Ambiente de *data warehousing* espacial para tomada de decisão sobre dados de crimes

Juliana Bernardes Freitas, João Paulo Clarindo, Cristina D. Aguiar

[1] Instituto de Ciências Matemáticas e Computação (ICMC)
Universidade de São Paulo (USP)
13566-590 – São Carlos – SP – Brazil

{julianabfreitas, jpcsantos}@usp.br, cdac@icmc.usp.br

Abstract. *The occurrence of crimes is a worrying reality, with countless cases registered daily. Analyzing data from police reports can provide valuable information for public safety managers, assisting in making strategic decisions for public safety and the fight against crime. Based on this motivation, we propose the creation of a spatial data warehousing environment to help public safety managers identify patterns and make more effective decisions to fight crime. We present some examples of spatial analytic queries with data from the Secretariat of Public Security of the state of São Paulo. We also show how the query results can enhance decision-making.*

Resumo. *A ocorrência de crimes é uma realidade preocupante, com inúmeros casos registrados diariamente. A análise dos dados provenientes dos boletins de ocorrência pode fornecer informações valiosas para os gestores de segurança pública, auxiliando na tomada de decisão estratégica para a segurança pública e o combate à criminalidade. Dentro deste contexto, este artigo propõe a criação de um ambiente de* data warehousing *espacial para auxiliar gestores de segurança pública a identificar padrões, direcionar recursos e tomar decisões mais eficazes para combater a criminalidade. São apresentados exemplos de consultas analíticas espaciais com dados disponibilizados pela Secretaria de Segurança Pública do estado de São Paulo, e como elas podem ser utilizadas na tomada de decisão.*

1. Introdução

A ocorrência de crimes é uma realidade preocupante em todo o mundo, com inúmeros casos registrados diariamente. No Brasil, por exemplo, segundo o Instituto de Pesquisa Econômica Aplicada (IPEA), houve um aumento de 5% no número de homicídios em 2019 em relação ao ano anterior [Cerqueira et al. 2020]. Com isso, as autoridades governamentais têm trabalhado para reduzir esses números, com iniciativas que incluem a disponibilização de dados abertos relacionados a este tema, nos níveis municipais, estaduais e federais [BRASIL 2022].

A quantidade de dados gerados diariamente pelo poder público é enorme. Desde que esses dados são gerados em grande volume, velocidade e variedade, pode-se utilizar soluções existentes para lidar com *big data* para auxiliar na tomada de decisão por gestores [Al-Sai and Abualigar 2017]. Uma das soluções que podem ser utilizadas é a disponibilização dos dados em um *data warehouse* espacial (ou *spatial data warehouse*,

SDW), que é uma coleção de dados convencionais e espaciais caracterizada por ser histórica, orientada a assunto, integrada e não-volátil. Um SDW provê o suporte para consultas de *Spatial On-Line Analytical Processing* (SOLAP), que são consultas analíticas que utilizam dados espaciais, ou geográficos [Han et al. 1998, Rivest et al. 2001].

A criação e o gerenciamento um ambiente de *data warehousing* espacial (ou *spatial data warehousing*, SDWing) utilizando dados abertos de criminalidade é um desafio. Este desafio refere-se ao fato de que os dados abertos governamentais não possuem padronização de atributos, tamanhos, e formatos [Macedo and Lemos 2021], além de serem caracterizados como *big data*, sendo necessário lidar com diversas tecnologias para extração, transformação, carga e visualização destes dados. Na literatura, existem várias propostas para lidar com a tomada de decisão sobre dados governamentais de segurança pública [Sá et al. 2021, Santos and Oliveira 2018, Kasprzyk and Donnay 2016]. Porém, essas propostas referem-se a aplicações específicas para uma cidade ou estado.

Para auxiliar na tomada de decisão sobre dados de criminalidade, este artigo propõe a criação de um ambiente de SDWing. O ambiente disponibiliza um esquema lógico baseado em um Boletim de Ocorrência (BO), com dados sobre local de ocorrência do crime, tipo do crime, e pessoas envolvidas. Também é apresentada uma *pipeline* de fluxo de dados baseada em uma arquitetura típica de um SDW, com fases de extração, transformação e carga, utilizando tecnologias de código-aberto. Como resultado, gestores de segurança pública podem utilizar esse modelo de ambiente de SDWing para implementar suas soluções específicas, e prover a tomada de decisão sobre os dados disponíveis. Outra contribuição é a especificação de um estudo de caso que usa dados fornecidos pela Secretaria de Segurança Pública do estado de São Paulo (SSP/SP) sobre roubos e furtos. São mostradas a execução e a visualização de consultas SOLAP sobre esses dados.

Este artigo está estruturado da seguinte forma. Na seção 2 são discutidos trabalhos relacionados. Na seção 3 são apresentados o esquema lógico e a *pipeline* para a criação do ambiente de SDWing proposto. Na seção 4 é apresentado um estudo de caso que implementa SDWing. Na seção 5 são descritos considerações finais e trabalhos futuros.

2. Trabalhos Relacionados

Na literatura existem diversos trabalhos que investigam ambientes de *data warehousing* modelados sobre dados de criminalidade. Nesta seção são discutidos trabalhos voltados aos ambientes de *data warehousing* convencionais e espaciais.

Sá et al. (2021) apresentam um *data warehouse* de trajetória para viabilizar o desenvolvimento e testes de abordagem de geração de rotas policiais em centros urbanos. A partir de dados disponibilizados pela SSP/SP, os autores criam tabelas de dimensão relacionadas aos segmentos de vias. Um conjunto destes segmentos forma grafos que podem ser usados para definir rotas para patrulhamento policial. Os autores disponibilizam um conjunto de dados baseado nesta modelagem e apresentam consultas em mapas.

O trabalho de Santos e Oliveira (2018) descreve a proposta de uma abordagem "que visa integrar dados das diferentes organizações associadas à segurança pública para prover um ambiente de apoio à decisão". Para isso, os autores desenvolvem um *data warehouse* baseado em dados disponibilizados pela Secretaria Estadual de Segurança Pública do estado do Rio de Janeiro (SESEG/RJ). No esquema lógico, há uma tabela de fatos relacionada às ocorrências de roubo de veículos e de transeuntes, e outra tabela de fatos

correspondente às metas que a SESEG determina para prevenir e controlar a ocorrência de crimes por regiões. Os autores mostram exemplos de consultas OLAP agrupados por crimes, por ano, e por batalhão.

Kasprzyk e Donnay (2016) introduzem um SDW que utiliza dados espaciais do tipo *raster*, que são representações geográficas baseadas em imagens. No estudo de caso, os autores utilizam dados de criminalidade da cidade de Seattle, Estados Unidos, com dimensões de tempo, tipo de crime, e dimensões espaciais relacionadas à localidade em que o crime ocorreu.

Diferentemente dos trabalhos anteriormente discutidos, este trabalho objetiva auxiliar gestores na tomada de decisão sobre dados de criminalidade a partir da proposta de um ambiente de SDWing sobre dados vetoriais. São apresentados um esquema lógico que pode ser empregado para dados neste contexto, e uma *pipeline* com sugestões de tecnologias, que visa auxiliar o gestor a implementar este ambiente. Também é detalhado um estudo de caso baseado nos dados de criminalidade do estado de São Paulo.

3. Ambiente de *data warehousing* espacial

Para possibilitar a criação de um ambiente de SDWing, um gestor deve: (i) realizar a modelagem multidimensional conforme os dados; e (ii) extrair, transformar e carregar os dados em um SDW com base no esquema lógico da modelagem multidimensional. Nas seções 3.1 e 3.2 são detalhados os passos que possibilitam a modelagem e a construção deste ambiente de SDWing no contexto de criminalidade.

3.1. Modelagem multidimensional

SDWs e sistemas SOLAP são baseados na modelagem multidimensional para visualizar dados segundo várias perspectivas [Rivest et al. 2001]. No modelo relacional, essa modelagem usa uma representação lógica com esquemas dispostos em tabelas de fatos e de dimensão. Nestes esquemas, conhecidos como esquemas-estrela, a tabela de fatos contém medidas numéricas e espaciais, além de referências às tabelas de dimensão, as quais podem conter atributos convencionais e espaciais [Han et al. 1998].

A Figura 1 ilustra um esquema lógico que pode ser utilizado no contexto de criminalidade. Os atributos espaciais no esquema-estrela ilustrado são representados conforme a especificação de [Vaisman and Zimányi 2014]. O fato deste esquema é um BO, com uma medida espacial indicando a posição espacial da ocorrência. São associadas dez tabelas de dimensão a esse fato, que foram dispostas a partir do levantamento dos campos mais comuns relacionados a BOs registrados em diversos estados brasileiros: (i) Data e Tempo, que indicam o dia e a hora em que o crime ocorreu; (ii) Crime, que define o tipo de crime e o código penal correspondente, além do ambiente em que ocorreu; (iii) Pessoa, que contém uma descrição da pessoa envolvida no crime; (iv) Objeto, que descreve o objeto mencionado no BO, caso o crime seja relacionado a roubo ou furto; (v) Delegacia, referente ao departamento e seccional de delegacias, e (vi) Municipio, Rua e Bairro, que armazenam dados espaciais relacionados a essas divisões territoriais. A tabela de fatos contém duas chaves-estrangeiras para a tabela de dimensão Delegacia: `delegaciaElaboração`, que indica em qual delegacia o BO foi elaborado, e `delegaciaCircunscricao`, que indica qual delegacia é responsável pela região em que o crime ocorreu.

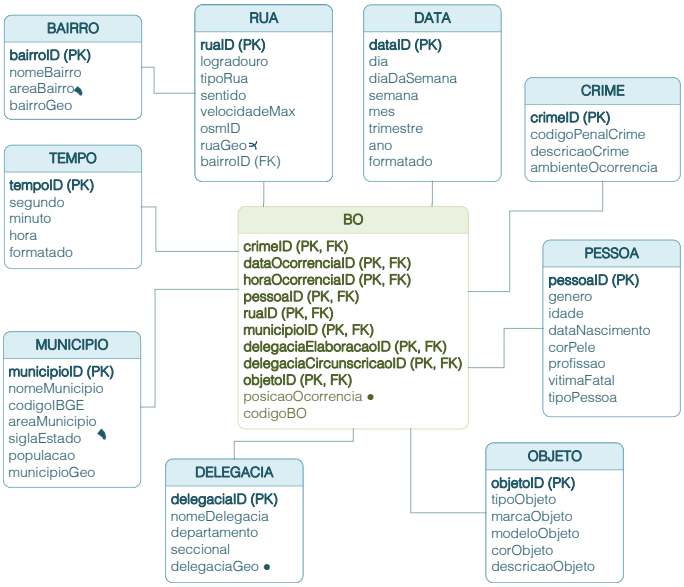

Figura 1. Esquema lógico no contexto de criminalidade.

3.2. Criação de um *data warehouse* espacial

O fluxo de uma arquitetura típica de um SDW consiste nos seguintes passos; (i) extrair os dados de fontes externas; (ii) transformar estes dados conforme a representação lógica do modelo multidimensional; (iii) carregar os dados em um *data warehouse*, e (iv) utilizar um servidor SOLAP para possibilitar a visualização e análise dos dados contidos no SDW [Chaudhuri and Dayal 1997, Vaisman and Zimányi 2014].

Com base nesta arquitetura, é possível modelar *pipelines* contendo as tecnologias necessárias para viabilizar os passos da arquitetura. A Figura 2 ilustra um exemplo de *pipeline* que pode ser utilizado para os dados de criminalidade com tecnologias gratuitas e de código aberto. As fontes de dados são variadas, com dados de divisões administrativas disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE)[1], dados sobre ruas e pontos de interesse disponibilizados pelo OpenStreetMap[2], e dados de BOs disponibilizados pelas secretarias estaduais de segurança pública. A Open Knowledge Brasil oferece um catálogo com links para estes dados nos níveis municipais, estaduais e federais[3]. A extração é feita utilizando *Application Programming Interfaces* (APIs), com as tarefas sendo gerenciadas pelo Apache Airflow, ou usando técnicas de *web scrapping* com a biblioteca da linguagem Python pyAutoGUI, caso não haja disponibilidade de APIs. A transformação é feita utilizando bibliotecas da linguagem Python apropriadas para a manipulação e análise de dados, como Pandas e pySpark. Os dados são carregados no SDW, que inclui o PostgreSQL com a extensão espacial PostGIS. Por fim, a visualização dos dados convencionais e espaciais é feita utilizando bibliotecas da linguagem Python, com o folium, e o sistema de informação geográfica QGIS.

Outras *pipelines* podem ser criadas com base na arquitetura típica de um SDW, a depender do ambiente em que o gestor deseja implementar. Numa plataforma Google

[1] `https://www.ibge.gov.br/geociencias/downloads-geociencias.html`
[2] `https://www.openstreetmap.org/`
[3] `https://go.ok.org.br/dados-segp`

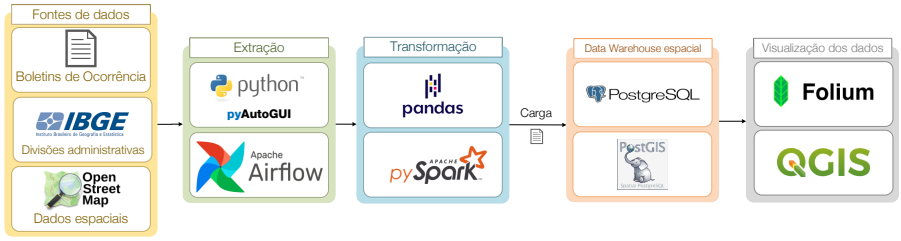

Figura 2. *Pipeline* baseada na arquitetura típica de um SDW, no contexto de criminalidade.

Cloud[4], por exemplo, pode ser utilizado o BigQuery para *data warehousing* e o Looker Studio para visualização dos dados.

4. Estudo de caso

Seguindo o esquema lógico e a *pipeline* descritos na seção 3, foi criado um ambiente de SDWing que utiliza dados de BOs disponibilizados pela SSP/SP[5]. Os BOs registrados incluem crimes como roubo e furto de veículos e celulares, homicídios e feminicídios. Os dados extraídos foram de BOs registrados a partir de 2011, pois não há dados de geolocalização de BOs antes deste ano. Com isso, foram extraídos dados de 2.193.338 BOs registrados no estado de São Paulo. O processo de extração e transformação aplicado aos dados da SSP/SP para o carregamento destes no esquema lógico da Figura 1 apresenta grande complexidade, devido a problemas relacionados à padronização, consistência e de integração com outras bases. Portanto, o detalhamento deste processo está fora do escopo deste artigo. Ele pode ser encontrado em [Freitas et al. 2023], no qual o *dataset* gerado é disponibilizado forma aberta para o público.

As seções 4.1 e a 4.2 descrevem consultas SOLAP de interesse por parte de gestores de segurança pública, cujos resultados estão dispostos em gráficos ou mapas.

4.1. Distribuição de roubos de veículos no município de Santa Bárbara d'Oeste

Nessa consulta SOLAP de *slice*, utilizou-se as tabelas de dimensão Crime, Municipio e Data para verificar a localização dos crimes de roubo de veículos na cidade de Santa Bárbara d'Oeste no ano de 2022. Os resultados estão ilustrados na Figura 3. Verifica-se uma concentração elevada de roubos de veículos em bairros como Jardim das Orquídeas e Jardim São Fernando, e no centro da cidade. Com isso, um gestor de segurança pública de Santa Bárbara d'Oeste pode prover a melhoria no patrulhamento da Polícia Militar ou da Guarda Municipal nestes bairros, com ênfase em períodos nos quais há uma predisposição maior para que estes tipos de crimes ocorram, como o período noturno.

4.2. Quantidade de furtos na cidade de São Paulo

Nessa consulta de *roll-up* e *slice*, utilizou-se as tabelas de dimensão Crime, Municipio e Data para verificar a quantidade de furtos classificados por tipo entre os anos de 2018 e 2022. A partir do gráfico de barras ilustrado na Figura 4, nota-se que os furtos em veículos, no transporte público e a transeuntes em vias públicas foram recorrentes. Na elaboração do BO, embora exista um campo que indique o ambiente em que o furto ocorreu, nem sempre esse campo é preenchido, ocasionando que o crime seja categorizado

[4]https://cloud.google.com
[5]http://www.ssp.sp.gov.br/transparenciassp/

Figura 3. Mapa de calor dos pontos em que ocorreram roubos de veículo no ano de 2022 no município de Santa Bárbara d'Oeste

como "outros". Nota-se também uma diminuição de furtos em 2020 devido ao isolamento social causada pela pandemia da COVID-19. Com esses resultados, um gestor de segurança pública pode planejar ações que diminuam os crimes ocorridos com maior frequência, alocando agentes da segurança pública em locais estratégicos, como estacionamentos, para evitar furtos em veículos, por exemplo.

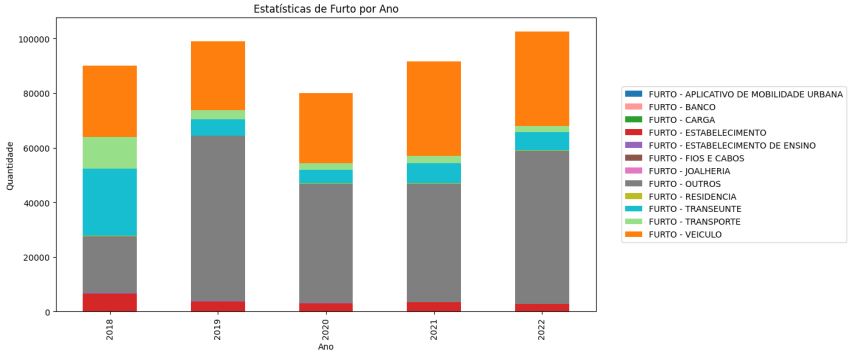

Figura 4. Gráfico da quantidade de furtos na cidade de São Paulo entre os anos de 2018 e 2022

5. Conclusão

Este trabalho propôs um ambiente de SDWing sobre dados de criminalidade para auxiliar gestores na tomada de decisão para combater crimes. Foi apresentado um esquema lógico de um SDW e uma *pipeline* baseada na arquitetura típica de um SDW, com tecnologias que podem ser usadas para a implementação do ambiente. Foi realizado um estudo de caso utilizando dados da SSP/SP, com a execução de consultas SOLAP dentro do ambiente proposto. Os resultados identificaram padrões interessantes para possíveis aplicações no contexto de criminalidade.

Trabalhos futuros incluem a adição de novas tabelas de dimensão e fatos relacionados ao esquema lógico do SDW. Pretende-se adicionar novas fontes de dados que sejam relacionadas aos dados de segurança pública, como o Sistema de Informações sobre Mortalidade (SIM). Utilizando o esquema lógico estendido, pretende-se formular novas consultas analíticas espaciais e discutir como os resultados dessas consultas podem ser empregados para melhorar a tomada de decisão por gestores.

Agradecimentos

Este trabalho foi apoiado pela Universidade de São Paulo, a partir do Programa Unificado de Bolsas de estudo para apoio à formação de estudantes de graduação (PUB-USP), pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

Al-Sai, Z. A. and Abualigar, L. M. (2017). Big Data and E-government: A review. In *8th International Conference on Information Technology (ICIT)580*, pages 580–587.

BRASIL (2022). Plano de Dados Abertos 2022-2024. Technical report, Ministério da Justiça e Segurança Pública, Brasília.

Cerqueira, D., Bueno, S., Palmieri Alves, P., Sergio de Lima, R., R. A. da Silva, E., Ferreira, H., Pimentel, A., Barros, B., Marques, D., Pacheco, D., de Oliveira Accioly Lins, G., dos Reis Lino, I., Sobral, I., Figueiredo, I., Martins, J., Chacon Armstrong, K., and da Silva Figueiredo, T. (2020). Atlas da Violência 2020. *Relatório Institucional*, pages 1–91.

Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1):65–74.

Freitas, J. B., Clarindo, J. P., and Aguiar, C. D. (2023). SPSafe: um dataset sobre dados de criminalidade no estado de São Paulo. *XXXVIII Simpósio Brasileiro de Banco de Dados: V Dataset Showcase Workshop (DSW), SBBD 2023 Companion*.

Han, J., Stefanovic, N., and Koperski, K. (1998). Selective materialization: An efficient method for spatial data cube construction. In *LNCS*, volume 1394, pages 144–158, Berlin, Heidelberg, Germany. Springer.

Kasprzyk, J.-P. and Donnay, J.-P. (2016). A Raster SOLAP for the Visualization of Crime Data Fields. *8th International Conference on Advanced Geographic Information Systems, Applications, and Services*, pages 121–125.

Macedo, D. F. and Lemos, D. L. d. S. (2021). Dados abertos governamentais: iniciativas e desafios na abertura de dados no Brasil e outras esferas internacionais. *AtoZ: novas práticas em informação e conhecimento*, 10(2):14.

Rivest, S., Bédard, Y., and Marchand, P. (2001). Toward better support for spatial decision making: defining the characteristics of Spatial On-Line Analytical Processing (SOLAP). *Geomatica*, 55(4):539–555.

Sá, B. C., Muller, G., Banni, M., Santos, W., Lage, M., Rosseti, I., Frota, Y., and Oliveira, D. d. (2021). PolRoute-DS: um Dataset de Dados Criminais para Geração de Rotas de Patrulhamento Policial. *Anais do Dataset Showcase Workshop (DSW)*, pages 117–127.

Santos, W. and Oliveira, D. d. (2018). Um Ambiente de Apoio à Decisão baseado em Data Warehouse para a Área de Segurança Pública do Estado do Rio de Janeiro. *Anais do Workshop Brasileiro de Cidades Inteligentes (WBCI)*.

Vaisman, A. and Zimányi, E. (2014). *Data Warehouse Systems: Design and Implementation*. Springer Publishing Company, Incorporated, Berlin, Heidelberg, Germany.