WHITE PAPER

Beyond Cosmetic Compliance in Data Analytics: A Guide to CPRA

2021

IMMUTA™

Table of Contents

The California Consumer Privacy Act (CCPA), is the most important state-level privacy framework in the US. While it has often been described as a "light" version of GDPR, it introduces a similar list of privacy rights (information, access, deletion, and opt-out rights, among others).

On November 3, 2020 however, the California Privacy Rights Act (CPRA), passed with 56.1% of the vote in California, meaning it is on track to become effective in California in January 2023 and amend CCPA only three years after its entry into force. Of note, CPRA contains a 12-month lookback period. This means that compliance should start on January 1, 2022, so that covered businesses can respond to consumers' requests once CPRA goes into effect on January 1, 2023, with enforcement scheduled to begin on July 1, 2023.

On the whole, CPRA brings CCPA closer to GDPR, at least for large organizations. A few things stand out:

- **First**, CPRA sets up a new regulator called the California Privacy Protection Agency (PPA). While individuals maintain the right of private action, PPA will be responsible for primary enforcement. PPA has the power to impose administrative fines, up to $2,500 *per* violation, and triples them — $7,500 per violation — for violations concerning under-16 minors.

- **Second**, key data protection principles under GDPR are now expressly acknowledged or strengthened by CPRA. This includes data minimization, which requires that organizations "only process the amount of data that is reasonably necessary and proportionate to achieve your purpose," and purpose limitation, which compels data teams to "only process the data for a predetermined or compatible purpose." Of note, for under-16 minors, opt-in consent must be obtained for "*narrowly defined particular purposes*."

- **Third**, CPRA enlarges the list of rights granted to consumers. It introduces new rights, such as rectification and restriction, and extends the opt-out right to apply to data exchanges characterized as either "sales" or "sharing." Processing of sensitive information for legitimate business purposes is more restricted than under CCPA, and the definition of consent is now similar to GDPR's definition.

Since CPRA will now drive compliance efforts in the US and privacy laws are progressively converging at the global level, it is crucial to understand that 'cosmetic compliance' will cease to be sufficient. There have been countless examples proving internal compliance structures do not necessarily prevent unlawful conduct and, in many instances, only serve as a 'window-dressing function.'[1]

Modern privacy laws imply a restructuring of organizational processes in such a way that they are driven from the start by a series of privacy goals, including purpose limitation, data minimization, storage limitation, security (confidentiality, availability, integrity), transparency, and intervenability. They thus require clear and actionable or enforceable compliance requirements.

Within data analytics environments, the main challenge with setting forth such requirements is that doing so requires a delicate balance in order to avoid creating tension and frustration among data scientists. Data scientists want access to data to do their jobs and are likely to become nervous when put on a data diet. Data engineers, acting as mediators, thus play a key role in activating meaningful rules and 'coding' them within the data analytics environment itself to avoid process bottlenecks and rules rejection.

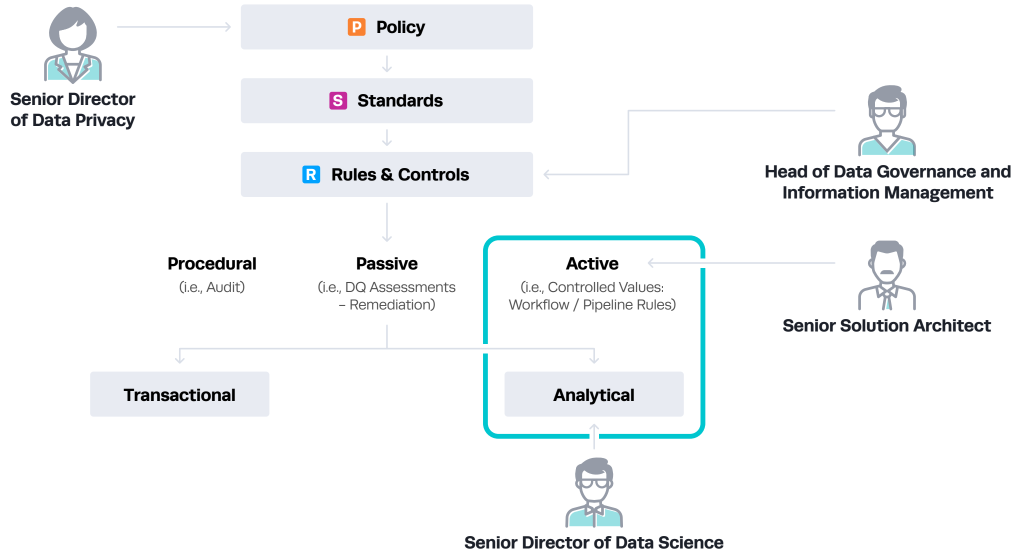
This white paper addresses data engineers and illustrates how to leverage automation to softly but surely force CPRA compliance within data science environments. First, we'll show how self-executing policies can be designed to meet a variety of privacy goals, as well as support consumer rights under CPRA. Second, we'll explain how smart reporting can improve the quality of monitoring and auditing, which are often neglected functions within organizations that see compliance as purely cosmetic. Yet, these functions are essential for privacy teams to be able to demonstrate compliance to regulators, judges, or more simply to consumers and the wider public through privacy policies and notices.

[1] See e.g., Kimberly D Krawiec, 'Cosmetic Compliance and the Failure of Negotiated Governance' (2003) 81 Washington University law quarterly. Washington University. School of Law 487.

1.0 Self-Executing Policies

Diagram 1 illustrates what usually happens within analytics and data science environments, i.e., a three-step process starting with the drafting of policies, followed by the formulation of standards, and then the creation of rules and controls.

This is what we call the PSR (Policy-Standard-Rule) process. Policies are usually set by the legal or compliance team, most likely under the supervision of the Director of Data Privacy or their equivalent. Generic recommendations such as "*data should be anonymized as often as possible*" will usually be produced at this stage. More specific standards will then be expressed at a second stage, such as a list of available anonymization techniques.



Senior Director of Data Privacy
P Policy
S Standards
R Rules & Controls
Head of Data Governance and Information Management
Procedural
(i.e., Audit)
Passive
(i.e., DQ Assessments – Remediation)
Active
(i.e., Controlled Values: Workflow / Pipeline Rules)
Senior Solution Architect
Transactional
Analytical
Senior Director of Data Science

Diagram 1: A typical compliance journey within data-driven organizations.

These high-level recommendations should then be translated by an intermediate team in order to generate executable rules and controls for each data science project. This intermediate team is often led by the Data Governance Team, acting under the supervision of the Head of Data Governance. Rules and controls will finally have to be enforced, which is also the point at which confusion suddenly grows because privacy goals often appear undermining to data scientists.

By empowering data engineers with a set of 'automated policies,' it is possible to accelerate the mediation between the privacy and data science teams and make compliance much more effective. An automated policy is a policy that is both meaningful for compliance personnel and data scientists and granular enough to be immediately enforceable. With an automated policy, the PSR process is shortened. Once the automated policy is activated negotiation ceases and data scientists are not be able to claim ignorance of the rule anymore.

Let's go back to CPRA and give examples of useful automated policies in this context.

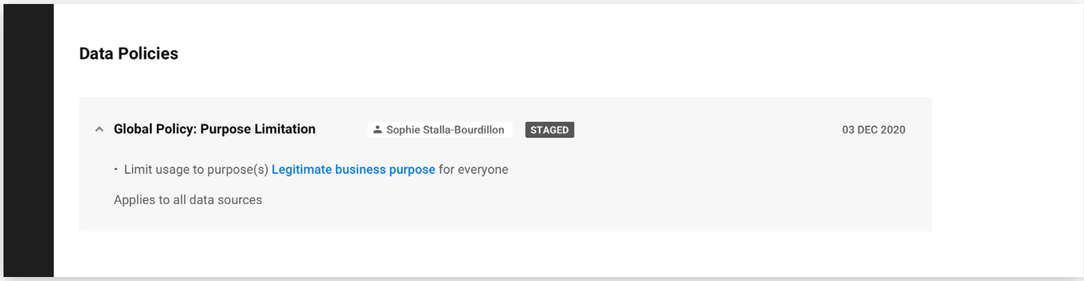
1.1 Purpose limitation

CPRA, compared with CCPA, expresses the goal of purpose limitation in a much stronger fashion. According to the introductory section of the CPRA (section 3):

> “Businesses should only collect consumers' personal information for specific, explicit, and legitimate disclosed purposes, and should not further collect, use, or disclose consumers' personal information for reasons incompatible with those purposes.”

In section 4 (new section 1798.100(a)(2) of the Civil Code) it is provided that *“A business shall not collect additional categories of personal Information or use personal information collected for additional purposes that are incompatible with the disclosed purpose for which the personal information was collected, without providing the consumer with notice consistent with this section.”*

Limiting data processing to legitimate and specific purposes is thus a must-have. Imagine a data governance interface which would allow a data engineer to activate an automated policy. This is what the policy could say to effectuate the privacy goal of purpose limitation:



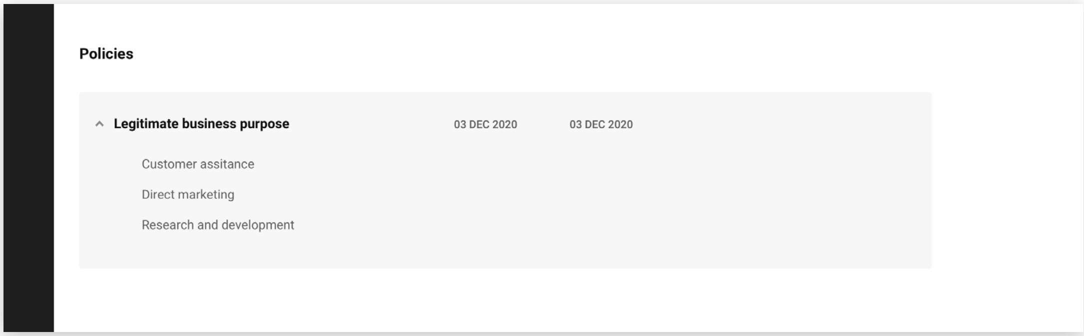
Data Policies

Global Policy: Purpose Limitation — Sophie Stalla-Bourdillon — STAGED — 03 DEC 2020

- Limit usage to purpose(s) Legitimate business purpose for everyone

Applies to all data sources

The list of legitimate business purposes could also be pre-populated so that data scientists can only choose predetermined purposes. In the screenshot below purposes are organized by domains of activity.



Policies

Legitimate business purpose — 03 DEC 2020 — 03 DEC 2020

Customer assitance

Direct marketing

Research and development

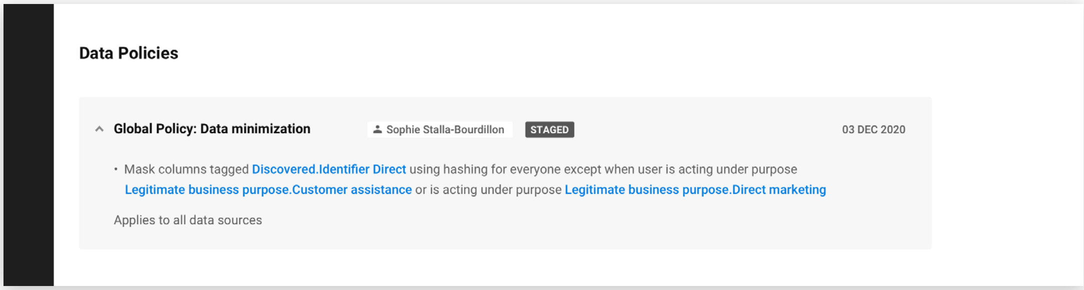
Data scientists could also be asked to add a description when attaching a purpose to their project in order to make it more specific.

1.2 Data minimization

CPRA introduces a requirement of data minimization in section 4 (new section 1798.100(c)):

> “A business's collection, use, retention, and sharing of a consumer's personal Information shall be reasonably necessary and proportionate to achieve the purposes for which the personal information was collected or processed, or for another disclosed purpose that is compatible with the context in which the personal information was collected, and not further processed in a manner that is incompatible with those purposes.”

There are different ways to meet the data minimization goal through an automated policy. One possibility is to set the following rule:



Data Policies

Global Policy: Data minimization Sophie Stalla-Bourdillon STAGED 03 DEC 2020

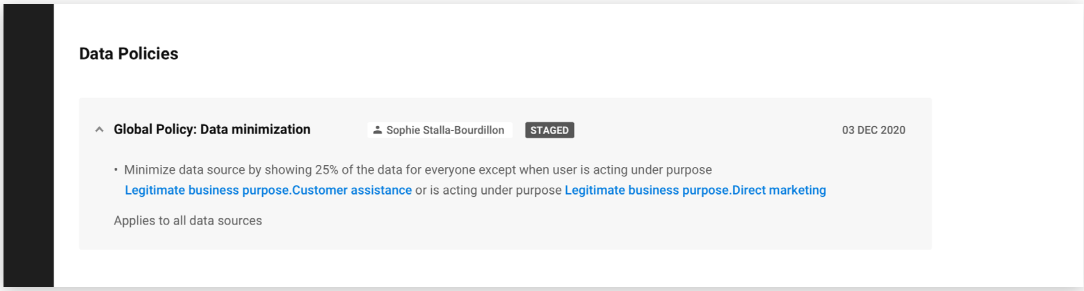
- Mask columns tagged Discovered.Identifier Direct using hashing for everyone except when user is acting under purpose Legitimate business purpose.Customer assistance or is acting under purpose Legitimate business purpose.Direct marketing

Applies to all data sources

With this policy, data scientists would never be able to access data tagged as direct identifiers when conducting projects for research and development purposes.

To add to this first data minimization layer, a data engineer could decide that data scientists could only analyse a certain percentage of the data.

This is what such an automated policy could read:



Data Policies

Global Policy: Data minimization Sophie Stalla-Bourdillon STAGED 03 DEC 2020

- Minimize data source by showing 25% of the data for everyone except when user is acting under purpose Legitimate business purpose.Customer assistance or is acting under purpose Legitimate business purpose.Direct marketing

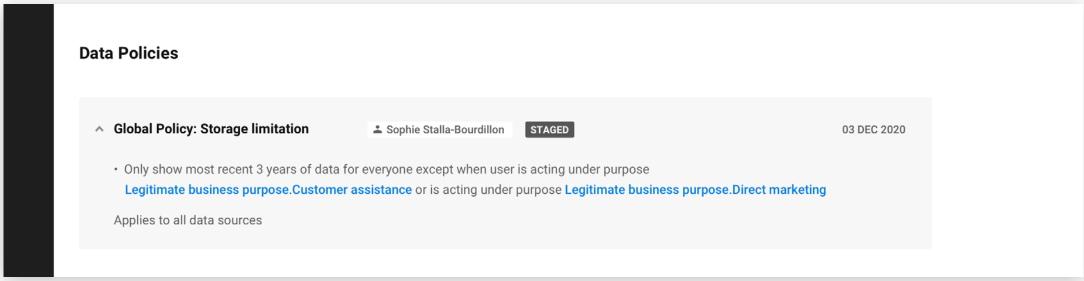
Applies to all data sources

1.3 Storage limitation

CPRA also introduces a requirement of storage limitation as per section 4 (new section 1798.100(d)), which requires setting time-based policies for processing activities:

> "a business shall not retain a consumer's personal information or sensitive personal information for each disclosed purpose for which the personal information was collected for longer than is reasonably necessary for that disclosed purpose."

An automated policy enforcing such a privacy goal could read as follows:



Data Policies

Global Policy: Storage limitation Sophie Stalla-Bourdillon STAGED 03 DEC 2020

- Only show most recent 3 years of data for everyone except when user is acting under purpose Legitimate business purpose.Customer assistance or is acting under purpose Legitimate business purpose.Direct marketing

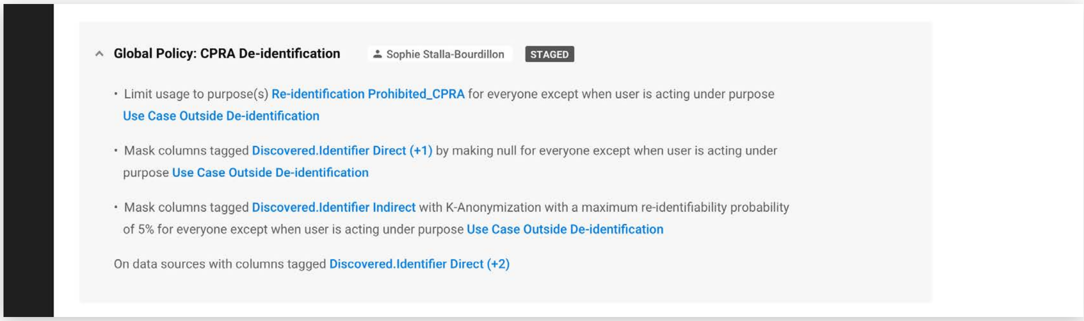
Applies to all data sources

It should be possible to tailor even more time-based policies per project by making sure that once a data science project is finished, access to the data is terminated.

1.4 Pseudonymization, de-identification, and aggregation

CPRA confirms the relevance of pseudonymization, de-identification, and aggregation techniques to protect the data against privacy threats. It is worth mentioning that the definition of de-identified data has been slightly amended in an attempt to offer greater flexibility to data controllers. With this said, even though both data controls (data transformation techniques) and context controls (e.g., business processes) must still be combined, there is room for automation on that front.

Once a data transformation technique has been selected, it is possible to author an automated policy to enforce the CPRA de-identification requirement. Let's assume that k-anonymization, which enables individuals to hide in groups of a k number of individuals, is supported by the data governance interface the data engineer is using. This is what the automated policy could read:



Global Policy: CPRA De-identification Sophie Stalla-Bourdillon STAGED

- Limit usage to purpose(s) Re-identification Prohibited_CPRA for everyone except when user is acting under purpose Use Case Outside De-identification
- Mask columns tagged Discovered.Identifier Direct (+1) by making null for everyone except when user is acting under purpose Use Case Outside De-identification
- Mask columns tagged Discovered.Identifier Indirect with K-Anonymization with a maximum re-identifiability probability of 5% for everyone except when user is acting under purpose Use Case Outside De-identification

On data sources with columns tagged Discovered.Identifier Direct (+2)

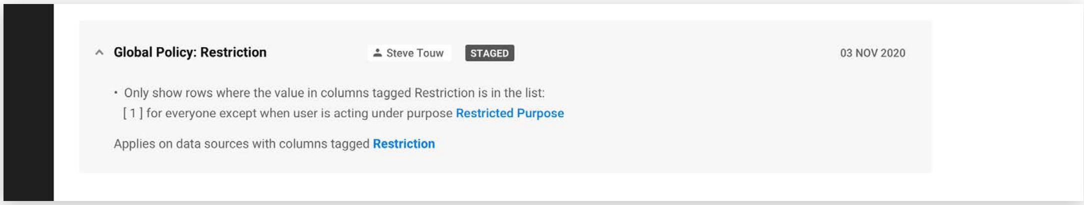
This technique relieves data engineers of having to start from scratch or build policies manually for each new addition to a data set, enhancing data privacy while accelerating speed to access.

1.5 Consumer rights

Creating automated policies to effectuate consumer requests is also an option data engineers should closely consider.

As much as the right to opt out and the right to deletion can be usefully supported by automated policies, the rights to rectification and restriction can also be expressed in automated policies easily digestible by compliance personnel.

Here is an example:



Global Policy: Restriction

Steve Touw

STAGED

03 NOV 2020

- Only show rows where the value in columns tagged Restriction is in the list: [1] for everyone except when user is acting under purpose **Restricted Purpose**

Applies on data sources with columns tagged **Restriction**

This increases transparency and collaborative communication between data and compliance teams, in addition to building consumer confidence in their data's use.

2.0 Smart Reporting

Once automated policies are created and activated, a mediating data engineer mindful of compliance requests will want to produce evidence of the actions taken to meet privacy goals within the data science environment being regulated.

This is where working with a data governance interface that is able to both directly impact the way the data is being accessed by data scientists and produce reports to document that impact in a way that is meaningful to compliance and privacy teams makes all the difference.

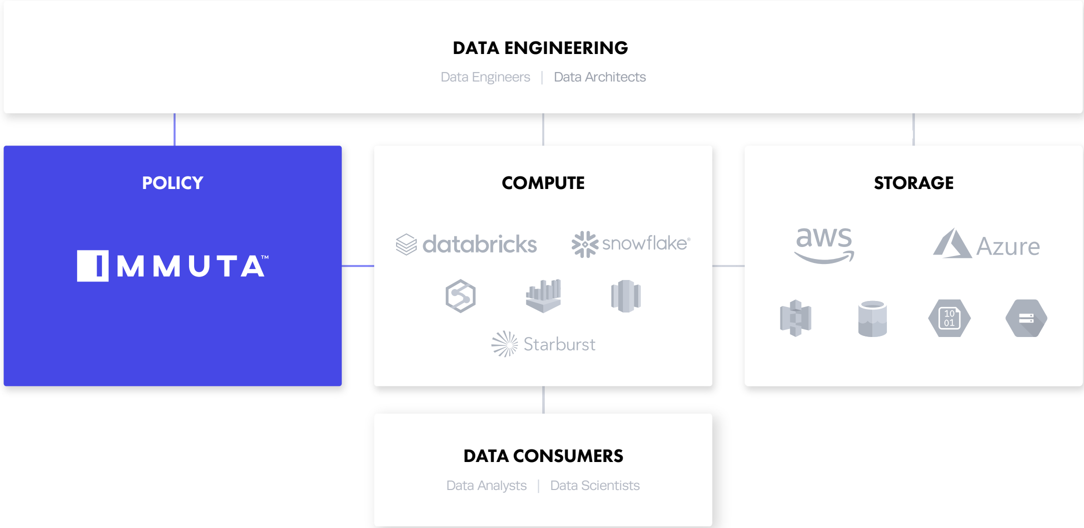
Reporting is, however, usually much more convoluted in practice. Privacy and compliance teams tend to rely upon their own software interface to document processing activities and usually ask data engineers to fill in long questionnaires without fully understanding the true implications of the descriptions inputted and, more importantly, without being kept up to date.

In a CPRA age, smart reporting is becoming essential. Contrary to CCPA, CPRA is much more prescriptive with regard to the content of the contract concluded between a covered business and a data recipient. In fact, even if a data recipient is not subject to CPRA, a covered business subject to CPRA must impose similar obligations upon its data recipients through contract. What is more, CPRA acknowledges that covered businesses can legitimately claim a contractual right to monitor the data recipient's compliance with the contract through “measures including, but not limited to, ongoing manual reviews and automated scans, and regular assessments, audits, or other technical and operational testing at least once every twelve months.” PPA has also been asked to appoint a Chief Privacy Auditor to conduct audits of businesses to ensure compliance with CPRA.

Importantly, as CPRA introduces a new category of sensitive personal information, it behooves data teams to have an efficient, comprehensive system in place for demonstrating that these data items have been properly identified and taken care of. Sensitive personal information is broadly defined and includes, among other things, information that reveals: a consumer's precise geolocation; a consumer's racial or ethnic origin, religious or philosophical beliefs, or union membership; the contents of a consumer's mail, email and text messages, unless the business is the intended recipient of the communication; a consumer's genetic data; the processing of biometric information for the purpose of uniquely identifying a consumer; personal information collected and analyzed concerning a consumer's health; personal information collected and analyzed concerning a consumer's sex life or sexual orientation.

Imagine a data governance interface which could transform data on the fly to make sure data scientists only get access to what they should have access to and that is able to produce documentation that directly speaks to compliance and privacy teams. With such an interface, global reports could be produced to support the creation of data inventories, which need to be generated before privacy policies can be drafted, and risk assessments. Local reports could also be produced per project to evaluate, for example, the quality of the risk assessment at the project level — not only at the platform level — or to demonstrate that consumer requests have been acted upon. These global and local reports could then be automatically updated over time as data usage evolves.

Does such an all-encompassing data governance interface exist? It does. Immuta can act as that interface, making it possible for data engineers to author and activate self-executing, automated policies to meet privacy goals within data science environments and produce documentation appealing to compliance and privacy teams.


DATA ENGINEERING
Data Engineers | Data Architects
POLICY
IMMUTA
COMPUTE
databricks
snowflake
Starburst
STORAGE
aws
Azure
DATA CONSUMERS
Data Analysts | Data Scientists