SKRIPSI

STUDI DAN IMPLEMENTASI R PADA SISTEM TERSEBAR HADOOP UNTUK ANALISIS BIG DATA

Adrian Stefanus Tanuwijaya

NPM: 2015730014

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2020

UNDERGRADUATE THESIS

STUDY AND R IMPLEMENTATION ON HADOOP DISTRIBUTED SYSTEM FOR BIG DATA ANALYSIS

Adrian Stefanus Tanuwijaya

NPM: 2015730014

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

2020

PERNYATAAN

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

STUDI DAN IMPLEMENTASI R PADA SISTEM TERSEBAR HADOOP UNTUK ANALISIS BIG DATA

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 Juni 2020

Adrian Stefanus Tanuwijaya
NPM: 2015730014

LEMBAR PENGESAHAN

STUDI DAN IMPLEMENTASI R PADA SISTEM TERSEBAR HADOOP UNTUK ANALISIS BIG DATA

Adrian Stefanus Tanuwijaya

NPM: 2015730014

Bandung, 10 Juni 2020

Menyetujui,

Pembimbing

Dr. Veronica Sri Moertini

Ketua Tim Penguji	Anggota Tim Penguji
Mariskha Tri Adithia, P.D.Eng	Vania Natali, 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:

STUDI DAN IMPLEMENTASI R PADA SISTEM TERSEBAR HADOOP UNTUK ANALISIS BIG DATA

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 Juni 2020

Adrian Stefanus Tanuwijaya
NPM: 2015730014

ABSTRAK

Perkembangan internet membuat pertumbuhan data setiap tahunnya semakin banyak. Data yang diperoleh semakin tidak terstruktur, banyak, dan semakin sulit diproses hanya menggunakan pemrosesan tradisional, sehingga muncul istilah big data. Big data dapat dianalisis untuk mendapatkan informasi dan pengetahuan yang berguna untuk menarik kesimpulan, dibutuhkan teknologi untuk menangani data yang jumlahnya sangat besar dan dilakukan komputasi statistik untuk menarik kesimpulan. R merupakan salah satu bahasa yang dapat digunakan untuk melakukan komputasi statistik dan Apache Hadoop digunakan untuk melakukan penyimpanan dan komputasi secara terdistribusi.

RHadoop merupakan salah satu cara untuk mengintegrasikan R dan Hadoop. RHadoop merupakan projek *open source* yang terdiri dari 5 *package* untuk menganalisis data dengan Hadoop melalui R. Pada skripsi ini, akan dilakukan studi bahasa R untuk komputasi statistika dan RHadoop sebagai integrasi antara R dan Hadoop. Terdapat 2 *package* dari RHadoop yang akan digunakan yaitu rmr dan rhdfs. Rhdfs menyediakan konektivitas antara HDFS dengan R, sehingga pengguna dapat melakukan operasi baca, tulis, dan modifikasi data yang tersimpan pada HDFS. Rmr menyediakan konektivitas antara MapReduce dengan R, sehingga pengguna dapat melakukan analisis statistik dalam R melalui MapReduce.

Pada skripsi ini, telah berhasil dibangun beberapa fungsi statistik dalam bentuk R *script* untuk analisis big data menggunakan RHadoop. Pada eksperimen yang telah dilakukan menggunakan *cluster* Hadoop. Eksperimen dilakukan dengan tujuan mengukur kinerja dan skalabilitas dari setiap fungsi yang telah dibangun terhadap data csv dengan variasi ukuran 1GB, 5GB, 10Gb, dan 20GB. Dari eksperimen yang telah dilakukan diketahui bahwa terdapat 1 fungsi yang kinerjanya tidak baik karena tidak menjamin skalabilitas yaitu fungsi kuartil, sedangkan fungsi lainnya memiliki kinerja yang baik sehingga dapat digunakan untuk analisis big data. Fungsi kuartil tidak menjamin skalabilitas karena data harus diurutkan terlebih dahulu sebelum dicari nilai kuartilnya. Pada skripsi ini juga telah berasil dilakukan analisis studi kasus menggunakan fungsi-fungsi yang telah dibangun.

Kata-kata kunci: R, Hadoop, RHadoop, Sistem Tersebar, Statistika

ABSTRACT

The development of the internet makes data growth every year more and more. The data obtained are increasingly unstructured, numerous, and increasingly difficult to process using only traditional processing, so that the term big data appears. Big data can be analyzed to obtain information and knowledge that is useful for drawing conclusions, technology is needed to handle very large amounts of data and statistical computing to draw conclusions. R is one of the languages that can be used to do statistical computing and Apache Hadoop is used to store and distribute computing.

RHadoop is one of the ways to integrate R and Hadoop. RHadoop is an open source project that consists of 5 packages to analyze data with Hadoop through R. In this thesis, an R language study will be carried out for statistical computing and RHadoop as an integration between R and Hadoop. There are 2 packages of RHadoop that will be used, namely rmr and rhdfs. Rhdfs provide connectivity between HDFS and R, so users can perform read, write and modify data stored on HDFS. Rmr provide connectivity between MapReduce and R, so users can do statistical analysis in R through MapReduce.

In this thesis, several statistical functions have been successfully built in the form of R scripts for big data analysis using RHadoop. In experiments that have been carried out using the Hadoop cluster. Experiments carried out with the aim of measuring the performance and scalability of each function that has been built on CSV data with a size variation of 1GB, 5GB, 10Gb, and 20GB. From the experiments that have been carried out, it is known that there is 1 function whose performance is not good because it does not guarantee scalability, which is a quartile function, while other functions have good performance so that it can be used for big data analysis. Quartile functions do not guarantee scalability because the data must be sorted before the quartile value is searched. This thesis also analyzes case studies using functions that have been built.

Keywords: R, Hadoop, RHadoop, Distributed System, Statistic

Dipersembahkan kepada keluarga dan teman.

KATA PENGANTAR

Puji syukur kepada Tuhan atas seluruh berkat yang telah diberikan kepada penulis selama penulisan skripsi sehingga skripsi dengan judul Studi dan Implementasi R pada Sistem Tersebar untuk Analisis Big Data dapat diselesaikan. Penulis juga mengucapkan terima kasih kepada pihak-pihak yang telah membantu penulis dalam proses penulisan dan penyelesaian skripsi, yaitu:

- Keluarga yang selalu memberi dukungan kepada penulis.
- Ibu Dr. Veronica Sri Moertini selaku pembimbing yang telah membimbing penulis selama penulisan skripsi ini dari awal hingga akhir.
- Yuni Asmara Gunawan yang telah membantu dan menyemangati penulis selama proses pembuatan skripsi.
- Matthew Alvredo, Stephen Senjaya, Glenn Reysan, Hengky Surya, Vincent Joel Sinatra, Yosua, Kevin Pratama yang telah membantu memberikan dukungan dan sindiran selama penulisan skripsi.
- William Stefanus, Jeremmy Owen, Aldy Raynaldo, Anthony Fernando yang telah menemani maupun memberikan semangat saat penulisan dokumen skripsi.
- Reza Valentino, Steven Souw, Stephen Lunardi, Renard Junio, Glenn Biondi, Alexander Gunawan, William Akira, Darrel Nolan, dan teman-teman Massa Coffeshop yang telah membuat penulis kerepotan dan menghambat dalam penulisan skripsi.

Bandung, Juni 2020

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE PROGRAM

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Perkembangan internet yang cepat berpengaruh terhadap perkembangan data. Data yang tersimpan semakin banyak dan cepat terkumpul. Dengan bertambahnya jumlah data dan terkumpulnya data dengan cepat istilah big data muncul. Big data adalah kumpulan data yang berukuran sangat besar sehingga sudah tidak bisa ditangani oleh sebuah sistem saja. Mengumpulkan informasi dari menyimpan data yang sangat besar telah menjadi suatu masalah untuk analisis big data, pada saat yang sama cara mempresentasikan informasi yang mudah untuk diakses semakin sulit.

Hadoop merupakan salah satu *framework* solusi untuk masalah big data. Hadoop merupakan *framework* atau *platform open source* berbasis Java yang menggunakan teknologi *Google MapReduce* serta *Google File System (GFS)* sebagai fondasinya [**?**]. Dengan teknologi Hadoop maka data yang jumlahnya besar dapat ditangani dan dianalisis dengan cara mendistribusikan data pada beberapa komputer dan dilakukan komputasi secara terdistribusi sehingga mengurangi waktu komputasi. Hadoop menggunakan beberapa komputer atau biasa disebut *cluster* untuk menyimpan dan memproses data secara terdistribusi. Hadoop memiliki karakteristik *fault tolerance* yang berarti Hadoop menyimpan data pada beberapa komputer sehingga jika suatu komputer rusak, maka suatu data tidak akan hilang karena terdapat salinan data pada komputer lain.

Metode statistika sudah lama dimanfaatkan untuk menganalisis data tetapi belum diterapkan pada lingkungan big data. Statistika digunakan sebagai alat bantu peneliti untuk mengungkapkan dan memecahkan masalah penelitian melalui analisis data [**?**]. Ditinjau menurut variabelnya, analisis statistika dibagi menjadi tiga yaitu analisis univariat, analisis bivariat, dan analisis multivariat. Analisis univariat bertujuan untuk menjelaskan suatu variabel pada penelitian dengan mencari nilai rata-rata, kuartil, standar deviasi, skewness, atau kurtosis. Analisis bivariat merupakan analisis pada 2 variabel, dapat digunakan untuk mencari hubungan antara 2 variabel seperti mencari nilai koefisien korelasi. Analisis multivariat adalah analisis pada 3 atau lebih variabel seperti memprediksi nilai berdasarkan 3 atau 4 variabel bebas dengan membuat model menggunakan fungsi linear regresi berganda.

R merupakan bahasa pemrograman dan perangkat lunak *open source* terkenal untuk analisis data [**?**]. R memiliki banyak *package* yang didukung oleh banyak komunitas dan pengembang. R dapat berjalan hampir di seluruh sistem operasi seperti Windows, Linux, dan MacOs. R merupakan program yang komprehensif karena menawarkan banyak teknik analisis data yang dikembangkan oleh banyak komunitas di seluruh dunia, sehingga pengguna dapat langsung menggunakan fungsi-fungsi yang telah dikembangkan.

R merupakan perangkat lunak yang baik untuk analisis data tetapi memiliki kelemahan yaitu seluruh objek pada R disimpan pada memori fisik. Memori fisik pada suatu komputer sangat terbatas, Hadoop merupakan salah satu solusi untuk menangani kelemahan yang dimiliki R dengan cara mendistribusikan setiap data ke dalam beberapa komputer. RHadoop adalah salah satu cara untuk melakukan integrasi antara R dan Hadoop [**?**]. RHadoop terdiri dari 5 *package* yang sebagi kontektivitas terhadap Hadoop. Komponen utama dari Hadoop adalah MapReduce dan HDFS, pada skripsi ini akan digunakan 2 *package dari RHadoop* yaitu rmr2 dan rhdfs sebagai penghubung

antara MapReduce dan HDFS menggunakan R. rmr2 merupakan penghubung antara R dengan MapReduce. Dengan menggunakan rmr, maka pengguna dapat menulis kode dalam bahasa R dan melakukan eksekusi MapReduce *job* dalam R. Rhdfs merupakan penghubung antara R dan HDFS sehingga pengguna dapat melakukan operasi baca, tulis dan modifikasi data pada HDFS melalui R.

RHadoop merupakan solusi untuk analisis big data. RHadoop menggunakan bahasa R untuk membuat kode program pada MapReduce, sehingga pengguna dapat memanipulasi data menggunakan dataframe, vector, dan struktur lainnya pada R. Hal ini memudahkan dalam membuat program dalam bahasa R. RHadoop merupakan kumpulan *package* yang berjalan diatas Hadoop sehingga pengguna harus mengerti keterbatasan dan kelebihan dari Hadoop itu sendiri. Rhadoop dapat menerima berbagai format data untuk analisis big data, format yang digunakan pada penelitian dan eksperimen memiliki format csv. Pada skripsi ini akan dilakukan analisis big data menggunakan RHadoop untuk dengan cara menerapkan beberapa fungsi statistika menggunakan RHadoop. Pada skripsi ini juga akan dilakukan eksperimen kinerja dari fungsi-fungsi statistika yang telah dibangun menggunakan RHadoop untuk mengetahui kinerja dari RHadoop dalam menangani big data.

1.2 Rumusan Masalah

Masalah yang diangkat dalam skripsi ini adalah:

1. Bagaimana tentang konsep statistika untuk analisis data?
2. Bagaimana menjalankan fungsi-fungsi R untuk menganalisis data dengan pendekatan statistik?
3. Bagaimana integrasi RHadoop pada sistem terdistribusi untuk analisis big data studi kasus?
4. Bagaimana membuat modul-modul program RHadoop untuk menganalisis big data studi kasus?
5. Bagaimana mengukur kinerja fungsi pada RHadoop?

1.3 Tujuan

Tujuan yang ingin dicapai dari penulisan skripsi ini adalah:

1. Mempelajari konsep statistika untuk analisis univariat, bivariat, dan multivariat.
2. Menjalankan fungsi-fungsi R untuk menganalisis data dalam pendekatan statistik.
3. Melakukan integrasi RHadoop pada sistem terdistribusi untuk analisis big data studi kasus.
4. Membangun modul-modul program RHadoop untuk menganalisis big data studi kasus.
5. Mengukur kinerja fungsi pada RHadoop dengan membandingkan waktu eksekusi dengan ukuran data yang bervariasi.

1.4 Batasan Masalah

Batasan-Batasan masalah yang digunakan dalam penelitian ini adalah:

1. Data studi kasus dan eksperimen memiliki format csv.
2. *Package* RHadoop yang digunakan adalah rmr2 dan rhdfs.
3. Fungsi-fungsi statistika yang diimplementasikan menggunakan RHadoop adalah rata-rata, kuartil, count, standar deviasi, kurtosis, skewness, koefisien korelasi, linear regresi, residual standard error, adjusted R-squared.

1.5 Metodologi

Metodologi yang digunakan dalam skripsi ini adalah:

1. Mempelajari konsep statistik untuk analisis univariat, bivariat, dan multivariat.
2. Menginstal dan mempelajari konsep R, melakukan eksperimen awal untuk mencoba fitur-fitur pada R.
3. Mempelajari fitur-fitur pada RHadoop untuk analisis big data.
4. Menginstal dan mempelajari sistem tersebar Hadoop.
5. Melakukan eksplorasi terhadap RHadoop dengan data berukuran kecil.
6. Merancang modul program pada RHadoop untuk analisis big data.
7. merancang eksperimen pada *cluster* Hadoop untuk big data studi kasus.
8. Menguji kinerja modul program yang telah dibuat dari eksperimen.
9. Menulis dokumen skripsi.

1.6 Sistematika Pembahasan

Sistematika pembahasan dari skripsi ini adalah sebagai berikut:

1. Bab 1 Pendahuluan
   Bab 1 berisi tentang latar belakang, rumusan masalah, batasan masalah, dan sistematika pembahasan pada skripsi ini.
2. Bab 2 Landasan Teori
   Bab 2 membahas tentang teori-teori mengenai konsep dan cara kerja RHadoop, statistika dasar, dan sintaks dasar penulisan dalam bahasa R.
3. Bab 3 Eksplorasi R dan RHadoop
   Bab 3 membahas tentang eksperimen-eksperimen dimulai dari langkah-langkah instalasi hadoop, RHadoop, eksperimen dengan RHadoop, dan eksperimen fungsi-fungsi pada RHadoop.
4. Bab 4 Perancangan dan implementasi
   Bab 4 membahas perancangan dan implementasi fungsi-fungsi yang telah dibangun menggunakan RHadoop.
5. Bab 5 Eksperimen dan Analisis RHadoop
   Bab 5 membahas eksperimen dan analisis dari fungsi-fungsi yang telah dibangun dengan menggunakan data eksperimen dan melakukan eksperimen menggunakan data studi kasus.
6. Bab 6 Kesimpulan dan Saran
   Bab 6 membahas mengenai kesimpulan-kesimpulan dan saran-saran dari penulis yang berhasil didapatkan dari penelitian ini.