Hardware y Software de Google

Blas Varela López

Índice

Historia

Larry Page y Sergey Brin comenzaron Google como un proyecto universitario en enero de 1996 en la Universidad de Stanford. El nombre original del buscador era BackRub, pero en 1997 los creadores decidieron cambiar el nombre a Google inspirados en el término “gugol”, que se refiere al número 10 elevado a 100, en referencia de su objetivo de organizar la enorme cantidad de información de la web.

El 4 de septiembre de 1998 funda la compañía Google Inc, que estrena su motor de búsqueda el 27 de septiembre. En ese momento contaban con un armario lleno de servidores (en total unos 80 procesadores) y dos routers HP. Este motor acabó superando al más popular de la época, AltaVista, que había sido creado en 1995.

En el 2000 Google presentó AdWords, su sistema de publicidad y la barra de google.

En febrero de 2001 compra el servicio de debate Usenet Deja News y lo transforma en Grupos de Google, un servicio de foros de discusión. En marzo del mismo año Eric Schmidt es nombrado presidente de la junta directiva. En julio de 2001 lanza su servicio de búsqueda de imágenes.

En febrero de 2002 lanza Google Search Appliance, un sistema de búsqueda universal que permite a las organizaciones incorporar información procedente de una variedad de fuentes externas e internas. En mayo lanza Google Labs, un lugar que mostraba proyectos de google y que fue cerrado nueve 9 más tarde. En septiembre se lanza Google Noticias. Ese mismo diciembre se lanza Froogle, un servicio de búsqueda de productos, actualmente se conoce como Google Products.

En febrero de 2003 adquiere Pyra Labs y con ello el servicio de creación de blogs Blogger. En abril se presenta Google Grants, un servicio de publicidad gratuito para empresas sin ánimo de lucro. En diciembre se lanza Google Prints y posteriormente Google Libros.

En enero de 2004, lanza la red social orkut. En marzo se lanza Google Local que más tarde se integraría con Google Maps. En abril presentó Gmail, su servicio de correo electrónico el cual disponía de 1GB de almacenamiento. El 19 de agosto Google Inc. salió a la bolsa de valores, con una oferta inicial de 25,7 millones de acciones y con un rango de precios de 85 a 95 dólares. En octubre se lanza Google Desktop y Google Académico.

En 2005 lanzaron Google Maps y Google Earth, ese mismo año Google compra Android Inc y Vint Cerf es contratado. Google también lanzó Google Code y se desarrolla el primer Summer of Code. En agosto presenta Google Talk y en octubre lanza RSS Google Reader. En noviembre presenta Google Analytics y en diciembre Google Transit.

En 2006 lanza Picasa, una herramienta para edición de imágenes. En marzo adquiere Writely para más tarde lanzar Google Docs y presenta este mismo mes Google Finance. En

abril se lanza Google Calendar y en agosto Google Apps. En octubre Google adquirió por 1650 millones de dólares Youtube y ese mismo mes adquiere Jotspot el cual se convertiría en Google Sites.

En 2007 Google lanzó el sistema operativo abierto para móviles Android y creó la Open Handset Alliance. En abril de 2007 compra DoubleClick por 3100 millones de dólares. Este mismo mes Google se convirtió en la marca más valiosa del mundo, alcanzando los 66000 millones de dólares y superando a empresas como Microsoft o Coca-cola. En junio lanza Google Gears que más tarde sería abandonado. En julio compró Panoramio, una web dedicada a exhibir fotografías que los usuarios crean y geoposicionan. En noviembre lanza OpenSocial.

En febrero de 2008 se lanza Google Sites. El 2 de septiembre de presenta Google Chrome y el proyecto de código abierto Chromium. En diciembre se lanza Google Friend Connect y en febrero Google latitude. En marzo se presenta Google Voice.

En septiembre de 2009 adquiere reCAPTCHA. En noviembre presenta el sistema operativo de código abierto Chromium OS y en diciembre lanza Google Public DNS.

En enero de 2010 Google presenta su primer teléfono móvil, el Nexus One. En febrero adquiere Aadvark y en marzo Picnik. En mayo lanza Google TV. En octubre presenta su proyecto de vehículos autónomos y en diciembre presenta su segundo teléfono, el Nexus S, fabricado por Samsung.

En enero de 2011 Larry Page es nombrado CEO. En mayo se presentan los Chromebooks. En junio se presenta la nueva red social llamada Google+. En agosto adquiere Motorola Mobility por 8800 millones de euros y presenta su tercer teléfono, el Galaxy Nexus, fabricado por Samsung.

En abril de 2012 presenta Project Glass, un proyecto para crear gafas de realidad aumentada. En el evento de Google I/O 2012 se anunció que la versión para desarrolladores de Google Glass estaría disponible en 2013 y la de consumidores en 2014, en ese mismo evento se anunció la primera tableta de Google, la Nexus 7, fabricada por Asus.

En febrero de 2014 Google compra SlickLogin, una compañía compuesta por desarrolladores expertos en seguridad. En junio compra Skybox Imaging para proveer imágenes satélite al servicio de mapas en línea.

El 10 de agosto de 2015, Google se convierte en la principal subsidiaria de Alphabet Inc., compañía creada para la mejor administración de todos los productos y servicios de Google.

Hardware

El Primer Hardware

Cuando los servidores todavía se encontraban en Stanford y el buscador se hacía llamar Backrub, las especificaciones eran las siguientes:

- Sun Ultra II con procesador 200MHz dual y 256 MB de RAM. Esta era la máquina principal.
- Dos servidores Pentium II duales a 300 MHz donados por Intel que incluían 512 MB de RAM y 9 discos de 9 GB entre los dos servidores. En estos servidores se ejecutaba la mayor parte de la búsqueda.
- F50 IBM RS/6000 donado por IBM que incluía 4 procesadores, 512 MB de memoria y ocho discos duros de 9GB.
- Dos armarios adicionales con tres discos duros de 9 GB y seis de 4 GB que estaban conectados al Sun Ultra II.
- Un armario de expansión de discos de IBM con otros ocho discos duros de 9 GB donados por IBM.
- Armario con discos duros caseros que tenía 10 discos duros de 9 GB SCSI.

Hardware Actual

Los servidores de Google cuentan con baterías de 12V por si hay algún problema con el suministro de energía. Están almacenados en *containers* donde hay 1.160, pudiendo consumir cada *container* hasta 250Kw.

Los servidores utilizados no son maquinas especificas, son ordenadores montados a medida con componentes que podrían comprarse en cualquier tienda, los componentes de cada servidor son los siguientes:

- Dos procesadores.
- Dos discos duros.
- 8 *slots* para memoria RAM.
- Procesadores x86 AMD e Intel.

Topología de Red

Se desconocen las cifras exactas, pero el Google mantiene unos 2.000.000 de servidores los cuales están ordenados en racks de clusters y distribuidos por el mundo, esta dispersión geográfica de los servidores permite a Google ofrecer un servicio más rápido a los usuarios.

Cuando alguien se conecta a Google los servidores DNS traducen la dirección www.google.com a varias IP´s distintas permitiendo que se distribuya la carga entre varios

clusters. El orden en el que los servidores DNS van traduciendo las direcciones IP´s sigue un algoritmo round-robin.

Cada cluster de Google tiene miles de servidores, por lo que cuando alguien se conecta a un cluster la carga es distribuida mediante el hardware del mismo enviando la consulta al servidor web que esté menos ocupado en ese momento.

Los racks usados por Google pueden contener entre 40 y 80 servidores, cada uno de los racks tiene una conexión ethernet a un router local que a la vez se conecta al central mediante una conexión de 1 Gigabit.

Infraestructura de Servidores

Lo servidores de Google están divididos en varias categorías:

- **Distribuidores de carga**: aceptan la petición del cliente y la reenvían a uno de los servidores web mediante servidores proxy Squad.
- **Servidores proxy Squad**: aceptan la petición de los distribuidores de carga y en caso de ser posible devuelven el resultado desde la caché local y si no reenvían la petición al servidor web.
- **Servidores web**: coordinan la ejecución de las consultas enviadas por los usuarios y formatean el resultado usando HTML. La ejecución consiste en enviar peticiones a los servidores de índices, fusionar resultados, calcular su rango usando PageRank, elaborar un resumen para cada resultado, preguntar por sugerencias a los servidores de ortografía y finalmente obtener una lista con anuncios del servidor de publicidad.
- **Servidores de recolección de datos**: Están dedicados a navegar por Internet al estilo “araña”. Van actualizando el índice y las bases de datos de documentos con las páginas web que van encontrando y calculan el rango de cada página.
- **Servidores de índices**: contienen un conjunto de trozos de índice. Devuelve una lista de id´s de documentos, llamados “docid”, de modo que los documentos a los que identifican contienen la palabra que el usuario está buscando.
- **Servidores de documentos**: sirven para almacenar los documentos. Cada documento se almacena en docenas de servidores de documentos. Cuando se realiza una búsqueda, el servidor devuelve un resumen de la página basado en las palabras buscadas por el usuario.
- **Servidores de anuncios**: gestionan la publicidad de los servicios AdWords y Adsense.

Centros de Datos

Seguridad

El acceso a los recintos está altamente controlado, no se permiten turistas o visitas, el perímetro está vallado y vigilado, también disponen de una red de videocámaras que monitorizan el acceso al recinto las 24 horas. Para acceder al recinto hay que identificarse

en los puntos de acceso y una vez dentro del edificio se procede a realizar una segunda identificación. Una vez en las oficinas antes de entrar a la sala de servidores se procede a una última verificación.

Los datos se guardan en múltiples localizaciones para asegurar su disponibilidad, los datos también son protegidos mediante algoritmos de encriptación, por lo tanto no están en texto plano, si no cifrados.

Los datos son guardados en discos duros que siguen un ciclo de vida, se monitoriza tanto su estado como su localización para evitar la pérdida o la sustracción de alguno de ellos. Cuando algún disco duro empieza a fallar es formateado y si supera una serie de pruebas puesto a funcionar de nuevo. En caso de no superar la prueba los datos serán sobrescritos y a continuación el disco será destruido.

Eficiencia Energética y Sostenibilidad

Una parte importante de la energía que es consumida en los centros de datos va dirigida a la refrigeración de los equipos. Google ha conseguido reducir este consumo de energía de forma notable, por ejemplo aprovechando el entorno en el que se construyen sus centros de datos.

En la bahía de Finlandia aprovecharon el agua marina para enfriar los equipos debido a que esta se encuentra a muy baja temperatura, una vez utilizada esta se mezclaba con nueva agua para suavizar su temperatura y evitar así el impacto ecológico.

Ubicaciones

La siguiente imagen muestra la ubicación de los centros de datos de Google.

Como se puede observar se encuentran principalmente en Estados Unidos y Europa, encontrando solo uno en Sudamérica y dos en Asia.

Google Search Appliance

Es un sistema de búsqueda que permite a las organizaciones incorporar información de fuentes externas e internas. El software es producido por Google y el hardware por Dell Computers, basándose éste en Dell PowerEdge R710.

Existen tres modelos del sistema, uno con la capacidad para indexar hasta 300.000 documentos, otro hasta 10.000.000 documentos y el de mayor capacidad con hasta 30.000.000 documentos. Las versiones más actuales permiten además ofrecer una búsqueda de hasta mil millones de documentos mediante la conexión de varios dispositivos.

Las características de los sistemas son las siguientes:

- Soporta funciones de Google Analytics y Google Sitemaps.
- Permite buscar en sitios web, 220 tipos de archivos, bases de datos y sistemas de gestión de contenidos.
- La indexación del contenido a buscar puede ser configurada mediante una URL de rastreo. Los patrones de búsqueda también puede ser incluidos para limitar la información que se busca y puede ser personalizado mediante la API de OneBox.
- Lo resultados tienen la apariencia de la búsqueda de Google, pero pueden ser personalizados utilizando XSL Transformations.
- Permite utilizar palabras claves que devuelvan un resultado específico.
- Los sinónimos devuelven términos alternativos de búsqueda.
- Cada enlace incluye un enlace llamada cache. Al hacer click mostrará una versión HTML del documento sin necesidad de abrirlo.
- Se ponen de relieve los términos de búsqueda y permite ver las palabras en contexto sin tener que abrir los documentos.
- Se agrupan los resultados similares para evitar duplicados.
- Los resultados pueden ser ordenados por fecha o relevancia.

Un ejemplo de lo mostrado en la búsqueda es el siguiente:

Software

Los servidores Google hacen uso de diferentes lenguajes de programación favoreciendo C++, Java, Python y Go sobre los demás. Por ejemplo el back end de Gmail está escrito en Java y el de la búsqueda de Google en C++.

Google Web Server

Google Web Server (GWS) es el servidor que Google utiliza en sus infraestructuras y servidores, este se ejecuta en sistemas UNIX como GNU/Linux. Existen especulaciones sobre que GWS es una versión adaptada y modificada de Apache HTTP Server.

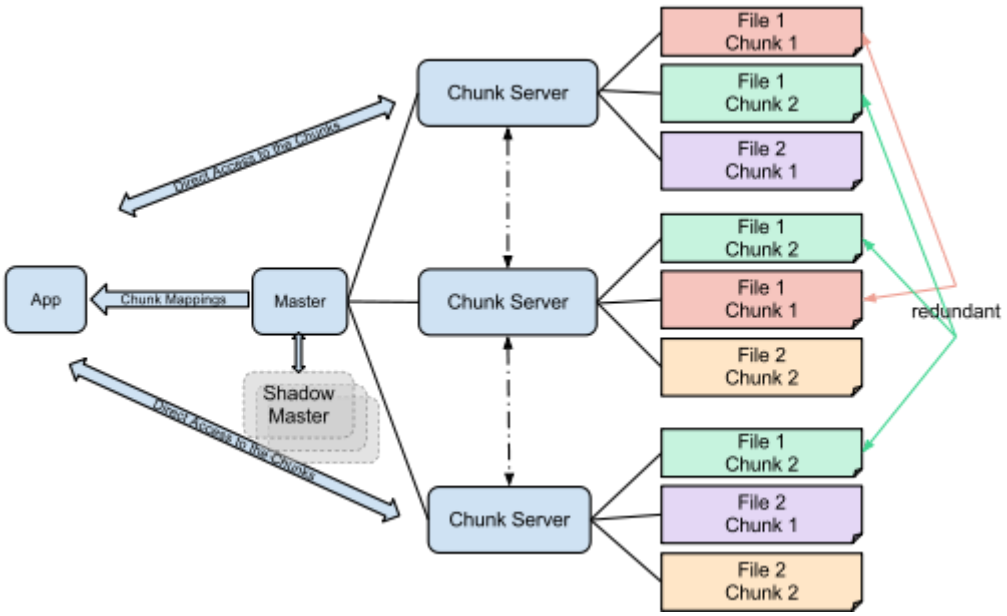
Sistemas de Almacenamiento

Google File System

Es un sistema de archivos distribuido desarrollado por Google, soporta toda su infraestructura de procesamiento en la nube. Está diseñado para proveer eficiencia y fiabilidad de acceso a datos haciendo uso de sistemas masivos de cluster de procesamiento en paralelo.

En este sistema los archivos son divididos en porciones de tamaño fijo de 64 megabytes, los archivos rara vez son sobreescritos o reducidos, por lo general se adicionen o se leen. También está diseñado y optimizado para funcionar con los clusters de Google por lo que se toman precauciones contra un alto índice de fallos que pueden llevar a la pérdida de algunos datos. El sistema también apunta al manejo de grandes caudales de datos y a la resolución de problemas de latencia.

La siguiente imagen muestra cómo se distribuyen los nodos del sistema:



Principalmente se divide en dos tipos de nodos, un nodo maestro y un gran número de Chunk Servers. Los archivos son divididos en porciones de tamaño fijo y los Chunk Server almacenan estas porciones, a cada porción almacenada se le asigna una identificación única de 64 bits en el nodo maestro. Cada porción es replicada en al menos tres servidores, siendo replicadas en mas servidores cuanto mayor sea su demanda.
Los programas acceden a las porciones mediante consultas al nodo maestro, en caso de que las porciones no se encuentren activas el nodo maestro responde con su ubicación, la aplicación contacta y recibe los datos desde el nodo alojamiento directamente.

Por último este sistema no está implementado en el kernel del sistema operativo, funciona con una librería en el espacio de usuario.

BigTable

Es un sistema de gestión de bases de datos creado por Google con las características de ser distribuido y de alta eficiencia, está construido sobre Google File System, Chubby Lock Service y otros programas de Google.

La información es almacenada en tablas multidimensionales cuyas celdas están en su mayoría sin utilizar. Estas celdas disponen de versiones temporales de sus valores, con lo que se puede hacer un seguimiento de los valores que se han tomado históricamente.

Las tablas se dividen en columnas y son almacenadas como tabletas de unos 200 Megabytes cada una. Cada máquina almacena 100 tabletas, mediante el sistema Google File System. Esta disposición permite un sistema de balanceo de carga pues si una tableta recibe una gran cantidad de peticiones la máquina puede desprenderse del resto de tabletas o trasladar la tableta en cuestión a otra máquina. También permite una rápida recomposición del sistema en caso de que este caiga.

Spanner

Es una base de datos NewSQL, que pretende añadir las capacidades NoSQL a las bases de datos relacionales, distribuida por Google. Está descrita como una una base de datos relacional no del todo pura pues cada tabla tiene que tener una columna de llave primaria.

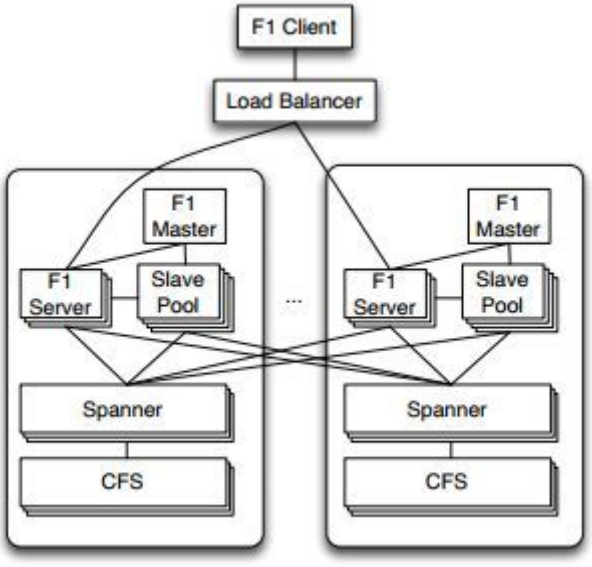
Hace uso del algoritmo Paxos para fragmentar los datos a través de los centros de datos, hacen un uso intensivo de la sincronización de tiempo asistida por hardware usando relojes GPS y relojes atómicos.

Google F1

Es una base de datos distribuida relacional creada por Google. Es una base de datos híbrida que combina alta disponibilidad, la escalabilidad de los sistema NoSQL como BigTable y la consistencia y usabilidad de las bases de datos SQL tradicionales.

Está construida sobre Spanner, el cual provee de replicación síncrona entre los centros de datos y una fuerte consistencia. Esta replicación supone una mayor latencia, pero es mitigada haciendo uso de un modelo jerárquico con tipos de datos estructurados y diseño de aplicaciones inteligentes.

También incluye un motor de consultas SQL y un sistema automático de seguimiento de cambios y publicaciones.
La siguiente imagen muestra como estaría distribuido este sistema:



Puede verse como el cliente hace las peticiones al Servidor F1 y este a su vez realiza consultas al servidor Spanner que trabajara sobre el sistema de archivos de Google.

Servicio de Bloqueo

Los sistemas operativos usan gestores de bloqueo para organizar y serializar el acceso a los recursos. Google hace uso de un gestor de bloqueo distribuido llamado Chubby el cual se ejecuta en cada una de las máquinas de los clusters. Este gestor proporciona por lo tanto aplicaciones software que se distribuyen en todas las máquinas para sincronizar los accesos a los recursos compartidos.

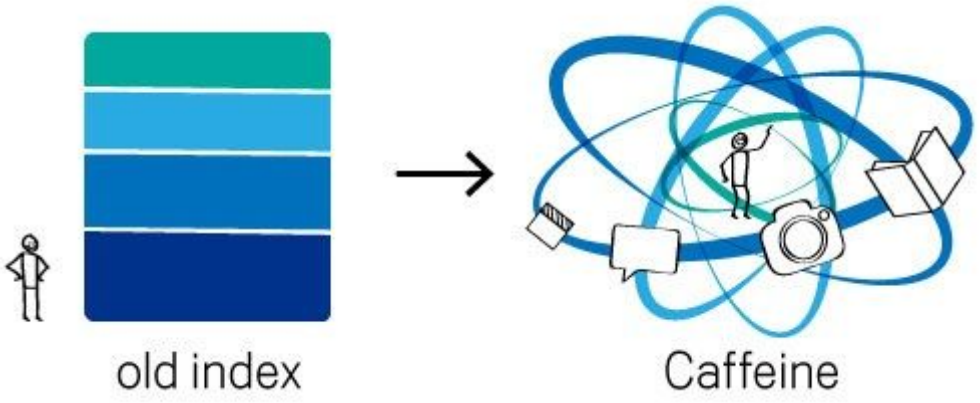
Mediante este sistema de archivos unificado se consiguen ventajas significativas en rendimiento y disponibilidad. El principal beneficio en rendimiento se produce al solucionar el problema de la coherencia de caché de disco entre los equipos participantes.

A pesar de que Chubby fue diseñado como un servicio de bloqueo, ahora es muy utilizado por Google como un servicio de nombres, suplantando a DNS.

Sistemas de Indexación/Búsqueda

Caffeine

Este fue el nuevo sistema de indexación de la web que Google creó en 2009, la siguiente imagen nos muestra como fue el paso desde el antiguo sistema:



El antiguo sistema contaba con varias capas, algunas de estas capas eran refrescadas con una tasa de actualización más rápida, la capa principal por ejemplo tenía que ser actualizada cada dos semanas. Para refrescar una de estas capas Google tenía que analizar la web entera, lo que producía una demora entre que Google encontraba la información y esta se hacía disponible.

Con Caffeine, Google analiza la web en pequeñas porciones ya actualiza el índice de búsqueda de forma continua a nivel mundial. Tan pronto Google encuentra una nueva web o nueva información en las existentes lo adiciona directamente al índice.

Caffeine permite a Google indexar páginas web a una escala enorme. Cada segundo procesa cientos de miles de páginas en paralelo. Caffeine ocupa cerca de 100 millones de Gigabytes de almacenamiento en una base de datos y adiciona nueva información a una velocidad de cientos de miles de Gigabytes por día.

Hummingbird

Fue lanzado en septiembre del 2013 y sustituye a Google Caffeine. Fue llamado así puesto que es rápido y preciso como lo sería un colibrí. La principal novedad de este algoritmo fue la capacidad para responder a preguntas, generando por lo tanto búsquedas más precisas. Esto supuso un avance hacia la web semántica.

Buffers de Protocolo

Es un método de serialización de las estructuras de datos. Es útil para desarrollar programas que se conectan entre sí o para guardar datos. El método implica un lenguaje de descripción de la interfaz que describe la estructura de algunos datos y un programa que genera código de la descripción generada o analizando un conjunto de bytes que representa el dato estructurado.

Bibliografía

https://en.wikipedia.org/wiki/Google_Data_Centers
https://es.wikipedia.org/wiki/Buscador_de_Google
https://www.velaio.com/google-caffeine-seo.html
http://www.seocom.es/blog/hummingbird-que-como-porque-de-este-nuevo-algoritmo-de-google
https://en.wikipedia.org/wiki/Protocol_Buffers
https://www.xataka.com/otros/los-servidores-que-utiliza-google
https://hipertextual.com/2009/04/revelan-datos-de-los-servidores-de-google
http://omicrono.elespanol.com/2012/07/todo-sobre-los-servidores-de-google-donde-y-como-almacenan-toda-la-informacion/
https://es.wikipedia.org/wiki/Google
https://es.wikipedia.org/wiki/Google_File_System
https://es.wikipedia.org/wiki/Google_Search_Appliance
https://es.wikipedia.org/wiki/Google_Caffeine
https://es.wikipedia.org/wiki/Buscador_de_Google
https://research.google.com/pubs/pub41344.html
https://es.wikipedia.org/wiki/Plataforma_de_Google
https://www.google.com/about/datacenters/inside/
https://es.wikipedia.org/wiki/Centros_de_datos_de_Google