**Latar belakang**

Lebih dari perusahaan lain, Google mungkin bertanggung jawab untuk memperkenalkan kami pada manfaat menganalisis dan menafsirkan Data Besar dalam kehidupan kita sehari-hari.

Saat kami melakukan pencarian Google, kami memanipulasi Big Data. Ukuran indeks Google - arsipnya dari setiap halaman Web yang dapat ditemukannya, yang digunakan untuk mengembalikan hasil pencarian - diperkirakan mencapai sekitar 100 petabyte (atau 100 juta gigabytes!) - tentu saja Big Data, menurut standar siapa pun.1

Tetapi seperti yang telah kita lihat selama dekade terakhir, menyatukan semua informasi di Internet untuk memudahkan kita menemukan sesuatu hanyalah awal dari rencana mereka. Google telah meluncurkan browser Web, email, sistem operasi ponsel dan jaringan periklanan online terbesar di dunia - semuanya dengan kuat berpijak pada teknologi Big Data yang dengannya mereka menjadikan diri mereka sebagai nama keluarga.

**Apa Masalah Big Data Yang Membantu Untuk Dipecahkan?**

Internet adalah tempat yang besar - karena kami bergerak online secara massal di tahun 1990-an, internet tumbuh dengan laju yang luar biasa dan tidak menunjukkan tanda-tanda akan melambat. Ukuran ini sendiri merupakan masalah: ketika kita memiliki akses secara topikal segala sesuatu yang pernah diketahui siapa pun, bagaimana kita menemukan apa yang kita butuhkan untuk membantu kita memecahkan masalah kita?

Tidak hanya sebesar itu, Internet juga sangat luas. Informasi diunggah ke server yang mungkin berlokasi di mana saja di dunia, yang berarti siapa pun yang ingin menelusuri data yang tersedia bagi mereka terhubung ke komputer yang kadang-kadang dihubungkan ribuan mil terpisah satu sama lain. Mendapatkan setiap bit data spesifik melalui pengguna tidak butuh waktu lama, dengan kecepatan di mana informasi dapat berjalan sepanjang kabel tembaga atau serat optik - hanya beberapa detik. Tapi itu mengandaikan pengguna tahu di mana data berada di tempat pertama. Menelusuri seluruh Internet bahkan untuk sepotong informasi yang sangat sederhana, jika Anda tidak tahu alamat IP komputer yang disimpannya akan membutuhkan waktu yang sangat, sangat lama jika Anda tidak memiliki indeks.

Namun, dengan miliaran halaman informasi online, membangun indeks bukanlah hal yang sepele. Butuh pasukan manusia selamanya untuk menemukan apa pun yang mendekati basis data komprehensif konten Internet. Jadi itu harus dilakukan secara otomatis - oleh komputer. Ini menimbulkan masalah lain: bagaimana komputer tahu apa informasi yang baik dan suara tidak berguna? Secara default, komputer tidak dapat menentukan ini sendiri: mereka tidak memiliki konsep perbedaan antara berguna dan tidak berguna, kecuali kita mengajar mereka dan, lagi pula, apa yang tidak berguna bagi satu orang mungkin penting bagi orang lain untuk menyelesaikan masalah mereka

**Bagaimana Big Data Digunakan Dalam Praktek?**

Google tidak menciptakan konsep mesin pencari, atau indeks Web. Tetapi dengan sangat cepat setelah diluncurkan pada tahun 1997, mereka menetapkannya sebagai anjing teratas - sebuah judul yang sudah ada selama hampir 20 tahun.

Konsep yang menjadikannya sebagai nama rumah tangga di setiap sudut dunia, sementara pesaing awalnya seperti Alta Vista atau Ask Jeeves nyaris tidak diingat, dikenal sebagai Google PageRank. (Google memiliki kesukaan untuk membuat nama baru untuk berbagai hal dengan menyatukan dua kata, tetapi tetap menggunakan kedua kata tersebut seolah-olah mereka masih dua kata yang terpisah!)

PageRank dikembangkan oleh pendiri Google Larry Page dan Sergey Brin sebelum mereka membentuk perusahaan, selama penelitian di Universitas Stanford. Prinsipnya adalah bahwa semakin banyak halaman yang terhubung ke halaman tertentu, semakin tinggi “otoritas” halaman tersebut - karena situs-situs yang menghubungkannya cenderung mengutipnya dengan cara tertentu. Google menciptakan algoritma pencarian pertama mereka untuk menetapkan peringkat pada setiap halaman dalam indeksnya berdasarkan pada berapa banyak situs lain yang menggunakan kata kunci yang serupa (dan kemungkinan besar berada pada topik atau subjek yang sama) yang terhubung dengannya, dan pada gilirannya bagaimana "berwibawa" (sangat Linked-to) halaman-halaman yang terhubung itu sendiri. Dengan kata lain, ini adalah proses yang melibatkan mengubah data yang tidak terstruktur (isi halaman Web) menjadi data terstruktur yang diperlukan untuk mengukur informasi itu, dan memeringkatnya untuk kegunaan.

Google membuat indeks Web-nya dengan mengirimkan robot perangkat lunak - sering disebut crawler atau labah-labah - yang mengumpulkan semua teks dan informasi lainnya, seperti gambar atau suara, yang terkandung di situs web dan menyalinnya ke arsip Google sendiri yang luas - datanya pusat dikatakan menyumbang 0,01% dari semua listrik yang digunakan di planet ini!

Dengan informasi yang sekarang semuanya disimpan di satu tempat, itu dapat dicari jauh lebih cepat - daripada menjaring seluruh dunia untuk menemukan dokumen yang mengandung informasi yang dicari para pencari, semuanya berada di bawah satu atap yang sangat besar. Dikombinasikan dengan PageRank dan perkembangan selanjutnya seperti Grafik Pengetahuan (lebih lanjut tentang ini di bawah ini), itu kemudian melakukan yang terbaik untuk mencocokkan permintaan kami dengan informasi yang kami temukan berguna

**Apa Hasilnya?**

Pada saat penulisan, Google menyumbang 89% dari semua penggunaan pencarian Internet. Di antara mereka, pesaing terdekat Yahoo, Bing dan Baidu menyumbang hampir semua dari 11% sisanya, dalam urutan itu.2

**Data apa yang digunakan?**

Google menggunakan data dari indeks Web-nya untuk mencocokkan permintaan dengan hasil yang berpotensi bermanfaat. Ini ditambah dengan data dari sumber tepercaya dan situs lain yang telah diberi peringkat untuk akurasi oleh algoritma pembelajaran mesin yang dirancang untuk menilai keandalan data.

Akhirnya, Google juga menggabungkan informasi yang ia ketahui tentang pencari - seperti riwayat pencarian sebelumnya, dan informasi apa pun yang mereka masukkan ke dalam profil Google Plus, untuk memberikan sentuhan pribadi pada hasilnya.