BAB 6

KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari awal hingga akhir penelitian beserta saran untuk penelitian selanjutnya.

6.1 Kesimpulan

Berdasarkan penilitan Spark Streaming, Structured Streaming, dan Kafka, berikut kesimpulan yang dapat diambil:

1. Data stream adalah data yang tidak pernah berhenti dan datang terus menerus. Manfaat dari pengolahan data stream adalah bisa mentrasnformasi dan mengambil informasi dari aliran data secara real-time karena itu hal yang paling penting untuk dicapai adalah ketepatan data. Data yang diterima label waktunya harus sesuai. Teknik pengumpulan yang baik digunakan adalah *unbounded processing* karena *micro-batch* langsung memproses data yang masuk tanpa membuat *batch* kecil-kecil terlebih dahulu dan bisa melakukan *event-time processing*. Sehingga label waktu yang tersimpan tepat karena tidak terjadi delay. Tidak ada desain arstiektur universal untuk mengumpulkan data stream. Tapi, harus mengerti masalah dan kebutuhan apa yang ingin diselesaikan, data apa yang ingin dikumpulkan, format seperti apa, bagaimana transformasinya, dan data aka disimpan di mana. Sehingga Arsitektur optimal dan efisien bisa diterapkan.

2. Spark Streaming cocok untuk melakukan monitoring terhadap data yang datang secara real-time dan mengamati perkembangan data. Karena berdasarkan eksperimen, data yang ditransformasi hasilnya langsung terlihat dan *Insight* bisa langsung terlihat pada perkembangan datanya. Tetapi, perangkat lunak harus terus diawasi karena data diolah berdasarkan waktu masuk data ke sistem sehingga waktu yang tersimpan tidak akurat dan tidak bisa mengatasi data yang terlambat. Agregasi data terbatas tidak bisa melakukan perhitungan kompleks secara real-time. Lalu, ketika dijalankan berhari-hari tingkat akurasi pengelompokkan data berdasarkan waktu terus menurun karena terjadi delay. Karena itu, Spark Streaming tidak cocok untuk melakukan transformasi kompleks terhadap aliran data yang sangat besar.

3. Structured Streaming yang diintegrasikan dengan Kafka cocok melakukan transformasi terhadap data yang memiliki skema rumit(*preprocessing*), seperti mengubah data tidak terstruktur menjadi terstruktur. Hal ini disebabkan karena Structured Streaming mengolah data dengan dataframe yang berbasis SQL sehingga perintah-perintah SQL bisa digunakan secara real-time. Structured Streaming juga bekerja dengan sistem trigger, data yang masuk ke sistem akan langsung dikelompokkan berdasarkan waktu ketika data itu dibuat. Sehingga bisa mengatasi data yang terlambat dan melakukan *preprocessing* terhadap aliran data dengan *throuhput* yang besar.Bisa berjalan selama berhari-hari atau berbulan-bulan tanpa pengawasan. Tetapi, tidak cocok untuk monitoring data secara real-time karena hasil transformasi tidak ditampilkan perkembangannya. Biasanya data di transformasi dulu formatnya menjadi terstruktur dan akan dianalisis kemudian.

6.2 Saran

Berdasarkan eksperimen dan kesimpulan di atas, berikut saran yang dapat penulis berikan:

1. Data Stream Twitter yang digunakan menggunakan akun Twitter *Free* yang hanya menyediakan 1% sampel data stream dari keseluruhan data stream. Akan lebih baik jika akun Twitter di-*upgrade* menjadi akun *premium* yang menyediakan 10% sampel data stream dari keseluruhan data stream dan menyediakan data-data tambahan yang tidak tersedia di akun *Free*

2. Data stream dari konektor Kafka throughputnya masih sedikit karena masih dari sumber yang terbatas. Untuk melihat kemampuan Kafka dan structured streaming, akan lebih baik, jika membuat konektor yang lebih banyak dan variatif, mengambil data dari sumber yang berbeda dengan format dan kecepatan yang berbeda dan menjalankan konektor-konektor tersebut secara paralel.

3. Penelitian tidak dilakukan sampai pemodelan *machine learning* akan lebih baik jika pada penelitan selanjutnya melakukan pemodela *machine learning* secara real-time.

DAFTAR REFERENSI

[1] Kapil, G., Agrawal, A., dan Khan, P. R. (2016) A study of big data characteristics, . 10, pp. 1–4.

[2] Akidau, T. (2018) Streaming 101. Bagian dari Roumeliotis, R. dan Bleiel, J. (ed.), *Streaming System*. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[3] Akidau, T. (2018) The what, where,when, and how of data processing. Bagian dari Roumeliotis, R. dan Bleiel, J. (ed.), *Streaming System*. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[4] Miner, D. dan Shook, A. (2013) *Map Reduce Design Pattern*, 1st edition. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[5] Wampler, D. dan Payne, A. (2009) *Programming Scala*, 1st edition. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[6] Karau, H., Konwinski, A., Wendell, P., dan Zaharia, M. (2015) Introduction to data analysis with spark. Bagian dari Spencer, A. dan Beaugureau, M. (ed.), *Learning Spark*. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[7] Karau, H., Konwinski, A., Wendell, P., dan Zaharia, M. (2015) Programming with rdds. Bagian dari Spencer, A. dan Beaugureau, M. (ed.), *Learning Spark*. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[8] Karau, H., Konwinski, A., Wendell, P., dan Zaharia, M. (2015) Spark streaming. Bagian dari Spencer, A. dan Beaugureau, M. (ed.), *Learning Spark*. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[9] Narkhede, N., Shapira, G., dan Palino, T. (2017) Meet kafka. Bagian dari Spencer, A. dan Beaugureau, M. (ed.), *Kafka: The Definitive Guide Real-Time Data and Stream Processing at Scale*. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[10] Narkhede, N., Shapira, G., dan Palino, T. (2017) Kafka producers: Writing messages to kafka. Bagian dari Spencer, A. dan Beaugureau, M. (ed.), *Kafka: The Definitive Guide Real-Time Data and Stream Processing at Scale*. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[11] Narkhede, N., Shapira, G., dan Palino, T. (2017) Kafka consumers: Reading data from kafka. Bagian dari Spencer, A. dan Beaugureau, M. (ed.), *Kafka: The Definitive Guide Real-Time Data and Stream Processing at Scale*. O?Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.