

Differential Privacy for Statistical Data of Educational Institutions

Ivan Podsevalov(✉), Alexei Podsevalov, and Vladimir Korkhov

Saint-Petersburg State University, St. Petersburg, Russia
i.podsevalov@spbu.ru

Abstract. Electronic methods of managing the educational process are gaining popularity. Recently, a large number of user programs have appeared for such accounting. Based on this, the issue of personal data protection requires increased attention. The coronavirus pandemic has led to a significant increase in the amount of data distributed remotely, which requires information security for a wider range of workers on a continuous basis.

In this article, we will consider such a relatively new mechanism designed to help protect personal data as differential privacy. Differential privacy is a way of strictly mathematical definition of possible risks in public access to sensitive data. Based on estimating the probabilities of possible data losses, you can build the right policy to "noise" publicly available statistics. This approach will make it possible to find a compromise between the preservation of general patterns in the data and the security of the personal data of the participants in the educational process.

Keywords: Differential privacy · Security · Statistics · Education

1 Theoretical Foundations Review

1.1 Introduction

The term "differential privacy" (hereinafter referred to as DP) appeared and formulated into an industrial standard in the works of several authors: Dwork, Nissim, Blum, McSherry, and Smith [1–4]. At the same time, in the classic presentation, definitions were given for entities that can be operated on during the process of working with the DP: databases, data sets, metrics for determining the "proximity" of data sets and distributions. Currently, many practical implementations have their own rules for determining the distance between the database or data sets, the sensitivity of functions, DP mechanisms, however, all implementations undergo a complete "test" for compliance with academic definitions. This indicates the complete success of the developed theory, which has been tested by practice for almost 20 years.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13380, pp. 603–615, 2022.
https://doi.org/10.1007/978-3-031-10542-5_41

1.2 The Concept of Differential Privacy

Differential privacy is a collection of methods based on "noise" of output data through randomization or randomness. The general idea behind this mechanism is that the information should be distributed regardless of whether the user of the information is an adversary or not. It is worth noting immediately that this is not some specific process, but, rather, a property that the process can have. The differential privacy property ensures that the result of statistical analysis of two sets of output data from one algorithm will be almost indistinguishable regardless of whether there is data of an individual in the set or not. Visually (Fig. 1), this property can be represented as follows (D_1 and D_2 differ by one line):

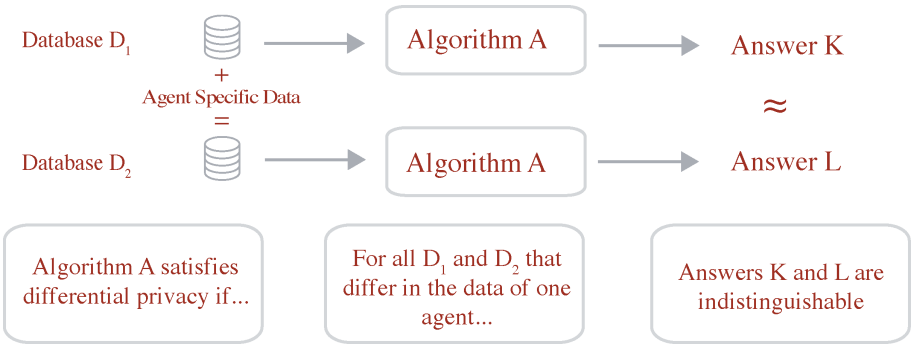

Fig. 1. Differential privacy

The meaning of the DP is the idea that any similar data sets should produce similar results based on the chosen distribution of a random variable. Reasoning more strictly and generally, the Lipschitz condition must be met for mapping "close" data sets in a random variable distribution.

The most important is the concept of the distance between two sets of data, which is necessary to determine a certain measure of "proximity" or "contiguity." We say that two sets are adjacent if the distance between them is equal to one. How close the two resulting distributions are to each other can be "measured" or determined by means of a positive parameter, called the privacy loss coefficient. In a certain, broader sense, we can refer to as a budget of the privacy loss, that is, a certain amount that can be allocated to a differential-private series of requests to the DB [14].

1.3 Local and Central Differential Privacy Models

The architecture of systems for differential privacy needs to provide both privacy and security. Privacy mechanisms control what can be excluded from data we

release. Security refers to controlling who is allowed to access the database, but makes no guarantees about what can be extracted from that database.

To date there are two main architectural solutions for differential privacy, central and local models. The central privacy model is the most popular system, where data is stored centrally and each result computed on the database is differentially private by adding noise to query results. Each user of the system can send any data without any additional noise. Main industrial differential privacy solutions include Microsoft's PINQ [5], Uber's FLEX (exploring its internal analytics) [6], LinkedIn's PriPeARL (for its ad analytics) [7], Google's differential privacy open source project [8,9], and the 2020 U.S. Census disclosure avoidance system[10]. Figure 2 depicts the central model of differential privacy.

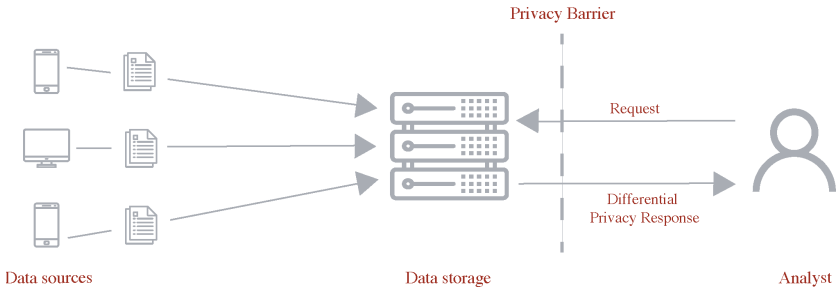

Fig. 2. The central model of differential privacy

An alternative approach is the local model, where the data curator never operates the sensitive data, so does not need to be trusted. With currently implemented systems based on local mode data is individually "noised" before being stored on a central "untrusted" server. The main implementations of local differentially private systems include Google's RAPPOR on Chrome browser [11], Apple's iOS and MacOS diagnostics [12], and Microsoft's telemetry data system in Windows 10 [13]. The Fig. 3 below demonstrates the local model.

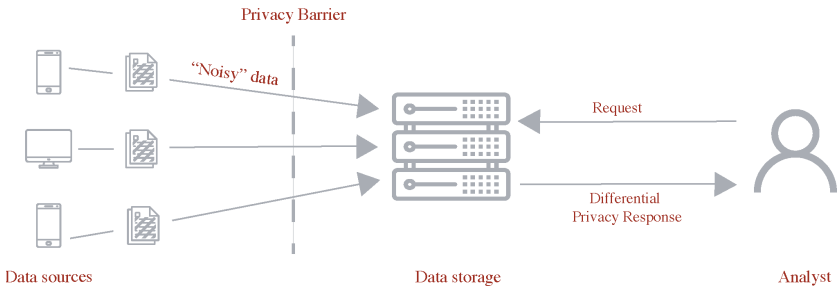

Fig. 3. Local model of differential privacy

The privacy model we are interested in for this article is the central privacy, where data is stored centrally, but we want to ensure each result requested on the data is privatized properly.

1.4 Classic Case of Differential Privacy

Let's consider an analysis on databases containing personal information about students. The analysis may be as simple as determining the average age of the students in the data, or it may be more sophisticated and involves performing some computation on data and outputting the result.

For example, consider counting queries to compute the number of objects that have some property, say how many students live in a dormitory in the first 99 records. Doing the same with the first one hundred records and determining the increase in result by one adversary can definitely find out that one hundredth record contains information about the dorm's resident. If the adversary has access to information on the hundredth record that the student came from the Kemerovo region, then in a certain case identification will happen almost instantly, because there may be a single student from the Kemerovo region in the database. That is, the organization of the process of issuing statistical data requires careful planning and determining and limiting the set of data that should be issued in response to requests.

Let's now make some notation and definitions that will be used to describe a prospective differential privacy system. To begin with, let's determine that the interaction between the analyst and the DP mechanism occurs through some kind of transformation or transcript. First of all, it is quite enough to consider the transcript as a single request to the database and the response to this request (we will consider a series of requests later).

Roughly speaking, a non-interactive mechanism M is differentially private if for any transcript and for any two sets of databases y and y' differing by one entry, the probabilities of obtaining results differ by $1+\epsilon$ (for sufficiently small ϵ) times [4]. It is technically more convenient to use exp^{ϵ} instead of $(1+\epsilon)$, because the former behaves more nicely under multiplication ($exp^{\epsilon_1} \cdot exp^{\epsilon_2} = exp^{\epsilon_1+\epsilon_2}$).

$$Pr(M(y) \in S) \leq exp^{\epsilon} Pr(M(y') \in S)$$

For greater clarity, it is necessary that the ratio of the probability distribution lies in the interval $[exp^{-\epsilon}, exp^{\epsilon}]$. The parameter ϵ is selected based on privacy policy considerations [21].

The mechanism is most often chosen based on the probability density distribution, such as Laplace or Gauss. We will consider the Laplace distribution, although this is not mandatory. Moreover, in general terms, differential privacy can be considered as a certain family of definitions parameterized by data sets and metrics used [5], which implies significant research opportunities.

So, the basic method for achieving differential privacy is to choose a Laplace mechanism that uses the exact calculation of the function f with the subsequent addition of noise. The noise is calculated according to the selected distribution

and scaled according to the sensitivity of the function f. In this case we use sensitivity equal to 1, $\Delta = 1$.

$$M(x) = f(y) + Y,\ where\ Y \sim Lap(1/\epsilon)$$

Note an important definition: the Laplace distribution $Lap(\lambda)$ has a probability density function $h(y) = \frac{1}{2\lambda} exp(-\frac{|y|}{\lambda})$, mean equal 0, and standard deviation $\sqrt{2}\lambda$.

Suppose we are dealing with a set D, whose elements can take the values $\{0, 1\}$, and we want to count the number of units in the subset X. It is clear that if we get a set X' from a set X by adding or removing one entry, then the number of units will change by a maximum of one, $||X - X'||_1 = 1$. In this case, we can talk about calculating the difference between two data sets using the L1 norm. It is worth noting that no one restricts us in using other metrics for a database of a larger dimension.

We can define that a deterministic function of a set of real data $f : D \rightarrow R$ has a sensitivity of at least Δ if for all adjacent D and D' the condition $|f(D) - f(D')| \leq \Delta$ is satisfied.

Then the distribution mechanism takes the following form $\widetilde{f}_\epsilon(D) = f(D) + Y,\ where\ Y \sim Lap(\Delta/\epsilon)$. The Fig. 4 below demonstrates the distribution mechanism.

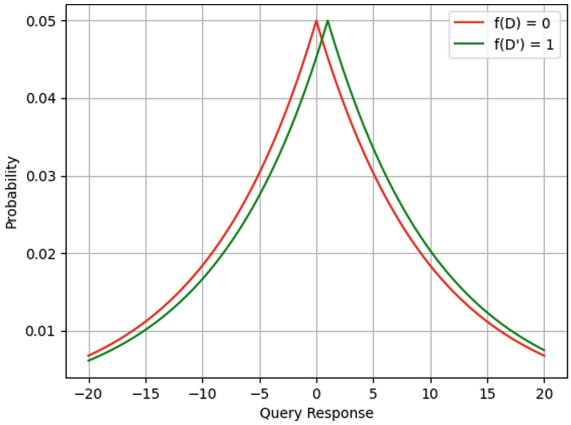

Fig. 4. Distribution of $\widetilde{f}_\epsilon(D) = f(D) + Lap(\Delta/\epsilon)$ over two inputs: $f(D) = 0$, $f(D') = 1$, sensitivity $\Delta = 1$, privacy parameter $\epsilon = 0.1$

It is clear that using continuous distribution as a mechanism can significantly reduce the performance of the program, so we should use a sample of discrete distribution. The most suitable method of generating Y is based on the generic procedure, called the inverse sampling method.

In our case, we know the probability distribution function for the Laplace distribution and this is $F(y) = Pr(Y \leq y) = 1 - exp^{-y/\lambda}$.

Thus, having random variables U in the range from 0 to 1 and an inverse function $F^{-1}(\cdot)$ at the input, one can obtain a discrete distribution: $Y \sim F^{-1}(U)$. This computing results in the procedure for sampling from the exponential distribution with parameter lambda: $\lambda Ln(U)$. Such similar transformations have already been implemented in the form of library functions for many popular programming languages.

Getting back to our example with students, we can get the following discrete distributions (Fig. 5).

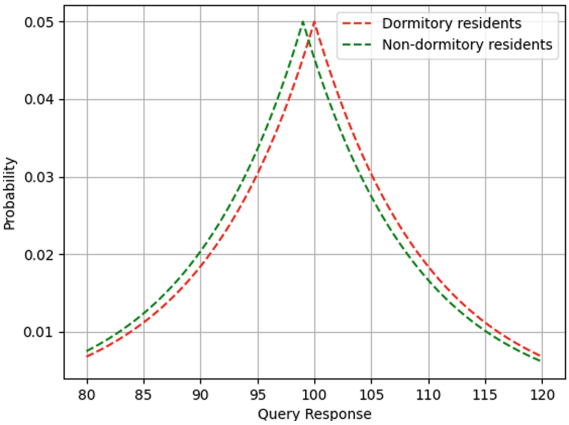

Fig. 5. Distribution of dormitory residents

1.5 Spending of the Differential Privacy Budget

In the previous section, we considered the option of non-interactive use of differential privacy, which is quite understandable and applicable for single use. However, in practice it is necessary to deal with multiple user requests to the system, moreover, with different requests. In this case, the mechanism of differential privacy should be defined as a sequence of randomized functions. In this sense, it may be interesting to apply useful consequences of the theoretical foundations of DP.

Differential privacy does not indicate the nature of the distribution used and has no relations with the auxiliary data that the adversary can get in the process from different sources. Also, the result of the DP mechanism itself is differentially private and it will not lose this property during subsequent post-processing [20]. However, when interacting with the system, it is necessary to set the total amount of allowed privacy leakage or the "privacy budget".

In fact, the privacy budget determines how many statistical queries can be made to the entire system and how accurate the results will be. In this case, it is necessary to maintain a balance: by setting the value too high, confidential

data will leak, by setting it too low, the accuracy of the data will quickly tend to zero. Based on this, in many cases, the privacy budget is obtained empirically.

The quantification of privacy loss budget permits the analysis and control of cumulative privacy loss over multiple computations. Understanding the behavior of differentially private mechanisms under composition enables the design and analysis of complex differentially private algorithms from simpler differentially private building blocks [20].

A simple composition of independent differential private algorithms is differentially private and the final value of ϵ is the sum of the coefficients of ϵ for each individual algorithm [21].

It is possible to define and prove the concept of differential-private "cumulative" composition as an extension of a simple composition when the maximum value of ϵ is set, at which the database manager (or curator) stops issuing data [14]. In practice, this is expressed in the fact that we can apply the differentially private mechanism to separate parts of the database and get a simple way to account for the total loss of privacy.

Suppose we are considering class registers for two different classes and the mechanisms M_1 and M_2 are used to work with them, with coefficients ϵ_1 and ϵ_2, respectively. Both of these logs (or tables) can be combined into one large log (or table), which is a common practice in working with a database. The concept of cumulative composition DP allows us to assert that the combined mechanism $M(M_1, M_2)$ is $(\epsilon_1 + \epsilon_2)$-differentially private. This process is depicted below (Fig. 6).

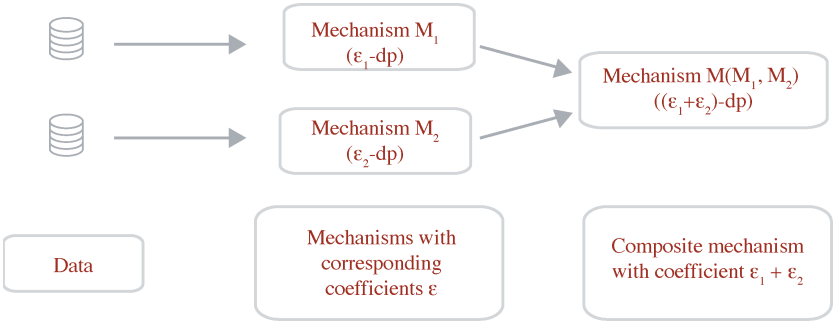

Fig. 6. Composite mechanism of differential privacy

However, such an approach is not without drawbacks related to the very nature of such a strict approach to differential privacy, when probabilities in the range $[-\epsilon, \epsilon]$ are considered, that is, tails are "lost".

To weaken such a condition, the concept of (ϵ, δ)-differential privacy is introduced, when the added coefficient δ makes sense to correct the situation in an emergency situation when the strict condition of differential privacy is violated.

That is, with a probability of $1-\delta$, the mechanism works as a "pure" differential privacy.

Note an important definition: an algorithm M is called (ϵ, δ)-differentially private if, for any adjacent sets D, D' and any measurable S, the following is true: $Pr(M(D) \in S) \leq exp^{\epsilon} Pr(M(D') \in S) + \delta$.

However, with probability δ, as they say in the community, "all bets burn out", which is very inconvenient. In addition, the possibility of composing sequential mechanisms is seriously complicated [14,16].

Ilya Mironov proposed a different mechanism - Renyi-differential privacy [15], based on the concept of Renyi divergence. Suppose we have two sets X and Y, each of which consists of n random variables. Accordingly, the probability of occurrence of each value is denoted as $p_i \in X$ and $q_i \in Y$, respectively. Then the Renyi divergence can be defined as

$$D_{\alpha}(X||Y) = \frac{1}{\alpha - 1} log(\sum_{i=1}^{n} \frac{p_i^{\alpha}}{q_i^{\alpha-1}}), \; where \; \alpha > 0 \; and \; \alpha \neq 1.$$

Thus, we determine the measure of similarity (or difference) of two distributions. Using this definition, we can say that $D_{\alpha}(A(D)||A(D')) \leq \epsilon$, i.e. the divergence of the two generated distributions should be less than the privacy loss parameter for large values of α (almost infinity). In this case, we can talk about (α, ϵ)-differential privacy when the coefficient α is finite. A remarkable feature of this mechanism is a simple way of composition, when the privacy coefficients ϵ_1 and ϵ_2 are simply summed up without any unpleasant consequences. All this makes it possible to manage the privacy budget quite simply in systems where you need to monitor a lot of requests.

2 Educational Institutions Case Study

Access to student data in Russia is regulated by the Federal Legislation of the Russian Federation, which defines basic concepts, for example, the concept of personal data. "Personal data - any information relating directly or indirectly to a certain or identifiable individual (subject of personal data)" (Federal Law No. 152-FZ of 27.07.2006 "On Personal Data").

For example, a student's personal data includes: information contained in a birth certificate, passport or other identity document; information contained in the student's personal file; information contained in the personal file of a student deprived of parental care; information contained in the class journal; information contained in the Student's Health Card; information about the state of health; a document about the place of residence; photographs; other information necessary to determine the relationship of education and upbringing. The right of access to the student's personal data may have:

- Director of an educational institution;
- Teaching staff;
- Social staff;

– Doctors/health staff;
– Other.

The section discusses 2 options - service requests for employees of educational institutions, and requests for external users. Requests for official purposes are broader in nature, requests from third-party users may be limited to a small range of functionality.

2.1 Data Structure

Let's describe our data as relational database tables in a normalized form, somewhat simplify the data schema without loss of generality. At the same time, it should be taken into account that the process of bringing to a normalized form leads to "fragmentation" of information. On the one hand, such presentation of information is convenient and familiar for developers of information systems, on the other hand, it complicates our work. However, for the sake of simplicity of understanding the principles, we will leave everything as it is.

The Fig. 7 shows the key independent data models, namely: students, teachers, subjects, classes and grades.

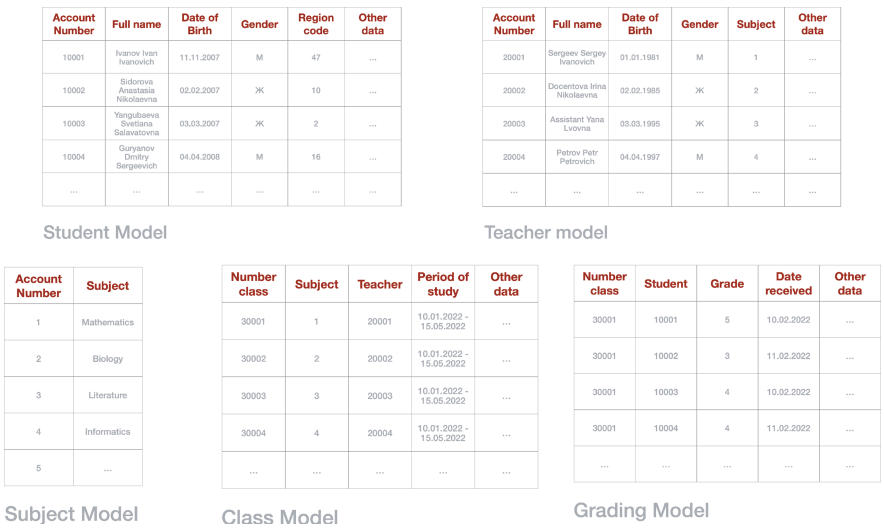

Fig. 7. Data models

Thus, such a structure allows to keep a very simplified accounting of the data of the educational process and organize the issuance of statistics according to user requests. We will consider the first and last tables, the rest models are explanatory.

2.2 Basic Requests

The main requests include all kinds of calculations of statistics on personal affairs of students, for example, by age, by region, by benefits. Such requests are classic from the point of view of the DP, since the personal data table contains one record per student and individual sets.

Requests for academic achievement can be considered as one of the main ones, let us focus on them. Users may be interested in statistical queries regarding academic performance in terms of subjects, class periods, etc.

If an adversary has access to accurate statistics, such as average, number, variance, then his work on identifying a particular student is significantly simplified. Having only limited access to personal data (for example: last name, first name, account number) and sufficient to the mentioned statistical data on the "classbook," it is quite realistic to restore its contents.

2.3 Histograms

Usually we talk about the database as a collection of records, each containing information about a particular person. Sometimes it is very useful to present the DB as histograms for different keys. For example, if a record of each student contains d attributes, the histogram will contain all possible combinations of these d attributes and the number of students having records with the corresponding combination.

However, a histogram with grouping by all attributes is not always informative. The most common may be histograms for a single attribute or a shorter combination of attributes, or fragments thereof. In fact, we are talking about a GROUP BY operation, often not clearly limited, and subsequent aggregation by the resulting unions. As we mentioned at the beginning about the only student from the Kemerovo region, even with the proper noise of the histogram, this will not be enough and the set of unions itself should be made differential-private [17].

If you consider the table with students' grades, you can note that each student in most cases appears in the classbook many times, these can be records of different grades for different times in different subjects. Traditional differential privacy is based on the recognition of two datasets as adjacent in the case of a difference of sets with only one record. In the case of our classbook, this has a very limited effect and potentially does not properly protect the privacy of the student. Instead, it is logical to consider adjacent datasets that either contain data about an individual or not. Naturally, such a definition of adjacency (or neighborhood) will lead to changes in calculations and performance. Suppose that we compute a histogram from a certain key combination k and previously get a list of length r. At the same time, the rules for calculating the privacy coefficient over the entire composition should also be preserved, and it is possible to evaluate ϵ_k and δ_k for each individual cluster set as the ϵ/r and δ/r, respectively, and obtain a (ϵ, δ)-differentially private mechanism as a result.

In the development of such systems, as a rule, several common problems arise:

– Restriction on the number of occurrences of an individual in a cluster;
– Predefined data grouping keys;
– Scale the solution to larger databases.

There are theoretical and practical solutions, such as Google's Plume system [18]. Another approach can be illustrated by the LinkedIn's Audience Engagement API [19], which does not impose any restrictions on entries into the cluster and significantly changes the approach to calculating the privacy loss budget, when deviations in histogram values are taken into account and the accumulation of the used budget is non-linear [6].

2.4 Architectural Solution

The strategy of spending the privacy budget can be implemented by adding an intermediate service with the calculation, accumulation and accounting of the current budget. Final implementations are no longer so obvious and require a lot of effort from developers.

The implementation of the DP can be described in the form of an independent module, which will be an addition to the software product that distributes statistics. Architecturally, the differentially private module can be divided into two services: privacy service and budget spending service. The privacy service acts as a coordinator or instructor who receives the request, processes and returns the response. The most important in this module is the budget spending service, which stores the current state of the budget, calculates the new value and regulates access to data based on the established privacy restrictions. In fact, the implementation of such a service is the main and most labor-intensive part of building a general architecture of differential privacy.

It is worth noting that budget spending implies the use of a certain starting budget value, which can be calculated from the bottom up, starting from the simplest components of the system.

Schematically, the module can be represented as follows (see Fig. 8). The privacy service does not have an internal state, it coordinates, makes the necessary calculations and gives the final result. Thus, the final pipeline will look like this:

1. Get the current value of the privacy budget and private data if the budget has not yet been spent. Otherwise, the data is not distributed, it is necessary to notify the agent;
2. Based on the data obtained, the result of the deterministic function from the set of real data is calculated $f : D \rightarrow R$;
3. Based on the obtained value of the privacy budget and the initial data, noise is calculated (use, for example, Laplace distribution), which is added to the result from paragraph (2);
4. The result of paragraph (3) is distributed with the agent.

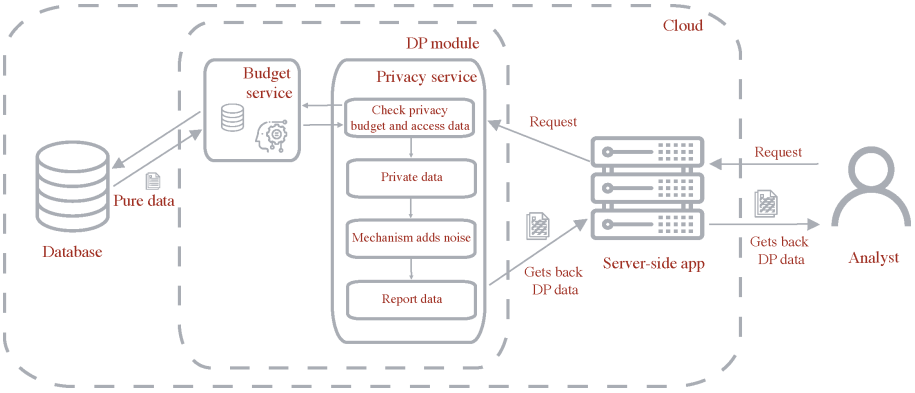

Fig. 8. Architecture for differential privacy module

Such an architectural solution is not a standard, but clearly demonstrates the direction of further work in the study of differential privacy by the community.

3 Conclusion

As part of this study, we considered a relatively new mechanism - differential privacy. We tried to demonstrate that by evaluating the probabilities of possible data losses, we can build the right policy to "noise" publicly available data. This approach makes it possible to find a compromise between maintaining the accuracy and privacy of the distributed data. We demonstrated the possibilities of practical application of this mechanism, and also described the possible architectural solution of the differentially private module. It should be noted that the community still has a lot of work to do in standardizing the principles of differential privacy; on our part, we hope that we have been able to demonstrate directions for further development.

References

1. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2003, pp. 202–210 (2003)
2. Dwork, C., Nissim, K.: Privacy-preserving datamining on vertically partitioned databases. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 528–544. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28628-8_32
3. Blum, A., Dwork, C., McSherry, F., Nissim, K.: Practical privacy: the SuLQ framework. In: Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 128–138. ACM (2005)
4. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14

5. McSherry, F.D.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, ser., SIGMOD 2009, pp. 19–30. Association for Computing Machinery, New York (2009). https://doi.org/10.1145/1559845.1559850
6. Johnson, N., Near, J.P., Song, D.: Towards practical differential privacy for SQL queries. Proc. VLDBEndow. **11**(5), 526–539 (2018). http://arxiv.org/abs/1706.09479
7. Kenthapadi, K., Tran, T.T.L.: PriPeARL: a framework for privacy-preserving analytics and reporting at LinkedIn. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ser., CIKM 2018, pp. 2183–2191. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3269206.3272031
8. Guevara, M.: Google developers, September 2019. https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html
9. Wilson, R.J., Zhang, C.Y., Lam, W., Desfontaines, D., Simmons-Marengo, D., Gipson, B.: Differentially private SQL with bounded user contribution. In: Proceedings on Privacy Enhancing Technologies, vol. 2020, no. 2, pp. 230–250 (2020). https://content.sciendo.com/view/journals/popets/2020/2/article-p230.xml
10. Dajani, A.N., et al.: The modernization of statistical disclosure limitation at the U.S. Census Bureau (2017). https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf
11. Erlingsson, U., Pihur, V., Korolova, A.: RAPPOR: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, ser., CCS 2014, pp. 1054–1067 (2014). https://doi.org/10.1145/2660267.2660348
12. Apple Differential Privacy Team: Learning with privacy at scale (2017). https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html
13. Ding, B., Kulkarni, J., Yekhanin, S.: Collecting telemetry data privately, December 2017. https://www.microsoft.com/en-us/research/publication/collecting-telemetry-data-privately/
14. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. J. Priv. Confid. **7**(3), 17–51 (2017). https://doi.org/10.29012/jpc.v7i3.405
15. Mironov, I.: Renyi Differential Privacy (2017). https://arxiv.org/pdf/1702.07476.pdf
16. Bun, M., Steinke, T.: Concentrated differential privacy: simplifications, extensions, and lower bounds (2016). https://arxiv.org/abs/1605.02065
17. Gopi, S., Gulhane, P., Kulkarni, J., Hanwen Shen, J., Shokouhi, M., Yekhanin, S.: Differentially private set union. arXiv preprint arXiv:2002.09745 (2020)
18. Amin, K., Gillenwater, J., Joseph, M., Kulesza, A., Vassilvitskii, S.: Plume: differential privacy at scale. arXiv:2201.11603 (2022)
19. Rogers, R., et al.: LinkedIn's audience engagements API: a privacy preserving data analytics system at scale. J. Priv. Confid. **11**(3) (2021). https://doi.org/10.29012/jpc.782
20. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(34), 211407 (2013). https://doi.org/10.1561/0400000042. ISSN 1551-305X
21. Vadhan, S.: The complexity of differential privacy. In: Tutorials on the Foundations of Cryptography. ISC, pp. 347–450. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57048-8_7