ISSN: 0974-5823 Vol. 7 No. 5 May, 2022

International Journal of Mechanical Engineering

Trust Prediction Tree: Use of Decision Tree in Cloud Scenarios

Dr Archana B. Saxena, Jagan Institute of Management Studies, Delhi

Dr Deepti Sharma, Jagan Institute of Management Studies, Delhi

Dr Deepshikha Aggarwal, Jagan Institute of Management Studies, Delhi

Abstract: A continuous and steep increase has been noticed in the cybersecurity budget allocation every year. Even though, the data theft and data breach cases are regular entries in the cyber world periodic. It has been noticed through data statistics that to a certain extend data breach liability relay on cloud provider as well. This research piece is working towards this direction. In current work researchers have worked on aspects to make use of predictive tools in the current scenario. There focus is on options where already available information can be used to predict future. This paper is commissioning the same idea, to predict about the trust value of the provider on the basis of already available information. Decision tree is used as a tool for trust prediction. Among the various prediction algorithms Decision Tree is considered as very simple and useful tool in supervised learning technique. The ease of implementation is the one aspect but the best part is interpretation of the results from the tree that can be derive in form of rules. Among the various algorithms to implement tree, ID3 is applied in current research work.

Keywords: Prediction, Classification, Decision Tree, Trust Prediction, Cloud, Cloud Provider, Machine Learning Algorithm, Certification Attainment Status.

1. **Introduction and Related Work**: A close look at the cloud usage statistics gives an immense pleasure that public cloud service market has worth of $623.3 billion in 2020. Now more than 80% of an organizations workload is dealt through various cloud services [1]. As the cloud adopted among the masses, so there is an increase in the cloud concerns like: Security breach, data theft or leakage, and trust and privacy related concerns. To over these issues researchers are working in this direction and monetary allocations are done every year to overcome such hassles. In the recent years one can register a perceptible change in cyber security budget. Cybersecurity market was $3.5 billion in 2004 and with many fold increased its now at the position of $1trillion in year 2021 [2] [3] [4]. Along with an increase in the Cybersecurity expenditure there is an increase in the cyber-crime graph as well [5]. Now a major discussion in this aspect is regarding who is liable for these security breach cases. There are major three stakeholders in public cloud environment: Cloud Service Provider, Business utilizing the cloud service, and customers of that business [6]. All of them have contribution in making the complete cloud process a secure environment to operate. When any breach occurs over the cloud they all suffer monetary terms or goodwill aspects. To overcome this loss they all should work in coordination and do their best to ensure secure operations. If cloud providers aspect is considered, it should offer all possible secure means in terms of security tools, safe transactions, reliable admin staff and compliance with the government or regulatory bodies. This way it can assure it consumers a safe

and storage of data and operations [7]. The business utilizing the cloud services has to play its part in verifying security and concerned aspects about the service provider. The current research piece is contributing in this aspect. Business utilizing the cloud service should carefully check the back ground and trust factor of the cloud provider. If the trust factor is meeting the satisfaction threshold then business should proceed and save its client/ consumers information or data over the internet.

2. **Problem Description and Solution Proposed:** The current cloud literature is flooded with “Cloud Trust Models”. These trust models are generating trust value but there is no further direction how to use these values. In the current work authors are trying to use these trust values [generated through trust model] for trust prediction by implementing prediction algorithm.

Multiple predictive modeling techniques like statistical Analysis, Machine Learning are available that can predict on the basis of already available data set. Sinking with current topic, authors are using ML technique to predict a trust value for a provider. Concerning this context: Collected data set [section 4.1] and implemented Algorithm [section3.3] are explained in the next section.

3. **Machine Learning Algorithm Implemented:** This paper is implementing**,** Machine Learning (ML) technique of predictive modeling. Algorithm used in this work carry out Classification technique of Machine Learning. Classification algorithms are two step process:

3.1. Learning step: Learning step uses training data and develop a Model.

3.2. Prediction step: Prediction step, uses that model and make predictions about unknown dataset. Current paper is using decision tree classification algorithm and its implementation on provider’s certification related dataset.

3.3. **Decision Tree**: Decision tree algorithms build tree or flow chart like structure that follows a top down approach. We start from a root or root attribute and moves down till leaf node and produces a rule that can be used to make predictions about unknown dataset.

There are different algorithms Like C4.5, ID3, CART that can used to build tree structure, this paper is using ID3 that is an extension of D3. ID3 uses Entropy and Information at each level since the root till the leaf node. The attribute used in this "Trust Prediction Tree" are: Security, Governance, SLA (Service Level Agreement) and Audit. For model training, dataset is collected about the providers from the website. More details about the application and dataset are explained in the next section.

3.4. **Why Decision Tree**: Numerous Classification Machine Learning algorithms are available, authors have decided to use Decision tree because:

3.4.1. It requires less efforts in data pre-processing like removing missing value, normalization of data and many more.

3.4.2. It works with both categorical and continuous variables.

3.4.3. As name suggests it reflects tree like structure and rules can be retrieved from the structure that can be easily interpreted by stakeholders.

3.4.4. It takes less time and can be easily implemented on large datasets. Although it has over fitting, under fitting and tree pruning issues but they are not affecting much in the problem concern.

4. Model Specification and Dataset Collection:

4.1. Dataset collection

4.1.1. **Data Collection**: Decision tree is predictive modelling technique where known data set with class label is required to train the model. To train the current trust model: Trust aspects and contribution is calculated through a survey. Survey is conducted among free and paid cloud consumers and their responses are recorder and analyzed in the excel tool. Table1 lists the concluded result which indicates security has the major share in trust building. Other important factors that are considered by the consumers are: Governance, Audit and SLA [8].

Table1: percentage contribution of various components in trust

Component	% contribution in Trust
Security	57%
Governance	21%
SLA (Service Level Agreement)	10%
Audit	10%
Diverse	2%

4.1.2. **Cloud Certifications**: Cloud is an IT (Information Technology) based industry. It is regulated through standards and certifications. Various advocacy groups like CSCC, CSA, CSCC, NSIT, OCC, DMTF, CSIG, STAR and CSIG works for the betterment of the cloud. Different organizations are operational in different field. To ensure the smooth execution of the cloud industry, these bodies recommend certifications to ensure various aspects. The certifications recommended by CSCC and CSIG are used in this paper to analyze the status of cloud provider. Table2 lists the various certifications recommended by the CSIG and CSCC concerning the components mentioned in table1. This work is an extension of authors already executed and published

model to evaluate cloud providers trust value based on above mention parameter and their respective certification attainment by the provider [9].

Table 2 lists the certifications that are recommended against above mentioned aspects.

Components	Security	Governance	SLA	Audit
List of Recommended standards & Certifications	ISO/IEC 27001	ISO/IEC 27014	ISO/IEC 27004	C
	ISO/IEC 27002	ISO/IEC 38500	ISO/IEC 27002	ISO/IEC 27007
	FIPS 140-2	ISO 20000	ISO/IEC 19086 Part I	ISO/IEC 27008
	ISO/IEC 27017	COBIT 5	ISO/IEC 19086 Part II	STAR
	ISO/IEC 27040	ITIL v3	ISo/IEC 19086 Part IV	SSAE
	HTTPS/SFTP/VPN	SSAE 16		
	LDAP	ISO/IEC 27005		
	SAML 2.0	NSIT 800-173		
	OAuth 2.0	ISO/IEC 19086		
	ISO/IEC 29134			
	ISO/IEC 27018			
	ISO/IEC 29100			
	ISO/IEC 27033			
	NIST SP 800-53			
	SC7-SC8			
	FIPS99 & FIPS200			
	ISO/IEC 27034			

On the basis of the certifications mentioned in table2, data set is prepared by checking the status of each provider against each certification recommended by CSIG and CSCC. In the information reduction process these certifications are further scale to components level. The final attributes used in this dataset are: Security, Governance, SLA and Audit. The sample record-set is given in the table 3. The dataset mentioned in the table3 is used to train the model of decision tree.

Table3: Dataset components after information reduction and aggregation.

Company	security	Governance	SLA	Audit
Provider1	12	0	0	2
Provider2	10	1	0	0
Provider3	8	0	0	3
Provider4	10	0	0	3

4.2. **Trust Model:** The trust model that is used in current scenario collects all required inputs and compute trust value also known as OTF (Overall Trust Fact or) through a compute engine. The complete utilization of collected data can be understood from figure 1. The complete model and its working is explain by authors in their earlier publications [9].

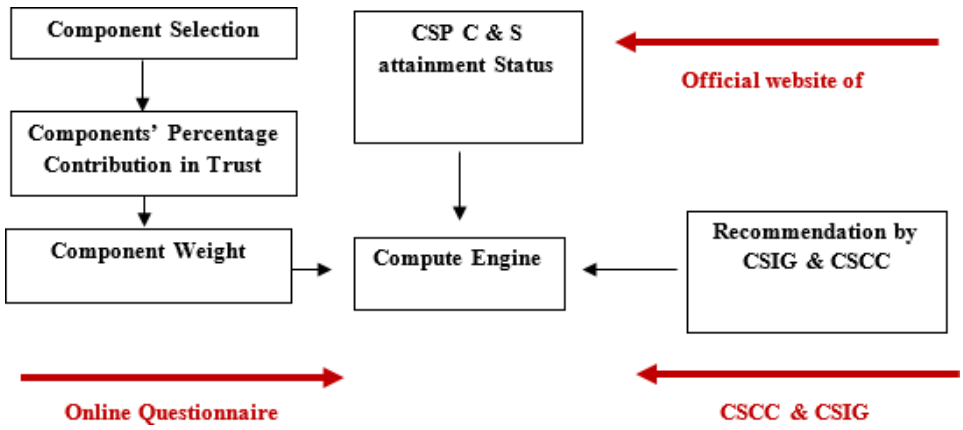

Figure 1: Execution process of Compute Engine for evaluating OTF (Overall Trust Factor)

Using the above trust model and collected dataset, "Trust Prediction Tree" is built to get and quantitative value for a cloud provider.

5. Trust Prediction Tree:

5.1. **Application Description:** In this paper authors would like to analyze the "Trust Value" of the cloud provider. The value is generated on the basis of the aspects like Security, Governance, SLA (Service Level agreement) and audit. To analyze the competence of provider in these aspects, Certification attainment status of the provider is checked.

- **Applied Tool**: Once the dataset collection is complete, dataset is concerted into

.csv file to be used with "Anaconda Tool". Anaconda is very good tool that offers very good packages for data analysis, Machine Learning and Predictive analysis. Although it supports both R and python but primarily python commands are used to complete this research work.

Python commands Window

```
Import panadas as pd

Data=pd.read_csv("providers.csv")
```

Figure 1: Code Window1, Python command import panadas and load dataset

5.2. **Tree construction:** ID3 decision tree algorithm that is used in algorithm uses two types of variables:

521. **Predictor Variable/Independent Variable/Decisive Attribute:** For this dataset predictor variables are: Security, Governance, SLA and Audit.

522. **Class Label/Dependent Variable/Decision Attribute**: For this dataset class label is "Trust value / Trust".

Once the Independent and dependent variables are decided the next step is to create tree through Anaconda tool using python as language. The dataset file is saved in the .csv format to use it with python code. [Ref: Python Command Window1]

Table4: Sample dataset used for Decision tree model training.

Security_ Value	Governance_ Value	SLA_ Value	Audit_ Value	Trust_Value
2	0	0	1	Low
1	1	0	0	Medium
1	0	0	1	Medium
1	0	0	1	Medium
1	0	0	1	Low
1	0	0	0	Medium
1	1	0	0	Medium
0	0	0	0	High
1	0	0	1	Medium
1	0	0	2	High

ID3 was introduced by J.R.Quinlan. Quinlan has used Entropy and Information Gain in the tree construction process. The steps of tree construction as per Quinlan are:

- Calculate entropy for the dataset, entropy value is calculated for each attribute/feature/independent variable[python command, Reference Figure :]
- Once the dataset is loaded into the python, then commands are used to calculate entropy an Information gain for each independent variable.
- Entropy(label)
- InfoGain (data,split_attribute_name,target_name="class"):
- Calculate information gain for each attribute/feature/independent variable
- The attribute that has "Maximum" information gain will be chosen as root node. [output: Reference, figure:]

```
Information Gain (Security_Value) =  0.8817516073670975    [Maximum]
Information Gain (Governance_Value) = 0.07018755653590647
Information Gain (Audit_Value) = 0.1678445503819601
Information Gain (SLA_Value) = 0.0033633613131536233
```

Figure2: Information gain value of different attributes/independent variable

- Decision tree in current scenario will have security as root node [Ref: figure].

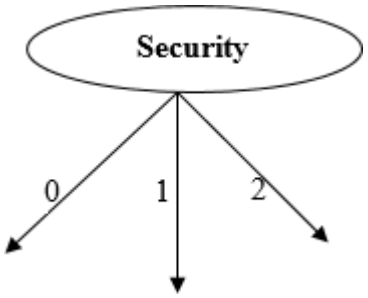

Figure3: Level 0: decision tree: Security with maximum information gain as root node

- Now the complete dataset will be split on basis of security values. Security has three possible values:0,1,2.
- If the the rows of the split dataset has same value

Index	Company	Governance	SLA	Audit	Trust-value
0	Amazon	0	0	1	2
26	Cloud stack	0	1	0	2
27	Kamatera	0	1	0	2
28	CSC	0	0	0	2

Figure 4: **Dataframe split according to Security =2**

Index	Company	Governance	SLA	Audit	Trust-value
0	Amazon	0	0	1	2
1	IBM	1	0	0	2
2	Microsoft	0	0	1	1
3	Google	0	0	1	2

Figure 5: **Dataframe split according to Security =1**

Index	Company	Governance	SLA	Audit
7	Red Hat	0	0	0
10	SAP	1	0	1
11	Verizon	0	0	0
12	Navisite	0	0	1

Figure 6 : Figure 4: **Dataframe split according to Security =0**

- If the target variable (Trust) has same values for the rows in the spitted data set then that value will become the leaf node for that branch.[Ref: Figure 4] [if security=2, trust value is 2].
- If the target variable (Trust) has different values for the rows in the spitted data set then dataset is not a pure dataset and further splitting is required. [Ref: Figure 5,6]
- Splitting is further done by calculation Information gain of remaining attribute/variable/independent variable.

Governance: InfoGain_sec0_Gov : 0.06767468402298782
SLA: InfoGain_sec0_SLA: 0.005230994456215887
Audit: InfoGain_sec0_Audit: 0.306565839128495 [Maximum]

Figure 7: Information gain for rest 3 variables to decide the splitting node for left side of tree.

Governance: InfoGain_sec1_Gov : 0.035889543041326744 [Maximum]
SLA: InfoGain_sec1_SLA: 0.0033011362730746008
Audit: InfoGain_sec1_Audit: 0.0

Figure 8: Information gain for rest 3 variables to decide the splitting node for right side of tree.

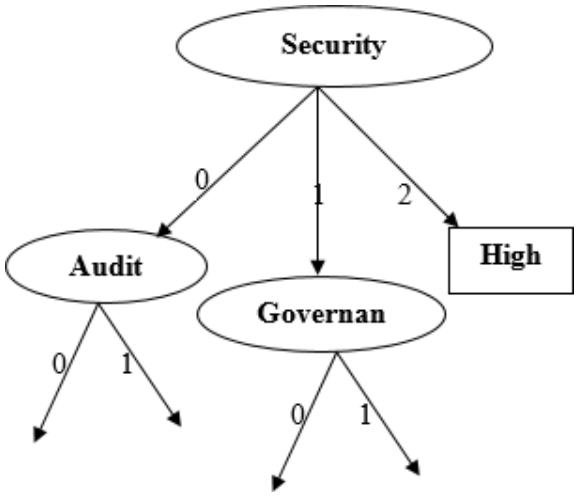

Figure 9: level 1 decision tree

Since among the rest 3 variable, "Audit" has highest information gain (Figure 7) so it the splitting node at this level

Since among the rest 3 variable, "Governance" has highest information gain (Figure 8) so it the splitting node at this level

To complete the tree, now the same process is repeated (Extracting data frames) and then decide the trust value re further split the node.

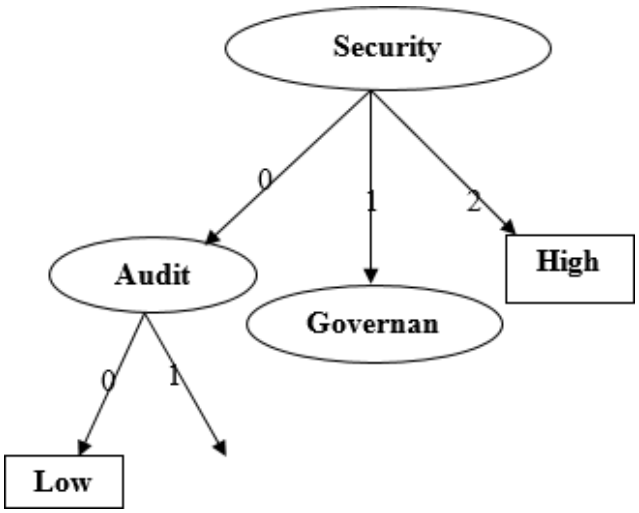

Figure 10: Level 2 decision tree

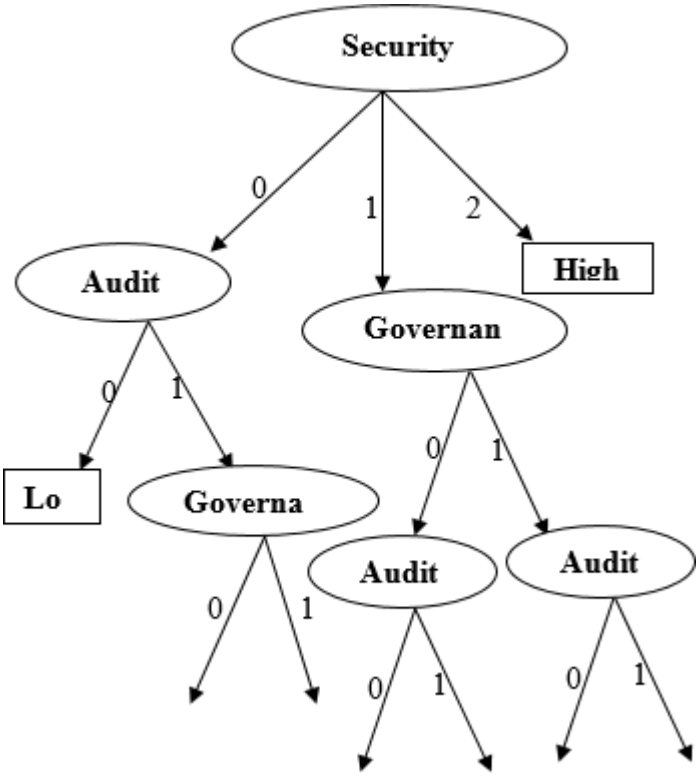

Figure 11: Level 3 decision tree.

- Same process is repeated till all the attributes are place in the tree or the leaf node is generated for each branch [Ref: Figure 11].\

6. **Results:** To check the accuracy of the model and its implementation through algorithm accuracy score and confusion matrix are used in the current paper:

6.1. **Accuracy Score: It g**ives an idea about the correct predictions made by the algorithm. We calculate it using syntax:

Accuracy Score= Correct Predictions /Total number of instances

The present classifier working on provider's data has accuracy score of 87.5% calculated through python command.

6.2. **Confusion Matrix**: confusion matrix also known as error matrix is very significant in statistical; classification related algorithms. This matrix lets you visualize the performance of the algorithm:

N= 24	Predicted LOW	Predicted MEDIUM	Predicted HIGH	The marginal sum of actuals
Actual LOW	9	0	0	9
Actual MEDIUM	3	10	0	13
Actual HIGH	0	0	2	2
The marginal sum of predictions	12	10	2	T=24

7. **Conclusion:** The main objective of this work is to get an authenticated trust value for a provider through information available on its public domain. By collecting the data through public domain, authors managed to get a trust value (OTF) for the provider. The true positive score of this algorithm is more than 80%. We have implemented decision tree through ID3 algorithm. Same decision tree can be implemented by using C4.5, C5.0 and CRT. Same data can be implemented through these algorithms and more accurate predictions can be generated for contribution in "Trust" issue.

References:

[1] N. Galov, "Incredible Cloud Adoption statistics," 9 August 2021. [Online]. Available: https://hostingtribunal.com/blog/cloud-adoption-statistics/#gref. [Accessed 4th March 2022].

[2] S. Morgan, "Cybercrime Magzine," 10 6 2019. [Online]. Available: https://cybersecurityventures.com/cybersecurity-market-report/. [Accessed 15 2 2022].

[3] S. Madan and P. Goswami, "Hybrid Privacy Preservation Model for Big Data," *Publishing on Cloud, IJAIP,* no. 10.1504/IJAIP.2018.10025582.

[4] L. Kharab, "A Comprehensive Study of Security in Cloud Computing," *International Journal of Engineering & Technology*, no. vol 7, no 4, 2018.

[5] . K. Sandhya, "Number of cyber crimes reported across India," 25 10 2021. [Online]. Available: https://www.statista.com/statistics/309435/india-cyber-crime-it-act/. [Accessed 15 2 2022].

[6] D. Kolevski, K. Michael, R. Abbas and M. Freeman, "Stakeholders in the cloud computing value-chain : A socio-technical review of data breach literature," in *International Symposium on Technology and Society (ISTAS)*, 2020.

[7] S. Ramgovind, M. M. Eloff and E. Smith, "The management of security in Cloud computing," in *Information Security for South Africa (ISSA)*, 2010.

[8] A. B. Saxena and M. Dawe, "Consumer's Perception on Cloud Trust: Evaluation Based on Trust Components and Sub Elements," in *Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020*, 2020.

[9] A. B. Saxena, D. Sharma and D. Aggarwal, "Consuming IT Certifications & Customer‟s Preferences: Theoretical Framework To Evaluate," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH,* Vols. VOLUME 10,, no. ISSUE 01, pp. 202-207, 2021.

[10] "2020 Data Breaches," December 2020. [Online]. Available: https://www.identityforce.com/blog/2020-data-breaches#:~:text=May%2020%2C%202020%3A%20Over%2040,advertised%20in%20a%20hacking%20forum..

[11] "10 Advantages of Cloud Computing for Business," [Online]. Available: https://www.salesforce.com/au/blog/2017/06/10-advantages-of-cloud-computing- for-small-businesses.html.

[12] "cloud-computing-history.html," 17 April 2018. [Online]. Available: https://www.javatpoint.com/history-of-cloud-computing. [Accessed 4 May 2019].

[13] "history-of-cloud-computing," 17 April 2018. [Online]. Available: https://www.computerweekly.com/feature/A-history-of-cloud-computing. [Accessed 23 May 2019].

[14] A. B. Saxena, "Contribution of Various Components in Cloud Trust: A Cloud Consumer's Perspective," *ICT for Competitive Strategies*, March 2020.

[15] S. Pearson, "Privacy, Security and Trust in Cloud Computing," in *Privacy and Security for Cloud Computing*, Verlag, London, 2013, pp. 3-42.

[16] P. Li, J. Li, Z. Huang and c. Zhi Gao, "Privacy-preserving outsourced classification in cloud computing," in *Cluster computing*, US, 2018.

[17] A. B. Saxena and M. Dawe, "Cloud Trust: A Key to attain Competitive Advantage," in *ICEBM-2019," International Conference on Evidence-Based Management"*, BITS, Pilani, Rajasthan, 2019.

[18] R. Shaikh and S. M., "Trust framework for calculating security strength of a cloud service," in *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*, Mumbai, India.

[19] M. Alhanahnah, P. Bertok, Z. Tari and S. Alouneh, "Context-Aware Multifaceted Trust Framework For Evaluating Trustworthiness of Cloud Providers," *Future Generation Computer Systems*, pp. 488-499, 2018.

[20] M. Alhamad, T. Dillon and E. Chang, "SLA-Based Trust Model for Cloud Computing," in *International Conference on Network Based Information System* , Takayama, 2010.

[21] P. Manuel, "A Trust Model of Cloud Computing Based on Quality of Services," *Annals of Operation Reserach*, pp. 1-12, 2013.

[22] . K. Johnson and M. Kuhn, Applied Predictive Modeling, New york: Springer, 2016.

[23] T. M. Mitchell, Machine Learning, Online: Springer, 2017.

[24] "Cloud Service Level Agreement Standardisation Guidelines," CSIG Memebers, Brussels, 2014.

[25] A. Ali and . C. Baudoin, "Practical Guide to Cloud Computing Version 3.0," Cloud Standards Customer Council., 2017.

[26] "Who is responsible when cloud security breach occurs," 1st April 2021. [Online]. Available: https://www.dpp-gdpr.com/news/cloud-security-data-breach/. [Accessed 4th March 2022].