

Apache MADlib Tutorial

Apache MADlib Tutorial

Introduction

Apache MADlib is an open-source library for scalable in-database analytics. It provides data-parallel implementations of mathematical, statistical and machine learning algorithms.

Apache MADlib is primarily for data scientists (that is you) who are working with very large datasets. MADlib really shines when datasets are really large and need MPP (Massively Parallel Processing) Architecture in which to operate them.

MAD in MADlib mean : Magnetic Agile & Deep



Magnetic



All your data in one place

Agile



Fast, progressive queries and code

Deep



Sophisticated data science with large data sets

Three features of MADlib

- **Magnetic** – Bringing all your analysts and organisational data to one place.
- **Agile** – Enabling the Analyst to quickly develop hypothesis and test them with statistical algorithms and build code in an iterative fashion over the statistical methods.

- **Deep** – The statistical methods enable quite deep drill down with their sophisticated machine learning probabilistic methods.

How does MADlib handle large data sets ?

MADlib is developed in such a way that it can take advantage of the distributed nature of an MPP Database. And also to implement the algorithms in an efficient way w.r.t. processing, scale and network bandwidth.

MADlib is build in SQL for the simple reason that people have data in SQL databases and they are working with that data through SQL interfaces. MADlib brings machine learning technology to data.

Where is MADlib used ?

Because its a general purpose library, MADlib is used across a wide variety of industries from manufacturing, financial services, government and health care and many more.

Where does MADlib fit with the big data community ?

MADlib brings machine learning technology to environment where we have SQL, storage and scalability. And in the big data open source community you can get all these SQL, storage and scalability now decoupled into different pieces and put them together in interesting ways.

Apache Big Data Tutorials

- ◆ [Learn Apache Hadoop](#)
- ◆ [Learn Apache Spark](#)
- ◆ [Learn Apache Flink](#)
- ◆ [Learn Apache Beam](#)
- ◆ [Learn Apache MXNet](#)

Apache Database Projects

- ◆ [Apache CouchDB Tutorial](#)

Apache Other Projects

- ◆ [Apache Tomcat Tutorial](#)
- ◆ [Apache PDFBox Tutorial](#)

