RESUMO EXPANDIDO
Anais do Congresso Interdisciplinar Internacional – CONIÚNICA
08 a 11 de setembro. Híbrido, Ipatinga – MG
doi.org/10.53426/con.v5i0.002

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

ALÉM DO PANDAS E DO COLAB: APACHE SPARK, DATABRICKS E PIPELINE MEDALHÃO PARA BIG DATA ESCALÁVEL — ESTUDO DE CASO COVID-19

BEYOND PANDAS AND COLAB: APACHE SPARK, DATABRICKS AND MEDALLION PIPELINE FOR SCALABLE BIG DATA — A COVID-19 CASE STUDY

MÁS ALLÁ DE PANDAS Y COLAB: APACHE SPARK, DATABRICKS Y PIPELINE MEDALLÓN PARA BIG DATA ESCALABLE — ESTUDIO DE CASO COVID-19

doi.org/10.53426/con.v5i0.002

Ibisen de Brito Gonçalves[1]
[1] Faculdade Uniúnica, Ipatinga-MG, [PREENCHER: seu-email@dominio.com]
Grupo de Trabalho: GT 2 – Educação a Distância, Ensino Híbrido e Tecnologias Educacionais
Trabalho premiado: Menção honrosa.

RESUMO

A formação acadêmica em análise de dados frequentemente recorre a Pandas, Google Colab, R e MATLAB para exploração, estatística e visualização em um único nó computacional — abordagem adequada a datasets de porte moderado que cabem em memória, tipicamente na ordem de gigabytes (McKINNEY, 2022). Essas ferramentas não implementam, por si sós, processamento distribuído. Em contraste, ecossistemas lakehouse baseados em Apache Spark e Databricks suportam arquiteturas escaláveis para big data em produção, conforme documentado na literatura e em guias de referência (ZAHARIA et al., 2016; DATABRICKS, 2023). Este trabalho, conduzido por estudante de Engenharia Elétrica da Faculdade Uniúnica, relata a implementação de um pipeline ETL em Arquitetura Medalhão (Bronze, Silver, Gold) no Databricks Community Edition, utilizando o subconjunto latest/aggregated.csv do COVID-19 Open Data V3 (GOOGLE, s.d.; WALMSLEY et al., 2022). O estudo processou 22.712 registros, 737 variáveis e arquivo fonte de 24,1 MB, com abrangência geográfica global. Foram documentadas regras de ingestão, tipagem na camada Silver, validação por inspeção SQL e seleção de 99 variáveis com taxa de nulos ≤50% para análises confiáveis, excluindo 638 colunas esparsas. A conversão CSV → Parquet/Delta reduziu o volume em ≈4,3× em teste equivalente. Consultas SQL por país retornaram 233 linhas agregadas em 2,08 s. Os resultados incluem KPIs epidemiológicos, visualizações comparativas entre nações (Figuras 2 e 3) e proposta de evolução em Azure (Figura 4, não implementada). A orquestração foi realizada por notebooks encadeados (nt0–nt6) em cluster Spark 14.3 LTS. O repositório público goncalves-ibisen/covid19_open_data_v3 (acessível em https://github.com/goncalves-ibisen/covid19_open_data_v3, verificado em set./2026) contém notebooks nt0–nt6, instruções de execução e materiais necessários à reprodução do pipeline. Conclui-se que a experiência demonstra viabilidade pedagógica de introduzir stack corporativa de big data no contexto universitário, aproximando estudantes de práticas documentadas de engenharia de dados, vinculando cada afirmação aos resultados efetivamente obtidos, sem extrapolar para volumes ou efeitos educacionais não mensurados neste estudo.

Palavras-chave: big data; Apache Spark; Databricks; Pandas; COVID-19

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

INTRODUÇÃO

O mercado de ciência de dados emprega ecossistemas distribuídos e plataformas lakehouse em nuvem. Contudo, muitas disciplinas ainda se apoiam em Pandas, Google Colab, R e MATLAB — ferramentas úteis para datasets que cabem em memória, tipicamente gigabytes (McKINNEY, 2022). Para volumes muito superiores, organizações adotam Apache Spark e Databricks (ZAHARIA et al., 2016; DATABRICKS, 2023); benchmarks da literatura contextualizam essas soluções, mas não foram replicados neste estudo (dataset de 24,1 MB).

O COVID-19 Open Data V3 (GOOGLE, s.d.; WALMSLEY et al., 2022) combina relevância epidemiológica e porte compatível com demonstração de pipeline acadêmico. Objetivos: (a) implementar ETL em Arquitetura Medalhão; (b) executar análises epidemiológicas; (c) apresentar resultados quantitativos; (d) discutir limites e evolução futura.

METODOLOGIA

Trata-se de estudo de caso aplicado e relato de experiência. O repositório público goncalves-ibisen/covid19_open_data_v3 (acessível em https://github.com/goncalves-ibisen/covid19_open_data_v3, verificado em set./2026) contém notebooks nt0–nt6, instruções de execução e materiais necessários à reprodução do pipeline.

Fonte: COVID-19 Open Data V3, subconjunto latest/aggregated.csv (737 variáveis; 22.712 registros; 24,1 MB; abrangência global). Snapshot com último registro não nulo por localidade (GOOGLE, s.d.), distinto da série histórica completa (aggregated.csv.gz, ≈1,3 GB).

Ambiente: Databricks Community Edition; Spark Runtime 14.3 LTS; Python, Pandas, PySpark, Spark SQL; DBFS. Execução documentada em ago./2024.

Pipeline Medalhão. Landing: download CSV via URL pública. Bronze: leitura Spark, colunas year/month/day/data_carga_lake, gravação Delta. Silver: tipagem Pandas (pd.to_numeric, errors='ignore'). Gold: passagem conforme pipeline. Orquestração: notebook nt0 (nt1–nt6). CI/CD e Azure não implementados (Figura 4 = proposta futura).

Qualidade e validação. View analítica: 99 variáveis com nulos ≤50% (638 excluídas). Validação Silver/Gold: display, LIMIT 10 e contagem 22.712 linhas. coalesce(..., 0) em totalizadores nulos. Análise: KPIs epidemiológicos e comparação entre países (Figuras 2 e 3).

RESULTADOS E DISCUSSÃO

A Tabela 1 contrasta ecossistemas de ensino e mercado (McKINNEY, 2022; ZAHARIA et al., 2016; DATABRICKS, 2023). A Tabela 2 consolida resultados desta implementação, confirmando volume moderado (24,1 MB) e arquitetura escalável — distinta de execução em petabytes.

Tabela 1. Ferramentas acadêmicas versus tecnologias corporativas para escala de dados.

Dimensão	Pandas / Colab / R / MATLAB	Spark / Databricks / Nuvem
Modelo de execução	Memória de um nó	Cluster distribuído
Escala prática típica	GB a poucos TB (chunking)	TB a escala distribuída
Orquestração	Manual / células	Workflows, Airflow, CI/CD (produção)
Armazenamento	CSV, XLSX	Parquet, Delta Lake
Governança	Limitada	ACID, versionamento, RBAC
Uso predominante	Prototipagem e ensino	Produção analítica em escala

Fonte: elaboração própria com base em McKINNEY (2022), ZAHARIA et al. (2016) e DATABRICKS (2023).

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

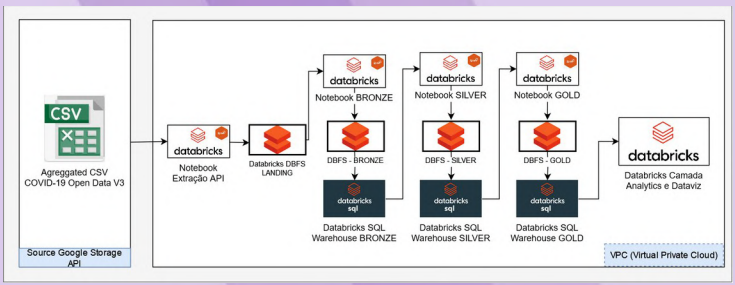
Tabela 2. Resultados quantitativos da implementação.

Indicador	Resultado observado neste estudo
Dataset (V3, latest)	22.712 registros; 737 variáveis; 24,1 MB
Abrangência / corte	Global; snapshot latest (último registro por localidade)
Bronze → Gold	22.712 registros; metadados temporais e data_carga_lake
Redução CSV → Parquet	≈4,3× (24,1 MB → 5,6 MB)
View analítica	99 variáveis (nulos ≤50%); 638 colunas excluídas
Consulta por país	233 países; tempo: 2,08 s

Fonte: elaboração própria a partir da execução documentada no repositório.

A Figura 1 apresenta o pipeline Medalhão implementado. A conversão CSV→Parquet/Delta reduziu o volume em ≈4,3× (24,1 MB → 5,6 MB). Benchmarks da literatura não foram apresentados como medições próprias.

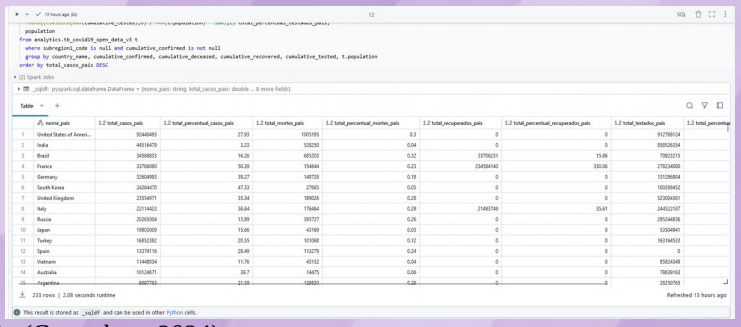
Figura 1. Pipeline em Arquitetura Medalhão implementado no Databricks Community Edition (Landing, Bronze, Silver, Gold e camada analítica).



Fonte: elaboração própria (Gonçalves, 2024).

A Figura 2 exibe KPIs por país (233 países; 2,08 s). A Figura 3 compara os 20 países com mais mortes acumuladas — visualização transversal, não série temporal. A Figura 4 propõe evolução futura em Azure (não implementada).

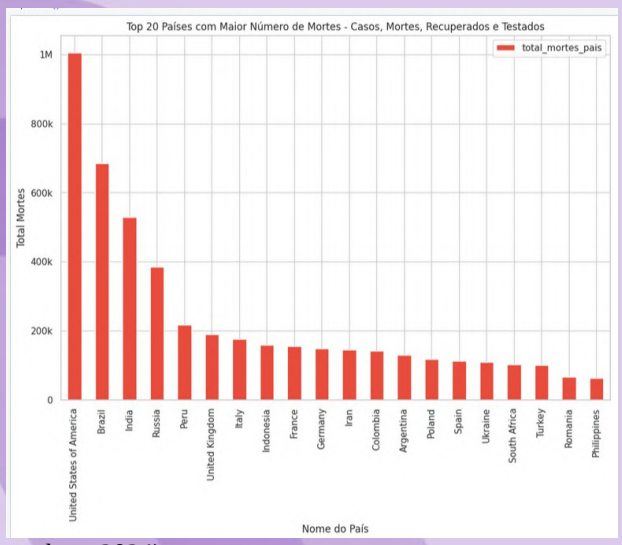
Figura 2. KPIs epidemiológicos por país obtidos da camada analítica (233 países; consulta SQL executada em 2,08 s).



Fonte: elaboração própria (Gonçalves, 2024).

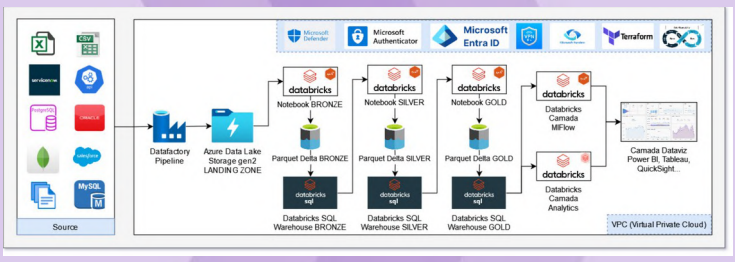
ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

Figura 3. Comparação entre países: top 20 com maior número acumulado de mortes reportadas no recorte analítico.



Fonte: elaboração própria (Gonçalves, 2024).

Figura 4. Proposta de evolução futura: arquitetura Azure Databricks com segurança corporativa, CI/CD e MLOps (não implementada neste estudo).



Fonte: elaboração própria (Gonçalves, 2024).

CONSIDERAÇÕES FINAIS

O estudo demonstra viabilidade de implementar pipeline Medalhão, Delta Lake e SQL analítico no Databricks Community Edition, com resultados quantitativos na Tabela 2.

Contribuições: (1) relato reproducível com pipeline documentado; (2) contraste fundamentado entre ferramentas single-node e stack lakehouse; (3) resultados quantitativos de ingestão, compressão e análise COVID-19; (4) proposta de evolução futura para Azure, CI/CD e MLOps (Figura 4).

Limitações: dataset de porte moderado; ausência de avaliação educacional; Azure, CI/CD e MLOps apenas propostos. Conclusões sobre empregabilidade ou aprendizagem exigiriam estudos longitudinais não conduzidos aqui.

REFERÊNCIAS

DATABRICKS. Medallion architecture. Databricks Documentation, 2023. Disponível em: https://www.databricks.com/glossary/medallion-architecture. Acesso em: 2 set. 2026.

FONSECA, Ronald Assis; RAFAEL, Clélio Rodrigo Paiva; VICTOR, Emerson Francisco; SANTOS, Núbia Cristina Geraldo dos; OLIVEIRA, Amanda Xavier de; SILVA, Cleiciane Ferreira da. O uso do Mentimeter como metodologia ativa virtual no curso de Gestão Ambiental da Faculdade Única de Ipatinga. **Revista Multidisciplinar de Educação e Meio Ambiente**, v. 2, n. 4, p. 1–7, 2021. DOI: 10.51189/rema/1939.

GONÇALVES, I. B. covid19_open_data_v3: pipeline Databricks COVID-19 Open Data V3. GitHub, 2024. Disponível em: https://github.com/goncalves-ibisen/covid19_open_data_v3. Acesso em: 2 set. 2026.

GOOGLE. COVID-19 Open Data. Google Cloud Platform, s.d. Disponível em: https://github.com/GoogleCloudPlatform/covid-19-open-data. Acesso em: 2 set. 2026.

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

MCKINNEY, W. **Python for Data Analysis**. 3. ed. Sebastopol: O'Reilly Media, 2022.

RAFAEL, Candyce Mabelle Paiva et al. O geoprocessamento, a territorialização e o cadastramento como ferramentas de unificação do trabalho de agentes de saúde e potencialização das práticas de cuidado no território. **RENOTE**, v. 22, n. 2, p. 422–429, 2024. DOI: 10.22456/1679-1916.142583.

RAFAEL, Clélio Rodrigo Paiva; TESSAROLE, Cristina de Lacerda; FONSECA, Ronald Assis. A importância das tecnologias da informação e comunicação (TIC's) no processo de ensino e aprendizagem do ensino fundamental: uma revisão. **Anais do Encontro Virtual de Documentação em Software Livre e Congresso Internacional de Linguagem e Tecnologia Online**, v. 11, n. 1, 2023. Disponível em: https://ciltec.textolivre.pro.br/index.php/CILTecOnline/article/view/1131. Acesso em: 30 set. 2026.

WAHLTINEZ, Oscar et al. COVID-19 Open-Data a global-scale spatially granular meta-dataset for coronavirus disease. **Scientific Data**, v. 9, art. 162, 2022. DOI: 10.1038/s41597-022-01263-z.

ZAHARIA, M. et al. Apache Spark: a unified engine for big data processing. **Communications of the ACM**, v. 59, n. 11, p. 56-65, 2016. DOI: 10.1145/2934664.