BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi menigkatkan kecepatan pertumbuhan data di berbagai bidang. Untuk menanggulangi masalah tersebut, banyak perusahaan meninggalkan sistem manajemen basis data relasional dan berpindah ke basis data terdistribusi yang lebih banyak memuat berbagai jenis data dan dapat tersebar dalam banyak mesin. Selain itu, karena alasan ketersediaan dan reliabilitaslah kenapa basis data terdistribusi lebih banyak digunakan saat ini. Sekalipun terjadi kegagalan pada suatu *node* atau komunikasi diantaranya, *node* lain dapat dengan mudah melakukan *backup* dan tetap membuat basis data berfungsi dan memenuhi kebutuhan aplikasi yang menggunakannya seolah tidak ada kendala. Kelebihan basis data terdistribusi yang paling utama adalah kecepatan pemroses kuerinya yang sangat diandalkan dalam menangani data berjumlah besar.

Namun, masalah muncul terkait dengan data yang berukuran besar. Saat ini kecepatan data yang dihasilkan menuntut peningkatan kecepatan pada pemrosesannya. Data yang tak kunjung dianalisis atau membutuhkan waktu yang lebih lama dalam pemrosesannya akan terdesak oleh data baru yang muncul. Kecepatan pada proses analisis sangat dibutuhkan untuk menanggulangi hal tersebut. Di sisi lain, data yang dihasilkan dari berbagai sumber bersifat multi-dimensi atau memiliki banyak atribut yang membuatnya sulit untuk dilakukan analisis. Salah satu contohnya adalah data dokumen teks. Data jenis tersebut tidak dapat langsung diketahui keberadaan pola dan kemiripan yang terdapat di dalamnya dan diperlukan metode khusus seperti ekstraksi fitur untuk mengolahnya sehingga dapat dimanfaatkan kemudian. Hasil ekstraksi fitur dapat digunakan sebagai masukan untuk berbagai analisis data, seperti pengelompokkan. Pengelompokkan data berfungsi untuk mengetahui keterkaitan antardata dalam suatu kumpulan.

Terlebih lagi data dokumen teks apabila dilakukan pengelompokkan dapat dilihat adanya kemiripan pada beberapa subset data, semisal membahas topik yang sama.

Akan tetapi, beberapa data dokumen teks tidak memiliki label sehingga tidak dapat dikelompokkan menggunakan algoritma *supervised learning*. Dalam penelitian ini dipakai algoritma *Self Organizing Map* karena salah satu algoritma pengelompokkan ini melatih data tanpa membutuhkan label dalam data sehingga dapat menangani kumpulan dokumen teks yang belum diketahui kategori topiknya. Beberapa penelitian menunjukkan, algoritma SOM dapat mengelompokkan data berdimensi tinggi dan menguranginya menjadi dimensi yang lebih kecil sehingga dapat diketahui kelompok subset yang memiliki banyak kemiripan. Algoritma SOM juga mengadaptasi pembelajaran kompetitif dimana selama pelatihan neuron masukan akan mendekat ke arah klaster yang menang yang memiliki jarak terdekat dengannya. Penelitian lain juga menunjukkan bahwa SOM terkadang memiliki waktu komputasi yang lebih baik daripada algoritma pengelompokkan data lain seperti *k-means* (Abbas, 2008). Beberapa metode pengelompokkan pun tak sedikit yang menghasilkan klaster yang buruk dimana terdapat ketimpangan jumlah data antara satu klaster dengan klaster yang lain. Oleh karena itu, dilakukan pula uji hasil klaster untuk melihat apakah algoritma SOM dapat menghasilkan pesebaran data pada klaster dengan baik.

Untuk menanggulangi kecepatan pertumbuhan data, diperlukan pula kecepatan dalam proses impor data dan lingkungan yang mendukungnya. Salah satu sistem *big data* yang bersifat *cloud-based* adalah HDFS (*Hadoop Distributed File System)* yang dapat menyimpan data berukuran hingga ratusan Terabita dan dapat melakukan pemrosesan dengan cepat menggunakan metode *map-reduce* sedangkan untuk basis data terdistribusi peneliti menggunakan HBase, yaitu basis data non-relasional yang bersifat *open-source* dan terdistribusi yang dimodelkan dari Google Bigtable. Data

yang telah masuk ke dalam HDFS kemudian diimpor ke dalam HBase menggunakan *map-reduce* lalu dilakukan pengelompokkan menggunakan *Self-Organizing Map* (SOM).

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang dan permasalahan yang telah disebutkan sebelumnya maka perumusan masalah utama yang akan dibahas pada penelitian ini adalah masih banyak pengguna basis data terutama basis data terdistribusi yang belum mengelompokkan data khususnya yang bersifat multi-dimensi sehingga analisis implementasi salah satu algoritma pengelompokkan data, yaitu *Self Organizing Map* pada basis data terdistribusi perlu dilakukan untuk menanganni masalah tersebut..

1.3 Batasan Masalah

Batasan masalah dari penelitian ini adalah:

1. *Dataset* yang digunakan untuk training berupa file data abstrak penelitian yang diambil dari *Elecctronic Theses & Dissertations* (ETD) UGM pada tanggal 29 Mei 2017 serta beberapa data yang bersumber dari Internet berupa *tab separated value* (.tsv)
2. Sistem big data yang akan digunakan adalah *Hadoop Distributed File System* (HDFS) yang dikonfigurasi secara *Fully-Distributed Mode*
3. *Tools* untuk basis data adalah Apache Hbase yang dikonfigurasi secara *Fully-Distributed Mode*
4. Parameter yang akan diuji berupa jumlah perulangan (*epoch*), waktu komputasi dan sebaran data antarklaster.

1.4 Tujuan Penelitian

Penelitian ini memiliki tujuan untuk mengimplementasi algoritma *Self-Organizing Map* (SOM) pada basis data terdistribusi HBase di lingkungan *map-reduce* untuk mengelompokkan data abstrak penelitian berdimensi tinggi ke dalam klaster-klaster sesuai topik penelitiannya.

1.5 Manfaat Penelitian

Secara ilmiah penelitian ini memberikan manfaat untuk mengurangi pekerjaan proses analisis data dalam sistem *big data* karena data yang tersimpan telah terkelompok sesuai dengan kemiripannya masing-masing berdasarkan kata kunci yang telah diekstraksi serta dapat memberikan referensi untuk penelitian selanjutnya yang memiliki keterkaitan sama.

1.6 Metode Penelitian

Langkah-langkah yang akan dilakukan selama penelitian adalah sebagai berikut.

1. Menentukan permasalahan, tujuan penelitian, serta batasan masalah dengan melakukan diskusi bersama dosen pembimbing.
2. Melakukan studi literatur untuk mengetahui perkembangan terkini dari topik penelitian yang diusulkan. Sudah sejauh mana penelitian-penelitian serupa yang dilakukan. Topik yang akan dipelajari mengenai sistem *big data*, *Self-Organizing Map* (SOM), *Hadoop Distributed File System* (HDFS), Apache HBase serta teori-teori lainnya yang mendukung penelitian. Literatur yang akan dipelajari dapat berupa buku, jurnal, *paper*, skripsi dan karya ilmiah lainnya baik yang didapatkan secara *offline* di perpustakaan maupun *online* di internet.
3. Melakukan identifikasi terhadap spesifikasi kebutuhan sistem. Mengidentifikasi data yang akan digunakan sebagai *data training* serta menentukan *tools* yang akan digunakan untuk membangun sistem baik perangkat lunak maupun keras.
4. Berdasarkan hasil analisis, dilakukan perancangan dengan memodelkan sistem dan dari pemodelan inilah nantinya akan diperoleh gambaran dari pemecahan masalah yang telah diidentifikasi sebelumnya. *Dataset* untuk *training* berupa file berisi data ilmiah yang diperoleh dari *Electronic Theses & Dissertations* (ETD) UGM yang bersifat multidimensi setelah diimpor ke dalam basis data akan dipetakan menggunakan *Self-Organizing Map* (SOM). Setelah didapatkan data yang telah terkelompok

berdasarkan kata kunci yang diekstraksi kemudian disimpan kembali dalam basis data.

5. Mengimplementasikan sesuai dengan rancangan yang telah dibuat berdasarkan analisa terhadap data yang diuji menggunakan *tools* Apache HBase sebagai basis data dan *map-reduce* sebagai *tool data ingestion* dan *data extraction*. Setelah sebelumnya dilakukan instalasi perangkat lunak yang dibutuhkan dalam penelitian.
6. Melakukan pengujian dengan menghitung waktu komputasi untuk proses impor dan pencarian kata kunci serta menghitung akurasi dari model untuk proses *data analysis* menggunakan SOM.

1.7 Sistematika Penulisan

Sistematika penulisan dalam laporan skripsi ini akan terdiri atas tujuh bab, yaitu:

1. BAB I PENDAHULUAN

   Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan, manfaat, metode dan sistematika penulisan dari penelitian yang dilakukan.
2. BAB II TINJAUAN PUSTAKA

   Bab ini memuat penelitian terdahulu yang memiliki keterkaitan dengan penelitian yang dilakukan serta dijadikan sebagai bahan referensi. Di bab ini pula akan dijelaskan perbandingan antarpenelitian agar didapatkan perbedaan dan persamaan dengan penelitian yang dilakukan. Selain itu juga memuat penjelasan mengenai perbedaan penelitian yang dilakukan dengan penelitian-penelitian terkait sebelumnya.
3. BAB III LANDASAN TEORI

   Bab ini berisi tentang teori-teori yang digunakan sebagai dasar untuk melakukan penelitian dan acuan untuk penulisan laporan. Dalam hal ini teori yang dipakai tidak jauh dari *Self-Organizing Map* (SOM), *data analysis*, *Hadoop Distributed File System* (HDFS), Apache Hbase, dan lain-lain.

4. BAB IV ANALISIS DAN PERANCANGAN SISTEM

   Bab ini akan membahas mengenai analisis: permasalahan, *dataset* dan algoritma, spesifikasi perangkat yang digunakan baik lunak maupun keras, rancangan arsitektur sistem dan rancangan pengujian.

5. BAB V IMPLEMENTASI

   Pada bab ini akan dijelaskan implementasi dari sistem yang telah dibangun, dalam hal ini implementasi *Self-Organizing Map* (SOM) berdasarkan pada analisis dan perancangan yang telah dilakukan dan disampaikan pada bab sebelumnya.

6. BAB VI HASIL DAN PEMBAHASAN

   Bab ini akan menerangkan mengenai hasil implementasi sistem yang telah dilakukan dan akan dibahas mengenai keberhasilan dari pengujian sistem.

7. BAB VII KESIMPULAN DAN SARAN

   Di bab ini akan dipaparkan mengenai kesimpulan dari penelitian yang telah dilakukan serta saran-saran untuk penelitian terkait selanjutnya. Kesimpulan diambil berdasarkan analisis dan pengujian yang telah dilakukan.