## 1. Pendahuluan

### Latar Belakang

Hadoop adalah Suatu *software framework* (kerangka kerja perangkat lunak) open source berbasis *Java* di bawah lisensi *Apache* untuk aplikasi komputasi data besar secara intensif [1]. Hadoop mampu menyelesaikan masalah yang berhubungan dengan data yang sangat besar. Dengan banyaknya perkembangan data saat ini Hadoop menjadi solusi y/ang tepat untuk menangani, menyimpan dan mengolah data yang sangat besar. Hadoop juga sudah sangat banyak digunakan di berbagai prusahaan seperti yahoo, facebook dll [2].

HDFS adalah *system file* yang dirancang untuk menyimpan *file* yang sangat besar dengan akses data *streaming* dan berjalan pada kelompok perangkat keras komoditas [3]. HDFS adalah tempat atau direktori di komputer dimana data Hadoop disimpan. HDFS akan melakuakan proses pemecahan *file* yang besar menjadi bagian yang lebih kecil. *Hadoop distributed file system* dirancang untuk menangani data dalam jumlah banyak yang ukuran *file* nya hingga petabyte, exabyte.

Namun terdapat masalah yang di temukan ketika HDFS menangani *file* kecil dalam jumlah banyak [4], sehingga terdapat beberapa solusi yang ditawarkan yaitu dengan menggunakan metode HAR(*Hadoop Archive*) dan *Combine File Input*. Dengan kedua metode ini maka file kecil di HDFS teratasi. Kemudian akan di lakukan perbandingan, yang mana metode yang digunkan menghasilkan penggunaan block dan waktu pemrosesan yang di gunakan pun sedikit [5].

### Topik dan Batasannya

Terdapat masalah yang di temukan ketika HDFS menangani file kecil dalam jumlah banyak sehingga tidak bekerja dengan baik. File kecil dapat didefinisikan sebagai file apa pun yang secara signifikan lebih kecil dari ukuran blok HDFS [6]. Ukuran blok Hadoop di *set* ke 128 dan 512 MB. Sehingga untuk mengatasi masalah file kecil di HDFS maka digunakan metode Combine File Input Format dan Hadoop Archive. Metode ini memproses setiap file kecil di hdfs hingga menghasilkan *output* file yang sudah digabungkan.

Batasan pekerjaan yang dilakukan di dalam TA ini berupa Inputan file text yang berisikan 2000 dan 4000 file karena keterbatasa Komputer yang tersedia Program yang dibangun hanya dapat memproses small file yang akan di merge, tidak untuk mengeluarkan file atau data tertentu.

### Tujuan

Didalam masalah yang terjadi di hdfs maka tujuan daripada Tugas Akhir ini adalah dapat mengimplementasikan HDFS yang mampu menangani *file* berukuran kecil dalam jumlah banyak. Maka untuk mengatasi file berukuran kecil dalam jumlah banyak yaitu Dengan menggunakan metode *combine file input format* dan HAR (*Hadop Archive*) maka masalah *file* kecil yang di proses di HDFS bisa teratasi. Kemudia akan di hitung waktu pemrosesan *file* dari setiap metode dan block yang digunakan.

### Organisasi Tulisan

Di dalam studi terkait terdapat penjelasan mengenai teori-teori yang terkait atau berhubungan dengan tugas akhir yang dibuat. Yang didalamnya terdapat teori hadoop, hadoop distributed file system, mapreduce, combine file input format, dan hadoop archive. Sistem yang dibangun mejelaskan tentang alur alur atau cara kerja sistem yang dibuat. Yang didalamnya terdapat alur kerja hadoop archive, combine file input format, dataset, skenario pengujian, skenario pengujian combine file input format, skenario pengujian hadoop Archive, dan spesifikasi sistem. Evaluasi menjelaskan tentang penilaian untuk menentukan, apakah pengujian yang dilakukan sesuai dengan tujuan yang ingin dicapai. yang didalamnya terdapat hasil pengujian dan hasil analisis pengujian. Kemudian

terdapat kesimpulan yang menjelaskan kesimpulan daripada pengujian yang telah dilakukan. Dan terdapat daftar pustaka sebagai acuan paper yang berkaitan dengan tugas akhir yang telah dibuat. Kemudian yang terakhir terdapat lampiran, lampiran berupa detail data, detail hasil pengujian, analisis hasil pengujian dan screenshot tampilan system.