Study online security & privacy issues in configuring Cloud resources

Nabeera Hussain[1], Uzair Afzal[2], and Mustafa Alam[3]

[1] 22100146@lums.edu.pk, SBASSE, LUMS
[2] 22100031@lums.edu.pk, SBASSE, LUMS
[3] 19030004@lums.edu.pk, SBASSE, LUMS
[#] Corresponding author; 22100146@lums.edu.pk

Abstract*— While the rapid evolution of cloud computing has accelerated innovation and provided several computing breakthroughs in scalability and efficiency, it has also brought with it new and challenging risks to security and privacy. In this report, we analyze 4 developer discussion platforms: Stack Overflow (SO), Information Security Stack Exchange (IS), AWS Developer Forum, and Azure Community Support to understand challenges and confusions developers face while dealing with online security-related issues in configuring cloud resources. We use ParseHub to scrape AWS and Azure Forums and use SO and IS data dumps to create our datasets and then apply topic modeling techniques to 5563 SO, 739 IS, 372 AWS, and 871 Azure questions related to cloud security. We use Latent Dirichlet Allocation (LDA) to cluster the data into topics and then use those clusters to make further analyses. Identified topics include accessing cloud instances using SSH, security groups, fraudulent certificates in OpenSSL, secure data transmission over the internet (TLS, RSA, AES, etc.), securing APIs as well as Active Directory. In our study, we have identified that developers using AWS and GCP generally tend to use Stack overflow to post their questions and queries. In the IS dataset, the questions are mostly independent of the platform, but topics are relevant to security questions encountered in configuring cloud instances. From the SO dataset we have found Google Cloud Platform (GCP) to be the most popular among developers followed by Amazon AWS.*

***Keywords*— Stack Overflow (SO), Information Security Stack Exchange (IS), AWS Developer Forum, Azure Support Forum, Latent Dirichlet Allocation (LDA)**

I. INTRODUCTION

Cloud computing is rising rapidly in popularity as it evolves to provide new benefits to users, companies, and institutions. This rapid transition towards the clouds has fuelled concerns on a critical issue for the success of information systems, communication, and information security. From a security perspective, a number of uncharted risks have been introduced, deteriorating much of the effectiveness of traditional protection mechanisms [3]. As a result, cloud developers face several security-related challenges when configuring cloud resources. According to a Verizon report, misconfiguration accounts for 70% of all errors in the cloud industry [30]. Developers located in large organizations may have access to dedicated staff with training in cloud security to assist them, but many developers are expected to incorporate security measures while configuring cloud resources on their own. Thus, it becomes pertinent to explore the avenues cloud developers use to overcome security vulnerabilities and the ways in which they interpret and think about security-related configuration issues.

Security and privacy can be challenging for developers to get right, even with the support of tools [6]. Vulnerabilities have been highly correlated with developer errors [23] due to several causes such as APIs with poor developer support [28, 11]. While there is a lot of existing research related to security conversations on Stack Exchange, more particularly, Stack Overflow, there is a dearth of literature that is (i) specific to security challenges in configuring cloud resources and that (ii) analyzes data from discussion platforms provided by popular cloud service providers such as AWS Developer Forum and Azure Community Support.

In this paper, we employ a research methodology that combines techniques from the literature on SO analysis with questions about the security-related tasks of developers. Our research questions are: (1) What topics are covered by cloud security-related questions asked on cloud support platforms? (2) Which topics are most popular? (3) Which platform is the most relevant for discussions revolving around cloud security?

To answer our research questions, we collect questions from four developer discussion platforms: Stack Overflow (SO), Information Security Stack Exchange (IS), AWS Developer Forum, and Azure Community Support via a mixture of web-scraping and utilizing existing data dumps that mention terms related to cloud security in the title or tags and then apply topic modeling. We find that developers ask questions when dealing with key management, access control, encryption, and Schannels.

II. Related Work

A. *Stack Exchange*

Stack Exchange [29] launched in 2008 with Stack Overflow as its first Q&A site for programming and software engineering questions. Stack Exchange continues to grow and adds more specific Q&A sites that cover diverse topics including 70 Technical communities, e.g., Ubuntu, Server Fault, and Information Security.

Stack Overflow [22] (SO) is one of the largest developer Q&A platforms and defines itself as “an open community for anyone that codes.” It has an Alexa rank of 56 [8] and more than 50 million unique visitors per month (as of December 2021) [9]. It attracts a wide range of developers who ask questions about programming, security, and data management. SO’s dataset has been heavily used for research on such topics as: security issues developers face and how they interact and build knowledge around it [24, 3], and the negative impact of SO code snippets in software security [17].

Given the opportunity, it has been shown that developers turn to Stack Overflow to find solutions to security problems, however the code samples taken from security posts may not be as robust or correct as other information sources like books and vendor supplied documentation [26]. Another recent study has investigated other types of on-line sources of guidance about secure software development that are available to software developers, finding that developers must rely on diverse sources of information because there are gaps in coverage [27].

There are mainly five pieces of information associated with the questions: the title of the question, the description of the question, a list of answers, the number of views, and the date when it was posted.

Prior topic modeling of security SO questions found five main categories: web security (51%), system security (19%), cryptography (17%), software security (9%), and mobile security (4%) with popular subjects including: password,hash, signature and SQL injection (out of 30,054 posts) [3]. Such outcomes can help both industry and researchers to understand better the challenges developers are facing. For example, injection (such as SQL, NoSQL, LDAP) and broken authentication such as passwords, keys, and session tokens are the two top risks in OWASP’s ten most critical web application security risks [20], which are similar to the findings of Yang et al. who also studied SO questions [25].

Similar studies have analyzed Stack Overflow and Stack Exchange to understand challenges and confusions developers face while dealing with privacy-related topics [18] as well as blockchain technologies [29]. Information Security Stack Exchange, which is a platform for Information Security professionals has been largely unexplored in literature.

B. Cloud Developer Platforms

In general, a cloud provider supports a developer forum for their client developers to discuss challenges regarding the use of cloud resources. For e.g AWS Developer Forum [14] is a support community for AWS developers to share their development experience with their fellow peers. Similar forums include Azure Support, GCP Support Forum as well as Alibaba Cloud Developer Forum. Typically, developer forums have internal technical experts, employed to answer questions from client developers, to offer a fast and right-to-the-point response to client developers. Much like SO, the metadata associated with a question is its: title, body, answers, number of views, and date posted.

Currently no studies exist that leverage these forums to analyze cloud based discussion. Vinkatesh et al. [21] performed an exploratory analysis on discussions from developer forums and Stack Overflow to understand the dominant questions that developers ask when integrating Web APIs.

Squire [19] shows that client developers tend to prefer Developer Forums, because moderators of Stack Overflow tend to close valid questions as off-topic questions. Hence, we decided to collect discussion from both Developer Forums and Stack Exchange.

C. Cloud Security and Developers

Cloud security has been studied extensively. Cloud data and computing data brings on new security threats and precautions which include the threat to: (i) availability, (ii) integrity, (iii) confidentiality, (iv) multi tenancy, (v) elasticity, (vi) privacy and (vii) non-repudiation [5]. Sinanc and Sagrigoglu discussed cloud computing properties, security issues and security models. They analyzed literature in terms of the popular cloud security subjects due to the relevance of cloud users’ and reviewers’ in order to develop more available and manageable cloud security systems. They concluded that the risks and challenges in cloud computing can be easily overcome with making clear agreement between stakeholders and taking precautions before using a cloud computing solution.

Yan et al. [15] surveyed potential risks on cloud security and privacy including communication security, abuse and use of cloud computing resources, virtualisation and multi-tenant, availability, insecure interface APIs, authentication and authorization, data encryption, data security and keys loss. While existing solutions do exist for cryptographic cloud storage, new cloud architectures and identity based management schemes, further problems to be addressed are on standards, cloud resource monitoring, secure communication channels, identity management, access control, and encryption.

D. Latent Dirichlet Allocation (LDA)

LDA [4] is an unsupervised method of topic modeling where topics are not labeled by humans but are discovered naturally through patterns of clustering in the data. For example, LDA might discover that documents fall into two topics, one in which typical words include (key, cipher, public), and another in which these common words are (‘election’,’ vote’,’ parliament’). A human annotator is needed to label these topics as “encryption” and “politics”, as the model does not assign labels. Vocabulary is not exclusive to a single topic but has different distributions according to the topic. LDA models text generation as a two-step process: first, a mixture of topics is sampled through the Dirichlet distribution, then a mixture of vocabulary items is sampled from the Dirichlet distribution associated with each topic. The model assumes that the words in a document are sampled by selecting a topic from the mixture of topics and a word from the mixture of words associated with that topic. We interpret these topics by inspecting the words most indicative of each topic. We take advantage of this automation to analyze a larger dataset than is feasible with human annotation. LDA is commonly used to find topics in SO questions [2, 25]. We use a python package pyLDAvis [7] which extracts information from a fitted LDA topic model to inform an interactive web-based visualization to aid us in this process. The package offers a novel measure called relevance, by which terms within topics can be ranked to aid in the task of topic interpretation.

III. Contributions

Through this paper we make the following contributions:

(i) Perform the first large scale study of security-related questions on cloud developer platforms AWS and Azure

(ii) Identify the topics covered by cloud security-related questions asked on developer support platforms

(iii) Identify the most popular support forum for cloud security-related discussions

(iv) Provide information to develop guides and manuals for common security and privacy issues in configuring cloud resources

IV. Datasets

We have taken four datasets in this study i.e. StackOverflow, security.StackExchange, AWS, and Azure Support forums. For the StackOverflow dataset, we have used the dataset from the archive.org website where the datasets are uploaded quarterly. For AWS and Azure support forums, we have used a web-crawler to scrape the cloud discussion forums. For web crawlers, we considered custom PHP scrapers, selenium, and ParseHub [10], but eventually, ParseHub was selected as it provides a user-friendly interface for web-crawling. From the AWS support forums, the terms “Cloud security” and “Cloud Privacy” were filtered, which gave us 373 questions. From Azure support forums we searched for “Cloud security and privacy issues in configuring cloud resources”. Azure had an extensive set of questions on this topic and more than 88,000 questions were reported relating to this topic from which 894 questions were recorded in the dataset.

Other cloud platforms like GCP, Alibaba, and IBM cloud were also considered initially but later were discarded as GCP had only 40 questions related to this topic. GCP support forums also had references to StackOverflow and StackExchange for most of their questions. Alibaba and IBM support forums didn't have enough questions reported. This could be due to the fact that Alibaba and IBM have a small market share in the public cloud domain, hence not many questions are reported on public support forums. Moreover, Alibaba and IBM cloud platforms operate in niche markets and specialize in one domain or the other. None of those domains for either of these two cloud platforms is security or privacy. Hence, we could not collect data from these two platforms. Stackoverflow discourages web crawlers, as it has advanced CAPTCHAs in place, which makes circumvention difficult using ParseHub.

Parsehub was considered a viable option, however, it has its own challenges. Parsehub is a paid web-crawling tool that also has a free-tier plan which comes with limited projects and a limit on how much data can be scraped in on go. It also has problems with pagination pattern detection, which results in faulty pattern detection. To overcome this problem, manual intervention was required to go over files to make sure the results are correct. Also, in order to overcome the problem with pagination, we had to inject our own code in JavaScript inside our scraper, ParseHub. The conditional code goes over the HTML DOM tree in the browser, picks up internal HTML elements and text, and conditionally clicks on the “Next” button to go to the next page and logically avoids clicking the “previous” button, sitting adjacent to the “next” button

Stackoverflow is a widely used platform for developers, hence it has millions of questions and posts. The dataset size on archive.org was 85.6 GB (uncompressed). The data was also in XML format. We have observed two challenges (i) pandas and ElementTree data frames readers usually give MemoryError when reading large dataset files. (ii) high-performance Python libraries like vaex [11] and Dask [2] for reading large datasets, do not have reader functions for XML files. To overcome these challenges, we used Google Big Query [12]. We have found Google Big Query very effective in filtering out cloud security and privacy-related posts from the StackOverflow Posts dataset, using a SQL-like querying interface. The SQL query was crafted in such a way that the 'Titles' of the posts should always contain the term "cloud" along with terms like "VM", "key-pair", "instance", "privacy", "security groups", "ACL", "encryption", "decryption", "configuration management", "ssh", "identity access management", "VPC", "inbound rules", "outbound rules", "tunnel", "SSL" and "certificate management" should also be included. The Stackoverflow dataset contains a total of 21,641,802 records spanning from July 2008 to September 2021. The filtered Stackoverflow dataset contains 5564 records which are 0.0257% of the overall records. The SQL query is available on GitHub [13].

V. Data Analysis pipeline

Once the datasets are made available in .csv format, they are read into a pandas data frame. For our analysis, currently, the posts "body" is considered for topic extraction. In the Data Cleaning step, while analyzing different datasets, we have learned that there is no one size fits all approach to filter and clean the datasets, however, there are some steps that are common e.g. in security.StackExchange dataset, there are a total of 175900 records (after removing the NaN rows). After filtering for relevant terms like cloud, AWS, Azure & GCP in the titles and tags of the posts, we get 739 records which are 0.42% of the overall records. For StackOverflow, AWS, azure datasets, the data filtration has already been done during the data collection step, hence we get 5563, 372 & 871 records (after removing the NaN rows) respectively. Punctuations and special characters are also removed from the posts and finally converted to lowercase so that the topics should include only natural language data.

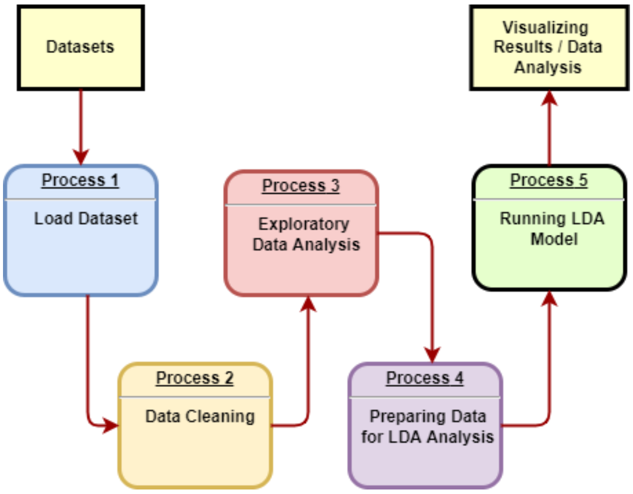


Figure 1. LDA Data Analysis Pipeline

For the Exploratory Data Analysis step, we visualize the terms appearing in the posts "body" with respect to the relative frequency using the python "word cloud" library. This gives us an insight into the terms that are not relevant, which are included in our stop words list, for removal from the text. Word Clouds from our four datasets under consideration are below:

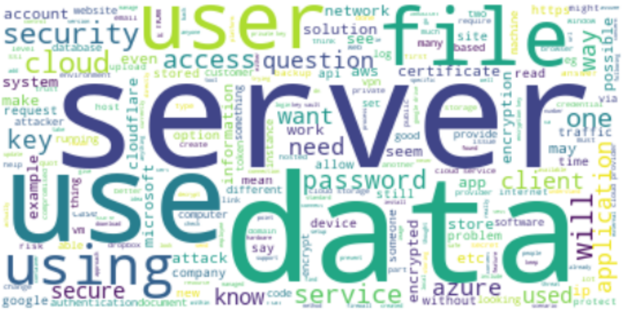


*Figure 2: **Security.StackExchange** WordCloud*

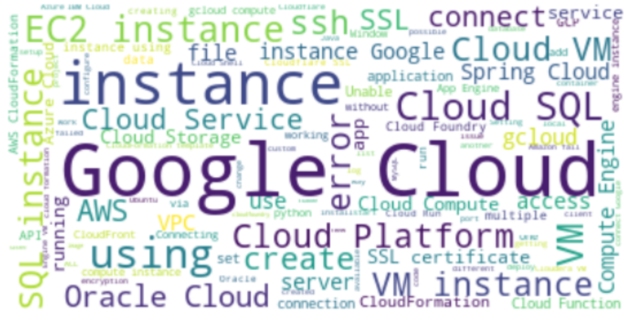


*Figure 3: **Stack Overflow** WordCloud*

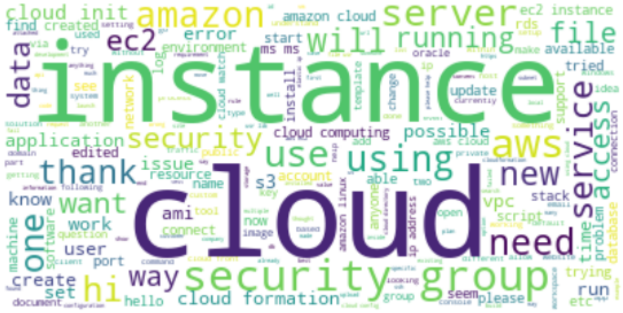


*Figure 4: **AWS Developer Forum** WordCloud*

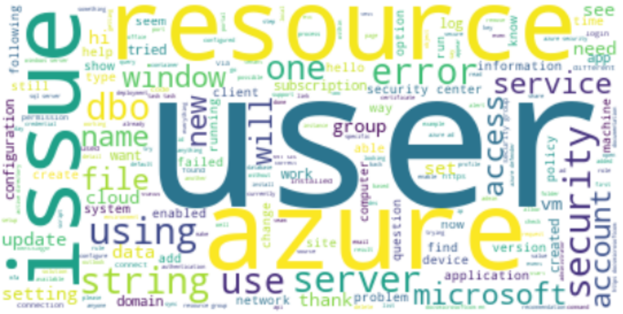
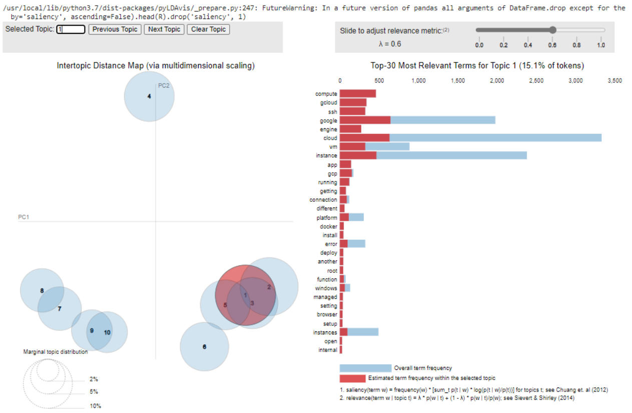


*Figure 5: **Azure Support Forum** WordCloud*

In preparing the dataset for the LDA step, we are first converting the text document into a list of lowercase tokens using the gensim.utils.simple_preprocess function and then removing the stop words that we have built in the previous step. Later on, a mapping between words and their integer ids is created using gensim.corpora.dictionary function. In our unoptimized implementation, we have prepared a corpus (stream of document vectors), mapping of word IDs to words, and set the number of topics to 10 to be fed to the gensim.models.ldamulticore function. This model infers the topic distribution from the documents.

Once the LDA model has been learned, pyLDAvis library is used to visualize the topics in the topic model that has been fit to a corpus of text data. The pyLDAvis package extracts the information from a fitted LAD model to inform an interactive web-based visualization. The visualization shows an Intertopic Distance Map and Top-30 Most Relevant Terms for each topic. It has a configurable relevance metric (λ), which denotes the degree to which a term appears in a particular topic to the exclusion of others. Relevance is based on another metric, lift, which is the ratio of a term's probability within a topic to its margin probability across the corpus. When λ = 1, the terms are ranked by their probabilities within the topic (the 'regular' method) while when λ = 0, the terms are ranked only by their lift. The interface allows you to adjust the value of λ between 0 and 1. In this study, we will use λ = 0.6 which is considered an optimal value, and it resulted in an estimated 70% probability of correct identification, whereas for values of λ = 0 and λ = 1, the estimated proportions of correct responses were closer to 53% and 63%, respectively [1]. Once we have the Top-30 terms for each cluster, we will interpret the topics using the term frequencies. This step was performed manually. A visual representation of the pyLDAvis visualization is shown in the figure:



```
/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the
  by='saliency', ascending=False).head(R).drop('saliency', 1)
```

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6: The layout of LDAvis, with the 10 topics on the left, and the term frequency on the right. Here λ = 0.6.

V. Data Analysis

To perform the data analysis, the goal of our research is to identify (i) What are the popular topics (across all datasets) & (ii) Which is the most relevant forum for cloud discussions. To answer the first research question, the most popular topics are extracted from security.StackExchange, StackOverflow, AWS, and Azure datasets as below:

Table 1. LDA topics and the top terms in the topic (security.StackExchange)

Cluster	Topic	Top terms
1	Web application security and key management in AWS	cloud, data, IP, aws, access, server, key, security, web, provider, public, server
2	Data encryption and information security in Azure	azure, service, cloud, files, data, encryption, need, security, information, storage, Microsoft
3	uploading and encrypted data on the cloud (google, dropbox, azure)	key, cloud, server, security, store, google, data, certificate, file, encrypt, upload, access
4	cloud user management (password, keys, etc.)	cloud, password, security, data, azure, service, file, server, VM, user, encrypted, key, users, access
5	Secure channel/SSL/TLS in Azure cloud platform	data, schannel, security, access, azure, somecompanycom, server, company, token, network, encrypted, user, TLS, connection
6	Accessing cloud machine in Azure cloud platform using public key	access, cloud, machine, azure, server, Cloudflare, key, data, virtual, public, however, ssl, security, site, keys
7	Security compliance in AWS and Azure cloud platforms	security, cve, certificate, aws, azure, application, script, need, server, service, cloud, secure
8	Users files and applications getting encrypted by ransomware	user, file, encrypted, files, data, app, security, application, ransomware, access, machine, refresh, cloud
9	SSH Access With Cloudflare Argo and Access	Cloudflare, server, password, key, argo, tunnel, access, azure, user, traffic, data, get, need, keys, attacker, ssh
10	Restricted access and serving private content on Amazon CloudFront (cache/CDN)	server, data, user, access, password, key, CloudFront, files, cloud, request, item, device

Table 1 shows the top 10 LDA topic clusters generated from the security.StackExchange data. The topic labels have been inferred from the frequent terms appearing in the post's body. The topics include various areas of security e.g. Web application and key management, data encryption, uploading, and encrypting data on the cloud, user management, Schannel, etc.

Table 2. LDA topics and the top terms in the topic (Stack Overflow)

Cluster	Topic	Top terms
1	Accessing google cloud instance via ssh	compute, gcloud, ssh, google, engine, cloud, VM, GCP
2	Google Cloud SQL related issues	SQL, instance, cloud, google, create, connect, IP, creating, groups, security, id,
3	Google cloud platform VM and container-based environments	Google, cloud, platform, VM, address, instance, default, run, stack
4	Using Amazon cloud formation templates to configure EC2 instances	Cloudformation, EC, AWS, template, VPC, instance, group, existing, Cloudwatch,

5	Configuring cloud instances resources (disk, network, images, etc.)	Cloud, formation, get, size, image, python, disk, console, public, node, static, VMs, update
6	Cloud foundry with Oracle Cloud database MySQL	Oracle, database, cloud, cannot, MySQL, foundry, set, load, files, storage, user, find, balancer, password
7	Issues related to running google cloud SQL on the Cloudera platform	access, unable, file, cloud SQL, command, Linux, config, already, Cloudera, reference, error, failed, cli
8	SSL certificate issues with Cloudflare and Cloudfront CDN/Caching servers	SSL, Cloudflare, certificates, CloudFront, port, name, elastic, domain, issue, terminate, custom, tags
9	autoscaling support in NoSQL databases	NoSQL, script, auto, scaling, change, init, volume, spring, windows, JSON, non, volumes
10	Issues related to data types and data structures	data, type, list, service, key, delete, instance, multiple, stop, account, supported, encryption

Table 2 shows the top 10 LDA topics clusters generated from the StackOverflow dataset. The top-3 topics are related to the Google Cloud Platform. As it was initially found, GCP support forums have references to Stackoverflow. This insight revealed from running LDA on the SO dataset hence proves our initial findings. The next two topics are related to AWS cloud formation, which shows that developers frequently use this platform to provision cloud resources.

Table 3. LDA topics and the top terms in the topic (AWS Developer Forum)

Cluste r	Topic	Top terms
1	General security questions related to configuring EC2 instance	cloud, security, aws, group, instance, amazon, ms, VPC, oracle, user, way, packages, new, ec, rds
2	Questions related to AWS but with a polite tone	cloud, AWS, search, service, hi, server, thanks, computing, instance, new, regards, oracle, please
3	Data security on amazon cloud instances EC2	cloud, instance, EC, amazon, please, data, security, summit, hi, config, tried, business, possible, go, get, stack, help, need, group, web, development
4	Configuring security groups on the cloud instance	security, group, cloud, account, server, IP, data, EC, instance, init, since, euw, name, cidr, allows, formation
5	Security groups and vpc related discussions	cloud, AWS, group, EC, security, vpc, private, thanks, server, java, instances, wrong, files, process, work, amazon
6	Questions related to Hashicorp Consul deployed on AWS	ms, request, consul, httpd, running, deploy, cloud, awsiotpythonsdk, cpd, pip, urlib, windows, autoscaling, instance
7	Issues related to amazon elastic load balancer and cloud directory	Amazon, group, elb, security, mp, clouddirectory, schemafacets, create, sq, level, data, sgs, traffic, internal
8	Performance testing surveys are done using Apache Jmeter	schema, survey, cloud, development, process, JMeter, migration, directory, groups, may, data, participate, appreciate, great, complete
9	Questions related to AWS CloudHSM	gb, aws, document, cloud, log, regions, HSM, across, attackers, bastion, SDF, end, vms, stuck, physical, file
10	General topics on HTML, malware, software, IPSec	HTML, survey, malware, software, IPsec, header, video, tnx, archiva, applications, complete

Table 4. LDA topics and the top terms in the topic (Azure Support Forum)

Cluste r	Topic	Top terms
1	Securing web app with Azure App service	azure, server, security, app, resources, new, log, user, sql, help, windows, need, thanks, name, hi, services
2	Java application development on Spring Cloud	spring, error, system, server, domain, security, releasejar, windows, access, users, get, release, cannot, cloud
3	Azure cloud vm security	azure, security, cloud, microsoft, error, get, vm, task, https, help, subscription, group, hi, update, resources
4	Azure windows server security issues with configurations	azure, service, server, windows, issues, error, get, configuration, https, resource, resources, app, security
5	Azure Active Directory users and group security	server, group, windows, azure, security, site, user, ad, access, error, exchange, users, get, issue, update, cloud
6	Coding syntax related questions	string, security, replace, tostring, azure, cloud, false, true, name, storage, shell, array, qualifier, application, please
7	Security groups other Azure vm configurations for instantiation	azure, data, resource, issue, group, still, security, address, email, vm, ip, network, level, please, peripheral, disk
8	Access management in azure AD	user, domain, users, set, new, azure, able, ad, dns, server, error, access, resources, tried, file, admin, cloud, microsoft
9	Manage database ownership and policies in Azure MSDB	dbo, task, movenext, security, issue, inline, smart_admin, center, azure, access, policy, initiative, modules, users, msdb
10	Security in Mobile device management, Data loss prevention, Mobile App Management	security, mdm, policy, frontpage, thanks, unitofwork, dlp, mam, cu, microsoft, hi, ir, access, please, azure, devices

Table 4 shows the top 10 LDA topics clusters generated from the dataset extracted from Azure support forums. The top question is related to securing web apps. The second most popular topic is related to Spring Cloud, and the third and fourth are related to configuring VM security/security groups in the cloud and securing windows server VM instances. The fifth most popular topic is related to Microsoft Azure Active Directory, which is one of the widely used directory services in organizations and enterprises [17].

An observation in analyzing support forums of AWS and Azure is that there can be terms that are common in various topics e.g. aws will appear in multiple topics as one of the most frequent terms. In that case, we look further beyond the top-5 most frequent terms to infer the topics being discussed in the posts.

V. LIMITATION

Throughout our research, several challenges surfaced. We dealt with most of them. Still, there are a few that we are surely aware of but could not circumvent, owing to the short span of the temporal timeline we were operating in

i. First of all, we could not have structural consistency in our datasets. AWS cloud platform did not provide us with any tags related to the questions. Hence, tags were not found in AWS scraped data. So, we could not have filtered all the data based on tags. This did not raise sufficient problems, as we later resorted to filtering datasets based on the question title, not the tags or the body text

ii. Second of all, we may have failed to interpret all the questions correctly. This is due to the fact that the research team is proficient in English, whereas there is a possibility that not all questions, especially the ones obtained from the online datasets, were in English. But given the fact that the datasets and the scraped data are English-dominated, this problem should not be that big of an issue for us

iii. Thirdly, we scraped all the data available on AWS and did not apply any temporal filter. This is to say that questions pertaining to some old and obsolete data and technologies might have ended up in our scraped data. But given the fact that the higher-level problem for all those users was the same, this limitation, once again, did not stop us or impede our progress in any way

iv. Fourthly, data filtering bias creeps into the project whenever researchers manually apply filters, just as we did. We used keywords, like “GCP”, “Azure”, “ AWS”, “Cloud” etc. for filtering.

v. Finally, we broadly collected security and privacy data from the internet and did not make any distinction between the two topics. Hence, our results present a picture of security and privacy issues in configuring cloud resources combined, and we implemented no separation of the two concerns in any way. Also, our entire data is security-heavy, i.e., most of the questions that we collected, scraped, and analyzed are related to the issues related to security, not privacy

VI. FUTURE WORK

As mentioned before, owing to the rather shorter temporal timeline, we could not carry out the project to its full potential. Those elements can certainly be incorporated by those that build upon our work

i. Firstly, user analysis can be incorporated into the project. While downloading and scraping datasets, one can always use the user metadata section to get an insight into the professional role and job title of the questioners. This will help us understand what group of people (managers, administrators, students, academic workers, etc.) is primarily concerned with what issues and problems. This can offer a lot of insights into the challenges that workers in different roles face on a day-to-day basis

ii. Secondly, one can work on creating extending guides for the community that focuses on providing information about the problems as well as secure solutions to them. This, however, will demand significant effort and energy. The main reason for this is that as technology progresses, the nature and details of the problems that the community faces evolve as well, making our guides outdated. So, to maintain the guides well enough, one will have to update them regularly, and for that, one might have to go over the entire process all over again.

iii. To uncover more posts from the StackExchange datasets, one approach could be to concatenate titles with the body of the posts to include more records in the analysis. This can be considered as future work.

VII. Conclusion

In conclusion, we have identified that developers face challenges in accessing cloud instances using SSH, configuring security groups, dealing with fraudulent certificates in OpenSSL, secure data transmission over the internet (TLS, RSA, AES, etc.), securing APIs as well as Active Directory. In our study, we have identified that developers using AWS and GCP generally tend to use Stack Overflow while those using Azure and AWS tend to use security.StackExchange to post their questions and queries. In the IS dataset, we have seen questions encountered by developers are common on multiple cloud platforms. From the SO dataset we have found Google Cloud Platform (GCP) to be the most popular among developers followed by Amazon AWS.

VIII. ACKNOWLEDGEMENTS

We acknowledge the open source communities for all the effort they have put into collecting and archiving all the data we downloaded and obtained with just one click. Mainly, Stack Exchange and StackOverflow were used for this purpose

Also, we recognize Microsoft and Amazon for structuring their discussion forums in a neat and clean way that it was not only easy for us to get to the relevant data but also scrape it

We thank the creators of the ParseHub scraping tool. We were able to collect all the data without even subscribing to the premium pricing plans, using only the free-tier plan. The quality of the software combined with the in-tool step-by-step tutorials and the online documentation is the best that the team has come across yet

Last but certainly, certainly not least, we express our sincere and deepest gratitude towards Dr. Mobin Javed — Assistant Professor Computer Science, Lahore University of Management Sciences, LUMS – for overseeing the entire project and supporting us through all the stages. Special thanks to her for providing us with the cloud resources from her lab to help us move forward with the project in a timely manner.

I. References

[1] Christoffer Rosen and Emad Shihab. 2016. What are mobile developers asking about? A large scale studies using stack overflow. Empirical Software Engineering 21, 3 (Jun 2016), 1192–1223. DOI: http://dx.doi.org/10.1007/s10664-015-9379-3

[2] Dask: Scalable analytics in Python

[3] D. Zissis, D. Lekkas, "Addressing cloud computing security issues", Future Generation Computer Systems, 28, 583–592, 2012

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[5] Duygu Sinanc and Seref Sagiroglu. 2013. A review on cloud security. In Proceedings of the 6th International Conference on Security of Information and Networks (SIN '13). Association for Computing Machinery, New York, NY, USA, 321–325. DOI:https://doi.org/10.1145/2523514.2527013

[6] Hala Assal, Sonia Chiasson, and Robert Biddle. 2016. Cesar: Visual representation of source code vulnerabilities. In 2016 IEEE Symposium on Visualization for Cyber Security (VizSec). 1–8. DOI: http://dx.doi.org/10.1109/VIZSEC.2016.7739576

[7] https://pyldavis.readthedocs.io/en/latest/readme.html

[8] https://www.alexa.com/siteinfo/stackoverflow.com

[9]https://stackoverflow.com/advertising/audience#:~:text=How%20many%20developers%20visit%20Stack,visits%20from%20developers%20on%20average.

[10] https://www.parsehub.com/

[11] https://vaex.io/

[12] https://cloud.google.com/bigquery/

[13]https://github.com/mustafaalam1991/Stackoverflow-Google-Big-Query/blob/main/Filter_CloudSecPriv

[14] https://forums.aws.amazon.com/index.jspa

[15] Li Yan, Xiaowei Hao, Zelei Cheng, and Rui Zhou. 2018. Cloud Computing Security and Privacy. In Proceedings of the 2018 International Conference on Big Data and Computing (ICBDC '18). Association for Computing Machinery, New York, NY, USA, 119–123. DOI: https://doi.org/10.1145/3220199.3220217

[16] Manuel Egele, David Brumley, Yanick Fratantonio, and Christopher Kruegel. 2013. An Empirical Study of Cryptographic Misuse in Android Applications. In Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security (CCS '13). ACM, New York, NY, USA, 73–84. DOI: http://dx.doi.org/10.1145/2508859.2516693

[17] Microsoft Active Directory Market Share and Competitor Report | Compare to Microsoft Active Directory, OpenSSL, SEDO Parking (datanyze.com)

[18] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. 2020. Understanding Privacy-Related Questions on Stack Overflow. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376768

[19] M. Squire, "Should we move to stack overflow?: Measuring the utility of social media for developer support," in 37th International Conference on Software Engineering, 2015, pp. 219–228

[20] OWASP. 2017. Top 10 - 2017 The ten most critical web application security risks. Technical Report. The OWASP Foundation. https://www.owasp.org/index.php/Category:OWASP_Top_Ten_Project

[21] P. K. Venkatesh, S. Wang, F. Zhang, Y. Zou and A. E. Hassan, "What Do Client Developers Concern When Using Web APIs? An Empirical Study on Developer Forums and Stack Overflow," 2016 IEEE International Conference on Web Services (ICWS), 2016, pp. 131-138, doi: 10.1109/ICWS.2016.25.

[22] Stack Overflow. 2021d. Where Developers Learn, Share, & Build Careers. Retrieved December 2021 from https://stackoverflow.com

[23] Stefanie Beyer and Martin Pinzger. 2014. A Manual Categorization of Android App Development Issues on Stack Overflow. In 2014 IEEE International Conference on Software Maintenance and Evolution. 531–535. DOI: http://dx.doi.org/10.1109/ICSME.2014.88

[24] Tamara Lopez, Thein Tun, Arosha Bandara, Mark Levine, Bashar Nuseibeh, and Helen Sharp. 2019. An Anatomy of Security Conversations in Stack Overflow.In Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS '10). IEEE Press, Piscataway, NJ, USA, 31–40. DOI: http://dx.doi.org/10.1109/ICSE-SEIS.2019.00012

[25] Xin-Li Yang, David Lo, Xin Xia, Zhi-Yuan Wan, and Jian-Ling Sun. 2016. What Security Questions Do Developers Ask? A Large-Scale Study of Stack Overflow Posts. Journal of Computer Science and Technology 31, 5 (Sep 2016), 910–924

[26] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. 2016. You Get Where You're Looking for: The Impact of Information Sources on Code Security. In 2016 IEEE Symposium on Security and Privacy (SP). 289–305. DOI:http://dx.doi.org/10.1109/SP.2016.25

[27] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. 2017. How Internet

Resources Might Be Helping You Develop Faster but Less Securely. IEEE Security Privacy 15, 2 (March 2017), 50–60. DOI: http://dx.doi.org/10.1109/MSP.2017.24

[28] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. 2017. Comparing the Usability of Cryptographic APIs. In 2017 IEEE Symposium on Security and Privacy (SP). 154–171. DOI: http://dx.doi.org/10.1109/SP.2017.52

[29] Z. Wan, X. Xia and A. E. Hassan, "What Do Programmers Discuss About Blockchain? A Case Study on the Use of Balanced LDA and the Reference Architecture of a Domain to Capture Online Discussions About Blockchain Platforms Across Stack Exchange Communities," in IEEE Transactions on Software Engineering, vol. 47, no. 7, pp. 1331-1349, 1 July 2021, doi: 10.1109/TSE.2019.2921343.

[30] 2021 Data Breach Investigations Report (DBIR) Verizon 20