BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi semakin pesat dan layanan yang disediakan semakin banyak. Terdapat layanan yang berkembang sangat cepat yaitu teknologi *internet*. Dimana dengan *internet* semua perangkat elektronik yang mempunyai alamat *Internet Protocol* maka dapat saling terkoneksi. Seiring berkembangnya *internet* terdapat banyak layanan *software* yang disediakan. Layanan *software* dapat digunakan oleh setiap pengguna secara berbayar ataupun gratis. Semakin berkembangnya teknologi pada *software* yang kompleks menyebabkan konsumsi data penyimpanan semakin besar. Oleh karena itu, Segala data yang terdapat pada suatu aplikasi pada *internet* akan berkembang dari data yang kecil kemudian menjadi data yang besar.

Data dalam skala besar yang disebut dengan *Big Data*. *Big Data* adalah kumpulan data yang sangat besar, sangat variatif, sangat cepat pertumbuhannya dan mungkin tidak terstruktur. Proses komputasi yang terjadi pada *big data* dapat berjalan lambat apabila komputer yang digunakan untuk memproses data tersebut tidak memenuhi standar yang dibutuhkan oleh suatu *big data*. Maka, diperlukan suatu algoritma khusus sehingga informasi yang mendalam mudah didapatkan dan dapat membantu pengambilan keputusan yang lebih baik [1]. Untuk memanfaatkan kekuatan *Big Data*, maka akan diperlukan sebuah infrastruktur yang dapat mengelola dan memproses data dalam *volume* besar, kecepatan yang tinggi, serta kompleks dan berjalan secara *realtime*.

Solusi untuk *Big Data* yaitu Hadoop. Hadoop adalah *framework open source* di bawah lisensi Apache untuk mensupport aplikasi yang jalan pada *Big Data*. Asal mula Hadoop muncul karena terinspirasi dari makalah tentang *Google MapReduce* dan *Google File System (GFS)* yang ditulis oleh ilmuwan dari Google, *Jeffrey Dean* dan *Sanjay Ghemawat* pada tahun 2003. Penamaan menjadi Hadoop adalah diberikan oleh *Doug Cutting*, yaitu berdasarkan nama dari mainan gajah anaknya [2].

Hadoop dijalankan pada lingkungan yang menyediakan *storage* dan komputasi secara terdistribusi atau bisa disebut sebagai *Hadoop Distributed File System* (HDFS). *Hadoop* mendistribusikan klaster-klaster dari komputer/*node* menggunakan suatu model pemrograman.

Mapreduce adalah paradigma pemrograman yang berjalan di latar belakang Hadoop untuk menyediakan skalabilitas dan mudah solusi pengolahan data [3]. Pengolahan data dapat pada sebuah data yang terstruktur, semi-terstruktur, dan tidak terstruktur [4].

Data yang diolah terdapat pada *cloud* yang mempunyai sistem penyimpanan *disk storage* atau biasa disebut *hardisk*. Aplikasi yang membutuhkan operasi *write once read many* (WORM) memerlukan *Latency* yang rendah dan nilai *Throughput* yang tinggi. Dengan mengganti disk yang tidak efisien dengan cache memori *low latency*, *high-throughput* yang terdistribusi untuk mengoptimalkan proses pengiriman data dari tugas *map* ke tugas *reduce* [5]. Untuk memenuhi kebutuhan tersebut proses komputasi menggunakan fungsi yang terdapat pada Apache Flink.

Apache Flink yaitu salah satu *platform open source* yang dapat digunakan untuk merancang program Mapreduce dengan *in-memory batch processing* dan *stream processing* [6]. Ada dua API inti di Flink: API DataSet untuk memproses kumpulan data yang terbatas (sering disebut sebagai pemrosesan *batch*), dan API DataStream untuk memproses data *stream* yang tak terbatas (sering disebut sebagai pemrosesan *stream*) [7]. Pada penelitian ini, akan dilakukan pengolahan data pada sebuah data yang tidak terstruktur dalam bentuk teks. Mengimplementasikan metode Mapreduce *batch processing* berbasiskan HDFS yang dijalankan pada sistem operasi linux.

1.2 Rumusan Masalah

Rumusan masalah dalam pembuatan Tugas Akhir ini adalah rendahnya efisiensi dalam pengaksesan suatu data pada *big data* dan proses komputasi metode Hadoop Mapreduce memerlukan waktu yang lama. Terdapat kumpulan data dalam skala besar dan beragam variasinya untuk dikelola. Data yang bertumbuh secara cepat (*up to date*).

1.3 Tujuan

Dengan merujuk pada rumusan masalah diatas, maka tujuan yang dibahas pada Tugas Akhir ini:

a) Merancang aplikasi HDFS untuk mengolah kumpulan data pada *Big Data*.
b) Implementasi metode Mapreduce berbasis HDFS.
c) Menganalisa response time dan penggunaan resource dari sistem berupa *memory*, *processor*, serta *disk* pada metode Mapreduce.

d) Mengkaji efisiensi metode Hadoop Mapreduce dengan metode Mapreduce *in-memory batch processing* Apache Flink.

1.4 Batasan Masalah

Tugas Akhir ini mempunyai batasan masalah yaitu:

a) Menjalankan program Mapreduce *wordcount* berbasiskan HDFS pada sistem operasi Ubuntu Server 16.04 LTS.
b) Aplikasi dijalankan pada single node.
c) Program Mapreduce *wordcount* menggunakan *library* yang terdapat pada Apache Flink dan dirancang menggunakan bahasa Java.
d) Data yang dipakai adalah data yang tidak terstruktur dalam bentuk teks dengan ekstensi .txt dan .csv.

1.5 Metodologi Penelitian

1. **Pengumpulan data**

   Mencari *dataset* yang akan digunakan sebagai data uji dari *internet*, *dataset* yang digunakan adalah kumpulan nama-nama orang yang ada pada facebook yang berformat .txt.

2. **Studi literatur**

   Mencari dan mempelajari teori, konsep serta implementasi *platform* yang digunakan dari jurnal, buku, materi dari internet.

3. **Perancangan Sistem**

   a. Perancangan perangkat keras, spesifikasi perangkat keras dari notebook yang digunakan adalah *processo*r core i5 2,8 GHz dan RAM sebesar 4 GB.
   b. Perancangan perangkat lunak, perangkat lunak yang digunakan adalah Ubuntu versi 16.04 LTS serta *platform* untuk memproses *big data* yang digunakan adalah Apache Hadoop dan Apache Flink.

4. **Pengujian**

   Pengujian dilakukan ketika semua sistem selesai dibangun.

5. **Hasil Pengujian**

   Setelah dilakukan pengujian maka selanjutnya analisis keluaran dari sistem tersebut dialalisis untuk mengetahui apakan keluaran tersebut sesuai seperti yang diharapkan.

1.6 Sistematika Penulisan

Tugas akhir ini dibagi menjadi lima bab bahasan, ditambah dengan lampiran. Dibawah ini merupakan masing-masing dari bahasan tiap babnya:

BAB 1 PENDAHULUAN

Bab ini menjelaskan tentang permasalahan serta solusi dari masalah tersebut.

BAB II LANDASAN TEORI

Bab ini berisikan beberapa teori yang mendukung dan menjadi dasar dari pembuatan tugas akhir ini.

BAB III ANALISIS DAN PERANCANGAN

Bab ini berisi tentang perancangan system yaitu system perangkat keras serta system perangkat lunak yang digunakan.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Bab ini berisi tentang implementasi dan pengujian system sehingga didapat keluaran yang diinginkan.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi tentang kesimpulan dan saran dari system yang dibuat.