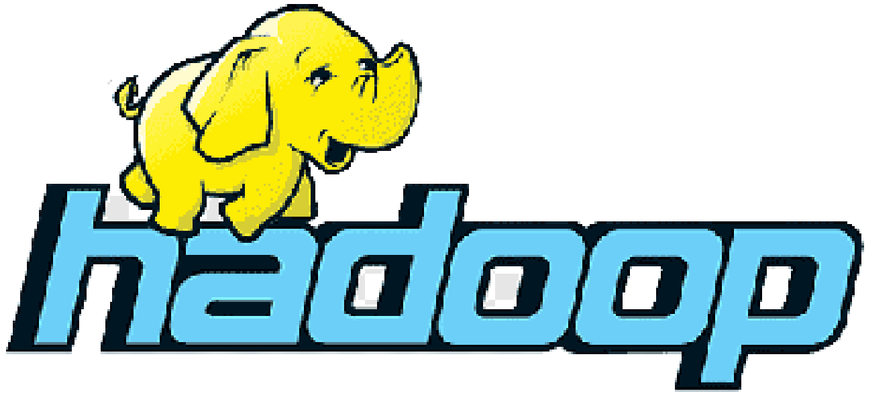
Learning Progress Review

WEEK 7

Charlie's Angels Team

DigitalSkola



Salah satu software platform yang bisa digunakan untuk mengelola Big Data adalah Hadoop. Hadoop adalah library software yang menyerupai framework open source dari bahasa pemrograman Java di bawah lisensi Apache yang digunakan untuk melakukan pemrosesan Big Data menggunakan model pemrograman sederhana.

Kelebihan Hadoop

- **Volume**, adanya keperluan untuk menyimpan dan mengelola data dalam jumlah yang sangat besar dan terus bertambah dari waktu ke waktu
- **Velocity**, adanya keperluan untuk bisa mengakses data dalam jumlah besar dengan cepat
- **Variety**, semakin bervariasinya data saat ini, sehingga teknologi Relational Database Management System (RDBMS) sudah tidak mungkin menanganinya lagi

Kekurangan Hadoop

- Tidak cocok untuk OLTP (Online Transaction Processing)
- Tidak cocok untuk OLAP (Online Analytic Processing)
- Tidak cocok untuk DSS (Decission Support System)

Prinsip-Prinsip pada Hadoop

Hadoop mampu menggabungkan banyak komputer menjadi satu kesatuan, dimana dengan banyaknya penggabungan ini maka data akan disebar ke seluruh komputer yang ada untuk saling menjaga data di dalamnya agar tetap aman.

Hadoop memiliki sistem yang dapat membagi proses perhitungan atau komputasi yang biasanya memakan waktu yang sangat lama. Secara teknis, pada proses ini Hadoop menggunakan teknik map reduce yang dikoordinasikan dengan job tracker.

Sistem pada Hadoop mampu membagi beban penyimpanan ke berbagai komputer guna menyelamatkan data jika ada komputer yang mati. Sistem tersebut biasa dikenal dengan sebutan Hadoop Distributed File System (HDFS).

Data Flow Hadoop

1 Mendapatkan Big Data

Sumber data dapat berupa daftar ekstensif yang terstruktur, semi terstruktur, dan tidak terstruktur, beberapa streaming, sumber data realtime, sensor, perangkat, data yang diambil oleh mesin, dan banyak sumber lainnya.

2 Proses dan Struktur

Membersihkan, memfilter, dan mengubah data dengan menggunakan kerangka kerja berbasis MapReduce atau kerangka kerja lain yang dapat melakukan pemrograman terdistribusi di ekosistem Hadoop. Kerangka kerja yang tersedia saat ini adalah MapReduce, Hive, Pig, Spark dan sebagainya.

Data Flow Hadoop

3 Mendistribusikan Hasil

Data yang diproses dapat digunakan oleh BI dan sistem analitik atau sistem analitik data besar untuk melakukan analisis atau visualisasi.

4 Umpan Balik dan Pemeliharaan

Data yang dianalisis dapat diumpankan kembali ke Hadoop dan digunakan untuk peningkatan dan audit.

Komponen (Modul) Utama Framework Hadoop

Hadoop Distributed File System (HDFS)

File-system terdistribusi yang memfasilitasi penyimpanan data secara terdistribusi dalam kluster komputer.

Hadoop MapReduce

Sebuah sistem yang ditujukan untuk memproses data berukuran besar secara paralel.

Hadoop Common

Common utilitas yang digunakan untuk mendukung modul-modul Hadoop yang lainnya.

Hadoop YARN

Framework yang berperan dalam job scheduling dan resource management pada kluster Hadoop.

HDFS

HDFS berdasarkan konsep dari Google File System (GFS) dan oleh karenanya sangat mirip dengan GFS baik ditinjau dari konsep logika, struktur fisik, maupun cara kerjanya.

Sebagai layer penyimpanan data di Hadoop, HDFS adalah sebuah sistem file berbasis Java yang fault-tolerant, terdistribusi, dan scalable. Dirancang agar dapat diaplikasikan pada kluster dan dapat dijalankan dengan menggunakan proprietary atau commodity server

Sistem penyimpanan terdistribusi pada HDFS melakukan proses pemecahan file besar menjadi bagian-bagian lebih kecil dan kemudian didistribusikan ke kluster-kluster sehingga memungkinkan pemrosesan secara pararel

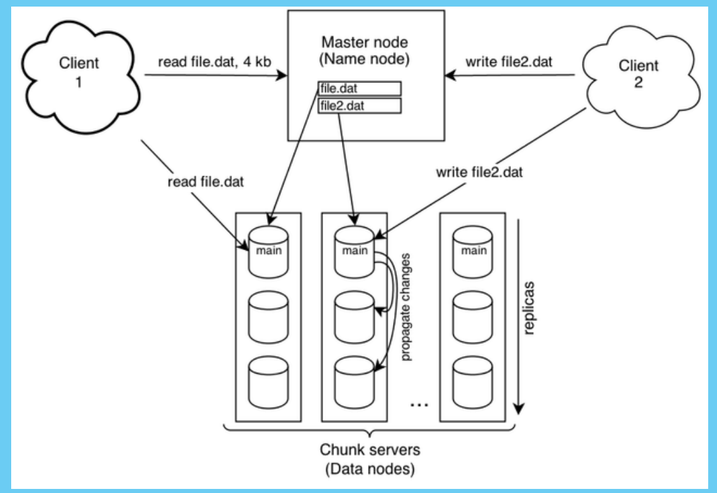
Feature HDFS

- Sangat sesuai untuk penyimpanan, pengelolaan dan pemrosesan dataset yang besar secara terdistribusi.
- Hadoop menyediakan antarmuka perintah untuk berinteraksi dengan HDFS.
- Heartbeat memudahkan pemeriksaan status kluster.
- Akses data melalui MapReduce streaming.

- HDFS menyediakan file permissions and authentication.
- Fault detection dan recovery.
- Lokasi komputasi berada dekat dengan data untuk mengurangi traffic jaringan dan meningkatkan throughput.

Cara Kerja HDFS

Sebuah kluster HDFS yang terdiri dari namenode sebagai pengelola metadata dari kluster, dan datanode yang menyimpan data.



Pada komputer yang sudah terhubung dengan kluster atau disebut sebagai client, penyimpanan data dilakukan dengan mengetikkan baris perintah pada console, kemudian file akan dikirim ke kluster dan disimpan pada node-node yang tersebar dalam kluster yang bertindak sebagai datanode.

Pada saat perintah membaca dieksekusi, client akan berkomunikasi dengan namenode untuk memperoleh nama dan alamat datanode yang harus diakses untuk mendapatkan data yang diinginkan.

DigitalSkola

MapReduce

MapReduce adalah sebuah model pemograman yang didesain untuk dapat melakukan pemrosesan data dengan jumlah yang sangat besar dengan cara membagi pemrosesan tersebut ke beberapa tugas yang indipenden satu sama lain.

Manfaat MapReduce

1. Skalabilitas. Bisnis dapat memproses petabyte data yang disimpan di Hadoop Distributed File System (HDFS).
2. Fleksibilitas. Hadoop memungkinkan akses yang lebih mudah ke berbagai sumber data dan berbagai jenis data.
3. Kecepatan. Dengan pemrosesan paralel dan pergerakan data minimal, Hadoop menawarkan pemrosesan data dalam jumlah besar dengan cepat.
4. Sederhana. Developer dapat menulis kode dalam pilihan bahasa, termasuk Java, C++ dan Python.

MapReduce Concept

UserID	BookID	Date
111	10001	20200101
112	10001	20200102
113	10002	20200103
114	10002	20200104
114	10003	20200105
115	10004	20200106
111	10005	20200107

MAPPER

111:10001	112:10001	113:10002	114:10002	114:10003	115:10004	111:10005

REDUCER

111:2	112:1	113:1	114:2	115:1

MapReduce Concept

Fungsi Map

Apa tugas fungsi map? Map berfungsi mengumpulkan semua informasi yang dimuat oleh potongan-potongan data dalam cluster di komputer. Nantinya, kumpulan informasi tersebut akan disalurkan ke proses Redue.
Dalam mengumpulkan data, Map juga akan membaca input ‘key/value’ lalu menghasilkan output yang sama: ‘key/value’. Pasangan ini disebut juga dengan ‘key/value intermediate’. Key/value intermediate lah yang akan disalurkan pula ke fungsi Reduce

MapReduce Concept

Fungsi Reduce

Reduce adalah fungsi setelah Map, di mana hasil dari Reduce akan dikirimkan ke user (pengguna). Fungsi reduce membaca pasangan Key/Value intermediate yang dihasilkan dari Map. Setelah membaca informasi, fungsi reduce akan mengelompokkan dan menggabung setiap Value dengan Key yang sama menjadi 1 kelompok. Pun output yang dikirimkan ke user tetap dalam bentuk key/value.

Aggregation on MapReduce

Agregasi pada MapReduce terdiri dari :

- Average
- Minimum
- Maximum

Aggerate Average

Contoh Kasus:

Kita memiliki data nilai siswa sekolah dasar. berisi nomor_induk, nama, mata_pelajaran, nilai

Kita ingin mencari rata-rata nilai pada setiap mata pelajaran yang ada

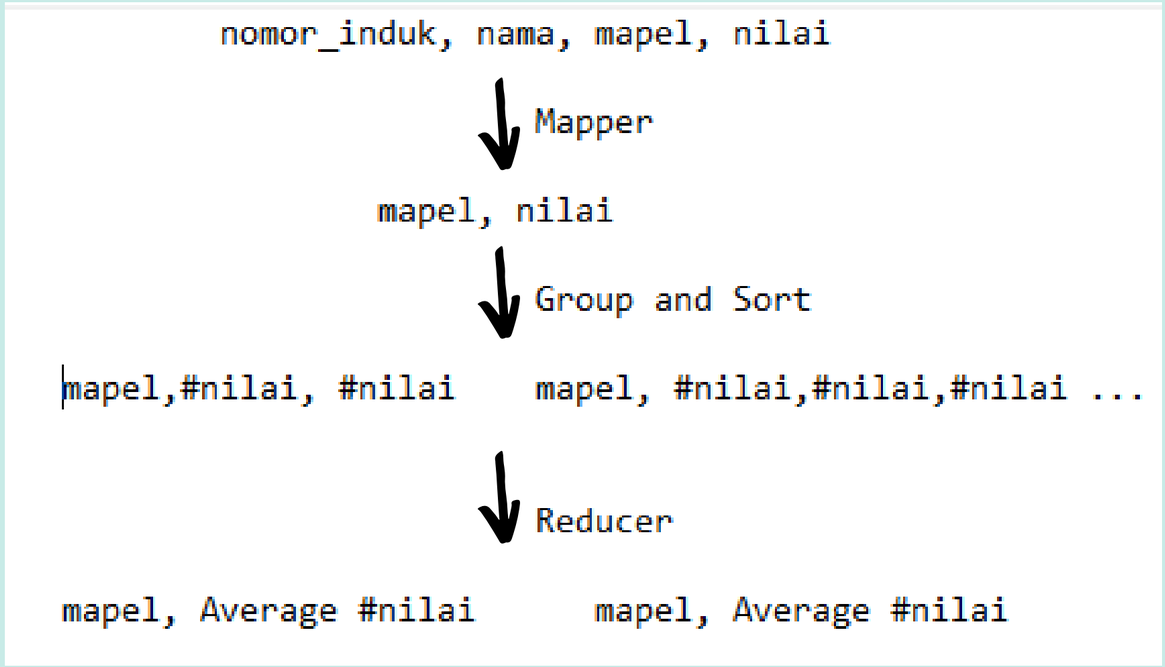
```
001, Yona, 1, 10
001, Yona, 2, 9
002, Sinta, 1, 9
003, Jojo, 2, 8
004, Fadil, 1, 10
005, Lola, 2, 9
005, Lola, 1, 10
006, Meli, 1, 8
```

Aggerate Average

Mapper akan mengambil informasi yang dibutuhkan saja yaitu mata pelajaran dan nilai

MapReduce akan melakukan Grouping dan Sorting untuk setiap mata pelajaran dan nilai

Reducer akan merata-rata nilai untuk setiap mata pelajaran



Aggerate Minimum

Contoh Kasus:

Kita memiliki data nilai siswa sekolah dasar. berisi nomor_induk, nama, mata_pelajaran, nilai

Kita ingin mencari nilai terendah pada setiap nama siswa yang ada

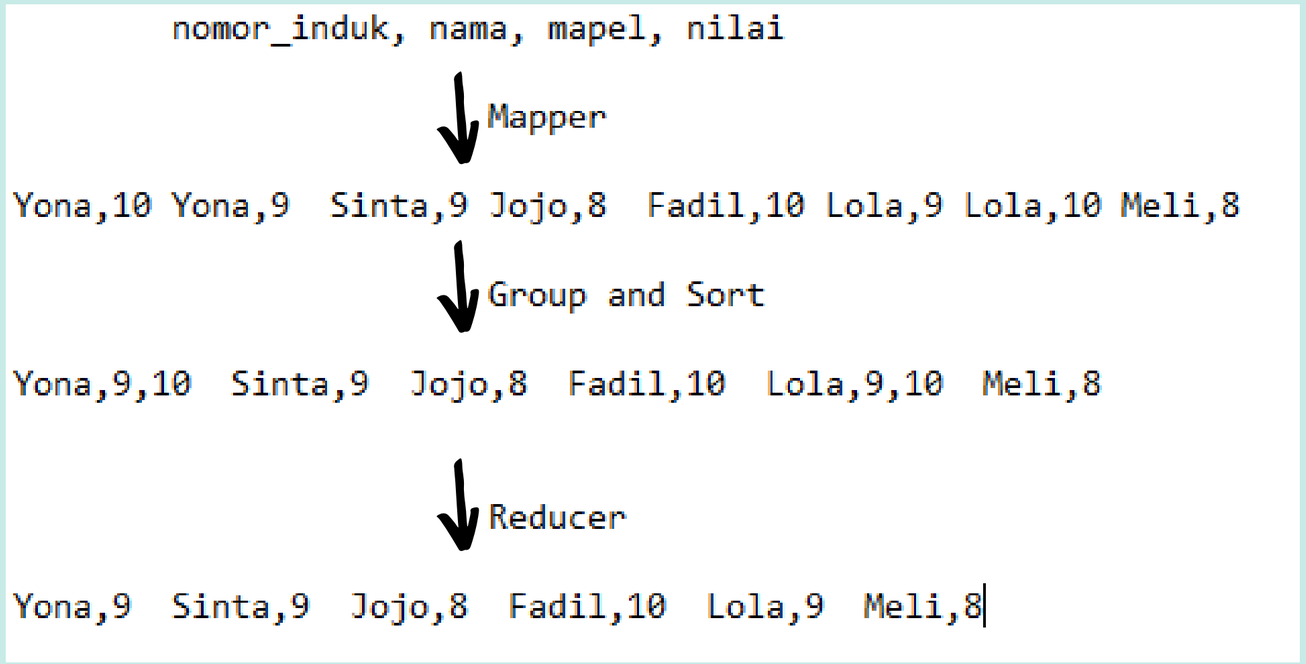
```
001, Yona, 1, 10
001, Yona, 2, 9
002, Sinta, 1, 9
003, Jojo, 2, 8
004, Fadil, 1, 10
005, Lola, 2, 9
005, Lola, 1, 10
006, Meli, 1, 8
```

Aggerate Minimum

Mapper akan mengambil informasi yang dibutuhkan saja yaitu nama dan nilai

MapReduce akan melakukan Grouping dan Sorting untuk setiap nama dan nilai

Reducer hanya akan mengambil nilai terendah untuk setiap siswa



Aggerate Maximum

Contoh Kasus:

Kita memiliki data nilai siswa sekolah dasar. berisi nomor_induk, nama, mata_pelajaran, nilai

Kita ingin mencari nilai tertinggi pada setiap nama siswa yang ada

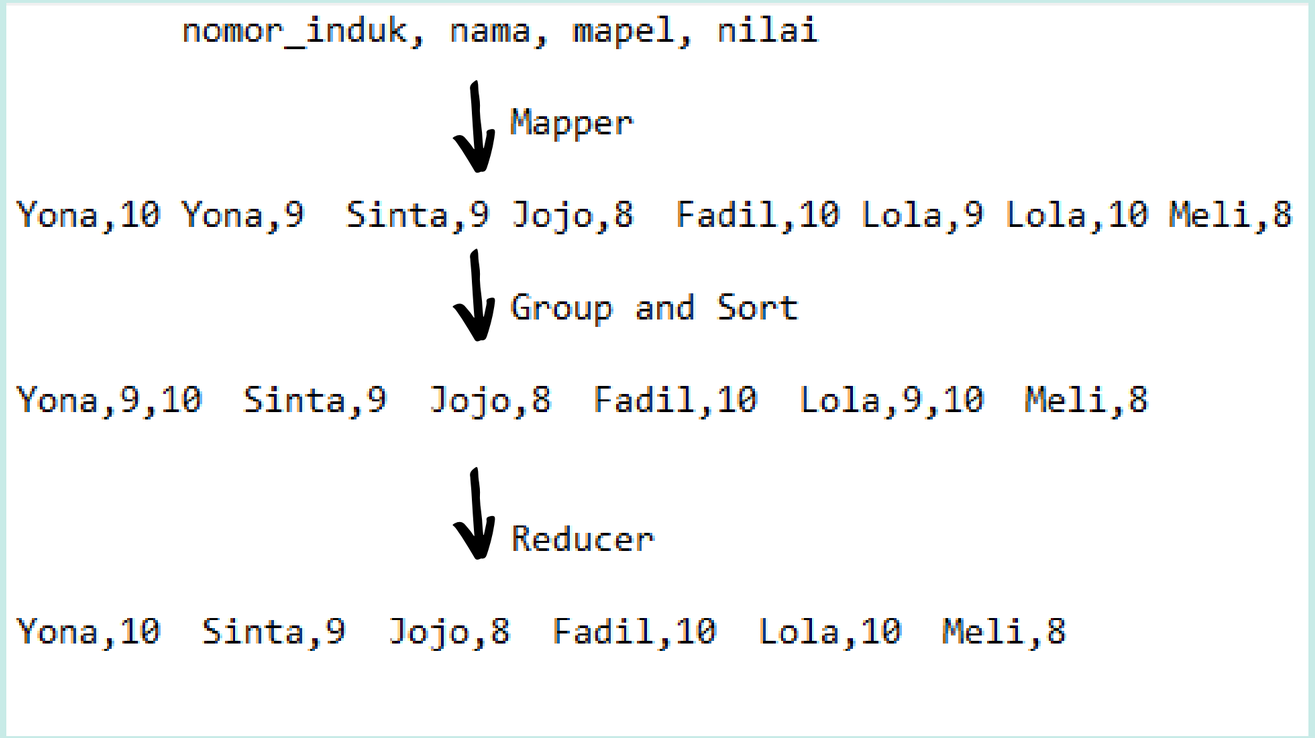
```
001, Yona, 1, 10
001, Yona, 2, 9
002, Sinta, 1, 9
003, Jojo, 2, 8
004, Fadil, 1, 10
005, Lola, 2, 9
005, Lola, 1, 10
006, Meli, 1, 8
```

Aggerate Maximum

Mapper akan mengambil informasi yang dibutuhkan saja yaitu nama dan nilai

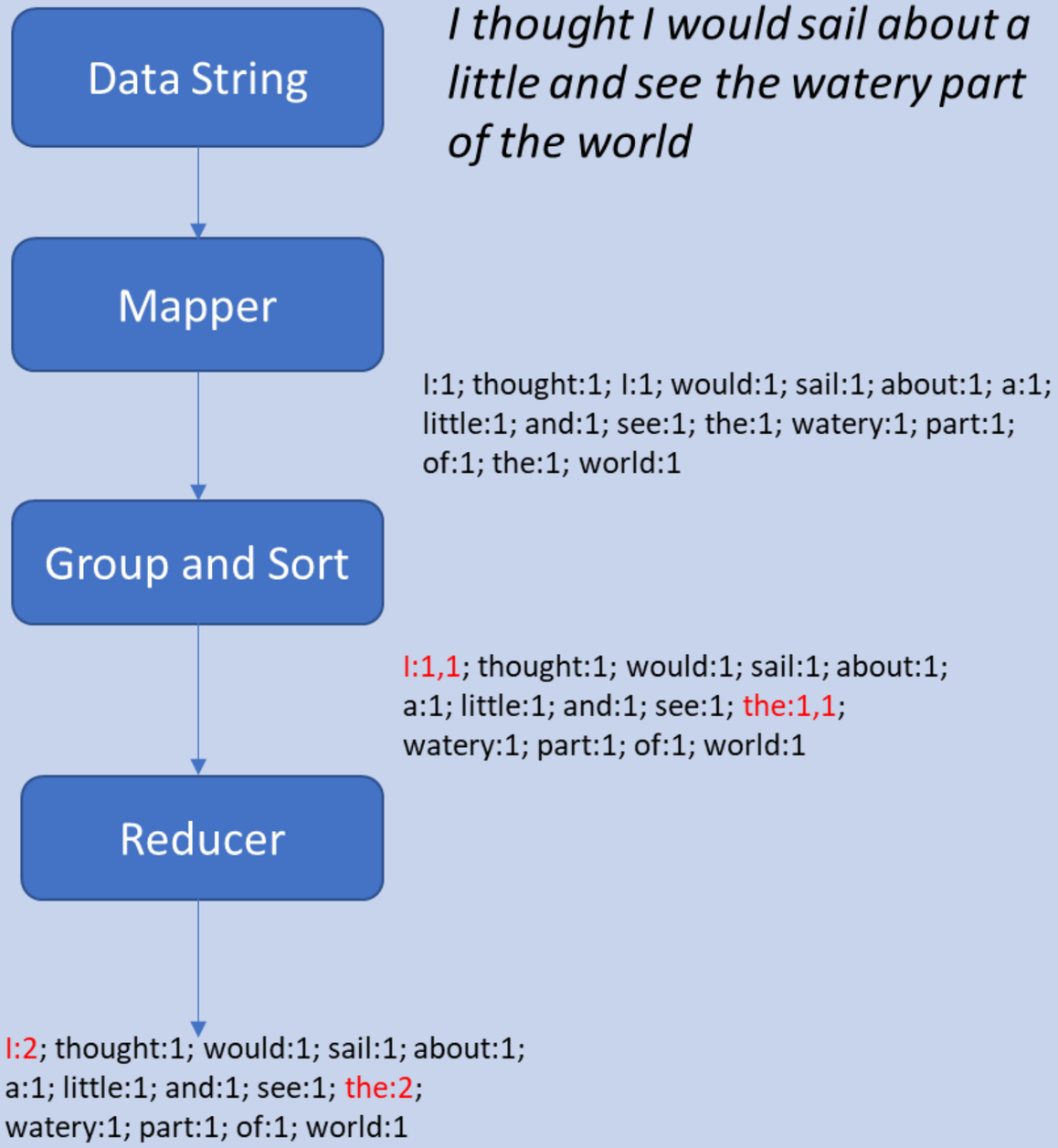
MapReduce akan melakukan Grouping dan Sorting untuk setiap nama dan nilai

Reducer hanya akan mengambil nilai tertinggi untuk setiap siswa



MapReduce Regex

- Regex (Regular Expression) : metode untuk mengenali suatu pattern pada tertentu pada string
- MapReduce Regex dapat digunakan untuk menghitung seberapa sering sebuah string muncul pada data yang berupa artikel ataupun buku



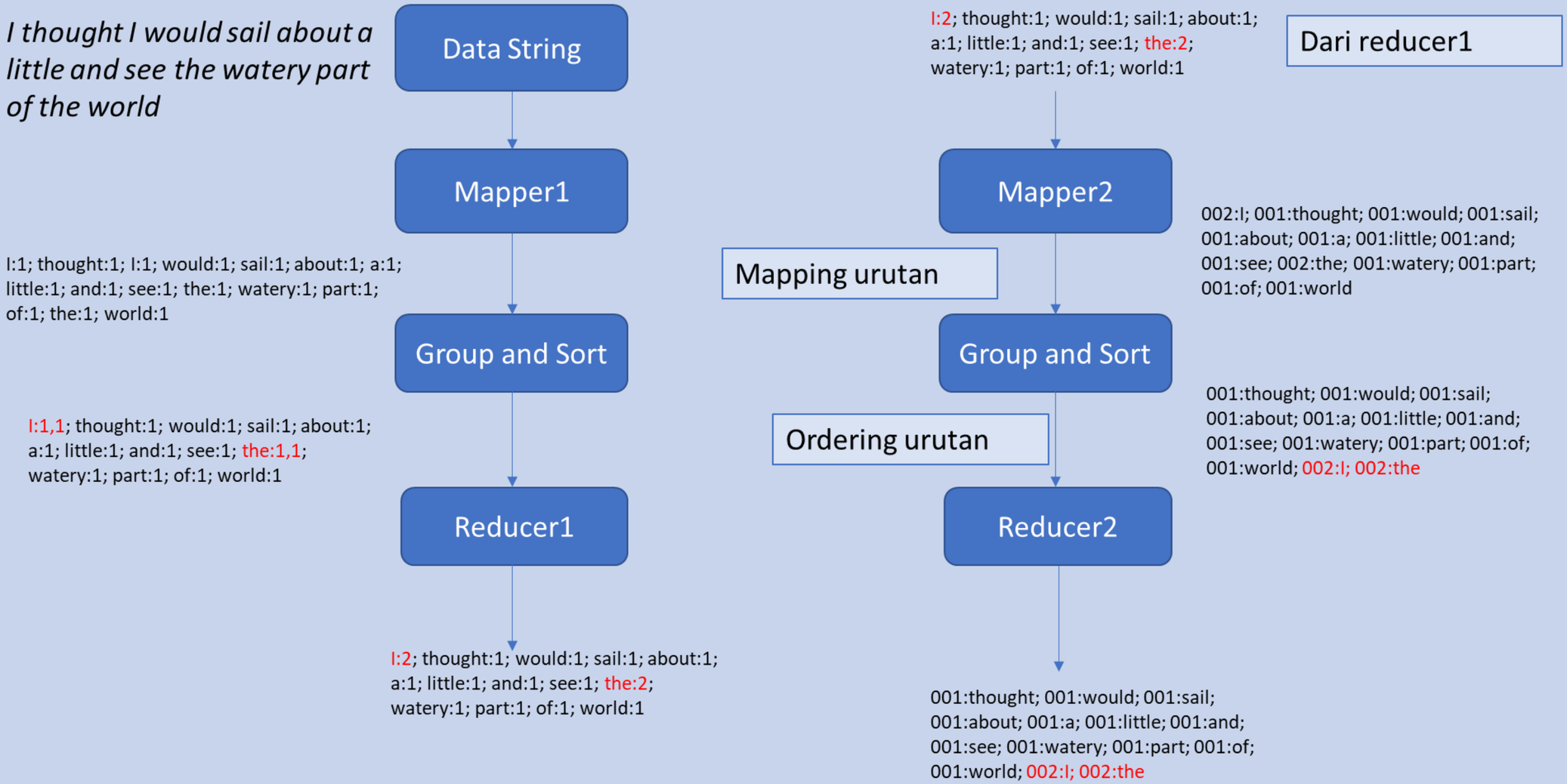
How MapReduce Regex Works

- Mapper akan melakukan pemetaan pada tiap kata
- MapReduce kemudian akan melakukan grouping and sorting untuk tiap kata
- Tahap akhir, reducer akan menjumlahkan total dari masing-masing kata

Multi Stage MapReduce Jobs

- Penggunaan lebih dari satu mapper dan reducer dalam sebuah job
- Pada multi stage, output dari reducer pertama akan digunakan sebagai input mapper kedua, biasanya dilakukan untuk kebutuhan sorting, grouping, dan ordering data

How Multi Stage MapReduce Jobs Works



DigitalSkola

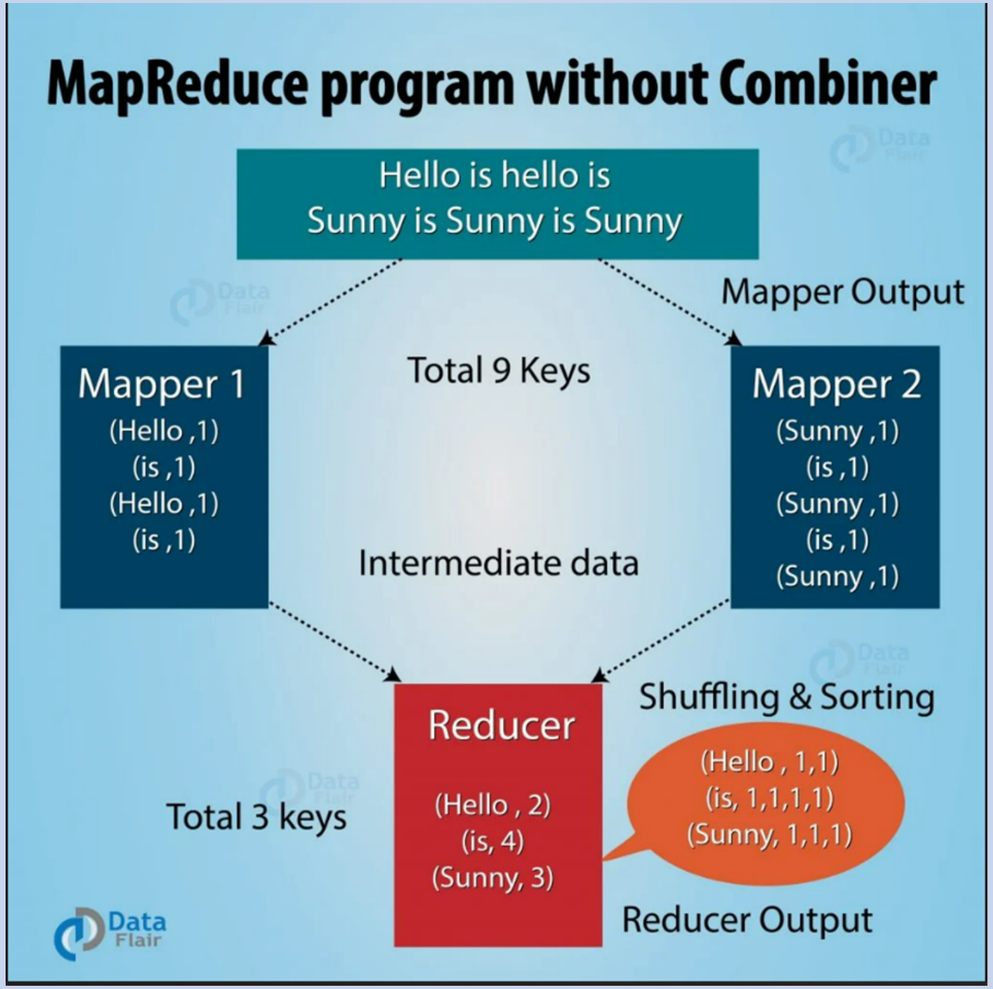
MapReduce Combiner

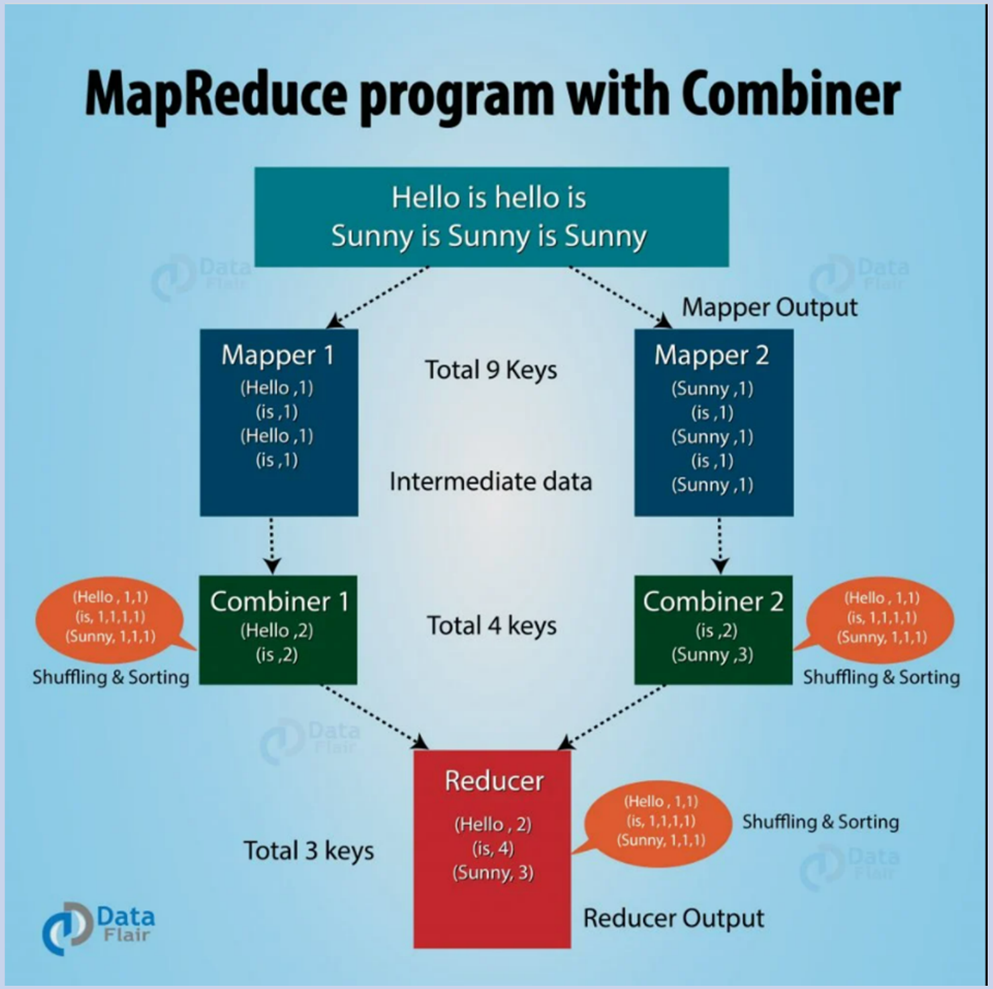
- Dikenal juga sebagai “mini reducer”
- Berfungsi untuk summarize hasil dari reducer pertama sebelum dilanjutkan ke mapper/reducer kedua

Fungsi dan Kelebihan Combiner

- Mengurangi waktu yang dibutuhkan dalam transfer data dari Mapper ke Reducer
- Mengurangi size output dari mapper
- Meningkatkan keseluruhan performa reducer

With vs Without Combiner





Source: https://data-flair.training/

DigitalSkola

Terima kasih!

Sampai jumpa di Learning Progress Review kami berikutnya!

Hilda Meiranita Prastika Dewi

Nur Indrasari

Rezha Sulvian

Thasha Dinya Ainsha

DigitalSkola