Apache Mahout

Functional Area: Data Analytics

Overview:

Provides scalable machine learning algorithms for three primary applications: classification, clustering and recommendation mining. Within each of these areas, different algorithms are provided. For example, within Classification, Mahout offers algorithms for Naïve Bayes, Hidden Markov Models, Logistic Regression and Random Forests.

Mahout is intended as a scalable, distributed solution, but also includes single node contributions. While many of the Mahout algorithms originally utilized the Apache Hadoop platform, the community has announced that all future development will be done in a new Domain Specific Language for linear algebra designed to run in parallel on Apache Spark.

Mahout began as part of the Apache Lucene information retrieval project, and became an independent project in 2010. The original goal of Mahout (not yet complete) was to implement the 10 algorithms included in the paper “Map-Reduce for Machine Learning on Multicore” [1]

1. Chu, C. T., et al. (2006). Map-Reduce for Machine Learning on Multicore. NIPS. B. Schölkopf, J. C. Platt and T. Hoffman, MIT Press**:** 281--288.