Os desafios e soluções para a implementação de Big Data Analytics em cidades inteligentes

Felipe F. Vasconcelos[1], **Vinicius T. Ramos**[1], **Fábio J. Coutinho**[1]

[1]Instituto de Computação – Universidade Federal Alagoas (UFAL)

ffv,vtpr,fabio@ic.ufal.br

Abstract. *In recent years, the growth of smart cities has required solutions for managing massive and heterogeneous urban data in order to transform raw data into information. These solutions need to deal with particular characteristics of data access, format and manipulation, requiring different approaches for data processing, integration and analysis. In this work, we discuss the aspects involved in the construction of Big Data Analytics in smart cities, highlighting the characteristics of the available solutions. In order to demonstrate the tools discussed in this work, an initial experiment was carried out, which built a data pipeline focused on the analysis of bus traffic in the city of São Paulo.*

Resumo. *Nos últimos anos, o crescimento das cidades inteligentes tem exigido soluções para a gerência de dados urbanos massivos e heterogêneos com o objetivo de transformar dados brutos em informação. Essas soluções precisam lidar com características particulares de acesso, formato e manipulação dos dados, requerendo abordagens distintas para o processamento, integração, análise de dados. Neste trabalho, discutimos os aspectos envolvidos na construção de Big Data Analytics em cidades inteligentes, ressaltando as características das soluções disponíveis. No intuito de demonstrar as ferramentas discutidas neste trabalho, foi realizado um experimento inicial, que construiu uma pipeline de dados voltada para a análise do tráfego dos ônibus da cidade de São Paulo.*

1. Introdução

Os recentes avanços tecnológicos permitiram um crescimento na utilização de dispositivos IoT para diversas finalidades. Considerando o desenvolvimento de cidades inteligentes, os dispositivos IoT exercem um papel estratégico e possuem presença cada vez maior nos grandes centros urbanos. A expansão do uso desses dispositivos eleva significativamente o volume de dados gerados. Em 2023, estima-se que mais de 100 zetabytes de dados sejam produzidos [Reinsel et al. 2017]. Neste contexto, transformar uma quantidade massiva de dados brutos em informação útil representa um desafio importante, que abrange diferentes etapas dentre as quais se destacam a ingestão, limpeza e transformação de dados; a análise/predição e visualização de dados.

Em uma típica arquitetura de cidade inteligente, encontram-se múltiplas fontes espalhadas por diferentes regiões, produzindo dados heterogêneos relativos a diversas áreas de conhecimento como administração, saúde, meteorologia, mobilidade, logística, segurança, etc. Tal cenário exige lidar com características particulares de acesso, formato e manipulação dos dados, requerendo abordagens distintas para o processamento,

integração, análise e visualização de dados. As propostas de solução buscam criar diferentes fluxos de dados e processos a partir de uma infraestrutura capaz de executar procedimentos de ELT (extract, load and transform) através de ferramentas de ingestão de dados, armazenamento distribuído e processamento analítico [Clarindo et al. 2021].

Neste trabalho são discutidos os principais desafios e soluções existentes para a construção de *Big Data Analytics* no contexto de cidades inteligentes. Os desafios envolvem típicos problemas de aplicações big data inseridos no âmbito do processamento analítico de dados urbanos massivos e heterogêneos, tais como lidar com fontes heterogêneas, processar dados espaciais e prover um armazenamento eficiente. Na literatura, existem trabalhos que exploram diversas ferramentas como soluções para essas problemáticas [Rathore et al. 2018] [Ahad et al. 2020] e outros que propõem novas arquiteturas para implementar o ambiente analítico [Clarindo et al. 2021], [Massobrio et al. 2018].

Este trabalho tem como objetivo principal identificar as características inerentes à análise de dados de cidades inteligentes no intuito de encontrar as possíveis soluções para as etapas de ingestão, armazenamento, processamento paralelo, visualização e análise de dados. As ferramentas discutidas neste trabalho são aplicadas em um experimento inicial, que construiu uma pipeline de dados voltada para a análise do tráfego de ônibus da cidade de São Paulo.

O artigo encontra-se organizado da seguinte forma: na seção 2 são apresentados os desafios para implementação de Big Data Analytics em cidades inteligentes; a seção 3 discute, por meio de uma análise comparativa, as possíveis soluções para os desafios propostos e, finalmente, a seção 4 traz as considerações finais, trabalhos futuros e o experimento realizado.

2. Desafios para Big Data Analytics em cidades inteligentes

A análise de dados de cidades inteligentes abrange características e demandas específicas tais como processamento de dados espaciais, resposta em tempo real, tolerância a falhas, escalabilidade, entre outras. Essas necessidades, quando implementadas neste contexto, devem considerar o emprego de técnicas como ingestão de dados de dispositivos IoT, armazenamento distribuído, processamento paralelo, análise espaço-temporal e outras. As seções a seguir discutem esses tópicos.

2.1. Ingestão de Dados

O ecossistema de uma cidade inteligente produz dinamicamente dados volumosos e heterogêneos a partir de múltiplas fontes provenientes de diferentes sistemas como monitoramento do tráfego de veículos, previsão de tempo, alerta de segurança, mapeamento de áreas, controle sanitário, entre outros. Neste contexto, o processo de ingestão de dados torna-se complexo principalmente devido ao volume, a heterogeneidade e a velocidade de geração dos dados. Ao construir uma solução, deve-se considerar a adição de novas fontes de dados com diferentes formatos; a capacidade de ingestão de dados em tempo real via streaming; a tolerância a falhas, dado que erros na transmissão dos dados podem causar distorções nas análises [Meehan et al. 2017]; a segurança, de modo a proteger dados sensíveis como a privacidade dos cidadãos [Mătăcuţă and Popa 2018].

2.2. Armazenamento Distribuído e Data Lakes

O processamento analítico de dados de uma cidade inteligente deve considerar aspectos relevantes com respeito à solução de armazenamento. Características como *elevado volume*, *variedade* e *baixa latência* inviabilizam a aplicação de soluções tradicionais como Data Warehouses (DW) e SGBDs relacionais. Desse modo, faz-se necessário encontrar soluções de armazenamento que possam prover escalabilidade, flexibilidade e tolerância a falhas.

A abordagem distribuída dispõe de diferentes nós de armazenamento, sendo capaz de prover suporte ao processamento analítico de cidades inteligentes visto que oferta escalabilidade e tolerância a falhas através de técnicas como a fragmentação e replicação de dados [Chang and Cui 2021]. Essas demandas são importantes, dado o constante crescimento urbano e a necessidade de alta disponibilidade para os serviços destinados a cidadãos, gestores e cientistas de dados. Neste contexto, os data lakes representam um paradigma de armazenamento que busca implementar um repositório de dados de baixo custo, escalável, e capaz de armazenar informações provenientes de fontes heterogêneas. Para lidar com a heterogeneidade, os data lakes utilizam schemas on-read, retirando a necessidade de modelagem prévia dos dados. Dessa forma, os data lakes realizam a ingestão e o armazenamento de dados em seus formatos nativos, priorizando a extração de informações de acordo com as necessidades dos usuários. Essa abordagem tem ganhado destaque na indústria, especialmente em soluções voltadas para dispositivos IoT [Hai et al. 2021].

2.3. Processamento Paralelo

Cidades inteligentes produzem grandes volumes de dados a partir de dispositivos IoT de modo que se faz necessário implementar processos de Extract, Load e Transform (ELT) a fim de obter informações úteis. Porém, realizar tais operações em dados volumosos de natureza diversa demanda um alto pode computacional. O processamento paralelo e distribuído apresenta-se como uma possível solução capaz de reduzir os custos operacionais por meio do uso mais eficiente da capacidade computacional [Clarindo et al. 2021]. As duas principais plataformas gratuitas para processamento paralelo são os frameworks Apache Hadoop e Apache Spark, os quais serão melhor discutidos na seção 3.

2.4. Dados geoespaciais

Dados espaciais são tipos de dados relacionados a algum tipo de referência geométrica. Em específico, dados geoespaciais são dados espaciais que têm sua referência na superfície do planeta, contendo informações como latitude e longitude. No contexto de cidades inteligentes, os dados geoespaciais oferecem um grande potencial de informações sobre o espaço físico urbano, o que é relevante para a tomada de decisões acerca do aumento da qualidade de vida urbana em cidades inteligentes [Alablani and Alenazi 2020].

O armazenamento e processamento de dados espaciais impõem desafios devido às características de dados coletados por dispositivos IoT, como o grande volume [Chen et al. 2014] e a heterogeneidade [Liu et al. 2016], além da necessidade de um processamento eficiente em diferentes operações geoespaciais, como predicados topológicos (e.g. contém, no interior, e intersecciona), operações de conjuntos geométricos (e.g. união, intersecção), e operações numericas (e.g. área, distância)

[de Carvalho Castro et al. 2020]. Soluções para essas problemáticas serão discutidas na seção 3.

3. Soluções para Big Data Analytics em cidades inteligentes

Diante dos desafios apresentados na seção anterior, a escolha das soluções adequadas à criação de um ambiente de Big Data Analytics em cidades inteligentes torna-se uma tarefa complexa. Neste sentido, a construção da pipeline de dados deve considerar especificidades técnicas como: escalabilidade; capacidade de distribuição; interoperabilidade; análise de dados Big Data; análise de dados geoespaciais. Esta seção discute as principais soluções gratuitas disponíveis, realizando uma análise comparativa baseada em trabalhos como [Mătăcuţă and Popa 2018], [Pereira et al. 2018] e [Veiga et al. 2016].

3.1. Ingestão de Dados

Considerando os desafios apresentados na seção 2.1 e os resultados publicados em [Mătăcuţă and Popa 2018], nosso trabalho considerou duas ferramentas para a ingestão de dados: Apache Kafka e Apache NiFi. Kafka é uma solução distribuída que fornece uma plataforma de alta taxa de transferência e baixa latência para lidar com grandes fluxos de dados em tempo real. Enquanto o Apache NiFi é uma plataforma de integração e processamento de dados em tempo real, com interface visual e recursos de processamento e monitoramento de dados.

[Mătăcuţă and Popa 2018] evidenciam que o Kafka apresenta uma melhor interoperabilidade, escalabilidade, tolerância a falhas e melhor desempenho do que concorrentes como o Nifi. Kafka também se destaca como uma solução mais abrangente, permitindo a integração com uma maior variedade de dispositivos e ferramentas. Tal fator é relevante no contexto de cidades inteligentes, onde existe a presença de diferentes dispositivos (e.g. sensores de velocidade, câmeras de vigilância e outros). Cabe ressaltar que o NiFi pode ser utilizado em conjunto com o Kafka, sendo o Kafka responsável pela transferência de dados em tempo real enquanto o NiFi trata das operações em lote (e.g. backups diários). Considerando esses aspectos, além da capacidade de atender eficientemente às demandas de dados via streaming em cidades inteligentes, o Apache Kafka é indicado como principal solução para ingestão de dados.

3.2. Armazenamento Distribuído

Data warehouses tradicionais não conseguem atender eficientemente às demandas e propriedades características das aplicações Big Data [Panwar and Bhatnagar 2020]. Neste contexto, surge o data lake com a proposta de armazenar dados não-relacionais, provendo escalabilidade a baixo custo e o uso de schemas on-read. Essas características, em conjunto com a capacidade de armazenar dados em diferentes estágios (e.g. dados crus, dados pré-filtrados, dados processados) e diferentes formatos, fator importante para o cenário de cidades inteligentes, tornam os data lakes uma solução de armazenamento viável. Neste sentido, os data lakes podem atuar como serviço principal de armazenamento, a partir do qual outros serviços, como dashboards e modelos de aprendizagem de máquina poderiam implementar DW especializados, que fariam parte do repositório data lake original, aproveitando das vantagens desse paradigma.

A seguir, duas soluções de armazenamento capazes de implementar um data lake são discutidas: HDFS e Couchbase. HDFS é o sistema de armazenamento distribuído

do framework Hadoop, sendo bastante utilizado por aplicações big data, logo, representa uma solução validada pela indústria. Suas principais características são: escalabilidade; interoperabilidade com outras ferramentas além do Hadoop; tolerância a falhas. Por sua vez, o Couchbase é uma solução de armazenamento representante dos SGBDs NoSQL e que possui desempenho superior a concorrentes como MongoDB e RethinkDB [Pereira et al. 2018]. O SGBD Couchbase utiliza os modelos chave-valor e orientado a documentos JSON. Outras características relevantes providas pelo SGBD são a boa escalabilidade horizontal e vertical, o suporte à indexação e às tarefas MapReduce.

A escolha da solução irá depender dos dados e experimentos a serem utilizados pelos gestores das cidades. Todavia, em termos gerais, o HDFS é a solução mais indicada em nosso contexto devido à ferramenta de armazenamento fazer parte do ecossistema Hadoop, ofertando boa integração com o Spark, MapReduce e outros serviços de interesse.

3.3. Processamento e Análise

As duas principais plataformas gratuitas para o processamento paralelo de dados em larga escala são os frameworks Apache Hadoop e Apache Spark. O Apache Hadoop constitui uma suíte de softwares que implementam o paradigma MapReduce, sendo uma das soluções pioneiras para processamento big data, capaz de prover alta disponibilidade, tolerância a falhas e escalabilidade [Veiga et al. 2016].

Apache Spark é um framework para computação distribuída que pode atender, por meio da biblioteca SparkStreaming, às demandas de processamento de dados em tempo real, tal como a análise de dados de semáforos de trânsito. Spark também oferece bibliotecas como a MLLib e o SparkSQL, que facilitam operações como consultas e algoritmos de aprendizagem de máquina. Veiga et al. (2016) chegaram à conclusão de que o Spark apresenta um desempenho até 77% superior ao Hadoop [Veiga et al. 2016]. Com base nesses fatores, o Spark é indicado como ferramenta de processamento, servindo como solução tanto para dados em streaming quanto para dados em lotes.

3.4. Dados geoespaciais

O armazenamento de dados espaciais dispõe de soluções conhecidas, tais como PostGIS e SpatiaLite, extensões dos SGBDs Relacionais PostgreSQL e SQLite, respectivamente. Ambos oferecem suporte a diversos tipos de dados espaciais, podendo ser utilizados, por exemplo, para a criação de dashboards de dados urbanos.

Todavia, esse tipo de solução de armazenamento não conseguem lidar com Urban Big Data. Neste caso, soluções como GeoMesa oferecem consultas e análises geoespaciais em larga escala mediante integração com SGBDs NoSQL como HBase e Cassandra. Outras opções interessantes para a análise de dados geoespaciais são o GeoPandas, uma solução Python que estende a biblioteca Pandas, e o Apache Sedona, uma solução que estende o Apache Spark. Para a visualização de dados espaciais, é possível utilizar ferramentas como QGIS e Matplotlib.

4. Experimento, Trabalhos Futuros e Considerações Finais

A fim de validar as soluções indicadas neste trabalho, foi realizado um experimento inicial a partir da construção de uma pipeline, representada na figura 1, que teve como estudo de caso os dados de transporte público da cidade de São Paulo. A pipeline inicia com

a ferramenta Kafka responsável pela ingestão dos dados obtidos a partir de dispositivos instalados em ônibus da cidade de São Paulo. Em seguida, os dados são armazenados no HDFS para posterior consulta e processamento com o Spark. E finalmente, os dados são utilizados pelo QGIS para a geração dos mapas de calor representados nas figuras 2 e 3. A partir dos mapas, pode-se verificar a concentração de ônibus no bairro da Vila Leopoldina, às 15 horas e às 18 horas de uma terça-feira, indicando alteração de fluxo no decorrer do dia.

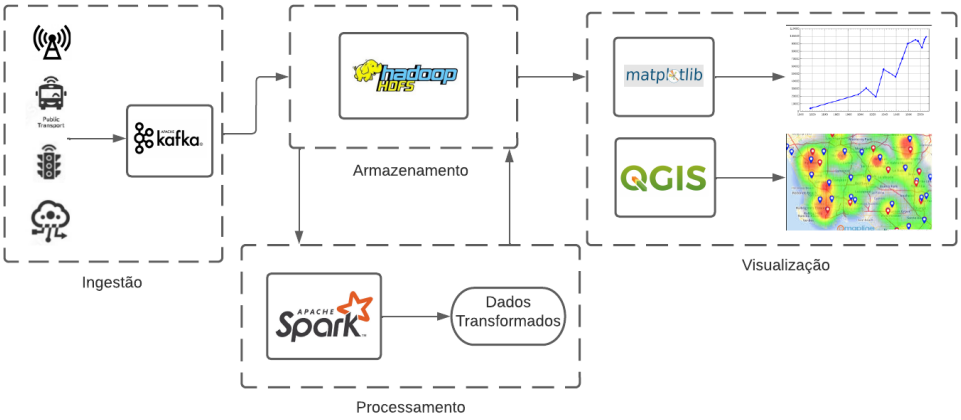

Figura 1. Diagrama da pipeline de dados construída.

Este trabalho elaborou uma discussão acerca dos desafios e soluções para a implementação de Big Data Analytics em cidades inteligentes mediante a identificação de suas principais demandas. Neste sentido, foram realizadas análises comparativas entre as ferramentas disponíveis gratuitamente a fim de auxiliar stakeholders nos desafios do processo de construção de ambientes para análise de dados urbanos massivos.

Como trabalhos futuros pretende-se considerar serviços de nuvem como possíveis soluções, além de expandir a discussão para abarcar outras temáticas inseridas no contexto de cidades inteligentes, tais como: a segurança e privacidade dos dados; qualidade e governança dos dados; data discovery e outros.

Figura 2. Mapa de calor na Vila Leopoldina, às 15 horas da terça-feira.

Figura 3. Mapa de calor na Vila Leopoldina, às 18 horas da terça-feira.

Referências

Ahad, M. A., Paiva, S., Tripathi, G., and Feroz, N. (2020). Enabling technologies and sustainable smart cities. *Sustainable cities and society*, 61:102301.

Alablani, I. and Alenazi, M. (2020). EDTD-SC: An IoT sensor deployment strategy for smart cities. *sensors*, 20(24):7191.

Chang, X. and Cui, H. (2021). Distributed storage strategy and visual analysis for economic big data. *Journal of Mathematics*, 2021:3224190.

Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, 19(2):171–209.

Clarindo, J. P., Castro, J. P. d. C., and Aguiar, C. D. d. (2021). Combining fog and cloud computing to support spatial analytics in smart cities. *Journal of Information and Data Management-JIDM*, 12(4):342–360.

de Carvalho Castro, J. P., Chaves Carniel, A., and Dutra de Aguiar Ciferri, C. (2020). Analyzing spatial analytics systems based on hadoop and spark: A user perspective. *Software: Practice and Experience*, 50(12):2121–2144.

Hai, R., Quix, C., and Jarke, M. (2021). *Data lake concept and systems: a survey*. CoRR, abs/2106.09592.

Liu, S., Peng, L., Chi, T., and Wang, X. (2016). Research on multi-source heterogeneous data collection for the smart city public information platform. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 623–626. IEEE.

Massobrio, R., Nesmachnow, S., Tchernykh, A., Avetisyan, A., and Radchenko, G. (2018). Towards a cloud computing paradigm for big data analysis in smart cities. *Programming and Computer Software*, 44(3):181–189.

Mătăcuţă, A. and Popa, C. (2018). Big data analytics: Analysis of features and performance of big data ingestion tools. *Informatica Economica*, 22(2).

Meehan, J., Aslantas, C., Zdonik, S., Tatbul, N., and Du, J. (2017). Data ingestion for the connected world. In *CIDR*, volume 17, pages 8–11.

Panwar, A. and Bhatnagar, V. (2020). Scrutinize the idea of hadoop-based data lake for big data storage. *Applications of Machine Learning*, pages 365–391.

Pereira, D. A., Ourique de Morais, W., and Pignaton de Freitas, E. (2018). Nosql real-time database performance comparison. *International Journal of Parallel, Emergent and Distributed Systems*, 33(2):144–156.

Rathore, M. M., Paul, A., Hong, W.-H., Seo, H., Awan, I., and Saeed, S. (2018). Exploiting iot and big data analytics: Defining smart digital city using real-time urban data. *Sustainable cities and society*, 40:600–610.

Reinsel, D., Gantz, J., and Rydning, J. (2017). Data age 2025: The evolution of data to life-critical. *Don't Focus on Big Data*, 2.

Veiga, J., Expósito, R. R., Pardo, X. C., Taboada, G. L., and Tourifio, J. (2016). Performance evaluation of big data frameworks for large-scale data analytics. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 424–431. IEEE.