International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Survey on Data Management in Cloud Computing

Nagaveena N K [1], Chandrani Chakravorty[2]

[1]*PG Student, Master of Computer Applications, R V College of Engineering®, Karnataka, India*
[2]*Assistant Professor, Master of Computer Applications, R V College of Engineering®, Karnataka, India*

Abstract - *Data is something that is more precious than anything in this fast-moving world driven by technology. Every second thousands of TB of data is being released from multiple sources on the internet like doing WhatsApp messages, searching something on the internet(basically the keywords used while searching), using multiple platforms to fulfil the needs like Login into the e-commerce websites also experiences a huge amount of data traffic per day. The Storage of data in order to make it available for user 24*7 has always been a tough task. Because there is no such estimation made till now that how much amount of data they can tend to receive per day it may be different day by day, season per season like in every festival season the e-commerce websites experience a large amount of traffic as compared to the other seasons or the other regular days. Examples of such situations may be Friday Sale in America, Big Billion Day Sale in India on e-commerce websites generates a large amount of data in terms of user credentials, searching patterns, address details, payment details. These all come under the confidential data of customer. There are multiple platforms being utilized by many companies in order to manage the data efficiently. Cloud Computing is one of the up-to-the-minute subjects in the fast-growing industry to store data remotely and then applying multiple algorithms to the data in order to make it competent to respond. Cloud Computing data management norms and conditions have always been a keen interest for companies in order to shift their database to cloud to scale the database depending upon the number of users at different intervals of time and make it available to usage round the clock.*

Key Words: Database, NoSQL, SQL, Paxos Algorithm, query optimization, serialization of jobs, Cloud Data Management Infrastructure

1.INTRODUCTION

Data management in Cloud Computing deals with majorly two different but enormous areas one is Data and secondly Cloud Computing. Both have their own importance as it can be easily observed in this emerging new industry trends. **Necessity is the mother of invention**. This was something that lead to the invention of Cloud Computing by Joseph Carl Rapnet Licklider in 1960s with his work on ARPANET to connect people and data at any given time. Cloud computing is defined as the provision of computer services such as network, servers, databases, storage, virtualization, storage, software, business analytics. It provides product innovations and flexible resources for the business, such as payment for usage from the cloud. Benefits of Cloud Computing: Flexible Resources - Depending on the needs of the service, users can quickly or under-utilize resources. The Meter service gives you the responsibility to pay for your use. Self Service You can access all IT resources without any help. Cloud Computing is mainly comprising of two words Cloud and Computing. Cloud means connecting remote devices in order to fulfil certain requirements. Cloud Computing is basically an on-demand management service offered by multiple cloud service providers in order to store and scale up their infrastructure from scratch. Cloud Computing comes with multiple services for customers like choosing the infrastructure as per their own requirement , choosing hardware requirement writing rules about transmission of data generating proper secured pathway and then designing database according to requirement and then finally deploying their application on cloud so that as much as possible the services can be reached out. Data Management is one of the key aspects of Cloud computing which assures the developer about the traffic of data generated per second, per day and then regulating the servers as per the requirement. Cloud Data Management considers that data stored on the premises and in the cloud may be subject to completely different practices and that data stored in the cloud has its own rules for data integrity and security. Traditional data management practices may not apply to the cloud, so management designed for the specific needs of the cloud is important. Cloud is helpful as an information stockpiling level for catastrophe recuperation, reinforcement, and long-haul documenting with the information the board in the cloud, assets can be bought varying Data can likewise be shared across private and open mists, just as in on-premises stockpiling.

2. LITERATURE SURVEY

Investigates the focal points and inconveniences of sending database frameworks in the cloud. The Author observes how the regular properties of financially accessible distributed computing stages influence the selection of information the executives' applications to send in the cloud. Due to the ever-expanding requirement for more examination over additional information in the present corporate world, alongside a building match in right now accessible sending choices, the reason that read-for the most part logical information the board applications are more qualified for arrangement in the cloud than value-based information the board applications. Accordingly, layout an exploration plan

for the enormous scope of information examination in the cloud, demonstrating why as of now accessible frameworks are not in a perfect world appropriate for cloud arrangement and contending that there is a requirement for a recently planned DBMS, architected explicitly for distributed computing stages [1].

Explains about the administration of huge sensor information produced by fabricating gadgets is a significant issue in Cloud Assembling. This needs a structure supporting disseminated the capacity of semi-or un-organized fleeting information and proficient equal handling of ultra-enormous informational collections. This paper presents the structure supporting monstrous sensor information the executives based on Hadoop innovation. This system can give an answer to compose the gigantic sensor information viably and figure it out equal preparation productively, which will make up for the restrictions of the customary connection database [2].

Briefly explains about various problematic and transformative enormous information advancements and arrangements that are quickly radiating furthermore, developing so as to give information-driven knowledge and advancement. The essential object of the examination was the manner by which to group mainstream large information assets the board frameworks. The study was likewise planned for tending to the determination of up-and-comer asset suppliers dependent on explicit large information application prerequisites. The author studied diverse enormous information assets the board structures and explored the points of interest and burdens for every one of them [3].

Introduced a writing audit and classification of procedures to settle asset the board issues in cloud DCs. The objective of this study has been to centre on the over-underloaded host and VM arrangement issues while introducing the streamlining objectives, utilized techniques, and goal models. Revealed work has been characterized and arranged based on this grouping. Added to that, the process also utilized a cloudsim test system to execute two kinds of target models: mono goal what's more, multi-objective. Through test results, certain facts are also indicated that while upgrading one more target have accomplished preferable exhibitions over one goal, in terms of vitality utilization and asset use [4].

Presents a key administration scientific categorization for distributed computing. The author thinks about key administration techniques by taking parameters, for example, versatility, adaptation to internal failure, and so forth. In order to apply key administration ways to deal with different cloud situations. for information stockpiling, recognize the applications and appropriate key the board draws near. A different eye at different key administration approaches by applying them to different cloud conditions. At last, the author dissected which symmetric key calculation is quicker with the goal that it may not abuse the qualities of distributed computing, for example, on-request access, versatility, and estimated administration, and so on. The Author also explains about future work which is to actualize a reasonable key administration strategy for a specific cloud condition and do execution examination [5].

Explains about distributed computing which is imagined as the cutting-edge design of IT Enterprise is an all the rage nowadays. The manner in which the cloud has been commanding the IT advertise, a significant move towards the cloud can be normal in the coming years. Distributed computing offers genuine advantages to organizations looking for a serious edge in the present economy. A lot more suppliers are moving into this region, and the opposition is driving costs even lower. Appealing evaluating, the capacity to let loose staff for different obligations, and the capacity to pay for —as needed‖ administrations will keep on driving more organizations to consider distributed computing. Versatile distributed computing is relied upon to rise as perhaps the greatest market for cloud specialist co-ops and cloud engineers. Distributed computing can possibly turn into a leader in advancing a protected, virtual, and monetarily feasible IT arrangement later on. As the advancement of distributed computing innovation is still at a beginning period, this exploration exertion will give a superior comprehension of the structural difficulties of distributed computing, and make ready for additional examination here [6].

Briefly explains the rise of Cloud processing that presents numerous challenges that may restrict the selection pace of the Cloud worldview. As information volumes handled by applications running on Clouds increment, the requirement for effective and secure information the executives develop as an essential prerequisite. This work expects to empower BlobSeer, enormous scope information the executives the framework, as a Cloud information administration, by tending to a progression of self-administration necessities. The initial move towards autonomic conduct was to prepare the BlobSeer stage with contemplation abilities, which can fill in as information for a self-versatile motor planned to address such objectives as self-arrangement, self-advancement or on the other hand self-insurance. The Author also explains about the built up for self-assurance course inside conventional security the board system permitting suppliers of Cloud information the executive's frameworks to characterize and uphold complex security strategies. Also, the author structured an expressive approach depiction language empowering framework overseer to characterize an enormous exhibit of security assaults and to implement different kinds of limitations upon the recognized vindictive customers [7].

Experiment gives information the executives are getting basic, given the wide scope of capacity areas and plenty of

cell phones. The information has gotten away from IT division control into the more extensive ranges of cloud-based administrations, cell phones, and social organizing. The multiplication and development of cloud registering will proceed to reinforce and drive the requirement for solid cloud database administrations. The year 2012 will see the turn of events and development of increasingly more NoSQL databases [8].

States a general system for giving Information Encryption as a Service supporting diverse security arrangements in a multi-cloud condition. The system is adaptable to help any security arrangement in any cloud stage through the execution of explicit modules. It gives full mechanization of information encryption administration to clients in a complete life-cycle from information encryption to information unscrambling. An evidence of-idea model of the system was actualized adjusted to a subset of the necessities evoked by the BT use-instance of ESCUDO-CLOUD undertaking and utilizing a cloud commercial center created by BT in the STRATEGIC to demonstrate the practicability of plan [9].

Explains about cloud computing environment, without the virtualization technique, it would not be possible to use a single hardware device among the users. It is the basic service of any development in cloud computing. Data management in cloud computing shows the rapid growth of deployment in remote servers for the purpose of storage and cloud services. Cloud BigTable is mainly used for the non-transactional data where it does not give any redundancy for the data. It can be used for data analytics where you can get the results by querying historical data. Cloud DataStore is built on BigTable but they are completely different from each other, where it supports ACID properties of the transaction and it is used on transactional data. It features are similar to SQL but it cannot perform some operations [10].

2.1 SUMMARY OF LITERATURE SURVEY

The creator examines the central focuses and burdens of sending database systems in the cloud. The opportunity yields a glimpse at how the customary properties of monetarily open circulated processing stages impact the determination of data the administrators' applications to send in the cloud. Due to the ever-extending prerequisite for more assessment over extra data in the present corporate world, nearby a structure coordinate in right now available sending decisions, and provide an explanation that read-generally coherent data the board applications are more equipped for the course of action in the cloud than esteem based data the board applications. The needs be formatting an investigation plan for the gigantic extent of data assessment in the cloud, exhibiting why starting at now open structures are not ideally proper for cloud course of action and battling that there is a necessity for an as of late arranged DBMS, architected unequivocally for appropriated processing stages. Yuan Bao clarifies about the organization of immense sensor data delivered by manufacturing contraptions is a huge issue in Cloud Assembling. This needs a structure supporting dispersed the limit of semi-or un-sorted out temporary data and capable equivalent treatment of ultra-gigantic instructive assortments. This paper presents the structure supporting colossal sensor data the administrators dependent on Hadoop advancement. This framework can offer a response to create the huge sensor data suitably and make sense of its equivalent planning beneficially, which will compensate for the limitations of the standard association database. There are different hazardous and transformative colossal data headways and game plans that are rapidly transmitting moreover, growing in order to give data-driven information and progression. The basic object of the assessment was the way to bunch standard enormous data resources the board systems. The investigation was in like manner made arrangements for keeping an eye on the assurance of up-and-comer resource providers subject to express huge data application requirements. The creator considered various tremendous data resources the board structures and investigated the focal points and weights for all of them. In this paper, the writer has presented a composing review and characterization of systems to settle resource the board issues in cloud DCs. The goal of this investigation has been to focus on the over-underloaded host and VM course of action issues while presenting the smoothing out targets, used procedures, and objective models. Uncovered work has been described and organized dependent on this gathering. Added to that, the procedure likewise used a cloudsim test framework to execute two sorts of target models: mono objective what's more, multi-objective. Through test outcomes, certain realities are additionally demonstrated that while overhauling one more objective has achieved ideal shows more than one objective, as far as imperativeness usage and resource use. Amar R. Buchade, Rajesh Ingle have presented key organization logical order for circulated registering. The creator considers key organization strategies by taking parameters, for instance, adaptability, adjustment to interior disappointment, etc. So as to apply key organization approaches to manage diverse cloud circumstances. for data amassing, perceive the applications, and proper key the board moves close. An alternate eye at various key organization approaches by applying them to various cloud conditions. Finally, the creator analyzed which symmetric key count is snappier with the objective that it may not mishandle the characteristics of appropriated registering, for instance, on-demand access, flexibility, and evaluated organization, etc. The Author likewise clarifies about future work which is to realize a sensible key organization methodology for a particular cloud condition and do execution assessment. Dispersed figuring, envisioned as the front-line structure of IT Enterprise is extremely popular these days. The way wherein the cloud has been instructing the IT publicize, a critical move towards the

cloud can be typical in the coming years. Appropriated figuring offers real preferences to associations searching for a genuine edge in the current economy. Significantly more providers are moving into this district, and the restriction is driving expenses even lower. Engaging assessing, the ability to let free staff for various commitments, and the ability to pay for as needed organizations will continue driving more associations to think about conveyed registering. Adaptable circulated figuring is depended upon to ascend as maybe the best market for cloud master communities and cloud engineers. Conveyed processing can transform into a pioneer in propelling an ensured, virtual, and fiscally attainable IT course of action later on. As the headway of circulated registering advancement is still at a starting period, this investigation effort will give a predominant perception of the auxiliary troubles of conveyed figuring, and prepare for extra assessment here. The ascent of Cloud handling presents various difficulties that may confine the choice pace of the Cloud perspective. As data volumes took care of my applications running on Clouds increase, the necessity for successful and secure data the administrators create as a fundamental essential. This work hopes to engage BlobSeer, colossal extension data the administrators the structure, as a Cloud data organization, by keeping an eye on a movement of self-organization necessities. The underlying move towards autonomic lead was to set up the BlobSeer stage with thought capacities, which can fill in as data for a self-adaptable engine intended to address such destinations as self-course of action, self-progression or then again self-protection. The Author likewise clarifies the developed for confidence course inside traditional security the board framework allowing providers of Cloud data the official's structures to portray and maintain complex security systems. Additionally, the creator organized an expressive methodology portrayal language enabling the system administrator to describe a gigantic display of security ambushes and to actualize various types of impediments upon the perceived pernicious customers. Information about the officials is getting fundamental, given the wide extent of limit regions and a lot of mobile phones. The data has escaped from IT division control into the broader scopes of cloud-based organizations, mobile phones, and social sorting out. The augmentation and improvement of cloud enlisting will continue to fortify and drive the prerequisite for strong cloud database organizations. The year 2012 will see the new development and improvement of progressively more NoSQL databases. In this paper, the creator has presented a general framework for giving Information Encryption as a Service supporting assorted security game plans in a multi-cloud condition. The framework is versatile to help any security course of action in any cloud stage through the execution of express modules. It gives full motorization of data encryption organization to customers in a total life-cycle from data encryption to data unscrambling. A proof of-thought model of the framework was realized changed in accordance with a subset of the necessities evoked by the BT use-occurrence of ESCUDO-CLOUD undertaking and using a cloud business focus made by BT in the STRATEGIC to show the practicability of plan .In a distributed computing condition, without the virtualization procedure, it would not be conceivable to utilize a solitary equipment gadget among the clients. It is the essential help of any improvement in distributed computing. Information the executives in distributed computing shows the fast development of sending in remote servers with the end goal of capacity and cloud administrations. Cloud BigTable is mostly utilized for the non-value-based information where it doesn't give any repetition for the information. It tends to be utilized for information investigation where you can get the outcomes by questioning verifiable information. Cloud DataStore is based on BigTable however they are totally not the same as one another, where it bolsters ACID properties of the exchange and it is utilized on value-based information. It highlights are like SQL yet it can't play out certain tasks.

3. APPROACHES FOR CLOUD COMPUTING

The Paxos algorithm runs on a messaging model with asynchronous and n / 2 crash failures (but not byzantine failures, at least in the original algorithm). As always, we want to work with random algorithms and cryptographic algorithms to check contract, validity and cancellation and random algorithms including greedy algorithms, segmentation and successes, dynamic programming, network flow, approximation algorithms.
Amazon Web Services is one of the leaders in cloud computing and data management solutions. Amazon EMR is available in 14 locations worldwide.

•Amazon S3 only allows the customer to choose from the US and EU data storage options, the customer has minimal choice but should be treated as the worst and can access the data until the data is encrypted using the key on the host. Third party without customer knowledge.

•Amazon's S3 cloud storage service replicates data in "regions" and "availability zones" so that data and applications can continue despite failures across the space.

•Amazon's Elastic Compute Cloud (EC2) calculates resources for small, large, and extra-large virtual private server instances, the largest of which is no more than four cores. If an application cannot take advantage of additional server instances, load some of its required functions onto the new instance running in parallel with the old instance.
Presto can be installed with any implementation of Hadoop, and Amazon EMR is packaged in Hadoop distribution.
Technology:
Cassandra being a column store NoSQL database, distributed database which is specifically intended to handle large amount of data across commodity hardware. It also supports high availability with no single point of failure. Cassandra supports multiple clusters in terms of datacenters.

Hive is the data warehouse built on Apache Hadoop in order to process and store large amount of data because if the data is complex in nature. Apache Hive executes the large requests with the help of SQL like Hive Query language also known as HQL in order to break the large amount of data in clusters and then running that as per the MapReduce job. Apache Hive plays an important role in circumstances of analysis of data because if the firm is thinking in terms of data management then data analysis will play an important role in analyzing the storage space, execution time and many other factors.

SQL also known as Structured Query language plays an important role fulfilment of user requirement. One cannot deny the fact that apart of many NoSQL databases available many firms relies upon the relational databases in order to store data and manipulate data. Data Management is done in databases using query optimization, creating routine procedures, using indexing in order to fetch data as fast as possible. Taking the necessary steps at right time will help in dealing with data management. Cloud service providers also provide a wide range of databases, RDMS also finds its place in that list. Cloud infrastructure provide an easy go through medium in order to manage the data with respect with fasten the process of response time.

BigQuery is a completely overseen, serverless information distribution centre that empowers versatile, financially understanding, and quick investigation over petabytes of information. It is a serverless Software as a Service that supports questioning utilizing ANSI SQL. It additionally has worked in AI abilities. BigQuery is a web administration from Google that is utilized for taking care of or investigating enormous information. It is a piece of the Google Cloud Platform. As a NoOps (no activities) information investigation administration, BigQuery offers clients the capacity to oversee information utilizing quick SQL-like inquiries for constant examination. BigQuery is a totally overseen arrangement meaning you do not setup any servers or introduce any software - you fair send it information and after that inquiry specifically.

non-relational sources like the Hadoop Distributed File System (HDFS), MongoDB, and HBase.

4. ARCHITECTURE DIAGRAM OF DATA MANAGEMENT IN CLOUD

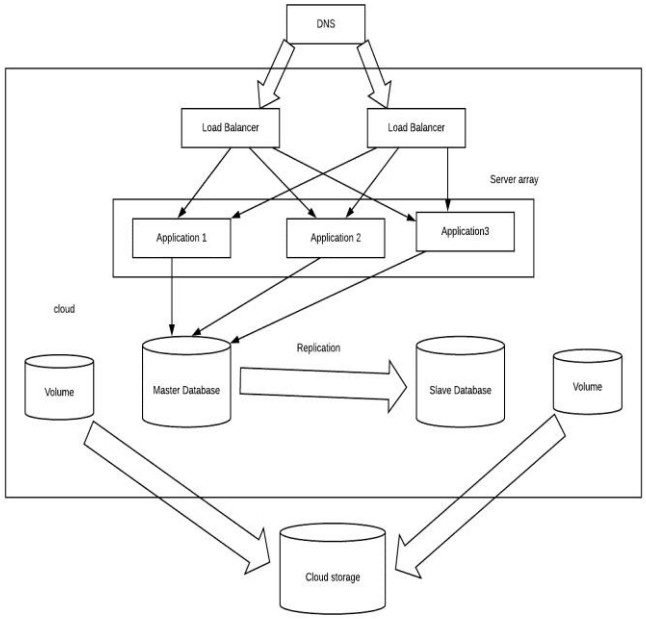

Fig -1: Architecture Diagram [11]

Cloud transitions can be difficult to begin. Transitions can be difficult to design and plan, as much of the diligence now falls on the customer side. This change is a double-edged sword; it cuts both ways. It enables the customer to have significantly more control over designs, technical choices, economics, and risk. It also places the significantly more of the cloud computing architecture burden on the customer, who may not have the level of solution design experience that many service providers do. A key benefit of cloud computing is the ability to consume what is needed when it is needed. Autoscaling describes the ability to scale horizontally (that is, shrink or grow the number of running server instances) as application user demand changes over time. Autoscaling is often utilized within web/app tiers within the baseline architectures mentioned. In the following figure, an additional server is dynamically added based on demand and threshold settings. Load balancers must be preconfigured or configured dynamically to handle the new servers that are added. The following Architecture design collapses both web and app onto the same virtual or physical server. Load balancers are added to the design to delegate the load across multiple servers. Database servers are shown as primary-backup with replication between them. This redundant architecture can protect against issues with applications due to system unavailability and downtime. Resiliency considerations may include RAID (Redundant Array of Independent Disks) configurations for database drives, how databases are backed up and restored, how applications and devices handle state and session information, and how databases rebuild after data or drive loss.

5. COMPARITIVE STUDY OF PNUTS AND GOOGLE BIG TABLE

Yahoo and Google are two prominent players in terms of dealing with large amount of data, processing it with a fraction of second and fulfilling the requirements of the user. Millions of users interact with these platforms for different purposes. Big Data, a new revolution in terms of amount of data encountered in terms of type of data, speed of data and amount of data. For this both firms have different largely scaled databases are available like for Yahoo it is PNUTS, and for Google it is BigTable in order to scale up data and process it as soon as possible in an optimized way. There are some points explained below in order to find which one is better for which purpose in this real competitive world.

•PNUTS bid for stringent consistency (database after summing up all the operations will be updated) as compared to Google BigTable as in terms of per-key CC.

•PNUTS communication between replicas of databases is much faster than other databases as it is designed with the principle of geo-replicated deployment. In this type of design implementation, one cluster stores a complete copy of the whole database.

•PNUTS also consists of multiple reads like READ-LATEST and READ-ANY which helps developer to take consequent decisions based upon the consistent and latency perspective whereas Google BigTable wins the battle by providing highest latency i.e. average read latency of less than 100ms of 65% of whole production dataset.

•Google BigTable uses ACID along with PAXOS algorithms where transactions services occur in ACID compliance which uses 2-phase commit lock, also solves multiple CAP theorem trade-offs. It also supports high precision clock synchronization.

•PNUTS by Yahoo also supports automatic load balancing, also uses failovers in order to reduce on-development complexity. Google BigTable runs the queries in terms of a map reduce job in order to execute the requests as fast as possible.

6.CONCLUSION

Data management in cloud computing environment, without the virtualization method it would not be conceivable to utilize single equipment hardware among the clients. It is the fundamental help of any advancement in distributed computing. Information the executives in data management in cloud computing shows the quick development of arrangement in remote servers with the end goal of capacity and cloud administrations. Cloud BigTable is basically utilized for the non-transactional information where it doesn't give any repetition for the information. It very well may be utilized for information examination where you can get the outcomes by querying the information. Cloud DataStore is based on BigTable however they are totally not the same as one another, where it supports ACID properties of the exchange and it is utilized on value-based information. It highlights are like SQL however it can't play out certain tasks and PNUTS supports a fundamental key-esteem information model in which esteems are essentially records as in a customary RDBMS. Records are naturally put away in allotments, which are re-adjusted by the framework for scale-out, and clients can control information association by alternatively determining a composite key for arranging a given record assortment (table).

REFERENCES

[1] Daniel J. Abadi," Data Management in the Cloud: Limitations and Opportunities", IEEE Computer Society Technical Committee on Data Engineering,2009

[2] Yuan Bao, Lei Ren, Lin Zhang, Xuesong Zhang, Yongliang Luo," Massive Sensor Data Management Framework in Cloud Manufacturing Based on Hadoop", 978-1-4673-0311-8/12/$31.00 ©2012 IEEE

[3] Saeed Ullah, M. Daud Awan, and M. Sikander Hayat Khiyal," Big Data in Cloud Computing: A Resource Management Perspective", Hindawi Scientific Programming Volume 2018, Article ID 5418679, 17 pages

[4] Khaoula Braiki, Habib Youssef, "Resource Management in Cloud Data Centres: A
Survey", 978-1-5386-7747-6/19/$31.00 ©2019 IEEE

[5] Amar R. Buchade, Rajesh Ingle," Key Management for Cloud Data Storage: Methods and Comparisons", 2014 Fourth International Conference on Advanced Computing & Communication Technologies

[6] Mohsin Nazir," Cloud Computing: Overview & Current Research Challenges", IOSR Journal of Computer Engineering (IOSR-JCE) ISSN: 2278-0661, ISBN: 2278-8727Volume 8, Issue 1 (Nov. - Dec. 2012), PP 14-22

[7] Alexandra Carpen-Amarie," Towards a Self-Adaptive Data Management System for Cloud Environments", 2011 IEEE International Parallel & Distributed Processing Symposium

[8] Deka Ganesh Chandra, Ravi Prakash, Swati Lamdharia," A Study on Cloud Database", 2012 Fourth International Conference on Computational Intelligence and Communication Networks

[9] Maurizio Colombo, Rasool Asal,et.all," Data Protection as a Service in the Multi-cloud Environment", 2019 IEEE 12th International Conference on Cloud Computing (CLOUD)

[10] K. Yogitha Lakshmi, S. Dhanalakshmi, B.G. Obula Reddy," An Overview of Data Management in Cloud Computing", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5C, February 2019

[11]https://www.networkcomputing.com/cloud-infrastructure/guide-cloud-computing-architectures