Modelagem dimensional do Cadastro Geral de Empregados e Desempregados

André Augusto da Silva Pereira[1], **Marcos César da Rocha Seruffo**[1], **Marcelino Silva da Silva**[2]

[1]Instituto de Tecnologia – Universidade Federal do Pará (UFPA)
Belém – PA – Brazil

[2]Instituto de Engenharia e Geociências – Universidade Federal do Oeste do Pará (UFOPA)
Santarém – PA – Brazil

{andresp,seruffo}@ufpa.br, marcelino.ss@ufopa.edu.br

Abstract. *The Brazilian Ministry of Labor and Employment (MTE) has a permanent registry of data regarding admissions and dismissals of employees under the Brazilian Consolidation of Labor Laws (CLT) regime, called CAGED (General Register of Employed and Unemployed). Its information is used by the Unemployment Insurance Program, among other social programs, and is published for society's reuse. In this paper, CAGED's microdata from January 2020 to April 2023 is enriched and consolidated on a dataset, using Kimball's dimensional modeling technique, and is made available on a binary, open, and columns-oriented data file type, in order to allow simple and efficient data storage and retrieval. This approach was adopted because it is a massive dataset, with more than 123 million records, which presents a technical challenge for manipulation using traditional formats tabular text (CSV).*

Resumo. *O Ministério do Trabalho e Emprego (MTE) conta com um registro permanente de dados referentes a admissões e dispensas de empregados sob o regime da Consolidação das Leis do Trabalho (CLT), denominado CAGED (Cadastro Geral de Empregados e Desempregados). Suas informações são utilizadas no Programa de Seguro-Desemprego e em outros programas sociais, e são divulgados para reutilização pela sociedade. Neste trabalho, os microdados do CAGED de janeiro de 2020 a maio de 2023 são enriquecidos e consolidados em um conjunto de dados, utilizando a técnica de modelagem dimensional, e disponibilizados em arquivos de dados em formato binário, aberto e orientado a colunas, com objetivo de permitir o armazenamento e recuperação de dados de maneira simples e eficiente. Esta abordagem foi adotada por tratar-se de um volume de dados massivo, com mais de 123 milhões de registros, o que representa um obstáculo técnico para manipulação utilizando formatos tradicionais de texto tabular (CSV).*

1. Introdução

O Ministério do Trabalho e Emprego (MTE) conta com um registro permanente de dados referentes a admissões e dispensas de empregados sob o regime da Consolidação das Leis do Trabalho (CLT), denominado CAGED (Cadastro Geral de Empregados e Desempregados), cuja coleta de dados é realizada através de declarações mensais prestadas em

caráter compulsório por todos os empregadores brasileiros. Suas informações são utilizadas no Programa de Seguro-Desemprego e outros programas sociais, e são divulgados para reutilização pela sociedade [MTE 2023].

Neste trabalho, os microdados do CAGED, a partir de Janeiro de 2020, são consolidados em um conjunto de dados utilizando modelagem dimensional e disponibilizados no formato *parquet*, um formato de arquivos de dados aberto, binário e orientado a colunas, projetado com objetivo de prover armazenamento e recuperação de dados de maneira eficiente. Esta abordagem foi adotada por tratar-se de um volume de dados massivo, e que conta no momento da escrita deste trabalho com mais de 123 milhões de registros. Alternativamente, o conjunto também é publicado em formato texto separado por virgulas (CSV), apesar de seu uso ser desencorajado. Os dados são ainda enriquecidos através da utilização de outros conjuntos de dados públicos brasileiros, utilizados na criação de dimensões utilizadas na modelagem.

Na Seção 2 será apresentada a metodologia utilizada na elaboração deste trabalho e as ferramentas e técnicas que a compõem. A seguir, na Seção 3, são apresentadas as fontes de dados utilizadas. Na Seção 4 será apresentada a modelagem realizada neste trabalho, aplicada ao conjunto de dados do CAGED. Na Seção 5 são detalhadas as estratégias de divulgação do conjunto de dados produto deste trabalho, além da apresentação de alguns exemplos de utilização deste. Por fim, a Seção 6 apresenta uma revisão do conteúdo abordado e as perspectivas de trabalhos futuros.

2. Metodologia para Aquisição e Tratamento dos Dados

O conjunto de dados tratado e publicado neste trabalho é parte de uma iniciativa que visa simplificar e estimular a reutilização de dados públicos brasileiros por membros da sociedade que atuem nas mais diversas áreas do conhecimento e com diferentes níveis de formação técnica. A abordagem utilizada na busca por este objetivo é a de modelar um Armazém de dados, conhecido como *Data Warehouse* (DW), conforme detalhado na Seção 2.1.

A motivação para definição de tal objetivo decorre da constatação de que grande esforço é feito pela iniciativa pública no sentido de publicar dados abertos, contudo ainda nota-se um relativo baixo volume de iniciativas de utilização destes por parte da sociedade. Possíveis fatores contribuintes para este cenário são a complexidade envolvida na obtenção e manipulação de grandes bases de dados públicas, principalmente por profissionais de áreas não relacionadas à tecnologia, a disponibilização de dados em contextos específicos e isolados e a consequente complexidade na correlação de diferentes bases de dados públicas.

Neste trabalho é implementado, portanto, um processo de extração, tratamento e publicação da base de dados do CAGED, disponibilizada pelo MTE, seguindo o modelo dimensional, tal como será detalhado na Seção 3.1. Esse processo é conhecido na literatura como ETL (*Extract Transform and Load*) [Kimball and Caserta 2011].

O processo consiste na extração automatizada dos dados disponíveis no FTP do MTE[1], através da ferramenta `extractds`[2], seguida por um processo de transformação

[1] ftp://ftp.mtps.gov.br/pdet/microdados/NOVO%20CAGED/

[2] Desenvolvida pelos autores, disponível em https://github.com/andrespp/dw-br

com limpeza e enriquecimento de dados, e finalizada com o armazenamento e publicação[3] dos dados resultantes em arquivos do tipo *Apache parquet*[4]. Esses arquivos são o produto final deste trabalho e servirão de insumos para análises por terceiros, podendo ser utilizados de diversas formas, tal como será brevemente discutido na Seção 5.

O *Apache parquet* é um formato de arquivos de código aberto, com armazenamento orientado a colunas, desenvolvido para o ecossistema *Hadoop*. O formato lança mão da compressão de dados por colunas, com possibilidade de utilização de técnicas de codificação e compressão específicas para os tipos de dados armazenados nestas, sendo muito eficiente nos quesitos de utilização de espaço em disco e recuperação de informações [Vohra and Vohra 2016]. Sistemas de gerenciamento e armazenamento de dados orientados a coluna utilizam esta abordagem como forma de eliminar a necessidade de recuperação de linhas completas de registros quando deseja-se apenas alguns atributos destes, além de permitir a utilização de técnicas de codificação e compressão especificas para o tipo de dado armazenado em cada coluna, conforme citado anteriormente. A contrapartida é uma penalidade no desempenho em operações de escrita. Contudo, este formato se mostra ideal para operações de análise, onde operações de consulta em um número limitado de atributos em um grande volume de registros são frequentes, [Abadi et al. 2009], tornando-se um padrão da indústria em soluções de *Big Data* e *Analytics* [Abeykoon and Fox 2023].

Alternativamente, este trabalho também disponibiliza as tabelas modeladas no formato CSV, apesar do seu uso ser desencorajado em detrimento ao formato *parquet*, por conta da necessidade da utilização de recursos computacionais consideravelmente superiores para recuperação dos dados a serem analisados.

2.1. Modelagem Dimensional

Proposta por [Kimball 1997], a Modelagem Dimensional é uma técnica para modelagem de dados comumente utilizada em projetos de DW, e que possui uma ampla adoção. Dentre os seus principais benefícios estão o fato de possuir um bom desempenho em operações de consulta e, principalmente, o fato de representar dados de forma intuitiva, o que torna a consulta por parte dos usuários do DW mais simples.

A simplicidade na representação das informações é um fator chave, pois permite aos usuários uma melhor possibilidade de entendimento e análise dos dados, o que é em si o verdadeiro objetivo de um DW. Na modelagem dimensional, a informação é organizada em topologia de estrela, e possui dois elementos principais: fatos e dimensões.

As *tabelas fato* em modelos dimensionais registram medidas relacionadas a um processo organizacional. Nestas tabelas devem ficar armazenadas as informações na menor granularidade disponível e de maneira centralizada, evitando-se ao máximo a reprodução de dados de um determinado processo em mais de uma tabela. Uma linha em uma tabela fato corresponde a uma medição, e todas as medições em uma tabela fato devem possuir a mesma granularidade. Os fatos mais úteis são os numéricos e somáveis. A possibilidade de soma é crucial porque aplicações clientes do DW raramente buscam

[3]https://zenodo.org/record/8075924

[4]Neste trabalho ambos os termos *parquet* e *Apache parquet* são utilizados para referenciar o formato de arquivos *Apache parquet*.

observar um registro individualmente, ou seja, apenas uma linha, mas sim dezenas, milhares ou milhões de registros da tabela fato, visando obter uma visão geral do processo que está sendo analisado.

As *tabelas dimensão*, por sua vez, contém descrições textuais associadas aos eventos registrados nas tabela fato. Elas costumam ter muitos atributos, logo tendem a ter muitas colunas. Esses atributos são utilizados para restringir o escopo e/ou agrupar resultados de consultas realizadas nas tabela fato.

As tabelas fato e dimensões são então associadas através de suas chaves primárias, utilizando uma topologia estrela, conforme detalhado na Seção 4.

3. Descrição dos Conjuntos de Dados de Origem

Conforme apresentado na Seção 2, uma tabela fato no contexto da modelagem dimensional representa um processo de negócio, armazena indicadores numéricos deste e cada registro corresponde a uma medição de sua ocorrência. As tabelas de dimensão, por sua vez, contêm descrições textuais de características associadas aos eventos registrados nas tabela fato.

Neste trabalho, a tabela fato é derivada do conjunto de dados do Cadastro Geral de Empregados e Desempregados (CAGED), apresentado na Seção 3.1, e as suas respectivas tabelas dimensão são identificadas a partir da análise dos atributos disponíveis nesse. Em tal análise, foram identificadas oportunidades de enriquecimento dos dados de duas formas: a partir de fontes próprias e a partir de fontes externas. A primeira forma consiste da definição de uma dimensão de data, que é apresentada na Seção 3.2. Já a última inclui a utilização de outros conjuntos de dados públicos e é detalhada na Seção 3.3.

Os dados do CAGED são detalhados na Seção 3.1 e o modelo final desenvolvido neste trabalho é apresentado na Seção 4.

3.1. Novo CAGED

O Cadastro Geral de Empregados e Desempregados (CAGED) foi criado pela Lei 4.923 de 23/12/1965, quando instituiu-se a obrigatoriedade da declaração de informações sobre admissões, desligamentos e transferências. Em Janeiro de 2020 o uso do CAGED foi substituído pelo Sistema de Escrituração Digital das Obrigações Fiscais, Previdenciárias e Trabalhistas (eSocial) [Almeida et al. 2020]. Tal transição foi feita em etapas, onde diferentes grupos de empregadores passaram a ser obrigados a declarar pelo novo sistema de forma gradual, seguindo um calendário específico.

Como forma de dirimir o impacto das diferentes formas de captação de dados relativos ao emprego formal, o MTE realizou um trabalho de consolidação dos dados disponíveis nos sistemas eSocial, CAGED e Empregador WEB em um único conjunto de dados chamado de *Novo CAGED*[5] [MTE 2020].

O conjunto de dados do CAGED é distribuído em arquivos por competência de declaração. Cada mês do ano é considerado uma competência de declaração e em todo mês-calendário devem ser declarados os dados de movimentações referentes ao mês anterior. Por exemplo, a competência `202305` contém as informações de empregados e

[5]Neste trabalho, referências feitas a "conjunto de dados do CAGED", "dados do CAGED", ou simplesmente "CAGED", referem-se a esta base consolidada, também conhecida como "Novo CAGED".

desempregados ocorridas no mês de abril de 2023 e deve ser informada pelos empregadores no mês de junho de 2023. Este trabalho inclui as competências de janeiro de 2020 até abril de 2023, por serem as competências disponíveis no momento da elaboração deste documento. Tais dados estão divididos em 40 arquivos, contendo total de 123.008.571 registros, que consomem de 5,5 Gb de espaço de armazenamento em disco.

Dentre os atributos presentes nos dados destacam-se 4 tipos:

- **Medições**: campos numéricos inerentes ao processo de negócio representado. Nesta categoria encontram-se os campos `saldomovimentacao`, `idade` e `horascontratuais`, que representam respectivamente o saldo de vagas criadas ou destruídas naquele registro, bem como a idade do trabalhador e o número de horas semanais pactuadas naquele contrato de trabalho. Nesta categoria enquadram-se também os atributos indicativos de salário;
- **Dimensões externas**: campos que indicam categorias que podem ser enriquecidas através de conjuntos de dados externos ao que está sendo utilizado na tabela fato. Aqui foram identificados atributos de localização geográfica, classificação das atividades empresariais dos empregadores, classificação da ocupação dos empregados e tipo de vínculo empregatício. Essas categorias e as bases de dados externas utilizadas na elaboração das dimensões são detalhadas na Seções 3.3;
- **Dimensões internas**: atributos que (1) indicam categorias inerentes ao processo da tabela fato, ou que (2) são de natureza genérica, não podendo ser atribuídos diretamente a um conjunto de dados externo administrado por terceiros. No primeiro caso pode-se citar os atributos `tipoempregador`, `tipoestabelecimento` e `tipomovimentacao`. Já relacionados ao último tem-se atributos como `sexo`, `racacor` e `tipodeficiencia`.
- **Dimensões temporais**: atributos de classificação temporal dos fatos medidos, como data de ocorrência, registro, exclusão, dentre outros. Os campos de dimensão temporal dos dados do CAGED são detalhados na Seção 3.2.

Neste trabalho, os atributos de medição são representados na tabela `fato_caged`. As dimensões externas possuem suas respectivas tabelas, e são geradas a partir de conjuntos de dados externos ao CAGED, e ligados à tabela fato por uma chave estrangeira, tal como preconizado na metodologia de modelagem dimensional. Com relação às dimensões internas, optou-se por manter os dados de código e descrição na tabela fato. Esta abordagem, apesar de incorrer em uma penalidade no espaço em disco para armazenamento, foi utilizada como forma de evitar um excesso de dimensões ligadas à tabela fato, principalmente aquelas com dados estritamente inerentes ao fato, e com baixa probabilidade de reutilização em outras tabelas fato a serem desenvolvidas. Essa decisão de projeto visa simplificar consultas, sem que haja impacto no custo computacional destas, uma vez que é utilizado formato de arquivo orientado a colunas para armazenamento das tabelas. No que tange a trabalhos futuros, as dimensões internas de natureza genérica são candidatas a tornarem-se tabelas dimensão, uma vez que, diferentemente das anteriores, possuem maior probabilidade de serem comuns a outras tabelas fato. Esta e outras possibilidades de melhorias e trabalhos futuros são discutidos na Seção 6.

3.2. Dimensão Data

O conjunto de dados do CAGED possui três atributos relacionados a data: `competenciamov`, `competenciadec` e `competenciaexc`. Todos eles campos

de seis dígitos, na forma "AAAAMM", e que representam ano e mês para os seus respectivos registros.

Acontece que possíveis análises no registro de empregados e desempregados podem facilmente ser realizadas em diferentes subconjuntos de períodos. Por exemplo, um analista pode desejar analisar tendências em anos, trimestres, semestres, dentre outros. Tal análise, ainda que possível no conjunto de dados original, adiciona complexidade à implementação, além de custo computacional em tempo de análise, tornando o processo mais lento e propenso a falhas.

Desta forma, foi definida uma dimensão de data onde, a partir de uma chave primária, é possível obter-se diversos outros atributos relacionados ao período. Neste trabalho optou-se por utilizar uma granularidade de dia, como forma de permitir uma reutilização desta dimensão em outras tabelas fato, futuramente.

3.3. Fontes Externas

O conjunto de dados do CAGED possui três atributos que identificam a localização geográfica às quais seus respectivos registros referem-se, que são: `regiao`, `uf` e `municipio`, formadas por 1, 2 e 6 dígitos respectivamente, e seus valores representam os códigos IBGE da região, unidade da federação e município. Portanto, a utilização do conjunto de dados de municípios brasileiros[6] é uma escolha natural. A partir desta decisão de modelagem, obtém-se não só uma economia no armazenamento de dados, uma vez que três campos de texto no tamanho combinado de pelo menos 9 bytes são combinados em apenas um campo inteiro, mas também a possibilidade de enriquecimentos adicionais nos dados disponibilizados, a partir da inclusão de mais atributos na tabela a ser definida. De fato, um segundo conjunto de dados públicos é utilizado na composição desta dimensão, que é o conjunto de municípios SIAFI[7], que inclui o código de municípios brasileiros definidos no Sistema Integrado de Administração Financeira do Governo Federal, do Tesouro Nacional e que é utilizado em diversos outros conjuntos de dados públicos. Esta decisão de projeto visa ampliar ainda mais a capacidade de reutilização desta dimensão em outras tabelas fato, bem como ampliar e simplificar a capacidade de agrupamento e correlações realizadas por clientes do conjunto de dados. Além destes, deseja-se ainda incluir outros dados censitários nesta dimensão, tal como será detalhado na Seção 6, onde discute-se trabalhos futuros.

Avançando com a análise dos atributos do CAGED, temos os campos `secao` e `subclasse` de 1 e 9 bytes respectivamente, que representam códigos de Seção e Subclasse da Classificação Nacional de Atividade Econômica[8] (CNAE 2.0). Apesar de o conjunto de dados do CAGED incluir apenas estes dois campos do CNAE, esta classificação é composta de 4 campos: seção, divisão, grupo e classe. Desta forma, a utilização desta fonte externa em conjunto com a Modelagem dimensional permite o enriquecimento das possibilidades de agrupamento, além de incluir as respectivas descrições dos códigos de classificação, as quais não estão originalmente presentes dos dados do CAGED.

Outra referência a conjuntos de dados externos feita nos dados do CAGED é o

[6]https://www.ibge.gov.br/explica/codigos-dos-municipios.php

[7]https://www.tesourotransparente.gov.br/ckan/dataset/lista-de-municipios-do-siafi

[8]https://cnae.ibge.gov.br/classificacoes/por-tema/atividades-economicas/classificacao-nacional-de-atividades-economicas

atributo `cbo2002ocupacao`, que representa o código da ocupação para o emprego medido no registro, segundo a Classificação Brasileira de Ocupações[9] (CBO 2002) [Nozoe et al. 2003]. As ocupações do CBO são organizadas de maneira hierárquica nas seguintes categorias: grande grupo, subgrupo principal, subgrupo, família e ocupação. Nos dados do CAGED apenas o código da ocupação é definido, e suas respectivas descrições devem ser consultadas manualmente pelo usuário em uma tabela auxiliar fornecida juntamente com conjunto de dados. Com a definição de uma dimensão a partir destes dados externos, o pesquisador poderá ter acesso a todas as categorias da classificação, bem como suas descrições, sem a necessidade de consultas manuais a fontes externas e utilizando apenas um campo inteiro na tabela fato.

A última fonte de dados externa é de categoria de trabalhador segundo a classificação do eSocial[10], que segue a mesma abordagem das fontes citadas anteriormente, definindo uma tabela de dimensão externa com as descrições textuais dos atributos definidos no CAGED, sendo referenciada neste a partir de uma chave estrangeira representada por um inteiro, em detrimento dos campo de três caracteres dos dados originais.

4. Modelagem

A Figura 1 ilustra a modelagem obtida a partir da aplicação da metodologia apresentada na Seção 2 sobre os dados apresentados na Seção 3.

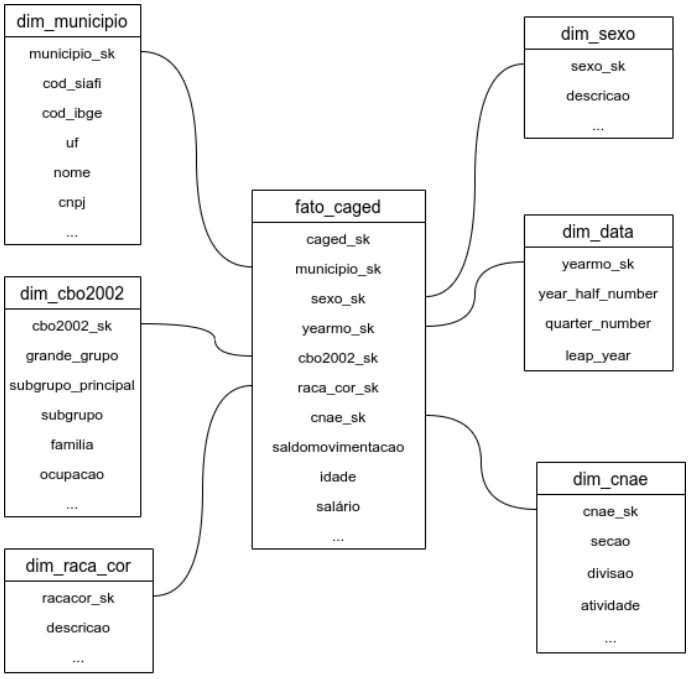

Figura 1. Modelagem Dimensional do Conjunto de Dados do CAGED.

Por uma questão de clareza, nem todos os campos das tabelas estão representados, mas apenas os seus principais, bem como as chaves que relacionam as tabelas. Os dados finais disponibilizados neste trabalho contêm uma tabela auxiliar descrevendo todos os atributos das tabelas de dimensões e fato.

[9]https://cnae.ibge.gov.br/classificacoes/por-tema/ocupacao/classificacao-brasileira-de-ocupacoes

[10]https://www.gov.br/esocial/pt-br/documentacao-tecnica/manuais/leiautes-esocial-v-1-1-beta/tabelas.html

5. Disponibilização e Utilização do Dataset Modelado

O conjunto de dados desenvolvido neste trabalho encontra-se disponível publicamente através da plataforma *Zenodo*[11], para fins de reutilização pela comunidade científica. Os *scripts* de ETL implementados são de código aberto e também estão disponíveis ao público interessado[12]. Desta forma, não só é possível a reutilização dos dados já tratados, como a evolução do modelo, através da inclusão de novas tabelas fato, novas tabelas de dimensão, novos atribuídos à dimensões existentes, dentre outras possibilidades de melhoria.

Ademais, este trabalho também disponibiliza uma ferramenta de visualização dos dados tratados, onde além da disponibilização de algumas análises nestes dados, também é possível acessar pré-visualizações das tabelas que compõe o DW, bem como realizar consultas sobre estas, utilizando linguagem SQL, de forma simples e direta. Esta ferramenta chama-se `Dash-br`[13] e também possui código aberto[14], permitindo a qualquer interessado a reutilização, melhoria e contribuição, com análises, painéis, dentre outros.

Uma outra forma direta de utilização do conjunto de dados a partir de uma pilha de tecnologias comumente utilizada em análise de dados é a utilização da biblioteca Dask[15], escrita em *Python*, que reproduz API da biblioteca *Pandas* otimizada para processamento de grandes volumes de dados através de técnicas de execução concorrente e paralela.

6. Considerações finais e Trabalhos Futuros

Neste trabalho foi realizada a modelagem dimensional do conjunto de dados do CAGED, compreendendo o período de janeiro de 2020 a abril de 2023. Neste processo foram identificadas dimensões relacionadas ao fato medido (empregos formais no Brasil), que foram enriquecidas com conjuntos de dados externos.

A abordagem utilizada consiste em um extenso trabalho de obtenção de dados de diversas fontes, tratamento, limpeza e disponibilização em um formato de fácil acesso, e o sucesso desta abordagem pode ser constatado a partir da possibilidade de execução de consultas em um volume massivo de dados em poucos segundos, utilizando a biblioteca *Dask* ou mesmo linguagem SQL, tal como apresentado na Seção 5.

Conforme citado ao longo deste trabalho, a disponibilização destes dados faz parte de um objetivo maior, de construir um armazém de dados públicos brasileiros, contendo tabelas fato de diversos órgãos do governo, devidamente relacionados com as suas dimensões, que deverão ser reutilizadas em tantas tabelas fato quanto possível, e que também prevê outras ferramentas de apresentação dos dados tratados, além dos arquivos consolidados, como por exemplo painéis, APIs, ferramentas de análise gráfica, dentre outras.

Especificamente no contexto dos dados do CAGED, deseja-se enriquecer a tabela `dim_municipio` com outros dados censitários disponibilizados pelo IBGE, tais como população do município, renda per-capita, expectativa de vida da população, faixa

[11]https://zenodo.org/
[12]https://github.com/andrespp/dw-br
[13]Em desenvolvimento pelos autores, disponível em https://dwbr.andrepereira.eng.br
[14]https://github.com/andrespp/dash-br
[15]https://docs.dask.org/

etária, dentre outras. Este é um bom exemplo de como essa abordagem é interessante para extração de novos conhecimentos a partir de dados públicos: uma vez incluídas essas informações em uma pequena tabela do DW, ou seja, dimensão município, um pesquisador poderia facilmente analisar a relação entre geração de empregos e renda per-capita por região em um determinado período, por exemplo. E tal análise seria possível sem a necessidade de alteração na tabela `fato_caged`.

Agradecimentos. Este trabalho foi realizado com apoio financeiro do CNPq.

Referências

Abadi, D. J., Boncz, P. A., and Harizopoulos, S. (2009). Column-oriented database systems. *Proc. VLDB Endow.*, 2(2):1664–1665.

Abeykoon, V. and Fox, G. C. (2023). Trends in high performance data engineering for data analytics.

Almeida, M. E., Dias, T. S., Farias, R. J. d., Albuquerque, A. V. S. M., Torres, S. L. R., and Oliveira, L. F. B. d. (2020). Substituição da captação dos dados do caged pelo esocial: implicações para as estatísticas do emprego formal.

Kimball, R. (1997). A dimensional modeling manifesto. *Dbms*, 10(9):58–70.

Kimball, R. and Caserta, J. (2011). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley.

MTE (2020). Nota técnica: Substituição da captação dos dados do caged pelo esocial.

MTE (2023). Cadastro geral de empregados e desempregados (caged).

Nozoe, N. H., Bianchi, A. M., and Rondet, A. C. A. (2003). A nova classificação brasileira de ocupações: anotações de uma pesquisa empírica. *São Paulo em perspectiva*, 17:234–246.

Vohra, D. and Vohra, D. (2016). Apache parquet. *Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools*, pages 325–335.