International Journal of Computer Science and Business Informatics

IJCSBI.ORG

Cloud Architecture for Search Engine Application

A. L. Saranya
School of Information Technology & Engineering,
VIT University, Vellore-632014, Tamil Nadu, India

B. Senthil Murugan
Assistant Professor (Senior)
School of Information Technology & Engineering,
VIT University, Vellore-632014, Tamil Nadu, India

ABSTRACT

Cloud computing has become popular because of its on demand self services capability and business benefits. This paper presents design of search engine application developed and deployed using Google app engine. The application uses pattern-matching and regular expression language processing across millions of web document and returns the matching web documents. To facilitate large dataset processing the application makes use of Apache Hadoop suite, which is distributed data processing framework that brings up hundreds of virtual servers on-demand, runs a parallel computation on them, then shuts down all the virtual servers releasing all its resources back to the cloud. The MapReduce concept is used to implement the system to do the parallel computation and give efficient result to user. The application is efficient and scalable to any number of users in quick response time. The Google app engine uses cloud SQL instance to store data virtually in a cloud database.

Keywords

MapReduce, Pattern matching, SQL instance, Google app engine, Apache Hadoop suite.

1. INTRODUCTION

Using cloud architecture the software application can be effectively designed and online databases are used on-demand. Cloud infrastructure used for software application is utilized on need and returned it back to cloud providers after its usage to make it available for other application. Cloud architecture can handle large number of data's easily. Physical location of the application infrastructure is determined by the provider, so that there are many business benefits in cloud architecture, such as business people no need to invest for infrastructure, quick infrastructure when needed, resources is utilized efficiently, pay only for what using, through parallelization processing time of the job is reduced. The main objective of this paper is to develop efficient, scalable search engine application based on cloud architecture which will give responses to many users. This application should be loosely coupled so that it is available to all user community and can access concurrently.

2. BACKGROUND STUDY

The new computing model of cloud computing provide resource, storage and online application as service to the user. Cloud computing is dynamic, reliable, scalable, low cost and secure so that it provides virtual service to any number users. The cloud computing provide three type of services such as , software as a service where the application software can used by any one as on demand resource, platform as a service and infrastructure as a service. The internet users are more interest in searching data's and getting needed information. For quick and efficient result, large computing resources are needed. Cloud infrastructure is used get the resources needed, to get data after processing the data and resources is given back. Using Google apps engine implementation of search engine cloud application is explained is this paper. The application use Hadoop mapreduce concept to get large data from the cloud and map the process request on that data and reduce the result set to give the searched result. Mapping of millions of result has been done parallel and quick response to request is generated so that application is more efficient.

3. RELATED WORKS

Chunzhi Wang and Zhuang Yang [1] of Hubei University of Technology, explain the cloud search engine process based on user interest. They showed that demand of user can be known by introducing user interest model. Push mechanism used to get result for search and close all exciting sever on demand to user. This lets the user to get relevant information on time. They compare the traditional search model with user interest based search model. The user interest model has accurate rate of giving relevant information on user demand.

Lingyging Zeng and Hao Wen Lin [2] of Harbin Institute of technology explain the concept of existing MapReduce and modified MapReduce to perform parallel computing to collect the hardware performance information from the virtual machine. The existing MapReduce will have master slave process, when the client request is generated master node will create a new job and assign to a new processor and is ready to perform. The master node always checks the salve process status is working based on that it will split and assign work to all available process and get combine all task. They used this concept in cloud computing which is dynamic and server will generate the request to the persistence independent storage device to collect and information.

Jinessh varia [3], technology Evangelist, Amazon Web services explained the cloud Architecture in June 2008. Varia explained how to develop an efficient, reliable, scalable, distributed parallel application using Amazon

Web Service which is loosely coupled system. Explained development of application with GrepTheWeb Hadoop implementation based search engine deployed using Amazon Web service. He also explained Amazon web service such Amazon S3 which is used to get input and output, Amazon SQS act as message passing, Amazon SimpleDB a database to get status, Amazon EC2 a controller.

Gaizhen Yang [4] in 2011 explained the application of MapReduce in Cloud Computing. Hadoop is the frame work for cloud programmers and Map Reduce is the parallel computing large scale programming model. He analyses the Hadoop and MapReduce model and described how this both can perform together that’s Map Reduce program in distributed cloud computing programming.

Kejiang Ye, Xiaohong Jiang, Yanzhang He, Xiang Li, Haiming Yan, Peng Huang [5] in 2012 discusses A Scalable Hadoop Virtual Cluster Platform for MapReduce-Based Parallel Machine Learning with Performance Consideration. Big data processing is increasing its important because of increasing data. Efficiently process large data virtual infrastructure is not clear at present. He clearly explained based on the performance of Hadoop and vHadoop. The performance is measured based on clustering, k-means, on vHadoop.

Zhiqiang Liu, Hongyan Liu, Gaoshan Miao [6], in 2010 proposed MapReduce-based Backpropagation Neural Network over Large Scale Mobile Data. MapReduce-based Backpropagation Neural Network is proposed to process classifications on large-scale mobile data. MapReduce-based framework on cloud computing platform is discussed to improve the efficiency and scalability over large scale mobile data. MapReduce framework is well known as a parallel programming model for cloud computing. It supports the parallelization of data processing on large clusters and built on the top of a distributed file system. However the research of how to design a neural network on MapReduce framework is rarely touched nowadays especially over large scale mobile data.

Closed frequent Itemset mining [7] plays important role in many real world applications. Cost and handling of large dataset is challenging issues of such data mining. A parallelized AFOPT-close algorithm is proposed and implemented based on the cloud computing framework MapReduce in 2012 by Su Qi Wang, Yu Bin Yang, Guang Peng Chen, Yang Gao and Yao Zhang.

4. METHODOLOGY

4.1 OVERVIEW OF SYSTEM

The overview of the proposed system is shown in Figure 1.

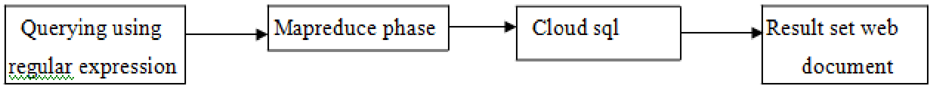

Figure 1. Overview of Engine for Extraction of Similar Resultant Web Document

Design of search engine is divided into modules. The process of sending request and getting result has four steps. First is launching of request, here the input query is validated and Hadoop is initiated. Second is map data and reduce based on matched input data from the cloud data base. Third is billing of used data for processing Hadoop and stop the Hadoop process. Fourth one is giving back the resource to cloud database by cleaning all data used in the application.

4.2 SEARCH ENGINE ARCHITECTURE

The system architecture depicted in Figure 2 implies that the GAE design will get the query from the user as regular simple expression, then process the request to the mapreduce phase which split the expression data set into small sub set and request is sent to all different database machines. After the extraction of resultant web document which matches the expression it will combine into a single resultant set and produce it to the user as web document.

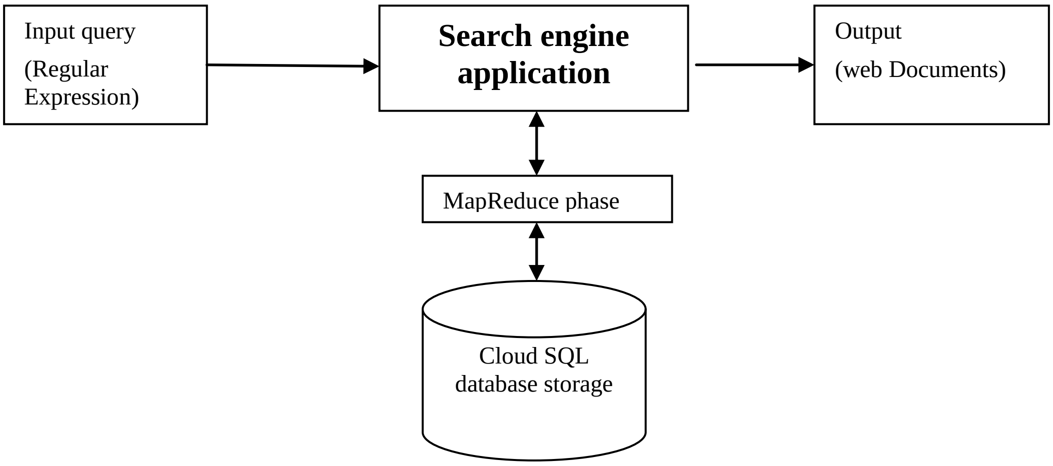
Input query (Regular Expression)
Search engine application
Output (web Documents)
MapReduce phase
Cloud SQL database storage

Figure 2. Search Engine System Architecture using cloud sql

Search engine application is developed to provide the user software as a service (SaaS). This application is developed based on to give efficient web search to user. Search engine uses regular expression as a query to search into the cloud database. This regular expression is run over millions of web document using Hadoop map reduce concept. It uses matching pattern to retrieve the document which is matched at most of the user entered regular expression query. The challenges in designing search engine is that complex regular expression, if there are many web document which matches, or else pattern is unknown. This application is overcomes all of such difficulties and gives result to number of users even with large dataset, with quick response and cost of usage is less. This is done over because of mapping is done parallel in number of processor then reduce and combine into smaller needed information.

4.3 HADOOP MAPREDUCE IMPLEMENTATION

The Mapreduce implementation is pictorially shown in Figure 3.

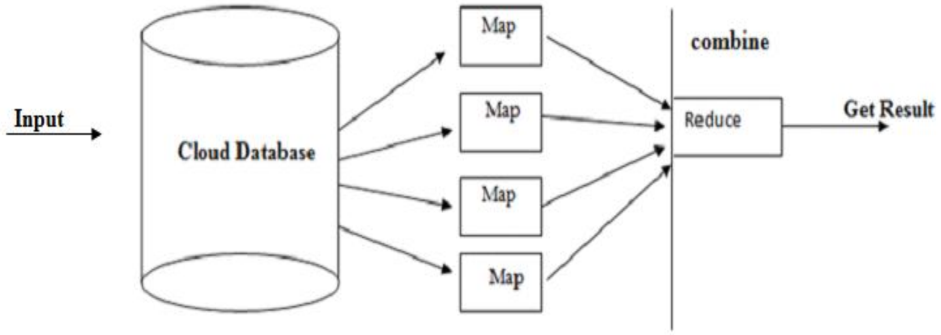

Figure 3. Mapreduce phase implementation in cloud database

Hadoop split dataset into manageable data and give it to many machines, job launched and processed in different machine which is located physically wide somewhere because of its open source and distributed which can manage large dataset. After that the result of all are aggregated as final output of job. It works in three phases to implement this. Map phase will map the data which is matched with the regular expression from the cloud database. Reduce phase will produce intermediate result of the web document. Map and reduce phase is done independent of each other in separate processor. Combine phase will combine all the extracted data from different machine. Thus needed data will be computed from all over the cloud data base and processed parallel to give efficient search result.

Hadoop use the master slave process, master process will run in separate node and see all the slave process which runs in some other separate node. Salve process all workers which extract data from different machine if any failure in worker or any problem will be take care by master process.

5. RESULTS

Application implementation in Figure 4 shows the start up page of search engine application which is developed and deployed using Google app engine and web tool kit. The application ask the user to enter search string and shows web document which match the search string based on Map Reduce concept. Because Map Reduce concept uses parallel computation search result will be mapped and computed fast so that response time of application will increase. Cloud SQL instance is used to access the cloud database and get all resource need for result and after the process is over resource is released back to the cloud.

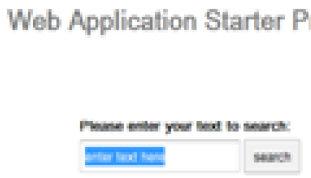
Web Application Starter Project
Please enter your text to search:
enter text here
search

Figure 4. Search engine application start up page

6. CONCLUSION

In this paper Search engine application is successfully designed, developed and deployed using Google Apps engine and cloud instance sql database. Search engine performs pattern matching across millions of web document using Apache Hadoop Map-Reduce for regular expression inputted by the user for query processing. Because of using map reduce concept, millions of documents are pattern matched in parallel at a time and result is combined and given to user as a web document. The process uses parallel distributed processing across many dataset gives the quick response to the user and also scale for any number of users. Application uses cloud sql data base using instance created for the application, so that billing of used resources from cloud computing data base can be easily maintained.

References

[1] Wang C., Yan Z., Chen H., 2010. Search engine concept based on user interest model and information push mechanism. 8th International Conference on computer science and education, Sri Lanka.

[2] Zeng L. and Lin H. W. 2012. A modified mapreduce for cloud computing. International conference on computing, measurement, control and sensor networks.

[3] Jinesh Varia, explained Cloud Architectures in Technology Evangelist Amazon Web Services in June 2008

[4] Gaizhen Yang, "The Application of MapReduce in the Cloud Computing", International Symposium on Intelligence Information Processing and Trusted Computing in 2011.

[5] Kejiang Ye, Xiaohong Jiang, Yanzhang He, Xiang Li, Haiming Yan, and Peng Huang, "vHadoop: A Scalable Hadoop Virtual Cluster Platform for MapReduce-Based Parallel Machine Learning with Performance Consideration", IEEE International Conference on Cluster Computing Workshops in 2012.

[6] Zhiqiang Liu, Hongyan Li , Gaoshan Miao, "MapReduce-based Backpropagation Neural Network over Large Scale Mobile Data", Sixth International Conference on Natural Computation (ICNC 2010) in 2010.

[7] Su Qi Wang, Yu Bin Yang, Guang Peng Chen, Yang Gao and Yao Zhang, "MapReduce-based Closed Frequent Itemset Mining with Efficient Redundancy Filtering in IEEE 12th International Conference on Data Mining Workshops in 2012.

This paper may be cited as:

Saranya, A. L. and Murugan, B. S., 2014. Cloud Architecture for Search Engine Application. *International Journal of Computer Science and Business Informatics, Vol. 10, No. 1, pp. 1-7.*