1. Pendahuluan

1.1 Latar belakang

Seiring dengan bertambahnya permintaan layanan teknologi informasi saat ini, ketersediaan kekuatan komputasi yang tinggi dari sebuah sistem komputer yang *highly-available* semakin dibutuhkan. Peningkatan jumlah pengguna layanan termasuk jenis layanan yang ditawarkan (misalnya layanan informasi *online*, interaktif, dan tersedia 24/7, seperti layanan Google, Multiply, dan Facebook) akan membutuhkan *server* yang *fault-tolerant*, artinya layanan *server* akan tetap tersedia walaupun terjadi kegagalan dalam sistem. Pengelolaan data-data dengan skala besar harus mampu diolah dalam waktu yang relatif singkat karena termasuk kedalam kinerja sebuah sistem.

Pada kenyataannya data yang berkembang dalam di dalam *digital-universe* hingga 2006 mencapai 0,18 zettabytes dan diperkirakan akan terus berkembang hingga 1,8 zettabytes pada tahun 2011[1]. Namun infrastruktur dari sebuah media penyimpanan memiliki kendala, apabila *hard drive* tahun 1990 memiliki kapasistas 1370 MB dan kecepatan transfer sekitar 4,4 MB/s, sehingga dapat membaca keseluruhan data hanya dalam kurun waktu lima menit. Hampir 20 tahun setelah itu, mundul *hard drive* dengan kapasistas 1 TB, namun kecepatan transfer sekitar 100 MB/s, sehingga membutuhkan waktu dua setengah jam untuk membaca keseluruhan data [1]. Sehingga teknologi yang berhubungan dengan peningkatan kinerja sistem dalam skala besar benar-benar dibutuhkan.

Map Reduce dan *Hadoop Distributed File System* (HDFS) hadir sebagai sebuah solusi terhadap kebutuhan-kebutuhan tersebut. *mapreduce*dikemukakan oleh Jeffrey Dean dan Sanjay Ghemawat tahun 2004 melalui publikasinya yang berjudul *Mapreduce: Simplied Data Processing on Large Clusters* melalui Google Labs. Dalam publikasinya, mereka mengatakan bahwa *mapreduce*merupakan model pemrograman yang diimplementasikan untuk melakukan pemrosesan atau pengolahan data elektronik dalam jumlah besar [2]. Aplikasi yang dibuat menggunakan model pemrograman *mapreduce*akan secara otomatis paralel, hal ini membuat pengembang yang tidak berpengalaman

terhadap komputasi paralel dalam sistem terdistribusi akan mudah membangun aplikasi pada sistem terdistribusi yang besar [2]. Sedangkan HDFS merupakan sistem berkas terdistribusi yang memiliki kemampuan dalam mengelola tingkat ketersediaan data yang tinggi. HDFS dapat menangani data-data dalam jumlah skala besar.

Namun keterbatasan infrastruktur menjadi kendala dalam membangun sebuah sistem *cluster*. Karena secara teori semakin banyak jumlah *node* yang bekerja pada sebuah *cluster* maka kinerjanya semakin meningkat, diperlukan suatu cara untuk meningkatkan jumlah *node* untuk meningkatkan kinerja dari *cluster* tanpa menambah *node* secara fisik. Pada tugas akhir ini akan dirancang sistem Hadoop *cluster* yang akan berjalan pada *node* virtual sebagai cara untuk menambah jumlah *node* guna meningkatkan kinerja *cluster* dalam menyelesaikan suatu proses. Akan dilakukan pengukuran kinerja Hadoop *cluster* apabila berjalan dalam mesin virtual. Diharapkan dengan bertambahnya *node* secara virtual dapat menaikan kinerja *cluster* secara keseluruhan.

1.2 Perumusan masalah

Beberapa permasalahan pada Tugas Akhir dapat didefinisikan sebagai berikut :

1. Bagaimana merancang sistem Hadoop *cluster* dalam mesin virtual?
2. Apakah penggunaan mesin virtual sebagai cara untuk meningkatkan jumlah *node* dalam sistem dapat meningkatkan kinerja *cluster*?
3. Parameter apa yang berpengaruh secara dominan terhadap kinerja Hadoop *cluster?*

Sementara yang menjadi batasan masalah dalam Tugas Akhir ini adalah:

1. Sistem operasi yang digunakan adalah Ubuntu 10.04.
2. Dalam implementasi digunakan peralatan jaringan yaitu media transmisi kabel UTP, *switch* dan NIC *Gigabit Ethernet* 1000Mbps untuk penghubung antar *node*.
3. Jumlah komputer yang digunakan dalam implementasi sistem yaitu 4 buah komputer.

4. Bagian dari Hadoop Framework yang digunakan dalam penelitian ini hanya *mapreduce* dan HDFS.
5. Parameter yang diukur pada pengujian performansi yaitu nilai *throughput* untuk pengujian HDFS dan waktu eksekusi untuk pengujian sort untuk melihat kinerja *mapreduce*.
6. Variabel parameter konfigurasi yang diuji yaitu jumlah *node*, nilai replikasi, ukuran blok, dan kapasitas *taskmap*.

1.3 Tujuan

Tujuan dari penulisan tugas akhir dengan judul ”Analisis Performansi MapReduce Pada Mesin Virtual Berbasis Hadoop Cluster” adalah:

1. Mengimplementasikan *mapreduce* pada sistem berbasis Hadoop cluster pada mesin virtual.
2. Menganalisis pengaruh jumlah *node* virtual terhadap performansi *cluster*.
3. Menganalisis pengaruh perubahan parameter replikasi, ukuran blok dan kapasitas *taskmap* terhadap nilai *throughput* dari HDFS dan waktu eksekusi *mapreduce* pada *cluster*.
4. Membandingkan performansi *cluster* pada skala fisik maupun skala virtual.

1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan dalam memecahkan masalah di atas adalah dengan menggunakan langkah-langkah berikut:

1. Tahap studi literatur

   Melakukan studi berdasarkan literatur dan diskusi materi dengan dosen pembimbing maupun dengan orang yang berkompeten mengenai konsep proses terdistribusi.
2. Tahap perancangan

   Melakukan perancangan dan pemodelan pada sistem yang akan diuji. Hal ini berkaitan dengan relevansinya di lapangan dan kemungkinannya untuk diimplementasikan.
3. Tahap implementasi dan pengumpulan data

Mengumpulkan data-data dari parameter yang telah ditentukan dari hasil pengujian pada implementasi

4. Tahap analisis dan penarikan kesimpulan

   Melakukan analisis dari data yang telah didapatkan dari hasil pengujian.

5. Tahap penyusunan laporan tugas akhir