IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

USER PRIVACY PRESERVING WITH CRYPTOGRAPHY ON AD BLOCKING E-COMMERCE

Mr.K.S.ARUN,M.Tech(IT)[1], VIGNESH M[2],THAMARAIKANNAN D[3], VINOTH P[4],

1 ASSISTANT PROFESSER, 2,3,4 UG STUDENTS
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAHENDRA ENGINEERING COLLEGE, TAMILNADU, INDIA

ABSTRACT

The growing number of mobile and IoT devices has nourished many intelligent applications. In order to produce high-quality machine learning models, they constantly access and collect rich personal data such as photos, browsing history and text messages. However, direct access to personal data has raised increasing public concerns about privacy risks and security breaches. To address these concerns, there are two emerging solutions to privacy-preserving machine learning, namely local differential privacy and federated machine learning. The former is a distributed data collection strategy where each client perturbs data locally before submitting to the server, whereas the latter is a distributed machine learning strategy to train models on mobile devices locally and merge their output (e.g., parameter updates of a model) through a control protocol. In this paper, we conduct a comparative study on the efficiency and privacy of both solutions.

Keywords : Blocking E-commerce, Browsing history and Text messages

INTRODUCTION

The growing number of mobile and IoT devices has nourished many intelligent applications. In order to produce high-quality machine learning models, they constantly access and collect rich personal data such as photos, browsing history and text messages. However, direct access to personal data has raised increasing public concerns about privacy risks and security breaches. To address these concerns, there are two emerging solutions to privacy-preserving machine learning, namely local differential privacy and federated machine learning. The former is a distributed data collection strategy where each client perturbs data locally before submitting to the server, whereas the latter is a distributed machine learning strategy to train models on mobile devices locally and merge their output (e.g., parameter updates of a model) through a control protocol. In this paper, we conduct a comparative study on the efficiency and privacy of both solutions. Our results show that in a standard population and domain setting, both can achieve an optimal misclassification rate lower than 20% and federated machine learning generally performs better at the cost of higher client CPU usage. Nonetheless, local differential privacy can benefit more from a larger client population (> 1k). As for privacy guarantee, local differential privacy also has flexible control over the data leakage.

SYSTEM ANALYSIS

EXISTING SYSTEM

In existing system, the user behavior analysis will be done using item set mining.Thus the user privacy will be known to anyone who mines data from big data.4The mining implements pattern based technique to mine the data.

PROPOSED SYSTEM

In our proposed system, a privacy based ad blocking system is implemented.The user searched pattern are stored in the server as encrypted format using ECC algorithm.So that the analyst can get only the count values but not the searched details.Here the unwanted user analysis ad will be blocked.The privacy of the user is maintained by using the data mining and the security approach.

FEASIBILITY STUDY

To support data collection, a web application called AWS has been developed using PYTHON and MySQL as a database server that fulfills specifications coming from domain experts. Using AWS through home Web access, each patient can access using a personal username and password and compile a daily report of the affecting toxicities by choosing and grading any of them from the user interface. If too much time has passed since the last patient report, AWS will send the doctors a communication about the missing data. This decision also takes into account suggestions about reminders and clinician feedback discussed . Physicians access to a specific part of the site where they can perform managing operations on patients' data, as inserting or updating database information and visualize flowsheets of patient toxicities by means of graphing functions.

2 PRELIMINARIES

Local Differential Privacy

LDP [4] extends the notion of differential privacy by per- turbing local data with noise determined by a predefined parameter. In a nutshell, a perturbation algorithm prob- abilistically modifies a local raw value v_i to another value in the same domain of possible outputs κ. The modified value is then submitted to the server. A learning task on the statistical features (e.g., frequency and mean) of such data retains certain accuracy after the server collects all perturbed values. Meanwhile, each individual can have plausible privacy guarantee bounded on a privacy budget of s.

Formally, the perturbation algorithm suffices s-LDP prin- ciple if and only if for any two individuals' inputs v_i and v_j, we have

$$Pr[A(v_i) = s] \leq e^s \cdot Pr[A(v_j) = s],$$

where s κ. Obviously, perturbed data is closer to the original data with a larger privacy budget s and user population. Since the noises are applied to the data set directly, this strategy may have a strong impact on model performance when the budget is low.

Federated Machine Learning

In a task of federated machine learning, each mobile de- vice initializes its own training using the shared model downloaded from the server and builds a new model using its local data. The updated model parameters will then be returned to the server, averaged with other peer devices and merged as the new shared model. This process is repeated

..................Column Break..................To train such an objective, a straightforward gradient descent algorithm can be applied to estimate model param- eters using the iterative rule below:

$$W_{t+1} \leftarrow W_t - \eta \nabla g(W),$$

which is a full-batch gradient descent using all client data to generate an update in round t. However, this is not practical since it takes a long time for each iteration and even multiple times longer under the case of potentially high latency and limited bandwidth of the mobile network. To improve communication efficiency, federated machine learning commonly increases individual client computation by asking each mobile device to iterate over local data sev- eral times with stochastic gradient descent before submit- ting the parameter updates to the server for averaging [5].

SYSTEM REQURIMENT

FUNCTIONAL REQURIMENT

- **Operational Cost**: The architected solution for an application should have a minimum monthly operational cost as nothing is free on the cloud. The solution at the minimum should meet the minimum requirements for scalability, availability, fault tolerance, security, replication, and disaster recovery.
- **Scalability cloud infrastructure**: The cloud infrastructure should scale the application up or down by adding/removing application nodes from the network, depending on the load on the application.
- **Scalability application**: The architected solution should be designed in a decoupled and stateless manner, which lends itself to support scaling.
- **High availability**: The architected solution will be designed in a manner which avoids single point failures in order to achieve high availability.
- **Fault tolerant**: The application should be coded to handle cloud services' related failures to the extent possible.
- **Application security**: The application should use an encrypted channel for communications. All the confidential data should be stored in an encrypted format. All the files at rest should be stored in an encrypted format.
- **Cloud infrastructure security**: The cloud infrastructure should be configured to close all the unnecessary network ports with the help of the firewall. All the compute instances on the cloud should be secured with SSH keys.
- **Replication**: All the data should be replicated in real time to a secondary location to reduce the window for data loss.
- **Backups**: All the data from the databases shall be backed up on a daily basis.
- **Disaster recovery**: The architected solution should be designed in a manner that it is easy to recover from an outage with minimal human intervention, with the help of automated scripts.
- **Design for failure**: The architected solution should be designed for failure; in other words, the application should be designed, implemented, and deployed for automated recovery from failure.
- **Should be coded** using open source software and open standards to prevent vendor lock-in and to drive costs down.
- **AWS Shelid**

SYSTEM REQURIMENT

Using AWS, you will gain the control and confidence you need to securely run your business with the most flexible and secure cloud computing environment available today. As an AWS customer, you will benefit from AWS data centers and a network architected to protect your information, identities, applications, and devices. With AWS, you can improve your ability to meet core security and compliance requirements, such as data locality, protection, and confidentiality with our comprehensive services and features.

AWS allows you to automate manual security tasks so you can shift your focus to scaling and innovating your business. Plus, you pay only for the services that you use. All customers benefit from AWS being the only commercial cloud that has had its service offerings and associated supply chain vetted and accepted as secure enough for top-secret workloads.

SOFTWARE CONFIGURATION

PHP

PHP (recursive acronym for PHP: Hypertext Preprocessor) is a widely-used open source general-purpose scripting language that is especially suited for web development and can be embedded into HTML.

HTML

HTML stands for Hyper Text Markup Language. HTML is the standard markup language for creating Web pages. HTML describes the structure of a Web page. HTML consists of a series of elements. HTML elements tell the browser how to display the content.

SQL

SQL is Structured Query Language, which is a computer language for storing, manipulating and retrieving data stored in a relational database. SQL is the standard language for Relational Database System.

LINUX

Linux has long been the basis of commercial networking devices, but now it's a mainstay of enterprise infrastructure. Linux is a tried-and-true, open-source operating system for computers, but its use has expanded to underpin systems for cars, phones, web servers and, more recently, networking gear.

NGINX

IT is a well known open source project originally written by Igor Sysoev, a Russian engineer. Igor started the project in 2002 and made it public in 2004. Since that time NGINX has become a de-facto standard for high-performance, scalable websites. Tens of millions of active websites use NGINX, including 1 million busiest websites in the world. Companies like Airbnb, Box, Dropbox, Netflix, Tumblr, WordPress.com, and many others deploy NGINX for scalability and performance reasons.

AWS Account Security

AWS is equipped with numerous measures such as access control, creation of IAM user accounts, data encryption, and the trusted advisor security checks to secure information from any attack. Security measures exist at each level of software and for each application. Hence, every application can implement advanced security in order to secure data.

GITHUB

It’s every developer’s worst nightmare – you’re working on a new project feature and you screw up. Enter version control systems (VCS) – and more specifically, GitHub.By rolling out your project with the service, you can view any changes you’ve made or even go back to your previous state (making pesky mistakes a thing of the past). The repository hosting service also boasts a rich open-source development community (making collaboration between teams as easy as pie), as well as providing several other components such as bug tracking, feature requests, task management, and wikis for every project. Many employers will look for finely honed Git skills, so now’s the perfect time to sign up – plus it’s a great way to get involved and learn from the best with a wide array of open-source projects to work on.

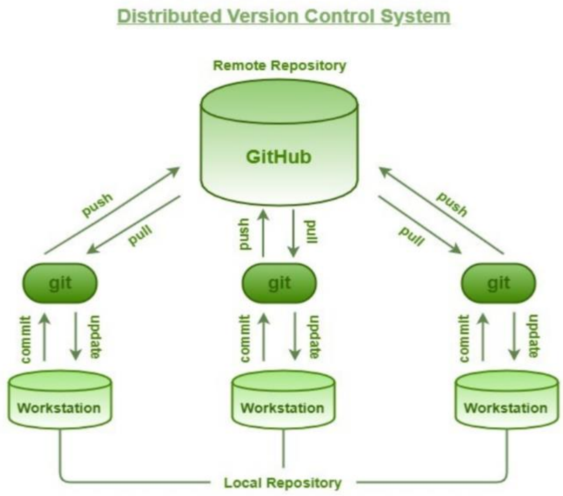
SYSTEM DESIGN

GIT HUB

Git is a free and open-source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.
Git relies on the basis of distributed development of software where more than one developer may have access to the source code of a specific application and can modify changes to it which may be seen by other developers. It allows the user to have “versions” of a project, which show the changes that were made to the code over time, and allows the user to backtrack if necessary and undo those changes.

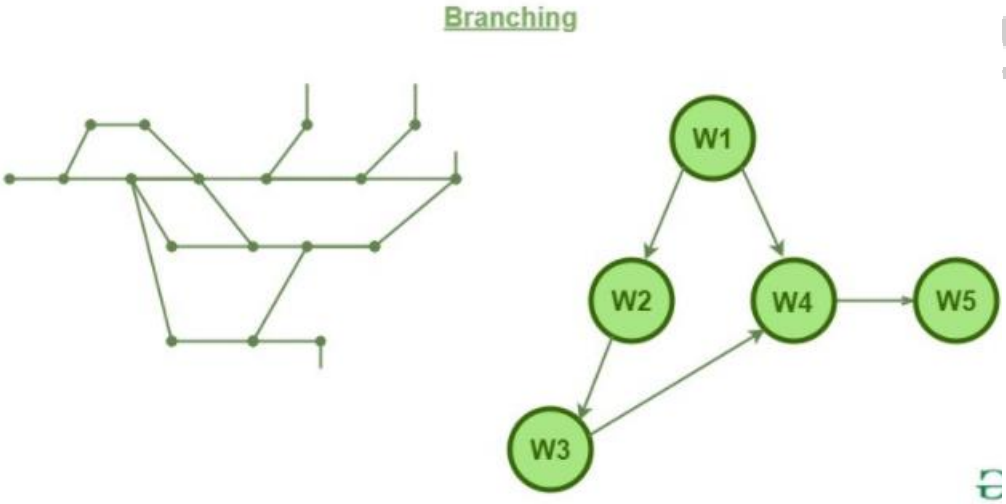
Distributed

Distributed systems are those which allow the users to perform work on a project from all over the world. A distributed system holds a Central repository that can be accessed by many remote collaborators by using a Version Control System. Git is one of the most popular Versions Control System that is being used nowadays. Having a Central Server results in a problem of Data Loss or Data disconnectivity in case of a system failure of the central server. To tackle such kind of a situation, Git mirrors the whole repository on each snapshot of the version that is being pulled by the user. In this case, if the central server crashes, then the copy of repositories can be gained back from the users who have downloaded the latest snapshot of the project.



Branching

Git allows its users to work on a line that runs parallel to the main project files. These lines are called branches. Branches in Git provide a feature to make changes in the project without affecting the original version. The master branch of a version will always contain the production quality code. Any new feature can be tested and worked upon on the branches and further, it can be merged with the master branch.
Branching and merging can be done very easily with the help of a few Git commands. A single version of a project may contain n number of branches a per the user's requirement.



Lightweight

Git stores all the data from the central repository on to the local repository while cloning is done. There might be hundreds of users working on the same project and hence the data in the central repository might be very huge. One might be worried that cloning that much data into local machines might result in system failure but Git has already taken care of such a problem. Git follows the criteria of lossless compression that compresses the data and stores it in the local repository occupying very minimal space. Whenever there is a need for this data, it follows the reverse technique and saves a lot of memory space.

Speed

Since Git stores all the data related to a project in the local repository by the process of cloning, it is very much efficient to fetch data from the local repository instead of doing the same from the remote repository. Git is very fast and scalable compared to other version control systems which results in the handling of large projects efficiently.

The fetching power from a local repository is about 100 times faster than what possible with the remote server. According to a test conducted by Mozilla, Git is one order of magnitude faster, which is about 10 times faster than other VCS tools. This is because Git is actually written in C language which is unlike other languages, very close to machine language and hence it makes processing very fast.

Open-Source

Git is a free and open-source distributed version control system designed to handle everything from small to very large projects with speed and efficiency. It is called open-source because it provides the flexibility to modify its source code according to the user's needs. Unlike other Version Control Systems which provide paid functionalities like the repository space, privacy of codes, accuracy, and speed, etc. Git is all open-source software that provides these functionalities in free and even in a better way than others. Being open-source Git allows multiple people to work on the same project at the same time and collaborate with each other very easily and efficiently. Hence, Git is considered to be the best Version Control System available nowadays.

Secure

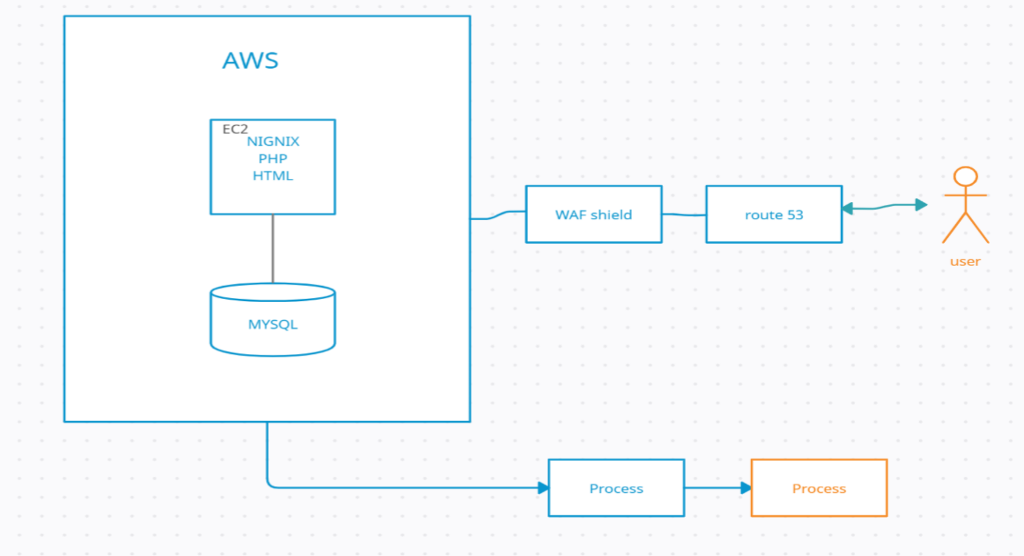
Git keeps a record of all the commits done by each of the collaborators on the local copy of the developer. A log file is maintained and is pushed to the central repository each time the push operation is performed. So, if a problem arises then it can be easily tracked and handled by the developer. Git uses SHA1 to store all the records in the form of objects in the Hash. Each object collaborates with each other with the use of these Hash keys. SHA1 is a cryptographic algorithm that converts the commit object into a 14-digit Hex code. It helps to store the record of all the commits done by each of the developers. Hence, easily diagnosable that which commit has resulted in the failure of the work.

USE CASE DAIGRAM

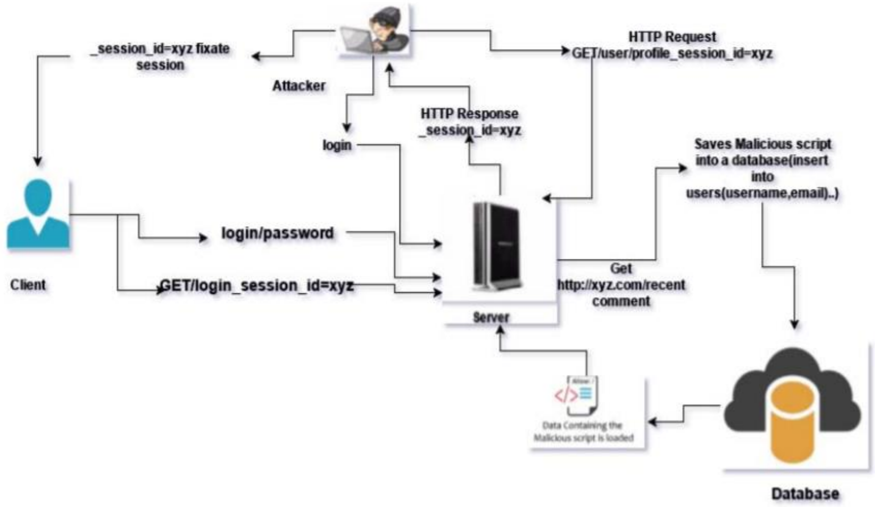
use case diagram consists ec2 linux and software configured in the linux by linux command

and data are storedin the mysql and the are portected by the cloud watch and waf,route 53

and used IAM ascces admin.



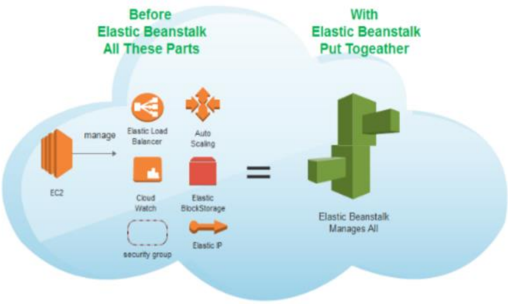
SYSTEM ARCHITECTURE


_session_id=xyz fixate session
Attacker
HTTP Request
GET/user/profile_session_id=xyz
HTTP Response
_session_id=xyz
login
Saves Malicious script into a database(insert into users(username,email)..)
login/password
Client
GET/login_session_id=xyz
Get http://xyz.com/recent comment
Server
Data Containing the Malicious script is loaded
Database

In our architecture we describe the all sessions. Client: The client of a web browser is effectively making client requests for pages from servers all over the web. In this Attack Goal % Stealing Sensitive Information 42% Defacement 23% Planning Malware 15% Unknown 08% Deceit 03% Blackmail 02% Link Spam 03% Worm 01% Phishing 01% Information Warfare 01% 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Sep. 20-22, 2017, AIIT, Amity University Uttar Pradesh, Noida, India 453 article client login to system normally, client sends request to server and gets response. This happens only in normal scenario .

Attacker: Attacker is a unauthorized user. Typically this kind of attacker would be a proficient programmer or engineer with sufficient technical knowledge to understand the weak points in a security system. In this article Attacker attacks the website through SQL injection and XSS. Uses of SQL injection and XSS by the attacker is mentioned below.

DATA FLOW DAIGRAM



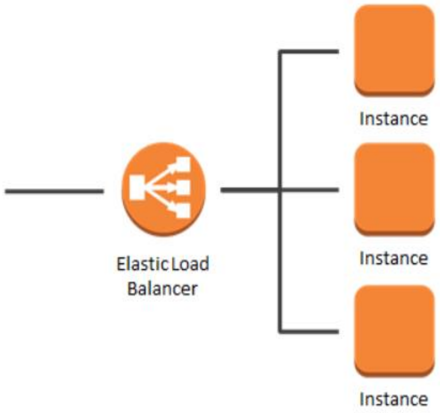
EC2

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) Cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic.

CLOUD WATCH

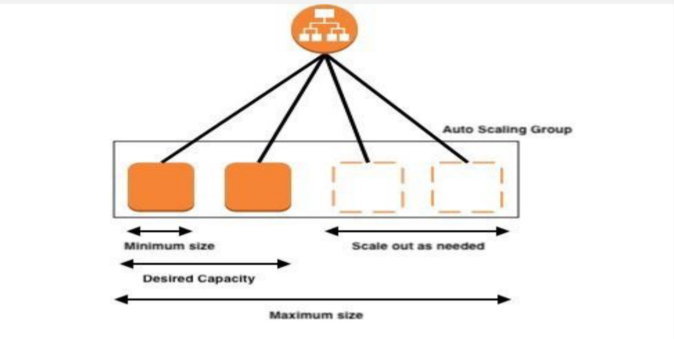
CloudWatch provides you with data and actionable insights to monitor your applications, respond to system-wide performance changes, optimize resource utilization, and get a unified view of operational health. CloudWatch collects monitoring and operational data in the form of logs, metrics, and events, providing you with a unified view of AWS resources, applications, and services that run on AWS and on-premises servers. You can use CloudWatch to detect anomalous behavior in your environments, set alarms, visualize logs and metrics side by side, take automated actions, troubleshoot issues, and discover insights to keep your applications running smoothly.

LOAD BALNCER



A load balancer distributes the incoming application or network traffic across multiple targets, such as Amazon EC2 instances, containers, and IP addresses, in multiple Availability Zones.It uses health checks to detect which instances are healthy and directs traffic only across those instances.

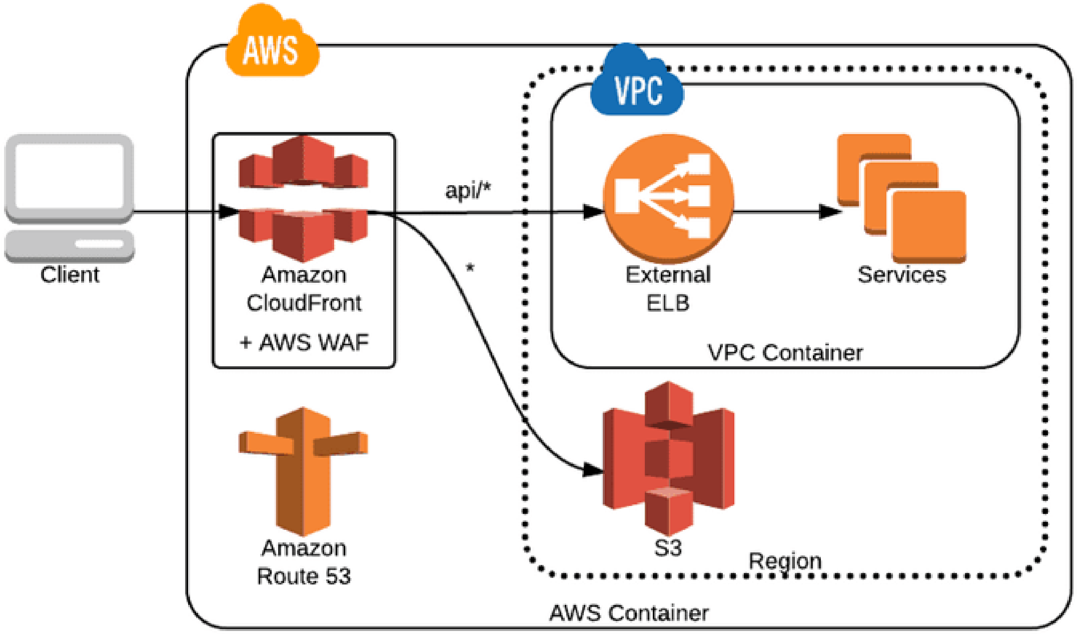
AUTOSCALING



Auto Scaling monitors your applications and automatically adjusts the capacity to maintain steady, predictable performance at the lowest possible cost.

EASTIC IP An *Elastic IP address* is a static, public IPv4 address designed for dynamic cloud computing. You can associate an Elastic IP address with any instance or network interface in any VPC in your account. With an Elastic IP address, you can mask the failure of an instance by rapidly remapping the address to another instance in your VPC.

CONTROL FLOW DAIGRAM



CLOUD FRONT

CLOUD FRONT

Amazon CloudFront is a fast content delivery network (CDN) service that securely delivers data, videos, applications, and APIs to customers globally with low latency, high transfer speeds, all within a developer-friendly environment.

WAF

AWS WAF is a web application firewall that lets you monitor the HTTP and HTTPS requests that are forwarded to CloudFront, and lets you control access to your content. Based on conditions that you specify, such as the values of query strings or the IP addresses that requests originate from, CloudFront responds to requests either with the requested content or with an HTTP status code 403 (Forbidden). You can also configure CloudFront to return a custom error page when a request is blocked

The S3 in Amazon S3 stands for **Simple Storage Service**. As the name implies it is a web service provided by Amazon Web Services which provides storage for the internet. This storage is **highly-scalable and secure in the cloud**. Having data stored in the cloud eliminates the need for in-house storage and customers can opt for unlimited storage or buy more as it is needed.

S3 is an incredibly helpful product which allows users to store and retrieve data from anywhere on the web, at any time. This is done though the AWS Management Console which is an easy to use web interface.

PREVENTION AND SECURITY

SSM

AWS Systems Manager Agent (SSM Agent) is Amazon software that can be installed and configured on an EC2 instance, an on-premises server, or a virtual machine (VM). The agent processes requests from the Systems Manager service in the AWS Cloud, and then runs them as specified in the request.

AWS service that you can use to view and control your infrastructure on AWS. Using the Systems Manager console, you can view operational data from multiple AWS services and automate operational tasks across your AWS resources.

HTTPS & SSL

SSL stands for Secure Sockets Layer and, in short, it's the standard technology for keeping an internet connection secure and safeguarding any sensitive data that is being sent between two systems, preventing criminals from reading and modifying any information transferred, including potential personal details. The two systems can be a server and a client (for example, a shopping website and browser) or server to server (for example, an application with personal identifiable information or with payroll information).

It does this by making sure that any data transferred between users and sites, or between two systems remain impossible to read. It uses encryption algorithms to scramble data in transit, preventing hackers from reading it as it is sent over the connection. This information could be anything sensitive or personal which can include credit card numbers and other financial information, names and addresses.HTTPS (Hyper Text Transfer Protocol Secure) appears in the URL when a website is secured by an SSL certificate. The details of the certificate, including the issuing authority and the corporate name of the website owner, can be viewed by clicking on the lock symbol on the browser bar.

CONCLUSION

We investigate two promising data analytic strategies for distributed setting while preserving user privacy. Both strategies are adopted in the same real machine learning problems and evaluated with extensive experiments under various system settings. The results show that local dif- ferential privacy mainly benefits from a large user popu- lation and consumes less CPU/battery on mobile devices while maintaining a rigorous privacy guarantee. Federated machine learning can adapt itself quickly for a moderate number of users and produce a learning model with higher quality while the fine-grained update is vulnerable to infer- ence. Nonetheless, the data submitted with local differential

.................Column Break.................privacy can be reused indefinitely for other tasks such as marginal release or itemset mining, while the model trained by FL is specified for one type of prediction task. As for future work, we plan to evaluate different unified solutions again each other using similar empirical framework. We also plan to propose new privacy-preserving method based on the comparative study.

REFERENCES

- U. F. T. Commission, "F.T.C. commissioners back privacy law to regulate tech companies," 2019. [Online]. Available: https://www.nytimes.com/2019/05/08/business/ftc- hearing-facebook.html
- E. Commission, "Data protection in the EU," 2018. [On- line]. Available: https://ec.europa.eu/info/law/law-topic/data- protection/data-protection-eu
- A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *S&P*, 2008, pp. 111–125.
- R. Dewri, "Local differential perturbations: Location privacy un- der approximate knowledge attackers," *IEEE Transactions on Mo- bile Computing*, vol. 12, no. 12, pp. 2360–2372, 2013.
- H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017.
- P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *Journal of Machine Learning Research*, vol. 17, pp. 17:1–17:51, 2016.
- N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, and
- G. Yu, "Collecting and analyzing multidimensional data with local differential privacy," in *ICDE*, 2019, pp. 638–649.
- N. TLC, "Trip record data," 2016. [Online]. Available: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- IPUMS, "Harmonized international census data for social science and health research," 2018. [Online]. Available: https://international.ipums.org/international/index.shtml
- D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
- A. Privacy, "Our approach to privacy," 2019. [Online]. Available: https://www.apple.com/privacy/approach-to-privacy/
- A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and
- M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *NDSS*, 2019.
- K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan,
- S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *CCS*, 2017, pp. 1175–1191.
- V. Pihur, A. Korolova, F. Liu, S. Sankuratripati, M. Yung, D. Huang, and R. Zeng, "Differentially-private "draw and discard" machine learning," *CoRR*, vol. abs/1807.04369, 2018.
- U´ . Erlingsson, V. Pihur, and A. Korolova, "Rappor: Random-
- ized aggregatable privacy-preserving ordinal response," in *CCS*. ACM, 2014, pp. 1054–1067.
- S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.
- :1–12:19, 2019.