aircloak

# Aircloak Anonymized Analytics: Better Data, Better Intelligence

An Aircloak White Paper

**Data is the new gold. The amount of valuable data produced by users today – through smartphones, wearable devices, and connected "things" – is exploding. The companies that can best collect, mine, and monetize that data will be the winners. At the same time, public concerns over privacy are on the rise, and tough new regulations are just around the corner. Aircloak's anonymized analytics lets companies maximize the benefits of user data, while satisfying the public's and governments' need for privacy.**

## Data is the New Gold

A WIFI hot-spot provider receives location data about millions of users. The provider's clients – restaurants, stores, public buildings - would benefit from the insights this data provides: in the aggregate, where do their customers live and work? What other businesses do they frequent? How do these correlate with the customers' purchases?

Health clinics increasingly store patient care data electronically. Combining this data with that of other health clinics could provide tremendous benefits in the quality and cost of health care: For a given quality of health care, are some clinics more cost effective than others? Are the differences due to operational factors (nurse/patient ratio), or due to patient demographic factors (age or economic welfare)?

Smart watch makers hold detailed information about peoples' activities and certain health metrics. Smart scale makers track their customers' weight and body mass index. Supermarkets track people's food purchases. Combining this data would provide unprecedented insights into diet and health for individuals and researchers alike.

Privacy concerns stand in the way of these use cases. For example, the WIFI provider must certify that its data is anonymized before releasing information to its clients – a lengthy, costly, and error-prone process. Using traditional anonymization methods, it is typically not possible to anonymize complex data like health data without destroying the utility of the data. Combining data from different data silos also cannot be done with traditional anonymization methods. In these cases, data is most often simply not shared. When it is shared, it is done only in high-trust settings supported by costly legal and technical oversight.

These use cases are just a few of many examples, from transportation, automotive, health, Smart City, Internet of Things, finance, and government. Across the spectrum, vast amounts of useful data are going to waste because privacy concerns prevent easy and safe access to data.

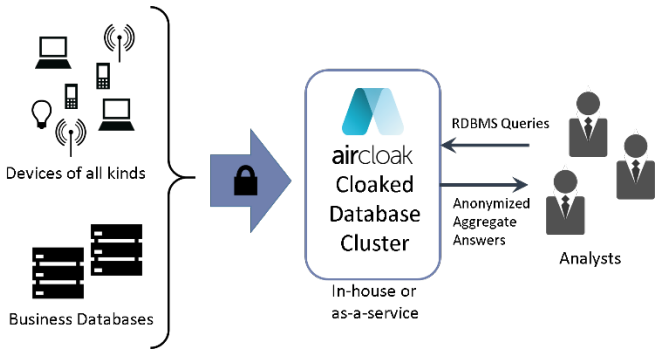
## Aircloak's Anonymized Analytics Solution

Aircloak's patented-pending solution enables anonymization for use cases where it was traditionally not possible. Aircloak is certified for all anonymization[i] use cases by TÜViT, the data protection division of TÜV, the largest and best-known certification and testing organization in Germany.

Aircloak immensely simplifies data anonymization. For the first time, certified data anonymization is as simple as uploading the data into an Aircloak SQL-based database cluster.

## Scalable, Flexible, Robust

Aircloak's core technology is a *scalable relational database* deployed on a cluster of servers called "cloaks". Sensitive data (data about users or their devices) is uploaded into a cloaked database, queries are submitted to the database, and anonymized answers are returned.




**Deployment**: Aircloak's database cluster can be used *as-a-service* deployed by Aircloak in a data center in the client's country of choice, or as *appliances*

deployed in-house. When in-house, the cluster may be administered remotely by Aircloak, or may be administered by the client. Either way, the data is anonymous: nobody, including system administrators, can view individual user data.

Data Upload: Data is uploaded into the cloaked database through a simple RESTful API (JSON over attested HTTPS). The cloaked database can handle structured and unstructured data types, including free text. Data may be uploaded from existing databases, or directly from user devices. The data is anonymous from the moment it leaves the existing database or user device.

Data may be uploaded in *batches*, or *streamed* and analyzed in real-time. For instance, a company that collects GPS location data for a car routing service may be legally obligated to delete that data after a short while. The company can insert that data into a cloaked database in daily batches, and keep the data indefinitely for later analysis, or offer analysis to third parties like city road planning. As another example, a company that monitors crowd movement in train stations through its own app may stream WIFI-triangulated location data directly into a cloaked database. In this case, the data is anonymous end to end – there is virtually no individual data exposure at any point.

Data may be uploaded from multiple sources and combined in the cloaked database.

Data Query: Data may be queried from the cloaked relational database through the RESTful API, or via a simple and intuitive web interface. Via the web interface, the analyst may select tables, columns in tables, and ranges over the values in the columns. The analyst may select from a library of pre-configured aggregation functions such as sum, average, median, number of rows, and so on. The analyst may also write his or her own aggregation functions in the scripting language Lua.

Queries may be *one-time batch*, *periodic batch*, or *streaming*. A one-time batch query runs once, for instance over a time range selected by the analyst. A periodic batch query runs the same query periodically over a shifting time window, for instance a daily query that runs over the last 24 hours of data. A streaming query runs every time data is uploaded into a cloaked database, producing continuous real-time output.

A cloaked database only returns aggregate results: it is prohibitively difficult, and in most cases impossible, to infer anything about individual users in the database.

For example, the crowd monitoring application could insert a streaming query that produces a heat-map of crowd density (number of people per area) in real time. Although the output heat map displays approximately accurate density in relatively more crowded areas, individuals cannot be tracked even in areas

where for instance there is only a single person. In another example, a medical researcher inserts a one-time batch query that outputs the correlation between various treatment protocols and the result patient reactions. Where there are a statistically significant number of patients for a given correlation, the cloak produces an approximately accurate result. Where the number of patients is too small or even one, the cloak will not output a meaningful value. This protects individual patient privacy.

Query Answers: Answers to queries may be received through the RESTful API, may be viewed on Aircloak's web GUI, or can be downloaded from the web GUI as a csv file for direct import into virtually any existing analytics package, including Excel, Tableau, MySQL, R, SPSS, and many others.

Scalable: The cloaked database cluster is built on a sharded PostgreSQL substrate. By adding machines, the cluster can grow to accommodate virtually any sized database.

## The Failure of Traditional Anonymization

The opinion on anonymization written by the Article 29 Data Protection Working Party of the EU discusses a number of traditional anonymization approaches, such as K-anonymity, L-diversity, and data permutation[ii]. One problem with these approaches is that they simply don't work for complex data, for instance health records and geo-location data. Even where they do work, these traditional approaches preclude joining data from multiple sources. As a result, the norm today is not *anonymization*, but rather *pseudonymization*. Pseudonymization involves removing *personally identifying information* like names, addresses, and account numbers, and replacing them with a random identifier.

As the working party 29 opinion makes clear, a common error organizations make in private analytics is mistakenly believing that pseudonymization is anonymization. In spite of several past high-profile data leaks[iii], organizations continue to make the same mistake[iv]. It is easy to see why pseudonymization doesn't work. Imagine a database of GPS location traces for millions of people, but where names have been replaced by random numbers. To uniquely identify a person in the database requires knowing only a few locations and times where that person has been[v].

To mitigate the risks in sharing pseudonymized data, reputable organizations today minimize their data sharing, and typically share only under strict contractual arrangements. Many opportunities for better analytics and data monetization are passed over.

## A New Approach to Anonymization

Aircloak takes a different approach to anonymization. Based on research done at the Max Planck Institute for Software Systems in Germany, Aircloak's approach combines three unique features:

- *"Zero Password, Zero Access"* data protection.
- Advanced *"post query"* answer anonymization.
- Trusted third-party oversight, enforced and remotely attested by trusted secure hardware (TPM).

The first feature, Zero Password, Zero Access data protection, ensures that nobody, not even system administrators or system developers, can view sensitive data residing in a cloaked database. The second feature, post query answer anonymization, ensures that queries run over the raw data do not expose individual user data. Nevertheless, Aircloak's unique answer anonymization limits answer distortion to a minimum, resulting in sufficiently accurate answers in most cases. By allowing queries to run over raw data, cloaked databases overcome the limitations of traditional anonymization. Assurance is provided by a trusted third-party, enforced and remotely attested by secure hardware TPM.

## Secure Hardware and Super-Hardening

*Hardening* a computer means disabling every unnecessary function so as to minimize both the attack surface and the possibility of human error. Cloaked databases are *super-hardened.* The *only* interfaces enabled are those for uploading data, submitting queries, receiving answers, and reporting system performance. This hardening is enforced by Security Enhanced Linux (SELinux). The cloaked database has *no password*, because there is no system login. There are no user or system accounts, no SSH or telnet; no way to log into the system at all. This approach to hardening protects against human error such as poorly protected passwords, and as well protects against insider attacks.

All data is encrypted while on disk. The crypto-disk key is protected by the *sealing* function of the secure hardware TPM. The secure hardware TPM doesn't use a password to encrypt and decrypt the crypto-disk key. Rather, when the TPM first encrypts the crypto-disk key, it takes a snapshot of certain cloak software and remembers it. When the TPM is asked to decrypt the crypto-disk key, it takes another snapshot of the same cloak software. If the software has changed, the TPM will refuse to decrypt the key, and no user data can be read. In essence, the software itself is the password.

While in memory (RAM), data is protected by the proprietary Data Scrambling feature of the Intel CPU.

## Trusted Third-Party Oversight, with Teeth

Aircloak uses the TPM sealing function to give trusted third parties enforceable oversight over software upgrades: the system won't boot unless the system software is cryptographically signed by the trusted third-party. The oversight function works as follows. The software included in the TPM snapshot includes the entire boot sequence as well as the software that checks the third-party signature. If the third-party signature does not match the system software, the checking software will refuse to let the system boot. If the third-party checking software is changed, the TPM will refuse to decrypt the crypto disk key.

Currently Aircloak uses TÜViT as its trusted third-party. Aircloak can work with any trusted third-party or combination of trusted third-parties as required.

Data sources uploading to a cloaked database can get *cryptographic proof* that the cloaked database has been signed off by the trusted third-party. This cryptographic proof is provided by the secure hardware TPM using a function called *remote attestation*. Overall, the TPM supports a cryptographically proven chain-of-trust extended all the way from a trusted third-party to the data source.

Finally, the cloaked database produces a *tamper-proof log* of all analyst interactions (queries and answers). Any attempts by the analyst, or for that matter by Aircloak, to infer individual user data through the query interface will be logged. In certain settings, such as medical research, ethics committees can use the logs to verify appropriate behavior by analysts and Aircloak.

This level of oversight is unmatched in the analytics industry.

## Dynamic Stateful Answer Anonymization

Aircloak's patent-pending anonymization dynamically filters and distorts query answers to prevent disclosure of individual user data while still providing sufficiently accurate answers for almost all analytics tasks. The distortion is less than 1% for answers involving roughly 400 or more users. The distortion, however, is never less than roughly plus or minus five. This defeats queries that target individual users.

Aircloak anonymization adds a small amount of zero-mean random noise to answers (Gaussian, standard deviation of 2). In this one regard, Aircloak's anonymization is similar to Differential Privacy, regarded by many researchers to be the gold standard in anonymization. Aircloak's anonymization, however, goes much further. Aircloak's anonymization adds a second layer of *fixed noise* whose value depends on the exact set of users that contributed to the answer. This prevents simple repeated answers from averaging away the first layer of random noise.

Aircloak's anonymization probabilistically suppresses answers with very low user counts. This prevents an analyst from inferring individual user data based on

the answer's label. For instance, a query that says "If target user has the secret attribute, output an answer labeled 'yes', otherwise output an answer labeled 'no'" will result in no output what-so-ever. The analyst learns nothing.

Aircloak's anonymization maintains a history of all previous answers reported by the cloaked database, at least until the data that produced the answer is deleted from the cloaked database. Every answer derived from some given user data is compared with every past answer derived from the same data. The comparisons take place at lightning speed, millions of comparisons per second, so query latency does not suffer. These comparisons detect *combinations* of answers that, taken together, could allow inference of individual user data. When such combinations are found, which rarely occur in normal usage, Aircloak's anonymization further modifies the answers to defeat the attack.

In spite of the above mechanisms, there may still exist unique combinations of tens or hundreds of answers that can leak small amounts of user data that are not prevented. These combinations, however, exhibit a certain analyst "fingerprint" that very rarely occurs with normal analyst behavior. Cloaked databases detect these fingerprints, and block analysts from further activity until the tamper-proof logs can be examined for possible misbehavior.

## Example Use Cases

Aircloak analytics enables data collection, storage, sharing, and monetization anywhere user privacy is important. Here we highlight just three of the many possible use cases.

Smart City / Transport: Smart city technology promises to reduce cities' operating and infrastructure costs, make them more attractive for citizens and businesses, and even provide new revenue sources. With Aircloak, cities can:

- Improve intelligence for both real-time services (safety and law enforcement) and city planning by combining data from different sources: public transportation and taxi services, car-sharing, socio-demographic data, city web services, indoor geo-location tracking, and smart streetlights.
- Safely provide data to third parties for monetization or Open Data initiatives.
- Instantly comply with national data privacy laws and reduce citizens' concerns over a "surveillance society".

Medical Research: Sharing medical data for research is a costly and risky process. Huge opportunities for improved medical research are lost because medical institutions leave data locked-up in protected silos. Aircloak enables:

- Safe and anonymous transmission of hospital and clinical research data to central research repositories.

- Improved quality of analytics by preventing the loss of data accuracy normally associated with traditional forms of anonymization.
- Compliance with medical research anonymization requirements, thus eliminating the need for new patient opt-ins.

**Monetization of Geo-location Data:** Cellular operators obtain outdoor geo-location data over a wide area from cell towers. Many businesses, for instance malls or airports, collect indoor geo-location data from WIFI access points. By combining indoor and outdoor geo-location data, Aircloak can:

- Improve insights for businesses using indoor geo-location applications by providing information about where customers live and work, as well as their demographics.
- Provide better monetization opportunities for cellular operators that offer aggregate marketing data.
- Avoid the need for indoor geo-location user opt-in, and provide compliance for cellular operators.

## Aircloak Professional Services

Aircloak's professional services team can support every facet of your analytics project to ensure a successful outcome. Aircloak can provide basic privacy consulting so that you can assess you need for anonymized analytics. Aircloak can work with you to understand your analytics goals and consult on how to achieve those goals using our solution: what data to collect (and not collect), what kinds of queries to run, how to interpret and display the results to obtain the insights you require. Aircloak professional services can develop any custom tools needed to move your data from existing storage to the cloaked database, develop a data plan and query plan to analyze the data, and develop any custom tools needed to import the results into your existing analytics ecosystem. Aircloak can work with your Data Privacy Officer (DPO) or any certification bodies you require to ensure that the proper approvals and oversights are in place. If you choose to deploy a cloaked database in-house, Aircloak can assist in equipment purchase, installation, and provide remote management. Finally, Aircloak professional services can provide complete training.

## Contact Us

Contact us today, and let Aircloak show you how anonymized analytics can help you collect and retain better data, and lead the way to better insights and data monetization.

[i] For its certification, TÜViT uses the definition of anonymization from the German Bundesdatenschutzgesetz (BDSG), §3(6) http://www.gesetze-im-internet.de/bdsg_1990/__3.html

[ii] ARTICLE 29 DATA PROTECTION WORKING PARTY "Opinion 05/2014 on Anonymisation Techniques", http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[iii] America Online, the state of Massachusetts, and Netflix: see section B.1 of Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", 57 UCLA L. Rev. 1701, http://www.uclalawreview.org/?p=1353.

[iv] http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546

[v] "Unique in the Crowd: The Privacy Bounds of Human Mobility," http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

We are sponsored by:

EUROPÄISCHE UNION

Gefördert durch:
Bundesministerium für Wirtschaft und Technologie
aufgrund eines Beschlusses des Deutschen Bundestages

ESF Europäischer Sozialfonds für Deutschland

eXIST Existenzgründungen aus der Wissenschaft

aircloak

Paul-Ehrlich-Str. 26
67663 Kaiserslautern
Germany

www.aircloak.com
**solutions@aircloak.com**