



SERTIFIKAT

Diberikan Kepada

Ade Jamal

Atas peran sertanya sebagai

Pemakalah

Dalam

Seminar Nasional Teknologi Informasi XI
Fakultas Teknologi Informasi
Universitas Tarumanagara

Jakarta

13 Desember 2014

Dekan

Prof. Dr. Dyah Erny Herwindiaty, M.Si.

Ketua SNTI XI

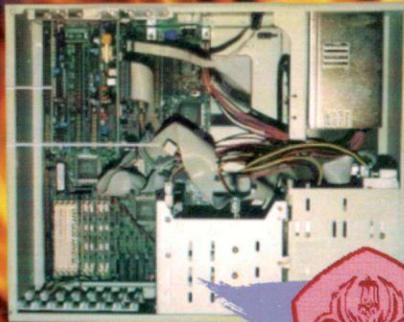
Dra. Ery Dewayani, MMSI.

Prosiding

ISSN : 1829-9156
Vol. 11 No. 1 Tahun 2014

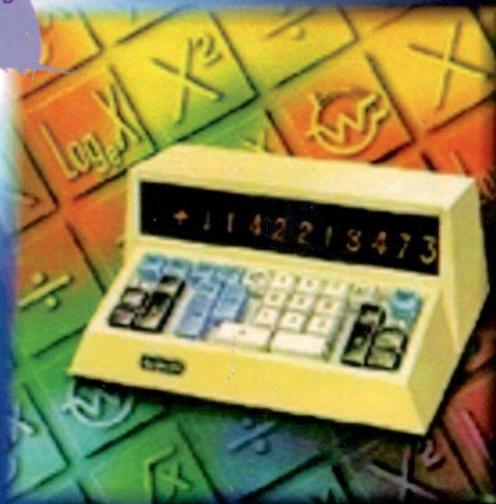
SNTI 2014

13 Desember 2014



SNTI

Seminar Nasional Teknologi Informasi



**Fakultas Teknologi Informasi
Universitas Tarumanagara
Jakarta**



**KATA SAMBUTAN
KETUA PANITIA SNTI 2014**

Kami mengucapkan selamat pagi dan selamat datang di Seminar Nasional Teknologi Informasi (SNTI) ke XI. Puji syukur kepada Tuhan Yang Maha Esa atas rahmat dan karunia Nya, sehingga dapat terlaksananya seminar pada hari ini dan kita dapat berkumpul bersama-sama dalam keadaan sehat.

SNTI merupakan kegiatan tahunan Fakultas Teknologi Universitas Tarumanagara (FTI Untar) yang diselenggarakan untuk pertama kalinya sejak tahun 2004. Seminar kali ini mengusung Tema: “*Enterprise Information Systems*”, dengan mengundang 2 orang *keynote speaker* : Dr. Teoh Ai Ping dari Graduate School Of Business Universiti Sains Malaysia dan Mohamad Rizal Arrayadi, CPIM,CSCP selaku Principal Solution Consultant Oracle Corporation. Selanjutnya Seminar ini dilanjutkan dengan presentasi hasil penelitian dari para peneliti akademisi maupun profesional yang terbagi dalam 3 ruang sesi paralel.

Kami ucapkan terima kasih kepada para pemakalah yang telah berpartisipasi mengirimkan makalah pada seminar ini, dengan jumlah 58 makalah dari berbagai kota di Indonesia, dan setelah dilakukan *review* sejumlah 42 makalah yang dapat kami terima untuk didokumentasikan dalam Prosiding SNTI XI 2014 dengan ISSN: 1829-9156 dan akan dipresentasikan pada seminar hari ini.

Kami ucapkan terima kasih kepada pimpinan FTI Untar, Ibu Desi Arisandi S. Kom., MTI sebagai Wakil Ketua panitia dan seluruh anggota panitia SNTI XI tahun 2014 yang telah bekerja keras atas terselenggaranya seminar ini. Tidak lupa kami ucapkan pula terima kasih kepada Asosiasi Sistem Informasi Indonesia atas dukungannya sebagai sponsor.

Semoga seminar ini dapat bermanfaat bagi kita semua.

Terima Kasih

Jakarta, 13 Desember 2014

Ketua SNTI 2014

(Dra. Ery Dewayani, MMSI)



Sambutan Dekan
Fakultas Teknologi Informasi
Universitas Tarumanagara Pada Seminar
Teknologi Informasi ke XI,
Sabtu, 13 Desember 2014

Yang terhormat,

*Bapak Ketua dan Pengurus Yayasan Tarumanagara
Bapak Rektor dan Wakil Rektor Universitas Tarumanagara
Bapak/Ibu Pembicara Utama dan Pembicara Undangan SNTI 2014
Bapak/Ibu Pemakalah SNTI 2014
Bapak/Ibu Dosen dan Karyawan FTI-Untar
Para Mahasiswa FTI-Untar yang kami banggakan dan cintai,
serta Undangan lainnya.*

Puji syukur kami ucapkan pada Tuhan Yang Maha Kuasa, atas karuniaNYA pada hari yang berbahagia ini Fakultas Teknologi Informasi-Universitas Tarumanagara (FTI-Untar) dengan bangga menyelenggarakan Seminar Teknologi Informasi (SNTI) ke XI yang bertepatan dengan rangkaian upacara Hari Ulang Tahun Untar ke- 55 dan Dies Natalis FTI-Untar ke-13. Bertemakan '*Enterprise Information System*', SNTI 2014 mengundang seorang pembicara utama **Dr. Teoh Ai Ping**, beliau adalah seorang dosen senior dari **Graduate School of Business, Universiti Sains Malaysia** sekaligus seorang pakar terkenal dalam bidang teknologi informasi.

Topik dan kajian ilmu yang diangkat pada perjalanan SNTI hingga sebelas tahun ini cukup beragam, mencakup topik-topik yang menarik dan sedang dikembangkan oleh para peneliti dalam bidang teknologi informasi, antara lain: *intelligence system and application, networking and distributed system, applied computational science, game development, database* serta *information System*.

Pada kesempatan ini perkenankan saya mengucapkan terimakasih kepada ketua dan sekretaris SNTI 2014, ibu Ery Dewayani, MMSI dan ibu Desy Arisandi, MTI atas waktu dan upaya kerasnya agar SNTI 2014 terlaksana, tak lupa juga saya ucapkan terimakasih pada seluruh dosen, karyawan dan mahasiswa panitia SNTI 2014 sehingga seminar ini terlaksana dengan baik.

Semoga seminar ini bermanfaat untuk kita semua, khususnya peserta SNTI 2014. Selamat mengikuti seminar, terimakasih atas kehadirannya dan sampai jumpa di acara SNTI 2015 pada tahun mendatang.

Jakarta, 13 Desember 2014
Dekan Fakultas Teknologi Informasi

Prof. Dr. Dyah Erny Herwindiati

SUSUNAN PANITIA SNTI 2014

Program Committee

1. Prof. Dr. Ir. Aniati Murni A., M.Sc (UI)
2. Dr. Eng. Wisnu Jatmiko (UI)
3. M. Ivan Fanany, Ph.D (UI)
4. Agus Hardjoko, M.Sc., Ph.D (UGM)
5. Prof. Jazi Eko Istiyanto, M.Sc., Ph.D (UGM)
6. Retantyo Wardoyo, M.Sc., Ph.D (UGM)
7. Prof. Dr. Budi Nurani Ruchyana (UNPAD)
8. Dr. Ir. Agus Buono, M.Si., M.Kom (IPB)
9. Dr. Ir. Aji Hamim Wigena, M.Sc (IPB)
10. Prof. Dr. Ir. Eko Sedyono, M.Kom (UKSW)
11. Dr. Teoh Ai Ping (USM Malaysia)
12. Prof. Dr. Ir. Dali S. Naga, MMSI (Untar)
13. Ir. Jap Tji Beng, MMSI., Ph.D (Untar)
14. Dr. Ir. Dyah Erni Herwindiati, M.Si (Untar)
15. Dr. Sani M. Isa (Untar)
16. Lina, ST., M.Kom., Ph.D (Untar)

Organizing Committee (Universitas Tarumanagara)

1. Ery Dewayani, MMSI
2. Desi Arisandi, S.Kom., MTI
3. Agus Budi Dharmawan, M.T., M.Sc
4. Bagus Mulyawan, S.Kom., MM
5. Dra. Chairisni Lubis, M.Kom
6. Dedi Trisnawarman, M.Kom
7. Ir. Jeanny Pragantha, M.Sc.
8. Lely Hiryanto, ST., M.Sc
9. Tony, M.Kom
10. Viny Christanti, M.Kom
11. Wasino, S.Kom., M.Kom
12. Zyad Rusdi, ST, M.Kom

DAFTAR MAKALAH YANG DITERIMA SNTI 2014*

Kata Sambutan Ketua Pelaksana			ii
Kata Sambutan Dekan Fakultas Teknologi Informasi			iii
Susunan Panitia			iv
Daftar Isi			v
Keynote Speaker			ix
<u>A. Algorithm, Intelligent Systems, Computational</u>			
A1	Penerapan Algoritma C4.5 Terhadap Pemilihan Mitra Kerja Penyedia Jasa Transportasi di Jakarta	Harry Dhika Tri Yani Akhirina Surajiyo	1
A2	Perancangan Model Simulator Rantai Pasok	Erna Mulyati, Dini Hamidin, Mohamad Nurkamal Fauzan	9
A3	Analisa Efektifitas Jalur Terpendek Menggunakan Algoritma Semut Dan Pendekatan Leadtime	Erlin Windia Ambarsari, Dwi Marlina, Arif Susanto	15
A4	Penghitungan Indeks Prestasi Kumulatif Mahasiswa Dengan Pendekatan Metode Fuzzy Takagi-Sugeno	Zaenal Abidin, Y. Agus Nurhuda, Sabam Parjuangan	21
A5	Perbandingan Ciri Tekstur GLCM Citra Radar Pita P Dengan Citra Radar Pita L Dari Wilayah Laut : Kasus Pengamatan Sudut 0 Derajat	Hasan Basri, Prastika Indriyanti, Harwikarya	28
A6	Pererapan Algoritme Viterbi Pada Hidden Markov Model (HMM) Untuk Prediksi Struktur Sekunder Protein	Dian Puspitasari, Toto Haryanto	34
A7	<i>Internet Of Things</i> Pada Keamanan Rumah Berbasis Arduino Dengan Labview	Asep Andang, Nurul Hiron, Nundang Busaeri, Iwan Septiawan	41
A8	Strategi Memperbaiki Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik	Herry Sujaini , Arif Bijaksana P.N	47

A9	Aplikasi Sistem Pakar Berbasis Android untuk Diagnosis dan Resiko Obesitas	Puspa Setia Pratiwi,	52
A10	Algoritma Baru dalam Menyelesaikan Vehicle Routing Problem With Time Window (Vrptw)	Diska Permana Sari Sarngadi Palgunadi , Elkagianda Listya Arma Putri	57
A11	Implementasi <i>Process Mining Multi Choice</i> Dan <i>Deferred Choice Workflow</i> Dengan <i>Petri Net</i>	Sarngadi Palgunadi,	66
A12	Penggunaan Augmented Reality Berbasis Mobile Untuk Pembelajaran Perangkat Keras Komputer	Arofiah Hidayati	72
A13	Membangun Padanan Kata Dengan Korpus Paralel	Adam Arif Budiman, Timor Setyaningsih	76
A14	Penghitungan Jumlah Kendaraan Pada Citra Dari Kamera Pemantau Dengan Metode Fuzzy C Means	Jeanny Pragantha, Viny C. Mawardi, Novita	82
A15	Model Keputusan Penilaian Kinerja Kompetensi Guru Sd Menggunakan Metode Ahp	Joko Lianto Buliali , Chastine Fatichah, Ahmad Saikhu , SilvesterTena	87
A16	Klasifikasi Fase Pertumbuhan Padi Dengan <i>Bee Algorithm</i> Dan K-Nearest Neighbor	Sugesti, Heru Soetanto Putra	93
A17	Prediksi Panen Padi Menggunakan Analisis Robust Regression	F.X. Adrian, Sendy M. Wiguna, Dyah E. Herwindiati, Sidik Muljono	98
A18	Aplikasi Expert Advisor D3fx Menggunakan Frame Work Meta Trader 4 Dengan Metode Scalping	Felix Indra Kurniadi, Sani M. Isa, Sidik Muljono	106
		Iwan Rijayana,	
		Arief Hertadi Rustam	

B. Information Systems

B1	Perancangan Sistem Informasi Jasa Pengiriman Barang Berbasis Web	Harry Dhika	1
		Aswin fitriansyah	
		Lukman	
B2	Peta Interaktif Berbasis Web Untuk Visualisasi Informasi Geografis Industri Kreatif	Aviarini Indrati,	5
		Murniyati Murniyati, Rooswhan Budhi Utomo	
B3	Perancangan Sistem Informasi Pengukuran Kinerja Kesehatan BPR Jabotabek Berbasis Web	Imelda Sinaga,	11
		Yulistia Devi, Ria Arisandi	

B4	Aplikasi Sistem Informasi Untuk Memprediksi Financial Distress Berbasis Web	Yulistia Devi, Imelda Sinaga, Aprilia Maya Wulandari	22
B5	Analisis Kinerja Rad Studio Xe3 Untuk Membangun Sistem Berbasis Android	Agus Mulyanto, Antonius Junaidi	30
B6	Perencanaan Sistem Informasi Menggunakan Framework Sc2m-Sme	Dini Hamidin, Akhmad Yunani, Azizah Zakiah	35
B7	Service Patterns In The Process Of Service-Oriented Development	Suryo Bramasto Putri Pratiwi Nur Hasan Ervian Gelar Winahyu	41
B8	Purwarupa Pemetaan Lingkungan Kampus Universitas Padjadjaran Berbasis Augmented Reality	Erick Paulus, Stanley P. Dewanto, Juli Rejito	49
B9	Pengembangan E-Marketing Jasa Paket Wisata Pada Cv Omnino Nusantara	Arum Rakhmasari , Rusliyawati, Nina Suryani	54
B10	Penerapan Change Management Pada Pembangunan Sistem Informasi Akademik Itenas	Yusup Miftahudin	64
B11	Analisis Dan Implementasi Restful Webservice Untuk Mobile Environment Pada GIS Lahan Pangan Kabupaten Minahasa Tenggara	Angga Purwoko, Eko Sedyono, Adi Setiawan	70
B12	Rancang Bangun Aplikasi Lelang Barang Berbasis Web Pada Lingkungan Kampus	Nurul Hiron, Ruly Rasid Mujadid	75
B13	Pengembangan Aplikasi Penjualan Mobil Nissan Pada Pt Wahana Prima Berbasis Web	Himawan, Kusendang Fitriyanti, Mery Mayang Sari, Dewi	82
B14	Perancangan Sistem Informasi Desa (Sim-Desa) Sebagai Aplikasi Manajemen Dana Pembangunan Desa Berbasis Web	Nur Widiyasono, Nurul Hiron, Firmansyah , Aldy Putra Agus Salim,	87
B15	Elisitasi Kebutuhan Kms Dengan Menggunakan Metode Gap Analysis Pengetahuan	Teddy Siswanto , Juita Tuhfati R	93

B16	Implementasi Kerangka Kerja I.T.I.L (Service Support) Pada Pengelolaan Infrastruktur	Mira Musrini Barmawi, Asep Nana Hermana, Chashif Syadzali	100
B17	Sistem Informasi Geografis Lokasi Lembaga Pendidikan di Kota Metro	Sabam Parjuangan, Ridho Nugroho	108
B18	Analisis Dan Perancangan Sistem Informasi Pajak Perhotelan Pada Dinas Pendapatan Daerah Kabupaten Manokwari	Ismail Ibrahim, Suyoto, Sapti	113
B19	Penerapan Teknologi GIS dan Analisis Spasial Pada Pengembangan Klaster Industri Eko-Wisata	I Ketut Gunarta, Setyo Nugoroho, Edwin T. Hermawan, Gede Ardi Adzandi, Farid Kurniawan, Muhammad Reza Suerman, Satria Oktafanus Sarwoko	118

C. Network, Distributed, Instrumentations

C1	Implementasi Arsitektur Pertukaran Data Perpustakaan Daerah Dalam Membangun Gorontalo Library Network	Moh.Hidayat Koniyo, Arip Mulyanto, Rochmad M. T. Yassin	1
C2	Performance Monitoring Server System Berbasis Web, Email dan Sms : STUDI KASUS PT. BONET UTAMA	Prihandini Astuti, Wahyu Pramusinto, Ferdiansyah	7
C3	Pencarian <i>Sequence</i> DNA Genbank Menggunakan <i>Hadoop Distributed Filesystem</i>	Ade Jamal, Winangsari Pradani, Nida'ul Hasanati, Muhammad Sukaya Nugraha	13
C4	Perancangan Paralel Prosesor MD5 CRACKING	Andi Yusuf, Claudia Dwi Amanda	19
C5	APLIKASI REMOTE KONTROL MP3 PLAYER BERBASIS HANDPHONE	Jos Timanta Tarigan, Ertina Sabarita Barus	27

PENCARIAN SEQUENCE DNA GENBANK MENGGUNAKAN HADOOP DISTRIBUTED FILESYSTEM

Ade Jamal¹⁾, Winangsari Pradani²⁾, Nida'ul Hasanati³⁾, Muhammad Sukaya Nugraha⁴⁾

¹⁾²⁾³⁾⁴⁾Program Studi Teknik Informatika, Universitas Al-Azhar Indonesia
Kampus UAI, Jl. Sisingamangaraja, Jakarta Selatan
email : adja@uai.ac.id

ABSTRACT

Telah dilakukan penelitian tentang pengolahan terdistribusi data genbank menggunakan Hadoop Distributed Filesystem (HDFS) dengan tujuan mengetahui efektifitas dan efisiensi pengolahan data Genbank khususnya pada pencarian sequens dengan data masukan yang berukuran besar. Penelitian dilakukan di Laboratorium Jaringan Universitas Al-Azhar Indonesia dengan menggunakan 6 komputer dan satu server dimana dalam Hadoop menjadi 7 node dengan rincian 1 namenode, 7 datanode. Pengukuran waktu proses pencarian sequens data genbank menggunakan jumlah node yang berbeda dari 1 sampai 7 node menunjukkan percepatan (*SpeedUp*) yang cukup baik. Tiga skenario eksperimen dengan banyak luaran yang berbeda, yaitu luaran sedikit, sedang dan banyak untuk memperlihatkan efek besarnya penulisan luaran data terhadap waktu keseluruhan. Evaluasi menunjukkan bahwa proses pencarian sequens data genbank menggunakan HDFS dengan 7 node adalah 4 kali lebih cepat dibandingkan dengan menggunakan 1 node. Dan besarnya luaran data berpengaruh sangat kecil terhadap proses pencarian secara keseluruhannya.

Key words

Sistem Berkas Terdistribusi, HADOOP, sequence, pencarian.

1. Pendahuluan

National Center for Biotechnology Information (NCBI) merupakan institusi pemerintah yang menyediakan sumber informasi mengenai biologi molekuler. NCBI memiliki salah satu *database* yang dapat diakses oleh publik berhubungan dengan *database* sequens DNA lebih dari 100.000 organisme yang berbeda, yaitu Genbank. Untuk melakukan pencarian sequens data Genbank diperlukan *web* NCBI dan aplikasi pendukung lain yang harus digunakan. Aplikasi tersebut merupakan aplikasi tersendiri

dan tidak disediakan oleh NCBI. Namun untuk melakukan ini merupakan hal yang merepotkan karena membutuhkan koneksi internet yang cepat dan stabil, serta tidak mudah bagi yang belum terbiasa menggunakannya. Solusinya adalah memindahkan *database* Genbank ke dalam *Local Filesystem* agar tidak memerlukan koneksi internet lagi. Dibutuhkan spesifikasi komputer yang tinggi untuk dapat melakukan pencarian sequens data Genbank yang sedemikian besar. Hal ini tentu merugikan dari segi biaya karena komputer dengan spesifikasi tinggi dan terbaik mempunyai harga yang mahal.

Dibutuhkan suatu teknologi dalam mengakomodir kebutuhan ini dengan biaya yang minimal namun dengan hasil optimal. Maka solusi terbaik adalah menggunakan teknologi *distributed computing* dengan sistem *computer cluster*. *Distributed computing* merupakan suatu teknologi yang memecahkan problem berukuran besar ke dalam proses-proses kecil yang akan dilakukan pengolahannya oleh banyak komputer. Kemudian komputasi dari banyak proses kecil itu diselesaikan secara simultan dan apabila sudah didapatkan solusi-solusi kecil maka disatukan kembali dalam satu solusi yang besar dan terintegrasi.

Hadoop merupakan *framework software* berbasis *Java* dan *open source* yang berfungsi untuk mengolah data yang sangat besar secara terdistribusi dan berjalan di atas *cluster* yang terdiri dari beberapa komputer yang saling terhubung. *Hadoop* dapat mengolah data dalam jumlah yang sangat besar hingga berukuran *petabyte* dan dijalankan di atas ribuan komputer. Dengan menggunakan *Hadoop* melalui HDFS, maka pencarian sequens DNA dari data Genbank dapat dilakukan tanpa menggunakan koneksi internet yang cepat dan stabil serta dapat menggunakan beberapa komputer personal yang terdapat pada Laboratorium Jaringan Universitas Al Azhar Indonesia.

2. Hadoop

Hadoop adalah *framework software* berbasis Java yang berfungsi untuk mengolah data berukuran besar secara terdistribusi dan berjalan di atas *cluster* komputer yang saling terhubung serta jumlah komputer dapat ditambahkan tanpa batasan^[1]. *Hadoop* memiliki 2 inti, yaitu^[2]:

1. HDFS (Hadoop Distributed Filesystem)
2. MapReduce

A. HDFS (Hadoop Distributed File System)

HDFS merupakan *filesystem* yang menyimpan data sangat besar, berjalan di atas *cluster* dengan spesifikasi komputer pada umumnya dan disimpan secara terdistribusi di banyak komputer dalam satu *cluster* yang saling berhubungan. Terdapat tiga keunggulan dalam HDFS, yaitu:

- File yang sangat besar; HDFS dapat digunakan untuk data sebesar ratusan megabyte, gigabyte, terabyte, bahkan saat ini sudah ada *cluster Hadoop* yang menggunakan data sampai *petabyte*.
- *Streaming data access*; HDFS dibangun dengan pola *streaming data access*, yaitu data cukup ditulis satu kali ke dalam HDFS, namun dapat dibaca atau diproses lebih dari sekali.
- Penggunaan *hardware* biasa; Tidak dibutuhkan *hardware* dengan spesifikasi khusus dalam menggunakan dan mengimplementasikan HDFS.

HDFS juga memiliki *block size* sebesar 64 MB. *Block size* pada HDFS sebesar itu karena untuk meminimalisir waktu membaca suatu data. Dengan *block size* yang cukup besar, waktu yang dibutuhkan untuk transfer data dari *disk drive* akan lebih besar secara signifikan dibandingkan dengan *single disk drive*. Ada keuntungan apabila menggunakan HDFS yang memiliki *block size* besar, yaitu file yang disimpan dalam HDFS dapat berukuran lebih besar dibandingkan dengan satu *hard disk drive*. Misalnya, file berukuran 10TB tidak akan bisa disimpan kedalam *hard disk drive* berukuran 3TB yang merupakan *hard disk drive* dengan kapasitas terbesar saat ini. Namun dengan HDFS, penyimpanan file sebesar itu dapat dilakukan^[2].

HDFS *cluster* memiliki dua tipe *node* yang beroperasi, yaitu *namenode* (yang bekerja sebagai *master*) dan *datanode* (yang bekerja sebagai *slave*)^[3]. *Namenode* bertugas untuk mengatur *filesystem namespace*. *Filesystem namespace* memelihara *filesystem tree* dan *metadata* untuk semua file dan direktori dalam *tree* yang ada dalam *cluster*. Informasi ini disimpan dalam *disk* dengan format dua file, yaitu *namespace image* dan *edit log*^[2].

Filesystem image adalah file yang berisi tentang informasi seluruh direktori dan file yang terdapat dalam HDFS. Informasi yang ada meliputi level replikasi,

modifikasi dan *access time*, *access permission*, *block size*, dan apa yang terdapat di dalam *block* tersebut. *Edit log* adalah hasil sinkronisasi setelah setiap *write* sebelum kode sukses dikembalikan ke *datanode*. Saat *namenode* melakukan *write* dengan multi direktori, maka *write* harus disinkronisasi sebelum kode sukses dikembalikan ke *datanode*. Ini untuk menjaga dan meyakinkan bahwa tidak ada operasi yang hilang karena komputer *failure*^[4].

Datanode adalah pekerja atau *slave* dalam HDFS. *Datanode* menyimpan dan menerima *block* ketika *datanode* diberitahu oleh *datanode* lain atau *namenode* dan *datanode* melaporkan kembali ke *namenode* secara periodik dilengkapi dengan daftar *block* apa saja yang disimpan dalam *datanode* tersebut^[3].

Tanpa *namenode*, maka HDFS tidak bisa digunakan sama sekali. Apabila *Hadoop Cluster* telah dijalankan namun *namenode* terhapus, semua data yang

tersimpan dalam HDFS akan hilang karena tidak ada cara lain untuk mengetahui keberadaan data dan merekonstruksi file yang terbagi di *block - block datanode* dalam HDFS selain oleh *namenode*. Itu merupakan alasan betapa pentingnya *namenode* sampai tidak terjadi kegagalan dan *Hadoop* mempunyai solusi untuk masalah ini. Pertama adalah melakukan *backup file* yang terdapat dalam HDFS secara periodik. *Hadoop* bisa dikonfigurasi untuk melakukan *backup* ke dalam *Local Filesystem* dari HDFS. Kedua adalah dengan menggunakan *secondary namenode*. Aturannya adalah melakukan *merge namespace* secara periodik berdasarkan *edit log* yang tersimpan dalam *namenode*. Biasanya *secondary namenode* disimpan ditempat yang berbeda dengan *namenode* namun memiliki spesifikasi komputer yang sama untuk dapat melakukan *merge* yang sempurna. *Secondary namenode* menyalin *namespace image*, agar bisa digunakan seketika ketika *namenode* gagal melakukan tugasnya. Namun *secondary namenode* tidak bisa menggantikan posisi *primary namenode* karena *secondary namenode* hanya menyimpan *metadata* terbaru agar bisa digunakan oleh *namenode* lain untuk melanjutkan *job* baru berdasarkan *edit log* terakhir menggantikan *namenode* yang *failed*^[2].

B. MapReduce

MapReduce adalah model pemrograman yang diperkenalkan oleh Google dan digunakan untuk mendukung *distributed computing* yang dijalankan di atas *data set* yang sangat besar dan dijalankan secara simultan dibanyak komputer. Model pemrograman ini terinspirasi oleh konsep fungsi *map* dan *reduce* yang biasa digunakan di *functional programming*^[5].

1. Proses "Map" : *Master node* menerima *input*, kemudian *input* tersebut dipecah menjadi beberapa sub problem yang kemudian didistribusikan ke *worker nodes*.

Worker nodes ini akan memproses sub problem yang diterimanya untuk kemudian apabila problem tersebut sudah diselesaikan, maka akan dikembalikan ke *masternode*.

2. Proses "Reduce" : *Masternode* menerima jawaban dari semua sub problem dari banyak data nodes, menggabungkan jawaban-jawaban tersebut menjadi satu jawaban besar untuk mendapatkan penyelesaian dari permasalahan utama.

Keuntungan dari *MapReduce* ini adalah proses *map* and *reduce* dijalankan secara terdistribusi. Dalam setiap proses *mapping* bersifat independen sehingga proses dapat dijalankan secara simultan dan paralel. Demikian pula dengan proses *reducer* dapat dilakukan secara paralel di waktu yang sama, selama *output* dari operasi *mapping* mengirimkan *key value* yang sesuai dengan proses *reducer*-nya. Proses *MapReduce* ini dapat diaplikasikan di *cluster server* yang jumlahnya sangat banyak sehingga dapat mengolah data dalam jumlah *petabyte* hanya dalam waktu beberapa jam^[2].

Didalam *Hadoop*, *MapReduce engine* ini terdiri dari satu *jobtracker* dan satu/banyak *tasktracker*. *Jobtracker* adalah *server* penerima *job* dari *client*, sekaligus akan mendistribusikan *job* tersebut ke *tasktracker* yang akan mengerjakan *sub job* sesuai yang diperintahkan *jobtracker*. Strategi ini akan mendekatkan pengolahan data dengan datanya sendiri, sehingga ini akan sangat signifikan mempercepat proses pengolahan data.

3. Sequence DNA pada HADOOP

Data sekuens DNA yang digunakan dalam penelitian ini diunduh dari *National Center for Biotechnology Information* (NCBI) yang disimpan dalam sebuah basisdata yang dikenal dengan nama GenBank^[6,7,8]. Pada penelitian sebelumnya, data sekuens DNA dan juga sekuens Protein dari sumber yang sama disimpan dalam satu server tunggal^[9]. Ketika upaya pencarian dilakukan untuk melakukan proses pencarian pada *sequence alignment* pada data yang lengkap ditemukan kendala waktu yang sangat lama^[10]. Dengan menggunakan *Hadoop* maka proses pencarian sekuens data dari Genbank diharapkan dapat dilakukan dengan cepat dan efisien dibandingkan dengan proses pencarian menggunakan *single processing*. *Hadoop* memiliki 2 inti, yaitu *MapReduce* dan HDFS. Data Genbank yang telah diunduh dari NCBI, diunggah ke HDFS (*Hadoop Distributed Filesystem*). Kemudian dilakukan proses pencarian sekuens data berdasarkan *field - field* dalam data Genbank sesuai kata kunci yang diinginkan (misalnya *Author*) dengan menggunakan pustaka *MapReduce*. *MapReduce* merupakan *framework software* yang digunakan untuk

mendukung *distributed computing* pada platform *Hadoop* dan dijalankan dalam data dengan jumlah besar yang tersimpan dalam HDFS.

Hadoop mengintegrasikan *MapReduce* dan HDFS untuk mencapai tujuan yaitu melakukan proses pencarian sekuens data Genbank yang telah diunggah kedalam HDFS untuk menampilkan hasil pencarian berdasarkan *record* dan *field - field* yang telah ditentukan sebelumnya.

A. Bahan Penelitian

Bahan yang diperlukan untuk melakukan penelitian tentang *Hadoop* ini adalah data Genbank yang terdapat di FTP NCBI, yaitu <ftp://ftp.ncbi.nih.gov/genbank>. Data Genbank yang diambil adalah *Gbbct1.seq.gz* sampai dengan *Gbbct100.seq.gz* sebesar 6 GB.

B. Perangkat Penelitian

Perangkat dalam penelitian ini meliputi *software* dan *hardware* yang digunakan untuk pembuatan infrastruktur *Hadoop Cluster*.

1) Komputer Server (Master) :

- Dell Power Edge R200
- Intel Xeon Quad Core X3210 @ 2.13 GHz
- VGA Controller ATI ES1000 32MB Memory
- RAM 4GB 667MHz
- Hard disk drive Seagate 250GB 7200 rpm 6Gbps
- Network Card Broadcom NetXtreme BCM5721 Gigabit Ethernet Adapter
- DVD-ROM OEM

2) Komputer Slave; 6 PC dengan spesifikasi:

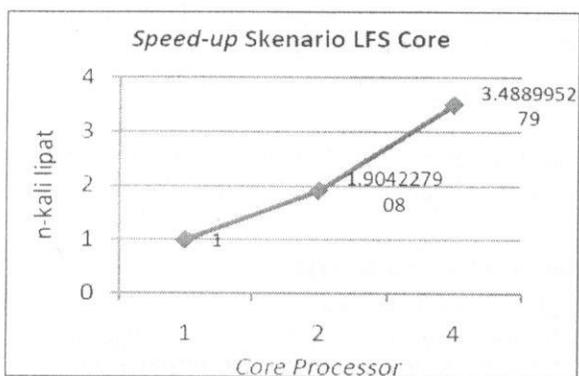
- Intel Pentium Dual Core E2160 @ 1.8 GHz
- VGA Controller SiS771/671
- RAM 1GB
- Hard disk drive Maxtor 80GB 7200 rpm 6Gbps
- Network Card SiS191 Gigabit Ethernet Adapter
- DVD-ROM Samsung
- Network: • Switch 16 port 10/100 Mbps D-Link DES-1016D; • UTP Cable Belden Cat5e

4. Pengujian Kecepatan Pencarian Sekuens DNA

Pengujian ini meliputi pengukuran waktu menggunakan aplikasi pencarian sekuens genbank untuk mendukung sistem *Hadoop Cluster* yang telah dibuat dalam penelitian ini dan dibandingkan dengan aplikasi pencarian sekuens data genbank menggunakan *Local Filesystem*.

Sebelum memperlihatkan kinerja dari pencarian menggunakan HDFS, berikut adalah tabel waktu pencarian

data sekuens DNA pada data Genbank yang diunggah pada satu basisdata tunggal dalam *Local File System* di server Dell Power Edge R200 yang spesifikasinya disebut sebelum ini. Waktu pencarian dengan menggunakan satu core dari 4 (Quad) processor core yang dimiliki server Dell Power EdgeR200 adalah 5247 detik atau hampir 1,5 jam. Dengan menggunakan dua core waktu yang diperlukan adalah 2755 detik atau percepatan (*speed up*) hampir 2x. Waktu yang diperlukan dengan menggunakan semua 4 core yang dimiliki adalah 1503 detik atau sekitar 25 menit. Gbr 2 dibawah memperlihatkan percepatan waktu pencarian antara penggunaan multi core dibandingkan dengan core tunggal.



Gambar 1 Percepatan waktu pencarian pada basisdata tunggal (Local File System) dengan penggunaan multicore yang berbeda.

C. Skenario Eksperimen

Skenario eksperimen pada penelitian ini adalah pencarian sequens data Genbank sebesar 6 GB dengan menggunakan HDFS pada *Hadoop Cluster* dan sebagai perbandingan menggunakan *Local Filesystem* pada satu komputer.

Eksperimen ini melakukan pencarian sekuens data genbank dengan menggunakan *Hadoop Cluster* yang telah dikonfigurasi sebelumnya, yaitu:

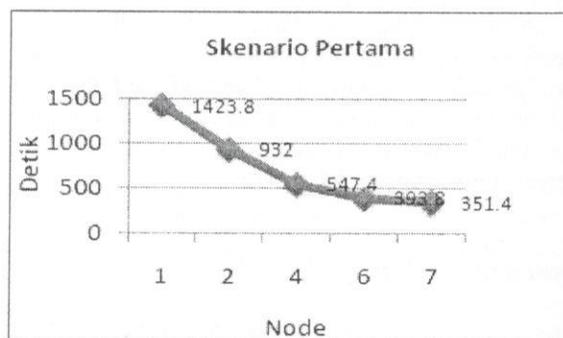
- 1) Mencari Authors, bernama Harano dengan luaran sedikit yaitu field Locus, Definition, dan Authors dari data genbank^[9].
- 2) Mencari Authors bernama Harano dengan luaran sedang yaitu field Locus, Authors, dan Origin dari data genbank. Dimana Origin adalah data sekuens DNA^[9].
- 3) Mencari Authors bernama Harano dengan luaran yang lebih banyak yaitu semua field yang ada di genbank format maupun sub-field pada data Genbank, yaitu: Locus, Definition, Accesion, Version, Keywords, Source,

Organism, Reference, Authors, Title, Journal, Pubmed, Comment, Features, dan Origin^[9].

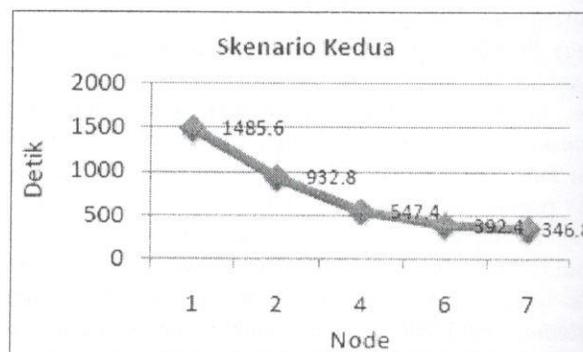
Pada eksperimen dengan basisdata tunggal menggunakan *Local Filesystem* yang telah ditampilkan sebelum ini merupakan data yang sama dengan yang digunakan oleh Hadoop Cluster. Pada komputer tunggal dengan *Local File System* juga dilakukan pencarian berupa Authors bernama Harano dengan output yang diinginkan berupa semua field maupun sub-field pada data Genbank, yaitu: Locus, Definition, Accesion, Version, Keywords, Source, Organism, Reference, Authors, Title, Journal, Pubmed, Comment, Features, dan Origin^[9]. Eksperimen ini dilakukan sebagai perbandingan Local Filesystem dalam satu komputer dengan HDFS pada *Hadoop Cluster* pada akhir artikel ini.

D. Hasil Pengujian

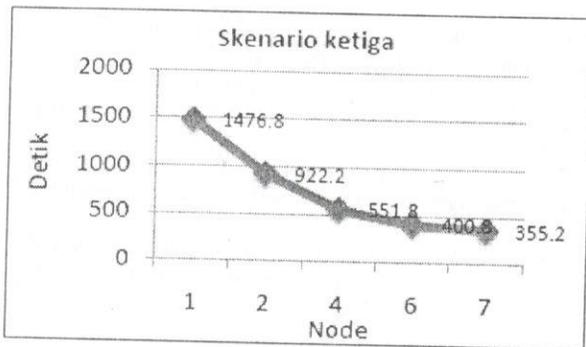
Untuk mengetahui rata-rata waktu yang dibutuhkan pada skenario eksperimen pertama, kedua, dan ketiga terdapat pada grafik dibawah ini:



Gambar 2 Rerata waktu pencarian pada skenario pertama (luaran sedikit) dengan banyaknya node pada HDFS diubah dari 1 sampai dengan 7 node



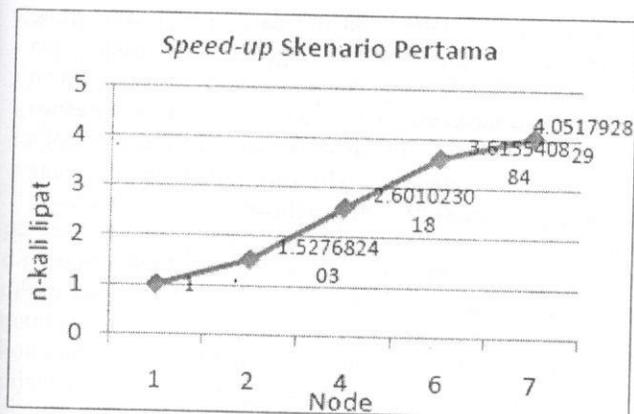
Gambar 3 Rerata waktu pencarian pada skenario kedua (luaran sedang) dengan banyaknya node pada HDFS diubah dari 1 sampai dengan 7 node



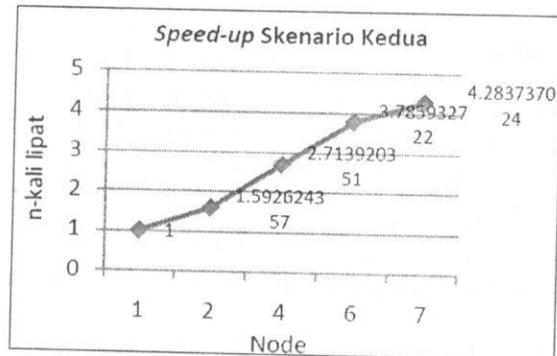
Gambar 4 Rerata waktu pencarian pada skenario ketiga (luaran banyak) dengan banyaknya node pada HDFS diubah dari 1 sampai dengan 7 node

Pada grafik diatas terlihat bahwa terjadi penurunan waktu yang dibutuhkan untuk melakukan pencarian sekuens data genbank dengan menambahkan jumlah node dalam HDFS mulai dari 1 node, 2 node, 4 node, 6 node sampai dengan 7 node. Sedangkan perbedaan dalam waktu pencarian tidak tampak terlihat antara skenario satu, dua dan tiga, Hal ini menjelaskan bahwa banyaknya luaran yang berarti waktu proses penulisan luaran tidak berarti terhadap lamanya waktu pencarian atau pembacaan data.

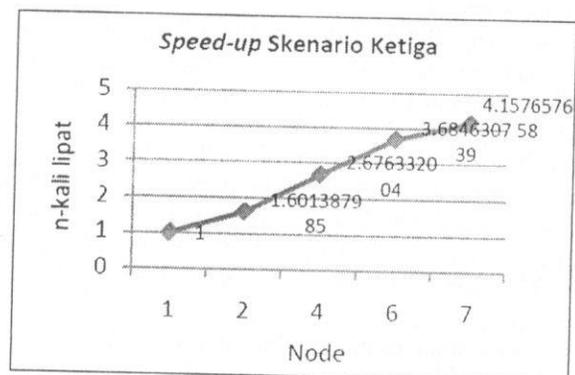
Sedangkan untuk perbandingan dari *speed-up* dari skenario eksperimen pertama, kedua dan ketiga dapat dilihat pada grafik dibawah ini. Dari Gambar 5, 6 dan 7 terlihat percepatan akibat penambahan node tidak selinear percepatan atau *speed-up* karena penambahan pemakain processor core pada quad core serve pada eksperimen dengan LFS di Gambar 2.



Gambar 5 SpeedUp pada skenario pertama (luaran sedikit) untuk pencarian sekuens pada HDFS dengan jumlah node 1, 2, 4, 6 dan 7



Gambar 6 SpeedUp pada skenario kedua (luaran sedang) untuk pencarian sekuens pada HDFS dengan jumlah node 1, 2, 4, 6 dan 7



Gambar 7 SpeedUp pada skenario ketiga (luaran banyak) untuk pencarian sekuens pada HDFS dengan jumlah node 1, 2, 4, 6 dan 7

Penggunaan 7 node 4,15x lebih cepat dibandingkan dengan penggunaan 1 node. Walaupun secara teoritis seharusnya penggunaan 7 node dapat lebih cepat 7x lipat dibandingkan dengan penggunaan 1 node. Hal ini kemungkinan terjadi diakibatkan beberapa faktor, diantaranya:

- *Overhead* yang terjadi pada proses *input/output* (I/O) antara prosesor dengan memori dan dengan HDFS
- *Overhead* yang terjadi pada proses I/O antara *master* dengan *node* saat *namenode* melalui *jobtracker* memberikan delegasi pekerjaan kepada *datanode* melalui *tasktracker*
- Penggunaan *Switch* yang berspesifikasi 10/100 Mbps, bukan *Gigabit Switch*
- Menggunakan spesifikasi komputer yang dibawah *minimum requirement* dari *Hadoop Cluster*
- Menggunakan kabel *UTP cat5e*, bukan *UTP cat6* untuk spesifikasi *Gigabit Network*

4. Penutup

Berdasarkan implementasi eksperimen dan evaluasi yang telah dilakukan terhadap pencarian sekuens data Genbank menggunakan *Hadoop Distributed Filesystem*, maka dapat dirangkum hasil pengujian ini sebagai berikut:

- *Hadoop Cluster* berhasil diimplementasikan dalam Laboratorium Jaringan Universitas Al Azhar Indonesia menggunakan 1 komputer sebagai *master namenode* dan 7 komputer sebagai *slave datanode* (termasuk *master namenode* didalamnya merangkap sebagai *datanode*) dengan topologi jaringan yang digunakan adalah topologi *star*.
- *Hadoop Cluster* dapat digunakan dengan *single node* ataupun *multi node*, disesuaikan dengan eksperimen yang telah ditentukan sebelumnya.
- Waktu yang dibutuhkan untuk melakukan proses pencarian sekuens data Genbank menggunakan 7 *node* rata-rata sebesar 351 detik dan rata-rata sebesar 1462 detik menggunakan 1 *node*.
- Proses pencarian data menggunakan 7 *node* adalah rata-rata 4x lebih cepat dibandingkan dengan menggunakan 1 *node*, memperlihatkan efisiensi Hadoop sekitar 57%.
- Banyaknya luaran tidak banyak berpengaruh pada waktu pencarian, sehingga dapat diambil kesimpulan bahwa proses penulisan pada Hadoop sangat efisien.
- Proses pencarian menggunakan HDFS dengan 1 *node* memiliki perbedaan sebesar 1.8% lebih cepat dibandingkan dengan menggunakan LFS dengan 4 *core processor*.

Adapun saran untuk pengembangan selanjutnya dari penelitian ini adalah:

- Penggunaan spesifikasi komputer dan perangkat jaringan yang sesuai atau melebihi dari kebutuhan minimum untuk melakukan implementasi *Hadoop Cluster* sehingga dalam melakukan pemrosesan suatu data dapat berjalan lebih cepat.
- Melakukan pengunggahan data ke dalam HDFS tidak dalam format yang tidak terkompresi untuk melihat perbedaan waktu yang dibutuhkan akibat proses *uncompressed*.
- Memperbanyak jumlah *node* untuk memeriksa faktor skalabilitas dari implementasi pencarian sekuens DNA pada sistem Hadoop yang dikembangkan.

REFERENSI

- [1] K. Shvachko, H. Kuang, S. Radia, R. Chansler, "The Hadoop Distributed File System" Proceeding of MSST2010, Proceeding The 26th IEEE Symposium on Massive Storage Systems and technologies, May 2010

- [2] T. White, *Hadoop: The Definitive Guide*, Third Edition. United State of America: O'Reilly, 2012.
- [3] J. Venner, *Pro Hadoop*. New York: Apress, 2009.
- [4] E. Sammer, *Hadoop Operations*. United State of America: O'Reilly, 2012.
- [5] P. Russom, "HADOOP- revealing Its True Value for Business Intelligence", TDWI Checklist Report, December 2011
- [6] D.L. Wheeler, T. Barrett, et al. "Database Resources of The National Center for Biotechnology Information", *Nucleic Acids Research*. 2008, Vol.36, Database issue, D13-D21, 2008
- [7] E.W. Sayers, T. Barrett, et al., "Database Resources of The National Center for Biotechnology Information", *Nucleic Acids Research*, Vol.37, Database issue, D5-D15, 2009
- [8] D.A. Benson, I. Karsch-Mizrachi, K. Clark, D.J. Lipman, J. Ostel, E.W. Sayers, "Genbank", *Nucleic Acids Research*, Vol. 40, Database issue, D48-D53. 2012.
- [9] A. Jamal, I. Chabib, V. Narita, "Pengembangan Database GenBank UAI-BioInformatics", *Prosiding e-Indonesia Initiatives (eII) IX*, ITB Bandung, 2013.
- [10] A. Jamal, K.K. Wardani, V. Narita, "Pairwise Sequence Alignment untuk Phylogenetic tree dari genbank UAI-Bioinformatics", *Prosiding e-Indonesia Initiatives (eII) IX*, ITB Bandung, 2013.

Ade Jamal, lulus S3 dari Delft University of Technology, Belanda pada tahun 1992, dalam bidang mekanika komputasi. Saat ini sebagai staf Pengajar program studi Teknik Informatika Universitas Al Azhar Indonesia.

UCAPAN TERIMA KASIH

Artikel ini adalah hasil dari penelitian yang mendapatkan dana Hibah Bersaing dari Dikti pada tahun kegiatan 2013/2014, untuk itu penulis mengucapkan banyak terima kasih atas dukungannya. Selain itu, apresiasi yang tinggi juga disampaikan kepada LP2M, lembaga penelitian Universitas Al-Azhar Indonesia yang mendukung banyak kegiatan penelitian.

ABS
C
infor
discu
or h
proce
loopi
string
gener
proce
so th
crack
Integr
Langu
based
The r
parall
than 8

Key v
Br

1. Pe

Ke
perken
dirasak
dikarer
mampu
komun
dijawal
dikatak
bagi se
paling
teknolo
Info
Kebera
menem
ini, keb
teknolo