SYSTEMIC
Vol. 04, No. 01, Agustus 2018, 11-16

IMPLEMENTASI SENTIMEN ANALYSIS PENGOLAHAN KATA BERBASIS ALGORITMA MAP REDUCE MENGGUNAKAN HADOOP

Fawaidul Badri[1)]

[1] Jurusan Teknik Informatika Universitas Nahdlatul Ulama Sidoarjo
e-mail: fawaid90.ti@unusida.ac.id[1)]

Abstrak

Sentimen analysis merupakan bidang riset komputasi berbasis text mining dari opini sentimen dan emosi yang diekpresikan secara tektual. Dokumen teks yang digunakan dalam penelitian ini berasal dari web tentang opini masyarakat mengenai tingkat kualitas pendidikan secara umum. Metode yang digunakan dalam penelitian ini menggunakan algoritma map reduce untuk menghitung dari sebuah kata secara kesuluruhan sehingga akan menemukan kata yang sering muncul untuk dijadikan sebuah acuan dalam menyimpulkan opini masyarakat. Algoritma map reduce mengambil set data dan mengubah menjadi satu data set, unsur-unsur individu data set dipisah menjadi tupel. Tahapan dari algoritma map reduce membaca inputan data yang berupa text yang tersimpan dalam HDFS (Hadoop Distributed File System) kemudian akan di akan diproses sesuai pasangan dari key dan value yang sudah dirubah kedalam bentuk tuple. Tahap berikutnya dilakukan proses shuffel dan reduce yang kemudian akan diproses sehingga akan menghasilkan sebuah keputusan dari data set yang diproses.Hasil sistem penelitian implementasi sentimen analysis pengolahan kata dengan menggunakan algoritma map reduce menghasilkan penghitungan kata dengan sangat baik dengan parameter yang sama.

Kata Kunci: *Sentimen Analysis, text mining, opini, dokumen, map reduce.*

Abstract

Sentiment analysis is a field of text and information based research. Text documents in this language come from the web about socialization issues. The method used in this study uses algorithmic maps to calculate from a word that will be used to find a meaning in the context of public opinion. The map algorithm reduces the retrieval of data sets and converts them into a data set, data collection of individuals separated into tuples. The stages of the map algorithm reduce reading input data in the form of text stored in HDFS (Hadoop Distributed File System) then it will be processed according to the key and the value has been changed into tuple form. The next step is to process the shuffel and reduce it which will then produce a process from the data set that is processed. Furthermore, the research data uses sentiment analysis by using a map algorithm to reduce the amount of data that is very good.

.

Keywords: *Sentimen Analysis, text mining, opinion, document, map reduce.*

1. PENDAHULUAN

Perkembangan dokumen berbasis teks pada saat ini perkembangan yang sangat pesat sehingga menjadi salah satu sebab percarian dokumen berbasis text menjadi hal yang sulit. Teruma teks yang beredar melalui internet. Perkembangan ini menjadi hal yang sangat menarik dibidang informatika khususnya dibidang pemrosesan dokumen teks bahasa indonesia. Data mining merupakan salah cabang ilmu komputer yang mempelajari tentang penggalian pola data dengan tujuan untuk memperoleh informasi yang lebih berharga, penambangan pengetahuan dari sebuah data dengan jumlah yang sangat besar [1]. Salah satu fungsi dari data mining untuk mencari pola dari sebuah dokumen sehingga data yang akan dicari akan diketahui [2].

Data mining memiliki tujuan sebagai pengelompokan data, pencarian, klasifikasi dan menemukan pola dari data, sehingga mempunyai nilai informasi yang penting untuk dijadikan sebuah pohon keputusan untuk membantu mengambil sebuah keputusan. Dokumen teks merupakan salah satu dokumen yang sering diproses untuk menghasilkan sebuah informasi

yang lebih berharga dan mempunyai pola sehingga bisa dijadikan sebuah penelitian untuk membantu mengetahui informasi yang mempunyai nilai lebih [3].

Dalam sentimen analysis data yang diposes berkelanjutan, dimulai dari data mentah yang kemudian akan diproses dengan menggunakan algoritma map reduce dengan mengubah data set menjadi beberapa data tupel untuk mengolah perkata, yang kemudian akan diproses didalam *HDFS* (*Hadoop Distributed File System*) untuk menghasilkan key dan value dari setiap kata mentah yang diproses [4][5]. Inputan dalam bentuk pasangan key atau value untuk menghasilkan sebuah output berupa pasangan key atau value juga. Pasangan key atau value map ini dinamakan intermediate[6][7]. Kemudian, fungsi recude akan membaca pasangan key atau value intermediate hasil fungsi map. Proses selanjutnya tiap value yang memiliki yang sama akan digabungkan dalam menghasilkan satu kelompok fungsi recude untuk menghasilkan sebuah output berupa pasngan key atau value dari data yang diproses.

2. LANDASAN TEORI

2.1 Sentimen Analysis

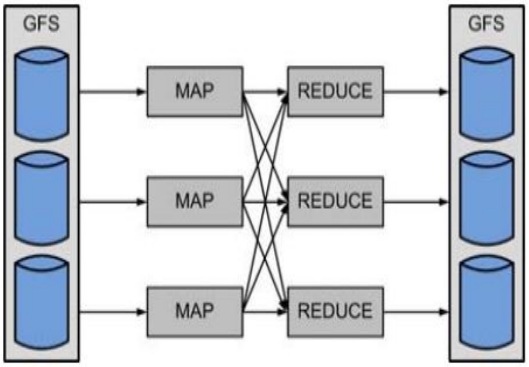
Sentiment analysis merupakan bidang ilmu komputasi yang menjelaskan emosi, sikap, penilaian, opini, dari sekumpulan teks yang fokusnya untuk mengindentifikasi, mengekstraksi, atau menemukan pola sentimen didalam unit teks dengan menggunakan metode *Natural Language Processing* atau *NLP*, machine learning atau sentiment analysis merupakan sebuah proses untuk klasifikasi text dokumen ke dalam beberapa kelas seperti sentimen negatif atau positif serta besarnya pengaruh atau manfaat dari sentimen analisis dapat menyebabkan penelitian ataupun memperkaya literatur penelitian dibidang sentimen analis.

Penelitian sentimen analisis mengalami perkembangan yang cukup pesat hal ini ditandai dlebih dari 20 sampai 30 bidang perusahaan dan layanan publik menggunakan sentimen analisi. Pada dasarnya sentiment analysis merupakan metode klasifikasi atau mencari pola dari sebuah data, namun dalam implementasi penerapan sentimen analisi sulit hal ini disebabkan proses klasifikasi berkaitan dengan text bahasa, terdapat tata bahasa yang ambigu didalam penggunaan kata, bukan hanya tata bahasa dalam sebuah teks, namun seiiring dalam perkembangan disebabkan oleh aturan bahasa itu sendiri.

Perkembangan penelitian dibidang sentimen analisis mangalami perkembangan yang cukup pesat bahkan dinegara super power amerika serikat memfokuskan dengan menggunakan layanan sentimen analysis, baik di perusahaan ataupun layanan publik. Pada dasarnya sentiment analysis merupakan metode untuk klasifikasi klaster dari sebuah data text mining, namun dalam implementasi proses klasifikasi tidak mudah hal ini disebabkan oleh tatanan bahasa dan ambigu dalam dalam penggunaan kata, dan terdapat tatanan bahasa yang semakin berkembang.

2.2 Map Reduce

Map Reduce merupakan salah satu software framework untuk digunakan sebagai mendukung sistem terdistribusi dengan penolahan data yang cukup besar serta mesin framework ini diperkenalkan pertama oleh perusahaan Google. Berikut gambar 2.1 merupakan cara kerja secara umum dari framework map reduce :



Gambar 2.1 Map recude

2.2.1 Map

Merupakan sebuah proses ketika node master menerima inputan data berupa file, kemudian data inputan tersebut dipecah menjadi beberapa individu-individu yang kemudian akan didistribusikan ke job node. job node akan memproses beberapa bagian data individu-invidu yang diterima yang kemudian data node invidu tersebut sudah diproses, selanjutnya akan dikembalikan ke master node semula untuk diproses kembali.

2.2.2 Reduce

Reduce merupakan sebuah proses dimana node master menerima notifikasi dari semua bagian data individu dari banyaknya node, serta menggabungkan master node tersebut menjadi satu individu besar untuk penyelesaian dari sebuah permasalahan utama. Map reduce mempunyai keuntungan yaitu memproses map dan reduce dijalankan secara terdistribusi dalam sistem map reduce. Dari setiap proses mapping data yang sifatnya independen sehingga membuat proses dijalankan secara paralel dan simultan untuk menghasilkan pengurangan dari data yang cukup besar. Pada proses reduce juga dilakukan secara paralel pada waktu bersamaan, keluaran data mapping akan mengirimkan value atau key ke

dalam reduce untuk diolah sesuai dengan logic sentimen analisi. Pada proses mapreduce dapat diimplementasikan pada cluster server dengan jumlah yang cukup besar sehingga sehingga proses komputasi dari pengolahan sangat cepat dan akurat.

Sistem hadoop dapat mereduce engine yang mempunyai banyak job tracer sehingga banyak proses bisa diselesaikan dalam waktu bersamaan job tracer merupakan sebuah server yang mengolah mapreduce dari beberapa client untuk meminimalkan resource yang ada.

2.3 *HDFS* (*Hadoop Distribused File System*)

HDFS merupakan sebuah file sistem yang dapat menyimpan data yang cukup besar dan saling saling terintegrasi antar data dengan cara mendistribusikan ke banyak komputer yang saling terhubung antara komputer satu dengan yang lain. File yang masuk kedalam sistem akan dipecah dalam bentuk blok-blok dengan ukuran kapasitas file yang akan diproses pada HDFS. Data akan mereplikasi sesuai dengan banyaknya data kedalam bentuk node, yang kemudian akan disimpan ke beberapa file secara terstruktur kedalam beberapa rak untuk menjadi kestabilan dari file yang ada dalam HDFS.

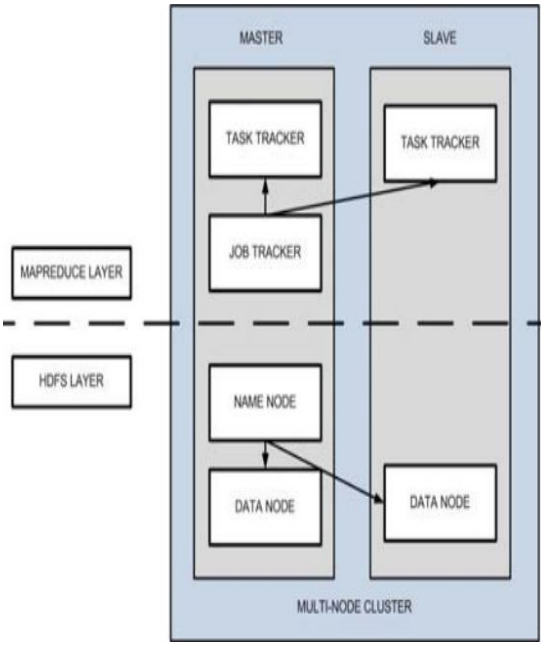
Dengan demikian node data yang berada dalam HDFS membutuhkan server sprimer untuk untuk menyimpan beberapa metadata yang kemudian akan didistribusikan kedalam protokol server untuk menjaga kestabilan node yang berjalan secara realtime dalam waktu yang bersamaan sehingga dibutuhkan sebuah protokol HTTP supaya komunikasi antar node saling berkomunikasi secara baik serta menjaga proses replikasi berjalan dengan baik.

Kelemahan dari HDFS yaitu menjaga master node untuk menjaga kestabilan agar node-node dalam sistem tidak hilang walaupun server primer mati. Sehingga file akan tersimpan secara otomatis kedalam secondary dan akan menyimpan informasi yang baru pada node direktori sistem HDFS. Dengan demikian dibutuhkan sebuah cloning node sistem dimana kalau terjadi server mati atau mengalami gangguan akan menyimpan dan mengkloning sistem secara langsung, sehingg file yang ada HDFS tidak ada informasi yang hilang.

Kelebihan dari model HDFS yaitu untuk menjadwal dan menentukan peta pola peta dari jobtracer untuk meminimalkan dan mengurangi pekerjaan yang berjalan pada pola data yang sama. Misal contoh data pada node C (a, b, c) dan data yang berada node dengan variabel D (e, f, g). Jobtracer pada node C akan mengurangi/peta lintasan transfer laju data yang tidak perlu, sehingga akan mengurangi penggunaan resource node yang terpakai. Node terintegrasi dalam HDFS akan meminimalkan jumlah pemakaian file sistem dalam hadoop, serta dapat mengurangi lintasan pola data yang tidak penting yang berjalan dalam sistem secara paralel.

2.4 Hadoop

Hadoop salah satu software framework open source dengan teknologi terbaru berbasis java. Struktur hadoop sebagai pengolahan data yang sangat besar, data-data diolah secara terintegrasi dan terdistribusi dalam beberapa komputer dimana satu dengan lainnya saling terhubung. Proses komputasi pada sistem hadoop sangat cepat sehingga sangat bagus untuk mengolah data yang sangat besar seperti sistem cloud.



Gambar 2.2 Model Hadoop

Rancangan kinerja sistem hadoop menyediakan beberapa fungsi common hadoop serta terdapat file sistem untuk mendistribusikan file-file node yang berjalan dalam jaringan sehingga kinerja dalam file sistem akan lebih cepat. File sistem hadoop memiliki sistem yang lebih kompatibel dalam mengatur penjadwalan node yang akan diproses supaya node bisa bekerja secara optimal. Slave node merupakan contoh dari cluster hadoop yang memliki struktur master node untuk memproses node secara bersamaan.

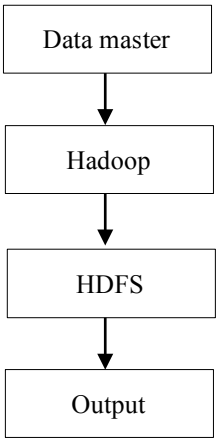
Node master dalam struktur hadoop terdapat beberapa bagian yaitu name node, node data, task tracker, jobtracer. Node name dan node data menjalankan tugas sesuai dengan peta node kemudian tracker dan jobtracer akan mengirimkan

ke server untuk diolah sesuai output logic dalam hadoop.

Dalam cluster sistem yang cukup besae, file HDFS akan menjalankan node name untuk didistribusikan ke indek hos sistem HDFS, sehingga mengurangi proses file yang rusak. Pada mapreduce cluster hadoop file sistem akan mengirimkan ke cloud untuk menggantikan node sekunder sehingga file yang dijalankan akan lebih spesifik.

3. METODE PENELITIAN

Metode penelitian yang dilakukan pada penelitian ini melalui beberapa tahapan, tahapan penelitian dapat dilihat pada gambar 3.1.



Gambar 3.1 Blok Diagram penelitian

3.1 Data Master

Data master merupakan data yang digunakan dalam penelitian ini, data didapat dari website edukasi kompas, data yang akan diproses berupa data text yang akan menjadi inputan ke sistem, data mempunyai record sebanyak 3000 record, dimana record data ini akan diproses sehingga akan menghasilkan sebuah output.

3.2 Hadoop

Data master sebanyak 3000 record akan di masukkan ke dalam sistem hadoop dimana data akan distribusikan kedalam beberapa komputer untuk di cluster dan diintegrasikan kedalam sistem komputer secara bersamaan. node slave akan memecah node sehingga terdapat beberapa node yang kemudian akan diproses dalam server sistem HDFS untuk memudahkan mengirimkan name node ke indek server dengan demikian akan mengurangi proses terjadinya pengurangan data node yang hilang, dan memaksimalkan kinerja dari server file HDFS untuk proses komputasi yang cepat dan akurat. Name node sekunder dapat menghasilkan snapshot dari struktur memori name node, sehingga mencegah resource sistem file dan mengurangi data yang hilang. Node-node pada sistem akan memproses sesuai dengan logic HDFS secara terstruktur.

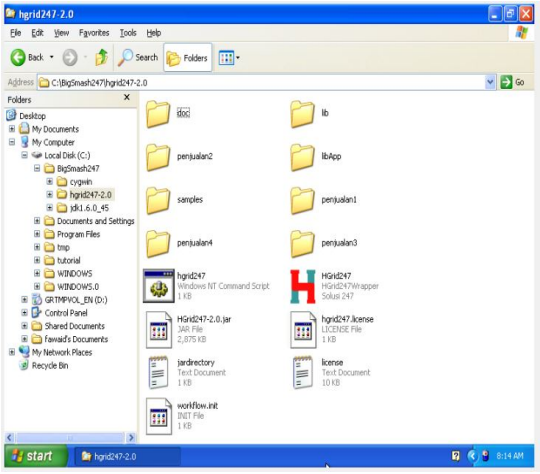
3.3 *HDFS (Hadoop Distributed File System)*

File data yang didapat dari hadoop akan ditransfer ke dalam file HDFS dimana node akan menyimpan informasi terbaru serta menjalankan sesuai direktori node name logic pada sistem HDFS. Kemudian node-node akan menggabungkan untuk menyelesaikan sebuah proses yang besar yang berada pada file sistem node mapreduce, dalam mapreduce terdapat proses mapping independen proses dimana key value mengirimkan sesuai dengan proses reducenya. Proses mapping pada node mane akan berjalan secara bersamaan dan simultan untuk mendistribusikan kedalam bentuk reduce yang mempunyai keunggulan mempunyai kemampuan proses komputasi yang cepat dan akurat dalam mereduce sebuah data text.

Node data yang berupa text setelah diproses sesuai dengan node map akan dimasukkan ke proses tahapan shuffle yang kemudian akan proses ke tahapan HDFS kembali. Proses akan berulang key value akan mengirimkan node data dan kemudian akan berjalan sesuai tahapan pada node map. Node-node akan diproses oleh server. kemudian masukan tersebut dipecah menjadi beberapa bagian permasalahan yang kemudian didistribusikan ke worker nodes. Worker node sini akan memproses beberapa bagian permasalahan yang diterimanya untuk kemudian apabila masalah tersebut sudah diselesaikan, maka akan dikembalikan ke master node.

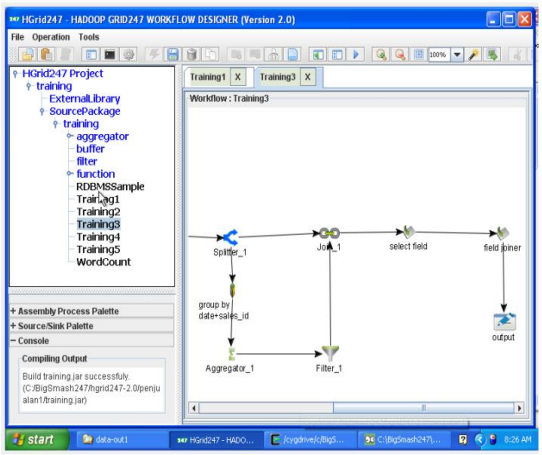
4. HASIL DAN PEMBAHASAN

Dari penelitian yang dilakukan didapatkan hasil penelitian sentimen analysis berbasis algoritma map reduce dengan hadoop. Pada penelitian ini mempunyai beberapa proses tahapan, proses pertama input data master, data master berupa data text, tahap berikutnya hadoop, dimana data text akan di konversi ke dalam bentuk node-node, sehingga akan memudahkan untuk mendistribusikan ke dalam server. Tahapan selanjutnya *HDFS* untuk memproses node-node sesuai dengan logic. Logic pada HDFS terdiri dari beberapa node input data text, node splitter, node aggregator, node filter, node join, node select field, field joiner, dan output. Tahapan terakhir yaitu output.



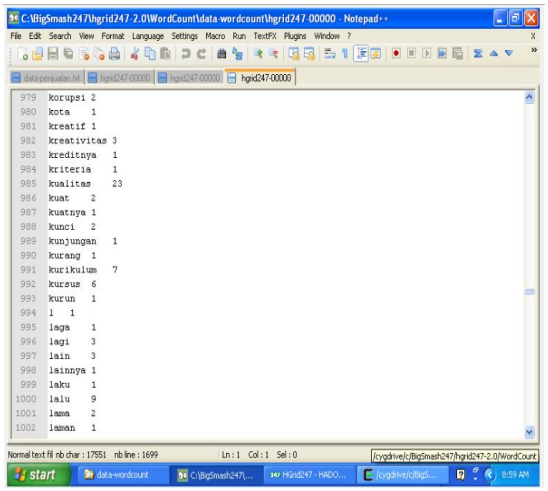
Gambar 4.1 Hadoop versi hgrid247.2.0

Pada gambar 4.1 terlihat hadoop dengan versi hgrid247. 2.0 terdapat beberapa library untuk mengolah data mentah dari penelitian ini.



Gambar 4.2 hadoop versi hgrid247.2.0

Pada gambar 4.2 menunjukkan hasil dari model *HDFS (Hadoop Distributed System)* dimana model ini akan memproses file ke server. Dalam sistem *HDFS* terdapat 7 node. Node pertama adalah input data, dimana data yang dimasukkan kedalam sistem ini berupa data text, kemudian node splitter node ini untuk memecah text yang jumlahnya besar, sehingga proses selanjutnya akan lebih mudah dan proses komputasi lebih cepat. proses selanjutnya node aggregator untuk mengunpulkan semua node-node terklasifikasi memudahkan untuk mencari text yang jumlahnya besar. Node filter fugsi dari node ini untuk memfilter tiap kata yang diproses, proses filter kata menjadi hal penting agar menemukan kata yang mewakili dari setiap kalimat. Node join untuk menggabungkan setiap kata yang terklasfikasi, sehingga dilanjutkan ke node berikutnya node filed untuk mencari filed kata yang mempunyai pasangan, kemudian dari pasangan kata tersebut akan di filter lagi dengan node filed joiner, dimana proses node ini merupakan proses yang terakhir untuk mendapatkan infornasi dari tiap kata yang diproses, yang kemudian akan ditampikan di node output menampikan hasil dari seluruh kata yang diproses, output dari poses dapat dilihat pada gambar 4.3.



Gambar 4.4 Output Sistem

Gambar 4.4 menunjukkan hasil dari proses sistem, dimana setiap kata yang yang terdapat diwebsite akan dihitung, kemudian akan menyimpulkan kata yang sering muncul akan mewakili dari opini, bahwa analisis sentimen dari sistem akan diketahui. Kata yang banyak mncul merepresentasikan opini yang terdapat dalam sistem, sehingga bisa dianalisa bentuk sentimen masyarakat tentang pelayanan pendidikan di indonesia.

5. KESIMPULAN

Dari penelitian yang dilakukan maka didapat sebuah kesimpulan sebagai berikut :

1. Semakin banyak data yang dimasukka ke sistem hadoop proses komputasi semakin lama.
2. Node-node dalam memecah text masih terjadi kesalahan, diperlukan node-node yang baru untuk proses seleksi text sehingga proses menghasilkan dengan baik.
3. Diperlukan data yang lebih banyak untuk menguji keakuran sistem untuk menganalisa dari tiap kata yang akan akan diproses.
4. Proses aggregator menunjukkan performa yang baik, sehingga cocok untuk sistem yang skala besar.

DAFTAR PUSTAKA

[1] AlHakami. H., dkk., '*Comparison Between Cloud And Grid Computing: Review Paper*' , International Journal on Cloud Computing: Services and Architecture (IJCCSA),Vol.2, No.4, August 2012

[2] Bart Jacob, Michael Brown, dkk., "*Introduction to Grid Computing", International Technical Support Organization*.

[3] Casanova. Henri., '*Distributed Computing Research Issues in Grid Computing*', University of California at San Diego, La Jolla, CA 92093-0505.

[4] Javier Conejero, Peter Burnap, Omer Rana, Jeffrey Morgan, 2013. *Scaling Archived Social Media Data Analysis using a Hadoop Cloud*.

[5] Visalakshi P and Karthik TU,2011 *MapReduce Scheduler Using Classifiers for Heterogeneous Workloads*,

[6] Keke Cai, Scott Spangler, Ying Chen, Li Zhang, 2006, *Leveraging Sentiment Analysis for Topic Detection*.

[7] Neethu M S, Rajasree R, 2013 *'Sentiment Analysis in Twitter using Machine Learning Techniques'*.