

Google's Secret Formula

by Paul Smalera | September 2007 Issue

In the past 12 months, Google doubled its staff, tinkered with its search engine to speed up results, and now answers more queries than Microsoft and Yahoo combined. But there's one query we had to answer ourselves: How does Google work?



Blame spell-check. Ten years ago this September, so the story goes, some Stanford grad students were helping Larry Page choose a name for his search engine. "Googolplex," said Sean Anderson. (They'd already sensed how big this could become.) "Googol," Page replied. Anderson, checking to see if the name was taken, typed g-o-o-g-l-e into his browser and made the most famous spelling mistake since p-o-t-a-t-o-e. Page registered the name within hours, and today, *Google* isn't a typo, it's a verb, one with a market cap of about \$160 billion. Here, then, is a guide to what happens during a typical Google search—now, of course, with automatic spell-check.

1. Query Box

It all starts with somebody typing in a request for information about the safest dog food, what time the D.M.V. closes, or what the prime rate is in China.

2. Domain-Name Servers

"Hello, this is your operator..."

The software for Google's domain-name servers runs on computers in leased or company-owned data centers all over the world, including one in the old Port Authority headquarters in Manhattan. Their sole purpose is to shepherd searches into one of Google's clusters as efficiently as possible, taking into account which clusters are nearest to the searcher and which are least busy at that instant.

3. The Cluster

The request continues into one of at least 200 clusters, which sit in Google-owned data centers worldwide.

4. Google Web Server

This program splits a query among hundreds or thousands of machines so that they can all work on it at the same time. It's the difference between doing your grocery shopping all by yourself and having 100 people simultaneously find one item and toss it into your cart.

5. Index Server

Everything Google knows is stored in a massive database. But rather than waiting for one computer to sift through those gigabytes of data, Google has hundreds of computers scan its "card catalog" at the same time to find every relevant entry. Popular searches are cached—held in memory—for a few hours rather than run all over again. That means you, Britney.

6. Document Server

After the index server compiles its results, the document server pulls all the relevant documents—the links and snippets of text from its massive database. How does Google search the Web so quickly? It doesn't. It keeps three copies of all the information from the internet that it has indexed in its own document servers, and all those data have already been prepped and sorted.

7. Spelling Server

Google doesn't read words; it looks for patterns of characters, be they in English or Sanskrit. If it sees your requested pattern a thousand times but finds a million hits for a similar pattern that's off by one character, it connects the dots and politely suggests what you probably meant, even while it provides you the results, if any, for your fat-fingered query for "hwedge funds."

8. Ad Server

Each query is simultaneously run through an ad database, and matches are fed to the Web server so that they're placed on the results page. The ad team is in a race with the search team. Google vows to deliver all searches as quickly as possible; if ad results take longer to pull up than search results, they won't make it onto the page—and Google won't make money on that search.

9. Page Builder

The Google Web server collects the results of the thousands of operations it runs for a query, organizes all the data, and draws Google's cunningly simple results page on your browser window, all in less time than it took to read this sentence.

10. Results Displayed

Often in 0.25 seconds or less.

Cluster Control

Google's genius lies in its networking software, which helps thousands of cheap computers in a cluster act like one huge hard drive. Those inexpensive computers allow Google to replace parts without stopping the whole show: If a computer drops dead, there are at least two others ready to take its place while an engineer swaps out the busted machine.

Power Power

Just about the only thing limiting Google's performance is how much electricity the company can buy. One of its newest data centers (code name: Project 02) is near the Columbia River in The Dalles, Oregon, which has access to 1.8 gigawatts of cheap hydroelectric power; not coincidentally, this is where major internet hookups from Asia connect to U.S. networks. The byte factory has two computing centers, each the size of a football field.

Petabytes

Based on the few numbers Google releases, experts guess that at least 20 petabytes of data are stored on its servers. But Googleytes are famous for understatement; *Wired* says Google may have 200 petabytes of capacity. So how much is that? If your iPod were just 1 petabyte (one million gigabytes), you'd have about 200 million songs to shuffle. And if you started downloading a petabyte over your high-speed internet connection, your great-great-great-great-grandchild might still be around when the last few bytes get transferred, in 2514.

Page Rank

Google decides how reliable a site is—and thus how important the site's content will be when Google forms a list of search results—by considering more than 200 factors as it analyzes content. But the secret sauce is Google's patented formula for following and scoring every link on a page to learn how different sites connect, which means a site is deemed reliable based largely on the quality of the sites that link to it.

Googlebots

Google deploys programs called spiders to build its copies of the internet. On popular sites, Googlebots may follow every link several times an hour. As they scour the pages, the spiders save every bit of text or code. The raw data are pulled back into the cluster, run through the mill, and scheduled to incrementally replace the older data already on the index and doc servers, ensuring that results are fresh, never frozen.