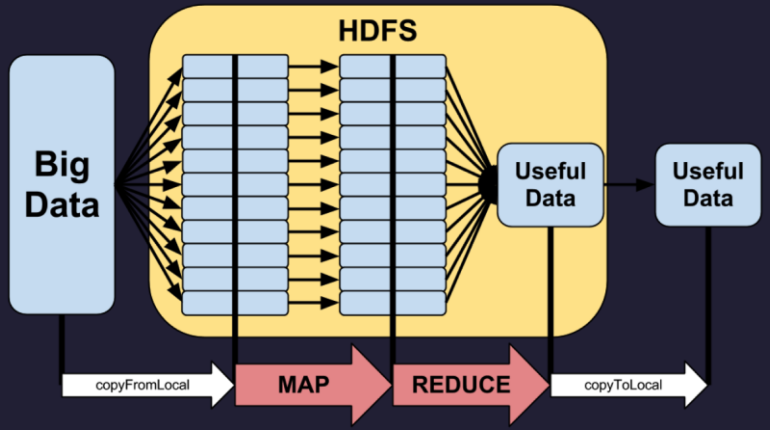
Runtime Analysis for Wordcount Using Hadoop and Python

Hadoop Works

HDFS (Hadoop Distributed File System)

Data besar akan didistribusikan menjadi blok kecil sebesar block size yang telah diatur ke seluruh node dalam cluster hadoop.

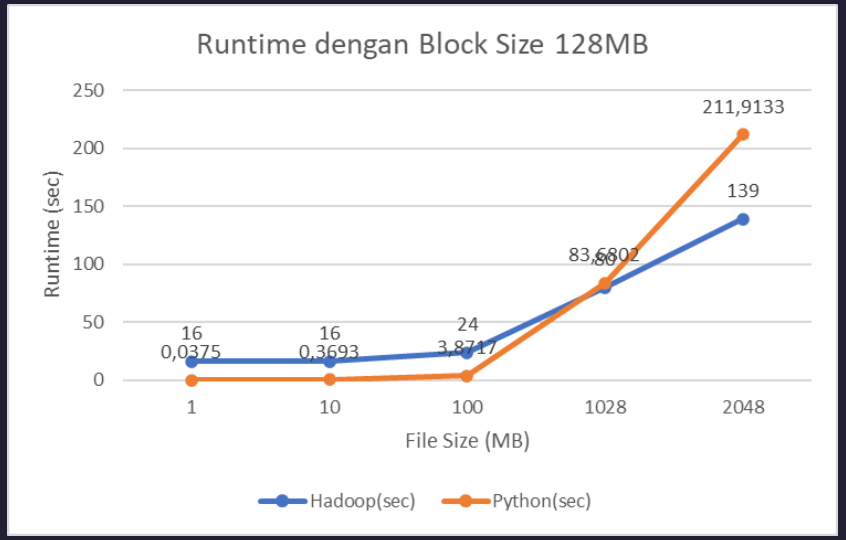


MapReduce

Memproses data secara paralel dengan fungsi Map untuk menerima input berupa key-value pair dan fungsi Reduce untuk penggabungan key-value pair berdasarkan kunci dan output akhir

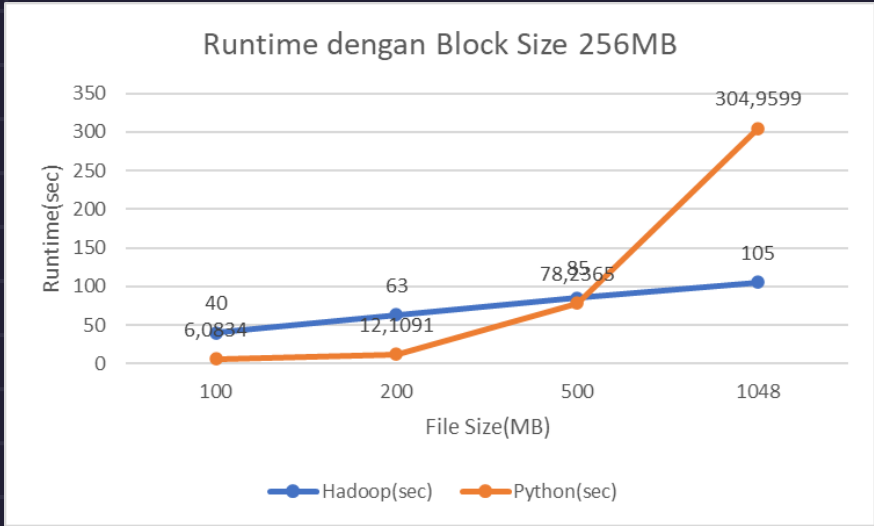
Hadoop (Block Size 128MB dan 256MB) VS Python

Hasil Runtime



Python tanpa hadoop lebih cepat pada file ukuran kecil (1-10 MB), dan terjadi peningkatan secara linear hingga sampai di posisi file 1 - 2 GB, dari sini python dengan hadoop waktu eksekusinya lebih cepat.

Hasil Runtime



Python tanpa hadoop, berjalan lebih cepat pada file kurang dari 256 MB, sedangkan pada file diatasnya python dengan hadoop bekerja lebih cepat

Analisis (1)

Berdasarkan hasil runtime, file dengan ukuran di bawah block size pada masing-masing percobaan, akan menunjukkan hasil yang lebih lama jika program dijalankan dengan hadoop. Hal tersebut berkaitan dengan penanganan hadoop terhadap block memori tersebut.

Jika sebuah file kurang dari sebuah block size, maka clustering akan tetap dilakukan, dan blok kosong akan ditangani. 1 blok ini juga akan ditangani sebagai blok yang utuh sehingga juga melibatkan proses pembacaan dan pertukaran data ke dalam HDFS, juga di dalam node dalam cluster. Proses tersebut akan memperlambat proses eksekusi program

Analisis (2)

Berdasarkan hasil runtime, pada ukuran file yang sama dan berukuran besar seperti 1 - 2 GB, hasil runtime dengan hadoop dengan block size lebih tinggi, akan menghasilkan waktu runtime yang lebih cepat.

Hal tersebut juga merupakan salah satu akibat dari clustering, dan pemrosesan data yang lebih sedikit, sehingga overheadnya lebih sedikit. Pada block size yang lebih besar, block program yang dicluster akan lebih sedikit, sehingga overhead berkurang, yang menyebabkan waktu runtime semakin cepat.