A CORP white paper snippet sample

By Susan Kraft-Yorke

Table of contents

Context **3**
Executive summary **3**
CSP Viewpoint **4**
Walking the tightrope 5
CORP data privacy and security solutions 5
Strategy for preserving the privacy of data analytics and product innovation **6**
How CORP collects data 6
How CORP uses data 7
How CORP classifies information for data management 7
How CORP refines data-use cases for product improvements 8
How CORP implements personal data de-identification 9
How CORP anonymizes data 10
How long CORP retains data for analytics and product innovation The length of time we keep personal data depends on the following criteria: ● business needs for which it was collected, 11
Conclusion **12**
Appendix A **13**
Glossary 13
Appendix B **16**
References 16
Appendix C **17**
List of figures 17

Context

This paper describes how CORP Design Inc. (CORP) protects and utilizes customer's real-world production data. Further it describes how we continually improve our product line while keeping personal and business data private and secure.

Executive summary

As a smart home experience company focused on data management and business intelligence, CORP considers it mission critical to ensure data is always secure. We use the data we collect to benefit the user, such as improving online protection and digital well-being. Our privacy preserving strategies are in compliance with industry frameworks.

In response to market needs for an improved smart home management platform, we added cloud-based experience and data-insight services. New strategies directly correlate with the amount of time we store—and analyze—real-world and de-identified data. The time we spend developing new solutions with protected data is informed by the:

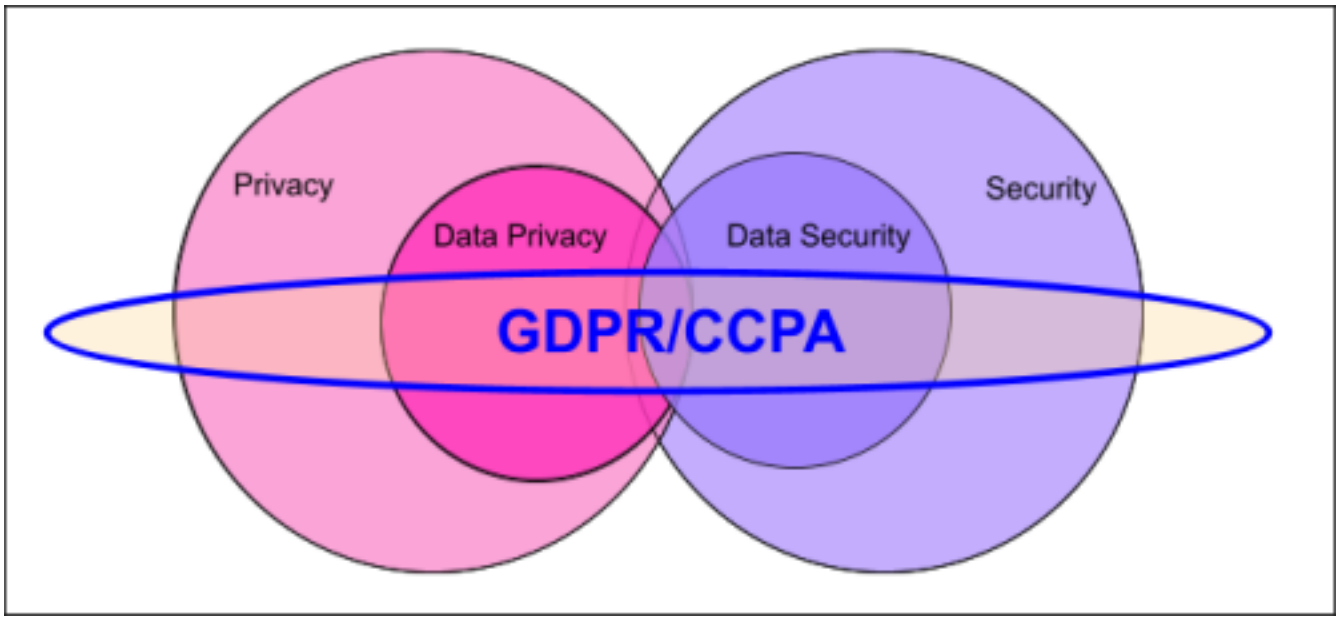
- business need for which it was collected,
- contractually required duration, and
- legally required retention for certain transactional records.

Our commitment to protecting personal and business data extends to all individuals who interact with us: Communications Service Providers (CSPs), business partners, vendors, and end-customers.

CSP Viewpoint

CORP created the first SaaS experience platform dedicated to CSPs and their subscribers (users). As a SaaS provider, our top priority has always been to protect data with security tools and individuals with data privacy (privacy) tools. We baked in strong encryption, so that data is always secure. This applies to data at rest on a device or in the cloud, and data traffic from the internet.

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 they have to be integrated with each other.



CORP security and privacy within regulation guidelines

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

Walking the tightrope

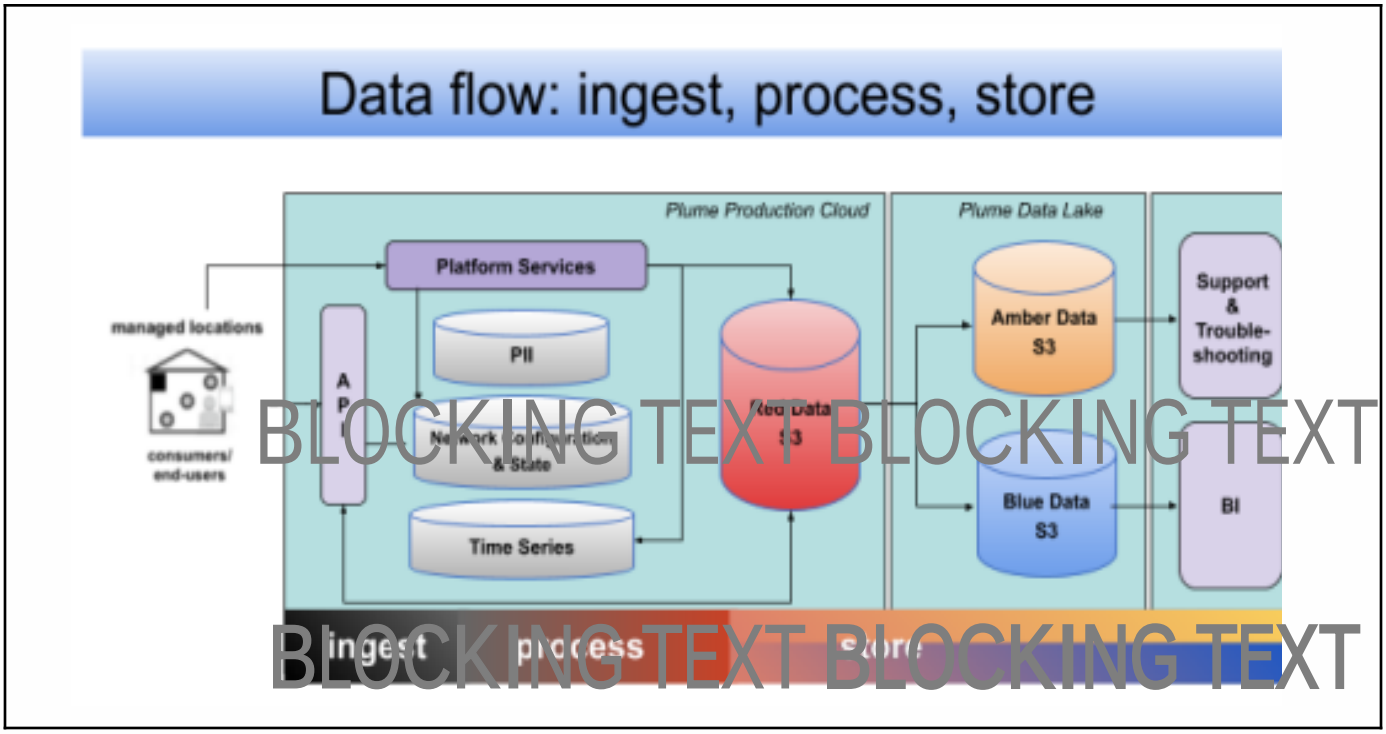
Rules and regulations can keep us up at night as we walk between what is right for our customers and what next new thing needs technical investments. The quandary is that the research required to develop that next great solution/feature depends on working with real-world data for analytics and machine learning algorithms. To ease the tension of working with real-world data collected by doing business, we protect the data with de-identification and anonymization solutions so that personal attributes in real-world data are not traceable to any one person.

CORP data privacy and security solutions

CORP's strategy for bringing personal data protection up to contractual and regulatory requirements is to fold in advanced processes on top of tried and true methodologies. The following table shows how these requirements are addresse

Strategy for preserving the privacy of data analytics and product innovation

Our privacy-preserving solutions keep personal attributes in real-world data untraceable to any one person. In this way we can continue delivering reliable and high quality insights and visualizations.



Dozens of pages removed at this point

for Susan Kraft-Yorke samples viewing purposes only.

Data flow to ingest, process, and store

How CORP collects data

OpenSync-enabled CSP Gateways and CORP Pods (network nodes) create a CORP managed WiFi network in home or small business locations. When end-users connect to the WiFi network, the network nodes report the account, network, and internet connectivity data (with personal identifiers). This information is required for monitoring and optimizing WiFi performance, providing online protection and motion features, as well as enhancing services for CSPs.

Platform services ingest and store production data from managed locations in a CORP Cloud deployment, while API services ingest and store account-profile and WiFi configuration data

from the customer mobile app at managed locations in a CORP Cloud deployment. The ingested data is stored within the PII, Time Series, Network Configuration, and State databases.

How CORP uses data

A data ingestion pipeline periodically copies select production data sets; these sets inform product innovation algorithms in the CORP data lake. This data is further prepared with data enrichment and aggregation pipelines to create dimensional models and reporting tables that:

● we deliver to consumers through authenticated mobile and Web Apps, like HomePass or WorkPass for online protection and digital well-being,
● deliver personalized trends with email or push notifications,
● are available through business intelligence dashboards, such as Panorama, CorpDash, etc.,
● we publish as aggregated and anonymized global trends on our plume.com resources page, and
● CSP customers can receive daily data exports.

How CORP classifies information for data management

CORP manages all information in accordance with CORP'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. We color-code assignments of red, amber, blue, and green to simplify associating data access conditions and rules.

The CORP personal data categories are:

● RED raw data

How long CORP retains data for analytics and product innovation **The length of time we keep personal data depends on the following criteria:**

- business needs for which it was collected**,**
- contractually required durations,
- legally required retention for certain transactional records

.

Every 30 days, or when a customer account is terminated, raw personal data is deleted from our production clouds. Anonymized and de-identified data is retained for more than a month for long-term analytics in our development and test environments.

Our personal data categories and the associated retention approach we have taken for the CORP data lake is:

- AMBER pseudonymized data (**Internal**)

is protected and encrypted; it is retained in the CORP data lake for a shorter duration as agreed by the customer, and not for more than 24 months.

- BLUE pseudonymized data (**Internal**)

is protected and reasonably de-identified; it is retained in the CORP data lake for longer durations than Amber data, commensurate to CORP-acceptable risk levels.

- GREEN anonymized data (**Public**)

is protected and anonymized; it is retained in the CORP data lake and on public websites for longer durations than Blue data, commensurate to CORP-acceptable risk levels.

Conclusion

In this paper we describe how CORP, in a high-growth industry, constructed a careful path to ensure our product line continues the path of innovation, problem solving, and ultimate ease of use for clients and end-users. By designating and color-coding four categories of personal data protection solutions and safety, it’s clear the course is set with internal and external expectations and requirements.

Our goals are to continue on this path of privacy-preserving strategies for data analytics and product improvements.

top

Appendix A

Glossary

Aggregate criteria of regression analysis is to investigate the common influence of several potential influence factors on the target parameter. For example, Cox regression or Poisson regression can be used for the data analysis of cohort studies, depending on the target parameter

AWS key-management service (KMS) is a secure and resilient service that uses hardware security modules that have been validated under FIPS 140-2, to protect CORP keys. AWS KMS is integrated with AWS CloudTrail to provide logs of all key usage to help meet any regulatory and compliance needs.

Anonymize is the complete and permanent removal of personal identifiers from data, such as converting personal, 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 persona identifying elements, those elements can never be re-associated with the data or the underlying individual.

Data handling occurs after the conclusion of a business function or project ensuring data is stored, used, archived or disposed in a secure and private manner ~~during.~~ This includes policy development and procedures to manage data handled electronically or by non-electronic means.

Data ingestion is a short-hand way of saying that data is being prepared for analysis. This usually includes steps to extract (taking the data from its current location), transform (cleansing and normalizing the data) and load (placing the data in a database where it can be analyzed).

Data lake is a storage repository that can store large amount of structured, semi-structured, and unstructured data.

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 personal identifying information in order to protect personal privacy. In some definitions, de-identified data may not necessarily be anonymized data (as defined in this document). This may mean that the personal identifying information may be able to be re-associated with the data at a later time.

In such cases, anonymized data is a subset of de-identified data. 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).).

NOTE: The de-identify definition excludes references to health information in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

Dimensional database is a relational database that uses a dimensional data model to organize data. This model uses fact tables and dimension tables in a star or snowflake schema. A dimensional database is the optimal type of database for data warehousing.

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.

Direct personal identifiers (*commonly known as direct personal data*) includes any pieces of information whereby an individual is directly identifiable using nothing but the information one possesses.

Host hardening has several meanings in computer security such as limiting network access to a system by turning off unnecessary network services, by firewalling, or by enforcing authentication to use a service.

Personal data is any information that relates to an **identified or identifiable living individual**. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data.

Indirect personal identifiers (*commonly known as indirect personal data*) includes any pieces of information whereby an individual is not directly identifiable using the information alone, but one may be able to identify the individual by using other information obtained from a *reasonably* accessible source.

PII database is secured such that links between it and the rest of our cloud data can only be accessed by our API software.

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

S3 or Amazon Simple Storage Service is a service offered by Amazon Web Services (AWS) that provides object storage through a web service interface. Amazon S3 uses the same scalable storage infrastructure that Amazon.com uses to run its global e-commerce network.

S3 data lake—*see* S3 and data lake definitions.

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

State databases hold the last known committed value for any given key. They are populated when each peer validates and commits a transaction. The state database can always be rebuilt from re-processing the ledger.

Time series databases store data as a sequence of data points collected over time intervals, giving the ability to track changes over time. By storing data in this way, it makes it easy to analyze time series, or a sequence of points recorded in order over time, efficiently and continuously add, process, and track massive quantities of real-time data with speed and precision.

Tokenized data has no meaningful value. It is generated by a process wherein personal identifiers are replaced by randomly-generated values, known as tokens, without having any mathematical relationship with the original identifiers. Data in the “Blue” zone criteria has preserved statistical qualities in addition to being tokenized data. In some cases, the statistical quality to join datasets across multiple tables is preserved by generating consistent tokens for the

Remaining pages removed after this point

for Susan Kraft-Yorke samples viewing purposes only.