Kilimanjaro

TECHNICAL WHITEPAPER

DATA PROTECTION & SOVEREIGNTY

Global Analytics Without Global Leaks

Statistical Disclosure Control for per-user data residency: why a unified analytics surface never becomes a covert cross-border PII channel.

July 2026 · v1.0 · 16 min read

01 The tension, stated fairly
02 The reframe: authorization and residency are orthogonal axes
03 How an aggregate leaks an individual
04 The subtler failure: counting rows instead of people
05 Worked example: the count that almost named five people
06 Worked example: the average that was really one number
07 The graduated model: what a cleared operator actually gets
08 The drill-down: the cleared operator loses nothing
09 The one further lever: residency-flexible, on the customer's own declaration
10 Why this beats the alternatives

Executive summary

A cleared operator, authorized to see personal data and trusted to run cross-region reports, asks a fair question: why does the platform still put statistical controls in the way of a query they're allowed to run? It looks like bureaucratic friction on someone who has already been vetted.

It isn't, and the reason is a distinction worth making precisely: being authorized to see a piece of data is a different question from whether that data is allowed to leave its jurisdiction. Clearance answers the first. Nothing an operator holds answers the second — because a data-localization law doesn't have an exception for people with the right role.

And a global aggregate that skips disclosure control isn't exempt from that question either: a sum, an average, or a top-ten list computed over a small enough group is just that group's personal data wearing a statistics label, and it can re-identify someone to anyone who sees the number afterward — not just to the operator who ran the query.

This paper works through exactly how an aggregate leaks an individual, walks two worked examples end to end, and lays out the graduated model that gives a cleared operator everything their clearance entitles them to — full records in-region, accurate statistics globally — without ever letting a global number become a side channel for one person's data.

01 The tension, stated fairly

A regulated multinational wants what every data team wants: one dashboard that shows the whole business, not four regional ones stitched together by hand. Its operators are cleared — role-based access control, background-checked, audited. And the data those operators want to aggregate is, by definition, spread across jurisdictions that don't let personal records cross a border.

The natural objection from a cleared operator is not unreasonable: "I'm already authorized to see this data. Why is the platform slowing down a query I'm allowed to run?" Any serious residency story has to answer that question on its merits, not wave it away — because the direct answer is that clearance and residency are answering two different questions, and conflating them is exactly how a well-intentioned analytics feature turns into an unintentional PII channel.

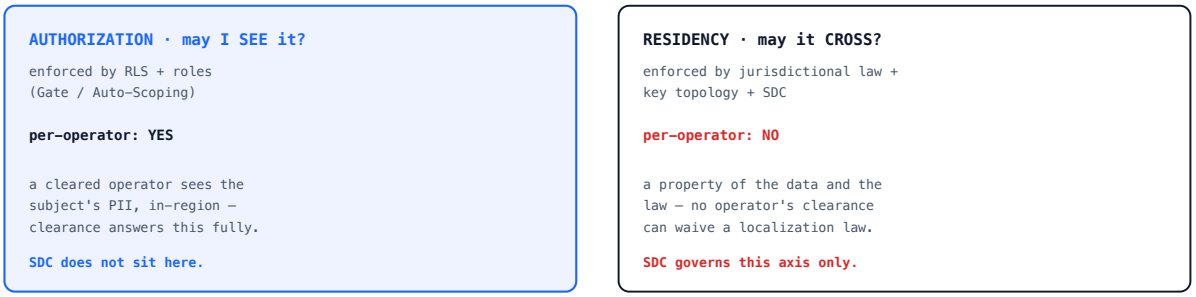
Question	In this section's own words
What the operator wants	One dashboard for the whole business, not four regional ones stitched together by hand
What the operator is	Cleared — role-based access control, background-checked, audited
The operator's objection	"I'm already authorized to see this data. Why is the platform slowing down a query I'm allowed to run?"
The direct answer	Clearance and residency answer two different questions; conflating them turns analytics into an unintentional PII channel

The tension, stated in the terms this section itself uses.

02 The reframe: authorization and residency are orthogonal axes

The objection conflates two controls that answer different questions. Authorization — may this operator see this subject's data — is answered by role-based access and row-level security, and it is genuinely a per-operator question: two operators on the same team can have different clearance. Residency — may this subject's raw data leave its region — is answered by jurisdictional law and the customer's own residency declaration, and it is not a per-operator question at all. It is a property of the data and the law that governs it.

A cleared operator's role cannot waive a data-localization statute. An operator with full clearance to view an EU subject's record is still not permitted to move that record to a US region — the clearance was never the relevant permission for that action. So a control that governs whether data may cross a boundary is not friction on legitimate access; it is friction on a different question entirely, one that clearance was never designed to answer.



Authorization asks whether an operator may see a subject's data — answered by clearance, per operator. Residency asks whether that data may cross a jurisdiction — answered by law and enforced by key topology, a property of the data, not the viewer. Clearance never answers the second question.

Statistical Disclosure Control governs only the residency axis. It never sits between a cleared operator and personal data they're authorized to see in-region — that access is unrestricted, exactly as their clearance says it should be. What SDC restricts is one specific path: deriving an individual's data out of what is supposed to be a non-personal, cross-region statistic. That path is a leak for anyone who later sees the result, cleared or not — which is the subject of the next section.

03 How an aggregate leaks an individual

“It’s just an aggregate” is not a safe assumption, and the ways it fails are concrete rather than theoretical. A SUM computed over a single contributing row is, arithmetically, that row’s exact value — there was never any aggregation to hide behind. A COUNT of one paired with a MIN, MAX, or AVG identifies the one record those functions were computed over.

Two aggregates that differ by a single record — a total, and the same total with one person excluded — subtract to that person’s exact contribution, a classic reconstruction technique that works regardless of how the underlying values are protected in storage. A top-N ranking over a cohort smaller than N simply names everyone in it.

None of this is a novel discovery — it is decades-old official-statistics methodology. National statistics offices have published census and survey data without exposing individuals for generations, using exactly this discipline: suppress small cells, guard against differencing, count subjects rather than records. Applying Statistical Disclosure Control to a cross-region analytics platform is applying a rigorous, well-understood field to a new setting — not inventing friction to slow operators down.

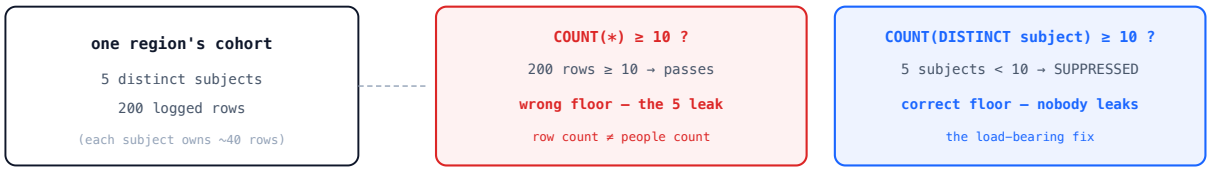


A sum over one row is that row’s exact value. A count of one, combined with a min, max, or average, names the individual. Differencing two aggregates recovers the delta between them. And a top-N ranking over fewer than N users names all of them.

04 The subtler failure: counting rows instead of people

The leaks above are about the shape of the arithmetic; this one is about what the floor counts in the first place. It is tempting to floor an aggregate on row count — "at least ten rows contributed" — but one data subject can generate many rows: orders, support tickets, logged actions, transactions. A cohort of five people behind two hundred logged actions clears a row-count floor of ten with room to spare, while still being exactly five identifiable people.

The floor has to count distinct data subjects — COUNT(DISTINCT subject), never COUNT(*) — or it protects nothing for exactly the collections where this matters most: event logs, transaction histories, support-ticket streams, anywhere one person generates many records. This single correction is what the next worked example turns into an end-to-end walkthrough.



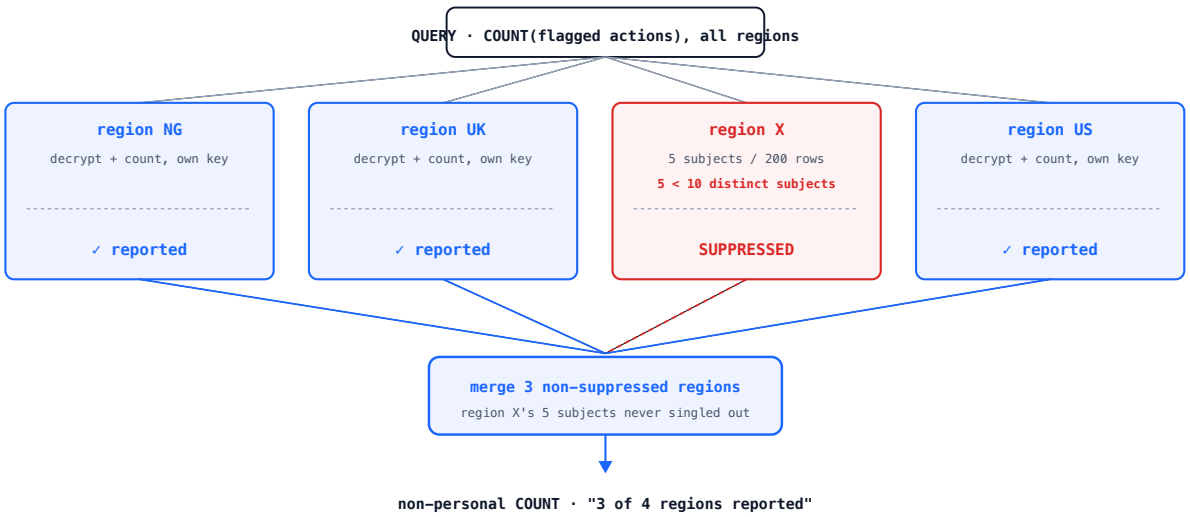
One data subject can own many rows — orders, tickets, logged events. A cohort of **five people can clear a row-count floor of ten with room to spare and still be exactly five identifiable people.**

A cohort of five data subjects can generate two hundred logged rows. A floor set on row count clears easily and still exposes exactly five identifiable people. A floor set on distinct data subjects correctly measures five and suppresses the cohort — the load-bearing fix.

05 Worked example: the count that almost named five people

Consider a compliance team running a global scalar aggregate — a COUNT of flagged actions this quarter, across every region. The query fans out to all four regions; each one decrypts its own data under its own key and computes its local count. Region X returns two hundred flagged actions, comfortably clearing a naive floor of ten contributing rows. But those two hundred actions belong to five individuals — checked against the distinct-subject floor described above, region X's contribution is five, which falls below the default floor of ten.

Region X returns a first-class SUPPRESSED result rather than a number — the region knows its own subject count and withholds before the coordinator ever sees anything from it. The other three regions, each with a healthy distinct-subject count, contribute normally. The coordinator merges the three that passed into a single global count and reports coverage explicitly — "three of four regions reported" — rather than presenting the result as complete. The five individuals behind region X are never singled out by a query that was never supposed to be able to name them in the first place.



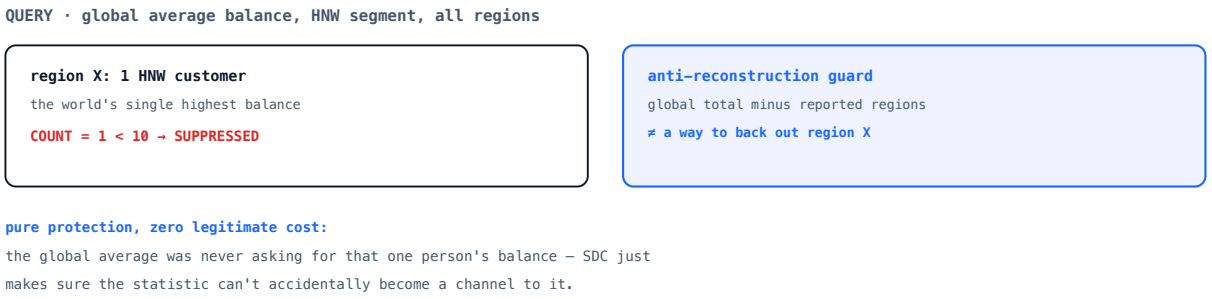
A global COUNT of flagged actions fans out to every region; each region checks its distinct-subject count against the floor before contributing. Region X has five tracked individuals behind two hundred logged actions and is suppressed — never a value. The coordinator merges the three regions that passed into a single non-personal count, reporting three of four regions explicitly.

This is one of the ungrouped scalar aggregates — count, sum, min, max, average — that the platform serves cross-region. A ranked or grouped cross-region query — a top-ten list, a group-by-region breakdown — works differently: it returns per-region pages rather than a merged global ranking, covered next.

06 Worked example: the average that was really one number

A second case makes the same point with zero ambiguity about whether the protection costs anything. A finance team asks for the global average account balance across the platform’s high-net-worth customer segment. Region X happens to hold exactly one customer in that segment — who also happens to hold the single largest balance on the entire platform. An “average” over one record is not an average; it is that customer’s balance, unrounded.

The distinct-subject count for region X is one, which is suppressed under any reasonable floor without argument. The remaining protection is arithmetic rather than statistical: the coordinator that merges regional contributions into a global figure is explicitly barred from computing a grand total and then subtracting every reported region to back out the one that was suppressed — the same reconstruction technique described above, applied at the coordination layer instead of the query layer.



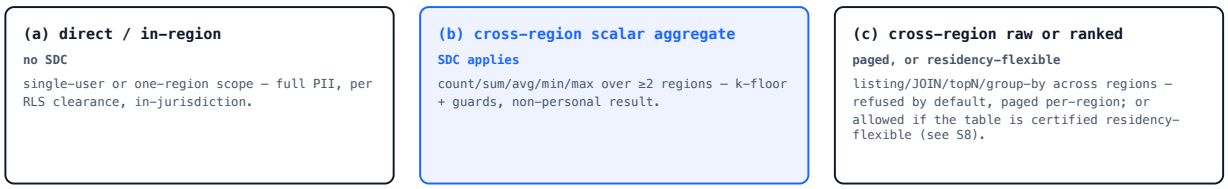
A global average balance for high-net-worth customers; region X holds the single highest balance in the world. Suppressing region X — and refusing to let the coordinator reconstruct it by subtracting the reported regions from a grand total — keeps the average from ever becoming a channel to that one figure.

Between the two guards, the global average genuinely never becomes a way to learn that one customer’s balance. This is the case where Statistical Disclosure Control is pure protection and costs a cleared operator nothing they were legitimately trying to get: nobody needed region X’s single data point to know the platform-wide average.

07 The graduated model: what a cleared operator actually gets

The point of this model is that a cleared operator does not lose access to individuals — they get individuals a different way than reconstructing them from a global number. Direct, single-region access carries no SDC at all: an operator querying one user's records, or every record in one region, gets full personal data exactly as their role-based clearance permits, in that jurisdiction. A cross-region scalar aggregate — count, sum, min, max, average — is where the distinct-subject floor and the reconstruction guards apply, because that is the one path where a number computed globally can quietly become an individual's data.

A cross-region raw listing or join — "every customer, every region, one table" — is refused by default and served instead as a sequence of per-region pages, because no disclosure-control gate makes a raw row non-personal; paging keeps each page inside its own jurisdiction. A ranked or grouped cross-region query — a top-ten list, a group-by-region breakdown — returns the same way: per-region pages, not a merged global ranking, because ranking and grouping are not a variant of the scalar aggregate above. The precise framing here is positive, not a limitation: a cleared operator ranks or groups by drilling into a region, in-jurisdiction, with the same clearance that already governs their access there — the next section walks through exactly that.



Three access patterns, by whether data crosses a region: direct in-region access carries full PII per clearance and no SDC; a cross-region scalar aggregate is SDC-gated; a cross-region raw listing, join, or ranked/grouped query is refused by default and paged per region.

08 The drill-down: the cleared operator loses nothing

The practical shape of the graduated model, for a dashboard, is a drill-down: the top-level view is cross-region aggregates, SDC-safe by construction. Clicking into a single region switches to that region's raw records, gated by the same clearance the operator already has, with no SDC on the way in — because a single-region view was never the side channel in the first place.

An operator who genuinely needs to see the five individuals behind a suppressed regional statistic gets them by navigating into that region directly, exactly as their role allows; they simply cannot reconstruct them by algebra on a number that was supposed to be a statistic. This is also where a ranked or grouped view lives: an operator ranks or groups the records in a region however they need, in-jurisdiction, under the same clearance — they rank where they're cleared, not across a boundary that was never theirs to cross.

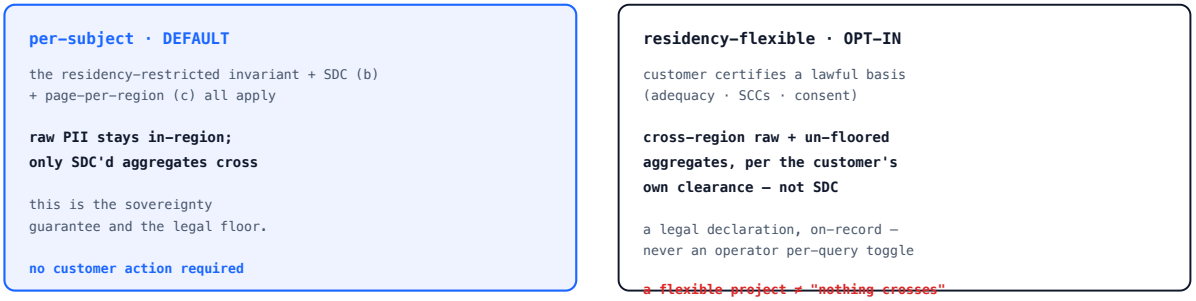


A dashboard's top-level view is cross-region aggregates, SDC-safe by construction. Clicking into a single region switches to that region's raw records, gated by the same clearance the operator already has, with no SDC on the way in.

09 The one further lever: residency-flexible, on the customer’s own declaration

One further lever exists for data that is not, in fact, legally bound to stay in place. Not every table a project marks as per-subject data is actually restricted by law from crossing a border — a customer may hold an adequacy decision, standard contractual clauses, or explicit consent that makes cross-border movement lawful for a specific dataset.

For those tables, and only on the customer’s own explicit, certified, audited declaration, cross-region raw queries and un-floored aggregates are permitted — governed by the customer’s own clearance model, not by SDC. That declaration is a statement of legal fact the customer makes about their data; it is not a per-query switch an individual operator can flip, and a project that has opted some tables into it is never accurately described as having “nothing crosses” — the precise claim names exactly which tables the certification covers.



The per-subject class is residency-restricted by default: raw PII stays in-region, only SDC’d aggregates cross. residency-flexible is an opt-in, on-record legal declaration covering specific tables — never an operator per-query toggle, and never grounds for calling a project “nothing crosses.”

10 Why this beats the alternatives

Each alternative gets one part of the problem right and misses another. Role-based access control is the correct tool for the question it answers — who may see this — and the wrong tool for a cross-jurisdiction data-movement question, because no role can grant an exception to a residency law.

Trusting the operator ignores that a value, once it has crossed a boundary, is visible to every system and person downstream of that operator, not just to the person who ran the query. Letting aggregates cross without control treats "it's a statistic" as a synonym for "it's anonymous," which the small-cohort and differencing cases above show is false. And running a separate project per jurisdiction solves residency by giving up the thing a platform is supposed to provide — one product.

Approach	What it gets right	Where it fails
RBAC / clearance only	Correctly governs who may see data in-region	Can't waive jurisdictional law, and can't stop reconstruction — a fully cleared operator's "global average" query still leaks a small region by differencing, because the leak is in the arithmetic, not the authorization
Trust the operator	Simple, and matches how in-region access already works	The crossed value is visible to everyone downstream of the operator — logs, a breach, a less-cleared colleague, a subpoena in the receiving jurisdiction. Trust in one person doesn't extend to everywhere the number travels afterward
Aggregates cross freely	No added latency or suppressed results	"An aggregate is anonymous" is false for small cohorts — a freely-crossing aggregate is a personal-data channel wearing a statistics label
Separate project per region	Residency is unambiguous — nothing crosses because nothing is unified	Loses the single project, API, and analytics surface — residency becomes the customer's engineering problem again, the exact cost the platform exists to remove

Four alternative approaches to cross-region analytics, and the specific failure each one has that Statistical Disclosure Control does not.

Statistical Disclosure Control is the only approach in the table that keeps the unified product, respects the operator's legitimate clearance in the one place it actually applies (in-region), and closes the one path — cross-region aggregation — where an authorized operator's trusted query could otherwise become an unintended way to move personal data across a border it was never supposed to cross.

What we disclose

Statistical Disclosure Control is a boundary control, not a general-purpose anonymization or access-control system. It applies to exactly one path — a cross-region aggregate — and does not apply to, restrict, or slow down direct, in-region access to personal data, which is governed entirely by the customer's own role-based access and row-level security. A cleared operator's ability to see records in their own region is unaffected by anything in this paper.

The k-anonymity floor that gates a cross-region aggregate counts distinct data subjects, not database rows, and defaults to ten, configurable upward only — it is never presented as a per-operator dial, and lowering it below the default is a separately gated, audited exception with a hard floor of five, not a routine clearance bypass. An aggregate over a handful of subjects is re-identifiable by anyone who later sees the number, regardless of who ran the query that produced it.

The residency-flexible opt-out is a customer's own declaration of a legal fact about a specific dataset — an adequacy decision, standard contractual clauses, or documented consent — certified and recorded on their account, never an operator-level toggle available per query. A project that has certified some tables as residency-flexible cannot be accurately described with an unqualified "nothing crosses"; the precise claim names which tables the certification covers and which remain fully restricted. Erasure, custody tiers and hardware attestation are scoped in the Zero-Residue and Sovereign Key Custody papers.

See the disclosure-control math for your data model

Talk to a Sovereign Engineer about which of your tables are residency-restricted by law versus lawfully movable, and how the k-anonymity floor applies to your cross-region reporting.

Talk to a Sovereign Engineer →

Read the residency whitepaper