A CRITICAL ANALYSIS OF THE CHALLENGES AND OPPORTUNITIES TO OPTIMIZE STORAGE COSTS FOR BIG DATA IN THE CLOUD

Authors: **Ameya Shastri Pothukuchi[1], Lakshmi Vasuda Kota[2] & Vinay Mallikarjunaradhya[3]**

Sr. Product Manager, Microsoft, Redmond, USA[1], Senior IT Risk Auditor, Voya Services, Windsor, USA[2], Principal Product Manager, Thomson Reuters, Canada[3]

ABSTRACT:

Businesses are increasingly adopting cloud computing solutions and the economic implications of cloud storage have become a significant topic of study. The research paper explores how organizations manage Big Data storage on cloud platforms and the inherent challenges and opportunities involved in achieving cost efficiency.

The paper provides a brief outline of foundational concepts covering primary cost components associated with traditional on-premises storage solutions. Subsequently, the paper examines how big data came to the forefront in the mid 2000's and how cloud computing platforms can be leveraged to handle the exponential growth of complex and mega volumes of data in modern digital landscapes. This paper serves as a basis for a comparative evaluation of key challenges organizations face in managing big data and comprehensively examines economic benefits and factors associated with hosting big data in the cloud.

Ultimately, this research presents insights into the impact of the AI ecosystem on Bigdata management. The recent development and explosion of increasingly powerful AI technologies offer capabilities such as advanced analytics powered by ML, automation, and insights extraction. Foremost among these applications of AI are how these developments are helping the software industry currently and how it is capable of transforming the processing of big data generated worldwide.

Asian Journal of Multidisciplinary Research & Review (AJMRR)
ISSN 2582 8088
Volume 3 Issue 1 [January February 2022]

Keywords: *Cloud Storage, Big data, AI in cloud storage, Cloud computing*

I. INTRODUCTION:

'Big data' became a popular term in the first decade of the 21st century, although the concept of big data can be traced back to the initial days of the internet in the 1960s and to the first data centers that dealt with large data sets. Four decades later, the explosion of social networking sites and online streaming services generated on a continuous basis, complex data sets which are too large to be handled by traditional systems or data processing applications.

This necessitated the development of new frameworks to address the challenges arising from vast amounts of data generated such as capturing the data from the source; storage, analysis, transfer of data; querying, updating, sharing and visualization of data, administering information privacy policies to name a few.

II. RESEARCH METHODOLOGY

2.1 Participant Sampling and Selection:

The selection of interview and survey participants was based on purposive sampling.A pool of participants was created by purposely selecting software engineers and PMs from different types of organizations who worked or are currently working in the cloud environment domain, with an average work experience of 12.6 years.

2.2 Interview Response Coding:

Regarding data analysis, we utilized an interpretivist approach. Our process involved the initial coding of data, followed by categorization and identification of thematic development.

A total of 30 Tech industry professionals whose experience was in the range of 3 to 28 years were interviewed. The interviews and participant surveys were conducted in August 2021, with each interview lasting approximately 45-60 minutes and at least two researchers hosting it.

Verbatim recordings were transcribed with participant consent, and PersonallyIdentifiable Data (PII) was cleaned from transcription, and participants were referred to using pseudonyms.

Table 1:Interview Questions guide

1.	Do you work with Big Data in your current role? If so, to what extent?
2.	Is cloud storage leveraged to handle the Big Data in your organization?
3.	If the answer to 2 is yes, is AI being used to assist with Big Data management in the cloud environment?
4.	Which of the below trends pertaining to the deployment of AI in cloud storage do you think will likely materialize in the coming years?: A) Increased adoption of AI in cloud storage B) The level of AI deployment being the same in the future as it is now. C) Reduced AI usage in cloud storage of Big Data.
5	Is there a lack of understanding currently about how AI can realistically be used in the cloud storage, and what its limitations are?
6	Is there any additional information you've seen regardingthe adoption of AI in Cloud Storage in the backdrop of big data in the software industry that we didn't ask about, that you think is important for us to know?

II. Modeling and Analysis:

For data analysis, we evaluated our findings using the criteria described in Lincoln and Guba (1985). The 'Discussion' and Results sections below present the findings of our interviews with respondents about the impact of SDLC on AI.

To ensure reliability of our findings, we conducted an external peer review of our results and process. The data and analysis process were made available to an external peer reviewer to scrutinize our interpretations and conclusions. This step served to strengthen the dependability of our results.

III. DISCUSSION AND KEY RESULTS

3.1 Characteristics of Big Data:

To be classified as big data, data should have the qualities listed below

Volume:The size of Big Data is usually larger than TBs and PBs and involves mostly low-density, unstructured data.

Variety:Unstructured and semi-structured data tested the limitations of earlier technologies such as RDBMS since Big Data draws from text, images, audio and video. This requires additional preprocessing to derive meaning and support metadata.

Velocity:Velocity is the rate at which data is received and acted on. The former is called the frequency of data generation and the latter is known as the frequency of handling, recording and publishing.

Big data is generated and processed at a speed to meet the growth and development of data lifecycle. Highest velocity of data streams directly into memory versus being written to disk. Big Data is often available in real-time as data is produced more continually.

The list is not exhaustive since some technology practitioners include in features of big data, ***Value*** - worth of information derived from processing and analysis of large data sets and

Scalability*:* The rate at which the storage system can expand rapidly.

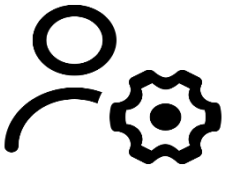
Storage costs pertaining to data farms and servers are one of the biggest Capex expenses for any organization handling large amounts of data on the cloud in Petabytes (1 Petabyte or PB is equal to 1 million GB) or Exabytes (1 PB = 1000 PB).

Facebook spent a staggering $28 billion on data centers and servers in 2021. Research firm Gartner estimates that spending on data center systems in 2021 was an estimated $207 billion on a global basis. Forrester study 2021 states that the growth in public cloud is expected to increase to one trillion USD by 2026. According to Cloud cost savings statistics report, companies using cloud computing save 20% annually on infrastructure costs.

The primary costs associated with traditional on-premises storage solutions include:

Hardware costs (costs incurred for storage devices such as servers, disk arrays, storage area network (SAN) switches, and storage controllers),

Infrastructure (server racks, cooling systems, power supplies, and physical space requirements.)



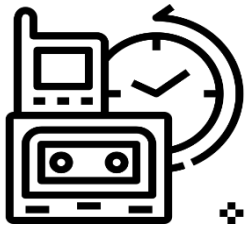
Maintenance and support costs (Regular maintenance, updates, hardware replacements, physical space and security, Power and cooling costs)

Staffing (Training and compensation for IT personnel to manage, maintain and monitor the storage infrastructure)

Software Licensing costs - costs incurred on cloud service provider for software solutions and for cybersecurity such as data protection software, firewalls, encryption, and access controls to protect stored data

Backup and Disaster recovery costs (unplanned downtime due to disasters and expenses incurred for data recovery and restoration, costs associated with recovery efforts include additional hardware for site backup, software licenses for cloud backup, and management efforts.)

Scalability costs (additional costs incurred to add hardware and infrastructure components in scaling the storage environment)



Obsolete technology:

hardware and software become outdated and may need to be replaced, resulting in costs associated with technology refresh cycles.

Organizations storing large amounts of customer data such as Facebook, AWS (Amazon Web services) and Google are constantly looking for ways to reduce and optimize storage costs. In this article, we shall review some of the common challenges encountered by organizations and the techniques to reduce storage costs for big data and the most popular strategies to achieve this goal:

IV. KEY CHALLENGES:

Here are some of the key challenges in optimizing storage costs for big data:

1. Rapid growth of big data: Most big organizations offering cloud storage services like the Big 4 (Google, Amazon, Facebook and Microsoft) produce at least hundreds of Petabytes of new data every month.

The uncontrolled growth in the data volumes stored in data centers owned by corporations or academia puts stress on storage resources. This leads to a lack of cohesive strategy in the management of data storage resulting in scattered data silos. Also known as data sprawl, this is a major concern with Big Data as data originates from various sources and in diverse formats. Data sprawl could result in inefficiency due to duplicated efforts as multiple teams collect and store the same data independently.

Amidst these challenges, the work by M. Muniswamaiah, T. Agerwala, and C. C. Tappert in the paper titled 'Approximate Query Processing for Big Data in Heterogeneous Databases' sheds light on strategies to optimize query processing in heterogeneous data environments, which can play a pivotal role in addressing data sprawl and enhancing cost-effective data management in the era of rapid data growth.

***2. High cost of storage*:** Storage is one of the most expensive components of a big data infrastructure. For perspective, Amazon has spent $35 billion on data centres for AWS in just one region (Virginia) in the last few years

3.*Data Retention policies*: Organizations typically have specific data retention policies necessitated by regulatory compliance or business requirements. Walking the tightrope between retaining data for legal or analytical purposes can be quite challenging.

4. *Data Redundancy*: Replicating data is one of the key ways that organizations adopt to achieve fault tolerance and high availability. However, this strategy also increases storage costs and is a necessary cost most organizations have accepted as an 'essential evil'.

5. *Networking costs:* The cost of networking is another notable cost factor to be considered when planning to optimize for costs associated with storing and accessing big data. When the data is in the order of Petabytes, the costs of bandwidth for moving data internally within an organization's network or between the cloud and an on-premise infrastructure is not insignificant in the least.

Transferring large volumes of data between on-premises systems and the cloud or between different cloud services can incur significant data transfer costs and latency. Optimizing data movement strategies is essential to minimize these costs.

6. *Lack of expertise:*There is a shortage of skilled professionals who have expertise in developing and implementing complex algorithms and machine learning models for Big Data in Cloud environment. To extract insights from Big Data requires expertise and experimentation. The cost of research and development of these models can be substantial.

***7. Cost of obtaining insights*:** Resource Intensive Processing requires significant computational resources, including high-performance servers, clusters, and specialized hardware like Graphics Processing Units (GPUs). The cost of maintaining these resources in cloud can be substantial. Inconsistent or redundant data can lead to conflicting analyses and inaccurate insights.

8. ***Metadata Management:***Metadataprovides context about the data and is crucial for understanding data lineage, quality, and usage. Ensuring consistent metadata across distributed cloud resources is challenging.

9. ***Vendor Lock-In:*** Locking into a specific cloud provider's service can limit flexibility and increase dependencies. Data and applications need to be designed with portability and interoperability in mind to avoid vendor lock-in.

V. COMMON METHODS OF OPTIMIZING STORAGE COSTS:

80% of our respondents said they leverage cloud storage for everyday tasks. But the usage of cloud storage presents its own set of challenges. Despite these challenges, there are a number of ways to optimize storage costs for big data. Some of the most common techniques include:

1.De-duplication: De-duplicating data to avoid storing redundant information is a common technique to decrease the cost and size of stored data.
For instance - Dropbox successfully saved significant storage costs by implementing data de-duplication techniques. As a cloud storage provider, Dropbox deals with a vast amount of user data, and de-duplication helped them eliminate hundreds of Petabytes of redundant data using deduplication – thus significantly reducing their storage costs.

2.Data compression:Data compression is another very popular technique that is effectively used to reduce the storage cost and space for storing large data lakes and warehouses.
Netflix famously developed the open-source data compression tool called "Brotli" which achieved compression ratios substantially higher than existing methods. This enabled the media giant to store its humongous volume of video-heavy data in an efficient way and save millions of dollars in storage costs

3.Tiered storage: To optimize storage costs, organizations frequently implement data storage strategies that are determined by the importance and frequency of data access. For instance, data that is very occasionally accessed will be stored in cold storage in magnetic tapes, but frequently accessed data will be stored in HDD or even SDD.

4.Hybrid storage: On-premise infrastructure has some inherent advantages such as security and control, while cloud infra is readily scalable and usually cost-effective too. Hybrid storage seeks to combine the best elements of both these worlds, which is what most organizations do in order to save costs while having the most important data in their own data centers.
As a real-world example, the Mayo Clinic saved an estimated $10 million per year by using cloud storage.

5. Data Partitioning: To reduce the amount of network traffic and cost overheads associated with sending large amounts of data from one part of an organization to another (which could be thousands of miles away), data partition is an effective way to optimize costs enabling select retrieval of data. Data partitioning is commonly done to fragment data into smaller, more manageable chunks, basedon criteria like geography, time or any other key attribute.
Social media giant Facebook (now Meta) notably implemented a combination of compression, de-duplication, and encoding techniques to decrease their storage costs by millions of dollars every year.

***6. Use data analytics to understand your data usage*:** Data analytics help organizations to understand their data usage patterns, peak periods when their business experiences high online traffic and identify areas where they can optimize their storage costs. For example, businesses can analyze their data usage to identify data that is no longer needed and can be deleted.

7.Use of data lifecycle policies to minimize the storage overhead*:* Classification of data on the basis of regulatory compliance, sensitivity, access frequency and importance to business can help in avoiding the costs arising from storing unnecessary data. These factors can in turn direct the purging and Data archival, data replication/backup policies. Setting up data governance policies and implementing centralized data management through data warehouses or data lakes can consolidate data sources and establish guidelines for data creation, storage, and management.

8.Monitoring the usage metrics:Reviewing the effectiveness of data policies can act as a pointer for autoscaling policies that dynamically adjust storage resources based on demand. When predetermined storage thresholds are met, the policy triggers an increase in the storage capacity. This ensures the resources are allocated as needed and minimizes the risk of overprovisioning.

VI. RESULTS:

The aim of this research is to understand how the recent advances in AI helps hosting Big Data in Cloud Storage. After conducting an extensive review of the literature and an empirical study through our interviews, we have observed that AI will have following impacts and characteristics:

Use of ML and AI in Big Data Cloud Storage:

In recent years, many organizations are using AI (Artificial Intelligence) and ML (Machine Learning) to optimize data storage costs on the cloud. More than half of the IT professionals believed that the pace of software deployments will accelerate significantly in the next 5 years with the advent of Generative AI. (Pothukuchi, L. V. Kota et al, *Impact of Generative AI on the Software Development Life Cycle (SDLC)*, IJCRT). As such, AI can help organizations process, analyze and make meaningful use of vast amounts of data.

Data Processing and Cleaning:AI algorithms can automate the process of data cleansing and preprocessing, identifying and rectifying errors, missing values, and inconsistencies in large datasets. 60% of the subjects indicated that they deal with Big Data Analytics on a day-to-day basis.

Predictive Analytics: AI-driven predictive analytics models leverage historical data to make future predictions. Machine learning algorithms analyze Big Data to identify patterns and trends, enabling organizations to make informed decisions. 30% of the survey respondents said that AI is already at a stage carrying out crucial operations such as automating and enhancing data preparation and performing other complex analytical tasks

Anomaly Detection: AI-powered anomaly detection techniques identify deviations from expected patterns in data, helping detect fraud, faults, and unusual events.

Natural Language Processing (NLP): NLP techniques enable machines to understand, interpret, and generate human language, which is useful for analyzing unstructured textual data in Big Data.

DeepLearning: Deep learning architectures, such as neural networks, are utilized to extract complex patterns and representations from Big Data, enhancing the accuracy of tasks like image and speech recognition.

Automated Insights: AI algorithms can automatically extract insights and generate meaningful reports from Big Data, enabling decision-makers to understand trends and make data-driven choices.As much as 80% of the respondents shared a view that there will be increased adoption of AI in cloud environment specifically in the storage and analysis of Big Data.

Further, an IT expert stated that latency in the cloud environment coupled with bias and inherent limitations of algorithms may result in outcomes which are not always best for the enterprises. While AI capabilities are multifold, it suffers from constraints owing to the nature and training of AI algorithms.

Below are some real-world examples of organizations employing AI in the cloud:

- Microsoft Azure Storage uses machine learning to tier data between hot, warm, and cold storage tiers
- Amazon S3 uses ML (Machine learning) to identify and remove duplicate data.
- Google Cloud Storage uses ML to compress data more efficiently.
- IBM Cloud uses machine learning to forecast data storage needs up to 12 months in advance.
- VMware vRealize Operations uses machine learning to automate data management tasks across on-premises and cloud environments.

VII. CONCLUSION:

The correct strategy to optimize storage costs for big data will vary depending on the data architecture of the organization and the specific needs of business. There is no one approach that would fit all organizations, but the cost savings achieved can be immense if the right strategy is executed effectively.

These findings have important implications for both researchers and practitioners. For researchers, this study highlights the need to plan for application of AI in Big Data processing. We have already seen in paper [3] that Generative AI is changing the very basics of the Software Development Life Cycle. While this study offers valuable insights into the issue of AI in cloud environment, it is important to acknowledge its limitations. The relatively small sample size of this study may restrict the extent to which the findings can be generalized to a broader population or industries. Nevertheless, the results of this study can still provide significant implications for future research and practice in addressing the impacts of Generative AI on software development.

The researchers note that the use cases of AI are multifold and AI technologies are very resourceful in hosting big data in cloud storage irrespective of the size, scale and nature of the enterprises adopting it.

REFERENCES:

1. SS Gill, S Tuli et al, *Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges,* 2019, Internet of Things.
2. Yang, Chaowei, Qunying Huang, Zhenlong Li, Kai Liu, and Fei Hu. "Big Data and cloud computing: innovation opportunities and challenges." International Journal of Digital Earth 10, no. 1 (2017): 13-53.
3. L Inonescu, *Big Data, Blockchain, and Artificial Intelligence in Cloud-based Accounting Information Systems*, 2019

4. Gupta, Rajeev, Himanshu Gupta, and Mukesh Mohania. "Cloud computing and big data analytics: what is new from databases perspective?." In International conference on big data analytics, pp. 42-61. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
5. Elazhary, Hanan. "Cloud computing for big data." MAGNT Res Rep 2, no. 4 (2014): 135-144.
6. Muniswamaiah, Manoj, Tilak Agerwala, and Charles C. Tappert. "Approximate query processing for big data in heterogeneous databases." *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020.
7. Zhang, Liang-Jie. "Big services era: Global trends of cloud computing and big data." IEEE Transactions on Services Computing 5, no. 04 (2012): 467-468.
8. Das, Madhusmita, and Rasmita Dash. "Role of cloud computing for big data: A review." Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2 (2021): 171-179.
9. Purcell, Bernice M. "Big data using cloud computing." *Journal of Technology Research* 5 (2014): 1