A case study of Google Search Engine and Bigtable: Distributed Systems

Abdul Rehman (811298)
MS(Information Technology)
College of Arts and Sciences (CAS)
Universiti Utara Malaysia (UUM)
Sintok, Kedah, Malaysia

Abstract - Interaction of autonomous computers on a network so as to achieve a common goal is called a “distributed system” [8]. The goal of distributed system is to divide resources/tasks among different machines on a network either to solve or to achieve the same objective with fewer burdens on machines with maintenance of the high performance of systems so that these free resources can handle other users at same time. Similarly, a very famous search engine “Google” uses the same concept of distributed system to answer the queries of its over 10 billion users a month [1] maintaining its speed. Projects like Google earth, Google maps, Google finance, and many other projects of Google need the independent big storage place; and Bigtable, a distributed storage system, functions to provide sufficient storage to all projects of Google by using its Google App engine.

Keywords— Distributed System, Google, Search Engine, Bigtable, crawler

I. INTRODUCTION:
The world nowadays is reckoned to be a global village with an acute advancement of technology where people can switch themselves to any corner of the world without moving physically with the help of internet. The emergence of Internet has facilitated people in different dimensions. Nowadays, Internet facilities are used for different purposes like starting business using online facilities, applying for university admissions and job around the world, either purchasing or selling things are also done using Internet facilities in almost all countries of the world. But the ground realities are totally inverse what people think on surface as due to ever increasing number of internet users have caused much complexities among I. Tans to keep all systems with required speed to maintain all projects function normally.

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 [2], or different components located on network, able to pass messages, and the motivation to share resources is called Distributed system [1]. Favoring the definition in [1], 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 [1]. 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.

The motive of this article is to explain the ways “Google” search engine works, and to elaborate the involvement of different distributed systems and to show the support of distributed systems to make a finest performance for search engine.

II. Google Search Engine: A Distributed System

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 [3]. The first and foremost indispensible 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.

Generally, whenever user types a query while entering in Google.com by mapping address or IP, Google starts load-balancing DNS search, the closer active cluster server from user’s location. A Google hardware load balancer has responsibility to transfer the Google request to Google Web Server (GWS). Whereas, Google Web Server (GWS) is responsible to execute the given query and formulate the response in Hyper Text Markup Language (HTML).

Google distributes the query into three parts. Firstly, a Googlebot, web crawler, is an intelligent program that finds and fetches the page being searched. The second part is indexers, which stores
the results in resulting index pages (index shard) from every found word. Thus, these index results in the last become as order list of document identified with ‘docid’. In each docid the Google Web Server (GWS) keeps the record comprising of Uniform Resource Locator (URL), page title, and relevant text. In this way, all processes related to docid process are maintained by docserver. Finally, query processor compares the query with indices and recommends the most relevant result [3]. Additionally, Google Web Server (GWS) checks spellings and adds advertisements while indexing and checking documents from docserver.

A Google Crawlers find the links by two ways i.e., one from addurl(www.google.com/addurl.html) and second through finding links as mentioned above, Google crawlers can work faster than one imagines, but Googlebot/ Crawler deliberately makes some late to avoid over crowded search [4].

After Googlebot finds out the pages it passes to Google indexer, and Google indexer then alter the search results like it removes the stop words for example; is, or, an, of, how, why and single letters and single digits as well. Stop words are most common so they don’t put much effect in searching. Lastly, these all searches are given indices and stored in indexer’s database alphabetically by terms of search. [4].

To ranking the search result pages, finding the relevance in a query with popularity, locality, size, and search terms in the searched pages, and preparing the user interface are the last phases of engine. Query processor in Google engine is responsible of these operations [4]. In the user interface or displaying and organizing the results are remained main task. Might people from Malaysia want to see the different results when they type “songs”, may be by locking their locality, language, frequently accesses or search results, and interest. The Google’s query processor organized a different interface than the people from Pakistan write the same letters like “songs”. It is just possible because the load balancer shares a nearest Google Web Server (GWS) to users’ query, in the start of searching, to maintain the performance and reliability in results [3][4].

III. Bigtable: A Distributed Storage System:

The complexity of Google search engine in terms of searching query, keeping millions of bytes of data for one query, keeping the history of searches, handling billion of users in a month, and hard them all to provide the “storage” to billions user is a bigger challenge in Google search engine. Even Google is providing the backup/ support to many applications for instance; Google Analytics, Web indexing, Google Finance, Google Earth, personalized search, and so on.

It can be as much complex as we imagine the functionality of our brain in terms of memorizing things for long time or responding may be thousand phone calls at a time. Google and its applications require data storage not only in megabytes or gigabyte as we use PC in our houses, but it requires the petabytes to thousand servers. It is, therefore, for managing structured data Google uses a distributed storage system called Bigtable [5]. Bigtable supports high performance, availability, scalability, and applicability, and it has taken in use in more than sixty Google products for instance; Google maps, Google earth, Orkut, and Google finance. Bigtable clusters are used to provide storage up to hundreds of terabyte of data to these products [5].

It is a distributed storage system because it uses the Google App engine. Any machine can be added without configuring or reconfiguring with it at any time to use its resources [7]. It is a main reason of its outrage support of users and flexibility in access. Main memory-databases and parallel databases are main reason for great performance and its scalability, but with different interfaces [6][9]. Locality of data can be control by clients easily, only if the choices in schemas are considered carefully. Even clients are given a wide range of option to server the data by using disks or may be client can be offered to take off data from memory to serve [5].
IV. CONCLUSION:

In overall discussion of the article it would not wrong to add that if a system running with the backbone of distributed approach can be more efficient, effective and productive. In terms of performance, availability and dealing large number of requests by high number of user implementation of distributed system can be the best solution. As discussed in this article that Google search engine and bigtable smoothly interact with many applications and billions of user and result query within milliseconds.

REFERENCES