**MLlib**

[MLlib](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=https://spark.apache.org/mllib/&usg=ALkJrhi8v7SFqBcM-Vk47Lv7-H_Q_FwW9g) adalah perpustakaan pembelajaran mesin yang menyediakan berbagai algoritma yang dirancang untuk skala cluster untuk klasifikasi, regresi, clustering, penyaringan kolaboratif, dan sebagainya (lihat artikel Toptal tentang [pembelajaran mesin](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=https://www.toptal.com/machine-learning/machine-learning-theory-an-introductory-primer&usg=ALkJrhh3sTcRT_x01ioCmnacGOg_Szkxkw) untuk informasi lebih lanjut mengenai topik itu). Beberapa algoritma ini juga bekerja dengan data streaming, seperti regresi linier menggunakan kuadrat terkecil biasa atau k-means clustering (dan lebih banyak lagi di jalan). Apache [Mahout](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=http://mahout.apache.org/&usg=ALkJrhjICT3NuDFroB8PzGglU3efS8StNQ) (mesin belajar perpustakaan untuk Hadoop) telah berpaling dari MapReduce dan bergabung pada Spark MLlib.

**MLlib** adalah perpustakaan pembelajaran scalable Apache Spark.

Kemudahan penggunaan

Digunakan di Jawa, Scala, Python, dan R.

MLlib cocok dengan API [Spark](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=https://spark.apache.org/&usg=ALkJrhgxvvqU67K1RWpICM3An6hWyd8efA) dan beroperasi dengan [NumPy](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=http://www.numpy.org/&usg=ALkJrhgogfcqiMe4_bAYGbdJkwYu3Gl3ow) dengan Python (seperti Spark 0.9) dan perpustakaan R (seperti Spark 1.5). Anda dapat menggunakan sumber data Hadoop (misalnya file HDFS, HBase, atau lokal), sehingga mudah dihubungkan ke alur kerja Hadoop.

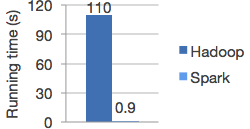
data = spark.read.format ( "libsvm" ) \   
.load ( "hdfs: // ..." )   
  
model = KMeans (k = 10) .fit (data)

Memanggil MLlib dengan Python

Kinerja

Algoritma berkualitas tinggi, 100x lebih cepat dari MapReduce.

Spark unggul pada perhitungan iteratif, memungkinkan MLlib berlari dengan cepat. Pada saat yang sama, kita peduli dengan kinerja algoritmik: MLlib berisi algoritme berkualitas tinggi yang memanfaatkan iterasi, dan dapat menghasilkan hasil yang lebih baik daripada perkiraan satu-pass yang terkadang digunakan pada MapReduce.



Regresi logistik di Hadoop dan Spark

Mudah disebarkan

Berjalan pada kelompok dan data Hadoop yang ada.

Jika Anda memiliki cluster Hadoop 2, Anda dapat menjalankan Spark dan MLlib tanpa pra-instalasi. Jika tidak, Spark mudah dijalankan secara [standalone](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=https://spark.apache.org/docs/latest/spark-standalone.html&usg=ALkJrhhuq8GhT1fvuk4anJiFUHGOeiUokw) atau [EC2](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=https://spark.apache.org/docs/latest/ec2-scripts.html&usg=ALkJrhgQ6n8Hhy0k_p3VxIqcbBfha7h0QQ) atau [Mesos](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=https://mesos.apache.org/&usg=ALkJrhj1bhORtWsk55HEYv753aHU7EA-vQ) . Anda bisa membaca dari sumber data[HDFS](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html&usg=ALkJrhhui2JtDGZ3eIUDDQCq8jucJYQNLw) , [HBase](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=https://hbase.apache.org/&usg=ALkJrhif2XY4G0JGBYvSSjxnP0AH_-b98A) , atau Hadoop lainnya.



Algoritma

MLlib berisi banyak algoritma dan utilitas.

Algoritma ML meliputi:

* Klasifikasi: regresi logistik, naif Bayes, ...
* Regresi: regresi linier berganda, regresi survival, ...
* Pohon keputusan, hutan acak, dan pohon berpendukung gradien
* Rekomendasi: alternating least squares (ALS)
* Clustering: K-means, Gaussian mixtures (GMMs), ...
* Pemodelan topik: alokasi Dirichlet laten (LDA)
* Frequent itemset, aturan asosiasi, dan pertambangan pola berurutan

Utilitas alur kerja ML meliputi:

* Transformasi fitur: standardisasi, normalisasi, hashing, ...
* Konstruksi pipa ML
* Evaluasi model dan tuning hyper-parameter
* ML ketekunan: model saving dan loading dan Pipelines

Utilitas lainnya meliputi:

* Aljabar linear terdistribusi: SVD, PCA, ...
* Statistik: statistik ringkasan, pengujian hipotesis, ...