PRIVITAR LABS

Federated Analytics: How to Conduct Safe Intercompany Analytics

Regulated industries face legal and ethical hurdles when sharing raw data. However, data analytics and data science teams can yield significantly improved accuracy and outcomes by doing so. Federated analytics allows the analysis of data split across multiple organizations without centralizing the data. In this document, we highlight the main challenges in performing federated analytics, briefly summarize our evaluation of the technologies to address these challenges, and provide guidance on evaluating commercial tools for federated analytics.

PRIVITAR Labs

Intercompany Data Sharing for Better Analytical Outcomes

Different organizations collect different information about the same people. While regulations and laws can make sharing data across geographical and organizational boundaries challenging and even impossible, the potential value from an analytical perspective is high. Analyzing this data together can be beneficial in detecting patterns, predicting trends, creating novel services, and providing enhanced value to customers.

Two example use cases that would benefit from cross-company data sharing include:

- A private, multinational healthcare service provider with patient history data and a pharmaceutical company with clinical trial data collaboratively study how prior patient conditions or previously used medications affect the performance of a drug.
- Different brands for a range of customer budgets under the same retail group collaborate to analyze how customer loyalty transfers from one brand to another helping to create more opportunities for cross-sell and up-sell.

Obstacles to Sharing Data

Consolidating cross-company datasets by pooling data from all the individual organizations for collaboration can be difficult, or even impossible due to:

- **Data protection regulations** – legal restrictions may prevent the sharing of personal data between different organizations or across international borders.
- **Commercial interests** – the data may contain sensitive information or intellectual property that the organization does not wish to share.
- **Reputational concerns** – organizations risk losing the trust of their customers when they share personal data with other organizations (see Privitar's 2020 Consumer Trust and Data Privacy Report). Also, it creates a single point of failure, and a breach of this centralized data will be more damaging to the organization, as it contains more detailed information about individuals.
- **Increased privacy risk** – individuals in the dataset are more vulnerable to reidentification because direct identifiers need to be retained in order to join the datasets. Also, the new dataset contains more information about each individual. When taken together, this information could allow attackers to single out individuals and crossreference them against other data they have access to.

Overcoming The Obstacles With Federated Analytics

Federated analytics allows the analysis of data split across multiple organizations without centralizing the data. The main challenges in performing federated analytics responsibly and legally are:

- Protecting privacy while performing analytics
- Ensuring privacy-safe analytics outputs
- Linking individuals

Challenge I: Protecting Privacy While Performing Analytics

Organizations participate in federated analytics in order to extract useful statistical insights from joint datasets: they are not interested in the record-level data concerning individuals. To give an example, an analysis of how a pre-existing medical condition affects the performance of a drug does not require knowledge of each patient and their medical history—we are looking for statistically significant conclusions about large groups of patients. In order to perform the joint analytics, however, some representation of the data "—be it plain text, aggregated, or encrypted"—must be shared.

It is tempting to assume that sharing mathematical representations of data instead of the actual data is sufficient for protection. This is wrong. Experience tells us that even mathematical representations of the data can reveal sensitive information. For example, federated protocols that share only updates to a machine learning model were initially assumed to be safe. Yet new attacks have now shown that these updates can reveal the underlying data. Similarly, perturbation methods used to obfuscate data while sharing have also been repeatedly broken.

Multiple privacy enhancing technologies (PETs) have been developed to provably protect data while performing analytics so that only the output of the analytics is revealed. Secure multiparty computation (MPC) and fully homomorphic encryption (FHE) are two such technologies. Unfortunately, at this time, the additional compute, storage, and bandwidth costs for MPC and FHE compared to approaches that offer no privacy guarantees are prohibitively high to work with large datasets or perform complex analytics.

With some popular libraries for FHE, 1Mb of data can result in more than 10Gb of encrypted data. And even state-of-the-art academic research attempts at building a generic solution for answering SQL queries with MPC takes somewhere between 10,000x to 1,000,000x more time compared to working with plain text. Moreover, most libraries for MPC and FHE require cryptographic expertise to work with them, limiting the number of people that can use these tools.

We do not intend to say that the picture with MPC and FHE is all grim. Although a general-purpose tool that will provably protect the data during computation comes with huge additional costs, the good news is that we have very efficient solutions for some of the specific problems that are encountered frequently in practice. Examples include calculations involving only sums and counts.

Solutions for Private Set Intersection are already deployed at large scale in applications, such as password monitoring and private contact discovery. To efficiently run inference on machine learning models, Microsoft® has recently released EzPC (Easy Secure Multi-Party Computation), which does not require any expertise in cryptography. Google has a solution deployed for computing ad conversions with partial homomorphic encryption. Technologies, such as secure enclaves and trusted execution environments, relying on specialized hardware, can also be used to protect data while performing analytics. Multiple advances and custom optimizations in these technologies are making their use in different applications practical. The trick is to pick the right set of technologies for your use case.

Challenge II: Ensuring Privacy-Safe Analytics Outputs

Protecting data only while performing analytics is not enough. It's important to be aware that the output of the analytics can reveal sensitive information about individuals. Carry out the appropriate checks before revealing the output. At a basic level, ensure that the output is never information at an individual level that permits enhanced profiling. Only allow aggregate statistics and patterns in the data about groups of individuals.

However, releasing too many aggregate statistics can also be a privacy risk as it might be possible to reconstruct sensitive information about the original data from these statistics. If possible, you should use controls around other aspects beyond the data to minimize the privacy risk.

For example, this could entail proper vetting of analysts before allowing access to the outputs and monitoring all the interactions of analysts with the data. Alternatively, if no controls can be placed around the environment and the people accessing the output (for example, when releasing data to the public), statistical disclosure controls, such as differential privacy can be applied to the output.

Challenge III: Linking Individuals

Linking individuals in data distributed across organizations is not straightforward, as sharing the consistent identifier as clear text can raise privacy concerns. The linkage should ideally happen without disclosing the identifier values, either via a semitrusted third party that performs the matching or through some secure cryptographic protocols. Solutions that use commutative encryption, such as Privitar SecureLink, can help achieve this by allowing linking datasets without revealing the actual identifiers.

In the absence of a shared consistent identifier to link individuals across the datasets, attributes such as names, addresses and email IDs can be used. Unfortunately, these are prone to noise; for example, due to typographical errors or an individual changing names/addresses at different points of time. Although record linkage methods (generally based on bloom filters) exist to account for such errors, they are not cryptographically secure. Hence, either a semitrusted third party is required for performing the matching or one needs to opt for methods that use bloom filters with MPC, which are slow and computationally expensive.

Privitar Labs

At Privitar Labs, we work on state-of-the-art privacy enhancing technologies for solving practical business problems. We are building technologies for performing federated analytics on sensitive data that is distributed across multiple organizations. Our solutions match the techniques discussed here—including secure multiparty computation, homomorphic encryption, and trusted execution environments—to your business needs. Put together with our expertise in differential privacy and products to securely link datasets, we are developing solutions that offer strong privacy protection for cross-organizational data sharing. Reach out to us if you are looking for solutions in this space and want to learn more about the landscape and Privitar's work in this area.

Recommended Questions to Ask When Evaluating a Commercial Solution for Federated Analytics

Descriptions for features at a high level can sound very similar for many solutions. When evaluating a commercial solution for federated analytics, ask detailed questions on how data is handled while performing the analytics and what information is disclosed as outputs.

Here are some starting questions that can help you to dig deep when evaluating a solution:

- How is the data shared with external parties while performing analytics?
- Keep asking this question for different kinds of data, and check whether you are satisfied with the level of protection being offered.
- How are direct identifiers like passport numbers shared with external parties?
- How are sensitive numerical attributes like salaries shared with external parties?
- How are sensitive categorical attributes like race and gender shared with external parties?
- Is the data of any individual customer revealed as the output of the analytics?
- Is the output aggregate statistics and trends in data or row-level information about specific individuals?
- Who is the output of the analytics revealed to, and what are the protections around it?
- How are records from different organizations about the same individual matched?
- Does this involve sharing direct identifiers in clear text form?
- Can the solution match individuals when there is no shared consistent identifier?
- When there is no shared consistent identifier, how are the matching attributes shared?

About Privitar

Privitar empowers organizations to use their data safely and ethically. Our modern data provisioning solution builds collaborative workflows and policy-based data protection into data operations. Only Privitar has the right combination of technology, domain expertise, and best practices to support data-driven innovation while navigating regulations and protecting customer trust.

PRIVITAR Labs

Reference: P1013-PL-Federated Analytics-DP.