SKRIPSI

STUDI DAN IMPLEMENTASI SPARK STREAMING UNTUK MENGUMPULKAN *BIG DATA STREAM*

Muhammad Ravi

NPM: 2016730041

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2020

UNDERGRADUATE THESIS

STUDY AND IMPLEMENTATION OF SPARK STREAMING TO COLLECT BIG DATA STREAM

Muhammad Ravi

NPM: 2016730041

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2020

PERNYATAAN

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

STUDI DAN IMPLEMENTASI SPARK STREAMING UNTUK MENGUMPULKAN *BIG DATA STREAM*

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 9 Juni 2020

Muhammad Ravi
NPM: 2016730041

LEMBAR PENGESAHAN

STUDI DAN IMPLEMENTASI SPARK STREAMING UNTUK MENGUMPULKAN *BIG DATA STREAM*

Muhammad Ravi

NPM: 2016730041

Bandung, 9 Juni 2020

Menyetujui,

Pembimbing

Dr. Veronica Sri Moertini

Ketua Tim Penguji	Anggota Tim Penguji
Pascal Alfadian, Nugroho, M.Comp.	**Natalia, M.Si.**

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 SPARK STREAMING UNTUK MENGUMPULKAN *BIG DATA STREAM*

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 9 Juni 2020

Muhammad Ravi
NPM: 2016730041

ABSTRAK

Big Data bukan hanya tentang data dengan ukuran yang besar. Tetapi, data juga dihasilkan dengan sangat cepat. Seiring perkembangan teknologi, internet, dan pengguna internet, data datang dari berbagai macam sumber yang berbeda-beda dengan format yang berbeda-beda pula. Selain itu data harus bisa ditransformasi dengan cepat untuk meningkatkan analisis dan keputusan bisnis. Sehingga data tidak bisa diolah secara konvensional lagi; menyimpan seluruh data dengan format yang berbeda pada *data lake*.

Dengan menggunakan konsep-konsep pengolahan data stream, proses ekstraksi data dari sumber, transformasi format data, dan penyimpanan data bisa dilakukan secara *real-time*. Konsep-konsep data stream akan diimplementasikan dengan *Apache Spark*, sebuah *framework* untuk mengolah data secara terdistribusi. Untuk mengolah data stream, *Apache Spark* akan menggunakan API *spark streaming* yang mengumpulkan data dengan potongan-potongan kecil (*windows*) dan dengan *Spark SQL* yang menambahkan aliran data sebagai baris baru pada tabel, *Structured Streaming*. Selain itu *Apache Spark* akan diintegrasikan dengan teknologi lain, Kafka, sistem yang digunakan untuk menghubungkan sistem-sitem lain dari berbagai sumber data dan meningkatkan performa *streaming*.

Penelitian ini dilakukan untuk mengumpulkan data stream dan membandingkan performa *Spark Streaming* ketika mengumpulkan data stream secara langsung dengan *Structured Streaming* yang terintegrasi dengan Kafka. Eksperimen dilakukan untuk melihat performa *Spark Streaming* dan *Structured Streaming* dalam mengumpulkan data stream.

Langkah-langkah penelitian untuk *Spark Streaming* adalah dengan mempelajari konsep data stream, mempelajari teknologi *Apache Spark* dan *Spark Streaming*, meng-instal *Apache Spark* pada klaster dan meneksplorasi dengan menggunakan API *Spark Streaming*; membuat *Dstream*; dan menghitung kata-kata yang masuk pada sistem; mengumpulkan data *twitter* secara real-time dengan *library* yang disediakan oleh *Spark*.

Langkah-Langkah penelitian untuk *Structured Streaming* hampir sama dengan *Spark Streaming*. Tetapi, ada langkah-langkah tambahan yaitu mempelajari teknologi *Apache Kafka* beserta konsepnya, menginstal Kafka dan Zookeeper pada klaster, meneksplorasi dengan membuat topik, menentukan partisi, dan membuat konektor, dan mengintegrasikan Kafka dengan *Structured Streaming*. Adapun, data yang dikumpulkan oleh Kafka ada dua yaitu data dari *twitter* dan data cuaca yang keduanya dikumpulkan oleh Kafka Connect sebelum ditulis ke topik dan diambil oleh *Structured Streaming*.

Bentuk dari perangkat lunak kedua eksperimen ini akan berupa jar-jar yang akan dieksekusi di klaster dan performanya bisa dilihat di antar muka web dan hasil eksekusi akan disimpan di HDFS

Kesimpulan pertama yang didapat dari penelitian ini adalah *Structured Streaming* jauh lebih baik dalam mengumpulkan data stream dibandingkan *Spark Streaming* karena *Structured Streaming processing time* maka pengumpulan jadi lebih cepat. Kesimpulan kedua yang didapat adalah, *Kafka* jauh lebih baik untuk mengumpulkan data stream karena bisa mengumpulkan data stream dari berbagai sumber dengan cepat dan *fault tolerant*. Tetapi, *Structured Streaming* baik dalam mentransformasi data tidak terstruktur menjadi terstruktur dan ekstraksi fitur.

Kata-kata kunci: *Big Data*, *Data Stream*, *Apache Spark*, *Spark Streaming*, *Structured Streaming*, Kafka

ABSTRACT

Big Data is not only about size.But also, data is coming really fast. As the internet and its user gradually grow, data comes in different shapes and sources. Data must be transformed quickly, so we can get insights and make business decision faster. Thus, data cannot be computed in a conventional way by dumping all of the data in data lake.

By using Streaming concepts, extracting data from sources, transforming data format, and loading the data can be done in real-time manner. Streaming Concepts will be implemented using Apache spark, cluster-computing framework that compute data in a distributed way. Apache Spark will use API, spark streaming that collect data by making small batches called window and Spark SQL treating each data stream as a row that continuously appended into larger table. Structured Streaming will aso be integrated to another system, kafka, that connecting many source and target systems and handling data transfer between system.

This research is done by collecting Data stream and then comparing the performance of spark streaming and structured streaming that integrated with kafka. The procedures taken to do this research are studying streaming concept such as processing time;event time;and windowing, studying cluster-computing framework such as *Apache Spark* and *Apache Kafka*, installing *Apache Spark* on cluster environment and exploring Spark Streaming by creating Dstream and counting words, installing Apache Kafka and exploring it by creating topics, set a partition, and creating connectors. The result of this research are executable jars that run on clusters. The output, performance graph and result, of this jars can be seen on a web interface.

In conclusion, structured streaming is far more efficient than spark streaming. Due to its lack of processing time. Thus, it compute data faster and more in real-time manner. Moreover, structured streaming is well integrated with another system and library such as MLlib and Kafka. When it comes to collecting data Kafka is better than spark or structured streaming because kafka is written and save in binary, it can collecting data from many system at once and different format. But, when it comes to transforming data and feature extraction structured streaming is better because it can deconstruct complex data format.

Keywords: Big Data, Data Stream, Apache Spark, Spark Streaming, Structured Streaming, Kafka

Untuk Ibu dan Ayah...

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Kuasa karena telah mengatur seluruh kejadian, kesempatan, dan hal-hal lain yang di luar kendali penulis berada pada pihak penulis. Sehingga, skripsi ini dapat selesai dan kerja keras yang selama ini penulis lakukan membuahkan hasil. Pada Kesempatan kali ini penulis ingin mengucapkan terima kasih kepada:

1. Kedua orang tua yang telah membiayai kuliah selama ini
2. Ibu Veronica Sri Moertini selaku dosen pembimbing yang telah banyak membantu, mengarahkan, dan memberi masukan selama proses pembuatan skripsi ini
3. Pak Kristopher David Harjono yang selalu bersedia menjawab pertanyaan jika ada kesulitan dan pelajaran *Data Mining*-nya yang sangat membantu.
4. Teman-teman seperjuangan kuliah dan skripsi
5. Orang-orang yang saya temui di internet yang membuat video dan artikel untuk menerangkan konsep-konsep
6. pihak-pihak lain yang terlalu banyak untuk disebutkan satu per satu yang telah membantu skripsi ini

Akhir kata, penulis menyadari bahwa skripsi ini tidak lepas dari kekurangan dan penulis berharap skripsi ini bisa membantu pembelajaran dan penelitian selanjutnya

Bandung, Juni 2020

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE PROGRAM

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Data Berkembang sangat cepat dalam beberapa tahun terakhir. Semakin banyak orang yang terhubung ke perangkat internet, semakin banyak juga akses terhadap website dan media sosial. Selain itu, sensor-sensor perangkat elektronik yang saling terhubung ke internet semua meninggalkan jejak digital berupa data. Data yang terakumulasi ini berukuran besar dengan format yang bervariasi dan aliran data yang datang sangat cepat. Big data akan lebih mudah untuk dianalisis ketika seluruh datanya telah terkumpul dan direduksi. Reduksi data adalah pengecilan ukuran data dengan mengambil rangkuman dari sekelompok data. Tetapi, masalah muncul ketika aliran data datang secara cepat, terus menerus, dan data tersebut harus diolah secara *real-time* untuk mendapatkan informasi. Masalah-masalah yang muncul tersebut antara lain; bagaimana data tersebut bisa ditangkap, dianalisis, dan disimpan.

Teknik yang digunakan untuk menangkap dan mengumpulkan aliran big data berbeda dari teknik mengumpulkan big data biasa karena pola aliran data berbeda, aliran data berupa potongan-potongan data yang datang secara terus-menerus tidak seperti big data biasa yang berupa file-file ukuran besar sekaligus.

Aliran data tersebut bisa langsung diolah dan dianalisis, banyak informasi-informasi bermanfaat yang bisa didapat. Contohnya, data bisa menjadi bahan pertimbangan untuk pengambilan keputusan bisnis. Tetapi, tidak semua data memiliki nilai dan sifat yang sama. Ada data yang memiliki nilai lebih ketika bisa langsung dianalisis ketika didapatkan. Kebutuhan untuk langsung mendapatkan dan menganalisis data secara *real-time*menjadi sangat penting. Selain itu, teknik pengumpulan data yang digunakan untuk pola data yang datang secara terus menerus berbeda dengan teknik yang digunakan untuk mengumpulkan dan mengolah data biasa. *Big Data* yang perlu diakses secara *real-time* adalah page views pada sebuah website, sensor pada IoT (*Internet of Things*.

Selain itu kebutuhan untuk mengolah data dengan cepat semakin penting karena nilai suatu data cenderung menurun seiiring bertambahnnya waktu. Banyak Perusahaan dan Organisasi yang membutuhkan data untuk diolah secara cepat. Semakin cepat data bisa diambil, dianalisis, dimanipulasi, dan semakin banyak througput yang bisa dihasilkan maka sebuah organisasi akan lebih *agile* dan responsif. Semakin sedikit waktu yang digunakan untuk ETL (*Extract, Load, Transform*) pekerjaan akan semakin fokus untuk melakukan analisis bisnis.

Untuk menjawab masalah di atas, *Spark Streaming* merupakan teknologi yang menjadi salah satu solusi terhadap adanya kebutuhan untuk menganalisis *big data* secara *real time*. Data hasil streaming kemudian dapat dianalisis dengan teknik-teknik komputasi sederhana dan divisualisasikan agar lebih mudah dimengerti.

1.2 Rumusan Masalah

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

1. Bagaimana Karakteristik *data stream* dan contoh-contoh analisisnya?

2. Bagaimana cara kerja *Spark Streaming*?
3. Bagaimana cara mengintegrasikan *Spark Streaming* untuk mengumpulkan data?

1.3 Tujuan

Berikut ini adalah tujuan yang ingin dicapai pada skripsi ini.

1. Melakukan studi tentang definisi, teknik pengumpulan, arsitekstur, dan manfaat analisis dari data stream
2. Mempelajari konsep, arsitektur, cara kerja Spark Streaming dan integrasinya dengan teknologi-teknologi lain
3. Mengimplementasikan *Spark Streaming* pada sebuah sistem untuk mengumpulkan data stream dengan kasus-kasus tertentu.

1.4 Batasan Masalah

Batasan masalah untuk skripsi ini adalah sebagai berikut.

1. Data yang diolah bisa berubah dan memiliki batasan akses sesuai penyedia data tersebut
2. Data akan diolah secara terdistribusi pada 10 komputer saja.
3. Pengembangan hanya fokus di ekstraksi data stream dan analis data stream agregasi saja tidak sampai pembuatan model *machine learning*

1.5 Metodologi

Metodologi yang digunakan dalam pembuatan skripsi ini adalah sebagai berikut.

1. Mempelajari pola, arsitektur, dan sumber dar *Big Data Stream.*
2. Mempelajari arsitektur, cara kerja, dan komponen-komponen *Spark.*
3. Mempelajari Distribusi data pada *Hadoop distributed file System.*
4. Mempelajari arsitektur dan cara kerja *Spark Streaming* pada *Spark.*
5. Mempelajari Bahasa pemrograman *Scala.*
6. Mempelajari *Kafka* dan *Structured Streaming.*
7. Mengimplementasikan *Spark Streaming* pada klaster.
8. Mengintegrasikan *Structured Streaming* dengan Kafka.
9. Melakukan eksperimen dan pengujian *Spark Streaming* dan *Structured Streaming* pada klaster

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 penilitian, dan sistematika pembahasan.

2. Bab Landasan Teori
   Bab 2 membahas tentang teori-teori mengenai *Big Data*,*Big Data Stream*, Sistem terdistribusi *Spark*, *Spark Streaming*, dan *Kafka*.

3. Bab Studi Eksplorasi
   Bab 3 membahas tentang langkah-langkah untuk melakukan konfigurasi klaster pada *hadoop*, konfigurasi klaster untuk *Spark* dan *kafka*, hasil studi eksplorasi *Spark Streaming*.

4. Bab Analisis dan Perancangan
   Bab 4 membahas tentang analisis perangkat lunak *Spark*, analisis data uji, analisis masukan dan keluaran, rancangan Arsitektur, dan diagram kelas.

5. Bab Implementasi dan Eksperimen
   Bab 5 membahas tentang implementasi perangkat lunak, eksperimen performansi 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 lanjut.