MoreData: Enriquecimento Semântico para Grandes Volumes de Dados Geolocalizados*

Germano B. dos Santos[1], Leonardo J. A. S. Figueiredo[2], Fabrício A. Silva[1], Antonio A. F. Loureiro[2]

[1]Laboratório de Inteligência em Sistemas Pervasivos e Distribuídos (NESPeD-Lab)
Universidade Federal de Viçosa - Campus Florestal (UFV – CAF)

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

germano.santos@ufv.br, leonardo.alves@dcc.ufmg.br

fabricio.asilva@ufv.br, loureiro@dcc.ufmg.br

Abstract. *The volume of collection and availability of geolocated data has increased significantly. Such data can be used to enrich databases in different contexts to extract new information and enhance the results of various research. However, in most cases, raw geospatial data does not have semantic relevance, and the enrichment task with this type of data can be expensive. In this context, tools that facilitate such a task, such as MoreData, are essential to reduce the time spent and allow resources to be better allocated in other steps, such as creating models. However, in its first version, MoreData was not prepared to receive large volumes of data efficiently. Thus, the current work presents an update of MoreData to enrich databases with millions of data. With this update, it was possible to make MoreData about 2,500 times faster and use it for large volumes of data.*

Resumo. *O volume da coleta e disponibilização de dados geolocalizados aumentou significativamente nos últimos anos. Este tipo de dado pode ser utilizado para enriquecer bases em diferentes contextos para extrair novas informações e potencializar resultados de diversas pesquisas. Contudo, o dado geolocalizado bruto na maioria dos casos não possui relevância semântica e a tarefa de enriquecimento com este tipo de dado pode ser bastante dispendiosa. Nesse contexto, ferramentas que facilitam tal tarefa, como o MoreData, são fundamentais para diminuir o tempo gasto e permitir que os recursos sejam melhor alocados em outras etapas, como a criação de modelos. Entretanto, em sua primeira versão o MoreData não foi preparado para receber grandes volumes de dados de maneira eficiente. Assim, o trabalho atual apresenta uma atualização do MoreData para enriquecer bases com milhões de dados. Com essa atualização foi possível tornar o MoreData cerca de 2.500 vezes mais rápido e utilizá-lo para grandes volumes de dados.*

*https://www.youtube.com/watch?v=GzWpCl6OwC0

1. Introdução

Nos últimos anos, houve um aumento considerável na coleta e disponibilidade de dados geoespaciais que podem ser provenientes de dispositivos móveis e redes sociais [Domingues et al. 2020]. Por sua vez, estes dados geralmente são encontrados na forma de tupla $\langle latitude, longitude, timestamp \rangle$ ou como uma chave de identificação como $\langle código\_postal \rangle$. Porém, somente estas informações brutas não possuem um valor semântico, e consequentemente, são insuficientes nas tomadas de decisão. Logo, faz-se necessário enriquecer semanticamente estes dados brutos com fontes externas.

Entretanto, a tarefa de enriquecer os dados pode despender muitos recursos computacionais e humanos para ser realizada. Uma das soluções possíveis é utilizar ferramentas prontas, como [Gubert and Silva 2022] que facilitam a coleta e o enriquecimento de dados utilizando fontes externas de dados do *Google Places API*. Já [Rettore et al. 2020], exploram o enriquecimento de dados de tráfego de sistemas de transporte. Afim de trabalhar genericamente, sem considerar contextos específicos como os trabalhos mencionados, o *MoreData* [Figueiredo et al. 2021] é um *framework* capaz de enriquecer dados georreferenciados utilizando quatro conectores diferentes: *OpenStreetMap* (OSM), *API*, banco de dados relacional e o *Elasticsearch*.

Em sua primeira versão, o *MoreData* pôde ser útil em trabalhos como [Figueiredo et al. 2022] onde enriquecem localizações brutas de 46.439 usuários móveis com dados de locais do *OpenStreetMap* para gerar perfis de mobilidade considerando diferentes perspectivas. Outro trabalho beneficiado foi [Souza et al. 2022] no qual, para obter mais assertividade na recomendação de aplicativos personalizada, enriqueceram as localizações aproximadas de 87.903 usuários com dados de setores censitários utilizando o *MoreData*. No entanto, a primeira versão do *framework* trabalha somente com arquivos do tipo *JSON*, que restringe a escalabilidade por não ser preparado para enriquecer grandes volumes de dados. Além disso, os volumes dos dados nos trabalhos que utilizaram o framework não foram suficientes para validar sua robustez em relação ao tamanho da entrada. Ainda, no trabalho original não foram apresentados testes de eficiência considerando memória e tempo gasto conforme a variação dos tamanhos de entrada. Por fim, até o momento, o código fonte do *MoreData* não havia sido disponibilizado publicamente.

Portanto, o objetivo deste trabalho é melhorar o *MoreData* gerando uma nova versão disponível publicamente[1] capaz de suportar grandes volumes de dados. Para isso, o primeiro passo foi melhorar a manipulação dos dados testando dois diferentes tipos de representação de dados: *GeoPandas DataFrame* e *Dask GeoPandas DataFrame*. Além disso, será apresentada uma análise quantitativa do desempenho comparando o uso de memória e o tempo gasto de acordo com a variação do tamanho da entrada ao enriquecer dados utilizando *JSON*, *GeoPandas Dataframe* e o *Dask GeoPandas Dataframe*.

O restante deste trabalho está organizado da seguinte forma. Na Seção 2 é apresentada a versão inicial do *MoreData* e a melhoria realizada. Uma demonstração é apresentada na Seção 3. Por fim, a Seção 4 apresenta as conclusões e trabalhos futuros.

[1]https://github.com/NESPEDUFV/more-data

2. *MoreData*

2.1. Descrição

O *MoreData* foi proposto para facilitar o enriquecimento semântico de dados geoespaciais com fontes externas. Na Figura 1 é apresentada uma visão geral dos componentes, sendo os módulos Enriquecedor e Conversor, os principais. O *framework* foi totalmente desenvolvido em *Python*.

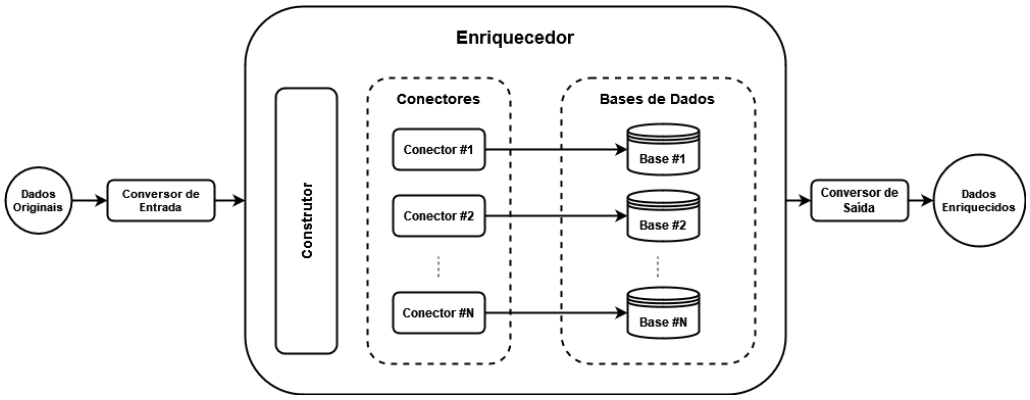


Figura 1. Visão geral do *MoreData* (adaptado de [Figueiredo et al. 2021]).

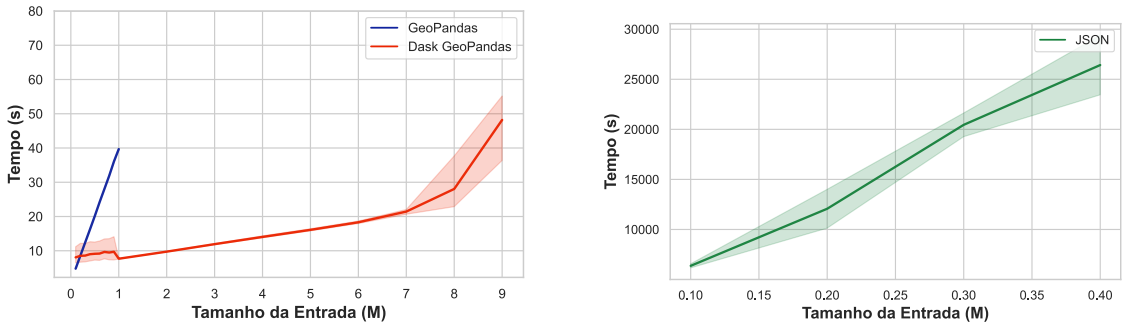
O componente principal do *MoreData* é o Enriquecedor que pode ser dividido em dois módulos menores: Conector e Construtor. O Conector já possui implementadas quatro conexões: *SQL*, *API*, *Elasticsearch* e *OpenStreetMap*. Para facilitar a criação de novos conectores, foi utilizado o padrão de projeto *Strategy*, logo é necessário somente implementar a interface e usar na classe Enriquecedor. O Construtor é o responsável por realizar o enriquecimento e foi desenvolvido utilizando o padrão de projeto *Builder*. Dessa forma, é possível enriquecer sequencialmente um dado bruto utilizando diferentes conectores, o que possibilita a combinação das estruturas.

Outro componente é o Conversor que tem como objetivo generalizar os suportes aos diversos tipos de dados que o ecossistema *Python* possui. Assim, dados fornecidos como entrada, como CSV e Parquet, podem ser convertidos para, por exemplo, o JSON. Similarmente, os dados enriquecidos podem ser convertidos para o mesmo tipo da entrada ou serem transformados para outro tipo suportado pelo *framework*.

2.2. Melhoria para Grandes Volumes de Dados

A fim de resolver o problema de escalabilidade, desenvolveu-se novas abstrações *GeopandasData* e *DaskGeopandasData* que incorporam as representações de dados *GeoPandas DataFrame* [Jordahl et al. 2019] e *DaskGeopandas DataFrame* [Van Den Bossche et al. 2023], respectivamente. A primeira abstração possibilita o enriquecimento de dados geoespaciais estruturados, e a segunda permite o enriquecimento em grande escala.

Para avaliar as novas alterações foram realizados testes comparando a primeira versão que utilizava somente *JSON*. O computador de testes no qual os enriquecimentos foram realizados é formado por uma CPU Intel Core i7 - 1165G7, que conta com 8 núcleos, 2 *threads* e frequência máxima de 4.70 GHz. Sua memória RAM é composta por um módulo de 16 GB DDR4 3200 MHz.

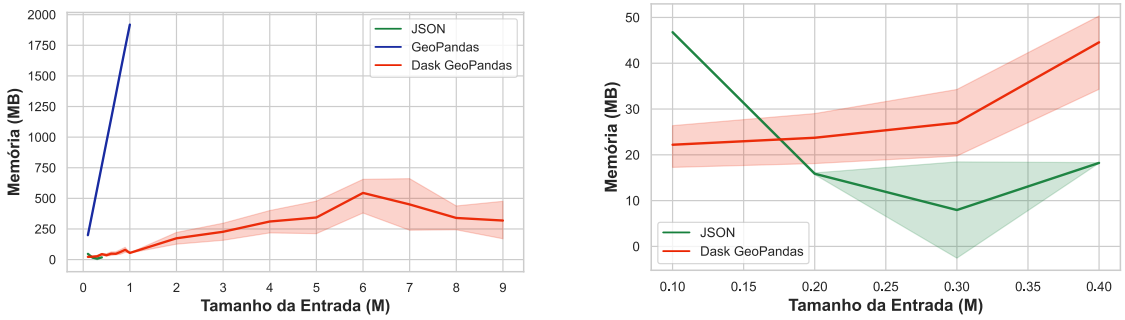


(a) *GeoPandas* e *Dask GeoPandas*. **(b)** *JSON*

Figura 2. Tempo gasto para enriquecer cada tamanho de entrada.

Para os testes foram utilizados dados reais com cerca de 11 milhões de localizações de usuário móveis de São Paulo fornecidos por uma empresa parceira. Foram utilizados tamanhos de entrada crescentes, começando de 100 mil até 1 milhão, aumentando de 100 em 100 mil, depois de 1 milhão, aumentando de 1 em 1 milhão até alcançar os 9 milhões. Contudo, o alto tempo gasto pelo *JSON* forçou sua parada em 400 mil registros, já o *GeoPandas* teve seu teste finalizado com 1 milhão de registros pela alta quantidade de memória utilizada. Além disso, restaurantes de São Paulo foram coletados para que as localizações fossem enriquecidas corretamente. Ademais, o processamento a partir do conector *OSMPlacesConnector* com raio igual a $25\ m$.

A Figura 2 apresentam os tempos gastos no enriquecimento conforme o tamanho da entrada. Com os resultados é possível observar que o *Dask Geopandas* obteve os melhores resultados gastando aproximadamente 50 segundos para enriquecer 9 milhões de registros, enquanto o *JSON* requer cerca de 7 horas para enriquecer somente 400 mil registros.



(a) *GeoPandas* e *Dask GeoPandas*. **(b)** *JSON* e *Dask GeoPandas*.

Figura 3. Memória gasta para enriquecer cada tamanho de entrada.

Em relação à memória gasta nos enriquecimentos a Figura 3 apresenta os resultados dos testes realizados. Os resultados mostram que o *Dask Geopandas* possui os melhores resultados gastando cerca de 4 vezes menos para enriquecer 9 vezes mais dados que o *Geopandas*. Apesar da solução original consumir pouca memória, ela só foi capaz de enriquecer, no máximo, 400 mil registros.

3. Demonstração

O enriquecimento semântico pode ser útil em diferentes contextos para extrair novos conhecimentos dos dados. Considerando um cenário no qual se tenha as localizações e os valores dos imóveis da região oeste da Rússia[2], deseja-se enriquecer com comércios contidos em um certo raio para correlacionar com o preço do imóvel[3]. Essa base de dados foi escolhida para a demonstração do *framework* por ser pública e possuir cerca de 11 milhões de registros, uma quantidade significativa para testar a escalabilidade dos conectores. Além disso, mostra a generalização e a facilidade do uso do *MoreData*, visto que os testes foram feitos com dados reais privados e a demonstração com dados públicos.

Para a coleta dos comércios foi utilizado o *OpenStreetMap* aplicado na mesma região dos dados da Rússia para o enriquecimento. Para simplificação, somente os dados de identificação, categoria, nome e geometria foram utilizados. Como o *Dask Geopandas* obteve os melhores resultados, ele foi selecionado para o exemplo de código abaixo.

```
1  from moredata.enricher.osm.osm_places_connector import OSMPlacesConnector
2  from moredata.models import DaskGeopandasData
3  from moredata.enricher.enricher_builder import EnricherBuilder
4  import pandas, geopandas, numpy
5  from distributed import Client, LocalCluster
6  dask.config.set("distributed.nanny.environ.MALLOC_TRIM_THRESHOLD_":0})
7  cluster = LocalCluster(n_workers=2, threads_per_worker=1, memory_limit='4GB')
8  client = Client(cluster)
9  df = pandas.read_csv("russia-real-estate.csv").reset_index()
10 gdf = geopandas.GeoDataFrame(df, geometry=geopandas.points_from_xy(x=df.geo_lon, y=df.
       geo_lat))
11 data = DaskGeopandasData.from_geodataframe(gdf,npartitions=128)
12 restaurant_enricher = OSMPlacesConnector(
13     files=["restaurant.csv"], radius=1000, geometry_intersected=True
14 )
15 df_enriched = restaurant_enricher.enrich(data, join_how="inner")
16 df_enriched = df_enriched.data.compute()
```

Código 1. Exemplo do *MoreData* com Dask GeoPandas

Na linha 10 do código, um *cluster* local do *Dask* é criado, com 2 *workers* com limite de memória de 4GB para cada *worker*. Note que esse *cluster* é básico, porém é possível criar um *cluster* em um ambiente de nuvem com mais recursos computacionais. Na linha 15, a variável *restaurant_enricher* utiliza o *OSMPlacesConnector* que possui a seguinte semântica: dado um imóvel presente na base utilizada, quais locais em um raio de 1 km existem? Portanto, a saída desse enriquecimento será uma base que possui as características do imóvel, como preço e localização, e os atributos dos restaurantes: identificação, categoria, nome e a geometria.

Para esse experimento, a base de dados foi reduzida para 1.5 milhão de registros, visto que o objetivo dessa seção é demonstrar o *MoreData*. O resultado desse enriquecimento tem como saída uma base de dados com cerca de 4 milhões de registros. Além disso, esse processamento teve como tempo total de execução de 2 minutos. Note que um registro pode ter mais de um restaurante dentro de um raio de 1 km, por isso a base de dados de saída é maior que a base de dados de entrada. Ademais, esse enriquecimento é diferente do processo feito na seção 2.2, pois o raio é 40 vezes maior, aumentando a quantidade de registros da base de saída.

[2]https://www.kaggle.com/datasets/mrdaniilak/russia-real-estate-2021

[3]https://colab.research.google.com/drive/1PeYUOHrwyrH4Ly27bzwGcauB6Gx7YvM8?usp=sharing

4. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma atualização do *MoreData* para enriquecer grandes volumes de dados. Os resultados apresentados mostraram que as atualizações para suportar *GeoPandas* e *Dask GeoPandas* são mais eficientes, principalmente o último que conseguiu ser melhor em termos de memória e tempo gasto no enriquecimento. Além disso, com a demonstração é possível perceber a facilidade de uso do framework que com poucas linhas de código é capaz enriquecer milhões de dados.

Esta pesquisa possui muitas oportunidades de expansão do framework para serem exploradas. Assim, como trabalhos futuros, pretende-se implementar novos conectores, como informações de segurança da região visitada pelo usuário. Além disso, aprimorar os conectores já existentes para enriquecerem com mais dados, como utilizar o *OpenStreetMap* para caracterizar as funcionalidades da região visitada pelo usuário. Por fim, realizar testes comparando a eficiência com outros *frameworks* da literatura.

Referências

Domingues, A., Silva, F., Santos, L., Souza, R., Coimbra, G., and Loureiro, A. A. F. (2020). Dados geoespaciais: Conceitos e técnicas para coleta, armazenamento, tratamento e visualização. *Sociedade Brasileira de Computação*.

Figueiredo, L., Santos, G., Souza, R., Silva, F., Silva, T., and Loureiro, A. (2022). Score: um serviço de classificação de usuários móveis com base em seus aplicativos e suas cidades. In *Anais do XIV Simpósio Brasileiro de Computação Ubíqua e Pervasiva*, pages 21–30, Porto Alegre, RS, Brasil. SBC.

Figueiredo, L. J. A. S., dos Santos, G. B., Souza, R. P. P. M., Silva, F. A., and Silva, T. R. M. B. (2021). Moredata: A geospatial data enrichment framework. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '21, page 419–422. Association for Computing Machinery, New York, NY, USA.

Gubert, F. and Silva, T. (2022). Google places enricher: A tool that makes it easy to get and enrich google places api data. In *Anais Estendidos do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 91–94, Porto Alegre, RS, Brasil. SBC.

Jordahl, K., Van den Bossche, J., Wasserman, J., McBride, J., Gerard, J., Fleischmann, M., Tratner, J., Perry, M., Farmer, C., Hjelle, G. A., et al. (2019). geopandas/geopandas: v0.6.0. *Zenodo*.

Rettore, P. H. L., Santos, B. P., Rigolin F. Lopes, R., Maia, G., Villas, L. A., and Loureiro, A. A. F. (2020). Road data enrichment framework based on heterogeneous data fusion for its. *IEEE Transactions on Intelligent Transportation Systems*, 21(4):1751–1766.

Souza, R. P., Figueiredo, L. J., Silva, M. P., Silva, F. A., Silva, T. R., and Loureiro, A. A. (2022). Investigating the impact of demographic and device information in the recommendation of mobile applications. *Journal of Internet Services and Applications*, 12(1):21–32.

Van Den Bossche, J., Fleischmann, M., Statham, T., Daniel Jahn (Dahn), Augspurger, T., Signell, J., Gadomski, P., Bell, R., Lumnitz, S., Zaidi, A. A., Bunt, F., Rose, I., Truong, I., Bourbeau, J., Baker, J., Morris, M., Hagen, R., RichardScottOZ, and Bernardpazio (2023). geopandas/dask-geopandas: v0.3.1.