Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

Request for Information (RFI) on

Advancing Privacy Enhancing Technologies

NowVertical Group Inc

NowVertical Group Inc. (NOW) is a big data technology company that helps businesses, government institutions, and large enterprises conduct vertically intelligent (VI) transformations through industry-specific software and services.

NOW possesses deep expertise in developing and bringing to commercial and government markets a flexible platform-agnostic software solution that transforms data to open standards and automates data processes.

Respondent Organization	**NowVertical Group Inc.**
Respondent Points of Contact	**David Whitmire** **President, NOW Solutions** **Farid Kassam** **President, NOW Origin**
Respondent Type	**Industry**

Our RFI Response addresses: *Specific technical aspects or limitations of PETs*

There are currently two primary strategies to conduct data collaboration between data owners, namely:

- Peer-to-Peer sharing
- Data Clean Rooms

Both these strategies have been around for years and have their advantages and disadvantages. However, with emerging privacy legislation and focus on personal data rights, organizations need to find new, scalable strategies to enable turn-key, privacy-safe data collaboration strategies that will satisfy the needs for sharing data without violating the privacy of individuals. Privacy-preserving computation (PPC) refers to strategies that enable data privacy to be protected while still being able to model and generate insights. **As part of this RFI, we will consider the existing strategies of data collaboration and PPC models and propose an architecture that we believe is scalable and efficient that will satisfy the needs for future privacy-safe data collaboration requirements.**

Peer-to-Peer Data Collaboration

Traditionally, peer-to-peer data sharing was the only way to perform analytics between multiple parties (and continues to be used today in second party data exchanges). It requires copies of raw data to be sent directly between collaborators with privacy restrictions managed by both parties independently and through legal agreements between parties. Typically, the goal is to “join” the shared data to an existing party’s data, and thus often requires the data to be row-level and in raw format (no encryption, etc.). Although encryption can be used to secure data in transit, encryption keys are often shared between parties to revert back to raw state or a common encryption method is used between both parties so the data can be joined (however, this still

enables either party to translate the encrypted data back to an original data point - and thus potentially an individual).

The benefits of this model are:

- Full control over what data is share to which partner

The primary down-falls to this model are:

- Engagements are delayed by lengthy legal/privacy reviews
- Raw data is used (or encrypted data can be reversed back to an individual)
- Lacks scalability (to add another party increases the privacy risks per project)
- Requires a heavy lift on behalf of the data teams to manage, join, and analyze the data
- Copies of the data can easily be made, shared, and exposed increasing liability risks

Data Clean Rooms (First Generation)

For the purposes of scale, commercial organizations began offering “data clean rooms” which were primarily created in order for partners to share and collaborate using their first party customer data. Since this application was the driving force behind data clean rooms, most of the early partners offering the solution had an underpinning identity graph that was used to 1) translate customer Personal Identifiable Information (PII) to a single common ID, 2) provide insights around the percentage of segment overlap between collaboration parties, 3) identify an ID set that in the overlap for media activation. This model required all participating and collaborating parties to share copies of their raw PII data with the data clean room providers, creating a hub-and-spoke model where the agency offering the service is the “hub” and all participating parties represent the spokes.

The early iterations of this model have challenges from a privacy perspective since raw PII and sensitive data is being shared into the centralized hub agency (and is thus exposed to the raw PII of all participating spoke parties). In addition, it creates a single point of failure since the entire model is controlled by the centralized hub organization, creating yet another walled garden of sorts.

The benefits of this model are:

- Scale enabling multi-party collaboration across different partners leveraging a common ID
- Privacy from “spoke” partners - since raw data was never shared to all other partners, the privacy risk does not increase with the number of data collaboration projects

The challenges of this model are:

- One agency owning access to all private data (single point of failure)
- Private data exposed to the centralized “hub” agency
- Control - insights and flexibility of data collaboration restricted by the centralized agency

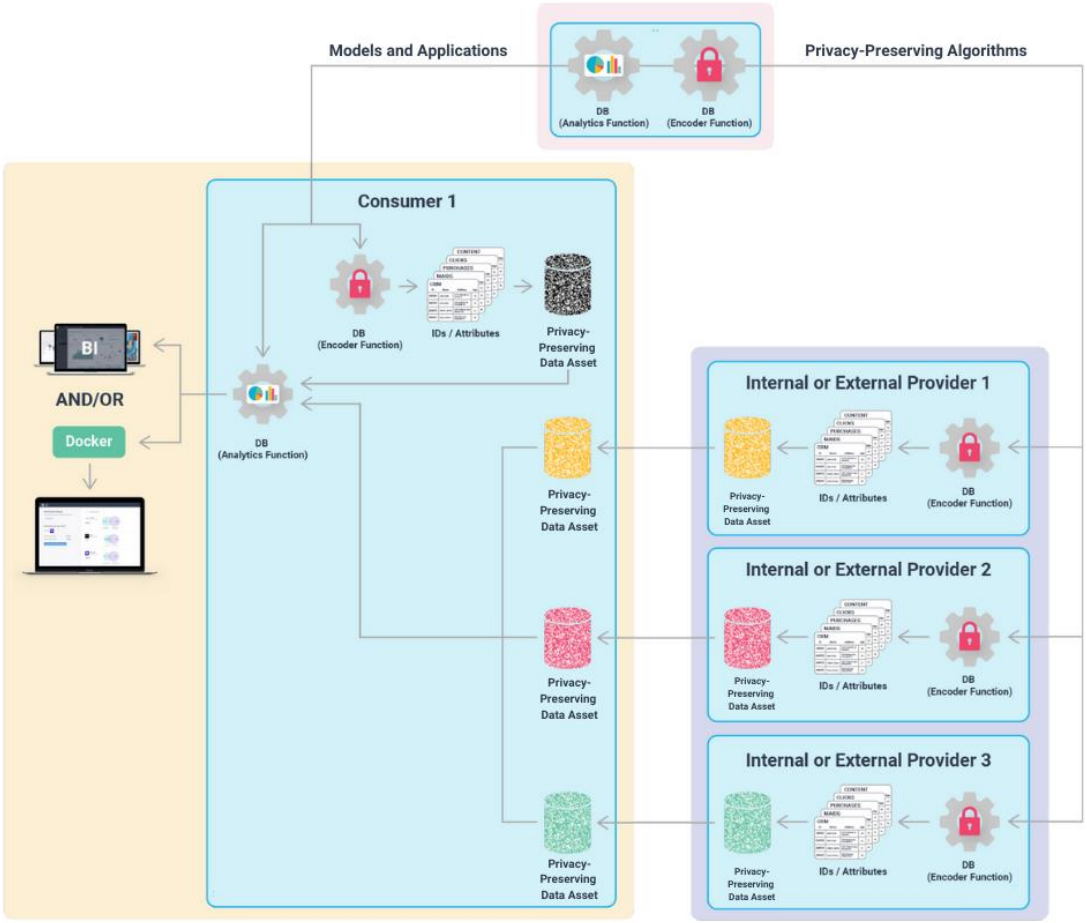
Data Clean Rooms (2nd Generation)

Over the last few years, a new generation of data clean rooms has emerged that leverages privacy-protecting strategies at the heart of the model. With this new model, private (PII) data of the “spoke” partners is shared to a “bunker” owned by the central agency where privacy-protecting algorithms are applied to the data to create a new data asset that can be used safely for analytics between “spoke” partners. This strategy reduces the privacy risk of the “at rest” data inside the bunkers since the original raw/private data is flushed after the privacy algorithm has been applied to the data set. This strategy has been proven to reduce legal and privacy due diligence, but still suffers from many of the challenges highlighted in the first-generation model.

Recommended Architecture for Privacy-Safe Multiparty Data Collaboration

The future of privacy-safe, multiparty data collaboration needs to leverage the benefits of the aforementioned models but change architecturally. This model needs:

- **A Distributed Model**: no one party "owns" the data, nor is there a centralized agency that manages all collaborations – each party owns their own data and has full control over what they share and with what party (similar to the peer-to-peer model)
- **Compatible Privacy-Preserving Data Assets**: no raw, private data should ever be shared. The model requires a unified privacy-preserving algorithm and strategy so that all participants use the same algorithm so that all privacy-safe data assets are compatible. This algorithm should be non-reversible and use privacy methods like noise injection to prevent privacy risks of "drilling down" to a single ID (and thus potentially exposing PII).
- **Secure Data Sharing**: copies of private data should not be shared; the architecture should enable secure data sharing without moving data. This simplifies data sharing and access revocation.
- **Modeling and Application Infrastructure:** beyond sharing the data, this model requires the ability for each participating party to access productized models and applications that can be applied to the privacy-safe data assets for insights and actionability.



A key to this architecture is where the privacy-preserving algorithms are being run. There are a host of well-known algorithms, each that often serve unique purposes, but in order for this future state to scale, there must be an agreement across all parties relating to which algorithm(s) is used.

We recommend that this algorithm(s) be set as an open standard to ensure compatibility across all collaborating parties.

This standardized algorithm(s) is applied to the data on the private servers of each participating party, creating a new privacy-safe data asset. All data sets are indexed/encoded/encrypted in the same way creating a foundation of compatible privacy-safe data assets. Regardless of which cloud or servers this new data asset has been created on, it next needs to be copied to a centralized platform or data warehouse for data sharing.

Once the data asset is on a centralized platform or data warehouse, it can be shared to other participating parties with no concerns about data privacy (since the new asset is non-reversible and anonymized). The data owner can choose who they share the data with and have full control to revoke that access if necessary.

With data sharing there is now a network of data assets accessible to all participants. The final feature of this future state is a common set of models and applications that can unlock aggregate insights and intel from the combined privacy-safe data assets.

We recommend that these models and applications can be user-contributed, but vetted (for security), authorized and managed as a library available to all participating parties.

These models and applications provide a means of creating pre-canned insights for specific needs (i.e., finding the percentage of overlap between two data sets). These models and applications also provide turn-key productized solutions so that data scientists are not needed to extract insights, and it can be done quickly and easily by citizen data scientists.

In summary, in order to create this distributed privacy-safe data collaboration model each participant must:

1. Create a privacy-safe data asset leveraging the agreed-upon algorithm.
2. Place a copy of this data asset on servers or a data warehouse that enables data sharing (without moving data).
3. Share this data asset with participating parties of its choosing (and likewise receive these data assets from others)
4. Leverage models and apps to compare, analyze and model multiple data assets in a secure environment.