Studi dan implementasi *Spark Streaming* untuk Mengumpulkan *Big Data Stream*

Muhammad Ravi–2016730041

1 Data Skripsi

Pembimbing utama/tunggal: **Veronica Sri Moertini**
Pembimbing pendamping: -
Kode Topik : **VSM4702**
Topik ini sudah dikerjakan selama : **1** semester
Pengambilan pertama kali topik ini pada : Semester **40 - Ganjil 19/20**
Pengambilan pertama kali topik ini di kuliah : **Skripsi 1**
Tipe Laporan : **B -** Dokumen untuk reviewer pada presentasi dan **review Skripsi 1**

2 Latar Belakang

Dalam beberapa tahun terakhir,Perkembangan data melonjak secara cepat. Hal ini disebabkan karena semakin banyak orang yang mengakses ranah digital. Website yang diakses, media sosial yang dijelajahi, atau sensor-sensor dari barang-barang elektronik yang terhubung ke internet semua meninggalkan jejak digital berupa data. Data yang terakumulasi ini berukuran besar dengan format yang bervariasi dan berkembang dengan sangat cepat. Jika data yang terakumulasi tersebut diolah dan dianalisis, banyak informasi-informasi bermanfaat yang bisa didapat. Contohnya, data bisa menjadi bahan pertimbangan untuk pengambilan keputusan bisnis. Tetapi masalahnya bukan hanya itu saja, Tidak semua data memiliki nilai yang sama. Ada data yang memiliki nilai lebih ketika bisa langsung dianalisis ketika didapatkan. Spark Streaming merupakan teknologi sebagai solusi terhadap adanya kebutuhan untuk menganalisis big data secara real time. Big data yang perlu dianalisis secara real-time misalnya: page views (data klik). Data hasil streaming kemudian dapat dianalisis dengan teknik-teknik analisis data berbasis statistik maupun machine learning/data mining.

3 Rumusan Masalah

- Bagaimana Karakteristik *datastream* dan contoh-contoh analisisnya?
- Bagaimana cara kerja *Spark Streaming* untuk mengumpulkan *datastream*?
- Bagaimana cara mengintegrasikan *Spark Streaming* dengan sistem pengumpul data lain?
- Bagaiaman cara menganalisis data yang telah terkumpul?

4 Tujuan

- Melakukan studi tentang definis, pola-pola, arsitekstur, dan manfaat analisis dari data stream
- Mempelajari konsep, arsitektur, cara kerja Spark Streaming dan integrasinya dengan sistem lain seperti; *Kafka, Flume, atau Kinesis.*
- Mengimplementasikan Spark Streaming pada sebuah sistem untuk mengumpulkan data stream dengan kasus-kasus tertentu.
- Menganalisis data yang sudah terkumpul

5 Detail Perkembangan Pengerjaan Skripsi

Detail bagian pekerjaan skripsi sesuai dengan rencan kerja/laporan perkembangan terkahir :

1. Pada Bab I: Pendahuluan, saya mempelajari dan menulis latar belakang dari skripsi ini, mengapa topik ini diangkat dan gambaran besar tentang topik ini, merumuskan masalah apa yang ingin dijawab dan menentukan tujuan, proses, yang ingin dicapai. Serta memberi batasan masalah terhadap topik ini.

2. Bab 2: landasan Teori mempelajari dan menulis hal-hal apa saja yang akan menjadi dasar untuk pengerjaan skripsi di Bab 4 nanti.

3. Selama bab 2 saya mempelajari Karakteristik dan masalah-masalah yang muncul dari *Big Data*. Lalu, saya belajar tentang bentuk Big data yang lain yaitu; Big data Stream atau data yang datang secara real-time dan terus menerus. Karena itu perlu ada penanganan khusus berupa penerapan arsitektur.

4. Arsitektur yang diterapkan pada data stream tidak mengenal satuan waktu, artinya data yang pertama masuk ke sistem hanya akan diberi timestamp terlepas waktu tersebut sama dengan waktu awal data dibuat. Penerapan timestamp bertujuan untuk memudahkan agregasi pada sistem. Sistem yang digunakan harus bisa menganalisis data secara real time

5. Untuk mencapai hal tersebut, ada 2 arsitektur yang bisa digunakan untuk menangani data stream yaitu arsitektur kappa dan arsitektur lambda. Arsitektur kappa sepenuhnya menggunakan real-time processing. Sedangkan arsitektur lambda menggunakan real-time dan batch-processing untuk meningkatkan akurasi.

6. Agar bisa menerapkan kedua arsitektur tersebut saya mempelajari hadoop dan map reduce dan Spark yang memungkinkan untuk memproses data dengan cepat dan fault tolerant

7. Terakhir di Bab 2 saya mempelajari Spark Streaming, yaitu sebuah API Spark yang bisa memproses data yang besar secara real-time. Spark Streaming bisa berintegrasi dengan sistem pengumpul data lain atau sumber data.

8. Pada Bab 3: Eksplorasi, saya mulai menyetel environment untuk spark streaming; seperti menginstak hadoop, spark, zookeeper, dan kafka, serta mencoba mengintegrasikan data dengan twitter, kafka, dan Amazon kinesis. Serta mengolah data dengan operasi- operasi sederhana seperti menyimpan tweet ke HDFS dan menghitung banyaknya log.

9. Pada bab 4: Analisis, saya mencoba menerapkan teori bab 2 dan melakukan analisis pada kasus-kasus seperti menghitung hashtag dan menghitung jumlah error pada weblogs

10. terakhir saya sedang menulis sebagian Bab 2 dan 3.

6 Pencapaian Rencana Kerja

Langkah-langkah kerja yang berhasil diselesaikan dalam Skripsi 1 ini adalah sebagai berikut:

1. Mempelajari teknik-teknik analisis dan manfaat analisis dari *datastreams*

2. Mempelajari konsep-konsep, arsitektur, dan cara kerja *Spark Streaming*

3. Mengintegrasikan *Spark Streaming* dengan sistem lain yaitu; *Twitter, Kafka, dan TCP Socket* untuk mengambil *datastream*

4. menulis sebagian dokumen skripsi Bab 1, Bab 2, Bab 3

7 Kendala yang Dihadapi

Kendala - kendala yang dihadapi selama mengerjakan skripsi :

- Kesulitan meng-instal environment di PC.
- Sering ada versi yang tidak saling cocok.
- terlalu banyak environment yang di-instal
- Mengalami kesulitan untuk menerapkan kode karena bahasa baru.

Bandung, 20/11/2019

Muhammad Ravi

Menyetujui,

Nama: Veronica Sri Moertini
Pembimbing Tunggal