**Apa itu Apache Spark? Sebuah pengantar**

[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) adalah proyek Apache yang diiklankan sebagai "cluster cluster cepat kilat". Ini memiliki komunitas open-source yang berkembang dan merupakan proyek Apache yang paling aktif saat ini.

Spark menyediakan platform pengolahan data yang lebih cepat dan lebih umum. Spark memungkinkan Anda menjalankan program hingga 100x lebih cepat di memori, atau 10x lebih cepat pada disk, dari pada Hadoop.Tahun lalu, Spark mengambil alih Hadoop dengan menyelesaikan kontes TB Daytona GreySort 100 TB lebih cepat pada sepersepuluh dari jumlah mesin dan juga menjadi [mesin sumber terbuka tercepat untuk memilah petabyte](https://translate.googleusercontent.com/translate_c?depth=1&hl=id&prev=search&rurl=translate.google.co.id&sl=en&sp=nmt4&u=http://databricks.com/blog/2014/10/10/spark-petabyte-sort.html&usg=ALkJrhg1onK_9GMySIjbUb70LiLyTgRVDw) .

Spark juga memungkinkan untuk menulis kode lebih cepat karena Anda memiliki lebih dari 80 operator tingkat tinggi yang Anda inginkan. Untuk menunjukkan ini, mari kita lihat "Hello World!" Dari BigData: contoh Word Count. Ditulis di Jawa untuk MapReduce memiliki sekitar 50 baris kode, sedangkan di Spark (dan Scala) Anda dapat melakukannya sesederhana ini:

sparkContext.textFile("hdfs://...") .flatMap(line => line.split(" ")) .map(word => (word, 1)).reduceByKey(\_ + \_) .saveAsTextFile("hdfs://...")

Aspek penting lainnya saat belajar bagaimana menggunakan Apache Spark adalah shell interaktif (REPL) yang menyediakan kotak out-of-the-box. Dengan menggunakan REPL, seseorang dapat menguji hasil setiap baris kode tanpa terlebih dahulu harus melakukan kode dan melaksanakan keseluruhan pekerjaan. Jalan menuju kode kerja jauh lebih singkat dan analisis data ad-hoc dimungkinkan.

Fitur kunci tambahan dari Spark meliputi:

* Saat ini menyediakan API di Scala, Java, dan Python, dengan dukungan untuk bahasa lain (seperti R) di perjalanan
* Terintegrasi dengan ekosistem Hadoop dan sumber data (HDFS, Amazon S3, Hive, HBase, Cassandra, dll.)
* Bisa berjalan di cluster yang dikelola oleh Hadoop YARN atau Apache Mesos, dan bisa juga berjalan standalone

Inti Spark dilengkapi oleh seperangkat perpustakaan tingkat tinggi yang kuat yang dapat digunakan secara mulus dalam aplikasi yang sama. Perpustakaan ini saat ini meliputi SparkSQL, Spark Streaming, MLlib (untuk pembelajaran mesin), dan GraphX, yang masing-masing diperinci lebih lanjut dalam artikel ini. Perpustakaan dan ekstensi Spark tambahan saat ini sedang dikembangkan juga.

