Universidade Federal de Santa Catarina
Centro Tecnológico
Departamento de Informática e estatística
Bacharelado em Sistemas de Informação
Discriplina INE5643-07238 (20231) - Data Warehouse
Alunos: Cinthia Carolina Shiratori, Daniel Roberto de Aguiar, Levi Andrew de Aquino Paz, Marco César da Silva, Rafael Begnini de Castilhos, Ramon Seugling

Proposta de criação de um Data Warehouse:

Análise de admissões e dispensas no CAGED

1. INTRODUÇÃO

O presente trabalho visa realizar a construção de um Data Warehouse contendo os dados de admissões e dispensas disponibilizados pelo Cadastro Geral de Empregados e Desempregados (CAGED) e ocorre com a motivação de entender os setores que possuem os maiores saldos de contratações e suas faixas salariais por cada período e localidade no território brasileiro.

Atualmente os dados são disponibilizados em tabelas com dados mensais, além de tabelas adicionais de dados de exclusões e entregas fora do prazo, e ainda, os dados necessitam de documentação para entender cada uma das informações, dificultando a realização de análise dos dados. O data warehouse vai facilitar a análise dos dados, agrupando os dados de séries históricas e traduzindo os dados, facilitando e agilizando a realização de análises dos mesmos.

O público alvo deste trabalho é composto por Consultores e Profissionais de Recursos Humanos, que buscam entender as médias salariais na contratação por categoria e os volumes de contratações por localidade, e o Orgãos do Poder Público, que buscam informações de saldos de contratações por períodos e localidade, para entender o desenvolvimento das regiões e onde focar políticas públicas para estímulo ao emprego, além de buscarem o desenvolvimento ou estimular o desenvolvimento de cursos profissionalizantes.

2. COLEÇÃO DE DADOS

A coleção de dados do CAGED é disponibilizada em forma de microdados não identificados em formato de texto (.txt) e com delimitador ";" (codificado em UTF-8). O acesso aos dados pode ser feito via navegador, sendo recomendado pela fonte dos dados o uso do Microsoft Edge configurado para o modo Internet Explorer ou via servidor FTP, ajustando a conexão ao servidor ftp://ftp.mtps.gov.br/pdet/microdados/.

Para cada período de disponibilização dos dados, que ocorre de forma mensal, são disponibilizados três arquivos, o CAGEDEXCYYYYMM.txt, que traz os dados de exclusão de registros gravados em arquivos de períodos anteriores, CAGEDFORYYYYMM.txt, que traz dados de registros de períodos anteriores que foram entregues fora do prazo e o CAGEDMOVYYYYMM.txt, que traz os dados gerais das movimentações que ocorreram no período e foram entregues dentro do prazo.

Abaixo segue a tabela 1, que traz um descritivo dos dados disponibilizados pela fonte e seus conteúdos. As três tabelas disponibilizadas mensalmente pela fonte possuem a mesma estrutura de dados, o que facilita na integração dos mesmos.

TABELA 1: Descrição dos dados e seus conteúdos

Variável	Descrição	Código
competênciamov	Competência da movimentação (anteriormente competência)	<AAAAMM>
região	Região geográfica de acordo com o código do IBGE	<99>
uf	Unidade da federação de acordo com o código do IBGE	<99>
município	Código do Município	<999999>
seção	Código da seção do CNAE 2.0	<N>
subclasse	Código da subclasse do CNAE 2.0	<9999999>
saldomovimentação	Valor da movimentação em termos de saldo	<99>
categoria	Categoria de trabalhador	<999>
cbo2002ocupação	Código da ocupação do trabalhador de acordo com a CBO 2002	<999999>
graudeinstrução	Grau de instrução do trabalhador	<99>
idade	Idade do trabalhador	<999>
horascontratuais	Horas contratuais semanais	<99>
raçacor	Raça ou cor do trabalhador	<9>
sexo	Sexo do trabalhador	<9>
tipoempregador	Tipo de empregador	<9>
tipoestabelecimento	Tipo de estabelecimento	<9>
tipomovimentação	Tipo de movimentação	<99>
tipodedeficiência	Tipo de deficiência do trabalhador	<9>
indtrabintermitente	Indicador de trabalhador intermitente	<9>
indtrabparcial	Indicador de trabalhador parcial	<9>
salário	Salário mensal declarado	<999999999,99>
tamestabjan	Faixa de Emprego no início de Janeiro do Estabelecimento	<99>
indicadoraprendiz	Indicador de trabalhador aprendiz	<9>
origemdainformação	Origem da informação da movimentação	<9>
competênciadec	Competência da declaração	<AAAAMM>
competênciaexc	Competência da exclusão	<AAAAMM>
indicadordeexclusão	Indicador de Exclusão	<9>
indicadordeforadoprazo	Indicador de informação declarada fora do prazo	<9>
unidadesalariocódigo	Unidade de pagamento da parte fixa da remuneração	<99>
valorsalariofixo	Salário base do trabalhador, correspondente à parte fixa	<999999999,99>

A natureza dos dados da fonte trata-se de registro administrativo, captados dos sistemas eSocial, Caged e Empregador Web. A periodicidade da sua disponibilização é mensal e sua abrangência é para todo o território nacional, onde suas desagregação geográfica se dá por: Brasil, Regiões, Unidades Federativas e Municipais. O número de estabelecimentos declarantes é cerca de 900 mil por mês. Os rendimentos representam os salários de fluxo dos admitidos e desligados, e não da totalidade do estoque de trabalhadores, isto é, correspondem aos salários que constam na Carteira de Trabalho.

3. CONTEXTUALIZAÇÃO DO PROBLEMA

De acordo com o IBGE, no 4º trimestre de 2022, tivemos 8,6 milhões de desempregados, uma taxa de desemprego de 7,9%, 4 milhões de desalentados, e uma taxa de subutilização de 18,5%. Ainda segundo o IBGE, na pandemia tivemos 400 mil empregos perdidos e 100 mil empresas encerraram suas atividades. Entender de forma mais detalhada os setores que sofreram as maiores quedas e as regiões que sofreram mais com estes problemas é de grande importância para os tomadores de decisão do setor público, no intuito de gerar subsídios para as decisões de políticas públicas para o desenvolvimento do emprego no país.

A base do CAGED trata-se de um conjunto de informações que possibilita o cálculo do índice de emprego, taxa de rotatividade e a flutuação de emprego, desagregados em nível geográfico, setorial e ocupacional. Permite igualmente a obtenção de dados sobre os atributos dos empregados admitidos e desligados: gênero, grau de escolaridade, faixa etária, salários e tempo de emprego.

Sendo assim, as principais perguntas estratégicas que se buscará responder com este trabalho são:

- Qual o saldo de contratações anual de 2020 a 2022?
- Qual o número de contratações por área profissional, região?
- Qual a média salarial por área profissional, região e grau de escolaridade?
- Qual a média salarial por fatores socioeconômicos?
- Qual o saldo de contratações de pessoas com deficiência?

4. ESCOPO

Este projeto se justifica pela importância para a sociedade no geral com os ganhos obtidos por políticas públicas, visando expandir os níveis de empregos no país e gerando assim desenvolvimento econômico. Além disso, também traz benefícios para as organizações, garantindo

um norte para a realização das contratações, entendendo onde estão regionalmente sendo contratados os profissionais das áreas e funções que buscam contratar e as médias salariais pagas aos mesmos nas contratações.

Sendo assim, o escopo do trabalho será composto por dados históricos desde janeiro de 2020, período em que o CAGED começou a divulgar os dados no formato atual, a atualização dos dados será mensal, assim como a granularidade dos dados também será mensal, em relação ao período, pois é a menor granularidade entregue pela fonte. Em questões de localidade, a granularidade será por municípios, menor granularidade de localidade entregue pela fonte.

Como não existem dados sensíveis na fonte, não serão previstas exclusões de dados. Porém os dados só serão considerados a partir de 2020, pois os dados anteriores a este período possuem métricas diferentes, dificultando a realização de análises de séries históricas junto aos dados atuais.

Os fatores críticos de sucesso considerados para este projeto serão:

- Prover fonte única de disponibilização de dados do CAGED para a realização de análises pelo público alvo definido;
- Reduzir o tempo gasto pelo público algo nas análises de dados sobre admissões e dispensas do CAGED;
- Utilização efetiva por órgãos públicos na tomada de decisões relacionadas a políticas públicas referentes ao emprego;
- Utilização efetivas por consultores e profissionais de RH nas suas análises por profissionais e faixas salariais;
- Geração de pelo menos três políticas públicas por órgãos públicos baseadas em dados disponibilizados por este projeto no período de 18 meses após a publicação do mesmo.

Os riscos relacionados a este projeto que deverão ser considerados em sua construção e seus planos de contingência são os seguintes:

- Perda de dados por problemas em seu processamento;
    - Que deverão ser mitigados com o projeto cuidadoso da arquitetura utilizada.
- Descontinuidade de disponibilização dos dados pela fonte;
    - Os dados deverão ser baixados logo em seguida da sua disponibilização. Como atualmente não temos outra fonte que disponibilize os mesmos dados, não temos como garantir um plano de contingência para este risco específico, mas pela atual lei Brasileira de disponibilização de dados por órgãos públicos, há grandes possibilidades que a fonte não seja descontinuada.
- Não utilização efetiva dos dados pelo público alvo;

- O projeto deverá ser divulgado para demonstrar sua importância e ganhos ao público alvo, e estar em fácil acesso.

- Falta de entendimento pelo público alvo de como usar efetivamente os dados para obtenção das informações desejadas;
  - Será disponibilizada documentação e vídeo-aulas para explicar a forma de utilização dos dados para a obtenção de informações pelos públicos alvo.

5. PLANO DE DESENVOLVIMENTO

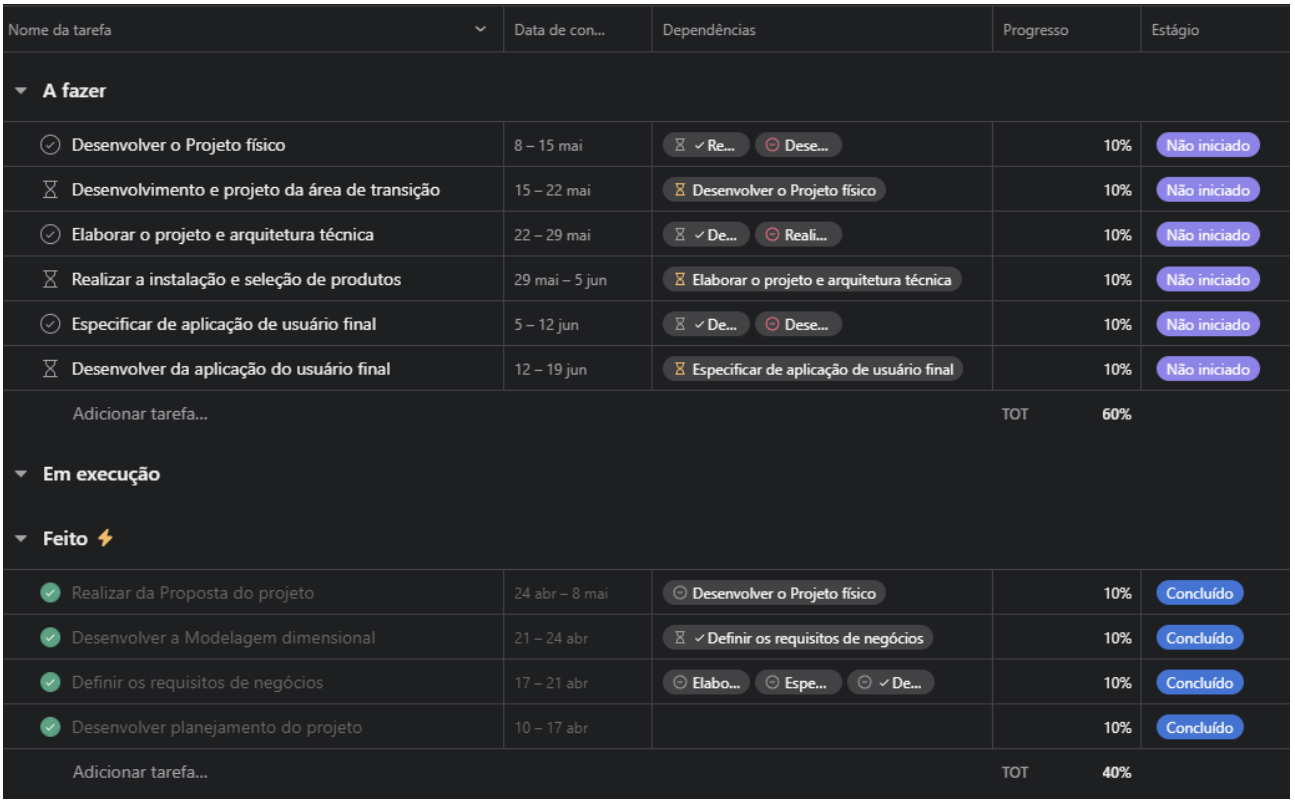
Nome do projeto: Análise de admissões e dispensas no CAGED.

Sigla do Projeto: ADC

Cronograma:

- Estimativa de esforço: 70 dias
- Data de início: 10/04/2023
- Estimativa de data de término: 19/06/2023
- Situação atual: 40% concluído
- Ferramenta para acompanhamento do cronograma: Será utilizado o aplicativo Asana em sua versão free. Abaixo segue imagem do cronograma atual:

IMAGEM 1: Cronograma do projeto



Nome da tarefa	Data de con...	Dependências	Progresso	Estágio
A fazer				
Desenvolver o Projeto físico	8 – 15 mai	✓ Re... Dese...	10%	Não iniciado
Desenvolvimento e projeto da área de transição	15 – 22 mai	Desenvolver o Projeto físico	10%	Não iniciado
Elaborar o projeto e arquitetura técnica	22 – 29 mai	✓ De... Reali...	10%	Não iniciado
Realizar a instalação e seleção de produtos	29 mai – 5 jun	Elaborar o projeto e arquitetura técnica	10%	Não iniciado
Especificar de aplicação de usuário final	5 – 12 jun	✓ De... Dese...	10%	Não iniciado
Desenvolver da aplicação do usuário final	12 – 19 jun	Especificar de aplicação de usuário final	10%	Não iniciado
Adicionar tarefa...			TOT 60%	
Em execução				
Feito ⚡				
Realizar da Proposta do projeto	24 abr – 8 mai	Desenvolver o Projeto físico	10%	Concluído
Desenvolver a Modelagem dimensional	21 – 24 abr	✓ Definir os requisitos de negócios	10%	Concluído
Definir os requisitos de negócios	17 – 21 abr	Elabo... Espe... ✓ De...	10%	Concluído
Desenvolver planejamento do projeto	10 – 17 abr		10%	Concluído
Adicionar tarefa...			TOT 40%	

Equipe:

- Diretor: Marco César da Silva
- Gerente: Cinthia Carolina Shiratori
- Arquiteto de DW: Daniel Roberto de Aguiar
- Analista de Negócios: Ramon Seugling
- Suporte/Treinamento: Rafael Begnini de Castilhos
- Arquiteto de Segurança: Levi Andrew de Aquino Paz

Custos:

- Por se tratar de um projeto acadêmico, não haverá custos relacionados a pagamento de mão de obra para a execução do projeto, todos os softwares utilizados serão open source. Tendo apenas custos relacionados ao despendidos pelos membros da equipe.

6. MODELAGEM DIMENSIONAL

Diante dos objetivos propostos para a criação de um data warehouse e da coleção de dados apresentada anteriormente, foi desenvolvida uma modelagem de dados do tipo estrela, sendo composta por uma tabela fato chamada de FATO_CAGED, além de cinco tabelas de dimensões, que são: DIM_EMPREGADOR, DIM_PERIODO, DIM_TRABALHADOR, DIM_LOCALIDADE e DIM_MOVIMENTACAO.

A FATO_CAGED receberá as chaves que garantem os relacionamento com as demais tabelas de dimensões, além das medidas que serão usadas nos cálculos. A DIM_EMPREGADOR receberá os dados referentes ao contratante, mas com dados anonimizados, isto é, não permitindo a identificação específica do contratante, apenas os dados referentes ao tamanho do estabelecimento, área de atuação, e tipo de empregador e estabelecimento. A DIM_PERIODO traz as informações de ano e mês. A DIM_TRABALHADOR traz as informações socioeconômicas do trabalhador, não possibilitando a identificação específica do mesmo. A DIM_LOCALIDADE traz as informações de região do país, uf e município onde ocorreu a contratação, e não o endereço do trabalhador ou da empresa. E a DIM_MOVIMENTACAO traz as informações da movimentação em si, se foi uma contratação ou desligamento, a categoria da movimentação, e a ocupação do trabalhador na movimentação. Abaixo segue o dicionário de dados com as descrições dos atributos.

TABELA 2: Dicionário de dados

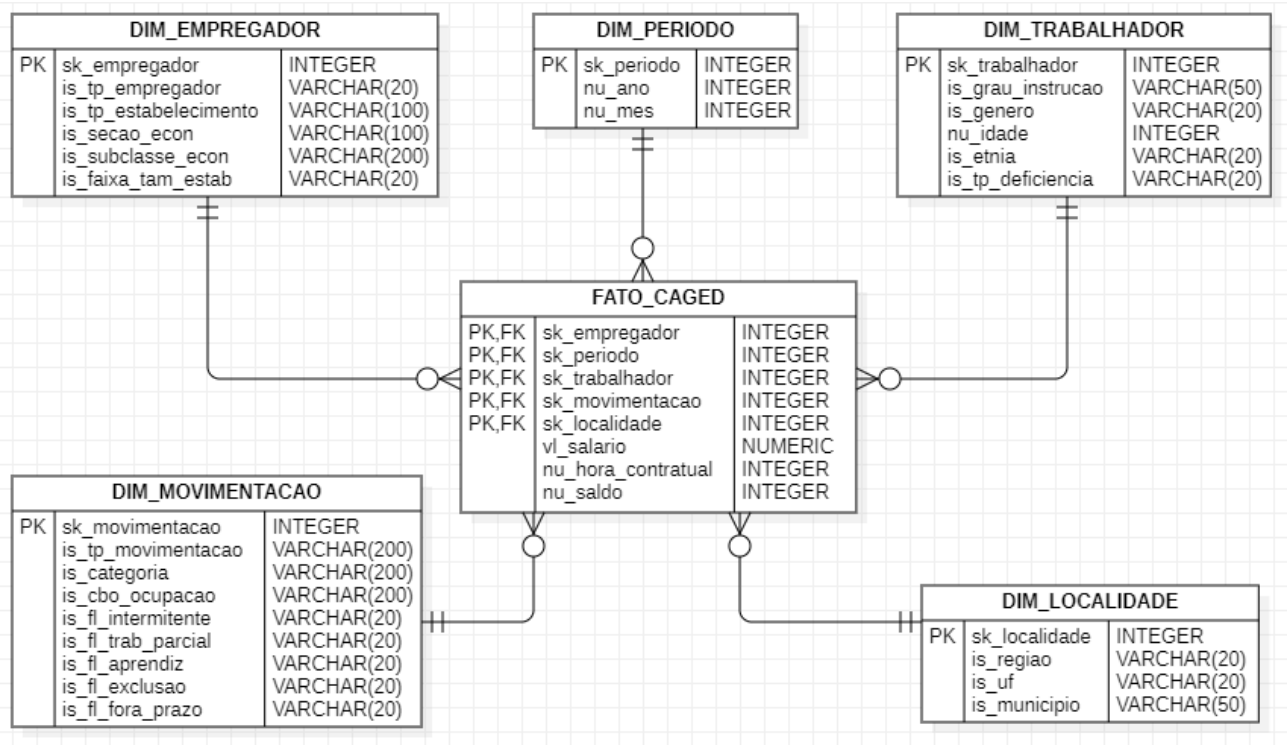
Tabela	Atributo	Descrição	dado fonte
FATO_CAGED	sk_empregador: int	Chave estrangeira para a DIM_EMPREGADOR	Gerado na ETL
FATO_CAGED	sk_periodo: int	Chave estrangeira para a	Gerado na ETL

		DIM_PERIODO	
FATO_CAGED	sk_trabalhador: int	Chave estrangeira para a DIM_TRABALHADOR	Gerado na ETL
FATO_CAGED	sk_movimentacao: int	Chave estrangeira para a DIM_MOVIMENTACAO	Gerado na ETL
FATO_CAGED	sk_localidade: int	Chave estrangeira para a DIM_LOCALIDADE	Gerado na ETL
FATO_CAGED	vl_salario: nun	Salário mensal declarado	salário
FATO_CAGED	nu_hora_contratual: int	Horas contratuais semanais	horascontratuais
FATO_CAGED	nu_saldo: int	Valor da movimentação em termos de saldo	saldomovimentação
DIM_PERIODO	sk_periodo: int	Chave primária da DIM_PERIODO	Gerado na ETL
DIM_PERIODO	nu_ano: int	Ano da movimentação	competênciamov
DIM_PERIODO	nu_mes: int	Mês da movimentação	competênciamov
DIM_LOCALIDADE	sk_localidade: int	Chave primária da DIM_LOCALIDADE	Gerado na ETL
DIM_LOCALIDADE	is_região: var(20)	Região geográfica da contratação	região
DIM_LOCALIDADE	is_uf: var(20)	Unidade da federação da contratação	uf
DIM_LOCALIDADE	is_município: var(50)	Município da contratação	município
DIM_TRABALHADOR	sk_trabalhador: int	Chave primária da DIM_TRABALHADOR	Gerado na ETL
DIM_TRABALHADOR	is_grau_instrucao: var(50)	Grau de instrução do trabalhador	graudeinstrução
DIM_TRABALHADOR	is_genero: var(20)	Gênero do trabalhador	sexo
DIM_TRABALHADOR	nu_idade: int	Idade do trabalhador	idade
DIM_TRABALHADOR	is_etnia: var(20)	Etnia do trabalhador	raçacor
DIM_TRABALHADOR	is_tp_deficiencia: var(20)	Tipo de deficiência do trabalhador	tipodedeficiência
DIM_EMPREGADOR	sk_empregador: int	Chave primária da DIM_EMPREGADOR	Gerado na ETL
DIM_EMPREGADOR	is_tp_empregador: var(20)	Tipo de empregador	tipoempregador
DIM_EMPREGADOR	is_tp_estabelecimento: var(100)	Tipo de estabelecimento	tipoestabelecimento
DIM_EMPREGADOR	is_secao_econ: var(100)	Seção do CNAE 2.0	seção
DIM_EMPREGADOR	is_subclasse_econ: var(200)	Subclasse do CNAE 2.0	subclasse
DIM_EMPREGADOR	is_faixa_tam_estab: var(20)	Faixa de tamanho do estabelecimento	tamestabjan
DIM_MOVIMENTACAO	sk_movimentacao: int	Chave primária da DIM_MOVIMENTACAO	Gerado na ETL
DIM_MOVIMENTACAO	is_tp_movimentacao: var(200)	Tipo de movimentação	tipomovimentação
DIM_MOVIMENTACAO	is_categoria: var(200)	Categoria de trabalhador na	categoria

		movimentação	
DIM_MOVIMENTACAO	is_cbo_ocupacao: var(200)	Ocupação do trabalhador de acordo com a CBO 2002	cbo2002ocupação
DIM_MOVIMENTACAO	is_fl_intermitente: var(20)	Indicador de trabalhador intermitente	indtrabintermitente
DIM_MOVIMENTACAO	is_fl_trab_parcial: var(20)	Indicador de trabalhador parcial	indtrabparcial
DIM_MOVIMENTACAO	is_fl_aprendiz: var(20)	Indicador de trabalhador aprendiz	indicadoraprendiz
DIM_MOVIMENTACAO	is_fl_exclusao: var(20)	Indicador de Exclusão	indicadordeexclusão
DIM_MOVIMENTACAO	is_fl_fora_prazo: var(20)	Indicador de informação declarada fora do prazo	indicadordeforadoprazo

Para visualizar de forma gráfica, abaixo segue esquema dimensional com o relacionamento entre as tabelas de dimensões com a tabela fato.

IMAGEM 2: Modelagem Dimensional



Para o data warehouse proposto, serão necessárias hierarquias de dados em duas tabelas de dimensões, uma na tabela de DIM_PERIODO, onde os dados de ano (nu_ano) e mês (nu_mes) poderão ser agregados e analisados de forma hierárquica, como é representado imagem 3, assim como na tabela de DIM_LOCALIDADE, onde haverá hierarquia entre os dados de região (is_regiao), UF (is_uf) e município (is_municipio), como é representado na imagem 4.

IMAGEM 3: Hierarquia de período

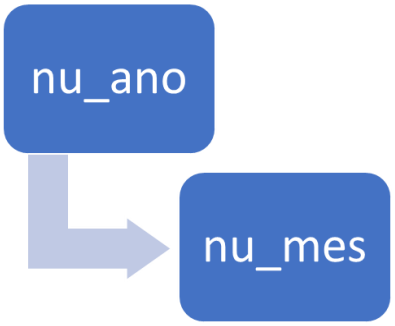
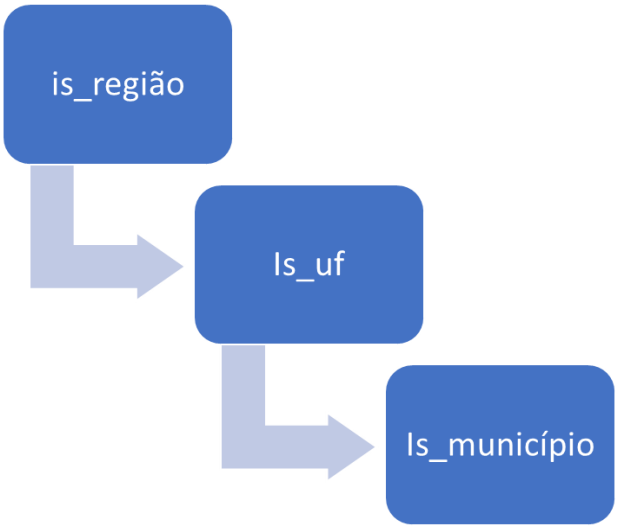


IMAGEM 4: Hierarquia de localidade



7. CONCLUSÕES PARCIAIS

Analisar os dados do CAGED pode ser de grande importância para a sociedade em geral, possibilitando informações para o embasamento de decisões de políticas públicas ou facilitando o trabalho de profissionais de RH.

E para isso, construir um data warehouse de forma planejada e cuidadosa, entendendo os requisitos de negócios, quais perguntas queremos responder, pensando na arquitetura para a obtenção de sucesso na execução e em como documentar para que o usuário final posso entender a sua utilização é de extrema importância para o comprimento dos objetivos propostos.