SKRIPSI

KONSTRUKSI DENDROGRAM DAN PENGELOMPOKAN DENGAN ALGORITMA AGGLOMERATIVE PADA SISTEM TERDISTRIBUSI HADOOP

GDE WIRADITYA SUARJANA

NPM: 2012730042

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

2017

UNDERGRADUATE THESIS

DENDROGRAM CONSTRUCTION AND CLUSTER ANALYSIS WITH AGGLOMERATIVE ALGORITHM ON HADOOP DISTRIBUTED SYSTEM

GDE WIRADITYA SUARJANA

NPM: 2012730042

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2017

LEMBAR PENGESAHAN

KONSTRUKSI DENDROGRAM DAN PENGELOMPOKAN DENGAN ALGORITMA AGGLOMERATIVE PADA SISTEM TERDISTRIBUSI HADOOP

GDE WIRADITYA SUARJANA

NPM: 2012730042

Bandung, 6 Januari 2017

Menyetujui,

Pembimbing

Dr. Veronica Sri Moertini

Ketua Tim Penguji

Vania Natali, M.T.

Anggota Tim Penguji

Luciana Abednego, 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:

KONSTRUKSI DENDROGRAM DAN PENGELOMPOKAN DENGAN ALGORITMA AGGLOMERATIVE PADA SISTEM TERDISTRIBUSI HADOOP

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,

Gde Viraditya Suarjana
NPM: 2012730042

ABSTRAK

Big data adalah istilah untuk set data yang sangat besar atau kompleks sehingga aplikasi pengolahan data tradisional tidak memadai untuk mengolahnya. Analisis big data bisa sangat berguna untuk berbagai bidang ,misalnya analisis *trend* bisnis, mencegah penyakit ,memerangi kejahatan dan sebagainya. Penelitian ini bertujuan untuk mengembangkan perangkat lunak yang dapat melakukan analisis *cluster* dalam *big data* menggunakan algoritma *agglomerative clustering*. Algoritma agglomerative yang digunakan akan menghasilkan sebuah dendrogram, yaitu sebuah pohon yang merepresentasikan *hierarchy* dari sebuah set data.

Perangkat lunak yang dikembangkan dalam makalah akan mencoba mengolah big data dengan cara membagi data set ke beberapa partitisi . Untuk membagi data menjadi beberapa partisi digunakan *framework* MapReduce. Di setiap partisi data akan diproses secara terdistribusi, kemudian hasilnya akan dikumpulkan dan diproses lebih lanjut untuk mendapatkan hasil akhir. Proses terdistribusi ini diharapkan dapat mengurangi waktu komputasi perangkat lunak. Perangkat lunak ini akan dikembangkan dengan menggunakan kerangka MapReduce dalam sistem terdistribusi Hadoop.

Perangkat lunak yang telah dibangun menerima masukan berupa satu atau lebih file teks yang berisi data yang akan diproses dan mengeluarkan keluaran berupa file teks yang berisi *cluster - cluster*. Perangkat lunak memiliki waktu komputasi yang lebih kecil jika dibandingkan dengan perangkat lunak yang tidak terdistribusi. Dari segi keakuratan, perangkat lunak dapat membuat *cluster - cluster* yang cukup akurat dari sebuah set data.

Kata-kata kunci: Big Data,Dendrogram,Agglomerative clustering,MapReduce Hadoop

ABSTRACT

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on. This research aims to develop a software that can do a cluster analysis in a large data set using the agglomerative clustering algorithm . The agglomerative clustering algorithm will produce a dendrogram, a dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.

The software developed in this paper try to solve the big data problem by dividing the data set into partitions. The MapReduce framework is used to divide the data into partitions, then each partition is processed in a distributed system. The results of each partition are then collected and used to calculate the final result. This process is done to reduce the computation time of the software. The software will be developed using the MapRecude framework in a distributed Hadoop System.

The software developed in this paper will receive input in the form of text file(s) containing the data set, and will produce output in the form of a text file containg the result. The software have a smaller computation time compared to a software that are not run in a distributed system. From an accuracy standpoint, the software can do a cluster analysis from a data set with accepatable accuracy.

Keywords: Big Data,Dendrogram,Agglomerative clustering,MapReduce Hadoop

Skripsi ini dipersembahkan kepada keluarga tercinta

KATA PENGANTAR

Puji syukur kami panjatkan kehadirat Tuhan Yang Maha Esa karena dengan rahmat dan karuniaNya, penulis dapat menyelesaikan skripsi berjudul Konstruksi Dendrogram dan Pengelompokan dgn Algoritma Agglomerative pada Sistem Terdistribusi Hadoop ini dengan baik meskipun masih banyak kekurangan didalamnya. Dan juga kami berterima kasih pada Ibu Veronica Sri Moertini selaku Dosen pembimbing yang telah membimbing penulis dalam menyelesaikan skripsi ini. Baik teguran, perkataan, saran, pendapat serta candaan beliau sangat berkesan dan berarti bagi penulis dalam menyelesaikan skripsi ini. Penulis juga sangat berterima kasih kepada keluarga tercinta yang selalu menyemangati penulis disaat penulis putus asa, teman-teman seperjuangan dalam mengerjakan skripsi, dan admin-admin lab komputer FTIS yang selalu setia menunggu kita menggunakan lab. Penulis sangat berharap skripsi ini dapat berguna dalam rangka menambah wawasan serta pengetahuan kita mengenai pengolahan data berukuran besar dengan sistem terdistribusi Hadoop. Penulis menyadari sepenuhnya bahwa di dalam makalah skripsi ini terdapat kekurangan dan jauh dari kata sempurna. Oleh sebab itu, penulis berharap adanya kritik, saran dan usulan demi perbaikan makalah skripsi yang telah penulis buat di masa yang akan datang. Semoga makalah skripsi ini dapat dipahami bagi siapapun yang membacanya. Sekiranya makalah skripsi yang telah disusun ini dapat berguna bagi penulis sendiri maupun orang yang membacanya. Sebelumnya kami mohon maaf apabila terdapat kesalahan kata-kata yang kurang berkenan dan penulis memohon kritik dan saran yang membangun dari Anda demi perbaikan makalah ini di waktu yang akan datang. Terima kasih.

Bandung, Januari 2017

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

big data adalah istilah untuk suatu set data yang ukurannya sangat besar sehingga sulit diproses untuk mendapatkan informasi yang berguna dari set data tersebut. Saat ini umumnya hampir semua informasi elektronik seperti transaksi, lokasi, website log dan sebagainya dicatat ke dalam suatu *database*, karena *database* akan terus diubah dengan data baru maka ada kemungkinan suatu saat jumlah data akan menjadi terlalu besar untuk diolah dengan metode konvensional. Oleh karena itu teknik untuk mengolah *big data* menjadi salah satu tantangan di bidang *data mining*.

Dari data yang jumlahnya besar ini sebenarnya bisa didapatkan informasi - informasi berguna jika datanya diolah menjadi bentuk yang dapat dipahami oleh analis. Salah satu contoh bentuk pengolahan data yang populer adalah *cluster analysis*. Tujuan dari *cluster analysis* adalah mengelompokkan setiap anggota data ke dalam kelompok - kelompok dengan kriteria tertentu. Informasi ini bisa berguna misalnya sebuah bank dapat mengelompokkan nasabahnya ke dalam kelompok - kelompok dan membuat promosi yang mensasar ke kelompok nasabah tertentu.

Tiga faktor yang membuat *big data* sulit untuk diproses adalah *volume*,*velocity* dan *variety*[1]. Volume adalah besar jumlah data, sebuah *big data* bisa berukuran beberapa terabyte sampai petabyte. Velocity adalah kecepatan perubahan data, karena data baru selalu dicatat dalam *real time* maka data akan terus berubah seiring dengan waktu. Variety adalah jenis - jenis data yang dapat menyusun sebuah set *big data* misalnya text, numeric, gambar dan sebagainya.

Selain itu, karena *big data* berukukan sangat besar dan kompleks maka perlu dilakukan pra-olah sebelum *big data* tersebut dapat diproses. Salah satu contoh hal yang harus di pra-olah adalah data yang tidak lengkap, data yang rusak (*corrupted*), format data yang tidak sesuai dengan algoritma dan lain-lain.

Hierarchical clustering adalah metode *cluster analysis* yang sering digunakan untuk mengelompokkan data. *Hierarchical clustering* dibagi menjadi dua yaitu *divisive clustering* dan *agglomerative clustering*. Algoritma *agglomerative* adalah algoritma *clustering* dengan metode *bottoms up*, yang dimaksud dengan metode *bottoms up* adalah pada awalnya semua objek dalam set data dianggap sebagai sebuah *single cluster*. Kemudian dua *cluster* yang terdekat akan digabungkan menjadi sebuah *cluster*. Proses ini dilakukan secara *iterative* sampai seluruh set data berada di dalam satu *cluster*.

Kelemahan dari algoritma *agglomerative* adalah waktu komputasinya meningkat secara kuadratik terhadap jumlah data. Oleh karena itu algoritma *agglomerative* sulit digunakan untuk melakukan *cluster analysis* jika jumlah datanya besar.

Salah satu teknik yang dapat digunakan untuk membantu memproses data dalam jumlah besar adalah *MapReduce*. *MapReduce* membagi *big data* menjadi beberapa set data yang ukurannya lebih kecil. Set - set data ini kemudian dapat diproses secara paralel oleh beberapa komputer agar waktu dan kompleksitas dari setiap proses lebih kecil dari pengolahan *big data* secara langsung.

Framework *MapReduce* yang populer digunakan adalah Apache Hadoop. Hadoop terdiri dari dua bagian utama, yaitu *MapReduce* dan Hadoop Distributed File System (HDFS). *MapReduce* adalah model pemrograman yang ditujukan untuk memproses data berukuran raksasa secara terdistribusi dan paralel dalam *cluster* yang terdiri atas banyak komputer. HDFS adalah sistem file terdistribusi yang dirancang untuk berjalan pada beberapa perangkat keras.

Hadoop berjalan pada sebuah sistem terdistribusi yang disebut sebagai Hadoop *cluster*. Hadoop *cluster* terdiri atas satu komputer *master* yang bertugas mengatur *slave* dan beberapa komputer *slave* yang bertugas menyimpan data dan memproses data secara paralel.

Pada skripsi ini, akan dirancang sebuah perangkat lunak memanfaatkan *MapReduce* yang berjalan dalam lingkungan sistem terdistribusi Hadoop yang dapat melakukan *cluster analysis* dengan menggunakan *hierarchical clustering* dengan algoritma *agglomerative* di data dengan ukuran besar. Perangkat lunak tersebut dapat membuat visualisasi dalam bentuk *dendrogram* dari kelompok - kelompok yang terdapat di dalam data dan memberikan daftar kelompok dan semua anggotanya.

Kelemahan dari algoritma *agglomerative* adalah waktu komputasinya meningkat secara kuadratik terhadap jumlah data. Oleh karena itu faktor yang paling berpengaruh dalam kasus ini adalah *volume*. Dengan perangkat lunak yang dibangun dengan *MapReduce* dan berjalan pada sistem terdistribusi Hadoop diharapkan algoritma *agglomerative* dapat digunakan untuk melakukan *cluster analisys* dari sebuah set data dengan *volume* yang besar dengan waktu yang lebih singkat dari metode *standalone*.

1.2 Rumusan Masalah

Perumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana cara membuat *dendrogram* dengan algoritma *agglomerative*?
2. Bagaimana cara menggunakan *MapReduce* dalam sistem terdistribusi Hadoop?
3. Bagaimana cara merancang dan mengaplikasikan algoritma *agglomerative* untuk konstruksi *dendrogram* dalam lingkungan sistem terdistribusi Hadoop?
4. Bagaimana cara mengukur kinerja dari algoritma yang diimplementasikan dan menguji kebenarannya?

1.3 Tujuan

Tujuan dari penelitian ini adalah:

(a) Melakukan studi algoritma *agglomerative* untuk konstruksi *dendrogram*.

(b) Melakukan eksperimen-eksperimen awal untuk memahami kerja *MapReduce* dalam sistem terdistribusi Hadoop.

(c) Merancang dan mengimplementasikan algoritma *agglomerative* untuk konstruksi *dendrogram* dan pengelompokan *big data* dalam sistem terdistribusi Hadoop.

(d) Melakukan eksperimen untuk mengukur performa algoritma dalam sistem terdistribusi Hadoop dan menguji keakuratan hasil *clustering* dengan cara membandingkan hasil *clustering* algoritma dalam sistem terdistribusi Hadoop dengan algoritma *agglomerative* konvensional(versi *standalone*) untuk pengelompokan *big data*.

1.4 Batasan Masalah

Batasan masalah dari tugas akhir ini adalah:

(a) Atribut dari data hanya bilangan real

1.5 Metodologi

Penyusunan tugas akhir ini menggunakan metodologi sebagai berikut:

(a) Melakukan studi literatur yang berkaitan dengan *clustering* data dengan metode untuk konstruksi *dendrogram* dan *clustering* dengan algoritma *agglomerative*

(b) Melakukan studi literatur tentang sistem terdistribusi Hadoop.

(c) Melakukan eksperimen pada sistem terdistribusi Hadoop dengan membuat program - program kecil seperti wordcount

(d) Merancang teknik-teknik *clustering* data dengan metode algoritma *agglomerative* dalam sistem terdistribusi Hadoop untuk membuat perangkat lunak secara keseluruhan.

(e) Mengimplementasikan rancangan yang telah dibuat menjadi perangkat lunak.

(f) Melakukan pengujian dan eksperimen terhadap perangkat lunak yang telah dibuat.

(g) Melakukan analisis dari hasil pengujian dan eksperimen yang telah dilakukan.

(h) Menyusun dokumen skripsi.

1.6 Sistematika Pembahasan

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

Bab 2 Dasar Teori
Bab 2 berisi teori-teori mengenai data mining, konstruksi *dendrogram* , *clustering* dengan algoritma *agglomerative* dan Hadoop

Bab 3 Analisis
Bab 3 berisi analisis masalah *agglomerative clustering* pada *big data*, rancangan algoritma dan desain perangkat lunak

Bab 4 Perancangan
Bab 4 berisi format input, perancangan perangkat lunak *MapReduce*, dan perancangan perangkat lunak untuk menampilkan hasil akhir.

Bab 5 Implementasi, Pengujian dan Eksperimen
Bab 5 berisi implementasi algoritma konstruksi *dendrogram* dan *clustering* dengan algoritma *agglomerative* dengan *MapReduce* pada sistem terdistribusi Hadoop, pengujian kebenaran perangkat lunak, eksperimen perangkat lunak dengan *big data*, analisis serta kesimpulan eksperimen.

Bab 6 Kesimpulan dan Saran
Bab 6 berisi kesimpulan dari hasil analisis ekperimen, hasil penelitian dan saran.