SKRIPSI

REDUKSI BIG DATA DENGAN ALGORITMA CLUSTERING AGGLOMERATIVE UNTUK SISTEM TERDISTRIBUSI SPARK

Matthew Ariel

NPM: 2015730010

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2019

UNDERGRADUATE THESIS

BIG DATA REDUCTION USING AGGLOMERATIVE CLUSTERING ALGORITHM FOR SPARK DISTRIBUTED SYSTEM

Matthew Ariel

NPM: 2015730010

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2019

LEMBAR PENGESAHAN

REDUKSI BIG DATA DENGAN ALGORITMA CLUSTERING AGGLOMERATIVE UNTUK SISTEM TERDISTRIBUSI SPARK

Matthew Ariel

NPM: 2015730010

Bandung, 10 Desember 2019

Menyetujui,

Pembimbing

Dr. Veronica Sri Moertini

Ketua Tim Penguji

Mariskha Tri Adithia, P.D.Eng

Anggota Tim Penguji

Kristopher David Harjono, M.T.

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

REDUKSI BIG DATA DENGAN ALGORITMA CLUSTERING AGGLOMERATIVE UNTUK SISTEM TERDISTRIBUSI SPARK

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.
Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 10 Desember 2019

Meterai
Rp. 6000

Matthew Ariel
NPM: 2015730010

ABSTRAK

Big data adalah istilah yang menggambarkan kumpulan data dalam jumlah yang sangat besar, baik data yang terstruktur maupun data yang tidak terstruktur. Kumpulan data tersebut menyimpan informasi yang bisa dianalisis dan diproses untuk memberikan wawasan kepada organisasi atau perusahaan. *Big data* dapat mencapai *petabyte* dan menghabiskan banyak tempat penyimpanan.

Big data perlu direduksi untuk menghemat tempat penyimpanan. Algoritma *Hierarchical Agglomerative Clustering* dapat digunakan untuk mereduksi data. Dengan bantuan sistem terdistribusi seperti Hadoop, proses reduksi data dapat dilakukan secara paralel dan lebih cepat. Sayangnya, teknologi Hadoop masih dapat dikatakan 'terlalu lambat' dalam melakukan proses reduksi data karena hasil sementara dari setiap tahap akan disimpan di *disk* sampai dibutuhkan kembali di tahap selanjutnya.

Untuk mempercepat proses reduksi data, Hadoop dapat digantikan dengan Spark. Spark adalah sistem terdistribusi, mirip seperti Hadoop. Tetapi, yang membedakan antara Hadoop dengan Spark adalah pada cara penyimpanan sementara saat melakukan proses reduksi data. Hadoop menggunakan *disk* sebagai tempat penyimpanan sementaranya, sedangkan Spark menggunakan memori sebagai tempat penyimpanan sementaranya. Pembacaan dan penulisan akan lebih cepat saat menggunakan memori dibandingkan dengan menggunakan *disk*, sehingga Spark akan lebih cepat dibandingkan dengan Hadoop.

Perangkat lunak dibuat untuk mengimplementasikan algoritma *Hierarchical Agglomerative Clustering* dalam Spark. Pengujian juga dilakukan dengan membandingkan waktu eksekusi algoritma *Hierarchical Agglomerative Clustering* saat diimplementasikan pada Hadoop dan saat diimplementasikan pada Spark. Waktu eksekusi dicatat untuk ukuran data 1GB, 2GB, 3GB, 5GB, 10GB, 15GB, dan 20GB.

Berdasarkan hasil pengujian, Spark memiliki waktu eksekusi yang lebih cepat dibandingkan dengan Hadoop pada jumlah partisi yang besar. Waktu eksekusi Spark menurun ketika jumlah partisi ditingkatkan, sedangkan waktu eksekusi Hadoop menurun ketika jumlah partisi ditingkatkan. Waktu eksekusi terbaik Spark masih lebih cepat dibandung waktu eksekusi terbaik Hadoop.

Kata-kata kunci: *Big Data*, Reduksi Data, *Hierarchical Agglomerative Clustering*, Spark, Hadoop

ABSTRACT

Big data is a term that describes the large volume of data, both structured and unstructured. The data set stores informationcan be analyzed and processed to provide insight to organization or company. Big data can reach up to petabytes and takes a lot of storage spaces.

Big data need to be reduce to save storage space. The Hierarchical Agglomerative Clustering algorithm can be used to reduce data. With the help of distributed systems such as Hadoop, reduction process can be done in parallel with less execution time. Unfortunately, Hadoop can still be said to be 'too slow' in the process of data reduction because temporary results from each stage will be stored on the disk until it is needed again at a later stage.

To speed up the data reduction process, Hadoop can be replaced with Spark. Spark is a distributed system, similar to Hadoop. However, what distinguishes Hadoop from Spark is the way Spark temporarily store data. Hadoop uses disk as its temporary storage, while Spark uses memory as its temporary storage. Readi and write process will be faster when using memory than using disks, Spark will be faster than Hadoop.

The Hierarchical Agglomerative Clustering algorithm is implemented in the software. Experiment were done by comparing the execution time of the Hierarchical Agglomerative Clustering algorithm when implemented on Hadoop and Spark. The execution time is recorded for 1GB, 2GB, 3GB, 5GB, 10GB, 15GB, dan 20GB of data.

Based on the experiment, Spark has a faster execution time compared to Hadoop on a large number of partitions. Spark execution time decreases when the number of partitions is increased, whereas Hadoop execution time decreases when the number of partitions is increased. Spark best executiron time is still much better than Hadoop best execution time.

Keywords: Big Data, Data Reduction, Hierarchical Agglomerative Clustering, Spark, Hadoop

Dipersembahkan untuk Tuhan YME, keluarga tercinta, Ibu Veronica sebagai dosen pembimbing, teman-teman yang telah berperan dalam pembuatan skripsi ini, dan diri sendiri.

KATA PENGANTAR

Puji dan syukur kehadirat Tuhan Yang Maha Esa atas berkat rahmat serta kasih-Nya sehingga penulis dapat menyelesaikan skripsi ini yang mengambil judul Ŕeduksi Big Data dengan Algoritma Clustering Agglomerative untuk Sistem Terdistribusi Spark. Penulisan skripsi ini diajukan untuk memenuhi salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika Universitas Katolik Parahyangan. Pada penyusunan dan penulisan skripsi ini, penulis menyadari bahwa penyusunan dan penulisan skripsi ini juga tidak terlepas dari bantuan berbagai pihak, baik langsung maupun tidak langsung. Secara khusus, penulis ingin berterima kasih kepada:

1. Keluarga yang selalu memberi dukungan dan mengurus penulis selama mengerjakan skripsi.
2. Ibu Veronica Sri Moertini selaku dosen pembimbing yang telah membimbing penulis dan memberikan dukungan dan bantuan kepada penulis dalam proses penyusunan skripsi ini.
3. Sahabat yang selalu memberi dukungan dan semangat, khusunya Edick Zakari, Felicia Christiany, Irvan Wijaya, Emanuel Yudistira, Hapsari Laksmi, Kezia, Raymond Nagawijaya, Cornelius David, Aditya Putra, Thoby, dan teman-teman yang tidak dapat saya sebutkan.
4. Sahabat seperjuangan dari jurusan lain Rensky Picco dan Christian Stefano.
5. Sahabat SMURF -rm -rf, Wanted To Resign, sad, Mati Lari Sudah, dan shitos.
6. Teman dan kerabat lainya yang tidak bisa saya sebutkan satu-persatu.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna. Oleh karena itu, penu- lis memohon maaf jika terdapat kesalahan. Penulis juga mengharapkan kritik dan saran yang membangun untuk menyempurnakan skripsi ini. Semoga skripsi ini dapat memberi informasi yang bermanfaat dan menjadi inspirasi untuk penelitian-penelitian berikutnya.

Bandung, Desember 2019

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Big data adalah sebuah istilah yang menggambarkan volume data yang besar, baik data yang terstruktur maupun data yang tidak terstruktur. Data-data tersebut memiliki potensi untuk digali menjadi informasi yang penting. Dalam bidang *big data* ada beberapa tantangan seperti volume data yang besar, kecepatan aliran data yang masuk, dan variasi data dengan format yang berbeda. Tantangan tersebut membuat aplikasi pemrosesan data tradisional tidak bisa memproses dan menganalisis *big data*. Muncul teknologi-teknologi seperti Hadoop dan Spark yang dirancang khusus untuk menangani *big data*.

Big data akan lebih mudah dianalisis dan diterapkan teknik-teknik *data-mining* ketika volume *big data* tersebut telah direduksi. Reduksi data adalah proses mengecilkan ukuran data dengan mengambil rangkuman dari sekelompok data. *Big data* akan dipecah menjadi beberapa bagian. Dari bagian tersebut akan diolah menggunakan algoritma *Hierarchical Agglomerative Clustering* untuk menghasilkan *cluster-cluster*. Dari *cluster-cluster* yang dihasilkan, akan diambil rangkuman yang bisa menggambarkan karakterisik setiap *cluster*. Rangkuman yang menggambarkan karaketristik *cluster* disebut sebagai pola. Pola mengandung jumlah total objek, nilai minimum, maksimum, nilai rata-rata, dan standar-deviasi dari setiap atribut pada *cluster*. Nilai-nilai tersebut diambil karena dapat menggambarkan karakteristik *cluster*. Dengan mereduksi data, kita bisa menghemat biaya pengiriman data, *disk space*, dan jumlah data yang diproses. Hasil dari reduksi *big data* harus bisa mewakili data yang belum direduksi secara akurat.

Algoritma *Hierarchical Agglomerative Clustering* membangun hirarki dari sekelompok data. Setiap objek akan ditempatkan kepada *cluster*-nya tersendiri pada awalnya. Kemudian, *cluster* terdekat akan digabung menjadi satu *cluster*. Penggabungan *cluster* terdekat akan diulang sampai hanya satu *cluster* yang tersisa. Proses pembangunan hierarki dapat digambarkan dengan *dendrogram*. *Dendrogram* akan dipotong untuk menghasilkan *cluster-cluster*. Dari setiap *cluster* hasil perpotongan, akan dicari polanya dan disimpan sebagai hasil akhir.

Untuk mempercepat proses reduksi, pekerjaan dapat dipecah dan dikerjakan secara paralel dengan bantuan Hadoop. Dengan bantuan Hadoop, proses reduksi data akan lebih cepat. Implementasi algoritma *Hierarchical Agglomerative Clustering* sudah dilakukan pada Hadoop [1]. Hasil dari penelitian tersebut membuktikan bahwa algoritma *Hierarchical Agglomerative Clustering* sudah dapat mereduksi data dengan menyimpan pola hasil reduksi. Tetapi ada beberapa limitasi yang dimiliki Hadoop.

Walau Hadoop dapat melakukan proses reduksi secara paralel, waktu yang dibutuhkan Hadoop untuk melakukan reduksi data masih terlalu lambat. Hadoop banyak melakukan proses penulisan dan pembacaan kepada disk. Dari satu tahap ke tahap lainya, Hadoop akan menulis dan membaca hasil sementara kepada disk. Hadoop perlu digantikan dengan Spark untuk mencepat proses reduksi.

Spark adalah *distributed cluster-computing framework* yang bisa menggantikan MapReduce beserta kekurangannya. *In-memory processing* pada Spark dapat mengalahkan kecepatan pemrosesan pada Hadoop MapReduce. Karena data disimpan pada RAM, kecepatan pemrosesan akan jauh lebih cepat. Spark membaca data yang akan direduksi dari RAM. Pembacaan data dari RAM akan lebih cepat dibanding disk.

Pada skripsi ini, dibangun sebuah perangkat lunak yang dapat mereduksi *big data*. Perangkat lunak tersebut akan dibangun menggunakan *framework* terdistribusi Spark dan mengimplementasikan algoritma *Hierarchical Agglomerative Clustering* yang khusus dirancang untuk lingkungan Spark. Perangkat lunak

dapat menampilkan hasil reduksi dalam format tabel. Dengan menggunakan Spark, waktu proses reduksi data menjadi lebih cepat dibanding MapReduce.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, dapat dibentuk rumusan masalah sebagai berikut:

1. Bagaimana cara kerja algoritma *Hierarchical Agglomerative Clustering* berbasis MapReduce untuk mereduksi *big data*?
2. Bagaimana cara mengkustomisasi dan mengimplementasikan algoritma *Agglomerative Clustering* pada sistem tersebar Spark?
3. Bagaimana mengukur performa hasil dari implementasi algoritma *Agglomerative Clustering* pada sistem tersebar Spark?

1.3 Tujuan

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

1. Mempelajari cara kerja algoritma *Hierarchical Agglomerative Clustering* berbasis MapReduce untuk mereduksi *big data*.
2. Mengkustomisasi dan mengimplementasikan algoritma *Hierarchical Agglomerative Clustering* pada lingkungan Spark.
3. Melakukan eksperimen pada Spark dan Hadoop untuk membandingkan waktu eksekusi kedua perangkat lunak.

1.4 Batasan Masalah

Batasan masalah pada skripsi ini adalah sebagai berikut:

1. Pola yang dikomputasi hanya sebatas jumlah objek, nilai minimum, maksimum, rata-rata dan standar deviasi.
2. Perangkat lunak yang dibangun tidak menangani pembersihan masukkan *big data*.

1.5 Metodologi

Metodologi yang digunakan dalam pembuatan skripsi ini adalah:

1. Melakukan studi literatur Hadoop hanya mempelajari konsep dasar dari Hadoop dan *Hadoop Distributed File System* (HDFS).
2. Melakukan studi literatur tentang konsep Apache Spark.
3. Melakukan studi literatur bahasa pemrograman Scala.
4. Melakukan studi literatur tentang algoritma *Hierarchical Agglomerative Clustering*.
5. Melakukan instalasi dan konfigurasi Apache Spark.
6. Melakukan eksperimen dengan bahasa pemrograman Scala.
7. Melakukan eksperimen dengan Spark RDD.

8. Melakukan kustomisasi algoritma *Hierarchical Agglomerative Clustering* untuk Spark.
9. Mencari dan mengumpulkan data uji coba yang bervolume besar.
10. Merancang dan mengimplementasikan perangkat lunak.
11. Melakukan eksperimen terhadap perangkat lunak dan menganalisis hasil eksperimen.
12. Menulis dokumen skripsi.

1.6 Sistematika Pembahasan

Laporan penelitian tersusun ke dalam enam bab secara sistematis sebagai berikut:

- Bab 1 Pendahuluan
  Berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika pembahasan.
- Bab 2 Dasar Teori
  Berisi dasar teori tentang *big data*, *Hierarchical Agglomerative Clustering*, Hadoop, Spark, dan Scala.
- Bab 3 Studi dan Eksplorasi Apache Spark
  Berisi percobaan-percobaan yang dilakukan pada Spark.
- Bab 4 Analisis dan Perancangan
  Berisi analisis masalah, diagram alur, *use case* dan skenario, diagram kelas, dan perancangan antarmuka.
- Bab 5 Implementasi dan Pengujian
  Berisi implementasi antarmuka perangkat lunak, pengujian eksperimen, dan kesimpulan dari pengujian.
- Bab 5 Implementasi dan Pengujian
  Berisi kesimpulan awal sampai akhir penelitian dan saran untuk penelitian selanjutnya.