Markus Blomvall

Automaattisten varmuuskopioiden toteuttaminen osana Spamrankings.net-projektia

Metropolia Ammattikorkeakoulu

Insinööri (AMK)

Tietotekniikka

Insinöörityö

30.4.2014

Metropolia

Tekijä(t) Otsikko Sivumäärä Aika	Markus Blomvall Automaattisten varmuuskopioiden toteuttaminen osana Spamrankings.net-projektia 27 sivua + 2 liitettä 30.4.2014
Tutkinto	Insinööri (AMK)
Koulutusohjelma	Tietotekniikka
Suuntautumisvaihtoehto	Sulautettutietotekniikka
Ohjaaja	Tutkintovastaava Anssi Ikonen

Tämän insinöörityön tarkoituksena oli toteuttaa Spamrankings.net – projektin käyttöön järjestelmä, jolla automaattisesti pystytään varmuuskopioimaan päivittäin sisään saapuva data sekä tarvittaessa suorittamaan myös koko käyttöjärjestelmän varmuuskopiointi. Työ toteutettiin keväällä 2013.

Spamrankings.net – projektin palvelimet toimivat Linux-käyttöjärjestelmällä, joten työ toteutettiin Bash-skriptausta käyttäen. Toteutus tehtiin pääosin Cron- ja Rsync-ohjelmien avulla niiden monipuolisuuden ja toimivuuden takia.

Alun perin suunnitellut tavoitteet saavutettiin, kuten oli haluttu, ja lopputuotos on osittaisessa käytössä. Työn aikana erinäisistä syistä johtuen tavoitteet kuitenkin muuttuivat useasti ja haluttu lopputulos muuttui isommaksi, kuin oli suunniteltu. Työstä löytyy siis potentiaalia jatkokehittämiselle.

Avainsanat	Linux, Roskaposti, Varmuuskopio

Author(s) Title Number of Pages Date	Markus Blomvall Automaattisten varmuuskopioiden toteuttaminen osana Spam-rankings.net-projektia 27 pages + 2 appendices 30 April 2014
Degree	Bachelor of Engineering
Degree Programme	Information Technology
Specialisation option	Embedded systems
Instructor	Anssi Ikonen, Head of degree programme

The purpose of this was to build up a system for Spamrankings.net – project that would automatically back up the daily incoming data and also could be used to back up the operating systems if needed. The work was done in the Spring of 2013.

The work was done by bash programming, because all the servers of Spamrankings.net – project work with Linux operating systems. It was mostly done by using the Cron- and Rsync-programs which are both very diverse and functional.

The original goals were achieved the way wanted. During the work the objectives changed a lot, because of various reasons. Mostly the objectives got bigger and wider range of things was supposed to be backed up. So potential for further development can be found from this project.

Keywords	Linux, Spam, Backup

Sisällys

Lyhenteet

ARPANET	Yhdysvaltain sotilaallista tutkimusta varten kehitetty tietoverkko, josta myöhemmin kehittyi Internet.
BASH	*Bourne again shell*; Tekstipohjainen tietokoneohjelma (komentotulkki), jonka avulla käytetään tietokoneen käyttöjärjestelmää.
IP-osoite	IP-osoite on yksilöllinen numerosarja, jonka avulla voidaan tunnistaa jokainen Internet-verkkoon liitetty tietokone.
CBL	*Composite Blocking List*; Internetin nimipalvelujärjestelmään pohjautuva mustalista, johon on kerättynä väärin käyttäytyviä IP-osoitteita.
PSBL	*Passive Spam Block List*; Roskapostitietokanta, johon on listattuna roskapostia lähettäneitä IP-osoitteita.
BOTTI	Itsenäisesti toimiva tietokoneohjelma, joka voidaan laittaa suoriutumaan halutulla tavalla haluttuna aikana.
RSYNC	Avoimenlähdekoodin tietokoneohjelma, jonka avulla voidaan synkronoida tiedostoja ja kansioita yhdestä kohteesta toiseen.
SUDO	*Superuser Do*; Komentoa tarkentava osa, joka antaa oikeuden suorittaa haluttu toiminto *sudoers*-tiedostossa määritetyn käyttäjän oikeuksin.
SSH	*Secure Shell*; Ohjelmisto, jolla voidaan muodostaa salattu etäyhteys yhdestä laitteesta toiseen.

1 Johdanto

Roskapostin määrä sähköpostiliikenteestä maailmalla oli vuonna 2012 noin 70 %. Määrä on pienessä laskussa monien yritysten siirtyessä mainostamaan tuotteitaan sähköpostiviestien sijaan eri sosiaalisiin medioihin, mutta tästä huolimatta roskaposti on dominoiva osa päivittäisestä sähköpostiliikenteestä. Jotta tietokoneista ja sähköpostin käytöstä saataisiin irti suurin mahdollinen hyöty jokapäiväisessä elämässä sekä vältyttäisiin erilaisilta huijausyrityksiltä, on tärkeää pyrkiä vähentämään roskapostin määrää. Roskapostin määrän ollessa melkein kolme neljäsosaa maailmalla liikkuvasta sähköpostista voidaan todeta, että määrää vähentämällä pystyttäisiin saavuttamaan suuria hyötyjä tietokoneiden ja palvelimien toiminnoissa varsinkin yritysmaailmassa. [1.]

Spamrankings.net on vuoden 2009 tienoilla alkanut projekti, jonka tarkoituksena on lisätä ihmisten tietoisuutta roskapostin määrästä sekä tätä kautta pyrkiä vähentämään roskapostin määrää maailmalla. Spamrankings.net tilastoi kuukausittain maailmalla lähetettävän roskapostin määrän sekä sitä lähettävät palvelimet ja esittelee nämä tilastot internetsivuillaan, jotta ihmiset voisivat olla valveutuneempia asian suhteen. Tilastot tietenkin sisältävät suuria määriä tärkeää dataa ja sen olemassa olon turvaaminen ja arkistoiminen on erittäin tärkeää projektin toimivuuden suhteen.

Insinöörityön tarkoituksena on esitellä Spamrankings.net-projektille tekemäni työ, jossa suunnittelin ja toteutin järjestelmän, joka automaattisesti tallentaa päivittäin tulevan datan sekä myös muun projektin palvelimilla olevan tärkeän datan. Palvelimet toimivat Linux-käyttöjärjestelmällä ja tuottamani järjestelmä onkin toteutettu hyödyntämällä Linuxin BASH-komentotulkkia ja sen monipuolisuutta sekä helppokäyttöisyyttä. Yksi syy tiedostojen ja datan varmuuskopiointiin oli myös edessä häämöttävä laitteistojen käyttöjärjestelmien päivitys.

Kerron eri työvaiheista, jotka projektini sisälsi, sekä esittelen ongelma kohtia, joita projektini aikana kohtasin. Tuon esille myös kehitysideoita, joilla tekemääni työtä voidaan laajentaa ja tehdä siitä entistä monipuolisempi. Lisäksi kerron roskapostista, sen historiasta, nykypäivästä ja vaikutuksesta maailmalla, sekä esittelen projektia, jonka osana tämän insinöörityöni toteutin.

2 Roskaposti

Yksi maailman ensimmäisistä roskapostin kaltaisista massaviesteistä lähetettiin 1978 silloisessa Arpanetissä. Tällöin yksi Arpanetin käyttäjistä lähetti yrityksensä uutta tuotetta mainostavan viestin 400:lle käyttäjälle tavoittaen suurimman osan Arpanetin käyttäjistä. Monet eivät tästä pitäneet ja palauteryöpyn aikaansaama viestiliikenne oli kaataa Arpanetin pienillä muistimäärillä toimivat palvelimet. Varsinkin yritysmaailmassa ongelma roskapostin kanssa on juuri sen aiheuttama ylimääräinen verkkoliikenne, joka vie resursseja niin palvelimilta kuin myös työntekijöiltä, mikä taas aiheuttaa suuria taloudellisia menetyksiä yrityksille. Lisäksi sen kautta leviävät haittaohjelmat ovat suuri riski tietoturvalle. [2.] [3.] [4.]

Roskapostiksi luokitellaan massapostitetut viestit, joiden lähettämiseen ei ole saatu vastaanottajalta lupaa, ja niitä pyritään nykyään levittämään niin viruksien kuin myös erilaisten haittaohjelmien avulla. Näitä markkinointiin ja huijausyrityksiin tarkoitettuja viestejä on pitkään yritetty kitkeä pois ja taistelu niiden lähettäjien ja torjujien välillä käy kiivaana. Tällä hetkellä noin 70 % maailmalla liikkuvasta sähköpostista on niin sanottua roskapostia, mutta muutamia vuosia sitten määrä oli jopa 90 %. Syitä vähenemiseen ovat roskapostittajien siirtyminen sosiaalisen median maailmaan, mutta myös sitä vastaan käytävän taistelun toimiminen. Vaikka roskapostin määrä on lyhyessä ajassa huomattavasti laskenut, on sen osuus sähköpostiliikenteessä tästä huolimatta kuitenkin edelleen valtava. [5.] [6.]

Roskapostia vastaan on kehitetty monia apukeinoja, kuten erilaisia suotimia, joilla pyritään estämään roskapostia tukkimasta käyttäjien saapuvien viestien kansioita. Näissä suotimissa hyödynnetään erilaisia avainsanoja, jotka havaittuaan sähköpostiohjelmat luokittelevat viestit roskapostiksi ja ne poistetaan tai siirretään tiettyyn kansioon. Näiden käytössä on kuitenkin riskinsä, kuten esimerkiksi se, että tavallisissa viesteissä saattaa olla kyseisiä avainsanoja, eivätkä ne täten saavutakaan vastaanottajaa. Nämä suotimet eivät myöskään niin paljoa edesauta roskapostiliikenteen vähentämisessä tai ehkäisemisessä vaan ennemminkin palvelevat loppukäyttäjää. [6.]

Nykyaikana roskaposti lähetetään pääosin bottiverkoista, jotka ovat jopa satojentuhansien bottien kaappaamien koneiden muodostamia verkostoja. Botit ovat haittaohjelmia, jotka ottavat käyttäjän koneen haltuun käyttäjän tietämättä asiasta. Päästyään koneelle botit voivat pysyä pitkäänkin piilossa, kunnes viimein käskyn saatuaan ne toimivat sen

mukaan. Roskapostin levityksen lisäksi niitä käytetään esimerkiksi palvelunestohyökkäyksiin sekä erilaisten tietojen, kuten luotto- ja pankkitietojen, varastamiseen. Bottiverkostoja vastaan on taisteltu viime vuosina kiivaasti ja sen myötä on saavutettu hyviä tuloksia. Tähän on monia syitä, kuten viranomaisten valveutuneisuus asioiden suhteen ja heidän panostuksensa saada rikolliset kuriin sekä tietoturvayhtiöiden tutkimus- ja kehitystyön onnistuminen. Lisäksi perinteisten roskapostisuotimien lisäksi on alettu kehittelemään menetelmiä, joilla voitaisiin estää roskapostiliikennettä lähtemästä liikkeelle alun alkaenkaan. Yksi näistä menetelmistä on Spamrankings.net -projekti, jossa roskapostiongelmaa vastaan on lähdetty siitä näkökulmasta, että puututaan ongelmaan jo ennen kuin se ehtii syntyä. Toisin sanoen pyritään alun alkaen estämään roskapostia lähtemästä palvelimilta taikka saastuneilta koneilta muille käyttäjille. [7.] [8.]

3 Spamrankings.net

3.1 Spamrankings.net-projekti

Austinin yliopistossa Teksasissa sijaitsee tutkimusyksikkö Center for Research in Electronic Commerce, jonka kanssa Metropolia Ammattikorkeakoulu on ollut yhteistyössä jo pari kymmentä vuotta. Vuosittain siellä on käynyt useita opiskelijoita suorittamassa työharjoitteluja erilaisten projektien parissa. Oma työtehtäväni Austinissa liittyi Spamrankings.net-projektiin.

Spamrankings.net-projekti alkoi noin neljä vuotta sitten ja sen tarkoituksena on taistella sähköpostitse liikkuvaa roskapostia vastaan. Spamrankings.net keskittyy tilastoimaan ja tuomaan julki maailmalla liikkuvan roskapostin määrän sekä yritykset ja organisaatiot, joiden servereiltä sitä levitetään. Lisäksi projektin tarkoituksena on tuoda julkisuuteen tietoisuutta siitä, että myös suurilla yrityksillä on ongelmia ja parannettavaa tietoturvallisuudessaan.

> Projektin johtajan, Professori Andrew B. Whinstonin, mukaan nämä tilastot edustavat roskapostia vastaan taistelevien tahojen tietokantoja ja antavat yrityksille maineeseen liittyviä houkutteita parantaa tietoturvallisuuksiaan.
>
> Hänen mukaansa tarkoituksena ei myöskään ole niin ikään taistella roskapostia vastaan, vaan tarkoituksena on tuoda esiin suurempia kysymyksiä, kuten pitäisikö yritysten julkisesti kertoa omista tietoturva ongelmistaan. [9.]

Spamrankings.net ei siis suoraan pyri estämään roskapostin levittämistä tai keksimään ratkaisua sitä vastaan, vaan projektin tarkoituksena on saada yritykset ja organisaatiot valveutuneemmiksi roskapostin suhteen. Yritykset ovat usein tietämättömiä omilla palvelimillaan olevista boteista, jotka käyttävät niiden palvelimia hyväksi ja levittävät roskapostia ja samalla myös hidastavat yritysten palvelimien toimintaa. Mikäli yritysten palvelin on saatu kaapattua roskapostin lähettämistä varten, se on usein myös merkki muista haavoittuvuuksista yrityksen tietoturvassa. Projektin avulla pyritään siihen, että yritykset ja organisaatiot olisivat paremmin perillä omien palvelimiensa vääränlaisesta toiminnasta ja osaisivat siten ajoissa puuttua asiaan. Projektilla pyritään myös luomaan tämänkaltaisista tietoturvallisuusongelmista avointa ja julkista keskustelua, joka edesauttaisi asioiden hoitamista ja kehittymistä parempaan ja turvallisempaan suuntaan. [9.]

3.2 Spamrankings.net-projektin julkaisemat tilastot

Roskapostia liikkuu maailmalla paljon ja sitä on myös paljon erilaista. Osa siitä on mainoksia, osa huijausviestejä, osa sisältää viiruksia ja niin edelleen. Myös tapoja tilastoida roskapostia löytyy useita erilaisia. Yritysmaailmassa maine on yksi isoimpia ja tärkeimpiä asioita, ja tätä faktaa hyväksi käyttäen myös Spamrankings.net pyrkii tilastonsa kokoamaan. Spamrankings.net sivustolla yrityksiä ja organisaatioita listataan järjestykseen maakohtaisesti kahteen eri listaan niiden palvelimilta lähtevän CBL- ja PSBL-roskapostidatan määrien mukaan, joita käsitellään tarkemmin luvussa 3.2. Mitä ylemmäksi listalla yritys sijoittuu, sitä enemmän sen palvelimilta lähtee roskapostia ja sitä huonomalta se näyttää yrityksen julkisuuskuvassa. Toivon mukaan yritykset, jotka listalle sijoittuvat, pyrkivät parantamaan omaa tietoturvallisuuttaan ja täten vähentämään palvelintensa ja koneidensa sisältämiä botteja ja parantamaan näin ollen yrityksensä turvallisuutta. Yritykset, jotka eivät sijoitu listalle ollenkaan tai pääsevät sieltä pois, voivat käyttää tätä tietoa hyväkseen omassa markkinoinnissaan ja julkisuuskuvansa parantamisessa. [9.]

Spamrankings.net sivustolla on erilaisia listoja, joissa on kerättynä kuukausittain eniten roskapostia lähettäneet yritykset sekä lääketieteelliset organisaatiot. Näiden lisäksi kuukausittain listataan kaksikymmentä eniten viestejä lähettänyttä valtiota. Listauksien lisäksi esillä on muutama kaavio, joista näkee tarkemmin, miten viestit jakautuvat kullekin päivälle ja miten suuren osan kukin organisaatio viestittää kokonaismääristä. [9.]

July 2013 Monthly Countries ∀ All Sp@mRankings.net from CBL Volume
(Last Month)

Most countries stayed the same or shuffled around a few places, but Belarus got markedly better while Sweden got a lot worse.

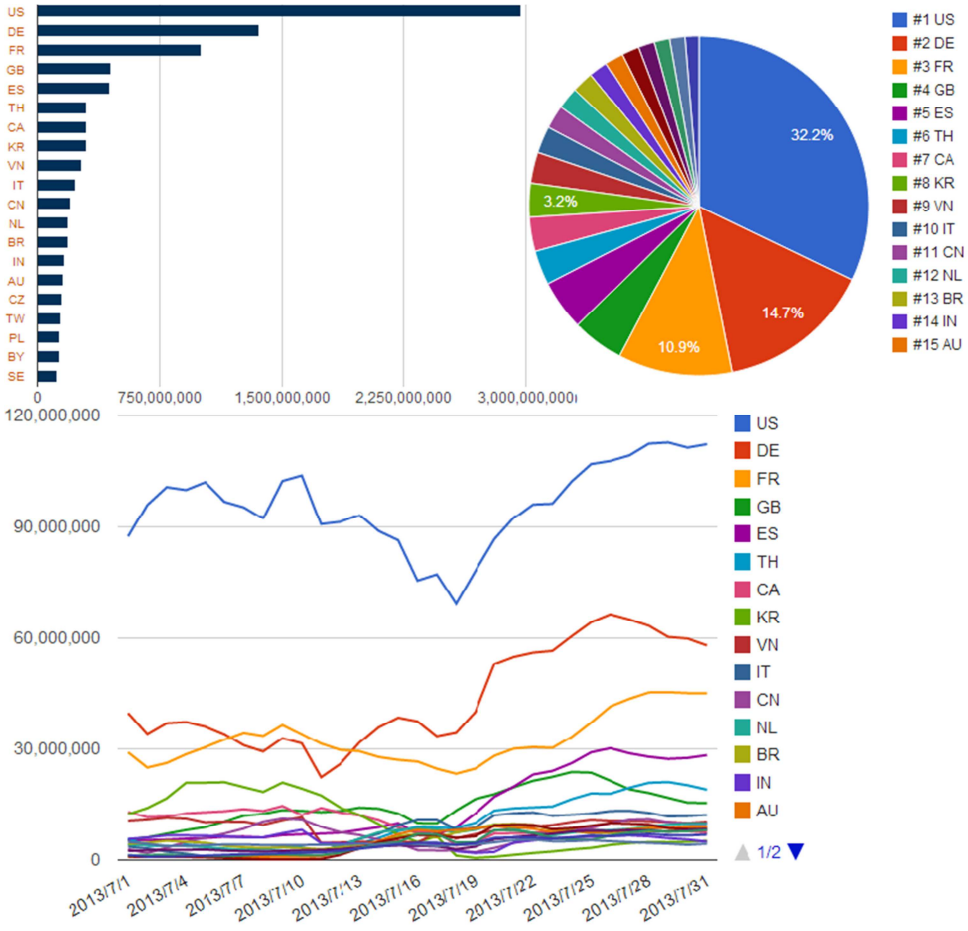
Rank This	(Last)	Country	Population	Volume This (Last)	Vol%
			Population		
1	(1)	US	310,232,863	2,971,311,474 (2,035,306,731)	32.2%
2	(2)	DE	81,802,257	1,358,941,434 (805,186,851)	14.7%
3	(3)	FR	64,768,389	1,006,500,734 (608,900,662)	10.9%
4	(5)	GB	62,348,447	448,600,726 (307,685,126)	4.86%
5	(8)	ES	46,505,963	443,413,116 (201,969,016)	4.8%
6	(10)	TH	67,089,500	304,063,855 (130,561,047)	3.29%
7	(6)	CA	33,679,000	302,445,410 (217,521,440)	3.27%
8	(4)	KR	48,422,644	297,610,213 (441,338,953)	3.22%
9	(7)	VN	89,571,130	271,824,864 (213,386,247)	2.94%
10	(16)	IT	60,340,328	235,651,458 (99,618,443)	2.55%
11	(20)	CN	1,330,044,000	206,469,983 (73,857,469)	2.23%
12	(12)	NL	16,645,000	191,719,952 (125,280,853)	2.07%
13	(14)	BR	201,103,330	188,364,678 (121,004,858)	2.04%
14	(11)	IN	1,173,108,018	163,722,261 (126,097,104)	1.77%
15	(35)	AU	21,515,754	158,560,650 (32,072,731)	1.72%
16	(52)	CZ	10,476,000	153,516,249 (11,701,618)	1.66%
17	(21)	TW	22,894,384	143,461,605 (73,123,755)	1.55%
18	(17)	PL	38,500,000	135,780,814 (87,998,943)	1.47%
19	(9)	BY	9,685,000	135,438,237 (174,801,466)	1.47%
20	(44)	SE	9,045,000	122,466,272 (16,997,756)	1.33%
		Total		9,239,863,985	100%
		In	Previous	Ranking	
27	(13)	SA	25,731,776	89,311,746 (123,298,044)	
25	(15)	UA	45,415,596	91,393,406 (100,018,315)	
24	(18)	CO	44,205,293	92,574,853 (82,766,817)	
26	(19)	RS	7,344,847	89,416,951 (78,439,594)	

Kuva 1. Heinäkuussa 2013 tilastoitu maakohtainen CBL - roskapostidatan määrä. [3.]

Kuvassa 1 on esitelty heinäkuun 2013 maakohtainen CBL–roskapostidatan määrä. Siinä on ilmoitettu kyseisen kuukauden sijoitus, edellisen kuukauden sijoitus, maatunnus, kansalaisten määrä, kyseisen kuukauden viestien määrä sekä suluissa edellisen kuukauden määrä ja lopuksi prosentuaalinen osuus kahdenkymmenen eniten viestejä lähettäneen maan yhteistuloksesta. Näiden alla ovat maat, jotka ovat edellisellä kuu-

kaudella vielä olleet listauksessa mukana, mutta ovat päässeet pois joko onnistuttuaan vähentämään viestien määrää tai, kuten harmillisen usein käy, sen seurauksena että jokin toinen valtio kasvattaa omaa roskapostidatan määräänsä vähentänyttä valtiota suuremmaksi. Väreistä, joilla tulokset on ilmoitettu, voidaan nähdä, onko luku kasvanut vai vähentynyt. Kellertävällä värillä ilmoitetaan kasvaneesta luvusta ja sinertävällä vähentyneestä.

Kuten kuvasta 1 nähdään, tuotti USA heinäkuussa 2013 maailmalla liikkuvasta roskapostista melkein kolmasosan, ja yksinomaan melkein kolme miljardia roskapostiviestiä, kun vertaillaan kahtakymmentä eniten roskapostia tuottanutta valtiota. Tästä saadaan päivittäiseksi keskiarvoksi melkein 300 miljoonaa sähköpostia per päivä pelkästään yhdysvaltalaisilta palvelimilta. Kuvasta 1 voidaan myös nähdä, että heinäkuussa 2013 Saudi-Arabia, Ukraina, Kolumbia ja Venäjä pystyivät pitämään oman roskapostinsa sen verran alhaisena, että ne pääsivät pois kahdenkymmenen eniten roskapostia lähettävän valtion listauksesta. [9.]



Kuva 2. Heinäkuun 2013 maakohtaisen CBL-roskapostidatan määriä kuvaavia kaavioita. [3.]

Kuvassa 2 ylempänä esitellyistä kaavioista nähdään kahdella eri ilmaisutavalla 20 eniten roskapostia lähettäneen valtion roskapostimäärien suhteita niin ympyrä- kuin myös pylväsdiagrammin muodossa. Alempana kuvassa on esiteltynä kaavio, josta voidaan nähdä roskapostien määrät päiväkohtaisesti kustakin 20 valtiosta.

3.3 Tilastoitava data

Tällä hetkellä Spamrankings.net-sivustolla näytetään julki CBL- ja PSBL-tietokantoihin perustuvien roskapostien määrät. Näiden lisäksi Spamrankings.net kerää dataa myös kuudesta muusta tietokannasta, mutta näitä tuloksia ei vielä julkisuuteen ole päätetty levittää. CBL- ja PSBL-tietokannat keräävät molemmat tietonsa niin sanottujen hunajapurkkien avulla.

Hunajapurkit, jotka usein esiintyvät nimellä Honey Pot, ovat ansoja, joiden avulla pyritään saamaan kiinni roskapostittajien käyttämiä harvester-botteja. Harvester-botit on sovelluksia, jotka on tehty lukemaan läpi websivujen sisältöä ja etsimään sieltä merkkijonoja, jotka sisältävät @-merkin. Nykyään on olemassa monia keinoja, joilla puolustautua näitä botteja vastaan, kuten käyttämällä @-merkin sijasta esimerkiksi (at)-merkkiparia. Käytössä on myös monia muita tapoja, joilla sähköposti voidaan merkitä ihmisille selkeäksi, mutta boteille vaikeasti luettavaksi. Näiden harvester-bottien käyttö on lainopillisesti kaksijakoista, koska toisissa maissa käyttö katsotaan rikolliseksi ja toisissa taas ei. [10.] [11.] [12.]

Hunajapurkit ovat verkkosivujen koodiin piilotettuja sähköpostiosoitteita, jotka eivät näy tavallisille käyttäjille. Näihin osoitteisiin tulevat viestit voidaan välittömästi mieltää roskapostiksi, sillä nämä sähköpostiosoitteet näkyvät vain harvester-boteille ja näin ollen ne päätyvät vain henkilöille, jotka käyttävät niitä roskapostin levitykseen. Hunajapurkkeja on pystytty kehittämään niin paljon, että nykyään kun harvester-botti käy keräämässä ansalla viritetyn sähköpostiosoitteen, jää tästä käynnistä ylös kellonaika sekä IP-osoite, josta käynti on suoritettu. Kun tähän ansalla viritettyyn sähköpostiosoitteeseen tulee roskapostia, voidaan tunnistaa, milloin ja kuka on osoitteen haltuunsa saanut. Tämän jälkeen tämä IP-osoite voidaan asettaa niin sanotulle mustalle listalle ja rajoittaa sen pääsyä verkkosivuille, rajoittaa siitä lähtevää sähköpostia sekä pyrkiä saamaan roskapostittajia edesvastuuseen. Vuonna 2004 käynnistetyn Project Honey

Pot -hankkeen myötä voivat myös tavalliset kuluttajat ladata omille internetsivuilleen hunajapurkkeja ja näin ollen auttaa roskapostin pysäyttämisessä. [13.] [14.]

Composite Blocking List

Composite Blocking List (jäljempänä ”CBL”) on toinen Spamrankings.netin tilastojen päälähteistä. CBL on Internetin nimipalvelujärjestelmään pohjautuva mustalista, englanniksi Domain Name Server Blockin List, joka listaa suurien roskapostiansojen kautta kerättyjä IP-osoitteita sekä IP-osoitteita, joiden tiedetään kuuluvan bottiverkkojen alaisuuteen. Sen listoilta löytyy arvioiden mukaan noin viisi miljoonaa IP-osoitetta. Esimerkiksi IPv4-osoitteiden määrä maailmassa on noin neljä miljardia. CBL-listaa voidaan käyttää pisteytykseen perustuvana mustana listana tai välittömään estoon perustuvana. Esimerkiksi sähköpostipalvelin voi käyttää CBL-tietokantaa hyödykseen käsitellessään omaa sähköpostivirtaansa. Jos jokin IP-osoite, johon on lähdössä tai joka on lähettämässä postia, löytyy CBL-listalta, voi sähköpostipalvelin estää sähköpostin lähetyksen suoralta kädeltä tai käyttää osoitteen saamaa pisteytystä arvioidessaan, voiko sähköpostin välittää eteenpäin. Sähköpostiosoitteen pisteytykseen vaikuttavat monet asiat, kuten esimerkiksi minkälaisia sanoja sen lähettämät viestit sisältävät. CBL:n ylläpitäjät itse suosittelevat CBL-tietokantaa käytettäessä käytettävän välitöntä estoa. [15.] [16.] [31.]

CBL-tietokantaan listatut IP-osoitteet eivät pääse sinne sattumalta, sillä ylläpitäjillä on enemmän syitä joiden takia osoite ei sinne päätyisi, kuin syitä joiden takia se sinne päätyisi. Tietokantaan ei myöskään listata tunnettujen roskapostittajien IP-osoitteita. CBL-tietokantaan päätyvät ainoastaan IP-osoitteet, joiden tiedetään olevan saastuneita tahtomattaan, kuten koneet, jotka ovat joutuneet bottiverkon alaisuuteen, tai viruksien tai vastaavien avulla kaapatut koneet. Aivan tarkkoja tietoja CBL:n toimintatavoista ja toiminnoista ei ole kuitenkaan julkisuuteen paljastettu, sillä mikäli roskapostittajat saisivat haltuunsa tärkeitä tietoja CBL:n toiminnasta, voisivat he käyttää näitä tietoja hyödykseen ja koettaa vahingoittaa CBL:n toimintaa. Jos oman IP-osoitteensa on CBL:n listoille saanut, ei sen pois saamiseksi tarvitse nähdä suurta vaivaa. Puhdistettuaan oman koneensa saastumisilta voi henkilö lähettää ylläpidolle tukipyynnön, että se tarkastaisivat käyttäjän koneen ja hän saisi osoitteensa listalta pois, ja mikäli osoite ei ole enää saastunut, se poistetaan. [15.] [16.] [31.]

Passive Spam Block List

Spamrankings.netin toinen tietolähde, Passive Spam Block List (jäljempänä ”PSBL”) on helppo ja yksinkertainen roskapostitietokanta. Se listaa tietokantaansa kaikki IP-osoitteet, joista on johonkin sen hunajapurkkiin tullut sähköpostiviesti. Listalta pääsee pois tietyn ajan kuluessa tai, mikäli käyttäjä on huomannut listalle joutuneensa, voi hän poistaa IP-osoitteensa sieltä myös itse. Tällä tavoin pyritään varmistamaan, että jos käyttäjän IP-osoitteesta on tullut vain kertaluontoisesti roskapostia, ei häntä rangaista siitä pitkällä estolla. PSBL ylläpitää myös sisäistä niin sanottua valkoista listaa, johon on listattuna sähköpostipalvelimia, jotka eivät missään tapauksessa joudu mustalle listalle. Monet roskapostia lähettävät ovat kuitenkin botteja, jotka toimivat automaattisesti, ja niiden alaisuudesta tulee tuhansia viestejä. Näin ollen nämä IP-osoitteet viettävät listalla aikaa pitkään ellei botteja havaita ja niiden toimintaa estetä. [17.]

3.4 Tulevaisuus

Tällä hetkellä Spamrankings.net on vielä niin sanotussa rakentamis- ja tutkimusvaiheessa. Tulevaisuudessa projektilla on kuitenkin tarkoitus laajentua ja levitä suuren yleisön tietoisuuteen. Suunnitteilla on monia uusia ominaisuuksia sekä sivuston uudelleen rakentaminen. Tällä hetkellä työn alla on järjestelmä, jolla automaattisesti saadaan yrityksille tieto, mikäli ne on rankattu sivuston listoille. Sen avulla ne voivat nopeasti tulla tietoisiksi asiasta ja pystyvät tekemään sen eteen välittömästi jotain. Jotta järjestelmä saadaan toimimaan kunnolla, joudutaan keräämään erittäin suurta tietovarastoa yrityksistä ja niiden yhteystiedoista. Projektille on myös lanseerattu oma blogi sekä on ollut puhetta projektista ja tilastoista kertovan tiedon levittämisestä sosiaalisessa mediassa. Tulevaisuudessa on lisäksi tarkoitus laajentaa sivuston listauksia käsittämään muistakin tietokannoista saatuja tietoja.

4 Varmuuskopioiden toteuttaminen

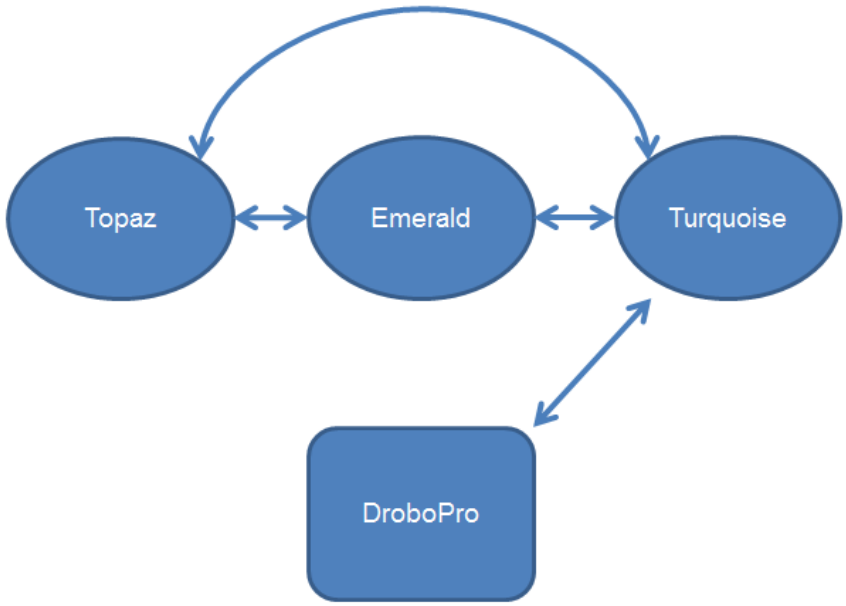
4.1 Prosessi

Projektin tavoitteena oli tutkia erilaisia tapoja toteuttaa luotettava ja toimiva varmuuskopiointijärjestelmä, jolla turvattaisiin jo olemassa olevan datan olemassa olo sekä päivittäin tulevan datan säilyvyys, sekä rakentaa ja ottaa käyttöön parhaaksi katsottu järjestelmä. Työ aloitettiin kartoittamalla lähtötilanne, eli mitä tiedostoja on jo tallennettu ja mitä tiedostoja ei sekä mihin mikäkin tieto on tallennettu ja kuinka moneen kertaan. Koska aikaisemmin varmuuskopioiden hoitamista ei ollut nimetty kenellekään, oli varmuuskopiointi jäänyt monilta osin puutteelliseksi, ja muutaman vuoden ajalta dataa jopa puuttui kokonaan. Edellä mainitun lisäksi työssä oli tavoitteena tutkia erilaisia vaihtoehtoja, joilla varmuuskopioidut tiedostot pystyttäisiin järjestämään ja arkistoimaan selkeästi sekä tarvittavan useasti ja tehokkaasti. Tavoitteena oli myös kehittää koodit, joilla päivittäinen varmuuskopiointi saataisiin suoriutumaan automaattisesti.

Työn varrella eteen tuli useita esteitä, ongelmia ja sekaannuksia. Työn tavoitteet ja ohjeet muuttuivat useita kertoja projektin aikana, monesti sen jälkeen, kun jotain saatiin valmiiksi ja huomattiin, että toisella tavalla se voidaan suorittaa järkevämmin ja tehokkaammin tai, että jotain oli järkevä lisätä siihen. Projektin aikana vaivasivat monesti myös erilaiset laitteisto-ongelmat. Useaan otteeseen toiminnan kannalta elintärkeät verkkoyhteydet toimivat hitaasti tai olivat katkenneet kokonaan. Tähän suurin syy oli paikka, johon palvelimet oli fyysisesti asennettu. Palvelimet oli laitettu tutkimuskeskuksen nurkkaan sermin taakse isoon kasaan muiden projektien palvelimien sekaan. Ongelma saatiin lopulta ratkaistua sijoittamalla laitteet uudestaan loogisempaan järjestykseen. Eri projektien palvelimet järjestettiin omiin riveihinsä ja johdotukset laitettiin kulkemaan piilossa ja poissa työntekijöiden jaloista. Myös vaihtamalla kuluneet ja rikkinäiset johdotukset uusiin saatiin palvelimet toimimaan aikaisempaa paremmin ja luotettavammin. Lisäksi kiintolevypalvelin, jolle varmuuskopiot tallennettiin, hajosi projektin aikana muutamaan otteeseen ja tutkimustyö uuden laitteen hankkimisesta toi mukanaan uusia haasteita.

4.2 Laitteisto

Alun perin Spamrankings.net-projektin palvelimet koostuivat kolmesta tietokoneesta, nimeltään Topaz, Emerald ja Turquoise, sekä varmuuskopioon tarkoitetusta ulkoisesta tallennuslaitteesta, DroboPro:sta. Näistä kolmesta tietokoneesta kaksi, Topaz ja Emerald, on tarkoitettu datan käsittelyyn ja Turquoise toimi linkkinä ulkoisen tallennuslaitteen kanssa, kuten kuvasta 3 nähdään.



Kuva 3. Alkuperäiset laitteet ja niiden topologia.

Yksi tietokoneista kerää päivittäin datan eri roskapostitietokannoista, prosessoi sen ja siirtää prosessoidun tiedon toiselle tietokoneelle. Päivittäin sisään tuleva data tulee niin sanottuna raakadatana ja se pitää prosessoida yksinkertaisempaan muotoon, jotta sen julkaiseminen on helpompaa ja järkevämpää sekä palvelimien suorituskyvylle kevyempää. Prosessoinnin jälkeen data tilastoidaan suljettuun websivustoon, josta joka kuukauden jälkeen se sitten käydään läpi lisäten tilastojen ja kaavioiden yhteyteen kommentteja ja tulkintoja. Edellisen kuukauden tilastoitu data pyritään aina saamaan julkisuuteen heti seuraavan kuukauden ensimmäisellä viikolla.

Päivittäin tulevaa dataa saapuu useita gigatavuja. Ja koska tietokoneissa oleva kovalevytila on rajallista, myös sille tallennettavan datan määrä on rajallista. Tietokoneisiin

onkin koodattu niin, että uuden datan saapuessa vanhimmat datatiedostot poistuvat automaattisesti uudempien tieltä. Tietokoneilta löytyy dataa suurin piirtein viimeisen vuoden ajalta, kun taas dataa on kerätty jo noin kolmen vuoden ajan.

Aloittaessani työt olivat kaikki tietokoneet noin kolme, neljä vuotta vanhoja ja niiden kaikkien käyttöjärjestelmänä oli Debian Linuxin Lenny-versio, joka oli jäänyt jo kaksi versiota vanhaksi. Ulkoinen tallennuslaite, DroboPro, oli vajaa kolme vuotta vanha ja sen takuuaika oli raukeamassa laitteen saavuttaessa kolmen vuoden iän. DroboPro sisälsi aloittaessani noin vuoden vanhat varmuuskopiot tietokoneista sekä roskaposti-datan viimeisen kolmen vuoden ajalta. Suuret määrät datasta oli kuitenkin useaan kertaan tallennettu sekä sekavasti arkistoitu.

Työni loppuvaiheilla DroboPro:n virtalähde hajosi ja jouduin takuuhuollon kautta hankkimaan kokonaan uuden laitteen. DroboPro on rakennettu niin, että kaikki komponentit on integroitu siihen eikä käyttäjä pysty vaihtamaan siihen itse mitään osia kiintolevyjä lukuun ottamatta. Ei ainakaan takuun puitteissa. Takuulaitteen kanssa tuli eteen myös teknisiä ongelmia, joten sain toisen takuulaitteen, mutta sekään ei toiminut, kuten oli tarkoitus. Sen kanssa tuli eteen ongelmia saada se pysymään verkossa siirrettäessä suuria määriä dataa, mikä vaikeutti tiedostojen siirtoa. Syitä outoon käyttäytymiseen yritettiin etsiä kuitenkaan siinä onnistumatta. DroboPro:n toimiessa ainoana varmuuskopiopäätteenä ja huomattuani sen luotettavuuden kärsineen tulin siihen tulokseen, että on hankittava uusi luotettavampi tallennuslaite, johon saadaan turvallisesti varmuuskopiot sijoitettua. Markkinoilla on monia eri laitteita, jotka olisivat sopineet Spamrankings.net-projektin käyttötarkoituksiin, joten vaati selvittelytyötä, jotta löysin parhaiten juuri tähän projektiin soveltuvan laitteen.

Verkkoon liitetty levypalvelin

Etsiessäni uutta laitetta varmuuskopiopäätteeksi oli pääpainona sen toimiminen verkon yli. Tällöin siihen käsiksi pääsyyn ei vaikuta ainoastaan yhden tietokoneen toiminta vaan sen toiminta riippuisi verkon toiminnasta. Tällä saadaan helpotettua myös yhdelle koneelle muodostuvaa tietoliikenteen kuormitusta ja tiedostojen käsittely muilta koneilta on helpompaa ja tehokkaampaa.

Network Attached Storage-palvelimet (jäljempänä ”NAS”) ovat juuri tähän tarkoitukseen tehtyjä tallennusjärjestelmiä. Vapaasti suomennettuna NAS tarkoittaa tietoverkkoon

liitettyä tallennusmuistia. NAS-palvelimien hienous on siinä, että ne eivät tarvitse erillistä näyttöpäätettä, hiirtä tai näppäimistöä toimiakseen, vaan kun ne on liitetty ethernet-johdolla verkkoon, pääsee niitä IP-osoitteen kautta käyttämään oman koneensa web-selaimella. Tätä kautta pystytään kaikki asetukset ynnä muut tarvittavat toimenpiteet suorittamaan miltä tahansa tietokoneelta, kunhan tarvittava IP-osoite sekä salasanat ovat tiedossa. [18.] [19.]

NAS-palvelimilla ei myöskään mene levytilaa hukkaan niiden sisällä pyörivän käyttöjärjestelmään vuoksi vaan ne ovat varustettuja yksinkertaisella sulautetulla käyttöjärjestelmällä, jonka avulla ne saadaan toimimaan juuri niiden suunnitellun käyttötarkoituksen mukaan eli tiedostojen kopiointiin ja tallennukseen. NAS-palvelimet ovat myös luotettavia ja toimintavarmoja. Esimerkiksi niiden ollessa kytkettynä verkkoon, eikä johonkin tietokoneeseen, vältytään ongelmilta, jos tietokoneisiin tulee käyttöjärjestelmän kaatumisia, hitautta tai muuta haittaa. Niiden suurimpina etuina perinteisempiin tiedostojen jako menetelmiin on niiden hinta, turvallisuus, nopeus ja helpompi hallittavuus ja ylläpito. [18.] [19.]

Nykyään NAS-palvelimia on markkinoilla suuria määriä. Niitä on erikokoisia, eri tehoisia, eri käyttötarkoituksiin, eri hintaluokista ja monilta eri valmistajilta. Jotta löysin juuri Spamrankings.net-projektille parhaiten soveltuvan laitteen, oli ensin kartoitettava ominaisuuksia, joita tulevalta laitteelta kaivattiin. Koska laite oli tulossa vain ja ainoastaan varmuuskopioiden tallentamiseen, haluttiin ensisijaisesti laite, jonka kapasiteetin laajennusmahdollisuudet ovat mahdollisimman suuret, sillä sisään saapuvan datan määrä on erittäin suuri. Lisäksi laitteen tuli olla nopea, toimintavarma, hinnaltaan tutkimusprojektin budjettiin sopiva sekä mielellään helposti ylläpidettävä. Loppuen lopuksi päädyin vertailemaan Qnap TS-869 Pro-, Synology Diskstation DS1812+- ja Drobo B800fs-malleja. Ne ovat suunnilleen samanhintaisia ja jokaiseen näistä kolmesta voidaan asentaa kahdeksan kiintolevyä, jolloin tallennuskapasiteetista saadaan mahdollisimman suuri.

Taulukko 1 - Taulukko verkkoon liitettyjen levypalvelimien ominaisuuksista. [20.] [21.] [22.]

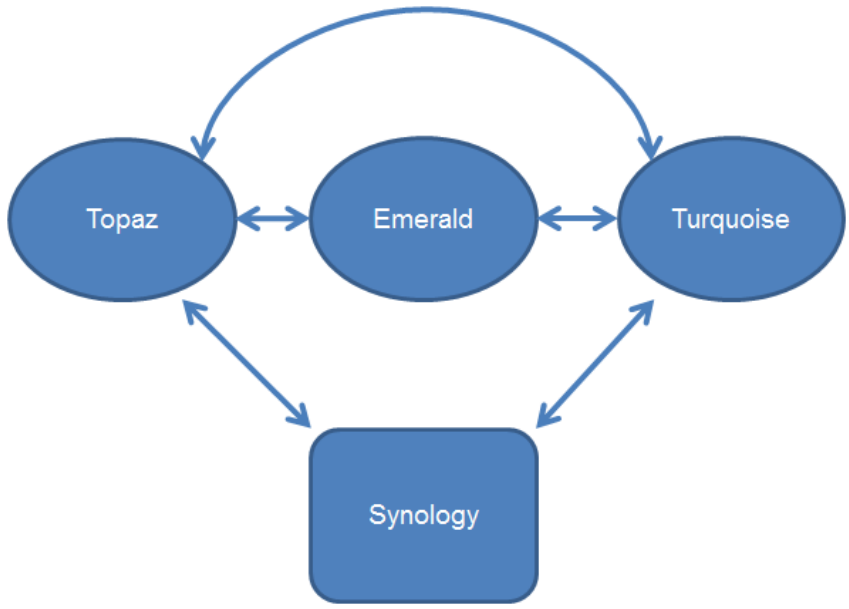
	Drobo B800fs	Synology DS1812+	Qnap TS-869 Pro
CPU		Intel D2700 2.13GHz Dual Core	Intel® Atom™ 2.13GHz Dual-core
Memory		DDR3 1 GB (Expandable up to 3 GB)	1GB (Expandable up to 3GB)
Drive Bays	8	8 (18 with expansion unit)	8
Internal HDD/SSD	3.5" SATA I 3.5" SATA II 3.5" SATA III	3.5" SATA HDD 2.5" SATA HDD 2.5" SATA SSD	3.5" SATA 6Gb/s + 3Gb/s + SSD 2.5" SATA 6Gb/s + 3Gb/s 2.5" SATA SSD
Maximum Internal Storage	25TB	32TB	32TB
Networking	2 x Gigabit LAN	2 x Gigabit LAN	2 x Gigabit LAN
External ports	——	2 x USB 3.0 Port 4 x USB 2.0 Port 2 x eSATA Port	2 x USB 3.0 Port 5 x USB 2.0 Port 2 x eSATA Port
Supported RAID Type	BeyondRAID	Synology Hybrid RAID JBOD RAID 0 RAID 1 RAID 5 RAID 6 RAID 10	JBOD RAID 0 RAID 1 RAID 5, 5 + Hot Spare RAID 6, 6 + Hot Spare RAID 10, 10 + Hot Spare
Size (HxWxD)	139 x 309 x 358 mm	157 x 340 x 233 mm	185 x 298 x 235 mm
Weight	7,34 Kg	5,21 Kg	7,3 Kg
Noise Level	30.4 dB	23.1 dB	27.5 dB
Power Consumption	82W (Operating) 13W (Idle)	71W (Operating) 28W (Idle)	59W (Operating) 30W (Idle)
Warranty	1 Year in the US	3 Years	2 Years

Tein taulukon (Taulukko 1) kuvaamaan kaikkien kolmen laitteen niitä ominaisuuksia, jotka olivat tärkeitä Spamrankings.net-projektin kannalta, jotta laitteiden eroista olisi helpompi muodostaa käsitys. Taulukon pohjalta kävi nopeasti ilmi, että Drobo B800fs ei millään tavalla yllä samalle tasolle kuin Synologyn tai Qnapin vastaavat laitteet ja pystyin sulkemaan sen pois lopullisesta valinnasta. Suurimpina syinä Drobon poissulkemiseen oli sen soveltuvuus RAID-protokollien kanssa ja takuun kesto sekä sen sisäisen muistinkapasiteetti on pienempi kuin Synologyssa tai Qnapissa.. Myöskään tietoja sen sisältämästä suorittimesta tai muistinmäärästä ei ollut saatavilla, mikä edelleen vähensi kiinnostusta vaihtoehtoa kohtaan. Drobon jäätyä pois aloin vertailla Synologyn ja Qnapin palvelimia, sillä niiden ominaisuudet ovat pääpiirteittäin samanlaiset. [20.] [23.]

Päädyin lopulta valitsemaan Synology Diskstation DS1812+ NAS-levypalvelimeen, joka on kerännyt paljon kehuja luotettavuutensa, ominaisuuksiensa, helppokäyttöisyytensä ja laajennettavuutensa vuoksi sekä se oli hiukan edullisempi hinnaltaan kuin

Qnap TS-869 Pro. Synology NAS-palvelimen lisäksi hankin siihen neljä kolmen teratavun kiintolevyä. Hankkimaani Synology DS1812+:aan pystyy asentamaan kahdeksan maksimissaan 4TB 3.5” SATA-kiintolevyä, ja kahden ulkoisen DX513 levykotelon avulla kiintolevyjen määrä voidaan kasvattaa jopa kahdeksaantoista, jolloin kokonaiskapasiteetin saa laajennettua jopa 72 teratavuun. Synology toimii myös erittäin hyvin erilaisten RAID-kokonaisuuksien kanssa ja sillä on myös oma RAID-järjestelmänsä, jota päätin käyttää. [20.] [23.]

Alustin levyt niin, että yhden kiintolevyn hajotessa voidaan vielä levyillä oleva data pelastaa. Toisena alustusvaihtoehtona olisi mahdollisuus datan pelastamiseen kahden levyn hajottua, mutta tällöin kiintolevyjen kapasiteetistä saadaan paljon pienempi osa hyötykäyttöön. Valitulla vaihtoehdolla kiintolevyillä oleville tiedostoille saatiin hyvä ja toimiva suojaus sekä jokaisen levyn käytettävissä oleva kapasiteetti saatiin pysymään 2,01 teratavussa. Koska Synologya käytetään verkossa, pystytään se haluttaessa liittämään jokaiseen projektin käytössä olevaan tietokoneeseen. Tällä hetkellä se on liitettynä kahteen koneeseen, kuten kuvasta 4 voidaan nähdä. Synologya liittäessä pitää niin Synologylle kuin kohdetietokoneelle lisätä sallittuihin yhteyksiin kummankin laitteen IP-osoitteet, jonka jälkeen määritellään vielä kohde tietokoneelle kansio, jonka alaisuudessa Synology toimii. [20.] [23.]



Kuva 4. Tämänhetkinen laitteisto ja niiden topologia.

4.3 Käytännöntoteutus

Projektin alussa vietin paljon aikaa selvittäessäni vanhojen varmuuskopioiden sisältöjä, puutteita ja arkistointia. Lisäksi järjestelin tiedostot loogisempaan järjestykseen. Alkuperäinen suunnitelma oli luoda automaattiset varmuuskopiot vain roskapostidatasta, mutta muutamien kuukausien sisällä suunnitelmat muuttuivat useasti ja lopulta varmuuskopiot laajenivat käsittämään lähestulkoon kaiken koneilla olevan tiedon. Suunnitelmien laajeneminen toi mukanaan erinäisiä haasteita, joten helpottaakseni työn toteutusta, tein listauksen jokaisen koneen sisällöstä ja siitä, mitä on järkevää varmuuskopioida ja mitä ei. (Liite 2.) Loppujen lopuksi tavoitteena oli saada varmuuskopioitua päivittäin saapuva data sekä sen lisäksi varmuuskopioitua käyttöjärjestelmän ja muun toiminnan kannalta tärkeät tiedostot jokaiselta tietokoneelta. Kaikki varmuuskopiointi suoritettiin ”cronuser”-käyttäjätilin alaisuudessa. Käyttäjätili toimii yhteisenä ylimääräisenä käyttäjätunnuksena projektiin osallistujien kesken, ja kaikki pystyvät näin ollen muokkaamaan sen alaisuuteen liitettyjä komentoja.

4.3.1 Järjestelmän varmuuskopiointi

Koska oli tiedossa, että koneiden sisältö muuttuu päivittäin esimerkiksi verkkosivujen päivityksien myötä, päätettiin varmuuskopiointi toteuttaa käyttäen Rsync-sovellusta, joka on suunniteltu tiedostojen synkronointiin. Rsync-sovellus oletusarvoisesti kopioi ainoastaan muuttuneet tiedostot sekä suurista tiedostoista vain muuttuneet osat, jolloin kopioinnin saa pidettyä nopeana eikä isompia kokonaisuuksia kopioidessa tarvitse erikseen murehtia kopioitavien tiedostojen valitsemisesta. Rsync suoritettiin SSH:n avulla, ja koska jokaiseen projektin käytössä olevaan tietokoneeseen tarvitaan salasana aina yhteyttä muodostaessa, otettiin avuksi SSH-avainpari, jonka avulla salasanojen kysely voitiin välttää. SSH on ohjelmisto, jolla voidaan luoda suojattu etäyhteys laitteesta toiseen. Yhteyden ansiosta voidaan toiselta koneelta käyttää esimerkiksi yhteistä palvelinta ja helposti suorittaa haluttuja toimintoja ilman, että tarvitsisi fyysisesti mennä palvelimen luokse ja kirjautua sisälle siihen. Ohjelmiston kehitti 1990-luvulla suomalainen Tatu Ylönen, ja nykyään se on erittäin yleisessä käytössä oleva työkalu. [24.] [25.]

Tietoturvallisuuden takia kaikille tietokoneille kirjautuessa tulee käyttäjän aina syöttää salasanansa. Suoritettaessa Rsync-sovelluksen avulla tiedonsiirtoa tietokoneelta toiselle ottaa käyttäjä yhteyden toiseen tietokoneeseen ja näin ollen joutuu syöttämään salasanan päästäkseen sisään. Tarkoituksena oli kuitenkin saada tiedonsiirto suoriutu-

maan automaattisesti, mikä tuotti haasteita ja oli löydettävä keino, jolla salasanan pakollisuus saataisiin kierrettyä. Lähdekoodissa 1 esiteltyjen ohjeiden avulla pystyttiin luomaan salasanaton SSH-yhteys eri tietokoneiden välille ja “cronuser”-käyttäjätunnus pystyi näin ollen siirtymään koneelta toiselle ilman salasanaa. Näissä ohjeissa a ja b korvattiin ”cronuser”-käyttäjätunnuksella ja A ja B kohdeasemien tunnuksilla.

```
“a@A:~> ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/a/.ssh/id_rsa):
Created directory '/home/a/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/a/.ssh/id_rsa.
Your public key has been saved in /home/a/.ssh/id_rsa.pub.
The key fingerprint is:
3e:4f:05:79:3a:9f:96:7c:3b:ad:e9:58:37:bc:37:e4 a@A
a@A:~> ssh b@B mkdir -p .ssh
b@B's password:
a@A:~> cat .ssh/id_rsa.pub | ssh b@B 'cat >> .ssh/authorized_keys'
b@B's password:”
```

Lähdekoodi 1. Koodi salasanatoman SSH-yhteyden salausavaimen luontia varten. [26.]

Suoritettaessa tietokoneiden järjestelmien varmuuskopiointeja ei tarkoituksena ollut kopioida ihan jokaista kansiota. Jokaiselle tietokoneelle tehtiin Rsync-skriptin yhteyteen selkeytyksen vuoksi *exclude.txt*-tiedosto, lähdekoodi 2, jonka sisälle pystyttiin listaamaan kaikki tiedostot ja kansiot, joita ei haluttu synkronoitavan. Kansiot, joita ei haluttu sisällytettävän synkronointiin, sisälsivät pääosin tiedostoja, jotka käsittelivät väliaikaisia tiedostoja.

```
### Made by MarkusB - 07/18/13
### Here are all the folders than we don't want to be rsynced, when running the
rsyncTurquoise.sh file.
/dev
/mnt
/opt
/proc
/selinux
/srv
/sys
/tmp
/var/tmp
/lib/init/rw
```

Lähdekoodi 2. Turquoiselle luotu Exclude.txt-tiedosto.

Itse Rsync-skriptissä pystyttiin täten kutsumaan *exclude.txt*-tiedostoa ja näin ollen koodi pystyttiin pitämään yksinkertaisempana ja synkronoitavien kansioiden lisääminen ja poistaminen selkeämpänä.

Lähdekoodissa 3 oleva skriptin määre ”-avze” tulee komennoista *archive, verbose, compress* ja *rsh=COMMAND*. Näistä *-a(archive)* määrittelee synkronoinnin sisältämään kaikki alihakemistot sekä tiedostojen symboliset linkit, omistussuhteet, käyttöoikeudet ja niin edelleen. *-v(verbose)* määrittelee skriptin tulostamaan ruudulle tiedon siitä, missä tiedostossa synkronointi on menossa, *-z(compress)* määrittelee tiedostot pakattavaksi pienempää muotoon synkronointia varten toimintojen nopeuttamiseksi ja *-e(rsh=COMMAND)* määrittelee käytettävän SSH-yhteyden. Lisäksi määritetään synkronointi suoritettavan ”superuser”-käyttöoikeuksin ”cronuser”-käyttäjänä, ”*sudo -u cronuser”,* käytettävän SSH-yhteyden parametrit sekä kohdekansio. *Sudo*-komennolla pystytään antamaan toiselle käyttäjätunnukselle oikeus suorittaa toiminto toisen käyttäjätunnuksen oikeuksin, joka oli *root*- eli pääkäyttäjä. Tällä tavalla pystyttiin *cronuser*-käyttäjätunnukselle antamaan käyttöjärjestelmän täysimuokattavuus, joka oli pakollista monien tiedostojen ja kansioiden omistusoikeuksien erilaisuuksien takia. [27.]

```
##!/bin/sh
### Rsyncs the whole filesystem to Synology, except the folders listed in ex-
clude.txt
cd / && sudo rsync -avze 'sudo -u cronuser ssh -p 31311' --exclude-from
'/backupfiles/exclude.txt' . topaz:/mnt/synology2/turquoise/
```

Lähdekoodi 3. Koko käyttöjärjestelmän varmuuskopioiva koodi.

Huomattuani synkronoinnissa esiintyvän useasti ongelmia ja sen välillä kestävän normaalia pitempään päätin lisätä loki-tiedoston, johon tallentuu tieto synkronoiduista tiedostoista, kuten lähdekoodissa 4. Näin ollen ongelmien ilmestyessä pystyin helpommin lähestymään ongelmaa ja paikantamaan, minkä tiedoston kohdalla ongelmat ilmenivät.

```
#!/bin/bash
### Markus Blomvall - 07/18/2013

date=`date +%Y-%m-%d`

### Runs the rsync script and writes a log file containing the information about
the rsynced files.
bash rsyncTurquoise.sh > /var/log/backupfiles/$date.rsync_turquoise.log 2>&1
```

Lähdekoodi 4. Koodi synkronoinnin etenemistä seuraavan loki-tiedoston lisäykselle.

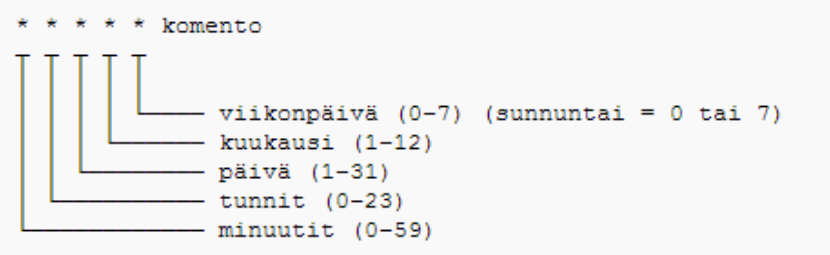
Varmuuskopioitavasta tietokoneesta riippuen sekä *exclude.txt*-tiedoston sisältö että kohdekansio vaihtelivat. Muuten samaa pohjaa pystyttiin käyttämään synkronoitaessa muutkin tietokoneet varmuuskopioita varten hankitulle NAS-palvelimelle. Synkronointia ei laitettu suoriutumaan automaattisesti vaan se pitää tehdä manuaalisesti. Tietokoneiden synkronoinnin automatisointi oli kylläkin suunnitelmissa, mutta synkronoinnin ajoituksesta ja toistuvuudesta ei päästy tyydyttävään lopputulokseen ja näin ollen sitä ei ole lisätty automaattisesti suoritettavien toimintojen listalle.

4.3.2 Päivittäiset varmuuskopiot

Päivittäisiä varmuuskopioita lähdettiin toteuttamaan siitä näkökulmasta, että ensin pitää turvata sivustolla julkaistava data. Tähän kuuluvat tällä hetkellä PSBL- ja CBL-data. Näistä PSBL-data tulee raakadatana, ja se pitää prosessoida ennen kuin sen esittäminen sivustolla onnistuu, joten tällöin oli tarvetta tallentaa niin PSBL-datan raaka kuin myös prosessoitu versio. Aivan tarkkoja tietoja ei ollut siitä, kuinka pitkään kunkin päivän PSBL- ja CBL-datat ovat saatavilla, mutta oletuksena oli, että PSBL-datan voi hankkia vain kyseisenä päivänä ja CBL-datan muutama päivä jälkikäteen. Näin ollen, jos PSBL-dataa ei joka päivä varmuuskopioida ja datan sisältävään tietokoneeseen tulee joku vika, on data kokonaan menetetty. Spamrankings.net-sivuston tilastot, taulukot ja kaaviot perustuvat kaikki kerättyyn dataan, joten datan menettämisen seurauksena menetettäisiin myös tilastot, joita sivustolla julkaistaan. Yhden päivän datan menetys ei yleensä vaikuta katastrofaalisesti tilastoihin, mutta on kuitenkin kuukausia, jolloin yhtenä tai muutamina päivinä on esimerkiksi roskapostittajien toimesta suoritettu hyökkäys tietoja keräämiä tahoja vastaan, ja nämä päivät ovat tilastoinnin kannalta oleellisia ja tärkeitä. Kaiken kaikkiaan PSBL- ja CBL-dataa on kerättynä jo noin kolmen vuoden takaa, ja se kaikki pyritään saamaan julkaistua tilastoituna Spamrankings.net-sivustolla, joten on erittäin tärkeää saada turvattua myös aiemmin kerätty data kokonaisuudessaan.

PSBL- ja CBL-datan noutaminen ja prosessointi suoritetaan automaattisesti aamuyöstä paikallista aikaa. Näin ollen automaattiset varmuuskopiot päätettiin ajastaa suoriutumaan iltapäivästä, jotta data olisi varmasti saatu jo tietokoneillemme. Varmuuskopioiden automaattinen suoriutuminen toteutettiin crontab-ohjelman avulla. Jokaisen käyttäjätunnuksen alaisuudesta löytyy oma tiedostonsa, johon käyttäjä voi kirjata haluamansa ajastetut toiminnot. Tiedostoa voidaan muokata crontab-ohjelman avulla ja sillä voidaan ajastaa toiminnot suoriutumaan haluttuna kellonaikana sekä haluttuina päivinä ja

kuukausina. Omaa tiedostoaan pääsee muokkaamaan ”*corntab -e*”-käskyllä. Tiedoston sisältämät komennot suoriutuvat ajastetusti cron daemonin toimesta, joka on itsenäisesti taustalla suoriutuva prosessi, jota käyttäjä ei voi itse suoranaisesti hallita. Kuvassa 5 on esitetty crontab-ohjelman ajastuksen periaate. [28.]



Kuva 5. Crontab-ohjelman ajastusperiaate. [29.]

Ensin ilmoitetaan haluttu aika asettamalla minuutit, tunnit, päivät ja kuukaudet sekä sen jälkeen kirjoitetaan suoritettava komento. Mikäli ilmoitettavan numeron sijasta syötetään ”*”, saadaan toiminto suoriutumaan kyseisen valinnan jokaisena hetkenä. Pilkkua käyttämällä voidaan halutulle hetkelle antaa monta eri arvoa ja väliviivan avulla voidaan listata kaikki arvot haluttujen numeroiden välillä. Lähdekoodista 5 voidaan nähdä, että varmuuskopiot on kirjattu suoriutumaan päivittäin kello 18:00, 18:10 ja 19:10.

```
# Script to do automatic backups of PSBL-data to Synology
0 18 * * * bash /backupfiles/syno_psbl_data.sh

# Script to do automatic backups of CBL-data to Synology
10 18 * * * bash /backupfiles/syno_cbl_data.sh

# Script to do automatic backups of PSBL-Rawdata to Synology
10 19 * * * bash /backupfiles/syno_raw_psbl_data.sh
```

Lähdekoodi 5. Crontab-ohjelmaan syötetyt skriptit.

Cronuser-käyttäjätunnuksen crontab-ohjelmalla muokattuun tiedostoon sijoitettiin kolme skriptiä, yksi CBL-dataa ja kaksi PSBL-dataa varten, joista toinen raakadataa ja toinen prosessoitua dataa varten. Koska kaikki projektin työntekijät tiesivät Cronuser-käyttäjätunnuksen salasanan, pääsivät kaikki myös muokkaamaan tarvittaessa näitä ajastettuja toimintoja. Näistä kolmesta datasta PSBL-datan raakaversion kopioiminen oli haastavin. Data on tallennettu kansioittain päivämäärän mukaan ja jokainen kansio sisältää keskimäärin 100 000-300 000 tiedostoa. Näin ollen kopioitaessa päivittäin tullut kansio sisältöineen tuli eteen ongelmia tiedostojen metadatan hallinnan kanssa. Perusajatuksena jokaista tiedostoa kohtaan on olemassa itse data sekä metadata. Meta-

data sisältää tiedot muun muassa tiedoston muokkauspäivästä, nimestä ja omistusoikeuksista. Metadatan sekä tiedoston sijaintidatan muodostavaa tietorakennetta kutsutaan inodeksi. [30.]

Kohdeasema, johon päivittäiset datat varmuuskopioitiin, voi sisältää enintään 137 miljoonaa inodea eli se voi maksimissaan sisältää 137 miljoonaa eri tiedostoa. Enimmäismäärä inodeja on määritelty alustettaessa kovalevyä ja yleinen käytäntö on varata niitä yhtä prosenttia kovalevyn kokonaiskapasiteetista vastaava määrä. Luku voi kuulostaa suurelta, mutta jos ajatellaan, että päivittäin näistä menisi käyttöön 100 000-300 000, niin muutaman vuoden kuluttua ne olisivat kaikki käytössä. Näin ollen päädyttiin siihen ratkaisuun, että päivittäinen PSBL-raakadata pakattaisiin yhdeksi tiedostoksi, jolloin ei olisi enää ongelmaa inode-tietorakenteiden riittävyyden kanssa, ja satojen tuhansien sijaan päivittäin niistä tulisi käyttöön vain yksi. [30.]

PSBL-raakadataa varten kirjoitetussa skriptissä, lähdekoodi 6, määritellään ensin kolme muuttujaa, päivä, kuukausi ja vuosi sen hetkisen päivämäärän mukaan. Sen jälkeen siirrytään kansioon, jossa sen hetkisen päivän datat sijaitsevat. Tämän jälkeen kaikki kansiossa olevat tiedostot kootaan yhdeksi tiedostoksi ja sen jälkeen pakataan gzip-sovelluksella. Seuraavaksi määritellään, että ne kopioidaan NAS-palvelimelle vuotta edustavaan kansioon, jonka alla on vielä kuukautta edustava kansio, johon tiedostot tallennetaan päivämäärän sisältävän tiedostonimen kera.

```
#!bin/bash
### Markus Blomvall - 06/19/2013

date=`date +%Y-%m-%d`
month=`date +%m`
year=`date +%Y`

cd /var/opt/spam/psbl/raw/$date

### Copy and pack the daily RAW PSBL-data to Synology, as cronuser, which is at-
tached to Topaz. Runs daily under cronusers crontab

tar cf - * | gzip -c | ssh -p 31311 cronuser@turquoise cd
/mnt/synology4/topaz/var/opt/spam/psbl/raw/$year/$month "&&" cat
">raw.psbl.$date.tar.gz"
```

Lähdekoodi 6. PSBL-raakadatan automaattista varmuuskopiointia varten kirjoitettu koodi.

Skriptit, joita käytetään kopioimaan prosessoitu PSBL-data ja CBL-data, ovat lähestulkoon identtisiä, vain tiedostonimet ja kohdekansiot eroavat, lähdekoodit 7 ja 8. Kummatkin kopioidaan Rsync-ohjelman avulla. Kummankin skriptin alussa määritellään

samat muuttujat kuin PSBL-raakadatan kopioivassa koodissa. Sen jälkeen käsketään kopioimaan sen hetkisen päivän tiedosto NAS-palvelimella olevaan kohdekansioon. Koodissa myös määritetään säilyttämään lähdetiedostojen kaikki ominaisuudet sekä ajettaessa skripti manuaalisesti se on määritetty tulostamaan näytölle, mitä kopioidaan ja miten kopiointi sujuu. Toisin kuin PSBL-raakadata, nämä kaksi sijoitetaan kummatkin yhden oman kansion alle vuodesta ja kuukaudesta riippumatta.

```
#!/bin/bash
### Markus Blomvall - 06/19/2013

date=`date +%Y-%m-%d`
month=`date +%m`
year=`date +%Y`

### Rsync the daily PSBL-data to Synology, as cronuser, which is attached to To-
paz. Runs daily under cronusers crontab

/usr/bin/rsync -av -progress --inplace --rsh='ssh -p 31311'
/var/opt/spam/volume/psbl/$date cronus-
er@turquoise:/mnt/synology4/topaz/var/opt/spam/volume/psbl/
```

Lähdekoodi 7. PSBL-datan automaattista varmuuskopiointia varten kirjoitettu koodi.

```
#!/bin/bash
### Markus Blomvall - 06/19/2013

date=`date +%Y-%m-%d`
month=`date +%m`
year=`date +%Y`

file=cbl.volume.$date

### Rsync the cbl.volume.(date) file to Synology, as cronuser, which is attached
to Topaz. Runs daily under cronusers crontab

/usr/bin/rsync -av -progress --inplace --rsh='ssh -p 31311'
/var/opt/spam/volume/cbl/$file cronus-
er@turquoise:/mnt/synology4/topaz/var/opt/spam/volume/cbl/
```

Lähdekoodi 8. CBL-datan automaattista varmuuskopiointia varten kirjoitettu koodi.

Prosessoidun PSBL- ja CBL-datan skriptit on molemmat tehty niin, että ne pystytään helposti ajamaan jälkikäteen, mikäli on esiintynyt jotain ongelmia kyseisenä päivänä tiedostojen saannissa tai prosessoinnissa. Niin PSBL- kuin myös CBL-datan prosessoidut versiot varmuuskopioituvat kokonaisuudessaan myös silloin, kun suoritetaan

koko järjestelmän synkronointia, joten jos jokin päivä on huomaamatta jäänyt väliin, voidaan näin olla varmempia, että kaikki data tulee varmuuskopioitua.

PSBL-raakadatan sisältävä kansio on sen sijaan listattu *exclude.txt*-tiedostoon. Syynä tähän on se, että PSBL-raakadata on pakattava ennen siirtoa, eikä sitä näin ollen voi suoraan synkronoida. Näin ollen oli tarpeen kehittää skripti, jolla voidaan kopioida myös aikaisempien päivien tiedostot, mikäli tiedostojen siirrossa on jonain päivinä ollut ongelmia, lähdekoodi 9. Skripti on sisällöltään suurimmalta osalta samankaltainen kuin päivittäin ajettava versio, ainoastaan alussa olevat muuttujat määritellään toisin. Skriptissä on sisäkkäin kaksi for-silmukkaa, jossa ensimmäisessä määritellään halutut kuukaudet ja sen sisällä olevassa silmukassa halutut päivät.

```
#!/bin/bash
### Markus Blomvall - 06/17/2013

### Choose the months you want to include in the copying
## for i in 01 02 03 04 05 06 07 08 09 10 11 12
for i in 07
do
### Goes in to the folder where the data folders are
cd /var/opt/spam/psbl/raw/

### Choose the days you want to copy
## for j in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
24 25 26 27 28 29 30 31
for j in 05 06 07 08 09 10 11 12 13 14 15 16 17 18
do

### Copies and packs the folders wanted to Synology under year 2013 and month
choosed in start.
tar cf - 2013-$i-$j | gzip -c | ssh -p 31311 turquoise cd
/mnt/synology4/topaz/var/opt/spam/psbl/raw/2013/$i "&&" cat ">raw.psbl.2013-$i-
$j.tar.gz"
done

done
```

Lähdekoodi 9. PSBL-raakadatan manuaalista varmuuskopiointia varten kirjoitettu koodi.

5 Kehitysideat

Tällä hetkellä käytössä olevat skriptit toimivat kuten haluttiin, mutta niissä on vielä muutamia kohtia, joita voitaisiin parantaa. Ensinnäkin koodi voitaisiin muokata hiukan selkeämmäksi ja helppolukuisammaksi sekä järjestelmän varmuuskopiointi voitaisiin lisätä crontabin alle. Myös muille päivittäin tuleville datoille voitaisiin luoda automaattiset varmuuskopiot sekä rakentaa sähköpostihälytykset, jotka ilmoittaisivat, jos jokin data ei jostain syystä ole varmuuskopioitunut.

Vaikka kirjoittamani koodit tällä hetkellä ovat samankaltaisia keskenään, voi niitä vielä kirjoittaa yhteneväisemmiksi, ja näin ollen niiden työstäminen taikka kopioiminen jatkossa olisi yksinkertaisempaa. Tallennuspalvelimena toimiva Synology on verkkoon liitetty komponentti, joten sen voi liittää halutessaan jokaiselle käytössä olevalle tietokoneelle. Näin ollen sen voi liittää myös siihen tietokoneeseen, johon päivittäiset datat tulevat. Tämän avulla voidaan käytössä olevaa koodia muokata niin, että kopioitaessa ei tarvitsisi luoda SSH-yhteyttä. Kirjoittamani koodit on pääosin tehty ennen kuin Synologya oltiin edes hankkimassa käyttöömme ja näin ollen kaikki kopiointi ja synkronointi piti suorittaa DroboPron hallinnoijana toimivan tietokoneen kautta.

Lähdekoodissa 10 esitettyä CBL-datan kopioivaa koodia on saatu muokattua kevyempään ja selkeämpään muotoon menettämättä kuitenkaan mitään ominaisuuksia, joita kopioinnilta halutaan. Alussa määritellään muuttujia: käsiteltävä data, kansio, päivämäärä, kuukausi, vuosi sekä tiedostonimi. Näiden avulla saadaan komento, joka suorittaa itse kopioinnin, pidettyä selkeämpänä. Siinä määritellään säilyttämään -a(archive), alkuperäisen tiedoston ominaisuudet, kuten tiedostojen symboliset linkit, omistussuhteet ja käyttöoikeudet. Mikäli skripti ajetaan manuaalisesti, on koodissa myös määritelty tulostettavaksi ruudulle prosessin eteneminen, -v(Verbose) sekä -progress. Aiempaan koodiin verrattuna on tässä päästy eroon SSH-verkkoyhteyden muodostuksen tarpeelta. Koodissa määriteltävät muuttujat on myös selkeästi eroteltu alussa, joten muokattaessa taikka käytettäessä samaa koodia toisten tiedostojen kopiointiin, on muuttujia helppo käsitellä.

```
#!/bin/bash

source=cbl
dir=/var/opt/spam/volume/$source

date=`date +%Y-%m-%d`
month=`date +%m`
year=`date +%Y`

file=$source.volume.$date

### Rsync the cbl.volume.(date) file to Synology, as cronuser, which is attached
to Topaz.
/usr/bin/rsync -av --progress $dir/$file /mnt/synology4/topaz/$dir
```

Lähdekoodi 10. Jatkokehitelty CBL-datan varmuuskopioiva koodi.

PSBL-datan kohdalla voidaan skriptiin tehdä lähes samankaltaiset muutokset, lähdekoodi 11. Ainoina eroina on tiedostopolkujen eroavaisuus sekä *file*-muuttujan puuttuminen, sillä PSBL-data on arkistoitu suoraan päivämäärän nimellä ja näin ollen sitä ei tarvitse määrittää.

```
#!/bin/bash

source=psbl
dir=/var/opt/spam/volume/$source

date=`date +%Y-%m-%d`
month=`date +%m`
year=`date +%Y`

### Rsync the daily PSBL-data to Synology, as cronuser, which is attached to To-
paz.
/usr/bin/rsync -av --progress $dir/$date /mnt/synology4/topaz/$dir
```

Lähdekoodi 11. Jatkokehitelty PSBL-datan varmuuskopioiva koodi.

Myös suorittaessa järjestelmien varmuuskopiointeja, voidaan luopua SSH-yhteyden tarpeesta ja keventää skriptiä, lähdekoodi 12.

```
#!/bin/bash

### Rsyncs the whole Topaz to Synology, except folders listed in exclude.txt
cd / && sudo rsync -avz --exclude-from '/backupfiles/exclude.txt' .
/mnt/synology4/topaz/
```

Lähdekoodi 12. Jatkokehitelty koko käyttöjärjestelmän varmuuskopioiva koodi.

Tällä hetkellä varmuus siitä, ovatko päivittäiset datat kopioituneet, on käyttäjän itsensä käsissä. Silloin tällöin kävi niin, että datan saannin kanssa oli ongelmia ja se jouduttiin

hakemaan myöhemmin päivästä manuaalisesti sekä sen jälkeen manuaalisesti prosessoimaan, ja näin ollen tämä sykli saattoi tulla valmiiksi myöhemmin kuin varmuuskopiot oli ajoitettu. Näin ollen käyttäjän on itse katsottava, että varmuuskopiot päivittäin onnistuvat. Välillä ilmeni myös ongelmia verkkoyhteyksien kanssa, jolloin verkonvälityksellä liitetty ulkoinen kovalevy ei ollut enää yhteydessä tietokoneisiin. Tällöin eivät myöskään automatisoidut varmuuskopiot onnistuneet. Tätä voitaisiin parantaa kirjoittamalla skripti, joka seuraisi varmuuskopioiden toimimista ja hälyttäisi sähköpostiviestillä, mikäli joltain päivältä data ei ole varmuuskopioitunut tai mikäli yhteys ulkoiseen kovalevyyn on jostain syystä katkennut. Tällä hetkellä on jo käytössä ilmoitusjärjestelmä, joka lähettää päivittäin sähköpostiviestillä tiedon kyseisenä päivänä tulleista datoista ja niiden määristä kaikille projektin jäsenille. Samankaltaista järjestelmää voitaisiin siis soveltaa myös varmuuskopioiden kanssa.

Tietokone, jolle päivittäiset datat tulevat, säilyttää sisällään dataa vain noin vuoden ajan, minkä jälkeen vanhimmat datat poistetaan uusien tieltä. Syynä tähän on inodejen riittämättömyys. Näin ollen käytössä oleva Synology on ainoa toimiva laite, johon on tallennettuna kaikki data alku ajoista lähtien. DroboPron levykkeillä on myös vanhaa dataa, mutta itse laitteen toiminta on erittäin epävarmaa, joten sen päivittäinen käyttäminen on tuskallista ja epäluotettavaa. Tästä syystä olisi hyvä hankkia toinen ulkoinen tallennuslaite taikka ottaa käyttöön pilvipalvelu, johon varmuuskopiot voitaisiin myös tallentaa. Tämänhetkisenä tarkoituksena on julkaista kaikki vuosien saatossa kerätty informaatio Spamrankings.net -sivustolle ja tästä syystä on erittäin tärkeää, että myöskään vanhempi data ei katoa.

6 Yhteenveto

Insinöörityön tarkoituksena oli luoda Spamrankings.net -projektille automaattisesti toimiva varmuuskopiointijärjestelmä. Työskentelin kyseisen projektin parissa puoli vuotta ja sinä aikana pystyin seuraamaan kehittämäni järjestelmän toimivuutta ja varmuutta.

Tuottamani järjestelmä toimii ja sisältää kaiken, mitä pyydettiin. Työskennellessäni projektissa oli se suurien muutosten kourissa ja moni asia oli vielä työn alla eikä monistakaan asioista oltu vielä aivan varmoja. Tästä syystä pyrin kirjoittamaan koodini niin, että ne olisivat mahdollisimman helposti muokattavissa ja kopioitavissa palvelemaan myös muiden tiedostojen ja tietokoneiden varmuuskopioinnissa. Järjestelmää tullaan varmasti kehittämään jatkossa, mutta aikataulusta on vaikea sanoa, sillä moni asia oli vielä vasta suunnitteluvaiheessa.

Olen itse tyytyväinen toteutukseen monilta osin. Asia, jonka parissa tuli mielestäni hiukan hätiköityä, oli varmuuskopioiden tallennukseen hankitun laitteen valinta. Markkinoilla on todella laaja skaala erilaisia laitteita ja niiden ominaisuudet vaihtelevat valtavasti. Valittaessa sopivaa laitetta oli sen saamisen kanssa erittäin kova kiire ja mielestäni valintaan ei pystytty paneutumaan sen vaatimalla tavalla. Hankkimani laite kuitenkin toimii halutulla tavalla ja tehokkaasti, joten en koe, että se olisi kuitenkaan ollut huono tai väärä valinta.

On mielenkiintoista seurata, kuinka suuriin mittoihin Spamrankings.net -projekti kasvaa, kun se vihdoin saadaan kunnolla julkiseen levitykseen ja ihmisten tietoisuus sen tekemästä työstä kasvaa. Mielenkiintoista on myös nähdä, miten yritykset tulevaisuudessa reagoivat sivuston kautta tulevaan informaatioon omien tietoturviensa kehittämisessä ja suunnittelussa. Spamrankings.net -projektin kaltainen lähestymistapa roskapostin levittämisen ehkäisemiseksi on muista eroava ja tuo hyvää vaihtelua ja uutta näkökulmaa erittäin suuren ongelman poiskitkemiseen.

Lähteet

1 Spam in 2012: Continued Decline Sees Spam Levels Hit 5-year Low. 2012. Verkkodokumentti. Kaspersky Lab. < http://www.kaspersky.com/about/news/spam/2013/Spam_in_2012_Continued_Decline_Sees_Spam_Levels_Hit_5_year_Low>. 21.1.2013. Luettu 6.9.2013.

2 Karl Muhlbach. 2010. The Effect of Spam on Business. Verkkodokumentti. SiteProNews. <http://www.sitepronews.com/2010/08/19/the-effect-of-spam-on-business/>. 19.8.2010. Luettu 10.10.2013.

3 Kate Stoodley. 2004. Father of Spam Speaks Out on His Legacy. Verkkodokumentti. eSecurityPlanet.com. <http://www.esecurityplanet.com/trends/article.php/3438651/Father-of-Spam-Speaks-Out-on-His-Legacy.htm>. 19.11.2004. Luettu 9.10.2013.

4 Roskapostin lyhyt historia. Verkkodokumentti. Mext. <http://www.mext.fi/stimarkkinointi/roskapostin-lyhyt-historia>. Luettu 9.10.2013].

5 Niko Rinta. 2011. Roskapostin määrä romahtanut - haittaohjelmat leviävät nyt Facebookissa ja Twitterissä. Verkkomedia. MikroPC. <http://www.mikropc.net/kaikki_uutiset/roskapostin+maara+romahtanut++haittaohjelmat+leviavat+nyt+facebookissa+ja+twitterissa/a735148>. 7.12.2011. Luettu 9.10.2013.

6 Roskaposti - Usein kysyttyä. 2012. Verkkomedia. Helsingin yliopisto - Helpdesk. <http://www.helsinki.fi/helpdesk/ohjeet/sahkoposti/roskaposti_ja_kalasteluviestit/spam_faq.html>. Luettu 9.10.2013.

7 Botit ja Bottiverkot - Kasvava uhka. Verkkomedia. Norton by Symantec <http://fi.norton.com/botnet>. Luettu 10.10.2013.

8 Petrus Laine. 2013. Symantec teki ison loven maailman suurimpaan bottiverkkoon. Verkkomedia. Muropaketti / OtavaMedia. <http://muropaketti.com/symantec-teki-ison-loven-maailman-suurimpaan-bottiverkkoon>. 2.10.2013. Luettu 15.10.2013.

9 Spamrankings.net. 2013. Verkkomedia. Center for Research of Electronic Commerce. <http://www.spamrankings.net/>. Luettu 6.9.2013.

10 Heinz Tschabitscher. How Spammers Get Your Email Address. Verkkomedia. About.com - Email. < http://email.about.com/od/spamandgettingridofit/a/spam_finds_you.htm>. Luettu 7.1.2014.

11 The Law of Spam Harvesting. Verkkomedia. Project Honey Pot.

<http://www.projecthoneypot.org/law_of_harvesting.php>. Luettu 7.1.2014.

12 How to Avoid Spambots. Verkkomedia. Project Honey Pot. <http://www.projecthoneypot.org/how_to_avoid_spambots.php>. Luettu 7.1.2014.

13 Samuli Kotilainen. 2005. Harvester-palvelimien ip-osoitteet talteen - Hunajapurkki paljastaa roskapostittajan. Verkkomedia. Tietokone.fi. <http://www.tietokone.fi/artikkeli/arkisto/2005/hunajapurkki_paljastaa_roskapostittajan>. 17.1.2005. Luettu 17.10.2013.

14 About Project Honey Pot Verkkomedia. Project Honey Pot. <https://www.projecthoneypot.org/about_us.php>. Luettu 17.10.2013.

15 The Composite Blocking List. Verkkomedia. What Is My Ip Address. <http://whatismyipaddress.com/blacklist/cbl>. Luettu 7.1.2014.

16 Internet toimii IP-osoitteilla. 2013. Verkkomedia. Viestintävirasto. <https://www.viestintavirasto.fi/internetpuhelin/internet/ip-osoitteet.html>. Päivitetty 1.3.2013. Luettu 7.1.2014.

17 PSBL - Passive Spam Block List. Verkkomedia. PSBL. <http://psbl.org/>. 2013. Luettu 9.9.2013.

18 Bradley Mitchell. Introduction to NAS - Network Attached Storage. Verkkomedia. About.com - Wireless / Networking. <http://compnetworking.about.com/od/itinformationtechnology/l/aa070101a.htm>. Luettu 18.1.2014.

19 Bradley Mitchell. Introduction to NAS - Network Attached Storage. Verkkomedia. About.com - Wireless / Networking. <http://compnetworking.about.com/od/itinformationtechnology/l/aa070101b.htm>. Luettu 18.1.2014.

20 Synology Diskstation DS1218+. 2013. Verkkomedia. Synology Inc. <http://www.synology.com/en-us/products/overview/DS1812%2B. Luettu 12.9.2013.

21 TS-869 Pro. Verkkomedia. QNAP Systems, Inc. <http://www.qnap.com/en/indexphp?lang=en&sn=822&c=351&sc=513&t=517&n=9789&g=2>. Luettu 1.28.2014.

22 Drobo B800fs. Verkkomedia. DROBO, INC. <http://www.drobo.com/products/business/b800fs/>. Luettu 28.1.2014.

23 IntraIpsum. 2012. Setting up Linux access to the Synology NAS shared folders. Verkkomedia. Intra Ipsum. <http://www.intraipsum.se/blog/2012/07/09/setting-up-linux-access-to-the-synology-nas-shared-folders/>. 9.7.2012. Luettu 8.10.2013.

24 Teppo Oranne ja Jani Markkanen. 2012. Rsync. Verkkomedia. Linux.fi.

<http://linux.fi/wiki/Rsync>. 10.3.2012. Luettu 12.9.2013.

25 Teppo Oranne ja Jani Markkanen. SSH. Verkkomedia. Linux.fi. <http://linux.fi/wiki/SSH>. 1.1.2013. Luettu 24.10.2013.

26 Mathias Kettner. SSH login without password. Verkkomedia. The Linux Problem Base. <http://www.linuxproblem.org/art_9.html>. Luettu 12.9.2013.

27 Paul Mackerras ja Andrew Tridgell. Rsync. Verkkomedia. Rsync web pages. <http://rsync.samba.org/ftp/rsync/rsync.html>. Päivitetty 28.7.2013. Luettu 12.9.2013.

28 Muutamia Unix-työkaluja. 1999. Verkkomedia. Jyväskylän Yliopisto. <http://www.mit.jyu.fi/opiskelu/kurssit/unix99/lecture6/index.html>. 4.10.1999. Luettu 29.1.2014.

29 Cron. Verkkomedia. Wikipedia. <http://fi.wikipedia.org/wiki/Cron>. 2.5.2013. Luettu 13.9.2013.

30 Jeffrey B. Layton. 2011. What’s an inode?. Verkkomedia. Linux Magazine. <http://www.linux-mag.com/id/8658/>. 9.6.2011. Luettu 16.9.2013.

31 CBL - Composite Blocking List. Verkkomedia. The Spamhaus Project Ltd. <http://cbl.abuseat.org/>. 21.03.2013. Luettu 9.9.2013.

Varmuuskopioiden toimintaperiaatteen dokumentaatio

Here's brief instruction how the backups to Synology work and where they are located.

Remember to use *screen* when running these rsyncs, because they might run for quite a long time and also to prevent from any interrupts, if something happens to your computer. Also check the files when using them, so you will know if they rsync to Synology that's attached to Topaz or Turquoise. If needed, the target location can be edited the way you prefer.

Topaz // Synology4 (turquoise:/mnt/synology4/topaz or topaz:/mnt/synology4/topaz)

In Topaz the files used for backups can be found from *topaz:/backupfiles/*
There you can find;
excluded.txt → File containing all the folders** that needs to be excluded in the rsync process of Topaz
rsyncwithlogfile.sh → File that runs the syno_rsync_topaz.sh and makes a log file**** of the results
syno_cbl_data.sh → Daily* backup to collect CBL data
syno_manual_raw_psbl_data.sh → File to manually collect raw PSBL data, when backups have failed.***
syno_psbl_data.sh → Daily* backup to collect PSBL data
syno_raw_psbl_data.sh → Daily* backup to collect raw PSBL data
syno_rsync_topaz.sh → File to run manually to rscyn Topaz to Synology (excluding exclude.txt)
~~syno_ut_data.sh → Daily* backup to collect UT data~~

* Runs under cronusers crontab daily around 6 PM.
** /dev, /mnt, /opt, /proc, /selinux, /srv, /sys, /tmp, /var/tmp, /lib/init/rw and /var/opt/spam/psbl/raw (/var/opt/spam/psbl/raw is excluded, because rsyncing it would cause that we would run out of inodes).
*** Remember to edit this file, so you'll collect the exact data you want.
**** Log file is located in /var/log/backupfiles/ and the file gets the date when it was ran, in to the file-name.

Emerald // Synology1 (turquoise:/mnt/synology1/emerald or topaz:/mnt/synology1/emerald)

You can find the files to do the backups of Emerald from *emerald:/var/opt/spam/bin/*. There you can find:
backupdata.sh
backupetc.sh
backuphome.sh
backup_var_www.sh
→ Files backupdata.sh, backupetc.sh, backuphome.sh and backup_var_www.sh are made to backup just /data, /etc, /home and /var/www/ folders.

For rsyncing the whole Emerald, use *rsyncEmerald.sh.* (Also located in *emerald:/var/opt/spam/bin/*).
In that file there are few folders that are excluded, that's because they are not necessary files/folders. These folders are */proc*, */dev*, */sys* and */lib/init/rw*.

How it looks like:
##!/bin/sh
cd / && sudo rsync -ravze 'sudo -u cronuser ssh -p 31311' --exclude '/proc' --exclude '/sys' --exclude '/dev' --exclude '/lib/init/rw' . turquoise:/mnt/synology1/emerald/

None of these files on Emerald run automatically at the moment.
To run these files, you need to type: *emerald:/var/opt/spam/bin$ bash rsyncEmerald.sh*

Turquoise // Synology2 (turquoise:/mnt/synology2/turquoise or topaz:/mnt/synology2/turquoise)

The files to rsync Turquoise can be found from *turquoise:/backupfiles/*. The files are:
excluded.txt → File containing all the folders* that needs to be excluded in the rsync process of Turquoise
rsyncTurquoise.sh → File to run manually to rscyn Turquoise to Synology (excluding exclude.txt)
rsyncwithlogfile.sh → File that runs the rsyncTurquoise.sh and makes a log file** of the results

* /dev, /mnt, /opt, /proc, /selinux, /srv, /sys, /tmp, /var/tmp and /lib/init/rw
** Log file is located in /var/log/backupfiles/ and the file gets the date when it was ran, in to the filename.

What the *rsyncTurquoise.sh* actually looks like:
##!/bin/sh
cd / && sudo rsync -ravze 'sudo -u cronuser ssh -p 31311' --exclude-from '/backupfiles/exclude.txt' . topaz:/mnt/synology2/turquoise/

None of these files on Turquoise run automatically at the moment.
To run these files, you need to type: *turquoise:/backupfiles$ bash rsyncTurquoise.sh* or *turquoise:/backupfiles$ bash rsyncwithlogfile.sh*

Made by Markus Blomvall – 07/16/2013
Edited by Markus B – 07/19/2013

Lista varmuuskopioitavista tiedostoista

TOPAZ			
Folder / File	DO	DON'T	Empty folder
/bin	X		
/boot	X		
/dev		X	
/etc	X		
/home	X		
/initrd.img	X		
/lib	X		
/lost+found	X		
/media	X		
/mnt		X	
/opt		X	X
/proc		X	
/root	X		
/sbin	X		
/selinux		X	X
/srv		X	X
/sys		X	
/tmp		X	
/usr	X		
/var	X		
/var/tmp	X		
vmlinuz	X		

EMERALD			
Folder / File	DO	DON'T	Empty folder
/bin	X		
/boot	X		
cdrom	X		
/data	X		
/dev		X	
/emul	X		
/etc	X		
/home	X		
/initrd.img	X		
/lib	X		
lib32	X		
lib64	X		
/lost+found	X		
/media	X		
/mnt	X		
/opt		X	X
/proc		X	
/root	X		
/sbin	X		
/selinux		X	X
/setups	X		
/srv		X	X
/sys		X	
/tmp		X	
/usr	X		
/var	X		
/var/tmp		X	
vmlinuz	X		
www	X		

TURQUOISE			
Folder / File	DO	DON'T	Empty folder
03	X		
/bin	X		
/boot	X		
cdrom	X		
/dev		X	
/emul	X		
/etc	X		
/home	X		
/initrd.img	X		
/lib	X		
lib32	X		
lib64	X		
/lost+found	X		
/media	X		
/mnt		X	
/opt	X		
/proc		X	
/root	X		
/sbin	X		
/selinux		X	X
/srv		X	X
/sys		X	
/tmp		X	
/usr	X		
/var	X		
/var/tmp		X	
vmlinuz	X		