 # Bulk Insert

kadang kita punya data duluan, baik data dari open data maupun dari aplikasi lain yang kita miliki sebelumnya dan mau dimigrasikan ke ES. sangat tidak praktis kalau datanya kita masukkan satu per satu menggunakan cara di atas. karena itu di ES disediakan fungsi insert secara bulk (langsung banyak). Baik dituliskan datanya sebagai badan pesannya maupun dari file.

ada file yang bisa digunakan (account.zip), bisa didownload dari https://download.elastic.co/demos/kibana/gettingstarted/7.x/accounts.zip (ukurannya 50an Kb, saat diextract jadi 250an Kb.

(buka file accounts.json)

data akunnya punya skema seperti ini :

{
"account_number": INT,
"balance": INT,
"firstname": "String",
"lastname": "String",
"age": INT,
"gender": "M or F",
"address": "String",
"employer": "String",
"email": "String",
"city": "String",
"state": "String"
}

isinya adalah identitas akun bank seseorang (data dummy), dari identitas pemiliknya, tempat tinggal, hingga saldo yang dimilikinya

data seperti ini bisa dimanfaatkan untuk meningkatkan penjualan. misalnya kita punya produk di daerah tertentu, dan demografi marketnya jelas, maka kita bisa secara spesifik menawarkan ke beberapa orang yang sudah terfilter. Daripada blasting ke semua orang (effort dan dana besar). Sales yang tertarget lebih murah secara operasional dan juga peluang konversinya lebih besar.

Contoh, produk properti. Misalnya saya pengembang perumahan baru, yang targetnya eksekutif muda, usia 25-40 tahun yang sudah berkeluarga. Dari data yang dimiliki bisa dikerucutkan targetnya ke beberapa orang untuk diprospek lebih jauh.

Untuk memasukkan data dengan banyak sekaligus, kita pakai fungsi ***_bulk***

```
$ curl -H 'Content-Type:application/x-ndjson' -XPOST 'http://localhost:9200/bank/account/_bulk?pretty' --data-binary @accounts.json
```

Figure 1: bulk insert menggunakan curl

itu ada beberapa bagian :
- content type kita state berupa x-ndjson
- metodenya POST
- kita post datanya ke index bank doc_type account , makanya dikirim ke alamat /bank/account/
- nama file yang mau dimasukkan (accounts.json) ditaruh sebagai parameter dari --data-binary

kalau pakai postman tipe datanya diganti, bukan raw tapi binary. dan link ke filenya diberikan di situ

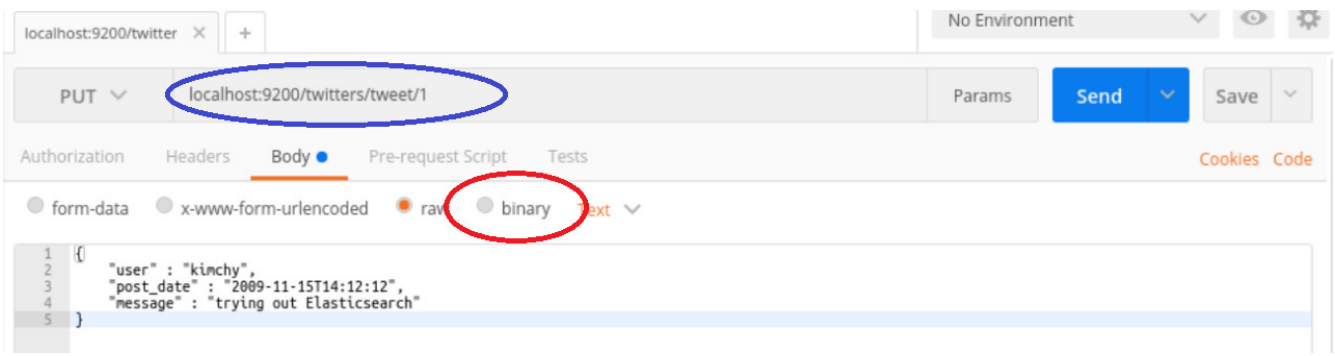

*Figure 2: bulk insert dengan postman. dipilih yang binary (lingkaran merah) dan tuliskan path ke filenya. jangan lupa alamat tujuannya diubah ke *http://localhost:9200/bank/account/_bulk* (lingkaran biru)*

di sini silahkan coba dimasukkan dulu datanya. Kalau ada kesulitan boleh dicari penyelesaiannya di internet/ditanyakan di grup. Karena setelah ini akan kita gunakan untuk mempelajari fungsi search dan query DSLnya.

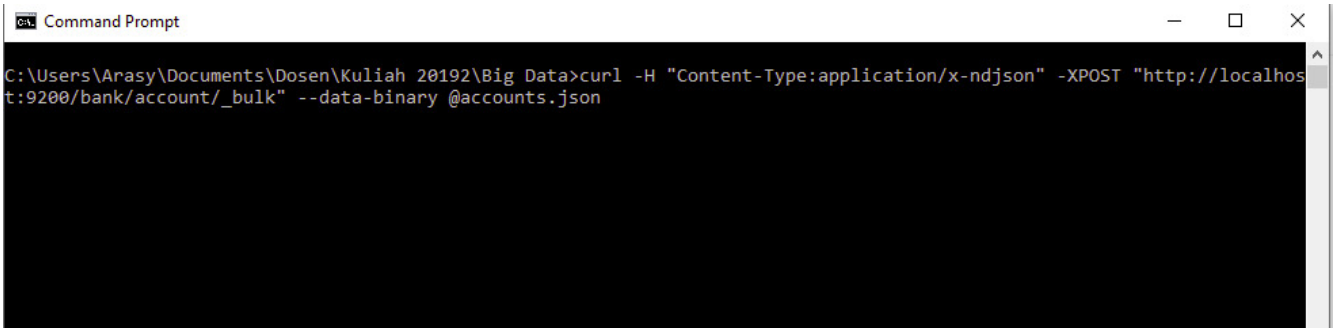

Figure 3: memasukkan data bulk menggunakan curl di windows. kalo di windows, pakenya petik dua ya, bukan petik satu

Figure 4: reply dari servernya. menandakan datanya sukses diinput. berantakan karena tidak di prettify

kalau mau mastiin lagi datanya beneran masuk atau nggak, coba akses localhost:9200/bank/account/_search . Nanti akan muncul datanya

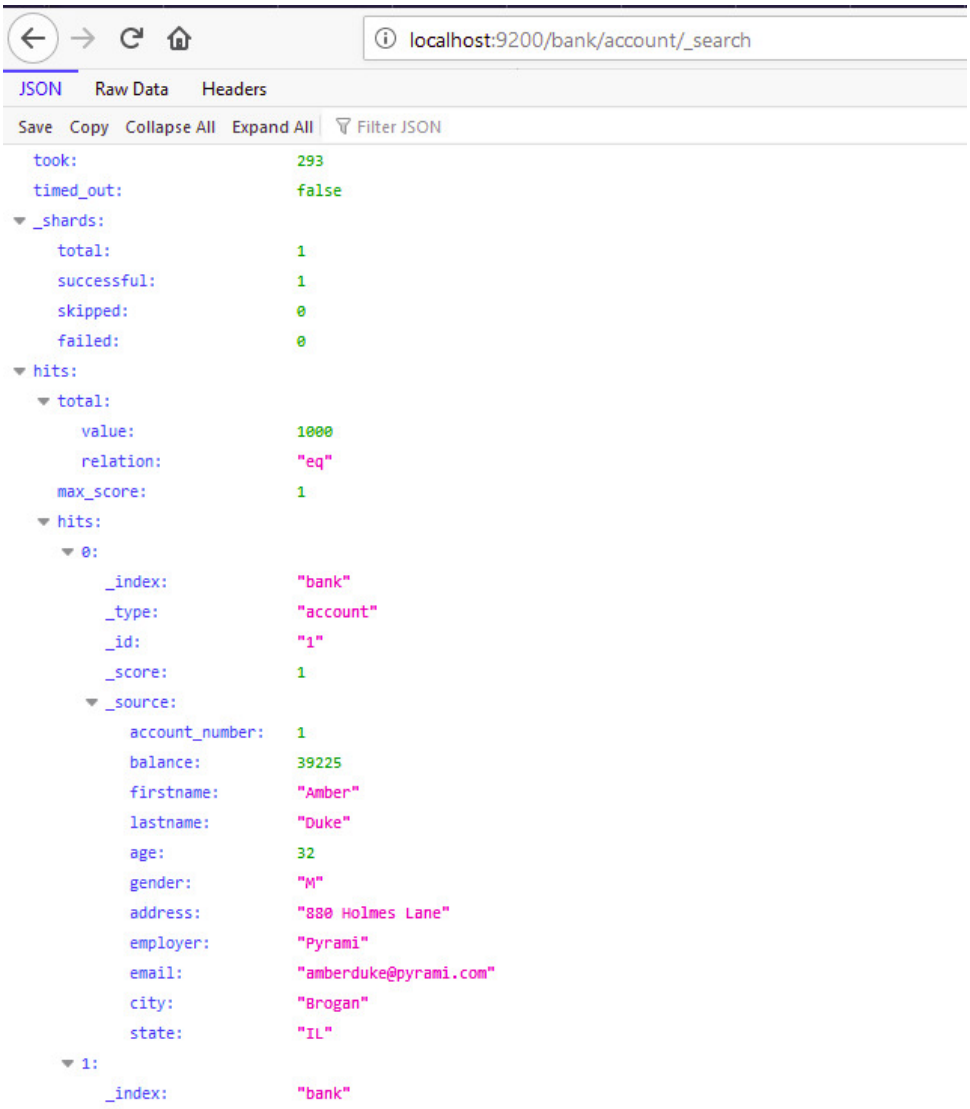

Figure 5: hasil akses ke localhost:9200/bank/account/_search

kalau munculnya monochrome (hitam putih) atau berantakan gak masalah. Yang penting datanya betul. Itu browser/client REST menampilkan datanya raw. Kalau mau nampilin bentuk json dengan rapih, bisa download addon untuk json viewer.

Setelah datanya dipastikan masuk, coba gunakan untuk mencari data tertentu, misal :
1. cari akun yang ada kaitannya dengan John (bisa namanya, nama keluarganya, atau yang lainnya adalah john)
2. cari akun yang nama depannya John
3. cari orang yang punya balance di atas 40.000
4. berapa banyak dari 1000 data tersebut yang laki-laki dan perempuan?

Tidak harus John ya, bisa eksperimen dengan nama/kriteria yang lain. Penjelasannya ada di slide dan grup wa pekan lalu.

Query DSL

kadang kita butuh pencarian yang lebih mendetail yang tidak bisa disupport dengan inline query seperti di atas. Karena itu kita butuh mekanisme lain.

misal :
saya mau tahu siapa aja laki laki yang usianya kurang dari 35 tahun, yang tinggalnya di kota Sunnyside atau Chestnut, yang uangnya (balance) di atas 40.000

kita bisa manfaatkan query DSL di sini. Penjelasan singkat tentang query DSL ada di halaman 29. yang di slide sebagian besar sudah dijelaskan (ngejelasin ulang sama aja kayak ngeprintscreen slidenya). tinggal dicobain biar kelihatan efek pencariannya seperti apa. Kalau butuh penjelasan lebih lanjut bisa lihat di web elasticsearchnya dulu.

paling yang saya tambahkan adalah tentang **fuzzy search**

kadang kita nyari kata-kata tertentu, tapi yang kita ketik dan kita maksudkan berbeda. maksudnya mau ngetik John, tapi ketulis posisi hurufnya kebalik jadi jhon. Maksudnya beras tapi ada huruf yang gak keketik jadi bers. Dll

kalo kita pakai fuzzy, servernya akan mencari kata yang kemiripannya tinggi dengan apa yang kita inputkan. jadi meskipun kita ngetiknya Jhon, yang John akan ikut dianggap betul. Kalo nyarinya Ahmad, servernya akan mengembalikan Achmad, Akhmad, dan Ahmat juga. Bahkan Rahmat juga bisa saja dimunculkan.

ini memanfaatkan Levenshtein Distance (dipelajari di MK NLP). Detail penggunaannya bisa dibaca di https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-fuzzy-query.html

Note :
kalau data account dirasa kurang banyak, saya pernah dapat data resep masakan dalam bentuk json juga. Ukurannya lebih dari 100MB. Cuma sumbernya lupa, jadi gak bisa ngasih link asalnya. Kalau nanti ada yang nemu atau nemu data lain yang menarik boleh juga dikasih tahu ke saya.

Data resep bisa didownload di
https://drive.google.com/file/d/1NhtqmwpxWMURJ7EZtcuIKzrMLZsWUo5C/view

bisa dicoba insert ke server lalu dicari hal tertentu. Apa yang sudah dilakukan bisa dishare di grup.