

SKRIPSI

STUDI DAN IMPLEMENTASI DESAIN POLA MAPREDUCE 1



CALVIN SATRIA TORSIA SARAGIH

NPM: 2014730076

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2018**

UNDERGRADUATE THESIS

**STUDY AND IMPLEMENTATION MAPREDUCE DESIGN
PATTERN 1**



CALVIN Satria Torsa Saragih

NPM: 2014730076

**DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2018**

LEMBAR PENGESAHAN



STUDI DAN IMPLEMENTASI DESAIN POLA MAPREDUCE 1

CALVIN SATRIA TORSIA SARAGIH

NPM: 2014730076

Bandung, 25 Mei 2018

Menyetujui,

Pembimbing

Dr. Veronica Sri Moertini

Ketua Tim Penguji

Claudio Franciscus, M.T.

Anggota Tim Penguji

Elisati Hulu, 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 DESAIN POLA MAPREDUCE 1

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 25 Mei 2018



CALVIN SATRIA TORSA SARAGIH
NPM: 2014730076

ABSTRAK

Perkembangan internet membuat perkembangan kebutuhan informasi yang semakin kompleks juga. Data yang diolah semakin beragam, besar dan lama untuk diakses bila menggunakan basis data tradisional sehingga diperlukan teknologi baru untuk mengatasi masalah ini. Salah satu solusi yang dapat digunakan untuk mengatasi permasalahan tersebut adalah dengan menerapkan sistem terdistribusi. Apache Hadoop merupakan salah satu sistem terdistribusi. Hadoop dikembangkan karena penyimpanan dan pengolahan data yang tradisional tidak dapat menangani perkembangan data saat ini.

Perangkat lunak yang dibangun pada sistem terdistribusi harus dapat memiliki kemampuan untuk melakukan analisis atau pengolahan data. Model pemrograman MapReduce dapat digunakan dalam sistem terdistribusi Hadoop untuk melakukan analisis data, dimana map berperan sebagai memetakan input data dan reduce melakukan operasi summarisasi dan agregasi.

Desain pola MapReduce berperan sebagai kerangka saat pembuatan pekerjaan MapReduce. Kerangka ini membantu dalam pengimplementasian perangkat lunak analisis data dalam Hadoop. Terdapat beberapa pola yang ada yaitu, pola summarisasi, organisasi data dan Metapattern. Pola summarisasi berguna untuk operasi statistik, pola organisasi data berguna untuk mengubah struktur data dan pola metapattern berguna untuk optimisasi desain pola MapReduce.

Pada penelitian ini dilakukan implementasi desain pola MapReduce untuk analisis data yang bergantung terhadap tipe data, masalah yang dihadapi dan hasil akhir yang ingin dicapai. Pola satu dengan yang lain memecahkan masalah yang berbeda-beda, sehingga waktu eksekusi program MapReduce dapat berbeda-beda. Semakin banyak objek data yang digunakan berdampak pada waktu komputasi dalam mengeksekusi program akan semakin besar.

Kata-kata kunci: Data, Apache Hadoop, MapReduce, data besar, summarisasi, organisasi data, metapattern

ABSTRACT

The development of the Internet makes the development of information needs increasingly complex as well. The data processed more diverse, large and long to be accessed when using traditional databases, so new technology is needed to solve this problem. One solution that can be used to overcome these problems is to implement a distributed system. Apache Hadoop is one of the distributed systems. Hadoop is developed because traditional data storage and processing can not handle the current data development.

Distributed systems must have the ability to perform data analysis or processing. MapReduce programming models can be used in Hadoop distributed systems to perform data analysis, where map task serve as mapping of data inputs and reduce task performing sumarization and aggregation operations.

MapReduce pattern design acts as a template for making the MapReduce job. This template helps in implementing data analysis software in Hadoop. There are several existing patterns, namely: sumarization patterns, data organization and Metapattern. sumarization pattern is useful for statistical operations, data organization pattern useful for transforming data structures and metapattern patterns useful for optimizing the MapReduce pattern design.

In this research, MapReduce pattern design implementation for data analysis depends on data type, problem faced, final result to be achieved. One Patterns and each other solve different problems, so that the execution time of the MapReduce program may vary. The more data objects used impact the computation time in executing the program will be greater.

Keywords: Data, Apache Hadoop, MapReduce, big data, summarization, data organization, metapattern

Untuk orang tua dan orang-orang terdekat

KATA PENGANTAR

Puji dan syukur kepada Tuhan Yang Maha Esa atas berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi ini sesuai dengan waktu yang diharapkan. Penulis menyadari bahwa di dalam skripsi ini masih terdapat kekurangan yang disebabkan oleh keterbatasan dari segi kemampuan dan pengetahuan yang dimiliki penulis.

Pada Kesempatan ini, penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada:

1. Orang tua penulis yaitu Saur Saragih dan Triwanti Tiolida yang menyediakan dukungan secara materi dan psikologis.
2. Ibu Veronica Sri Moertini sebagai dosen pembimbing yang memberikan arahan agar skripsi ini dapat selesai berupa nasihat, kritik dan saran.
3. Sahabat-sahabat seperjuangan saya (Ceem,Obrien,Zenso,Irfan,Aldo) yang juga menghadapi pembuatan skripsi secara bersama-sama.
4. Para admin lab. komputer dan teman-teman seperjuangan dengan topik yang sama, dimana membantu saya dalam penggunaan lab. skripsi.
5. Rekan-rekan Teknik Informatika 2014 yang tidak dapat disebutkan satu-persatu.
6. Teman-teman sepermainan yang membantu dan mendukung baik secara langsung maupun tidak langsung.

Akhir kata penulis berterima kasih kepada pihak-pihak yang membantu tetapi tidak dapat disebutkan satu-persatu. Semoga skripsi ini dapat berguna bagi diri saya dan orang lain yang membutuhkan pengetahuan mengenai skripsi ini.

Bandung, Mei 2018

Penulis

DAFTAR ISI

KATA PENGANTAR	xv
DAFTAR ISI	xvii
DAFTAR GAMBAR	xxi
DAFTAR TABEL	xxiii
1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
1.4 Batasan Masalah	2
1.5 Metodologi	3
1.6 Sistematika Pembahasan	3
2 LANDASAN TEORI	5
2.1 Big Data	5
2.2 Apache Hadoop	7
2.2.1 HDFS (Hadoop Distributed File System)	8
2.2.2 MapReduce	10
2.3 Pola Desain MapReduce	12
2.3.1 Pola Sumarisasi	13
2.3.1.1 Numerical Summarization	13
2.3.1.2 Inverted Index Summarization	15
2.3.1.3 Counting With Counter	15
2.3.2 Data Organization Pattern	16
2.3.2.1 Struktural menjadi Hirarkial	16
2.3.2.2 Binning	18
2.3.2.3 Partitioning	19
2.3.2.4 Total Order Sorting	20
2.3.2.5 Shuffling	20
2.3.3 MettaPattern	21
2.3.3.1 Job Chaining	21
2.3.3.2 Chain Mapper	21
2.3.3.3 Job Merging	22
3 STUDI EKSPLORASI DAN EKSPERIMEN HADOOP	25
3.1 Ekplorasi Hadoop	25
3.2 Analisis Data Studi Kasus Uji	26
3.2.1 Set Data Comment	26
3.2.2 Set Data Posts	27
3.2.3 Set Data User	28

3.3	Eksperimen Desain Pola MapReduce	29
3.3.1	Numerical Summarization	29
3.3.2	Inverted Index Summarization	32
3.3.3	Counting With Counter	32
3.3.4	Struktural menjadi Hirarkial	33
3.3.5	Binning	34
3.3.6	Partitioning	34
3.3.7	Total Order Sorting	35
3.3.8	Shuffling	35
3.3.9	Job Chaining	36
3.3.10	Chain Folding	37
3.3.11	Job Merging	38
4	PEMANFAATAN POLA MAPREDUCE UNTUK ANALISIS DATA	39
4.1	Analisis Data Uji Kasus	39
4.1.1	Set Data Yelp Checkin	40
4.1.2	Set Data Yelp Review	40
4.1.3	Set Data Yelp Bisnis	41
4.1.4	Set Data Yelp User	42
4.2	Analisis Desain Pola dalam Memecahkan Masalah	42
4.2.1	Numerical Summarization	43
4.2.2	Counting with Counter	51
4.2.3	Struktural menjadi Hirarkial	52
4.2.4	Binning	53
4.2.5	Partitioning	55
4.2.6	Total Order Sorting	56
4.2.7	Shuffling	58
4.2.8	Job Chaining	60
4.2.9	Job Merging	63
5	ANALISIS, PERANCANGAN DAN IMPLEMENTASI PERANGKAT LUNAK DEMO	65
5.1	Analisis Perangkat lunak	65
5.1.1	Diagram Use Case dan Skenario	65
5.1.1.1	Diagram Use Case	65
5.1.1.2	Skenario	66
5.2	Perancangan Perangkat lunak	67
5.2.1	Kelas Diagram	67
5.2.1.1	Kelas Diagram Modul Pola Sumarirasi	67
5.2.1.2	Kelas Diagram Modul Pola Organisasi Data	70
5.2.1.3	Kelas Diagram Modul Pola MetaPattern	72
5.2.1.4	Kelas Diagram Keseluruhan	73
5.2.2	Perancangan Antarmuka Perangkat Lunak	74
5.2.2.1	Halaman Awal	74
5.2.2.2	Halaman Modul Pola Sumarisasi	74
5.2.2.3	Halaman Modul Pola Organisasi Data	75
5.2.2.4	Halaman Modul <i>MetaPattern</i>	75
5.3	Implementasi Perangkat Lunak	76
5.3.1	Deskripsi Perangkat Keras dan Perangkat Lunak yang Digunakan	76
5.3.2	Implementasi Antarmuka	77
5.3.2.1	Antarmuka Modul Desain Pola Sumarisasi	77
5.3.2.2	Antarmuka Modul Desain Pola Organisasi Data	77
5.3.2.3	Antarmuka Modul Desain Pola MetaPattern	78

5.3.3	Implementasi Package,Kelas dan Method	79
5.4	Pengujian Perangkat Lunak	79
5.5	Eksperimen untuk Uji Perfoma Desain Pola MapReduce	81
5.5.1	Deskripsi Perangkat Keras	81
5.5.2	Uji Pengaruh Ukuran Data Terhadap Desain Pola MapReduce	81
5.5.2.1	Eksperimen Desain Pola Sumarisasi Numerik	82
5.5.2.2	Eksperimen Desain Pola Sumarisasi <i>Inverted Indeks</i>	83
5.5.2.3	Eksperimen Desain Pola Struktural menjadi Hirarkial	83
5.5.2.4	Eksperimen Desain Pola <i>Binning</i>	83
5.5.2.5	Eksperimen Desain Pola <i>Partitioning</i>	83
5.5.2.6	Eksperimen Desain Pola <i>Shuffling</i>	84
5.5.2.7	Eksperimen Desain Pola <i>Total Order Sorting</i>	84
5.5.2.8	Eksperimen Desain Pola Sumarisasi <i>Job Chaining</i>	84
5.5.2.9	Eksperimen Desain Pola Sumarisasi <i>Chain Mapper</i>	84
5.5.2.10	Eksperimen Desain Pola Sumarisasi <i>Job Merging</i>	85
5.5.3	Kesimpulan	85
6	KESIMPULAN DAN SARAN	87
6.1	Kesimpulan	87
6.2	Saran Penelitian Lanjutan	87
	DAFTAR REFERENSI	89
	A KODE PROGRAM ANALISIS DATASET STACKOVERFLOW	91
	B KODE PROGRAM ANALISIS DATASET WEBSITE YELP	113
	C KODE PROGRAM PERANGKAT LUNAK DEMO	125

DAFTAR GAMBAR

5figure.2.1

2.2	Contoh data terstruktur	7
2.3	Arsitektur hadoop [1]	8
2.4	Arsitektur HDFS [2]	9
2.5	Skema Hadoop	10
2.6	Skema Mapper dalam Hadoop Cluster	11
2.7	Skema Reducer dalam Hadoop Cluster	12
2.8	Struktur dari pola Numerical [3]	13
2.9	Struktur dari pola Inverted Index [3]	15
2.10	Struktur dari pola Counting with Counter [3]	16
2.11	Struktur Pola Struktural menjadi Hirarkial [3]	17
2.12	Struktur Pola Binning [3]	18
2.13	Struktur Pola Partitioning [3]	19
2.14	Skema Chain Mapper dalam memecahkan masalah [3]	22
2.15	Skema Job Merging dalam memecahkan masalah [3]	23
3.1	Antarmuka HDFS	26
3.2	Skema Pola Job Chaining dalam Memecahkan Masalah	36
4.1	Website Yelp	39
5.1	Use Case Perangkat Lunak	66
5.2	Diagram Kelas Pola Sumarisasi	68
5.3	Diagram Kelas Pola Organisasi Data	71
5.4	Diagram Kelas Pola MetaPattern	72
5.5	Diagram Kelas Antarmuka dan Package modul-modul	73
5.6	Skema Job Merging dalam memecahkan masalah	74
5.7	Rancangan Antarmuka Modul Pola Sumarisasi	75
5.8	Rancangan Antarmuka Modul Pola Organisasi Data	75
5.9	Rancangan Antarmuka Modul <i>MetaPattern</i>	76
5.10	Antarmuka Desain Pola Sumarisasi	77
5.11	Antarmuka Desain Pola Organisasi Data	78
5.12	Antarmuka Desain Pola MetaPattern	79
5.13	Informasi Waktu Eksekusi	81

DAFTAR TABEL

5.1	Hasil Pengujian Perangkat Lunak Demo	80
5.2	Uji Desain Pola Sumarisasi Numerik (kasus Min, Maks dan Perhitungan	82
5.3	Uji Eksperimen Desain Pola Sumarisasi Numerik (kasus pencarian rata-rata	82
5.4	Eksperimen Desain Pola Sumarisasi Numerik (kasus Pencarian Standar Deviasi	82
5.5	Eksperimen Desain Pola Sumarisasi Numerik (Median	83
5.6	Uji Desain Pola Inverted Index	83
5.7	Hasil Eksperimen Pengujian Waktu Eksekusi Desain Pola struktural menjadi hirarkial	83
5.8	Hasil Eksperimen Pengujian Waktu Eksekusi Desain Pola <i>Binning</i>	83
5.9	Eksperimen Desain Pola Sumarisasi Numerik (Median	84
5.10	Eksperimen Desain Pola (Shuffling	84
5.11	Eksperimen Desain Pola <i>Total Order Sorting</i>	84
5.12	Eksperimen Desain Pola <i>Job Chaining</i>	84
5.13	Uji Desain Pola <i>Chain Mapper</i>	85
5.14	Uji Desain Pola <i>Job Merging</i>	85

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Data berasal dari segala macam aktivitas yang terkait dengan teknologi, seperti aktivitas dalam internet, data transaksi pada supermarket, data cuaca tiap harinya, data naik turun dalam jual-beli saham, dan lainnya. Data bermanfaat sebagai bahan yang akan diolah untuk dijadikan suatu informasi yang berguna. Saat ini data mengalami perkembangan dalam segi ukuran dan variasi yang menyebabkan perbedaan pada data biasa menjadi data besar (*Big Data*).

Big Data merupakan istilah umum untuk sekumpulan data yang sulit diproses menggunakan aplikasi basis data tradisional [4]. Kesulitan yang menjadi masalah dalam pemrosesan *Big Data* adalah : (1) Ukuran data yang sangat besar berpengaruh pada kebutuhan tempat penyimpanan, (2) Variasi tipe data yang bermacam-macam dalam satu himpunan berpengaruh pada kemampuan aplikasi analisis data, (3) Kecepatan pemrosesan data atau optimal berpengaruh pada perangkat lunak atau keras yang dibutuhkan [4]. *Big Data* seperti data umumnya dapat digunakan untuk memperoleh informasi yang berguna. Perusahaan yang terkait dengan *Big Data* dihadapkan dengan pertumbuhan *Big Data* tiap harinya yang harus mereka analisis, tentunya tidak dapat menggunakan teknik komputasi sederhana dengan satu komputer saja, karena waktu yang terlalu lama dan komputer yang dimiliki harus sangat canggih.

Apache Hadoop merupakan platform yang berguna untuk menangani masalah yang terkait dengan *Big Data*. Hadoop dapat berjalan diatas beberapa *cluster* komputer yang saling berinteraksi dalam melakukan analisis data secara paralel. Hadoop memiliki dua komponen utama yaitu, HDFS (*Hadoop Distributed File System*) dan Mapreduce. HDFS adalah sebuah sistem file yang terdistribusi dan dapat beroperasi dalam sekumpulan hardware. Kelebihan HDFS adalah dapat menyimpan data yang berukuran besar dengan memecahkannya menjadi ukuran kecil dan menyimpan dalam beberapa tempat penyimpanan.

MapReduce adalah sebuah model pemrograman yang berguna untuk memecahkan analisis *Big Data*. MapReduce dapat berjalan diatas platform hadoop dimana bekerja secara paralel dan terdistribusi. Fungsi MapReduce adalah menerima *record-record* data kemudian memetakan data agar dapat disaring dan dilakukan agregasi data [1]. Desain pola MapReduce merupakan kerangka dalam memecahkan analisis *Big Data*. Desain ini bermanfaat sebagai panduan dalam memecahkan analisis data dimana kasus yang dihadapi sangat beragam [3]. Kasus yang dihadapi dapat berupa perhitungan statistik, penyaringan data, pengubahan struktur data dan lainnya. Pola MapReduce terdiri dari bermacam-macam bergantung kepada tujuan pengolahan data yang kita punya. Saat ini telah banyak dikembangkan pola-pola MapReduce, Beberapa diantaranya yaitu:

- Pola Sumarisasi. Pola yang berfokus kepada mengelompokkan data bersama-sama untuk kemudian dilakukan operasi seperti menghitung statistik, menghitung dan membuat indeks [3].
- Pola Organisasi data. Pola untuk mengorganisasikan data dapat berupa menjadi struktur yang berbeda, urutan maupun menjadi terkelompok [3].

- Pola Mettapattern. Pola yang dapat mengoptimisasi pekerjaan MapReduce seperti pekerjaan beberapa tahap MapReduce dan optimisasi pembacaan input. [3].

Pola-pola ini belum banyak dimanfaatkan untuk analisis berbagai tipe *Big Data* yang ada dalam menyajikan suatu informasi. Setelah mengetahui deskripsi dan belum banyaknya dilakukan analisis data dengan menggunakan pola MapReduce, akan dicoba pembuatan aplikasi analisis data yang mengimplementasikan desain pola summarisasi, Organisasi data dan Metapattern menggunakan *Big Data* yang sebenarnya.

1.2 Rumusan Masalah

Berdasarkan deskripsi di atas, ditemukan beberapa rumusan masalah yang akan dituang dalam bentuk pertanyaan seperti berikut :

1. Bagaimana cara kerja pola summarisasi, organisasi data dan metapattern ?
2. Bagaimana implementasi dan manfaat pola summarisasi, organisasi data dan metapattern dalam lingkungan Hadoop ?
3. Bagaimana kinerja pola summarisasi, organisasi data dan metapattern ?
4. Bagaimana mengembangkan perangkat lunak demo yang mengimplementasikan pola sumarisasi, organisasi data dan metapattern ?

1.3 Tujuan

Tujuan dari penelitian sebagai berikut:

1. Mempelajari cara kerja pola summarisasi, organisasi data dan metapattern
2. Mengimplementasikan pola desain MapReduce dalam lingkungan Hadoop untuk menganalisis beberapa jenis *Big Data*.
3. Melakukan beberapa eksperimen untuk mengetahui kinerja desain pola MapReduce khususnya pola summarisasi, organisasi data dan metapattern.
4. Mengembangkan perangkat lunak demo yang mengimplementasikan desain pola MapReduce khususnya pola summarisasi, organisasi data dan metapattern

1.4 Batasan Masalah

Untuk membatasi cakupan masalah, berikut batasan masalah dari skripsi ini:

1. Perangkat lunak simulasi dibangun dengan memanfaatkan bahasa pemrograman Java dan *Framework* Hadoop.
2. MapReduce terdiri dari beberapa desain pola, tetapi pada skripsi ini hanya dibahas 3 pola yaitu, pola summarisasi, pola organisasi data, pola metapattern.
3. Perangkat lunak hanya dapat menganalisis data sesuai yang terdapat pada buku dan kasus masalah tertentu.
4. Data digunakan pada perangkat lunak diambil dari data dump website www.archive.org dan www.kaggle.com.

1.5 Metodologi

Metodologi yang digunakan dalam pembuatan skripsi ini adalah:

1. Melakukan studi pustaka mengenai cara kerja Hadoop,serta arsitektur yang ada dalam Hadoop.
2. Melakukan studi pustaka mengenai cara kerja desain pola MapReduce dalam memecahkan suatu masalah pengolahan data.
3. Mengimplementasikan setiap pola dalam lingkungan Hadoop untuk memecahkan masalah pengolahan data.
4. Mengimplementasikan sistem Hadoop Cluster atau sistem terdistribusi Hadoop.
5. Mengumpulkan data uji dengan kategori big data.
6. Merancang dan mengimplementasikan perangkat lunak dengan desain pola MapReduce untuk menganalisis big data yang sebenarnya.
7. Melakukan eksperimen dan analisis terhadap hasil pencarian data agar mengetahui kinerja dari suatu desain pola.
8. Membuat perangkat lunak demo yang memudahkan menjalankan pekerjaan MapReduce diatas platform hadoop.

1.6 Sistematika Pembahasan

Bab 1 menjelaskan latar belakang untuk penelitian ini dalam bentuk motivasi. Berdasarkan motivasi, ditemukan permasalahan-permasalahan dalam menyelesaikan penelitian ini dan akhirnya menemukan solusi. Solusi ini dijadikan panduan untuk membuat tujuan penyelesaian penelitian.

Bab 2 menjelaskan seluruh teori tentang framework Hadoop dan desain pola MapReduce yang digunakan pada pembuatan skripsi ini. Pada bab ini akan menjelaskan konsep Hadoop melalui arsitektur, modul-modul yang ada.Lalu akan dijelaskan konsep desain pola MapReduce serta struktur desain pola dalam memecahkan suatu masalah.

Bab 3 akan menjelaskan hasil studi eksperimen dari pola desain Mapreduce dalam lingkungan Hadoop. Dijelaskan juga analisis pembuatan program dan perangkat keras yang digunakan.

Bab 4 akan menjelaskan pemanfaat desain pola MapReduce untuk menganalisis data sebenarnya yang menjadi kasus uji. Dijelaskan jenis data yang digunakan serta analisis penyelesaian masalah menggunakan pola-pola MapReduce

Bab 5 akan menjelaskan tentang kebutuhan implementasi pada demo perangkat lunak serta analisis perangkat lunak. Dijelaskan rancangan antarmuka dan hasil implementasi dari perangkat lunak.Lalu dilakukan eksperimen untuk mengetahui kinerja desain pola MapReduce terhadap ukuran data.

Bab 6, terakhir, akan menjelaskan tentang kesimpulan dan saran dari penulis untuk para pembaca bila ingin mengembangkan aplikasi pengolahan data besar.