Bab 14
PENGENALAN HADOOP

Dani Anggoro, S.Kom., M.Kom

14.1 Pendahuluan

Hadoop adalah kerangka kerja perangkat lunak open source berbasis Java yang digunakan untuk aplikasi komputasi data besar secara intensif, dengan Hadoop File System (HDFS) sebagai bagian intinya. HDFS adalah sistem file terdistribusi yang sangat fault-tolerant dan dirancang untuk digunakan dengan hardware yang terjangkau.

Dalam konteks platform perangkat lunak, Hadoop adalah engine analitik yang memungkinkan pengguna untuk melakukan penulisan perintah dan menjalankan aplikasi pemrosesan data dalam jumlah besar. Hadoop terdiri dari dua komponen utama: HDFS (Hadoop Distributed File System) dan MapReduce.

14.2 HDFS

HDFS digunakan untuk menyimpan data dalam cluster, yang terdiri dari banyak node atau komputer/server. Cluster ini harus terinstalasi Hadoop pada setiap nodenya. Dalam Hadoop versi 1.x, ada beberapa jenis node dalam cluster, termasuk Name Node, Data Node, Secondary Name Node, dan lainnya. Namun, dalam Hadoop versi 2.x, terdapat perubahan signifikan, seperti lebih dari satu Name Node untuk meningkatkan ketersediaan tinggi.

Dengan perubahan Hadoop versi 2.x, terdapat peningkatan dalam ketersediaan tinggi dengan lebih dari satu Name Node, dan komponen seperti Secondary Name Node dan Backup Node menjadi opsional.

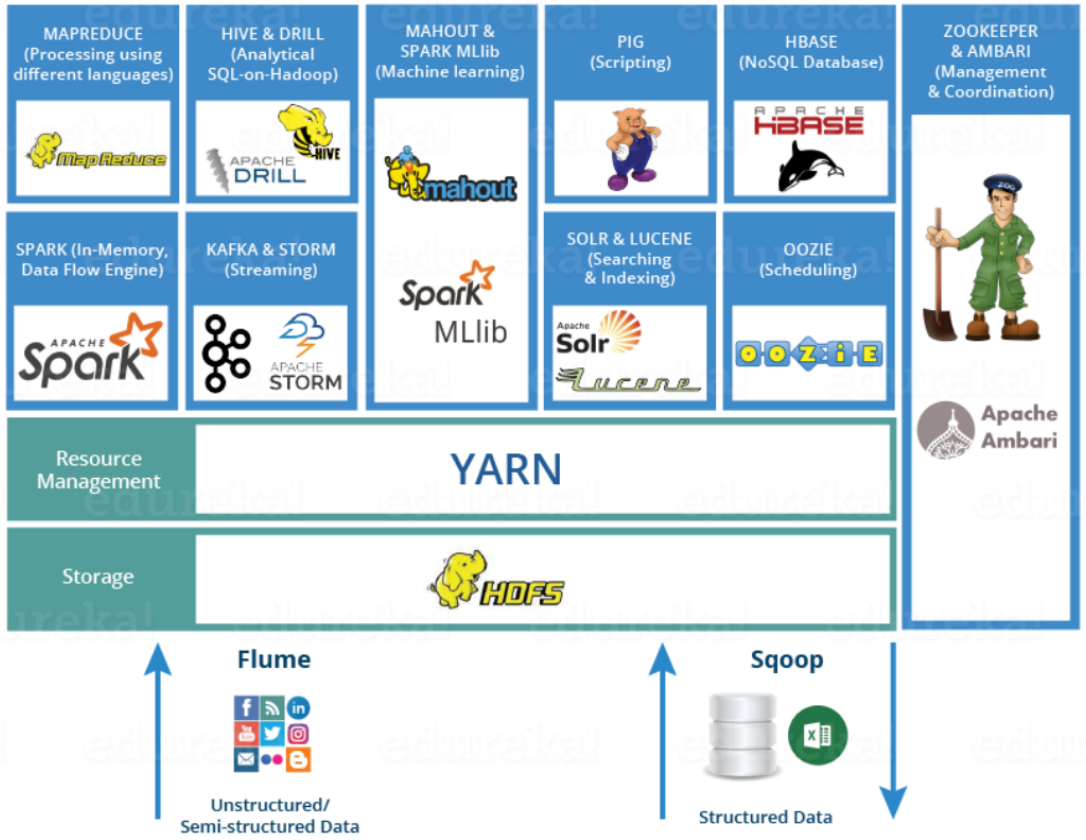
14.3 Map Reduce

MapReduce adalah model pemrograman yang dikembangkan oleh Google, yang awalnya digunakan untuk aplikasi pencarian dan pengindeksan internal Google. Model pemrograman ini termasuk gaya pemrograman fungsional yang dapat diparalelkan dalam sekumpulan besar workstation atau PC. Komponen utama dalam MapReduce adalah Job Tracker dan Task Tracker.

Dalam Job Tracker, tugas utamanya adalah menerima permintaan pekerjaan dari klien, mengelola lokasi data, menjadwalkan eksekusi program di berbagai mesin untuk pemrosesan data, mengatasi kegagalan mesin, dan mengelola komunikasi antar-mesin. Task Tracker, di sisi lain, menjalankan tugas yang diberikan oleh Job Tracker pada data yang ada di masing-masing mesin.

14.3 Ekosistem Hadoop

Untuk membangun Ekosistem Hadoop yang sukses, langkah pertama yang penting adalah mengidentifikasi kegunaan dan interaksi antara berbagai Tool Big Data. Hal ini mencakup memahami bagaimana setiap alat Big Data berinteraksi satu sama lain dan mengidentifikasi alat mana yang akan digunakan dalam implementasi di lingkungan Hadoop.



Gambar 14.1 : Ekosistem Hadoop

Gambar 14.1 memberikan gambaran tentang interaksi aplikasi, alat, dan antarmuka dalam ekosistem Hadoop. Ini membantu dalam kategorisasi alat berdasarkan fungsi mereka, seperti penyimpanan, pemrosesan, kueri, integrasi eksternal, dan koordinasi. Visualisasi ekosistem ini membantu dalam memahami interaksi antara alat-alat Big Data.

Ekosistem Hadoop juga dapat disesuaikan dengan berbagai alat Big Data yang berbeda, tergantung pada kebutuhan projek analitik dan kompleksitas kasus yang ditangani. Ekosistem dapat dirancang dan dibangun dengan menggabungkan berbagai

alat Big Data sesuai dengan kebutuhan, dan arsitektur dapat dimodifikasi atau diperbarui sesuai kebutuhan dari waktu ke waktu.

Ada beberapa distribusi Hadoop yang dapat digunakan, seperti Cloudera, HortonWorks, MapR Technologies, dan lainnya. Masing-masing memiliki karakteristik dan manfaatnya sendiri. Membangun Hadoop Distribution sendiri memerlukan pemahaman tentang konsep Big Data dan alat-alat yang digunakan untuk Big Data.

Dengan memahami interaksi alat Big Data dan membangun ekosistem Hadoop yang sesuai, kita dapat mengelola dan menganalisis data dalam skala besar dengan lebih efisien dan efektif. Hadoop terus berkembang dan menjadi alat yang sangat penting dalam bidang analitik data.

14.4 Kesimpulan

Dalam mengolah big data, Hadoop adalah kerangka kerja perangkat lunak open source yang sangat berguna. Hadoop menyediakan dua komponen utama, yaitu Hadoop Distributed File System (HDFS) dan MapReduce. HDFS digunakan untuk menyimpan data dalam cluster yang terdiri dari banyak node atau komputer/server. Dalam versi Hadoop 1.x, terdapat beberapa jenis node dalam cluster, termasuk Name Node, Data Node, Secondary Name Node, dan lainnya. Namun, dalam Hadoop versi 2.x, terdapat peningkatan dalam ketersediaan tinggi dengan lebih dari satu Name Node.

MapReduce adalah model pemrograman yang dikembangkan oleh Google untuk pemrosesan data secara terdistribusi. Komponen utama dalam MapReduce adalah Job Tracker dan Task Tracker. Job Tracker bertanggung jawab atas manajemen pekerjaan, pengelolaan lokasi data, dan penjadwalan eksekusi program di berbagai mesin. Task

Tracker menjalankan tugas-tugas yang diberikan oleh Job Tracker pada data yang ada di masing-masing mesin.

Dengan penggunaan Hadoop, perusahaan dapat mengelola dan menganalisis data dalam jumlah besar dengan efisien. Hadoop memungkinkan pengguna untuk melakukan pemrosesan data yang sangat intensif dan menangani pekerjaan yang terdistribusi. Meskipun Hadoop versi 1.x memiliki beberapa keterbatasan, seperti masalah ketersediaan jika Name Node mati, versi 2.x mengatasi masalah ini dengan lebih dari satu Name Node.

Keseluruhan, Hadoop adalah alat yang sangat penting dalam mengelola dan menganalisis big data, dan terus berkembang untuk meningkatkan kinerja dan ketersediaan.