BAB I
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan berkembangnya zaman *big data* merupakan suatu yang menjadi trend dalam dunia informasi. Bisa dibilang *big data* merupakan kumpulan data yang sangat besar yang di dalamnya mencakup berbagai jenis data. *Big Data* menjadi kata yang popular seiring dengan bagaimana dapat menyimpan data dalam jumlah yang besar, melakukan proses serta analisa. Sesuatu yang tidak dapat dihindari bagaimana *impact* dari *big data* ini dalam kehidupan sehari-hari. *Big Data* telah memberikan kesempatan atau peluang bisnis bagi banyak perusahaan. Hampir semua industri telah memanfaatkan atau baru melakukan identifikasi tentang pentingnya *big data* dalam menumbuhkan bisnisnya atau tetap dapat bersaing bahkan menjadi keunggulan dalam berkompetisi [1].

Dari sekian banyak manfaat dan peluang, *big data* dapat meninggalkan beberapa tantangan diantaranya adalah tantangan teknologi yang dapat menghandle *big data* ini, tantangan *skill* atau keahlian orang yang akan mengolah data sehingga data yang tersedia dapat menjadi informasi, *insight* yang bermanfaat. Dalam dunia akademik, istilah *big data* mengacu pada aplikasi teknologi informasi untuk menangani masalah data yang sifatnya besar [1].

Guna mengatasi masalah data yang terus bertambah besar pada tugas akhir ini dibuatlah sebuah sistem yang dapat memproses *big data*. Metode yang digunakan untuk memproses *big data* tersebut yaitu MapReduce, MapReduce adalah model pemrograman untuk menulis aplikasi yang dapat mengolah data besar. MapReduce memberikan kemampuan analitis untuk menganalisis volume besar data yang kompleks [2]. Algoritma MapReduce berisi dua tugas penting yaitu *Mapper* dan *Reducer*. Pada proses *Mapper* data yang masuk diurai berdasarkan jenisnya sehingga dapat dengan mudah dipilih, kemudian masuk proses *Reducer*, pada proses ini data dikelompokkan berdasarkan tipe data yang sama kemudian keluarlah output hasil data yang sudah diproses. MapReduce sendiri mempunyai

banyak *platform* misalnya yang penulis pakai di sini adalah Apache Spark, Apache Spark adalah teknologi komputasi klaster kilat-cepat, dirancang untuk perhitungan cepat. Hal ini didasarkan pada Hadoop MapReduce dan memperluas model MapReduce menggunakannya secara efisien untuk lebih banyak jenis perhitungan, yang mencakup *query* interaktif dan pengolahan aliran. Fitur utama dari Apache Spark adalah di memori klaster komputasi yang meningkatkan kecepatan pemrosesan aplikasi [3].

1.2 Rumusan Masalah

Rumusan masalah dalam pembuatan tugas akhir ini adalah sebagai berikut. Pertama adalah lamanya proses dalam memproses data pada suatu *big data,* hal ini dikarenakan besarnya data itu sendiri. Keanekaragaman jenis data dalam suatu *big data*, data dalam suatu *big data* sangat bervariasi misalnya data media sosial, data transaksi keuangan, dan lain sebagainya. Maka diperlukan suatu sistem yang dapat memproses suatu *big data* tersebut dengan cepat.

1.3 Tujuan

Berdasarkan rumusan masalah diatas maka tujuan dari tugas akir ini adalah sebagai berikut:

a. Mengimplementasikan *platform* Apache Spark yang berjalan pada *Hadoop Distributed File System* secara standalone sebagai alternatif dari Hadoop MapReduce dalam memproses suatu *big data.*
b. Menganalisa pemrosesan *big data* menggunakan *platform* Apache Spark dan membandingkannya dengan Hadoop MapReduce dalam hal kecepatan atau *response time.*
c. Menganalisa penggunaan *resource* dari sistem *memory*, *processor,* serta *disk* dalam menjalankan Apache Spark dan Hadoop MapReduce untuk memproses *big data.*

1.4 Batasan Masalah

Batasan masalah pada tugas akhir ini adalah sebagai berikut:

a. Aplikasi berbasis linux yang berjalan di OS Ubuntu 16.04 LTS.
b. Aplikasi menggunakan bahasa pemrograman Scala.
c. Aplikasi dijalankan di atas *Hadoop Distributed File System (HDFS)* untuk penyimpanan datanya.
d. Sistem berjalan pada *single node*.
e. *File input* berekstensi .txt.
f. Data uji yang digunakan tidak terstruktur.

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 di fabebook 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 2,4 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 Hadoop versi 2.9 dan Apache Spark versi 2.7.

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 I PENDAHULUAN

Bab ini menjelaskan tentang permasalahan serta solusi dari masalah tersebut.

BAB II KAJIAN PUSTAKA

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

BAB III PERANCANGAN

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

BAB IV PENGUJIAN DAN ANALISIS

Bab ini berisi tentang pengujian sistem serta analisis hasil dari keluaran sistem

BAB V KESIMPULAN DAN SARAN

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