IMPLEMENTASI KLUSTER KOMPUTER UNTUK EKSEKUSI ALGORITMA MACHINE LEARNING MENGGUNAKAN APACHE MAHOUT DAN HADOOP

Mohamad Naufal Ihsan[1] **Ismail[2]** **Yahdi Siradj[3]**

[1,2,3] Fakultas Ilmu Terapan –Telkom University
[1]mohamadnaufalihsan@gmail.com [2]ismail@tass.telkomuniversity.ac.id [3]yahdi@tass.telkomuniversity.ac.id

Abstrak

Saat ini teknologi server terus berkembang, demikian juga teknologi kluster komputer untuk berbagai keperluan dalam pengolahan data. Dengan hal ini penulis membuat komputer kluster serta mempelajari keunggulan komputer kluster untuk mengeksekusi algoritma machine learning. Dengan demikian dengan hadirnya komputer kluster ini akan menawarkan layanan untuk melakukan proses komputasi yang lebih cepat.

Untuk membuat komputer kluster diperlukan beberapa tahapan mulai dari instalasi Apache Hadoop sampai membagi node pada masing-masing komputer. Dilanjutkan dengan instalasi Apache Mahout sebagai perangkat lunak yang menyediakan algoritma machine learning untuk di eksekusi pada kluster komputer. Adapula Zabbix yaitu aplikasi monitoring yang digunakan untuk mengetahui sejauh mana kinerja CPU load dan trafik jaringan port ethernet yang digunakan selama proses eksekusi algoritma machine learning berlangsung.

Pada dataset 20 Newsgroups dengan ukuran file 36 MB, terlihat rata-rata lamanya waktu eksekusi dengan 1 node mencapai 22 menit 12 detik, 2 node mencapai 20 menit 42 detik dan 3 node hanya mencapai 18 menit 3 detik. Sedangkan pada dataset Wikipedia XML dengan ukuran file 1 GB, terlihat rata-rata lamanya waktu eksekusi dengan 1 node mencapai 22 menit 6 detik, 2 node mencapai 18 menit 39 detik dan 3 node mencapai 15 menit 6 detik. Hasil pengujian menunjukan, ada perbandingan waktu yang cukup signifikan mencapai rata-rata 2 sampai 7 menit lebih cepat bila node ditambahkan lebih banyak, sehingga 3 node lebih cepat dari 2 node dan 1 node.

Kata kunci: Komputer kluster, Apache Hadoop, Apache Mahout, *Machine learning*, Zabbix

Abstract

The current server technology continues to evolve, so does the computer cluster technology for various purposes in data processing. By this, author makes the computer clusters and study advantage cluster computer to execute a machine learning algorithm. Thus the presence of this cluster computer will offer services to make the process faster computing.

To create a cluster computer takes several steps ranging from the installation of Apache Hadoop to divide the nodes on each computer. Proceed with the installation of Apache Mahout as software that provides machine learning algorithms for execution on a computer cluster. Zabbix monitoring application that is used to determine the extent to which performance of the CPU load and network traffic ethernet port that is used for machine learning algorithm execution process takes place.

From the results of this final project, the dataset 20 Newsgroups with file size 36 MB, looks average length of time of execution with 1 node reached 22 minutes 12 seconds, 2 nodes reached 20 minutes 42 seconds and 3 nodes reached 18 minutes 3 seconds. While the Wikipedia XML dataset with a file size of 1 GB, seen the average length of time of execution with 1 node reached 22 minutes 6 seconds, 2 nodes reached 18 minutes 39 seconds and 3 nodes reached 15 minutes 6 seconds. The test results show, there are significant time ratio reached an average of 3 minutes and 12 minutes faster when more nodes are added, so three nodes faster than two nodes and one node.

Keywords: Komputer kluster, Apache Hadoop, Apache Mahout, *Machine learning*, Zabbix

1. Pendahuluan

Semakin meningkatnya kebutuhan akses data yang cepat dan akurat, dengan demikian akan terjadi masa peralihan dimana sebagian pekerjaan dikerjakan dengan menggunakan komputer dibandingkan secara manual oleh manusia. Namun dibutuhkan sumber daya yang besar pada sebuah komputer untuk menjalankan proses data secara cepat. Penggunaan super komputer merupakan salah satu solusinya. Mengingat harga dari super komputer sangat mahal, ini menjadikan tidak semua kalangan bisa menggunakannya.

Berkembangnya teknologi jaringan saat ini, telah muncul metode kluster komputer sebagai cara untuk meningkatkan kinerja komputer dalam pengolahan data di suatu jaringan. Kluster komputer merupakan sebuah kumpulan komputer yang saling terhubung dan bekerja sama sehingga membentuk sebuah komputer tunggal. Metode ini sangat bermaanfaat untuk menggabungkan sumberdaya perangkat keras komputer menjadi satu dan dapat menekan harga agar lebih murah karena tidak diperlukan komputer server yang berspesifikasi tinggi.

Berdasarkan pemaparan diatas, proyek akhir ini akan menerapkan sistem kluster komputer menggunakan Apache Hadoop dengan komputer biasa yang terbagi menjadi 1 node, 2 node dan 3 node untuk membandingkan lamanya waktu proses eksekusi algoritma machine learning pada Apache Mahout. Adapula monitoring CPU load dan trafik jaringan port ethernet menggunakan Zabbix pada masing-masing komputer.

2. Dasar Teori

2.1 Kluster komputer

Kluster komputer adalah suatu sistem perangkat keras dan perangkat lunak yang menggabungkan beberapa komputer dalam suatu jaringan sehingga komputer-komputer tersebut dapat bekerjasama dalam memproses data. Kluster komputer menawarkan sejumlah keuntungan, antara lain:

1. Menghemat pengeluaran biaya untuk sumber daya perangkat yang ada.
2. Kekuatan proses pengolahan data secara paralel dari kluster.
3. Kluster komputer dapat dengan mudah diperluas dengan menambahkan node tambahan ke jaringan.
4. Kluster komputer memiliki sistem penyimpanan yang baik karena data akan disebarkan ke setiap node.

2.2 Hadoop

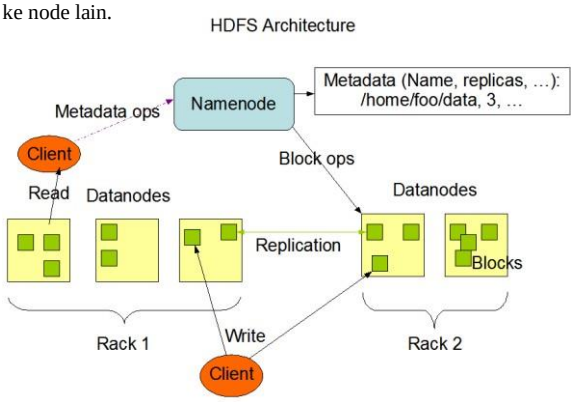
Hadoop adalah perangkat lunak berbasis java dan open source yang berfungsi untuk mengolah data yang sangat besar secara terdistribusi dan berjalan di atas kluster yang terdiri atas beberapa komputer yang saling terhubung. Hadoop terdiri dari sistem penyimpanan file terdistribusi yang disebut HDFS dan mengimplementasikan program pengolahan data yang disebut

MapReduce. Hadoop bisa digunakan untuk komputer apapun tanpa perangkat keras atau perangkat jaringan khusus [1].

2.2.1 HDFS

HDFS (Hadoop Distributed File System) adalah sebuah tempat untuk menyimpan direktori dan file di kluster komputer hadoop. HDFS ini tidak sama seperti dengan jenis sistem penyimpanan dari sistem operasi misalnya NTFS atau FAT32. HDFS ini menumpang di atas sistem penyimpanan milik sistem operasi.

Setiap direktori dan file yang disimpan di HDFS selalu memiliki lebih dari satu salinan yang disebut refplication factor. Suatu file disimpan di Datanode sehingga jika ada satu Datanode yang rusak, maka Datanode yang lain bisa memberikan datanya. Setiap 3 detik sekali, Datanode mengirim sinyal yang disebut heartbeat, ke Namenode untuk menunjukkan bahwa Datanode masih aktif. Bila dalam 10 menit Namenode tidak menerima heartbeat dari Datanode, maka Datanode tersebut dianggap rusak atau tidak berfungsi sehingga setiap request read/write dialihkan ke node lain.



Gambar 1 Arsitektur HDFS

Sumber: hadoop.apache.org

Ada tiga komponen penting pada HDFS dalam mengekseskusi program MapReduce yang dibuat, antara lain [2]:

1. Namenode merupakan node utama yang fungsinya mengatur penempatan data di kluster serta menerima job dan program untuk melakukan eksekusi pengolahan data dan analisis data melalui MapReduce. Namenode akan mengkoordinasikan ekseskusi dari Map Reduce terhadap seluruh data yang ada di kluster Hadoop. Namenode menyimpan metadata tempat data di kluster dan juga replikasi data tersebut. Namenode juga yang menjadwalkan eksekusi MapReduce sekaligus mengulang eksekusinya jika eksekusi berikutnya gagal.
2. Datanode yaitu fungsinya mengeksekusi MapReduce terhadap data yang ada di kluster Hadoop. Setelah selesai eksekusi, Datanode memberi tahu Namenode hasil dari eksekusi tersebut, gagal atau berhasil.
3. Secondary Namenode merupakan node yang bertugas untuk menyimpan informasi penyimpanan data dan pengolahan data yang ada di Namenode. Fungsinya adalah bila Namenode mati dan diganti dengan Namenode baru maka Namenode baru bisa langsung bekerja dengan mengambil data dari Secondary Namenode dialihkan ke node lain.

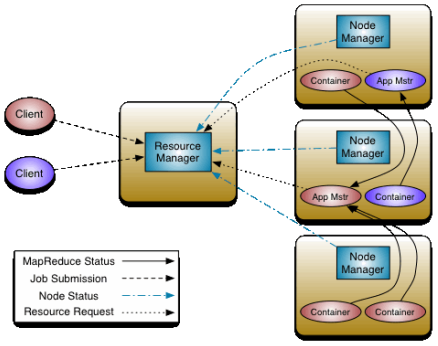
2.2.2 MapReduce

MapReduce merupakan eksekutor yang digunakan untuk menjalankan suatu proses pengolahan data sehingga akan didapatkan informasi dari data tersebut. MapReduce dibuat oleh Google lalu dikembangan oleh Apache untuk diterapkan pada Hadoop.

2.2.3 YARN

YARN (Yet Another Resource Negotiator) adalah perbaikan besar yang baru diperkenalkan di Hadoop v2. YARN adalah sistem pengelolaan sumber daya yang memungkinkan beberapa data didistribusikan pengolahan untuk secara efektif berbagi sumber daya komputasi dari kluster Hadoop dan untuk memanfaatkan data yang disimpan dalam HDFS.

MapReduce dengan sendirinya tidak cukup untuk mendukung pertumbuhan yang jumlah kasus penggunaan pemrosesan terdistribusi lain seperti perhitungan data real-time, perhitungan grafik, perhitungan looping, dan permintaan data real-time. Tujuan dari YARN adalah untuk memungkinkan pengguna untuk memanfaatkan beberapa aplikasi terdistribusi yang menyediakan seperti kemampuan berdampingan berbagi satu kluster dan file sistem HDFS.



Gambar 2 Cara Kerja Yarn

Sumber: hadoop.apache.org

Beberapa komponen yang terdapat pada YARN, antara lain [2]:

1. ResourceManager adalah scheduler sumber daya pusat yang mengelola dan mengalokasikan sumber daya untuk aplikasi yang berbeda (juga dikenal sebagai job) disampaikan kepada kluster. YARN NodeManager adalah per-node proses yang mengelola sumber daya dari menghitung node tunggal. Komponen scheduler dari ResourceManager mengalokasikan sumber daya dalam menanggapi permintaan sumber daya yang dibuat oleh aplikasi, dengan mempertimbangkan kapasitas kluster dan kebijakan penjadwalan lain yang dapat ditentukan melalui kerangka plugin kebijakan YARN.
2. Container merupakan unit dari alokasi sumber daya. Masing-masing dialokasikan container memiliki hak untuk menggunakan sejumlah CPU dan memori. Aplikasi dapat meminta sumber daya dari YARN dengan menentukan dibutuhkan jumlah container CPU dan memori yang dibutuhkan oleh masing-masing wadah.
3. ApplicationMaster adalah proses per-aplikasi yang mengkoordinasikan perhitungan untuk satu aplikasi. Langkah pertama mengeksekusi aplikasi YARN adalah untuk menyebarkan ApplicationMaster tersebut. Setelah aplikasi diajukan oleh klien YARN, ResourceManager mengalokasikan wadah dan menyebarkan ApplicationMaster untuk aplikasi tersebut. Setelah digunakan, ApplicationMaster bertanggung jawab untuk meminta dan negosiasi sumber daya container yang diperlukan dari ResourceManager. Setelah sumber daya yang dialokasikan oleh ResourceManager tersebut, koordinat ApplicationMaster dengan NodeManagers untuk meluncurkan dan memantau container aplikasi dalam sumber daya yang dialokasikan. Pergeseran tanggung jawab koordinasi aplikasi ke

ApplicationMaster mengurangi beban pada ResourceManager dan memungkinkan untuk fokus hanya pada pengelolaan sumber daya kluster. Juga memiliki ApplicationMasters terpisah untuk setiap permohonan yang diajukan untuk meningkatkan skalabilitas dari kluster untuk memiliki proses bottleneck tunggal untuk mengkoordinasikan semua contoh aplikasi.

2.3 Machine Learning

Machine learning merupakan metode komputasi untuk mempelajari dan membuat prediksi dari algoritma agar menghasilkan informasi yang akurat. Machine learning membutuhkan sampel algoritma dan dataset untuk bisa di analisis berdasarkan konsep dan kompleksitasnya.

2.3.1 Arctificial Intelligence

Artificial Intelligence merupakan cabang dari ilmu komputer yang telah lama mencoba untuk meniru kecerdasan manusia. Sebuah subset dari Artificial Intelligence, disebut sebagai machine learning, mencoba untuk membangun sistem cerdas dengan menggunakan data. Misalnya, sistem pembelajaran mesin dapat belajar untuk mengklasifikasikan spesies yang berbeda dari bunga atau berita-kelompok terkait bersama-sama untuk membentuk kategori seperti berita, olahraga, politik, dan sebagainya, dan untuk masing-masing tugas-tugas ini, sistem akan belajar menggunakan data. Untuk setiap tugas, algoritma yang sesuai akan melihat data dan mencoba untuk belajar dari itu.

2.3.2 Naive Bayes

Naive Bayes adalah classifier probabilistik berdasarkan teorema Bayes. Ini mengasumsikan kuat (naif) kebebasan antara fitur. Selama fitur yang tidak berkorelasi dan tidak berulang-

ulang, baik Naive Bayes dan regresi logistik akan melakukan dengan cara yang sama. Namun, ketika fitur tersebut berkorelasi dan berulang-ulang, algoritma Naive Bayes berperilaku berbeda karena asumsi kebebasan bersyarat.

Ini adalah persamaan matematika untuk teorema Bayes [3]:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (1)$$

Perihal A dan B:

1. P (A) dan P (B) adalah probabilitas A dan B, independen satu sama lain
2. P (A|B), probabilitas bersyarat, adalah probabilitas A mengingat bahwa B benar
3. P (B|A), adalah probabilitas B mengingat bahwa A benar

2.3.3 Confusion Matrix

Salah satu cara yang paling umum dan dasar mengevaluasi kinerja model adalah dengan menciptakan confusion matrix dan

komputasi berbagai metrik seperti accuracy, precision, recall, dan sebagainya. Contoh untuk dua kelas pengklasifikasi untuk memahami konsep-konsep, dan kemudian memperluas untuk masalah yang melibatkan lebih dari dua kelas.

Konsep perhitungan confusion matrix adalah sebagai berikut [4]:

Label satu kelas sebagai positif dan yang lain negatif. Untuk menggambarkan beberapa spesifik masalah dengan dataset yang tidak seimbang, maka akan mempertimbangkan contoh dengan label kelas tidak seimbang, di mana negatif adalah ham dan positif adalah spam.

Entri dalam confusion matrix memiliki arti sebagai berikut dalam konteks penelitian:

1. A adalah jumlah prediksi yang benar bahwa contoh negatif, di mana ham diprediksi sebagai ham.
2. B adalah jumlah prediksi yang salah bahwa sebuah contoh positif, di mana ham diperkirakan sebagai spam.
3. C adalah jumlah prediksi yang salah bahwa sebuah contoh negatif, di mana spam yang diprediksi sebagai ham.
4. D adalah jumlah prediksi yang benar bahwa contoh positif, di mana spam yang diperkirakan sebagai spam.

Kinerja metrik berasal dari confusion matrix adalah sebagai berikut:

1. Accuracy (AC) adalah proporsi jumlah prediksi yang benar. Hal ini ditentukan dengan menggunakan persamaan:

$$AC = \frac{A+D}{A+B+C+D} \qquad (2)$$

2. The recall or true positive rate (TP) adalah proporsi kasus positif yang diidentifikasi dengan benar, yang dihitung dengan menggunakan persamaan:

$$TP = \frac{D}{CD} \qquad (3)$$

3. The false positive rate (FP) adalah proporsi kasus negatif yang salah diklasifikasikan sebagai positif, yang dihitung dengan menggunakan persamaan:

$$FP = \frac{B}{AB} \qquad (4)$$

4. The true negative rate (TN) didefinisikan sebagai proporsi kasus negative yang diklasifikasikan dengan benar, yang dihitung dengan menggunakan persamaan:

$$TN = \frac{A}{AB} \qquad (5)$$

5. The false negative rate (FN) adalah proporsi kasus positif yang salah diklasifikasikan sebagai negatif, yang dihitung dengan menggunakan persamaan:

$$FN = \frac{C}{CD} \qquad (6)$$

6. Precision (P) adalah proporsi kasus positif meramalkan bahwa benar, yang dihitung dengan menggunakan persamaan:

$$P = \frac{D}{BD} \qquad (7)$$

Tabel 1 Confusion Matrix

Cara lain untuk mengukur model accuracy adalah untuk menghitung F-score. Keseimbangan F-score adalah mean harmonik precision dan recall:

$$F_1 = 2 \cdot \frac{\square \cdot \square}{\square + \square} \qquad (8)$$

Untuk masalah kelas tidak seimbang, akurasi mungkin tidak menjadi ukuran kinerja yang memadai ketika jumlah kasus negatif jauh lebih besar dari jumlah kasus positif. Misalkan ada 1000 kasus, 995 di antaranya adalah kasus negatif dan lima kasus positif; jika sistem mengklasifikasikan mereka semua negatif, akurasi akan 99,5 persen, meskipun classifier terjawab semua kasus positif. Dalam skenario ini, kita bisa menggunakan rumus umum, sebagai berikut:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \quad (9)$$

β memiliki nilai dari 0 hingga tak terbatas dan digunakan untuk mengontrol berat ditugaskan untuk TP dan P. Biasanya, kita dapat menggunakan rasio positif untuk kasus negatif sebagai nilai β.

2.4 Mahout

Apache Mahout merupakan proyek dari Apache Software Foundation yang memiliki implementasi algoritma machine learning. Mahout di mulai sebagai proyek dari Apache Lucene proyek pada tahun 2008. Setelah beberapa waktu, sebuah proyek open source yang bernama Taste, yang dikembangkan untuk collaborative filtering, dan itu diserap ke Mahout. Mahout di buat dengan Java dan memberikan skalabilitas algoritma machine learning. Mahout memberikan Java libraries dan tidak menyediakan antarmuka untuk pengguna atau komputer.

Saat ini, Mahout mendukung kasus penggunaan berikut [5]:

1. Recommendation yaitu mengambil data pengguna dan mencoba untuk memprediksi item bahwa pengguna mungkin ingin. Dengan kasus penggunaan ini, Anda dapat melihat semua situs yang menjual barang kepada pengguna. Berdasarkan tindakan sebelumnya, mereka akan mencoba untuk mengetahui item yang tidak diketahui yang bisa berguna. Salah satu contoh dapat ini: segera setelah Anda memilih beberapa buku dari Amazon, website akan menampilkan daftar buku-buku lain dengan judul, Customers Who Bought This Item Also Bought. Hal ini juga menunjukkan judul, What Other Items Do Customers Buy After Viewing This Item? Contoh lain dari recommendation adalah bahwa saat bermain video di YouTube, ia menyarankan agar Anda mendengarkan beberapa video lain berdasarkan pilihan Anda. Mahout memberikan dukungan API penuh untuk mengembangkan mesin rekomendasi Anda sendiri berbasis pengguna atau item berbasis.
2. Classification yaitu klasifikasi memutuskan berapa banyak item milik salah satu kategori tertentu. E-mail klasifikasi untuk menyaring spam adalah contoh klasik dari klasifikasi. Mahout menyediakan banyak set API untuk membangun model klasifikasi sendiri. Sebagai contoh, Mahout dapat digunakan untuk membangun classifier dokumen atau classifier e-mail.
3. Clustering yaitu teknik yang mencoba untuk item kelompok bersama-sama berdasarkan semacam kesamaan. Di sini, kita menemukan kelompok yang berbeda dari item berdasarkan sifat tertentu, dan kita tidak tahu nama cluster di depan. Perbedaan utama antara clustering dan klasifikasi adalah bahwa dalam klasifikasi, kita tahu akhir nama kelas. Clustering berguna untuk mencari tahu segmen pelanggan yang berbeda. Google News menggunakan teknik pengelompokan untuk berita kelompok. Untuk clustering, Mahout telah menerapkan beberapa algoritma yang paling populer, seperti k-means, fuzzy k-means, canopy, dan sebagainya.
4. Dimensional reduction yaitu proses mengurangi jumlah variabel acak dalam pertimbangan. Hal ini membuat data yang mudah digunakan. Mahout memberikan algoritma untuk pengurangan dimensi. Dekomposisi nilai singular dan Lanczos adalah contoh dari algoritma yang Mahout sediakan.
5. Topic modeling yaitu digunakan untuk menangkap ide ringkasan dokumen. Sebuah model topik adalah model

yang mengaitkan distribusi probabilitas dengan setiap dokumen pada topik. Mengingat bahwa dokumen adalah tentang topik tertentu, orang akan berharap

kata-kata tertentu muncul dalam dokumen lebih sering atau kurang. "Football" dan "goal" akan tampil lebih sering dalam dokumen tentang olahraga. Latent Dirichlet Allocation (LDA) adalah algoritma belajar yang kuat untuk pemodelan topik.

2.5 Zabbix

Zabbix merupakan perangkat lunak opensource untuk monitoring jaringan dan kinerja perangkat keras komputer sistem di linux. Dengan Zabbix kita bisa mudah mengetahui status server, kondisi jaringan dan mendapatkan peringatan berupa fitur notifikasi yang disebut dengan trigger apabila terjadi gangguan. Zabbix mudah dikonfigurasi karena menggunakan antarmuka berbasis web. Untuk mengetahui informasi monitoring dari suatu komputer, Zabbix bisa dikonfigurasikan dengan menggunakan Zabbix Agent dan SNMP Agent. Zabbix Agent berguna untuk memeriksa status komputer seperti kapasitas hardisk, beban memori, beban CPU dan beberapa proses lainnya. Sedangkan pada SNMP Agent kita bisa memantau lalu lintas TCP/IP pada perangkat jaringan LAN.

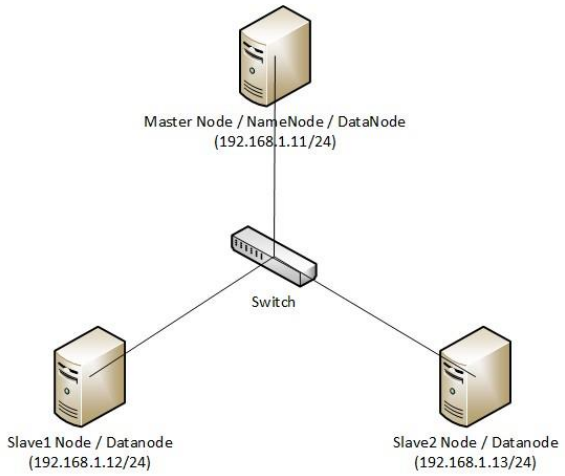
Fungsi dasar yang di harapkan dari monitoring, antara lain [6]:

1. Data gathering yaitu di mana semuanya dimulai. Biasanya data yang akan dikumpulkan dengan menggunakan berbagai metode, termasuk SNMP, agen, IPMI, dan lain-lain.
2. Alerting yaitu berkumpulnya data dapat dibandingkan dengan ambang batas dan tanda dikirim keluar bila diperlukan menggunakan saluran yang berbeda, seperti e-mail atau SMS.
3. Data storage yaitu setelah mengumpulkan data, aplikasi akan sering menyimpannya untuk analisis nanti.
4. Visualization yaitu manusia lebih baik saat membedakan data yang divisualisasikan dari angka baku, terutama ketika ada jumlah data yang besar. Aplikasi data yang telah dikumpulkan dan disimpan menghasilkan grafik sederhana.

3. Analisis Perancangan

3.1 Desain Jaringan Kluster Komputer

Pada proyek akhir ini kluster komputer diterapkan dengan 3 buah komputer yang terhubung dengan switch. Untuk meningkatkan performansi, maka komputer pada kluster ini dibagi menjadi 3 node. Hal ini bertujuan agar performansi yang dihasilkan mampu memberikan hasil yang lebih baik lagi.



Gambar 3 Desain Jaringan Komputer Kluster

Seluruh komputer menggunakan sistem operasi Ubuntu 14.04 dan sudah terinstal Hadoop. Mahout dan dataset hanya terinstal di Master node. Algoritma machine learning yang digunakan yaitu Naive Bayes karena jenis dataset 20 Newsgroups dan Wikipedia XML hanya bisa diklasifikasi oleh algoritma machine learning tersebut.

Aplikasi monitoring yang digunakan yaitu Zabbix yang terinstal pada Master node sebagai zabbix server dan Slave node sebagai zabbix agent. Monitoring berguna untuk mengetahui kinerja CPU load dan trafik jaringan port ethernet saat eksekusi algoritma machine learning berlangsung.

Pemberian IP pada setiap komputer ditentukan secara static yang berada dalam satu jaringan Network ID. Switch menggunakan Routerboard Mikrotik yang di ubah fungsi masing-masing port nya menjadi bridge.

3.2 Spesifikasi Perangkat Lunak dan Perangkat Keras

3.2.1 Spesifikasi Komputer Master

Berikut adalah spesifikasi laptop Master.

Tabel 2 Spesifikasi Komputer Master

Perangkat Keras	Tipe
OS	Ubuntu 14.04 64-bit
CPU Speed	2.4 GHz
CPU Model	Intel(R) Core(TM) i3-2370 CPU
Hardisk	HGST 500 GB
USB Controler	VIA Technologies, Inc. USB 3.0
RAM	6 GB VISIPRO DDR3
Audio Device	Realtek
GPU	Intel HD Graphics 3000
Motherboard	Acer BA40_HC

3.2.2 Spesifikasi Komputer Slave1

Berikut adalah spesifikasi laptop Slave1.

Tabel 3 Spesifikasi Komputer Slave1

Perangkat Keras	Tipe
OS	Ubuntu 14.04 64-bit
CPU Speed	3.0 GHz
CPU Model	Intel(R) Core(TM) i7-4510U CPU
Hardisk	1 TB
USB Controler	VIA Technologies, Inc. USB 3.0
RAM	4 GB DDR3
Audio Device	Dolby
GPU	AMD Radeon 2 GB
Motherboard	Lenovo

3.2.3 Spesifikasi Komputer Slave2

Berikut adalah spesifikasi komputer Slave2.

Tabel 4 Spesifikasi Komputer Slave2

Perangkat Keras	Tipe
OS	Ubuntu 14.04 64-bit
CPU Speed	3.5 GHz
CPU Model	Intel (R) Core (TM) i3-4150
Hardisk	500 GB
USB Controler	USB 2.0
RAM	4 GB DDR3
Audio Device	Realtek
GPU	Intel HD Graphics 4400
Motherboard	Lenovo Sharkbay

3.2.4 Spesifikasi Komponen Jaringan

Berikut ini adalah komponen jaringan yang digunakan.

Tabel 5 Spesifikasi Komponen Jaringan

Perangkat Keras	Tipe
Switch	Mikrotik RB951G
Kabel UTP	Belden Cat 5e

3.2.5 Spesifikasi Aplikasi

Berikut adalah aplikasi yang digunakan.

Tabel 6 Spesifikasi Aplikasi

Aplikasi	Versi
Apache Hadoop	2.6.0
Apache Mahout	0.10.0
Zabbix	2.2

4. Implementasi dan Pengujian

4.1 Implementasi

4.1.1 Instalasi Hadoop

Berikut ini adalah tahapan instalasi Hadoop.

1. Pertama, edit file "/etc/hosts/" tambahkan IP serta host master dan slave.

```
127.0.0.1       localhost
192.168.1.11    master
192.168.1.12    slave1
192.168.1.13    slave2
```

2. Lakukan update, instalasi SSH dan Java.

```
# apt-get update
# apt-get install ssh
# apt-get install default-jdk
```

3. Buatlah user dan group untuk Hadoop.

```
# adduser hduser
# usermod -a -G sudo hduser
# addgroup hadoop
# usermod -a -G hadoop hduser
```

4. Login ke hduser kemudian buat authorized key.

```
$ mkdir ~/.ssh
$ ssh-keygen -b 2048 -t rsa -f ~/.ssh/id_rsa -q -N ""
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

5. Melakukan ssh id ke user dan hostname slave.

```
$ ssh-copy-id -i ~/.ssh/id_rsa.pub hduser@slave1
$ ssh-copy-id -i ~/.ssh/id_rsa.pub hduser@slave2
```

6. Download hadoop versi 2.6.0 di website "hadoop.apache.org", lalu lakukan ekstrak dan instalasi.

```
# tar xvzf hadoop-2.6.0.tar.gz
# mv hadoop-2.6.0 /usr/local/hadoop/
```

7. Mangganti kepemilikan user pada folder hadoop.

```
# chown -R hduser:hadoop  /usr/local/hadoop/
```

8. Selanjutnya, lakukan edit pada file "~/.bashrc" untuk menambahkan konfigurasi variable aplikasi hadoop. Lalu, Reload file "~/.bashrc" dengan perintah "source ~/.bashrc".
9. Mausk ke direktori konfigurasi hadoop. Lakukan edit file "hadoop-env.sh" untuk menambahkan konfigurasi lokasi direktori Java.
10. Ubah konfigurasi pada file "core-site.xml"
11. Ubah konfigurasi pada file "hdfs-site.xml".
12. Ubah konfigurasi pada file "mapred-site.xml.template" lalu ubah nama file menjadi "mapred-site.xml".
13. Ubah konfigurasi pada file "yarn-site.xml"
14. Melakukan edit pada file "slaves" untuk menambahkan konfigurasi.

```
#localhost
192.168.1.11
192.168.1.12
192.168.1.13
```

15. Lakukanlah format pada namenode.

```
$ hadoop namenode –format
```

4.1.2 Instalasi Mahout

Instalasi Mahout hanya dilakukan pada Master Node. Berikut ini adalah tahapan instalasi Mahout.

1. Install aplikasi maven.

```
# apt-get install maven
```

2. Masuk ke direktori "/usr/local/", lalu buat folder mahout.

```
# mkdir mahout
```

3. Unduh mahout versi 0.10 di website "mahout.apache.org" dan ekstrak file tersebut ke direktori "/usr/local/mahout".

```
# cd /usr/local/mahout/
# tar –xvf mahout-distribution-0.10.0.tar.gz
```

4. Lakukan konfigurasi pada file "~/.bashrc" untuk menambahkan variabel Mahout.
5. Masuk ke direktori "/usr/local/mahout/, lalu lakukan instalasi mahout dan tes aplikasi mahout.

```
# mvn install -Dmaven.test.skip=true
```

4.1.3 Instalasi Dataset 20 Newsgroups

Dataset 20 Newsgroups merupakan dataset standar yang biasa digunakan untuk riset machine learning yang dibuat oleh 20 Usenet Newsgroups dari awal tahun 1990. Dataset ini akan dikonfigurasi menggunakan algoritma Naive Bayes yang merupakan algoritma dengan model statistik sederhana dan banyak digunakan di bidang akademisi. Pada algoritma ini pula dapat dipelajari tentang mengubah teks kedalam format vektor yang merupakan input untuk pengklasifikasian. Dataset 20 Newsgroups memiliki ukuran standar sebesar 36 MB. Pada pengujiannya, dataset ini akan dipecah menjadi 3 macam ukuran yaitu 10 MB, 21 MB dan 36 MB dengan cara mengurangi beberapa folder didalam dataset tersebut.

Berikut ini tahapan instalasi dataset 20 Newsgroups menggunakan algoritma Naive Bayes.

1. Langkah pertama, pastikan Hadoop sudah berjalan. Setelah itu unduh dataset 20 Newsgroups di halaman website "http://qwone.com/~jason/20Newsgroups/".
2. Buatlah direktori dan letakan dataset 20news-bydate.tar.gz pada direktori tersebut.

```
$ mkdir /tmp/mahout-work-hduser/
$ hadoop fs -mkdir /tmp/
```

3. Untuk melakukan proses pembelajaran pada mahout, Gunakan perintah berikut ini. Pilih "1. Cnaivebayes-MapReduce".

```
$ time $MAHOUT_HOME/examples/bin/classify-20newsgroups.sh
```

4. Proses pembelajaran pada suatu dataset akan memakan waktu cukup lama. Berikut ini tahapan proses pembelajaran yang dilakukan oleh Mahout untuk bisa mengklasifikasi suatu dataset.
    a. Membuat direktori pada HDFS untuk dataset dan semua folder input output.
    b. Melakukan ekstrak file 20news-bydate.tar.gz.
    c. Konversi dataset 20 newsgroups kedalam teks dan SequenceFile.
    d. Proses konversi dataset kedalam teks, VectorWritable dan SequenceFile.
    e. Menggabungkan dataset kedalam satu set training dan testing.
    f. Train klasifikasi.
    g. Test klasifikasi.
5. Bila hanya ingin melihat hasil tes dari hasil klasifikasi dataset 20 Newsgroups, maka dapat gunakan perintah berikut.

```
$ mahout testnb -i /tmp/mahout-work-hduser/20news-train-vectors -m /tmp/mahout-work-hduser/model -l /tmp/mahout-work-hduser/labelindex -ow -o /tmp/mahout-work-hduser/20news-testing
```

4.1.4 Instalasi Dataset Wikipedia XML

Wikipedia XML merupakan dataset yang berukuran standar 1 GB untuk di klasifikasi menggunakan algoritma Naive Bayes. Pada pengujiannya, dataset ini akan dipecah menjadi 3 macam ukuran yaitu 200 MB, 525 MB dan 1GB dengan cara melakukan edit script pada dataset tersebut. Berikut ini tahapan instalasi dataset Wikipedia XML.

1. Langkah pertama, pastikan hadoop sudah berjalan dan unduh dataset Wikipedia XML di halaman web "http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles10.xml-p000925001p001325000.bz2".
2. Buatlah direktori dan letakan dataset enwiki-latest-pages-articles10.xml.bz2 pada direktori tersebut.

```
$ mkdir /tmp/mahout-work-wiki/wikixml
$ tar –xvf enwiki-latest-pages-articles10.xml.bz2
$ hadoop fs -mkdir /tmp/
```

3. Untuk melakukan proses pembelajaran pada mahout, Gunakan perintah berikut ini. Pilih "1. CBayes".

```
$ time $MAHOUT_HOME/examples/bin/classify-wikipedia.sh
```

4. Proses klasifikasi dilakukan dengan metode klasifikasi Naive Bayes dan tahapannya sama dengan dataset 20 Newsgroups. Berikut ini tahapan proses pembelajaran

yang dilakukan oleh Mahout untuk bisa mengklasifikasi suatu dataset.

a. Membuat direktori pada HDFS untuk dataset dan semua folder input output.
b. Melakukan ekstrak file enwiki-latest-pages-articles10.xml.bz2
c. Konversi dataset Wikipedia XML kedalam teks dan SequenceFile.
d. Proses konversi dataset kedalam teks, VectorWritable dan SequenceFile.
e. Menggabungkan dataset kedalam satu set training dan testing.
f. Train klasifikasi.
g. Test klasifikasi.

5. Bila hanya ingin melihat hasil tes dari hasil klasifikasi dataset Wikipedia XML, maka dapat gunakan perintah berikut.

```
$ mahout testnb -i /tmp/mahout-work-wiki/testing
-m /tmp/mahout-work-wiki/model -l /tmp/mahout-
work-wiki/labelindex -ow -o /tmp/mahout-work-
wiki/output -c -seq
```

4.1.5 Instalasi Zabbix

Berikut ini adalah tahapan instalasi Zabbix.

1. Install beberapa paket aplikasi terlebih dahulu.

```
# apt-get update
# apt-get install apache2
# apt-get install mysql-server
# apt-get install php5 php5-cli php5-common
php5-mysql
```

2. Lakukan edit pada file "/etc/php5/apache2/php.ini".

```
date.timezone = ' Asia/Jakarta '
```

3. Download Zabbix dengan perintah berikut.

```
# wget
http://repo.zabbix.com/zabbix/2.2/ubuntu/pool/mai
n/z/zabbix-release/zabbix-release_2.2-
1+trusty_all.deb
```

4. Lakukan instalasi paket Zabbix. Ikuti instalasi nya.

```
# apt-get install zabbix-server-mysql zabbix-
frontend-php
```

5. Lakukan restart layanan apache dengan perintah.

```
# service apache2 restart
```

6. Lakukan restart layanan zabbix server dengan perintah.

```
# service zabbix-server restart
```

7. Buka browser dan lakukan instalasi zabbix via browser anda dengan ketikan "localhost/zabbix" pada pencarian.
8. Lakukan instalasi zabbix agent dengan perintah.

```
# apt-get install zabbix-agent
```

9. Lakukan konfigurasi zabbix agent pada file "/etc/zabbix/zabbix_agentd.conf". Edit pada bagian ini.

```
Server=192.168.1.11
Hostname=Master
ListenIP=192.168.1.11 #PC yang di monitoring
```

10. Lakukan restart zabbix agent dengan perintah.

```
# /etc/init.d/zabbix-agent restart
```

4.2 Pengujian

Proses eksekusi dibagi menjadi 1 node, 2 Node dan 3 node untuk membandingkan kecepatan eksekusi dengan perintah "time" di CLI. Adapun istilah-istilah dari perintah "time" pada CLI sebagai berikut:

a. Real yaitu lamanya waktu saat proses eksekusi berlangsung dalam hitungan detik.
b. User yaitu total jumlah waktu dari proses CPU yang digunakan secara langsung.
c. Sys yaitu total jumlah waktu dari proses CPU yang digunakan oleh sistem operasi.

Pengujian dilakukan dengan metode klasifikasi Naive Bayes pada suatu dataset. Klasifikasi merupakan proses pengambilan keputusan yang dipelajari dari data tersebut dan menyamakan keputusan secara otomatis. Langkah-langkah yang berbeda yang disebutkan dalam hasil untuk mengevaluasi model yang disebutkan di sini:

a. Kappa statistics yaitu metrik yang membandingkan akurasi diamati dengan akurasi yang diharapkan.
b. Reability yaitu sejauh mana pengukuran memberikan hasil yang konsisten.
c. Precision yaitu sebagian kecil dari contoh diambil yang relevan.
d. Recall yaitu fraksi contoh yang relevan yang diambil.
e. F1 measure yaitu ukuran yang menggabungkan precision dan recall.

Adapula monitoring yang dilakukan untuk mengetahui kinerja perangkat keras laptop dan trafik jaringannya. Gambarannya hanya berupa proses load CPU dan grafik jaringan eth0 yang akan update per-60 detik.

4.2.1 Pengujian Hadoop

Berikut ini tahapan untuk mengoperasikan Hadoop.

1. Setelah selesai format namenode, jalankan hadoop dengan perintah berikut.

```
$ start-all.sh
```

2. Cek apakah hadoop sudah benar-benar berjalan dengan perintah.

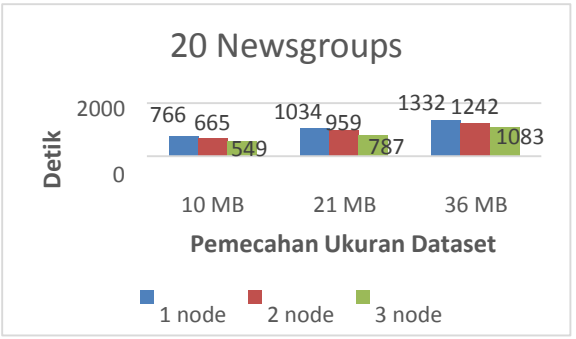
```
$ jps
$ ssh slave1
$ ssh slave2
```

3. Untuk mengecek node yang aktif, buka web browser dan masuk ke "master:8088/cluster/nodes".
4. Untuk mengetahui datanode yang aktif dan sistem penyimpanan HDFS, Buka web browser, lalu masuk ke "master:50070".
5. Untuk menghentikan Hadoop cukup lakukan perintah berikut.

```
$ stop-all.sh
```

4.2.2 Pengujian Eksekusi Dataset 20 Newsgroups

Pengujian eksekusi dataset 20 Newsgroups menggunakan metode klasifikasi data dengan algoritma Naive Bayes yang akan memunculkan confusion matrix sebagai hasilnya. Di bawah confusion matrix muncul statistik, yang menunjukkan tingkat akurasi dan proporsi jumlah prediksi yang diklasifikasikan dengan benar serta perbandingan waktu antara 1 node, 2 node dan 3 node.

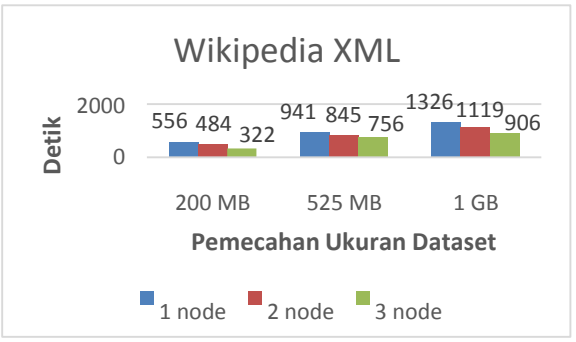


Gambar 4 Hasil Perbandingan Kluster Komputer 20 Newsgroups

Grafik diatas menggambarkan hasil perbandingan lamanya eksekusi antara 1 node, 2 node dan 3 node sesuai dengan pemecahan 3 macam ukuran dataset 20 Newsgroups.

4.2.3 Pengujian Eksekusi Dataset Wikipedia XML

Pengujian dilakukan dengan metode klasifikasi pada dataset Wikipedia XML menggunakan algoritma Naive Bayes juga. Hasil nya akan di tampilkan dalam bentuk confusion matrix.



Gambar 5 Hasil Perbandingan Kluster Komputer Wikipedia XML

Grafik diatas menggambarkan hasil perbandingan lamanya eksekusi antara 1 node, 2 node dan 3 node sesuai dengan pemecahan 3 macam ukuran dataset Wikipedia XML.

4.2.4 Masalah Saat Pengujian

Saat awal pengujian proyek akhir ini, terdapat beberapa masalah yang dihadapi secara teknis. Contohnya seperti tidak kompatibel nya versi mahout dan hadoop yang digunakan. Efeknya akan ada log error yang terjadi saat eksekusi algoritma machine learning berlangsung yang berarti proses MapReduce gagal dijalankan.

Masalah ini bisa diatasi dengan mengganti versi mahout ke 0.10. Karena Hadoop versi 2.x.x keatas hanya bisa bekerja dengan mahout 0.10.

Untuk menjalankan eksekusi algoritma machine learning, diharapkan untuk tidak menjalankan proses lain agar lamanya waktu eksekusi algoritma machine learning sesuai dan tidak ada overhead yang terjadi secara berlebihan yang akan memakan banyak kinerja dari perangkat keras komputer.

Walaupun ukuran standar dataset 20 Newsgroups dan Wikipedia XML berbeda jauh, lamanya waktu eksekusi algoritma machine learning tidak jauh berbeda karena algoritma machine learning tersebut tidak melakukan proses pembelajaran dataset secara keseluruhan atau sesuai banyaknya ukuran dari dataset tetapi hanya mempelajari konten tertentu yang tercantum pada script dataset tersebut.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan yang bisa di dapatkan dari Proyek Akhir ini, antara lain:

1. Hasil eksekusi algoritma machine learning menampilkan confusion matrix sebagai proses klasifikasi nya. Pada dataset 20 Newsgroups dengan ukuran file 36 MB, terlihat rata-rata lamanya waktu eksekusi dengan 1 node mencapai 22 menit 12 detik, 2 node mencapai 20 menit 42 detik dan 3 node hanya mencapai 18 menit 3 detik. Sedangkan pada dataset Wikipedia XML dengan ukuran file 1 GB, terlihat rata-rata lamanya waktu eksekusi dengan 1 node mencapai 22 menit 6 detik, 2 node mencapai 18 menit 39 detik dan 3 node mencapai 15 menit 6 detik. Hasil pengujian menunjukan, ada perbandingan waktu yang cukup signifikan mencapai rata-rata 2 sampai 7 menit lebih cepat bila node ditambahkan lebih banyak, sehingga 3 node lebih cepat dari 2 node dan 1 node.
2. Monitoring berhasil menampilkan grafik CPU load dan trafik jaringan port ethernet pada masing-masing PC yang merupakan hasil dari kinerja perangkat keras komputer dan jaringan yang digunakan saat eksekusi algoritma machine learning berlangsung.

5.2 Saran

Untuk mengembangkan kluster komputer maka dibutuhkan ukuran dataset yang lebih besar serta penambahan node dengan jumlah yang lebih banyak lagi untuk menambah performansi komputer dalam pengolahan data. Monitoring dilakukan dengan aplikasi yang lebih baik lagi agar hasil monitoring terlihat lebih spesifik dan akurat secara waktu eksekusi maupun kinerja dari komponen perangkat keras yang digunakan.

6. Daftar Pustaka

[1] S. Karanth, Mastering Hadoop, Birmingham: Packt Publishing, 2014.

[2] T. Gunarathne, Hadoop MapReduce v2 Cookbook Second, Birmingham: Packt Publishing, 2015.

[3] J. Withanawasam, Apache Mahout Essentials, Birmingham: Packt Publishing, 2015.

[4] C. Tiwary, Learning Apache Mahout, Birmingham: Packt Publishing, 2015.

[5] A. Gupta, Learning Apache Mahout Classification, Birmingham: Packt Publishing, 2015.

[6] R. Olups, Zabbix 1.8 Network Monitoring, Birmingham: Packt Publishing, 2010.