IF4044 Teknologi Big Data
Tugas Proyek Big Data

Semester 2 2023/2024

Hari/tanggal release: Selasa, 16 April 2023

Requirement

Dalam proyek ini, Anda akan membangun arsitektur big data menggunakan konsep dan teknologi yang dipelajari di kuliah Big Data.

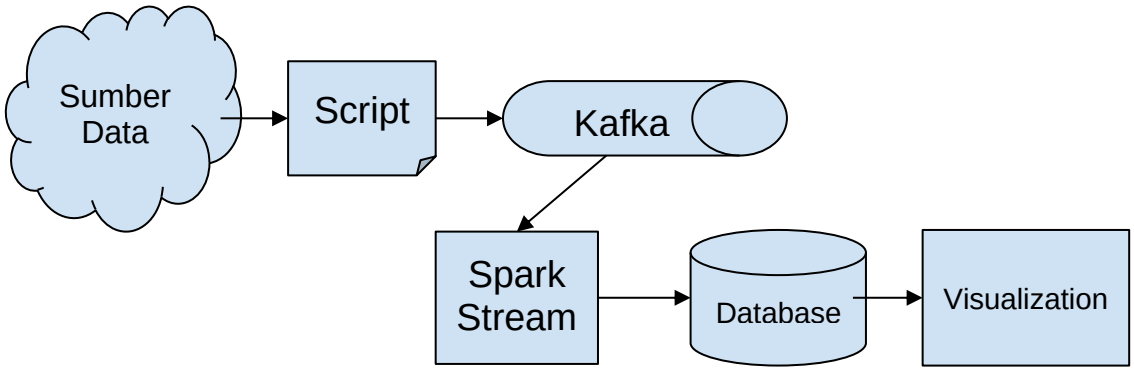
Arsitekturnya mencakup:

- Streaming menggunakan Kafka
- Stream Processing menggunakan Spark Streaming
- Data lake / data warehouse: Penyimpanan data dan query-nya
- Data visualization: Menampilkan datanya dalam bentuk visualisasi.

Sumber Data: Gunakan salah satu sumber data di halaman ini:
https://github.com/bytewax/awesome-public-real-time-datasets

Alur data:

1. Program melakukan pengambilan data dari sumber yang Anda pilih, lalu menuliskan ke Kafka secara *streaming*.
2. Pemrosesan data dari Kafka dengan menggunakan Spark Streaming dengan API RDD, dengan window time 1 menit (atau window time lain yang menurut Anda pantas dilakukan sesuai studi kasus)
3. Spark Streaming menuliskan hasil pemrosesan ke database dengan jadwal setiap 5 menit (atau jadwal lain yang menurut Anda pantas dilakukan sesuai studi kasus)
4. Menampilkan data hasil pemrosesan dalam visualisasi



Teknologi yang digunakan:

- Kafka sebagai *streaming db*
- Spark sebagai *stream processing*
- Database pilihan Anda.
- Visualisasi pilihan Anda (lihat opsi di bagian ‘Resource’, jika perlu).

Bonus poin:

- Menggunakan teknologi data warehouse sebagai teknologi basis data.
- Melakukan *machine learning*, misalnya prediksi atau *forecasting*.

Contoh kasus (hanya contoh saja):
Melakukan polling data twitter untuk keyword tertentu, tulis ke Kafka setiap 5 detik. Setiap 1 menit melakukan kalkulasi count untuk setiap kata kunci, kemudian dituliskan ke database setiap 5 menit. Melakukan visualisasi hashtag apa yang paling banyak dibicarakan, serta melakukan analisis sentimen pada tweet-nya (*machine learning classification*).

Deliverables

Berikut yang perlu dikumpulkan.

- Laporan - **kelompok**
  - Penjelasan terkait data sumber dan data final hasil pemrosesan, apa yang ingin ditunjukkan. Contoh: tren perubahan suhu selama 1 minggu
  - Penjelasan alur secara umum serta arsitektur yang digunakan
  - Penjelasan masing-masing komponen arsitektur dan alasan konfigurasinya
  - Lessons learned yang dipelajari selama implementasi dan eksplorasi
- Kode - **kelompok**
  - Instruksi instalasi dan eksekusi kode (README.txt)
  - Semua kode yang digunakan untuk menjalankan
- Demo (rekaman) maks. 10 menit - **individu, setiap orang masing-masing merekam**
  - Menunjukkan pemrosesan (bila lama bisa dipotong bagian menunggunya)
  - Menunjukkan hasil pemrosesan dengan query
  - Mendemokan visualisasi datanya

Pengumpulan:

- folder ‘laporan’ berisi file laporan
- folder ‘kode’ berisi readme file dan kode yang digunakan
- folder ‘presentasi’ berisi sebuah txt yang isinya link rekaman presentasi. Silakan gunakan YouTube unlisted.

Seluruh folder tersebut dizip dan dikumpulkan melalui LMS Edunex, pada modul/minggu 15.

Komponen Penilaian

- Solusi: 50%
    - Tugas selesai dan deliverables lengkap sesuai requirement
    - Hasil dapat diakses dan didemokan
    - Kodenya benar sesuai pemaparan (code review)
    - Demo video
- Laporan & Presentasi: 30%
    - Penjelasan dari arsitekturnya
    - Penjelasan masing-masing komponen dan konfigurasinya
    - Sintesis lessons learned
- Peer review: 20%
    - Masing-masing anggota menilai anggota lainnya (akan diberikan dalam bentuk survey)
- Bonus poin seperti disebutkan sebelumnya. Bonus poin akan menambah poin hanya bila ada yang kurang dari komponen di atas.

Deadline

Pukul 12.00 WIB, hari Senin, tanggal 13 Mei 2024

Presentasi

Presentasi tugas dilakukan pada hari Jumat, tanggal 17 Mei 2024, pukul 13.15

FAQ

- Apakah boleh jika dijalankan-nya di local laptop?
    - Boleh, selama bisa didemokan.
- Apakah boleh mengasumsikan use case untuk memilih arsitektur yang tepat?
    - Boleh, mohon asumsi dituliskan di report.

Resources

Sebagai petunjuk saja, silakan gunakan teknologi lain bila perlu.

- Cloud Data Warehouse
    - Snowflake free : https://signup.snowflake.com/
    - Bigquery free (sandbox) : https://cloud.google.com/bigquery/docs/sandbox
- Alat visualisasi
    - Metabase (install local /self host) : https://www.metabase.com/
    - Zeppelin (install local / self host) : https://zeppelin.apache.org/
    - Looker Studio free: https://lookerstudio.google.com/
    - Google colab free: https://colab.research.google.com/