CYBERNETICA
Infoturbeinstituut

Turvalist ühisarvutust kasutava käibemaksudeklaratsioonide riskianalüüsi süsteemi prototüüp

Dan Bogdanov, Marko Jõemets, Sander Siim, Meril Vaht

T-4-22 / 2014

Seda teadustööd toetasid:

1. Euroopa Regionaalarengu Fond läbi Eesti Arvuti-teaduse Tippkeskuse EXCS ja

2. Euroopa Komisjoni 7. raamprogrammi projekt PRACTICE (leping nr. ICT-609611).

Cybernetica teadusaruanded on Internetis saadaval aadressil
`http://www.cyber.ee/`

Postiaadress:
Cybernetica AS
Mäealuse 2/1
12618 Tallinn
Estonia

Secure multi-party computation system prototype for analyzing risks in value added tax declarations

Dan Bogdanov, Marko Jõemets, Sander Siim, Meril Vaht

June 9th, 2014

Abstract

Cybernetica AS and the Estonian Tax and Customs Board agreed to study methods for developing the KMD INF information system that would allow it to process the economic transactions of companies in an encrypted form. The goal of KMD INF is to collect value added tax (VAT) declarations and analyze them to detect VAT reimbursement fraud. The research and development work leading to this report was funded by the European Union Framework Programme 7 project PRACTICE (contract no 609611).

The study focused on four questions:

1. Can an information system with the complexity of KMD INF be implemented using today's secure multi-party computation technology?
2. Does the use of secure multi-party computation technology require changes to the processes or change the architecture of the information system?
3. What kind of restrictions does the use of secure multi-party computation technology enforce on the risk analysis methods?
4. What are the benefits of using secure multi-party computation in comparison with standard technologies?

We performed a systems analysis, implemented two prototypes of the KMD INF with a reduced scope on the Sharemind secure multi-party computation platform[1] and measured their performance. The results are as follows.

First, Sharemind supports the development of a simplified KMD INF information system with excellent confidentiality guarantees. This was confirmed through the development of two prototypes that supported the main activities of the KMD INF system. The economic transactions provided by honest tax paying companies would remain encrypted throughout the risk analysis process. Only company identifiers with an above-the-threshold risk score would be declassified to the Tax and Customs Board for further auditing. Secure multi-party computation also protects data from unauthorized access by both insiders and outside attackers. Sharemind significantly reduces the risk of compromise of the economic transactions, as that would require the theft of data from all three servers. If an attacker can combine the data from two of the three servers, it would still need to cheat the third server into participating in the risk analysis to recover the confidential data.

[1] Sharemind–`http://cyber.ee/sharemind`

Second, we found that changes to the processes and architecture will be needed. The changes are caused by the distributed nature of the Sharemind system. Sharemind stores information in a secret-shared form, distributed among three different servers. This also means that the taxpayer will be performing secret sharing at the source of the data, without ever sending unencrypted economic transactions to a third party. Fortunately, the secret sharing process can be integrated into web forms and accounting software.

Third, we found that secure multi-party computation also brings restrictions. The main restriction is that a secure multi-party computation application requires significantly more computing resources. Based on the calculations from the Tax and Customs Board, in Estonia, 80 000 companies will upload 50 million economic transactions every month. Our prototype could perform a preliminary risk analysis on the encrypted transactions of 80 000 companies in a week. This would need a Sharemind deployment on three servers with 24-core 3 GHz processors, all connected by 10 Gbit/s network connections.

As an additional restriction, it will be harder to hide the risk analysis algorithms. Today, the Tax and Customs Board performs risk analysis autonomously and unauthorized individuals have no knowledge of the kind of analyses that are performed. Secure multi-party computation with a system such as Sharemind will change that. All hosts of the secure computation system would need to agree to the risk analysis algorithms proposed by the Tax and Customs Board. Furthermore, Sharemind security guarantees hold when the servers are hosted by three independent organizations. Assuming that the Tax and Customs Board controls one of the servers, two more will be needed who will contribute computing resources and are trustworthy partners to the Tax and Customs Board.

Our research project came to the following conclusions:

1. Sharemind is a suitable platform for the risk analysis of confidential data and significantly improves the security guarantees of information systems that use it.
2. A KMD INF information system based on today's secure multi-party computation technology would be several times more expensive in both its resource use and computation time when compared to unencrypted solutions.
3. The use of Sharemind in an information system brings a shared responsibility for the data and improves the transparency of the system.

Cybernetica AS expects breakthroughs in the performance of secure multi-party computation systems in the coming years and believes that these will lower the cost of using the technology. Based on this prediction, we propose to consider the use of secure multi-party computation in systems such as KMD INF in two years time. Simpler systems can already be successfully implemented today.

Turvalist ühisarvutust kasutava käibemaksudeklaratsioonide riskianalüüsi süsteemi prototüüp

Dan Bogdanov, Marko Jõemets, Sander Siim, Meril Vaht

9. juuni 2014. a.

Kokkuvõte

Cybernetica AS ning Maksu- ja Tolliamet leppisid kokku uurimustöös, mis puudutab KMD INF süsteemi[1] loomist turvalise ühisarvutuse platvormi Sharemind abil nii, et see töötleks krüpteeritud kujul ettevõtete ostu- ja müügiarveid. Teadus- ja arendustöid teostati Euroopa Liidu 7. raamprogrammi teadusprojekti PRACTICE (grandileping nr 609611) raames. Vastuseid otsiti neljale küsimusele:

1. Kas infosüsteemi KMD INF keerukusega süsteemi saab ehitada praeguse turvalise ühisarvutuse tehnoloogia abil?
2. Kas turvalise ühisarvutuse tehnoloogia rakendamine esitab piiranguid protsessidele või muudab infosüsteemi arhitektuuri?
3. Millised piirangud seab riskianalüüsi meetoditele praeguse tehnoloogiatasemega turvalise ühisarvutuse kasutamine?
4. Millised eelised on turvalist ühisarvutust rakendaval süsteemil tavalise süsteemi ees?

Töö tulemusena viidi läbi süsteemianalüüs, teostati kaks piiratud käsitlusalaga KMD INF infosüsteemi prototüüpi turvalise ühisarvutuse platvormil Sharemind[2] ja mõõdeti nende prototüüpide jõudlust. Tulemused olid järgmised.

Esiteks, Sharemindi platvormi abil on võimalik ehitada väga heade turvagarantiidega KMD INF infosüsteem. Seda kinnitasid kaks edukalt teostatud prototüüpi, mis suutsid toetada KMD INF süsteemi tähtsamaid protsesse. Ausalt makse maksvate ettevõtete esitatud deklaratsioonide lisad koos arvetega jäävad krüpteeritud kujul varjatuks läbi kogu riskianalüüsi protsessi, kuna Maksu- ja Tolliametile avalikustatakse vaid nende ettevõtete nimed, kes on mõne riskianalüüsi käigus saanud kõrge riskiskoori. Lisaks on tugevaks turvagarantiiks kaitse väliste ja sisemiste ründajate vastu. Sharemindi kasutades väheneb märgatavalt risk, et käibedeklaratsioonide andmed varastatakse, kuna selleks tuleks varastada

[1] KMD INF–Mis ja milleks? `http://www.emta.ee/doc.php?34896`
[2] Sharemind–`http://cyber.ee/sharemind/`

andmed kõigist kolmest serverist või siis kahest serverist ja petta kolmas server endaga koos riskianalüüsi läbi viima.

Teiseks, muudatused infosüsteemi talitlusmallides ja arhitektuuris on vajalikud ning neid põhjustab Sharemindi hajutatud iseloom. Sharemind hoiab andmeid ühissalastatuna kolmeks osaks jaotatult. Sellest lähtudes peab ettevõte parima turvalisuse nimel andmeid üles laadides ühissalastuse ise läbi viima ning saadud osakud erinevatesse serveritesse laadima. Õnneks on ühissalastuse protsess integreeritav nii veebivormidesse kui raamatupidamissüsteemidesse.

Kolmandaks tuvastati, et turvalise ühisarvutuse rakendamine toob kaasa mitmeid kitsendusi. Esiteks kaasneb selle rakendamisega suurem arvutusressursside kulu. Kui eeldame, et igal kuul esitab deklaratsiooni 80 000 firmat ja kokku esitatakse 50 000 000 arvekirjet, siis pilootprojekti käigus arendatud prototüüp suudaks 80 000 firma krüpteeritud andmetel esialgsed riskianalüüsid ära teha nädalaga. Selleks on vaja Sharemind paigaldada kolmele serverile, milles on 24-tuumalised 3 GHz protsessorid ning serverite omavahelise ühendusvõrgu kiirus peab olema 10 Gbit/s.

Täiendavaks piiranguks loeme analüüsialgoritmide varjamise keerukust. Kui täna saab Maksu- ja Tolliamet riskianalüüse teha iseseisvalt ning volitamata isikud ei tea, milliseid analüüse tehakse, siis tänase turvalise ühisarvutuse tehnoloogiaga seda teha ei saa. Nimelt peaksid kõik ühisarvutuse serverite haldajad nõustuma selliste algoritmide käivitamisega. Sharemindi turvaeelduste kehtimiseks on vaja, et servereid haldaksid kolm sõltumatut osapoolt. Kui Maksu- ja Tolliamet kontrollib üht serverit, on vaja leida veel kaks serverit, mille puhul MTA on nõus nendega riskianalüüside algoritme jagama.

Kokkuvõtvalt on pilootprojekti järeldused järgmised.

1. Sharemind on sobiv platvorm konfidentsiaalsete andmete riskianalüüsiks ning tõstab oluliselt seda kasutavate infosüsteemide turvalisust.
2. Sharemindil põhinev KMD INF infosüsteem oleks tänast tehnoloogiat kasutades mitu korda kulukam nii ressursi- kui ka ajakasutuse mõttes.
3. Sharemindi kasutamine infosüsteemis loob jagatud vastutuse andmete eest ning tõstab andmeid töötleva süsteemi läbipaistvust.

Cybernetica AS hinnangul on turvalise ühisarvutuse tehnoloogia arengus lähema aasta-kahe jooksul oodata märgatavat läbimurret jõudluses ja see võib vähendada siin hinnatud kulusid. Viimasest lähtudes soovitame turvalise ühisarvutuse rakendamist KMD INF keerukusega infosüsteemide loomisel kaaluda kahe-kolme aasta perspektiivis. Lihtsamaid süsteeme on võimalik ehitada juba täna.

Sisukord

1 Määratlused ja lühendid

MTA	Maksu- ja Tolliamet
KMD	Käibedeklaratsioon. Kuna KMD INF on KMD lahutamatu osa, siis edaspidi selles dokumendis hõlmab lühend KMD ka lisade osa KMD INF. Käibedeklaratsiooni ilma lisadeta KMD INF on nimetatud KMD põhiosaks.
KMD INF	Käibedeklaratsiooni informatiivne lisa (KMD lisa), mis jaguneb A- ja B-osaks.
Süsteem KMD INF	Loodav süsteem, mis tegeleb KMD ja KMD INF dokumentide töötlemisega ning pakub liidest klientidele ja ametnikele nende dokumentide esitamiseks, parandamiseks ja vaatamiseks.

2 Taust

Maksu- ja Tolliameti (MTA) andmetel oli 2013. aasta novembri seisuga käibemaksukohuslasi 73 504 isikut. Igas kuus deklareerib käibemaksu enammaksmist ja küsib raha tagasi 26000–29000 ettevõtet. Keskmiselt küsitakse kuus tagasi 91 miljonit eurot sisendkäibemaksu.

2013. aasta juuli seisuga on Maksu- ja Tolliameti riskianalüüsi käigus välja selgitatud ligi 10000 käibemaksu riskiga isikut, kelle puhul on kahtlus, et nad on moonutanud oma käibedeklaratsioonidel andmeid maksukohustuse vältimiseks. Esialgsete hinnangute põhjal oli ettevõtete tekitatud käibemaksukahju 222 miljonit eurot aastas [4].

Ligi kahel kolmandikul nendest ettevõtetest ei ole sisulist majandustegevust ja neid võidakse kasutada pettuste läbiviimiseks. Ülejäänud on tegutsevad ettevõtted, kes püüavad käibemaksudeklaratsioonil andmeid näidata sellistena, et maksukohustus puuduks või oleks minimaalne. Sellise hulga ettevõtete kontrolliks kuluks Maksu- ja Tolliametil praeguste jõududega 11 aastat [4].

2013. aastal esitas Rahandusministeerium Riigikogu menetlusse käibemaksuseaduse muutmise eelnõu, mille eesmärk oli pettuste efektiivne vähendamine.

Parema kontrolli saavutamise eesmärgil lisandub käibemaksudeklaratsioonile kohustuslik lisa (KMD INF), mis jaguneb omakorda kaheks osaks – ostuosaks ja müügiosaks, milles ettevõtetel tuleb hakata kajastama nii ostu- kui ka müügiarveid (tehingupartnerite lõikes), kui tehingute summa tehingupartneriga on vähemalt 1000 eurot ilma käibemaksuta. Deklareerida on lubatud ka kõiki arveid, see vabastab 1000 euro piiri jälgimise kohustusest.

Muudatuse eesmärk on uuendada kontrollisüsteemi nii, et informatsiooni kogumine oleks arvutipõhine ning iga kontrollitava ettevõtte puhul ei peaks MTA eraldi nende ostu- ja müügiarvete infot pärima. Ostu- ja müügipoolte võrdlus annab MTA-le kergesti tehinguahela tuvastamise võimaluse ning muudab riskianalüüsi protsessi automatiseerituks.

MTA sõnul tõstab käibedeklaratsiooniga kogutav informatsioon oluliselt riskianalüüsi kvaliteeti ning kontrollimise kiirust ilma kontrollijate arvu suurendamiseta ning üldine kontrolliga kaasnev halduskoormus vastupidiselt üldlevinud arvamusele väheneb. Hetkel võrdleb MTA kuus keskmiselt ainult umbes 1300 isiku andmeid [4].

Käibedeklaratsiooni lisa võimaldab oluliselt kiiremini ja täpsemalt teha kindlaks isikud, kellele pöörata kõrgendatud tähelepanu. MTA usub, et arveandmete kogumise tulemusena paraneb varjatud käibe avastamine, lihtsustub fiktiivsete arvete ja puhvrite tuvastamine ning käibemaksu tagastusnõuete riskid täpsustuvad ja menetlemine kiireneb. Lisaks loodetakse seeläbi kindlustada võrdne konkurentsikeskkond ettevõtjatele ning tagada maksukoormust tõstmata riigitulude kasv.

Käibemaksuseaduse ja raamatupidamise seaduse muutmise seaduse eelnõu ettevõtjate heakskiitu ei saanud.

Eesti Maksumaksjate Liidule valmistab muret loodava „superandmebaasi“ turvalisus. Maksu- ja Tolliametis vahetub igal aastal sadu ametnikke ning ei saa välistada, et tulevikus nii mõnigi neist loob endale erasektorisse tööle mineku tarbeks korraliku andmebaasi tulevaste konkurentide hankijatest ja klientidest. Peamiseks riskiks loetakse siserünnet, mille tulemusena mõni KMD INF andmete juurde pääsev isik ei kasuta saadud informatsiooni eesmärgipäraselt.

Kolmanda lugemise tulemusena võttis Riigikogu 11. detsembril 2013 seaduse vastu, kuid Vabariigi President jättis seaduse välja kuulutamata, sest tema hinnangul on see vastuolus põhiseadusega, ning saatis ta tagasi Riigikogule uueks arutamiseks.

President Ilves rõhutas, et kõigi ettevõtjate koormamist täiendavate kulutuste ja kohustustega ning peaaegu kogu Eesti ärisaladust sisaldava andmebaasi loomist ei saa õigustada tõendamata oletusega, et nn maksuauk väheneb [5].

“Sedavõrd intensiivset põhiõiguste piirangut saab vaadelda põhiseaduspärasena üksnes juhul, kui suudetakse erinevaid käibemaksupettuse liike ja tüüpnäiteid analüüsides tõendada, et senised käibe varjamise ja riigilt käibemaksu väljapetmise skeemid muutuvad uue regulatsiooni tingimustes kas võimatuks või siis oluliselt kergemini avastatavaks ja neid ei ole võimalik kiiresti asendada uute tõhusate pettuseskeemidega.” [5]

Sellist analüüsi ei ole tehtud, märkis president Ilves. Tema hinnangul ei saa järelikult praegu usutavalt väita, et kavandatav ettevõtlusvabaduse piirang aitaks oluliselt parandada käibemaksu laekumist ning väldiks pettureile tegelikult tasumata käibemaksu tagastamist, samuti ei saa väita, et teisi võimalusi vähemalt samasuguses ulatuses käibemaksuaugu vähendamiseks ei ole [5].

Riigipea peab vajalikuks vältida olukorda, kus käibemaksuseaduse muudetud § 27 lg 1^2 jõustub, kuid tunnistatakse seejärel põhiseadusevastaseks ja kehtetuks, ent ettevõtjad on kulutused juba teinud või saanud karistada. Lisaks võimalikele kahju hüvitamise nõuetele, mis hakkaksid koormama riigieelarvet, kahjustaks taoline olukord Eesti ettevõtluskeskkonda ja mainet [5].

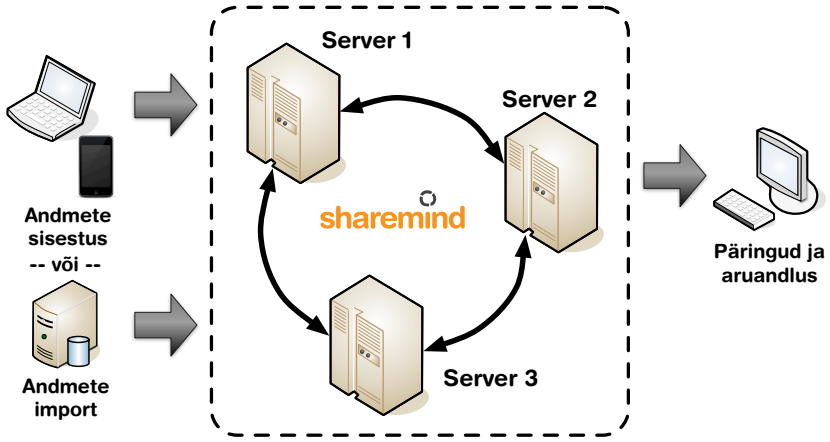
Algne plaan oli seadus jõustada alates 1. juulist 2014. Eelnõu võeti Riigikogus muudetud kujul taas vastu 7. mail 2014. Vabariigi President Toomas Hendrik Ilves kuulutas seadusemuudatuse välja 20. mail 2014.

3 Sissejuhatus Sharemind KMD INF pilootprojekti

Cybernetica AS sooritas käibemaksudeklaratsiooni ja selle lisade informatsiooni kogumise süsteemi (edaspidi KMD INF infosüsteem) analüüsijärgu ning projekteerimisjärgu 1. juulil 2013 sõlmitud raamlepingu 2.15/19 alusel. Projekt peatati pärast käibemaksuseaduse muutmise seaduse väljakuulutamata jätmist Vabariigi Presidendi poolt.

Kuna Cybernetica AS tegi samal ajal töid ka projektis PRACTICE ning Eesti Maksumaksjate Liidu ja President Ilvese poolt tõstatatud turvalisusprobleem oli täpselt selline, millele selles projektis lahendusi otsiti, kasutas Cybernetica AS võimalust proovida lahendada seda elulist probleemi projekti PRACTICE raames ning selgitada välja, kas Cybernetica AS väljatöötatud turvalise andmebaasisüsteemi Sharemind kasutamisega saab vähendada KMD INF infosüsteemi konfidentsiaalsusriski.

Sharemind on turvalisel ühisarvutusel põhinev andmete analüüsi lahendus, mis võimaldab mitmest allikast kogutud andmeid töödelda nii, et andmete omanike konfidentsiaalsus on tagatud [3]. Turvaline ühisarvutus on krüptograafiline meetod, millega saab digitaalsel kujul informatsiooni töödelda nii, et töötleja ei suuda eristada andmekirjete sisu ega ei saa kirjeid ka füüsiliste või juriidiliste isikutega siduda. Turvalise ühisarvutuse tehnoloogiat saab kasutada andmete kogumiseks, analüüsiks ja koondtulemuste avaldamiseks privaatsust säilitaval viisil (Joonis 1).



Joonis 1: Sharemindi rakenduste üldine mudel

Kuna Cybernetical oli akadeemiline huvi pilootprojekt läbiviimiseks ning Euroopa Komisjoni 7. raamprogrammi FP7/2007-2013 raames olid olemas ka finantsvahendid, lepiti MTA-ga kokku, et Cybernetica AS viib lahenduse katsetamiseks läbi pilootprojekti (edaspidi Sharemind KMD INF pilootprojekt või KMD INF

pilootprojekt), mille tulemusena soovitakse jõuda selgusele, kas arvutussüsteem Sharemind on piisavalt võimas selle probleemi lahendamiseks. Selleks tuleb leida vastused küsimustele, mida on kirjeldatud jaotises 4.

4 Pilootprojekti uurimisküsimused

Kas infosüsteemi KMD INF keerukusega süsteemi saab praeguse turvalise ühisarvutuse tehnoloogia abil ehitada?

Turvalise ühisarvutuse tehnoloogiat on katsetatud edukalt majandusandmete analüüsi süsteemi loomiseks, mida on ka reaalselt kasutanud Eesti Infotehnoloogia ja Telekommunikatsiooni Liit. Sharemindi abil on loodud ka mitmeid katserakendusi, mis näitavad, kuidas seda saab kasutada privaatsust säilitava statistilise analüüsi ja andmekaeve jaoks.

KMD INF infosüsteemi Sharemindi versiooni arendamisel võivad ilmneda sellised nõuded, mis Sharemindi rakenduste arendamisel pole seni ette tulnud. Pilootprojekti põhieesmärk on välja selgitada, kas olemasoleva ühisarvutuse tehnoloogia abil on võimalik KMD INF infosüsteemi ja teisi samalaadseid süsteeme edukalt luua.

Kas turvalise ühisarvutuse tehnoloogia rakendamine esitab piiranguid protsessidele või muudab infosüsteemi arhitektuuri?

Eelnevalt KMD INF analüüsi käigus modelleeritud protsesside kirjeldus ning kavandatud arhitektuur on vaja üle vaadata ning jõuda selgusele, kas ja kuidas Sharemindi turvalise ühisarvutuse tehnoloogia muudab nende struktuuri. Kui on vaja teha muudatusi, siis tuleb analüüsida, kas KMD INF infosüsteemi jaoks on selliste muudatuste rakendamine võimalik ja mõistlik.

Millised piirangud seab riskianalüüsi meetoditele praeguse tehnoloogiatasemega turvalise ühisarvutuse kasutamine?

Kuna KMD INF infosüsteem eeldab, et kogu andmebaasi peal peab olema võimalik läbi viia mitmesuguseid riskianalüüse, tuleb pilootprojekti käigus uurida, milliseid piiranguid seab turvalise ühisarvutuse tehnoloogia rakendamine riskianalüüsi protsessile ja meetoditele. Juhul, kui nimetatud tehnoloogia rakendamine seab riskianalüüside läbiviimisele piiranguid, tuleb otsustada, kas need piirangud on aktsepteeritavad.

Millised eelised on turvalist ühisarvutust rakendaval süsteemil tavalise süsteemi ees?

Turvaline ühisarvutus on krüptograafiline meetod, millega saab digitaalsel kujul informatsiooni töödelda nii, et töötleja ei näe andmeid ega oska neid omanikega siduda. See tähendab, et tundlike andmete töötlemine on võimalik nende omaniku konfidentsiaalsust rikkumata. Pealegi on turvaline ühisarvutus hajutatud protsess ja võimaldab rakendada kontrolli ja vastutuse jagamist ning vältida kõikvõimsa osapoole teket andmetöötluses.

Pilootprojekti käigus selgitatakse täpsemalt, millised eelised ja puudused kaasnevad KMD INF infosüsteemi teostamisega Sharemindi tehnoloogia abil.

5 Sharemindi tehnoloogial põhineva KMD INF infosüsteemi arhitektuuri ja funktsionaalse lahenduse erinevused tavalisest lahendusest

See jaotis kirjeldab pilootprojekti käsitlusala ning projekti läbiviimiseks ja protsesside lihtsustamiseks tehtavaid muudatusi. Ta annab ka ülevaate olemasoleva KMD INF süsteemianalüüsi tulemuse ja uue tehnoloogia rakendamisega kaasneva süsteemianalüüsi tulemuse erinevustest – kirjeldatud on erinevused algse ja pilootprojekti jaoks loodud komponendiskeemi vahel.

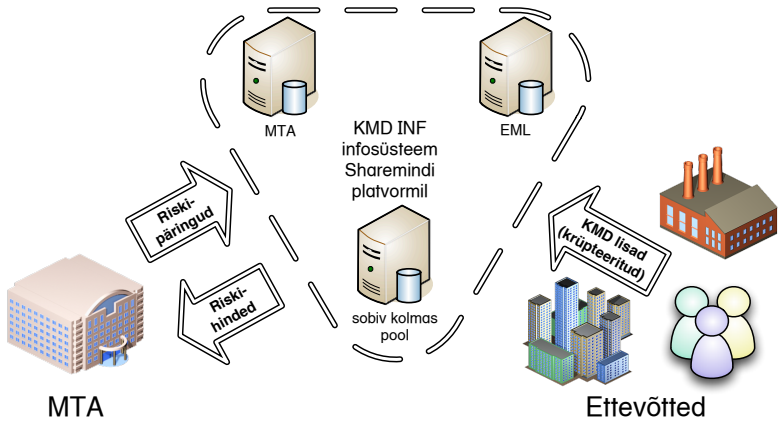
5.1 Käsitlusala ja muudatused

KMD INF pilootprojekti käsitlusala on tavalise lahendusega võrreldes vähendatud – teostatakse ainult deklaratsioonide esitamise, agregeerimise ja andmete analüüsiga seotud funktsioonid. Andmete muutmisega ja pärast riskianalüüsi tulemuste saamist otsuste vastuvõtmisega seotud protsesse pilootprojektis ei käsitleta. Pilootprojektis kasutatakse vaid ühe konkreetse perioodi andmeid. Pilootprojekti käigus realiseeritakse ka riskianalüüsi osa, mis tegelikult ei kuulu KMD INF projekti käsitlusalasse, ning viiakse läbi riskianalüüsid, selgitamaks välja, kas kasutatav tehnoloogia võimaldab kõigi vajalike riskianalüüside läbiviimist. Tunduvalt on vähendatud ja lihtsustatud ka andmemudelit. Seda on tehtud eesmärgiga vähendada pilootprojekti keerukust.

Kuna andmetöötlussüsteem Sharemind kasutab andmete konfidentsiaalsuse kaitseks ühissalastuse meetodit, peab Sharemindi rakendusi paigaldama kolme eraldi serverisse, mida haldavad eraldi asutused. See aitab oluliselt vähendada andmete lekkimise võimalust, sest isegi, kui ühe serverihoidja süsteemi suudetakse sisse murda, on andmete konfidentsiaalsus tänu ühissalastamise meetodile kaitstud.

Sharemindi-põhise KMD INF rakenduse üks võimalik teostusmudel näeks välja järgmine (Joonis 2). Kolmeks serverihoidjaks võiksid olla Maksu- ja Tolliamet, Eesti Maksumaksjate Liit ja mõni kolmas sobiva tegevusalaga asutus, näiteks Andmekaitse Inspektsioon vms. Rakendus võimaldaks Eesti ettevõtetel oma deklaratsioonid esitada nii, et andmed jaotataks ühissalastamise abil kolme Sharemindi serveri vahel ning ka serverihoidjad ise ei näe firmade pärisandmeid. Ühtlasi oleks võimalik MTA-l kogutud andmeid siiski analüüsida, et tuvastada võimalikke maksupettureid. Nende riskianalüüside tulemused saadetaks siis ainult MTA-le ning teised serverihoidjad neid tulemusi ei näe.

Pilootprojekti jooksul valmis prototüübist kaks versiooni. Esimese versiooni käsitlusala oli võimalikult lihtsustatud ega toetanud andmete paralleelset agregeerimist. Teises versioonis prooviti olemasolevaid arvutusi võimalikult tõhusalt rakendada, jaotades andmed andmebaasi tabelite vahel ja agregeerides neid kõigis tabelites paralleelselt, millega saavutati oluline võit jõudluses. Mõlemas versioonis kasutatud andmemudelid ja teostatud arvutused on kirjeldatud dokumendi lõpus lisades.



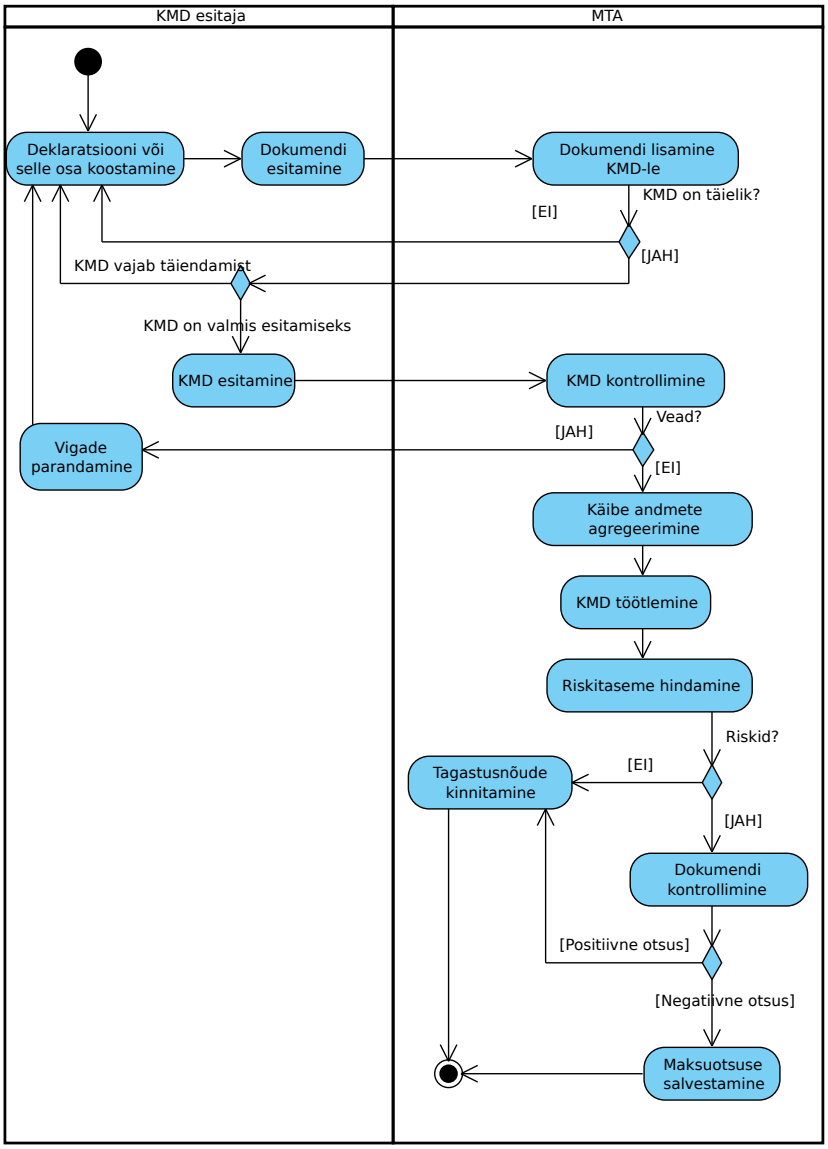
Joonis 2: KMD INF pilootprojekti prototüübi kujuteldav teostusmudel

Algselt oli kavas salvestada agregeeritud andmed kahte agregeeritud tabelisse – müügiandmete tabelisse ja ostuandmete tabelisse. KMD INF piloodis salvestatakse agregeeritud andmed samuti müügiandmete ja ostuandmete tabelisse, kuid arvutuste kiiruse tõstmiseks salvestatakse mõned vahetulemused lisatabelitesse (vt lisad 1 ja 2).

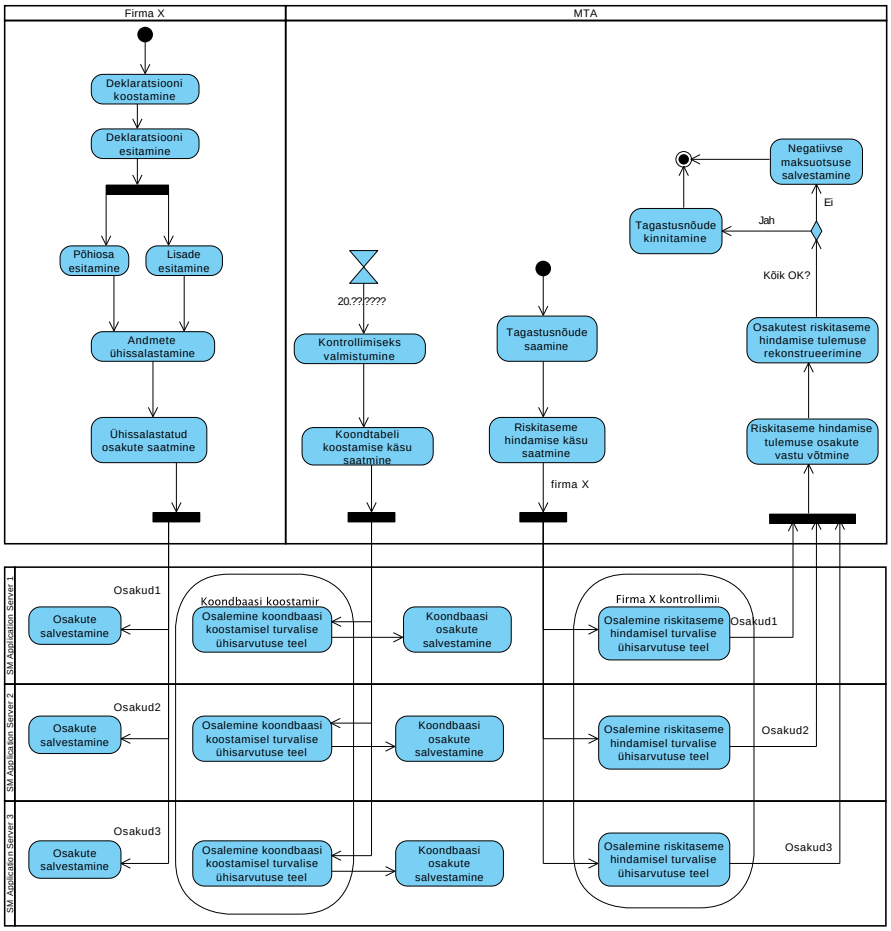
Pilootprojektis kasutatava riskianalüüsi meetodi nr 3 jaoks on samuti kasutusele võetud kaks lisatabelit. Vastavat lahendust kirjeldab jaotis 6.4.

Võrreldes algset tegevusskeemi (Joonis 3) skeemiga, mis on loodud pilootprojekti jaoks (Joonis 4) näeme, et turvalise ühisarvutuse tehnoloogia rakendamine ei muuda kliendi ega ametniku töökorraldust. Muudatused on süsteemisisesed ning kasutajale need ei paista.

Pilootprojekti raames ei teostatud paranduskannete esitamise ja deklaratsiooni mitmes osas esitamise võimalust. Lubatud on iga firma andmeid esitada ainult üks kord. Kui esitatud deklaratsiooni valideerimine ei õnnestu, antakse sellest kasutajale teada ja andmeid ei salvestata, seega on võimalus korrektne deklaratsioon uuesti esitada.



Joonis 3: Tavalise KMD INF infosüsteemi tegevusskeem [1]



Joonis 4: KMD INF pilootprojekti prototüübi tegevusskeem

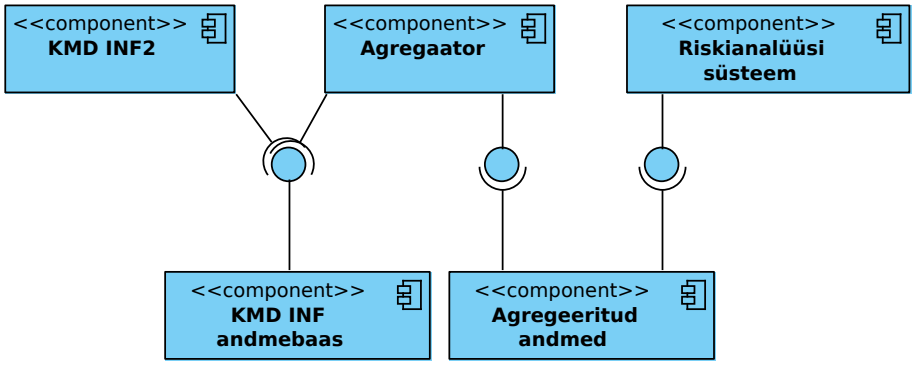
5.2 Komponendiskeem

Algselt realiseeris KMD INF infosüsteemi arhitektuur (Joonis 5) järgmisi põhimõtteid. Iga saabunud algdokument salvestatakse kohe KMD INF andmebaasi. Pärast saabunud dokumendi andmebaasi salvestamist algatatakse andmete agregeerimine ning lisatakse saadud tulemus agregeeritud müügi- ja ostutabelisse agregeeritud andmete andmebaasis. Agregeeritud andmed on kättesaadavad KMD INF süsteemist eraldiseisvale riskianalüüse sooritavale süsteemile, mis kasutab koondbaasi andmeid riskianalüüside läbiviimiseks.

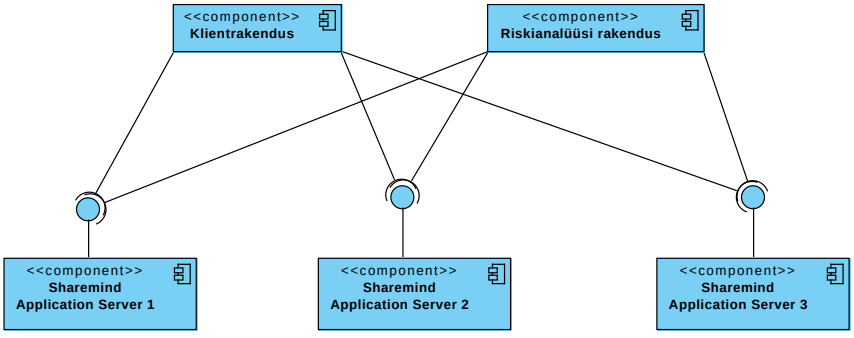
Turvalise ühisarvutuse tehnoloogia rakendamisega muutub süsteemi arhitektuur järgmiselt (vt Joonis 6). Selle asemel, et kõiki andmeid salvestada ja hoiustada ühel serveril, võetakse kasutusele kolm serverit. Andmed jaotatakse kolme

serveri vahel ühissalastuse abil. Sellega saavutatakse andmete konfidentsiaalne salvestus, sest ühissalastus saavutab sama turvaeesmärgi nagu krüpteerimine. Firmad saavad oma andmeid ühissalastatult üles laadida ka paralleelselt. Pilootprojekti prototüübi esimeses versioonis alustatakse andmete agregeerimist siis, kui kõigi ettevõtete andmed on laekunud ja salvestatud. Teises versioonis on teostatud sünkroonne andmete agregeerimine üleslaadimise ajal, nagu ka tavalises KMD INF infosüsteemis.

Agregeeritud andmed salvestatakse agregeeritud müügi- ja ostutabelis jaotatuna samade kolme serveri vahel. Agregeeritud andmetega viiakse läbi riskianalüüsi protsess ning tagastatakse nende firmade nimekiri, kellega oli seotud vähemalt üks risk.



Joonis 5: Tavalise KMD INF infosüsteemi komponendiskeem [2]



Joonis 6: KMD INF pilootprojekti prototüübi komponendiskeem

6 Andmete agregeerimine ja riskianalüüsi meetodid

See jaotis annab ülevaate andmete agregeerimisest ja riskianalüüsi meetoditest, mida pilootprojekti käigus kasutatakse. Kuna KMD INF pilootprojekti jaoks on vähendatud andmemudelit, on riskianalüüsi meetodid valitud selle põhjal, milliseid andmeid on võimalik selle vähendatud andmemudeliga kasutada.

Kuna teave riskianalüüside tegelike meetodite kohta on konfidentsiaalne, on pilootprojektis kasutatud näitlikke riskianalüüsi meetodeid, mis peegeldavad vähemalt osaliselt tegeliku riskianalüüsi protsessi keerukust. Näidismetoodika koostas Cybernetica AS MTA-lt saadud andmete põhjal.

Kuna Sharemind turvaliste arvutuste süsteem eeldab, et kõik arvutuskäigud tehakse samuti Sharemind tehnoloogiat kasutades, ei saa andmeid töödeldes kasutada tavalisi andmete analüüsi vahendeid, vaid vajalikud analüüsialgoritmid tuleb Sharemind tehnoloogiaga uuesti teostada.

6.1 Andmete agregeerimine

Müügi ja ostu koondtabelite loomise lihtsustamiseks salvestatakse juba andmete üleslaadimisel eraldi tabelitesse kõik unikaalsed tehingupartnerite paarid. Koondtabelitesse salvestataksegi iga tehingupartnerite paari kohta vastava perioodi agregeeritud andmed, mis arvutatakse nii deklaratsiooni põhiosa kui ka müügi- ja ostulisade andmete põhjal.

Müügi koondtabelisse salvestatakse riskianalüüsiks iga tehingupartnerite paari kohta järgnevad andmed:

- maksukohuslase (müüja) deklareeritud müügitehingute summa selle tehingupartneriga,
- tehingupartneri deklareeritud ostutehingute summa maksukohuslase suhtes (võib välismaa firma puhul olla tühi),
- kahe eelmise agregeeritud tulemuse vahe (deklareeritud müügi summast lahutatud partneri deklareeritud ostude summa),
- maksukohuslase (müüja) müügilisas deklareeritud kõigi tehingupartneritega sooritatud müügitehingute summa protsent põhiosas deklareeritud kogu müügisummast (ei sõltu tehingupartnerist),
- märge selle kohta, kas maksukohuslase tehingupartner (ostja) on deklaratsiooni esitanud,
- märge selle kohta, kas riskianalüüs on tuvastanud maksukohuslase tehingupartneril (ostjal) mõne riski (uueneb jooksvalt riskianalüüsi käigus, hoitakse eraldi tabelis).

Sarnaselt salvestatakse ostu koondtabelisse ostutehingute kohta käivad andmed:

- maksukohuslase (ostja) deklareeritud ostutehingute summa selle tehingupartneriga,
- tehingupartneri deklareeritud müügitehingute summa maksukohuslase suhtes,
- kahe eelmise agregeeritud tulemuse vahe (deklareeritud ostu summast lahutatud partneri deklareeritud müügi summa),
- maksukohuslase (müüja) ostulisas deklareeritud kõigi tehingupartneritega sooritatud ostutehingute summa protsent põhiosas deklareeritud kogu ostusummast (ei sõltu tehingupartnerist),
- märge selle kohta, kas maksukohuslase tehingupartner (müüja) on deklaratsiooni esitanud,
- märge selle kohta, kas riskianalüüs on tuvastanud maksukohuslase tehingupartneril (müüjal) mõne riski (uueneb jooksvalt riskianalüüsi käigus, hoitakse eraldi tabelis).

6.2 Riskianalüüsi meetod 1

Kõige lihtsama riskianalüüsi meetodi järgi võrreldakse maksukohuslase deklareeritud müügiarvete summat ühe tehingupartneri lõikes partneri deklareeritud ostuarvete summaga. Selleks kasutatakse müügi koondtabelis arvutatud kahe agregeeritud tulemuse vahet. Kui vahe on negatiivne, loetakse pilootprojektis seda tehingupartneri (ostja) suhtes riskiks.

Sarnane kontroll teostatakse ostu koondtabelil, kus kontrollitakse maksukohuslase deklareeritud ostu ja partneri deklareeritud müügi summade vahet. Kui tulemus on suurem nullist, arvestatakse seda maksukohuslase (müüja) riskina.

6.3 Riskianalüüsi meetod 2

Teise valitud riskianalüüsi meetodi järgi kontrollitakse maksukohuslase müügilisas deklareeritud müügitehingute summa osakaalu põhiosas deklareeritud kogu müügi summa järgi. Vastav protsent loetakse otse müügi koondtabelist. Pilootprojektis loetakse alla 30% osakaalu maksukohuslase riskiks.

Sarnaselt kontrollitakse ka maksukohuslase ostutehingute osakaalu ostukoondtabelis oleva protsendi järgi. Kui protsent on alla 30%, on see maksukohuslase risk.

6.4 Riskianalüüsi meetod 3

Viimasena väljavalitud riskianalüüsi algoritmi eesmärk on kontrollida, kas maksukohuslasega seotud partneritel on esinenud riske. Kui risk ilmneb, lisatakse maksukohuslase ja tema partneri registrikoodide paar eraldi selle riskianalüüsi meetodi jaoks loodud ostuarvete või müügiarvetega seotud tabelisse. Kui maksukohuslase partneril on esinenud riske, loetakse pilootprojektis seda võimalikuks riskiks maksukohuslase suhtes edaspidiste riskianalüüside jaoks.

7 Sharemindi-põhise infosüsteemi omadused võrreldes tavarakendusega

Sharemindi-põhise KMD INF infosüsteemi peamine eelis seisneb selles, et firmade esitatud deklaratsioonide andmed (tehingupartnerite nimed, tehingute suurused) salvestatakse süsteemi ühissalastatud kujul ja seetõttu on andmete konfidentsiaalsus tagatud serverihoidjate ning osaliselt ka väliste ründajate eest. Väline ründaja peaks andmete kättesaamiseks ründama korraga kolmes eraldi asutuses asuvaid Sharemindi servereid. See muudab välise ründe oluliselt keerukamaks, eeldusel et serverihoidjad kasutavad serverite kaitseks korrektselt nüüdisaegseid turvameetmeid. Kolme eraldi serveri haldamine suurendab aga kogu süsteemi ülalpidamise keerukust ja administratiivseid kulusid.

Välisel ründajal on ka võimalus teatavas koguses informatsiooni kätte saada siis, kui tal õnnestub mõne arvutuse teostamise ajal vähemalt kahte Sharemindi serverisse sisse murda. Sellisel juhul, kui ründaja on hõivanud kaks serverit kolmest, ei kehti enam Sharemindi garantiid andmete konfidentsiaalsuse kohta. Milliseid andmeid ründaja selles olukorras võib näha, sõltub nii teostatavast analüüsist kui ka ründe kestusest, see tähendab kui pikka aega õnnestub ründajal mõlemat serverit hõivata.

Andmete konfidentsiaalsus serverihoidjate eest on samuti tagatud, eeldusel, et ühelgi serverihoidjal ei ole juurdepääsu teistele arvutustes osalevatele Sharemindi serveritele. Kaks serverihoidjat võivad omavahel koostööd tehes saavutada samasuguse olukorra kui väline ründaja, kes on sisse murdnud kahte serverisse. Kõigi kolme serverihoidja koostöös on neil kohe võimalik kõiki salajasi andmeid näha. Seetõttu on oluline, et Sharemindi servereid haldaksid erinevad asutused, kellel poleks motivatsiooni konkreetse rakenduse kontekstis koostööd tehes konfidentsiaalsetele andmetele juurde pääseda.

Ühissalastuse eelis tavalise andmete krüpteerimise ees on see, et ühissalastatud andmetega saab hiljem arvutusi teostada ilma konfidentsiaalsust rikkumata. KMD INF pilootprojektis kasutatakse seda omadust riskianalüüsi tegemiseks ilma andmeid avalikustamata. Avalikustada saab riskianalüüsi tulemusi ainult kõigi osapoolte kokkuleppel. See tähendab, et kõigile serverihoidjatele lisandub kohustus auditeerida Sharemindi serverites jooksvate riskianalüüsi algoritmide koodi. Koodiauditid peaksid olema eri asutustes sõltumatud, et vähendada ebaturvalise koodi serveritesse sattumise ohtu.

Koodiauditeid aga lihtsustab oluliselt see, et Sharemindile kirjutatud rakendused on teostatud selleks spetsiaalselt Cyberneticas välja töötatud keele SecreC abil. SecreC eesmärk on turvalisi arvutusi sooritavate rakenduste kiire, lihtne ja arusaadav teostamine. SecreC-programmid kompileeritakse spetsiaalse baitkoodi kujule, mida saab käivitada Sharemindi platvormil. Ka käesoleva prototüübi Sharemindil töötavad turvalised arvutused on teostatud täielikult SecreC-s.

SecreC sarnaneb levinumate programmeerimise kõrgkeeltega nagu C, C++ ja Java, mis teeb SecreC koodist arusaamise programmerimistaustaga inimesele intuitiivseks. Kõige olulisem SecreC-keele omadus on konfidentsiaalsete ja ava-

like andmetüüpide selge eristus. Igas SecreC rakenduses kasutatavad konfidentsiaalsed andmed on selgesti deklareeritud vastavate võtmesõnadega. Samuti on SecreC-keeles igasuguste andmete avalikustamine alati selgelt eristatud protseduuriga declassify. Seega on kerge programmikoodi lugedes näha, mis hetkel ja milliseid andmeid programm avalikustab ning milline teave jääb salajaseks.

KMD INF rakenduse puhul leevendaks andmete konfidentsiaalsuse tagamine Sharemindi-põhise rakendusega oluliselt Eesti Maksumaksjate Liidu poolt tõstatatud loodava „superandmebaasi“ turvalisusega seotud probleeme, kuna firmade äriandmeid ei hoitaks enam loetaval kujul ühes kohas koos. See tähendab, et ühe serverihoidja käes olevatest ühissalastatud andmetest ei ole võimalik midagi välja lugeda tegelike andmete kohta. Kuigi põhiandmete konfidentsiaalsus on Sharemindi abil tagatud, on võimalikule ründajale ja serverihoidjatele siiski mingil määral arvutuste käigus näha kaudset teavet kaitstavate andmete kohta. Sellegipoolest tagab ühissalastatuse meetod krüptograafiliselt ettevõtete esitatud deklaratsioonide andmete konfidentsiaalsuse.

Teatud juhtudel on mõistlik Sharemindi rakendustes parema jõudluse saavutamiseks avalikustada arvutuse käigus seda teostavale programmile või serverihoidjale mõningaid kaudseid vahetulemusi. See tähendab, et avalikustatud tulemused ei ole enam ühissalastatud ning on soovi korral igale üksikule serverihoidjale näha. Loomulikult ei avalikustata sellisel moel konfidentsiaalseid sisendandmeid, kuid arvutuste jõudluse suurendamiseks on vahel kasulik avalikustada sellist teavet, mis on konkreetse rakenduse kontekstis piisavalt kaudne, et mitte rikkuda kaitstavate andmete konfidentsiaalsust.

Näiteks avalikustatakse projektis teostatud prototüübis mõnede analüüside käigus anonüümselt firma tehingupartnerite arv. See tähendab, et ei avalikustata konkreetse firma tehingupartnerite arvu, vaid seda, et andmetes esineb firma, kellel on niisugune arv tehingupartnereid. Isegi kui selle teabe järgi on mingi konkreetne firma identifitseeritav, selgub sellest ainult fakt, et see firma on oma deklaratsiooni esitanud, mida on nii või teisiti raske varjata, sest deklaratsiooni esitamine on kohustuslik. Sellised kaudse teabe avalikustamised arvutuse sooritajale võimaldavad krüpteeritud andmete analüüsil vältida kõigi võimalike andmeväärtuste läbivaatamist ja sellega kiirendada arvutusi.

Seega tuleb iga konkreetse rakenduse juures leida tasakaal jõudluse ja andmete konfidentsiaalsuse vahel. Mõnes kontekstis võib mingi teabe avalikustamine olla aktsepteeritav, aga teises mitte. Prototüübis on proovitud leida võimalikult mõistlik kompromiss jõudluse ja andmete avalikustamise vahel, lähtudes algse KMD INF infosüsteemi juures tõstatatud probleemidest. Andmete konfidentsiaalsuse küsimusi Sharemind KMD INF prototüübis kirjeldavad täpsemalt lisad 1 ja 2.

KMD INF pilootprojekti raames valminud prototüüp näitab ka seda, et sellise rakenduse teostamisel turvalise ühissalastuse abil ei ole tarvis rakenduse väliseid talitlusmalle muuta. Muutused on ennekõike süsteemisisesed ning võimaldavad teha tavarakendusele sarnaseid arvutusi ja analüüse, kuid saavutavad seda privaatsust säilitaval moel. Prototüübi teises versioonis on võimalik andmete üheaegne üleslaadimine ning pidev agregeerimine, nagu see oli planeeritud

ka algses rakenduses. Loodud koondtabelitelt on võimalik juba vajalikke riskianalüüsi algoritme kasutades leida firmade riskihinnanguid ja need volitatud osapoolele avalikustada.

Kuigi krüptograafiliselt on see võimalik, ei ole hetkel Sharemindi rakendustes võimalik käivitatavaid algoritme konfidentsiaalsena hoida. Samuti tähendaks see, et serverihoidjatel ei oleks võimalik käivitatavat rakendust auditeerida. See aga suurendab oluliselt riski, et salajane arvutus avalikustab andmeid, mida osapooled ei oleks nõus tegelikult avalikustama. Sellele probleemile praegu otsest lahendust ei ole. Seega hetkel oleksid KMD INF infosüsteemi puhul riskianalüüsi teostavad arvutused avalikud vähemasti kõigile kolmele serverihoidjale. See tähendab, et teada oleks käivitatavate algoritmide kuju, kuid teadmata jääksid andmed, millega nad töötavad.

Hetkel ei ole ka prototüübi teise versiooniga kõigi Eesti firmade andmete analüüs mõistliku ajaga sooritatav, küll aga on teises versioonis jõudlust väga oluliselt tõstetud võrreldes esimese versiooniga (vt lisa 1 ja lisa 2). Prototüübi teise versiooniga saaks võimeka riistvara peal kõigi Eesti firmade andmed mõne päevaga agregeerida (vt täpsemalt lisa 1 jaotis L2.4). Samuti ei tekiks andmete üleslaadimisel olulist jõudluse probleemi. Riskianalüüs ei ole aga hetkel nii suure jõudlusega, et oleks võimalik ta igakuine rakendamine. See nõuaks veel lisa-arendustöid prototüübiga. Ka on prototüübis teostatud riskianalüüsi käigus võimalik, et serverihoidjatele saab avalikuks mingi hulk firmasid, kellel leiti mõni risk. See on võimalik siis, kui firma, kellel risk leiti, on üheselt identifitseeritav oma tehingupartnerite arvu järgi (vt täpsemalt lisa 1 jaotis L1.3.3). Riskianalüüsi saaks teha ka nii, et ükski riskiga firma ei saaks avalikuks, kuid selles prototüübis oleks see tähendanud oluliselt suuremat lisatööd ja seetõttu jäi see käesoleva projekti võimalustest välja. Kuna Sharemindi platvorm on pidevas arengus, on kindlasti tulevikus rakenduste ehitamine ja suurte andmemahtude toetamine lihtsam ja ka praegu on see juba piisava arendustöö ja riistvaraga võimalik.

Tabelis 1 on esitatud lühike kokkuvõtte tavalise ja Sharemindi-põhise süsteemi omadustest.

	Tavaline lahendus	**Sharemindiga lahendus**
Turvalisus	Palju konfidentsiaalseid andmeid ühes kohas koos on potentsiaalne turvarisk	Andmete konfidentsiaalsus on tagatud, süsteemi hajutatuse tõttu on andmete leke sissemurdmise tagajärjel väga ebatõenäoline
Keerukus	Väike - tuntud tehnoloogiad ja kasutusmallid	Suurem keerukus – koodiauditid ja algoritmide realiseerimine uue tehnoloogia abil
Funktsionaalsus	Täidab funktsionaalsed nõuded	Tavalahendusele väga lähedal, arvutusalgoritmide peitmine pole hetkel võimalik
Jõudlus	Väga hea	Töökõlblik, kuid vajab oluliselt rohkem ressursse ja prototüübi edasiarendamist

Tabel 1: Tavalahenduse ja Sharemindi-põhise lahenduse võrdlus

8 Sharemindi vajalikud arengusuunad

Järgnevas on kirjeldatud Sharemindi olulisemad kitsendused ja vajalikud arendussuunad KMD INF laadsete infosüsteemide lihtsamaks arendamiseks tulevikus.

8.1 Paindlikum andmebaasiliides

Praegune Sharemindi andmebaasiteek on suhteliselt piiratud võimalustega. Üsna palju lihtsustaks andmete töötlemist võimalus muuta olemasolevaid andmebaasi ridu. Hetkel on võimalik ainult uusi ridu lisada või terve tabel kustutada. Andmebaasiteegi lisafunktsioonid on aga juba Sharemindi arendusplaani võetud. Selle prototüübi kontekstis saadi ka ilma nendeta hakkama, kuid mitme perioodi andmete käsitlemiseks oleks tarvis paindlikumat andmebaasiteeki.

8.2 Teabe avalikustamise aktsepteeritav määr

Sharemindi rakendustes pole lihtne teha kompromisse andmete konfidentsiaalsuse ja jõudluse vahel. On tarvis hoolikalt läbi mõelda, kas mingi teabe avalikustamine on ohtlik või aktsepteeritav. See nõuab koostööd andmete omanike, andmeid kasutavate poolte ja rakenduse arendajate vahel. Mingid kompromissid on praegu veel vajalikud, sest hetke tehnoloogiatega ei ole muidu võimalik reaalses elus ettetulevaid andmemahte toetada. Siinkohal aitaksid automatiseeritud

rakendusi auditeerivad ja turvaanalüüse sooritavad tööriistad. Selliste tööriistade väljaarendamine on hetkel Cyberneticas üks oluline teadustöö suund.

8.3 Rakenduse auditeerimine

Kõik serverihoidjad peavad hoolikalt auditeerima Sharemindi serverites käivitatavaid algoritme – see võime tuleb asutuses tekitada. Kui serverihoidjad lepivad kokku ebaturvalise koodi käivitamises, võivad konfidentsiaalsed andmed avalikuks tulla. Ka siin lihtsustaksid rakendusväliseid protsesse automaatsed tööriistad.

8.4 Arvutusalgoritmide varjamine

Kasutatavate algoritmide varjamine ei ole praegu veel Sharemindi peal võimalik, kuna on oluline, et serverihoidjad saaksid turvalisi arvutusi tegevat koodi auditeerida. Juba KMD INF rakenduses läheks aga MTA seisukohast sellist omadust vaja, et varjata KMD lisadel tehtavate riskianalüüside iseloomu turvalise ühisarvutuse partnerite eest. Sellele probleemile ei ole veel praegu selget lahendust ning see on kindlasti oluline Sharemindi tulevase edasiarenduse teema.

Viited

[1] Cybernetica AS. Süsteemi KMD2 talitlusmallimudel 2.1., 03.12.2013.

[2] Cybernetica AS. Süsteemi KMD2 arhitektuur 1.1., 29.01.2014.

[3] Dan Bogdanov. *Sharemind: programmeeritav turvaline arvutussüsteem praktiliste rakendustega.* Doktoritöö, Tartu Ülikool, 2013.

[4] Käibedeklaratsiooni lisa. `http://www.fin.ee/kaibedeklaratsiooni-lisa/`, 21.02.2014.

[5] Otsus 348. käibemaksuseaduse ja raamatupidamise seaduse muutmise seaduse väljakuulutamata jätmine. `http://president.ee/et/ametitegevus/otsused/9726-2013-12-18-11-37-04/index.html`, 18.12.2013.

Lisad

Lisa 1 Prototüübi esimese versiooni kirjeldus

Prototüübi esimeses versioonis on teostatud võimalikult lihtsa ülesehitusega rakendus, et näha, millise jõudluse saavutab kõige lihtsamini teostatav lahendus ning et oleks võrdlusvõimalus teise, optimeeritud versiooniga.

L1.1 Andmemudel

Prototüübi andmemudelit on võrreldes tavarakendusega oluliselt lihtsustatud, et vähendada prototüübi keerukust. Peamised lihtsustused on järgmised:

- Rakenduse käsitlusala on vähendatud nii, et ta katab ainult ühe perioodi andmeid. Selle perioodi nimetus kuskil andmetes ei kajastu, lihtsalt vaikimisi on eeldus, et kõik andmed on ühe ja sama perioodi kohta.
- Vähendatud on deklaratsiooni XML-kuju andmemudelit. Prototüüp käsitleb deklaratsiooni põhiosas olevaid terve perioodi summasid ning müügi- ja ostulisade arveridu.
- Oluliselt on vähendatud agregeeritud andmetega koondtabelite veergude arvu. Koondtabelis olevaid andmeid on kirjeldatud selle dokumendi jaotises 6.1.
- Kõiki rahalisi väärtusi hoitakse Sharemindi andmebaasis täisarvudena sentides, st. € 12.50 asemel on andmebaasis arv 1250. Täisarvudega arvutused on täpsed ja kiiremad kui ujukomaarvudega.
- Andmemudelit on lihtsustatud selle eeldusega, et firmade registrikoodidel on ainult numbrilised väärtused, st. neid saab otse täisarvuks teisendada ega pea sõnena käsitlema; see lihtsustab arvutusi.

Lisaks lihtsustustele kasutab KMD INF pilootprojekti prototüübi esimene versioon jõudluse tõstmiseks vahetabeleid.

Koondtabelite loomisel kasutatakse prototüübis lisaks kaht andmebaasitabelit unikaalsete tehingupartnerite paaride salvestamiseks. Esimesse tabelisse salvestatakse andmete üleslaadimisel ühissalastatult kõik firmade registrikoodide paarid (X, Y), kus firma X on esitanud deklaratsiooni ja seal kajastub müügitehing firmaga Y. Iga paar esineb tabelis ülimalt üks kord. Unikaalsed ostutehingute partnerid salvestatakse sarnaselt paaridena, kus X on sooritanud Y-ga ostutehingu. Need tabelid uuenevad iga esitatud deklaratsiooniga.

Riskianalüüsi protsessi ühe osana märgitakse ära kõik sellised olemasolevad tehingupartnerite paarid, kus tehingupartneril on avastatud risk. See teave talletatakse ühissalastatud andmebaasi edasiseks tarbeks. Tavarakenduses võiks

selle teabe lisada koondtabelitesse ühe veeruna, sest iga koondtabeli rida on üks-üheses vastavuses ühe tehingupartnerite paariga, kuid Sharemindi-põhises rakenduses tuleb need paarid eraldi tabelitesse salvestada. Hetkel ei võimalda Sharemindi andmebaasiliides tabelisse kirjutatud rida hiljem muuta, et aga koondtabeli koostamise hetkel riske leitud pole, salvestatakse leitud riskiga paarid riskianalüüsi käigus eraldi tabelisse. Olemasoleva rea salajane muutmine on aga teoreetiliselt täiesti võimalik operatsioon.

L1.2 Rakenduse põhiprotsesside kirjeldus

Prototüübi esimeses versioonis jaotub kogu rakendus kolme eraldiseisvasse etappi.

Andmete üleslaadimine — kõik firmad laevad oma deklaratsiooni XML-vormingus failina üles ja Sharemindi andmebaasi salvestatakse deklaratsioonist saadud väärtused ühissalastatud kujul. XML-faili ennast otseselt kuhugi ei salvestata, andmed ühissalastatakse kohe klientrakenduses ja Sharemindi serveritesse saadetakse juba ühissalastatud andmed. On võimalik mitut deklaratsiooni paralleelselt üles laadida. Eeldame, et iga firma esitab oma deklaratsiooni ainult üks kord ja paranduskandeid ei esitata.

Koondtabelite koostamine — kui kõik firmad on oma deklaratsioonid esitanud, koostatakse kõigi ühissalastatud müügi- ja ostuarvete pealt korraga nii müügi- kui ka ostukoondtabel. See on ühekordne arvutus, mis eeldab, et selleks ajaks on kõik deklaratsioonid esitatud.

Riskianalüüs — terve müügi- ja ostukoondtabeli peal rakendatakse riskianalüüsi meetodeid, mis on kirjeldatud jaotistes 6.2, 6.3 ja 6.4. Riskianalüüsi tulemusena leitakse nende firmade registrikoodid, kellel on mõni risk. Iga riski kohta, mida analüüsitakse (riskianalüüsi meetodid 1 ja 2 nii müügi- kui ka ostutabeli pealt) tagastatakse nende firmade registrikoodide nimekiri, kellel see konkreetne risk tuvastati. Samuti salvestatakse kõik sellised tehingupartnerite paarid, kus partneril on avastatud mingi risk, selleks eraldi ette nähtud tabelisse (vastavalt riskianalüüsi meetodile 3).

Iga etapi käivitamiseks on teostatud käsurea klientrakendus, mis suhtleb Sharemindi serveritega arvutivõrgu kaudu.

L1.3 Millist teavet analüüsi käigus avalikustatakse

Prototüübi ehitamisel on lähtutud sellest, et järgnevat teavet ei tohiks rakendus kellelegi avalikustada:

- firmade tehingute suurusi ja kõiki muid deklaratsioonis esinevaid rahalisi väärtused või nende väärtusi agregeeritud tulemusi,
- firmadevaheliste tehingute seostevõrku, st. mitte ühegi firma kohta ei tohi avalikustada, millised firmad on olnud tema tehingupartnerid.

Järgnevalt kirjeldatakse iga etapi kohta, millist informatsiooni on keerukuse vältimiseks või jõudluse parandamiseks prototüübis avalikustatud. See tähendab, et rakenduses ei avalikustata andmete kohta mitte mingit muud informatsiooni peale järgnevas kirjeldatu.

L1.3.1 Andmete üleslaadimine

Iga deklaratsiooni esitanud firma kohta avalikustatakse serverihoidjatele ja võimalikule süsteemi ründajale

- osturidade arv deklaratsioonis,
- müügiridade arv deklaratsioonis,
- tehingupartnerite arv ostu lisas,
- tehingupartnerite arv müügi lisas.
- Andmete esitamise etapi jooksul on ka serverihoidjatele ja võimalikule süsteemi ründajale teada selleks hetkeks deklaratsiooni esitanud firmade arv.

Deklaratsioonis oleva andmemahu järgi võib olla võimalik taustateabe abil kaudselt siduda deklaratsioon mingi konkreetse firmaga. See tähendab, et mõne firma kohta võib saada avalikuks, et ta on oma deklaratsiooni esitanud mingil konkreetsel ajal. Selle fakti salastamist ei ole aga võetud eesmärgiks, sest kõik Eesti firmad on kohustatud esitama iga kuu käibemaksu deklaratsiooni. Seega on selle teabe avalikustamine aktsepteeritav, sest üleslaetavad andmed ise avalikuks ei saa.

Samuti võib andmete üleslaadimisel võrguliikluse jälgimisega olla võimalik andmeid üleslaadiv firma identifitseerida IP-aadressi kaudu, kuid nagu öeldud, ei ole selle teabe varjamine prototüübi jaoks eesmärgiks seatud. Samuti ei ole hilisemates arvutustes võimalik teada saada, millise firma kohta arvutust tehakse, isegi kui üleslaadimisel on mõni firma identifitseeritud.

L1.3.2 Koondtabelite koostamine

Ostu ja müügi koondtabeli koostamise käigus avalikustatakse uuesti serverihoidjatele iga deklaratsiooni esitanud firma kohta tema müügi ja ostutehingute partnerite arv. Kui serverihoidja või süsteemi ründaja suudab tehingupartnerite arvu siduda mõne konkreetse firmaga, siis sarnaselt andmete üleslaadimise etapiga ei anna see talle teada muud, kui seda, et selle firma deklaratsioon on süsteemi üles laetud. Firma andmete või tehingupartnerite kohta mingisugust informatsiooni ei leki.

Kui andmete üleslaadimisel on firma juba identifitseeritud, saab siin andmete üleslaadimise järjestust teades leida ka selle firma tehingupartnerite arvu, kuid

tehingupartnerite arv on sellisel juhul juba tegelikult andmete üleslaadimisel teada. Seega kokkuvõttes koondtabelite koostamisel ei avalikustata informatsiooni, mida polnud avalikustatud andmete üleslaadimisel.

L1.3.3 Riskianalüüs

Riskianalüüsi käigus avalikustatakse serverihoidjatele ja võimalikule süsteemi ründajale järgmised väärtused.

- Kui palju on firmasid, kellel leiti müügikoondtabeli pealt vähemasti üks risk (1. ja/või 2.)?
- Kui palju on firmasid, kellel leiti ostukoondtabeli pealt vähemasti üks risk (1. ja/või 2.)?
- Kui suur oli tuvastatud riskiga firma unikaalsete tehingupartnerite arv?

Nagu eelnevates etappides, võib ka siin mõne firma puhul tehingupartnerite arv olla piisavalt unikaalne, et ainuüksi selle põhjal see firma identifitseerida. Kui sellisel firmal leitakse riskianalüüsi käigus mõni risk, võivad serverihoidjad ja süsteemi võimalik ründaja teada saada, et sellel firmal mõni risk leiti. Sellise teabe avalikuks tulemine võib alusetult kahjustada mõne firma mainet Eesti majanduskeskkonnas. Seega ei tohiks tegelik rakendus kindlasti sellist teavet avalikustada. Ka Sharemindi peal oleks võimalik riskianalüüs teha nii, et riskiga firmade tehingupartnerite arvu ei avalikustada, kuid hetkel nõuaks see andmebaasiteegi olulist täiendamist, mis jääb selle pilootprojekti võimalustest välja.

Riskianalüüsi tulemusena avalikustatakse riskianalüüsi algatanud klientrakendusele neli järjendit firmade registrikoodidega:

1) firmade registrikoodid, kellelt leiti müügikoondtabeli pealt risk 1,
2) firmade registrikoodid, kellelt leiti müügikoondtabeli pealt risk 2,
3) firmade registrikoodid, kellelt leiti ostukoondtabeli pealt risk 1,
4) firmade registrikoodid, kellelt leiti ostukoondtabeli pealt risk 2.

Need registrikoodid avalikustatakse ainult klientrakenduse käivitanud poolele (tüüpiliselt MTA), teistele mitte. Kui üks serverihoidjatest on ise riskianalüüsi käivitanud, siis temale analüüsi tulemus ka avaldatakse, teistele serverihoidjatele aga mitte.

L1.3.4 Jõudlustestid

Testisime esimese versiooni jõudlust Cybernetica Sharemindi klastri peal. Sharemindi klaster koosneb kolmest serverist, mida omavahel seob 1 Gbit/s läbilaskevõimega kohalik võrguühendus. Igal serveril on 48 GB RAM-mälu ning 12-tuumaline 3 GHz protsessor, mis toetab tehnoloogiat HyperThreading, seega kokku 24 paralleelset lõime. Jõudlusteste sooritasime seitsme testandmestikuga, mis on kirjeldatud tabelis 2.

Deklareerivaid firmasid	**Osturidu/ müügiridu kokku**	**Unikaalseid ostu-/müügipartnerite paare**
100	65 717/65 717	1089/1089
200	131 243/131 243	1990/1990
400	261 938/261 938	3990/3990
500	324 302/324 302	4990/4990
1000	649 242/649 242	9990/9990
2000	1 302 321/1 302 321	19 990/19 990
4000	2 602 486/2 602 486	39 990/39 990

Tabel 2: Kasutatud testandmestikud. Igal real on kirjeldatud üht testandmestikku.

Kõigi testandmestike struktuur on ühesugune. Igal firmal on müügiosas deklareeritud 500-800 müügirida (keskmiselt 650). Täpselt kümnel firmal on ostuosas deklareeritud umbes kümnendik kõigist osturidadest, teistel firmadel puudub ostuosa üldse. Kasutatud testandmestikud jäljendavad Eesti firmade andmetes esinevaid arveridade ja unikaalsete tehingupartnerite arvude suhteid nii, nagu need on MTA poolt Cyberneticale antud prognoosides kirjeldatud. Iga testandmestiku peal mõõtsime järgnevatele operatsioonidele kuluvat aega: 1) iga deklaratsiooni andmete üleslaadimine, 2) koondtabelite loomine, 3) riskianalüüsi tegemine.

Esimese versiooni jõudlustestide tulemused on tabelites 3 ja 4. Tabelis 3 on esitatud ühe deklaratsiooni üleslaadimise kiirus deklaratsioonide eri suuruste puhul. Kuna testandmetes olid deklaratsioonid küll suuruse järgi grupeeritud, kuid arveridade arv varieerus, on tabelis 3 näidatud iga deklaratsioonide grupi keskmist arveridade arvu. Tabelis 4 esitatud jõudlustulemused on ligikaudsed, sest igale operatsioonile kuluvat aega mõõtsime ainult üks kord. Kõige suurema testandmestiku peal ei õnnestunud agregeerimist lõpetada, kuid testi jooksul sooritatud agregeerimise mahu järgi sai hinnata kogu andmete agregeerimiseks kuluvat aega.

Jõudlustulemuste põhjal on näha, et prototüübi esimene versioon suudaks mõistliku ajaga teha arvutusi kuni 4000 firma andmetega perioodiliselt iga kuu. Riskianalüüs on agregeerimisega võrreldes kiire, kuid väga suurte andmemahtude

Arveridu deklaratsioonis kokku keskmiselt (osturead ja müügiread)	**Ühe deklaratsiooni üleslaadimise aeg**
660	0.44 s ± 1.68%
6600	3.91 s ± 0.89%
13 740	9.86 s ± 1.57%
26 830	25.48 s ± 0.91%
33 080	35.33 s ± 0.55%
65 590	109.61 s ± 0.48%
130 920	6 min 33 s ± 0.33%
260 900	13 min 57 s ± 0.36%

Tabel 3: Esimese versiooni andmete üleslaadimise kiirus ühe deklaratsiooni kohta. Tulemuste usaldusvahemik vastab usaldustasemele 95%.

Deklareerivaid firmasid	**Agregeerimine**	**Riskianalüüs**
100	9 min 15 s	6 s
200	32 min 59 s	14 s
400	2 h 7 min	40 s
500	3 h 11 min	55 s
1000	12 h 51 min	2 min 42 s
2000	50 h 44 min	9 min 6 s
4000	$>$ 120 h	—

Tabel 4: Esimese versiooni agregeerimise ja riskianalüüsi jõudlus

juures jääb tõenäoliselt liiga aeglaseks. Deklaratsioonide üleslaadimine on võrdlemisi kiire ka paljude arveridade korral. Sealjuures tuleb arvestada, et enamik ajast kulub andmete salvestamiseks andmebaasi, kui nad on juba kliendi poolt ühissalastatult Sharemindi serveritesse saadetud ja valideeritud. Seega andmete üleslaadija jaoks toimub see protsess tegelikult oluliselt kiiremini. Kui andmete salvestamisel juhtub mingi viga, võib sellest andmete üleslaadijat hiljem eraldi teavitada, ilma et ta peaks aktiivselt vastust ootama. Seega esimene versioon toetaks juba ka väga suurte deklaratsioonide (mis MTA hinnangul sisaldavad kuni 500 000 arverida) üles-laadimist mõistliku ajaga.

Lisa 2 Prototüübi teise versiooni kirjeldus

Prototüübi teises versioonis arendatakse edasi esimese valminud versiooni rakenduse skeemi ja põhimõtteid eesmärgiga oluliselt suurendada agregeerimise jõudlust, kasutades paralleelseid arvutusi. Firmade üleslaetud deklaratsioonide andmed jaotatakse mingi arvu ühesuguse struktuuriga andmebaaside vahel juhuslikult ning nende andmebaaside peal saavad agregeerimise protsessid töötada paralleelselt, iga protsess ühe andmebaasiga. Andmebaaside ja seega ka agregeerijate arvu saab hõlpsasti ümber konfigureerida, kuid mitte pärast seda, kui esimesed andmed on üles laetud. Lõpptulemus koondtabelitena on täpselt samasugune nagu esimeses versioonis ning ka riskianalüüsis ei ole muudatusi.

Järgnevalt kirjeldatakse prototüübi teise versiooni erinevusi võrreldes esimesega.

L2.1 Andmemudel

Firmade andmed jagatakse üleslaadimisel mitme ühesuguse struktuuriga andmebaasi vahel. Deklaratsiooni põhiosa andmed kirjutatakse kõik ühte andmebaasi, kuid ostu- ja müügiread jaotuvad andmebaaside vahel nii, et ühe firma tehingud on alati kõik ühes andmebaasis. Lisaks müügi- ja osturidadele salvestatakse igas andmebaasis ka unikaalsed müügi- ja ostupartnerite paarid, nagu ka esimeses versioonis, sest see lihtsustab agregeerimist. Lisaks salvestatakse iga firma andmete üleslaadimisel selle firma registrikood eraldi agregeerimisjärjekorra tabelisse. Kui firma andmed on ära agregeeritud, lisatakse registrikood teise vastavasse tabelisse. Neid tabeleid kasutatakse agregeerimisrakenduses, et kontrollida, kas on uusi andmeid, mis vajavad agregeerimist ja et samu andmeid kaks korda ei agregeeritaks.

Agregeerimine on uues versioonis jagatud kahte etappi: esmaseks agregeerimiseks, mis toimub paralleelselt mitme andmebaasi pealt ja võib toimuda ühel ajal andmete üleslaadimisega, ning lõppkoondtabelite koostamiseks, mis toimub pärast seda, kui kõikide firmade andmete peal on esmane agregeerimine tehtud. Lõppkoondtabelid ja prototüübi esimese versiooni koondtabelid on ühesuguse struktuuriga. Esmase agregeerimise tulemusena tekkivad algsed müügi- ja ostutabelid ei sisalda kõiki neid veerge, mis on lõppkoondtabelites, kuid neis on kõik vajalikud andmed, et lõppkoondtabelid ainuüksi nende põhjal koostada.

L2.2 Rakenduse põhiprotsesside kirjeldus

Järgnevas tähistab N agregeerimiseks kasutatavate andmebaaside arvu. Prototüübi teises versioonis jaguneb kogu rakenduse töö neljaks etapiks:

Andmete üleslaadimine — kõik firmad laadivad sarnaselt prototüübi esimese versiooniga oma deklaratsiooni andmed üles. XML-vormingus deklaratsiooni failist saadud väärtused salvestatakse ühissalastatud kujul ühte N andmebaasi seast. Iga firma saab oma deklaratsiooni ainult üks kord üles laadida. Prototüübi teine versioon veel paranduskandeid ega sama registrikoodiga firma andmeid

vastu ei võta. Nagu ka esimeses versioonis, võib mitut deklaratsiooni üheaegselt üles laadida.

Algsete koondtabelite koostamine — võib toimuda üheaegselt andmete üleslaadimisega ning paralleelselt kõigi N andmebaasi peal. Algsel agregeerimisel kontrollitakse, kas on üles laetud mõne uue firma andmeid, mis pole veel agregeeritud, ja need agregeeritakse. Agregeeritud firmade registrikoodid jäädvustatakse eraldi tabelis ning ühe firma andmeid kaks korda ei agregeerita.

Lõppkoondtabelite koostamine — toimub pärast seda, kui kõik firmad on oma andmed üles laadinud ja algsed koondtabelid on nende andmete pealt koostatud. Algsete koondtabelite põhjal tehakse vajalikud arvutused ning koostatakse lõppkoondtabelid.

Riskianalüüs — tehakse lõppkoondtabeleid kasutades täpselt samamoodi nagu eelmises versioonis.

Prototüübi teise versiooni demonstreerimiseks on ehitatud ka lihtne veebirakendus, mis liidestub Sharemindi serveritega ja võimaldab veebibrauseris kõiki eelnimetatud operatsioone käivitada ja tulemusi näha.

L2.3 Millist teavet avalikustatakse

L2.3.1 Andmete üleslaadimine

Serverihoidjatele ja süsteemi võimalikule ründajale avalikustatakse sama teave, mis eelmiseski versioonis. Lisaks avalikustatakse andmebaasi indeks, kuhu hetkel üleslaetava firma andmed salvestatakse. Indeks valitakse iga kord juhuslikult ja seega ei ole tal mingit olulist tähendust.

Samuti avalikustatakse see, kas mingi firma on oma deklaratsiooni juba esitanud või mitte. Kui on, siis uut deklaratsiooni vastu ei võeta.

L2.3.2 Algsete koondtabelite koostamine

Kehtib kõik sama, mis prototüübi esimeses versioonis koondtabelite koostamise osas. Lisaks avalikustatakse agregeerimise alguse seisuga see, mitme firma andmed on selleks hetkeks agregeeritavasse andmebaasi salvestatud ning mitmed nendest on juba agregeeritud. See teave ei ole kuidagi seotud deklaratsioone esitanud firmadega, vaid pigem rakenduse jõudlusega.

L2.3.3 Lõppkoondtabelite koostamine

Lõppkoondtabelite koostamisel segatakse enne arvutuse algust algsete koondtabelite read salajase järjestuse alusel, et eemaldada seosed tehingupartnerite identiteedi ja nende asukoha vahel koondtabelis. Lõppkoondtabelite koostamisel on see vajalik, sest siin seotakse algse müügitabeli rida vastava ostutabeli reaga. Teades andmete üleslaadimise järjestust ja vastavaid firmasid, oleks tabeliridade

segamata jätmise korral võimalik mõned tehingupartnerite paarid kindlaks teha, mis tähendaks, et mõne firma kohta saadaks teada, et tal on olnud tehinguid mingi teise konkreetse firmaga. Selle teabe kaitsmine on aga seatud prototüübi oluliseks eesmärgiks. Tabeliridade segamine kaotab ära seosed andmete üleslaadimise järjestuse ja nende tabelis esinemise asukoha vahel. Pärast seda ei ole enam kahe tabeli ridade sidumisel mingit sisulist tähendust.

Arvutuse ajal avalikustatakse selliste tehingupartnerite paaride arv, kes esinevad müügikoondtabelis ning kellele leidub ka vastav pöördpaar ostukoondtabelis, ehk kelle jaoks müüja deklaratsioonile vastav ostja on samuti mingis ulatuses oma oste deklareerinud selle partneriga. Ükski tehingupartnerite paar pole konkreetsete firmadega seostatav tänu tabeliridade segamisele. Samuti ei ole näha, millistel paaridel on ühiseid partnereid, kuna registrikoodid on ühissalastatud ja ka kaks ühesugust väärtust erinevad andmebaasides. Seega ei ole ühegi konkreetse tehingupartnerite paari kohta müügikoondtabelis võimalik öelda, kas selle paari pöördpaar leidub ostukoondtabelis või mitte. Avalikustatakse ainult selliste paaride arv, mille korral see on nii.

Samasugune informatsioon avalikustatakse ka ostukoondtabeli kohta, st. mitmel tehingupartnerite paaril ostukoondtabelis leidub vastav pöördpaar müügikoondtabelis.

Selle teabe avalikustamine aitab märgatavalt lihtsustada nii algsete kui ka lõplike koondtabelite koostamist ning ei avalikusta mitte midagi ühegi konkreetse firma äritegevuse kohta, seega aktsepteerime selle informatsiooni avalikustamist.

L2.3.4 Riskianalüüs

Kuna lõppkoondtabelid on identsed prototüübi esimeses versioonis koostatud koondtabelitega, toimub ka riskianalüüs siin täpselt samamoodi. Avalikustatakse sama teave, mis esimeses versioonis.

L2.4 Jõudlustestid

Teise versiooni jõudluse testimiseks kasutasime samu andmestikke, mis ka esimese versiooni puhul (tabel 5). Testid sooritati samas testimiskeskkonnas.

Teise versiooni puhul mõõtsime andmete üleslaadimise jõudlust mitme deklaratsiooni korraga laadimisel. Andmete üleslaadimise tulemused on tabelis 6. Need tulemused on võrreldavad esimese versiooni üleslaadimise jõudlusega, sest andmete üleslaadimise osas muutus ainult see, et andmeid salvestatakse erinevatesse andmebaasidesse juhuslikult. See aga jõudlust olulisel määral ei mõjutanud. Tulemused tähistavad seda aega, mis kulus esimese deklaratsiooni üleslaadimise algusest kuni viimase deklaratsiooni üleslaadimise lõpuni.

Agregeerimise osas mõõtsime eraldi algset paralleelset agregeerimist ning lõppkoondtabelite koostamist. Tulemused on esitatud tabelis 7. Agregaatorite arv näitab, mitme andmebaasi vahel andmed olid jaotatud ning mitu paralleelset

Deklareerivaid firmasid	**Osturidu/ müügiridu kokku**	**Unikaalseid ostu-/müügipartnerite paare**
100	65 717/65 717	1089/1089
200	131 243/131 243	1990/1990
400	261 938/261 938	3990/3990
500	324 302/324 302	4990/4990
1000	649 242/649 242	9990/9990
2000	1 302 321/1 302 321	19 990/19 990
4000	2 602 486/2 602 486	39 990/39 990

Tabel 5: Teise versiooni jõudluse testimiseks kasutatud testandmed. Igal real on kirjeldatud üht testandmestikku.

protsessi neid agregeeris. Lõppkoondtabelite koostamine toimub alati ühe protsessina. Meil ei õnnestunud paigalduse tehniliste puuduste tõttu kõigi andmemahtude juures kasutada ühepalju algoritmilist paralleelsust ning olime sunnitud suuremate andmemahtude puhul agregaatorite arvu vähendama. Teoreetiliselt toetab rakendus palju suuremat agregaatorite arvu, mis suurendaks jõudlust veelgi.

Arveridu deklaratsioonis kokku keskmiselt (osturead ja müügiread)	**Paralleelsete üleslaadimiste arv**	**Paralleelsed üleslaadimised kokku**
660	5	4.1 s
660	10	7.0 s
6600	5	8.4 s
6600	10	11.8 s
13 740	5	14.7 s
13 740	10	19.2 s
26 830	5	32 s
33 080	5	44 s
65 590	5	2 min 12 s
130 920	5	8 min 36 s

Tabel 6: Mitme deklaratsiooni korraga üleslaadimise jõudlus

Deklareerivate firmade arv	**Agregaatorite arv**	**Algne agregeerimine kokku**	**Lõppkoondtabelite koostamine**
100	5	47 s	16 s
100	10	43 s	16 s
200	5	92 s	36 s
200	10	64 s	36 s
400	5	6 min 45 s	104 s
500	5	9 min 58 s	2 min 31 s
1000	5	40 min 14 s	7 min 58 s
2000	5	1 h 56 min	25 min 5 s

Tabel 7: Algse paralleliseeritud agregeerimise ja lõppkoondtabelite koostamise jõudlus

Prototüübi teise versiooni jõudlustulemused näitavad juba Sharemindi võimet teha arvutusi mõistliku ajaga ka suuremate andmemahtude korral. Seda eriti siis, kui agregaatorite arvu oluliselt suurendada. Hinnanguliselt suudaks meie kasutatud testimiskeskkond riistvara poolest toetada ilma jõudlust piiramata kuni 30 paralleelset agregaatorit. Samuti oleks tehniliselt võimalik lõppkoondtabelite koostamist ja riskianalüüsi samuti paralleliseerida, mis suurte andmemahtude puhul suurendaks oluliselt nende jõudlust.

Deklaratsioonide paralleelne üleslaadimine annab võimaluse andmeid kiiremini süsteemi salvestada. Hetkel on aga raske hinnata, kuidas süsteem käituks deklaratsioonide esitamise tippajal, kui korraga esitatakse näiteks tuhandeid deklaratsioone.

Kui arvestada sellega, et MTA hinnangul esitab iga kuu deklaratsiooni 80 000 firmat kokku 50 000 000 arvereaga, tuleks selliste andmemahtude algseks agregeerimiseks prototüübi teise versiooniga kasutada 2 korda võimsamat riistvara kui meie testimiskeskkonnas. Kui kasutada 24-tuumalisi 3 GHz protsessoriga servereid ka 10 Gbit/s läbilaskevõimega võrguühendust, kuluks 80 000 firma andmete algseks agregeerimiseks 60 agregaatoriga hinnanguliselt nädal. Kuna algne agregeerimine töötab ka andmete üleslaadimisega üheaegselt, on võimalik oluline osa sellest tööst ära teha juba andmete üleslaadimisel, ning kokkuvõttes saaksid andmed mõned päevad varem agregeeritud.

Kuna hetkel ei ole lõppkoondtabelite koostamine ja riskianalüüs paralleelsed, ei ole nad sellise andmemahu juures mõistliku ajaga tehtavad. Teoreetiliselt oleks ka see paralleelsus võimalik, kuid vajab edasist tehnilist arendustööd.