SKRIPSI

STUDI DAN IMPLEMENTASI APACHE SPARK GRAPHX UNTUK ANALISIS BIG DATA GRAPH

Stephanie Tania

NPM: 2014730072

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

UNDERGRADUATE THESIS

STUDY AND IMPLEMENTATION OF APACHE SPARK GRAPHX FOR BIG DATA GRAPH ANALYSIS

Stephanie Tania

NPM: 2014730072

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

LEMBAR PENGESAHAN

STUDI DAN IMPLEMENTASI APACHE SPARK GRAPHX UNTUK ANALISIS BIG DATA GRAPH

Stephanie Tania

NPM: 2014730072

Bandung, 28 Mei 2018

Menyetujui,

Pembimbing Utama	Pembimbing Pendamping
Dr. Veronica Sri Moertini	Gede Karya, M.T., CISA, IPM

Ketua Tim Penguji	Anggota Tim Penguji
Rosa De Lima, M.Kom.	Luciana Abednego, M.T.

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

STUDI DAN IMPLEMENTASI APACHE SPARK GRAPHX UNTUK ANALISIS BIG DATA GRAPH

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 28 Mei 2018

Stephanie Tania
NPM: 2014730072

ABSTRAK

Data yang terkumpul di dunia saat ini memiliki berbagai macam tipe dengan setiap tipe tersebut diolah menggunakan cara yang berbeda, seperti data graf yang terdiri dari sisi dan simpul memerlukan algoritma-algoritma graf untuk mendapatkan informasi. Data graf dapat berupa data komunikasi telepon, hubungan pertemanan, atau rute transportasi yang berukuran besar. Untuk mengolah data berukuran besar tersebut dengan cepat, teknologi sistem terdistribusi seperti Hadoop dan Spark dikembangkan. Teknologi tersebut digunakan untuk membagi beban pemrosesan ke perangkat-perangkat lain. Hadoop dan Spark menggunakan sistem terdistribusi untuk mengolah data berukuran besar dengan cara pengolahan data yang berbeda.

Data graf besar dapat dianalisis dengan menggunakan *library* GraphX yang berupa *library* untuk menangani data graf. GraphX dapat digunakan untuk melakukan operasi dasar seperti membuat graf dari data yang disediakan dan menghitung banyak simpul dan sisi atau menjalankan algoritma analisis seperti *page rank*, *connected components*, *shortest paths*, *subgraph*, dan *triangle count*. Graf yang ditangani oleh GraphX terbentuk dari dua buah *resilient distributed dataset* atau RDD untuk himpunan simpul dan himpunan sisi.

Sebuah eksperimen dilakukan untuk mengukur performansi *library* GraphX dalam menganalisis data penerbangan dari Biro Transportasi Amerika Serikat. Eksperimen dilakukan dengan memanggil algoritma-algoritma graf yang sudah diimplementasikan pada GraphX pada set data uji dengan ukuran yang berbeda-beda. Hasil eksperimen menunjukkan bahwa waktu eksekusi untuk setiap algoritma tersebut berbeda-beda dan tergantung pada cara kerja algoritma serta ukuran data. Waktu eksekusi tersebut cenderung mengalami peningkatan seiring dengan bertambahnya ukuran data.

Kata-kata kunci: data besar, sistem terdistribusi, graf, Hadoop, Spark, GraphX

ABSTRACT

The gathered data of this world currently have various types with each of them processed in different ways, just like graph data, which consist of edges and vertices, need graph algorithms to gain information. The graph data can be telephone communication data, friend relations, or transportation routes in big size. To process data of such size quickly, distributed system technologies such as Hadoop and Spark are developed. Those technologies are used to divide the processing workload among machines. Both Hadoop and Spark use distributed system to process big data with different methods of processing.

Big graph data can be analysed by using Spark's GraphX library, which is a library used to handle graph data. GraphX can be used to do basic operations such as creating a graph from provided data and counting the number vertices and edges to running analysis algorithms such as page rank, connected components, shortest paths, subgraph, and triangle count. The graph handled by GraphX is made of two resilient distributed datasets or RDD for vertices and edges.

An experiment is done in order to measure the performance for Spark's GraphX library in analysing flight data from United States' Bureau of Transportation. The experiment is done by calling the graph algorithms implemented into GraphX on the test data with various sizes. The experiment shows the execution time for every algorithm differs and depends on how the algorithms work and the data size. The execution time tends to be slower as the data size grows.

Keywords: big data, distributed system, graph, Hadoop, Spark, GraphX

Skripsi ini dipersembahkan untuk semua orang terdekat penulis.

KATA PENGANTAR

Puji skyukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa yang telah memberikan berkat dan cinta kasih-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Penulis tidak akan dapat menempuh dan menyelesaikan pendidikannya tanpa bantuan secara langsung maupun tidak langsung dari berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada pihak-pihak yang telah membantu penulis:

- Ibu penulis, Tjong Suryani yang telah mendidik dan membesarkan penulis, serta memberikan banyak dukungan moral kepada penulis selama pembuatan skripsi ini.
- Ibu Dr. Veronica Sri Moertini dan Bapak Gede Karya, M.T., CISA, IPM yang telah mengajarkan hal-hal baru dan membimbing penulis dalam pembuatan skripsi ini.
- Ibu Rosa De Lima, M.Kom. dan Ibu Luciana Abednego, M.T. yang telah memberikan kritik dan saran yang berguna untuk membuat skripsi ini menjadi lebih baik.
- Teman-teman kuliah penulis, yaitu Lydia Febtriani, Gabriella, Vanessa Sukamto, Betari, dan Melinda Nur Abianti yang selalu ada dalam suka dan duka selama perkuliahan.
- Teman SMP penulis, Riska Wilian Triany Putri yang selalu memberikan dukungan dan hiburan walaupun sedang berjuang di jenjang pendidikan yang lebih tinggi di luar negeri.
- Admin-admin laboratorium komputer yang meminjamkan kunci laboratorium skripsi hampir setiap hari selama pembuatan skripsi.
- Lagu-lagu J-Pop dan permainan-permainan yang menghibur dan mendampingi penulis selama pembuatan skripsi.

Akhir kata, penulis memohon maaf apabila terdapat kesalahan pada skripsi ini. Penulis juga menerima semua kritik dan saran yang berguna untuk mengembangkan skripsi ini menjadi lebih baik. Semoga skripsi ini dapat menjadi referensi yang bermanfaat untuk penelitian-penelitian berikutnya.

Bandung, Mei 2018

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Seiring dengan perkembangan teknologi dan pertumbuhan penduduk dunia, ukuran data yang ada di dunia internet semakin bertambah besar. Data ini sebagian besar berformat graf, seperti data komunikasi telepon, data pertemanan pada media sosial, data rekan afiliasi antarorganisasi, dan lain-lain. Selain itu, masalah-masalah dalam dunia nyata juga dapat dimodelkan dalam bentuk graf seperti perancangan jadwal kegiatan, jalur tercepat pedagang keliling, dan bahkan solusi terbaik untuk permainan seperti catur. Oleh karena itu, setiap instansi atau organisasi di seluruh dunia harus dapat mengolah data yang berukuran besar atau *big data* dengan waktu yang sesingkat mungkin.

Pengolahan *big data* berformat graf tersebut dapat dilakukan dengan bantuan *framework*, seperti Hadoop maupun Spark. Ukuran dari *big data* sangat besar dan pengolahan data tersebut tidak jarang dilakukan. Selain itu, data yang diolah tersebut dapat menjadi sangat berharga pada waktu tertentu dan tidak berarti pada waktu lain. Oleh karena itu, pengolahan *big data* perlu dilakukan dalam waktu yang sesingkat mungkin untuk mendapatkan hasil analisis yang bernilai pada waktu yang dibutuhkan.

Data yang diolah dan dianalisis oleh perangkat lunak yang akan dibuat dalam skripsi ini adalah *big data* dengan format graf. Graf merupakan sebuah model visual yang menggambarkan relasi antara objek berupa simpul dengan simpul lainnya. Relasi yang terbentuk di antara simpul-simpul tersebut digambarkan dengan sebuah garis yang disebut dengan sisi atau *edge*. Graf tersebut digunakan untuk menyimpan informasi tertentu dan informasi tersebut dapat diambil dengan menggunakan algoritma-algoritma seperti *connected components*, *spanning tree*, *shortest paths*, *page rank*, dan *triangle count*.

Data graf yang diperoleh untuk dianalisis tersebut berukuran sangat besar dan melebihi batas kapasitas pemrosesan sistem basis data konvensional, sehingga disebut dengan istilah *big data* [8]. *Big data* merupakan sekumpulan data yang berukuran sangat besar dan dapat dianalisis untuk memperoleh informasi yang dapat digunakan dalam kegiatan manusia. Terkait dengan ukuran *big data* yang sangat besar, *framework* seperti Apache Hadoop dan Apache Spark dikembangkan untuk mempercepat waktu pengolahan data tersebut.

Hadoop merupakan sebuah *open source framework* untuk membuat dan menjalankan perangkat lunak terdistribusi yang mengolah data berukuran besar. Spark merupakan sebuah sistem untuk menganalisis data terdistribusi dan terdiri dari beberapa *library* yang mempunyai fungsi tersendiri tetapi tetap terhubung. GraphX merupakan salah satu dari *library* tersebut dan berfungsi untuk melakukan analisis pada data graf. Data graf tersebut dapat dianalisis dengan menggunakan algoritma-algoritma yang sudah diimplementasikan pada GraphX seperti *page rank*, *connected components*, *shortest paths*, *subgraph*, dan *triangle count*.

Ukuran data yang besar dapat menyebabkan waktu pemrosesan menjadi lebih lama, sehingga perlu digunakan sistem terdistribusi seperti Spark untuk mendapatkan hasil analisis dengan waktu yang lebih singkat. Oleh karena itu, pada skripsi ini dipelajari arsitektur dan cara kerja Spark dalam melakukan pengolahan data tersebut sehingga perangkat lunak Spark dapat diimplementasikan.

Hasil implementasi tersebut digunakan untuk melakukan pengukuran performansi perintah-perintah GraphX dalam menganalisis data uji yang berupa data penerbangan di Amerika Serikat.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dijabarkan, berikut adalah rumusan masalah untuk skripsi ini.

1. Bagaimana prinsip kerja Hadoop dan Spark dalam mengolah *big data*?
2. Bagaimana cara mengolah dan menganalisis *big data* graf menggunakan Spark dan *library* GraphX untuk beberapa data uji?
3. Bagaimana cara mengimplementasikan perangkat lunak demo dengan menggunakan Spark dan *library* GraphX?
4. Bagaimana hasil kinerja perintah-perintah pada *library* GraphX?

1.3 Tujuan

Berikut ini adalah tujuan yang ingin dicapai dalam skripsi ini.

1. Melakukan studi tentang prinsip kerja Hadoop dan Apache Spark dalam mengolah *big data*.
2. Mempelajari Spark dan *library* GraphX untuk mengolah *big data* graf untuk beberapa data uji.
3. Mempelajari Spark dan *library* GraphX untuk mengimplementasikan perangkat lunak demo.
4. Melihat hasil kinerja perintah-perintah pada *library* GraphX.

1.4 Batasan Masalah

Batasan masalah untuk skripsi adalah sebagai berikut.

1. Hadoop dan Spark dijalankan pada empat buah mesin dengan sistem operasi Ubuntu.
2. Data uji yang digunakan memiliki struktur graf berarah.
3. Pengembangan perangkat lunak untuk pemrosesan data dilakukan dengan menggunakan *library* Spark dan menggunakan bahasa pemrograman Scala.
4. Perintah-perintah GraphX yang diimplementasikan pada perangkat lunak adalah *page rank*, *connected components*, *shortest paths*, *subgraph*, dan *triangle count*.
5. Pengembangan antarmuka dilakukan dengan mengembangkan perangkat lunak web dan menggunakan bahasa pemrograman Java.

1.5 Metodologi

Metodologi yang digunakan dalam pembuatan skripsi ini adalah sebagai berikut.

1. Mempelajari arsitektur, cara kerja, dan komponen-komponen Hadoop.
2. Mempelajari arsitektur, cara kerja, dan komponen-komponen Spark.

3. Mempelajari distribusi data pada HDFS Hadoop.
4. Mempelajari distribusi data graf pada Spark.
5. Mempelajari *library* GraphX pada Spark.
6. Mempelajari bahasa pemrograman Scala.
7. Menguji dengan menggunakan MapReduce Hadoop.
8. Merancang perangkat lunak Spark yang memanggil fungsi-fungsi *library* GraphX.
9. Mencari data yang dapat dibuat menjadi struktur graf sebagai data uji.
10. Merancang dan mengimplementasikan perangkat lunak demo untuk mempraolah data uji untuk dibentuk menjadi graf dan dianalisis.
11. Membandingkan kinerja proses analisis graf yang dilakukan dengan *library* GraphX dan menggunakan data uji dengan ukuran yang berbeda.

1.6 Sistematika Pembahasan

Sistematika penulisan skripsi ini adalah sebagai berikut.

1. Bab Pendahuluan

   Bab 1 membahas tentang latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika pembahasan.

2. Bab Landasan Teori

   Bab 2 membahas tentang teori-teori mengenai graf, *big data*, bahasa Scala, sistem terdistribusi Hadoop, dan sistem terdistribusi Spark.

3. Bab Studi Eksplorasi

   Bab 3 membahas tentang langkah-langkah untuk melakukan konfigurasi klaster pada Hadoop, konfigurasi klaster untuk Spark, hasil studi eksplorasi MapReduce Hadoop, dan hasil studi eksplorasi Spark GraphX.

4. Bab Analisis dan Perancangan

   Bab 4 membahas tentang analisis perangkat lunak Spark, analisis data uji, analisis masukan dan keluaran, analisis antarmuka, diagram *use case*, skenario *use case*, rancangan proses praolah, rancangan proses analisis, diagram kelas, dan rancangan antarmuka.

5. Bab Implementasi dan Eksperimen

   Bab 5 membahas tentang implementasi perangkat lunak, eksperimen performansi perintah-perintah GraphX yang diimplementasikan, dan analisis hasil eksperimen.

6. Bab Kesimpulan dan Saran

   Bab 6 membahas tentang kesimpulan yang disampaikan penulis setelah melakukan penelitian ini dan saran-saran untuk pengembangan lebih lanjut.