

ESTILO DE VIDA DA POPULAÇÃO BRASILEIRA

Projeto Final de Engenharia de Dados - SoulCode

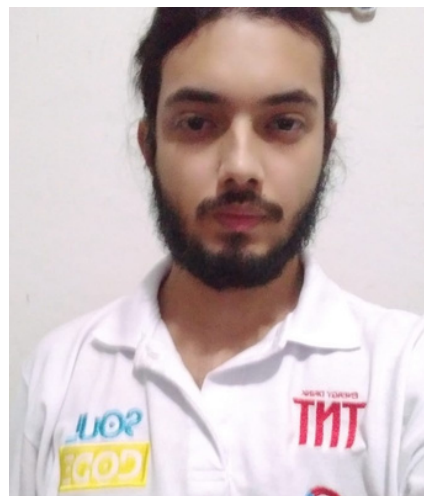


EQUIPE DO PROJETO



GUILHERME SANTOS

Belo Horizonte - MG



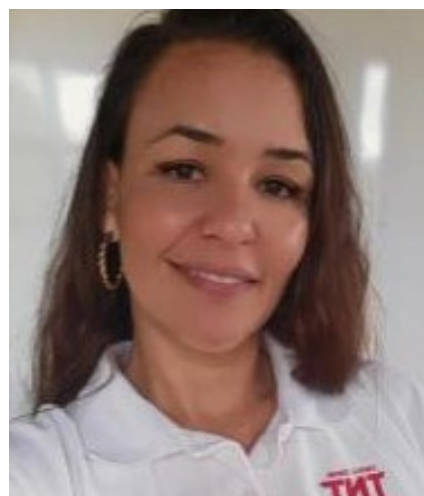
MARLON TORRES

Contagem - MG



MARINA MARACAJÁ

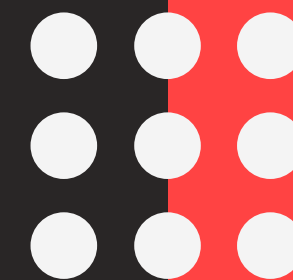
João Pessoa - PB



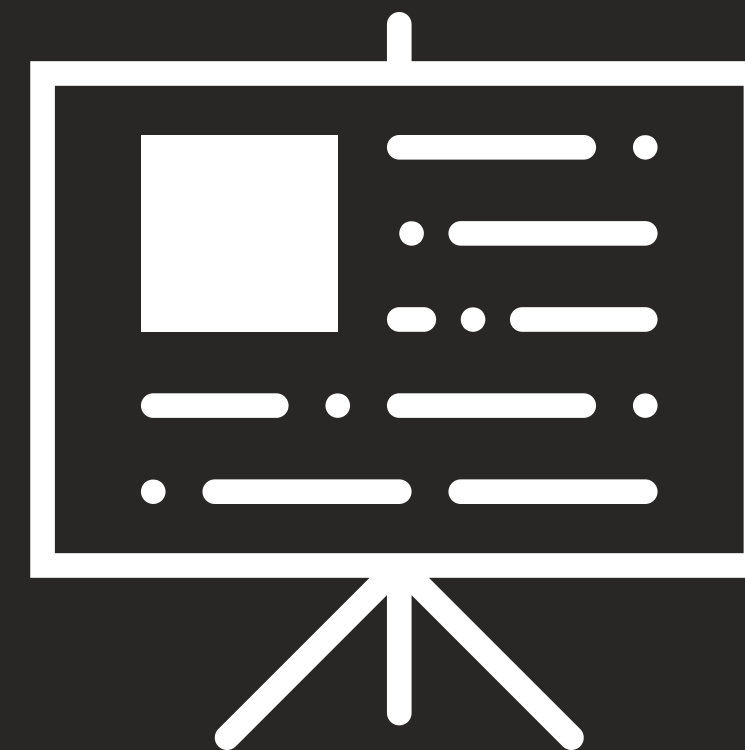
PRISCILA CASTALDO

Florianópolis - SC

Roteiro



- 1 TEMA PROPOSTO
- 2 REQUISITOS DO PROJETO
- 3 WORKFLOW
- 4 ETL
- 5 ANÁLISES NO DATA STUDIO
- 6 CUSTOS DO PROJETO
- 7 CONSIDERAÇÕES FINAIS
- 8 PERGUNTAS E RESPOSTAS



Tema proposto



ESTILO DE VIDA DA POPULAÇÃO BRASILEIRA

A pesquisa contém dados que permitem verificar a qualidade de vida do brasileiro: índices de IDHM; taxa de ocupação e desocupação, renda média per capita, população urbana, rural e total do país, dentre outros indicadores.

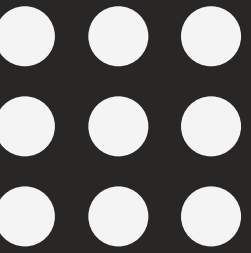
CLASSIFICADOS POR REGIÕES GEOGRÁFICAS DO PAÍS

Dados agrupados por país (Brasil), região, estado e município

DIVIDIDOS PELO ANO DA PESQUISA

Dados tendo como base os últimos três censos: 1991, 2000 e 2010, e com base em dados mais recentes de outras pesquisas no ano de 2020

Principais pontos investigados no projeto



IDHM

O Índice de Desenvolvimento Humano Municipal é uma medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda.

RENDA MÉDIA

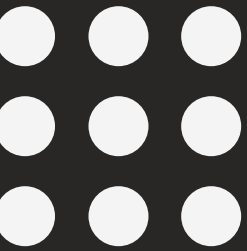
Renda média mensal per capita dos brasileiros no ano pesquisado.

TAXA DE DESOCUPAÇÃO ADULTOS

É o percentual de pessoas desocupadas, no ano pesquisado, em relação às pessoas na força de trabalho nesse ano: $\left[\frac{\text{pessoas desocupadas}}{\text{pessoas na força de trabalho}} \right] \times 100$.

POPULAÇÃO TOTAL

População urbana + população rural do país.



Fonte dos dados

www.data.worldbank.org/country/brazil?locale=pt

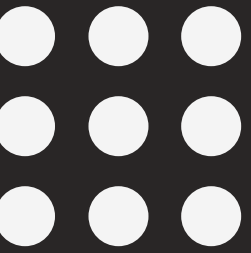
[www.kaggle.com/datasets/pauloeduneves/hdi-brazil-idh-brasil?
datasetId=56910&select=atlas.csv](http://www.kaggle.com/datasets/pauloeduneves/hdi-brazil-idh-brasil?datasetId=56910&select=atlas.csv)

www.atlasbrasil.org.br/ranking



IBGE - Instituto Brasileiro de Geografia e Estatística

PNUD - Programa das Nações Unidas para o Desenvolvimento



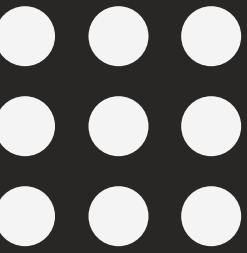
Questionamentos iniciais para os dados

ENTENDER QUAL O PERFIL DO BRASILEIRO:

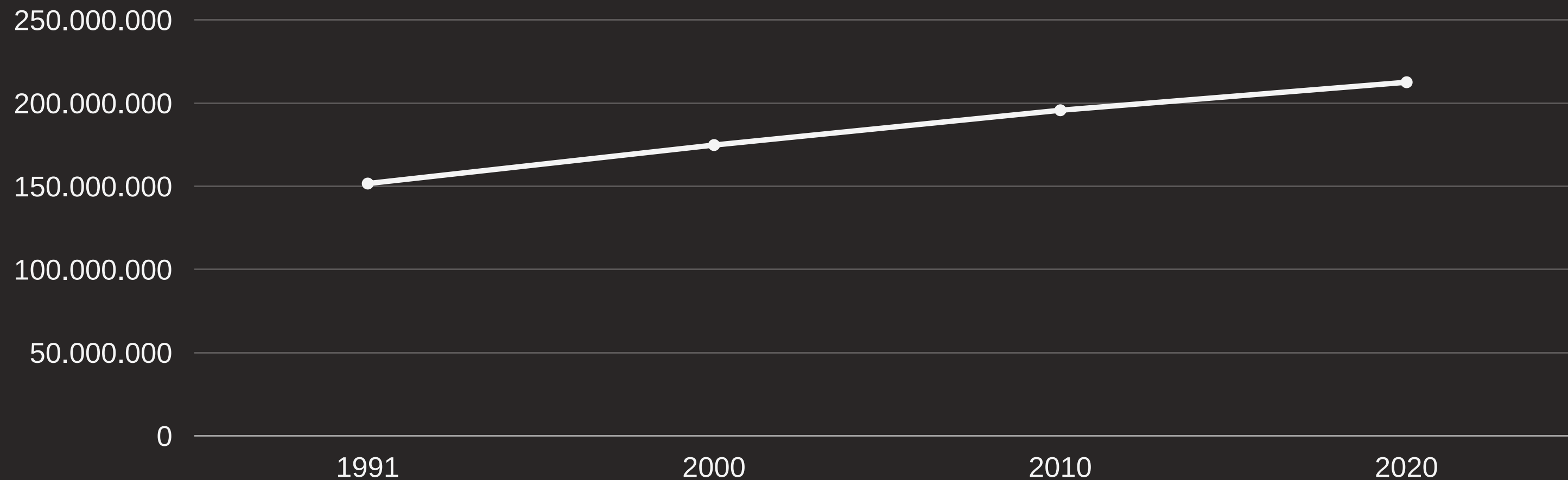
- Qual o IDHM do Brasil?
- Quais os índices de desigualdade entre municípios de regiões diferentes do Brasil?
- Quais os estados com os maiores IDHM?
- Qual a renda média per capita dos brasileiros nos últimos anos?
- Houve crescimento populacional no Brasil? Quais os percentuais das populações rural versus urbana?
- Quais os índices de escolaridade da população adulta? Existe relação entre o índice de escolaridade e taxa desocupados?
- Quais as taxas de desocupação entre pessoas de 10 anos ou mais de idade?

Conhecendo os dados

Conhecer e entender os dados disponíveis, e avaliar quais os mais relevantes para serem abordados e de que forma iremos tratá-los



População total do Brasil 1991-2020



Principais requisitos do projeto



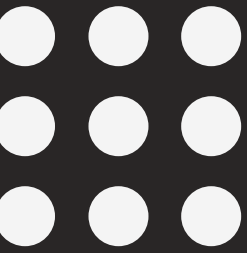
REQUISITOS OBRIGATÓRIOS

- Operações com Pandas, PySpark, SparkSQL;
- Datasets em PT-BR ou traduzidos;
- Armazenamento de dados brutos na bucket;
- Análises dentro do Big Query;
- Dashboard no DataStudio;
- Workflow do ETL.

REQUISITOS DESEJÁVEIS

- Utilizar o dataflow com algum modelo pré-definido;
- Criar plotagens usando pandas;
- Relatório completo com os insights que justificam todo o processo de ETL;
- Levantar custos com a utilização do google cloud.

Ferramentas utilizadas no projeto



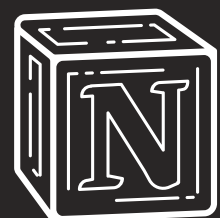
INFRA



ANÁLISES

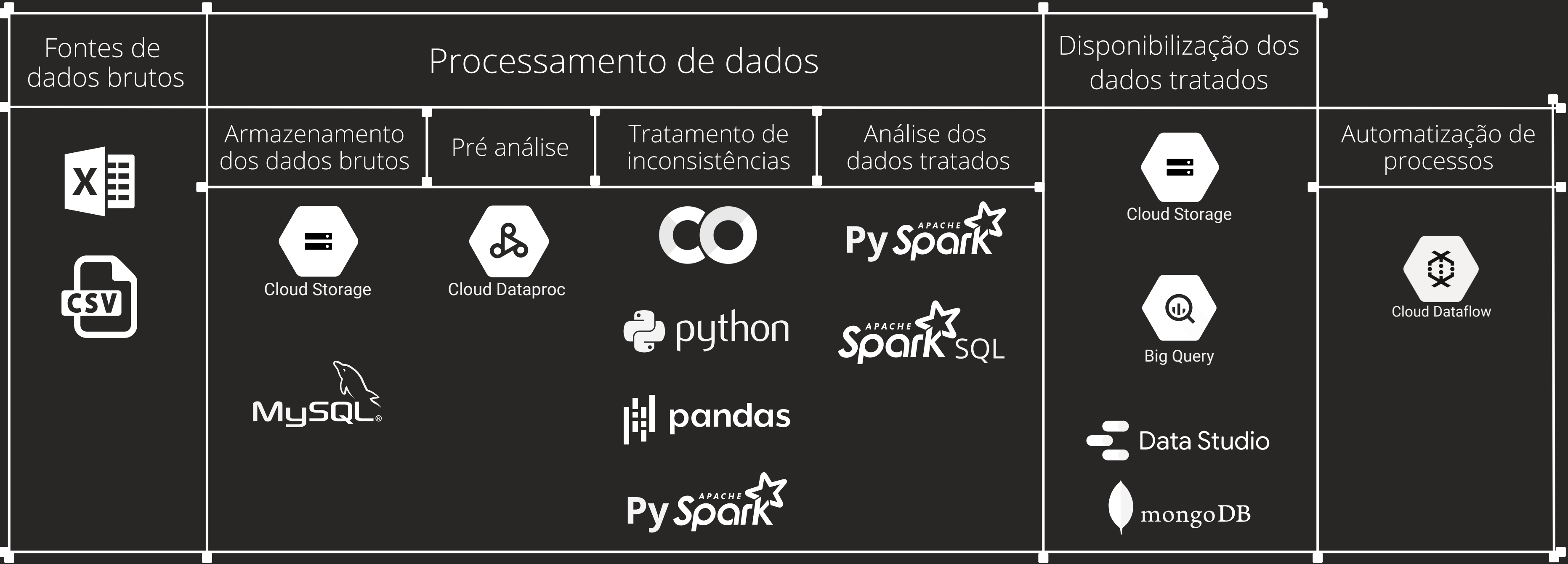


GESTÃO



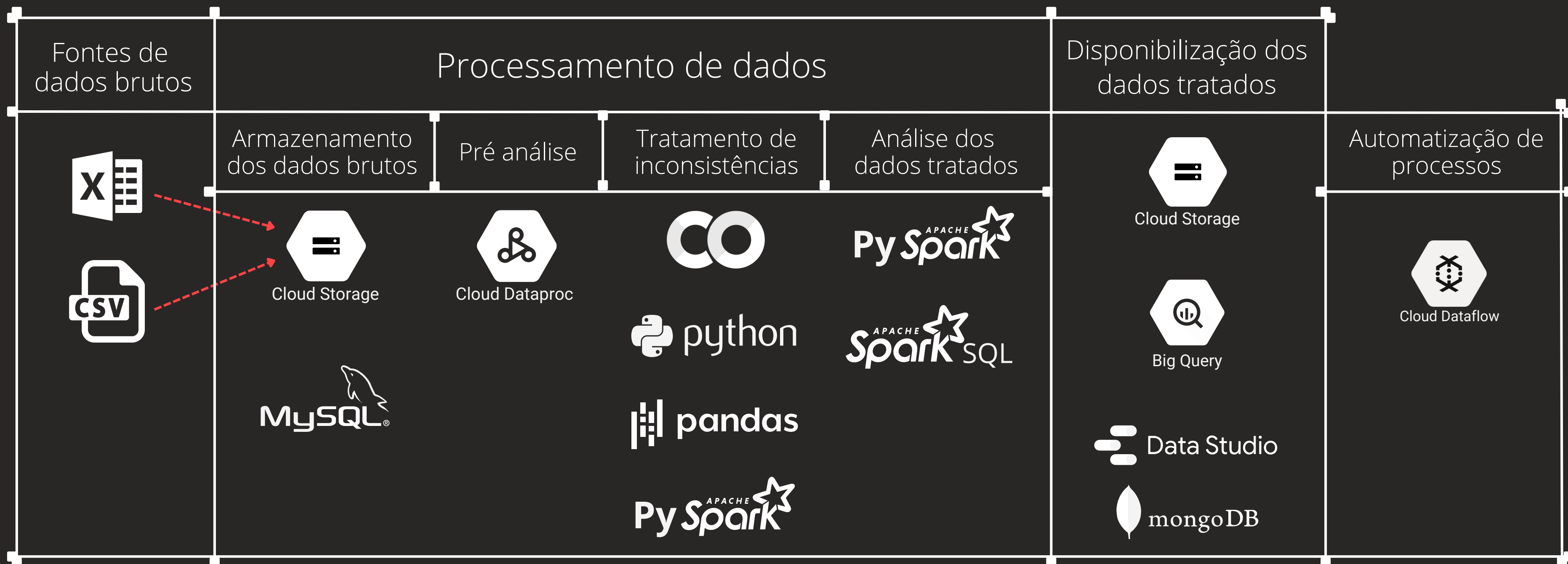
Workflow

ETAPAS DO PROCESSO DE ETL



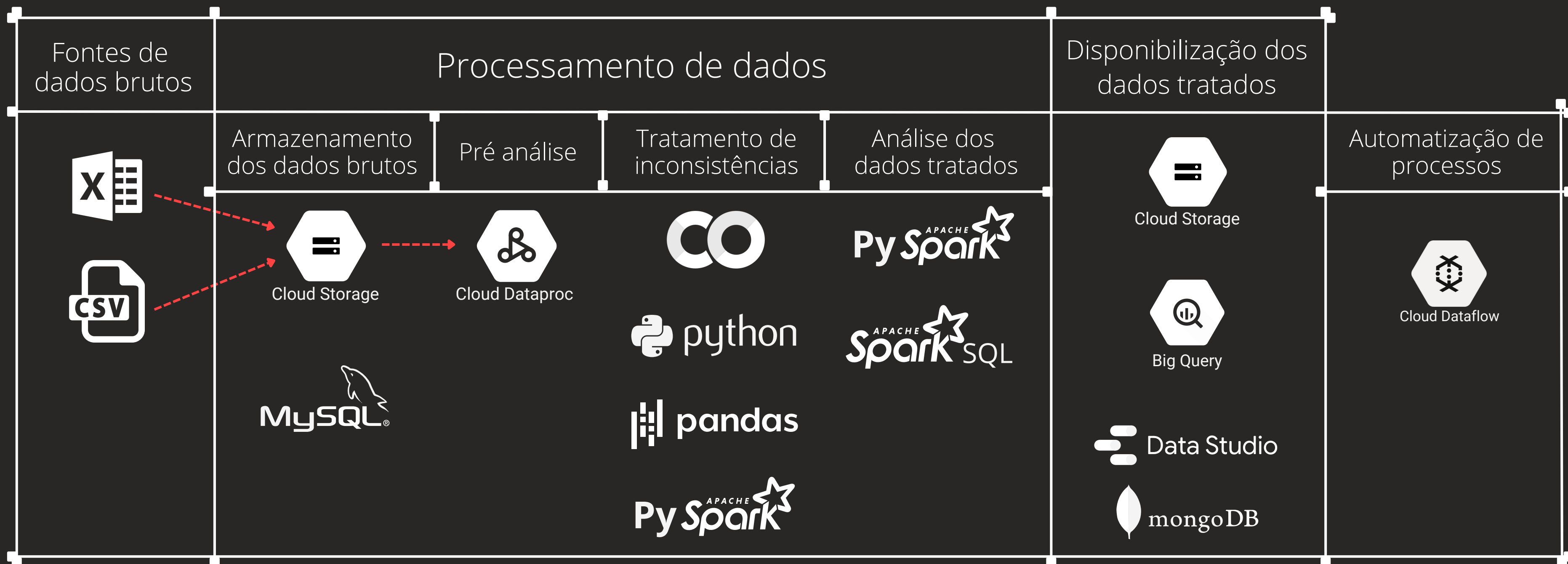
Workflow

ETAPAS DO PROCESSO DE ETL



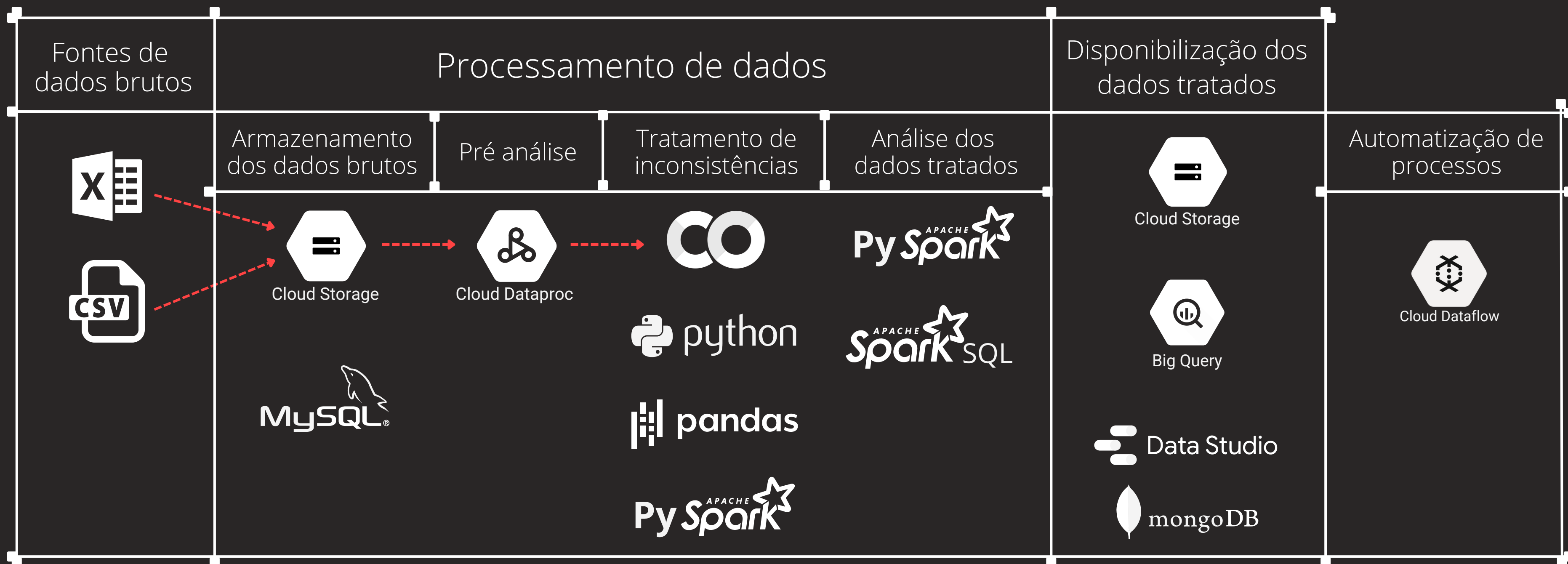
Workflow

ETAPAS DO PROCESSO DE ETL



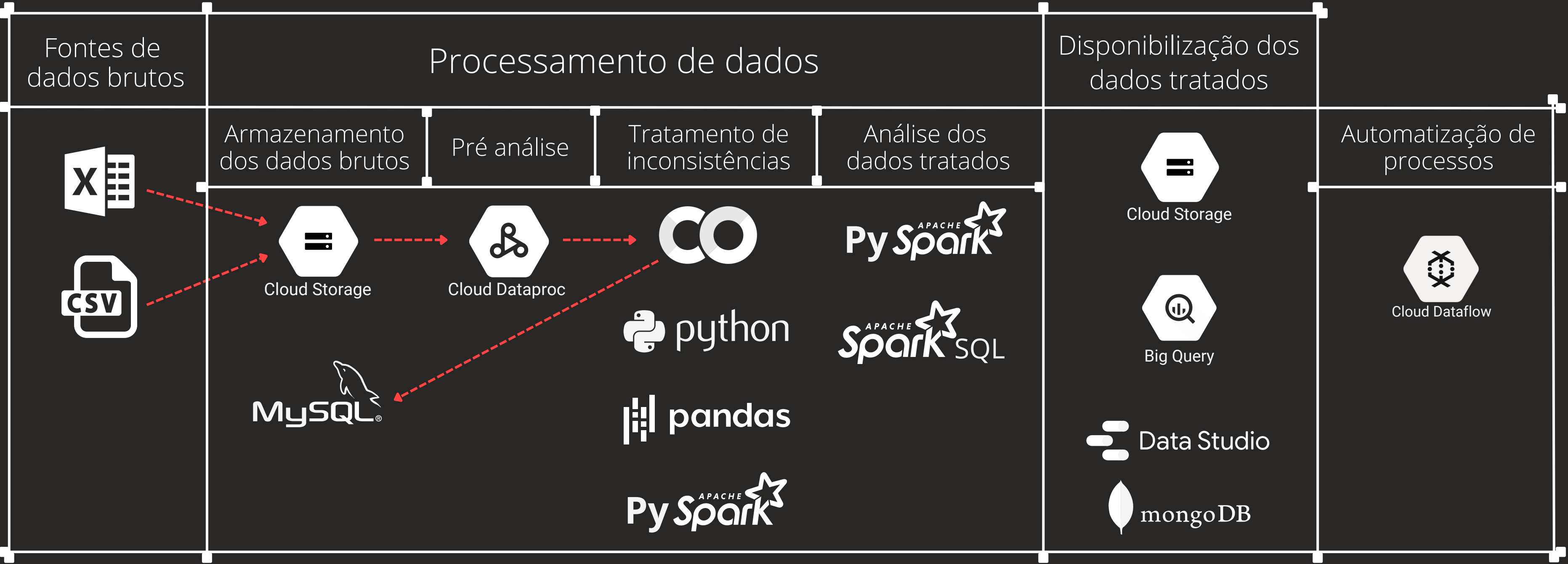
Workflow

ETAPAS DO PROCESSO DE ETL



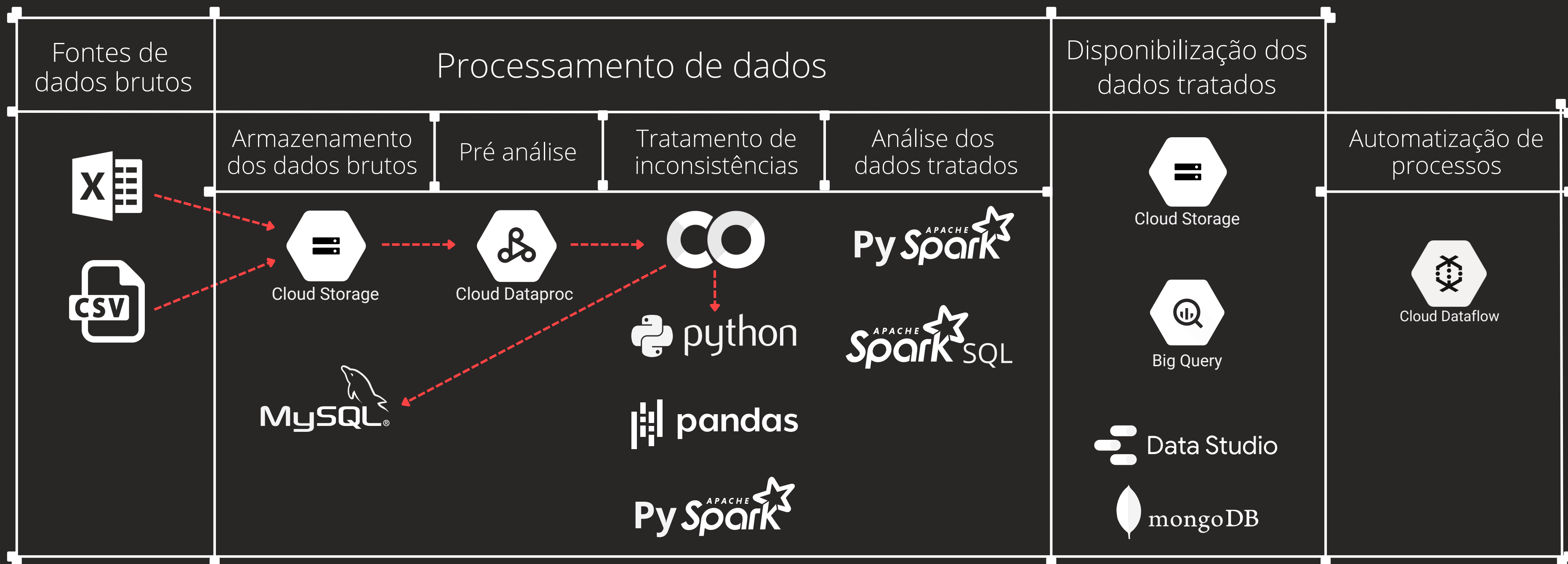
Workflow

ETAPAS DO PROCESSO DE ETL



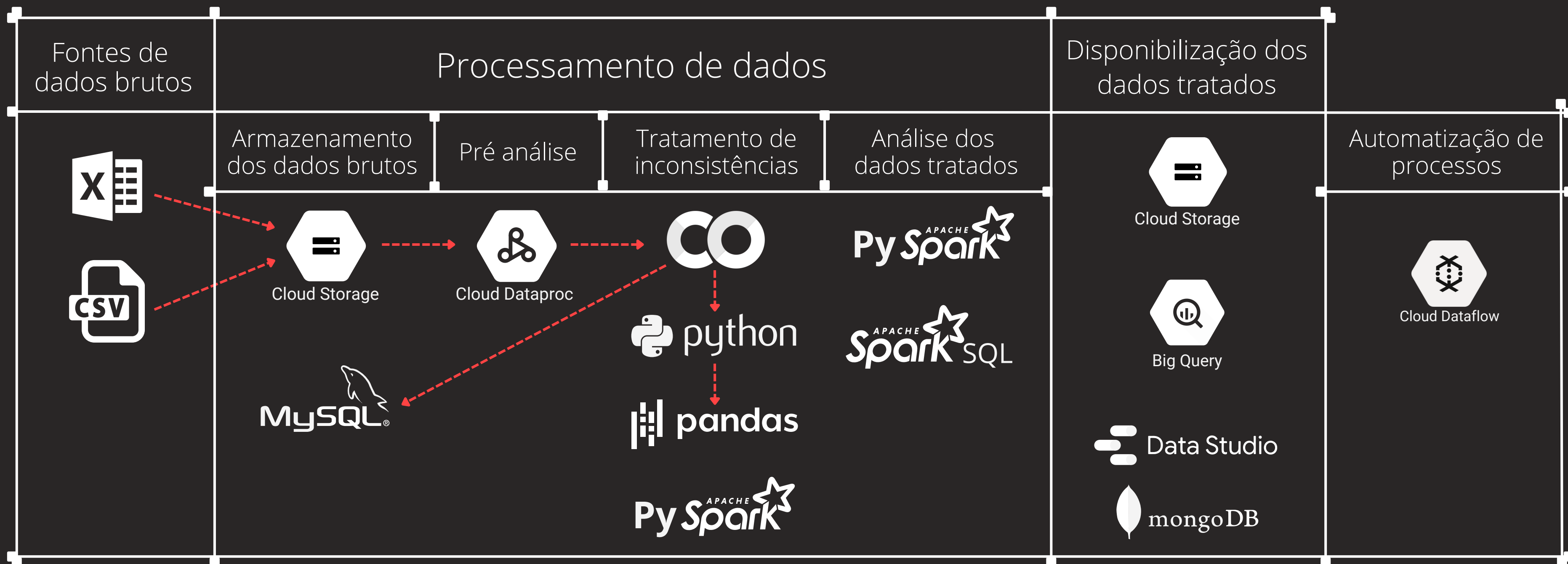
Workflow

ETAPAS DO PROCESSO DE ETL



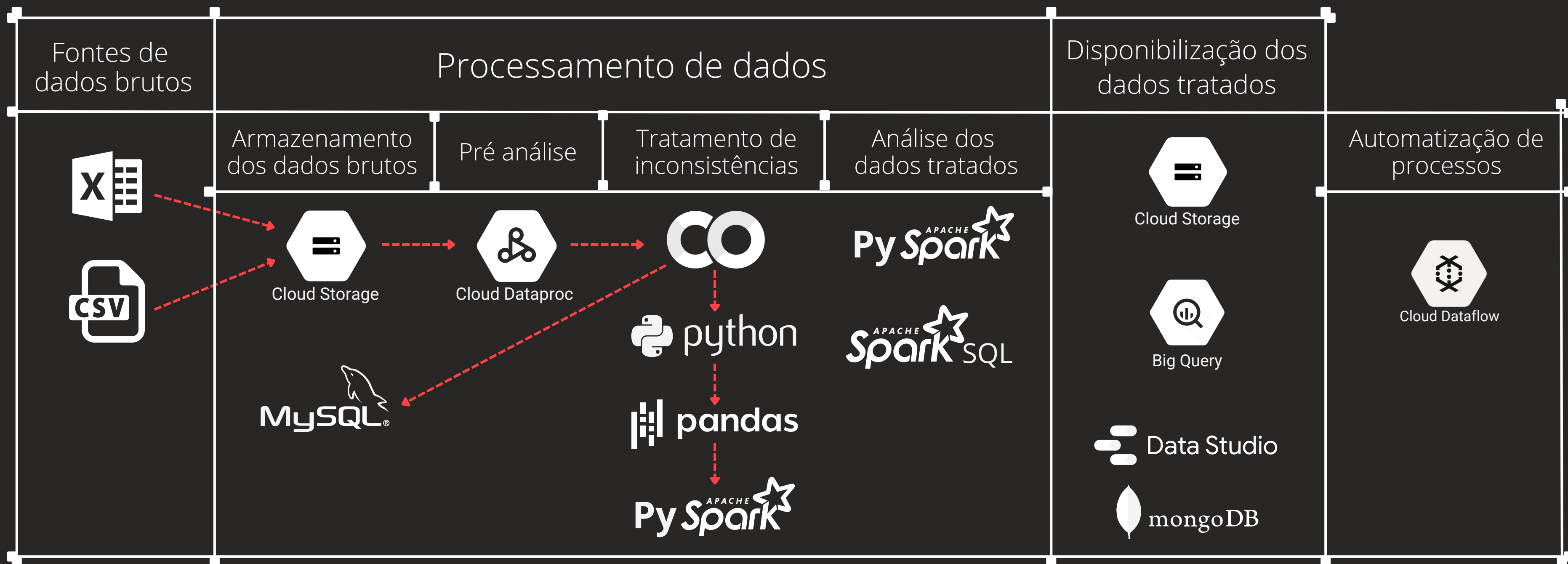
Workflow

ETAPAS DO PROCESSO DE ETL



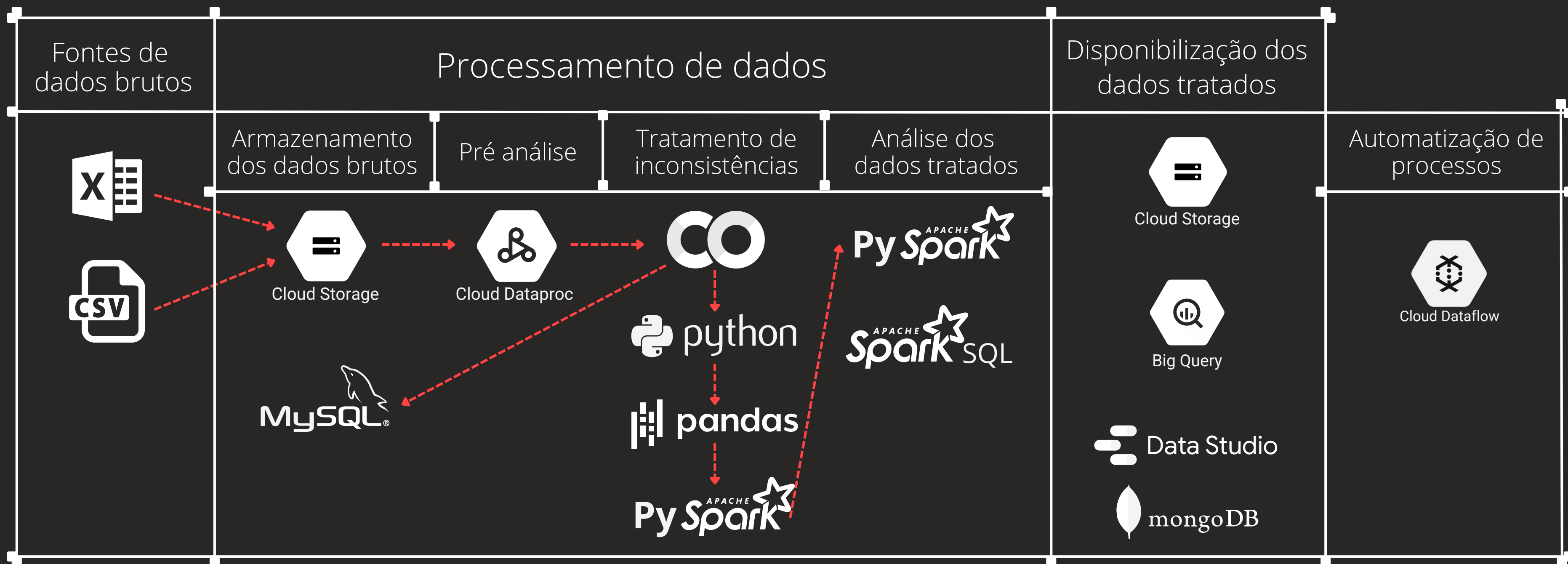
Workflow

ETAPAS DO PROCESSO DE ETL



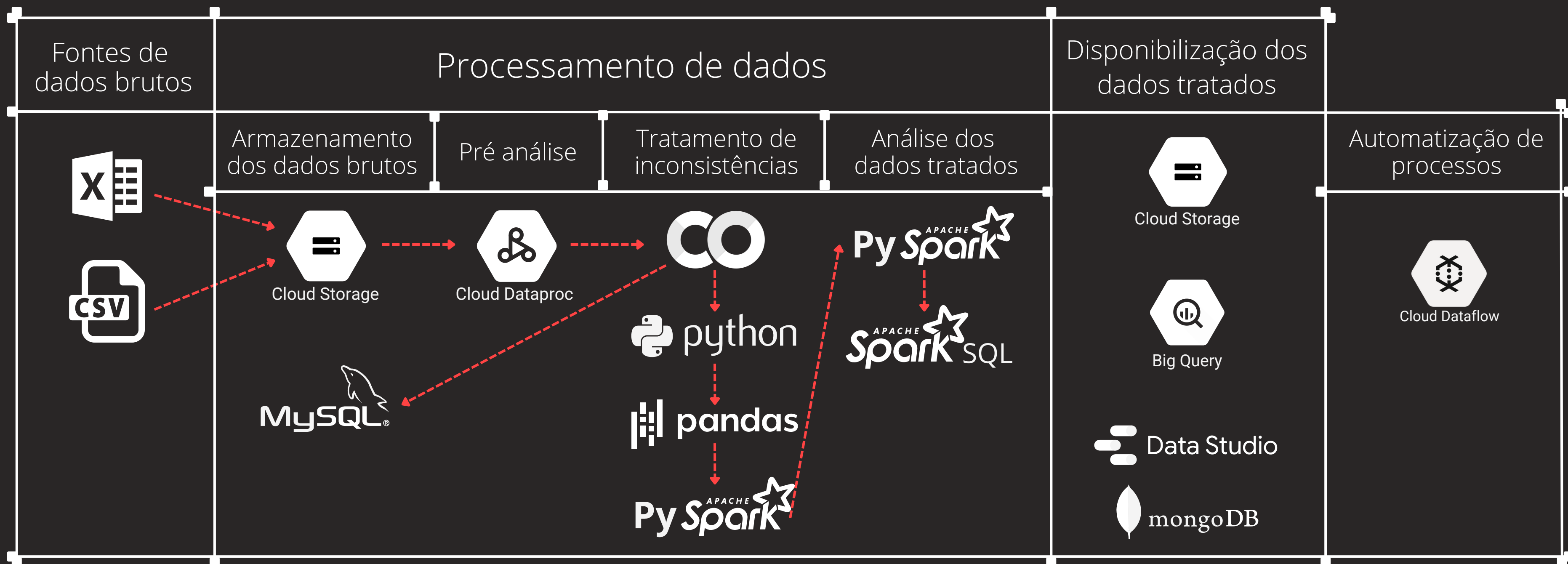
Workflow

ETAPAS DO PROCESSO DE ETL



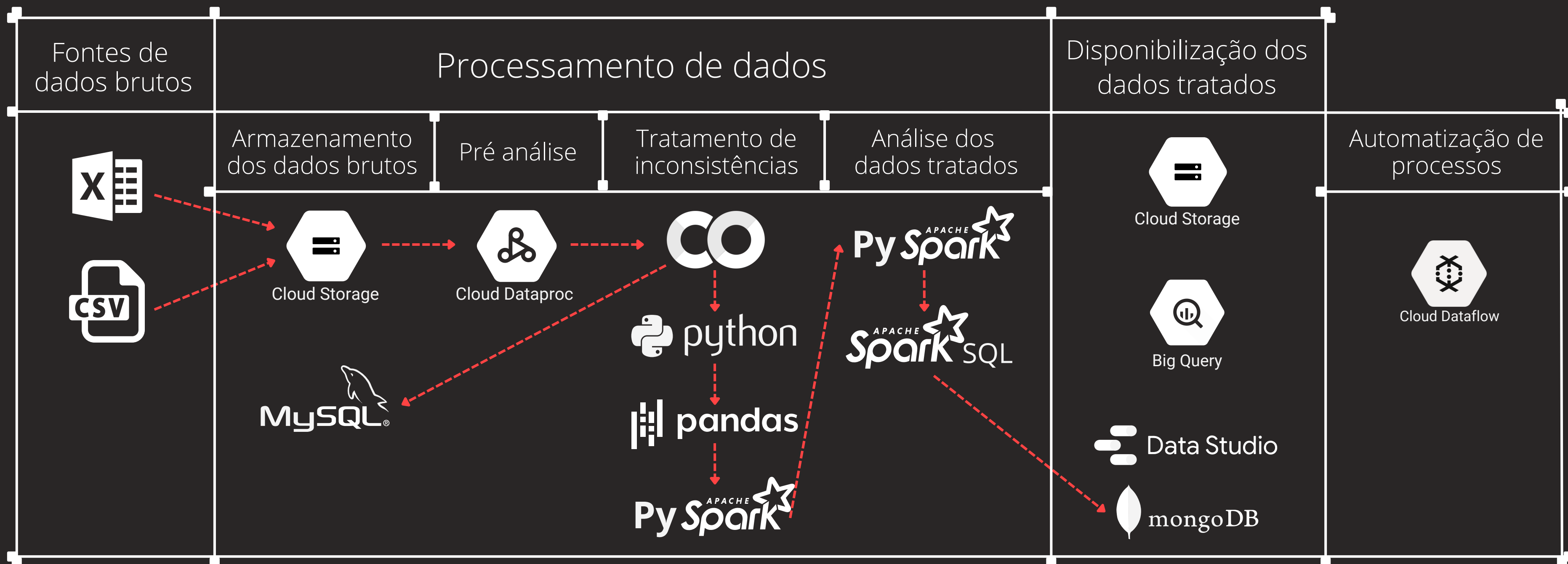
Workflow

ETAPAS DO PROCESSO DE ETL



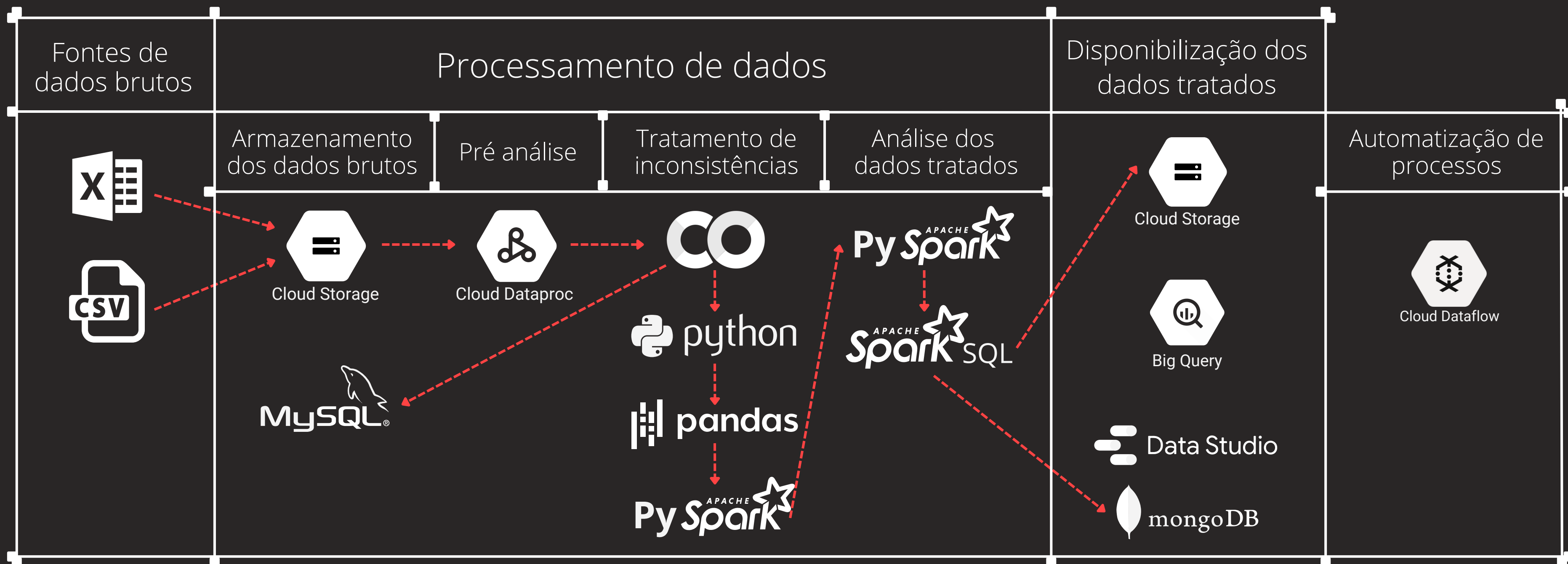
Workflow

ETAPAS DO PROCESSO DE ETL



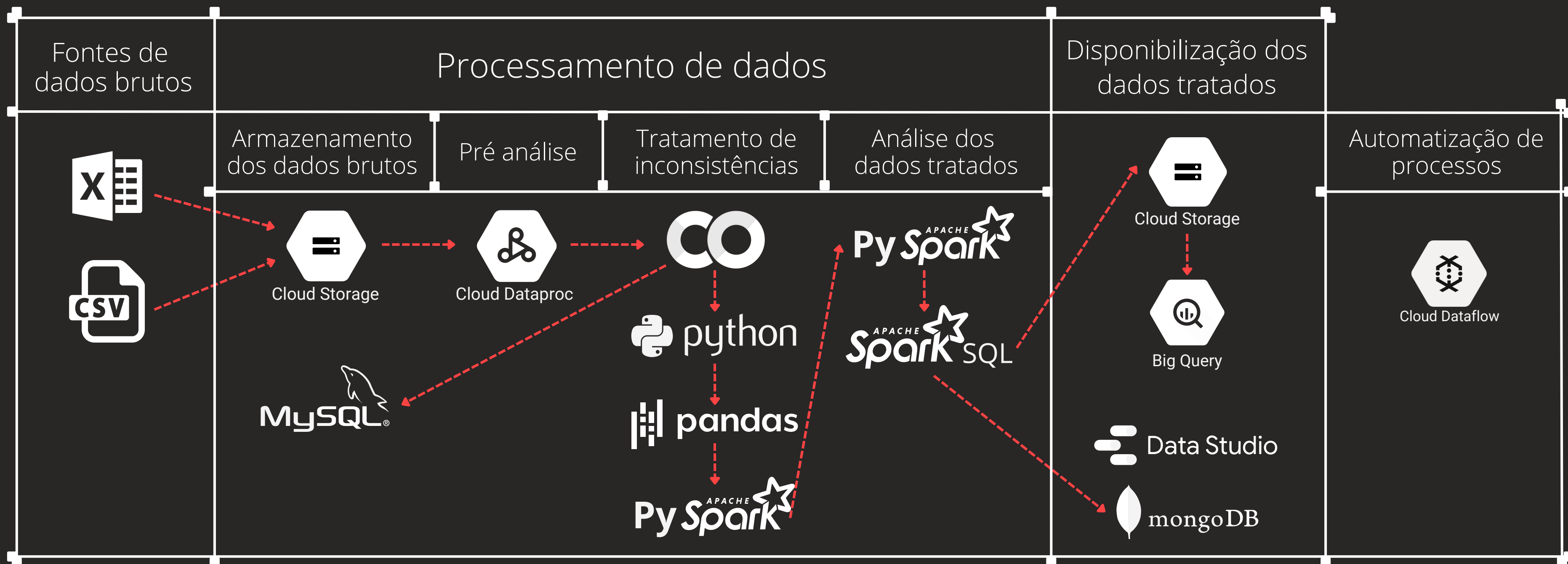
Workflow

ETAPAS DO PROCESSO DE ETL



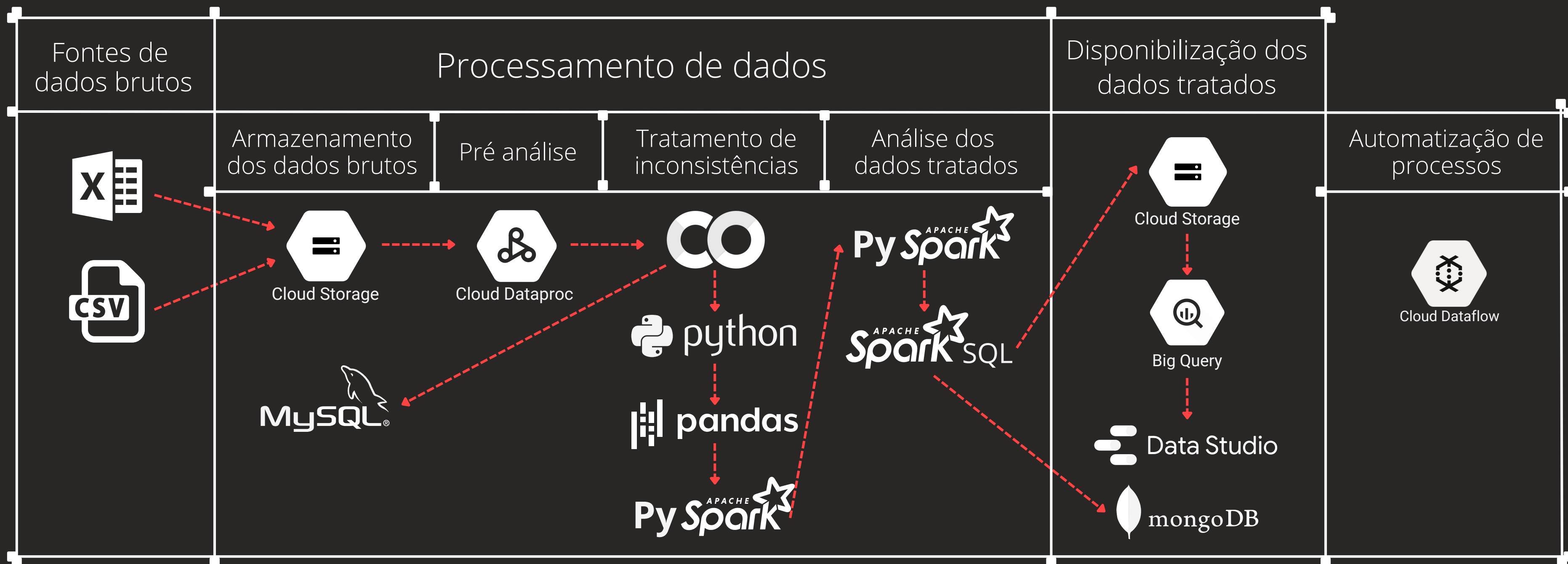
Workflow

ETAPAS DO PROCESSO DE ETL



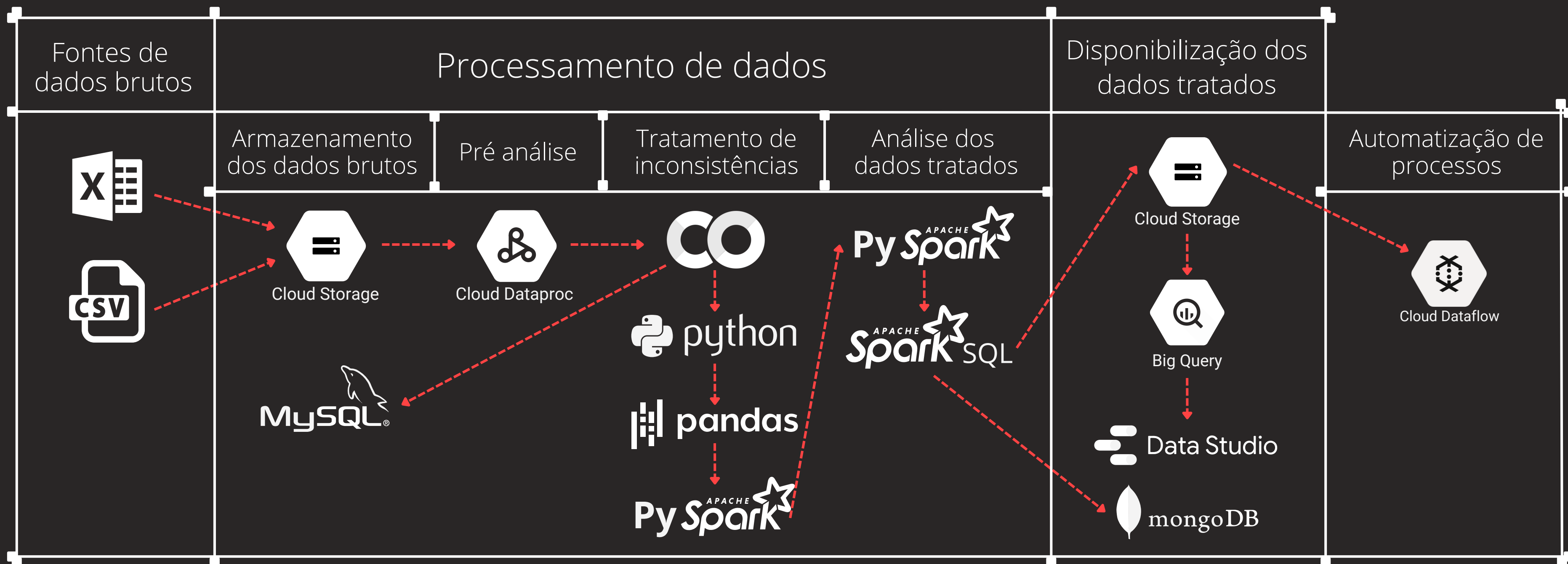
Workflow

ETAPAS DO PROCESSO DE ETL



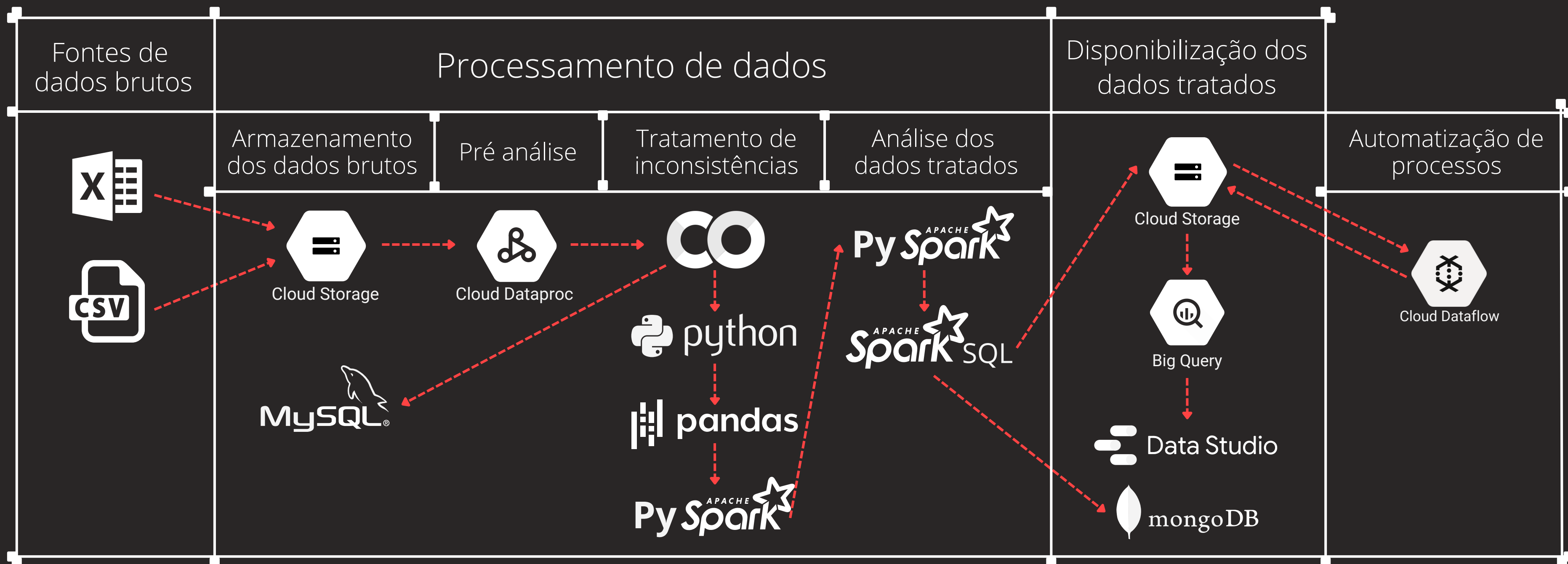
Workflow

ETAPAS DO PROCESSO DE ETL



Workflow

ETAPAS DO PROCESSO DE ETL



ETL



PROCESSO DE EXTRAÇÃO, TRATAMENTO E CARREGAMENTO DOS DADOS

- O termo ETL é uma sigla em inglês que significa **Extrac, Transform and Load** (Extrair, Transformar e Carregar), esse termo é utilizado para descrever como as empresas coletam dados, transformam e utilizam seus resultados para obter informações.
- O processo de ETL é utilizado há décadas por empresas para extrair dados de diversas fontes e disponibilizados como informação para seus stakeholders.



Extract (extração)

Nesta etapa fizemos a extração dos dados brutos utilizando o DATAPROC

DATAPROC - GOOGLE CLOUD STORAGE

n-champion-339219 > cluster-ca23

jupyter dataproc3103 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Python 3

Collecting xlrd
Downloading xlrd-2.0.1-py2.py3-none-any.whl (96 kB)
96.5/96.5 KB 4.0 MB/s eta 0:00:00
Installing collected packages: xlrd
Successfully installed xlrd-2.0.1
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: <https://pip.pypa.io/warnings/venv>
Note: you may need to restart the kernel to use updated packages.

In [5]: `import xlrd`

In [6]: `dff=pd.read_excel('gs://bucket-estilo-vida/dados_brutos/API_BRA_DS2_pt_excel_v2_3732408.xls')`

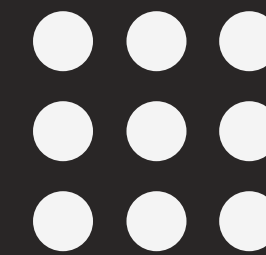
In [8]: `pd.set_option('max_column',None)`

In [10]: `dff.columns`

Out[10]: Index(['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code',
'1960', '1961', '1962', '1963', '1964', '1965', '1966', '1967', '1968',
'1969', '1970', '1971', '1972', '1973', '1974', '1975', '1976', '1977',
'1978', '1979', '1980', '1981', '1982', '1983', '1984', '1985', '1986',
'1987', '1988', '1989', '1990', '1991', '1992', '1993', '1994', '1995',
'1996', '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004',
'2005', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013',
'2014', '2015', '2016', '2017', '2018', '2019', '2020'],
dtype='object')

Transform (transformação)

pandas
matplotlib

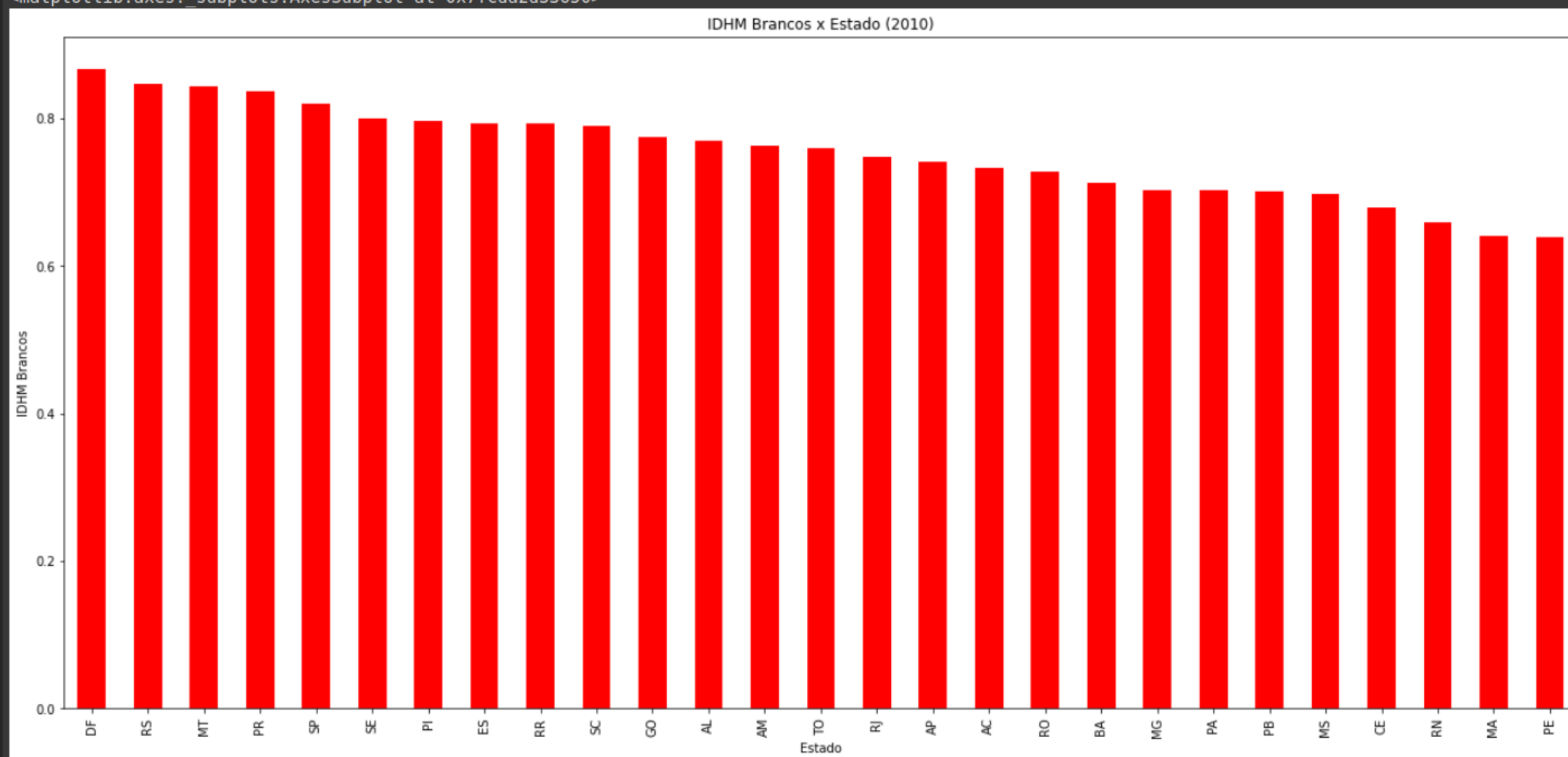


PANDAS

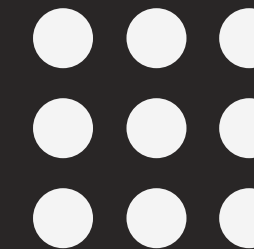
Criação de gráficos para visualização prévia

```
1 #Plotagem do gráfico de barras que representa o IDHM dos brancos por estado no ano de 2010
2 brancos2010["idhm_branco"].sort_values(ascending=False).plot.bar(figsize=(22,10),\
3 xlabel='Estado',ylabel='IDHM Brancos',title="IDHM Brancos x Estado (2010)",color='r')
```

<matplotlib.axes._subplots.AxesSubplot at 0x7fcda2d33650>



Transform (transformação)



PANDAS - PYSPARK

```
[22] 1 #TRANSPONDO AS LINHAS E COLUNAS DO DATAFRAME DFFSPARK UTILIZANDO O PANDAS E RECONVERTENDO PARA O PYSPARK  
      2 dffspark = dffspark.toPandas()
```

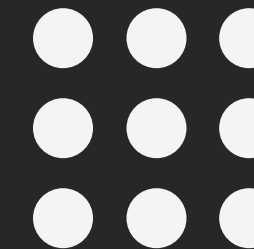
```
[23] 1 dffspark = dffspark.set_index('indicador')
```

```
[24] 1 dffspark = dffspark.T
```

```
[25] 1 dffspark['ano'] = dffspark.index
```

```
[26] 1 dffspark = spark.createDataFrame(data=dffspark)  
      2
```

Transform (transformação)

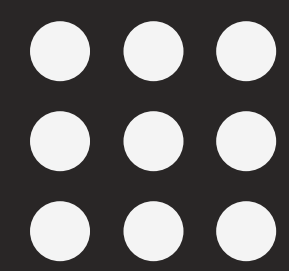


PYSPARK

A transformação com o Pyspark consistiu na validação e tratamentos de inconsistências, além da delimitação do dataset e a criação dos dataframes principais utilizados no projeto:

```
[42] 1 #Agrupando as cidades com base nos anos, de forma a definir suas rendas máximas, mínimas e médias
      2 dfs2 = ( dfs.groupBy(F.col("uf"),F.col("ano"))
      3           .agg(F.max("renda_media").alias("renda_maxima"), #Renda máxima
      4               F.min("renda_media").alias("renda_minima"), #Renda mínima
      5               round(F.avg("renda_media"),2).alias("renda_media")) #Renda média
      6 )
      7
```


Transform (transformação)



PYSPARK

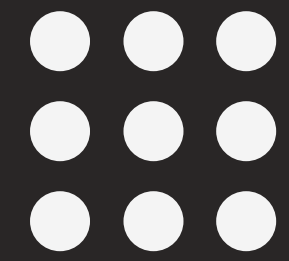
A transformação com o PySpark foi utilizada para gerar insights rápidos no próprio dataframe, por exemplo;

Qual a renda per capita do brasileiro nos últimos censos agrupados por região?

```
[163] 1 # filtros por ano
      2 ano1991 = dfregiao.filter(F.col("ano")==1991)
      3 ano2000 = dfregiao.filter(F.col("ano")==2000)
      4 ano2010 = dfregiao.filter(F.col("ano")==2010)

1 #Renda média per capita por região em 1991
2 ano1991.groupBy("regiao").agg(F.avg(F.col("renda_media"))).sort(F.col("avg(renda_media)").desc()).show()
```

regiao	avg(renda_media)
SUDESTE	313.2671462829734
SUL	300.8435101010101
CENTRO-OESTE	293.06987124463546
NORTE	175.49657015590202
NORDESTE	117.91146599777028



Load (carregamento)

MYSQL

Os dados originais foram armazenados em um banco de dados MySql na GCP

Google Cloud Platform

BC12-Estilo-de-Vida-Proj-Final

Pesquisa

Produtos, recursos, documentos (/)

?

🔔

⋮

Menu de navegação

INSTÂNCIA PRINCIPAL

Visão geral

Conexões

Notas de lançamento

<|

Visão geral

EDITAR

IMPORTAR

EXPORTAR

REINICIAR

INTERROMPER

EXCLUIR

CLONAR

Todas as instâncias > sqlbruto

✓ sqlbruto

MySQL 5.7

1 hora 6 horas ✓ 1 dia 7 dias 30 dias Personalizado

Gráfico

Uso da CPU

Abrir editor

CLOUD SHELL

Terminal

(bc12-estilo-de-vida-proj-final) × +

Database changed

mysql> SHOW TABLES;

+-----+
| Tables_in_bd-bruto |
+-----+
| df-atlas
| dfest2000
| dfest2010
| dff
+-----+
4 rows in set (0.03 sec)

mysql>

Load (carregamento)



Nesta etapa fizemos o armazenamento dos dados tratados no Google cloud storage, MongoDB e BigQuery

DATALAKE - GOOGLE CLOUD STORAGE

Google Cloud Platform

BC12-Estilo-de-Vida-Proj-Final

Pesquisa

Produtos, recursos, documentos (/)

ATUALIZAR

SAIBA

Cloud Storage

Navegador

Monitoramento

Configurações

Marketplace

Detalhes do bucket

bucket-estilo-vida

Local

us-central1 (Iowa)

Classe de armazenamento

Standard

Acesso público

Sujeito a ACLs de objeto

Proteção

Nenhum

OBJETOS

CONFIGURAÇÃO

PERMISSÕES

PROTEÇÃO

CICLO DE VIDA

Intervalos

bucket-estilo-vida

dados_tratados

FAZER UPLOAD DE ARQUIVOS

CARREGAR PASTA

CRIAR PASTA

GERENCIAR RETENÇÕES

FAZER O DOWNLOAD

EXCLUIR

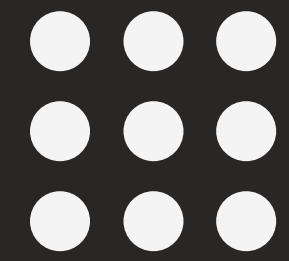
Filtrar apenas pelo prefixo do nome

Filtro

Filtrar objetos e pastas

Mostrar dados excluídos

<input type="checkbox"/>	Nome	Tamanho	Tipo	Criado	Classe de armazenamento	Última modificação	Acesso p
<input type="checkbox"/>	Datapro/	—	Pasta	—	—	—	—
<input type="checkbox"/>	Pyspark/	—	Pasta	—	—	—	—
<input type="checkbox"/>	Tratados_pandas/	—	Pasta	—	—	—	—



Load (carregamento)

Nesta etapa fizemos o armazenamento dos dados tratados no Google cloud storage, MongoDB e BigQuery

MONGODB

+ Create Database

Q NAMESPACES

▼ Projeto final-tratado

df_regiao

dff_spark

dfs2estado_pandas

Projeto final-tratado.dff_spark

STORAGE SIZE: 24KB TOTAL DOCUMENTS: 63 INDEXES TOTAL SIZE: 20KB

Find

Indexes

Schema Anti-Patterns 0

Aggregation

Search Indexes ●

INSERT DOCUMENT

FILTER { field: 'value' }

► OPTIONS

Apply

Reset

```
1991: 1.7594141960144043
2000: 1.4241673946380615
2010: 0.9379593133926392
2020: 0.7128728032112122
_id: ObjectId("624b1d848c13c79eaa3fb96c")
indicador: "Crescimento da população (anual %)"
idhm_1991: 0.381
idhm_2000: 0.523
idhm_2010: 0.659
idhm_2020: 0.765
```

< PREVIOUS

1-20 of many results

NEXT >

Load (carregamento)



Big Query

Nesta etapa fizemos o armazenamento dos dados tratados no Google cloud storage, MongoDB e BigQuery

BIGQUERY

The screenshot displays the Google BigQuery web interface. On the left, the 'Explorer' sidebar shows a project named 'bc12-estilo-de-vida-proj-fi...' with a tree view containing 'Consultas salvas (3)', 'BigqueryEstiloVida', and 'Insights'. Under 'Insights', several tables are listed, including 'renda_media_longevidade...'. The main area is the 'EDITOR' tab, which contains a SQL query. The query is as follows:

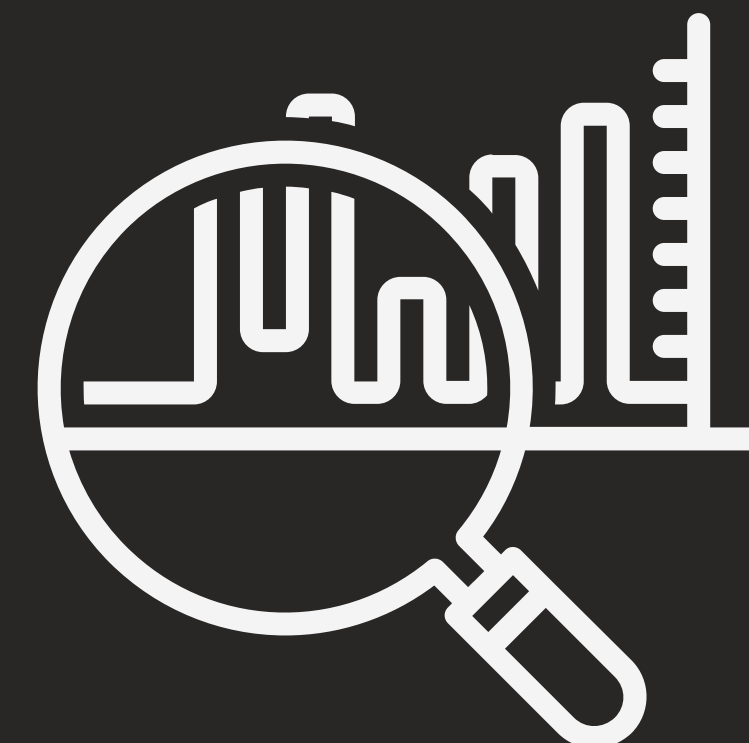
```
1 CREATE OR REPLACE TABLE bc12-estilo-de-vida-proj-final.Insights.renda_media_longevidade_2000 AS (  
2 SELECT dfe.ano, dfe.uf, dfe.idhm_longevidade, round(rm.renda_media, 2)  
3 AS renda_media FROM `bc12-estilo-de-vida-proj-final.BigqueryEstiloVida.dfs2estado` AS dfe  
4 FULL OUTER JOIN `bc12-estilo-de-vida-proj-final.BigqueryEstiloVida.renda_media_uf_1991` AS rm  
5 ON rm.uf = dfe.uf WHERE ano = 2000 ORDER BY(rm.renda_media)DESC)
```

At the top of the editor, there are buttons for 'EXECUTAR', 'SALVAR', 'COMPARTILHAR', 'PROGRAMAÇÃO', and 'MAIS'. A status bar at the bottom right indicates 'Esta consulta'.

Análises no Data Studio



Após o tratamento dos dados e definição dos insights realizamos as análises no Data Studio.



Custos do projeto



PREVISTO X REALIZADO

Detalhamento	Duração	Valor em Reais
Custo Total Previsto	15 dias	R\$ 709,33
Custo Parcial	8° dia	R\$ 299,36
Custo Final	15° dia	R\$ 666,29

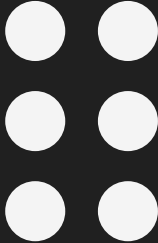
Lições aprendidas

- Backups dos dados tratados diariamente;
- Trabalhar em cópias dos documentos para posterior junção oferece mais liberdade e segurança;
- A importância de ferramentas de gestão para organizar e orientar os objetivos do projeto;
- Conferir os datasets;
- O trabalho em equipe faz toda a diferença.

Considerações finais

- Algumas das nossas inferências foram confirmadas, como por exemplo, o IDHM dos negros se apresentar mais baixo que o IDHM dos brancos em todos os estados;
- O aumento nas despesas públicas com educação e sua correlação direta com o aumento no IDHM não pode ser confirmada, no entanto o aumento das despesas com educação apresenta influência no pequeno aumento do índice de educação de todos os estados;
- A falta do censo em 2020 dificultou a obtenção de indicadores sociais mais atualizados.





**“ A COMBINAÇÃO DE
TRABALHO DURO E
TRABALHO INTELIGENTE
É TRABALHO EFICIENTE. ”**

ROBERT HALF

Perguntas?

