Impact Statement of "GeoSpark: A Cluster Computing Framework for Processing Large-Scale Spatial Data"

Jia Yu
Wherobots
jiayu@wherobots.com

Jinxuan Wu
Two Sigma
jinxuanw@apache.org

Mo Sarwat
Wherobots
mo@wherobots.com

1 Introduction

Over the past decade, GeoSpark has played a foundational role in advancing large-scale spatial data processing. Its impact spans academic research, industrial adoption, open-source ecosystems, and global usage. In what follows, we outline its contributions in accordance with the award selection criteria. Notably, the GeoSpark project was donated to the Apache Software Foundation in 2020 and was subsequently renamed Apache Sedona [1]. While many references now cite Sedona, they all trace their technical lineage back to the original GeoSpark paper.

2 Utility and impact on research

GeoSpark introduced the first distributed in-memory computing framework for large-scale spatial data. It pioneered the concept of spatial distributed datasets, spatial partitioning, and distributed spatial indexing, providing a scalable foundation for spatial analytics. It has been cited over 500 times [2], with nearly 800 citations across all related papers, including ICDE'16 [3], and Geoinformatica'19 [4]. The work is frequently referenced in top venues such as VLDB, ICDE, SIGMOD, and SIGSPATIAL.

The framework has been independently evaluated by researchers around the world. A SIGMOD 2020 paper [5] compared several spatial systems and found that GeoSpark's distance join significantly outperformed Simba, another Spark-based spatial computing system. Another PVLDB 2018 benchmarking paper concluded that GeoSpark was one of the most performant and complete distributed spatial analytics frameworks [6].

GeoSpark has been foundational for subsequent research on distributed geospatial processing, influencing many highly-cited research articles like Simba [7], LocationSpark [8], SparkGIS [9], ST-Hadoop [10], DITA [11], and JUST [12]. The paper is also central to the academic lineage of Apache Sedona (formerly GeoSpark), which is widely used in systems research, spatial machine learning, and spatiotemporal data analytics.

3 Practical impact and scientific reach

GeoSpark's scalable architecture has enabled practical adoption in a wide range of real-world applications involving spatial data. It supports fundamental operations such as spatial joins, nearest neighbor queries, distance joins, and trajectory analysis at distributed scale, which are essential across numerous domains:

(1) Urban computing: Used in traffic analysis [13], and traffic demand prediction [14], GeoSpark powers large-scale processing of GPS data in city-scale applications. (2) Climate science: GeoSpark have been used to join raster and vector datasets (e.g., zonal statistics over satellite imagery), enabling scalable climate and environmental analysis [15–17]. (3) Public health: During COVID-19, researchers employed GeoSpark to correlate anonymized mobility data with infection rates at regional levels, requiring high-throughput spatial joins [18, 19]. (4) Telecom and energy infra: Service providers in the telecom and energy sectors leverage GeoSpark to optimize infrastructure placement, coverage, and risk mitigation. Applications include cell tower coverage analysis [20] and facility management [21].

GeoSpark is highly attractive for interdisciplinary research, and it has been cited in venues across computer systems [5–8, 11–13], geoinformatics [9, 10, 14], remote sensing [16, 17], and communication & networks [15, 18].

4 Utility in industry

GeoSpark has had a substantial and lasting impact on industrial data infrastructure for spatial analytics. In 2020, we donated the open-source implementation of GeoSpark to the Apache Software Foundation, where it was renamed Apache Sedona [1, 22] and later graduated as a Top-Level Project. Today, GeoSpark (i.e., Apache Sedona) is the most widely adopted open-source distributed spatial processing engine in industry.

The Apache Sedona project has gained significant traction in the open-source and industrial communities, with over 2,000 GitHub stars and contributions from more than 130 developers worldwide. It has been downloaded over 50 million times across Maven Central and PyPI [23, 24], and is used by more than 10,000 organizations (see Figure 2) spanning logistics, e-commerce, financial services, agriculture, and cloud infrastructure sectors.

Each month, users launch more than 5 million Sedona-based clusters across all major cloud platforms, including AWS, Azure, Google Cloud, Databricks, and Snowflake. The project's official website receives over 10,000 monthly page views, with global traffic spanning all major countries across the Americas, Europe, Asia, and many parts of Africa, as illustrated in Figure 1.

Notable users of Apache Sedona include Amazon, which integrates it into daily last-mile delivery route planning; Allstate, which uses it to analyze traffic accident patterns; and Land O'Lakes, which applies it to large-scale agricultural analytics. Adoption has also been reported in organizations such as Mercedes-Benz, Shopee, SpaceX, JB Hunt, JPMorgan Chase, Capital One, HERE Technologies, and Mapbox, supporting use cases including routing optimization, asset tracking, and spatial risk modeling.

The technology introduced in GeoSpark continues to evolve through Wherobots [25], a company founded by the original authors of this paper. Wherobots offers a commercial geospatial lakehouse platform built on Apache Sedona's query engine, enabling

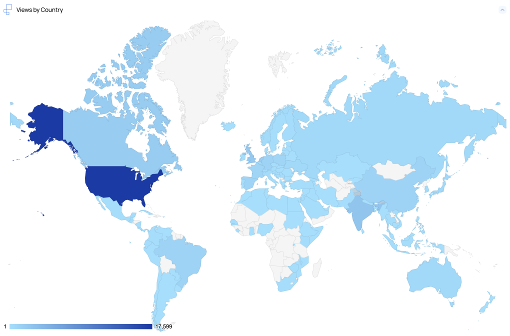
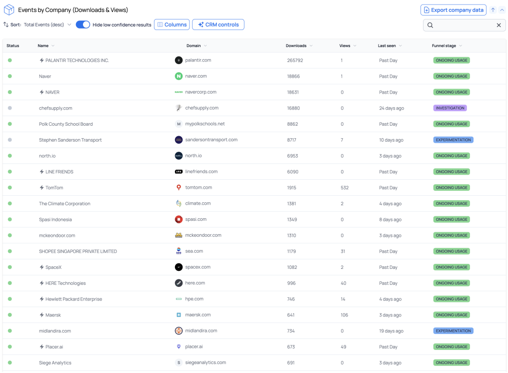


Figure 1: Geographic distribution of Sedona website traffic


Events by Company (Downloads & Views)
Sort: Total Events (desc)
Hide low confidence results
Columns
CRM controls
Export company data

Status	Name	Domain	Downloads	Views	Last seen	Funnel stage
●	⚡ PALANTIR TECHNOLOGIES INC.	palantir.com	265792	1	Past Day	ONGOING USAGE
●	Naver	naver.com	18866	1	Past Day	ONGOING USAGE
●	⚡ NAVER	navercorp.com	18631	0	Past Day	ONGOING USAGE
○	chefsupply.com	chefsupply.com	16880	0	24 days ago	INVESTIGATION
●	Polk County School Board	mypolkschools.net	8862	0	Past Day	ONGOING USAGE
○	Stephen Sanderson Transport	sandersontransport.com	8717	7	10 days ago	EXPERIMENTATION
●	north.io	north.io	6953	0	3 days ago	ONGOING USAGE
●	⚡ LINE FRIENDS	linefriends.com	6090	0	Past Day	ONGOING USAGE
●	⚡ TomTom	tomtom.com	1915	532	Past Day	ONGOING USAGE
●	The Climate Corporation	climate.com	1381	2	4 days ago	ONGOING USAGE
●	Spasi Indonesia	spasi.com	1349	0	8 days ago	ONGOING USAGE
●	mckeondoor.com	mckeondoor.com	1310	0	3 days ago	ONGOING USAGE
●	SHOPEE SINGAPORE PRIVATE LIMITED	sea.com	1179	31	Past Day	ONGOING USAGE
●	⚡ SpaceX	spacex.com	1082	2	Past Day	ONGOING USAGE
●	⚡ HERE Technologies	here.com	996	40	Past Day	ONGOING USAGE
●	⚡ Hewlett Packard Enterprise	hpe.com	746	14	4 days ago	ONGOING USAGE
●	⚡ Maersk	maersk.com	641	106	3 days ago	ONGOING USAGE
●	midlandira.com	midlandira.com	734	0	19 days ago	EXPERIMENTATION
●	⚡ Placer.ai	placer.ai	673	49	Past Day	ONGOING USAGE
●	Siege Analytics	siegeanalytics.com	691	0	3 days ago	ONGOING USAGE

Figure 2: Top 20 most active Sedona users in May 2025

high-speed spatial analytics and geospatial AI at enterprise scale. In the past 3 years, the company has raised 27 million dollars in two funding rounds (Seed and Series A) from world-renowned venture capital firms, including Felicis, Wing VC, Clear Ventures, Prosperity7, and JetBlue Ventures. Today, Wherobots has close to 40 employees and operates offices in San Francisco and Seattle, serving many customers across industries. In particular, these include the Overture Maps Foundation [26], one of the largest open map data providers, founded by Amazon, Meta, Microsoft, and TomTom.

5 Geographic diversity impact

GeoSpark and its successor, Apache Sedona, have demonstrated wide-reaching geographic impact across both academia and open-source adoption.

On the research side, the original GeoSpark paper has been cited by scholars affiliated with institutions in North America, Europe, Asia, and Africa, based on Google Scholar [2]. Notable citation activity comes from the United States [5, 7], Germany [6, 19], China [11], Tunisia [15], Turkey [16], Italy [18], Austria [20], and Sweden [17] with publications in areas such as geoinformatics, transportation, environmental modeling, and distributed systems.

On the open-source front, Apache Sedona enjoys global developer participation and production usage. The project's GitHub repository includes contributors from many countries, with frequent commits and issue activity from the U.S., Europe and Asia (see Sedona's GitHub [1]). The Sedona website receives more than 10,000 page views per month from users in tens of countries, spanning nearly all regions of the Americas, Europe, Asia, and many countries in Africa (see Figure 1).

References

[1] The Apache Software Foundation. Apache sedona github. https://github.com/apache/sedona, 2025. Accessed: 2025-06-08.

[2] Google Scholar. Citations for the geospark paper. https://scholar.google.com/scholar?cites=18142588456381599558&as_sdt=5,48&sciodt=0,48&hl=en, 2025. Accessed: 2025-06-08.

[3] Jia Yu, Jinxuan Wu, and Mohamed Sarwat. A demonstration of GeoSpark: A cluster computing framework for processing big spatial data. In *ICDE*, 2016.

[4] Jia Yu, Zongsi Zhang, and Mohamed Sarwat. Spatial data management in apache spark: the geospark perspective and beyond. *GeoInformatica*, 23:37–78, 2019.

[5] Ruby Y Tahboub and Tiark Rompf. Architecting a query compiler for spatial workloads. In *SIGMOD*, pages 2103–2118, 2020.

[6] Varun Pandey, Andreas Kipf, Thomas Neumann, and Alfons Kemper. How good are modern spatial analytics systems? *PVLDB*, 11(11):1661–1673, 2018.

[7] Dong Xie, Feifei Li, Bin Yao, Gefei Li, Liang Zhou, and Minyi Guo. Simba: Efficient In-Memory Spatial Analytics. In *SIGMOD*, 2016.

[8] MingJie Tang, Yongyang Yu, Qutaibah M. Malluhi, Mourad Ouzzani, and Walid G. Aref. LocationSpark: A Distributed In-Memory Data Management System for Big Spatial Data. *PVLDB*, 9(13):1565–1568, 2016.

[9] Furqan Baig, Hoang Vo, Tahsin Kurc, Joel Saltz, and Fusheng Wang. Sparkgis: Resource aware efficient in-memory spatial query processing. In *SIGSPATIAL*, pages 1–10, 2017.

[10] Louai Alarabi, Mohamed F Mokbel, and Mashaal Musleh. St-hadoop: A mapreduce framework for spatio-temporal data. *GeoInformatica*, 22:785–813, 2018.

[11] Zeyuan Shang, Guoliang Li, and Zhifeng Bao. Dita: Distributed in-memory trajectory analytics. In *SIGMOD*, pages 725–740, 2018.

[12] Ruiyuan Li, Huajun He, Rubin Wang, Yuchuan Huang, Junwen Liu, Sijie Ruan, Tianfu He, Jie Bao, and Yu Zheng. Just: Jd urban spatio-temporal data engine. In *ICDE*, pages 1558–1569. IEEE, 2020.

[13] Wentao Ning, Qiandong Tang, Yi Zhao, Chuan Yang, Xiaofeng Wang, Teng Wang, Haotian Liu, Chaozu Zhang, Zhiyuan Zhou, Qiaomu Shen, et al. Cheetahvis: a visual analytical system for large urban bus data. *PVLDB*, 13(12):2805–2808, 2020.

[14] Peiqi Zhang, Kathleen Stewart, and Yao Li. Estimating traffic speed and speeding using passively collected big mobility data and a distributed computing framework. *Transactions in GIS*, 27(4):1124–1144, 2023.

[15] Rim Moussa. Scalable analytics of air quality batches with apache spark and apache sedona. In *Proceedings of the 15th ACM International Conference on Distributed and Event-based Systems*, pages 154–159, 2021.

[16] Muhammed Oğuzhan Mete. Geospatial big data analytics for sustainable smart cities. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:141–146, 2023.

[17] Desta Haileselassie Hagos, Theofilos Kakantousis, Vladimir Vlassov, Sina Sheikholeslami, Tianze Wang, Jim Dowling, Claudia Paris, Daniele Marinelli, Giulio Weikmann, Lorenzo Bruzzone, et al. Extremeearth meets satellite data from space. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:9038–9063, 2021.

[18] Claudia Cavallaro, Armir Bujari, Luca Foschini, Giuseppe Di Modica, and Paolo Bellavista. Measuring the impact of covid-19 restrictions on mobility: A real case study from italy. *Journal of Communications and Networks*, 23(5):340–349, 2021.

[19] Jawad Tahir, Christoph Doblander, Ruben Mayer, Sebastian Frischbier, and Hans-Arno Jacobsen. The debs 2021 grand challenge: analyzing environmental impact of worldwide lockdowns. In *Proceedings of the 15th ACM International Conference on Distributed and Event-Based Systems*, pages 136–141, 2021.

[20] Georg Heiler. *Efficient temporal graph analytics: Using large scale telecommunication data for mobility modeling and infrastructure maintenance*. PhD thesis, Technische Universität Wien, 2022.

[21] Armir Bujari, Alessandro Calvio, Luca Foschini, Andrea Sabbioni, and Antonio Corradi. Ippodamo: A digital twin support for smart cities facility management. In *Proceedings of the Conference on Information Technology for Social Good*, pages 49–54, 2021.

[22] The Apache Software Foundation. Apache sedona website. https://sedona.apache.org, 2025. Accessed: 2025-06-08.

[23] PyPI. Download statistics of the geospark python package. https://pepy.tech/projects/geospark, 2025. Accessed: 2025-06-08.

[24] PyPI. Download statistics of the apache sedona python package. https://pepy.tech/projects/apache-sedona, 2025. Accessed: 2025-06-08.

[25] Wherobots. Wherobots website. https://wherobots.com/, 2025. Accessed: 2025-06-08.

[26] The Overture Maps Foundation. The overture maps foundation website. https://overturemaps.org/, 2025. Accessed: 2025-06-08.