BAB 1

PENDAHULUAN

1.1 Latar Belakang

Berkembangnya penggunaan sistem informasi di jaman sekarang mengakibatkan data dapat dihasilkan dalam jumlah yang sangat banyak. Data yang jumlahnya sangat banyak ini dikumpulkan dan disimpan dalam tabel basis data untuk keperluan analisis di masa yang akan datang. Data yang dikumpulkan dapat mengambil kapasitas penyimpanan yang besar, sehingga proses analisis data menjadi sangat lambat. Dampak yang ditimbulkan dari pertumbuhan data menyebabkan basis data konvensional menjadi kurang efektif untuk pengolahan data. Oleh karena itu, teknologi *big data* digunakan untuk mengurangi biaya penyimpanan dan komputasi data, sehingga kapasitas data dapat ditingkatkan dan data berukuran besar dapat lebih mudah untuk diolah.

Big data adalah kumpulan data yang telah dikumpukan dalam jumlah yang sangat besar pada rentang waktu tertentu. *Big data* disimpan, diolah, dan dilakukan analisis agar menghasilkan informasi yang bermanfaat sebagai dasar pengambilan keputusan atau kebijakan yang lebih tepat berdasarkan data sebenarnya. Karena *Big data* memiliki ukuran data yang besar, maka proses analisis *big data* harus dilakukan secara paralel. Caranya adalah dengan membagi data ke beberapa komputer untuk diolah masing-masing komputer tersebut. Konsep ini disebut dengan sistem terdistribusi. Sistem terdistribusi adalah solusi pengolahan *big data* karena terbukti dapat mengurangi biaya penyimpanan dan komputasi data dari pemrosesan data secara paralel.

Untuk melakukan proses analisis data, diperlukan teknik untuk mencari tahu kesamaan sifat yang dimiliki oleh sekumpulan data. Salah satu teknik yang dapat digunakan adalah *data mining*. *Data mining* adalah teknik untuk melihat kesamaan sifat yang terbentuk dari sekumpulan data. Teknik *data mining* dapat membantu proses analisis data pada lingkungan *big data*. Pemodelan data mining untuk *big data* dijalankan pada sistem terdistribusi, sehingga waktu komputasi dapat diminimalkan. Hasil *data mining* dipakai untuk berbagai macam kebutuhan. Umumnya, sebuah perusahaan meminta data dari perusahaan lain untuk kebutuhan analisis. Hal ini dapat menimbulkan kasus pelanggaran privasi ketika perusahaan lain melakukan teknik *data mining* pada sekumpulan data yang masih banyak mengandung data privasi. Oleh karena itu, diperlukan teknik khusus agar perusahaan masih dapat mencari informasi yang berharga dari data yang diberikan dan privasi data tetap terlindungi meskipun data tersebut dilakukan proses *data mining*.

Perlindungan privasi pada *data mining* dapat dicapai dengan menggunakan metode enkripsi dan anonimisasi pada data yang akan diberikan. Enkripsi adalah metode yang memanfaatkan pola atau kunci tertentu untuk melindungi data yang sifatnya sensitif. Anonimisasi adalah metode yang menyamarkan satu atau lebih nilai atribut data agar data seseorang tidak dapat saling dibedakan dengan data lainnya. Salah satu kekurangan dari metode enkripsi dibandingkan metode anonimisasi adalah keamanan enkripsi dapat diretas melalui penalaran hubungan nilai atribut yang unik untuk setiap baris data. Penalaran ini dicapai dengan menggabungkan seluruh nilai atribut yang unik pada masing-masing baris data untuk membentuk sebuah pola kelompok data. Penalaran ini sangat berbahaya karena menghubungkan nilai atribut data yang secara tidak langsung dapat mengungkapkan entitas pemilik data. Dengan menerapkan konsep anonimisasi diharapkan nilai keterhubungan antaratribut data diperkecil sehingga privasi dapat terlindungi.

Dengan melakukan metode anonimisasi pada sebagian nilai atribut data untuk kelompok data yang sama, maka bobot informasi yang diperoleh akan semakin kecil. Bobot informasi menunjukan seberapa besar peluang untuk mengetahui arti dari nilai data yang telah dianonimisasi. Oleh karena itu, semakin kecil bobot informasi yang diperoleh maka kelompok data yang dapat membentuk entitas data akan semakin kecil sehingga perlindungan privasi akan semakin aman. Akan tetapi dengan semakin kecil bobot informasi yang diperoleh, maka nilai akurasi yang dihasilkan oleh metode anonimisasi akan semakin kecil. Nilai akurasi menunjukan seberapa tepat model dapat menentukan sebuah data merupakan anggota dari kelompok data lain, sehingga kualitas informasi yang diperoleh semakin buruk. Oleh karena itu diperlukan cara untuk menyeimbangkan keamanan dan nilai akurasi informasi. Permasalahan *k-anonymity* melibatkan pencarian solusi dalam menyeimbangkan nilai akurasi informasi yang diperoleh dengan nilai informasi yang dapat dilindungi.

Metode *k-anonymity* dapat diuji menggunakan pendekatan generalisasi dan supresi. Hasil yang didapat dari penggunaan metode *k-anonymity* dinilai masih kurang untuk mendapatkan nilai akurasi data yang lebih baik, karena tingginya jumlah informasi yang hilang. Berdasarkan penelitian-penelitian yang telah dilakukan sebelumnya, permasalah metode *k-anonymity* dapat teratasi melalui penerapan algoritma *k-member clustering* untuk pengelompokan data. Penerapan algoritma *k-member clustering* pada algoritma *Greedy k-member clustering* menjadi baik karena algoritma *greedy* mencari solusi optimal untuk meminimalkan jumlah informasi yang hilang. Agar algoritma *Greedy k-member clustering* dapat dijalankan pada lingkungan *big data* dengan efisien, maka akan dipilih*framework* Spark untuk waktu komputasi yang lebih optimal.

Spark adalah *framework* yang tepat untuk melakukan proses anonimisasi data pada lingkungan *big data*, karena pekerjaan pengolahan data yang besar dapat dibagi ke beberapa komputer pada sistem terdistribusi. Penggunaan Spark dipilih karena Hadoop memiliki waktu pemrosesan *big data* yang lebih lama dari Spark karena melakukan komputasi pada *hardisk*, sedangkan Spark dapat melakukan komputasi pada memori. Selain itu Spark memiliki jenis *library* yang lebih beragam dibandingkan dengan Hadoop. Spark mampu melakukan pemrosesan teknik *data mining* pada lingkungan *big data* menggunakan *library* tambahan yaitu Spark MLlib. Spark MLlib menfasilitasi pemodelan *data mining* yaitu klasifikasi dan pengelompokan/ *clustering*. Kekurangan dari Spark adalah tidak memiliki penyimpanan yang tetap, sehingga membutuhkan mekanisme penyimpanan Hadoop, agar hasil pemrosesan data dapat tersimpan dalam *hardisk* komputer.

Pada skripsi ini, akan dibuat dua jenis perangkat lunak yaitu perangkat lunak anonimisasi data dan perangkat lunak analisis data. Perangkat lunak anonimisasi data menggunakan konsep *k-anonimity* dengan algoritma *Greedy k-member clustering* agar sebuah data tidak dapat dibedakan dengan $k-1$ data lainnya. Perangkat lunak anonimisasi data dibuat dengan bahasa Scala dan berjalan di atas Spark untuk meminimalkan waktu komputasi proses anonimisasi di lingkungan *big data*. Algoritma *Greedy k-member clustering* dinilai tepat untuk melakukan pengelompokan data karena meminimalkan jumlah informasi yang hilang saat proses *data mining* yang terbukti pada penelitian sebelumnya. Kedua jenis perangkat lunak ini menerima data input dalam format CSV. Untuk tampilannya, kedua perangkat lunak ini akan dibuat menggunakan GUI dari *library* Scala-swing. Penelitian ini memiliki tujuan utama yaitu membandingkan nilai akurasi dari hasil teknik *data mining* sebelum dan setelah dilakukan proses anonimisasi data.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah pada skripsi ini adalah sebagai berikut:

1. Bagaimana cara kerja algoritma *Greedy k-member clustering* ?
2. Bagaimana implementasi algoritma *Greedy k-member clustering* pada Spark?
3. Bagaimana hasil *data mining* sebelum dan setelah dilakukan anonimisasi?

1.3 Tujuan

Berdasarkan rumusan masalah di atas, tujuan dari skripsi ini adalah sebagai berikut:

1. Mempelajari cara kerja algoritma *Greedy k-member clustering*.
2. Mengimplementasikan algoritma *Greedy k-member clustering* pada Spark.
3. Menganalisis hasil *data mining* sebelum dan setelah dilakukan anonimisasi.

1.4 Batasan Masalah

Batasan masalah pada pengerjaan skripsi ini adalah sebagai berikut:

1. Perangkat lunak dapat berjalan diatas Spark.
2. Perangkat lunak dapat menerapkan algoritma *Greedy k-member clustering*.
3. Perangkat lunak dapat diimplementasikan menggunakan *library* Scala-swing.
4. Perangkat lunak hanya menerima input data semi terstruktur CSV dan XML.
5. Menggunakan teknik *data mining* yang tersedia pada *library* Spark MLlib.
6. Membandingkan hasil *data mining* sebelum dan setelah dilakukan anonimisasi.

1.5 Metodologi

Bagian-bagian pengerjaan skripsi ini adalah sebagai berikut:

1. Mempelajari dasar-dasar privasi data.
2. Mempelajari konsep *k-anonimity* pada algoritma *Greedy k-member clustering*.
3. Mempelajari teknik-teknik dasar *data mining*.
4. Mempelajari konsep Hadoop, Spark, dan Spark MLlib.
5. Mempelajari bahasa pemrograman Scala pada Spark.
6. Melakukan analisis masalah dan mengumpulkan data studi kasus.
7. Mengimplementasikan algoritma *Greedy k-member clustering* pada Spark.
8. Mengimplementasikan tampilan perangkat lunak menggunakan *library* Scala-swing.
9. Mengimplementasikan teknik *data mining* menggunakan *library* Spark MLlib.
10. Melakukan pengujian fungsional dan experimental.
11. Melakukan analisis hasil *data mining* sebelum dan setelah dilakukan anonimisasi.
12. Menarik kesimpulan berdasarkan hasil eksperimen yang telah dilakukan.

1.6 Sistematika Pembahasan

Pengerjaan skripsi ini tersusun atas enam bab sebagai berikut:

- Bab 1 Pendahuluan
  Berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika pembahasan.

- Bab 2 Landasan Teori
  Berisi landasan teori mengenai konsep privasi, teknik *data mining*, *privacy-preserving data mining*, *k-anonymity*, algoritma *greedy k-member clustering*, metrik *distance* dan *information loss*, teknologi *big data*, pemrograman scala, dan format penyimpanan data.

- Bab 3 Analisis
  Berisi analisis penelitian mengenai analisis masalah (dataset eksperimen, *personally identifiable information*, perhitungan *distance* dan *information loss*, algoritma *greedy k-member clustering*, *k-anonymity*, *domain generalization hierarchy*), eksplorasi spark (instalasi spark, pembuatan *project* spark, menjalankan program spark), studi kasus (eksperimen scala, eksperimen spark), dan gambaran umum perangkat lunak (diagram kelas dan diagram aktivitas).

- Bab 4 Perancangan
  Berisi perancangan antarmuka perangkat lunak anonimisasi data dan analisis data, diagram kelas lengkap, masukan perangkat lunak anonimisasi data dan analisis data.

- Bab 5 Implementasi dan Pengujian
  Berisi implementasi perangkat lunak anonimisasi data dan analisis data, pengujian fungsional, pengujian eksperimental, dan melakukan analisis terhadap hasil pengujian.

- Bab 6 Kesimpulan dan Saran
  Berisi kesimpulan penelitian dan saran untuk penelitian selanjutnya.