Brief Soal Data Challenge eFishery Batch 2

Mohon kerjakan salah satu soal data challenge sesuai dengan posisi yang Anda pilih, berikut soal data challenge:

Soal untuk posisi Data Engineer

Anda diminta untuk membuat aplikasi command line yang menggenerate data untuk disimpan dalam format file dengan ketentuan value column:

contoh: column_name: column_type

- id: uuidv4
- device_id: value diantara [00010...0001F] (hexadesimal counter)
- username: value diantara ['Andi', 'Budi', 'Taja']
- lokasi: value diantara ['Bandung', 'Jakarta']
- amount: value diantara [10...1000]
- timestamp: UNIX timestamp value diantara 01/01/2019-31/12/2019

Data berjumlah > 100.000 rows minimumnya. Format data dibebaskan bentuknya, ekstensinya ataupun representasinya (bisa berupa bytes, string, atau kodifikasi sendiri). Perlu diperhatikan ukuran file atau ratio jumlah data/ukuran file untuk optimasi. Semakin kecil, semakin bagus, karena tidak akan memakan banyak memori atau overhead I/O.

===== RELASI DATA (untuk generating data)

Andi (Bandung):

00010, 00011, 00012, 00013, 00014, 00015

Budi (Jakarta):

00016, 00017, 00018, 00019, 0001A, 0001B

Taja (Bandung):

0001C, 0001D, 0001E, 0001F

=====

Dari file yang sudah digenerate Anda, buatlah sebuah aplikasi CLI yang dapat membaca format file tersebut dan dapat menampilkan datanya. Fitur menampilkan data yang wajib ada:

- menampilkan 100 data awal (terurut dari timestamp)
- menampilkan 100 data terakhir (terurut dari timestamp)
- menampilkan data berdasar filter: device_id, username, lokasi, timerange (limit 100 data awal/akhir)
- menampilkan data berdasar specific rows_number (misal data urutan ke 102)
- menampilkan nilai agregasi data nilai amount grouped by (all, device_id, username, lokasi):
    - --- nilai max
    - --- nilai min
    - --- nilai avg
    - --- nilai sum
    - --- nilai count of rows
    - --- (opsional) menampilkan data dalam bentuk timeseries di granularity (weekly, monthly)

Goals:

1. Aplikasi CLI untuk menggenerate data
2. Aplikasi CLI untuk membaca data
3. Aplikasi disubmit ke repo github publik masing-masing dan sertakan link Githubnya untuk submit hasilnya (tanpa data hasil generate)

Hal yang perlu diperhatikan:

1. Besar ukuran file data
2. Efisiensi dalam pembacaan data
3. Commit message menggunakan format Karma (http://karma-runner.github.io/4.0/dev/git-commit-msg.html)
4. README untuk penggunaan dan penjelasan struktur data yang digunakan

 # Soal untuk posisi Data Analyst

Anda diberikan Dataset terkait perikanan nasional (data-perikanan-nasional.zip). Galilah insight dan buat report (dashboard/grafik/visualisasi). Contoh report & insight, semisal:

Report:

- Komoditas apa yang termasuk tertinggi dalam hasil budidaya?
- 3 daerah yang termasuk produsen tertinggi?
- ...lainnya silakan kreativitas Anda

Insight (pendalaman dari report):

- Kenapa daerah produsen itu tertinggi, apa komoditasnya dan karakternya?
- Korelasi antara produktivitas budidaya ikan dengan komoditas daerah penghasil ikannya?
- ...lainnya silakan kreativitas Anda

Anda perlu untuk menyatukan dan normalisasi data-data tersebut dahulu agar bisa diolah dengan semestinya. Anda dibebaskan eksplorasi yang memberikan trivia/insight dan korelasi yang menarik untuk dimasukkan ke laporan tertulis.

Goals:

1. File hasil normalisasi/ integrasi data
2. Hasil eksplorasi ditulis dalam format file dokumen teks digital atau JupyterNotebook/ RPubs

Hal yang perlu diperhatikan:

1. Hasil normalisasi data
2. README terkait struktur data hasil normalisasi dan integrase
3. Eksplorasi report analisis Anda dan kedalaman insightnya dari keterhubungan semua data