International Journal of Advance and Innovative Research
Volume 12, Issue 2 (XVII): April - June 2025
ISSN 2394 - 7780

LIFE CYCLE OF DATA IN CLOUD

[1]Riya Gupta and [2]Wilson Rao
[1]Student and [2]Assistant Professor, MSc. Big Data Analytics Department, Jai Hind College, Mumbai

ABSTRACT
Cloud computing has made a huge change in the big data management field. Cloud has come up with a technology that is flexible, scalable and it cuts down costs, thus making it possible to carry management and analysis of big data. The classic on-premises data management systems have always had a hard time handling the increasing amount of data that resembles big data in volume, variety, and velocity. Cloud environment like Microsoft Azure is capable of managing these huge datas with its robust infrastructure that allows companies to manage and work with large datasets with efficiency. By making use of such technologies, the organization can dynamically change the capacity of their working processors to suit the changing workload and be sure to use their resources to the maximum with the minimum cost. Another important advantage of making use of cloud computing technology in big data environments is that enterprises can exercise leverage on-demand resources. That is to say, organizations can swiftly increase or decrease the scale of their operations according to their dynamic needs, thus staying efficient. Nonetheless, in order to be completely effective in these assigning tasks to the cloud, one has to acquire the information regarding data transfer and processing which is vital for its proper functioning. This should involve knowledge of the data flow mechanisms, network latency, and the overall architecture of platforms like Azure. This research paper's primary focus has been on data processing and transportation in cloud environments, particularly on the Azure platform. It offers a thorough examination of the technological problems pertaining to the cloud ecosystem's data ingestion, processing, storage, and retrieval. It also examines security-related topics such data compliance, encryption, access management, and security standards in cloud systems. This clarifies the ways in which these factors impact data confidentiality and integrity in cloud-based big data management.

Keywords—Cloud, Data Flow, Data lifecycle, Big Data

INTRODUCTION
Organizations are turning to cloud computing as a more practical and scalable option as a result of the exponential expansion of data, which has overtaken traditional data management techniques due to the spread of digital apps and the Internet of Things (IoT). By allowing businesses to efficiently and flexibly store, process, and analyze large datasets, cloud computing provides a revolutionary approach to big data management. Cloud environments offer on-demand resources, enabling businesses to adjust their processing capacity to real-time demands, in contrast to traditional on-premises systems, which are frequently limited in scalability and require a significant upfront investment. For businesses that must handle workloads of different sizes without sacrificing efficiency or going over budget, this flexibility is crucial. In cloud computing, data flow encompasses distinct stages—data ingestion, storage, processing, and retrieval—that are essential to maximizing data management's functionality and security. Depending on the cloud service paradigm, each of these phases is handled differently. This paper explores the data lifecycle within the cloud service models, focusing on the unique mechanisms of each stage across IaaS, PaaS, and SaaS environments. (Boglaev, 2016)

OVERVIEW OF CLOUD SERVICE MODELS
Cloud computing users can request and obtain rented computing capabilities over a network that connects them to a cloud platform. A central server manages all communication between client devices and servers for data exchange. There is no one-size-fits-all approach to implementing cloud computing architecture. What works for one business might not work for another. Actually, one of the advantages of cloud computing is its adaptability and flexibility, which enables businesses to swiftly adjust to shifting measurements or markets. (Rieder, 2020)

This section presents and defines the three primary cloud service models: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). It also describes how the roles of data lifecycle management vary throughout these models. This fundamental knowledge helps readers understand how each model affects cloud environments' data management and flow.

DATA INGESTION
Data ingestion is the process of collecting and importing data files from various sources into a database for storage, processing and analysis. The goal of data ingestion is to clean and store data in an accessible and consistent central repository to prepare it for use within the organization. Data ingestion is the process of taking raw data from various sources and preparing it for analysis.

ISSN 2394 - 7780

Data discovery: The exploratory stage, during which all of the organization's data is found. The foundation for successful data ingestion is an understanding of the data environment, structure, quality, and possible applications.

Data acquisition: Data acquisition means gathering the data after the sources have been determined. Data can be retrieved from a variety of sources, including unstructured formats like spreadsheets and paper documents as well as structured databases and application programming interfaces (APIs). Maintaining data integrity during the acquisition process and managing the range of data types and perhaps high volumes provide challenges.

Data validation: Validation ensures that the data is accurate and consistent after it has been acquired. Data is examined for mistakes, discrepancies, and missing values. Through a variety of checks, including data type validation, range validation, and uniqueness validation, the data is cleansed, made trustworthy, and prepared for additional processing.

Data transformation: This is where verified data is transformed into an analysis-ready format. This could entail standardization (uniform formatting), aggregation (summarizing data), and normalizing (removing redundancies). Making the data easy to comprehend and evaluate is the aim.

Data loading: The modified data is then delivered to its assigned destination, usually a data lake or warehouse, where it is easily accessible for reporting and analysis. Depending on the particular requirements, this loading process can be carried out in real-time or in batches. When the data input process is finished, the data is prepared and ready for making well-informed decisions and producing useful business insight. This is known as data loading.

DATA STORAGE

Cloud storage is a cloud computing service in which data and files are stored offsite by a third-party provider and can be accessed through a public internet or dedicated private network connection.

- **How does cloud storage work?**

One way to store data is using on-premises networks, similarly, cloud storage also uses servers for storing data, but the data is placed on offsite servers. The majority of these servers are virtual machines (VMs) that work with a physical server. The provider generates additional virtual servers in response to the need for more storage.

A typical user gains access to a cloud storage through a web interface (web portal), website, or an app that communicates with an application programming interface. The server that you connect with, on the other hand, will redirect data to a group of servers situated in one or more big data centers, which rely on the scope of the cloud vendor's ab.

Providers that offer this service store the identical data on various machines as a backup measure. This implies that even though a server undergoes downtime for maintenance or faces an outage, the data would still be accessible by the users.

There are three main cloud storage types, each offering its own advantages.

→File storage

File storage is a technique that is used to save data in a pie file and folder structure. The data stays in the same format regardless of where it is, whether it is a cloud storage system or a client location, and the structure of the data makes it more accessible and intuitive to find and fetch it when necessary. (Boglaev, 2016)

Cloud storage for file storage is e a widely provided service allowing users to connect to the same group of files that are stored in the cloud.

→Block storage

This data model for storage arranges the information in the form of big "blocks" which are each a hard drive. The cloud storage companies utilize such blocks to divide a large amount of data into several storage nodes.

The block storage resources deliver an increased performance level across a network be it the low IO latency (which is the time taken to complete the connection between the system and the client). These resources are especially suitable for large databases as well as applications.

Block storage on the cloud can be easily scaled up to take care of the increased demand of an organization's databases and applications.

ISSN 2394 - 7780

→Object storage

Object storage is a method of data management called "object" storage, and it distinguishes and organizes data into distinct "objects." The data is the information in the file, the associated metadata, and the identifier of the object. The files contain the data in the same format that they are received and, at the same time, should allow creators to customize metadata to enable the data to be easily recognized and used. (Rieder, 2020)

The filing system is not made up of files or folder structures; they are instead stored in objects in repositories, which is a way to enable unlimited scalability. The lack of filing hierarchy and the possibility of personalization of metadata makes it possible for users to optimize storage resources at an affordable price through object storage.

Cloud-base object storage is a good way of preserving data - for the distant future. The importance of object storage has been growing with every passing day since more and more unstructured data (videos, audios, webpages, sensor data) is being stored and processed in an efficient and cost-effective way. In 2022, 90% of structured data was generated in the organizations.

DATA PROCESSING AND ANALYSIS

Data processing and analysis are critical stages in the cloud data lifecycle, where raw data is transformed into meaningful insights. This stage involves cleaning, organizing, and analyzing data to extract actionable intelligence, and it is highly dependent on the tools and services provided by the cloud service model in use (IaaS, PaaS, or SaaS).

In IaaS, data processing is largely dependent on the infrastructure set up by the users. The users can install any software packages of their choice on the virtual machines (VMs) and storage, which are in their complete control. For instance, users can deploy big data processing frameworks like Apache Hadoop, Spark, or even custom developed Extract, Transform, Load (ETL) pipelines into the VMs. While this lends the idea of elastic processing custom workflows as allowed by the cloud computing paradigm, it also means that the users have to take control of infrastructure management responsibilities as well as scaling of resources to the existing ones. IaaS supplies the required amount of computing capacity and storage but it is up to the users to optimize the processing systems for very large datasets where necessary.

When it comes to PaaS, data processing is much less concerned as the platform offers managed services that simplifies its engagement. For instance, many PaaS solutions come with native data processing options like managed DBs, serverless e.g. AWS Lambda Azure functions, and data analytics e.g. Google Big query among many others. Such platforms lend themselves for use by users concentrating on application development, with little or no need to concern about the infrastructures. For instance, a developer can apply ready-made services for data ingestion, transforming as well as analyzing without any effort for setting up the VMs or networks. This facilitates the quicker implementation of data processing and helps it to scale as the system's resources are optimally allocated and managed according to the user needs. There is, nonetheless, a downside in that the fine-grained aspects of the processing pipeline are exposed to the user, in that a lot of it is hidden by the platform.

In the case of SaaS, the entire process of data processing and analysis is left to the service provider. SaaS solutions usually include analytical tools, dashboards, and other reporting features, which are usable for specific business functions (like CRM analytics, forecasting finances, or business intelligence). Users also access the data through the application and are dependent on the SaaS provider to control the data and work on it as it comes in. For instance, Salesforce has an inbuilt analytics tool for sales and distribution and Power BI provides capabilities for visualisation of data drawing from varied sources. Also, SaaS solutions are typically user friendly in that the user does not have to be very technical in order to use advanced capabilities such as data analysis. But the unavailability of flexible options as well as the restrictions over the actual data processing systems could be a challenge for some organizations that have more precise or elaborate data processing needs.

DATA DISTRIBUTION AND ACCESS

Data distribution and access are fundamental components of the data lifecycle in cloud computing. They refer to how processed data is made available to users, applications, and systems, as well as how it is shared, stored, and retrieved across different platforms. Efficient data distribution ensures that data is accessible where and when it is needed, while robust access controls safeguard its confidentiality, integrity, and availability. This process varies significantly across IaaS, PaaS, and SaaS, with each model offering different degrees of control, security, and management capabilities.

ISSN 2394 - 7780

In the architecture of IaaS, the data distribution, as well as the access, is very extensible since the end users control the storage and the network layers. This implies that the users have to create their own basic storage systems whether it's an object storage, block storage or even file systems and also take responsibility on the way the data is spread over the various nodes or regions within the cloud infrastructure. Users can take advantage of Private clouds brough about by Amazon Web Services to store their content in different formats and many replicas also can be geographically dispersed for the purpose of ensuring availability in case one goes down. Besides that users are also afforded the freedom of geography, in that they can control fully the access of the data through allocation of roles as per the users , security protocols and management, with the addition of encrypting the data and determining who can change the information or who can access it. But all of this is possible only because the users are given a lot of power and with that, some extra responsibility of their own which is how they take care of the data control and distribution structures.

In Platform as a Service (PaaS), the data dissemination and access procedures are somewhat concealed. Such a platform includes consistent storage services which perform a lot of the distribution related activities, for instance, data backup copies, redundancy, and growth. For instance, Google's App Engine or Microsoft Azure App Services kind of platforms usually come with ready-built integrated databases, data storage, and data caching facilities which are elastic in nature for the application requirements. It is very easy for PaaS users to reassign data to different regions or nodes on the platform by leveraging the managed services that the platform provides, which take care of replication and load balancing. Nonetheless, the control over the strategies of data distribution is not as much as it is in Infrastructure as a Service. Access controls can still be imposed on certain data and platform-embedded security features like role-based access control (RBAC) can be employed to restrict data access to authorized persons only. Such as in PaaS the management of distribution of data becomes less cumbersome for the users thank the institution does PaaS wishing for more complex strategies of distribution requires trust on the institutions routing configurations.

In a SaaS model, the service provider completely takes over the responsibility of managing data distribution and access. Users only need to access the application through a web interface or an API and do not care about the location and mechanism of storing and distributing data. In addition, those service providers usually take care of data replication, and geographic distribution, and implementing failovers so that data is always accessible even in the case of hardware malfunctions or loss of server connections. How the data is accessed is strictly controlled by the access policies integrated into the application, for example, it may offer multiple user roles and support authentication and data scrambling to prevent unauthorized use. Following this model, the user is provided with the least control over the distribution and access of data, however, the provider takes the onus of making sure that data availability and secure access are assured.

To wrap things up, when it comes to understanding the distributions and access of data in IaaS, PaaS, and SaaS models, the main difference comes in control and abstraction levels. Among them, IaaS is secure and flexible enough for users to specify the desired data distribution and the access controls to be used. With PaaS, the distribution is made less complex with the use of managed services and offers a completely hands-off distribution and access model in SaaS. Therefore, the cloud model which an organization will adopt would depend on the level of control over the data distribution and the ease of operation and management.

DATA ARCHIVING AND DELETION

The last stage of any data lifecycle is data archiving and deletion, which is critical especially in cloud environments where there is lake of data. These strategies allow for the retention of data for future purposes whenever necessary, while still observing the data protection and retention policies. Within computing, the processes of archiving and deletion vary based on the cloud service model deployed since each has different capabilities, management approaches, and compliance systems.

→Data Archiving

Data archiving is a procedure to keep the already inactive computer data that would be still necessary for long-term storage or regulatory compliance purposes. Data archiving systems guarantee that files can be recovered at a later date and at the same time cut down on expenditures for memory by automatically moving to low-cost storage levels infrequently used files.

In IaaS, data archiving is very flexible. Users can store the archived data and manage it how they would like. For example, they can opt to use purpose-built dealt storage services such as Amazon Glacier or Azure Blob Storage Archive for cheap, long-term storage. Archiving in IaaS means defining custom policies, providing motion of data automation and management of storage classes so that data is archived properly. For instance, an organization may set a retention period in which the system will automatically transition aged, low access data

ISSN 2394 - 7780

to a cheaper class of storage. Furthermore, to protect the archived data, IaaS vendors usually provide encryption of the data and redundancy of the respective archived data.

In the case of PaaS, archiving is usually processed by the system with the provision of long-term data storage services. Such PaaS systems may connect to third-party systems or may have long-term storage facilities of their own. For example, Cloud Storage Nearline by Google and Cool Blob Storage by Microsoft Azure serve the purpose of archiving data. PaaS may include also extend to allowing users create automated processes for archiving data based on specific policies e.g. active data becomes stale for a certain amount of time and is archived. Some of these policies may be configurable by the users but the pain of designing and management is kindly concealed by the platform, which makes archiving quite easy but also less flexible than in IaaS.

When it comes to SaaS, the responsibility of managing archiving lies squarely with the service user. Most SaaS applications have automated archiving systems as part of the data management features of the application; especially those applications that are used for retaining data for a long period of time such as email applications, CRM systems, or document management systems. For instance, in Salesforce, some archived customer data can be available to users instead of allowing aged information to clutter the system. Health care Cloud based services SaaS solution provides managed archiving and data backup systems which preserve all data in accordance to health regulations (e.g., HIPAA, GDPR), and making sure that no loss of archived data is experienced. Yet users lack the distinction of how archiving is done because this aspect is taken up by the provider.

→Data Deletion

Data deletion is the action of permanently erasing any information from the system that is deemed unnecessary. In this case, the information should be permanently removed, and there should be no chance of recovering it which can lead to a compromise on security or privacy. Cloud environments demand appropriate data obliteration practices as a defense to many legal obligations (such as the "right to be forgotten" imposed in GDPR) and limits the amount of time an organization keeps certain types of data.

In IaaS, the implementation of data deletion policies falls in the hands of the user. As they own the entire infrastructure, users put in place deletion policies in which they make certain that all data is removed often using industry compliant data wiping mechanisms. Users can perform data deletion from virtual machines, storage volumes, and databases or alternatively create processes that will cause data deletion after certain duration. Data deletion in IaaS is performed hand in hand with data sanitization processes ensuring the storage devices have been rendered useless to the contained data. Additional services like provision of secure deletion tools, among others, may be provided by cloud service providers but the user is responsible for ensuring compliance.

In a typical PaaS (Platform as a Service) environment, data deletion is usually carried out via built-in services and policies of the platform, which are typically disruptive for the end user. For instance, several PaaS services make it possible for users to set up lifecycle management policies for the data they keep, which includes automating the process of deleting or archiving information after certain conditions are met, such as data becoming a certain age. In this case, users can still have some degree of influence on the way data is deleted and the specific time it is deleted, particularly when using particular database or object storage services. In many PaaS environments deleting of data tends to be of datapump's purging / versions control, which is tied to the application lifecycle, and additional features such as g.c. or data retention policies may be present in the platform to ensure that privacy policies of data regulation are adhered to. Even though much of the process is abstracted by the platform, the user's responsibility handles data that requires deletion.

In SaaS, data deletion is managed by the service provider according to the terms of the service agreement and applicable legal or compliance requirements. Many SaaS providers offer data retention and deletion features that automatically handle data erasure once it is no longer needed. For example, in platforms like Google Workspace, users can delete files manually, and the provider will handle the backend deletion processes, ensuring the data is completely removed from the servers. The service provider often adheres to best practices for secure deletion and may offer options for users to request permanent deletion of their data. However, as with archiving, users have limited control over the deletion process itself, and they must rely on the provider's security and compliance measures.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my mentor, Prof. Wilson Rao, for his invaluable guidance, insights, and continuous encouragement throughout this research. My sincere thanks to MSC - Big Data

ISSN 2394 - 7780

Analytics Department, Jai Hind College for providing the resources and supportive environment necessary for conducting this research.

REFERENCES

1. **Boglaev, I. (2016).** A numerical method for solving nonlinear integro-differential equations of Fredholm type, *Journal of Computational Mathematics*, 34(3), 262–284. https://doi.org/10.4208/jcm.1512-m2015-0241
2. **Lindberg, D. V., & Lee, H. K. H. (2015).** Optimization under constraints by applying an asymmetric entropy measure, *Journal of Computational and Graphical Statististics*, 24(2), 379–393. https://doi.org/10.1080/10618600.2014.901225
3. **Rieder, B. (2020).** *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press.