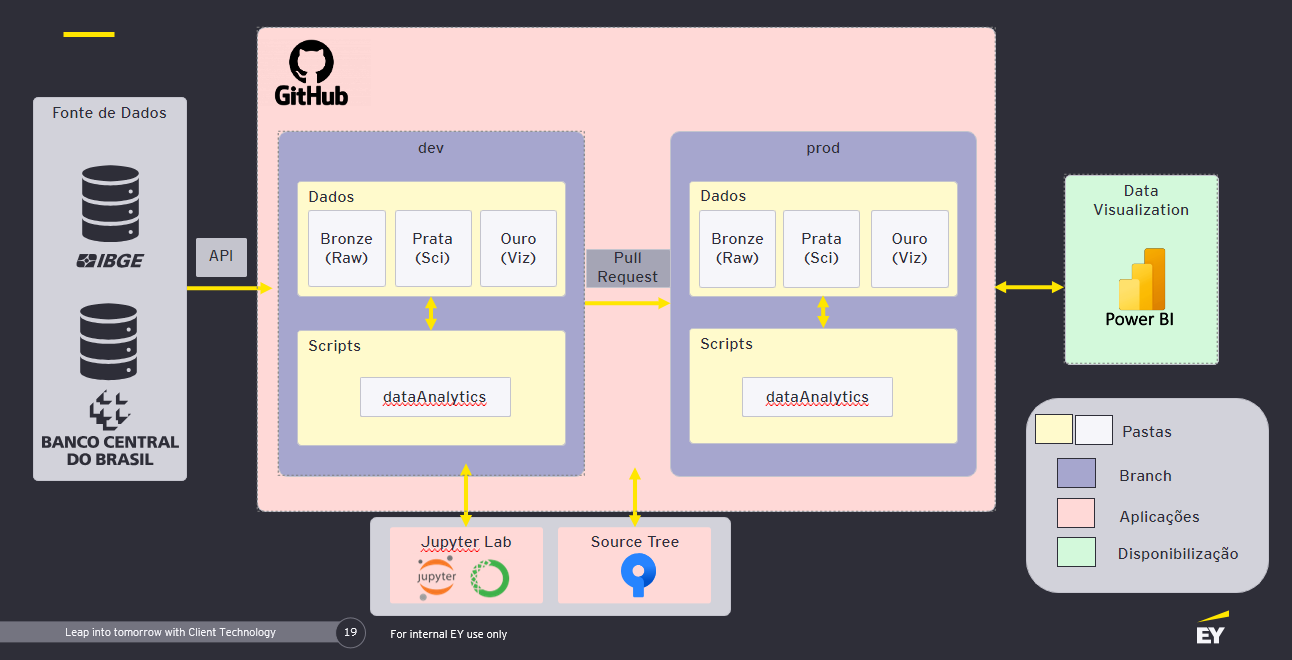
**Documentação - Data Engineering**

**1. Definição de Pipeline**

A primeiro etapa antes de reunir as bases foi a definição da Pipeline que iriamos implementar na ferramenta GitHub (usada como reposítorio). Com o auxilio do Power Point e as opinioes do time, foi desenvolvido um fluxo de como a pipeline do projeto iria se comportar:

* 1. **Fonte de Dados**

O primeiro ponto foi delimitar de onde nossas dados viriam e após isso desenvolver um script que possibilitasse a importação desses dados para nossa branch dev. O Sidra disponibiliza documentação sobre sua API além de existir uma biblioteca no Python própria para essa API do Sidra, chamada Sidrapy. No caso dos dados do BCB, não foi localizada nenhuma API capaz de extrair os dados perfeitamente, então a importação foi feita manualmente.

Dito isso, os dados extraidos foram os seguintes:

*- Sidra:*

- Relação de População x Estado x Idade (2022 e 2023)

- Relação de População x Nivel de Instrução x Sexo x Estado (2022)

- Relação de Populacão x Trabalho x Estado (2022)

*- BCB (Estban):*

- Relação de Municipio x Produtos (2022 e 2023)

* 1. **Branch Dev**

Dentro do ambiente GitHub, foi criada uma branch chamada “dev”, que tem como objetivo servi como um ambiente de experimentação. Dentro dessa branch, existem duas pastas: “Dados” e “Scripts”.

Na pasta “Dados”, temos três camadas: “Bronze”, “Prata” e “Ouro” e na pasta “Scripts” estão localizados todos os scripts desenvolvidos para alcançar os objetivos do projeto.

**1.2.1 Camadas de Tratamento**

Na camada “Bronze” é onde foi armazenado todos os dados brutos, ou seja, os dados que foram extraidos das Fontes de Dados foram alocados dentro da camada “Bronze”, e não foram alterados.

Na camada “Prata” foram disponibilizadas versões que sofreram adequações desses dados. No inicio tinhamos 24 fontes do BCB, após a adequação e pré-formatação via Python, foi disponibilizado apenas uma grande planilha que unifica toda a base extraida do BCB. Já em relação as bases retiradas do BCB, foram criados scripts de adequeção e pré-formatação via Python, resultando em planilhas mais legiveis e uteis para Data Science. (Nesta etapa Data Eng. contou com a ajuda da equipe de Data Sci.).

A camada “Ouro” recebeu apenas as planilhas que poderiam ser utilizadas pela equipe de Data Visualization para a criação do dashboard.

* 1. **Branch Prod**

A branch “Prod” é um espelho da branch “Dev”, porém ela era atualizada uma vez na semana para disponibilizar todo o trabalho completo durante as sprints. Nela era armazenado exatamente as mesmas coisas que na branch “Dev”.

* 1. **Aplicações e Disponibilização**

Durante esse processo, foram utilizadas três aplicações principais:

Jupyter Lab, SourceTree e GitHub. O Jupyter foi utilizado para a criação dos scripts de manipulação da base de dados e o SourceTree foi utilizado como uma interface mais amigavel para Commit, Pull e Push no GitHub.

O GitHub foi utilizado como repositorio do projeto, sendo ele, uns dos focos em nossa pipeline.

**2. Tratamento das Bases**

Após a Pipeline estruturada, o proximo passo realizado foi o tratamento das bases que foram alocadas na camada Bronze (Raw Data). Com auxilio da equipe de Data Science, foi-se delimitado todas as colunas que seriam preservadas em cada tabela. Após isso, focou-se em desenvolver os scripts que fariam essas formatações no dados, que resultou em novas tabelas que foram armazenadas na camada Prata:



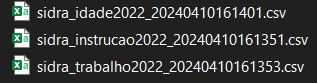
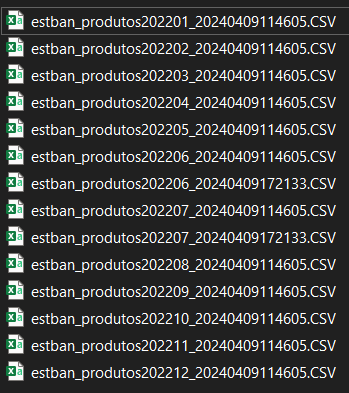
Os pontos principais considerados no momento da tratamento foram: a eliminação de colunas desnecessárias, a renomeação das colunas restantes, adequação da formatação de valores e agregação das tabelas ESTBAN e Sidra Nivel de Instrução.

Feito isso, todas as novas tabelas foram alocadas na camada Prata da branch Dev, para que a equipe de Data Science pudesse iniciar seus próprios tratamentos, visando os indicadores já delimitados pela equipe.

**3. Padronização**

Um dos pontos mais importantes acordado entre o grupo foi a Padronização que deveria ser seguida em varios aspectos:

*- Bases na Camada Bronze:* com auxilio do script e da API, foram extraidas as bases de dados e em comum acordo foi estipulado que nessa camada seria usado a nomenclatura Thesaurus para nomeação dos arquivos. O padrão Thesaurus consiste em fonteDados\_assunto\_timeStamp.csv:



*- Bases na Camada Prata:* quando as bases eram tratadas e movidas para a camada Prata, foi-se adotado uma padronização mais simples para o nome das bases, consistindo basicamente em fonteDados\_agreg.csv.



Outro ponto importante foi a padronização do t ipo de codificação (UTF-8) e delimitador ( | ).

- GitHub/SourceTree: Ao realizar os commits, estipulou-se uma padronização para os comentarios, considerando as equipes envolvidas (ENG, SCI, VIZ):







**4. Links**

Sidra:

[Tabela 7267: Pessoas de 14 anos ou mais de idade por sexo e nível de instrução (ibge.gov.br)](https://sidra.ibge.gov.br/tabela/7267)

[Tabela 7130: Pessoas de 14 anos ou mais de idade ocupadas na semana de referência, por grupamento de atividade no trabalho principal (ibge.gov.br)](https://sidra.ibge.gov.br/tabela/7130)

[Tabela 9514: População residente, por sexo, idade e forma de declaração da idade (ibge.gov.br)](https://sidra.ibge.gov.br/tabela/9514)

BCB:

[Estatística Bancária Mensal por município - ESTBAN (bcb.gov.br)](https://www.bcb.gov.br/estatisticas/estatisticabancariamunicipios)

GitHub:

[naomiteixeira/eyacademycrm (github.com)](https://github.com/naomiteixeira/eyacademycrm)