MAKALAH ANALISIS BIG DATA

TUGAS MAPREDUCE

Dosen Pengampu: Sevi Nurafni ST., M.Si., M.Sc.

Disusun Oleh Ryan Fadhilah Faizal Hakim

Program Studi Sains Data

Fakultas Sains dan Teknologi

Mengapa kita membutuhkan MapReduce?

MapReduce memungkinkan programmer yang terampil untuk menulis aplikasi terdistribusi tanpa harus khawatir tentang infrastruktur komputasi terdistribusi yang mendasarinya. Ini adalah hal yang sangat penting: Hadoop dan kerangka kerja MapReduce menangani segala macam kompleksitas yang tidak perlu ditangani oleh pengembang aplikasi. Misalnya, kemampuan untuk secara transparan meningkatkan skala kluster dengan menambahkan node dan failover otomatis dari subsistem penyimpanan data dan pemrosesan data terjadi tanpa dampak apa pun pada aplikasi.

MapReduce memungkinkan pemrosesan data secara paralel di kluster mesin dengan membagi data menjadi potongan-potongan dan memproses potongan-potongan tersebut secara bersamaan sehingga dapat meningkatkan kecepatan penanganan data bahkan hingga petabyte. MapReduce juga dapat menangani data yang sangat besar dengan menambahkan lebih banyak mesin dalam kluster sehingga memungkinkan MapReduce untuk memproses data dalam skala yang tidak dapat ditangani oleh sistem pemrosesan data tradisional.

Apa saja langkah-langkah utama dari MapReduce?

Terdapat langkah-langkah utama dari MapReduce di antaranya:

1. Data Collection (Input)
   Langkah pertama dalam proses MapReduce adalah mengumpulkan dan menyiapkan data yang akan diproses sebagai input. Data ini dapat diperoleh dari berbagai sumber seperti file log, basis data, atau hasil pengumpulan data dari aplikasi.

   Langkah-langkah dalam Data Collection:
   - Kumpulkan Data: Data dapat dikumpulkan dari berbagai sumber, seperti server log, data sosial media, atau basis data.
   - Simpan Data: Data biasanya disimpan dalam sistem file terdistribusi seperti Hadoop Distributed File System (HDFS), yang memungkinkan data dipecah menjadi potongan-potongan yang lebih kecil dan disebar ke banyak mesin.

2. Map Phase (Tahap Pemetaan)
   Pada tahap ini, data yang telah dikumpulkan diproses menggunakan fungsi Map. Fungsi ini menerima input dalam bentuk pasangan kunci-nilai dan menghasilkan pasangan kunci-nilai baru yang akan diproses lebih lanjut.

   Langkah-langkah dalam Map Phase:
   - Input Data: Data yang disimpan dalam HDFS dibaca oleh MapReduce framework. Data ini dibagi menjadi potongan-potongan kecil yang disebut split.
   - Terapkan Fungsi Map: Fungsi Map diterapkan ke setiap split data. Fungsi ini melakukan operasi seperti pemetaan data, penyaringan, atau transformasi. Misalnya, dalam analisis frekuensi kata, fungsi Map akan membaca kata dari setiap dokumen dan menghasilkan pasangan kunci-nilai berupa (kata, 1).

3. Shuffle and Sort Phase (Tahap Pengacakan dan Pengurutan)
   Setelah tahap pemetaan, pasangan kunci-nilai yang dihasilkan dari fungsi Map akan dipindahkan ke tahap berikutnya. Proses ini melibatkan pengacakan dan pengurutan data untuk persiapan tahap berikutnya.

   Langkah-langkah dalam Shuffle and Sort Phase:
   - Shuffle: Pasangan kunci-nilai yang dihasilkan selama fase Map dikumpulkan dan dikelompokkan berdasarkan kunci. Ini memastikan bahwa semua pasangan dengan kunci yang sama akan berada di tempat yang sama untuk diproses dalam tahap berikutnya.
   - Sort: Data dikelompokkan dan diurutkan berdasarkan kunci, sehingga pasangan kunci-nilai yang sama akan berada dalam urutan yang teratur.

4. Reduce Phase (Tahap Pengurangan)
   Pada tahap ini, data yang sudah dikelompokkan dan diurutkan dari tahap shuffle dan sort diproses menggunakan fungsi Reduce. Fungsi ini mengolah semua nilai yang memiliki kunci yang sama untuk menghasilkan hasil akhir dari proses MapReduce.

   Langkah-langkah dalam Reduce Phase:
   - Terapkan Fungsi Reduce: Fungsi Reduce menerima kunci dan kumpulan nilai terkait dengan kunci tersebut. Fungsi ini biasanya melakukan operasi agregasi, seperti menjumlahkan nilai atau menggabungkan data.
   - Hasilkan Output: Fungsi Reduce menghasilkan output akhir berupa pasangan kunci-nilai yang merupakan hasil akhir dari pemrosesan data.

5. Output Data (Output Generation)
   Setelah tahap Reduce, hasil pemrosesan data disimpan atau dikirim ke sistem lain untuk analisis lebih lanjut atau pembuatan laporan.