# bigDGeneral_IPCA Project

## DESCRIÇÃO:

O projeto bigDGeneral_IPCA tem como objetivo baixar todo mês a tabela de IPCA do endereço https://www.idinheiro.com.br/tabelas/tabela-ipca/, extrair os dados mensais de IPCA, transformar em tabela estruturada e gravar em um arquivo parquet, fazer a ingestão em um bucket s3 na AWS para posterior catalogação desse arquivo, através do AWS Glue Data Catalog, sendo os dados posteriormente liberados para os Data Analysts e Data Scientysts que usarão ferramentas de insights para geração de apresentações gráficas para a empresa.

## OBJETIVO:

### Motivação:

Teve como motivador para que esse projeto bigDGeneral_IPCA fosse feito: A necessidade de se ter o índice de inflação no Data Lake bigDVarejo e com isso conhecer a variação de preços praticados nas vendas ao consumidor.

### Aplicação Prática:

Conhecer qual foi a inflação de determinado período relacionada aos preços de produtos voltados ao consumidor final, ao consumidor do varejo. Através desse conhecimento usar esses índices para se descontar o que foi perdido para a inflação, fazendo o alinhamento de qualquer valor no cálculo de sua variação.

### Resultados Esperados:

Permitir a correção de qualquer valor ou índice descontando a variação da inflação do valor de produtos voltados ao consumidor.

## MÉTODOS:

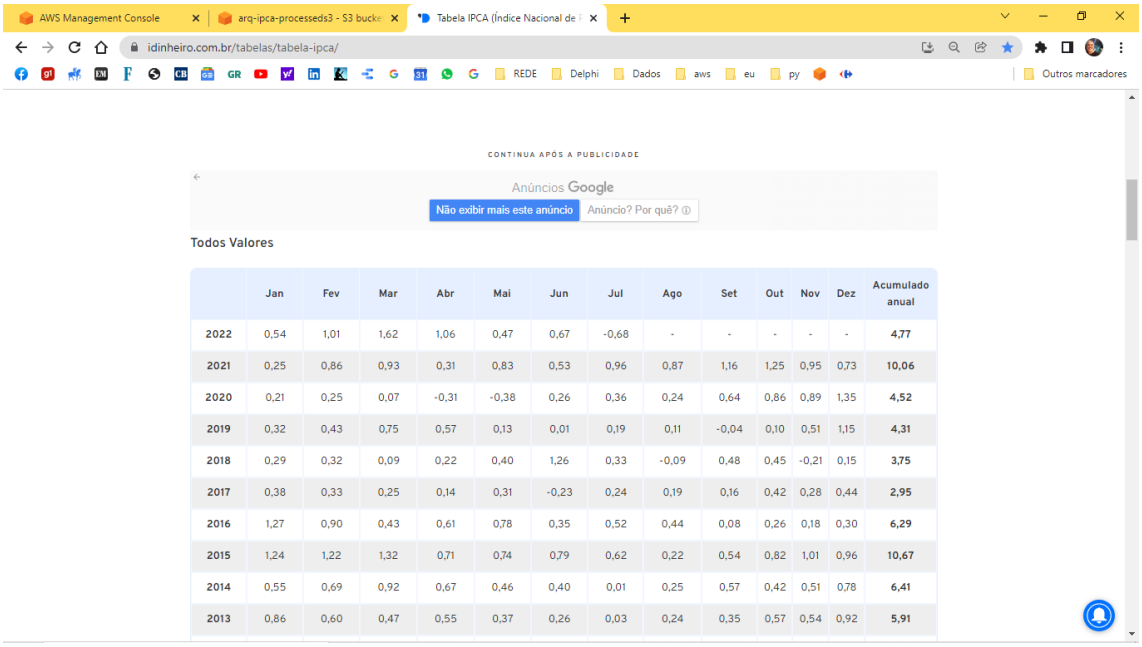
De acordo com a motivação já descrita anteriormente deu-se o início para que os primeiros passos fossem dados, conforme o problema apresentado, em direção ao estudo e implementação do bigDGeneral_IPCA:

- Verificar se as solicitações apresentadas são viáveis:
  - De acordo com as necessidades apresentadas fez-se uma análise preliminar para se certificar da viabilidade do projeto.

- Pesquisar tabelas de dados relativos à variação do IPCA:
  - Procurou-se encontrar na web uma tabela que resumisse todos os índices de IPCA existentes e atualizados do Brasil.

- Avaliar arquivos de dados disponibilizados:
  - Os dados da tabela table-all__value oferecidos pelo site https://www.idinheiro.com.br/tabelas/tabela-ipca/ foram

analisados e validados se poderiam entregar o resultado de informação que era necessário para a comparação da perda de inflação ocorrida no período.



CONTINUA APÓS A PUBLICIDADE

Anúncios Google

Não exibir mais este anúncio | Anúncio? Por quê?

Todos Valores

| | Jan | Fev | Mar | Abr | Mai | Jun | Jul | Ago | Set | Out | Nov | Dez | Acumulado anual |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2022 | 0,54 | 1,01 | 1,62 | 1,06 | 0,47 | 0,67 | -0,68 | - | - | - | - | - | 4,77 |
| 2021 | 0,25 | 0,86 | 0,93 | 0,31 | 0,83 | 0,53 | 0,96 | 0,87 | 1,16 | 1,25 | 0,95 | 0,73 | 10,06 |
| 2020 | 0,21 | 0,25 | 0,07 | -0,31 | -0,38 | 0,26 | 0,36 | 0,24 | 0,64 | 0,86 | 0,89 | 1,35 | 4,52 |
| 2019 | 0,32 | 0,43 | 0,75 | 0,57 | 0,13 | 0,01 | 0,19 | 0,11 | -0,04 | 0,10 | 0,51 | 1,15 | 4,31 |
| 2018 | 0,29 | 0,32 | 0,09 | 0,22 | 0,40 | 1,26 | 0,33 | -0,09 | 0,48 | 0,45 | -0,21 | 0,15 | 3,75 |
| 2017 | 0,38 | 0,33 | 0,25 | 0,14 | 0,31 | -0,23 | 0,24 | 0,19 | 0,16 | 0,42 | 0,28 | 0,44 | 2,95 |
| 2016 | 1,27 | 0,90 | 0,43 | 0,61 | 0,78 | 0,35 | 0,52 | 0,44 | 0,08 | 0,26 | 0,18 | 0,30 | 6,29 |
| 2015 | 1,24 | 1,22 | 1,32 | 0,71 | 0,74 | 0,79 | 0,62 | 0,22 | 0,54 | 0,82 | 1,01 | 0,96 | 10,67 |
| 2014 | 0,55 | 0,69 | 0,92 | 0,67 | 0,46 | 0,40 | 0,01 | 0,25 | 0,57 | 0,42 | 0,51 | 0,78 | 6,41 |
| 2013 | 0,86 | 0,60 | 0,47 | 0,55 | 0,37 | 0,26 | 0,03 | 0,24 | 0,35 | 0,57 | 0,54 | 0,92 | 5,91 |

- *Verificar a viabilidade de baixar essa tabela via python:*
    - Outra verificação importante foi a de se certificar que a tabela eleita para o projeto pudesse ser baixada, utilizando funções python que fazem web scraping e que estivessem disponíveis continuamente, inclusive com as atualizações.

- *Definir quais os formatos de arquivos serão utilizados no armazenamento:*
    - A ideia é armazenar a tabela de IPCA mensal em arquivo parquet, que será catalogado pelo Glue Data Catalog, podendo os dados e informações serem disponibilizados para o usuário final data analysts, data scientists ou mesmo data engineers.

- *Definir onde os arquivos serão armazenados, em local ou cloud:*
    - O local onde os arquivos ficarão armazenados tem como peso principal a continuidade do armazenamento com boa performance de acesso. Elegeu-se o aws S3 para armazenamento, que oferece toda a manutenção da estrutura, com armazenamento distribuído e alta escalabilidade, deixando os engenheiros de dados livres para se preocuparem somente com o planejamento e operação do ELT.

- *Definir quantos S3 bucket serão criados para o armazenamento:*
    - Ficou definido que será necessário somente um bucket s3 para o Data Lake com o nome de arq-ipca-processeds3.

- *Definir quantas camadas serão necessárias para o processamento:*

- O processamento será feito em 2 camadas, uma que usará o python fazendo a extração com web scraping da tabela de IPCA gravando em arquivo parquet e uma segunda camada que fará a catalogação dos dados através do aws Glue com suas ferramentas.

- *Definir qual a linguagem será utilizada para o processamento dos dados:*
  - Por conter uma enorme diversidade de bibliotecas para inúmeros fins, apresentar uma simplicidade de estrutura voltada para orientação a objetos, por ser uma linguagem que está sendo muito utilizada no mundo sendo uma tendência em manipulação de dados, apresentando funções voltadas para tal, por ser de fácil uso dentro da aws, optou-se em utilizar a linguagem python.

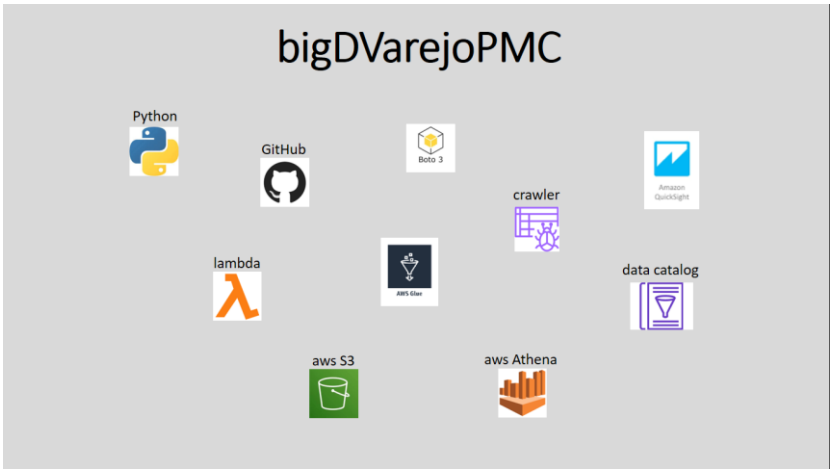
- *Definir onde serão executados os códigos:*
  - Os códigos serão processados utilizando-se o serviço da aws Lambda, que é orientada a eventos com computação sem servidor, não é necessário definir um servidor para executar uma aplicação ficando transparente para nosso processamento, sendo mais uma preocupação para a equipe da aws Amazon manter o serviço funcionando com escalabilidade.

- *Definir onde será o repositório de hospedagem dos códigos:*
  - Devido à experiência com o Git-Hub, por apresentar seus recursos de hospedagem e manutenção de versões de código com simplicidade e objetividade, pela sua divulgação e utilização na comunidade de desenvolvimento de software, definiu-se pela utilização dessa plataforma.

- *Definir as bibliotecas utilizadas:*
  - Para fazer upload dos arquivos baixados e gerados utilizar-se-á a biblioteca python boto3 (facilita o acesso aos serviços da aws).

- *Definir as características e configurações do scheduler:*
  - O processamento das 2 camadas ocorrerá no quinto dia do mês às 24:00.

## PRODUTOS, SERVIÇOS E SISTEMAS:



# ESCOPO DA ARQUITETURA:

## dgIPCAproject