Tugas Pendahuluan Proyek Akhir

IMPLEMENTASI APACHE SPARK STREAM CLUSTERING UNTUK ANALISA EVENT LOG PADA APLIKASI MATA GARUDA

DIMAS RIZKY HARSOYO PUTRO **2110141011**

D4 TEKNIK INFORMATIKA
DEPARTEMEN TEKNIK INFORMATIKA DAN KOMPUTER
POLITEKNIK ELEKTRONIKA NEGERI SURABAYA

2017

A. JUDUL PROYEK AKHIR

Implementasi Apache Spark Stream Clustering untuk Analisa Event Log pada Aplikasi Mata Garuda

B. PENDAHULUAN

Internet telah menjadi kebutuhan primer bagi berbagai kalangan manusia di dunia saat ini. Hampir 40% dari populasi manusia di dunia memiliki akses ke internet. Seiring dengan penggunaan yang masif ini, ancaman terhadap pencurian in-formasi yang dikirimkan melalui internet juga meningkat secara signifikan. Akibatnya, para ahli keamanan jaringan harus menggunakan sebuah sistem pencegah serangan untuk meminimalisir dampak buruk serangan tersebut. Sistem ini banyak di-kenal sebagai *Intrusion Detection System* (IDS).

Intrusion Detection System adalah sebuah sistem yang dapat digunakan untuk mendeteksi intrusi dalam sebuah sistem atau jaringan. Intrusi adalah sebuah aktivitas tidak sah atau tidak diinginkan yang mengganggu kerahasiaan, integritas dan atau ketersediaan dari data yang terdapat di sebuah sistem. Secara umum, sistem IDS berfungsi sebagai sebuah aplikasi yang memonitor lalu lintas data pada jaringan. Ketika sebuah intrusi terdeteksi, IDS akan menyimpan kejadian intrusi tersebut kedalam *log file* yang hasilnya dapat dianalisa lebih lanjut.

Mata Garuda merupakan *IDS (Intrusion Detection System)* yang memonitor lalu lintas data pada jaringan internet yang ada di Indonesia. Mata Garuda akan melaporkan suatu kejadian yang cirinya telah didefinisikan dalam sebuah rule. Kejadian-kejadian tersebut didapatkan melalui sensor yang terpasang di netiap *Network Access Point* yang ada di Indonesia. Sensor tersebut berfungsi untuk mengambil paket lalu meneruskannya ke *defense center*. Setiap harinya terdapat puluhan juta packet yang ditangkap oleh sensor Mata Garuda dan dengan terdeteksi rata-rata 2 juta serangan dideteksi per harinya.

Dengan Semakin berkembangnya teknologi dan semakin banyaknya pengguna internet di Indonesia maka akan semakin besar pula lalu lintas data yang melewati sensor mata garuda. Dengan semakin besarnya lalu lintas data pada jaringan akan menyebabkan semakin banyaknya *event* yang harus diproses dan dianalisa oleh *defense center* Mata Garuda. Hal ini berpengaruh secara langsung dalam kecepatan Mata Garuda dalam menganlisa data lalu lintas tersebut. Oleh karena itu,

pengembangan arsitektur sistem Mata Garuda harus dilakukan agar dapat sesuai dengan kondisi sekarang

C. PERUMUSAN MASALAH

Mata Garuda masih menggunakan *RDBMS* sebagai model databasenya dimana terdapat keterbatasan dalam segi volume data yang dapat disimpan dan kompleksitas query yang bisa ditangani, serta waktu komputasi yang kurang optimal yaitu kurang lebih terdapat *delay* kurang lebih 15 menit antara proses *logging* dan *pushing* ke databasenya dikarenakan besarnya biaya komputasi yang diperlukan jika digunakan untuk menganalisa *log data* yang berukuran besar dari sensor Mata Garuda.

D. TINJAUAN PUSTAKA

Tinjauan pustaka ini membahas tentang teori-teori penunjang dalam penyelesaian proyek akhir ini. Beberapa teori penunjang tersebut adalah :

1. *Intrusion Detection System*

   Intrusion Detection System atau yang dikenal sebagai IDS adalah sebuah ap-likasi yang memonitor jaringan dari aktivitas mencurigakan baik dari dalam maupun luar jaringan. Setiap aktivitas yang terdeteksi sebagai pelanggaran akan dilaporkan kepada Administrator melalui *Security Information and Event Management* (SIEM). Berdasarkan letak deteksinya ada dua tipe IDS yang ada, yaitu *Network-based IDS* (NIDS) dan *Host-based* (HIDS).

2. Snort

   Snort merupakan *Network-Based Intrusion Detection System* berlisensi *open source* yang dapat melakukan *packet logging* dan *traffic analysis* secara real-time. Dibuat oleh Martin Roesch pada tahun 1998. Snort merupakan *rule-based NIDS* yang artinya Snort menggunakan *rule* untuk mendeteksi adanya serangan pada jaringan. Snort akan melakukan tindakan yang sebelumnya telah ditentukan ketika mendeteksi adanya serangan yang sesuai dengan *ruleset* yang telah ditentukan.

3. Apache Hive

Apache Hive merupakan proyek *opens source* yang dilakukan oleh Apache Software Foundation dan dibangun di atas infrastruktur Hadoop. Hive adalah *tool* data warehouse untuk memproses data yang ada di Hadoop, Hadoop merupakan framework untuk menangani dataset berukuran besar. Hive memiliki tiga fungsi utama, *data summarization*, *query* dan *analysis*. Hive memiliki bahasa untuk mengekspresikan *query* yang dinamakan HiveQL. HiveQL akan menerjemahkan query SQL ke query Hadoop MapReduce untuk dilakukan pemrosesan, namun pemrosesan data tersebut juga dapat diintegrasikan dengan menggunakan *engine* lain, contohnya adalah Apache Sparko

4. Apache Spark

   Apache Spark adalah *engine* open source untuk melakukan pemrosesan big data. Dilengkapi dengan modul untuk melakukan *streaming,* SQL, *machine learning* dan *graph processing.* Apache Spark dapat berjalan berjalan di platform yang berbeda seperti Hadoop, Mesos, dan lainnya, serta dapat mengakses data dari sumber yang berbeda seperti HDFS, Cassandra, HBase, S3 ataupun sumber data yang lain. Spark memungkinkan untuk pembuatan aplikasi parallel dengan bahasa pemrograman yang berbeda-beda meliputi Java, Scala, Python, dan R.

5. Mata Garuda

   Mata Garuda adalah IDS berbasis Snort IDS. Mata Garuda memantau lalu lintas jaringan internet di Indoneisa dan mengenali sebuah serangan atau kejadian lain yang sesuai dengan *ruleset* yang ada lalu menampilkannya dalam bentuk laporan yang dapat dengan mudah dibaca.

E. PENELITIAN TERKAIT

Berikut adalah penelitian yang pernah dilakukan dan relevan dengan proyek akhir ini.

1. IDS Log Analisis Menggunakan Hadoop dan Mahout untuk Data Mining Pada Mata Garuda oleh M. Hisyam, F. A. Saputra and J. Akhmad [1]

   Penelitian ini mencoba untuk melakukan pemrosesan Snort *log file* dengan menggunakan prinsip *Big Data* di aplikasi Mata Garuda. Dengan

sistem terdistribusi, metode *data mining* dilakukan terhadap data geolocation untuk mendapatkan lokasi serangan yang terjadi. Penulis menggunakan UDTF untuk melakukan *query* serta membandingkannya dengan *join query*. Sedangkan algo-ritma yang diterapkan dalam proses *mining* tersebut adalah K-means *clustering* untuk mendapatkan *cluster* dari GeoIP serangan. Hasilnya, UDTF mampu meruduksi waktu komputasi menjadi 0.08 detik daripada *join query* yang me-makan waktu 3561 detik. Pada penelitian ini juga membuktikan bahwa penggunaan *distributed processing* untuk mengolah data Mata Garuda merupa-kan pilihan yang tepat karena data dapat diolah lebih cepat.

2. Real-Time Intrusion Detection System Using Multi-agent System oleh W. Laftah Al-Yaseen, Z. Ali Othman, dan M. Zakree Ahmad Nazri [2]

   Penelitian ini mencoba untuk meningkatkan performa dari sistem *IDS* dengan cara mengurangi waktu pemrosesan data pada saat *IDS* melakukan analisa serangan pada lalu lintas data yang ada menggunakan *Multi-agent System*. Jumlah *agent* yang dilibatkan bersifat adaptif dan dapat berubah-ubah sesuai dengan ukuran lalu lintas data dan ketersediaan *resource* pada sistem, sehingga penggunaan *multi-agent* tidak membebani performa dari host. Hasil yang didapatkan menggunakan *Multi-agent System IDS* mampu mengurangi waktu pemrosesan analisa serangan sebesar 81% dibandingkan dengan *IDS* tradisional dengan tetap mempertahankan akurasi dari analisanya

F. TUJUAN PROYEK AKHIR

Tujuan dari proyek akhir ini adalah untuk membangun arsitektur Mata Garuda yang dibangun menggunakan Apache Spark sebagai *engine* untuk pemrosesan data dengan prinsip big data dan Apache Hive sebagai data warehouse-nya.

G. KONTRIBUSI PROYEK AKHIR

Hasil dari proyek akhir ini dapat digunakan untuk mengoptimalkan analisa log Mata Garuda dan memperluas kapabilitas *query* pada Mata Garuda. Sehingga diharapkan kinerja analisa log Mata Garuda dapat berjalan secara optimal dalam melakukan pemrosesan data yang berukuran besar.

H. METODE PROYEK AKHIR

Untuk menyelesaikan proyek akhir ini langkah-langkah yang diambil ialah :

1. Studi Literatur

   Studi literatur merupakan langkah awal dalam pengerjaan proyek akhir ini. Tahapan ini merupakan tahap yang penting untuk memelajari teori-teori serta konsep teknis maupun non-teknis yang menunjang dalam pengerjaan proyek akhir ini.

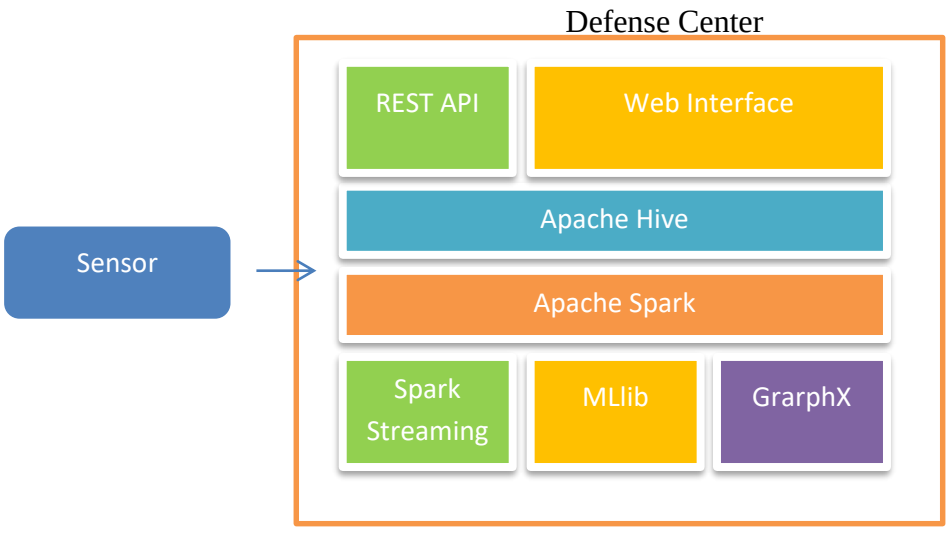
2. Pengumpulan data

   Langkah selanjutnya dari penelitian ini adalah pengumpulan data log file dari Bro sebagai bahan analisa awal. Setelah data terkumpul, maka data tersebut akan diolah dengan teknik *text extraction* untuk membentuk algoritma dalam log parser.

3. Perancangan sistem

   Setelah tahap studi literatur, langkah yang diambil selanjutnya adalah melakukan perancangan sistem.

   Secara umum, konfigurasi dari blok diagram sistemnya adalah sebagai berikut

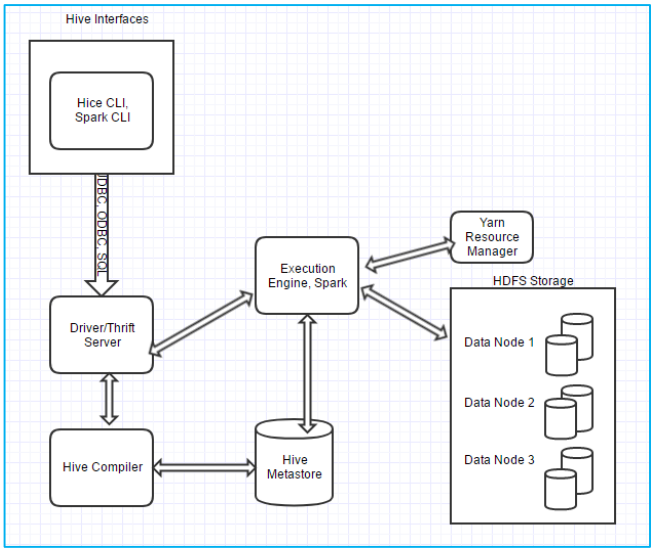


   a. Sensor

      Sensor merupakan sensor mata garuda yang menerima kejadian-kejadian yang ada di jaringan. Kejadian tersebut berupa *log file* yang berukuran besar. Untuk setiap kejadian yang terjadi akan dikirimkan *log file* ke *defense center* lalu akan diproses dan dianalisa disana.

b. Apache Hive

Apache Hive bekerja sebagai SQL untuk data yang dikirimkan pada sensor apache hive dikonfigurasikan agar dapat bekerja dengan engine spark. Architecture internal untuk Apache Hive sendiri adalah sebagai berikut



a. REST API

REST API merupakan service yang digunakan untuk melakukan pengambilan data dari database HDFS untuk ditampilkan ke web interface.

- Metric Mata Garuda

  Dalam service-nya digunakan metric mata garuda antara lain adalah, *event monitoring, event statistics, top signature, top protocols, sensor statistics, daily-monthly-annually report, attack trend, event analysis, user-role-profile-menu management*

b. Web Interface

Web interface akan digunakan untuk menampilkan hasil query yang dilakukan oleh REST API. Laporan yang ditampilkan adalah berupa laporan terkait serangan dan kejadian yang ditangkap oleh sensor

c. Apache Spark

Spark bekerja sebagai execution engine untuk pemrosesan data yang berasal dari sensor. Spark diintegrasikan dengan menggunakan SQL Hive. Spark bertugas untuk melakukan eksekusi query yang dilakukan oleh REST API

a. Streaming SQL

   Dapat memungkinkan spark memproses data secara stream, sehingga memungkinkan untuk melakukan *write job* secara streaming

b. Mllib

   Mllib merupakan library machine learning dari spark, bertugas untuk melakukan klustering data yang ada pada storage HDFS

c. GraphX

   Merupakan library spark untuk melakuka komputasi graph-paralel

4. Implementasi

   Dalam tahap ini akan dibangun arsitektur dan desain dari server *spark* dan *hive*. Dan juga implementasi dari rancangan aplikasi berdasarkan *ERD* dan *UML* yang telah dibuat sebelumnya disertai *testing* dan *debugging*

5. Pengujian dan analisa

   Pada tahap ini dilakukan pengujian terhadap implementasi desain sistem dengan menggunakan dataset yang disesuaikan dengan intrusi sesungguhnya dan komparasi dengan sistem sekarang. Pengujian dilakukan secara *side-by-side* dengan aplikasi Mata Garuda dengan arsitektur yang digunakan sekarang. Keluaran yang di-harapkan dari sistem ini adalah performa yang lebih optimal dari arsitektur mata garuda yang baru dibandingkan dengan yang lama.

I. JADWAL PELAKSANAAN

Kegiatan	Bulan												
	8	9	10	11	12	1	2	3	4	5	6	7	8
Proposal ProyekAkhir													
Pengambilan Data													
Implementasi/Coding													
Debugging													
Testing &Analisa													
Penyusunan Buku PA													
Sidang PA													

J. PERSONALIA PROYEK AKHIR

- Mahasiswa

  Nama : Dimas Rizky H.P.

  NRP : 2110141011

  Jurusan : Diploma IV - Teknik Informatika

  Agama : Islam

  Jenis kelamin : Laki-laki

- Dosen Pembimbing 1

  Nama : Ferry Astika Saputra, ST, M.Sc

  NIP : 197708232001121002

Departemen : Teknik Informatika

Bidang keahlian : Computer Network, Network Security

- Dosen pembimbing 2

Nama : Jauari Akhmad S.ST

NIP : 2000000052

Departemen : Teknik Informatika

Bidang keahlian : Rekayasa Perangkat Lunak

K. PERKIRAAN BIAYA PROYEK AKHIR

No	Uraian	Jumlah	Harga Satuan(Rp)	Total (Rp)
1	RAM 8GB	4 buah	800.000	3.200.000
2	Kertas A4 HVS 80gr	1 rim	40.000	40.000
3	Kertas A5	1 rim	40.000	40.000
4	Tinta Printer	4 buah	25.000	100.000
5	Internet	8 bulan	100.000	800.000
6	Dokumentasi Laporan	6 bundle	25.000	150.000
TOTAL PENGELUARAN				**4.330.000**

L. DAFTAR PUSTAKA

[1] M. Hisyam, F. A. Saputra and J. A. Nur Hasyim, IDS Log Analisis Menggunakan Hadoop dan Mahout untuk Data Mining Pada Mata Garuda, Surabaya: Politeknik Elektronika Negeri Surabaya, 2015.

[2] W. L. Al-Yaseen, Z. A. Othman and M. Z. Ahmad Nazri, "Real-Time Intrusion Detection System Using Multi-agent System," *IAENG International Journal of Computer Science,* pp. 1-11, 2016.