Bessemer
Venture
Partners

How data privacy engineering will prevent future data oil spills

Bessemer reveals the seven sins of data collection and the emerging markets that'll help companies protect and secure consumer data.

SEPTEMBER 2019

CO-AUTHORS
Alex Ferrara
Jules Schwerin
Mary D'Onofrio

The websites and apps we visit, the phones we use, and even the robot vacuums that feel like our friendly pets, all collect data. As consumers, we've become increasingly comfortable making small compromises, forking over tidbits of our personal data in exchange for "free" services. It's a trend that feels as common as the terms and conditions box on the latest app du jour. Simply put, tech companies have become unbelievably adept at collecting and analyzing our data, and that's quite literally by design. Data is the new oil. No matter how cliché the phrase may be, publicly listed technology firms now make up more than a quarter of the US stock market—so objectively speaking, it's true!

While data has permeated our lives and the economy, there are only a handful of leaders and entrepreneurs who are talking about the consequences of this new reality. And since data is the new oil, we believe it also has the potential to cause the next "data oil spill".

> While many tech companies were architected to collect data, they were not necessarily architected to safely store data. Today there's not just a rift, but a chasm between where data privacy technology, processes, and regulations should be and where they are, thus creating massive amounts of "privacy debt."

In 2018 alone, nearly half a billion records were exposed—that's not only half a billion passwords, credit card numbers and other relatively mundane pieces of our lives, but also medical test results, location history, home addresses, and all sorts of other deeply personal data. In what is beginning to seem like a never ending news cycle of breached consumer data, some of the most reputable brands from Marriott to Equifax to British Airways have been breached. (#DataOilSpill might not be trending, but we think it should be!)

If the brands above were not hints, then we should say explicitly that the problem isn't just relegated to Big Tech. Now that nearly every large enterprise would call itself a data company, enabled to collect and analyze consumer data by tools largely open sourced from Big Tech, even legacy enterprises are becoming saddled with "privacy debt."

Like technical debt, privacy debt requires reworking internal systems to adapt and build to the newest standards, which will not only make consumers happier but also make companies better. Thankfully, this is happening in an emerging field called Data Privacy Engineering. Data Privacy Engineering is not a term most consumers, or even technologists, are deeply familiar with, but it is a term that we believe is bound to enter the public lexicon as this problem comes to a head. Data Privacy Engineering represents the intersection of cybersecurity, big data analytics, legal, and compliance to address the requirements necessary to collect, safely store, and ethically use consumer data.

We believe Data Privacy Engineering will become a stand alone category which will soon be, or already is, top of mind for founders and C-level executives. In this roadmap we outline The Seven Deadly Sins of Data Privacy as a means to understand how we got here and to discuss the areas of investment that inspire us most.

The data privacy quagmire

If Facebook, with its deep bench of technical talent, isn't able to prevent massive breaches of user data, how is a more traditional company supposed to cope? The proliferation of consumer data collection shows no sign of relenting, and it is crystal clear that data breaches are accelerating, with 2019 being "the worst year on record for data breach activity" despite enterprises spending an estimated $124 billion on information security this year alone.

Put plainly, companies have failed to adequately protect consumer data. For the sake of consumer privacy and to prevent future "data oil spills," every company, not just tech companies, must apply a different approach to Data Privacy Engineering and privacy operations to keep us safe from these mostly preventable disasters. The most common mistakes both Big Tech and large enterprise make come down to these Seven Sins of Data Privacy Engineering.

The Seven Sins of Data Privacy Engineering

1. Collecting too much unnecessary data and storing it in perpetuity
2. Inadequately securing customer data
3. Not knowing what data is possessed or where it is stored
4. Sharing with third parties when the policies and practices of those third parties are unknown
5. Lack of timely data breach reporting
6. Not being responsive to consumer data access requests
7. Using AI/ML on customer data without proper consent, or in a manner that introduces biases

The most gripping and pervasive issue is the first sin — companies unnecessarily collect too much data. Data collection has been a default habit for engineers and database architects (DBAs) for the past few decades. And this practice has only accelerated because of exponentially shrinking costs associated with storing data driven by massively scalable data stores, cloud adoption, and Moore's Law. In addition, engineers tend to collect more data because they don't know if an AI model could potentially benefit from it in the future.

However there was not much thought put into the question of why to store this data, for how long, or whether end user consent was required.

There is a practice in privacy engineering known as 'minimization', which involves thinking through what is the minimum customer data set a company must collect. It's considered a best practice; however, most companies are product and engineering driven and engineers tend to keep as much data as they can.

This, in turn, leads to situations in which companies often do not know what data they have, so they don't properly secure it or they share it with third parties without knowing what their policies and procedures are. When data breaches do occur, policies aren't in place to handle communication, making a bad situation even worse.

We're already seeing the aftermath of #dataoilspills

The net result of all this mishandling of consumer data is a scandal like Cambridge Analytica where data on 87 million Facebook users were shared with a third party and ultimately used for malicious targeting purposes. This mishap not only earned Zuck a front row seat to a rather unhappy Congressional panel but also a negative sentence in the court of public opinion. Later, Facebook was fined $5 billion by the Federal Trade Commission and Zuckerberg agreed to potentially be held *personally liable* for future incidents. On a more global scale, corporate directors could be held personally accountable if their company fails to uphold GDPR.

Despite these Facebook fines, consumer behavior remains consistent and there's no noticeable change on Main Street. For instance, in the wake of the Facebook-Cambridge Analytica data scandal there were around 400,000 #DeleteFacebook tweets. Yet, in that same period Facebook active users grew by around 4%. Modern day services seem to be too intertwined into the daily lives of consumers to materially change.

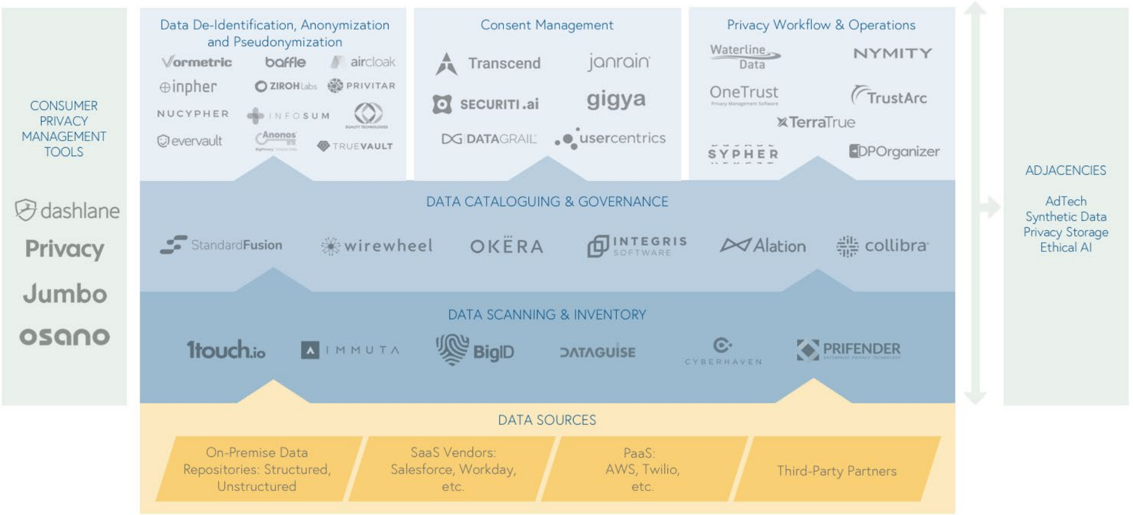
We doubt that consumer behavior is poised to change anytime soon either. Especially since there are undoubtedly benefits associated with consumer data collection, via ethical methods, to harness data into insights that advance innovation and improve the customer experience. New data-centric business models and services have emerged thanks to this proliferation of data, from the ability to deliver telemedicine services in time to save a life to ordering a burrito in time to watch The Office. It's not data that's the problem, it's unchecked data proliferation that causes issues.

Governments are stepping into the fray in an attempt to plug this leaky data oil well. Approved in 2016 and having gone into effect in 2018, the EU's General Data Protection Regulation (GDPR) was the first domino to fall that then spurred momentum amongst corporations to comply with newly established privacy rules. It is a law with sharp teeth, too; companies can potentially be fined up to 4% of their global revenue.

While the EU was an early champion in the data privacy movement, we are witnessing other national governments and individual US states follow suit. The California Consumer Privacy Act (CCPA) is the United States' first meaningful foray into modern day consumer data protection. While many other states have data privacy bills pending, the US Federal Government could (and, we hope, will) also pick up the mantle and pass a sweeping national regulation. Otherwise, companies are potentially left to navigate the nuances of 50 different state-sponsored bills. We believe this scenario would significantly impede businesses' ability to work across state lines.

Fortunately for consumers, a host of companies are emerging that allow companies to more ethically and responsibly leverage personal data. We see the data privacy market being divided into a few distinct categories, which would ideally represent an end-to-end set of solutions designed to identify, secure, and use data in a dynamic and ethical way. While this is our view now, we're admittedly in the early innings of what is sure to be a very long game.

Bessemer's Data Privacy Stack



This is the data privacy landscape as we currently envision it:

- Data Scanning & Inventory: Before a company can protect its sensitive data, it must first inventory all of its data to understand what it has and where it lives. And per most regulations, this must not only include personally identifiable information (PII) but also less structured personal information (PI). The difference between PI and PII is nuanced, but is an important distinction. PII (e.g. an email) is generally expressed in a well-defined way (e.g. SSNs) and is considered fairly straightforward to locate within structured data sets. PI on the other hand could include things like geo-location or product preferences; it is data that does not belong to an identifiable person but belongs to a given person and is equally protected. All of this data must be located and inventoried across all data stores in an enterprise both structured and unstructured. The need here is immediate, and we've witnessed companies like BigID experience tremendous success solving this acutely painful problem.

- Data Cataloging & Governance: Once a company has identified its sensitive data, it must understand the provenance of that data and where it is being stored, control who can access it and when, and apply certain rules to maintain this order. Companies like Okera are capitalizing on this opportunity, but we're seeing many others cropping up as well.

- Workflow Productivity: Companies must be able to respond to data subject access requests (DSARs) and other regulatory inquiries as well. As of a recent IAPP poll, only 2% of respondents claim to have currently automated these requests. While this might not be a costly problem now, the pace of these requests is sharply increasing and new software-defined solutions will be required for businesses to effectively manage these processes and other workflows. Companies like TrustArc and OneTrust have already raised at lofty valuations in this space, which is yet another sign this emerging category may have legs.

- **Consent Management:** Once a company has located its data, adequately determined its provenance and governance, and dealt with any associated statutory requests, it must then ensure it has and maintains adequate consent to actually use said data. As we mentioned earlier, data is a competitive advantage for most enterprises; consent management ensures that enterprises can use that data but do so in a consumer-friendly way. This is an emerging category with a rather nuanced problem to be solved, but we're seeing a host of companies in this space, though none seem to be established as a clear leader just yet.

- **Data De-identification / Pseudonymization / Anonymization:** Once an enterprise has inventoried it's data and has a handle on provenance, storage, DSARs and consent, it must then protect its most sensitive information. We're seeing companies like Privitar leverage elements of differential privacy, partial homomorphic encryption and other techniques to ensure private data stays that way, while others in the market are relying on multiparty compute and a slew of other techniques to achieve the same result. While the techniques may differ from company to company, it's clear to us that for any one company to be successful in this space it must leverage a host of different techniques to achieve not only the regulatory level of privacy (e.g. de-identification, pseudonymization or full on anonymization) required, but also to preserve the value of data while doing the same with consumer trust.

- **Consumer Privacy Tools:** We don't believe consumer behavior will change much in the near-term nor do we believe that consumers should even be personally obligated to protect their privacy — it is a fundamental right. But we do believe consumers will eventually want a more granular view of who has their data and how it is being used. Consumer password managers such as Dashlane have evolved their product to provide consumers with basic privacy features, alerting, and credit checks. However there is an opportunity for additional tools to help consumers take control of their privacy across a growing number of web services, social media sites, and apps. Companies like Jumbo Privacy are attempting to tackle this difficult problem by assisting consumers in grappling with the growing number of different privacy settings and policies.

While these are the major areas of the data privacy landscape that we find investable currently, we have also identified adjacencies in the adtech, synthetic data, privacy storage, and ethical AI markets which are currently even more nascent but that we believe may develop over time.

The future of Data Privacy Engineering

We believe that future data privacy platforms that will win the market will need elements of all of these categories in order to form an integrated suite that allows enterprises to identify, secure, and ultimately use their data in a manageable, legal, and ethical way. However, it's unclear if any one company will truly own the full stack or if we'll see specialization and best-in-breed solutions within categories, akin to what we've witnessed in cybersecurity. And while this is how we view the market today, we acknowledge that our views as well as the landscape — from a regulatory, technological and public opinion perspective — is constantly evolving.

We believe the net result of these collective efforts and solutions will help enterprises avoid The Seven Sins of Data Privacy Engineering, and while we're not nearly there today, companies must work towards a safer more consumer-friendly way to leverage data. Further, as data regulation continues to proliferate as quickly as the data itself, we firmly believe that Data Privacy Engineering will no longer be a nice-to-have but will morph into the new business imperative.

We're already investing against our Data Privacy Roadmap; Bessemer led BigID's $50 million Series C financing to help enterprises comply with global privacy regulations. If you're an expert or founder in the data privacy landscape and you'd like to get in touch, please email us at alex@bvp.com, jules@bvp.com, or mdonofrio@bvp.com.

Bessemer
Venture
Partners

San Francisco
Silicon Valley

Cambridge
New York

Bangalore
Herzliya