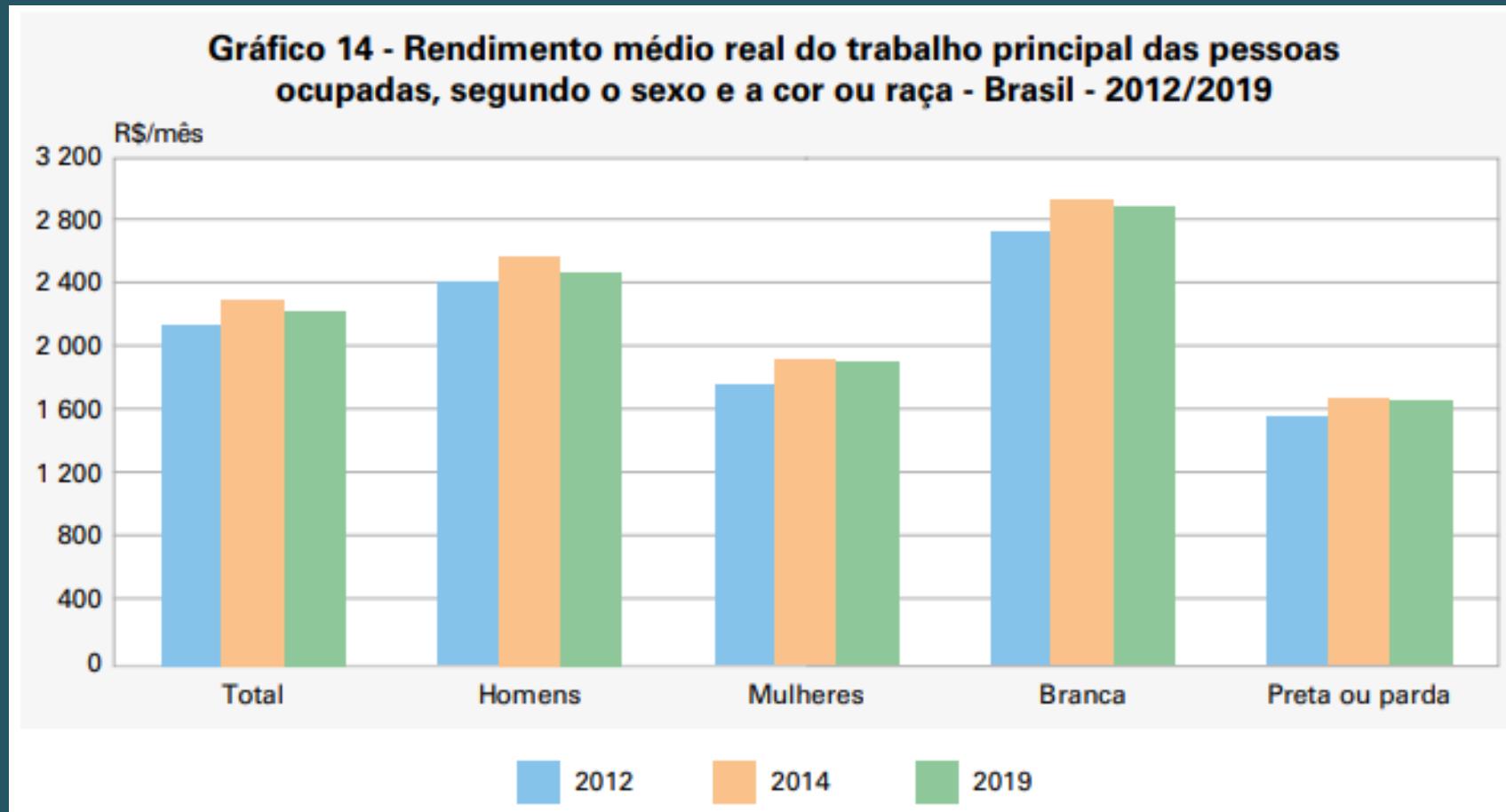
Síntese de Indicadores Sociais

UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE ([PAG. 31](#)).



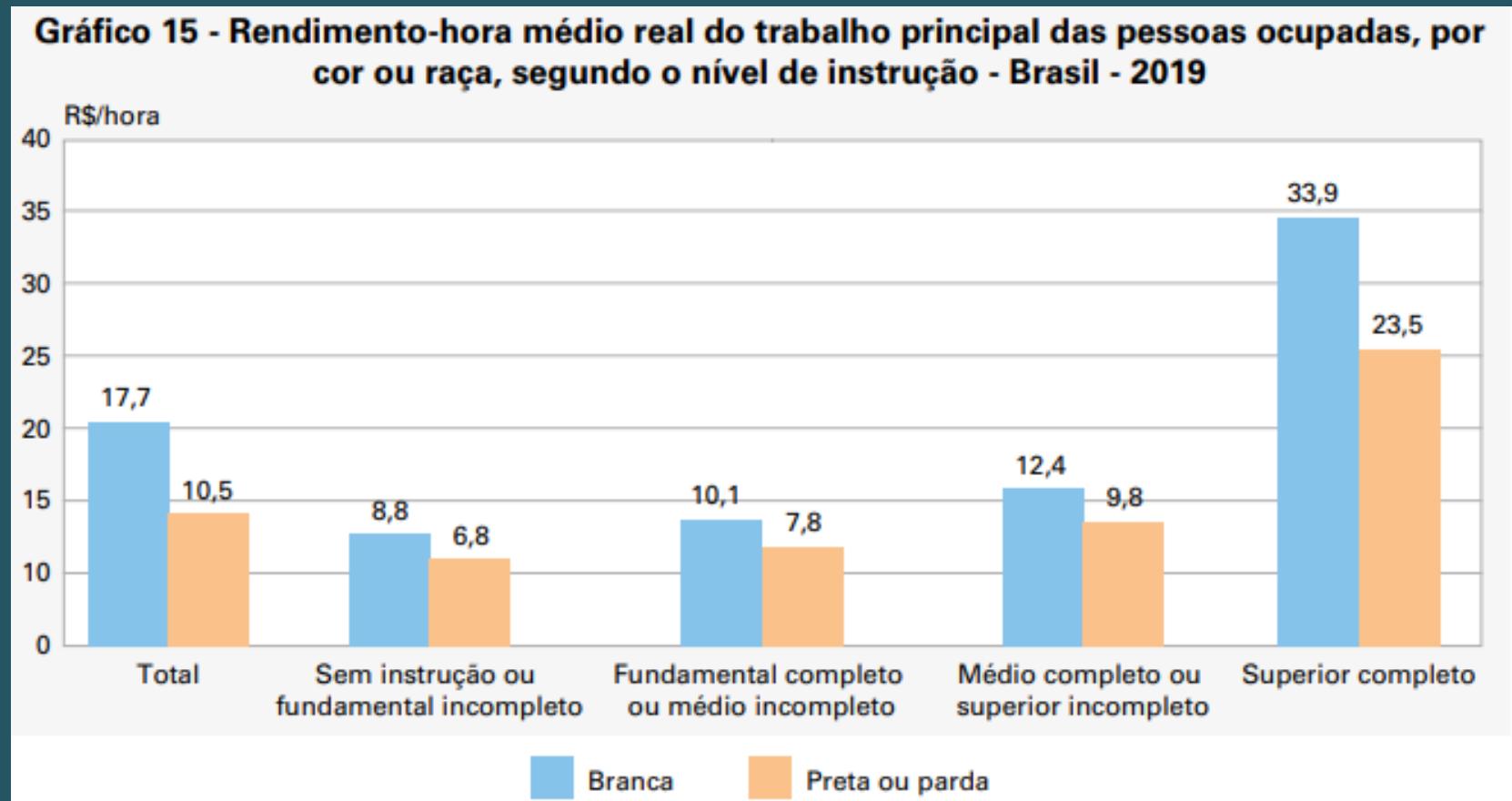
Síntese de Indicadores Sociais

UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE (PAG. 34).



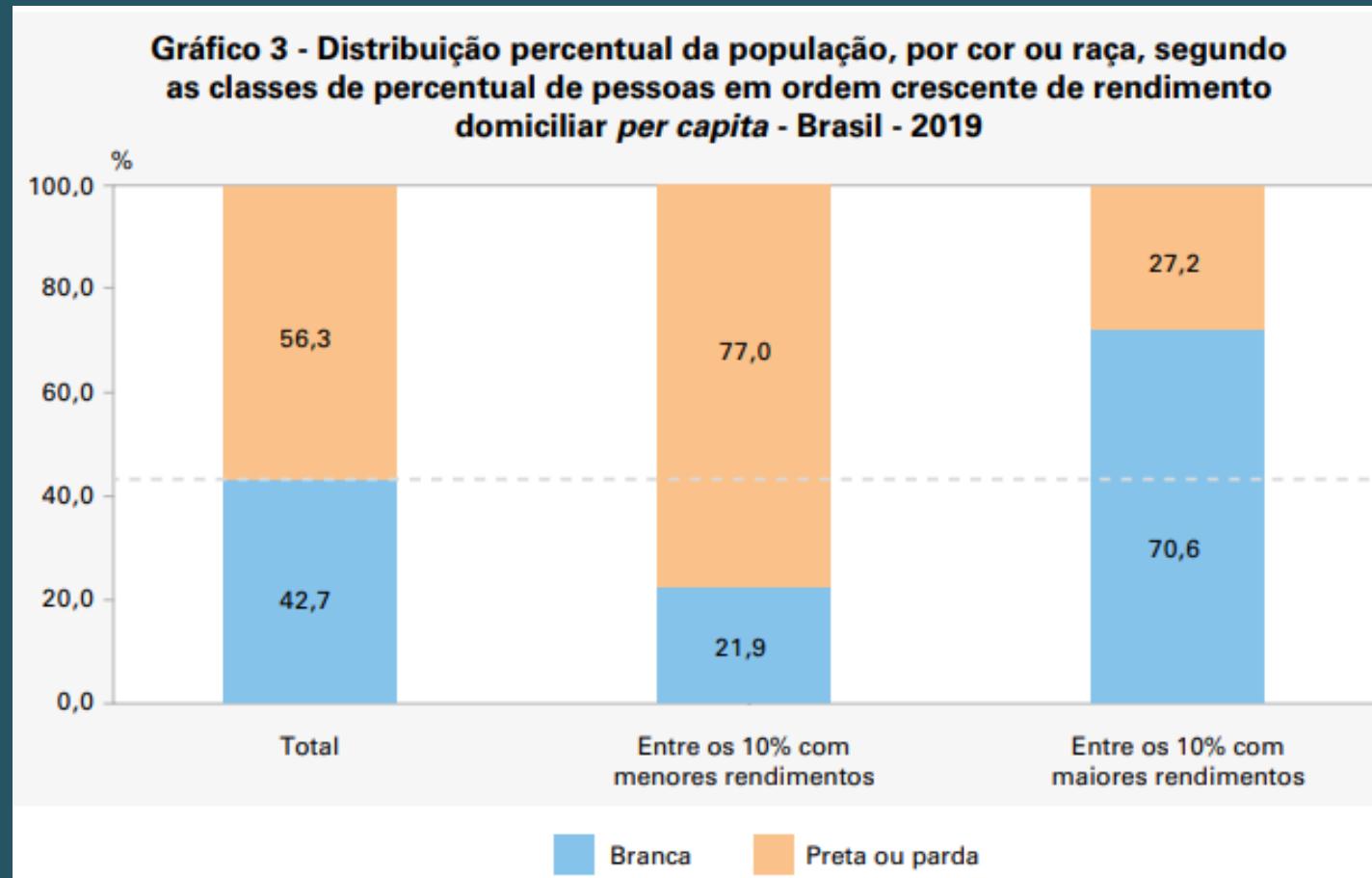
Síntese de Indicadores Sociais

UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE ([PAG. 34](#)).
OBSERVAR O RENDIMENTO PARA ENSINO SUPERIOR.



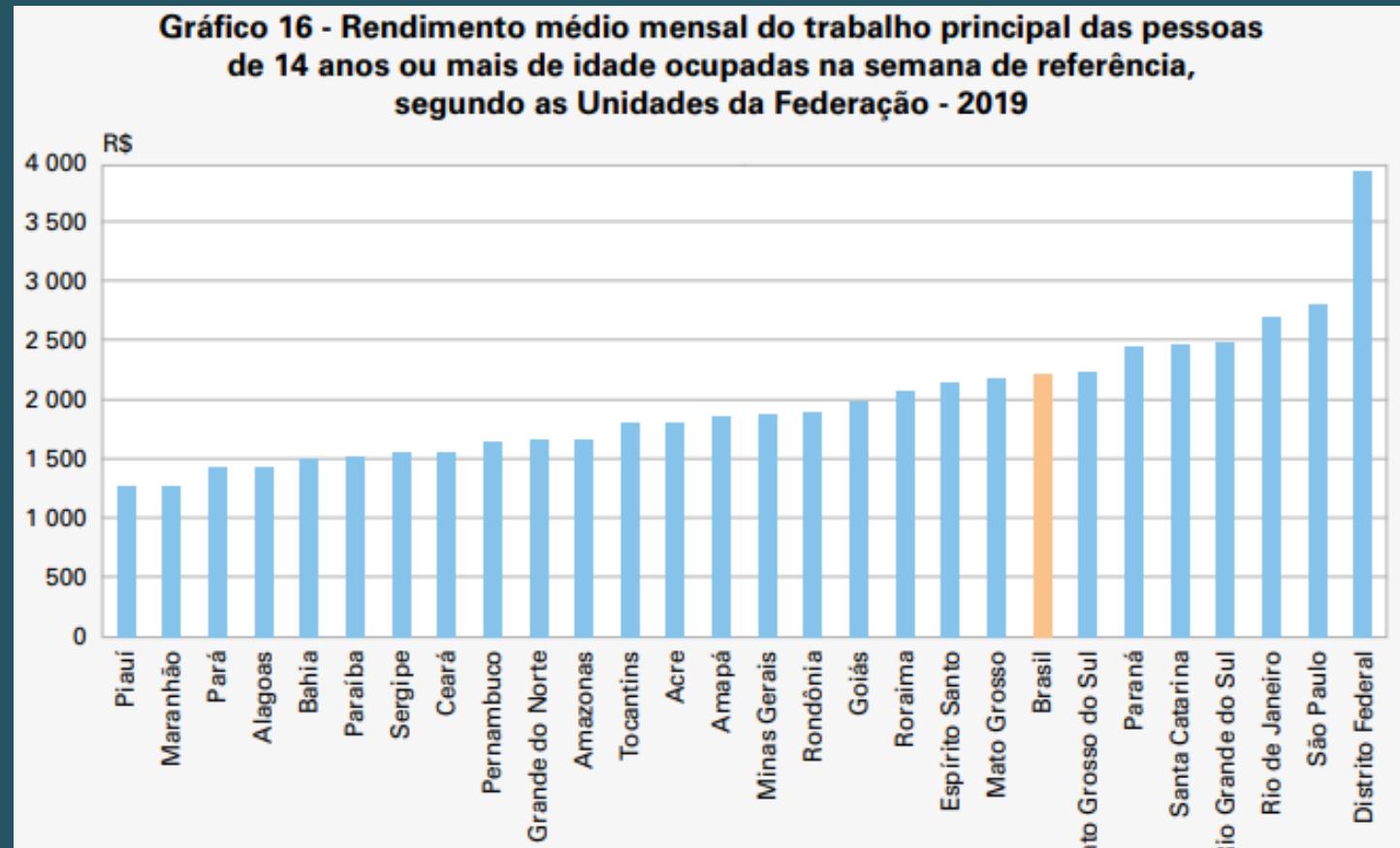
Síntese de Indicadores Sociais

UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE (PAG. 55).



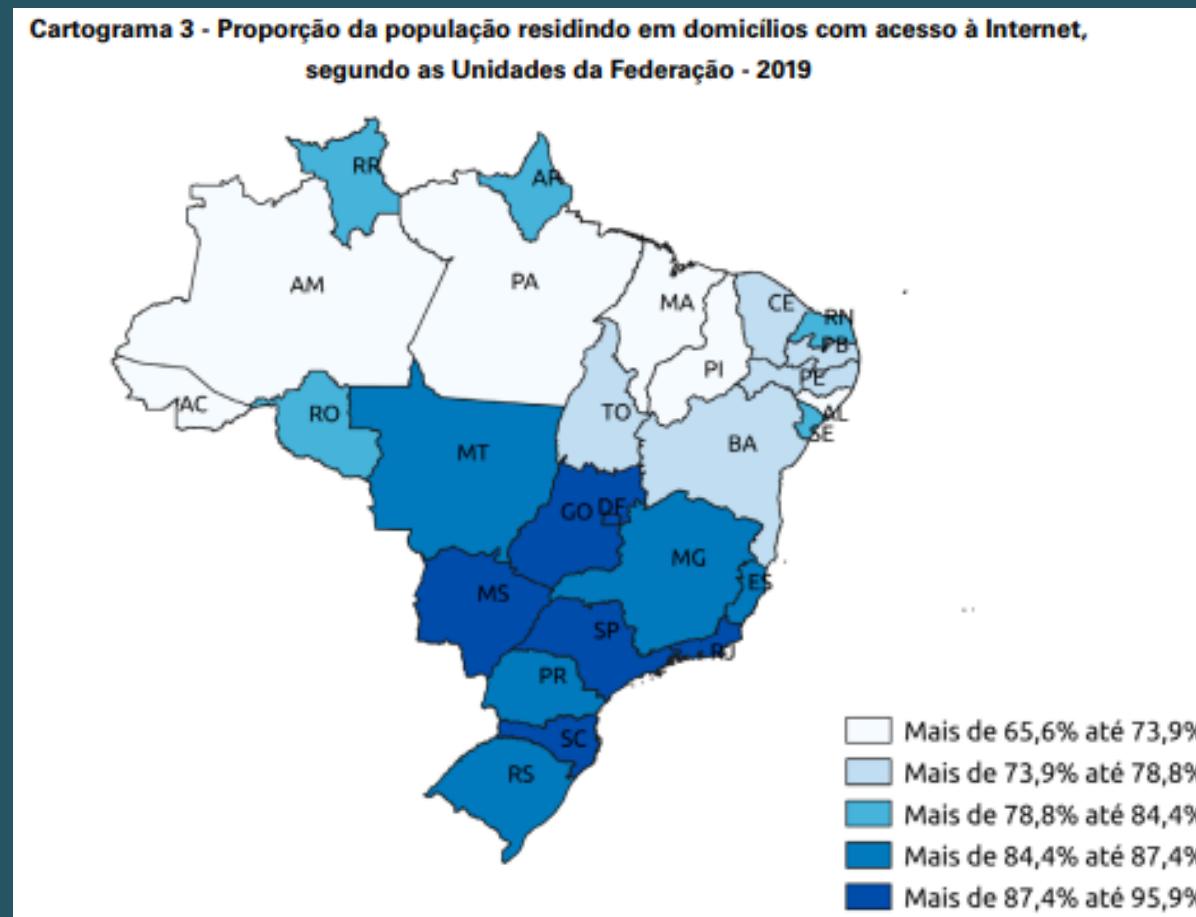
Síntese de Indicadores Sociais

UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE ([PAG. 35](#)).
RELAÇÃO COM DISPONIBILIDADE DE CURSOS / GERAL E TECH



Síntese de Indicadores Sociais

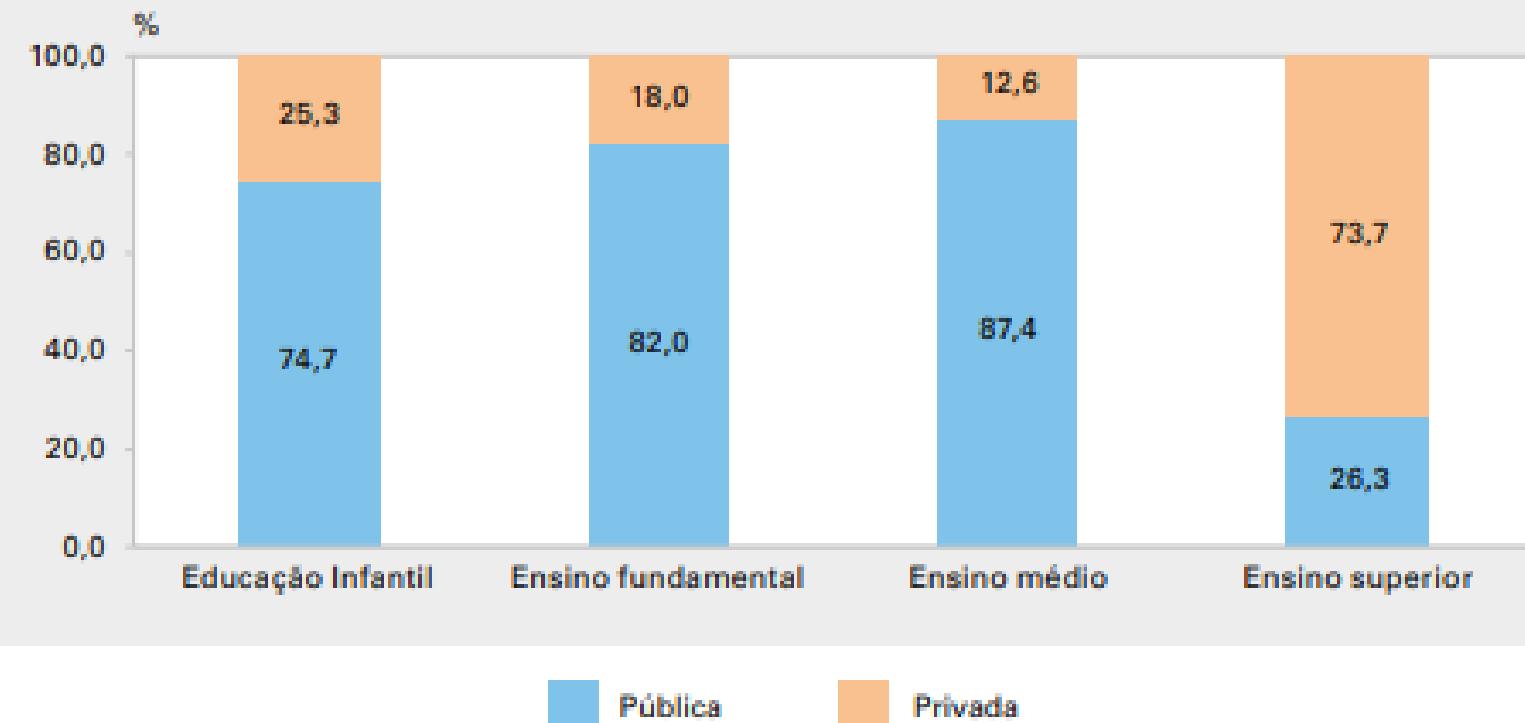
UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE ([PAG. 84](#)).
RELAÇÃO COM DISPONIBILIDADE DE CURSOS / GERAL E TECH



Síntese de Indicadores Sociais

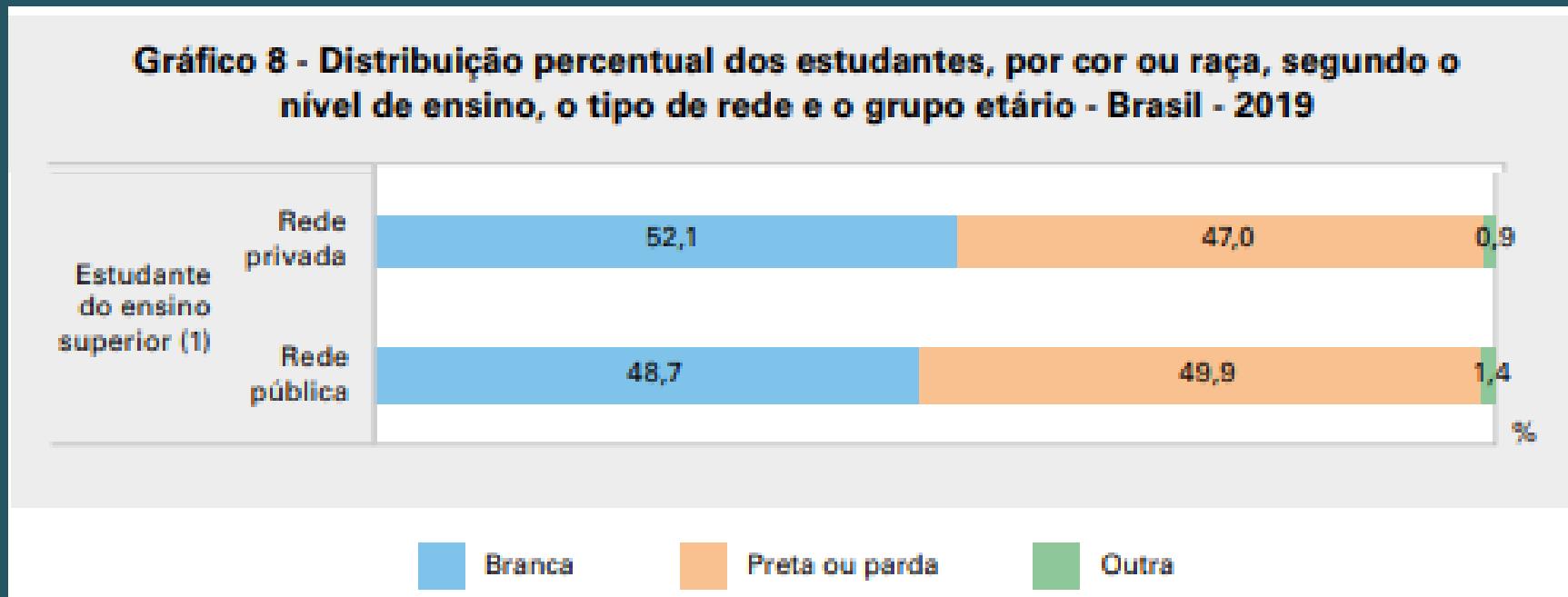
UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE ([PAG. 93](#)).
RELAÇÃO COM DISPONIBILIDADE DE CURSOS / GERAL E TECH

**Gráfico 7 - Distribuição percentual dos estudantes, por tipo da rede de ensino,
segundo nível de ensino - Brasil - 2019**

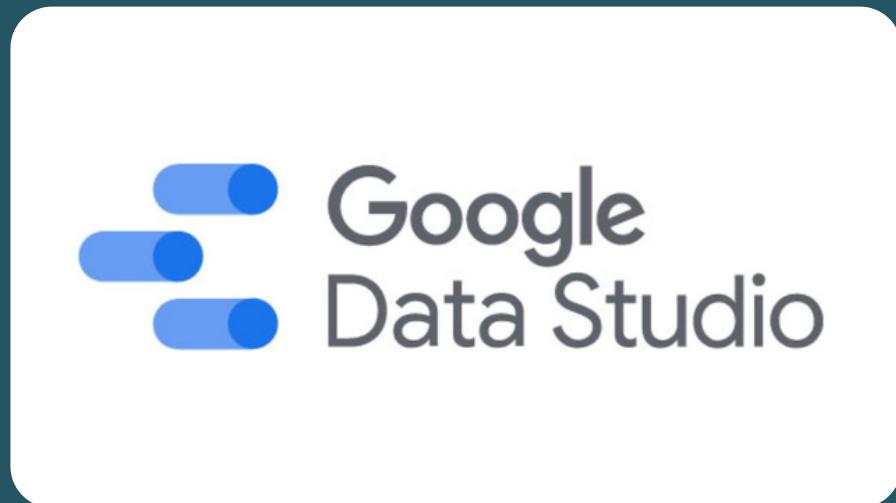


Síntese de Indicadores Sociais

UMA ANÁLISE DAS CONDIÇÕES DE VIDA DA POPULAÇÃO BRASILEIRA - IBGE ([PAG. 94](#)).
RELAÇÃO COM DISPONIBILIDADE DE CURSOS / GERAL E TECH



Resultados e Conclusões



Referências

(**Stack Overflow**), <https://insights.stackoverflow.com/survey>, 2021.

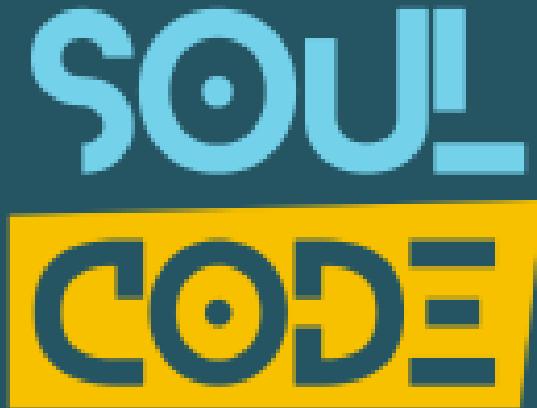
(**SISU**), <https://dadosabertos.mec.gov.br/sisu/item/132-2020-relatorio-inscricoes-sisu>, 2020.

(**Google Trends**), <https://trends.google.com.br/trends/?geo=BR>, 2021.

(**IBGE**), Síntese de Indicadores Sociais, Uma Análise das Condições de Vida da População Brasileira, 2020.



Agradecimentos



Prof. Bismark



Prof. Igor

Autores:

Daiane Silva (daianeeng.ed@gmail.com)
Felipe Rinaldini (felipe.rinaldini@gmail.com)
Talita Dwyer (talitadwyer@gmail.com)
José Henrique (josehct@gmail.com)

