# Overview of the Approaches to Managing Distributed Storage and Access to Cloud Data

Anton Kartashov and Larysa Globa
*Institute of Telecommunication Systems, Igor Sikorsky Kyiv Polytechnic Institute, Beresteiskyi Avenue 37, Kyiv, Ukraine*
*kartashov.anton.ukr@gmail.com, lgloba@its.kpi.ua*

Keywords: Cloud Storage, Cloud Access, Multi-Cloud.

Abstract: Currently, the main computing standard for hosting and delivering services over the global Internet has become the widely recognized Cloud Computing technology. Firms and end-users are constantly incorporating Cloud platforms into their technology stack because of their many advantages over traditional computing models. Key advantages of the cloud include nearly unlimited data storage, scalability, cost savings, high availability, and high fault tolerance. However, while cloud solutions offer tremendous opportunities and services to the industry, the landscape of cloud computing research is changing for several reasons, such as the emergence of data-intensive applications, multi-cloud deployment models, and more stringent non-functional requirements for cloud services. The proliferation of cloud computing providers on a global scale has surged significantly. Consequently, selecting an optimal provider that aligns with customer needs and defining appropriate evaluation criteria have become complex endeavors. Current trends reveal an increasing body of research focusing on appraising and ranking cloud-based services based on their ability to cater to user requirements. This paper synthesizes past investigations and recent studies, navigating the intricate landscape of modern cloud architectures. It delves into the dynamics of multiple providers, leveraging the advantages of decentralization. Additionally, it sheds light on the complexities of managing data storage and access within the dynamic context of multi-cloud environments, highlighting ongoing challenges. This comprehensive exploration culminates in a comparative summary of the existing approaches and a proposal for the forthcoming research, that includes a comprehensive set of criteria for the for multi-cloud data storage that involves a wide range of factors that can impact data placement, management, and retrieval across multiple cloud providers, contributing to the ongoing advancement of the field.

## 1 INTRODUCTION

As per the National Institute of Standards and Technology (NIST) interpretation, cloud computing can be defined as "a model of providing users with access to a shared pool of configurable computing resources, including networks, servers, storage, applications and services that can be rapidly provisioned (adapted for new services) with minimal management or interaction with service providers [1, 2, 3, 4]. In the same way, several sources state that Cloud Computing should be regarded solely as a way of remote access and management of computing resources [5]. The whole range of technologies that include Cloud Computing - these are already existing technologies, but as a way of combining them is a novelty [6]. To concentrate on the basic elements and functions of business, without worrying about maintenance or computer infrastructure and hiring specialized personnel, companies are increasingly introducing this technology into their business processes, which is essentially becoming another public service like electricity or the Internet [7]. It is necessary to highlight the main advantage of cloud computing, which is that due to a much more profitable use of computing resources in the cloud and their easy scalability, the results can also be obtained much faster. A good example of such an approach is comparing the cost of using one virtual machine for a long period of time (1000 hours) with the cost of using a large number of virtual machines (1000 VMs) for one hour. As a result, concurrently running machines can complete the task much faster, even though the monetary cost of these options may be identical.

In the following paper, we conduct the literature overview of existing approaches to managing distributed storage and access to Cloud data, review their pros and cons, and establish a basis that allows developing a set of criteria and recommendations to distribute data between the multiple cloud providers to achieve the best result according to the defined criteria. The goals of re-examining the existing literature encompass three main aspects: firstly, to grasp the historical role of evaluation methods in quantifying the performance of cloud services. Secondly, to demonstrate the utility of applying specific approaches in the identification of suitable cloud vendors capable of meeting user requirements. Lastly, to accentuate unresolved difficulties and engage in discourse regarding future research directions.

In Section 2 we provide an overview of the main features of cloud computing and existing modern cloud architectures related particularly to storage and access. Next, Section 3 is analysing the challenges of heterogeneous multi-cloud systems from several perspectives. Section 4 provides an overview of existing research papers, and we discuss the highlights of each paper that we use for the comparison later. Section 5 discusses the lessons learned from the literature and identifies challenges and future research directions. We also define cloud evaluation criteria conducted from the previous research.

## 2 MODERN CLOUD ARCHITECTURES

Now being a prominent field of research in computer science, many scholars have proposed definitions and explanations of different aspects of cloud computing. Let us consider the main features of cloud computing [8]:

- It is available to users remotely at any time without the need for any additional direct hardware maintenance.
- Many available interfaces for use and access including cell phones, laptops, and desktops.
- Cloud providers focus on multi-tenant service delivery models, where the same physical resources are used together to provide services to multiple users at the same time.
- Flexibility and elasticity of service usage following end users' requirements.
- Availability of transparent and easily accessible service accounting.

Absence of Capex (Capital Expenditure) or simply upfront hardware investments and thus lower operating costs, the possibility of easy access and reduced network and hardware support tasks, allows small and medium-sized companies to focus on the core elements of the business and makes the choice of cloud computing a priority strategy for the near future [9].

Three models of cloud computing service delivery dominate the industry today: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). In SaaS, customers can use the software on the provider's cloud system, usually over the Web [1, 4]. In PaaS, the provider allows the consumer to use the cloud network as a platform for their own developed or acquired software. In IaaS, virtual storage and machines are provided to the customer for greater control over the environment and the deployment of their applications [5].

Beyond basic considerations and specific use cases, the evolving nature of cloud solutions speaks to adoption at scale. Let us consider data storage: as data volumes increase exponentially, file-based storage now faces a challenge from new solutions, such as object-based frameworks, which empower the transfer of massive data sets at speed. Storage architecture is also evolving as options, such as hyper-converged and disaggregated, composable solutions emerge [10]. While hyper-converged offerings leverage software-defined clustering to provision storage resources on demand, disaggregated, composable systems decouple storage, network, and compute processes from physical hardware to create a shared pool of resources that can be individually assigned or used together. Open-source initiatives, meanwhile, are changing the fundamental nature of the cloud. From open-source software developments that allow organizations to customize key functions and frameworks to open-source architecture efforts-such as storage devices that are programmable and modifiable – the shift away from proprietary provisioning will impact long-term cloud decision-making. Organizations also must account for the movement of data and compute tasks away from central cloud services to edge infrastructure. Informed by increasingly complex processes handled by connected, intelligent devices at the point of origin-rather than being shunted to public or private clouds for analysis-edge computing modifies the value proposition of both public and private cloud services, as well as their impact on operational outcomes.

To better understand the motivation for multi-cloud, we segment the technical platform architecture into common scenarios for data distribution, by taking a suggested approach from [11] and further elaborating on it in terms of data management. The key observations of the multi-cloud models could be described within the following categories. On the figures, each type of shape - squares, triangles, diamonds, and circles - serves as a symbol or visual identifier for the arbitrary data or objects for our case (can also represent applications or workloads).

Squares: within the Figures, squares are used as visual representations of arbitrary data objects or files. The use of squares might be arbitrary, indicating that the choice of this shape is not inherently tied to a specific meaning. In this context, squares could represent distinct parts of data objects or files, possibly signifying that different elements within a dataset share certain commonalities.

Triangles, Diamonds, and Circles: similar to squares, they serve as visual symbols for elements within the Figures. These triangles may represent various aspects of data, applications, or workloads. Like squares, they could be used to depict parts of a whole, where similar elements are represented using the same shape. This repetition suggests that these elements have shared attributes or characteristics.

The central idea here is that the choice of shape is flexible and not inherently tied to specific meanings. Instead, the emphasis is on visually distinguishing different elements within the data objects (applications or workloads). The use of the same shape, whether square, triangle, diamond, or circle, suggests that these elements have similarities or are parts of a larger entity, such as a data object or file. The figures aim to convey complex information or relationships through the visual representation of these shapes, allowing for clearer visualization and understanding of the data.

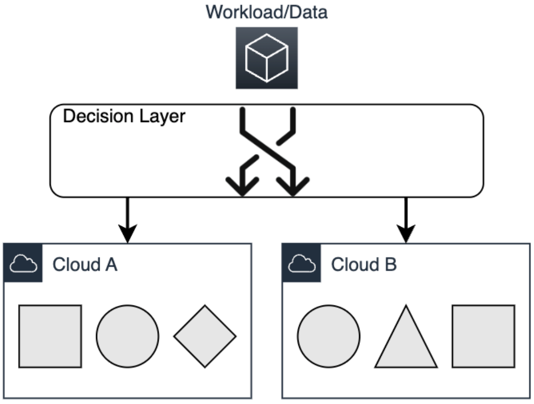

Figure 1: Random Data distribution scenario.

Figure 1 depicts a random scenario of a process for distributing data to a cloud provider. The decision layer, in this context, can be defined as an abstraction logic responsible for the selection of one cloud provider over another for data storage or deploying applications. The absence of robust governance and the influence of vendors are the primary factors leading to arbitrary decisions in choosing a cloud environment. Consequently, data is dispersed across multiple clouds, with some being processed in the first cloud, others in the second cloud, and additional portions in the third or fourth cloud. The lack of a well-defined process or criteria for determining where and what data should be stored contributes to this dispersed and less-than-optimal data distribution approach.

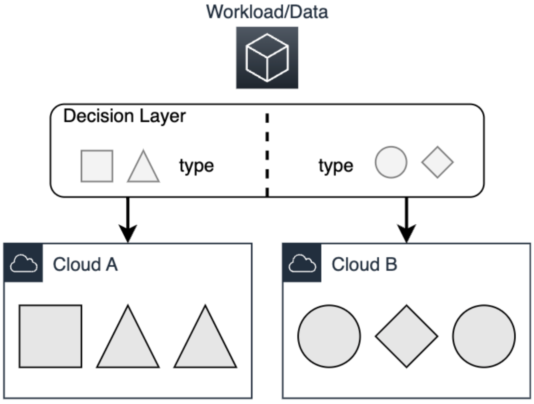

Figure 2: Segmented Data distribution scenario.

Figure 2 depicts segmenting data scenarios across different clouds, which is also common: specific types of data are being deployed to specific clouds. The following are the possible factors to decide on segmentation:

- Data Lifecycle (legacy or modern).
- Type of data (confidential vs. public).
- Type of product (compute vs. data analytics vs. collaboration software).

While considering this approach, we need to understand the traffic streams between clouds due to the possible excessive egress charges in case one half of the data ends up in one cloud and the other half in another one.

Often, the first two approaches might not be considered as true multi-cloud. The usual goal may be the ability to deploy data freely across cloud providers, thus minimizing vendor lock-in, usually by means of adding abstraction layers [12].

Figure 3 describes the choice scenario, which is common for large organizations' shared IT providers because they are expected to support a wide range of business units and their respective IT preferences. Often, such a setup involves a central commercial relationship and a common framework to create instances on the cloud provider of choice but with corporate governance and constraints tacked on.

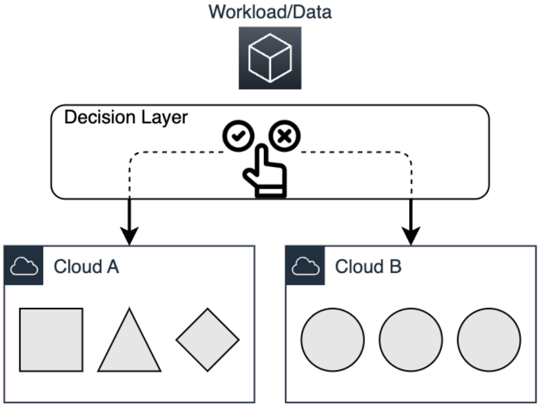

Figure 3: Choice Data distribution scenario.

The advantage of this setup is that projects are free to use proprietary cloud services, such as managed databases (depending on their preferred trade-off between avoiding lock-in and minimizing operational overhead). Hence, this setup makes a good initial step for "true" multi-cloud.

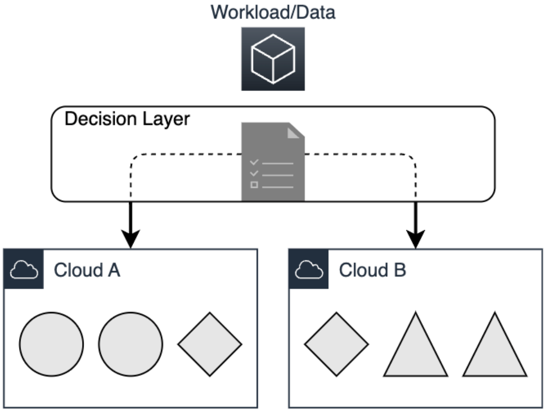

Figure 4: Parallel Data distribution scenario.

In Figure 4 we see the parallel data distribution approach, which corresponds to the notion of High Availability in the cloud, which is one of the main ensuring points for customers looking to store critical data across multiple clouds.

Being able to deploy the same application into multiple clouds requires a certain set of decoupling from the cloud provider's proprietary features. To achieve this, the following has to be considered [13]:

- Managing cloud-specific functions like identity management, deployment automation, or monitoring separately from the application in a cloud-specific manner.
- Using open-source components as much as possible – they will generally run on any cloud. However, it may reduce the ability to take advantage of other fully managed services, such as data stores or monitoring.
- Utilize a multi-cloud abstraction framework to perform a one-time development and be able to deploy to any cloud.
- Maintain two branches for those components of an application that are cloud provider-specific and wrap them behind a common interface. For example, there can be a common interface for block data storage.

The key aspect to watch out for is complexity, which can easily undo the anticipated uptime gain. Additional layers of abstraction and more tooling also increase the chance of a misconfiguration.

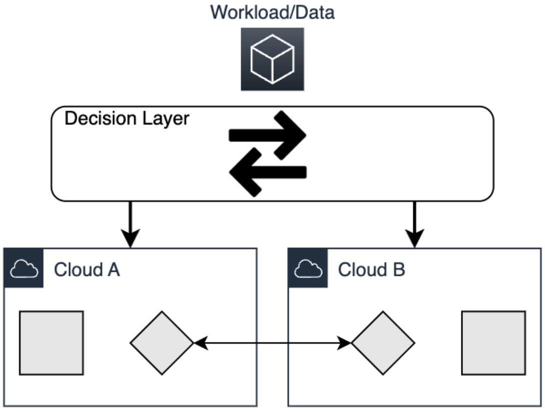

Figure 5: Portable Data distribution scenario.

Figure 5 depicts the best-case scenario for multi-cloud: free portability across cloud vendors. To the main benefits we assign avoiding vendor lock-in, and applications' placement based on resource needs. For example, day-to-day predictive and stable operations may be run in one cloud and burst excessive traffic into another.

The mechanism to enable this capability requires high levels of automation and abstraction away from cloud services. While for parallel deployments one could get away with a semi-manual setup or deployment process, full portability requires being able to shift the data at any time, so everything better be fully automated.

However, the cost of such a mechanism comes in the form of lock-in a specific vendor, product, and architecture plus a possible requirement for container distribution and orchestration [14]. Furthermore, these abstractions typically do not address data management, including the challenge of maintaining data synchronization. Even if data synchronization is effectively handled, concerns about egress data costs may arise.

The following Table 1 summarizes the choices, the main drivers, and the side-effects to be aware of when choosing a particular scenario.

Table 1: Summary of the Multi-cloud distribution scenarios.

| Common Challenges | Key Points to address in the research |
|---|---|
| Vendor Lock-In: In all scenarios, avoiding vendor lock-in is a primary concern. Organizations seek the flexibility to place their data across different cloud providers to minimize dependency on a single vendor. | Abstraction Logic: The need for a well-defined decision layer or abstraction logic is evident in all scenarios. This layer should facilitate the selection of cloud providers for data storage based on a comprehensive set of decision criteria. Research and development efforts should focus on creating a robust and vendor-agnostic decision layer. |
| Data Dispersal: A common challenge across the scenarios is the arbitrary distribution of data. Without well-defined processes and criteria, data ends up being dispersed across multiple clouds, leading to inefficiencies in data management and increased complexity. | Data Segmentation: Research should address the challenges of data segmentation, considering factors such as data lifecycle, data type, and the type of product. Efficient segmentation can help optimize data placement and reduce egress costs. |
| Egress Data Costs: The scenarios emphasize the potential for excessive egress charges when data is distributed across multiple clouds. Managing and optimizing data egress costs is a significant challenge, especially when data is scattered across different cloud environments. | High Availability: For scenarios involving parallel data distribution, research should explore methods for achieving high availability across multiple clouds. This includes managing cloud-specific functions, utilizing open-source components, and implementing multi-cloud abstraction frameworks while minimizing complexity. |
| Complexity: As organizations aim for multi-cloud solutions, they often encounter increased complexity. The need for additional layers of abstraction and tooling can lead to misconfigurations and operational challenges, potentially undermining the anticipated benefits of multi-cloud setups. | Free Portability: The best-case scenario involves free portability across cloud vendors. Research and development should focus on creating mechanisms for high levels of automation and abstraction, enabling seamless data migration and application deployment while addressing potential vendor lock-in and architectural considerations. |

Next, we continue elaborating on the challenges of multi-cloud computing technology.

# 3 CHALLENGES OF THE MULTI-CLOUD SYSTEMS

Although cloud computing provides more flexible usage models for businesses by eliminating the need for dedicated allocation and allowing small businesses to scale their computing needs to meet business requirements, there are some issues and challenges facing cloud systems that may be the subject of future research [15]. Even though cloud computing is still in its continuous stage of development, its widespread adoption has allowed many researchers and companies to begin to assess the real and further potential challenges facing this technology [16].

Data is no longer just created in local data centers. The amount of data being created in the cloud and from emerging technologies - such as IoT and edge computing - continues to grow. Yet, according to various reports, companies are only capturing around 55% of the data potentially available through operations.

Capturing all available data, however, would overburden existing IT infrastructure and drive up costs. This is one of the many reasons why companies need to rethink their data management. If data is identified and classified at the beginning of its lifecycle, for example, this enables faster data cleansing, which in turn leads to lower costs.

Data proliferation results in silos that complicate the work of data scientists and analysts who transform this data into insights for decision-makers. Corporate culture can lead to further silos. Competing groups pursue their own goals and therefore want the ability to control and keep certain data to themselves.

To make data accessible from silos, business owners must overcome both technological and human barriers. Automated tools, such as Unified Policy Mechanisms, can solve the technological aspect. Global data management and global standards can help bring teams in line.

Data security is always a major concern for both IT and business management. Multi-cloud security comes with its own set of problems, such as inconsistent visibility across different clouds and a lack of coordination between different security components [17].

Vulnerable environments carry the risk of data breaches. The consequences of this range from financial losses and fines to reputational damage and data breaches. But the importance of security goes beyond this. Strong security is important to unlock the full value of data, ensuring unhindered access to data and data integrity.

Successful data management requires a holistic view of data storage, both in local and cloud architectures. This does not simply mean data democratization, but storage unification and data management through a centralized view, no matter where the data is stored.

The widespread simultaneous use of different storage technologies means a high space requirement, which can become a problem for companies. Moreover, there is often a lack of a coherent data storage strategy. Figure 6 depicts a typical approach to addressing the data definition question from 3 perspectives.

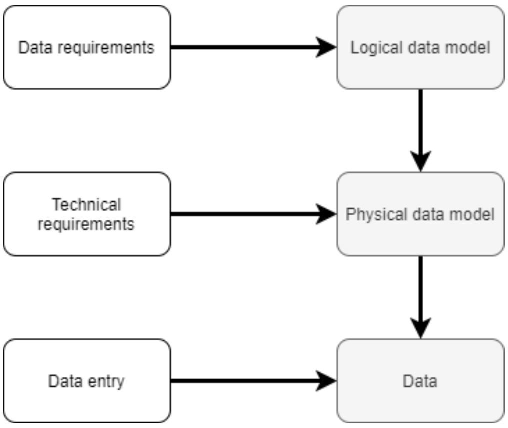

Figure 6: Holistic approach to data definition.

Often the data collected is stored in large repositories where it gathers dust. It is forgotten and businesses are missing out on a real treasure trove of information as a result. Smart data collection starts with understanding the business goals and insights that companies want to gain from their data. These goals make it clear what kind of data should be captured.

Organizing and sifting through massive amounts of data is also part of the data usability problem [18]. Companies must address factors such as complexity, overlapping tools, data integration, and more that impact the ability to obtain information valuable to the business.

A further challenge of data management in the cloud is the issue of steps and strategies for addressing possible catastrophic data loss. If, for example, a cloud provider goes insolvent or loses a data center due to a natural disaster, data may become unrecoverable. This would create a huge issue for large organizations that depend on data storage [18]. Compatibility or "the ability of different systems and organizations to work together (interoperate)" is the second most frequently mentioned cloud computing concern, after data security and trust issues. This is critical because it allows users to avoid vendor lock-in, in which the user becomes dependent on the Cloud Solution Provider (CSP) because they cannot migrate their services to other cloud platforms or clouds. The lack of standardization means that it can be difficult not only to transfer data between cloud systems but also to use them as one layer of data storage [19].

Another often-mentioned challenge is billing: pricing models are significantly different across providers, making it close to impossible to dynamically estimate the cost of the same task running on one cloud or another. To make the most of multi-cloud movement, enterprises need a three-step strategy:

1) Identify data architectures: before making the move to multi-cloud, it is critical for companies to identify current data architectures – including where data is stored, how it is used, and where processes could be improved – to define desired operational outcomes. With specific goals in mind, enterprises are better prepared to make multi-cloud decisions that deliver line-of-business benefits.
2) Assemble cloud resources: purpose-built clouds are now commonplace across both public and private deployments thanks to emerging solutions, such as customizable, composable infrastructure. As a result, it’s critical for companies to assemble multi-cloud network, compute, memory, and storage functions that are capable of addressing current needs and meeting future challenges.
3) Enable seamless orchestration: multi-cloud deployments only deliver on performance and process mandates when cloud elements operate seamlessly. Realizing Return of Investment (ROI) on this move to multi-cloud means building on defined architectures and assembled resources with robust data orchestration and management solutions capable of reducing friction across key cloud functions.

One of the pivotal challenges that cloud computing users face is selecting the most efficient and cost-effective data storage and access methods while circumventing vendor lock-in. Nonetheless, the extant criteria for optimizing data distribution in multi-cloud environments only encompass a subset of customer requirements. Despite prior efforts to delineate these criteria, the rapidly evolving cloud computing landscape necessitates a comprehensive reevaluation of approaches [20].

# 4 LITERATURE OVERVIEW

This section investigates the different solutions related to data storage and retrieval in a multi-cloud environment.

Yang and Ren [21] proposed a Virtual Framework for Cloud Storage Services (VCSS) to integrate diverse common open Cloud Storage Services to be a uniform virtual storage resources pool. Key concepts and technologies include a service metadata model with rich NFP, a service repository serving as a virtual storage resource pool to enable resources physically distributed stored, and logically centralized managed, and a service schedule model serving to estimate and select an appropriate service to store and replicate the file.

Megouache et al. [22] presented a novel module for solving security problems in multi-cloud platforms. This method contains 3 phases, the initial stage, is for proposing a private virtual network to secure the data transmission. Next, they utilized an authentication technique depending upon data encryption, for protecting the user's identity and information, and lastly, they created a method for knowing the reliability of data allocated on the several clouds of the scheme. The module attains identity verification and the capability to interoperate among processes run on distinct cloud providers. A data integrity method would also be established.

Colombo [23] encrypted the file as a whole and gave a choice to the user to select the cloud provider among the hybrid cloud. Irrespective of the sensitivity of data stored in the cloud, the entire file is encrypted. Also, to provide fine-grained access to data, multiple copies of data are kept in the cloud. High retrieval overhead is a problem in this solution.

Subramanian and Leo [24] aim to provide a framework that decreases malicious insiders and file risks and enhances data sharing security in multi-cloud storage services. This method would provide a secured platform where the data owner could retrieve and store data from the multi-cloud platform with no merging file conflict and prevent insider attacks from obtaining useful data. Research indicates that the recommended module is appropriate for making decision procedures for the data owners in an optimum acceptance of multi-cloud storage service to share their data safely.

Cao [25] implemented and designed a multi-cloud architecture to build Open Stack based environment for medicinal IoT, denoted by Tri-SFRS. For implementing this technique, they integrate various methods for attaining this decrease in efforts, comprising lower overhead native testing architecture, multi-cloud cascading framework, snapshot volume cascaded operation for b-ultrasonic data, and medicinal data storage backup method. Tri-SFRS can concurrently allow asset management. Tri-SFRS was implemented as a native element in the Open Stack environment, and it determines the degree of native Open Stack multi-cloud environment management using this presented cascading architecture.

Celesti [26] deliberated to improve the whole system regarding retrieval and data storage via validating and testing an MCS scheme consisting of 3 main Cloud Storage suppliers: Copy, Dropbox, and Google Drive. Research has shown that the selection of a Cloud storage provider for storing files according to data transfer efficiency depends on file chunk size.

Samundiswary [27] proposed an object storage architecture for unstructured data. Metadata about the object is used for searching in the object storage architecture. Though the object storage architecture proposed in this work assists in rapid retrieval, it is not secure against attacks.

Libardi [28] proposed a Multi-cloud Storage Selection Framework to automatically select a storage dispersal strategy. MSSF formalizes the selection process using a knapsack optimization problem using integer linear programming along with a rule-based system to select a multi-cloud storage strategy that fits the user's needs and requires only simple inputs from the user. Its main limitation is the requirement of user input parameter selection for each file upload.

Li et al. [29] proposed a privacy-preserving STorage and REtrieval (STRE) method that guarantees privacy and security however also offers consistency assurances for the outsourced searchable encrypted data. The STRE method allows the cloud user to distribute and search its encryption data over many independent clouds handled by distinct CSPs, and strong while a specific amount of CSP crashes. In addition to reliability, STRE provides the advantage of partly hidden search pattern. Though this scheme is secure and privacy-preserving, overhead of Shamir splitting and reassemble is very high for large data volume. With only a small percentage of sensitive information in the unstructured data, this overhead is too high.

Janviriya [30] approach consists of Multiple Cloud Storage Integration Systems based on RAID (Redundant array of independent disks) 0 stripping technology where the proposed application creates a single cloud storage combined out of multiple cloud storage accounts and also decreases the total time of file uploading and downloading to and from the cloud. Also, because of concerns about security on the cloud, the research suggests the security level enhancement of its cloud-stored files. The results from experiments using the application show an improvement in performance, storage capacity, and security.

Zhao et al. [31] proposed a middleware that enables any end-user application to automatically and securely store files in multiple cloud storage accounts. middleware that enables an application to use the cloud storage services securely and efficiently, without any code modification or recompilation. The proposed solution securely saves the user data to different cloud storage services to significantly enhance the data security, without the need to save it to the local disk. The solution is implemented as a shared library on Linux and supports applications written in different languages, supports various popular cloud storage services, and supports common user authentication methods used by those services.

Rios [32] proposed a new DevOps architecture intended to assist Cloud consumers in deploying, designing, and functioning (multi) Cloud systems that contain the required security and privacy controls to ensure law enforcement authorities, transparency for end users, and third-party in-service provisions. The architecture is based on the risk-driven requirement at the implementation time of security and privacy levels objective in the continuous enforcement and service level agreement and observing at run-time.

Tchernykh et al. [33] presented a multi-cloud-based storage framework named WA-RRNS which integrates threshold secret allocation redundant and weight access system remains number scheme with many failure recognition or recovery mechanism and homomorphic cipher. For optimum trade-offs between security and efficiency, WA-RRNS utilizes variables for adjusting data loss probability, redundancy and encryption/decryption speed. Investigational and Theoretical analyses with actual data displays that this method gives a secure manner for mitigating the uncertainty of untrusted and not consistent cloud storage.

Pravin [34] addressed the privacy and security risks of data in the multi-cloud storage. The proposed solution is based on the cryptographic technique with a dynamic file-slicing method for securing data in a multi-cloud environment. The stored data is fragmented into many slices. The number of slices is defined by the data owner. The sliced files are encrypted with 3 DES (data encryption standard) and elliptical curve cryptography (ECC) algorithm. The performance of the proposed technique was evaluated using latency time and the results revealed that the proposed technique outperformed the other methods discussed in the article.

Esposito [35] proposed three different methods for selecting the best cloud service set in order to maximize the quality of service and minimize cost. The three methods utilized are based on Fuzzy Logic, Theory of Evidence and Game Theory. The suggested solution mainly focuses on availability and cost.

Le [36] proposed a data partitioning model based on fragmentation, secret sharing, and encryption for medical data storage in clouds. Patient-centric information is represented as an Entity association and relation model. Access control to the sensitive information in the patient-centric model is enforced using cryptographic algorithms.

Valapula [37] proposed a secure and efficient data partitioning scheme for a hybrid cloud is proposed. The scheme is adaptive to unstructured text documents and achieves a fine balance between multiple objectives of maximizing security, increasing public cloud utilization, and minimizing the degradation of retrieval efficiency. The performance of the solution was tested against the Enron email dataset.

Vernik [38] presented an on-boarding federation mechanism for adding a special layer on cloud storage services allowing them to import data from other services. This was achieved without dependency on special functions from the other cloud vendors. The proposed solution suggested a design of a generic, modular onboarding architecture designed for content-centric data. This approach requires certain adaptability level from the cloud service provider becoming hard to implement. However, no fault tolerance mechanism is proposed.

Malensek [39] proposed a private and public cloud federation method in order to improve queries throughput in large datasets. Their distributed storage framework autonomously tuned in-memory data structures and query parameters to ensure efficient retrievals and minimize resource consumption. To avoid processing hotspots, they predicted changes in incoming traffic and federated their query resolution structures to the public cloud for processing. The efficacy of the suggested frameworks was demonstrated on a real-world, petabyte dataset. In addition, several approaches referred to Service Selection and Data Distribution were studied.

Sukmana [40] proposed a unified cloud access control model that provides the abstraction of Cloud Service Providers (CSP) services for centralized and automated cloud resource and access control management in multiple CSPs. Their proposal offered role-based access control for Cloud Storage Brocker (CSB) stakeholders to access cloud resources by assigning necessary privileges and access control lists for cloud resources and CSB stakeholders, respectively, following the privilege separation concept and least privilege principle.

Chang [41] suggested a mathematical formulation of the cloud service provider selection problem in which both the object functions and cost measurements are defined. The algorithms that are selected among cloud storage providers to maximize the data survival probability or the amount of surviving data are subject to a fixed budget, and a series of experiments demonstrated that the proposed algorithms were efficient enough to find optimal solutions in a reasonable amount of time, using price and fail probability taken from real cloud providers.

## 5 CRITERIA FOR MILTI-CLOUD

Based on the conducted literature review and modern cloud computing standards for storage and access we define a complex set of criteria for multi-cloud data storage that involves considering a wide range of factors that can impact data placement, management, and retrieval across multiple cloud providers. Table 2 depicts a comprehensive set of criteria to consider in the future research.

Future work should incorporate multi-cloud paradigms and involve an intricate exploration of each criterion, further refining their definitions and relevance within the ever-evolving cloud computing landscape. Initially, based on the suggested set of criteria we define a comprehensive framework for multi-cloud data storage strategies, encompassing technical, operational, and business considerations. Subsequently, we establish an ontology, serving as an abstract logical layer for the distribution of data based on the predefined set of criteria.

Following this, we devise an algorithm for the optimal allocation and storage of data, leveraging the ontology model created in the previous step. This algorithm ensures consideration of multiple criteria in alignment with the established ontology.

The final step involves developing a method for organizing optimal data access, taking into account the distributed and stored data following the predefined criteria. Concurrently, we introduce a pseudo-query language tailored for data retrieval, capable of describing algorithms for data retrieval based on selected operations. These operations may encompass full or partial data retrieval and filtering by date or data type, among other criteria.

Table 2: Complex set of criteria for multi-cloud data storage distribution.

| # | Criteria Category | Specific Criteria | Possible Measurement Metric |
|---|---|---|---|
| 1 | Data Accessibility Criteria | Latency Requirements | Milliseconds (ms) |
| 2 | | Redundancy and Availability | Availability Percentage (%) |
| 3 | | Data Consistency | Data Consistency Index |
| 4 | | Data Encryption | Encryption Strength (e.g., AES-256) |
| 5 | Cost and Resource Utilization Criteria | Cost Efficiency | Cost per GB/month ($) |
| 6 | | Resource Allocation | Resource Utilization (%) |
| 7 | | Data Lifecycle Management | Percentage of Archived Data (%) |
| 8 | Data Type and Format Criteria | Data Classification | Data Classification Score |
| 9 | | Data Format | Data Format Compatibility |
| 10 | Compliance and Security Criteria | Regulatory Compliance | Compliance Audit Score |
| 11 | | Data Ownership | Data Ownership Policy Adherence |
| 12 | | Security Protocols | Security Protocol Strength |
| 13 | Scalability and Performance | Scalability | Scalability Factor |
| 14 | | Performance Metrics | Throughput (requests/second) |
| 15 | Data Migration and Interoperability Criteria | Data Portability | Data Portability Index |
| 16 | | Interoperability | Interoperability Score |
| 17 | Vendor Lock-In and Vendor Criteria | Vendor Lock-In Mitigation | Lock-In Reduction Score |
| 18 | | Vendor Reputation | Vendor Reputation Rating |
| 19 | Disaster Recovery and Backup | Recovery Time Objective (RTO) | Recovery Time Objective (RTO, hours) |
| 20 | | Recovery Point Objective (RPO) | Recovery Point Objective (RPO, hours) |
| 21 | | Data Backup Frequency | Frequency (e.g., per day, per week) |
| 22 | | Backup Storage Redundancy | Redundancy Level (e.g., dual-site) |
| 23 | Monitoring and Reporting | Monitoring Tools | Tool Effectiveness (e.g., Score) |
| 24 | | Reporting | Reporting Accuracy (e.g., Percentage) |
| 25 | Sustainability | Environmental Impact | Carbon Emission Reduction (%) |
| 26 | | Energy Efficiency | Energy Usage (kWh) |
| 27 | | Resource Sustainability | Resource Conservation Index |

These research goals are strategically designed to systematically address the challenges of information environment representation, efficient data allocation and storage, and streamlined data access in a structured and methodical manner.

## 6 CONCLUSIONS

In summary, cloud computing remains a rapidly evolving technology with diverse applications across various industries, particularly in remote computing and storage. However, it is important to acknowledge that there are still many unresolved issues and untapped possibilities in this field.

In this paper, we have revisited the existing challenges associated with multi-cloud solutions and conducted a comprehensive review of solutions proposed in scientific literature. To address these challenges and expand the scope of potential solutions, we have introduced a comprehensive set of criteria that will serve as the foundation for future research. Additionally, we have outlined the primary directions for our upcoming research and highlighted the expected outcomes.

Our overarching goal is to address the complex challenges of multi-cloud data distribution, ultimately enhancing performance, minimizing costs, and ensuring data accessibility while catering to the diverse demands of customers.

## REFERENCES

[1] J. Hong, T. Dreibholz, J. Schenkel, and J. Hu, "An Overview of Multi-cloud Computing," In: Proceedings of the International Conference on Cloud Computing, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-15035-8_103.

[2] "Products in Cloud Infrastructure and Platform Services." Gartner, [Online]. Available: https://www.gartner.com.

[3] J. Alonso, L. Orue-Echevarria, V. Casola, et al., "Understanding the Challenges and Novel Architectural Models of Multi-cloud Native Applications – A Systematic Literature Review," Journal of Cloud Computing, vol. 12, no. 6, 2023. [Online]. Available: https://doi.org/10.1186/s13677-022-00367-6.

[4] M. Peter and G. Tim, "The NIST Definition of Cloud Computing," National Institute of Standards and Technology, Tech. Rep., 2011.

[5] T.B. Winans and J.S. Brown, "Cloud Computing: A Collection of Working Papers," Deloitte LLC, 2009.

[6] Z. Qi, "Cloud Computing: State-of-the-Art and Research Challenges," Journal of Internet Services and Applications, vol. 1, no. 1, 2010.

[7] F. Armando, G. Rean, J. Anthony, K. Randy, K. Andrew, et al., "Above the Clouds: A Berkeley View of Cloud Computing," Tech. Rep. UCB/EECS2009-28, EECS Department, University of California, Berkeley, 2009.

[8] R. Buyya, S.N. Srirama, G. Casale, R. Calheiros, Y. Simmhan, et al., "A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade," ACM Computing Surveys, vol. 51, no. 5, 2018.

[9] N. Antonopoulos and L. Gillam, "Cloud Computing: Principles, Systems, and Applications," London: Springer, 2010.

[10] D. Slamanig and C. Hanser, "On Cloud Storage and the Cloud of Clouds Approach," In: Proceedings of the International Conference on Internet Technology and Secured Transactions, 2012. [Online]. Available: https://doi.org/10.1109/ICITST.2012.6470979.

[11] G. Hohpe, "Multi Cloud Architecture: Decisions and Options," July 2019.

[12] Y. Ghanam, J. Ferreira, and F. Maurer, "Emerging Issues and Challenges in Cloud Computing – A Hybrid Approach," Journal of Software Engineering and Applications, vol. 5, no. 11A, 2012.

[13] D. Petcu, "Portability and Interoperability between Clouds: Challenges and Case Study," In: Towards a Service-Based Internet. Springer, Berlin/Heidelberg, 2011.

[14] T. Dreibholz, "Big Data Applications on Multi-Clouds: An Introduction to the MELODIC Project," Keynote Talk at Hainan University, College of Information Science and Technology, 2017.

[15] Y. Elkhatib, "Mapping Cross-Cloud Systems: Challenges and Opportunities," In: Proceedings of the 8th USENIX Conference on Hot Topics in Cloud Computing. Berkeley/United States, 2016.

[16] O. Tomarchio, D. Calcaterra, and G.D. Modica, "Cloud Resource Orchestration in the Multi-Cloud Landscape: A Systematic Review of Existing Frameworks," Journal of Cloud Computing, vol. 9, no. 49, 2020. [Online]. Available: https://doi.org/10.1186/s13677-020-00194-7.

[17] N. Vurukonda and B.T. Rao, "A Study on Data Storage Security Issues in Cloud Computing," Procedia Computer Science, vol. 92, 2016.

[18] T.G. Papaioannou, N. Bonvin, and K. Aberer, "Scalia: An Adaptive Scheme for Efficient Multi-Cloud Storage," In: Proceedings of the International Conference on High-Performance Computing, Networking, Storage, and Analysis, 2012. [Online]. Available: https://doi.org/10.1109/SC.2012.20.

[19] D. Petcu, "Consuming Resources and Services from Multiple Clouds," Journal of Grid Computing, vol. 12, pp. 321–345, 2014. [Online]. Available: https://doi.org/10.1007/s10723-013-9290-3.

[20] S. Bharany, S. Sharma, O.I. Khalaf, et al., "A Systematic Survey on Energy-Efficient Techniques in Sustainable Cloud Computing," Sustainability, vol. 14, no. 10, 2022. [Online]. Available: https://doi.org/10.3390/su14105712.

[21] D. Yang and C. Ren, "VCSS: An Integration Framework for Open Cloud Storage Services," In: Proceedings of the IEEE World Congress on Services, 2014. [Online]. Available: https://doi.org/10.1109/SERVICES.2014.36.

[22] L. Megouache, A. Zitouni, and M. Djoudi, "Ensuring User Authentication and Data Integrity in Multi-Cloud Environment," Human-centric Computing and Information Sciences, vol. 10, 2020. [Online]. Available: https://doi.org/10.1186/s13673-020-00254-7.

[23] M. Colombo, R. Asal, Q.H. Hieu, et al., "Data Protection as a Service in the Multicloud Environment," In: Proceedings of the IEEE 12th International Conference on Cloud Computing (CLOUD), 2019.

[24] K. Subramanian and J. Leo, "Enhanced Security for Data Sharing in Multi-Cloud Storage (SDSMC)," International Journal of Advanced Computer Science and Applications, vol. 8, 2017.

[25] R. Cao, Z. Tang, C. Liu, and B. Veeravalli, "A Scalable Multicloud Storage Architecture for Cloud-Supported Medical Internet of Things," IEEE Internet of Things Journal, vol. 7, no. 3, 2019.

[26] A. Celesti, A. Galletta, M. Fazio, and M. Villari, "Towards Hybrid Multi-Cloud Storage Systems: Understanding How to Perform Data Transfer," Big Data Research, vol. 16, 2019.

[27] S. Samundiswary and N.M. Dongre, "Object Storage Architecture in Cloud for Unstructured Data," In: International Conference on Inventive Systems and Control (ICISC), 2017.

[28] R.M.d.O. Libardi, S. ReiMarganiec, L.H. Nunes, et al., "MSSF: User-Friendly Multi-Cloud Data Dispersal," In: Proceedings of the IEEE 8th International Conference on Cloud Computing, 2015.

[29] J. Li, D. Lin, A.C. Squicciarini, J. Li, and C. Jia, "Towards Privacy Preserving Storage and Retrieval in Multiple Clouds," IEEE Transactions on Cloud Computing, vol. 5, no. 3, 2017. [Online]. Available: https://doi.org/10.1109/TCC.2015.2485214.

[30] P. Janviriya, T. Ongarjithichai, P. Numruktrakul, and C. Ragkhitwetsagul, "CloudyDays: Cloud Storage Integration System," In: Proceedings of the Third ICT International Student Project Conference (ICT-ISPC), 2014. [Online]. Available: https://doi.org/10.1109/ICT-ISPC.2014.6923233.

[31] R. Zhao, C. Yue, B. Tak, and C. Tang, "SafeSky: A Secure Cloud Storage Middleware for End-User Applications," In: Proceedings of the IEEE 34th Symposium on Reliable Distributed Systems (SRDS), 2015. [Online]. Available: https://doi.org/10.1109/SRDS.2015.23.

[32] E. Rios, E. Iturbe, X. Larrucea, M. Rak, W. Mallouli, et al., "Service Level Agreement-Based GDPR Compliance and Security Assurance in (Multi) Cloud-Based Systems," IET Software, vol. 13, no. 3, 2019.

[33] A. Tchernykh, M. Babenko, V. Miranda-López, A.Y. Drozdov, and A. Avetisyan, "WA-RRNS: Reliable Data Storage System Based on Multi-Cloud," In: Proceedings of the IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2018. [Online]. Available: https://doi.org/10.1109/IPDPSW.2018.00124.

[34] A. Pravin, T.P. Jacob, and G. Nagarajan, "Robust Technique for Data Security in Multicloud Storage Using Dynamic Slicing with Hybrid Cryptographic Technique," Journal of Ambient Intelligence and Humanized Computing, 2019.

[35] C. Esposito, M. Ficco, F. Palmieri, and A. Castiglione, "Smart Cloud Storage Service Selection Based on Fuzzy Logic, Theory of Evidence and Game Theory," IEEE Transactions on Computers, vol. 65, no. 8, 2016. [Online]. Available: https://doi.org/10.1109/TC.2015.2389952.

[36] D.-N. Le, B. Seth, and S. Dalal, "A Hybrid Approach of Secret Sharing with Fragmentation and Encryption in Cloud Environment for Securing Outsourced Medical Database: A Revolutionary Approach," Journal of Cyber Security and Mobility, vol. 7, 2018.

[37] S. Vulapula and H. Valiveti, "Secure and Efficient Data Storage Scheme for Unstructured Data in Hybrid Cloud Environment," Soft Computing, vol. 26, 2022. [Online]. Available: https://doi.org/10.1007/s00500-022-06977-1.

[38] G. Vernik, A. Shulman-Peleg, S. Dippl, C. Formisano, M.C. Jaeger, et al., "Data On-boarding in Federated Storage Clouds," In: Proceedings of IEEE International Conference on Cloud Computing CLOUD, 2013.

[39] M. Malensek and S. Pallickara, "Autonomous Cloud Federation for High-Throughput Queries Over Voluminous Datasets," IEEE Cloud Computing, vol. 3, no. 3, 2016. [Online]. Available: https://doi.org/10.1109/MCC.2016.65.

[40] M.I.H. Sukmana, K.A. Torkura, H. Graupner, F. Cheng, and C. Meinel, "Unified Cloud Access Control Model for Cloud Storage Broker (PS23)," In: Proceedings of the International Conference on Information Networking (ICOIN), 2019.

[41] C.W. Chang, P. Liu, and J.J. Wu, "Probability-Based Cloud Storage Providers Selection Algorithms with Maximum Availability," In: Proceedings of the 41st International Conference on Parallel Processing, 2012. [Online]. Available: https://doi.org/10.1109/ICPP.2012.51.