

Processamento de consultas analíticas espaciais sobre dados de cidades inteligentes

João Paulo Clarindo dos Santos¹, Cristina Dutra de Aguiar Ciferri¹

¹Programa de Pós-Graduação em Ciências de Computação
e Matemática Computacional
Instituto de Ciências Matemáticas e Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos, SP – Brasil

jpc santos@usp.br, cdac@icmc.usp.br

Nível: Mestrado

Ingresso no programa: Março/2019

Etapas concluídas: créditos em disciplinas; exame de qualificação

(apresentação em Setembro/2020); proposta de solução

Época esperada de conclusão e defesa: Março/2021

Abstract. *Spatial data generated by the Internet of Things (IoT) devices are important to assist the decision-making in issues related to smart cities. Spatial Data Warehouses (SDW) can be used to process analytical queries extended with spatial predicates (i.e. SOLAP queries – spatial on-line analytical processing). A smart city produces a huge volume of spatial data, thus processing SOLAP queries over spatial data generated by IoT devices is much expensive. Therefore, the processing of SOLAP queries may benefit from the use of frameworks aimed to provide parallelism and data distribution. Also, spatial analytics systems are developed on top of these frameworks to provide support to spatial data. In this paper, we fill a gap in the literature by investigating the processing of SOLAP queries to support the decision-making in the context of IoT and smart cities, using parallel and distributed processing and spatial analytics systems. We aim to introduce the following contributions: (i) specification of an architecture for a SDW environment in the context of IoT and smart cities; (ii) development of new methods related to the SOLAP query processing, using as basis the proposed architecture; and (iii) validation of the architecture and methods using real spatial data obtained from IoT devices.*

Palavras-Chave. IoT, SOLAP, data warehouses espaciais, sistemas analíticos espaciais, cidades inteligentes.

1. Introdução

Prover a infraestrutura necessária para comportar uma grande quantidade de pessoas em cidades pode ser um desafio para o poder público e empresas, já que necessitam de mecanismos de monitoramento de recursos para auxiliar a tomada de decisão. Logo, surgiu o conceito de cidades inteligentes, que são cidades cujas infraestruturas utilizam tecnologias visando o crescimento urbano e social. Exemplos são os dispositivos de Internet das Coisas (ou *Internet of Things* (IoT)), que são dispositivos interconectados a fim de fornecer uma base de informações [Yeh 2017, Atzori et al. 2017].

A tecnologia de IoT em uma cidade inteligente pode ser aplicada, por exemplo, no transporte público, cuja frota contém sensores que coletam dados relativos à quantidade de passageiros, tipo de veículo (ônibus, vans, etc.), rota realizada e velocidade máxima, visando aprimorar as linhas existentes. Outro exemplo de aplicação consiste em uma rede de sensores de poluentes espalhados pela cidade que coletam dados relativos à qualidade do ar naquela região, estimulando a melhora da qualidade do ar. Outras aplicações de dispositivos IoT em uma cidade inteligente incluem controle de tráfego, análise do consumo de água e controle de energia de fontes renováveis. [Atzori et al. 2017, Eldrandaly et al. 2019].

Consultas sobre dados gerados por dispositivos IoT em uma cidade inteligente podem auxiliar na tomada de decisão de gestores. Por exemplo, no cenário de dispositivos instalados no transporte público, pode-se determinar “*quantos passageiros foram transportados no último mês, por tipo de veículo, por rota, por região*”, sendo os resultados obtidos exibidos em um mapa. Uma consulta como esta pode ser respondida a partir da extração, transformação e carga dos dados em um *data warehouse* (DW), que é uma base de dados histórica, orientada a assunto, integrada e não volátil [Kimball et al. 2011]. Os dados são organizados como assuntos de interesse que podem ser analisados a partir de um conjunto de dimensões e que oferecem subsídios para o processamento analítico de consultas (*on-line analytical processing* – OLAP) [Chaudhuri and Dayal 1997].

Dispositivos IoT, além de gerar dados convencionais, tendem a gerar dados espaciais, que são composições que representam a geometria de objetos espaciais. Essas geometrias são usualmente representadas por pontos, linhas e polígonos, ou combinações destes [Güting 1994]. Nesse contexto, as consultas OLAP são limitadas por não considerarem os relacionamentos espaciais. Logo, surgiu o conceito de DW Espacial (DWE), que estende um DW convencional para lidar com dimensões não-espaciais, espaciais ou com ambas. Sobre o DWE incidem consultas Spatial OLAP (SOLAP), as quais oferecem suporte às consultas espaciais [Rivest et al. 2001].

No contexto de cidades inteligentes, DWE é muito volumoso, pois os dispositivos geram muitos dados espaciais [Bonomi et al. 2014]. Isso introduz problemas em relação ao processamento de consultas, uma vez que consultas SOLAP são caras pelo alto custo da junção-estrela e do processamento dos predicados espaciais [Rivest et al. 2001]. Logo, um DWE pode ser favorecido pelo uso de um *framework* de processamento paralelo e distribuído, como Hadoop e Spark, visando diminuir o custo do processamento dos dados espaciais. Este *framework* pode ser inserido em uma nuvem, permitindo a flexibilidade de acordo com a demanda exigida pela cidade inteligente. Entretanto, ele não oferece suporte nativo para o processamento de predicados espaciais, exigindo o uso de sistemas

analíticos espaciais (SAEs), que são expansões que permitem consultas e indexação destes dados espaciais [Castro et al. 2019].

A partir de revisão sistemática (seção 3), não foi detectado um trabalho que relate o processamento de dados espaciais obtidos de dispositivos IoT no contexto de cidades inteligentes em um DWE inserido em um ambiente paralelo e distribuído. Logo, este projeto de mestrado tem como objetivo investigar essa limitação existente na literatura, com as seguintes contribuições: (i) proposta de uma arquitetura relacionada a temática; (ii) proposta de métodos voltados ao processamento da consulta SOLAP considerando a arquitetura proposta; e (iii) validação da arquitetura e dos métodos propostos sobre dados reais obtidos de dispositivos IoT. O trabalho a ser desenvolvido, embora possua enfoque no contexto de cidades inteligentes, pode ser aplicado a qualquer tipo de dado espacial gerado por uma rede de dispositivos IoT, como exemplo dados espaciais gerados por *smartphones* e que são disponibilizados por mídias sociais.

Este artigo está estruturado da seguinte forma. Na seção 2 é apresentada a fundamentação teórica. Na seção 3 é descrita a revisão sistemática. Na seção 4 é detalhado o estado atual de desenvolvimento do trabalho. Por fim, na seção 5 são descritas as considerações finais e as próximas atividades a serem desenvolvidas.

2. Fundamentação Teórica

2.1. Rede de dispositivos IoT e Computação em Névoa

Existem vários dispositivos IoT como sensores de umidade e pressão, sensores de poluição, *tags* e Arduino. Para criar uma rede que conecte estes dispositivos a um servidor que possa manipular os dados gerados, utilizam-se várias tecnologias de transmissão, como *Radio Frequency IDentification* (RFID), *Global Positioning System* (GPS) e Wi-Fi [Atzori et al. 2017]. O envio de dados diretamente para um servidor próprio ou para um ambiente de computação em nuvem pode se mostrar ineficiente, devido à alta latência na transferência entre os dispositivos [Eldrandaly et al. 2019].

Com isso, [Bonomi et al. 2014] propuseram o paradigma computação em névoa (*fog computing*), que provê serviços de processamento, armazenamento e distribuição de dados próximos à borda da rede. Estes serviços são executados na camada da névoa, que consiste de dispositivos como roteadores, *gateways* e servidores locais capazes de processar, transmitir e armazenar temporariamente os dados recebidos pelos dispositivos IoT, possibilitando operações de extração, transformação e carga (ETL) e processamento analítico em tempo real. As consultas que demandam maior poder de processamento são realizadas na camada da nuvem, assim como é nesta camada que os dados históricos são armazenados. Algumas vantagens do uso de computação em névoa incluem a ampla distribuição geográfica dos serviços, redes de dispositivos distribuídos em larga escala, interações em tempo real, predomínio de acesso sem fio e heterogeneidade da rede de dispositivos IoT.

2.2. Dados espaciais e relacionamentos espaciais

Dados espaciais são componentes que representam a geometria de objetos espaciais. Estes dados podem ser categorizados na forma vetorial ou *raster*. Os dados espaciais podem se relacionar de forma métrica, topológica ou direcional [Güting 1994]. Em um DWE, além de operações OLAP convencionais, também são processadas consultas es-

paciais [Rivest et al. 2001]. As consultas espaciais são consultas nas quais ao menos um de seus predicados envolve um relacionamento espacial. Essas consultas podem se mostrar complexas, sendo necessário o uso de índices espaciais para execução eficiente dessas consultas. Uma estrutura popular de indexação é a *R-tree*, derivada da *B⁺-tree* [Gaede and Günther 1998].

3. Revisão Sistemática

A revisão sistemática refere-se a uma metodologia de pesquisa que visa reunir e avaliar as evidências disponíveis referentes a um determinado tópico a partir de passos, que vão desde a formulação de questões de pesquisa até a síntese dos resultados encontrados [Biolchini et al. 2005]. A revisão realizada propôs identificar publicações que atendessem às seguintes questões de pesquisa: (i) Como grandes quantidades de dados espaciais são gerados e manipulados por dispositivos IoT?; (ii) Como os dados espaciais gerados por dispositivos IoT podem ser gerenciados em um ambiente SOLAP?; (iii) Como consultas SOLAP podem auxiliar na tomada de decisão no contexto de cidades inteligentes?; (iv) Como um DW pode ser utilizado em ambientes de computação em névoa?; e (v) Existem estudos que relacionam SOLAP e computação em névoa no contexto de cidades inteligentes a fim de auxílio na tomada de decisão? Foram consideradas as fontes de busca IEEEXplore DL, ACM DL e Elsevier ScienceDirect.

As *strings* de busca utilizadas foram as seguintes (a) “*spatial data*” AND (“*internet of things*” OR *iot*) AND “*big data*”; (b) (*solap* OR *sdw* OR “*spatial data warehouse*”) AND (“*internet of things*” OR “*iot*” OR “*smart cities*”); e (c) (“*data warehouse*” OR *iot*) AND “*fog computing*”. Foram retornados 100 trabalhos, sendo selecionados cinco trabalhos que satisfaziam às questões de pesquisa e aos critérios de seleção, os quais incluíram trabalhos em inglês ou português e cuja publicação foi feita entre os anos de 2016 e 2020. Adicionalmente, dois trabalhos foram incluídos para a sintetização de resultados a partir de indicação de especialista. Esses dois trabalhos, embora não atendam ao critério de pesquisa relacionado ao período de tempo, possuem potencial para responder às questões de pesquisa (ii) e (iv). Os trabalhos foram agrupados em três grupos de estudo associados às *strings* de busca (a), (b) e (c), conforme ilustrado na Tabela 1.

Tabela 1. Tópicos abordados nos trabalhos selecionados na revisão sistemática

Trabalho	Grupo de Estudo	Dados Espaciais	SOLAP / OLAP	IoT	Cidades Inteligentes	Computação em Névoa	Satisfaz à Questão de Pesquisa
[Eldrandaly et al. 2019]	(a)	✓	✗	✓	✓	✓	(i)
[Iyer and Stoica 2017]	(a)	✓	✗	✓	✗	✗	(i)
[Jo et al. 2019]	(a)	✓	✗	✓	✗	✗	(i)
[Liu et al. 2019]	(b)	✓	SOLAP	✗	✓	✗	(iii)
[Khakimov et al. 2018]	(b)	✗	OLAP	✓	✗	✓	(iv)
[Yuan and Zhao 2012]	(c)	✓	SOLAP	✓	✗	✗	(ii)
[Bonomi et al. 2014]	(c)	✗	OLAP	✓	✓	✓	(iv)
Abordagem proposta	-	✓	SOLAP	✓	✓	✓	(v)

Embora os trabalhos analisados satisfaçam ao menos uma questão de pesquisa, nenhum deles responde à questão de pesquisa (v), que relaciona SOLAP e IoT no contexto de cidades inteligentes. Nos trabalhos de [Bonomi et al. 2014, Eldrandaly et al. 2019] são propostas arquiteturas que relacionam dados gerados por dispositivos IoT e computação em névoa. Entretanto, essas arquiteturas carecem de processamento analítico em um DWE inserido em um ambiente de processamento paralelo e distribuído. Adicionalmente, o trabalho de [Yuan and Zhao 2012] possui como objetivo a criação de um ambiente SOLAP no contexto de IoT, sendo o que possui maior aproximação com objetivo do mes-

trado. Entretanto, este trabalho não inclui os conceitos mais recentes em relação à IoT, como computação em névoa e processamento paralelo e distribuído. O trabalho a ser desenvolvido no mestrado pretende cobrir esta lacuna existente na literatura, conforme listado na Tabela 1.

4. Proposta e Estágio Atual de Desenvolvimento

4.1. Descrição da arquitetura

Para suprir a lacuna existente nas literatura e dado o objetivo deste trabalho, foi proposta uma arquitetura que relaciona dados obtidos por dispositivos IoT no contexto de cidades inteligentes com um DWE inserido em um ambiente de processamento e armazenamento paralelo e distribuído. Esta arquitetura, ilustrada na Figura 1, possui três camadas: terminal, processamento na névoa e processamento na nuvem.

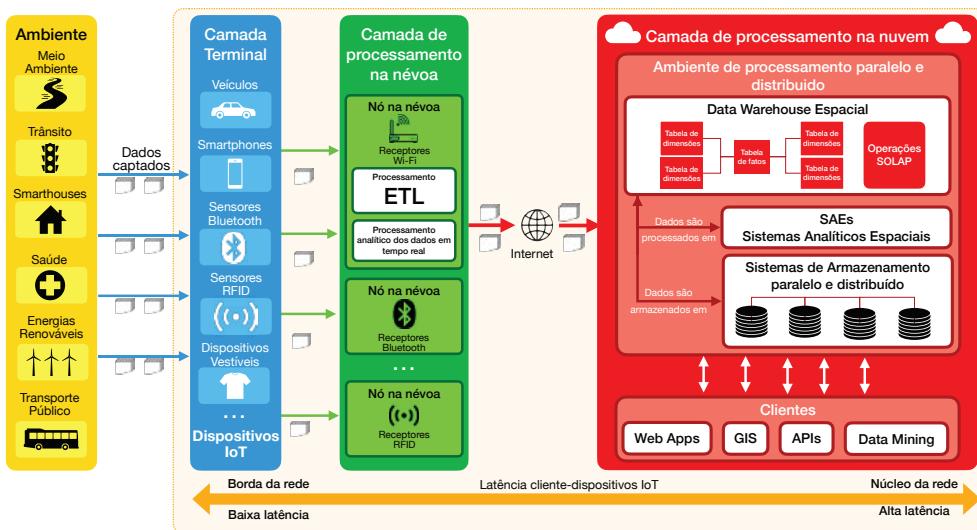


Figura 1. Visão geral da arquitetura proposta

Na camada terminal, os dados são coletados do ambiente a partir de dispositivos IoT, como sensores, *smartphones*, veículos e dispositivos vestíveis, dentre outros. Esses dispositivos conectam-se com a camada de processamento da névoa, que contém receptores com diversos protocolos de comunicação como exemplo Wi-Fi, RFID, Bluetooth e que possuem poder limitado de processamento e armazenamento dos dados. Estes receptores, que são os nós da névoa, são responsáveis pelas operações de ETL e pelo processamento analítico dos dados em tempo real, pois estes dispositivos encontram-se próximos à borda da rede, ocasionando na baixa latência entre o nó na névoa e o usuário.

Com relação ao processamento analítico dos dados históricos, os dados, após passarem pela etapa de ETL, são enviados para a camada de processamento em nuvem. Nesta camada, os dados são armazenados em um sistema de armazenamento paralelo e distribuído, dispostos em um DWE. As consultas SOLAP são processadas em SAEs vinculados ao sistema de processamento e armazenamento paralelo e distribuído. Pela natureza escalável da computação em nuvem, a quantidade de nós em um ambiente paralelo e distribuído pode ser aumentado ou diminuído dependendo das demandas de consultas da rede. Por fim, os clientes, como Web Apps e Geographic Information Systems (GIS) realizam consultas no DWE.

4.2. Detalhamento das Atividades

No momento atual do trabalho, está sendo investigada a instanciação da arquitetura com tecnologias disponíveis e de acordo com os objetivos do trabalho. O foco principal da pesquisa é a camada de processamento na nuvem. Portanto, a camada de processamento em névoa prevê o uso de tecnologias já existentes para oferecer subsídios para o trabalho. Essa camada utiliza o Apache Airflow para realizar operações ETL e o PostgreSQL para armazenamento dos dados antes de serem enviados para a nuvem. Na camada de processamento na nuvem, os dados do DWE são manipulados utilizando o *framework* de processamento paralelo e distribuído Spark e armazenados no *Hadoop File System* (HDFS). A escolha das tecnologias descritas foi feita com base nos avanços obtidos pelo grupo de pesquisa no qual o projeto de mestrado se enquadra.

Com relação ao processamento de consultas SOLAP considerando a arquitetura proposta, pretende-se desenvolver métodos para processar essas consultas eficientemente. Os métodos devem considerar os dados espaciais vetoriais e os relacionamentos topológicos de *intersection*, *closure* e *containment*. Para o processamento da junção-estrela, pretende-se usar os avanços descritos no trabalho de [Sangat et al. 2020], que representa o estado-da-arte neste assunto. Para o processamento do relacionamento topológico, pretende-se usar as funcionalidades providas pelo GeoSpark [Yu et al. 2015]. Também pretende-se investigar o projeto de diferentes esquema-estrela para a organização do DWE. Será analisado como estas disposições impactam na execução das consultas espaciais utilizando como base os estudos realizados por [Mateus et al. 2016]. É importante observar que os métodos a serem desenvolvidos visam integrar os componentes da arquitetura.

Para a validação da arquitetura e dos métodos propostos, serão usados *datasets* contendo dados gerados em dispositivos IoT no contexto de cidades inteligentes, fornecidos por portais de dados abertos governamentais e dados das cidades de Aarsus e Brașov gerados por [Ali et al. 2015]. Alguns tipos de dados encontrados nestes *datasets* incluem dados de poluentes, tráfego urbano, energia elétrica e estacionamentos. Os parâmetros para teste de desempenho da arquitetura incluem a definição do volume, ambiente, complexidade e seletividade das consultas, custo de processamento e tempo de execução destas consultas. Os resultados serão publicados em conferências e periódicos relacionados às áreas de banco de dados, geoinformática e IoT.

5. Conclusão

O projeto de mestrado tem como objetivo investigar o processamento de dados espaciais obtidos de dispositivos IoT no contexto de cidades inteligentes em um DWE inserido em um ambiente paralelo e distribuído, com ênfase em SOLAP. Neste contexto, este artigo descreve os conceitos de computação em névoa, dados espaciais e consultas espaciais, além de apresentar uma revisão sistemática relacionada aos temas do trabalho. O artigo também detalha os objetivos da pesquisa e descreve a arquitetura proposta, a qual será usada como base para o desenvolvimento do trabalho. A próxima etapa a ser realizada refere-se ao desenvolvimento dos métodos para o processamento de consultas SOLAP, conforme detalhado na seção 4.2. Na sequência, ocorrerá a validação desses métodos.

Referências

- Ali, M. I., Gao, F., and Mileo, A. (2015). CityBench: A configurable benchmark to evaluate RSP engines using smart city datasets. In *LNCS*, volume 9367, pages 374–389. Springer Verlag.
- Atzori, L., Iera, A., and Morabito, G. (2017). Understanding the Internet of Things: definition, potentials, and societal role of a fast evolving paradigm. *Ad Hoc Networks*, 56:122–140.
- Biolchini, J., Gomes Mian, P., Candida Cruz Natali, A., and Horta Travassos, G. (2005). Systematic Review in Software Engineering. Technical report, COPPE/UFRJ.
- Bonomi, F., Milito, R., Natarajan, P., and Zhu, J. (2014). Fog computing: A platform for internet of things and analytics. *Studies in Computational Intelligence*, 546:169–186.
- Castro, J. P. C., Carniel, A. C., and Ciferri, C. D. A. (2019). A User-centric View of Distributed Spatial Data Management Systems. In *GEOINFO*, pages 80–91.
- Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1):65–74.
- Eldrandaly, K. A., Abdel-Basset, M., and Shawky, L. A. (2019). Internet of Spatial Things: A New Reference Model With Insight Analysis. *IEEE Access*, 7:19653–19669.
- Gaede, V. and Günther, O. (1998). Multidimensional access methods. *ACM Computing Surveys*, 30(2):170–231.
- Güting, R. H. (1994). An introduction to spatial database systems. *The VLDB Journal*, 3(4):357–399.
- Iyer, A. P. and Stoica, I. (2017). A scalable distributed spatial index for the internet-of-things. In *ACM SoCC*, pages 548–560.
- Jo, J., Joo, I. H., and Lee, K. W. (2019). Constructing national geospatial big data platform: Current status and future direction. In *IEEE WF-IoT*, pages 979–982.
- Khakimov, A., Muthanna, A., and Muthanna, M. S. A. (2018). Study of fog computing structure. In *IEEE ElConRus*, pages 51–54.
- Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., and Becker, B. (2011). *The Data Warehouse Lifecycle Toolkit*, volume 3. John Wiley & Sons Inc, Hoboken, NJ.
- Liu, C., Wu, C., Shao, H., and Yuan, X. (2019). SmartCube: An Adaptive Data Management Architecture for the Real-Time Visualization of Spatiotemporal Datasets. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1.
- Mateus, R. C., Siqueira, T. L. L., Times, V. C., Ciferri, R. R., and Ciferri, C. D. A. (2016). Spatial data warehouses and spatial OLAP come towards the cloud: design and performance. *Distributed and Parallel Databases*, 34(3):425–461.
- 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.
- Sangat, P., Taniar, D., and Messom, C. (2020). Distributed ATrie Group Join: Towards Zero Network Cost. *IEEE Access*, 8:111598–111613.
- Yeh, H. (2017). The effects of successful ICT-based smart city services: From citizens' perspectives. *Government Information Quarterly*, 34(3):556–565.
- Yu, J., Wu, J., and Sarwat, M. (2015). GeoSpark: A cluster computing framework for processing large-scale spatial data. In *ACM GIS*, pages 1–4, New York.
- Yuan, L. and Zhao, J. (2012). Construction of the system framework of Spatial Data Warehouse in Internet of Things environments. In *5th IEEE ICACI*, pages 54–58.