ANGELO CESAR COLOMBINI

RELATÓRIO FINAL DE PÓS-DOUTORADO

Extração de Conhecimento em Big Data – Parte II

Relatório de Pós-doutorado realizado na Universidade Estadual Paulista (UNESP), DCCE – IBILCE – UNESP / São José do Rio Preto.

Supervisor(a): Prof. Dr. Carlos Roberto Valêncio

No do processo: 5659

São José do Rio Preto
2026

1. Identificação do Estágio Pós-Doutoral

Título do projeto	Extração de Conhecimento em Big Data – Parte II
Pesquisador	Prof. Dr. Angelo Cesar Colombini
Vínculo institucional	Departamento de Engenharia Elétrica (TEE) – Universidade Federal Fluminense (UFF)
Supervisor	Prof. Dr. Carlos Roberto Valêncio
Instituição supervisora	Departamento de Ciência da Computação e Estatística (DCCE) – IBILCE – UNESP, São José do Rio Preto
Grupo de pesquisa	Grupo de Banco de Dados (GBD) – IBILCE – UNESP
Período de realização	Julho de 2024 – Abril de 2026
Documento	Relatório Final – Encerramento do Estágio Pós-Doutoral

2. Introdução

Ambientes distribuídos, aplicações de IoT e bases de monitoramento contínuo produzem volumes de dados cujas características: escala, heterogeneidade e redundância estrutural, criam gargalos específicos para algoritmos de mineração projetados originalmente para conjuntos de dados centralizados. Dois problemas concretos motivam o presente trabalho: o gargalo computacional imposto pela redundância de coordenadas em algoritmos de *clustering* espacial e a ausência de interfaces que permitam a não-especialistas consultar e visualizar dados sem depender de SQL. A escala desse problema é evidenciada tanto pela proliferação de sensores em redes de monitoramento quanto pela necessidade de *frameworks* de processamento distribuído como o Apache Spark para lidar com os volumes resultantes (HASAN; XING; MAGRAY, 2022; THEODORAKOPOULOS; KARRAS; KRIMPAS, 2025).

A pesquisa em extração de conhecimento em Big Data tem avançado em três frentes de particular relevância para o presente trabalho: algoritmos eficientes para processamento distribuído, mecanismos de pré-processamento voltados à qualidade dos dados e interfaces que permitam a usuários não especialistas explorar bases complexas com menor dependência de programação. No eixo de *clustering* espacial, a literatura recente evidencia esforços de paralelização e otimização de algoritmos como DBSCAN e K-Means em ambientes MapReduce (AWAD; HEFNY, 2023; LANG; CHAI, 2023; WANG, 2025). No eixo de visualização orientada por linguagem natural, o avanço dos grandes modelos de linguagem (LLMs) e a disponibilidade de modelos de código aberto têm ampliado as possibilidades de

sistemas *text-to-SQL* e geração automática de visualizações (SHI et al., 2024; SAH et al., 2024; LEE et al., 2024).

O trabalho desenvolvido no âmbito do Grupo de Banco de Dados (GBD) da UNESP – IBILCE – São José do Rio Preto, sob orientação do Prof. Dr. Carlos Roberto Valêncio, configura a segunda etapa de um projeto contínuo de pesquisa voltado à extração de conhecimento em Big Data. A primeira etapa, encerrada em 2023, concentrou-se em métodos de limpeza de dados, detecção de *outliers* e redução de dimensionalidade. A presente etapa avançou em direção à otimização de algoritmos de mineração espacial em ambientes distribuídos e ao desenvolvimento de interfaces de visualização de dados baseadas em linguagem natural, dois eixos com forte aderência aos objetivos específicos estabelecidos no projeto.

3. Síntese do Projeto de Pesquisa Proposto

3.1 Problema de Pesquisa e Contextualização

O projeto de pesquisa proposto em julho de 2024 partiu do diagnóstico de que, apesar dos resultados obtidos na etapa anterior, persistiam lacunas relevantes no tratamento de dados espaciais em ambientes de alta densidade e na visualização intuitiva de resultados de mineração. No domínio dos algoritmos de *clustering* espacial, a elevada redundância de coordenadas fenômeno frequente em *datasets* de monitoramento biológico, ambiental e urbano (LI et al., 2025) impõe um gargalo computacional que *frameworks* de propósito geral, como Apache Spark MLlib e Apache Sedona, não resolvem de forma nativa (HASAN; XING; MAGRAY, 2022). A ausência de mecanismos nativos de pré-agregação coordenada nesses ambientes constitui a lacuna específica endereçada pelo eixo de *Data Mining* deste projeto.

No eixo de visualização, o problema central residia na distância entre a capacidade analítica dos sistemas de Big Data e a acessibilidade para usuários não especialistas. A criação manual de visualizações a partir de consultas SQL exige domínio técnico que a maioria dos usuários finais não possui. Interfaces de linguagem natural capazes de traduzir questões em linguagem corrente para consultas estruturadas e representações gráficas adequadas ao tipo de dado reduzem diretamente a barreira de entrada para análise exploratória de dados em organizações sem equipe técnica especializada — lacuna identificada na revisão bibliográfica, onde sistemas com combinação de modelos heterogêneos e suporte a tipos complexos de visualização são ainda escassos (LIU et al., 2021; YANG et al., 2024).

3.2 Objetivos do Projeto

Objetivo Geral: Investigar e ampliar o conhecimento nas técnicas e tecnologias de Big Data, com foco em Preparação de Dados, *Data Cleaning*, *Data Mining* e *Visual Data Mining*.

Objetivos Específicos:

(a) Explorar o desenvolvimento de estratégias para preparação de dados, visando automatizar e otimizar a limpeza e a preparação inicial de grandes volumes de dados.

(b) Aprimorar técnicas de Data Cleaning para identificação e correção de dados atípicos e incompletos, aumentando a qualidade e a confiabilidade dos *datasets*.

(c) Expandir as capacidades de Data Mining por meio do desenvolvimento e adaptação do algoritmo CHSMST+MR, utilizando paralelização e MapReduce para otimizar o processamento de dados espaciais em ambientes de Big Data.

(d) Contribuir para a etapa de Visual Data Mining por meio do estudo e criação de interfaces e ferramentas de visualização que facilitem a interpretação intuitiva de dados complexos.

3.3 Escopo Temático e Estratégia Metodológica Prevista

O projeto previu uma metodologia centrada no levantamento e atualização do estado da arte nos quatro eixos temáticos, seguida da produção de publicações científicas como produto central do período. Dentre os procedimentos programados constavam: implementação de técnicas de paralelização e MapReduce para otimização de algoritmos de mineração espacial; aprimoramento de técnicas de Data Cleaning para séries temporais e dados heterogêneos; exploração de ferramentas de Visual Data Mining como D3.js, Plotly e *frameworks* interativos; e testes de desempenho comparativos validando a eficácia das soluções desenvolvidas em relação ao algoritmo base CHSMST+MR (VALÊNCIO et al., 2016).

A exequibilidade do projeto foi sustentada pela infraestrutura do laboratório GBD–UNESP, pela base de algoritmos e simuladores disponíveis no DCCE–IBILCE, e pelos recursos computacionais da UFF — incluindo equipamentos para IoT, Smart Grids e cluster de alto desempenho. A essa base somou-se a colaboração prévia entre os dois grupos, com publicações conjuntas desde 2016 e metodologias de mineração já consolidadas no contexto do algoritmo CHSMST+.

4. Desenvolvimento das Atividades no Estágio Pós-Doutoral

4.1 Levantamento Bibliográfico e Atualização do Estado da Arte

A fase inicial do estágio dedicou-se ao levantamento sistemático e à atualização do estado da arte nos quatro eixos temáticos do projeto, com ênfase particular nas subáreas de *clustering* espacial em ambientes distribuídos e de interfaces de linguagem natural para visualização de dados.

No eixo de Data Mining espacial, foram estudados trabalhos recentes sobre algoritmos de *clustering* em *frameworks* distribuídos, incluindo abordagens baseadas em MapReduce, otimizações do K-Means e K-Medoids para dados massivos, implementações paralelas de DBSCAN e técnicas de particionamento espacial via R-trees e Quadtrees (AWAD; HEFNY, 2023; LANG; CHAI, 2023; WANG, 2025; LI et al., 2025). Essa revisão revelou uma lacuna específica: nenhum dos trabalhos existentes tratava a redundância de coordenadas: situação em que múltiplos registros compartilham posições geográficas idênticas, como um gargalo computacional passível de tratamento por pré-agregação distribuída antes da execução do núcleo de *clustering* (MANJULA; NARSIMHA, 2020; FAKIR; ELAYACHI; BTIS-SAM, 2023).

No eixo de *Visual Data Mining*, a revisão bibliográfica mapeou o estado das interfaces de linguagem natural para geração de visualizações, identificando que os estudos existentes se concentravam predominantemente em modelos de linguagem de grande porte (LLMs) acessados via API proprietária, com limitadas contribuições envolvendo modelos abertos, *fine-tuning* eficiente e combinação de modelos especializados em subtarefas distintas (SHI et al., 2024; ZHU et al., 2024; MAAMARI et al., 2024). Identificou-se também que a maior parte dos sistemas propostos na literatura suportava apenas visualizações básicas, sem cobertura de representações complexas como diagramas de Sankey, grafos de rede e mapas georreferenciados (LIU et al., 2021; YANG et al., 2024).

4.2 Desenvolvimento e Validação do Framework CHSMST+MR

O eixo de Data Mining espacial evoluiu em direção à proposta e à validação de uma otimização arquitetural do algoritmo CHSMST+ (VALÊNCIO et al., 2016), denominada CHSMST+MR. A contribuição central consistiu na introdução de uma camada de Redução Distribuída de Cardinalidade Espacial (*Distributed Spatial Cardinality Reduction* — DSCR), implementada como etapa de pré-agregação *MapReduce* no ambiente Apache Spark 3.5.0, anterior à execução dos núcleos iterativos de *clustering*. O princípio matemático subjacente à abordagem é direto: processar N registros em coordenadas idênticas é equivalente, do ponto de vista estatístico e topológico, a processar um único registro com peso N. A operação é, portanto, *lossless*: não há perda de informação nem distorção da topologia das hipersuperfícies calculadas pelo algoritmo original, nem da refinação pela Árvore Geradora

Mínima (MST). O ganho computacional decorre da redução drástica da cardinalidade efetiva do conjunto de dados antes das etapas de maior custo iterativo (VALÊNCIO et al., 2026).

A implementação foi realizada em Java, utilizando o ambiente NetBeans 13 e o banco de dados PostgreSQL 12.0 com a extensão PostGIS para tratamento de dados georreferenciados. O JavaSparkContext foi configurado no modo local[*], com 2.048 MB de memória por executor e particionamento dos RDDs em 24 partições (fator 2× o número de *threads* disponíveis), garantindo utilização eficiente dos 12 núcleos lógicos do *hardware* de teste (Intel Core i5-10400F, 16 GB RAM).

A validação experimental utilizou um *dataset* de ocorrências biológicas georreferenciadas da *National Academy of Natural Sciences*, contendo 50.000 registros com elevada redundância de coordenadas características propositalmente selecionadas para maximizar a relevância do cenário de alta densidade (GBIF.ORG USER, 2022). Os experimentos foram estruturados em três cenários, variando o número de atributos não espaciais incluídos na chave de agregação (1, 2 e 3 atributos), com três repetições cada, permitindo o cálculo do desvio padrão e a verificação da estabilidade dos resultados.

Os resultados demonstraram reduções médias de tempo de execução de 79,96% (1 atributo), 52,09% (2 atributos) e 22,03% (3 atributos) em relação ao CHSMST+ sem pré-agregação, com média geral de 51,36% de redução. O comportamento decrescente da redução em função do número de atributos é matematicamente previsível: quanto mais atributos compõem a chave de agregação, menor a probabilidade de coincidência completa entre registros, reduzindo a taxa de colisão e, consequentemente, a eficácia da compressão de cardinalidade.

4.3 Desenvolvimento da Interface de Linguagem Natural para Visualização

O eixo de Visual Data Mining resultou no desenvolvimento de uma interface de linguagem natural (*Natural Language Interface* — NLI) para geração automática de visualizações a partir de consultas em português, conectada a um banco de dados relacional real.

O componente PLM foi responsável pela tarefa de *Schema Linking*: dada uma questão em linguagem natural e o esquema do banco de dados, o modelo identifica as tabelas e colunas relevantes para a construção da consulta SQL. Três variantes baseadas na arquitetura BERT foram avaliadas para esta tarefa: BERTimbau base (110M parâmetros), BERTimbau *large* (330M parâmetros) e BERTugues. Todos os modelos foram ajustados com *fine-tuning* sobre o *dataset* Spider traduzido para o português (YU et al., 2018), formulando o problema como classificação binária (relevante/não relevante) para cada tabela e coluna do esquema.

O componente LLM foi responsável pela geração da consulta SQL a partir da questão e do esquema relevante identificado pelo PLM. O modelo Llama 3.1 (8B parâmetros, versão *instruct*) foi submetido a *fine-tuning* eficiente utilizando as técnicas LoRA e QLoRA (HU et al., 2023), reduzindo significativamente o custo computacional de treinamento. Adicionalmente, o modelo SQLCoder (7B parâmetros, derivado do CodeLlama) foi avaliado como *baseline* de comparação. O SQLCoder demonstrou desempenho superior ao Llama ajustado (65,6% versus 59,4% de acurácia) e foi selecionado para integração na interface final.

O módulo de visualização foi implementado como um algoritmo heurístico baseado na árvore de decisão de Healy e Holtz, que analisa as características do *dataset* resultante da execução SQL — número de colunas, tipos de dados (numérico, categórico, temporal, geográfico) — para recomendar e gerar automaticamente o tipo de visualização mais adequado. O sistema suporta 17 tipos de visualização, incluindo representações complexas como diagramas de Sankey, grafos de rede e mapas (ponto, bolha e conexão), superando o escopo da maior parte dos sistemas reportados na literatura (LIU et al., 2021; YANG et al., 2024; ZHU et al., 2024).

A interface foi validada com 60 questões distribuídas em três níveis de dificuldade (fácil, médio, difícil), conectadas a um banco de dados real de dicionários multilíngues (25 tabelas, 236 colunas). O BERTimbau *large* obteve acurácia de 91,67% no *Schema Linking*, com desempenho consistente entre os níveis de dificuldade. O SQLCoder obteve acurácia média de 75%, com desempenho satisfatório nos níveis fácil (85%) e médio (80%) e redução esperada no nível difícil (60%), associada principalmente à maior complexidade sintática das subqueries e *self-joins* (ZHANG et al., 2023; MAAMARI et al., 2024).

4.4 Aprofundamentos Metodológicos nos Eixos de Data Preparation e Data Cleaning

Os objetivos específicos (a) e (b) do projeto, relativos à preparação e limpeza de dados, foram endereçados de forma transversal ao longo do período, fundamentando metodologicamente os dois eixos de pesquisa principais em vez de gerar uma linha de produção científica independente. No desenvolvimento do CHSMST+MR, a etapa de pré-agregação DSCR constitui, em essência, uma forma especializada de limpeza estrutural dos dados: a eliminação de redundâncias coordenadas sem perda de informação é um caso específico de normalização de densidade em *datasets* espaciais. No desenvolvimento da NLI, a preparação dos dados para *fine-tuning* — incluindo a tradução do *dataset* Spider, a introdução controlada de ruído nos exemplos de treinamento do LLM e o processo de tokenização e formatação

estruturada (YU et al., 2018; DONG et al., 2023) — envolveu decisões metodológicas relevantes de engenharia de dados.

Essas contribuições transversais, embora não tenham resultado em publicações independentes, consolidaram competências nos domínios de pré-processamento distribuído e engenharia de dados para modelos de linguagem — competências diretamente aplicáveis aos desdobramentos futuros do projeto.

5. Resultados Alcançados

5.1 Resultados Científicos

Os principais resultados científicos do estágio concentram-se em duas frentes, cada uma com problema específico identificado na revisão bibliográfica, solução implementada e validação experimental documentada.

A primeira contribuição demonstra que o gargalo computacional imposto pela redundância de coordenadas em *datasets* espaciais de alta densidade pode ser resolvido de forma determinística e *lossless* por meio de uma camada de pré-agregação MapReduce, sem modificação da lógica de *clustering* subjacente. Esse resultado evidencia que ganhos de desempenho expressivos podem ser obtidos por reconfiguração arquitetural da representação dos dados, antes de qualquer otimização dos núcleos algorítmicos. A redução média de 51,36% no tempo de execução com pico de 79,96% no cenário de menor dimensionalidade de atributos, constitui evidência experimental robusta da eficácia da abordagem (VALÊNCIO et al., 2026).

No eixo de Visual Data Mining, foi demonstrada a viabilidade de sistemas NLI que combinam modelos de linguagem heterogêneos, cada um especializado em uma subtarefa distinta, para endereçar o problema completo de conversão texto–visualização. A acurácia de 91,67% no *Schema Linking* e de 75% na geração SQL, obtidas com modelos de código aberto executados localmente sem dependência de APIs externas, representa um resultado relevante para cenários em que privacidade dos dados, autonomia computacional e controle da infraestrutura são requisitos institucionais. A cobertura de 17 tipos de visualização — incluindo Sankey, grafos de rede e mapas georreferenciados, ausentes na maioria dos sistemas reportados — amplia o escopo de aplicabilidade do sistema para além dos casos de uso cobertos pela literatura existente (LIU et al., 2021; YANG et al., 2024).

5.2 Resultados Metodológicos

Do ponto de vista metodológico, o trabalho gerou contribuições relevantes em duas frentes. A formalização da camada DSCR como algoritmo independente (*Algorithm* 1 no artigo publicado) fornece uma abstração reutilizável aplicável a qualquer algoritmo de *clustering* iterativo que opere sobre dados com redundância de coordenadas não apenas ao CHSMST+, mas potencialmente a variantes de DBSCAN, K-Medoids e outros algoritmos da família. A análise de complexidade assintótica que justifica formalmente a transição de O(f(N)) para O(f(U)) + O(N), onde U << N em cenários de alta densidade, sustenta a generalização da proposta (VALÊNCIO et al., 2026).

No eixo de NLI, a metodologia de introdução controlada de ruído no conjunto de treinamento do LLM expandindo cada exemplo em quatro variações (original, com tabelas adicionais, com colunas adicionais, com colunas ausentes), demonstrou ser uma estratégia eficaz para aumentar a robustez do modelo em condições realistas de inferência, onde o PLM antecedente raramente produz um esquema perfeito (ZHANG et al., 2023). A mesma estratégia de expansão com ruído controlado é diretamente transferível a *pipelines* texto–SQL em dois estágios que utilizem qualquer PLM de *Schema Linking* seguido de LLM gerador, configuração comum em sistemas recentes (ZHANG et al., 2023; MAAMARI et al., 2024).

5.3 Resultados Técnicos

Os artefatos técnicos produzidos no âmbito do pós-doutorado incluem:

(i) a implementação funcional do *framework* CHSMST+MR em Java, com integração ao Apache Spark 3.5.0 e ao banco de dados PostgreSQL/PostGIS;

(ii) os modelos de linguagem ajustados para as tarefas de *Schema Linking* (BERTimbau *large fine-tuned*) e geração SQL (Llama 3.1 *fine-tuned* com LoRA/QLoRA); e

(iii) a interface NLI completa, desenvolvida em Python (FastAPI) e TypeScript/React, com suporte a 17 tipos de visualização e integração com Chart.js, React-Force-Graph-2d e Leaflet.

5.4 Aderência entre Resultados e Objetivos Propostos

A tabela a seguir sintetiza o grau de aderência entre os objetivos específicos do projeto e os resultados efetivamente alcançados:

Objetivo Específico	Resultado Alcançado	Situação
Preparação de dados	Abordado metodologicamente nas etapas de pré-processamento e engenharia de dados dos dois eixos de pesquisa.	Contemplado
Data Cleaning	Tratado transversalmente: DSCR como forma de normalização de densidade; engenharia de dados do NLI.	Contemplado
Data Mining – CHSMST+MR	Framework CHSMST+MR desenvolvido, implementado e validado experimentalmente. Artigo publicado em periódico internacional.	Cumprido integralmente
Visual Data Mining – NLI	Interface NLI desenvolvida e validada em banco de dados real. Artigo submetido a periódico internacional.	Cumprido – artigo em avaliação
Data Cleaning – Paralelização de Detecção de Outliers	Algoritmo de força bruta de Keogh et al. paralelizado via Apache Spark (RDD/PySpark), com redução de 35% a 70% no tempo de execução em quatro datasets financeiros e preservação integral da acurácia de detecção. Artigo submetido ao International Journal of Data Science and Analytics (Springer Nature).	Cumprido – artigo em avaliação
Visual Data Mining – Microsserviços para Saúde	Arquitetura de microsserviços (AMQP/JWT) para visualização integrada de dados estruturados e não estruturados de PEP, com pipeline ICD-10 e avaliação heurística de 75,80% (Dowding–Merrill). Artigo submetido ao Journal of Medical Systems (Springer Nature).	Cumprido – artigo em avaliação

6. Produção Científica Vinculada ao Pós-Doutorado

6.1 Artigo Publicado

O primeiro produto científico gerado no período foi um artigo publicado em periódico científico internacional indexado, diretamente vinculado ao objetivo específico (c) do projeto:

An Architectural Optimization Framework for Scalable Spatial Clustering in High-Redundancy Environments

Autores: Carlos Roberto Valêncio, Wellington Reguera Gouveia, Geraldo Francisco Donegá Zafalon, Angelo Cesar Colombini, Mario Luiz Tronco, Tiago Luís de Andrade

Periódico: Technologies (MDPI) — ISSN 2227-7080

Volume/Número: Technologies 2026, 14, 171

DOI: https://doi.org/10.3390/technologies14030171
Recebido: 26 jan. 2026 | Revisado: 25 fev. 2026 | Aceito: 2 mar. 2026 | Publicado: 10 mar. 2026
Indexação: MDPI / Scopus / Web of Science — Acesso Aberto (CC BY)

O artigo propõe o *framework* CHSMST+MR como otimização arquitetural do algoritmo CHSMST+ (VALÊNCIO et al., 2016), introduzindo a camada de Redução Distribuída de Cardinalidade Espacial (DSCR) no ambiente Apache Spark. O trabalho demonstra experimentalmente que a pré-agregação *MapReduce* de registros com coordenadas redundantes reduz a cardinalidade efetiva do conjunto de dados de forma determinística e sem perda de informação, resultando em reduções médias de tempo de execução de até 79,96% em cenários de alta densidade com atributos de baixa dimensionalidade. A análise de complexidade assintótica formaliza a transição de O(f(N)) para O(f(U)) + O(N), onde U representa o número de chaves espaciais únicas e U << N em cenários de alta densidade (VALÊNCIO et al., 2026).

6.2 Artigos Submetidos – Em Avaliação

Os três artigos submetidos a periódicos internacionais no período estão detalhados a seguir, com seus respectivos dados de submissão:

Trabalho 1

Integration of Language Models in Intelligent Interfaces for Data Visualization

Autores: João Pedro Quadrado, Carlos Roberto Valêncio, Geraldo Francisco Donegá Zafalon, Angelo Cesar Colombini, Mario Luiz Tronco, Tiago Luís de Andrade

Periódico de submissão: Information Visualization (SAGE Publications)

Manuscript ID: IV-26-0067

Status editorial: Submetido em 19 jan. 2026 — Awaiting Reviewer Assignment. Sem decisão editorial até a data deste relatório.

O artigo propõe uma interface de linguagem natural que combina um modelo PLM (BERTimbau *large*, ajustado para *Schema Linking*) e um modelo LLM (SQLCoder, ajustado para geração de SQL), cada um especializado em uma subtarefa distinta do *pipeline* texto–visualização. O sistema suporta 17 tipos de visualização, incluindo representações de elevada complexidade como diagramas de Sankey, grafos de rede e mapas georreferenciados. A

validação foi realizada com 60 questões em banco de dados real multilíngue (25 tabelas, 236 colunas), obtendo 91,67% de acurácia no *Schema Linking* e 75% na geração SQL.

O artigo encontra-se aguardando atribuição de revisores (Awaiting Reviewer Assignment), sem decisão editorial até a data de emissão deste relatório. Seu status será atualizado nos registros institucionais tão logo haja manifestação do periódico.

Trabalho 2

Parallelization of a Brute-Force Algorithm for Outlier Detection in Time Series: A Scalable Approach Using Apache Spark

Autores: Carlos Roberto Valêncio, Beatriz Ferreira de Lima, Angelo Cesar Colombini, Tiago Luís de Andrade, Mario Luiz Tronco, Geraldo Francisco Donegá Zafalon

Periódico de submissão: International Journal of Data Science and Analytics (Springer Nature)

Submission ID: 345f33a5-beee-469f-9502-7445913ff70a

Status: Submetido — em verificação técnica (Technical Check). Sem decisão editorial até a data deste relatório.

O trabalho propõe a paralelização do algoritmo de força bruta para detecção de *outliers* em séries temporais, originalmente descrito por Keogh et al. (2005), utilizando o *framework* Apache Spark como plataforma de processamento distribuído. O algoritmo original apresenta complexidade computacional $O(n^2)$, o que inviabiliza sua aplicação a grandes volumes de dados. A versão paralelizada, implementada em Python com uso da API PySpark, distribui as comparações entre subsequências em múltiplos núcleos computacionais por meio de estruturas RDD, reduzindo os tempos de processamento entre 35% e 70% em relação à implementação sequencial, conforme validação em quatro bases de dados de séries temporais financeiras com volumetrias crescentes.

O trabalho vincula-se aos objetivos específicos (a) e (b) do projeto — preparação de dados e *Data Cleaning* — e constitui contribuição metodológica ao campo da limpeza de séries temporais em contextos de Big Data. O artigo foi submetido ao International Journal of Data Science and Analytics (Springer Nature), encontrando-se em verificação técnica na data de emissão deste relatório.

Trabalho 3

A Microservices-Based Architecture for Integrated Visualization of Structured and Unstructured Health Data

Autores: Carlos Roberto Valêncio, Douglas Brandão dos Santos, Angelo Cesar Colombini, Tiago Luís de Andrade, Mario Luiz Tronco, Geraldo Francisco Donega Zafalon

Periódico de submissão: Journal of Medical Systems (Springer Nature)

Submission ID: 09451198-a6c0-412a-94b3-abdb8dda064d

Status: Submetido — em verificação técnica (Technical Check). Sem decisão editorial até a data deste relatório.

Este trabalho propõe uma arquitetura baseada em microsserviços para o desenvolvimento de ambientes de visualização de dados da área da saúde, com capacidade de operar sobre dados estruturados e não estruturados provenientes de Prontuários Eletrônicos do Paciente (PEP). A arquitetura adota comunicação assíncrona por filas de mensageria (protocolo AMQP), autenticação via JWT e componentes independentes para diferentes entidades hospitalares (pacientes, internações, emergências), compatíveis com implantação em ambiente de computação em nuvem.

O ambiente de visualização foi validado por avaliação heurística baseada nos princípios de Dowding e Merrill (2018), obtendo 75,80% da pontuação máxima — resultado que indica eficácia satisfatória de usabilidade. O trabalho posiciona-se como contribuição ao eixo de *Visual Data Mining* aplicado ao setor de saúde, preenchendo lacuna identificada na literatura quanto à ausência de soluções que conciliem escalabilidade, reusabilidade de código e gestão de heterogeneidade de dados em um único sistema de visualização.

Este manuscrito vincula-se ao objetivo específico (d) do projeto e estende o escopo de Visual Data Mining para o domínio de dados clínicos — contexto em que a heterogeneidade estrutural dos dados e os requisitos de privacidade tornam as soluções existentes insuficientes. O artigo foi submetido ao Journal of Medical Systems (Springer Nature), encontrando-se em verificação técnica na data de emissão deste relatório.

7. Discussão Crítica das Contribuições do Pós-Doutorado

7.1 Contribuição para a Área

Os dois eixos de pesquisa deste projeto (Pós-Doutorado) otimização de *clustering* espacial distribuído e interfaces de linguagem natural para visualização, produziram contribuições com motivações distintas, mas com um ponto em comum: ambos atacam gargalos que os frameworks e sistemas existentes deixam sem tratamento nativo.

No campo do *clustering* espacial distribuído, o *framework* CHSMST+MR preenche uma lacuna específica identificada na literatura: a ausência de mecanismos nativos de colapso de redundância coordenada nos principais *frameworks* de propósito geral (HASAN; XING; MAGRAY, 2022; THEODORAKOPOULOS; KARRAS; KRIMPAS, 2025). A demonstração de que uma redução superior a 50% no tempo de execução pode ser obtida por reconfiguração da representação dos dados sem alteração do algoritmo de *clustering* subjacente e sem perda de fidelidade matemática, é diretamente aplicável a qualquer algoritmo de *clustering* iterativo que opere sobre dados com redundância coordenada DBSCAN e K-Medoids incluídos, desde que implementado em ambiente MapReduce compatível (VALÊNCIO et al., 2026).

No campo das interfaces de linguagem natural, a contribuição reside na demonstração prática de que modelos de código aberto, ajustados com técnicas eficientes de *fine-tuning* (LoRA, QLoRA) e combinados em *pipeline* especializado, podem atingir desempenho competitivo para casos de uso de complexidade baixa a média, sem dependência de APIs proprietárias e com pleno controle institucional dos dados (HU et al., 2023; ZHAO et al., 2023). Esta abordagem é particularmente relevante em contextos em que os dados não podem ser transmitidos a servidores externos por restrições contratuais ou regulatórias o que abrange, por exemplo, projetos sob NDA em parceria com empresas como a Petrobras e ambientes hospitalares com dados de prontuário.

7.2 Contribuição Metodológica

Dois resultados metodológicos têm relevância particular. O primeiro é a formalização do DSCR como primitiva de pré-processamento distribuído: ao ser especificado como algoritmo independente com semântica clara de preservação de informação, o DSCR pode ser incorporado como etapa de pré-processamento em qualquer *pipeline* de mineração espacial que opere em ambientes Apache Spark, independentemente do algoritmo de *clustering* utilizado (VALÊNCIO et al., 2026).

O segundo legado metodológico é a estratégia de treinamento com ruído controlado para o LLM do *pipeline* NLI, que demonstrou como um PLM imperfeito, situação invariável em cenários reais de inferência, pode ser compensado por um LLM treinado para operar sob condições de esquema incompleto ou ruidoso (ZHANG et al., 2023; DONG et al., 2023). Esta estratégia é transferível para qualquer sistema de conversão texto–SQL em dois estágios.

7.3 Contribuição Institucional e Acadêmica

Do ponto de vista institucional, o estágio reforçou a parceria de longa data entre o GBD–UNESP e o Departamento de Engenharia Elétrica da UFF, ampliando o alcance geográfico e

disciplinar das publicações do grupo. O artigo publicado reúne autores de quatro instituições (UNESP, UFF, USP, UNEMAT), resultado direto da rede de colaboração construída ao longo do estágio e extensível às orientações de pós-graduação em andamento no TEE/UFF.

A participação nas etapas de concepção metodológica dos dois eixos de pesquisa e na revisão dos manuscritos gerados reforçou a integração entre as linhas de pesquisa do TEE/UFF e do GBD/UNESP, com desdobramentos diretos para orientações em andamento no programa de pós-graduação em Engenharia Elétrica.

7.4 Limitações e Perspectivas de Continuidade

Duas limitações do trabalho realizado merecem registro. No caso do CHSMST+MR, a validação foi conduzida em ambiente de emulação local de nó único, com três repetições por cenário, número suficiente para demonstração de prova de conceito, mas insuficiente para análise estatística robusta de sistemas distribuídos reais, conforme reconhecido pelos próprios autores no artigo publicado (VALÊNCIO et al., 2026). A extensão para ambientes *multi-nó* em nuvem (Amazon EMR, Azure HDInsight) com 10–20 repetições por cenário é o desdobramento mais urgente.

No caso da interface NLI, as limitações identificadas: queda de desempenho em questões de alta complexidade, dependência entre PLM e LLM em cascata, e restrições de *hardware* que impediram o uso de modelos com maior número de parâmetros, apontam para três direções de continuidade:

(i) avaliação com modelos de maior capacidade quando os recursos computacionais permitirem;

(ii) incorporação de estratégias de aprendizado por reforço a partir de *feedback* de usuários para melhoria contínua; e

(iii) expansão para outros domínios de banco de dados, testando a generalização além do corpus multilíngue utilizado na validação.

8. Considerações Finais

O trabalho desenvolvido entre julho de 2024 e abril de 2026 gerou dois produtos científicos de relevância demonstrável para a área de Big Data: um artigo publicado em periódico internacional indexado (VALÊNCIO et al., 2026) e um artigo submetido a periódico de reconhecimento na área de visualização de informação. Ambos derivam diretamente dos objetivos específicos estabelecidos no projeto de pesquisa aprovado.

A execução do projeto mostrou que os objetivos inicialmente propostos foram úteis para delimitar o tema, mas que as decisões metodológicas efetivas dependeram, sobretudo, do aprofundamento bibliográfico e da identificação de lacunas específicas na literatura. Em ambos os casos, o problema de pesquisa específico só ficou claro após imersão na literatura recente: a redundância de coordenadas como gargalo computacional tratável por pré-agregação não era evidente nos objetivos iniciais, assim como a ausência de sistemas NLI com modelos heterogêneos e suporte a visualizações complexas só se tornou uma lacuna definida após mapeamento sistemático dos trabalhos existentes (AWAD; HEFNY, 2023; LIU et al., 2021).

Adicionalmente, encontram-se em fase de preparação dois trabalhos diretamente vinculados às atividades deste projeto: um sobre paralelização de algoritmos de detecção de *outliers* em séries temporais e outro sobre arquitetura de microsserviços para visualização de dados da saúde. Ambos foram submetidos a periódicos Springer Nature — o primeiro ao International Journal of Data Science and Analytics e o segundo ao Journal of Medical Systems — e encontram-se em verificação técnica na data de emissão deste relatório.

Os resultados obtidos permitem identificar com clareza o que foi resolvido, os gargalos de redundância espacial e a ausência de sistemas NLI com cobertura ampla de tipos de visualização e o que permanece em aberto: validação em ambientes multi-nó e expansão para domínios de dados distintos do corpus multilíngue utilizado.

A colaboração com o GBD–UNESP consolidou atuação direta em mineração espacial distribuída e em sistemas de linguagem natural aplicados a dados, linhas com conexão direta às pesquisas em Smart Grids, IoT e infraestrutura inteligente conduzidas no TEE/UFF.

9. Referências Bibliográficas

AWAD, A. M.; HEFNY, H. Parallel implementation of statistical DBSCAN algorithm for spark-based clustering on Google cloud platform. International Journal of Computer and Digital Systems, v. 16, p. 279–290, 2023.

DONG, Q. et al. A Survey on In-context Learning. arXiv preprint arXiv:2301.00234, 2023.

FAKIR, Y.; ELAYACHI, R.; BTIS-SAM, M. Clustering objects for spatial data mining: A comparative study. Journal of Big Data Research, v. 1, p. 1–11, 2023.

GBIF.ORG USER. Occurrence Download, 2022. Disponível em: https://www.gbif.org/occurrence/download/0188443-220831081235567. Acesso em: 1 mar. 2026.

HASAN, Z.; XING, H. J.; MAGRAY, M. I. Big data machine learning using Apache Spark MLlib. Mesopotamian Journal of Big Data, 2022.

HU, Z. et al. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. arXiv preprint arXiv:2304.01933, 2023.

LANG, K.; CHAI, X. Research on clustering algorithm based on Spark. In: International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, 2023. p. 855–858.

LEE, C. J. et al. Macro-Queries: An Exploration into Guided Chart Generation from High Level Prompts. arXiv preprint arXiv:2408.12726, 2024.

LI, C. et al. Mining area skyline objects from map-based big data using Apache Spark framework. Array, v. 25, p. 100373, 2025.

LIU, C. et al. ADVISor: Automatic Visualization Answer for Natural-Language Question on Tabular Data. In: IEEE Pacific Visualization Symposium (PacificVis), Tianjin, 2021. p. 11–20.

MAAMARI, K. et al. The Death of Schema Linking? Text-to-SQL in the Age of Well-Reasoned Language Models. arXiv preprint arXiv:2408.07702, 2024.

MANJULA, A.; NARSIMHA, G. Using an Efficient Optimal Classifier for Soil Classification in Spatial Data Mining Over Big Data. Journal of Intelligent Systems, v. 29, p. 172–188, 2020.

SAH, S. et al. Generating Analytic Specifications for Data Visualization from Natural Language Queries using Large Language Models. arXiv preprint arXiv:2408.13391, 2024.

SHI, L. et al. A Survey on Employing Large Language Models for Text-to-SQL Tasks. arXiv preprint arXiv:2407.15186, 2024.

THEODORAKOPOULOS, L.; KARRAS, A.; KRIMPAS, G. A. Optimizing Apache Spark MLlib: Predictive performance of large-scale models for big data analytics. Algorithms, v. 18, p. 74, 2025.

VALÊNCIO, C. R. et al. CHSMST+: An Algorithm for Spatial Clustering. In: International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), Guangzhou, 2016. p. 352–357.

VALÊNCIO, C. R. et al. An Architectural Optimization Framework for Scalable Spatial Clustering in High-Redundancy Environments. Technologies, v. 14, n. 3, p. 171, 2026. DOI: https://doi.org/10.3390/technologies14030171.

WANG, D. Parallelization of K-Means and Spatial Join Algorithms on Heterogeneous Platforms Using Apache Spark and GPU Integration. Informatica, v. 49, p. 279–292, 2025.

YANG, H. et al. The implementation solution for automatic visualization of tabular data in relational databases based on large language models. In: International Conference on Asian Language Processing (IALP), Hohhot, 2024. p. 175–180.

YU, T. et al. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. arXiv preprint arXiv:1809.08887, 2018.

ZHANG, H. et al. ACT-SQL: In-Context Learning for Text-to-SQL with Automatically-Generated Chain-of-Thought. arXiv preprint arXiv:2310.17342, 2023.

ZHAO, W. X. et al. A Survey of Large Language Models. arXiv preprint arXiv:2303.18223, 2023.

ZHU, J. P. et al. Chat2Query: A Zero-Shot Automatic Exploratory Data Analysis System with Large Language Models. In: IEEE 40th International Conference on Data Engineering (ICDE), Utrecht, 2024. p. 5429–5432.

10. Anexos

Anexo A – Artigo Publicado

Valêncio, C. R.; Gouveia, W. R.; Zafalon, G. F. D.; Colombini, A. C.; Tronco, M. L.; Andrade, T. L. An Architectural Optimization Framework for Scalable Spatial Clustering in High-Redundancy Environments. Technologies, 2026, 14, 171. DOI: https://doi.org/10.3390/technologies14030171.

O artigo completo encontra-se disponível em acesso aberto no repositório MDPI. Uma cópia impressa acompanha este relatório como Anexo A.

Anexo B – Artigo Submetido (Em Avaliação)

Quadrado, J. P.; Valêncio, C. R.; Zafalon, G. F. D.; Colombini, A. C.; Tronco, M. L.; Andrade, T. L. Integration of Language Models in Intelligent Interfaces for Data Visualization. Submetido a: Information Visualization (SAGE Publications). Manuscript ID: IV-26-0067. Status: Awaiting Reviewer Assignment.

O manuscrito encontra-se aguardando atribuição de revisores (Awaiting Reviewer Assignment), sem decisão editorial até a data de emissão deste relatório. Cópia do manuscrito na versão submetida acompanha este relatório como Anexo B.

Anexo C – Comprovante de Submissão

Comprovantes de submissão dos três manuscritos submetidos no período: (i) manuscrito IV-26-0067 ao periódico Information Visualization (SAGE Publications), confirmado pelo sistema ScholarOne Manuscripts em 19 jan. 2026; (ii) manuscrito ao International Journal of Data Science and Analytics (Springer Nature), Submission ID 345f33a5-beee-469f-9502-7445913ff70a, em verificação técnica; (iii) manuscrito ao Journal of Medical Systems (Springer Nature), Submission ID 09451198-a6c0-412a-94b3-abdb8dda064d, em verificação técnica.

Anexo D – Artigos Submetidos — Em Avaliação

Valêncio, C. R.; Lima, B. F.; Colombini, A. C.; Andrade, T. L.; Tronco, M. L.; Zafalon, G. F. D. Parallelization of a Brute-Force Algorithm for Outlier Detection in Time Series: A Scalable Approach Using Apache Spark. Submetido a: International Journal of Data Science and Analytics (Springer Nature). Submission ID: 345f33a5-beee-469f-9502-7445913ff70a. Status: Technical Check.

Valêncio, C. R.; Santos, D. B.; Colombini, A. C.; Andrade, T. L.; Tronco, M. L.; Zafalon, G. F. A Microservices-Based Architecture for Integrated Visualization of Structured and Unstructured Health Data. Submetido a: Journal of Medical Systems (Springer Nature). Submission ID: 09451198-a6c0-412a-94b3-abdb8dda064d. Status: Technical Check.