Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 41 No. 4 (2022)

Security and Privacy Considerations in Cloud-Based Big Data Analytics

[1]Siddhant Benadikar, [2]Rishabh Rajesh Shanbhag, [3]Ugandhar Dasi, [4]Nikhil Singla,[5] Rajkumar Balasubramanian

[12345]*Independent Researcher, USA.*

Abstract

The paper aims at critically analysing the issues of security and privacy in cloud-based big data analytics as one of the innovative fields which employs the opportunities of cloud computing and big data analysis. The paper focuses on the specific issues concerned with implementation of these technologies, namely data security, data integrity, identity, and data protection legislation. It offers a comprehensive understanding of numerous security solutions and privacy-captivating strategies which include, cryptology, protected secure multi-party computation, and differential privacy. They also describe such new technologies as quantum computation and machine learning and connect them with threats and risks. Based on the analysis of the state of the art in literature and research applied in the cloud big data context, this study seeks to present benchmarks for cloud big data analytics security and present both recommendations to practitioners and directions for future research in the field.

Keywords: Cloud Computing, Big Data Analytics, Security, Privacy, Encryption, Data Protection, Regulatory Compliance, Emerging Technologies

1. Introduction

1.1 Background of Cloud Computing and Big Data Analytics

Cloud based computing and big data analysis has now emerged as one of the most effective technologies to process and analyse volumes of data. Cloud computing allows IT to scale up and out while big data analytics offers the set of tools and methodologies to analyse the collected and stored data effectively. How big the cloud computing market is today can be seen from a report by MarketsandMarkets, wherein the current global cloud computing market size is said to be $371. To $832 billion in 2020 However, it should be noted that the global outward FDI was also increasing steadily over the years and has grown from $4 billion in 2020. It is estimated that market reach $1 billion by 2025 growing at an average rate of 17 percent annually. 5% through the forecast period (MarketsandMarkets, 2020). This fast growth shows still continuous expansion of cloud solutions market among the companies of different sectors.

Whilst managing big data has become a nightmare, big data analytics however, has now become a strategic element of any business environment. Currently, IDC believes that the global market of big data and business analytics will grow up to $274. There is predicted average growth rate of 13% to reach $3 billion by 2020. 2 % from 2018 to 2020 (IDC, 2019). This growth is attributed to the rising number of data types, the rate at which this data is produced, and variety that arises due to data sources such as social media, IoT, and enterprise systems.

1.2 Importance of Security and Privacy in Cloud-Based Systems

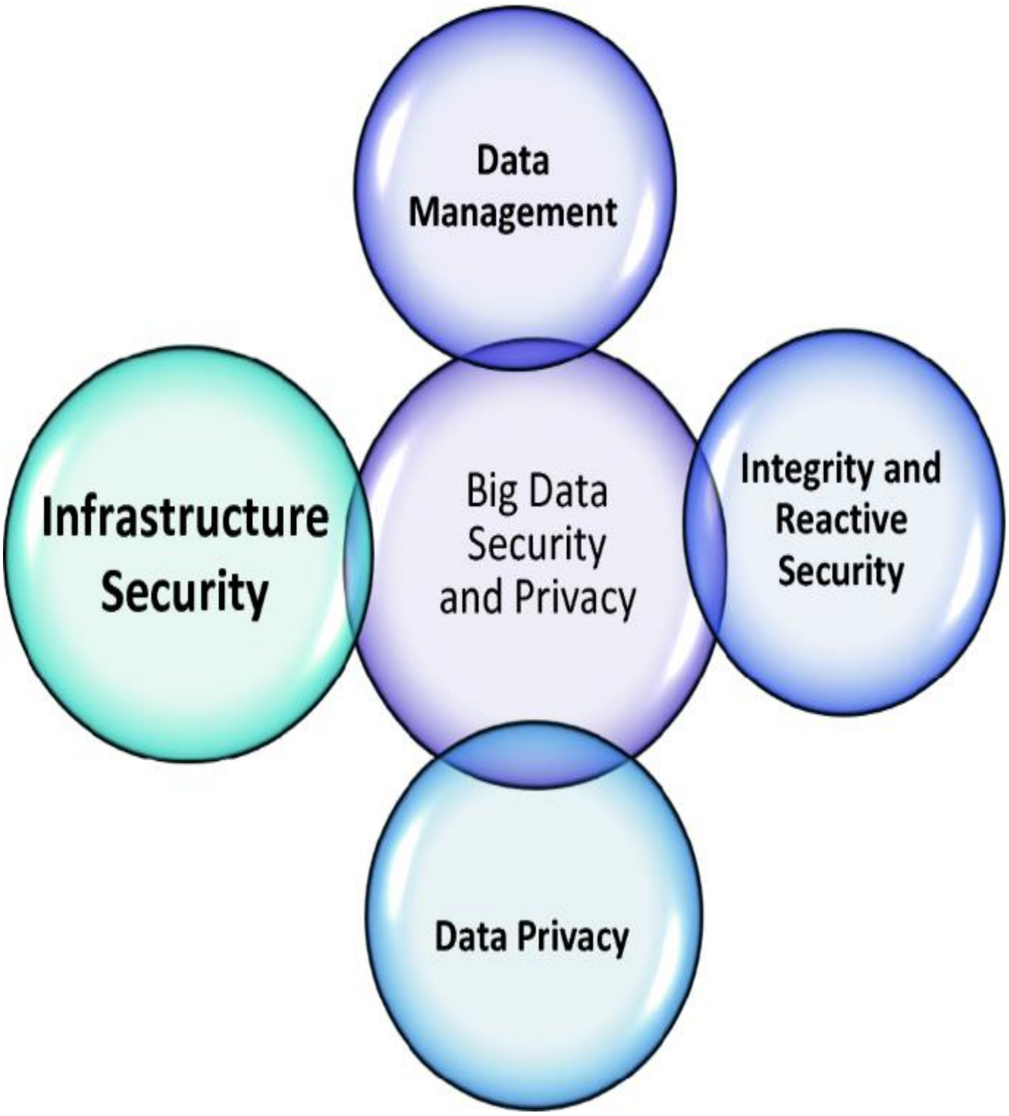
With more organizations utilizing cloud based big data analytics in decision making as well as business operations, data security and privacy become crucial. Because of the distributed nature of cloud computing as well as the huge volume of data in big data analysis, there are issues of data confidentiality, integrity, and availability. According to IBM, the global average cost of a data breach in 2020 was $3. 86 million, which indicates that companies face rather vast financial and reputational losses when using cloud-based systems (IBM, 2020).

In addition, the emerging legislations governing the protection of data and privacy such as the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States put more emphasis and expectation on cloud big data analytics security and privacy. It is imperative that organizations not only safeguard against these external threats and follow these regulations to mitigate fines as well as legal consequences (Zhang et al., 2014).

1.3 Scope and Objectives of the Study

Consequently, the objective of this research paper is to assess comprehensively the security and privacy issues of cloud based big data analytics.

1. To understand the construction and interaction between cloud computing and big data analysis.
2. To recognize and categorize the major security issues and privacy issues associated with cloud big data systems.
3. To compare current systems and approaches for security, privacy, and compliance evaluation.
4. To examine newer computing architectures and their implications on security and privacy of data in cloud based big data analyses.
5. To review outstanding matters and propose suggestions for experts and future research with respect to the field.


Data Management
Infrastructure Security
Big Data Security and Privacy
Integrity and Reactive Security
Data Privacy

In this research, the components under consideration include cloud computing models and architectures, big data processing methods, security and privacy measures, and legal policies. Using the framework developed from the analysed academic sources, industry reports, and technical standards, this paper aims to present the challenges and opportunities concerning the current state and the development of big data analytics security in the cloud environment.

2. Overview of Cloud-Based Big Data Analytics

2.1 Cloud Computing Architecture

Cloud computing architecture typically consists of three main service models: The major categories of cloud computing services are: IaaS, PaaS, and SaaS. These models are threefold with possibilities of public, private, and hybrid cloud setups. The definition of Cloud computing by the National Institute of Standards and Technology (NIST) is:

Cloud computing is a model for providing ‘easy, on-demand access to shared pools of configurable computing resources through “more secure network connections.

The key characteristics of cloud computing include:

1. On-demand self-service
2. Broad network access
3. Resource pooling
4. Rapid elasticity
5. Measured service

Table 1 provides an overview of the three main cloud service models and their characteristics:

Service Model	Description	Examples
IaaS	Provides virtualized computing resources over the internet	Amazon EC2, Microsoft Azure VMs
PaaS	Offers a platform for developers to build, run, and manage applications	Google App Engine, Heroku
SaaS	Delivers software applications over the internet	Salesforce, Google Workspace

2.2 Big Data Analytics: Concepts and Techniques

Big data analysis is the method of carrying out analysis on large and diversified data sets to constitute desirable information such as new trends or relations in the market, nature of consumption among customers, etc. The concept of big data is often characterized by the "5Vs”: Volume: First of all, it is worth stating that there is a great amount of data that is either produced or gathered.

1. Velocity: In the specific case the rate at which data is produced and analysed
2. Variety: The different characteristics of data gathered from several sources
3. Veracity: The reliability of the data collected and the subsequent analysis that is done on the data.
4. Value: It means that the ability to receive more meaningful information in the result of researches from the data is limited.

Some key techniques in big data analytics include:

1. Machine Learning and Artificial Intelligence: Models that can build knowledge that enables the computer to make some decision or prediction about the data.
2. Data Mining: Learning how to gain insights from big sets of data, correlation identification.
3. Predictive Analytics: A statistical approach of making predictions based on past performance and occurrences.
4. Natural Language Processing: The innate ability of man to process human language data.
5. Cluster Analysis: To put various similar data points together with a view of finding patterns and structures in the information available.

2. 3 Coupling of Cloud Computing and Big Data Analytics

The integration of cloud computing and big data analytics offers several advantages, including:

1. Scalability: The cloud infrastructure is very much flexible capable of accommodating the increasing amounts of data and the processing demands that come with it. This is especially so with the ever-challenging big data applications where there could be a sudden surge in the data or the analytics load.
2. Cost-effectiveness: In Pay-as-you-go models, the requirements of purchasing significant own funds of the equipment and programs do not arise: According to elasticity organizations can be able to use the cloud to address issues to do with resource utilization and cost (Xu et al., 2014).
3. Flexibility: As the solutions are cloud-based, the deployment period and availability of more reliable analytical tools are significantly reduced. This allows organizations to easily trial and deploy a variety of big data analytics processes without large amounts of necessary infrastructures.
4. Collaboration: One of the main benefits of cloud platforms is that it allows different groups of people simultaneously analyse data located on the platform irrespective of the distance. This is particularly useful in big data projects that involve multiple owners and/or input data sets.
5. Advanced Analytics Capabilities: Another advantage is that cloud providers usually provide managed services for big data analytics including ML platforms and DW solutions that can help to speed up the construction and implementation of complex analytics applications.

But at the same time, it also raises new security and privacy issues which should be managed well. The subsequent subtopics will expound on these challenges and possibly recommend the solutions.

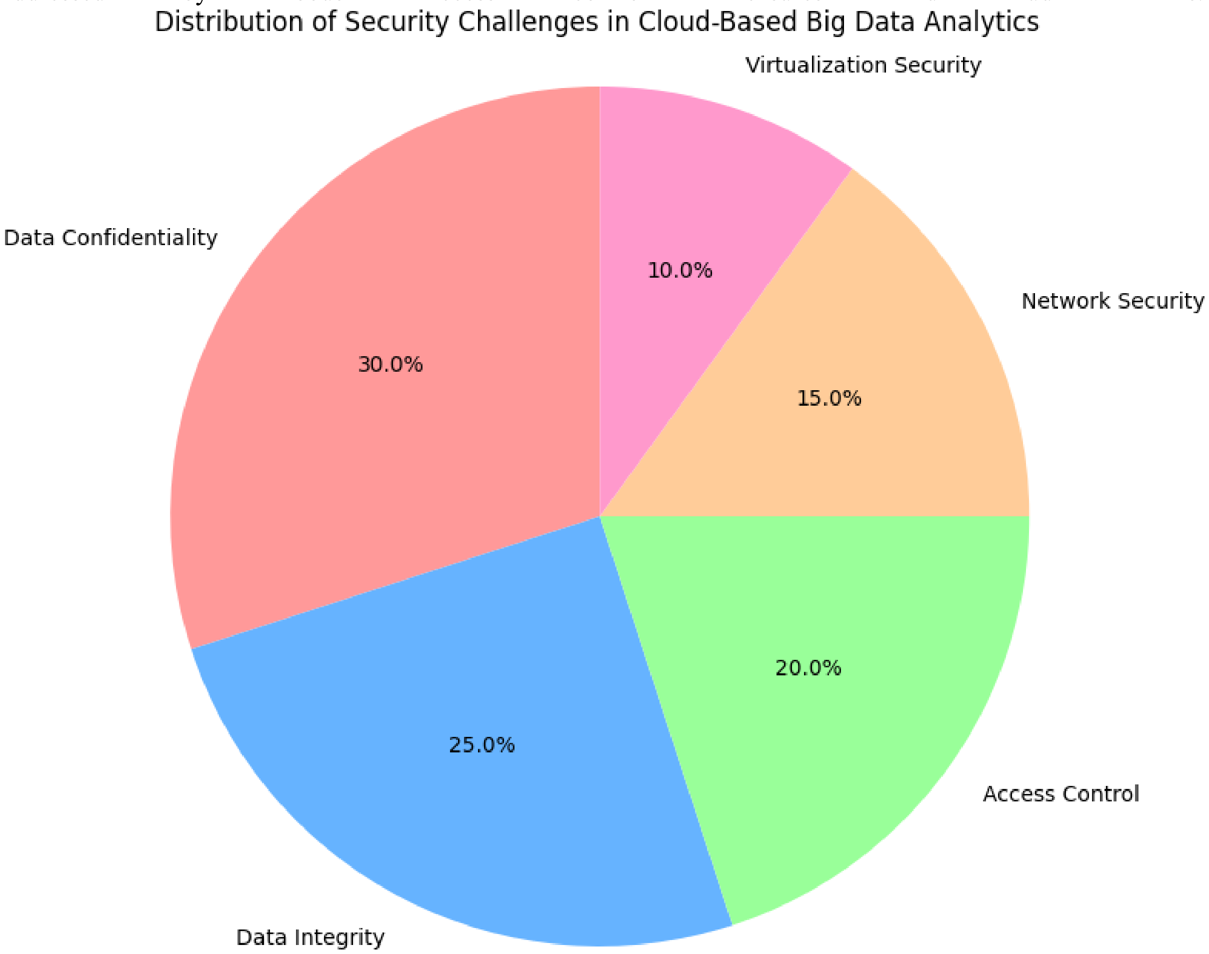
3. Security Challenges in Cloud-Based Big Data Analytics

3.1 Data Confidentiality

Data protection becomes a major challenge when handling big data in clouds because of the sharing of resources.

- Multi-tenancy: The scenario of using a common physical infrastructure where one tenant is an organization and another tenant is a competitor can cause sensitive information leakage between tenants. The need to protect confidentiality means that there must be measures in place that will hinder unauthorized individuals from accessing such data in different User Environments.
- Data Transit Security: The flow of data from premise-based systems to the cloud or between two or more services within the cloud exposes the data to interception. Privacy mechanisms are crucial to safeguard data transmission, and foolproof measures include communications security and cryptographic processes.
- Insider Threats: It is highlighted that employees of cloud service providers and the providers themselves can gain direct access to the customer’s data, which means there is a potential threat of intentional or unintentional leak of information. Extended access rights and poor security checks are one of the critical vulnerabilities that need to be

addressed by robust access control measures and audit trails.



3.2 Data Integrity

Preserving the data integrity of big data in cloud environments is not easy because data is dispersed over the cloud during handling and processing. Key concerns include:

- Data Corruption: Seamless data movement and the distribution of computation result in data degeneration. Therefore, there is need for strong methods of error-prone detection and correction that can be applied at various stages of Analytics process.
- Unauthorized Modifications: Another threat is that the data stored in cloud environments might be maliciously modified by the corresponding players. Protecting the material by greater access controls, use of audit trails, and change-versioning can assist in identifying modifications that are beyond the user's permissions (Wang et al., 2013).

3.4 Network Security

The issues of network security in the context of Cloud-based big data analytics are as follows, because of the dispersed distribution of resources and integration of large amount of data transmission. Due to integration and the dependency of the different services on each other and the constant data transfer between the different layers and components of the analytics pipeline, there are several possible vectors of attack.

The main issues that can be identified include the lack of proper security measures for the information being transferred. On the one hand, enormous data volumes are transmitted between on-premise infrastructures, cloud repositories, and processing elements; therefore, protecting these avenues is crucial. VPN and such Upper layer's protocols as TLS are used to secure the data in the transmission channel, whereas the volume of data can sometimes cause certain speed issue.

Another component of the network security is careful selection of firewalls and IDS for the protection of cloud resources from the outside world. They have to be designed to process the high velocity data streams characteristic to the big data systems while incurring rather low latency. SDN technologies are adopted gradually as the tools that can offer solutions for the flexible and scalable network security problems in the cloud.

Moreover, the adoption of public cloud services creates novel threats in terms of the exposure of Internet facing endpoints to threats and threats acting on them. Such channels of access require to be properly managed and secured since they present potential avenues of insecure access and therefore data leakage (Terzi, Terzi, & Sagiroglu, 2015). Ensuring that a robust authentication person is in place, encompassing multi-factor authentication for each attempt to connect to the firm's network, is important in warding off these threats.

3.5 Virtualization Security

Virtualization is another key technology of cloud computing since it allows the management of individual aspects of the physical hardware to be shared among many customers. However, it also creates several security issues concerning big data analytics within the framework of the given approach. The major threat is what is known as cross-VM where an attacker invades one VM and can access others running on the same physical server.

To overcome this, several isolation methods like the use of hardware assisted virtualization and a secure hypervisor are adopted by the cloud providers. Nevertheless, due to the highly complex big data workloads the mentioned isolation mechanisms can sometimes struggle and possibly introduce new points of weakness. The effectiveness, therefore, entails routine security assessments of virtualization as well as vulnerability assessment, commonly undertook via penetration testing.

Managing VM images and Snapshots is another issue in the security of virtualization. These can contain data and configuration information a programmer does not want a potential attacker to get their hands on. Among the big data workloads, VM images are also sensitive and must be guarded well; therefore, access control, encryption, and erasure should be used intensively.

Another source of trouble with cloud environments is in concerns to the VMs that are short-lived, provisioned, and decommissioned often; their security policies need to be consistent. The existence of automated security policy enforcement and the use of continuous monitoring tools as crucial components of any environment are critical when it comes to guaranteeing that all VMs are compliant with the security policies of an organization independently of the stage at which they are in the life cycle.

4. Privacy Concerns in Cloud-Based Big Data Analytics

4.1 Data Collection and Storage

A prime concern is the privacy of such large data collections and storage in the cloud environments. Companies tend to gather a lot of personal and private data, from the customers' simple nationality and age to their purchasing habits and preferences. While cloud storage an efficient way of storing the data, the fact that this information is stored in one location causes some concerns about vulnerability of this data to cyber criminals and uncertainty of who owns the data. One of the challenges therefore is to deal with the issues of data protection laws, many countries have laws that need to be complied with when collecting and storing any data that may be considered as personal data (Sookhak et al., 2017).

These concepts demand comprehensive data governance programs that cover data categorization, data provenance, and identification of data to be collected and retained in an organization. This is usually achieved through practices like applying data minimization measures and determining appropriate durations for the maintenance of data and their disposal.

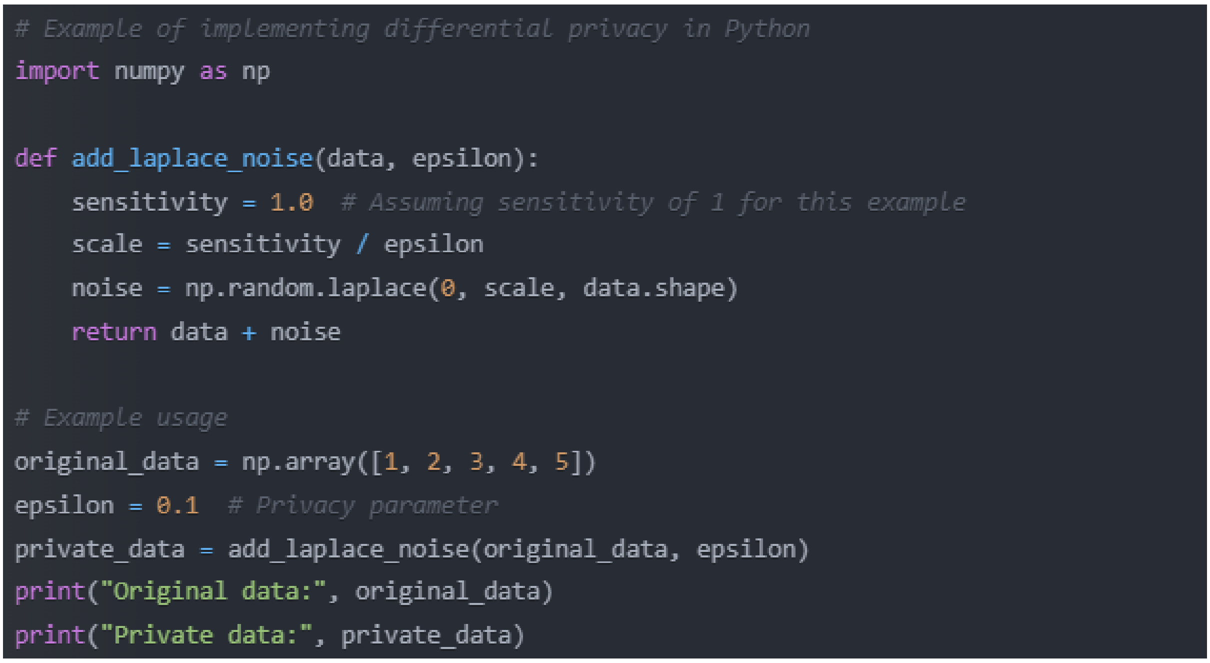
It may also imply that cloud service providers or their employees may have influence over the data and have unauthorized access to it. While most reputable providers ensure strict access controls and auditing standards, the fact remains that even access to such information may become a headache to organisations dealing with such datasets. The overall protection can be further enhanced with efforts like client-side encryption, where data is encrypted before being uploaded to the cloud.

4.2 Data Processing and Analysis

Challenges are also posed when big data records are processed and analysed in cloud surroundings. Machine Learning and Artificial Intelligence methods, have the potential of reaping very sensitive information about people’s lives even from the most harmless of figures.

Another problem which can be encountered is reidentification of data which was initially anonymized. Due to the growth of the volume and heterogeneity of data, the opportunity to identify people using supposedly anonymous data also increases. This is commonly referred to as the mosaic effect, and it is a highly dangerous situation for the privacy of the individuals and the non-applicability of the anonymization techniques (Sagiroglu & Sinanc, 2013).

Due to the above issues, companies have started applying privacy-preserving analytics solutions. Some of these are differential privacy where noise is introduced to datasets proportionally to the privacy level to deny a specific person’s information while facilitating research. Another promising approach is to use homomorphic encryption of reads so that data can be computed on while remaining encrypted all the time.



```python
# Example of implementing differential privacy in Python
import numpy as np

def add_laplace_noise(data, epsilon):
    sensitivity = 1.0  # Assuming sensitivity of 1 for this example
    scale = sensitivity / epsilon
    noise = np.random.laplace(0, scale, data.shape)
    return data + noise

# Example usage
original_data = np.array([1, 2, 3, 4, 5])
epsilon = 0.1  # Privacy parameter
private_data = add_laplace_noise(original_data, epsilon)
print("Original data:", original_data)
print("Private data:", private_data)
```

4.3 Data Sharing and Third-Party Access

Big data analytics is usually a cooperative process that results in the sharing of information with third parties including other businesses, academics or service providers among others. This data sharing has brought new risks and issues when it comes to controlling of personal information.

This means that there are reaches of data that certain third parties should not have access to no matter the level of the organization’s business relationship with them; this means there has to be policies regarding the sharing of data access to it with such third parties. Several procedures of data tokenization and masking can be applied in order to prevent the revelation of important value fields, while still making sense of the data.

Another factor is probability to have data moving across borders internationally, especially where global cloud service providers are being used. This may lead to compliance issues which entail a set of laws and regulations since the level of data protection across various regions may not be the same (Sarkar, 2020). Geographical location of data also becomes an important factor where international transfer of data does occur and adequate protection measures have to be put into practice.

4.4 Compliance with Data Protection Regulations

Privacy remains a significant concern in any cloud-based BDA, and meeting the requirements of data protection regulations is a crucial step. Laws that are currently in place, including GDPR in the EU and CCPA in the USA, set high standards concerning the gathering, utilization, and holding of personal information. Risk management requires the use of DPIAs that will need to outline thorough studies to ensure that privacy risks of big data analytics are well addressed. It concerns defining how personal data moves through various processes, recognizing risks, and applying relevant technological and managerial safeguards on personal data. Transparency and enabling individuals to have control over their data are some of the fundamental principles that data protection guidelines and laws dictate. Organizations must provide procedures for gaining express consent, processing data access requests from a subject, and offering options to transfer and delete data.

Table 2: Key Data Protection Regulations and Their Impact on Cloud-Based Big Data Analytics

Regulation	Key Requirements	Impact on Big Data Analytics
GDPR	Data minimization, Purpose limitation, Data subject rights	Stricter controls on data collection and processing, Need for explicit consent
CCPA	Consumer right to know, right to delete, right to opt-out	Enhanced data mapping and inventory processes, Implementation of consumer request handling mechanisms
HIPAA	Protection of health information, Strict access controls	Specialized security measures for healthcare data, Limitations on data sharing and analysis

This means that organizations should keep abreast with new and emerging data protection requirements and thus change their practices on cloud based big data analytics. This can include the use of privacy by design, where privacy is considered throughout the process of data analytics.

5. Security Measures and Best Practices

5.1 Encryption Techniques

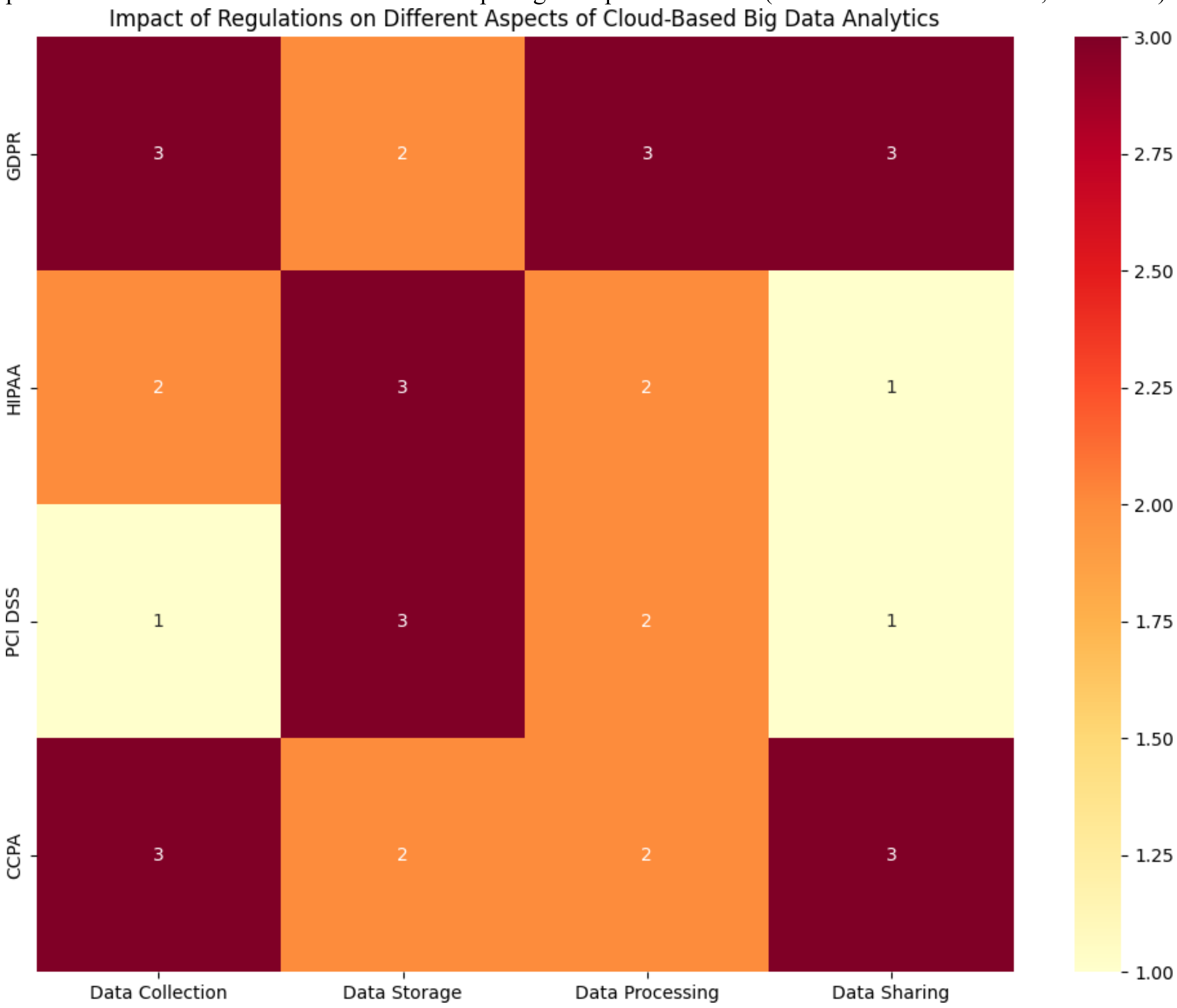
Encryption continues to prove itself as one of the most effective techniques of ensuring data security and confidentiality in big data analytics in the cloud environment. Further it is crucial in cases where information must be safeguarded while idle, in the process of transmission or mere processing. For data at rest, two major techniques used are Full-disk encryption and File system or File-level encryption (Mell & Grance, 2011). A majority of cloud providers have integrated encryption services and data segment encrypts extra measures of encryption for delicate information. In transit, data protection is often achieved by means of cryptographic protocols like Transport Layer Security (TLS) or Internet Protocol Security (IPsec).

Such protocols help to ensure that the data is always encrypted while in transit within the cloud structure or even while moving from the cloud environment to the traditional internal network. However, the question that might be hard to answer pertains to how to address this encryption at scale when dealing with big data streams. New approaches to computing are being developed and one of these is the attribute-based encryption that is recently being considered in cloud data big settings. This is especially helpful in multi-tenanted cloud systems where all the tenants' data must be highly guarded.

5.2 Secure Multi-party Computation

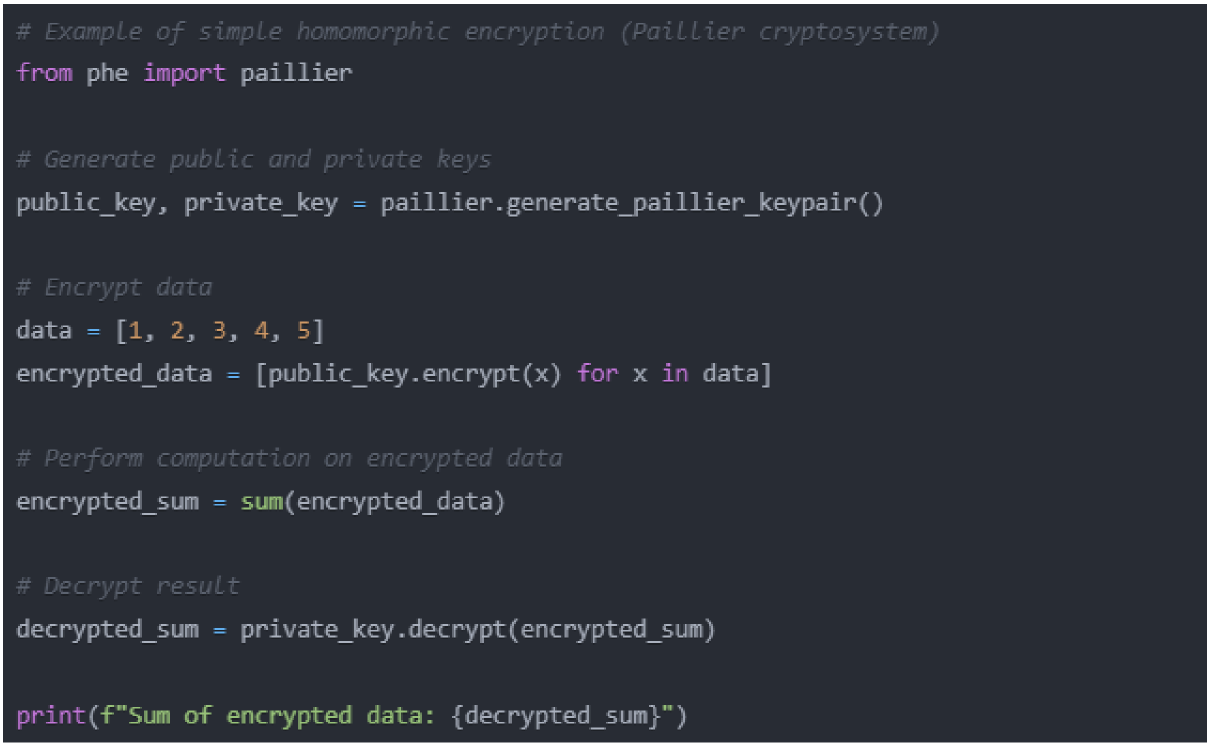
Secure Multi-party Computation (SMC) is a cryptographic protocol for several parties to perform an operation on the data they received without the other parties having to know what data the other party input. In the broader setting of cloud-based big data analysis, SMC can allow for joint analysis of big-data –containing sensitive information — without having the raw data available to any of the parties, the cloud provider included.

For instance, several hospitals can apply SMC as a means to analyse data on patients, searching for disease patterns or possible drug interactions, without compromising the patients' information. Although SMC provides strong levels of privacy and can work with very sensitive data, it is not very efficient for big data scenarios. Current research works are concerned with fine-tuning SMC protocols for the big data case and also in incorporating these protocols with distributed computing platforms (Mehmood et al., 2016).



5.3 Homomorphic Encryption

The technique of homomorphic encryption represents an innovation that makes it possible to carry out operations on encrypted information without decrypting it. This has a major impact on the paradigms of cloud-based big data analytics, as it could potentially enable organizations to offload complex calculations to the cloud providers while keeping the information safe from prying eyes. There are three types of homomorphic encryption: homomorphic encryption can be categorized into three levels including partially homomorphic encryption (PHE), somewhat homomorphic encryption (SHE), and fully homomorphic encryption (FHE). It should be noted that FHE scheme provides the maximum level of flexibility as it enables any computation on encrypted data; however, it is still time-consuming and the current state of FHE technology is not suitable for real-world data analytics. Regardless, PHE and SHE are useful tasks in more particular big data scenarios, including privacy-preserving machine learning (Li et al., 2017).



```python
# Example of simple homomorphic encryption (Paillier cryptosystem)
from phe import paillier

# Generate public and private keys
public_key, private_key = paillier.generate_paillier_keypair()

# Encrypt data
data = [1, 2, 3, 4, 5]
encrypted_data = [public_key.encrypt(x) for x in data]

# Perform computation on encrypted data
encrypted_sum = sum(encrypted_data)

# Decrypt result
decrypted_sum = private_key.decrypt(encrypted_sum)

print(f"Sum of encrypted data: {decrypted_sum}")
```

5.4 Blockchain for Data Integrity

As a technology that is conceptually decentralised and whose data cannot be easily altered, blockchain is gaining popularity as a means of implementing security for big data in the cloud. Due to the ability of developing an unalterable database of records of data transactions and changes, blockchain creates a proof of data authenticity and history modification.

When it comes to big data analytics, blockchain can be implemented for the generation of cryptographic hashes of a dataset at different levels of analysis. These hashes are then stored in the blocks, and other people can use them to check the data was not modified in transit. This is especially useful in circumstances which call for compliance to laws especially when it comes proof that the data has not been modified.

Certain cloud vendors are starting to launch services based on blockchains that can be smoothly incorporated into big data analysis applications. However, the application of blockchain solutions to large numbers of big data still presents a problem that is being actively researched by scholars.

5.5 Identity and Access Management

IAM is essential if one wants to ensure that cloud-based big data analytical platform is safe and secure. IAM systems need to be designed to be flexible and scale well to accommodate the more elaborate access mechanisms experienced in big data schema while at the same time enforcing robust security measures.

The regular use of multi-factor authentication (MFA) is becoming more common, although it is advisable to use it to access significant information or to perform any operations. Other identity management tools which are also being implemented in IAM include the use of fingerprint, or face recognition.

The two primary traditional models of permissions for big data are the role-based access control (RBAC) and the attribute-based access control (ABAC) (Kuner et al., 2012).

Two trends have been identified as important in present day IAM for big data cloud systems, which are constant monitoring and adaptive access control. These approaches involve the application of Machine Learning methods to detect user behaviour profile and modify the access rights based on proactive risk analysis. Since intrusions are immediately caught, it assists in preventing the attempts in real-time.

Thus, there is a need for the layered model, incorporating techniques like encryption, secure computation, and blockchain, as well as IAM best practices while organizations keep struggling with security challenges related to cloud-based big data analytics. It focuses on different ways and means of guaranteeing the security of big data with special emphasis made on the trade-off between protection measures and performance and expansion prerequisites of big data analysis processes.

6. Privacy-Preserving Techniques

6.1 Data Anonymization

Data anonymization is the core of the protection from the disclosure of sensitive information in a context of cloud-based big data analytics. It entails the process of stripping or masking PII from datasets in order to exclude any identifiable information of the people in those datasets. The traditional anonymization methods are k-anonymity for microdata, l-diversity, and t-closeness. However, these methods are nowadays more questionable in big data scenarios because of the correlation and inference attacks with respect to data identification.

Namely, in K-anonymity, every record in a dataset must be equally as similar to at least k-1 other records with regards to certain identifiable attributes. Although this affords a topographic degree of privacy it may not be adequate for intricate big data mining where distinct databases are fused to be analysed simultaneously. Scholars are now studying stronger forms of anonymization that are appropriate for big data environments, including aspects like high-dimensionality and continuous streams of data (Kaaniche & Laurent, 2017).

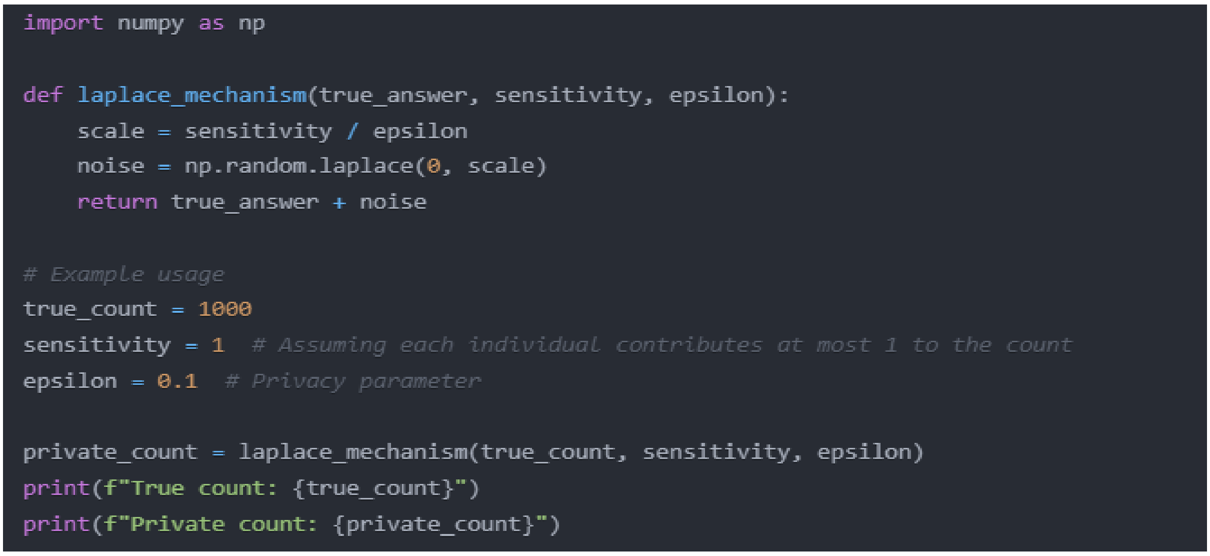
Privacy-preserving data publishing (PPDP) is another developing paradigm, which focuses on the idea of freeing the data so they could be used to peruse analysis and, at the same time, minimize private data revealing. PPDP solutions contain data perturbation, generalization, and suppression procedures determined according to the nature of data and its intended utilization.

6.2 Differential Privacy

Differential privacy has become one of the best mathematical methods used in providing privacy. Enhancing guarantees for databases and machine learning models. It does this in a way of adding noise that is selective, thus while making each individual record not proportional to the outcome, making the results non-singular.

For any environment where the results of computations involving big data have to be shared across a cloud, while at the same time preserving the privacy of individual users, differential privacy is being deployed more often in the context of cloud-based big data analytics in such as the context of sharing aggregate statistics or machine learning models. Most significant cloud services providers have developed differential privacy APIs and more that can be employed in the big data processing pipelines.

Differential privacy carries into effect the privacy level being set by the value of ε which specifies the level of privacy protection to be achieved on the results and utility. The value of ε has to be chosen wisely and usually depends on the nature of the data and the requirements of the given analysis (Jain, Gyanchandani, & Khare, 2016).



```
import numpy as np

def laplace_mechanism(true_answer, sensitivity, epsilon):
    scale = sensitivity / epsilon
    noise = np.random.laplace(0, scale)
    return true_answer + noise

# Example usage
true_count = 1000
sensitivity = 1  # Assuming each individual contributes at most 1 to the count
epsilon = 0.1  # Privacy parameter

private_count = laplace_mechanism(true_count, sensitivity, epsilon)
print(f"True count: {true_count}")
print(f"Private count: {private_count}")
```

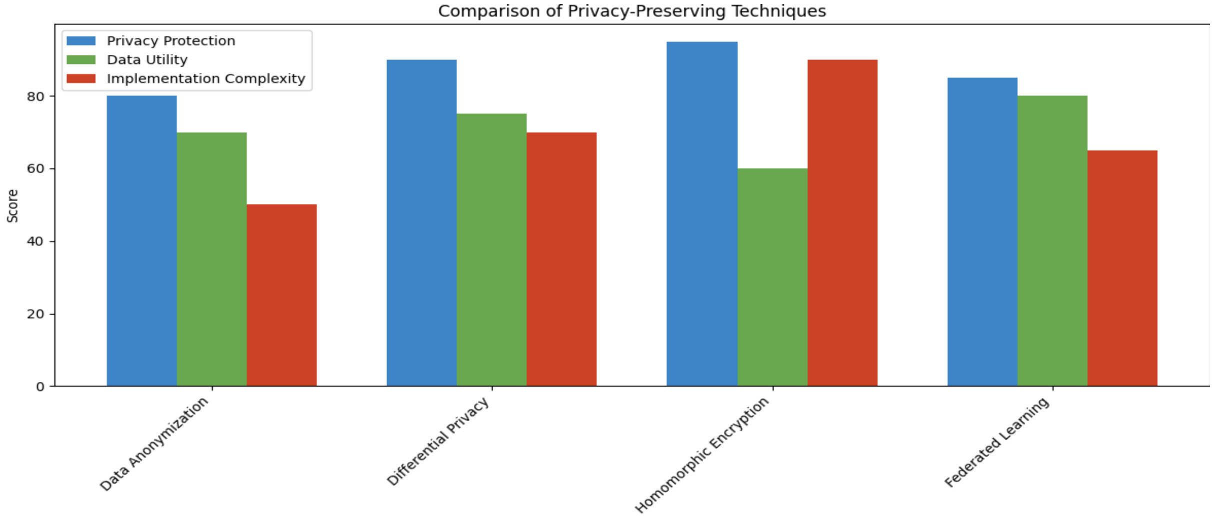
Here is how to perform Laplace noise addition the simplest mechanism of the differential privacy in Python by executing the following code:

6. 3 Privacy-Preserving Data Mining

Privacy-preserving data mining (PPDM) can be defined as a group of methodologies used for data mining with the purpose of maintaining the confidentiality of data. In big data context, especially cloud-based one, PPDM is important for allowing sharing of the large data sets among different stakeholders or for multiple datasets analyse without sharing the sensitive information of the individuals involved.

There are different methods to PPDM and one of them is SMC that enables computation of a function on correlated inputs of multiple parties while the inputs are kept private. For instance, in market share analysis, rival firms can combine their assessment without presenting their clients' data. Nonetheless, SMC is computationally expensive and the current research is directed towards improving these protocols for the big data regimes.

Another technique which is used in PPDM is Data perturbation where the original data is somehow distorted in order to provide better analysis. You can make alterations at the input level I. e., data manipulation or at the output level i.e., result manipulation. It hinges upon the trade-off between the amount of disturbance that has to be introduced to maintain privacy on one hand and the analysis that needs to be conducted on the other.



6. 4 Federated Learning

Federated learning is one of the contemporary approaches in machine learning which enables the models to train the data without collating the data. This approach is most applicable to cloud big data analytics situations where due to legal constraints or business competitiveness one is unable to consolidate the raw data.

In FL, every participant updates his local model using the data set then only model updates are sent to a central server but not data. The server then uses the importance values to build an update to enhance a global model. This is done consecutively until the model change is convergence or until the rounds specified reaches its limit.

Almost all these privacy issues are resolved by federated learning in that while the data remains with its owner, only the parameters of the model are shared. However, it also poses a novel set of difficulties, including non-IID data over participants and information leakage when updating the model (Hashem et al., 2015).

7. Regulatory Compliance and Legal Considerations

7.1 GDPR and Cloud-Based Analytics

The regulations of big data cloud analytics have changed through the emergence of the General Data Protection Regulation also known as the GDPR where it affects many organizations that deal with data of the European Union citizens. GDPR contains some of the most rigorous rules towards data protection and individual's rights to privacy, therefore, the application of analytics should be approached with caution.

The GDPR principles that apply to cloud-based analytics are data minimization, purpose limitation, and storage limitation. The data must also be collected and processed to a minimum extent possible, for the functions that are required and must not be stored for a longer period than necessary. This sometimes may require policies in areas such as data lifecycle management and automated data deletion in cloud space.

The first three rights of the GDPR notably are the rights of the data subject, for example, the right of access, right to rectification and right to erasure also referred to as the right to be forgotten raise a lot of challenges when it comes to big data environments. Cloud solutions need to be capable of identifying and updating or removing specific persons' details from distributed storage and processing infrastructures. This requirement has driven innovation of complex data cataloguing and lineage tracking solutions that are connected to cloud solutions.

7. 2 HIPAA compliance in health care information

The HIPAA in the United States has strict guidelines regarding the processing of PHI in cloud solutions. When working with big data and healthcare information, organizations need to address the HIPAA requirements before starting with the implementation of cloud-based big data analytics.

HIPAA also mandates protective safety measures, such as data encryption both at the during storage and transmission, user authentication, and logging. HIPAA compliant cloud services can be easily accessed due to service providers; however, compliance is usually the duty of the health care organization (Garg & Bawa, 2020).

Perhaps one of the biggest problems of utilizing big data analytics while adhering to the provisions of HIPAA, is the fact that there is tension between more utility and more privacy. There are commonly used procedures like de-identification and data masking which enables the analysis while keeping the privacy of the individuals involved. However, the re-identification risk in big data scenarios becomes a worthy studying issue and has to be investigated systematically.

7. 3 Industry-Specific Regulations

Different fields have their individual policies that affect the operation of cloud based big data analytics. For instance, industries in the financial services field have special legislation and standards like GLBA in the USA and PCI DSS around the world.

These types of regulations are normally very specific on issues to do with the protection of data, the access control of the data and the records of audit. They need to identify these requirements and align them with the cloud analytical setting, which, at times, may involve the use of the cloud provider security solutions and other tools and services acquired from third-party vendors (Elgendy & Elragal, 2014).

For instance, PCI DSS regulations and standards to protect cardholder data include designing controls on the data, where such controls include encryption, changes in access and the general security of any data, including security assessment. While the above controls are transparent and efficient in a centralised environment, there is may be a lot of complexity involved in applying these in a distributed big data architectural style and it calls for control architectures that are monitored in real time.

7. 4 Entry and Transfer Regulations

Cloud computing is defined by its ability to serve customers internationally and hence the issue of cross- border data transfer complicates the issue. Other laws like GDPR limit the transfer of the data of the individuals to third countries which are outside the EEA and do not offer sufficient protection to personal data.

To summarize, it can be stated that organizations utilizing the services of global cloud providers for big data analytics must be cautious about geographical location of data. It can be the utilization of the regional cloud services or the application of the data residency laws to meet the legal requirements on the protection of information.

The latter has emerged due to the nullification of EU-US Privacy Shield framework in the year 2020 (Schrems II decision) that has made the international data transfers even more challenging especially in the case of data analytics encompassing flow of data between EU and US. More measures including SCCs, must be endorsed while a comprehensive evaluation of the recipient country's data protection laws has become mandatory (Dwork, 2011).

Table 3: Key Regulatory Considerations for Cloud-Based Big Data Analytics

Regulation	Key Requirements	Impact on Analytics
GDPR	Data minimization, Purpose limitation, Data subject rights	Need for granular data control and lifecycle management
HIPAA	PHI protection, Access controls, Audit logging	Strict security measures for healthcare data analytics
PCI DSS	Cardholder data protection, Regular security assessments	Enhanced security for financial data processing
CCPA	Consumer privacy rights, Opt-out of data sales	Increased transparency and consumer control in data analytics

Currently, the compliance regime of cloud based big data analytics is structured in such a manner that calls for an inclusive compliance approach that is embedded into the formation and functioning of big data analytics system. This often involves:

1. Holding these assessments on a regular basis in order to evaluate the potential inconveniences having a negative impact on the protection of data.

2. Integrating the privacy by design techniques when developing the analytics solutions.

3. Developing sound protocols of the use of data.

4. Conducting awareness sessions with the staff at least once a year on issues relating to the protection of data and privacy.

5. Keeping records of all the compliance activities and having periodical checkups on the ward.

Thus, organizations are expected to remain abreast of the emerging regulatory structures regarding data protection laws, as well as apprise them to the risks that may impede performance of cloud-based analytics. It might also mean incorporating legal consultants, privacy officers and CSPs in ongoing compliance and risk management.

9. Challenges and Open Issues

9. 1 Proactivity And: Security, Privacy, And Utility

Another evidence from a cloud based big data analytics is that the major concern has been to balance security and privacy of information together with usefulness. To ensure that there is no leakage of the information contained and at the same time meet the legal requirements, the organization must implement effective security measures and strict regulations. However, these measures also reins-in some of the analytic capability that makes big data attractive in the first place.

For instance, data encryption is desirable for data confidentiality coverage, yet, it imposes serious constraints on data handling and restricts some analysis types. Like with other forms of data de-identification, there is the potential of compromising the quality of the data and thus the analytical results for the sake of individual privacy (Chen & Zhao, 2012).

These are trade-offs that organizations need to make and come up with ways that will ensure the data is utilizable while at the same time being secure and private enough. This may include the approach that uses specific technological features like the privacy-preserving data mining techniques and the guidelines involving the organizational functions like data governance frames or the ethical boards for the analytics projects.

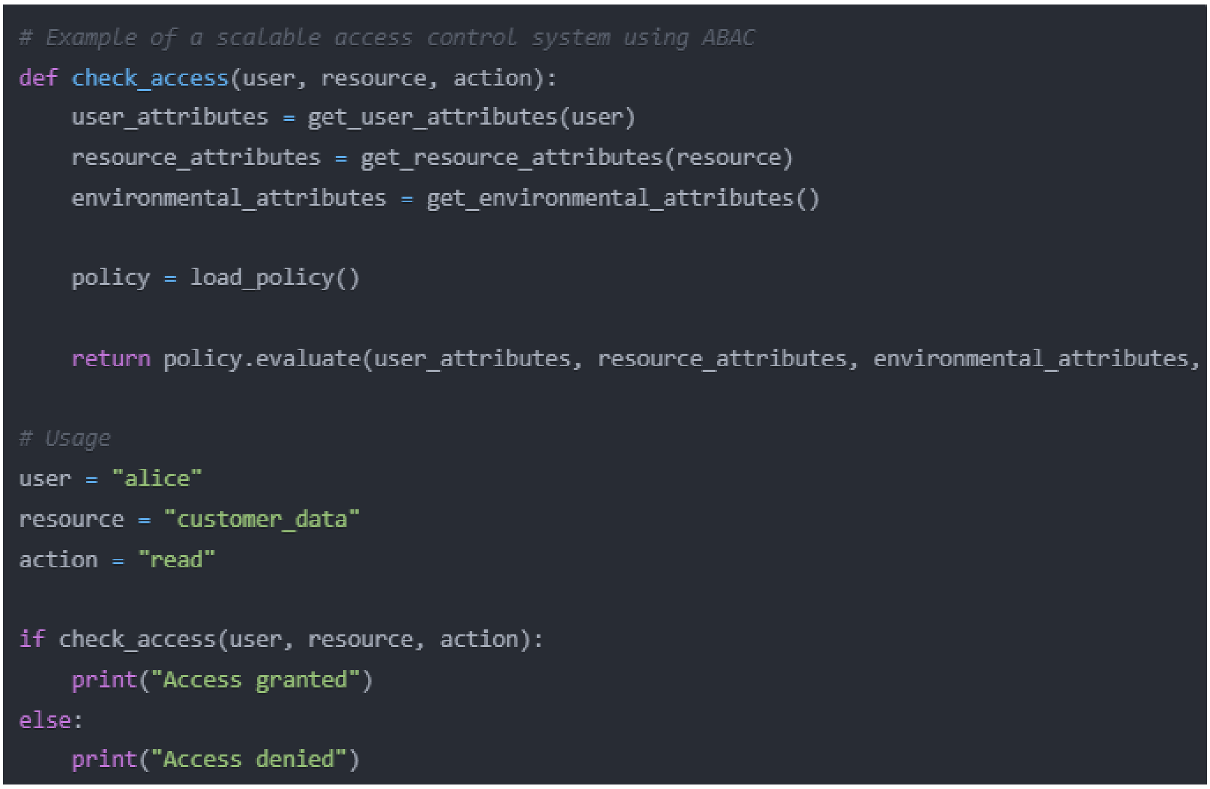
Moreover, the idea of the 'privacy loss budget' is appearing more often, especially in connection with differential privacy. This approach specifies the privacy loss related to any analysis of data or answering any query and then imposes some maximum tolerable risk. However, the creation of such a system for a modern, diverse environment with a large number of distributed big data sources remains a difficult task.

9. 2 Solving the Problem of Security Solutions' Scalability

With the expansion of the amount, speed and heterogeneity of data, the problem of the scalability of the security solutions becomes critical. Most of the conventional security mechanisms that are associated with cybersecurity, from data encryption to access control, were not created with the sampling scale of big data analysis in mind.

Encryption for instance increases large amounts of computation especially when applied on large data sets. This can result to performance hindrances when analysing large volumes of data that require real-time or near real time analysis. To counter this Researchers have developed what we call lightweight encryption algorithms and selective encryption that encrypts only a hard-coded list of heaviest weights only.

Likewise, the handling of the access control policies requires scalability particularly in today's dynamic cloud settings where resources are created and decommissioned frequently. Extensive implementation of RBAC often proves inefficient work with extensive numbers of users and objects, hence the consideration of more versatile access models like ABAC.



```
# Example of a scalable access control system using ABAC
def check_access(user, resource, action):
    user_attributes = get_user_attributes(user)
    resource_attributes = get_resource_attributes(resource)
    environmental_attributes = get_environmental_attributes()

    policy = load_policy()

    return policy.evaluate(user_attributes, resource_attributes, environmental_attributes,

# Usage
user = "alice"
resource = "customer_data"
action = "read"

if check_access(user, resource, action):
    print("Access granted")
else:
    print("Access denied")
```

This code snippet illustrates a basic framework for an ABAC system, which can be more scalable and flexible than traditional RBAC for large-scale cloud environments.

Another scalability challenge lies in security monitoring and threat detection. As the volume of log data and network traffic grows, traditional security information and event management (SIEM) systems may struggle to keep up (Bertino & Ferrari, 2018). This has led to the development of big data security analytics platforms that leverage distributed computing and machine learning to process and analyse vast amounts of security-related data in real-time.

9.3 Human Factors and Security Awareness

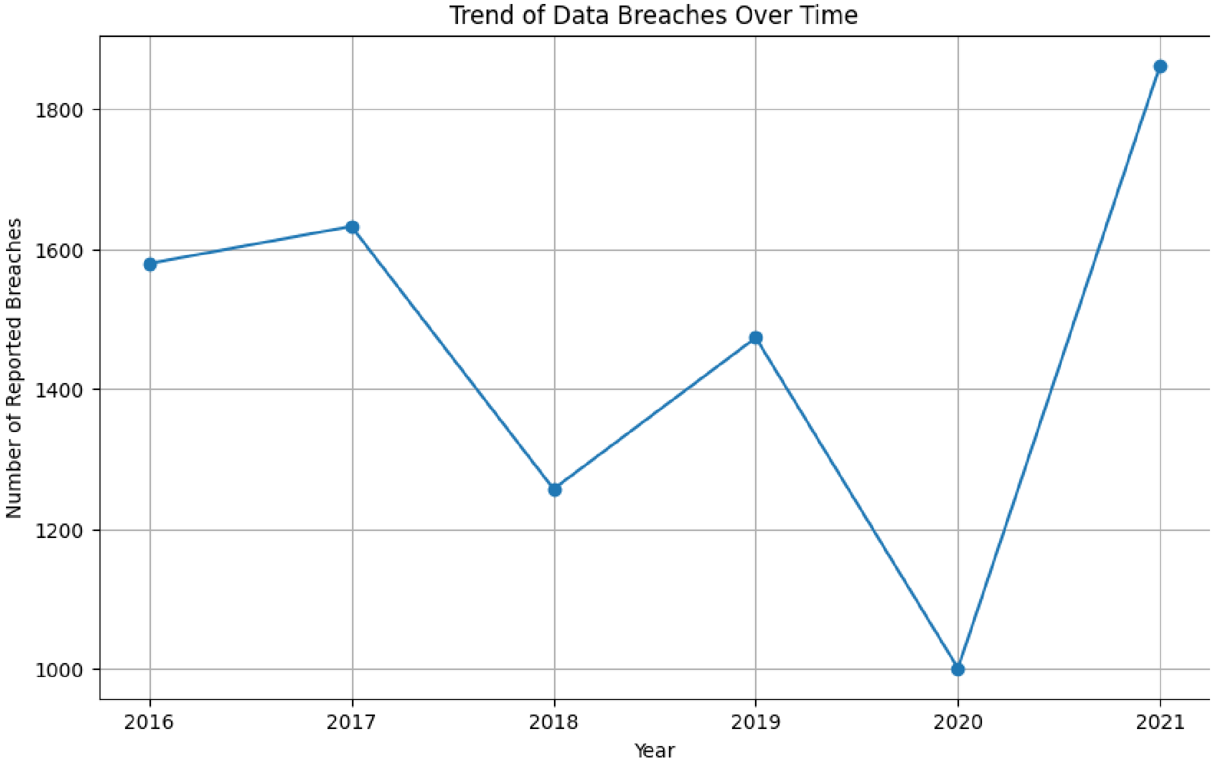
Despite the focus of technical means, the personnel factor has not lost its relevance in the protection of cloud-based big data analytics systems. Inadequate security measures Such issues as wrong configuration of the cloud services or weak passwords are some of the primary causes of data breaches.

Addressing this challenge requires a multi-faceted approach:

1. Training and Awareness: It means that all staff members who operate with the data or manage the cloud resources have to complete security awareness training frequently and in depth. This training should include fundamental measures of security but also the threats bound to big data analytics in the clouds.
2. Usable Security: The approach to security implementations should be user centred. If security controls are too cumbersome or complex, users may look for a way to start avoiding them and may bring in other problems. Security interfaces and related operational workflows should and must be based on the principles of user cantered design.
3. Security Culture: Entities have to acknowledge and promote security awareness where each user is tasked with the role of security. These include but are not limited to setting policies from the top management level while at the same time supporting more local initiatives for security and promoting security-oriented solutions with bonuses.

4. Incident Response Preparedness: Still the worst can happen irrespective of how much preventive measures have been taken. Every employee must know incident response strategies and practice how they execute them by frequently conducting exercises.

5. Ethical Considerations: Thus, the more effective big data analytics becomes, employees must consider the ethical consequences of their work. This is to include, for example, awareness about biases in algorithms, protection of data and social responsibilities of their analytics projects.



As for the discussed human factors, it is critical to note that their improvement is a never-ending process of inducing changes and advancements (Armbrust et al., 2010). Three novel technological aspects referring to human-centric security in context of cloud-based big data analytics have been discussed in this paper, and how they can be approached by responding to the novel technological trends.

10. Conclusion

10. 1 Results of the Study

Security and privacy issues of cloud based big data analytics unravelled a topic that is full of opportunities and challenges, thus needing further research. Given that big data is growing and organizations rely on cloud services for this purpose, security measures and privacy-preserving techniques are crucial.

Key findings from this research include:

1. The combination of cloud computing and big data analysis is very challenging because it brings specific security issues such as data confidentiality when data is distributed among multiple tenants, integrity of data in distributed applications, and the issue of access control.

2. Security of data is very important especially when data privacy laws such as GDPR and CCPA are in place. People's privacy is one of the critical concerns of AI; thus, organizations need to adopt measures like data masking techniques, differential privacy, and federated learning.

3. Advanced Trends Part new security measures and new threats appear in the perspective of the security industry elaborated from the subsequent key emerging technologies: quantum computing, AI security, and edge computing.

4. Challenges of compliance are still a major factor, as organizations are required to deal with a myriad of international laws, and legal requirements of many industries regarding data privacy and protection.
5. The role of people is still a centrally significant component of security: there is a need in extensive and efficient training, easy-to-understand, and practical security arrangements, as well as the development of a security-conscious culture.

10. 2 Recommendations for Practitioners

Based on these findings, the following recommendations are proposed for practitioners in the field of cloud-based big data analytics:

1. Both entities must adopt the following strategies to promote cloud security of big data: This should consist of strong measures for data storage, transmission, account controls and data monitoring along with analytical security tools.
2. Ensure that privacy-preserving analytic techniques become the normal way of carrying out data analysis. This may require the use of differential privacy techniques, investigate secure MP computing for big data analysis, and utilize federated learning when possible.
3. Monitor the appearances of new technologies and their effects on security and personal information confidentiality. Start preparing for the post-quantum cryptography and start researching the method to incorporate AI-based security technologies.
4. Introduce efficient standards of data governance compliance with regulations and requirements of a given field. This should cover commitments on data collection, processing, storage, and erasure as well as robust procedures to meet data subjects' requests.
5. It is recommended to ensure that employees receive training in security procedures and make them recognize the importance of data security in the organization. Make sure that each and every employee not just realizes that they have information that should not get out but they also know how to protect this information.
6. Carry out security audits and/or privacy impact assessments at reasonable intervals in order to establish possible risks to your big data analytics that are based on cloud environments.
7. Become a part of organizations that link up members who offer cloud services and solutions, security solutions providers and other industry players in order to share with the rest regarding new security threats and how to overcome them.

10. 3 Future Research Directions

As the field of cloud-based big data analytics continues to evolve, several areas warrant further research:

1. Efficient and flexible implementations of encryption functions that can offer high levels of security protection in big data frameworks which do not heavily affect the big data analysis procedures.
2. Solutions to machine learning that protect the privacy of users' data and are suitable for functioning in distributed cloud systems.
3. The proposal of new quantum-resistant cryptographic algorithms that can be applied to big data and the procedures of migrating from current systems to the novel ones.
4. Techniques to introduce better ways to provide access control and identity that are efficient enough to accommodate for the big data environments without compromising on its usability.
5. Developing methods for more accurate and efficient identification of data leaks in the cloud, especially in cases of unauthorized employees or stolen access codes.
6. Ethical guidelines in decision making systems including biases, fairness, and other possibilities of big data analytics.

7. Strategies for preserving the big data in cloud environments for long times as well as the future changes in data security and formats of stored big data.

References

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.

[2] Bertino, E., & Ferrari, E. (2018). Big data security and privacy. In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years (pp. 425-439). Springer, Cham.

[3] Chen, D., & Zhao, H. (2012, March). Data security and privacy protection issues in cloud computing. In 2012 International Conference on Computer Science and Electronics Engineering (Vol. 1, pp. 647-651). IEEE.

[4] Dwork, C. (2011). Differential privacy. Encyclopaedia of Cryptography and Security, 338-340.

[5] Elgendy, N., & Elragal, A. (2014). Big data analytics: a literature review paper. In Industrial Conference on Data Mining (pp. 214-227). Springer, Cham.

[6] Garg, N., & Bawa, S. (2020). Privacy preserving data mining techniques in big data: a comprehensive review. Multimedia Tools and Applications, 79(39), 28597-28623.

[7] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information systems, 47, 98-115.

[8] Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. Journal of Big Data, 3(1), 25.

[9] Kaaniche, N., & Laurent, M. (2017). Data security and privacy preservation in cloud storage environments based on cryptographic mechanisms. Computer Communications, 111, 120-141.

[10] Kuner, C., Cate, F. H., Millard, C., Svantesson, D. J. B., & Lynskey, O. (2012). The challenge of 'big data' for data protection. International Data Privacy Law, 2(2), 47-49.

[11] Li, Y., Gai, K., Qiu, L., Qiu, M., & Zhao, H. (2017). Intelligent cryptography approach for secure distributed big data storage in cloud computing. Information Sciences, 387, 103-115.

[12] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. IEEE access, 4, 1821-1834.

[13] Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. National Institute of Standards and Technology.

[14] Sarkar, B. K. (2020). Big data for secure healthcare system: A conceptual design. Complex & Intelligent Systems, 7(1), 35-51.

[15] Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In 2013 international conference on collaboration technologies and systems (CTS) (pp. 42-47). IEEE.

[16] Sookhak, M., Gani, A., Khan, M. K., & Buyya, R. (2017). Dynamic remote data auditing for securing big data storage in cloud computing. Information Sciences, 380, 101-116.

[17] Terzi, D. S., Terzi, R., & Sagiroglu, S. (2015). A survey on security and privacy issues in big data. In 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST) (pp. 202-207). IEEE.

[18] Wang, C., Chow, S. S., Wang, Q., Ren, K., & Lou, W. (2013). Privacy-preserving public auditing for secure cloud storage. IEEE transactions on computers, 62(2), 362-375.

[19] Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. IEEE Access, 2, 1149-1176.

[20] Zhang, X., Yang, L. T., Liu, C., & Chen, J. (2014). A scalable two-phase top-down specialization approach for data anonymization using MapReduce on cloud. IEEE Transactions on Parallel and Distributed Systems, 25(2), 363-373.

[21] Kaur, J., Choppadandi, A., Chenchala, P. K., Nakra, V., & Pandian, P. K. G. (2019). AI Applications in Smart Cities: Experiences from Deploying ML Algorithms for Urban Planning and Resource Optimization. Tuijin Jishu/Journal of Propulsion Technology, 40(4), 50-56.

[22] Case Studies on Improving User Interaction and Satisfaction using AI-Enabled Chatbots for Customer Service . (2019). International Journal of Transcontinental Discoveries, ISSN: 3006-628X, 6(1), 29-34. https://internationaljournals.org/index.php/ijtd/article/view/98

[23] Kaur, J., Choppadandi, A., Chenchala, P. K., Nakra, V., & Pandian, P. K. G. (2019). Case Studies on Improving User Interaction and Satisfaction using AI-Enabled Chatbots for Customer Service. International Journal

[24] ofTranscontinental Discoveries, 6(1), 29-34. https://internationaljournals.org/index.php/ijtd/article/view/98

[25] Choppadandi, A., Kaur, J., Chenchala, P. K., Kanungo, S., & Pandian, P. K. K. G. (2019). AI-Driven Customer Relationship Management in PK Salon Management System. International Journal of Open Publication and Exploration, 7(2), 28-35. https://ijope.com/index.php/home/article/view/128

[26] AI-Driven Customer Relationship Management in PK Salon Management System. (2019). International Journal of Open Publication and Exploration, ISSN: 3006-2853, 7(2), 28-35. https://ijope.com/index.php/home/article/view/128

[27] Big Data Analytics using Machine Learning Techniques on Cloud Platforms. (2019). International Journal of Business Management and Visuals, ISSN: 3006-2705, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76

[28] Shah, J., Prasad, N., Narukulla, N., Hajari, V. R., & Paripati, L. (2019). Big Data Analytics using Machine Learning Techniques on Cloud Platforms. International Journal of Business Management and Visuals, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76