

Measuring Similarities between Distributed and Cloud Computing

K. Devi Priya

Department of Computer Science and Engineering
Aditya Engineering College
Surampalem, East Godavari, AndhraPradesh,
India

K. Bhanu Rajesh Naidu

PG Scholar, Department of CSE
Kakinada Institute of Technology & Science
Divili (Tirupathi),
India

Abstract— Distributed Computing is to run two or more parts of the program executed concurrently on different machines. In distributed computing, many types of concepts are identified like distributed file system, distributed databases, distributed operating system etc. Cloud Computing is to offer services to the users on demand basis. In this paper, we analysed the architecture, computing differences between distributed computing and cloud computing and also analyzed distributed databases in the cloud. This analysing leads to researchers to develop effective computing architecture, programming models for both distributed and cloud computing.

Keywords— *Distributed computing; cloud computing; distributed database; cloud database; distributed file sytem;*

I. INTRODUCTION

Distributed Computing[1] is to run two or more parts of the program executed concurrently on different machines. In distributed computing, many types of concepts are identified like distributed file system, distributed databases, distributed operating system etc. Data base is a collection of interrelated data. Database management system is a software is to manage and retrieve the data which is stored in the database. A relation database which is having relationship between tables. A relation data base is specific to ACID properties. In relational database, the user must clearly need to specify the schema. Now a days huge amount of data is generated from the multiple sources per sec. Data[4] is not structured also, It very difficult task to store data in single database. For this reason distributed computing and distributed data base systems are required. Distributed data base is collection of databases are located in different places and connected through the internet. Cloud computing [2] is a growing technology which provides services to the users based on their requirements on pay per utilization. Data storage is one of the popular service provide by the cloud to store huge amount of the data..A distributed database system allows applications to access data from local and remote databases. Distributed database systems are two types. Those are homogenous distributed system and heterogeneous distributed database. In homogenous distributed database all the sites have identical software to access data. In heterogeneous distributed database system, at least one of the databases is a non-Oracle database. Various Cloud Providers implement different distributed databases like mongo db, Cassandra etc. A cloud database is a data base which is run on cloud .In cloud ,how to create database is based on two deployment models. First model, choosing the

predefined virtual machine, second one is purchasing database as a service. Choosing virtual machine is is the task with built in database or creating our own database in virtual machine. In second model, cloud providers provides data base as a service with out installing virtual machine. In this service ,the user chooses the required type of data base and pays according to utilization

II. COMPUTING PARADIGM DISTINCTIONS

The computing paradigm are mainly classified as centralized computing, parallel computing ,distributed computing and cloud computing. In centralized computing all the physical resources are centralized in one system and all the resources are tightly coupled in one integrated operating system. In parallel computing all processors are tightly coupled with shared memory or loosely coupled with distribute memory. In distributed computing, the problem is divided into multiple parts and each part is solved by different computer and connected through the network .Cloud computing is collection of resources available through the internet and these resources are centralized or distributed which is accessed by any one through a simple web browser”.

III. CLOUD COMPUTING OVER THE INTERNET

Cloud Computing[5] is a development paradigm which offers services to the users on pay per use. Cloud services are mainly divided in to three types. Those are Infrastructure as a Service (IAAS), Platform as a Service (PAAS) and Software as a Service (SAA).

IAAS: In this service storage, networks, processors are offered as a service. Google, Amazon, Azure etc cloud providers offers these service.

PAAS: In this service, development platform of the applications are offered as a service. Google, Amazon, GOGrid etc cloud providers offers these service.

SAAS: In this service application of particular task purpose is provided as a service. Drop Box providing a service called storing of files, CRM service etc.

IV. SOFTWARE ENVIRONMENTS FOR DISTRIBUTED SYSTEMS AND CLOUDS

In this ,discussed software environments[1] for using distributed and cloud computing systems.

1. Service Oriented Architecture
2. Distributed Operating Systems
3. Parallel and Distributed Programming Model

A. Service Oriented Architecture

A service oriented architecture is collection of services and communicated each other. The first SOA is ORB (Object Request Broker Architecture) CORBA is the acronym for Common Object Request Broker Architecture. It was developed under the auspices of the Object Management Group (OMG). It is middleware. A CORBA-based program from any vendor, on almost any computer, operating system, programming language, and network, can interoperate with a CORBA-based program from the same or another vendor, on almost any other computer, operating system, programming language, and network.

B. Distributed Operating Systems

Distributed operating system is software over a collection of independent, networked, communicating, and physically separate computational nodes. Each individual node holds a specific software subset of the global aggregate operating system. Each subset is a composite of two distinct service provisions.

C. Parallele and Distributed Programming Model

- **MPI**-Message Passing Interface mainly points the message-passing parallel programming model. In this model data is moved from the address space of one process to that of another process through supportive operations on each process .
- **Map Reduce** is a paradigm[3] which solves the problem with the help of set of maps and reducers.Mapper is a procedure which takes input in the format called<key, value>pair and generates <key, value>output. In this paradigm there is no communication between mappers.
- **Reducer**-The reducer[3] job is to process the data which is coming from the map and generates new output. The map reduce approach is used to solve big data problem.
- **Hadoop** is an open-source framework [3]that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. The following table shows parallel programming models with features

Table 1 : Parallel Programming model

| SN O | MODEL | DESCRIPTI ON | FEATURES |
|------|------------|---|--|
| 1 | MPI | Data moving from one process to other process | Practical Portable Efficient Flexible |
| 2 | MAP REDUCE | Handling and Processing large data sets | 1.<Key,Value paradigmim> 2.Map and Reducer functions |
| 3 | HADOOP | Framework | 1.Distributed file system 2. Hadoop ecosystem have map reduce,Hbase, Hive,PigLatin paradigms. |

Table 1 : Parallel Programming model

| SN O | MODEL | DESCRIPTI ON | FEATURES |
|------|------------|---|--|
| 1 | MPI | Data moving from one process to other process | Practical Portable Efficient Flexible |
| 2 | MAP REDUCE | Handling and Processing large data sets | 1.<Key,Value paradigmim> 2.Map and Reducer functions |
| 3 | HADOOP | Framework | 1.Distributed file system 2. Hadoop ecosystem have map reduce,Hbase, Hive,PigLatin paradigms. |

V. DISTRIBUTED DATABASES

A centralized distributed database management system integrates the data logically so it can be managed as if it were all stored in the same location. The DDBMS synchronizes all the data periodically and ensures that updates and deletes performed on the data at one place will be automatically reflect in the data stored elsewhere. Distributed databases can be homogenous or heterogeneous. In a homogenous distributed database system, all the physical locations have the same underlying hardware and run the same operating systems and database applications. In a heterogeneous distributed database, the hardware, operating systems or database applications may be different at each of the locations. There are two principal approaches to store a relation r in a distributed database system:

A. Replication

In replication, the system maintains several identical replicas of the same relation r in different sites. Data is more available in this scheme.

- Parallelism is increased when read request is served.
- Increases overhead on update operations as each site containing the replica needed to be updated in order to maintain consistency.

B. Fragmentation

The relation r is fragmented into several relations $r_1, r_2, r_3, \dots, r_n$ in such a way that the actual relation could be reconstructed from the fragments and then the fragments are scattered to different locations. There are basically two schemes of fragmentation:

- Horizontal fragmentation - splits the relation by assigning each tuple of r to one or more fragments.
- Vertical fragmentation - splits the relation by decomposing the schema R of relation r .

The following figure shows distributed database architecture.

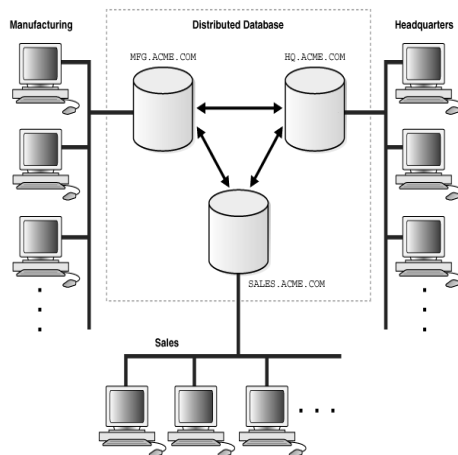


Fig 1: Distributed Database Architecture

VII. CLOUD DATABASES

Cloud database is a service provided by the third party and executed in the cloud. A traditional database system is installed on a server at an organization's site and data is stored and accessed directly or over a local area network (LAN). A cloud database management system, on the other hand, runs on a cloud provider's platform and data can only be stored or accessed when there is an Internet connection.

A cloud DBMS can be deployed in three different ways. The first way is as a virtual machine (VM) image. In this deployment model, the cloud provider sells virtual machine instances upon which a database management system can run. The provider is responsible for the infrastructure that supports the VM and the customer is responsible for uploading or purchasing the DBMS, making sure the DBMS is maintained properly and managing the databases it supports.

In the second deployment model, the cloud provider is responsible for supplying and maintaining the DBMS. The customer is responsible for managing the databases the DBMS supports and paying for storage and compute resources. This type of implementation is called Database as a service (DBaaS).

In the third deployment model, the cloud provider installs, maintains and manages the entire database implementation. This approach, which is called managed hosting, can provide a small organization with the benefits that a database provides without the administrative responsibilities and IT overhead typically required of DBMS usage.

VIII. CONCLUSION

In this paper, we analyzed distributed computing and cloud computing paradigms in terms of services, parallel and distributed programming models, software environments of distributed databases and clouds, distributed databases and cloud databases which is useful for researchers to find effective designing of databases, softwares in distributed and cloud computing model.

REFERENCES

- [1] Kai Hwang, Jack Dongarra, Geoffrey C. Fox. Distributed and Cloud Computing: From Parallel Processing to the Internet of Things
- [2] Mazhar Ali, Sameer U. Khan, Athanasios V. Vasilakos. Security in cloud computing: Opportunities and challenges. Information Sciences
- [3] hadoop.apache.org
- [4] From Databases to Big Data. S. Madden - IEEE Internet Computing, 2012 - search.ebscohost.com
- [5] <https://azure.microsoft.com/en-in/overview/what-is-cloud-computing/>
- [6] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy. Facebook Data Infrastructure Team. "Hive - A Petabyte Scale Data Warehouse using hadoop"

VI. HIVE: LARGE-SCALE, DISTRIBUTED DATA PROCESS

Hadoop is a framework to store and process huge amount of data in distributed environment. The Hadoop ecosystem contains several modules HDFS (Hadoop Distributed File System), MapReduce, Hive, Pig Latin etc., for processing huge amount of data. HDFS is used to store datasets. Map Reduce is used for to process structured, semi structured and un structured data. Pig Latin is used for processing semi structured data and Hive is used for processing structured data.

Hive [6] is a data ware house tool similar to sql. The commands of hive are create database/table/schema, alter database/table, load data, select queries etc.,