

Resource Summary Report

Generated by SPARC SAWG on Sep 17, 2026

Apache Spark

RRID:SCR_016557

Type: Tool

Proper Citation

Apache Spark (RRID:SCR_016557)

Resource Information

URL: <http://spark.apache.org/>

Proper Citation: Apache Spark (RRID:SCR_016557)

Description: Software tool as a unified analytics engine for large scale data processing. An open source distributed general purpose cluster computing framework.

Resource Type: data analytics software, data processing software, software application, software resource

Defining Citation: [DOI:10.1145/2934664](https://doi.org/10.1145/2934664)

Keywords: unified, analytics, engine, large, scale, data, processing

Availability: Free, Available for download, Freely available

Resource Name: Apache Spark

Resource ID: SCR_016557

License: Apache License, Version 2.0.

Record Creation Time: 20220129T080331+0000

Record Last Update: 20260912T200104+0000

Ratings and Alerts

No rating or validation information has been found for Apache Spark.

No alerts have been found for Apache Spark.

Data and Source Information

Usage and Citation Metrics

We found 43 mentions in open access literature.

Listed below are recent publications. The full list is available at [SPARC SAWG](#) .

Wlasnowolski M, et al. (2026) enhancer3D: 3D chromatin structures and enhancer-promoter distance profiles for archaic and modern human genomes. Nucleic acids research, 54(D1), D1046.

Blazyte A, et al. (2026) Fifteen Years of the Genome Analysis Toolkit as the De Facto Standard in Short-Read Variant Calling. International journal of molecular sciences, 27(9).

Boerman JP, et al. (2025) Data processing techniques to improve data integration from dairy farms. JDS communications, 6(3), 339.

Stojanov R, et al. (2025) Applicability Assessment of Technologies for Predictive and Prescriptive Analytics of Nephrology Big Data. Proteomics, 25(11-12), e202400135.

Nyström-Persson J, et al. (2025) Precise and scalable metagenomic profiling with sample-tailored minimizer libraries. NAR genomics and bioinformatics, 7(2), lqaf076.

Prasanna S, et al. (2024) A scalable tool for analyzing genomic variants of humans using knowledge graphs and graph machine learning. Frontiers in big data, 7, 1466391.

Hagan NKA, et al. (2024) SparkDWM: a scalable design of a Data Washing Machine using Apache Spark. Frontiers in big data, 7, 1446071.

Alonso R, et al. (2024) Improving digital twin experience through big data, IoT and social analysis: An architecture and a case study. Heliyon, 10(2), e24741.

Yönyül B, et al. (2024) MonARCh: an actor based architecture for dynamic linked data monitoring. PeerJ. Computer science, 10, e2133.

Özgüven YM, et al. (2023) Distributed messaging and light streaming system for combating pandemics: A case study on spatial analysis of COVID-19 Geo-tagged Twitter dataset. Journal of ambient intelligence and humanized computing, 14(2), 773.

Shih CC, et al. (2023) A five-safes approach to a secure and scalable genomics data repository. iScience, 26(4), 106546.

Umair A, et al. (2023) Vaccine sentiment analysis using BERT + NBSVM and geo-spatial approaches. The Journal of supercomputing, 1.

Perova Z, et al. (2023) PDCM Finder: an open global research platform for patient-derived cancer models. Nucleic acids research, 51(D1), D1360.

Almeida A, et al. (2023) Time series big data: a survey on data stream frameworks, analysis and algorithms. Journal of big data, 10(1), 83.

Almalki J, et al. (2022) A machine learning-based approach for sentiment analysis on distance learning from Arabic Tweets. PeerJ. Computer science, 8, e1047.

Chhetri TR, et al. (2022) Knowledge Graph Based Hard Drive Failure Prediction. Sensors (Basel, Switzerland), 22(3).

Brard R, et al. (2022) A Novel Walking Activity Recognition Model for Rotation Time Series Collected by a Wearable Sensor in a Free-Living Environment. Sensors (Basel, Switzerland), 22(9).

Perez-Riverol Y, et al. (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic acids research, 50(D1), D543.

Schubotz M, et al. (2022) Caching and Reproducibility: Making Data Science Experiments Faster and FAIRer. Frontiers in research metrics and analytics, 7, 861944.

Laurie S, et al. (2022) The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases. Human mutation, 43(6), 717.