# Google

Google is one of the largest distributed systems in use today. All Google, it is one of the most renowned terms in Internet world. Google's brand has become so universally recognizable that now days; people use it like a verb. For example, if someone asks "Hey what is the meaning of that word? The answer is "I don't know, goggle it". Google Inc. is an American public corporation specializing in Internet search technology and many products. Google's mission is based on the fundamentals of collaborative teamwork. Its main motive is to organize the world's information and make it universally accessible and useful. Google Company was founded by Larry Page and Sergey Brin while studying PHD at Stanford University in 1998[ (Abdul Rehman , 2013)].

The name is a play on the word googol, the number 10^100 (or 1 followed by a hundred zeros), emphasizing the sheer scale of information in Internet today.

Google was born out of a research project at Stanford University, with the company launched in 1998. Since then, it has grown to have a dominant share of the Internet search market, largely due to the effectiveness of the underlying ranking algorithm used in its search engine.

A familiar term in the world of computer science known "Distributed Systems", designed/developed to reduce the burden of main system with the help of sub-systems/nodes running on network. A set of different devices on a network work like single computer, or Different components located on network, able to pass messages, and the motivation to share resources is called Distributed system. Favoring the definition in, as we use an internet for many purposes like web search engines to find particular stuffs, GPS, we use maps to search a particular location, for educational purposes, in healthcare, Science, gaming or entertainment, for environment, communication, and even many more purposes. Even only one example of internet that is "Search engine" has more than 10 billion users per month throughout the world. Responding billion of users constantly without caring globalization or localization cannot be possible without the help of distributed system's concept. Because a distributed system is a system, that is concurrently available and executable, system that has no global time, and has the isolation while failure of node/host [ (Krishan Kant Lavania, Sapna Jain, Madhur Kumar Gupta, andNicySharma, April,2013)].

Google is now a major player in cloud computing which is defined as "a set of Internet-based application, storage and computing services sufficient to support most user's needs, thus enabling them to largely or totally dispense with local data storage and application software.

- **Software as a service:** Offering application-level software over the Internet as web application. A prime example is a set of web-based applications including Gmail, Google Docs, Google Talk and Google Calendar. Aims to replace traditional office suites.

- **Platform as a service:** Concerned with offering distributed system APIs and services across the Internet, with these APIs used to support the development and hosting of web applications. With the launch of Google App Engine, Google went beyond software as a service and now offers it distributed system infrastructure as cloud service. Other organizations to run their own web applications on the Google platform.

# Design Strategy

Google has diversified and as well as providing a search engine is now a major player in cloud computing. 88 billion queries a month by the end of 2010. The user can expect query result in 0.2 seconds. From a distributed systems perspective, Google provides a fascinating case study with extremely demanding requirements, particularly in terms of scalability, reliability, performance and openness.

Google is designed to scale well to keep up with the growth of web. It gives exactly what we want. For fast and efficient access, its data structures are optimized. In addition to smart coding, on the back end it developed distributed computing systems around that globe that ensure fast response times.

### Google infrastructure

The system is constructed as a set of distributed services offering core functionality to developers. This set of services naturally partitions into the following [ (Anon., n.d.)]:

- The underlying communication paradigms, including services for both remote invocation and indirect communication:
  - The protocol buffers component offers a common serialization format for Google, including the serialization of requests and replies in remote invocation:
  - The Google publish-subscribe service supports the efficient dissemination of events to potentially large numbers of subscribers.
- Data and coordination services providing unstructured and semi-structured abstractions for the storage of data coupled with services to support coordinated access to the data:
  - GFS offers a distributed file system optimized for the particular requirements of Google applications and services (including the storage of very large files).
  - Chubby supports coordination services and the ability to store small volumes of data.
  - Bigtable provides a distributed database offering access to semi-structured data.
- Distributed computation services providing means for carrying out parallel and distributed computation over the physical infrastructure :

- MapReduce supports distributed computation over potentially very large datasets (for example, stored in Bigtable.)

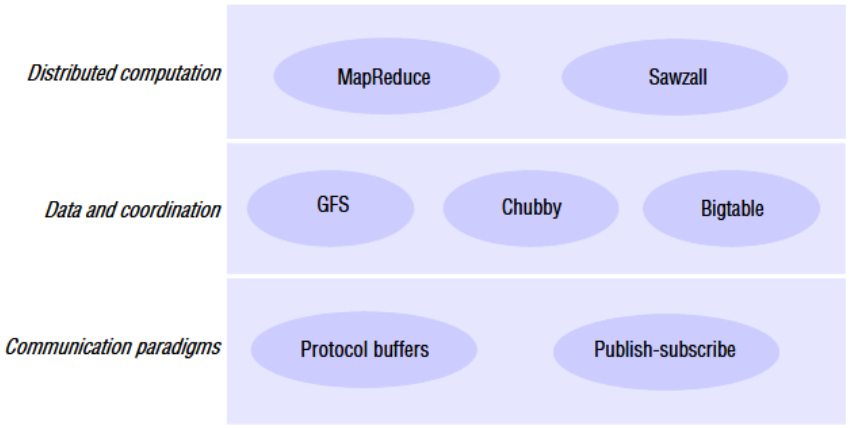



**Figure:** *Google Infrastructure.*

**Associated design principles**

To fully understand the design of Google infrastructure, it is important to also have an understanding of key design philosophies that pervade the organization:

- The most important design principle behind Google software is that of simplicity: software should do one thing and do it well, avoiding feature-rich designs wherever possible. For example, Bloch [2006] discusses how this principle applies to API design implying that API designs should be as small as possible and no smaller
- Another key design principle is a strong emphasis on performance in the development of systems software, captured in the phrase 'every millisecond counts'. In a keynote at LADIS'09, Jeff Dean (a member of the Google Systems Infrastructure Group) emphasized the importance of being able to estimate the performance of a system design through awareness of performance costs of primitive operations such as accessing memory and disk, sending packets over a network, locking and unlocking a mutex and so on, coupled with what he referred to as 'back of the envelope' calculations.
- A final principle is advocating stringent testing regimes on software, captured by the slogan 'if it ain't broke, you are not trying hard enough'. This is complemented by a strong emphasis on logging and tracing to detect and resolve faults in the system[ (Anon., n.d.)].

# Architecture

The environment of Google engine is quite complex phenomenon. It supports tens of thousands of query per second, reads hundreds of megabyte, and uses billions of CPU's cycles. The environment is fault tolerance and with efficient processing speed. The first and foremost indispensable thing to understand the Google search engine is that it doesn't use high power computer (like, super computer or main frame) to process complex functions; but on the contrary it basically makes the Clusters of low cost systems (workstations) to make up the high-end cluster. The high-end cluster works on an idea of parallel processing which means the processing can be taken simultaneously and efficiently to speed up the data processing[ (Abdul Rehman , 2013)].

Before examining the overall system architecture, it is helpful to examine the key requirements in more detail:

- Scalability: Deal with more data–deal with more queries–seeking better results.
- Reliability: There is a need to provide 24/7 availability. Google offers 99.9%service level agreement to paying customers of Google Apps covering Gmail, Google Calendar, Google Docs, Google sites and Google Talk. The well-reported outage of Gmail on Sept. 1st 2009 (100 minutes due to cascading problem of overloading servers) acts as reminder of challenges.
- Performance: Low latency of user interaction. Achieving the throughput to respond to all incoming requests while dealing with very large datasets over network.
- Openness: Core services and applications should be open to allow innovation and new applications.

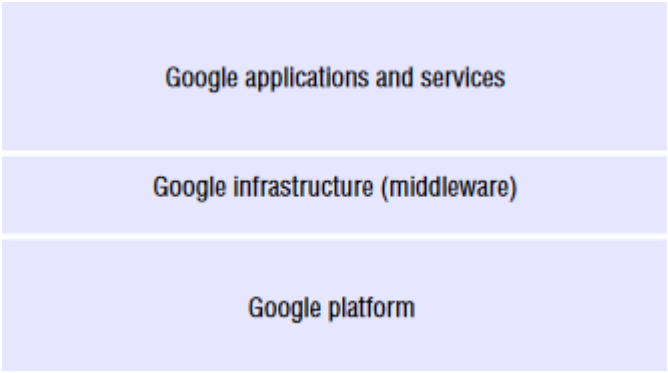



**Figure:** *The overall Google system architecture*

Google Search Engine consists of a set of services:

**Crawling:** to locate and retrieve the contents of the web and pass the content onto the indexing subsystem. Performed by a software called Googlebot.

**Indexing:** produce an index for the contents of the web that is similar to an index at the back of a book, but on a much larger scale. Indexing produces what is known as an inverted index mapping words appearing in web pages and other textual web resources onto the position

where they occur in documents. In addition, index of links is also maintained to keep track of links to a given site.

**Ranking:** Relevance of the retrieved links. Ranking algorithm is called PageRank inspired by citation number for academic papers. A page will be viewed as important if it is linked to by a large number of other pages[ (Coulouris, Dollimore, Kindberg and Blair, n.d.)].

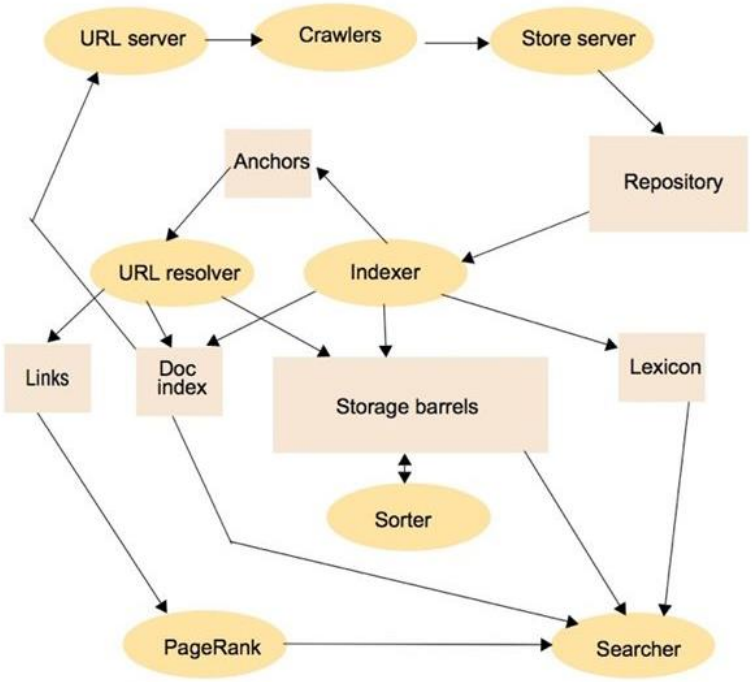



**Figure:** *Outline architecture of the original Google search engine*

## Layout

Google is continuously iterating on the layout and display of how search results are formatted and presented to users. Google launched their search tool in 1998 and their advertising platform Google AdWords debuted in 2000[ (Jennifer Kern, 2020)].

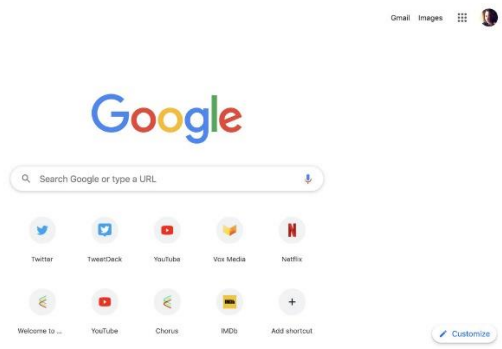



**Figure**: *Google search page*.

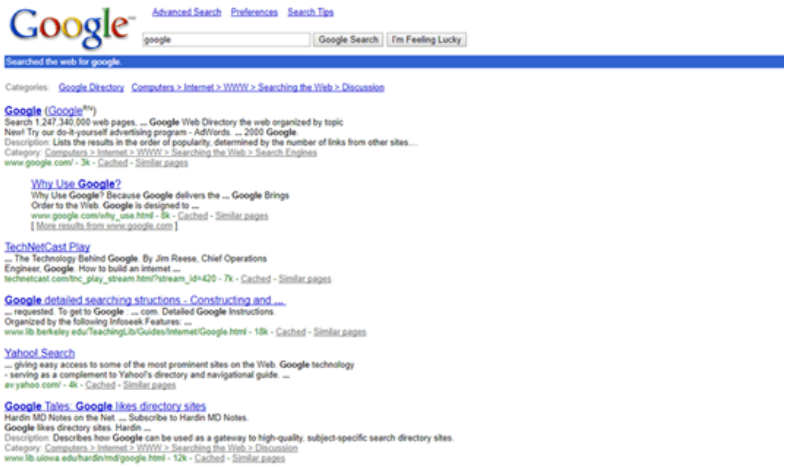
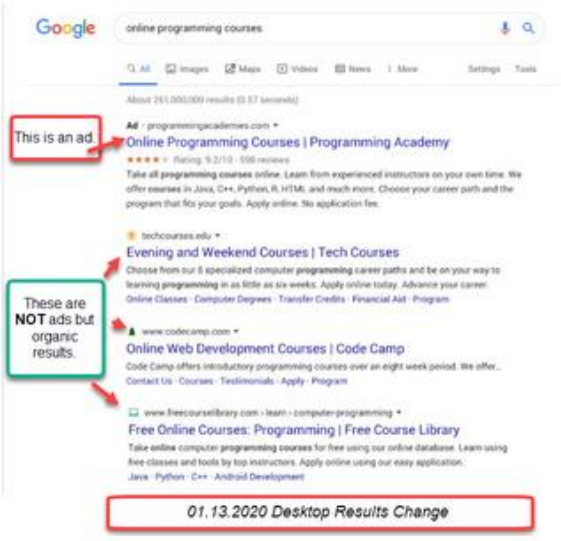



**Figure**: *Google search results page from 2008.*

A SERP looks much different in 2020 as Google moves more and more toward answering a users' questions and meeting their search intent directly on the SERP. In many cases, users no longer need to click on links and visit websites to have their search problem solved.

Recent changes to Google's search layout were rolled out in early January 2020. Those changes created quite a bit of controversy and resulted in an unusual turnaround by Google. The changes were even reported on by the New York Times in a piece January 31, 2020.

In short, Google released a desktop SERP update on January 13, 2020 with the stated intention to have the desktop SERP more closely mirror the look of mobile search results. The essential change placed a company's favicon and page URL above the traditional page title (blue link). See image below.




## Featured Snippets:

Featured snippets come in various forms and consist of content pulled from third-party websites. Featured snippets appear above the traditional Google list of 10 links. Google usually will give attribution to the website the content is pulled from.

Featured snippets are featured on 10%-15% of search queries and are highly coveted. Almost all the sites that earn a featured snippet are already ranking in the #1- #10 position of organic results. So be mindful before you consider optimizing your existing content to win a featured snippet that it should already be ranking in position 1-10.

The most common forms of content you'll see in featured snippets include:

- Paragraph extracts
- Numbered lists
- Bulleted lists
- Content already in a table format

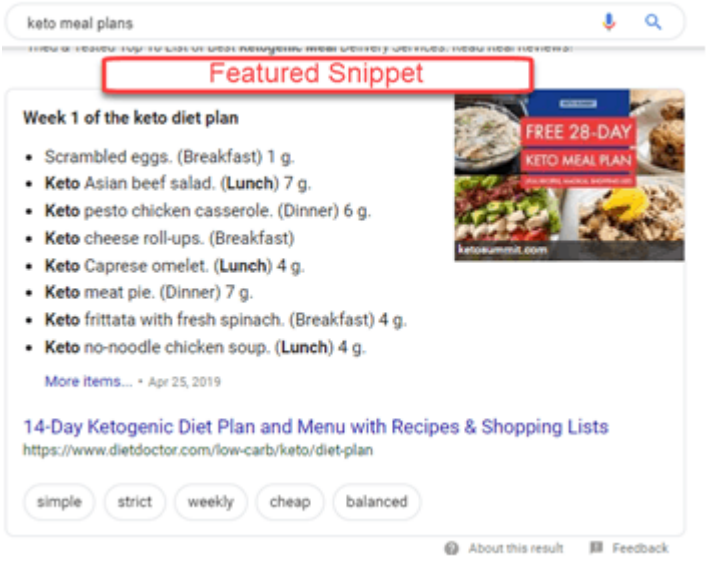



**Figure:** *Featured snippet*

Featured snippets are the results read aloud in response to a voice search. Voice search is projected to comprise more than half of all search queries in 2020! With mobile phones and digital assistants overtaking desktop searching, you'll need to make sure your pages are optimized for placement in a featured snippet.

## Answer Boxes/Instant Answers:

For succinct, factual questions, Google will often answer the question for you and populate it at the top of the SERP. Google has a massive database called the Knowledge Graph which contains billions of public domain facts. Below you'll see an example of an Instant Answer pulled from the Knowledge Graph, which also plays a big role in voice search results. This type of feature isn't one of primary concern for most businesses.

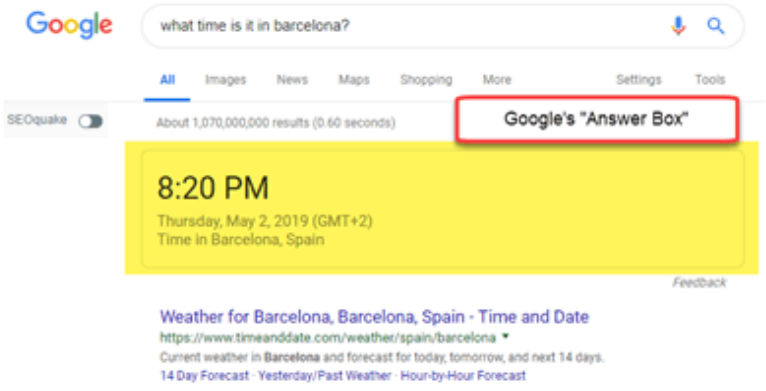



**Figure:** *Answer Boxes*

## Knowledge Cards/Knowledge Panels:

**Local Cards:-**If you own one or multiple local businesses, make sure you are listed on Google My Business. These local cards not only display address and phone number but also business photos, business hours, online reviews, and even questions & answers about your business. See an example below of a local card from a SERP:

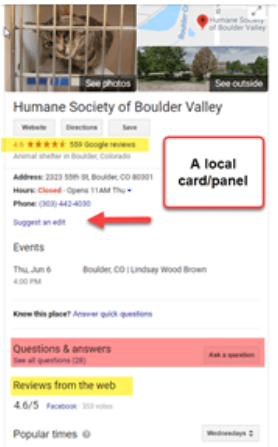



**Figure:** *Knowledge cards/panels*

**Branded/Personal Cards:-** Some examples of a branded Knowledge Card include a summary of a company with links to products, a brief summary of an historical event and a biography of a person. One of the best strategies to earning a Knowledge Card for your business is to make sure your website content is marked up with structured data. Unlike local cards which a company can populate themselves within Google My Business, branded & personal cards are something you can't edit directly**.**

Below is a company's knowledge card:

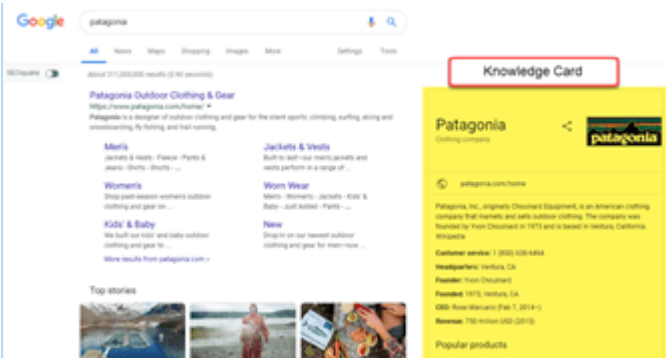



**Figure:** *Branded Knowledge cards/panels*

**Rich Product Snippets:-** Rich product snippets are incredibly important if you are selling products directly on your website. This Google feature again involves using structured data/schema markup to enhance your products and make it easier for all search engines to understand and display your product details. Some of the markup to take advantage of is:

- Limited time offers
- Product availability
- Review and ratings

Below is an example of a rich snippet:

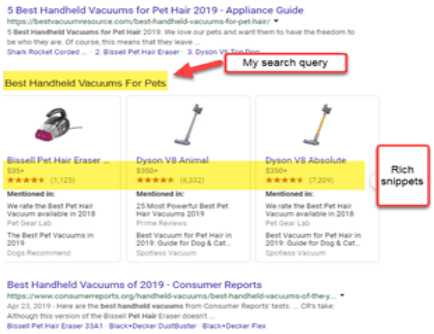



**Figure:** *Rich product snippet*

**People Also Ask (PAAs):-** One of my favorite SERP features is the People Also Ask (PAA) box. These are great to mine when you're doing keyword research or competitive research and they can help you gain insight into other ways your audience may be asking questions.

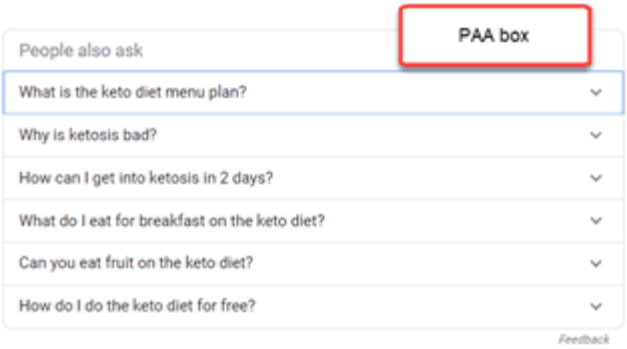



**Figure:** *People Also Ask (PAA) box*

# Working Procedure

Google search, known as one of the most popular search engines across the world, has its own mechanism of understanding an information asked. Every day, Google search engine processes billions of queries using its algorithm and tries to provide the best possible search results.

Now, the question comes, how Google search engine does its filtration process? Yes, there are few steps that Google does while getting the results.

Google search follows its algorithm and goes through the information available on its platform. Thus, Google search engine's algorithm helps to provide the best possible results related to a user's search (Anon., April 8, 2020).

Here are few steps which will help to understand how Google search engine works:

Google information sources include:

- Web pages
- User-submitted content such as Google My Business and Maps user submissions
- Book scanning
- Public databases on the internet
- Images, Videos, and many other sources

Google follows three basic steps to generate results from web pages:

## Step 1 Crawling:

When one types something in the search bar in the first place, it finds what pages exist on the web. As there is no central registry of the webpages on the internet, therefore it constantly adds them to its list. Some pages are known by the Google which are already crawled.

The crawler or spider takes note of all the keywords, web pages, descriptions, etc to learn your data search. This way the Google spider goes to the already visited page which the user has visited in the past for better search results.

Google does not accept payments to crawl a site or rank it higher.

## Step 2 Indexing:

After a page is discovered, Google understands the data of the page. Google analyses the content such as the videos, images, the catalogs, etc embedded on the webpage. This process is called Indexing. This information is then stored in the Google index, a huge date base stored in computers.

Step 3 Ranking:

When the user searches something, the Google then scans through the indexed pages to give the best result. It gives attention to factors like language, location, device, etc. With the same search query **restaurant near me**, search results for a user in India will be different from the person located in the USA. This is how the results are varied.

It ranks pages programmatically.

This is how the Google works to give the search results after going through simple, yet complex process.

# Algorithm

Google's algorithms are a complex system used to retrieve data from its search index and instantly deliver the best possible results for a query. The search engine uses a combination of algorithms and numerous ranking signals to deliver webpages ranked by relevance on its search engine results pages (SERPs)[ (Anon., n.d.)].

To be precise, the whole ranking system consists of multiple algorithms that consider various factors such as quality, relevance or usability of the page.

Google has issued five major algorithm updates, named (in chronological order) Panda, Penguin, Hummingbird, Pigeon, and Fred.

In general, we can divide Google algorithm updates into two categories:

- Minor updates
- Core updates

Google tweaks its algorithm quite often.

And by quite often we mean several times a day**.** Most of these changes are very small and people don't notice them.

Besides these small updates, Google rolls a couple of big core algorithm updates every year. These usually create a lot of buzz in the SEO community and often get catchy names.

Here's a list of the most well-known algo updates in the last decade that shaped the way Google algorithm works[ (Anon., n.d.)]:

**Panda (2011):** Google Panda is a filter focused on low-quality pages, thin content, keyword stuffing and duplicate content**.** It was incorporated into the core algorithm in 2016 and rolls out regularly.

**Penguin (2012):** An important algorithm update that focused on any kind of manipulative (low-quality, spammy, irrelevant, over-optimized) links**.**

**Hummingbird (2013):** Hummingbird update improved the way Google understands and interprets search queries; a shift from exact keywords to topics.

**Pigeon (2014):** Pigeon focused on the improvement of the local results both in terms of quality and accuracy.

**Mobile Update (2015):** This update is also known as Mobilegeddon in the SEO community. It favors mobile-friendly pages in mobile search results.

**RankBrain (2015):** As mentioned earlier, RankBrain is a machine-learning component of Google's Hummingbird algorithm that helps provide more relevant search results**.**

**Fred (2017):** Fred is an unconfirmed update that seemed to focus on low-quality, ad-centered content that violates Google Search Quality Guidelines.

**Medic (2018):** A broad core algorithm update that heavily affected the so-called YMYL (your money your life) pages, especially health-related content.

**Bert (2019):** Another machine learning algorithm focused on a better understanding of the context of a search query. It is based on the natural language processing model called BERT.

## Challenges and how to overcome them

The search giant brought information to the masses. But two decades after its founding, Google isn't the whimsical startup it once was.

Twenty three years ago on Sept. 4, 1998, Larry Page and Sergey Brin founded a small internet search company. The two met in grad school at Stanford, where they created a service called

Backrub that would crawl the web and rank its pages. Eventually, they renamed their search engine Google.

Two decades later, Google is one of the most powerful companies on the planet. It's made the internet, which in 1998 was still mostly just a hodgepodge of scattered content, more usable and relevant to the average person. It brought information to our fingertips and was one of the first dot-com companies to become so important to us that we made it a verb.

But as positive as its impact has been, today Google faces some of its biggest challenges.

“Companies fail because they do the wrong things or they aren't ambitious, not because of litigation or competition” - Larry Page, Google, 2013

The company, the world's largest digital advertiser, is being criticized more and more for its vast data-collection practices, which feed its powerful ad targeting. Misinformation runs rampant on YouTube. Employees are raising ethical concerns about the company's work in developing artificial intelligence for the US military and its reported efforts to create a censored search engine in China. And in just the past week or so, the search giant has become a favorite target for President Donald Trump. The company famous for its onetime "Don't be evil" mantra is under more scrutiny than ever.

**Google's 3 Biggest Challenges for 2020**

1. Growing Regulatory Risks. Alphabet formally acknowledged the government's antitrust probe earlier this year, but it isn't just U.S. federal law enforcement officials taking a harder look at Alphabet's business practices[9 ( Annie Gaus, Jan 7, 2019)].
2. Shoring Up Growth, Profits.
3. Skeptical Employees.

The development of Google distributed system, with an emphasis on addressing the main challenges of Google distributed system, including

- Heterogeneity
- Openness
- Security
- Scalability
- Failure handling
- Concurrency
- Transparency and
- Quality of service.

However, Google has overcome the challenges[ ( Andrew Beattie, May 19, 2020)].

- A key component of Google's success is the company's ability to launch a prototype or beta version of a product and continue to make improvements with each iteration.

- Google's initial business model focused on building a powerful search engine based on algorithms that help people sort through vast amounts of content to deliver accurate results for each search query.
- Google monetized its search engine capabilities through Google Adwords, an online platform that enables the company to earn revenue through pay-per-click advertising.
- The company has successfully launched a wide range of other products and services, including YouTube, Google Maps, Google Apps, and Google Cloud.

## Findings

- Google is now a major player in cloud computing which is defined as "a set of Internet-based application, storage and computing services sufficient to support most user's needs. That provides-
  - Software as a service
  - Platform as a service
- The system is constructed as a set of distributed services offering core functionality to developers. This set of services naturally partitions into the following:
  - The protocol buffers
  - The Google publish-subscribe service
  - GFS
  - Chubby
  - Bigtable
  - MapReduce
  - Sawzall
- The key requirements of the overall system architecture:
  - Scalability
  - Reliability
  - Performance
  - Openness
- Google Search Engine consists of a set of services:
  - Crawling
  - Indexing
  - Ranking

Google is used to upgrading its layout, algorithm and overcome its challenges.

## Summary

This case study concludes the key issue of how one very large Internet enterprise has approached the design of a distributed system to support a demanding set of real-world applications. This is a very challenging topic and one that requires a thorough understanding of the technological choices available to Google distributed system developers at all levels of system development, including system design, Architecture, Layout, Working Procedure, Algorithm, Challenges and how to overcome them. The inevitable trade-offs associated with the design choices demand a thorough understanding of the application domain.

## Remarks

Since inception of internet and development of information technology, Google's record is impressive in the way it has charmed people regardless of their ethnic, religious, and political affiliations.

The company has also reached out to different social and economic classes across the world through its numerous products.

Google identifies among the leading search engines available in the world market. Its reliability in terms of matching results and simple design of their website has attracted a respectable fraction of global population, which is increasingly warming up to the contemporary world of internet.

Some of the main competitors of Google are Yahoo, Amazon, MSN and Bing. Google has managed to fight off competition from these companies to command close to 85% of internet searches.

In 2005, Google's search engine was the best performing product from the company ahead of email services. Other products by Google include Google profiles, Google maps, Google talk, Google gadgets and Google trends.

# References

[1]Annie Gaus, Jan 7, 2019. *The Street.* [Online]
Available at: https://www.thestreet.com/investing/google-alphabet-3-biggest-challenges-in-2020

Andrew Beattie, May 19, 2020. *Investopedia.* [Online]
Available at: https://www.investopedia.com/articles/personal-finance/042415/story-behind-googles-success.asp

Abdul Rehman(811298)MS(Information Technology)College of Arts and Sciences(CAS)Universiti Utara Malaysia(UUM)Sintok, Kedah, Malaysia, 2013. *A case study ofGoogle Search Engine and Bigtable:Distributed Systems.*

Anon., April 8, 2020. *How google search works: In 3 steps.* [Online]
Available at: https://www.indiatoday.in/information/story/-how-google-search-works-in-3-steps-1664546-2020-04-08

Anon., n.d. DESIGNING DISTRIBUTED SYSTEMS: GOOGLE CASE STUDY. In: *INTRODUCING THE CASE STUDY: GOOGLE.* s.l.:s.n.

Anon., n.d. *Mangools Blog: Google algorithm.* [Online]
Available at: https://mangools.com/blog/seopedia/google-algorithm/

Anon., n.d. *Search engine journal.* [Online]
Available at: https://www.searchenginejournal.com/google-algorithm-history/

Coulouris, Dollimore, Kindberg and Blair, n.d. [Online]
Available at: http://cpulabserver.inaoep.mx/~lilrodriguez/files/google.pdf

Jennifer Kern, 2020. *Google's New Search Layout and Why It's Important to Keep Up.* [Online]
Available at: https://3aspensmedia.com/googles-new-search-layout-and-why-its-important-to-keep-up/

Krishan Kant Lavania, Sapna Jain, Madhur Kumar Gupta, andNicySharma, April,2013. 1. *Google: A Case Study (Web Searching and Crawling),* Volume 5.

.