

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação *Lato Sensu* em Analytics e Business Intelligence

**RELATÓRIO TÉCNCO**

PAINEL DE OPERAÇÕES DE CRÉDITO (SCR)

Alexandre Baudon

Belo Horizonte

2021

**SUMÁRIO**

[1. Introdução 3](#_Toc88148709)

[1.1. Contexto 3](#_Toc88148710)

[1.2. Objetivos 4](#_Toc88148711)

[1.3. Público alvo 5](#_Toc88148712)

[2. Modelo de Dados 5](#_Toc88148713)

[2.1. Modelo Dimensional 5](#_Toc88148714)

[2.2. Fatos e Dimensões 7](#_Toc88148715)

[3. Integração, Tratamento e Carga de Dados 9](#_Toc88148716)

[3.1 Ferramenta 9](#_Toc88148717)

[3.2 Staging area 10](#_Toc88148718)

[3.3 Detalhamento da origem dos dados 15](#_Toc88148719)

[3.4 Tratamento dos dados 23](#_Toc88148720)

[3.5 Dimensional 24](#_Toc88148721)

[3.6 Detalhamento cargas dimensional 27](#_Toc88148722)

[4. Camada de Apresentação 34](#_Toc88148723)

[4.1 Dashboard 34](#_Toc88148724)

[5. Registros de Homologação 38](#_Toc88148725)

[6. Análises Avançadas 38](#_Toc88148726)

[6.1. Machine Learning 38](#_Toc88148727)

[6.2. Resultados 39](#_Toc88148728)

[5. Conclusões 40](#_Toc88148729)

[6. Links 40](#_Toc88148730)

[REFERÊNCIAS 41](#_Toc88148731)

# 1. Introdução

## 1.1. Contexto

O Banco Central do Brasil – Bacen divulga mensalmente uma série de informações agregadas das operações de crédito recebidas através do Sistema de Informações de Créditos – SCR. Entende-se como operação de crédito toda transação relativa aos contratos de empréstimo, financiamento, títulos descontados, crédito rural e garantias contratadas junto a uma instituição financeira. Os dados são enviados pelas respectivas instituições e disponibilizados após 60 dias do fechamento contábil de cada período.

A principal finalidade do SCR é fornecer informações ao Bacen para fins de monitoramento do crédito do sistema financeiro e exercício das atividades de fiscalização. É uma forma também de troca de informações entre as instituições financeiras permitindo uma avaliação mais precisa e segura do cliente quanto a capacidade de pagamento das operações contratadas. É importante ressaltar que qualquer disponibilização dos dados referente ao SCR depende de autorização prévia do cliente.

O conjunto de dados disponibilizado pelo Bacen oferece um vasto leque de informações proporcionando uma visão geral do perfil da carteira de crédito de cada unidade da federação, porém não há uma correlação com outros indicadores socioeconômicos, que possibilite assim uma análise mais ampla e crítica do panorama financeiro das famílias brasileiras. Para traçar um perfil é necessário a inclusão de outras variáveis em um contexto mais abrangente, desta forma os painéis apresentados neste projeto se propõem a explorar também dados que estão disponíveis em outras esferas, como por exemplo o IBGE (Instituto Brasileiro de Geografia e Estatística).

Para este projeto em questão, 3 indicadores socioeconômicos foram levados em consideração:

1. PIB – Produto Interno Bruto: é um importante conceito referente à riqueza produzida pelas atividades econômicas de um determinado país. Seu valor corresponde a tudo o que foi produzido e devidamente consumido, seja esse consumo direto ou indireto.
2. Renda Per Capita: é a distribuição das riquezas produzidas no país pela sua população. É uma métrica mensurada a partir do [PNB (Produto Nacional Bruto)](https://www.preparaenem.com/geografia/produto-nacional-bruto-pnb.htm), que, resumidamente, é o valor do PIB subtraído pelo capital que deixa o país e somado ao capital que entra no país.
3. Desemprego: o desemprego é um dos principais problemas que podem ser enfrentados por um país, pois representa tanto uma ausência de renda por boa parte da população quanto a redução do mercado consumidor, o que gera menos lucro e, portanto, menos emprego. Não por acaso, as grandes crises econômicas sempre afetaram a população por meio da elevação das taxas de desemprego.
4. Taxa Selic: o valor da taxa Selic influencia diretamente na taxa das operações de crédito que as instituições financeiras repassam a seus clientes. Quanto maior a taxa Selic a maior taxa de juros de uma operação, em contra partida as aplicações financeiras tendem a render mais e vice-versa.

## 1.2. Objetivos

O projeto está estruturado em duas visões principais:

1. Uma primeira onde é possível observar um panorama geral de alguns indicadores a nível Brasil e unidades da federação, com dados do IBGE e Banco Central. São indicadores macros, que possibilitam aos usuários obterem respostas mais claras e objetivas das informações socioeconômicas e operações de crédito. Esta primeira visão possui informações anuais com um histórico a partir de 2012. Para alguns indicadores não foi possível recuperar os dados de alguns visto a ausência dos mesmos nos respectivos órgãos responsáveis pela disponibilização.
2. A segunda visão aprofunda mais nos dados do SCR onde é são disponibilizados dados mensais de carteira ativa, inadimplência, com possibilidade de detalhamento por tipo de cliente (pessoa física e jurídica), modalidade das operações de crédito, classificação nacional de atividades econômicas (CNAE – PJ), natureza da ocupação profissional (PF), porte/rendimento dos clientes, origem de recursos e indexador das operações.

O principal objetivo deste projeto é apresentar um panorama geral da carteira de crédito nas unidades da federação e associar a índices socioeconômicos de forma a apresentar um raio x bem como uma correlação.

## 1.3. Público alvo

Este trabalho destina-se a todas as pessoas e instituições que querem conhecer e explorar a situação econômica e social do Brasil e suas respectivos unidades da federação. Não é necessário ter nenhum conhecimento prévio em tecnologia. O painel será apresentado da forma mais didática possível.

# 2. Modelo de Dados

## 2.1. Modelo Dimensional

O modelo dimensional elaborado no formato *star schema* foi organizado em dois assuntos:

1. Fato do SCR (FAT\_SCR): contém as informações disponibilizadas nos arquivos das operações de crédito com periodicidade mensal e histórico a partir de junho de 2012 até o último mês fechado abril/2021. Segue modelo conforme figura 1:

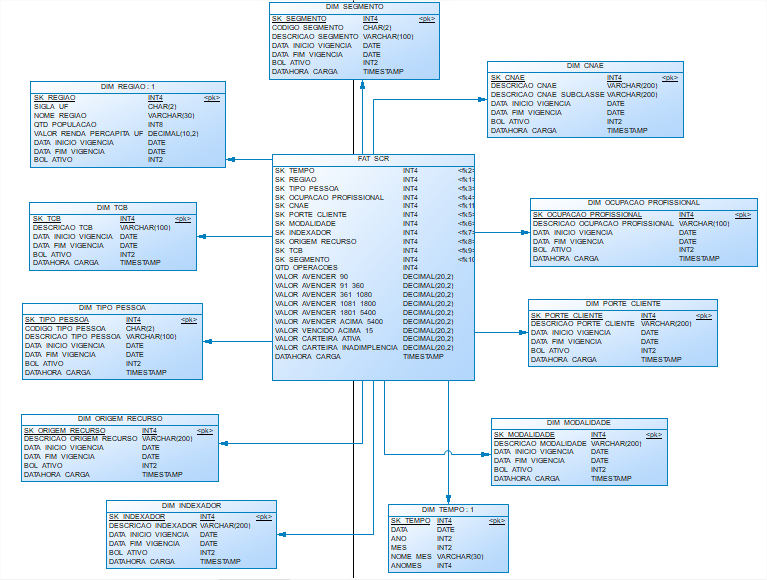


Figura 1: modelo dimensional SCR

1. Fato com os indicadores anuais socioeconômicos (FAT\_INDICADORES\_ANUAIS): contém os dados do SCR com o último mês de cada ano consolidados com os dados do IBGE. A periodicidade é anual com histórico a partir de 2012 até 2020, último ano fechado. Segue modelo conforme figura 2:

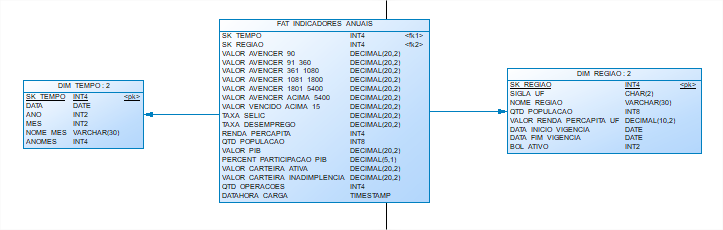


Figura 2: modelo dimensional indicadores anuais

## 2.2. Fatos e Dimensões

Segue um breve descritivo de cada uma das estruturas criadas no modelo dimensional:

* DIM\_SEGMENTO: a dimensão possui as informações do domínio dos possíveis segmentos que a instituição financeira pertence de acordo com a resolução do BACEN nº 4553/2017. Podem estar agrupadas da seguinte forma: S1 – representa os banco múltiplos com porte superior a 10% do PIB; S2 – bancos múltiplos com porte superior a 1% e inferior a 10% do PIB; S3 – instituições com porte superior a 0,1% e inferior a 1% do PIB; S4 – instituições com porte inferior a 0,1% do PIB; S5 - instituições com porte inferior a 0,1% do PIB que utilizem metodologia facultativa simplificada para apuração de patrimônio de referência.
* DIM\_CNAE: dimensão com a classificação nacional das atividades econômicas. Está estruturada nas sessões de A a U com suas respectivas subdivisões (vide imagem 3).
* DIM\_OCUPACAO\_PROFISSIONAL: natureza das ocupações profissionais das pessoas físicas agrupadas da seguinte forma: PF - Servidor ou empregado público; PF - Empregado de entidades sem fins lucrativos; PF - Empregado de empresa privada; PF - Aposentado/pensionista; PF – Autônomo; PF – Empresário; PF - MEI

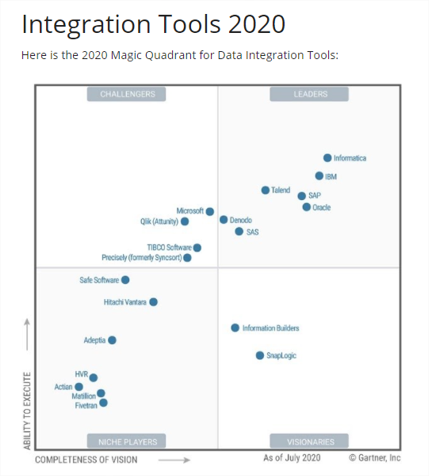
PF – Outros.

* DIM\_PORTE\_CLIENTE: porte dos clientes pessoa física agrupadas de acordo com faixas de salário-mínimo e para pessoa jurídica de acordo com o enquadramento da empresa.
* DIM\_MODALIDADE: as modalidades representam uma agregação das submodalidades disponíveis no cadastrado de uma nova operação de crédito. Alguns exemplos: PF – Cartão de Crédito; PF – Veículos; PF – Habitacional; PJ – Investimento.
* DIM\_INDEXADOR: informa por qual tipo de indexador ou taxa referencial a operação de crédito será reajustada. Exemplo: Prefixado; Pós-Fixado; TCR/TRFC.
* DIM\_ORIGEM\_RECURSO: a operação de crédito pode ter uma origem de recurso sem destinação específica ou com destinação específica. Quando possui destinação específica significa que o recurso teve como origem algum órgão regulador como BNDES, Finame, Fundos Estaduais.
* DIM\_TIPO\_PESSOA: informa se a operação de crédito é de uma pessoa física ou pessoa jurídica.
* DIM\_TCB: o tipo de consolidado bancário informado se a operação é originária de um agrupamento bancário (bancos comerciais, bancos múltiplos); agrupamento não bancário (instituições de pagamento e bancos não comerciais); ou cooperativas de crédito (singulares e centrais).
* DIM\_REGIAO: domínio das 27 unidades da federação e suas respectivas regiões (Norte, Nordeste, Sul, Sudeste e Centro-Oeste).
* FAT\_SCR: agrupamento de todas as operações de crédito enviadas pelas instituições financeiras ao Banco Central apresentadas mensalmente. As métricas geradas estão agregadas pelas dimensões informadas anteriormente. São elas: Quantidade de Operações; Valores a vencer em até 90 dias, valores a vencer entre 91 e 360 dias, valores a vencer entre 361 e 1080 dias, valores a vencer entre 1081 e 1800 dias, valores a vencer entre 1801 e 5400 dias, valor a vencer acima de 5400 dias, valores vencidos acima de 15 dias, valores da carteira ativa e inativa.
* FAT\_INDICADORES\_ANUAIS: representa um consolidado da FAT\_SCR enriquecido de dados anuais do IBGE: desemprego, renda per capita, população e PIB.

# 3. Integração, Tratamento e Carga de Dados

## 3.1 Ferramenta

Todo processo de ETL (*extract, transform and load*) foi realizado através da ferramenta Talend (https://www.talend.com). É uma solução open source, consolidada no mercado e muito bem avaliada pelo Gartner. A título de curiosidade segue abaixo o quadrante referente as soluções de integração de dados do ano de 2020.



Quadrante Gartner

As estruturas de banco de dados utilizadas nesta solução (controle de carga, *staging area* e *data warehouse*) foram criadas no PostgreSQL (https://www.postgresql.org).

## 3.2 Staging area

A primeira parte do processo consiste em buscar todos dados necessários para a solução deste projeto: Banco Central e IBGE. No site dos dados abertos do Banco Central (<https://dadosabertos.bcb.gov.br/>) foram recuperadas as informações do SCR e taxa Selic disponibilizadas mensalmente e no IBGE foram recuperados os dados socioeconômicos (população, desemprego, renda per capita e PIB (produto interno bruto) disponibilizado anualmente pela instituição. Os dados foram consolidados na área de staging que consiste em uma camada intermediária do processo ETL para acolher os dados que serão transformados para o Data Warehouse. Abaixo modelo de dados da stage conforme figura 04.

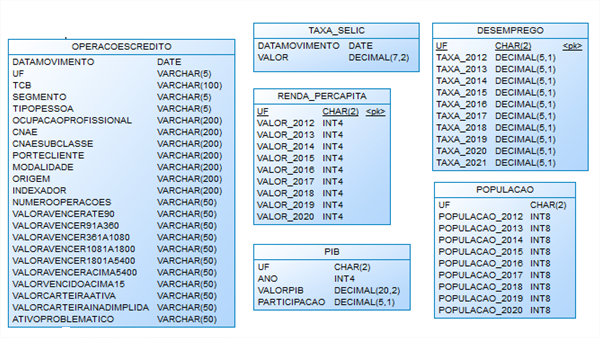


Figura 4: modelo stage

A solução foi construída para que todo o processo ocorra de forma automática, sem nenhum tipo de interferência manual. Para que isso ocorresse, foi necessário criar uma camada de controle onde está armazenado a última data de movimento carregada, conforme figura 05.

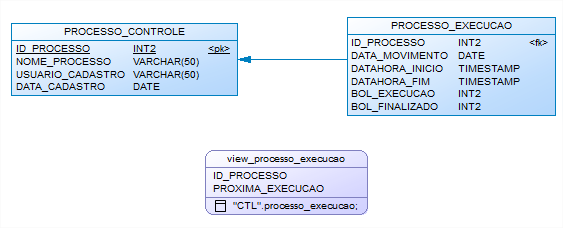
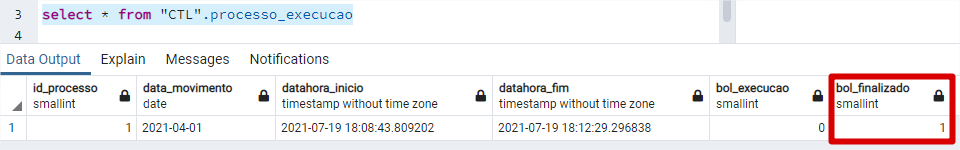


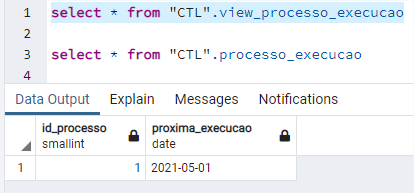
Figura 5: modelo controle

A view\_processo\_execucao tem por objetivo responder automaticamente a próxima data disponível para a execução. Isso não quer dizer necessariamente que existem dados disponíveis. A controle tem um caráter apenas informativo, sendo que o processo ETL é que faz a verificação se existem dados novos para serem carregos no DW. Na tabela PROCESSO\_EXECUCAO, pode-se verificar que existem dois campos do tipo *boolean* que informam o status do processo. É através destes dois indicadores que a *view* irá responder a próxima data de execução. Se o indicador BOL\_FINALIZADO for igual a 0, significa que o processo de carga foi finalizado com erro. Neste caso a *view* irá responder a data da tabela PROCESSO\_EXECUCAO. Caso o campo for igual a 1, ou seja, processo finalizado com sucesso, a view responderá a data com a próxima execução. Segue exemplo abaixo para ilustrar o mecanismo de resposta da *view*.

Data\_movimento = ‘2021-04-01’ e bol\_finalizado = 1 -> view irá responder com o próximo mês de execução. Neste caso a resposta no resultado do select é ‘2021-05-01’.

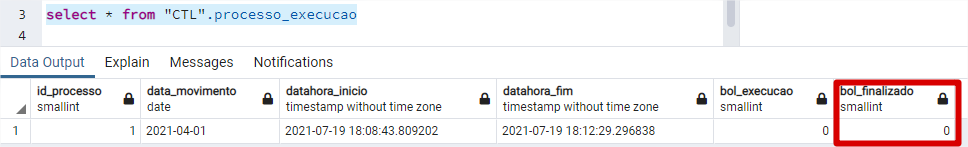


Controle – bol finalizado = 1

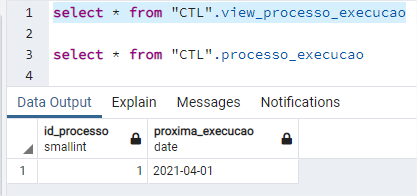


View processo\_execucao

Data\_movimento = ‘2021-04-01’ e bol\_finalizado = 0 -> view irá responder com a mesma data da tabela processo\_execucao, visto que o processo não foi finalizado com sucesso. Neste caso o retorno esperado do *select* é ‘2021-04-01’.

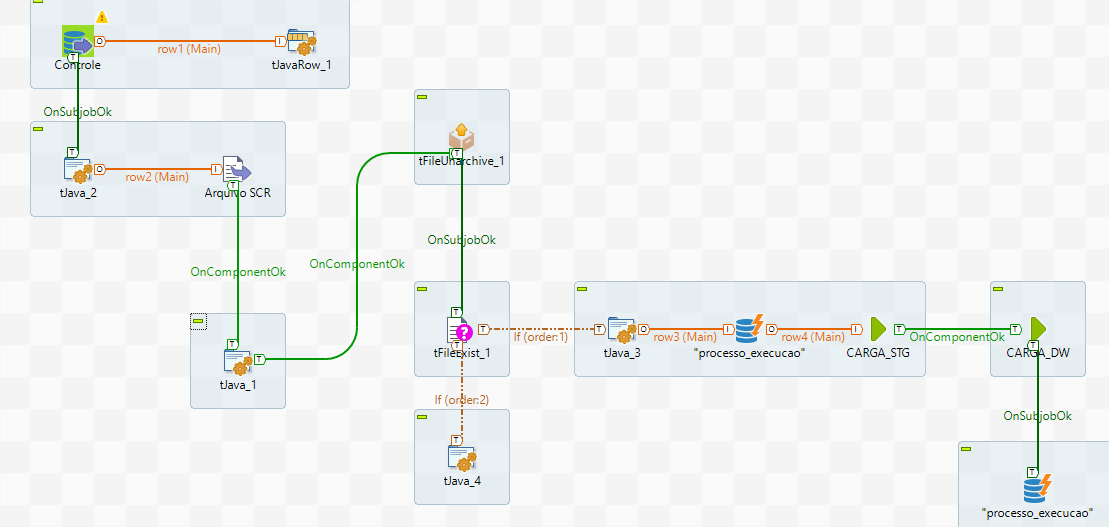


Controle – bol finalizado = 0



Controle view processo execucao

Segue abaixo fluxo master do processo de carga criado no Talend. Visto que a principal informação para este projeto são os dados do SCR, o gatilho para a execução do fluxo, consiste em verificar se existe dados mensais para o ano/mês que a controle responder. Caso não exista o processo é encerrado. Se existir, o fluxo segue para a execução dos *jobs* da *stage* e dimensional, representados neste fluxo com CARGA\_STG e CARGA\_DW.



Fluxo Talend

Uma informação importante para a execução deste fluxo, é que a URL disponibilizada pelo BACEN para recuperar os dados do SCR (<https://dadosabertos.bcb.gov.br/dataset/scr_data>), é referente ao ano. Desta forma para o processo ser automatizado, recupera-se o ano do retorno do select da view de controle ("https://www.bcb.gov.br/pda/desig/planilha\_"+context.ano+".zip") e verifica se existe arquivo para o ano/mês correspondente. No fluxo este processo é realizado pelo componente tFileExist.

Para os dados da taxa Selic, o BACEN disponibiliza uma API que retorna um arquivo do tipo *json, (*[*https://api.bcb.gov.br/dados/serie/bcdata.sgs.11/dados?formato=json*](https://api.bcb.gov.br/dados/serie/bcdata.sgs.11/dados?formato=json)*).* O processo extrai os campos (data e valor) e consolida na *staging area*  conforme as figuras 11 e 12.

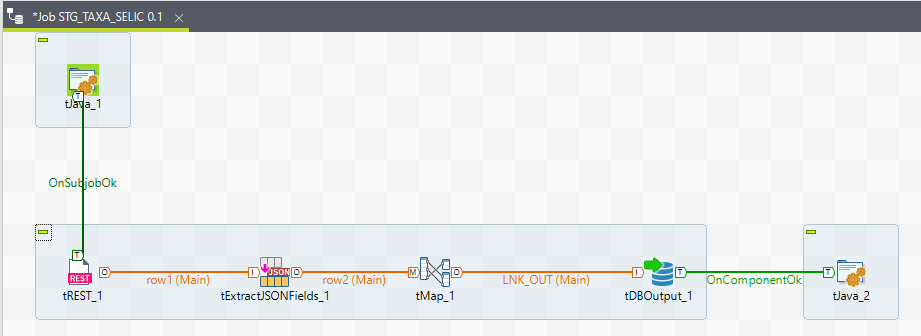


Figura 11: carga taxa selic

.

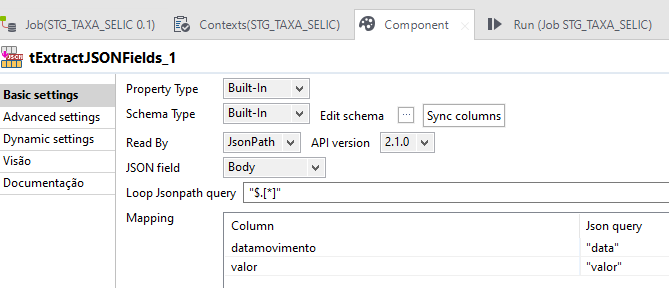
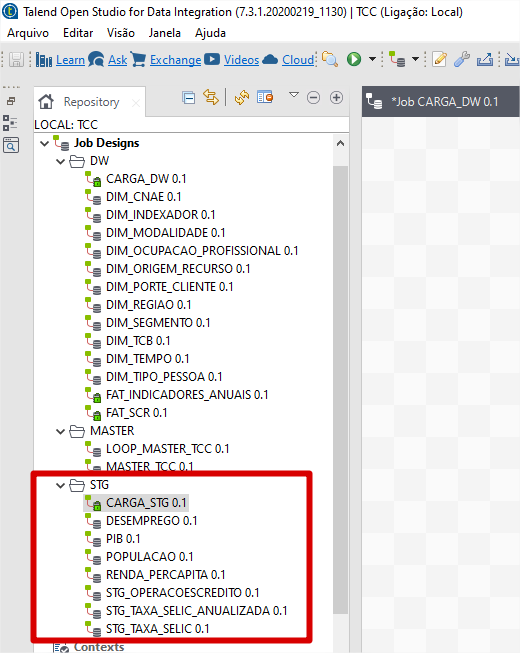


Figura 12: tratamento json

O restante das informações são planilhas excel, onde foi realizado o mapeamento dos campos para cada um dos assuntos da stage.



Jobs stage Talend

## 3.3 Detalhamento da origem dos dados

Segue abaixo prints de tela detalhando a origem das informações e devidas evidências dos dados carregados nas estruturas criadas para suportar a solução.

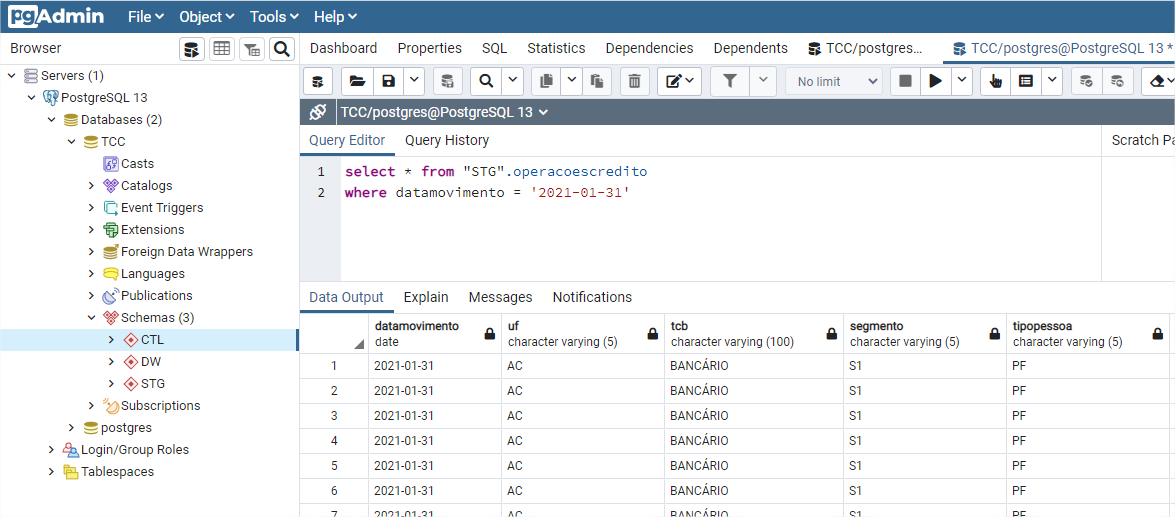
* SCR



Dados mensais SCR



Exemplo mensal 2012



Select operacaocredito



Group by data operacaocredito

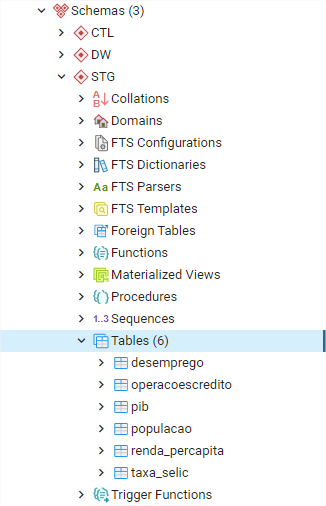
* IBGE



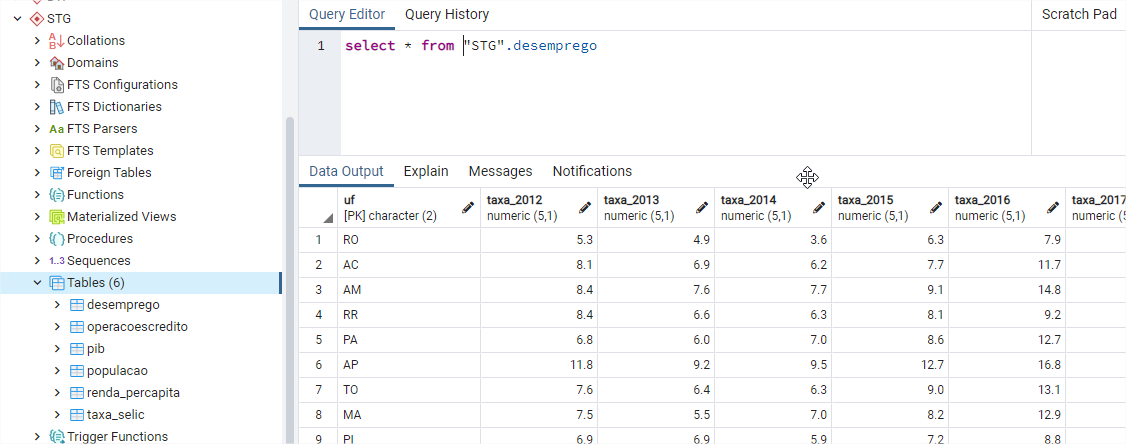
Dados IBGE



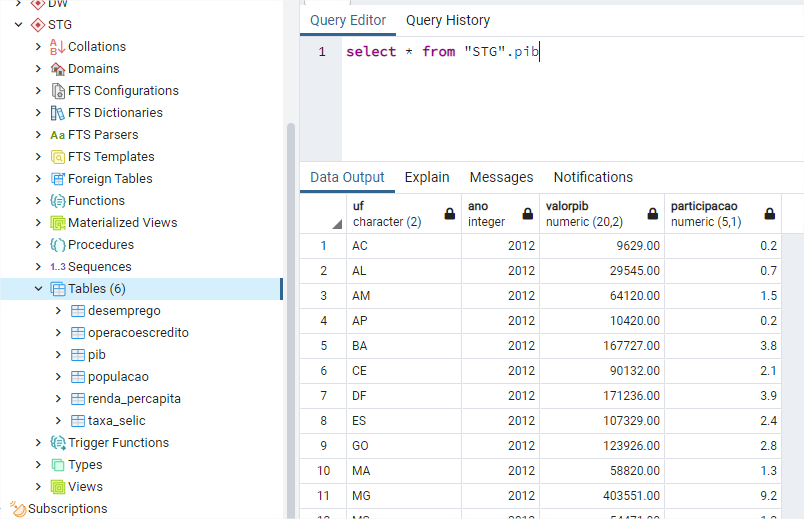
Dados IBGE



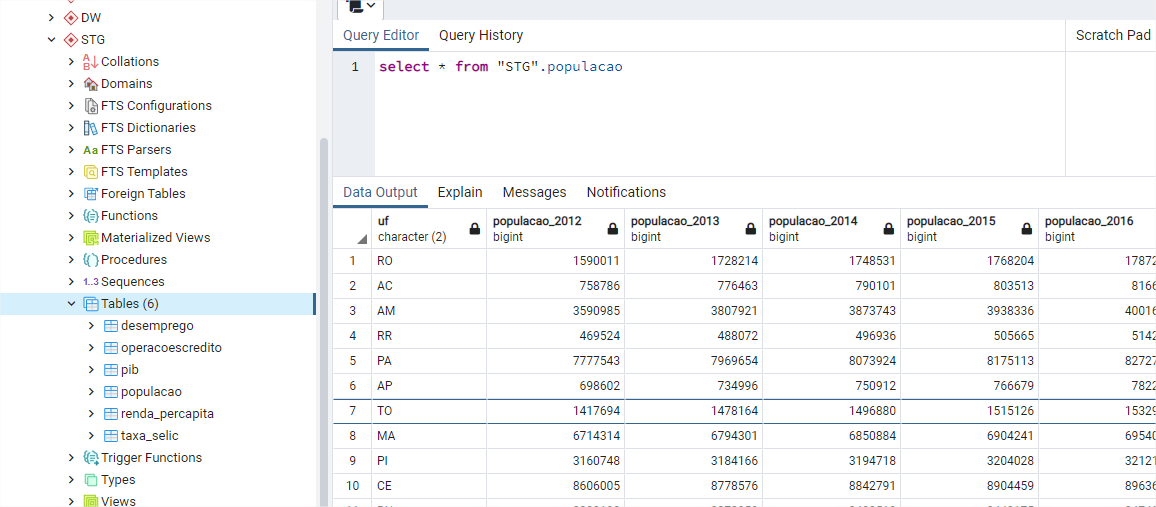
Tabelas stage



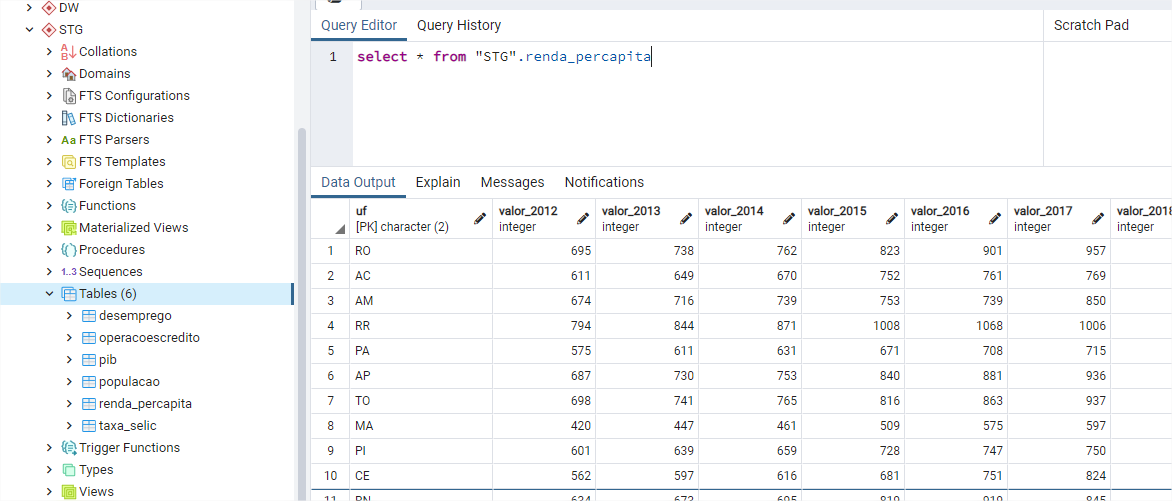
Select desemprego



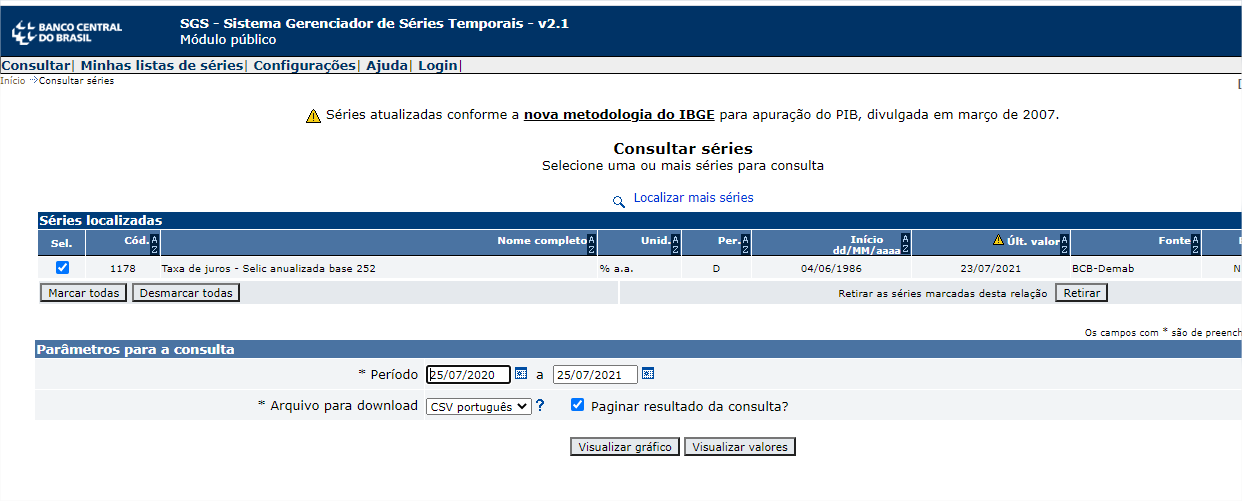
Select pib



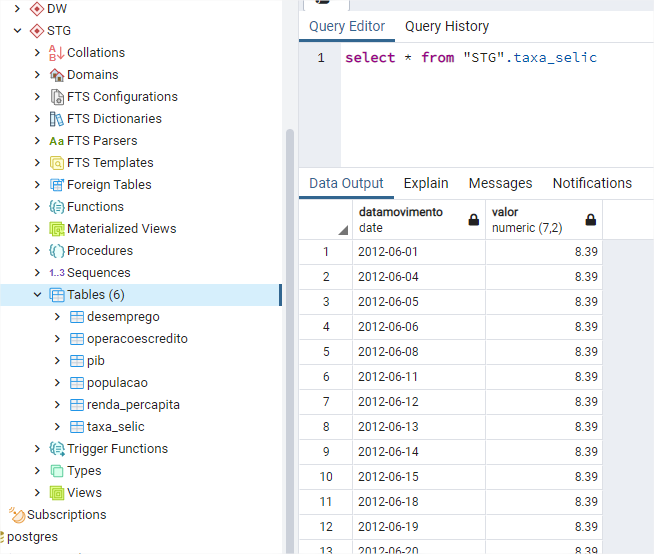
Select populacao



Select renda per capita



Consulta BACEN taxa Selic – tabela 1178



Select taxa selic

## 3.4 Tratamento dos dados

Para todas as interfaces trabalhadas na camada da *staging area*, foram realizados os seguintes tratamento nos dados:

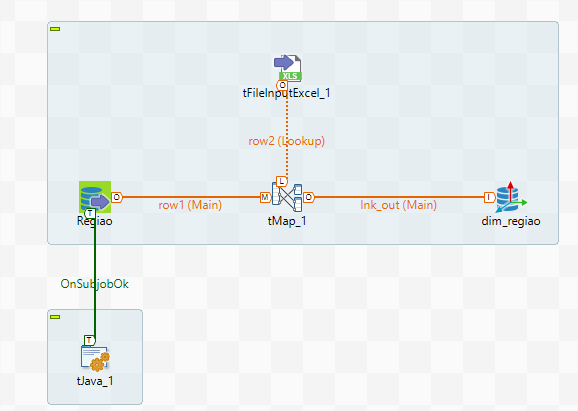
1. Retirar os espaços em branco que porventura ocorreu no final de cada registro;
2. Padronizar as datas no formato ano, mês e dia (YYYY-MM-DD);
3. Padronizar os campos numéricos com duas casas demais e separados de decimal com vírgula (,);
4. Higienização dos campos sem informação. Ora continha traços, espaços em branco, pontos. Para todos estes casos houve tratamento deixando o campo em branco, sem nenhum tipo de caractere.

## 3.5 Dimensional

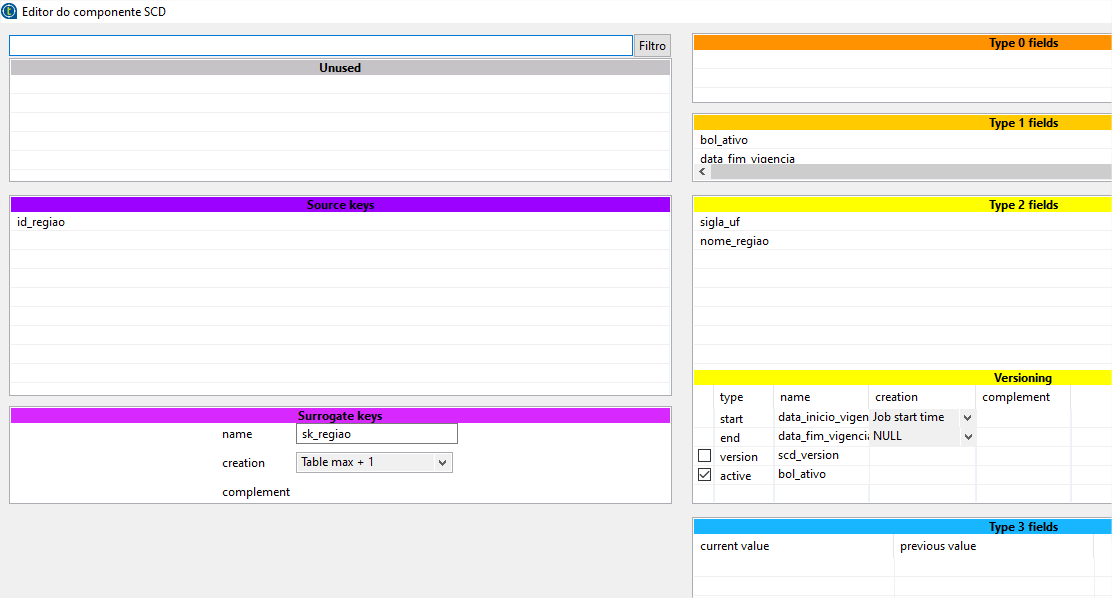
A segunda parte do processo de integração consiste na carga do modelo dimensional. Conforme comentado anteriormente, o modelo em questão foi confeccionado no formato *star schema,* sendo que para carga das dimensões foi utilizado a estratégia de SCD (*slowly changing dimension*). O SCD forma adotada para armazenar os dados nas dimensões, podendo ser realizado de 3 formas:

* Tipo 0: nenhuma ação é tomada no registro (nem atualiza, nem versiona);
* Tipo 1: o registro é atualizado, não armazenando histórico;
* Tipo 2: o registro é versionado, ou seja, caso a chave negocial definida já existir, fecha-se a vigência do registro e é então criado um novo registro, onde para cada um é definido sua respectiva data de início e fim de vigência.
* Existem também o tipo 3 onde, para alguns campos selecionados é armazenado os valores anteriores e atuais, tanto no registro fechado como no registro atual. Porém é uma técnica pouca usada em projetos de Data Warehouse.

No Talend é facilmente possível criar as dimensões utilizando técnicas de SCD pelos componentes já existentes. Abaixo é apresentado um exemplo de como a carga de uma dimensão é realizada. Nas images abaixo é apresentado um exemplo de job de carga de dimensão bem como a configuração do SCD.



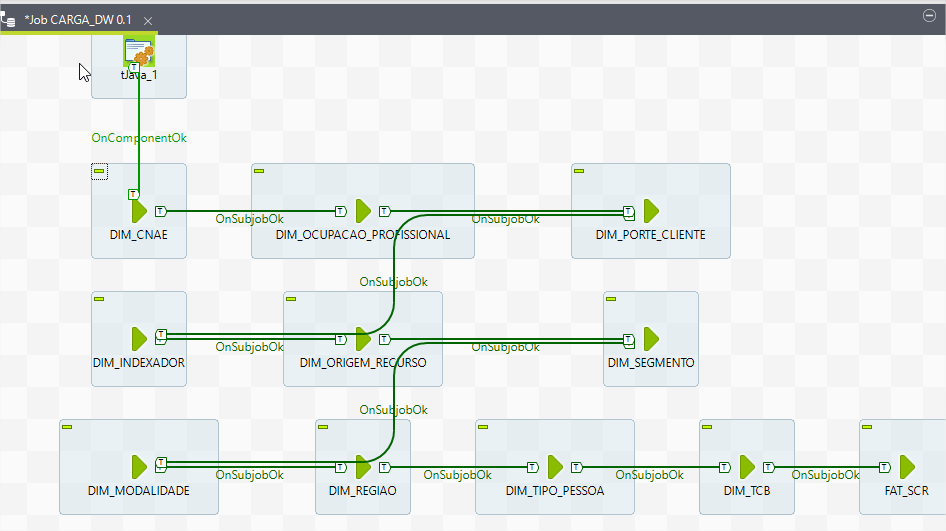
Carga dimensão região



Estratégia de SCD

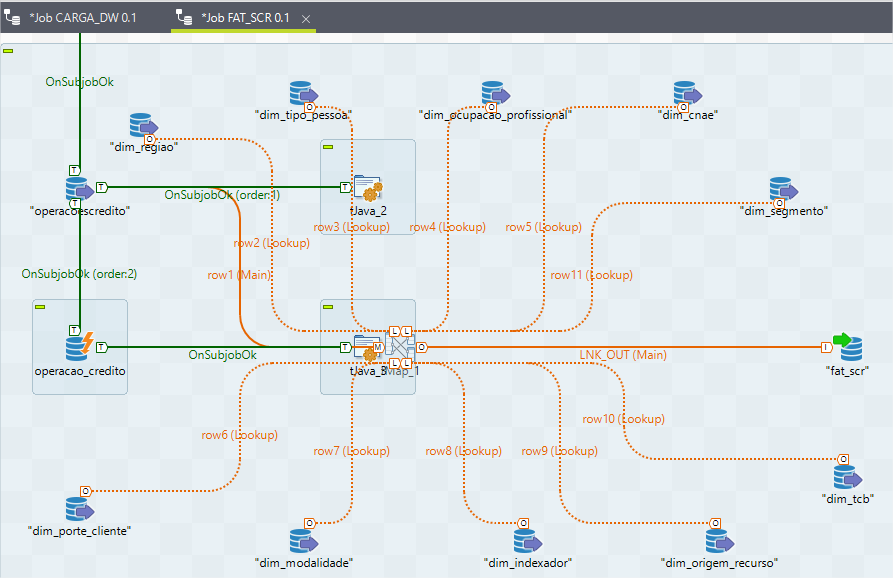
Para cada dimensão foi realizada também a inserção de valores *dummy* que consiste nas SKs (surrogate Keys) com valores -1 (não se aplica) e -2 (não informado). O objetivo e relacionar registros na fato com estas SKs em situação onde a informação não existe origem (não informado) ou não se aplica a situação do contexto (exemplo: estado civil para pessoa jurídica).

Com a carga das dimensões devidamente finalizadas, é realizado então carga das fatos SCR e indicadores anuais, armazenando o histórico mensal e anual respectivamente. A imagem abaixo representa o fluxo de carga do modelo dimensional. Não foi possível criar nenhum tipo de paralelismo devido a restrições na versão *free* da ferramenta.

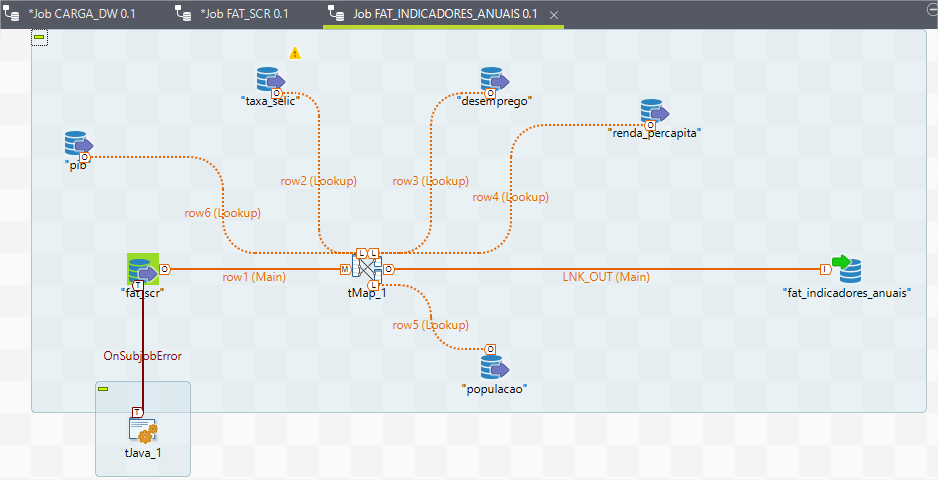


Fluxo carga DW

E na imagem abaixo carga das tabelas fato SCR e indicadores anuais, respectivamente imagens abaixo.



Carga fato SCR

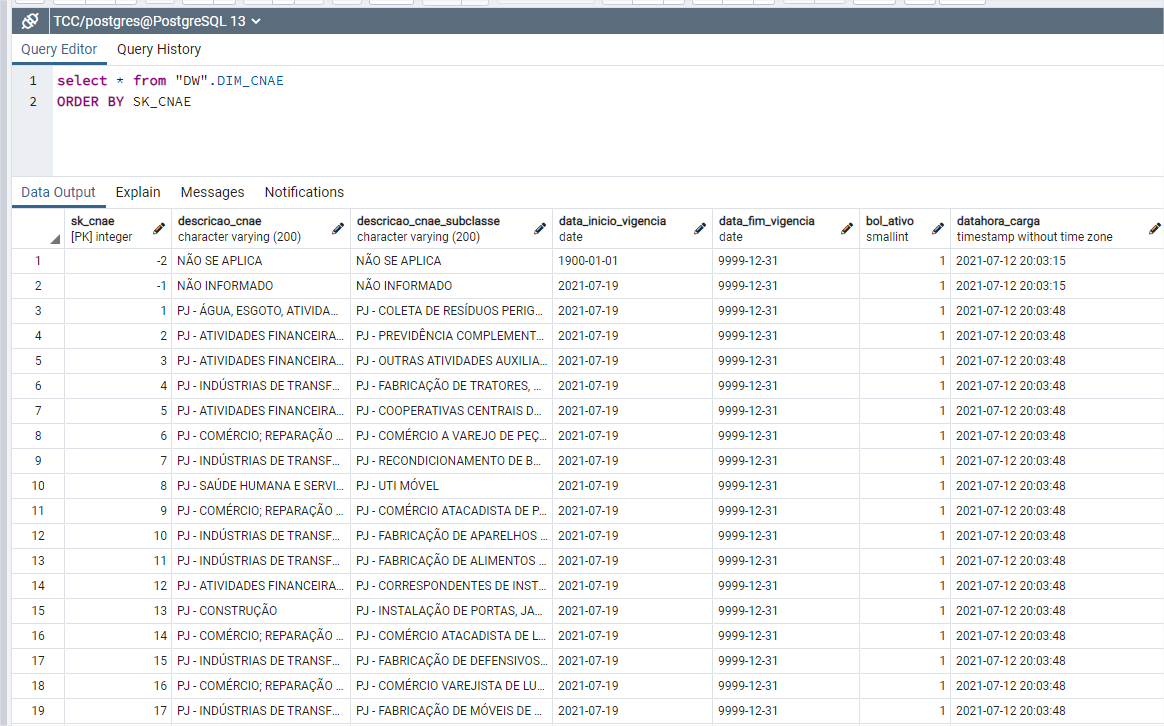


Carga fato indicadores anuais

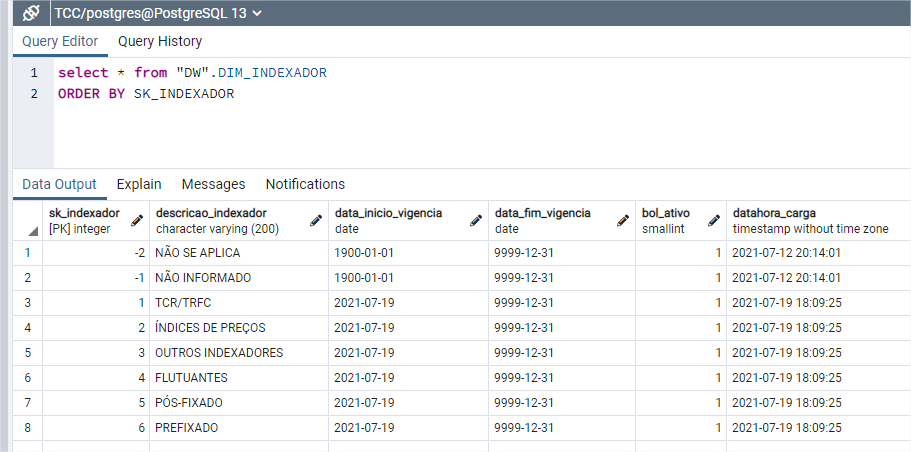
## 3.6 Detalhamento cargas dimensional

Segue prints com detalhamento das cargas realizadas nas estruturas criadas para suportar o modelo dimensional.

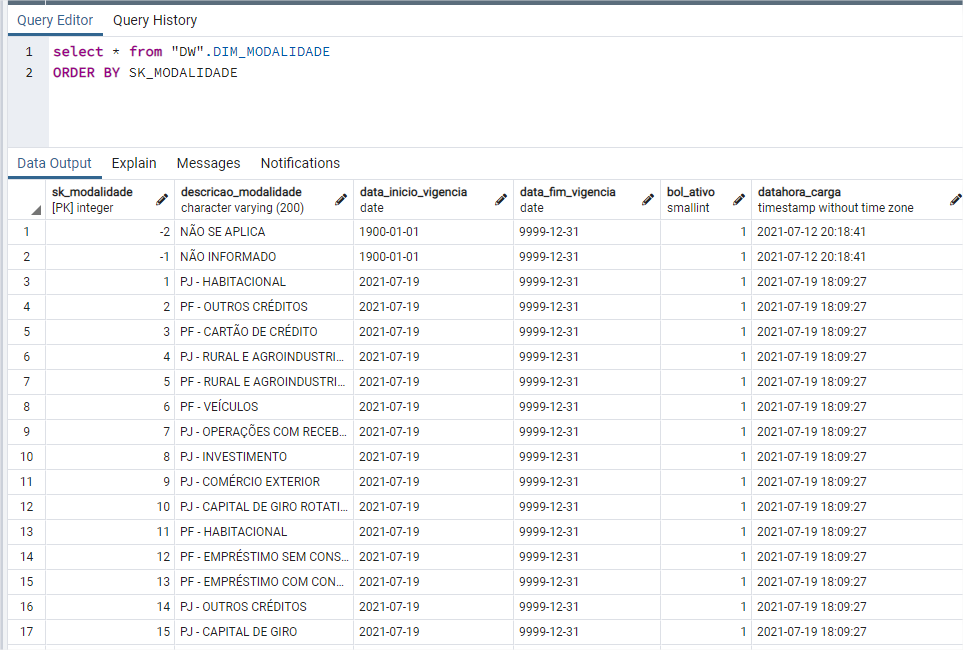
* Dimensões



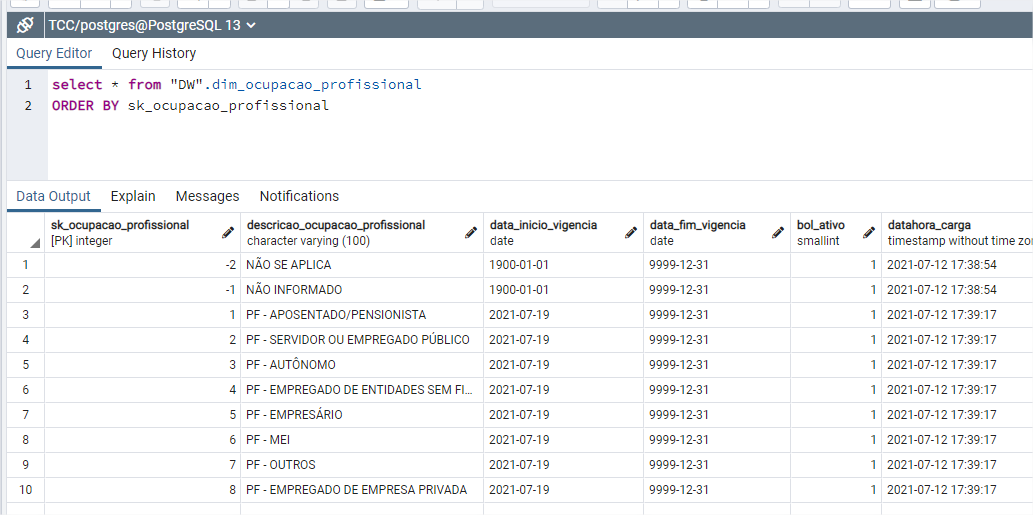
Dimensão CNAE



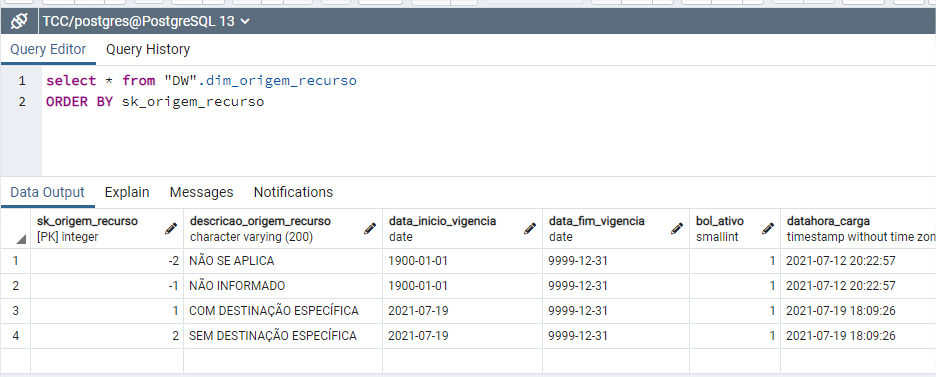
Dimensão indexador



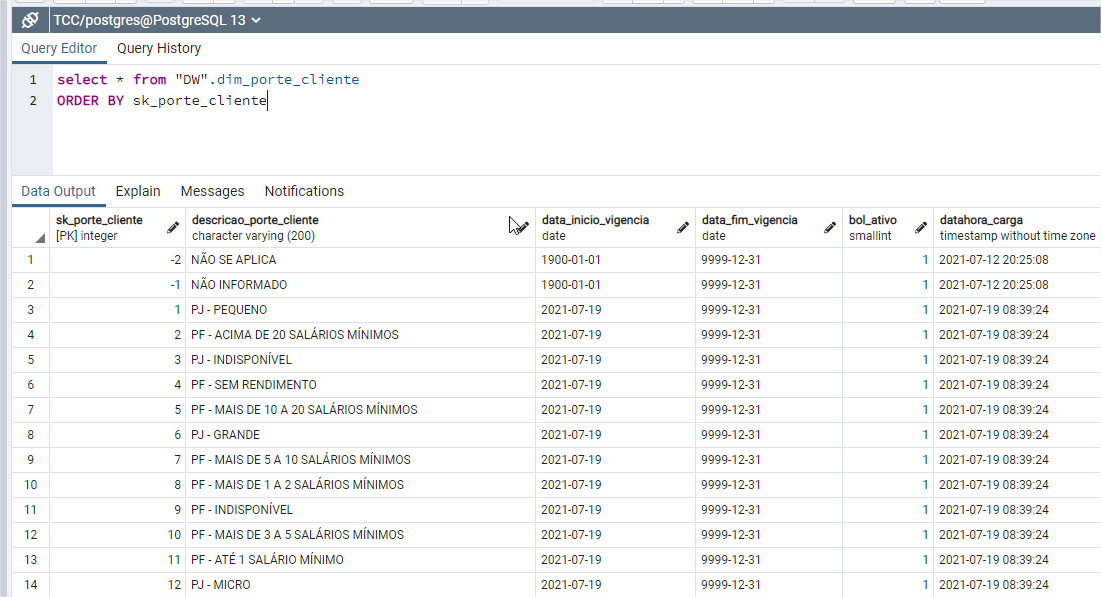
Dimensão modalidade



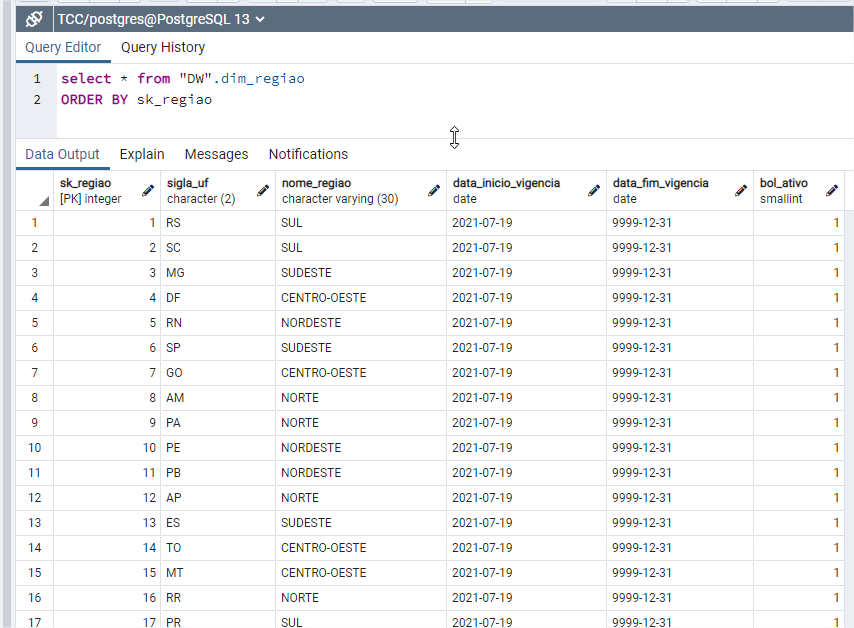
Dimensão ocupação profissional



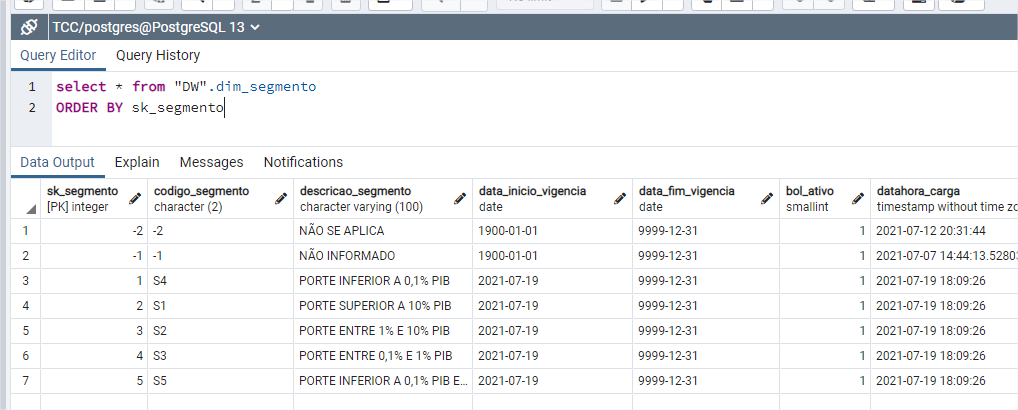
Dimensão origem recurso



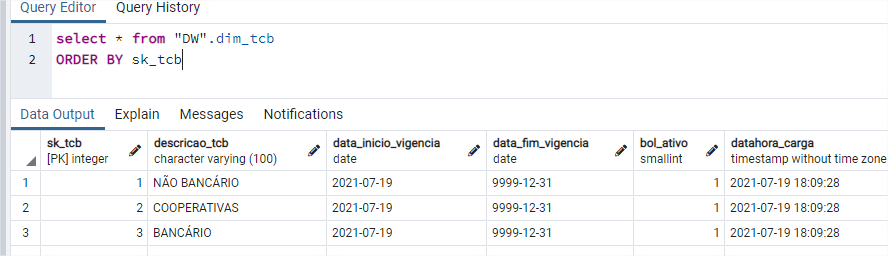
Dimensão porte cliente



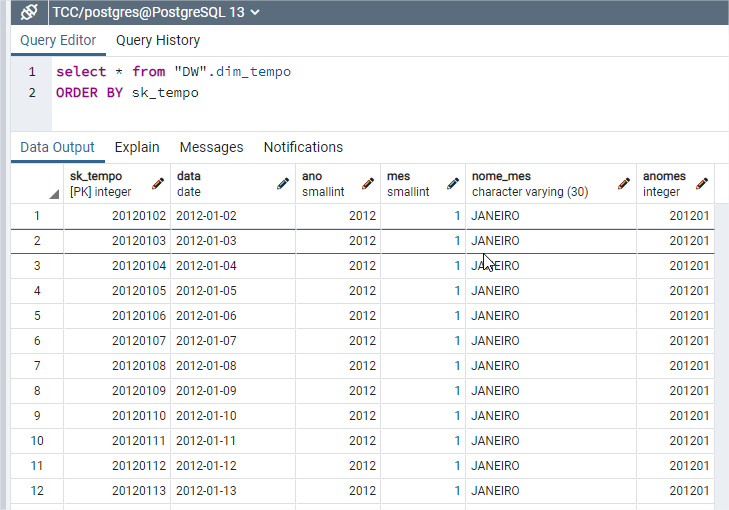
Dimensão região



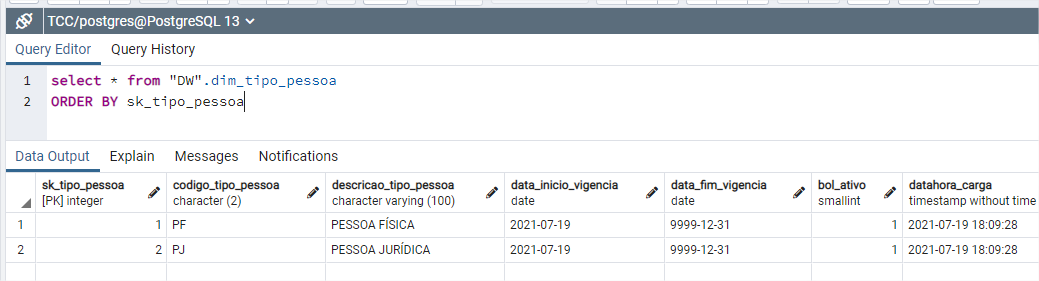
Dimensão segmento



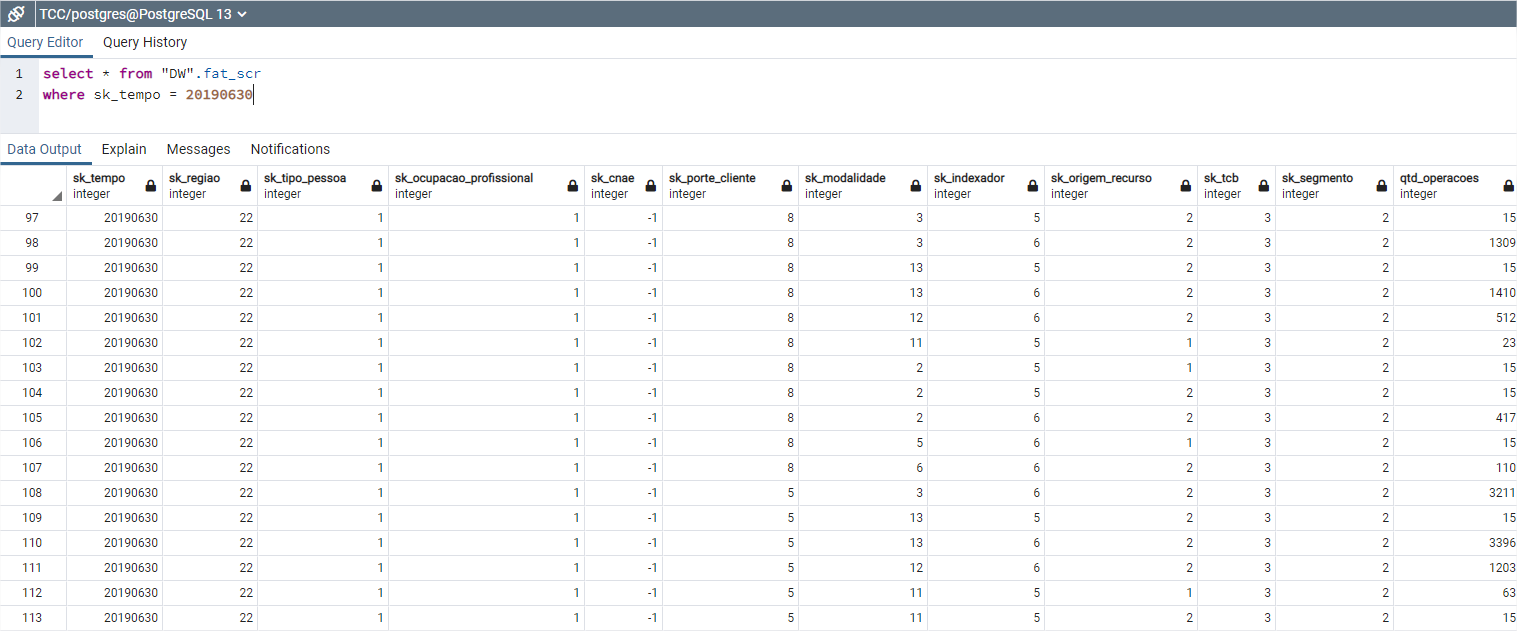
Dimensão tcb



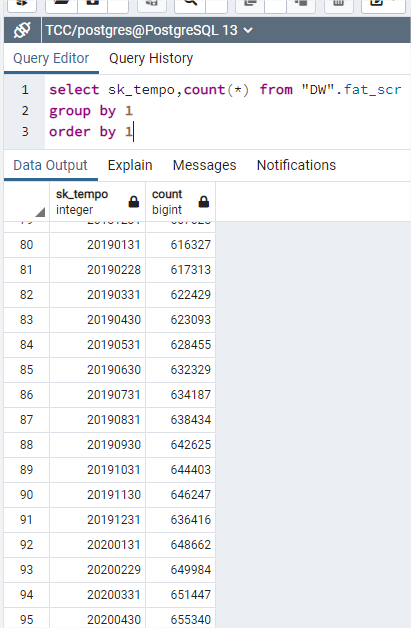
Dimensão tempo



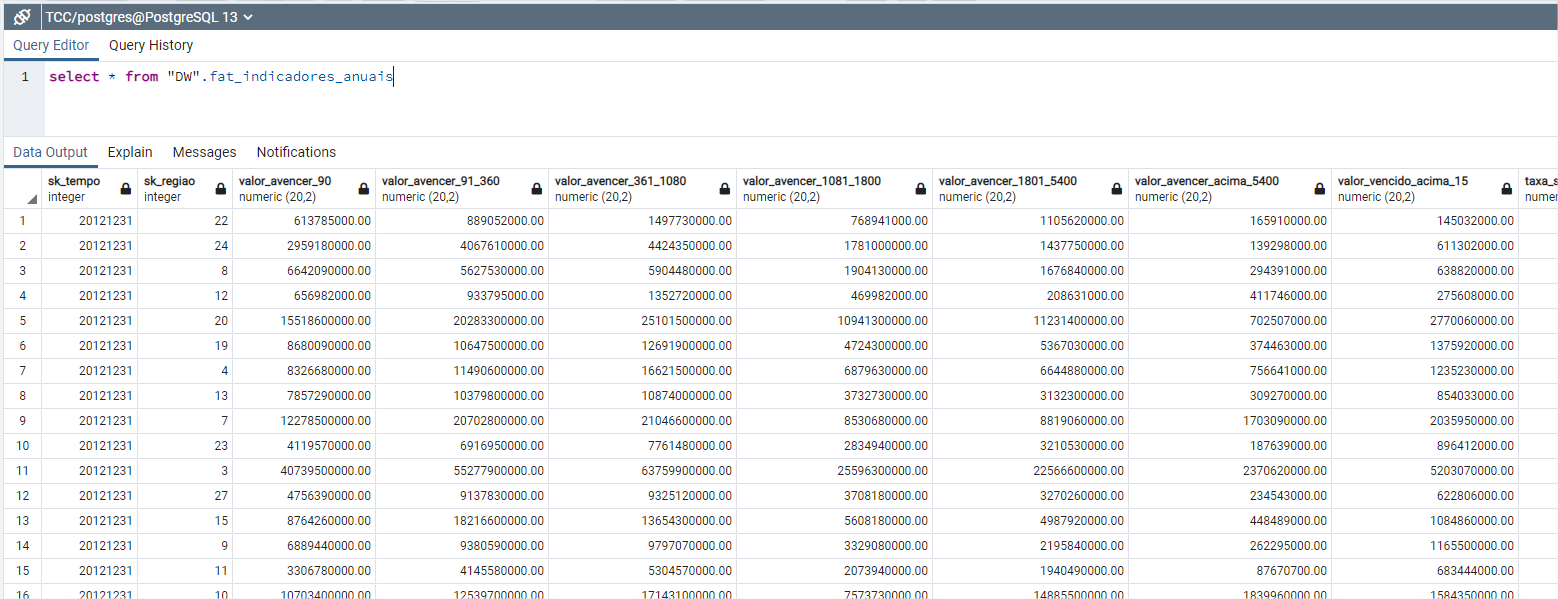
Dimensão tipo pessoa



Fato SCR



Quantitativo fato scr



Fato indicadores anuais

# 4. Camada de Apresentação

## 4.1 Dashboard

Para a elaboração dos dashboards, a escolha deste autor foi utilizar a ferramenta Power BI. A solução apresenta todos os recursos necessários para realizar as entregas idealizadas para este projeto.

Apesar da ferramenta permitir também tratar e transformar os dados, toda esta etapa já foi realizada anteriormente através do Talend, conforme evidências do item 3 deste trabalho. Desta forma, bastou importar o modelo para o Power BI e então desenvolver os dashboards.

Diagrama

Descrição gerada automaticamente com confiança baixa

Modelo de dados - Power BI

Os dashboards foram elaborados em cima dos dois modelos dimensionais criados para este projeto: SCR e indicadores anuais. O painel de indicadores anuais possui um viés estratégico onde são apresentados os principais indicadores socioeconômicos do Brasil. Já os painéis do SCR têm como objetivo principal apresentar uma visão tática dos principais números gerados pelas instituições financeiras que são reportadas ao Banco Central.

Através dos indicadores anuais, onde as informações estão agrupadas em um contexto mais amplo (anualizadas), é possível obter informações de PIB, inadimplência do SCR, taxa Selic, população, renda percapita e taxa de desemprego. Foi carregado um histórico de 09 anos, entre 2012 e 2020.

Gráfico

Descrição gerada automaticamente

Na visão SCR, foram disponibilizados dois painéis principais. O primeiro tem como objetivo apresentar uma visão com valores médios da taxa de inadimplência. Já o segundo painel permite uma navegação entre as informações mensais disponibilizadas pelo Banco Central.

Gráfico

Descrição gerada automaticamente

Uma imagem contendo Diagrama

Descrição gerada automaticamente

Foi disponibilizado também uma terceira visão, esta mais operacional, onde os dados são apresentados em um menor grão, podendo assim obter o detalhamento dos números utilizados para geração dos dashboards apresentados acima.

Tabela

Descrição gerada automaticamente

# 5. Registros de Homologação

Após todo o desenvolvimento concluído, foi realizado uma série de validações para confirmar que os indicadores apresentados estão de acordo com dados carregados na base TCC-DW. O documento encontra-se em anexo a esta entrega, intitulado “Relatório de homologação”.



# 6. Análises Avançadas

## 6.1. Machine Learning

Esta etapa do projeto tem como objetivo a aplicação de técnicas de *Machine Learning* para gerar prospecções futuras de alguns indicadores chaves, fruto de estudos deste projeto. Antes de detalhar os estudos realizados, segue brevemente algumas informações importantes a respeito do conceito de Machine Learning.

*Machine Learning*, termo em inglês para aprendizado de máquina, é um método de análise de dados que automatiza a construção de modelos analíticos. Podemos dizer que é um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com os dados de entrada e então identificar padrões e consequentemente tomar decisões com o mínimo de intervenção humana. Com as técnicas avanças existentes hoje em dia, é possível produzir respostas rápidas, modelos capazes de analisar dados de diversos tipos, de diversas complexidades e entregar resultados precisos. Com todos estes artifícios, as organizações tem maiores chances de identificar oportunidades tanto lucrativas ou evitar riscos desconhecidos. Existem dois métodos principais utilizados na aplicação de *Machine Learning*: o aprendizado supervisionado e não supervisionado. Não são os únicos, porém os mais populares.

Os algoritmos de aprendizado supervisionado são treinados por meio de exemplos rotulados, onde são recebidos um conjunto de dados de entrada juntamente com as saídas correspondentes onde o aprendizado ocorre na comparação da saída real com as saídas corretas para encontrar os erros. Este tipo de aprendizado utiliza padrões para prever os valores de rótulos em dados não rotulados e é comumente aplicados em dados de séries históricas com o objetivo de prever eventos futuros. As técnicas mais utilizadas são as classificações, regressões e *gradient boosting*.

Por outro lado, o aprendizado não supervisionado é utilizado com dados que não possuem rótulos históricos. O algoritmo é que tem a função de descobrir o que deve ser apresentado. É um tipo de aprendizado que é mais utilizado em dados transacionais.

Para este projeto foi aplicado algoritmo de aprendizado supervisionado utilizando técnicas de regressão linear. Esta técnica tem por objetivo estimar o valor de algo baseado em uma série de dados históricos, ou seja, prever o futuro baseado em eventos passados. As principais métricas de validação do modelo são:

* Soma dos quadrados dos resíduos: medida da variação que não pode ser explicada;
* R²: medida estatística de quão próximos os dados estão da linha de regressão;
* Erro médio absoluto: calcula o valor dos resíduos para cada um dos pontos;
* Médio dos erros ao quadrado: é a média dos erros elevada ao quadrado.

## 6.2. Escopo

O objetivo final da análise avançada foi estabelecer uma correlação entre as informações obtidas no SCR e dados socioeconômicos de forma que fosse possível realizar uma predição da taxa de inadimplência não apenas com os números de carteira ativa e inadimplência do SCR mas também através dos dados socioeconômicos. Os dados em questão para esta análise foram:

* Taxa de desemprego;
* Renda per capita;
* População;
* PIB;
* Taxa de alfabetização;
* Taxa Selic.

Para obter dados mais concretos e trabalhar na mesma granularidade, os dados foram agrupados por ano e região geográfica sendo Norte, Nordeste, Sul, Sudoeste e Centro-Oeste com histórico de 9 anos, 2012 a 2020.

## 6.3. Ambiente tecnológico

Para trabalhar com os dados e aplicar as técnicas de *Machine Learning*, foi criado em uma VM (*Virtual Machine*) um ambiente com sistema operacional Linux Ubuntu, Python 3.8.10 e Pyspark versão 3.2.0. Todas as linhas de código foram executadas através do Jupyter Notebooks e o arquivo .ipynb de nome “TCC” é parte também da entrega deste trabalho.

## 6.3. Detalhamento da solução

Os dados foram extraídos das fatos SCR e indicadores anuais apresentadas no capítulo 2 deste trabalho. Os dados foram agrupados por ano e região resultado no dataframe abaixo:

Tabela

Descrição gerada automaticamente

Foi verificado então a correlação entre as colunas e confirmar se seria possível seguir com a proposta de estudo deste trabalho. Para isso foi utilizado a função “corr” da classe “pyspark.sql.functions”.

Texto

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Uma imagem contendo Texto

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

A técnica de regressão linear no pyspark é acessada através da função “LinearRegression” pertencente a classe “pyspark.ml.regression”. Algumas premissas devem ser seguidas antes de utilizar a função:

1. Os campos utilizados na função devem estar em um vetor. Este passo foi realizado através da função “VectorAssembler” da classe “pyspark.ml.feature”.

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

1. Todos os campos utilizados na análise devem ser do tipo numérico.

Texto

Descrição gerada automaticamente

Após o passo da vetorização onde foi criado a variável “assembler” e o vetor armazenado na coluna “features” e então feito o merge com o dataframe “data” através da função “assembler”, resultando na seguinte saída:

Texto

Descrição gerada automaticamente

Uma vez os dados persistidos em um vetor, basta trabalhar com o vetor e a coluna do estudo, neste caso, “perc\_inad”, ou seja o percentual de inadimplência. Através do Pyspark é possível realizar a operação de *select* em um dataframe. Os dados foram então armazenados no dataframe final\_data.

Tabela

Descrição gerada automaticamente

Seguindo as boas práticas na criação de um modelo preditivo, os dados foram divididos em dois blocos: dados de teste, contendo 30% da massa de dados e dados de treinamento, contendo 70% dos dados. Essa divisão foi realizada através da função “randomSplit”. Com a massa de dados pronto, os dados foram então encaixados e validados. Estes passos foram executados através das funções “fit” e “evaluate”.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

O modelo foi então validado através da medida de variação, “rootMeanSquaredError” e R², onde chegou-se a uma accurace de 97% para a massa de teste.

Texto

Descrição gerada automaticamente

O mesmo procedimento foi realizado para a massa de treinamento, onde chegou-se a uma accurace de 87% e rootMeanSquareError de 0,26.

Texto

Descrição gerada automaticamente com confiança média

Validado o modelo preditivo, foi realizado o treinamento em cima de massa de dados para então comparar o percentual de inadimplência dos dados do SCR com a predição realizada. A chave para cruzar as informações são as colunas “ano” e id\_regiao, estas obtidas no desmembramento do vetor. Desta forma o resultado final apresentou o seguinte layout:

Tabela

Descrição gerada automaticamente

Os dados de inadimplência do SCR e do modelo preditivo foram transcritos para uma planilha para fins de comparação do resultado final chegando-se a uma margem de erro média de 8,9%. A validação pode ser verificada através da planilha “resultado\_final.xls”.

# 5. Conclusões

A elaboração deste trabalho trouxe um enorme aprendizado pela confirmação de que a melhor maneira de aprender é colocando literalmente a mão na massa. A busca dos dados abertos disponibilizados pelos órgãos também trouxe um grande desafio pois por algumas vezes os dados estavam incompletos, ou o link estava fora do ar ou nas inconsistências encontradas. Desta forma foi necessário realizar a busca em outras fontes para garantir que o projeto estava sendo realizado com os dados mais coerentes possíveis.

Válido destacar também a contínua disponibilização de dados abertos pelo governo federal pois estes são frutos de estudo e nos levam a perceber os resultados com olhares mais críticos. Este projeto não teria sido possível de sair do papel sem a disponibilização dos dados.

O estudo aqui realizado mostra a importância de uma sociedade com alto índice de educação e não apenas educação que aprendemos na escola ou universidade, mas sim de educação financeira. A inadimplência é um problema que persiste no Brasil e este cenário só será contornado quando a educação financeira por implementada nas escolas desde cedo, mostrando a importância de saber gerenciar o dinheiro, comprar aquilo que se pode e que seja necessário.

Por último, destacar os desafios tecnológicos para a construção deste projeto e a importância de ferramentas open source, aqui utilizadas Talend e Apache Spark.

# 6. Links

Todos os arquivos do projeto foram disponibilizados no GitHub. Podem ser acessados através do link abaixo:

<https://github.com/alexandrebaudon/TCC>

# 7. Referências

Como um projeto aplicado, as referências, levantadas no processo de revisão bibliográfica, são opcionais e estimulamos que o responsável pelo projeto apresente aquilo que for relevante para complementar o entendimento do trabalho. Dessa forma, relacione-as de acordo com o modelo a seguir.

<https://spark.apache.org/docs/latest/ml-classification-regression.html>

Introduction to Statistical Learning by Gareth Jamel, et al.

<https://dadosabertos.bcb.gov.br/dataset/scr_data>

<https://www.ibge.gov.br/acesso-informacao/dados-abertos.html>