Database on Cloud - Information Retrieval

Ashita Shetty
Computer Science
Santa Clara University
ashetty2@scu.edu

Ashutosh Somaiya
Computer Science
Santa Clara University
asomaiya@scu.edu

Mansi Tandel
Computer Science
Santa Clara University
mtandel@scu.edu

Shivani Tergaonkar
Computer Science
Santa Clara University
stergaonkar@scu.edu

Surya Kiran Udaya Kumar
Computer Science
Santa Clara University
sudayakumar@scu.edu

I. ABSTRACT:

Cloud Computing is one of the latest and emerging technology in the current IT field. One of the most common use of cloud computing is cloud database. Cloud database is a database service that normally runs on a cloud computing platform as service. Cloud databases offer the users the agility to transform business requirements by giving access to its data centers without the hassle of maintaining one. The ease of usage and retrieving information from the databases have made database on cloud a popular service. In this paper we will be covering the basics of database on cloud and will be focusing on the process of information retrieval from such services. Cloud database provides numerous advantages but there are also some drawbacks that needs to be addressed. Like, on one hand thousands of user's information become available on central servers and businesses are able to achieve location independency whereas on other the other hand this causes data security issue as well as impacts the performance of the information retrieval process. In this paper we will be addressing such issues and talk in detail about them.

II. INTRODUCTION:

In today's fast paced world, everything gets tied up to data. To store such massive amount of data generated by each user on the Internet is one of the reasons why Cloud Computing is becoming very popular. Cloud computing is based on the concept in which large groups of remote data servers are networked to keep a centralized data storage. This centralized data storage then can be accessed from anywhere in the world through internet, thus providing location independence. Dedicated companies are responsible to maintain such centralized data centers, which in turn allows clients which use these services to not worry about maintenance of their data servers. Cloud service offers three different service models which each satisfy unique set of requirements of a business. The three models are: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

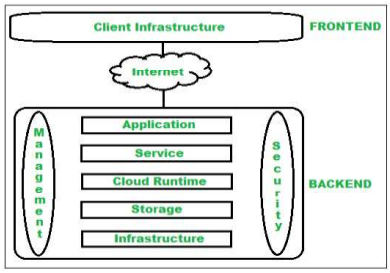

Figure 1: Architecture of cloud computing

Infrastructure as a Service (IaaS) is used to manage remote data center infrastructures. It is a self-service model which provides virtualized computing resources over the Internet which are usually hosted by a third party.

Platform as a Service (PaaS) allows organizations to build, run and manage applications without the IT infrastructure.

Software as Service (SaaS) replaces the traditional on-device software with software that is licensed on a subscription basis. This paper will cover Cloud Database, its advantages and process of information retrieval from such databases. A cloud database is a database deployed on a cloud, which provides location independence, and

maintenance free solution.

III. WHY CLOUD DATABASE

Accessibility, scalability, and disaster recovery are all significant features of cloud databases.

The scalability of cloud databases is one of its main advantages. A cloud database can scale with the size and complexity of applications. Hosting databases in the cloud lifts the limits on applications growth. Cloud databases are extremely flexible, as they can be set up and decommissioned in a matter of minutes. This flexibility makes testing, development, and validation much faster. Back-up and recovery are built-in to cloud databases, guaranteeing that your database is always available. Built-in backups help eliminate the danger of data loss if something goes wrong. Moving to a cloud database also eliminates the danger of downtime and provides consistent and dependable connectivity for those applications that rely on a constant connection to the databases that support them. The cloud service provider offers automatic security and features to limit the likelihood of a human error. Finally, cloud databases have the potential to reduce costs. A subscription- based approach is used by cloud service providers. This means that the firm or organization is charged a set fee based on the volume of data and how much it is used.

IV. DEPLOYMENT MODELS

Based on ownership, scale, and access, as well as the cloud's nature and purpose, the cloud deployment model determines the unique sort of cloud environment. A cloud deployment paradigm determines where servers are located and who has control over them. It describes how cloud architecture will look, what you can adjust, and whether you will be provided with services or must design everything yourself. Different types of cloud computing deployment models are:

Public Cloud:

This sort of cloud deployment architecture, as the name implies, supports all customers who want to subscribe to a computer resource, such as hardware (OS, CPU, memory, storage) or software (application server, database).

Private Cloud:

A private cloud, as the name implies, is primarily infrastructure used by a single company. Such infrastructure may be managed by the organization to support diverse user groups, or it may be handled by a service provider on-site or off-site

Hybrid Cloud:

In a hybrid cloud, an organization's private and public cloud infrastructure are interconnected. Many businesses use this technique when they need to quickly scale up their IT infrastructure, such as when using public clouds to supplement the capacity available in a private cloud.

Community Cloud:

It enables a group of organizations to access systems and services. It's a distributed system that's built by combining the services of many clouds to meet the demands of a community, industry, or enterprise.

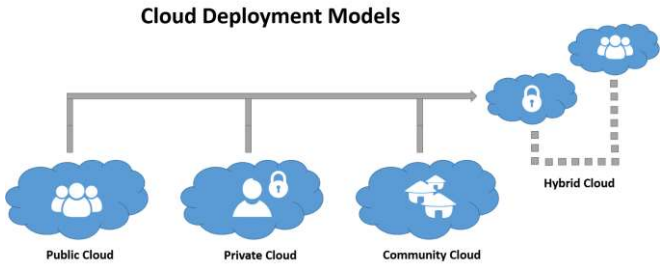

Figure 2: Cloud Deployment Models

Typically, cloud databases are installed in one of two ways.

Virtual Machine Image:

The first is the traditional paradigm, which entails a firm or individuals purchasing infrastructure and managing it privately. Virtual machines are typically purchased from cloud service providers as the infrastructure for the database to be housed on. The database, as well as anything related to it, is managed by the company or group of employees. The cloud service providers merely supply the hardware, such as a virtual machine, on which the organization can host a database.

Database-as-a-service (DBaaS):

Database as a Service, or DBaaS, is the other model. A corporation signs a contract with a cloud service provider to pay for a subscription service in this model. The service provider uses this paradigm to provide maintenance, administrative, and database management services, including data scalability, security, and recovery. Aside from these services, the supplier also supplies the infrastructure for the database to be hosted. This is useful for corporations and people who would rather not handle or deal with database setup and instead pay a company to do it.

V. ARCHITECTURE

The architecture of cloud databases is a stack of distinct service kinds. Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service are the three categories (SaaS). First and foremost, IaaS

provides the foundation for all software and can be thought of as the hardware equivalent in traditional stacks. Individual machine resources, whether physical or virtual, the operating system and network topology are all part of this infrastructure. IaaS is like a data center, but without the requirement for capacity planning or physical maintenance on the part of the customer. It's used to create and deploy PaaS and SaaS applications.

PaaS, or Platform as a Service, is the next step, which is a development and deployment environment. This is a cloud computing service that allows users to create and deploy their own cloud-based apps. PaaS is an extension of IaaS in that it includes infrastructure in addition to development and deployment environments. It is a step away from SaaS in that it is not a ready-made end-user application but rather a 'code environment' for creating such products, allowing for greater customizability.

Lastly SaaS gives a user interface that organizations may employ right away, as well as an unnoticed platform and infrastructure. Instead of being sold outright, this form of software is usually invoiced to customers on a monthly or yearly basis. The seller administers and maintains these products totally, limiting the consumer's ability to customize the product to their unique needs. This may or may not be an issue because SaaS products are very generic to appeal to a wide range of clients. The most well-known SaaS products include Gmail (for email), Salesforce (for CRM).

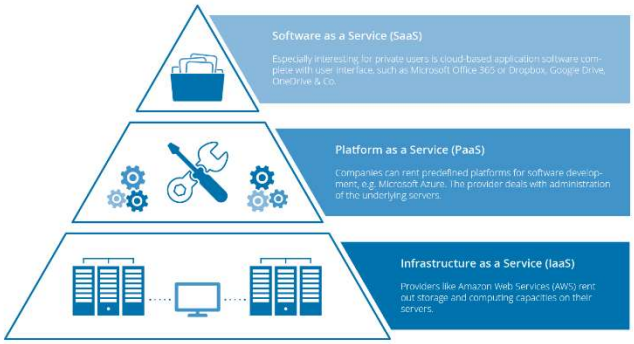

Figure 3: Cloud as service

VI. INFORMATION RETRIEVAL ON CLOUD DATABASE

Information retrieval in computing and IT is the process of obtaining information from a system like database or cloud. The information obtained is related to the query inputted by a user. Queries are statements of what the user needs, which when accepted by the system should retrieve the required information and present them back to the user. Information retrieval is an important aspect of database on cloud. Different techniques of information retrieval are used to store and extract data from the database as per the requirement of the user. With boom in technology development and need to store user's data in large scale, Information Retrieval has become one the most important aspect of cloud computing. Information retrieval system is a core support to internet-based services like cloud database and search engines. More and more advancements are taking place in the field of Information Retrieval to make the process of getting the data faster, reliable, and secure.

This paper will cover different techniques of retrieving information from the cloud database, its security aspect and advantages and disadvantages of the cloud database. The ever-increasing network bandwidth and reliable network connections make it possible for users to now subscribe to high quality data and software services from any part of the world. This increasing demand for cloud computing and databases due to its tremendous advantages makes it even more important to have a highly reliable information retrieval technique.

Hosting so much data over the internet has its own set of vulnerability and security concerns. In this paper we will also cover different ways on how the security of such high-risk data is maintained.

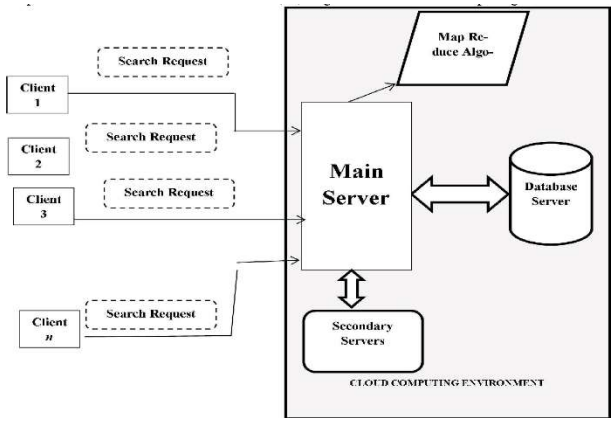

Figure 4: Implementation of Information Retrieval

VII. DIFFERENT WAYS OF INFORMATION RETRIEVAL

There are many methods used for information retrieval in the cloud. This paper discusses one such method.

Definition:

Cloud Information retrieval can be defined as follows:

{S, D, K, C, Q, F, R (Q, ci)} where

S [Security policy] -> defines the privacy and security policy.
D[documents] -> represents a group of documents.
K[keywords] -> is the list of keywords that are extracted from each document.
C [encrypted documents] -> is the list of encrypted forms of the documents D.Q is a list of queries for users.
F[Framework] -> is used for modeling queries, documents, and their relationships.
R [Ranking algorithm] -> is a function that determines the order in which the results are displayed based on the query Q given by the users.

If the above function works perfectly and all the security standards are maintained, a list of keywords are extracted. Sometimes only a few keywords or even one keyword is extracted. The users want their results using as less communication as possible, this can be a challenge because when the searchable keywords are so less, it is difficult to establish a relation between the searchable keywords in the documents and the query that the user enters. To overcome this a clous information retrieval framework can be used as shown in the diagram below. If a user wants to share some information with others using the cloud, the following protocols can be used:

a. The first protocol is called Document outsourcing protocol and the steps to upload these documents with searchable keywords are as follows:

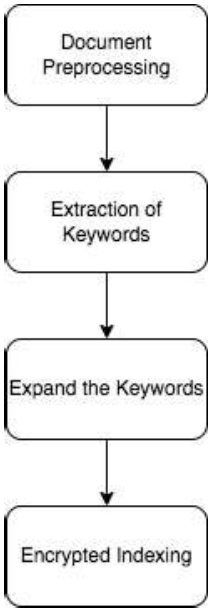

Figure 5: Document outsourcing

1. The first step is to preprocess the documents, in this step the document to be outsourced is collected and is converted to a list of tokens, this process is known as tokenizing the document. The document now is a list of tokens. This step also involves performing some language-based processing on the document which is known as linguistic preprocessing of the document. The policy to preserve privacy can be defined as follows: {Enc(sk,·), Setup(λ), T(·), Dec(sk,·)}
2. The next step involves getting the keywords from the documents di in D. From this set of keywords, a finalized set of top ranked searchable keywords are extracted. K is the set of all keywords ki. The process of extracting these important keywords from the document is dependent on the relevance of these keywords in the documents.
3. These keywords are then expanded with their nearest neighbors in the set K. This is denoted by the letter k+. This expansion is necessary to improve the keyword index. For each keyword a trapdoor is created which is represented as T(ti). To create this trapdoor, we consider the total number of documents that have the keyword ti, the total number of documents having the neighboring word tj and finally the total number of documents to be considered.
4. The encrypted document can then be indexed using a predefined encryption scheme and the documents can be outsourced to the cloud server to store them on the cloud. For this encryption, the following is considered: the trapdoor value of tj, the file id of each document di, finally the various statistics such as frequency, probability and inverse document frequency and other such important factors are taken into consideration. Using this information, the document and data can be outsourced to the cloud database or server for storage.

b. The second protocol is called Document retrieval protocol and the steps to retrieve these documents with the help of user queries are as follows:

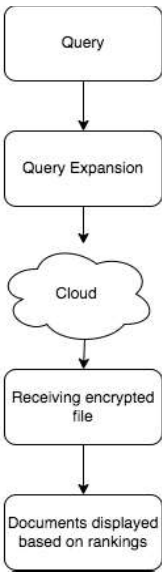

Figure 6: Document Retrieval

1. The first step is initiated by the user when they write the query in the form of the information, they want in the search box. Once the query is written by the users, these queries are translated, and keywords are extracted in the following steps.

2. The query is then expanded with its nearest neighbor in K+. This expanded query is denoted by the symbol Q+.

3. The encrypted files are then retrieved, and the query is compared with the searchable keywords to return a list of all possible files. These files are returned in the form of file IDs and the decryption algorithm follows as the next step.

4. The returned results are decrypted by the client and the file ids of the most relevant files are extracted and they are sent to the cloud service provider and a list of encrypted documents are received, they are decrypted, and the most relevant files are kept in the top results based on different Information Retrieval models.

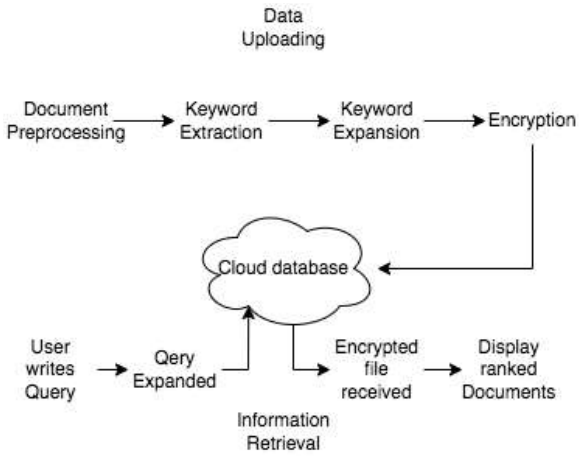

Figure 7: Document Retrieval 2

VIII. SECURITY

Adapting to a cloud-centric infrastructure provides numerous benefits such as quick setup, cost and time savings, but there are new challenges to protect your databases in the cloud. Many of the same dangers that impact cloud technology also harm cloud database systems. However, due to the nature of enormous amounts of potentially sensitive information held in databases, the consequences can be serious if left unchecked. While not exhaustive, these perils provide an indication of the types of risks that network managers may face as firms embrace large-scale cloud database storage systems.

Possible Threats with Cloud Database Security

Data breaches —Data breaches are the most common threat to cloud databases, according to media reports. In a data breach, hackers get access to sensitive data stored in the cloud, such as customer credit card information or mailing addresses, and use it for personal gain. Data breaches are becoming more catastrophic as more information is stored online in a single area, possibly harming millions of customers or employees at once.

Account hijacking - In a hijacking attempt, intruders attempt to acquire access to a user's account through phishing or exploiting security flaws in software. When burglars steal a user's account login credentials, they frequently change the password to lock them out of their accounts. Any files or other information saved in the user's cloud can now be accessed without restriction, including database information that offers data on multiple users at the same time.

Data loss - If an attacker acquires access to sensitive information, one possible conclusion is that the intruder will erase it to inconvenience the owner. If users do not retain up-to-date backups of their files, they may be permanently lost if they are tampered with. When all data are hosted on a single cloud-based server, deletions might spread to all user devices at the same time, resulting in files being lost everywhere at the same time.

In order to address the threats discussed above, security for information retrieval in the cloud can be achieved by the following.

1. Keep Data in Multiple Regions or Zones:

The data is kept on cloud servers, depending on the cloud provider you choose for your databases. Cloud storage systems can provide enterprise-grade storage at a cheap cost. In the same Availability Zone, the replication factor can be modified to protect against data loss due to server component failure, delivering high availability and reliability. The data cannot be recovered under extreme circumstances, such as when the entire availability zone is down. To protect data against such losses, it is best to distribute data across many availability zones, which is known as a multi-AZ configuration. So, while configuring a database infrastructure with DR (Disaster Recovery) and HA (High Availability) in mind, make sure the standby database is replicated in a different availability zone. Thus, if your primary instance /resource is under an outage, then the data can be retrieved from a secondary instance from different availability zone.

2.Safe Data Transit:

You may need to move significant amounts of data from on-premises storage or one cloud to another for an initial migration to the cloud or another replication

requirement. While in transit, the data is vulnerable to malfunctions, outages, or attacks, which can lead to data loss or compliance difficulties. All data in transit must be encrypted, including data volumes, boot disks, snapshot backups, and data archives and backups on cloud systems like Azure Blob and Amazon S3. Configure activity log to audit and log all storage occurrences. The data should be recoverable in the event of a data transmission fault or outage, and data synchronization schedules should be kept track of.

3. Data Access Control for a Database in the Cloud:

By giving complete control over the virtual networking environment, Virtual Private Cloud (VPC) enables you to run your database instances in logically isolated and private cloud environments. In addition to this, you can configure IP addresses, subnets, and network gateways with VPC.

You can configure security groups so that your public-facing apps are accessible from the internet but are protected by another security group that maintains databases and back-end application servers in a private cloud that is unreachable from the internet, avoiding security breaches. With VPC, you may create a hybrid environment in which your cloud and on-premises databases coexist in a single virtual private environment, allowing your data center to access cloud data directly and privately.

4. Cloud Access Control:

Create security groups for related permissions based on the stated policies as a best practice. For example, you could create a group with solely DBA permissions and then assign that group to database administrators while also enforcing a policy that prevents them from dropping the database. All your databases and data storage should be covered by an Identity and Access Management system so that you can keep complete control over who has access to what and identify any missed compliance procedures.

Summary:

Infrastructure and human resources must adapt and ensure data security, compliance, and protection. The key is to have a solid database protection policy in place from the beginning of any setup. To protect databases and information retrieval, you must have a trained storage system, regular compliance checks and database administration resources that are abreast of current vulnerabilities.

Once you understand the challenges of database security, it's a good idea to look at technologies that fulfills the necessary requirements.

IX. ADVANTAGES OF DATABASE ON CLOUD

1. Disaster recovery:

When data is stored on a cloud storage, there will be minimum risk of system failure as the data is stored as well as the data is backed-up on an external device which can be far from the initial location. This will elude the high cost of data retrieval which can be caused by common malfunctioning from the hard drive. The backup process of the data is provided by the cloud providers which aids the requirement for baking up the data on an external device.

2. Access your data anywhere:

When your team is diverse across different parts of the globe, cloud-storage helps to collaborate the work altogether. Employees can login and access the team's work smoothly, despite of working from different locations if the data is stored on the cloud. The restriction of working caused by place or medium will be hailed by cloud storage as the employees are able to work from home or across the globe.

3. Low cost:

All updates and software licenses are not needed to be paid as all of them are included in their timely plans if cloud storage is used. There is no need to invest in costly server infrastructure as the cloud storage providers deliver it to you off-site. There is no need to pay for a dedicated storage professional in-house because it can be outsourced by the cloud storage company.

4. Security:

Data security provided by the professional cloud storage firm is superior to those provided by small businesses. Password protected data storage is provided by professional cloud storage companies. For the transmission of the data, encryption technologies are used here, which ensures that highest security standards are maintained.

5. Scalability:

We can extend the amount of available storage as per our preferences that depends on how much we pay. 'Pay as you go' plans for payment are available as it supports scalable payment. Thus, all business sized, and needs are catered by cloud storage.

6. Mobile access:

Using cloud storage, the efficiency and security is preserved, and the software can be built to be built by globally dispersed teams.

7. Switching:

Different types of databases support various formats and switching which can increase the cost.

8. Multi-tenancy:

A cloud database which is public, is multi-tenant in

nature. It puts forward that an equivalent database is usually employed by multiple customers where a shared model is used, and we just have to pay for what we employ there.

X. DISADVANTAGES OF DATABASE ON CLOUD

1. Vendor lock – in:

There are different cloud service providers all over the world. If a business chose a cloud service provider, then it would be a complex process if you want to move to a different cloud service provider. So, it would be helpful if you pay attention to the cloud service structure if you want to change the provider soon if you have any proprietary applications which rely on the cloud platform.

2. Security and Privacy:

Although security is one of the major benefits of using DBaaS services. However, it is accompanied by concerns about data privacy. It may be safe when it comes to the computing equipment and/or software you're using to access your database instance, but you may never know how your data is handled or managed because you don't have control or access to restrict who has access.

3. Bandwidth problems:

To get ideal performance, plans should be made accordingly by the customers and customers should keep in mind not to pack huge numbers of servers and storage devices within a small set of data centers.

4. System vulnerabilities:

Leading vendors across the world provide strong general security of the cloud infrastructure. There is no guarantee that the entire system will not be at risk. If we generate sensitive data online, then there will always be a risk of data breach. The risk cannot be completely avoided but they can be mitigated. To mitigate the risk, we must follow better practices and a detailed cloud security policy should be communicated to all application developers.

5. Cost concerns:

The costs can be lower for hosting services for database deployments on the initial level than expanding existing servers of your business. Costs may rise, if your business grows, which can turn into higher costs depending on your business needs.

XI. DATABASE CHALLENGES IN THE CLOUD

There are several obstacles that cloud databases must overcome in order to become more viable. Everyone will be allowed to use cloud databases if they can solve these problems.

The following is a list of the challenges:

The organization loses control of their data in some ways because the cloud database is managed and administered by the cloud service provider. They can certainly save their data in a precise manner, but they cannot preserve it in the same manner. For example some cloud service providers do not supply all of the different database types and versions. When the organization has precise IT standards to follow, this could be a problem.

It can be difficult to estimate how much storage you will require when using a cloud service provider for the first time. You may underestimate the amount of storage available, limiting your options. You'll get greater performance if you go above your storage limit, but you'll pay more for the space you don't use.

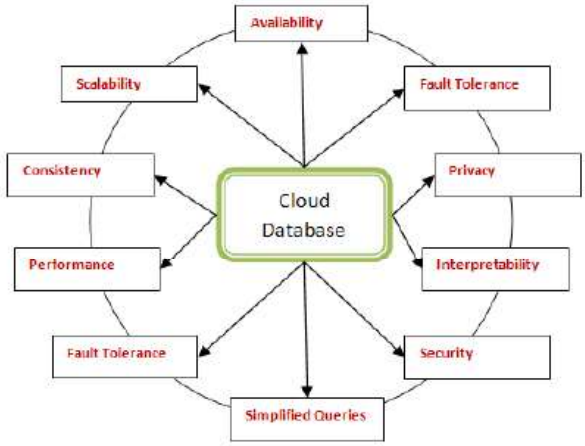

Figure 8: Cloud database overview

While our data is in transit, the first thing we should give is a reliable internet connection with adequate speed. Internet speed has a direct impact on cloud database performance and could be regarded as a performance barrier. We send queries and receive results in cloud databases; while submitting queries does not necessitate a lot of internet access, they will be sent to the database very quickly, the issue is with obtaining results; it takes time to retrieve data, which is directly dependent on the internet connection.

We too have a problem with query workloads and transactional workloads. While we can control transactions, we can't control query workload because it is dependent on the number of queries and we don't know how many users are waiting for queries to be executed.

Users that plan to migrate their databases to the cloud are particularly concerned about privacy. Because cloud databases are accessed and accessible via the internet, they are an attractive and important target for hackers attempting to disrupt the system, even if no sensitive data

is stored there.

XII. CONCLUSION

Database on cloud is a powerful tool that is being used extensively all over the world. The demand of database on cloud will keep on expanding in the next few years because of its ease of use and location independency. As the user base will increase, it will become more important to focus on all aspects of database, such as: performance, security, down time etc. Improvement in information retrieval techniques will in turn better the performance of database on cloud.

XIII. REFERENCES

1. https://ieeexplore.ieee.org/abstract/document/8370878?casa_token=ZigiuR9TRx8AAAAA:9DMOsc7J6_dkzGnDddIKVzQbtpWFZZneeMQXq2itHlSs7HA9Kl_zF2WX-Olkk8iY4k5IneXZ
2. https://cloud.netapp.com/blog/how-to-protect-a-database-in-the-cloud
3. https://www.villanovau.com/resources/cybersecurity/possible-threats-with-cloud-database-security/
4. https://severalnines.com/database-blog/advantages-and-disadvantages-using-dbaas
5. Yuan, J. L. (2015). An availability method for information retrieval in cloud service using work components. 2015 International Conference on Network and Information Systems for Computers. https://doi.org/10.1109/icnisc.2015.138
6. V.R.R.A., & D.R.S.R.P. (2017). A Novel Approach: Reliable and Secure Data Storage and Retrieval in a Cloud.
7. HaiBin, Y., & Ling, Z. (2016). A secure private information retrieval in cloud environment. 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS). https://doi.org/10.1109/incos.2016.85
8. Lu, S., & Abomakhelb, A. I. (2017). Secure cloud storage and quick keyword-based retrieval system. 2017 International Conference on Computing Intelligence and Information System (CIIS). https://doi.org/10.1109/ciis.2017
9. https://www.securestorageservices.co.uk/article/11/pros-and-cons-of-cloud-storage
10. https://www.mongodb.com/cloud-database/benefits
11. https://ieeexplore.ieee.org/document/8272322