Apache MADlib

MADlib graduated to an Apache Top Level Project on 7/19/17. Read the press release.

Apache MADlib® is an open-source library for scalable in-database analytics.

It provides data-parallel implementations of mathematical, statistical,

graph and machine learning methods for structured and unstructured data.

Quick Start Guides

- Installation Guide
- Quick Start Guide for Users
- Quick Start Guide for Developers
- Quick start Jupyter notebooks for many MADlib algorithms

General Information

- MADlib website
- Greenplum database YouTube channel with MADlib content
- Module and algorithm documentation
- FAQ

Developer Documentation

blocked URL

- Source code repo
- Contribution Guidelines
- Documentation Guide (Doxygen)
- Ideas for contribution

- Algorithm technical design document

Architecture

- Architecture
- Module Anatomy

Release Notes

- Release notes
- Downloads

Licensing

License information regarding MADlib and included third-party libraries can be found inside the license directory. ASF licensing guidance for MADlib pertaining to its pre-Apache history as an open source project with BSD licensing is described here.

Papers

- MAD Skills : New Analysis Practices for Big Data (VLDB 2009)
- Hybrid In-Database Inference for Declarative Information Extraction (SIGMOD 2011)
- Towards a Unified Architecture for In-Database Analytics (SIGMOD 2012)
- The MADlib Analytics Library or MAD Skills, the SQL (VLDB 2012)

Related Software

- PivotalR - lets the user run the functions of the open-source big-data machine learning package MADlib directly from R.
- PyMADlib - a nascent Python wrapper for MADlib, which brings you the power and flexibility of python with the number crunching power of MADlib.