ANALISIS KINERJA WORDCOUNT MAPREDUCE PADA *MULTI-NODE HADOOP CLUSTER*

SKRIPSI

Disusun Sebagai Salah Satu Syarat

Untuk Memperoleh Gelar Sarjana Komputer

Pada Departemen Ilmu Komputer/ Informatika

Disusun oleh:

ANDRE MANGARA

24010313130120

DEPARTEMEN ILMU KOMPUTER/ INFORMATIKA

FAKULTAS SAINS DAN MATEMATIKA

UNIVERSITAS DIPONEGORO

2019

PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan dibawah ini:

Nama : Andre Mangara

NIM : 24010313130120

Judul : Analisis Kinerja Wordcount Mapreduce pada *Multi-node Hadoop Cluster*

Dengan ini saya menyatakan bahwa dalam tugas akhir/ skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka.

Semarang, 15 Agustus 2019

METERAI TEMPEL

F93CEAFF911184965

6000

ENAM RIBU RUPIAH

Andre Mangara

24010313130120

HALAMAN PENGESAHAN

Judul : Analisis Kinerja Wordcount Mapreduce pada *Multi-node Hadoop Cluster*

Nama : Andre Mangara

NIM : 24010313130120

Telah diujikan pada sidang tugas akhir pada tanggal 24 Juli 2019 dan dinyatakan lulus pada tanggal 24 Juli 2019.

Semarang, 15 Agustus 2019

Mengetahui,

Ketua Departemen Ilmu Komputer/ Informatika FSM UNDIP

Dr. Retno Kusumaningrum, S.Si, M.Kom
NIP. 198104202005012001

Panitia Penguji Tugas Akhir

Ketua,

Satriyo Adhy, S.Si., M.T
NIP. 198302032006041002

HALAMAN PENGESAHAN

Judul : Implementasi Komputasi Terdistribusi pada *Multi-Node Hadoop Cluster*

Nama : Andre Mangara

NIM : 24010313130120

Telah diujikan pada siding tugas akhir pada tanggal 24 Juli 2019.

Semarang, 15 Agustus 2019

Pembimbing

Panji Wisnu Wirawan, ST, MT

NIP.198104212008121002

ABSTRAK

Perkembangan teknologi yang pesat menyebabkan pertumbuhan data meningkat secara eksponensial. Data yang berkembang sangat pesat tersebut akhirnya menjadi tantangan bagi banyak perusahaan teknologi seperti Google, Yahoo, IBM, dan Microsoft. Yahoo akhirnya menciptakan sebuah sistem untuk mengolah data berdasarkan MapReduce yang dinamakan Hadoop. Karateristik utama dari Hadoop adalah cara kerjanya yaitu mempartisi data dan melakukan komputasi diberbagai komputer *host* yang tergabung didalam *cluster* sehingga membuat pengolahan data menjadi lebih cepat dibandingkan dengan distribusi *file* secara konvensional. Hadoop dapat dibangun pada perangkat komoditas standar yang umum dipasaran dan dapat berasal dari berbagai vendor. Penelitian ini melakukan pengujian terhadap salah satu keunggulan Hadoop yaitu dapat diimplementasikan pada perangkat komoditas di laboratorium Informatika Universitas Diponegoro. MapReduce Wordcount digunakan sebagai program benchmarking untuk menganalisa kinerja dari cluster Hadoop yang dibuat pada penelitian. Tiga skenario pengujian dilakukan pada *single-node cluster* dan *multi-node cluster* Hadoop yang dibuat untuk melihat performa dari MapReduce Hadoop dalam mengolah sebuah *file*. Pengujian skenario pertama dilakukan dengan menggunakan empat data uji berupa file teks berukuran 64 MB, 256 MB, 512 MB, 1 GB, 2 GB dan 4 GB. Hasil yang didapatkan adalah Hadoop lebih cepat dalam mengolah data yang berukuran lebih besar dari *block size.* Pada skenario kedua dilakukan pengujian mengenai pengaruh jumlah node dalam performa eksekusi Wordcount. Data yang digunakan adalah *file* xml berukuran 50 MB, 256 MB, 512 MB dan 945 MB. Hasil yang didapatkan dari pengujian adalah terjadinya peningkatan performa dengan penambahan *node* pada data berukuran 945 MB. Pengujian skenario ketiga membandingkan pengaruh ukuran *block size* 128 MB dan 256 MB dan ukuran *file* uji terhadap performa MapReduce Hadoop. Pengujian menghasilkan ukuran *block size* mempunyai pengaruh yang berbeda terhadap setiap ukuran *file* uji. Semakin besar *file* yang diujikan, *block size* dengan ukuran yang lebih besar menghasilkan performa yang lebih baik.

Kata Kunci: Hadoop, MapReduce, *single-node cluster, multi-node cluster, block size.*

ABSTRACT

Rapid technological development causes data growth to increase exponentially. Data that is growing very rapidly has finally become a challenge for many technology companies such as Google, Yahoo, IBM, and Microsoft. Yahoo finally created a system for processing data based on MapReduce called Hadoop. The main characteristic of Hadoop is the way it works, namely partitioning data and doing computation in various host computers incorporated in the cluster, making processing data faster than conventional file distribution. Hadoop can be built on standard commodity hardware that are common in the market and can come from various vendors. This research tests one of the advantages of Hadoop, which can be implemented on commodity hardware that is located in Informatics Laboratory of Diponegoro University. MapReduce Word count is used as a benchmarking program to analyze the performance of the Hadoop cluster made in the study. Three test scenarios were carried out on a single-node cluster and a Hadoop multi-node cluster that was created to see the performance of MapReduce Hadoop in processing a file. Testing the first scenario is done by using four test data in the form of text files measuring 64 MB, 256 MB, 512 MB, 1 GB, 2 GB and 4 GB. Results obtained are Hadoop faster in processing data that is larger than the block size. In the second scenario a test is carried out on the effect of the number of nodes in the execution of Wordcount performance. The data used is an xml file measuring 50MB, 256MB, 512MB and 945MB. The results obtained from the test are an increase in performance with the addition of nodes on data measuring 945 MB. The third scenario testing compares the effect of block size size 128 MB and 256 MB and the size of the test file on MapReduce Hadoop performance. Testing results in a block size that has a different effect on each size of the test file. Larger file size required larger block size to have a better performance result.

Keywords: Hadoop, MapReduce, *single-node cluster, multi-node cluster, block size.*

KATA PENGANTAR

Segala puji syukur bagi Tuhan Yang Maha Esa atas karunia-Nya yang diberikan kepada penulis sehingga penulis dapat menyelesaikan penulisan laporan tugas akhir yang berjudul "Analisis Kinerja Wordcount MapReduce pada Multi-node Hadoop Cluster". Laporan tugas akhir ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana strata satu pada Departemen Ilmu Komputer/ Informatika Fakultas Sains dan Matematika Universitas Diponegoro Semarang.

Dalam menyusun laporan ini penulis mendapat bimbingan dan bantuan dari berbagai pihak. Untuk itu, pada kesempatan ini penulis mengucapkan rasa hormat dan terima kasih kepada:

1. Prof. Dr. Widowati, S.Si., M.Si., selaku Dekan FSM UNDIP.
2. Dr. Retno Kusumaningrum, S.Si, M.Kom., selaku Ketua Departemen Ilmu Komputer/ Informatika.
3. Panji Wisnu Wirawan, ST, MT., selaku Koordinator Tugas Akhir dan Dosen Pembimbing.
4. Semua pihak yang telah membantu kelancaran dalam penyusunan tugas akhir, yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa dalam laporan ini masih banyak kekurangan baik dari segi materi maupun dalam penyajiannya karena keterbatasan kemampuan dan pengetahuan penulis. Oleh Karena itu, kritik dan saran sangat penulis harapkan. Semoga laporan ini dapat bermanfaat bagi pembaca dan penulis pada umumnya.

Semarang, 15 Agustus 2019

Andre Mangara

NIM. 24010313130120

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademik Universitas Diponegoro, saya yang bertanda tangan di bawah ini:

Nama	: Andre Mangara
NIM	: 24010313130120
Program Studi	: Informatika
Departemen	: Ilmu Komputer/ Informatika
Fakultas	: Sains dan Matematika
Jenis Karya	: Tugas Akhir

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan **Hak Bebas Royalti Non-eksklusif** (***Non-exclusive Royalty Free Right***) kepada Universitas Diponegoro atas karya ilmiah saya yang berjudul:

Analisis Kinerja Wordcount MapReduce pada Multi-node Hadoop Cluster

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-eksklusif ini, Universitas Diponegoro berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis/ pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Semarang 15 Agustus 2019

Yang menyatakan

Andre Mangara

24010313130120

DAFTAR ISI

PERNYATAAN KEASLIAN SKRIPSI ii
HALAMAN PENGESAHAN iii
HALAMAN PENGESAHAN iv
ABSTRAK v
ABSTRACT vi
KATA PENGANTAR vii
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS viii
DAFTAR ISI ix
DAFTAR GAMBAR xi
DAFTAR TABEL xii
BAB I PENDAHULUAN 1
BAB II TINJAUAN PUSTAKA 4
2.1. Penelitian Terkait 4
2.2. Hadoop 6
2.2.1. Hadoop Distributed File System 6
2.2.2. MapReduce 9
2.2.3. YARN 10
BAB III METODOLOGI PENELITIAN 13
3.1. Kebutuhan Implementasi Komputasi Terdistribusi 13
3.1.1. Perangkat Keras 13
3.1.2. Perangkat Lunak 13
3.1.3. Rancangan Sistem 14
3.1.4. Wordcount MapReduce 15
3.1.5. Wordcount tanpa Hadoop MapReduce 22
3.2. Pembuatan Antarmuka 23
3.3. Skenario Pengujian 24
3.3.1. Skenario Pengujian Pertama 24
3.3.2. Skenario Pengujian Kedua 25
3.3.3. Skenario Pengujian Ketiga 26

3.4. Garis Besar Penyelesaian Masalah **26**
BAB IV HASIL DAN PEMBAHASAN **29**
4.1. Hasil Skenario Pengujian Pertama **29**
4.1.1. Analisis **29**
4.2. Hasil Skenario Pengujian Kedua **30**
4.2.1. Analisis **30**
4.3. Hasil Skenario Pengujian Kedua **33**
4.3.1. Analisis **33**
4.4. Hasil Pembuatan Antarmuka **34**
BAB V PENUTUP **37**
5.1. Kesimpulan **37**
5.2. Saran **37**
DAFTAR PUSTAKA **38**

DAFTAR GAMBAR

Gambar 2.1 Arsitektur HDFS 8
Gambar 2.2 Arsitektur MapReduce 10
Gambar 2.3 Arsitektur YARN 11
Gambar 2.4 Alur Menjalankan Aplikasi pada YARN 11
Gambar 3.1 Arsitektur Single-node Cluster 14
Gambar 3.2 Arsitektur Multi-node Cluster 15
Gambar 3.3 Contoh penyebaran block pada cluster Hadoop (Apache Hadoop, 2018) 15
Gambar 3.4 Gambar input splits pada proses MapReduce 16
Gambar 3.5 Pseudo-code Wordcount dengan Hadoop 16
Gambar 3.6 Kode Program Wordcount Teks menggunakan Hadoop 17
Gambar 3.7 Pseudo-code Wordcount parsing XML dengan Hadoop 19
Gambar 3.8 Penerapan Map Wordcount parsing XML dengan Hadoop 20
Gambar 3.9 Penerapan Reduce Wordcount Mapreduce parsing XML dengan Hadoop 21
Gambar 3.10 Alur Proses Wordcount MapReduce 21
Gambar 3.11 Pseudo-code Wordcount tanpa Hadoop 22
Gambar 3.12 Kode program Wordcount tanpa Hadoop MapReduce 23
Gambar 3.13 Pseudo-code MapReduce PHP 24
Gambar 3.14 Gambar Flowchart Skenario Pertama 25
Gambar 3.15 Flowchart Skenario Pengujian Kedua 27
Gambar 3.16 Flowchart Skenario Pengujian Ketiga 28
Gambar 4.1 Grafik Perbandingan antara Wordcount Mengunakan dan Tanpa Hadoop 30
Gambar 4.2 Grafik Hasil Pengujian Skenario Kedua 31
Gambar 4.3 Hasil Wordcount MapReduce Xml 32
Gambar 4.4 Contoh kegagalan task Hadoop 32
Gambar 4.5 Grafik Pengujian block size 128 MB 34
Gambar 4.6 Grafik Pengujian block size 256 MB 35
Gambar 4.7 Hasil Wordcount MapReduce Text 35
Gambar 4.8 Tampilan Antarmuka 1 36
Gambar 4.9 Tampilan Antarmuka 2 36

DAFTAR TABEL

Tabel 2.1 Tabel Hasil Perbandingan kinerja Wordcount pada Multi-Node Cluster 4
Tabel 2.2 Tabel Hasil Pengujian TestDFSIO pada Multi-node Cluster 5
Tabel 2.3 Tabel Hasil performa kecepatan MapReduce terhadap block size..... 6
Tabel 3.1 Spesifikasi Perangkat Keras 13
Tabel 4.1 Hasil Pengujian Wordcount pada single-node cluster dan multi-node cluster.... 30
Tabel 4.2 Hasil Pengujian Kedua dengan block size berukuran 256 MB 33
Tabel 4.3 Hasil Pengujian Kedua dengan block size berukuran 128 MB 33

BAB I
PENDAHULUAN

Bab ini membahas latar belakang, rumusan masalah, tujuan dan manfaat, dan ruang lingkup penelitian tugas akhir mengenai Analisis Kinerja *Word count* MapReduce pada *Hadoop Multi-node Cluster*.

1.1. Latar Belakang

Perkembangan teknologi yang pesat menyebabkan pertumbuhan data meningkat secara eksponensial. Data yang dihasilkan tersebut jika diolah dapat menghasilkan sebuah informasi yang nantinya dapat digunakan lebih lanjut sesuai kebutuhan. Data yang berkembang sangat pesat tersebut akhirnya menjadi tantangan bagi banyak perusahaan teknologi seperti Google, Yahoo, IBM, dan Microsoft. Google menjadi yang pertama untuk mengatasi masalah pengolahan data ini dengan menciptakan MapReduce. Yahoo juga akhirnya menciptakan sebuah sistem untuk mengolah data berdasarkan MapReduce yang dinamakan Hadoop.

Hadoop menyediakan sistem *file* terdistribusi dan menyediakan sebuah framework yang bernama MapReduce untuk mengolah dan menganalisis data set yang besar. Karateristik utama dari Hadoop adalah cara kerjanya yaitu mempartisi data dan melakukan komputasi diberbagai komputer *host* yang tergabung didalam *cluster* sehingga membuat pengolahan data menjadi lebih cepat dibandingkan dengan distribusi *file* secara konvensional, seperti RDBMS. Jenis *cluster* yang dapat dibangun pada Hadoop secara umum ada dua, yaitu *single-node cluster* dan *multi-node cluster.* (White, 2015)

Hadoop didesain agar dapat berjalan pada perangkat komoditas, yang berarti *cluster* Hadoop tidak harus dibangun pada perangkat yang mahal dan hanya pada satu vendor saja. Hadoop dapat dibangun pada perangkat standar yang umum dipasaran dan dapat berasal dari berbagai vendor. Solusi untuk menghadapi data dengan skala yang besar tidak lagi hanya dengan membangun server yang semakin besar, tetapi dengan menggabungkan *low-end/* perangkat komoditas sebagai sebuah sistem terdistribusi yang fungsional seperti Hadoop. (Lam, 2011)

Penelitian ini melakukan pengujian terhadap salah satu keunggulan Hadoop yaitu dapat diimplementasikan pada perangkat komoditas di laboratorium Informatika

Universitas Diponegoro. MapReduce *Word count* digunakan sebagai program *benchmarking* untuk menganalisa kinerja dari *cluster* Hadoop yang dibuat pada penelitian. MapReduce *Word count* dipilih sebagai sarana benchmarking karena program tersebut adalah salah satu program dasar yang terdapat pada Hadoop dan mudah untuk diimplementasikan serta diubah.

1.2. Rumusan Masalah

Dari latar belakang yang ada, dapat dirumuskan bahwa permasalahan adalah bagaimana menganalisis kinerja MapReduce *Word count* pada *single-node cluster* dan *multi-node cluster* Hadoop dalam mengolah data pada perangkat komoditas serta membangun sebuah antarmuka untuk menggunakan Hadoop.

1.3. Tujuan dan Manfaat

Tujuan dari tugas akhir yang ingin dicapai adalah menganalisis kinerja Wordcount MapReduce pada *cluster* Hadoop dalam mengolah data pada perangkat komoditas dan membangun antarmuka untuk pengguna. Manfaat yang diharapkan adalah mengoptimalkan waktu pengolahan data dan terciptanya sebuah antarmuka untuk mengakses Hadoop.

1.4. Ruang Lingkup

Ruang lingkup dalam pembuatan tugas akhir Analisis Kinerja Wordcount MapReduce pada Hadoop Cluster adalah sebagai berikut:

a. Uji coba dilakukan pada 6 komputer (1 NameNode dan 5 DataNode)
b. MapReduce yang digunakan adalah Wordcount dengan *dataset* berupa file xml berukuran 50 MB, 256 MB, 512 MB dan 945 MB dan file text berukuran 64 MB, 256, 512 MB, 1 GB, 2 GB dan 4 GB.
c. Sistem Hadoop *multi-node cluster* diimplementasikan pada Laboratorium Komputer C, Fakultas Sains dan Matematika, Universitas Diponegoro

1.5. Sistematika Penulisan

Sistematika penulisan laporan tugas akhir adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini menjelaskan mengenai teori-teori yang akan digunakan pada tugas akhir.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tentang metode pengumpulan data, kebutuhan *cluster*, langkah pengujian dan garis besar penyelesaian masalah.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil penelitian dan pembahasan hasil pengujian yang sudah dilakukan.

BAB V KESIMPULAN

Bab ini berisi kesimpulan dan saran dari tugas akhir.