Nama : Muhammad Rifky Harto Biantoro

Kelas : TI-3D

NIM : 2241720176

Mata Kuliah : Big Data

Dosen Pengampu : Muhammad Hasyim Ratsanjani S.Kom., M.Kom

Big Data

Seiring dengan bertambahnya tahun, kebutuhan akan teknologi dalam segala bidang semakin berkembang secara masif. Hal ini mendorong para ahli teknologi untuk berusaha memenuhi kebutuhan teknologi tersebut, dalam kasus ini adalah big data. Big data merupakan jawaban bagi kebutuhan dalam penyimpanan dan pengolahan data dimana big data timbul karena kebutuhan untuk penyimpanan dan pengolahan data dalam skala yang besar. Salah satu contoh penerapan penggunaan big data adalah pada platform Netflix, platform ini menggunakan big data untuk memenuhi kebutuhan usernya yang berada di seluruh dunia. Hal ini mencakup kebutuhan untuk menyimpan dan menampilkan ribuan rekomendasi film yang diberikan sesuai dengan selera user. Netflix menggunakan layanan cloud Amazon Web Services (AWS) yang dapat memenuhi kebutuhan Netflix untuk mendeploy filmnya agar dapat diakses dari seluruh belahan dunia. Netflix tidak hanya menggunakan layanan cloud untuk menyimpan, tetapi juga melakukan pengolahan data berdasarkan riwayat tontonan pengguna sehingga Netflix dapat memberikan rekomendasi yang sesuai dengan tontonan yang pernah diakses pengguna. Hal ini jelas menjadi nilai tambah bagi user dimana user merasa bahwa platform Netflix dapat memberikannya rekomendasi tontonan kesukaan usernya tanpa harus user mencarinya sendiri. Berbicara tentang big data juga tidak lepas dari *framework* yang dibuat untuk memudahkan dalam pengelolaan dan pengorganisiran data, berikut beberapa penjelasan sederhana terkait arsitektur *framework* yang biasa digunakan.

A. Apache Hadoop

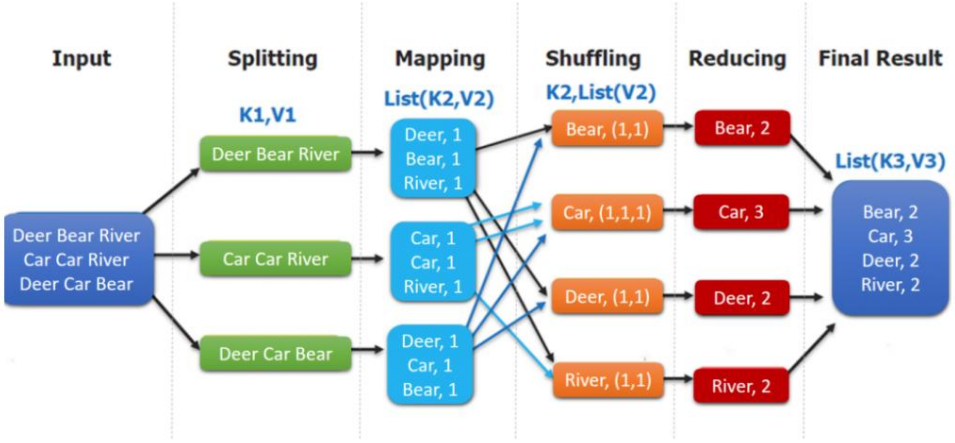
   Apache Hadoop merupakan *framework* yang memungkinkan untuk melakukan distribusi pengolahan data dalam jumlah yang besar lintas komputer menggunakan sintaks kode program yang sederhana. Didesain untuk meningkatkan skala komputer server dari skala satu sampai ribuan mesin komputer server, dimana setiap komputer servernya menawarkan penyimpanan dan komputasi lokal. Hadoop dikembangkan pada awal 2000-an oleh Doug Cutting dan Mike Cafarella yang terinspirasi oleh makalah yang dipublikasi oleh Google mengenai *Google File System* dan model pemrograman *MapReduce.* Didalam Apache Hadoop, terdapat modul ekosistem yang didesain untuk mengelola big data perusahaan, diantaranya:

   a. HDFS

      HDFS (Hadoop Distributed File System) adalah sistem file terdistribusi portabel yang dirancang untuk berjalan pada komoditas perangkat keras, ditulis dalam bahasa Java. Pada penggunaanya, HDFS menyediakan shell command dan antarmuka API Java yang digunakan untuk menyimpan data.

b. MapReduce

MapReduce merupakan modul lainnya pada Apache Hadoop yang digunakan untuk membantu dalam memaksimalkan pengolahan big data. Jika HDFS digunakan untuk menyimpan, maka MapReduce digunakan untuk mengolah data dalam skala besar. MapReduce memiliki 2 tugas utama yang dibagi berdasarkan fase, yaitu Map dan Reduce. Map bertugas untuk mendistribusikan pemrosesan data antara komputer yang berbeda untuk mendapatkan key-value, setelah proses mapping selesai, Reduce bertugas untuk meringkas data yang sama dan mengeluarkan value datanya. Secara detail, MapReduce memiliki beberapa langkah sepert gambar berikut:



- Splitting, Proses membagi data menjadi blok berdasarkan baris (dalam contoh diatas merupakan kalimat)
- Map, memetakan baris data menjadi sub data (dalam contoh diatas merupakan kata)
- Shuffling, mengumpulkan data yang sama (dalam contoh diatas, kata yang sama dikumpulkan dan menampilkan value nya)
- Reduce, menjumlahkan dan menampilkan total nilai data (dalam kasus diatas, menjumlahkan data berdasrkan informasi yang didapatkan pada proses Shuffling)