# Passo 1 - Criar cluster

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

# Passo 2 – Montar bucket s3 para o Databricks

Executar somente após subida do Cluster

Crie um notebook no seu Workspace

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

# Passo 3 – Implemente o código abaixo no Notebook (Notebook disponível no Repositório – pasta Notebooks)

# import libraries

import pyspark.sql.functions as f

import urllib

# Credentials to access data-lake-fia bucket on s3

ACCESS\_KEY = "AKIAQJ2IVF6JBGLXOBEF"

SECRET\_KEY = "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"

# Encoded secret key for security purposes

ENCODED\_SECRET\_KEY = urllib.parse.quote(SECRET\_KEY, safe="")

# variables used to mount drive

AWS\_S3\_BUCKET = 'data-lake-fia'

MOUNT\_NAME = '/mnt/data-lake-fia'

SOURCE\_URL = f"s3n://{ACCESS\_KEY}:{ENCODED\_SECRET\_KEY}@{AWS\_S3\_BUCKET}"

print(SOURCE\_URL)

# Mount the drive

dbutils.fs.mount(SOURCE\_URL, MOUNT\_NAME)

# view content of s3 bucket

display(dbutils.fs.ls(MOUNT\_NAME))

# Define location of parquet files (raw-data and context tier)

raw\_tier\_files = "/mnt/data-lake-fia/raw-data/datasus-imunizacao/"

context\_tier\_files = "/mnt/data-lake-fia/context/datasus\_db/covid\_dataset/"

# create Dataframe to make analysis

df\_covid = spark.read.parquet(context\_tier\_files)

df\_covid.display()

# view schema

df\_covid.printSchema()