SKRIPSI

STUDI DAN PERBANDINGAN APACHE SPARK SQL DAN HIVE DALAM KONTEKS ANALISIS BIG DATA

Lydia Febtriani

NPM: 2014730044

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

UNDERGRADUATE THESIS

STUDY AND COMPARISON OF APACHE SPARK SQL AND HIVE FOR BIG DATA ANALYSIS

Lydia Febtriani

NPM: 2014730044

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

LEMBAR PENGESAHAN

PERPUSTAKAAN UNIVERSITAS KATOLIK PARAHYANGAN

STUDI DAN PERBANDINGAN APACHE SPARK SQL DAN HIVE DALAM KONTEKS ANALISIS BIG DATA

Lydia Febtriani

NPM: 2014730044

Bandung, 21 Mei 2018

Menyetujui,

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

Ketua Tim Penguji	Anggota Tim Penguji
Vania Natali, M.T.	Claudio Franciscus, M.T.

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERPUSTAKAAN UNIVERSITAS KATOLIK PARAHYANGAN

PERNYATAAN

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

STUDI DAN PERBANDINGAN APACHE SPARK SQL DAN HIVE DALAM KONTEKS ANALISIS BIG DATA

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 21 Mei 2018

Lydia Febtriani
NPM: 2014730044

ABSTRAK

Seiring kemajuan dan pemanfaatan teknologi, data bertambah dengan cepat dalam waktu singkat sehingga ukuran data menjadi sangat besar atau sering disebut *big data*. Agar pengguna bisa mendapatkan informasi dari *big data*, *big data* harus diolah dan dianalisis dengan cepat. Beberapa teknologi yang dapat digunakan untuk mengolah dan menganalisis *big data*, yaitu Apache Hadoop dan Apache Spark. Hadoop adalah sistem terdistribusi yang dapat digunakan untuk mengolah dan menganalisis *big data* dengan cepat. Spark merupakan *framework* komputasi di *cluster*. Spark dioptimalkan untuk berjalan di memori sehingga pemrosesan data lebih cepat dibandingkan komponen komputasi Hadoop, MapReduce.

Hadoop dan Spark memiliki subproyek yang bekerja dengan data terstruktur, yaitu Apache Hive dan Spark SQL. Hive merupakan infrastruktur data *warehouse* berbasis Hadoop, sedangkan Spark SQL adalah *library* Spark untuk bekerja dengan data terstruktur. Hive dan Spark menyediakan pengerjaan kueri dengan atau mirip sintaks SQL. Spark SQL dan Hive yang digunakan untuk tujuan yang sama, yaitu mengolah dan menganalisis *big data* dengan dialek SQL, perlu dipelajari dan diteliti mengenai kesamaan dan perbedaan penggunaan sintaks SQL serta bagaimana performasinya dalam mengerjakan kueri dengan *big data*.

Penelitian ini dilakukan untuk mencari perbedaan Spark dan Hive. Perbedaan tersebut dapat berupa penggunaan sintaks SQL dan bagaimana performa Spark SQL dan Hive dalam mengerjakan kueri. Eksperimen dilakukan untuk melihat performa Spark dan Hive dalam mengerjakan kueri-kueri SELECT. Langkah penelitian yang dilakukan adalah mempelajari konsep dan pemanfaatan Spark SQL dan Hive, mengeksplorasi fitur SQL pada Spark dan Hive, menyiapkan *cluster* Hadoop dan Spark, mengumpulkan set *big data* untuk eksperimen, mengembangkan perangkat lunak berbasis *web* untuk eksperimen, dan melakukan eksperimen perbandingan kinerja Spark SQL dan Hive. Hasil dari penelitian ini berupa perangkat lunak berbasis *web* yang dapat menjalankan perangkat lunak dengan Spark untuk mengerjakan kueri-kueri Spark SQL dan Hive. Perangkat lunak *web* menampilkan daftar kueri untuk eksperimen dan dapat menampilkan hasil kueri serta waktu eksekusi kueri Spark SQL dan Hive. Kueri yang dipilih merupakan kueri-kueri SELECT yang diurutkan dari kueri sederhana hingga kueri yang lebih kompleks. Kueri-kueri yang digunakan sama baik untuk Spark SQL maupun Hive. Eksperimen dilakukan sebanyak tiga kali menggunakan perangkat lunak berbasis *web*. Data yang digunakan terdiri dari dua data, yaitu data Movies dengan ukuran 358 MB dan data Ratings dengan ukuran 2,6 GB. Hasil yang didapat dari eksperimen adalah waktu eksekusi kueri-kueri SELECT menggunakan Spark SQL dan Hive. Berdasarkan tiga kali eksperimen waktu eksekusi kueri-kueri SELECT dengan Spark SQL dan Hive, Spark SQL lebih cepat dalam menjalankan kueri-kueri SELECT dibandingkan Hive.

Kesimpulan pertama yang didapat dari penelitian ini adalah Spark SQL dan Hive digunakan untuk bekerja dengan data terstruktur. Tetapi, Spark SQL memiliki abstraksi RDD yang dirancang untuk penyimpanan data dalam memori sedangkan Hive menggunakan HDFS Hadoop sebagai komponen penyimpanan data. Hive juga mendukung lebih banyak pernyataan SQL dibandingkan Spark SQL. Kesimpulan dari eksperimen-eksperimen waktu eksekusi kueri Spark SQL dan Hive adalah Spark SQL lebih cepat dalam mengerjakan kueri dibandingkan Hive.

Kata-kata kunci: Apache Spark, Spark SQL, Apache Hadoop, Hive, *Big Data*

ABSTRACT

With current advanced technology and its utilization, in short period of time, the size of data has increased which results in what we know as big data. As users should be able to retrieve information from a significant amount of data, the system must be able to process and analyze it faster. Apache Hadoop is a distribution system which can efficiently handle the computation and analysis of big data. Meanwhile, Apache Spark is a computation framework that works on a cluster. One main advantage of Spark, compared to Hadoop's MapReduce, is in their processing speed since the computation runs in memory.

Both Apache Hadoop and Spark's subproject work with structured data, known as Apache Hive and Spark SQL. Hive is a data warehouse infrastructure with Hadoop as its basis, and Spark SQL is the Spark's library for working with structured data. Both Hive and Spark allow queries with SQL syntax or SQL-like syntax. Since both of them have the same purpose of processing and analyzing big data, Spark and Hive can be studied and analyzed further not only in the similarities and differences among usages of SQL syntax but also their performance on processing big data query.

This research aims at comparing Spark and Hive. The comparison can be in the form of SQL syntax usage and their performance on processing queries. The experiment demonstrates the performance of Spark and Hive on processing the SELECT queries. The research studies the concept and application of Spark SQL and Hive, explores the features of SQL in Spark SQL and Hive, prepares Hadoop and Spark clusters, collects big dataset for experiments, develops web-based application, and compares the performance of Spark SQL and Hive. The result of this research is a web application which can run Spark application to execute Spark SQL and Hive queries. The web application shows Spark SQL and Hive query lists for the experiments, the query results, and the execution time. The selected queries are SELECT queries, sorted from the simplest to more complex one. The queries for Spark SQL and Hive are same. The experiments are conducted three times using the web-based application. The data being used for this experiments are the Movies data, which is 358 MB in size, and the Ratings data, which is 2.6 GB in size. The result of this experiment is execution time SELECT queries for Spark SQL and Hive. Based on the experiments of SELECT queries, Spark SQL executes the commands faster than Hive.

The first conclusion of this research is that Spark SQL and Hive can be used to work with structured data. But the Spark SQL has RDD abstraction which is specifically designed for in-memory storage, while Hive uses HDFS Hadoop as the storage component. In contrast, Hive supports more SQL commands than Spark SQL. From the experiments of Spark SQL and Hive's query execution time, the Spark SQL runs faster than Hive.

Keywords: Apache Spark, Spark SQL, Apache Hadoop, Hive, Big Data

Dipersembahkan untuk orang tua, kedua kakak, dan keluarga penulis

KATA PENGANTAR

Puji dan syukur kepada Tuhan Yang Maha Esa atas dukungan-Nya sehingga penulis dapat menyelesaikan skripsi ini tepat pada waktunya. Pada kesempatan ini, penulis ingin menyampaikan rasa terima kasih kepada:

1. Kedua orang tua, kedua kakak, dan keluarga besar penulis yang telah mendukung penulis selama pengerjaan skripsi ini.
2. Ibu Veronica Sri Moertini sebagai dosen pembimbing utama dan Pak Gede Karya sebagai dosen pembimbing serta yang telah membimbing dan membantu penulis dalam mengerjakan skripsi ini.
3. Teman-teman seperjuangan yang saling membantu dan saling curhat, Tania dan Gabriella.
4. Teman-teman yang selalu menyemangati penulis, Vanessa, Tania, Gabriella, Betari.
5. Rekan-rekan mahasiswa bimbingan Ibu Veronica dan rekan-rekan dengan topik skripsi Spark yang telah saling membantu dan saling memberi informasi.
6. Rekan-rekan dengan topik skripsi Hadoop yang saling mengerti karena sama-sama menggunakan *cluster* Hadoop.
7. Rekan-rekan seangkatan Teknik Informatika yang saling memberi semangat.
8. Kakak perempuan dan tante yang selalu mendampingi dan menyemangati penulis.
9. Kakak laki-laki dan sepupu terdekat penulis, Gebbie De Nova, yang telah membantu dalam penulisan judul dan abstrak bahasa inggris.
10. Teman-teman SMA penulis yang memberi semangat ke penulis Jean, Nelly dan teman-teman lain yang tidak bisa disebutkan satu-per-satu.
11. Lagu-lagu Korea yang selalu menemani penulis dalam pengerjaan skripsi ini serta drama dan acara Korea yang menjadi penghibur penulis di saat jenuh dalam pengerjaan skripsi ini.

Akhir kata, penulis menyadari bahwa penelitian dan skripsi ini masih memiliki kekurangan. Penulis memohon maaf bila terdapat kesalahan pada penelitian atau skripsi ini. Penulis juga menerima kritik dan saran sehingga penelitian ini dapat berkembang menjadi lebih baik. Semoga penelitian ini dapat membantu penelitian-penelitian lainnya.

Bandung, Mei 2018

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

Pada bab ini dibahas tentang latar belakang penelitian dilakukan, rumusan masalah, dan tujuan dari penelitian ini. Selain itu batasan-batasan masalah, metodologi, dan sistematika pembahasan penelitian ini juga dijelaskan pada bab ini.

1.1 Latar Belakang

Seiring kemajuan dan pemanfaatan teknologi, data bertambah dengan cepat dalam waktu singkat sehingga ukuran data menjadi sangat besar. Ukuran data tersebut mencapai puluhan, ratusan *gigabyte*, hingga *petabyte*. Kumpulan data dengan ukuran besar tersebut disebut dengan *big data*. *Big data* perlu diolah dan dianalisis untuk memudahkan pengguna dalam memahami dan menyelesaikan masalahnya.

Beberapa teknologi, seperti Apache Hadoop dan Apache Spark, dibangun untuk mengolah dan menganalisis *big data*. Apache Hadoop merupakan *framework open source* untuk menulis dan menjalankan aplikasi terdistribusi yang memproses *big data*. Hadoop memiliki komponen penyimpanan, yaitu Hadoop Distributed File System (HDFS), dan komponen untuk komputasi, yaitu MapReduce. Hadoop mempartisi data dan melakukan komputasi *big data* secara paralel sehingga analisis dapat diselesaikan dengan cepat. Apache Hive merupakan infrastruktur data *warehouse* berbasis Apache Hadoop yang menyediakan dialek yang mirip dengan SQL. Dialek yang disediakan oleh Hive disebut HiveQL atau HQL.

Apache Spark merupakan *framework* komputasi di *cluster* untuk memproses dan menganalisis *big data*. Spark dioptimalkan untuk berjalan di memori sehingga membantu pemrosesan data lebih cepat dari pada MapReduce Hadoop. Spark memiliki komponen utama untuk menyimpan data dalam memori, yaitu RDD. Spark juga memiliki *library* untuk bekerja dengan data terstruktur, seperti data CSV, JSON, dan lainnya, yaitu Spark SQL. Spark SQL mendukung penggunaan bahasa SQL dan HiveQL.

Spark SQL dan Hive digunakan untuk tujuan yang sama, yaitu mengolah dan menganalisis *big data* dengan dialek SQL. Berdasarkan hal tersebut, Spark SQL dan Hive perlu dipelajari lebih lanjut. Penelitian ini diharapkan dapat membandingkan kinerja Spark SQL dan Hive dalam mengolah dan menganalisis *big data*.

1.2 Rumusan Masalah

Berikut rumusan masalah dari penelitian ini:

1. Bagaimana konsep dan pemanfaatan Spark SQL?
2. Bagaimana konsep dan pemanfaatan Hive?
3. Bagaimana perbandingan Spark SQL dan Hive dalam sintaks penulisan SQL dan kinerjanya untuk menganalisis *big data*?
4. Bagaimana rancangan dan implementasi perangkat lunak untuk Spark SQL dan Hive?

1.3 Tujuan

Berikut tujuan dilakukannya penelitian ini:

1. Mempelajari konsep dan pemanfaatan Spark SQL.
2. Mempelajari konsep dan pemanfaatan Hive.
3. Melakukan studi untuk membandingkan sintaks penulisan SQL pada Spark SQL dan Hive.
4. Melakukan eksperimen untuk membandingkan kinerja dalam menganalisis *big data* pada Spark SQL dan Hive.
5. Merancang dan mengimplementasikan perangkat lunak untuk menganalisis *big data* dengan Spark SQL dan Hive.

1.4 Batasan Masalah

Batasan masalah dari penelitian ini antara lain:

1. Data studi kasus yang digunakan untuk analisis merupakan data yang dapat diambil secara bebas di internet.
2. Data studi kasus terdiri dari 2 data berformat CSV, yaitu data *movies* dan data *ratings*.

1.5 Metodologi

Metodologi yang digunakan untuk penelitian ini antara lain:

1. Mempelajari konsep Spark, Hive, serta perintah SQL pada Spark SQL dan Hive.
2. Mempelajari bahasa pemrograman Scala.
3. Menginstalasi dan mengkonfigurasi Spark dan Hive *standalone* dan pada *cluster* Hadoop.
4. Mencoba perintah-perintah Spark dan Spark SQL.
5. Menulis program-program Spark SQL untuk mempelajari kemampuan Spark SQL.
6. Menulis skrip HiveQL untuk mempelajari kemampuan Hive.
7. Mencari dan mengumpulkan beberapa data studi kasus (berformat teks, CSV) dan menyimpannya ke HDFS.
8. Merancang antarmuka perangkat lunak.
9. Mengimplementasi perangkat lunak.
10. Menguji performa perangkat lunak dalam menganalisis *big data*.
11. Menulis dokumen skripsi.

1.6 Sistematika Pembahasan

- Bab 1 Pendahuluan
  Bab 1 membahas tentang latar belakang penelitian dilakukan, rumusan masalah, dan tujuan dari penelitian ini. Selain itu, batasan-batasan masalah dari penelitian yang dilakukan, metodologi penelitian, dan sistematika pembahasan dalam penelitian ini juga dijelaskan pada bab ini.

- Bab 2 Landasan Teori
  Bab 2 membahas tentang hasil studi literatur yang dilakukan. Bab ini terdiri dari penjelasan singkat mengenai *big data*, perintah-perintah, fungsi-fungsi dari bahasa Scala, deskripsi singkat SQL, basis data relasional, *primary* dan *foreign key*, serta operasi-operasi SQL. Selain itu, dijelaskan juga mengenai deskripsi dan komponen-komponen inti Hadoop, deskripsi dan operasi-operasi Hive, deskripsi dan komponen utama Spark, serta Spark SQL dan operasi-operasinya.

- Bab 3 Studi Eksplorasi
  Bab 3 membahas tentang cara instalasi dan konfigurasi Hive dan Spark, serta eksperimen-eksperimen Spark SQL dan Hive yang dilakukan. Eksperimen-eksperimen yang dilakukan berupa eksperimen sederhana melalui *console* maupun IDE.

- Bab 4 Analisis dan Perancangan
  Bab 4 membahas tentang analisis yang dilakukan mengenai perangkat lunak, data studi kasus, praolah data, dan kueri. Selain itu, bab ini juga berisi mengenai perancangan perangkat lunak dengan Spark beserta diagram kelas dan deskripsinya, serta perancangan perangkat lunak *web* beserta diagram *use case* dan diagram kelasnya.

- Bab 5 Implementasi dan Eksperimen
  Bab 5 membahas tentang deskripsi lingkungan perangkat yang digunakan, implementasi Spark dan Spark SQL, implementasi Hive, implementasi modul membaca kueri dan menulis hasil kueri ke HDFS. Selain itu, bab ini juga berisi pengujian kueri dan hasil kueri yang dilakukan. Eksperimen waktu eksekusi kueri dan kesimpulannya juga dibahas di bab ini.

- Bab 6 Kesimpulan dan Saran
  Bab 6 membahas tentang kesimpulan dari penelitian yang dilakukan. Bab ini juga berisi saran penulis untuk pengembangan perangkat lunak dengan Spark SQL selanjutnya.