sbbd:40

Diversidade de Dados Meteorológicos da Cidade do Rio de Janeiro: Uma Proposta de Arquitetura de Armazenamento e Fluxo de Dados para Modelos de Previsão

Bruno L. Freitas[1], **Augusto J. M. da Fonseca**[2], **Eduardo Bezerra**[2], **Flávia C. Bernardini** [1], **Mariza Ferro**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)

[2]Centro Federal de Educação Tecnológica do Rio de Janeiro (CEFET/RJ)

brunolf@id.uff.br

{mariza, fcbernardini}@ic.uff.br

ebezerra@cefet-rj.br, augusto.fonseca@aluno.cefet-rj.br

Abstract. *Precipitation nowcasting is an essential component of early warning systems and consecutive actions within crisis management for extreme weather events in urban areas. This article presents the work in progress for the collection, description of features, preparation, and use of multiple meteorological observations as data sources available for the city of Rio de Janeiro. The challenge is to bring together different sources, which are available openly or under restrictions of cooperation agreements, from different portals and websites, under the responsibility of different owners in different spheres (municipal, state and federal), with the objective of developing a data lake that brings them all together and can be used for the development of forecasting models.*

Resumo. *A precipitação nowcasting é um componente essencial dos sistemas de alerta precoce e das ações consecutivas no âmbito da gestão de crises para eventos climáticos extremos em áreas urbanas. Este artigo apresenta os trabalhos em andamento para a coleta, descrição de atributos, preparação e uso de múltiplas observações meteorológicas como fontes de dados disponíveis para a cidade do Rio de Janeiro. O desafio é reunir diferentes fontes, disponíveis abertamente ou sob restrições de acordos de cooperação, de diferentes portais e sites, sob a responsabilidade de diferentes proprietários em diferentes esferas (municipal, estadual e federal), com o objetivo de desenvolver um data lake que reúna todos e possa ser usado para o desenvolvimento de modelos de previsão.*

Introdução

Eventos climáticos e clima extremo podem afetar severamente sistemas naturais e humanos [Salles Civitarese et al. 2021], podendo causar danos materiais e perda de vidas. O aumento da frequência e da intensidade destes eventos, como a precipitação intensa, deve-se às mudanças climáticas detectadas em diferentes partes do mundo, de acordo com o Painel Intergovernamental sobre Mudanças Climáticas [Stocker et al. 2014].

A precipitação é o elemento meteorológico fundamental para definir o clima de uma região, sendo essencial para definir os desastres naturais mais frequentes

[Wanderley and Bunhak 2016]. No caso da cidade do Rio de Janeiro, desastres causados pelas chuvas na cidade do Rio de Janeiro mostram que a destruição faz parte do cotidiano da cidade há séculos [Marzban and Sandgathe 2006].

Neste contexto, há a necessidade de melhorar a capacidade de monitorar, predizer e emitir alertas de deslizamento de terra com antecedência para minimizar os danos relacionados e proteger vidas humanas e propriedades. No entanto, apesar dos esforços para melhorar a precisão da predição de eventos fortes, os resultados do Alerta Rio, um sistema do departamento do município responsável por lançar alertas de eventos extremos, precisam de melhorias. [Ferro et al. 2022] apresenta informações de um estudo interno realizado por um terceiro em que foram analisados 168 alertas de chuva, de fevereiro de 2019 a maio de 2019, indicando uma clara necessidade de melhorar as predições meteorológicas extremas em áreas urbanas e, em particular, para a cidade do Rio de Janeiro.

Neste contexto, tem-se o Projeto RioNowcast. Ele consiste em um grupo de estudantes e professores que buscam contribuir para esta questão de eventos climáticos extremos, sendo mais específico, para o caso de precipitações extremas. Este projeto busca realizar um levantamento de dados da cidade do Rio de Janeiro. Além disso, entender a ocorrência destes eventos, quais dados seriam relevantes para a predição e o desenvolvimento de um modelo de predição. Além disso, o conhecimento obtido ser tornará público visto que isso trará um grande benefício para a sociedade.

Entretanto, existe um desafio anterior ao desenvolvimento dos modelos de predição que é a coleta e a preparação de dados relevantes que permitam aprender sobre as condições que desencadeiam um evento de chuva extrema na cidade e permitam predizer com precisão este tipo de chuva.

Em relação à preparação de dados, os desafios envolvem a falta de dados processados disponíveis para conduzir os estudos necessários e incorporar várias fontes de dados para modelagem e treinamento de IA. O treinamento envolve aprender com várias fontes de dados meteorológicos em diferentes representações e granularidades (por exemplo, com diferentes grades espaciais e temporais) e poucos dados (poucos eventos extremos em uma longa série histórica). Outro desafio são os valores ausentes (por exemplo, em dados de precipitação devido a falhas nos sensores de captura de dados). Ainda assim, com relação à preparação dos dados, outro desafio é a falta de dados com curadoria para conduzir os estudos necessários. Embora a maioria das fontes de dados esteja disponível publicamente, há uma grande dificuldade em ter acesso a todos os dados. Além disso, as múltiplas fontes de dados estão espalhadas por diferentes aplicativos, sites ou repositórios e pertencem a diferentes proprietários em diferentes esferas do país (municipal, estadual e federal), o que torna o desenvolvimento de novas estratégias e abordagens para previsão e alertas muito mais difícil.

Logo, baseados nestas motivações, o objetivo deste trabalho é coletar e reunir em um só repositório as múltiplas fontes de dados sobre as observações meteorológicas na cidade do Rio de Janeiro, descrever cada uma delas e o conjunto de medidas disponíveis em cada uma delas. Elas estão sendo organizadas e neste artigo apresentamos resumidamente esta diversidade de dados, a complexidade da coleta e a proposta de um *data lake* como uma arquitetura de armazenamento e fluxo de dados para disponibilizar publicamente todo este conjunto de dados.

Diversidade de Fontes Dados do Projeto RioNowcast

Na tentativa de predizer e entender os eventos de chuva extrema na cidade do Rio Janeiro, múltiplas fontes de dados foram consideradas. Elas envolvem observações e medições de vários sensores os quais fornecem dados sobre as condições climáticas locais da cidade. Essas fontes de dados e seus provedores incluem o seguinte:

Trinta e três estações meteorológicas do Centro de Operações Rio (COR), as quais estão localizadas na cidade do Rio de Janeiro. A maioria dessas estações é pluviométrica e mede os níveis de precipitação, enquanto oito são meteorológicas e coletam dados de temperatura, umidade, vento e pressão. A resolução temporal dessas estações é de quinze minutos. Os dados serão coletados considerando as séries temporais de todas as estações quando os níveis de precipitação forem maiores ou iguais a 25 mm/h.

Dados de dois radares meteorológicos localizados em Guaratiba e Macaé do Instituto Estadual do Meio Ambiente (INEA). Cada um deles produz dados a cada cinco minutos. Os dois radares estão localizados no Pico do Couto e Sumaré e são de responsabilidade do Alerta-Rio.

Bóias meteorológicas do Sistema Brasileiro de Monitoramento Costeiro (SiMCosta) e da Marinha do Brasil as quais coletam variáveis como ponto de orvalho, temperatura da água, pressão, umidade, velocidade e direção do vento a cada hora.

Dados de satélite do Centro de Previsão do Tempo e Estudos Climáticos/Instituto Nacional de Pesquisas Espaciais (CPTEC/INPE) com os quais é possível coletar dados de atividade eletromagnética, aerossóis e vegetação.

Um balão meteorológico da Aeronáutica para coletar dados meteorológicos (umidade, temperatura, pressão do ar) em diferentes altitudes duas vezes ao dia.

Observa-se a existência de várias fontes de dados espaço-temporais. Uma fonte de dados espaço-temporal fornece medições que têm componentes no espaço e no tempo, sendo uma sequência de observações feitas em intervalos de tempo regulares [Jitkajornwanich et al. 2020]. O intervalo entre duas observações define a resolução temporal da fonte de dados.

Para que estas diferentes fontes de dados espaço-temporal, com diferentes grades temporais e espaciais, possam ser utilizadas para o treinamento de modelos de previsão uma etapa de pré-processamento deverá ser realizada para a sua integração. Estas fontes de dados devem ser conciliadas com relação às resoluções espaciais e temporais para formar uma fonte de dados agregada na qual cada elemento contém a união de todas as variáveis preditoras nas fontes de dados do elemento. Porém, a integração de fontes de dados com diferentes grades espaciais é desafiadora. Por exemplo, as estações meteorológicas estão em 33 locais no Rio, mas as estações de balões e boias estão localizadas em apenas um local como mostrado na Figura 1. Além disso, como pode ser observado pela descrição dos dados coletados, existem diferentes intervalos de medições entre certos equipamentos trazendo também uma resolução temporal variada entre as fontes de dados.

Outros desafios são os dados faltantes e a escassez inerente de dados de precipitação extrema. Ainda, alguns equipamentos/sensores foram instalados e começaram a coletar dados em diferentes datas, dificultando ainda mais a integração destes dados. Além disso, cada equipamento apresenta suas particularidades com relação

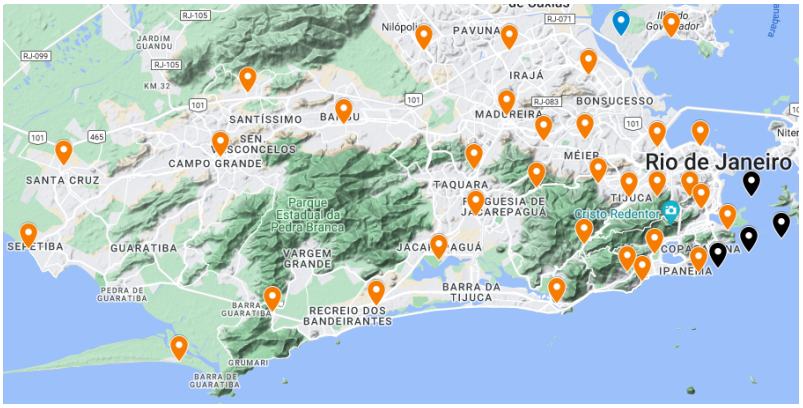

Figura 1. Fontes de dados na cidade do Rio de Janeiro. Estações meteorológicas, estação de balão e bóias indicados, respectivamente, pelos marcadores laranja, azul e preto.

a como a informação é disponibilizada e como ele apresenta dados faltantes. Assim, é possível observar a complexidade para a coleta, o processamento e a integração de todo esse conhecimento para que análises sejam realizadas e um modelo para predição de eventos extremos seja desenvolvido.

Arquitetura de Armazenamento e Fluxo de Dados

Data lake é uma abordagem de armazenamento de dados que permite implementar um repositório de arquivos estruturados, semi-estruturados e não estruturados, de diferentes fontes e formatos, de forma bruta e processada. Sua principal vantagem em relação a outras abordagens (*e.g.* bancos de dados relacionais, *data warehouses*) reside na capacidade de armazenar dados de forma flexível. Tais características tornam o *data lake* a abordagem ideal para o armazenamento e gestão dos dados apresentados na Seção anterior. Existem diversas tecnologias para a implementação de um *data lake* e, dentre elas, o *MinIO*[1] foi escolhido para o projeto *RioNowcast* devido à sua completa lista de recursos e licença *open source*.

Em geral, um projeto de *data science* segue um processo de desenvolvimento que pode iniciar na coleta dos dados, passando pelas etapas de análise exploratória e pré-processamento e finalizando com a construção e refinamento dos modelos. Tal processo é normalmente cíclico, já que os resultados produzidos por etapas posteriores podem gerar *insights* para melhoria no processamento de etapas anteriores ou até mesmo a detecção de erros de processamento. Visando a otimização de tal processo cíclico, o *data lake* do projeto *RioNowcast* foi projetado com uma arquitetura bem definida de *buckets* e fluxo de dados que permite reexecutar o processamento dos dados a partir de pontos específicos do fluxo e de forma segmentada em relação aos fornecedores e equipamentos. A arquitetura implementada é apresentada na Figura 2 e detalhada a seguir.

As etapas de processamento definidas no fluxo de dados são: *i) extração*; *ii) transformação*; *iii) pré-processamento* e; *iv) integração*. Para cada etapa do fluxo, para cada fornecedor e para cada equipamento é implementado um *script* de processamento separado. O resultado produzido por cada *script* é armazenado separadamente em um

[1]https://min.io/

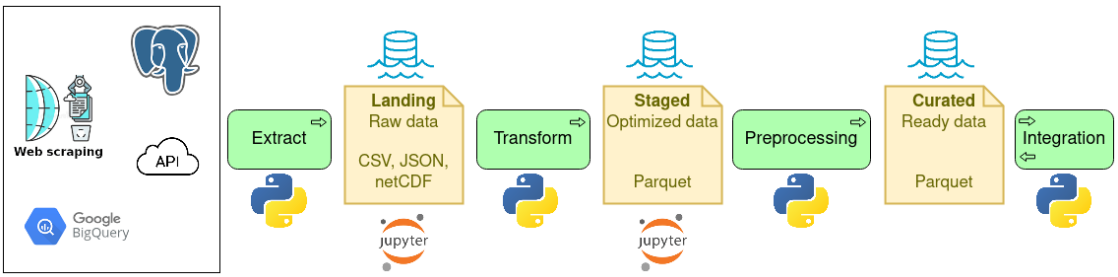

Figura 2. Arquitetura de armazenamento e fluxo de dados do *data lake* implementado para o projeto *RioNowcast*.

dos três *buckets*, sendo eles: *i) bucket landing*; *ii) bucket staged* e; *iii) bucket curated*. Tal estratégia permite reexecutar o processamento de forma segmentada e a partir de pontos específicos do fluxo, evitando o retrabalho que pode ser computacionalmente caro no caso de equipamentos que geram grandes volumes diários de dados. A exceção fica para o *script* final de *integração* que é justamente a etapa final de construção do *dataset* integrado e pronto para ser dado como entrada para uma rede neural de aprendizado.

O *script* de *extração* coleta os dados disponibilizados pelos fornecedores de acordo com as tecnologias disponibilizadas (e.g. *API*, *webscraping*). Os dados coletados são armazenados no *bucket landing* no formato mais próximo possível do original (*e.g.* CSV, JSON, netCDF). A exceção fica para os casos em que, por restrições de infraestrutura do projeto e por alta cobertura territorial do equipamento, o armazenamento do dado original implica em uma grande demanda de espaço de armazenamento. Como exemplo, dados de satélite podem cobrir uma área territorial a nível de continente e, nesse caso, os dados são filtrados para um limite de 250 km a partir do centroide da área de estudo do projeto *RioNowcast*.

Os dados armazenados no *bucket landing* sofrem uma pré-análise a fim de apoiar as decisões para a implementação do *script* de *transformação* dos dados. Nessa pré-análise são identificadas as variáveis disponíveis e quais otimizações podem ser aplicadas aos dados. Em seguida é implementado o *script* de *transformação* dos dados otimizando-os para o formato *Parquet*[2]. Os dados otimizados são armazenados no *bucket staged* com os respectivos tipos convertidos (*e.g. datetime, string, numeric*). Idealmente essa etapa não deve ser responsável por qualquer tipo de filtro ou agregação de dados. No entanto, devido às restrições de infraestrutura do projeto, exceções foram implementadas para os dados de radar e satélite por conta da alta resolução espacial e temporal destes equipamentos.

Em seguida é realizada a análise exploratória dos dados armazenados no *bucket staged*. A referida análise apoia as decisões para a implementação do *script* de *pré-processamento*. Uma vez que os *scripts* são implementados separadamente para cada fornecedor/equipamento, são aplicadas somente as operações que independem da integração dos dados. Como exemplos de operações processadas nesta etapa temos: *i)* remoção de duplicatas; *ii)* remoção de dados inconsistentes; *iii)* padronização de unidades de medida; *iv)* padronização de nomenclatura de variáveis; *v)* padronização do *timezone*; *vi)*

[2]https://parquet.apache.org/

transformação de variáveis cíclicas; *vii)* geração de *features* temporais e; *viii)* indexação por latitude e longitude. Os dados processados nesta etapa são armazenados no *bucket curated* ainda separados por fornecedor/equipamento. As operações de padronização de unidades de medida e nomenclatura de variáveis são particularmente importantes para viabilizar a correta integração dos dados na etapa seguinte de processamento.

A partir deste ponto do fluxo de processamento, os dados armazenados no *bucket curated* ficam disponíveis para a integração. Uma vez que os dados foram processados separadamente por fornecedor/equipamento, diferentes integrações podem ser processadas por meio da seleção de subconjuntos destes. Nesta última etapa são aplicadas as operações dependentes da integração dos dados, dentre elas: *i)* agregação espacial e temporal; *ii)* normalização e; *iii)* imputação de dados. O *dataset* resultante da integração é armazenado no próprio *bucket curated* e fica disponível para ser dado como entrada para uma rede neural de aprendizado.

Considerações Finais

A coleta e disponibilização de todas as fontes de dados preparadas não é uma tarefa trivial como demonstrado resumidamente neste artigo. Assim, a disponibilização de um *data lake*, como repositório de todas essas fontes de dados, é uma solução que trará benefícios para a comunidade cientifica e possibilitará um melhor entendimento de precipitações extremas, permitindo predições mais precisas.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo suporte financeiro e ao INEA, COR e Alerta-Rio por fornecerem os dados e cooperação no trabalho.

Referências

Ferro, M., Bezerra, E., Ogasawara, E., Moraes, N., and Porto, F. (2022). Towards a definition for extreme weather events in rio de janeiro city. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 181–186. SBC.

Jitkajornwanich, K., Pant, N., Fouladgar, M., and Elmasri, R. (2020). A survey on spatial, temporal, and spatio-temporal database research and an original example of relevant applications using sql ecosystem and deep learning. *Journal of Information and Telecommunication*, 4(4):524–559.

Marzban, C. and Sandgathe, S. (2006). Cluster analysis for verification of precipitation fields. *Weather and Forecasting*, 21(5):824–838.

Salles Civitarese, D., Szwarcman, D., Zadrozny, B., and Watson, C. (2021). Extreme precipitation seasonal forecast using a transformer neural network. *arXiv e-prints*, pages arXiv–2107.

Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M. (2014). Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of ipcc the intergovernmental panel on climate change.

Wanderley, H. and Bunhak, A. (2016). Alteration in precipitation and number of days without rain in the southern region of rio de janeiro state. rev brasil geog físic. 9. 7. 2341–2353.