BAB 1
PENDAHULUAN

1.1 Latar Belakang

Jumlah data di dunia kita telah meledak, dan menganalisis data berukuran besar tersebut *big data* [1] menjadi kunci dasar persaingan, mendasari gelombang baru pertumbuhan produktivitas, inovasi, dan surplus konsumen. Belum banyak yang mendefinisikan istilah *big data* secara pasti. Meskipun demikian, istilah "*Big Data*" sering digunakan oleh perusahaan untuk menguraikan jumlah data yang besar. Hal ini tidak mengacu pada jumlah khusus data, tetapi menguraikan suatu set data yang tidak dapat disimpan atau diproses menggunakan perangkat lunak database tradisional. Contoh *big data* mencakup Google *Search Index*, *database* Facebook (*user pro*file) [2].

Big data sering kali di distribusikan melalui banyak *storage device*, dapat dalam beberapa lokasi yang berbeda. Terdapat beberapa jenis berbeda dari solusi perangkat lunak *big data* yang berbeda, mencakup *platform* penyimpan data dan program analisa data. Produk yang paling umum dari perangkat lunak *big data* mencakup apache Hadoop, IBM's Big Data Platform, Oracle NoSql database, Microsoft HDInsight *dan* EMC Pivotal One [3].

Hadoop banyak dipakai untuk mengolah data yang sangat besar (Petabyte) secara terdistribusi dan berjalan di atas *cluster* yang terdiri dari beberapa komputer yang saling terhubung. Hadoop menggunakan *HDFS* yang tidak sama dengan jenis *file* system dari sistem operasi misalnya NTFS atau FAT32.

Penyimpanan *HDFS* adalah metadata, merupakan struktur direktori *HDFS* dan *file* dalam bentuk *tree*. Hal ini juga mencakup berbagai atribut direktori dan *file*, seperti kepemilikan, perizinan, kuota, dan faktor replikasi [4].

Hadoop merupakan salah satu produk yang menyimpan *big data* dan tersukses sampai saat ini. Beberapa contoh dari perusahaan yang menggunakan program Hadoop ini seperti Amazon, Facebook, Google, IBM, Spotify, Twitter, Yahoo, dan beberapa Universitas menggunakan Hadoop sebagai pencarian dan analisis data [5].

1.2 Rumusan Masalah

Berikut masalah–masalah yang mungkin akan terjadi dan perlu dijawab. Masalah–masalah yang ada sebagai berikut:

1. Apakah pengaruh ukuran *file* dibawah 500MB terhadap waktu saat perpindahan data ?
2. Bagaimana karakteristik penyebaran data dari *master* slave menuju *client server* ?
3. Bagaimana keterkaitan konfigurasi terhadap *block* data yang dikirim ke slave?

1.3 Tujuan

Tujuan yang dapat dari rumusan masalah adalah sebagai berikut :

1. Mengamati karakteristik penyebaran *file.*
2. Mengamati pengaruh ukuran *file* dibawah 500MB terhadap waktu pengiriman ke *client server*.
3. Mengamati dan mengukur besar *block* data yang dikirim ke *client server* berdasar konfigurasi pada *master*.

1.4 Ruang Lingkup Penelitian

Ruang lingkup penelitian memiliki batasan – batasan sebagai berikut :

1. Physical machine yang digunakan memiliki spesifikasi Intel i5-2320 3.00ghz dengan RAM 4 GB.
2. Mesin Virtual yang digunakan untuk *server* akan memiliki spesifikasi, OS Linux Ubuntu 14.03.3 *server*, 15GB HDD dan RAM 1GB.
3. Koneksi antara mesin tidak menggunakan password, maka SSH key akan dihapus atau diubah menjadi non pass SSH.
4. Program Virtual yang digunakan adalah Oracle VM Virtual Box v5.0.2, dan
5. Data percobaan yang digunakan berukuran 314MB, 235MB, dan 127MB dengan *file* format .mov, dengan *block* allocation 128MB.

1.5 Sumber Data

Penelitian dimulai dengan mencari informasi tentang Hadoop dan *HDFS.* Informasi–informasi yang didapat melalui buku, *e-book*, atau internet mengenai

Hadoop. menginstal dan mengkonfigurasikan Hadoop dengan beberapa *server*, menginstal *traffic* monitoring agar bisa memonitor jaringan antara *server* atau node yang ada dan dilakukan beberapa percobaan juga dibuat laporan.

1.6 Sistematika Penyajian

Laporan yang berisi hasil dari Tugas Akhir yang telah selesai dikerjakan selama Tugas Akhir berlangsung, bersistematik seperti berikut :

BAB I – Pendahuluan

Berisi mengapa Hadoop dipakai dan sejarah singkat, rumusan masalah yang berisi pertanyaan – pertanyaan yang akan terjawab dan diberi kesimpulan, tujuan yang berisi singkatan dari hasil, batasan – batasan pada ruang lingkup penelitian dan sumber data yang berisi darimana isi laporan didapat.

BAB II – Kajian Teori

Berisi penjelasan dari teori–teori yang didapat dalam sumber–sumber yang berisikan mengenai Hadoop dan ekosistem yang ada dalam Hadoop.

BAB III – Analisis dan Rancangan

Berisi permodelan rancangan dari penelitian, rancangan skenario yang akan dilakukan untuk mendapat hasil atau kesimpulan.

BAB IV – Implementasi

Berisi konfigurasi dan lingkup dari rancangan penelitian, yang digunakan sebagai tempat pengerjaan skenario – skenario pada bab III.

BAB V – Pengujian

Berisi hasil dari skenario rancangan penelitian, hasil dari skenario dan hasil dari tujuan yang dapat ditarik kesimpulan.

BAB VI – Simpulan dan Saran

Berisi simpulan dan saran dari keseluruhan hasil penelitian, hasil yang didapat mengacu kepada tujuan penelitian dan hasil akhir pada bab V. Saran yang diajukan diharap bisa memberikan peluang pengerjaan lebih baik untuk penelitian selanjutnya.