CYBERNETICA
Institute of Information Security

Privacy-preserving tax fraud detection in the cloud with realistic data volumes

Version 1.1

Dan Bogdanov, Marko Jõemets, Sander Siim, Meril Vaht

T-4-24 / 2016

This work has received funding from the Estonian Research Council through grant IUT27-1, ERDF through EXCS, and European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 609611 (PRACTICE).

Cybernetica research reports are available online at
`http://research.cyber.ee/`

Mailing address:
Cybernetica AS
Mäealuse 2/1
12618 Tallinn
Estonia

Privacy-preserving tax fraud detection in the cloud with realistic data volumes

Dan Bogdanov, Marko Jõemets, Sander Siim, Meril Vaht

Version 1.1

Abstract

Tax fraud detection is a suitable use-case for secure multi-party computation to be deployed in the cloud, allowing governments to detect tax fraud by analysing companies' business transactions, while enterprises maintain control of their private data. This report describes an efficient prototype application that analyzes tax declarations in a privacy-preserving way using the Sharemind® secure computation platform developed by Cybernetica. The prototype has been deployed and benchmarked in the Amazon EC2 cloud with realistic data volumes – the size of Estonia's economy. The benchmarks show unprecedented results in cost-efficiency of processing large amounts of data with secure computation techniques. In the cloud, we are now able to securely process 100 million business transactions in a matter of hours with less than $100.

Contents

1 Secure tax fraud detection prototype

A tax fraud detection system collects transaction data from companies, which is analysed by the government's tax authority to detect tax fraud. This scenario is an ideal use-case for secure multi-party computation (SMC). The companies act as input parties, who are concerned about the privacy of their business secrets. The tax board as the result party is only interested in identifying companies that are evading taxes. Using SMC, the business transactions can be analysed in a privacy-preserving manner, such that incorrectly declared transactions are found without requiring honest tax-paying enterprises to disclose their private data.

As part of the EU FP7 project PRACTICE[1], Cybernetica's researchers developed a prototype application that analyses tax declarations in a privacy-preserving manner [BJoSV15]. The prototype was built on the Sharemind® platform, which uses secret sharing to enable extracting meaningful information from private data while maintaining confidentiality.

The Sharemind secure computation system

A Sharemind® deployment consists of many computing servers, each hosted by a different entity. Private data is first secret-shared and then loaded into Sharemind® with each server-hosting party receiving a random share of the data. The parties can then jointly perform secure computations on the data, without actually seeing it. Afterwards, the result of the computation can be declassified only if all parties give their consent. This allows decision-makers to analyse data that cannot be accessed with traditional methods due to data protection regulations.

Currently, the most efficient protocols on Sharemind® require three non-colluding computing parties. If the parties are chosen with clearly non-collusive relations, the direct perception of security for data owners is greatly improved. For the tax fraud detection scenario, a possible deployment model in Estonia is depicted on Figure 1.

Risk analysis computations are performed jointly by the different organizations. Instead of sending the VAT declarations directly to the tax board, the VAT declarations are secret-shared between the computing parties. The performed computations are agreed upon beforehand when each party is satisfied that the algorithms do not disclose private information. Auditing and verification methods can be used to ensure that the servers do not deviate from these agreed-upon algorithms.

As output of the risk analysis, only the tax board receives the risk scores for companies with suspicion of fraud, and can then investigate further. The transactions of honest companies need not be revealed. The companies maintain a

[1] `https://practice-project.eu`

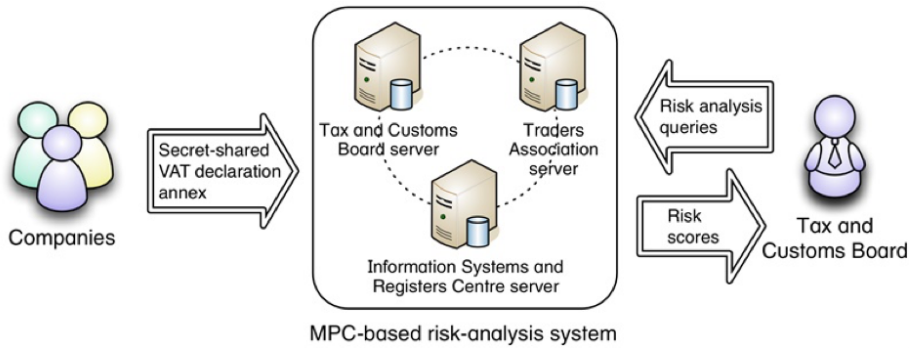

Figure 1: Deployment model of a tax fraud detection system using secure multi-party computation

degree of control over their data, since the Traders Association as a representative of the private sector is one of the server hosts. The third server host in this model acts as a neutral party.

Performance challenge

We estimated that the first version of the prototype system could perform risk analysis on a month of Estonian economy in 10 days using about 20 000 € worth of hardware in a local deployment. Following the interest from the PRACTICE project advisory board on assessing the viability of performing these computations in the cloud, we have optimized the prototype and prepared it for a large-scale cloud deployment.

We benchmarked the improved prototype on Amazon EC2 with data volumes according to the size of the Estonian economy, putting a realistic and surprisingly low price tag on running this system in the cloud.

2 Privacy-preserving fraud detection in the cloud

In the cloud setting, the computation servers would still be managed by the same three non-colluding organizations, however the actual physical servers would be hosted by one or many cloud service providers (see Figure 2).

The confidentiality of the data is similarly protected against the organizations managing the servers. However, additional trust assumptions about the cloud providers need to be made. Using different cloud service providers for hosting each party's servers provides the best security guarantees, but is not ideal for performance. Different possible deployment models are described in Table 1.

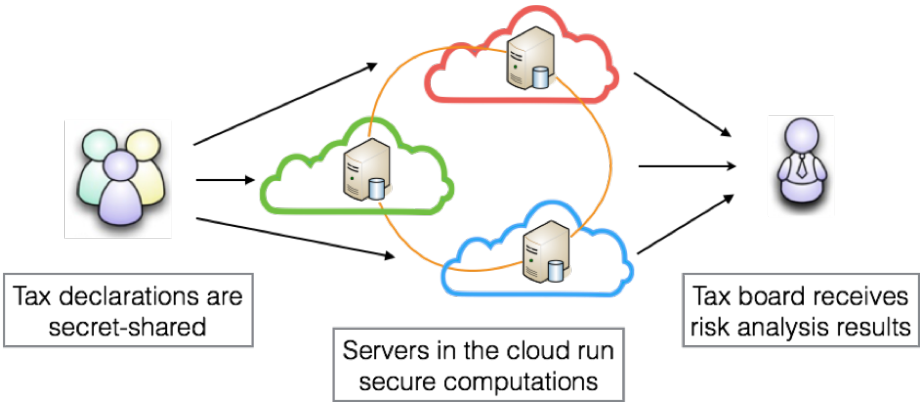

Figure 2: Tax fraud detection system deployed in the cloud. Different party's servers are hosted by one or many cloud providers.

Table 1: Descriptions of different possible cloud deployment models

Cloud deployment model	**Possible attacks**	**Security assumptions**	**Performance**
Single cloud provider – all Sharemind® servers are hosted by the same cloud provider	If the cloud provider has access to all computing servers, it can read all the private data	The cloud provider must be trusted not to access the data on the servers	The servers can be connected in a LAN, offering the highest performance
Two cloud providers – two out of three parties host their servers using one cloud provider, the third party uses a different cloud	The cloud provider hosting two servers can deduce the private data over time by reading the contents of encrypted network communication of both servers	The cloud provider hosting two party's servers must be trusted not to access the private keys of the servers' communication channels	Performance degrades due to latency as the physical distance of the two clouds increases
Three cloud providers – all parties use different cloud providers	If two cloud providers collude and monitor communication they can deduce the private data over time	The cloud providers must be non-colluding	Performance is dominated by the slowest connection between pairs of cloud providers

The different models offer a trade-off between security assumptions and perform-

ance. In the future, secure hardware solutions such as Intel® SGX[2] could help reduce the trust assumptions that need to be made about the cloud provider. In our benchmarks on Amazon EC2, we simulated all three of these models by running the computation servers in different Amazon EC2 regions to introduce latency.

3 Prototype evaluation in Amazon EC2

We now describe how the prototype was deployed into Amazon EC2 environment and give an overview of the performed computations. In our prototype, the whole computation process is divided into three distinct phases:

1. **Upload phase** – the secret-shared tax declarations are uploaded into Sharemind® and initial data validation is performed.
2. **Aggregation phase** – the data from each declaration is aggregated to enable risk analysis to be performed very efficiently. This is the most computation-intensive phase, however, data from each declaration can be processed independently, which allows for a high degree of parallelization.
3. **Risk analysis phase** – the results of the parallel aggregation are merged into a single large analysis table on which the risk analysis algorithms are performed. The output of this phase is the list of companies' registry codes with suspicion of fraud.

Since each company's data can be aggregated independently, we use a MapReduce approach by dividing the data between independently running Sharemind® processes during the upload phase. Then each process aggregates the data it receives in parallel and the results are merged into a single secret-shared database table in the risk analysis phase. This approach also allows scaling to larger volumes of data efficiently by adding hardware resources. This is one of the reasons why an elastic cloud-computing environment would be well-suited for deploying this kind of system.

To allow for a high number of parallel processes, we deployed each of the three party's servers as a group of four EC2 instances, totalling in 12 computing instances. Each set of 3 instances were running 20 Sharemind® processes in the parallel phases, whereas the risk analysis phase uses only a single process. An additional instance acted as the client that uploaded data into the computing nodes. Figure 3 illustrates this instance deployment using two EC2 regions.

For Sharemind® and most other methods of secure computation, a fast network connection is critical for performance. Thus, we chose to use c3.8xlarge instances in all our benchmarks, since it was the cheapest instance type having a 10 Gbps network connection and also supports Amazon's Enhanced Networking

[2]Intel® Software Guard Extensions – `https://software.intel.com/en-us/sgx`.

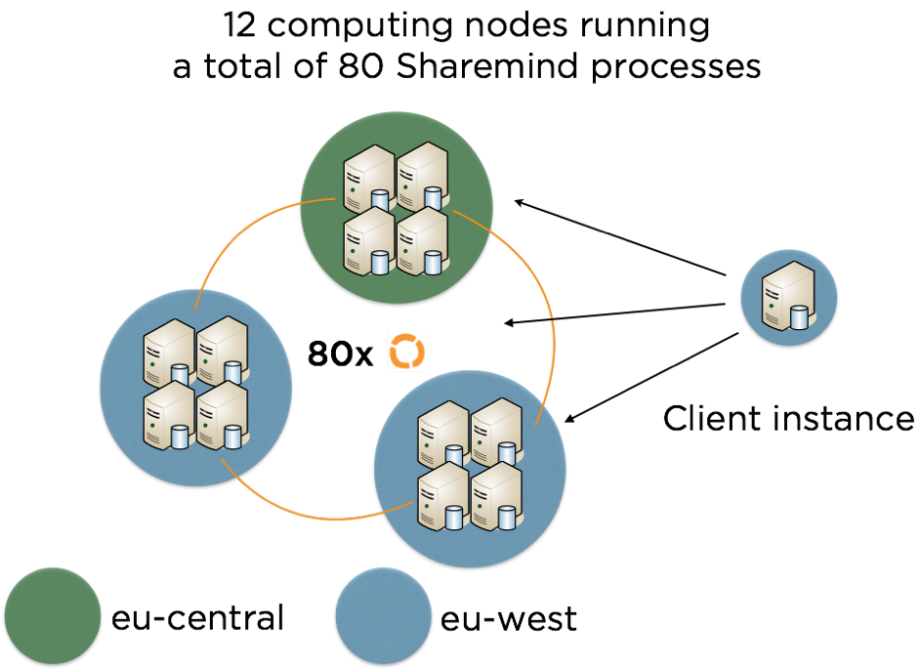

Figure 3: Amazon EC2 deployment within 2 Europe-based regions using a total of 80 Sharemind® processes to aggregate data in parallel

technology, improving overall network performance. The number of instances and parallel processes to use was then estimated by profiling the application with the largest used data set in a local deployment.

To be able to compare results, we used the same instance type setup in all the benchmarks. In Table 2 we summarize the deployment setup and network characteristics for different regional settings that we benchmarked. These correspond to the cloud deployment models described in Table 1.

Benchmark results

We used three input data sets with different size in our benchmarks (see Table 3). The largest data set corresponds to the estimates of Estonia's Tax and Customs Board on the number of taxable persons and performed business transactions in one month in Estonia. Each company's tax declaration is an XML-file consisting of a number of sales and purchase transactions with different business partners.

In the upload phase, declarations were uploaded to the 80 Sharemind® processes, each process receiving a single declaration at a time. After aggregating the data, the results were moved together into a single process running on three instances, and the remaining instances were closed. Note that each party

Table 2: The three regional instance deployments used, modelling one or many cloud providers

Regions	Client instance	Computing instances	Latency (round-trip)
1	us-east – c3.8xlarge	us-east – 12x c3.8xlarge	< 0.1ms between all nodes
2	eu-west – c3.8xlarge	eu-west – 8x c3.8xlarge eu-central – 4x c3.8xlarge	< 0.1ms between eu-west nodes 19ms – eu-west, eu-central
3	us-east – c3.8xlarge	us-east – 4x c3.8xlarge us-west – 4x c3.8xlarge eu-west – 4x c3.8xlarge	77ms – us-east, us-west 133ms – us-west, eu-west 76ms – us-east, eu-west

Table 3: Descriptions of the three data sets used in the experiments

No. of companies	No. of transaction partner pairs	Total no. of transactions	Total raw XML data size
20 000	200 000	25 000 000	8.61GB
40 000	400 000	50 000 000	17.26GB
80 000	800 000	100 000 000	34.51GB

only moves data shares between instances that it controls. The single process then merged the data and performed the risk analysis computations. We used Amazon's monitoring services to monitor the CPU, network and memory usage of the instances.

The running times of all computations are presented on Figure 4. The performance of the prototype has significantly improved compared to the earlier version and is well within practical limits as the analysis only needs to be performed once in a single tax period (each month). As can be expected, in multi-region deployments the computations are slower due to the increased latency. The aggregation phase is affected most, as the bulk of the computations are done there. Upload times are also affected since some secret-shared data validation is required. The risk analysis itself is very fast, since our risk analysis algorithm uses the assumption that the identity of a company can not be directly deduced from the number of its business partners. We also benchmarked a slower version that does not need this assumption for privacy (see Appendix A).

The total cost of a single run of the analysis is very low for a privacy-preserving computation of this scale (see Figure 5). The depicted costs include the price for running the instances and also data transfer between different EC2 regions (communication within a single region is free). Data transfer costs become increasingly important in multi-region deployments, forming up to 12% of the total cost. The depicted costs do not reflect expenses for data storage, which

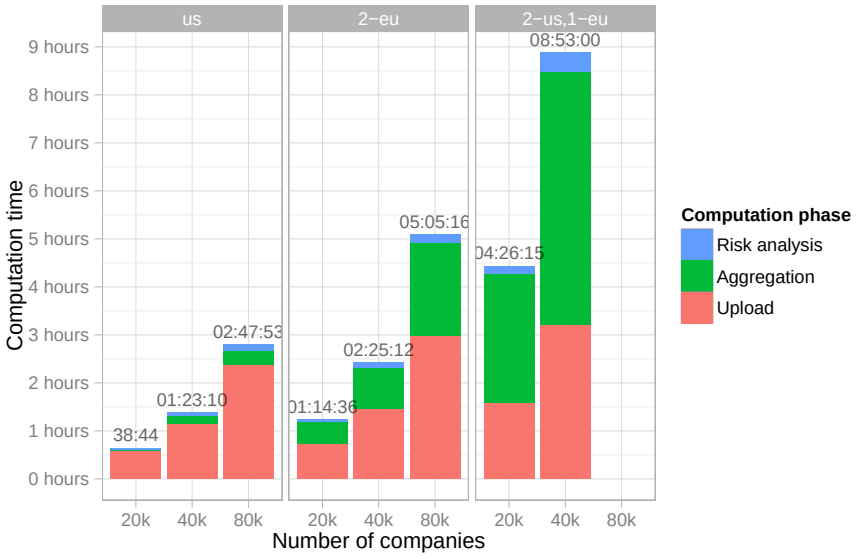

Figure 4: Running times of the computations in different deployments and varying amount of data

would be added for a persistently deployed system that stores all data from previous periods. Additionally, data transfer between different cloud service providers is more expensive than between EC2 regions.

In a real-life scenario, the data would be uploaded over a longer period of time and aggregation would also be continuous, processing new data as it is uploaded. An elastic cloud-computing environment would allow scaling the amount of hardware used dynamically, without requiring all the instances to run during the whole period. As such, the hardware costs would not differ much overall.

The c3.8xlarge instance type provides 32 CPU cores and 60GB of RAM. However, during aggregation, peak usage did not exceed 78% of total available CPU and 15% of RAM in any experiment. Average loads were 40% and 10% respectively for CPU and RAM. Maximum bandwidth used was measured only up to 4 Gbps for a single instance, which suggests more data could have been processed in a single process during aggregation to saturate the network connection without slowing down the computation, thus increasing cost-efficiency. With the largest dataset, a total of 1.2 terabytes of one-way communication was performed for the fast risk analysis implementation, which also stresses the importance of a fast network connection to achieve good performance.

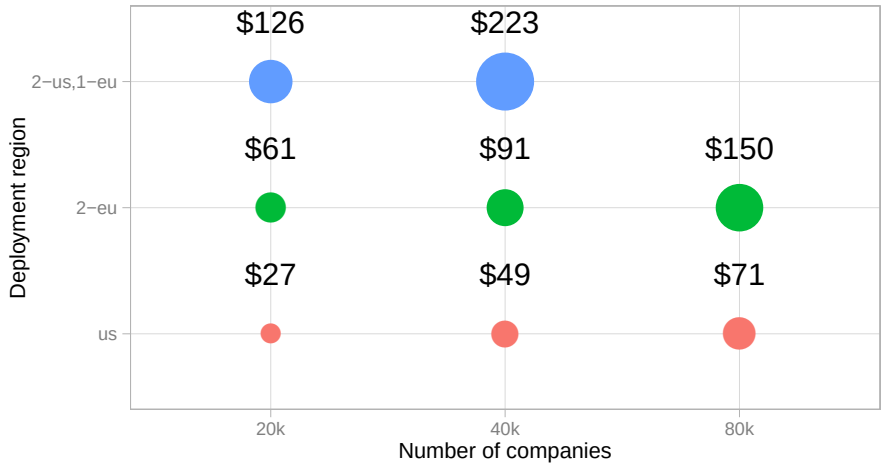

Figure 5: Total cost of the computation in different deployments and varying amount of data

4 Future work

The results of the benchmarks fully demonstrate that deploying and running a large-scale application performing secure computations, has become very cost-efficient, particularly in an elastic cloud computing environment. By performing these experiments, we gained much experience and insight into deploying large-scale SMC applications in the cloud. In the future, automatic service provision tools to deploy SMC in the cloud would make such deployments easier and bring SMC further toward the cloud.

Our prototype application could still further improved, especially the uploading process. Currently, we manually divided data between Sharemind® processes, but an automatic load balancer would be a more general solution for future applications and help make the current uploading phase faster. Also, tighter integration with cloud provider specific technologies and best practices could increase performance and practical security.

Building more sophisticated mathematical models for predicting the runtime and hardware usage of specific SMC applications in varying network conditions is another important future work to help estimate precise hardware requirements for future applications to optimise cost-efficiency.

A Risk analysis without admissible leakage

In addition to the fast risk analysis algorithm, we also implemented a slower version, which however is guaranteed to not reveal any side-information about companies who are found suspicious. The faster algorithm uses the assumption that the identity of a company cannot be directly deduced from the number of its business partners. Otherwise, the computing parties will learn if such a company is identified as fraudulent. Using this assumption, however, we can perform the analysis significantly faster, relying mostly on public operations with AES-encrypted values.

The running times and costs for using the slower algorithm are depicted on Figure 6 and 7.

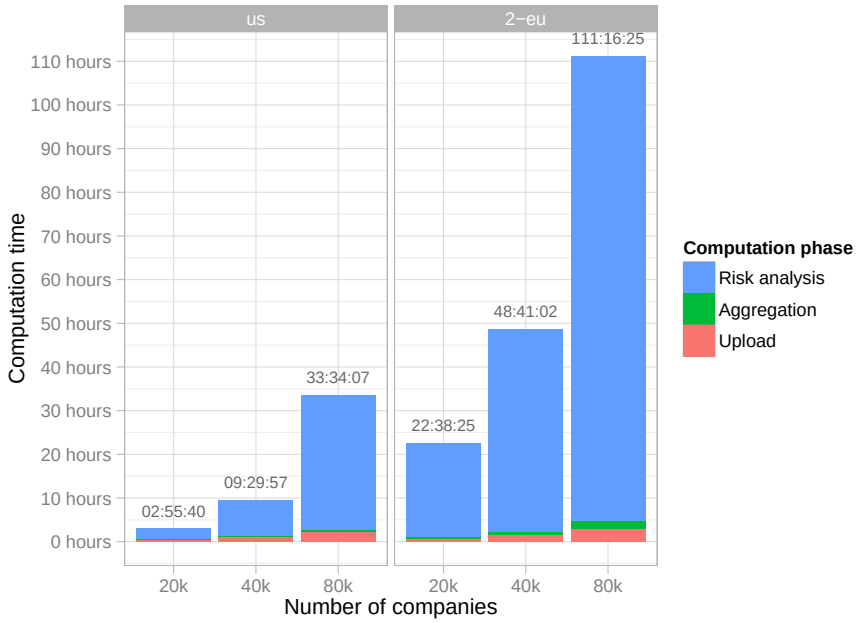

Figure 6: Running times of the computation using slower risk analysis algorithm that does not rely on admissible leakage

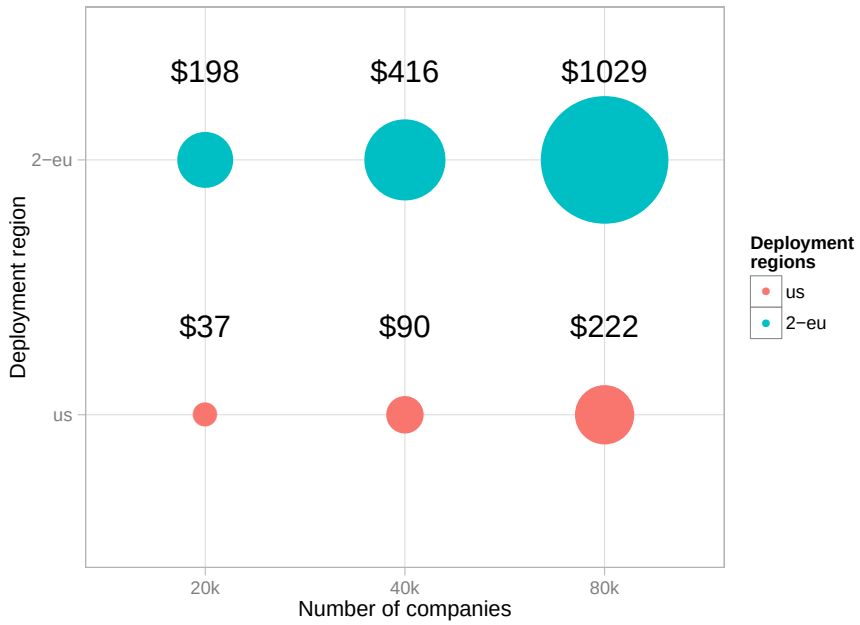

Figure 7: Total cost of the computation using slower risk analysis algorithm that does not rely on admissible leakage

B Detailed benchmark results

For reference, we bring out all measured performance statistics from the benchmarks. The network communication, CPU and RAM were measured using Amazon CloudWatch. CPU, RAM usage and bandwidth usage is calculated over means of one-minute periods. All reported metrics reflect the hardware/network usage for a single computing instance. Total communication refers to the total sum of one-way communication between all computing instances during the whole computation, measured by incoming network messages. Instance costs are calculated by charging to the full hour separately for the parallel phases (upload, aggregation) and the risk analysis step.

Data transfer costs reflect only inter-region communication. For example, in the eu-west, eu-central deployment, the servers in eu-west region communicate over private VPC IP addresses and this communication is free. All prices for different region instances and data transfer are taken as listed by Amazon on 2015/09/01.

Table 4: Running times, total exchanged communication and costs of benchmarks using the faster risk analysis algorithm

Deployment	Data size	Total comm. (GB)	Data transfer cost	Total time (h:min:s)	Instance cost	Total cost
us	20k	290.5	-	38:44	$26.88	**$26.88**
us	40k	587.8	-	01:23:10	$48.72	**$48.72**
us	80k	1202.2	-	02:47:53	$70.56	**$70.56**
2-eu	20k	307.1	$3.99	01:14:36	$56.82	**$60.81**
2-eu	40k	619.4	$8.05	02:25:12	$82.28	**$90.33**
2-eu	80k	1264.0	$16.43	05:05:16	$133.21	**$149.63**
2-us, 1-eu	20k	308.1	$6.13	04:26:15	$119.11	**$125.25**
2-us, 1-eu	40k	625.5	$12.46	08:53:00	$210.18	**$222.64**

Table 5: Running times, total exchanged communication and costs of benchmarks using the risk analysis algorithm without admissible leakage

Deployment	Data size	Total comm. (GB)	Data transfer cost	Total time (h:min:s)	Instance cost	Total cost
us	20k	1324.8	-	02:55:40	$36.96	**$36.96**
us	40k	4744.0	-	09:29:57	$89.04	**$89.04**
us	80k	17859.1	-	33:34:07	$221.76	**$221.76**
2-eu	20k	1383.2	$17.44	22:38:25	$180.46	**$197.90**
2-eu	40k	4958.3	$62.28	48:41:02	$353.13	**$415.41**
2-eu	80k	21643.3	$271.34	111:16:25	$757.34	**$1028.67**

References

[BJoSV15] Dan Bogdanov, Marko Jõemets, Sander Siim, and Meril Vaht. How the Estonian tax and customs board evaluated a tax fraud detection system based on secure multi-party computation. In *Financial Cryptography and Data Security - 19th International Conference, FC 2015, San Juan, Puerto Rico, January 26-30, 2015, Revised Selected Papers*, volume 8975 of *LNCS*, pages 227–234. Springer, 2015.

Table 6: Hardware and network usage metrics for the single region deployment (us)

Data size	Computation phase	Total time	CPU (mean)	CPU (max.)	Mean. bandwidth (Mbps)	Max. bandwidth (Mbps)	Mean RAM (MB)	Max. RAM (MB)
20k	upload	34:20	22.3%	26%	14.7	136.3	996	1029
20k	aggregation	02:26	45.9%	59.4%	802.0	1597.3	3174	5539
20k	fast risk analysis	01:58	4.9%	7.3%	462.6	686.2	897	1472
20k	risk analysis (total privacy)	02:17:51	5.1%	6.4%	342.1	538.1	979	1246
40k	upload	01:09:13	21.7%	23.8%	14.2	158.0	1005	1055
40k	aggregation	10:05	39.6%	54.8%	447.5	1685.5	3953	6699
40k	fast risk analysis	03:52	5.4%	8%	530.7	939.5	1054	1771
40k	risk analysis (total privacy)	08:08:39	5.5%	6.7%	382.8	642.5	1141	1670
80k	upload	02:23:18	21.8%	23.6%	15.8	155.1	1011	-
80k	aggregation	16:25	39.1%	77.7%	547.0	3494.8	4079	-
80k	fast risk analysis	08:10	5.3%	7.8%	494.7	1099.7	1050	-
80k	risk analysis (total privacy)	30:50:23	5.4%	7%	403.0	841.2	1178	-

Table 7: Hardware and network usage metrics for the two-region deployment (2-eu)

Data size	**Computation phase**	**Total time**	**CPU (mean)**	**CPU (max.)**	**Mean. bandwidth (Mbps)**	**Max. bandwidth (Mbps)**	**Mean RAM (MB)**	**Max. RAM (MB)**
20k	upload	43:32	30.4%	33.5%	10.4	67.2	1004	1031
20k	aggregation	28:10	42.5%	49.9%	89.9	957.9	4044	5303
20k	fast risk analysis	02:54	3.9%	6.5%	354.9	496.1	912	1482
20k	risk analysis (total privacy)	21:25:23	1.8%	4%	38.3	320.6	986	1250
40k	upload	01:27:37	27.3%	32%	15.5	432.9	1009	1499
40k	aggregation	51:17	42.2%	56.3%	93.5	2568.9	5068	6467
40k	fast risk analysis	06:18	4.6%	7.6%	376.2	583.7	1134	1778
40k	risk analysis (total privacy)	46:19:32	2.3%	4.8%	70.3	418.7	989	1496
80k	upload	02:59:19	30.4%	34.9%	15.4	455.0	1023	1880
80k	aggregation	01:55:09	42.6%	67.4%	85.4	3645.8	6285	8460
80k	fast risk analysis	10:48	4.7%	8.2%	437.5	804.9	1147	1544
80k	risk analysis (total privacy)	106:17:03	2.6%	5.5%	122.2	597.9	1251	1526

Table 8: Hardware and network usage metrics for the three-region deployment (2-us, 1-eu)

Data size	**Computation phase**	**Total time**	**CPU (mean)**	**CPU (max.)**	**Mean. bandwidth (Mbps)**	**Max. bandwidth (Mbps)**	**Mean RAM (MB)**	**Max. RAM (MB)**
20k	upload	01:34:42	32.3%	34.3%	7.1	204.7	1176	1816
20k	aggregation	02:41:37	38.2%	42.1%	15.0	1093.0	4186	5670
20k	fast risk analysis	09:56	2.5%	6%	119.0	184.5	1079	1602
40k	upload	03:12:19	35.9%	40.4%	6.7	219.6	1032	1533
40k	aggregation	05:16:48	41%	46%	15.8	1237.6	4994	6336
40k	fast risk analysis	23:53	2.3%	8.1%	102.9	187.2	1135	1839