[COMPANY] personal data de-identification whitepaper

Release date: July 2021

Legal disclosure:

This document was prepared by [COMPANY], Inc. for informational purposes only and not for any other purpose. Nothing contained in this document is, or should be construed as, a recommendation, promise or representation by the presenter or [COMPANY] or any officer, director, employee, agent or advisor of [COMPANY]. This document does not purport to be all-inclusive or to contain all of the information you may desire.

This document contains “forward-looking statements” that are based on our management’s current beliefs and assumptions and on information presently available to [COMPANY]. Forward-looking statements include information concerning our business plans and strategies, financial projections and projections of non-financial metrics, competitive position, industry environment, growth opportunities and addressable market.

Table of contents

Introduction

This paper outlines the [COMPANY] plans to protect and utilize real-world data so we can anticipate how to make our products better and keep personal and business data private and safe. All of the solutions outlined here follow current data protection and contractual requirements.

[COMPANY] created the first SaaS experience platform dedicated to Communications Service Providers (CSPs or customers) and their subscribers (users). Our SaaS top priority has always been to protect data with security and data privacy (privacy) tools. We baked in strong encryption so that data is always secure. This applies to data at rest on a local device or in the cloud. We also encrypt data traffic from network to network or from local storage to cloud storage. Our backup and recovery protocol performs continual backups as well as daily snapshots.

Recently, [COMPANY] added more products, experience management and data insight services for insights with enhanced data security and privacy measures for an improved smart home experience. Our commitment to protecting your personal and business data extends to all individuals who interact with us: CSPs, business partners, vendors, leads and prospects, and users.

Our data management security and privacy concepts are in compliance with the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA).

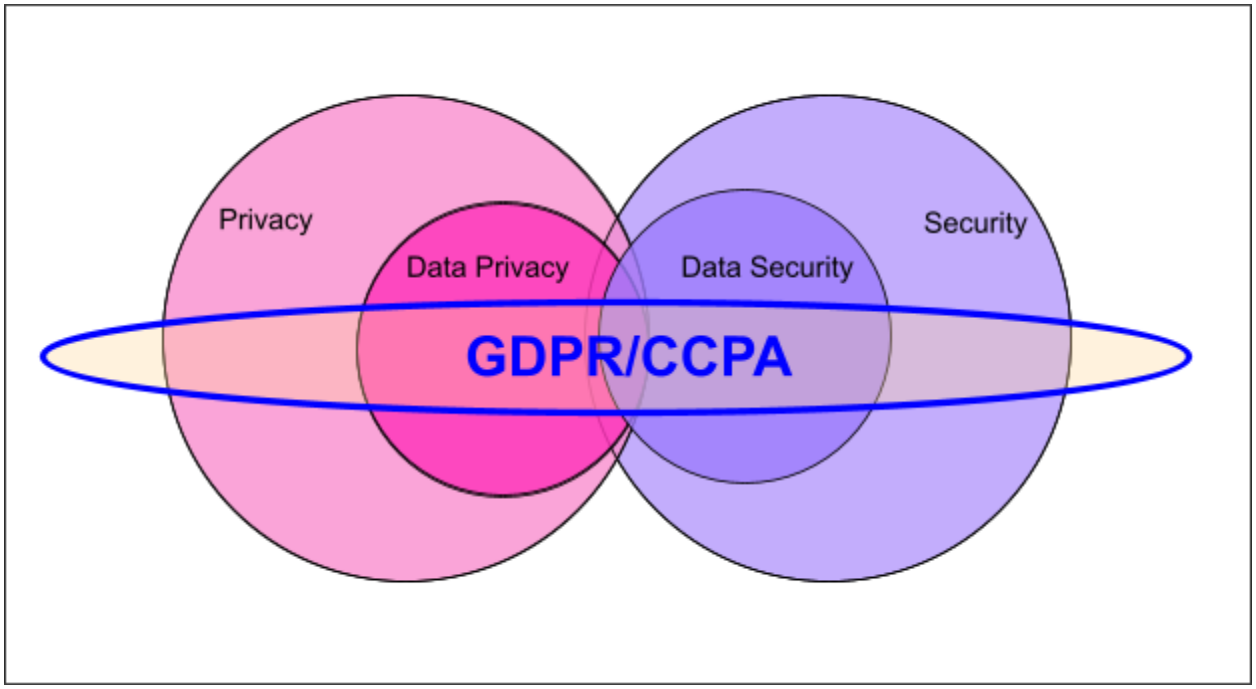

Figure: [COMPANY] security and privacy within regulation guidelines

Executive summary

As [COMPANY] expands from a smart home experience-focused company to a company responsible for handling personal data and using it to develop business intelligence, it is now mission critical to develop robust data handling solutions. We use real-world data for business needs, so our methods must be in compliance with contractual privacy and regulatory constraints. We will ensure data is secure and private, but on the other hand, we won't hamstring product functionality and impede people from doing support or research. In response to market needs for an improved smart home experience, we added experience management and data insight services with enhanced data security and privacy measures.

We preserve user privacy and adhere to data protection and retention regulatory requirements by de-identifying the personal data we collect from customer locations. We use the de-identified data for product improvement, analytics, and marketing initiatives.

The retention time we use de-identified data depends on requirements defined by:

- business need for which it was collected
- contractually required duration
- legally required retention for certain transactional records.

Our data privacy improvement plans include data anonymization on top of de-identifying the personal data. Our commitment to protecting your personal and business data extends to all individuals who interact with us: CSPs, business partners, vendors, leads and prospects, and users.

Industry Issues

Rules and regulations can keep us up at night as we navigate between what is right for you as individual business owners and what is the next new thing that needs technical investments. The quandary is that in order to develop that next great thing, our R&D depends on data analytics from real-world personal data. We know innovative ways to de-identify and anonymize data so that personal artifacts in real-world data are not traceable to any one person.

The guidance on security and privacy are complementary but come from distinct fields, with different goals. Successfully protecting data and privacy in the cloud means that both have to be integrated.

[COMPANY] security solutions

[COMPANY]'s earlier solution to personal data protection and contractual obligations was designed and architected for contemporary known issues, and for planned future developments. The following table shows how past requirements are resolved.

Security requirements	Initial privacy and security solutions
Ensure appropriate controls based on sensitivity of information	All personal data is treated with the same level of security and privacy protection.
All personal data must be encrypted at rest	A blanket encryption at rest is in all data stores using server side encryption and [COMPANY] managed keys.
All personal data must be encrypted in transit	Encrypt data in transit for internet traffic.
All forms of access to personal data must be authorized and audited	Periodic data-access audits were performed at a minimum of twice a year.
Appropriate technical and organizational measures to ensure a level of security for personal data appropriate to the risk	Host hardening in cyber security protects data in memory while processing data.
	Vendor third-party security risk assessments are implemented to understand data transfer risks. All [COMPANY] vendors are required to comply with data classification, privacy and de-identification practices.
	Customer or Vendor - [COMPANY] Data Processing Agreements (DPAs) are implemented to handle data transfer obligations.

[COMPANY] personal data protection solutions

[COMPANY]'s solution to personal data protection at a contractual and regulatory requirements level is 100% covered. The following table shows how the new requirements are resolved.

Data Protection Requirements	solutions
No personal identifiers should be retained in data stored for long term analytics or machine learning.	Tokenized, or encrypted data to create de-identified data sets.
Access to network and device performance data must be limited to the purposes of performing its obligations to customers and may be used in an aggregated and de-identified fashion to develop and improve algorithms, products and services, which developments and improvements may be provided to any customer.	Strictly defined user roles.
[COMPANY] will require logical and/or physical separation of test, development and production systems.	Backed into [COMPANY]'s engineering development cycle.
[COMPANY] will not use customer data in any	Retained only de-identified data.

Data Protection Requirements	solutions
development or testing environments	
Any access to personal data must be authorized and audited.	Strictly defined user roles.
[COMPANY] may collect Personal Data in order to provide the benefits of Cloud Services to CSPs and its end users.	Retained only de-identified data.
[COMPANY] may collect minimal personal data absolutely required for business and technical purposes, and use it only for the purpose disclosed to and consented by the CSP customer or consumer.	Retained and use only de-identified data for a maximum defined retention time.
Personal Data must be deleted upon customer account deletion.	This is part of the enforced retention definition rules.
Encryption of (personal) Data at Rest and Transit.	Tokenized, or encrypted data at rest or in transit.

How [COMPANY] de-identifies data

As data from several customer locations is ingested and stored in [COMPANY]'s production cloud, it is also ingested into [COMPANY]'s data lake at periodic intervals.

Data management classification: Red, Amber, Blue, and Green

[COMPANY] manages all information in accordance with [COMPANY]'s information classification and retention policy. This means that information must be classified and handled based on its value and sensitivity. The classification levels determine what baseline data protection safeguards are appropriate when handling information.

In order to easily apply data conditions and rules, [COMPANY] classifies personal data into different color categories:

- RED raw data
  Confidential—data with clear text personal identifiers that are collected from customer locations.
- AMBER pseudonymized data
  Internal—encrypted personal data using symmetric encryption algorithms; auditable records of customer authorization or reported customer service issues.
  Customer-defined data retention and residency requirements.

- BLUE pseudonymized data
  Internal—tokenized data that preserves statistical qualities of the original but cannot be re-identified.
  Data is protected and reasonably de-identified.

- GREEN anonymized data
  Public—aggregated and/or differentially private data that can be released to the public.
  Data is protected and anonymized.

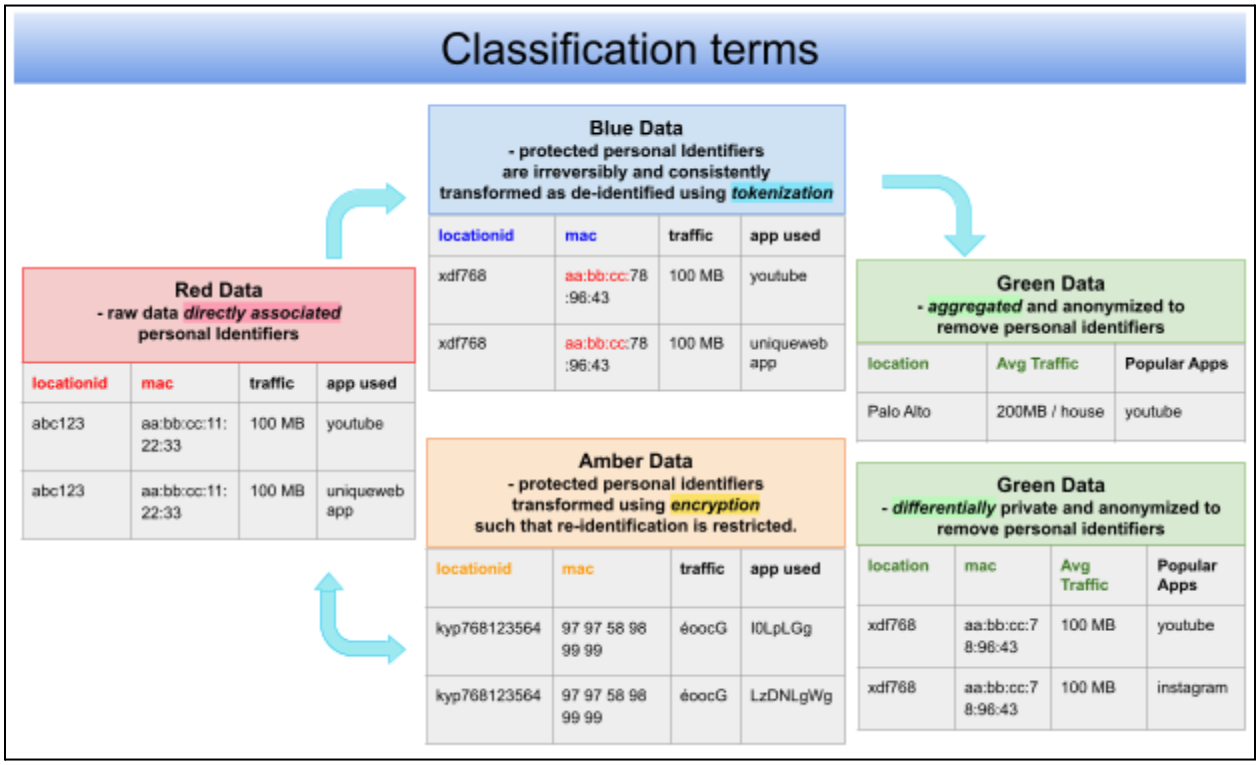

Figure: Classification terms and examples

Protected data use cases for enhancements and ML improvements

Our business use cases are for experience management (analytics) and transactions. Specifically we use:

- tokenized and de-identified (Blue-zone) data for data analysis for service insights, product improvement and machine learning, and
- encrypted and re-identification (Amber-zone) data for troubleshooting or transactional scenarios.

All transactional data is protected and either de-identified or re-identified using the third-party solution, Tonic.ai Data De-identification.

In the next phase, we plan to encrypt at a field level all personal identifiers which are ingested into the Amber-data zone, tightly control (system only or very limited human-access) the ability to decrypt personal identifier fields and map to Red. Audit access for decryption activity.

In the next phase, we plan to:

- encrypt Amber-data personal identifier data fields
- add an audit-access for decryption activity
- strict access controls to limit who can decrypt Amber-classified data
- strict access controls to limit who can reverse the data back into Red-classified data.

How [COMPANY] implements data protection

Dual ingestion pipelines to our data lake is critical to implement the Red-to-Blue classification criteria. In one of the pipelines:

- writes service and usage data into buckets within the “Amber” data zone wherein personal data is encrypted using a symmetric encryption algorithm with a customer managed key (CMK) managed in AWS key management service ~~(KMS)~~.
- CMKs are region specific ensuring logical isolation of data within AWS regions.

Access within the Amber-zone is restricted to machine users and shortlisted privileged “human” roles based on auditable customer support requests.

In the other pipeline:

- transforms personal identifiers in service and usage data using a data transformation policy.
- writes the tokenized data with preserved statistical qualities to the “Blue” data zone.

Access within the Blue-zone is authorized and granted to machine users and dev/test “human” roles based on auditable access records.

We process all raw personal identifiers (Red) so that data moves into the Amber-data zone. The Amber-data zone enforces tightly controlled (system only) access and audit controls thereby restricts mapping to Red.

- We create a closed loop for data aggregation and enrichment pipelines.
- We apply customer-defined retention and residency requirements to the Amber-data zone.

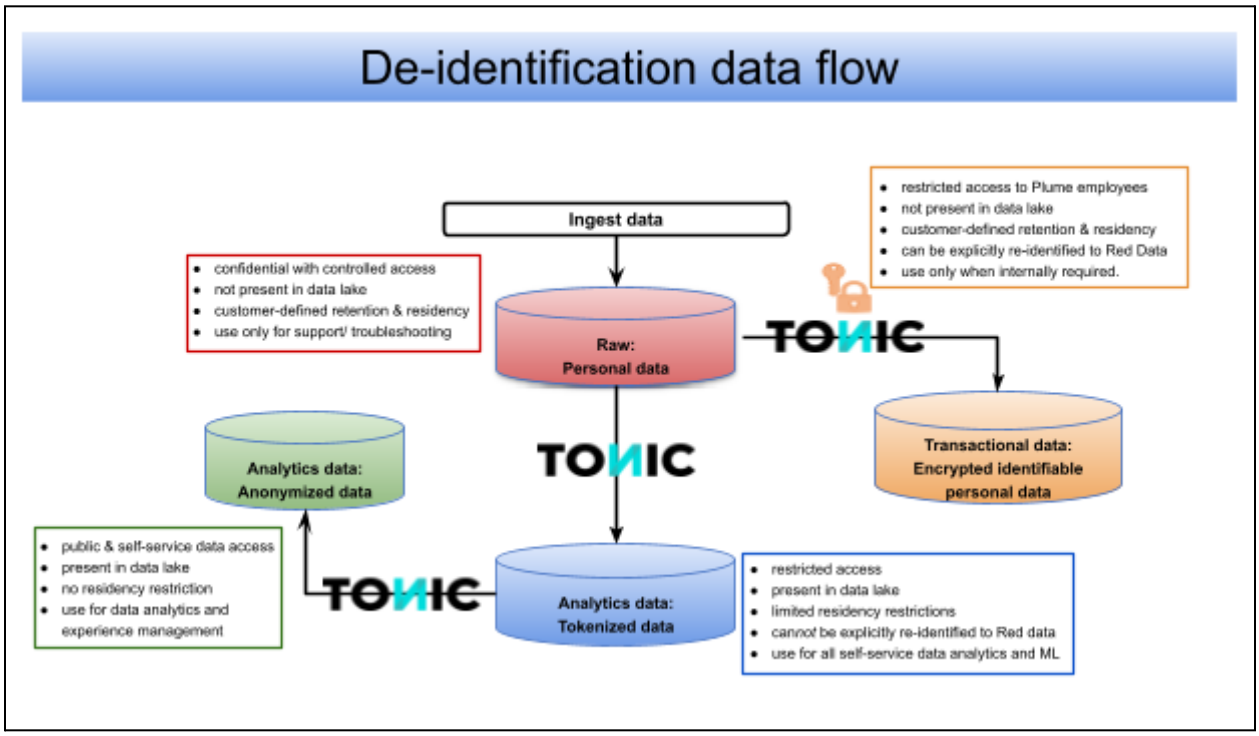

Figure: Data de-identification flowchart

Even after tokenizing personal identifiers (e.g. mac address, device ID, etc) in the blue data zone, two sources of risk still exist which can result in singling out an individual:

1. Quasi-identifiers such as device type, derived geoIP/location, person_profile_id, age or gender can be used alongwith public data to re-identify an individual. For example, Sweeney was famously able to re-identify Massachussette's Governor Bill Weld's data in anonymized medical records by comparing zip code and birthdates in the anonymized records to public census data.
2. Sparse event trajectories offer another path towards re-identifying an individual. The canonical example is the re-identification of users in the anonymized Netflix Challenge Dataset: where by it was observed that the sequence of movies that a user rates are often highly unique, and comparing these sequences with publicly available data — eg,

IMDB ratings — one can re-identify many users present in both datasets, even in the presence of noise added to dates and ratings.

We focused on the risk of singling out an individual from our tokenized datasets, which is implemented using differential privacy and methods tailored to [COMPANY]'s requirements. This includes converting tokenized Blue-data to anonymized Green-data using the following methods:

- Add randomized responses to the k-anonymized clusters— it's as if we flip a weighted coin for each user. If heads, that user's k-anonymized attributes are published, and if tails that user's attributes are replaced with a randomized response.
- Randomly permute events between users within a cluster— after users are clustered into groups of sanitized quasi-identifiers, we permute the events associated with that cluster between users within the cluster.

Following this Blue data transformation to Green we introduce self-service analytics as possible to operate on the Green data zone.

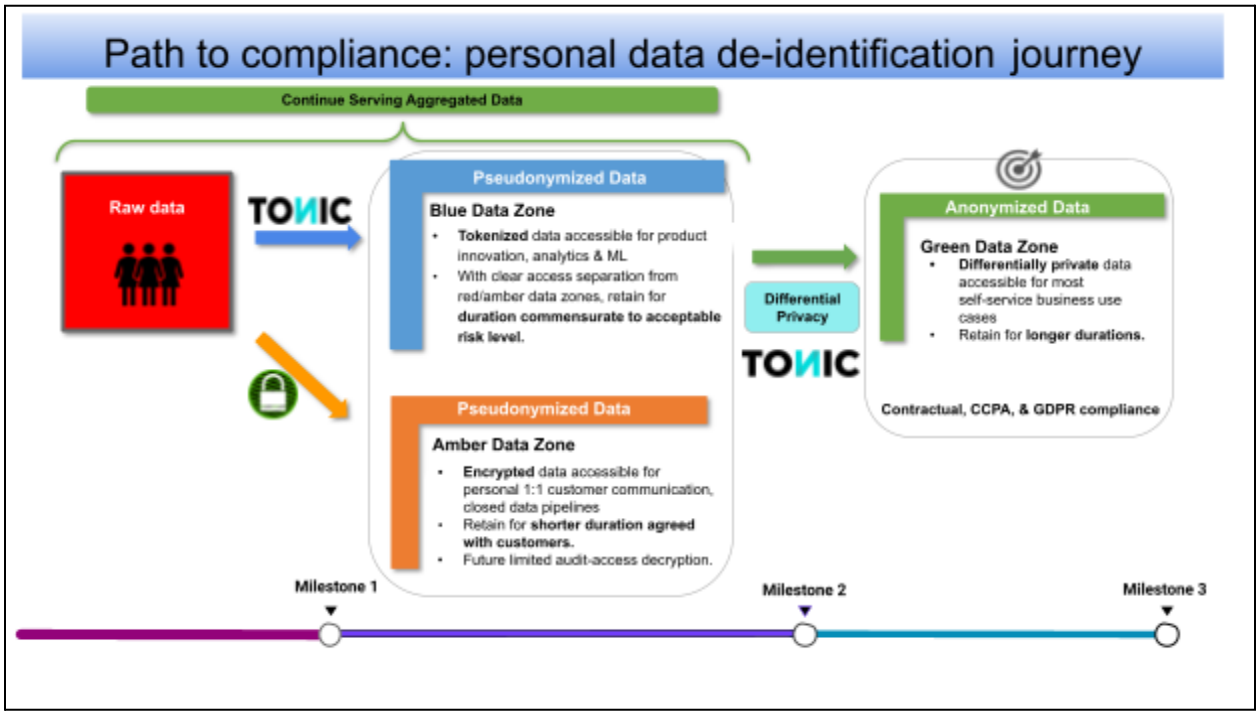

Figure: Plans for data protection compliance with our de-identification approach

Path to Compliance

So far we have discussed security solutions, personal data protection, and our data classification with methods to protect raw data. In this section we discuss another re-identification model based on risks due to quasi-identifiers and singling out.

About [COMPANY]'s data protection solution

[COMPANY] products power several millions of active households around the world which continues to push the scale of data we process at an accelerating pace.

Our solution to move streaming and batched service and usage data into a data lake, considers the following:

- **The Kickoff Data Transformation**
  We need to perform a point in time data transformation of all production data into the Amber and Blue data zones across different AWS regions. The destination bucket path and table schemas must be defined to correspond with the appropriate data zone and source schemas respectively. Personal data field conversion policies for each database table are mapped out and executed using scheduled Spark ingestion jobs.

- **Incremental Data Transformation**
  Following the Big Bang data transformation we incrementally convert raw data to Blue / Amber data based on any new writes into existing tables or when new schemas are provisioned in Production.

The Blue and Amber-data undergoes automated data quality regression tests before being made available for wider-consumption.

How long [COMPANY] keeps personal data

The length of time we keep personal data depends on the following criteria:

- business need for which it was collected
- contractually required duration
- legally required retention for certain transactional records. .

At a minimum, personal data is deleted from our production cloud every 30 days or when a customer account is terminated. At best, we retain de-identified data for longer term analytics.

Summary

Our data privacy improvement plans include data anonymization on top of de-identifying the personal data. Our commitment to protecting your personal and business data extends to all individuals who interact with us: CSPs, business partners, vendors, leads and prospects, and users.

personal data de-identification whitepaper

[COMPANY] Strategy WhitePaper Content -
https://confluence.[COMPANY].tech/display/SEC/Personal+Data+De-Identification+Strategy
Doc Plan (timeline)
https://drive.google.com/file/d/1P8TmxlpkI4PFp6s46I9KhbRPsOQfxs8d/view?usp=sharing
Solution Arch whitepaper diagram
https://confluence.[COMPANY].tech/display/SEC/For+white-paper%3A+red+to+blue+and+amber+data+lake+conversion?moved=true

Glossary

Anonymize is permanently and completely removing personal identifiers from data, such as converting personally identifiable information into aggregated data. Anonymized data is data that can no longer be associated with an individual in any manner. Once this data is stripped of personally identifying elements, those elements can never be re-associated with the data or the underlying individual.

Data privacy is a subset of privacy and refers to the rules we apply to handling personal data. Data privacy defines the policies that data protection tools and processes employ and is concerned with the proper handling of data (e.g. collection, consent, use, transfer to third parties etc.), particularly under regulatory obligation.

For protection, it is up to the companies handling data to ensure that it remains private. Data privacy is a legal issue and data protection is essentially a technical issue. Because this paper is on “data privacy” we refer only to *privacy* to imply *data privacy*.

Data protection is concerned with the unauthorized use, corruption, loss, availability of the personal data.

De-identify is removing personally identifying information in order to protect personal privacy. In some definitions, de-identified data may not necessarily be anonymized data (as we have defined that term in this document). This may mean that the personally identifying information may be able to be re-associated with the data at a later time. In such cases, anonymized data is a particularized subset of de-identified data. In this document, "de-identified" and "anonymized" will be considered synonymous terms.
NOTE: This definition excludes references to health information in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

Data is considered de-identified under the Privacy Rule when a number of specified data elements are removed. (45 C.F.R. §§ 164.502(d)(2), 164.514(a) and (b).) De-identified data is not regulated by HIPAA and may be shared without restriction.

Limited data sets are stripped of many categories of identifying information but retain information often needed for public health and research (such as birth dates, dates of treatment, and some geographic data). (45 C.F.R. § 164.514(e).).

Data Handling is a process of gathering, recording, and presenting information in a way that is helpful to others in using (read or fetch information) for instance, in graphs or charts. ... **It** is also used for comparing **data** and taking out mean, median, and mode. Which is useful for both maths and science.

Differential privacy transforms information shared with a company before it ever leaves the user’s device, such that the company can never reproduce the true data.

Encryption at rest and in **transit** are both data protection concepts. Data can be exposed to risks both in **transit** and at **rest** and requires protection in both states.

- For protecting data in **transit**, you can mount a file system so that all NFS traffic is encrypted in transit using Transport Layer Security 1.2 (TLS) with an industry-standard AES-256 cipher.
- For protecting data at **rest**, enterprises can **encrypt** sensitive files prior to storing them and/or choose to **encrypt** the storage drive itself.

Host hardening has several means of computer security such as limiting network access to a system by the traditional method of turning off unnecessary network services, by firewalling, or by enforcing authentication to use a service.

Indirect personal data differs from direct personal data. This includes data such as names, identity numbers, telephone numbers, email addresses, and in many cases even a person's postal address or bank account number.

This includes **data** such as names, identity numbers, telephone numbers, email addresses, and in many cases even a person's postal address or bank account number.

Privacy is a state in which one is not observed or disturbed by other people.

Security refers to preventing unauthorized access to personal information, through technologies like network security, firewalls, encryption, etc..

References

General Data Protection Regulation (GDPR)

Art. 5 GDPR Principles relating to processing of personal data.

California Consumer Privacy Act (CCPA)

Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)

Tonic.ai Data De-identification Solution

https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport

https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices/at_download/fullReport

https://tanker.io/tanker-whitepaper.pdf

personal data de-identification whitepaper

https://static1.squarespace.com/static/58793ca79de4bb30cf853894/t/5ad8ecb6f950b720c0d95fe9/1524165814712/Salesforce+Security+White+Paper+for+Salesforce+Government+Cloud_Final_02142017.pdf

https://a.sfdcstatic.com/content/dam/www/ocms/assets/pdf/platform/whitepaper-platform-shield.pdf

https://www.mckinsey.com/business-functions/risk/our-insights/the-consumer-data-opportunity-and-the-privacy-imperative#

https://iapp.org/news/a/design-considerations-for-building-privacy-protecting-analytics-services

Co-authored and designed by Susan Kraft-Yorke