Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačních technologií

Diplomová práce

Monitorování aktivit na internetu

Internet activities monitoring

Vedoucí diplomové práce: Ing. Martin Havránek

Vypracovala: Dana Papíková

Čestné prohlášení

Prohlašuji, že jsem tuto diplomovou práci na téma „Monitorování aktivit na internetu“ vypracovala samostatně s použitím pramenů uvedených v seznamu použité literatury a s využitím znalostí z odborných konzultací.

V Praze dne 29.3.2011

……

Dana Papíková

Poděkování

Velice ráda bych touto cestou poděkovala těm, kteří mi svými hodnotnými připomínkami, odbornými konzultacemi a mnoha dalšími radami velmi pomohli při lepší orientaci v dané problematice.

Jsou to zejména Ing. Martin Havránek, pracovníci zákaznické podpory společnosti InveaTech a.s., Ing. Zuzana Pacáková.

Souhrn

Diplomová práce se zabývá analýzou síťového provozu. Zpracována jsou NetFlow data z nasazené síťové sondy s podporou NetFlow protokolu pro monitorování provozu informačních systému. Ze získaných dat jsou vytvořeny statistiky a provedena analýza, která pomáhá určit atributy, které jsou podstatné pro detekci anomálie způsobené podezřelým provozem. Aplikována je dataminigová metoda rozhodovacího stromu. Rozhodovací strom je tvořen algoritmem C.5. Dále jsou získané znalosti implementovány do vytvoření systémového automatizovaného opatření, které bude schopno výskyt anomálie detekovat a zaznamenat. V případě výběru vhodného modelu a statistickém potvrzení správnosti, může být model použit jako predikční model.

Klíčová slova

datamining, vizualizace, NetFlow, sonda, anomálie, incident, skenování, rozhodovací strom,

Summary

This diploma thesis handles with analyzing the network traffic. Data are processed from network probe what is installed in network topology. The data are in the NetFlow protocol format. NetFlow is used for monitoring specific information system. The collected data are used to generate statistics and analysis what helps to determine the attributes that are essential to detect network anomalies caused by the suspect traffic. From data mining methods is applied a decision tree. The used decision tree algorithm is C.5. Furthermore, the knowledge gained from analysis and data mining are used to create a system implemented automated action to be able to detect network anomalies if occurrence will exists again. Selection an appropriate model and statistical confirmation of accuracy means that the model can be used as predictive model.

Keywords

datamining, vizualization, NetFlow, probe, anomaly, incident, scaning, decision tree

1. Úvod

V dnešní době již není žádnou novinkou, že budoucnost patří informačním a komunikačním technologiím. Jejich využíváním však vzniká velký objem dat, kterými jsme zcela obklopeni. Data jsou jednou z největších cenností každé firmy. Tato data je třeba analyzovat, zpracovávat a převést je na informace. Získaná data budou sloužit jako odpovědi na komplikované analytické dotazy a umožní nám tak zkvalitnit proces strategického a operativního rozhodování a pomáhat při snižování rizik.

Spolehlivá a dobře zabezpečená počítačová síť je podstatná pro úspěšné fungování každého podniku. I krátkodobý výpadek znamená narušení infrastruktury organizace a může způsobit nemalé a velmi nákladné škody, ztrátu důvěry zákazníků. Rovněž komplikace způsobující sníženou dostupnosti sítě a pomalou odezvu kritických aplikací se jistě podepíší na dobrém jménu společnosti.

Jednou z možností jak snížit výskyt takovýchto komplikací je zavedení průběžného monitorování síťového provozu. K tomu lze využít záznamy NetFlow, které jsou generovány aktivními síťovými prvky s podporou tohoto standardu. Po nasazení pravidelného sledování sítě lze rozeznávat anomálie a odchylky od běžného provozu, které nás mohou upozornit na případný útok na sledovanou síť.

2. Cíl práce a metodika

2.1 Cíl práce

Cílem této diplomové práce je provedení analýzy na objemných datech získaných z internetového provozu informačních systémů. Analýza je zaměřena na detekci anomálií souvisejících s provozem a vytížením sítě, přesněji detekcí možných skenování sítě, DoS útoků nebo pokusů o prolomení hesla služby SSH. Výstupem analýzy je zavedení takových automatizovaných opatření, která pomohou výskyt anomálie vizualizovat nebo jiným způsobem upozornit, že došlo k nezvyklé události v provozu.

2.2 Metodika

Kapitola 3 popisuje vybrané současné techniky pro dolování informací z dat. Je zde probrán postup tvorby modelu během zpracování dat a metody související se zpracováním velkých objemů dat.
V kapitole 4 je shrnut význam vizualizace výstupů a jsou zde zobrazeny ukázky, jakým způsobem je možné data prezentovat.
Obsahem kapitoly 5 je stručné představení několika vybraných nástrojů, které slouží účelům dataminingu.
Praktická část diplomové práce je zahrnuta v kapitole 6. Kapitola představuje protokol NetFlow, volně dostupné nástroje pro prácí s tímto protokolem, síťovou sondu. Dále je v kapitole uveden výčet bezpečnostních incidentů, na které se práce zaměřuje. Kapitola rovněž zahrnuje postup analýzy zpracování dat a její výsledky. V závěru kapitoly je uveden postup nastavení systémových opatření pro detekci anomálie.

Při tvorbě obsahu jsem vycházela ze zdrojů, které jsou uvedeny v seznamu literatury na konci této práce.

3. Metody dataminingových projektů

Dolování dat, datamining, je v současné době velmi rychle se vyvíjející obor v počítačovém průmyslu. Jednou z největších předností dolování dat se projevuje v jeho široké škále metod a technik, které lze aplikovat na celou řadu problémů sledovaných souborů.

V této kapitole je zdůvodněna potřeba zabezpečení webových aplikací v současné době a probrán obecný postup aplikovatelný na jakýkoliv projekt související se zpracováním velkého množství získaných informací. Obě tyto kapitoly úzce souvisí s tématem diplomové práce.

3.1 Aktivity na internetu

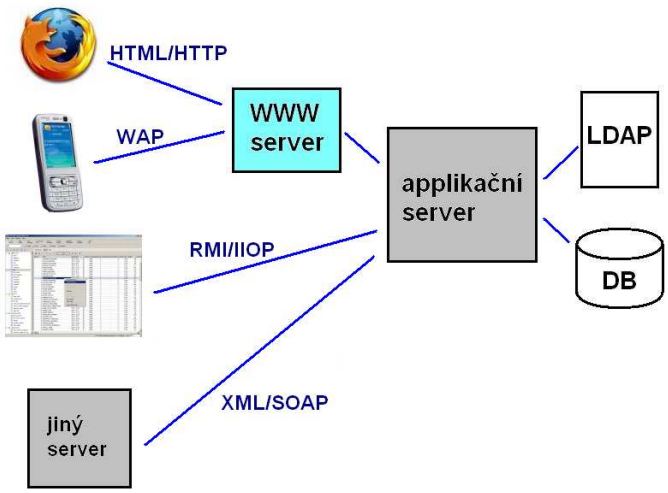
Na začátek si uveďme jednoduchou definici.

Internet je globální informační systém, který je logicky propojen jednoznačným adresovým prostorem založeným na Internet Protokolu (IP) a jeho rozšíření, podporuje komunikace za použití protokolů Transmission Control Protocol/Internet Protocol (TCP/IP) a jejich rozšíření, vykonává, užívá a zpřístupňuje soukromé i veřejné služby založené na komunikacích a popsané infrastruktuře. [5]

Za posledních patnáct let proběhl v oblasti internetu a internetových technologií opravdu dynamický vývoj. Největší růst ovšem zaznamenala především služba World Wide Web, síť vzájemně propojených webových stránek. Tato “síť” je často chápána laickou veřejností jako synonymum ke slovu internet.

Webové stránky jsou dnes v podstatě nejrozšířenějším zdrojem informací, poměrně nenákladným způsobem pro prezentaci firem, obchodů, šíření informací, zpracování marketingových průzkumů a v neposlední řadě i oblíbeným místem zábavy. Toto místo je však dnes oživené spoustou interaktivních, vzájemně propojených a uživatelsky přítulných aplikací tzv. webových aplikací.

Webová aplikace je aplikace poskytovaná uživatelům z webového serveru přes počítačovou síť internet, nebo její vnitropodnikovou obdobu (intranet). Webové aplikace jsou populární především pro značnou rozšířenost webového prohlížeče jako klienta. Ten se pak nazývá tenkým klientem, neboť sám o sobě logiku aplikace nezná.



Obr.č.1 – Schéma webové aplikace

Schopnost aktualizovat a spravovat webové aplikace bez nutnosti šířit a instalovat software na potenciálně tisíce uživatelských počítačů je hlavním důvodem jejich oblíbenosti. Webové aplikace jsou používány pro implementaci mnoha

podnikových i jiných informačních systémů, ale i freemailů, internetových obchodů, online aukcí, diskusních fór, weblogů.

Provozování webových aplikací má však i svá úskalí. Používání těchto aplikací je spojeno s přenosem velkého množství dat. Data jsou většinou uložena v databázi. Pokud aplikace využívající databázi obsahuje nějakou z bezpečnostních chyb, vzniklou nesprávným návrhem, programátorskou implementací nebo chybnou konfigurací systému, mohou být data v ohrožení. Webové aplikace tak fungují jako brány k databázím a potažmo tedy i citlivým údajům v nich uložených. Důraz kladený na vysokou míru jejich zabezpečení by měl být předním kritériem při jejich vývoji.

Monitorováním provozu webových aplikací, lze získávat potřebné informace sloužící dále ke zvýšení zabezpečení systému organizace. Ztráty způsobené jakýmkoli narušením z vnější mohou dosahovat značných rozměrů.

3.2 Postup dataminingu

Vzhledem k rostoucí složitosti a velikosti dnešních dat má datamining určitě své místo v informačních technologiích. V praxi má datamining dva primární cíle – popis a předpověď.

Popisný datamining byl vytvořen k vyhledávání skrytých, hodnotných a netriviálních závislostí ve velkých objemech dat a popisu nepřímé, automatické techniky analýzy dat, které využívají složitější a sofistikovanější nástroje než ty, které slouží k pouhé post analýze dat. Cílem je tedy pochopení analyzovaného systému, odhadování vzorců a vztahů ve velkých datových sadách.

Princip prediktivního dataminingu je založen na vytvoření primárního modelu k popisu faktických, fyzikálních, sociálních, behaviorálních a dalších vlastností systémů. Takový přístup používá jako základ formální model, na kterém se pak staví různé aplikace ve všech oborech lidské činnosti, které využívají informační a komunikační technologie.

Bez snahy pokrýt všechny existující názory a přístupy vztahující se k široké definici dolování dat, uveďme jeden z možných:

DEF: Datamining je opakující se proces objevování různých modelů, souhrnů a odvozených hodnot z daného sběru údajů [1]

Podstatou definice je cesta za poznáním zahrnující opakované procházení datových souborů s cílem nalezení odpovědí na nové dotazy. To se může opakovat mnohokrát. Každá technika je používána s mírně odlišnými aspekty dat, čímž mohou vznikat mírné odlišnosti v odpovědích. Účelem je pečlivě plánovaný proces rozhodování o tom, co bude velmi užitečné, perspektivní a odhalující.

Ve snaze o vytvoření standardního modelu procesu dobývání znalostí z dat vznikla metodika CRISP-DM (CRoss – Industry Standard Process for Data Mining). Vytvoření takovéto metodiky umožní řešit rozsáhlé úlohy dobývání znalostí rychleji, efektivněji, spolehlivěji a s nižšími náklady. Kromě návrhu univerzálního postupu má CRISP-DM nabízet průvodce potencionálními problémy a řešeními, které se mohou vyskytnout v reálných aplikacích.
Životní cyklus dataminingového projektu je podle metodiky CRISP-DM tvořen následujícími fázemi: [3]

1. POROZUMĚNÍ PROBLÉMU

Stanovení problému a formulace hypotézy.Vyžaduje kombinaci odborných znalostí a zkušenosti, jasnou specifikaci problému. V praxi to obvykle znamená vzájemné působení mezi datamining specialistou a specialistou dané aplikace.

2. POROZUMĚNÍ DATŮM

Týká se dat vytvořených i shromážděných. Důležité je si uvědomit, jak sběr dat ovlivňuje teoretické rozdělení a později i konečnou interpretaci výsledků.

3. PŘÍPRAVA DAT

Obecně zahrnuje zjišťování (a odstranění) odchylek, škálování, kódování a výběr funkcí pro převedení na stejné váhy atributů.

4. ODHAD MODELU

Hlavním úkolem této fáze je výběr příslušné datamining techniky. V praxi se vychází z několika vhodných modelů, z nichž se vybere ten nejlepší pro řešenou úlohu.

5. HODNOCENÍ A INTERPRETACE MODELU

Přesnost modelu a přesnost jeho interpretace jsou poněkud protichůdné zájmy. Obvykle, jednoduché modely jsou dobře vyhodnotitelné, nicméně na úkor jejich přesnosti. Od moderního dataminingu se očekávají velmi přesné výsledky pomocí multidimenzionálních modelů. Problém interpretace takových modelů je považován za samostatný úkol se specifickými technikami pro ověření výsledků. Bez ohledu na to, jak silnou dataminingovou metodu zvolíme, výsledný model nebude platný, pokud analyzované údaje nebudou správně shromážděny a předzpracovány nebo nebude-li formulace problému smysluplná. [1]

6. VYUŽITÍ V PRAXI

Konkretizace obsahu jednotlivých fází a doporučené výstupy jsou znázorněny v tabulce č.1Etapy a úlohy podle CRISP-DM [3]

Porozumění problému	*Porozumění datům*	*Příprava dat*	*Odhad modelu*	*Hodnocení*	*Využití v praxi*
Určení cíle *Základní kontext* *Cíle v rozhodování a poznání* *Kritérium úspešnosti řešení* ***Ohodnocení situace*** *Možné zdroje* *Požadavky, předpoklady a omezení* *Nebezpečí a nepředvídatelné výdaje* *Terminologie* *Náklady a zisky, hodnota řešení a přínos* ***Určení cíle projektu DM*** *Cíle a kritéria úspěšnosti řešení DM projektu* ***Tvorba plánu projektu*** *Plán projektu* *Výběr a hodnocení potřebných nástrojů a technik*	***Vstupní data- předběžný popis*** *Zpráva o vstupních datových zdrojích a jejich dostupnosti* ***Popis dat*** *Zpráva o zdrojích dat* ***Kontrola kvality dat*** *Zpráva o kontrole vstupních dat* ***Explorace dat*** *Pochopení významu dat* *Studium vztahů a informačních vazeb mezi proměnnými* *Struktury dat*	*Soubor dat* *Popis souboru dat* ***Výběr dat pro modelování*** *Posouzení výběru vhodných dat* ***Čištění dat*** *Zpráva o čištění dat* ***Konstrukce dat*** *Odvozené proměnné* *Generované záznamy* *Transformace* ***Intergrace dat*** *Spojení datových souborů* *Agregace* ***Formátování dat*** *Uspořádání sloupců i řádkově datové matice pro DM* *Vhodné kódování*	***Přístup k modelování – design postupu*** *Zavedení konstrukční a testovací fáze* ***Tvorba modelu*** *Nastavení parametrů* *Tvorba modelů* *Popis modelů* ***Hodnocení modelu*** *Ohodnocení modelu* *Revize nastavení parametrů*	***Vyhodnotit výsledky*** *Ohodnocení výsledků DM procesu vzhledem ke zvoleným kritériím úspěšnosti* *Přijetí modelu* ***Posouzení procesu*** *Posouzení celého datového procesu* ***Stanovení dalších kroků*** *Uvedení možných dalších kroků akce*	***Aplikace*** *Plán aplikace* ***Monitorování a údržba procesu aplikace*** *Plán monitoru a údržby* ***Tvorba závěrečné zprávy*** *Závěrečná zpráva* *Závěrečná prezentace* ***Posouzení projektu*** *Dokumentace praktických zkušeností*

Tab.č.1- Etapy a úlohy podle CRISP-DM

Zdroj [3]

3.3 Metody pro datamining

3.3.1 Rozhodovací stromy

Rozhodovací stromy jsou velmi často používanou technikou, oblíbenou zejména pro svou jednoduchou interpretaci a přehlednost, která umožňuje uživatelům rychle a snadno vyhodnotit získané výstupy.

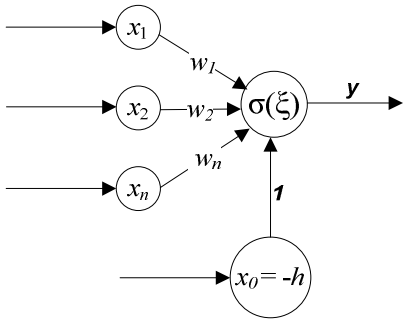
Jedná se o prediktivní model, který zobrazuje data v podobě stromu, kde každý uzel určuje kritérium pro následné rozdělení dat do jednotlivých větví. Strom tak rozděluje veškerá zdrojová data do segmentů, kde každý list odpovídá určitému segmentu definovanému předchozími uzly. Každý segment se tak vykazuje určitými shodnými vlastnostmi. Rozhodovací stromy mohou být založeny na použití množství algoritmů, zaměřme se na například C4.5, C5, CHAID (Chi-squared Automatic Interaction Detector), CART (Classification and Regression Trees), QUEST (Quick Unbiased Effecient Statistical Tree). [3][8]

3.3.2 Rozhodovací pravidla

Metoda rozhodovacích pravidel se využívá v případech, kdy je obtížné pochopit rozsáhlý rozhodovací strom, jehož uzly mají specifický kontext vycházející z předchozích uzlů. Převodem na rozhodovací pravidla lze učinit model rozhodovacího stromu přehlednějším. Cesta ke každému listu se transformuje na na IF – THEN pravidlo. Část IF obsahuje všechny testy pokrývající cestu k danému listu, a část THEN obsahuje konečnou klasifikaci.

3.3.3 Neuronové sítě

Rovněž jako rozhodovací stromy jsou neuronové sítě používány pro tvorbu prediktivních modelů. Jsou založeny na obdobných principech, které napodobují organizaci nebo způsob chování biologických struktur.

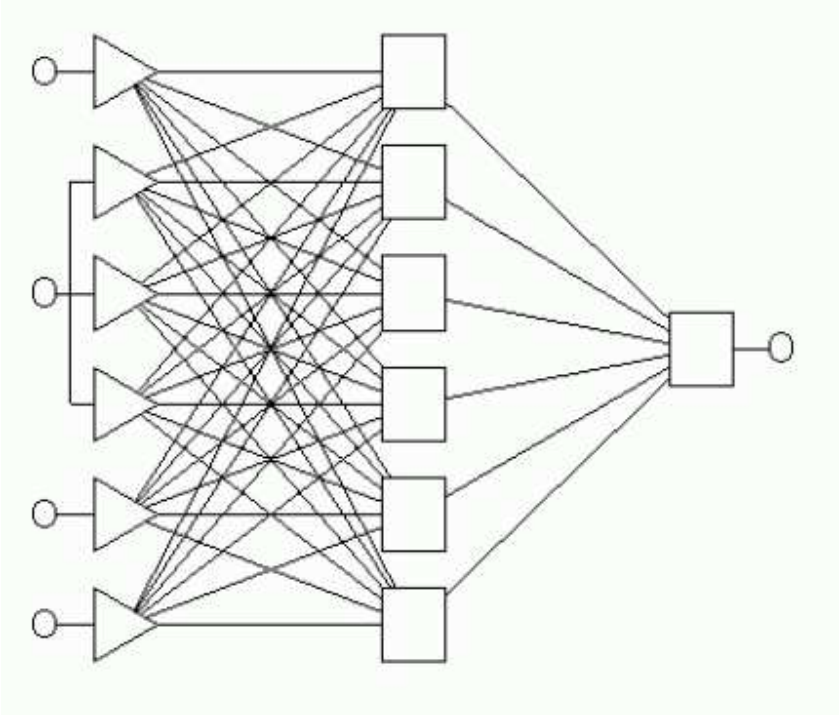


Obr.č. 2 - Schéma umělého neuronu

Skládá se z umělých (nebo také formálních) neuronů, jejichž předobrazem je biologický neuron. Neurony jsou vzájemně propojeny a navzájem si předávají signály a transformují je pomocí určitých přenosových funkcí. Neuron má libovolný počet vstupů, ale pouze jeden výstup.

Neuronové sítě se používají mimo jiné i pro rozpoznávání a kompresi obrazů nebo zvuků, předvídání vývoje časových řad, někdy i k filtrování spamu.

Neuronové sítě mají schopnost učení. Cílem učení neuronové sítě je nastavit síť tak, aby dávala přesné výsledky. V biologických sítích jsou zkušenosti uloženy v dendritech. V umělých neuronových sítích jsou zkušenosti uloženy v jejich matematickém ekvivalentu - váhách. Učení neuronové sítě rozlišujeme na učení s učitelem a učení bez učitele. Fáze učení neuronové sítě bývá nazývána adaptivní. Po naučení neuronové sítě je síť ve fázi vybavování.[1]



Obr.č.3 – Neuronová síť

Učení s učitelem

Podobně jako v biologických sítích je zde využita zpětná vazba. Neuronové síti je předložen vzor. Na základě aktuálního nastavení je zjištěn aktuální výsledek. Ten porovnáme s vyžadovaným výsledkem a určíme chybu. Poté spočítáme nutnou korekci (dle typu neuronové sítě) a upravíme hodnoty vah či prahů, abychom snížili hodnotu chyby. Toto opakujeme až do dosažení námi stanovené minimální chyby. Poté je síť adaptována.[1]

Učení bez učitele

Při učení bez učitele nevyhodnocujeme výstup. Při tomto učení nám výstup není znám. Síť dostává na vstup sadu vzorů, které si sama třídí. Buď si vzory třídí do

skupin a reaguje na typického zástupce, nebo si přizpůsobí topologii vlastnostem vstupu. [1]

3.3.4 Shluková analýza

Shlukování je technika, která se používá ke klasifikaci objektů. Slouží k třídění jednotek do skupin, shluků tak, aby si jednotky náležící do stejné skupiny byly podobnější než objekty ze skupin různých. Shlukovou analýzu je možné provádět jak na množině objektů, z nichž každý musí být popsán prostřednictvím stejného souboru znaků, které má smysl v dané množině sledovat, tak na množině znaků, které jsou charakterizovány prostřednictvím určitého souboru objektů, nositelů těchto znaků. [6]

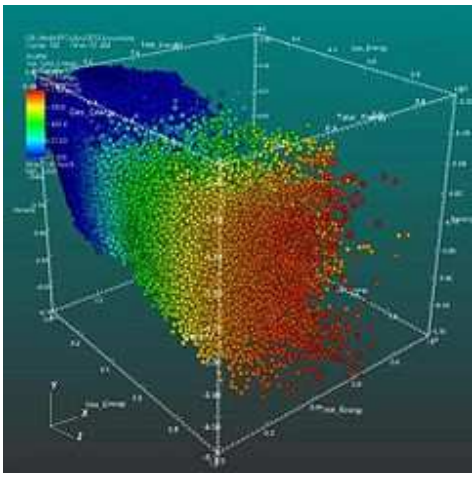
4. Vizualizace dat

Vizualizace dat je způsob objevování a pochopení vzorů v obrovském souboru dat pomocí vizuální interpretace a používá se pro vědeckou analýzu komplexních procesů. Jelikož se nástroje na zobrazení a interpretaci dat stále zdokonalují, umožňuje to manipulaci s daty v reálném čase a výzkumníci je tak mohou zkoumat způsoby, které dřív nebyly možné. Vizualizace dat, tedy jakási směs statistiky, vytěžování dat a jejich grafické interpretace, je nově vznikající obor, který by měl umožnit každému z nás procházet, zobrazit a pochopit složité koncepty a vztahy.[4]

Nedávno vědci objevili, že vidění a porozumění společně umožňuje lidem objevovat nové poznatky s hlubším pohledem vycházejícím ze zpracování z velkého množství dat. Sílu vizualizace, využívat lidské vnímání, nabízí výzvu i

příležitost. Úkolem je, aby se zabránilo vyvozování nesprávných vzorů, což by následně vedlo k nesprávným rozhodnutím a akcím.

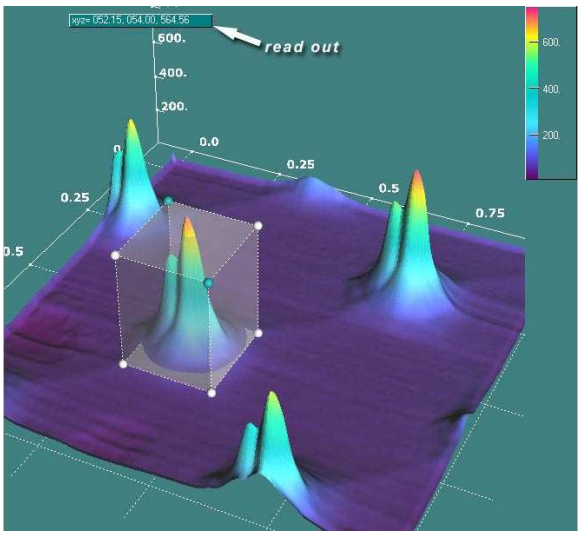
Uveďme si velmi zjednodušený příklad. Máme k dispozici data, které si zobrazíme pomocí 3D rozměrného bodového grafu. Díky vizualizačnímu nástroji uvidíme obrázek č.4 *Scatter-plot.* Rovnou můžeme provést shlukování pomocí vizualizace. Prvky, které budou blízko u sebe můžeme označit a provést shluky na základě vizuálního vjemu. Samozřejmě rovněž je možné provést shlukování formální cestou pomocí výpočtu, ale cesta zobrazení dat má již na první pohled vyšší interpretační schopnost.



Obr.č.4 – Vizualizace Scatter-plot

Technologie vizuálního dataminingu staví na vizuálních a analytických procesech vyvinutých v různých disciplínách, včetně vědecké vizualizace, počítačové grafiky, dolování dat, statistiky a strojového učení s vlastním rozšířením, které zvládá interaktivní zpracování velkého množství souborů vícerozměrných dat.

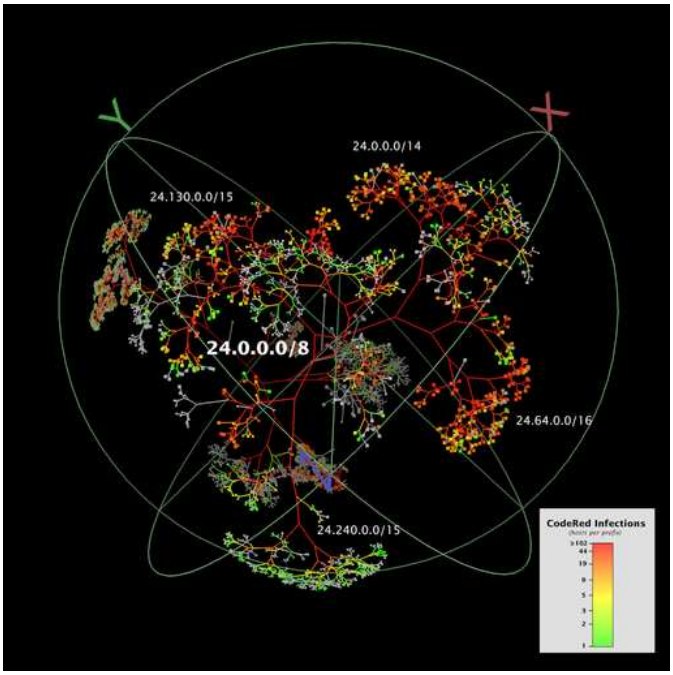
Výpočet, na základě těchto velkých datových souborů a databází vytváří obsah. Vizualizace umožňuje výpočet a jeho obsah přístupný člověku. Proto vizuální dolování dat používá vizualizace k rozšíření dataminingového procesu.



Obr.č.5 – Vizualizace 3D peak

Některé dataminingové techniky a algoritmy jsou obtížné k pochopení a používání. Vizualizace může učinit tyto údaje a následně i výsledky dolování přístupnější, což umožňuje porovnávání a ověřování výsledků. Vizualizace může být také použita k nasměrování dataminingového algoritmu.

Vizualizace na obrázku č.6 zobrazuje počet počítačů infikovaných červem CodeRed během 24 hodin od 19. července 2001, UTC. Počty infikovaných počítačů jsou agregovány podle IP prefixů.



Obr.č. 6 – Vizualizace červ na 24.0.0.0/8 dne 19.7.2001

5. Nástroje pro datamining

Nástrojů pro dolování znalostí z dat je díky dynamickému rozvoji tohoto oboru v současné době k dispozici již poměrně velké množství. Kromě toho, že takto zaměřené softwary obsahují běžně velký počet různých modelů výpočtu, analýz, algoritmů, statistik, snaží se předhánět v uživatelské přívětivosti formou intuitivního uživatelského rozhraní nebo multidimenzionálními vizualizacemi dat a jejich výstupů.

Systémy pro dobývání znalostí nabízejí jak malé firmy vzešlé z akademického prostředí (RuleQuest nebo Dialogis), tak význační producenti statistického software (SAS, SPSS nebo StatSoft). O rostoucím zájmu o dobývání znalostí z

databází svědčí fakt, že mezi výrobce softwaru se zařadily i firmy jako IBM nebo Silicon Graphics. Na druhé straně existují akademické, volně šířené systémy.
K nejznámějším z nich patří systém Weka.

Tab. č.2 - Nástroje pro dobývání znalostí z dat uvádí některé systémy pro dobývání znalostí z dat. Podrobný seznam systémů je možno nalézt např. na http://www.kdnuggets.com.

Systém	Výrobce	URL
CART	Salford Systems	http://www.salford-systems.com
IBM SPSS	SPSS	http://www.spss.com
Enterprise Miner	SAS Institute	http://www.sas.com/software/components/miner.html
Intelligent Miner	IBM	http://www-4.ibm.com/software/data/iminer
Kepler	Dialogis	http://www.dialogis.de
KnowledgeStudio	Angoss	http://www.angoss.com
LISp Miner	VŠE	http://lispminer.vse.cz
MineSet	Silicon Graphics	http://www-europe.sgi.com/software/mineset
See5	RuleQuest Research	http://www.rulequest.com/see5-info.html
Statistica Data Miner	StatSoft	http://www.statsoft.com
Weka	University of Waikato	http://www.cs.waikato.ac.nz/~ml/weka
WizWhy	WizSoft	http://www.wizsoft.com/why.html

Tab.č.2 – Nástroje pro dobývání znalostí z dat

V následujících podkapitolách uvedeme software dvou lídrů na trhu PASW Modeler (Clementine), Enterprise Miner dále pak Statistica Miner a Weka

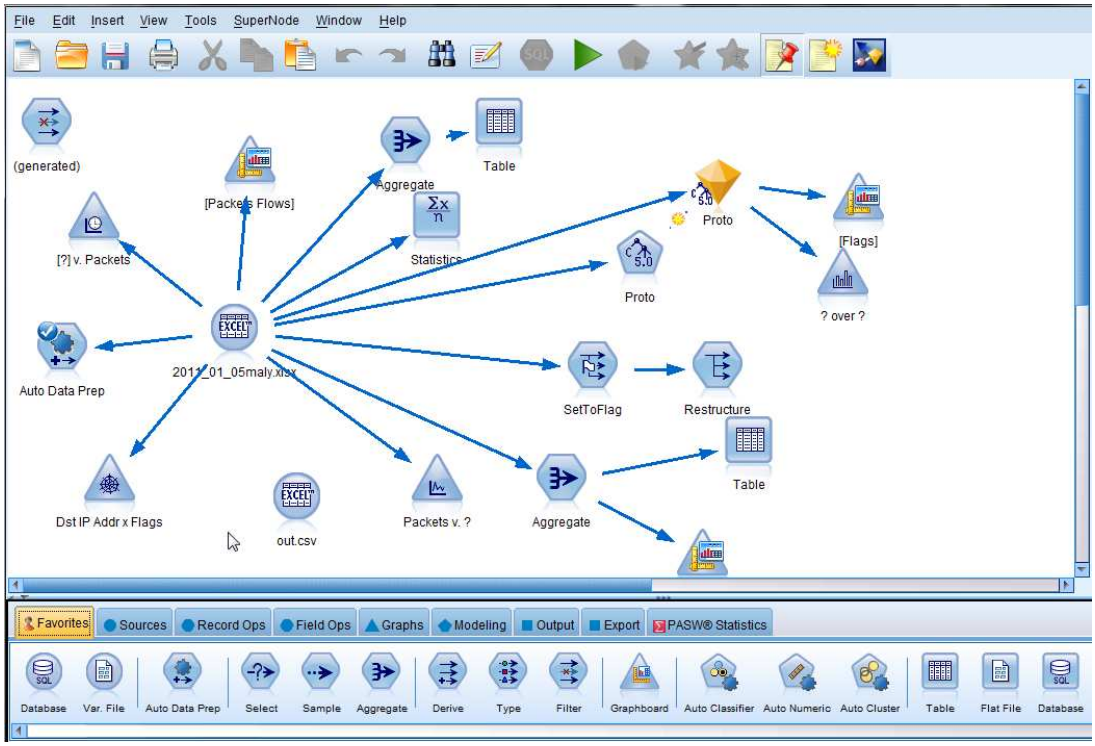
5.1 PASW Modeler (IBM SPSS)

Systém PASW Modeler (dříve Clemetine) vyvinula britská firma Integral Solutions Ltd. v polovině 90.let. K 1.lednu 1999 tuto firmu (a s ní i systém Clementine)

převzal přední výrobce statistického software, firma SPSS. Modeler patří mezi přední komerční systémy pro dobývání znalostí. Systém důsledně vychází z metodologie CRISP-DM, na jejímž vzniku se firma Integral Solutions podílela. Systém nabízí řadu metod pro klasifikační (predikční) i deskriptivní úlohy. Aplikace má velice propracovaný způsob ovládání, tzv. vizuální programování (vizual programming).

Systém nabízí analytikům tzv. Modeler External Module Interface pro přidávání vlastních algoritmů. Tak lze přidávat programy do jednotlivých palet nástrojů. Z implementačního hlediska je třeba zajistit správné začlenění nového programu do sekvence (streamu). Program tedy musí umět číst data a parametry ze sekvence a zapisovat do sekvence své výsledky.

Přenesení provedené analýzy ke koncovým uživatelům usnadňuje tzv. Modeler Solution Publisher, na jehož základě vznikne samostatná aplikace obsahující všechny provedené kroky v sekvenci (jako zdrojové programy v C, SQL příkazy apod.).



Obr.č.7 – Uživatelské rozhraní PASW Modeler

5.2 Enterprise Miner

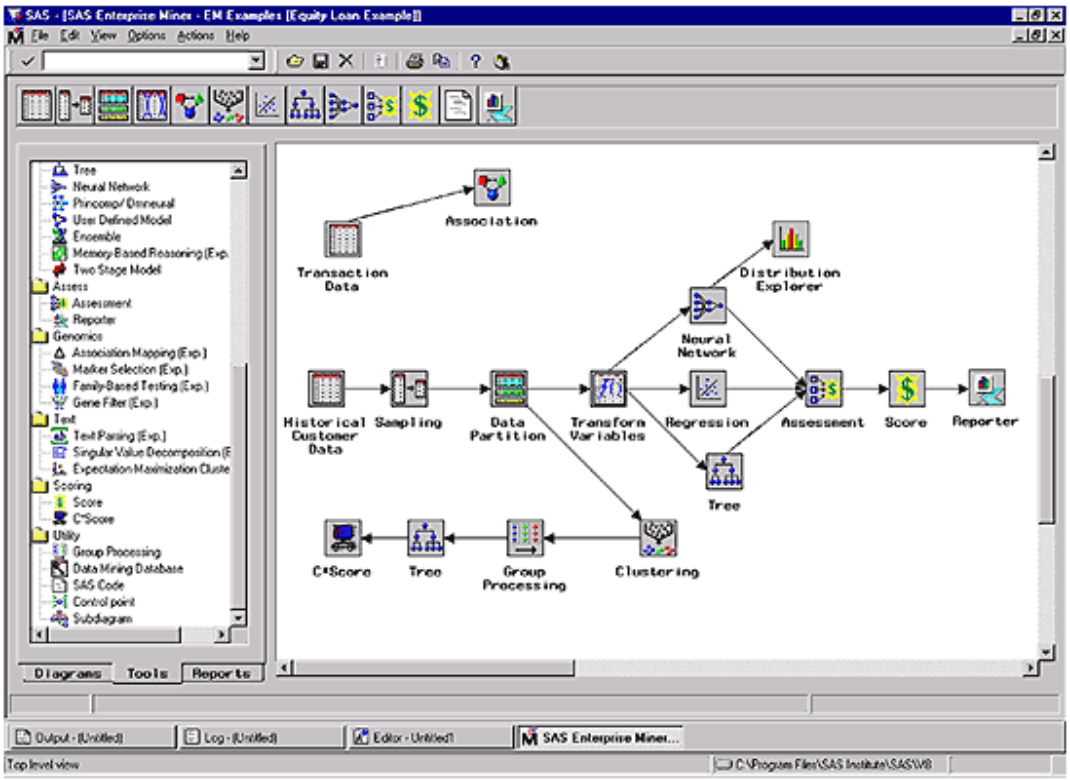
Enterprise Miner je produkt firmy SAS, jednoho z předních producentů statistického software. To je vidět i na algoritmech pro analýzu dat, které systém nabízí. Nejpropracovanější jsou statistické metody, které využívají již implementované procedury. Enterprise Miner použité metody integruje a nabízí uživatelsky příjemnější prostředí než je příkazový jazyk (SAS kód) jinak běžně používaný v „klasickém" SASu.

Enterprise Miner vychází z vlastní metodologie pro dobývání znalostí z databází. Název této metodologie SEMMA je akronym pro jednotlivé prováděné kroky:

- SAMPLE - *vybrání vhodných objektů,*

- EXPLORE - *vizuální explorace a redukce dat*,
- MANIPULATE - *seskupování objektů a hodnot atributů, transformace dat*,
- MODEL - *analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování,*
- ASSESS - *porovnání modelů a interpretace*.

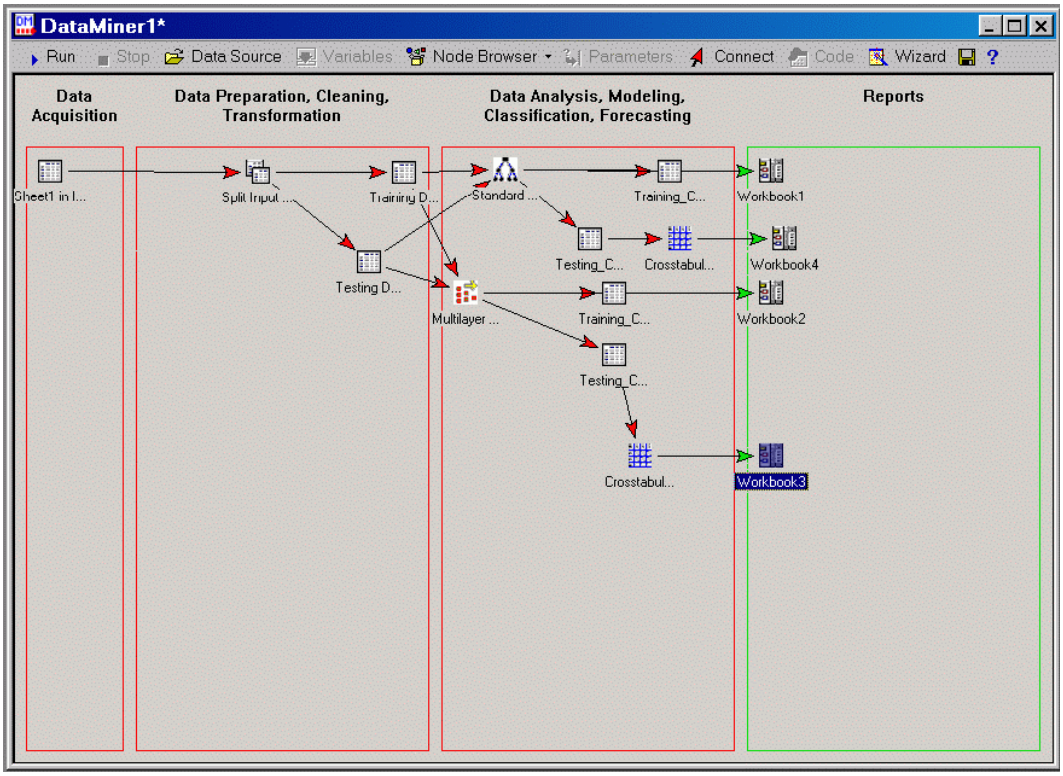
Důraz se klade na snadnou interpretaci výstupů ve formě srozumitelné obchodnímu uživateli. Proces dobývání znalostí pro danou úlohu se definuje („programuje") pomocí procesních diagramů (Process Flow Diagrams). Jde vlastně o analogii k vizuálnímu programování použitém v systému Modeler.



Obr.č.8 – Uživatelské rozhraní SAS Enterprise Miner

5.3 Statistica Data Miner

Statistica Data Miner je dalším příkladem systému pro dobývání znalostí, který vyvinula firma specializovaná na statistické programy. Systém opět pokrývá proces dobývání znalostí počínaje přípravou dat a využitím výsledků konče. Sekvence zpracování dat se vytvoří na pracovní ploše z jednotlivých nástrojů-uzlů.



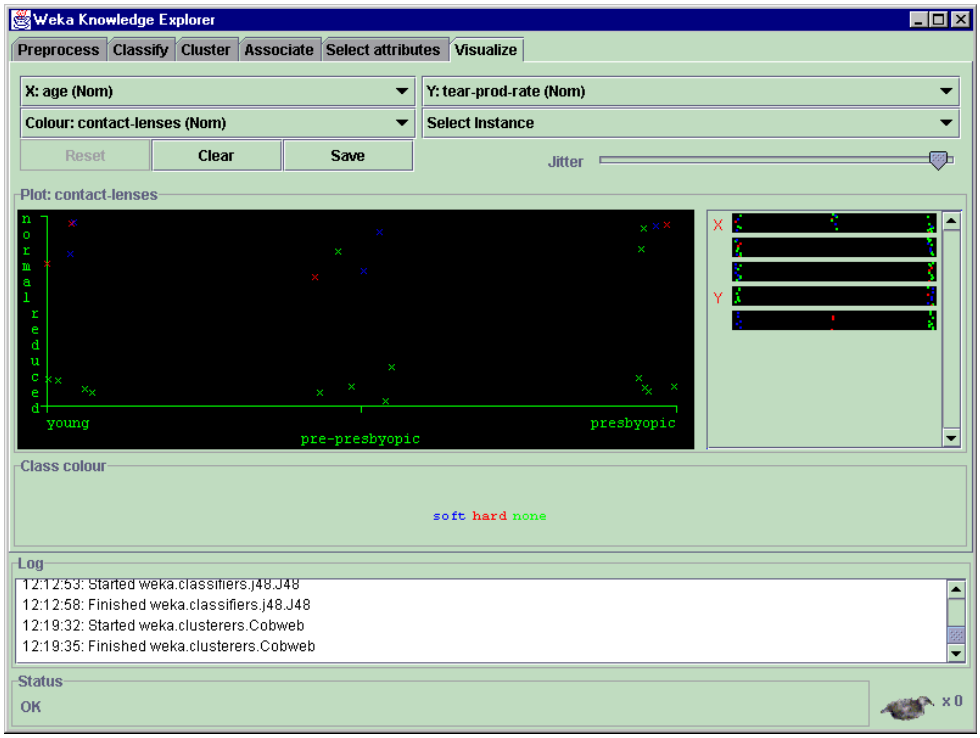
Obr.č.9 – Rozhraní Data Miner

Pro modelování nabízí Data Miner asociační pravidla, klasifikační a regresní stromy, shlukování, neuronové sítě, diskriminační analýzu a regresní metody. Z výsledných modelů (např.rozhodovacích stromů) lze vygenerovat spustitelný kód v C++, Visual Basicu nebo SQL. Visual Basic je ostatně přímo součástí systému. V systému najdeme i prvky OLAP a možnost kombinování modelů (bagging,

boosting, metaučení). Otevřená architektura umožňuje uživatelům přidávat do systému vlastní nástroje (uzly).

5.4 Weka

Weka je systém vyvinutý na universitě Waikato na Novém Zélandě. Přestože se jedná o freeware volně dostupný na Internetu, v ničem si nezadá s komerčními systémy. Weka nabízí celou řadu algoritmů pro učení i předzpracování známých v akademickém světě. K dispozici jsou i možnosti vizualizace a kombinování modelů. Systém je řešen jako knihovna programů v Javě volaných z jednotného (grafického) rozhranní. Většina modelů si ale ponechává původní textový výstup.[7]



Obr.č. 10 – Uživatelské rozhraní Weka - shlukování

6. Zpracování systémových protokolů

V této kapitole jsou zpracovány protokoly získané z provozu síťové sondy. První podkapitola popisuje princip fungování protokolu a sondy NetFlow. Sonda je hardwarové zařízení využívající protokol NetFlow. Pomocí tohoto protokolu je ukládán záznam o provozu sítě. Nad těmito daty bude provedena analýza pomocí dataminingového nástroje Statistica Data Miner uvedeného v kapitole 5.1.

V průběhu zpracování budou použity vybrané metody zmíněné kapitole 3.3 Metody pro datamining, jejichž pomocí bude zjištěna případná existence závislostí souvisejících s provozem a vytížením sítě, detekcí možných skenování sítě nebo případně i přímo útoku.

6.1 NetFlow

6.1.1 Protokol NetFlow

NetFlow je protokol, který vyvinula společnost Cisco. Jeho hlavní využití spočívá v možnosti monitorování síťového provozu na základě IP toků, které poskytují podrobný pohled do provozu na síti v reálném čase. Tvoří tak důležitou a nepostradatelnou součást zabezpečení každé počítačové sítě. K dispozici je specifikace tohoto protokolu v RFC 3954 (poslední verze 9) [10]. Díky dostupnosti specifikace je protokol implementován i na jiných platformách než Cisco např. Juniper (J-flow) nebo distribuce GNU/Linux. Vzniklo několik verzí a mezi v současné době nejpoužívanější patří verze 5. Zároveň se velmi rozšiřuje verze 9, která má podporu pro IPv6. Ukázka záznamů z NetFlow protokolu je uveden v příloze A.

Na základě protokolu NetFlow v9 vznikl v nedávné době nový IETF standard Internet Protocol Flow Information eXport (IPFIX). Lze očekávat, že se v blízké budoucnosti pravděpodobně stane průmyslovým standardem.

NetFlow pracuje s několika pojmy, které je třeba vysvětlit.[11]
IP TOK - *sekvence paketů se shodnou pěticí údajů: cílová/zdrojová IP adresa, cílový/zdrojový port a číslo protokolu, typicky TCP a UDP.*

NETFLOW ZÁZNAM – *detailnější informace ke konkrétnímu IP toku např. délka toku, počet přenesených bajtů, paketů, atd)*

EXPORTÉR - *zařízení, které monitoruje procházející provoz a vytváří z nich IP toky. Informace jsou z exportéru odesílány do kolektoru (např. router).*

NETFLOW KOLEKTOR - *do kolektoru jsou dedikovanou linkou posílány pakety z exportéru. Může sbírat data z více exportérů. Data jsou pak ukládány na disk.*

6.1.2 Nástroje pro zpracování NetFlow

Nástrojů pro NetFlow je celá řada. Dostupné je velké množství volně šiřitelných nástrojů. Většina těchto open-source nástrojů však podporuje především verzi 5, novější verze pro ně není zatím dostupná. Pro zpracování dat v této diplomové práci bylo použito nástrojů NFDUMP tools. Ty jsou distribuovány pod BSD licencí. Jak už název napovídá, jedná se o skupinu několika nástrojů. Všechny tyto nástroje podporují NetFlow verze 5, 7 a 9 a jsou určeny pro Unixové systémy. Pro získání a vizualizaci dat byly použity následující nástroje.

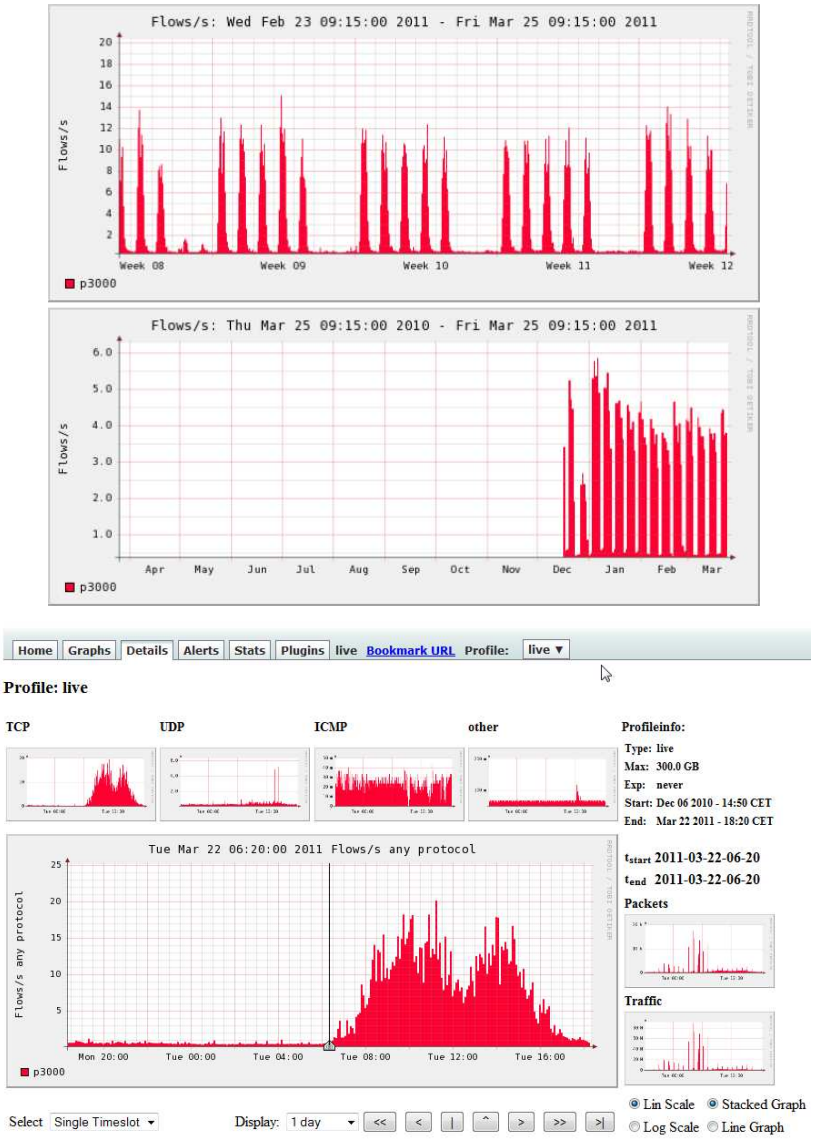
NFDUMP (NetFlowDump)

Jedná se o open-source nástroj, který slouží ke shromažďování a zpracování NetFlow záznamů na příkazovém řádku. Čte data ze souborů nasbíraných kolektorem. Soubory obsahují informace o provozu uložených za určitý časový interval, typicky 5 min. Kromě čtení dat dokáže nfdump i zpracovávat top N statistiky.[12]

NFSEN (NetFlow Senzor)

Nfsen je grafické rozhraní pro NetFlow nástroje. Prostředí webového rozhraní usnadňuje zobrazení a procházení NetFlow dat nasbíraných exportérem. Umožňuje všechny praktické výhody jako nástroj nfdump na příkazové řádce, ale s výhodou vykreslení grafů za různé typy provozů, jednoduchého ovládání, volení vlastních úseků k zobrazení. Další výhodou je možnost nastavení vlastních profilů a tak rychleji procházet specifická data. Grafické rozhraní umožňuje i nastavení výstrah tzv. alertů. Tyto alerty mohou upozornit v případě podezřelého provozu např. odesláním mailové zprávy.[13]

Na obrázcích viz níže je ukázka grafického rozhraní nfsen. Jednotlivé grafy odpovídají různých časovým úsekům (shora) den, týden, měsíc, rok. Grafy mohou zobrazovat počet toků, paketů či bajtů v závislosti na čase. V případě podezřelého provozu lze vybrat kritický časový úsek a podrobit jej detailnější analýze i za použití pole k vyplnění požadovaného filtru.



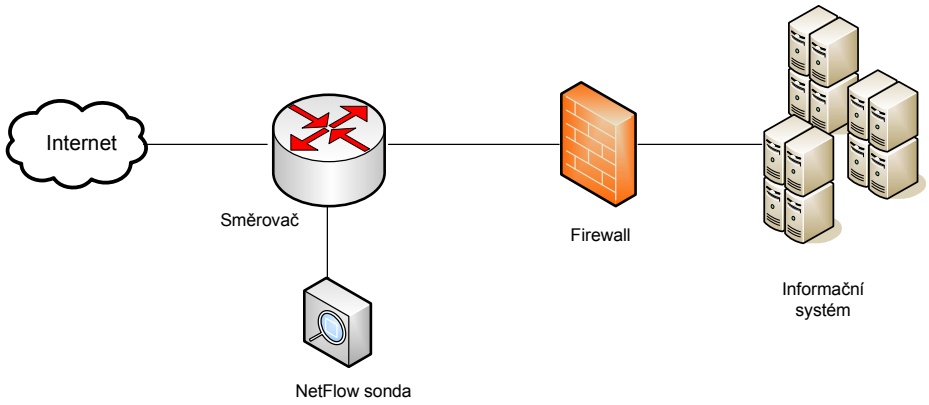
Obr.č.11 – Ukázka rozhraní nfsen

6.1.3 Sonda NetFlow

NetFlow sonda je pasivní hardwarové zařízení, které monitoruje IP provoz na síti a vytváří nad ním statistiky. Pod pojmem pasivní si představme, že data pouze

monitoruje a nijak do nich nezasahuje. Sonda se typicky umisťuje na vstupní a výstupní body sítě, kritická místa či linky s největším přenosem dat. Vlastní připojení do linky je možné realizovat pomocí mirror portu směrovače či přepínače, nebo přímým vložením do linky s pomocí externího či vestavěného Ethernet rozbočovače (TAPu).

Na obrázku níže je zobrazené schéma nasazení sondy pro účely této diplomové práce.



Obr.č.12 – Zjednodušené schéma nasazení sondy

6.2 Bezpečnostní incidenty v NetFlow záznamech

6.2.1 Typy bezpečnostních incidentů

Pro účely zpracování NetFlow dat budou bezpečnostní incidenty rozděleny do určitých skupin, tyto budou popsány pomocí pravidel (jejich typických vlastností), aby bylo možné je lépe identifikovat ve velkém objemu dat.

SKENOVÁNÍ SÍTĚ

Vlastnímu útoku na vybranou síť předchází nejprve její prozkoumávání. Aby útočník zjistil co nejvíce potřebných informací provádí tzv. skenování sítě. Skenování je tedy činnost, při které útočník zjišťuje, které porty jsou v jakém stavu. Porty mohou být v zásadě v následujících stavech [14]:

- OTEVŘENÝ – port je otevřený, běží na něm síťová služba a s portem lze navázat spojení
- ZAVŘENÝ – port je uzavřený, žádná služba na něm neběží. Při pokusu o připojení zašle TCP zpět paket s příznaky RST a ACK. U UDP portů je zpět zaslán ICMP paket typu 3 a kódu 3, což znamená port nedosažitelný.
- FILTROVANÝ – při pokusu o připojení nepřijde žádná odpověď. Znamená to, že počítač je chráněn firewallem.

Toto dělení však nemusí být jednoznačné. S pomocí nastavení filtrovacích pravidel může port záměrně zobrazovat jiný stav nebo předstírat provoz falešné služby s cílem nalákat útočníka.

Dívejme se však na problematiku z pohledu správce, který má alespoň základní přehled o tom, co mu na síti běží. Cílem je provést analýzu, zda na sledované síti někdo prováděl skenování, nikoliv jaké porty zobrazují falešný stav. O těch by správce měl mít přehled.

Uvádím vybrané skenovací techniky [15] [17]

VERTIKÁLNÍ SKENOVÁNÍ – *jedná se o skenování více portů na jednom cílovém IP.*

TCP

1. **Skenování SYN**

Tento typ skenování nedokončí TCP spojení. Předností je, že většina aplikací loguje pouze navázaná spojení.

Technika: Pro zjištění stavu portu se zašle paket s příznakem SYN. Je-li port uzavřený dostaneme odpověď RST, v případě, že je port otevřený v odpovědi získáme příznak SYN a ACK. Následně by měl být námi odeslán paket s příznakem ACK a došlo by k navázání TCP spojení. Tento paket se ale již neodešle a spojení je ukončeno bez záznamu naší IP adresy v logovacích souborech.

2. **Skenování CONNECT**

Technika je zde stejná jako u SYN skenování, ale dochází k úplnému navázání TCP spojení, tedy i k záznamu do logu.

Technika: Odešle se paket s příznakem SYN, v případě uzavřeného portu dostáváme RST. V případě otevřeného portu dostáváme SYN a ACK, odpovídáme ACK a naváže se TCP spojení, v zápětí se spojení ukončí zasláním paketu s příznakem RST.

Následující skeny (body 3 - 6) se používají podobné techniky. Při odeslání echo request, požadavku, se posílá pouze jeden paket a jako echo reply, odpověď, se očekává také pouze jeden paket. Níže uvedené typy skenů využívají definici RFC 793 o TCP [16], které specifikují vlastnost TCP, že je-li port ve stavu uzavřený, odešle se paket s příznakem RST. Je-li port otevřený reaguje pouze na zaslaná pakety s příznakem SYN,RST, ACK, ostatní zahodí.

Takto reaguje však pouze v případě operačních systémů (OS) jiných než Windows, kde je použita jiná implementace TCP. U OS Windows je tedy v případě

otevřeného či zavřeného portu reakce RST. Tím jsou skeny nepoužitelné na prostředí Windows, dochází tím ale k jednoduché identifikaci OS cílové stanice, jelikož pokud neobdržíme příznak RST, lze předpokládat, že se nejedná o OS Windows.

3. **Skenování FIN**

Využívá se příznaku FIN, který slouží k ukončení navázaného TCP spojení.

Technika: Odešle se paket s příznakem FIN, je-li port otevřený není zaslána žádná odpověď, v případě uzavřeného portu obdržíme RST.

4. **Skenování NULL**

Paket typu NULL je specifický tím, že v normálním provozu se nevyskytuje. Všechny příznaky má totiž vypnuté.

Technika: Odešle se paket s příznakem "00000000", otevřený port neodpoví, zavřený zašle RST.

5. **Skenování Xmas**

Paket typu Xmas má nastaven příznak na FIN, URG a PUSH.

Technika: Odešle se paket s příznakem FIN, URG, PUSH. Otevřený port neodpoví, zavřený zašle RST.

6. **Skenování ACK**

V tomto případě slouží sken pro zjištění, zda je port filtrován, nikoliv otevřen.

Technika: Odešle se paket s příznakem ACK, filtrovaný port neodpoví, zavřený i otevřený port zašle RST.

UDP

U protokolu UDP jde pouze o jeden druh skenovací techniky. Nevyužívá se zde komunikace na bázi příznaků, posílá se UDP paket na konkrétní port a podle typu odpovědi můžeme usuzovat o stavu portu.

Technika: Odešle se paket s hlavičkou, která neobsahuje žádná data. Pokud obdržíme zpět paket obsahující nějaká data, je port otevřený. Je-li port zavřený obdržíme ICMP paket typu 3, kódu 3. Je-li port filtrován vrátí se paket ICMP jiného kódu (1,2,9,10,13).

HORIZONTÁLNÍ SKENOVÁNÍ – *při tomto typu skenování zkoumá útočník jeden port na více IP, počítačích, spadající do jedné podsítě. Účelem je zjistit pod jakými IP adresami jsou možné cílové stanice pro útok.*

1. Skenování PING

Při této technice se využívá se klasického pingu. Jedná se o nejjednodušší nástroj pří skenování.

Technika: Odešle se ICMP Echo Request na cílovou IP adresu. Pokud je stanice dostupná odpoví ICMP Echo Reply, není-li dostupná nebo je-li cílová stanice filtrovaná neobdržíme žádnou odpověď. [17]

2. Skenování TCP SYN

Jde o stejné použití jako u vertikálního SYN skenu s tím rozdílem, že se skenuje více počítačů najednou.[17]

BLOKOVÉ SKENOVÁNÍ – *jedná se o kombinaci obou předchozích*

DOS ÚTOKY

Denial of Service (česky odmítnutí služby) nebo též Distributed Denial of Service (česky distribuované odmítnutí služby) je technika útoku na internetové služby nebo stránky, při níž dochází na síti k přehlcení požadavky a pádu nebo minimálně nefunkčnosti a nedostupnosti pro ostatní uživatele.
Cíle takového útoku jsou v zásadě dva:

- Vnucení opakovaného resetu cílového počítače
- Narušení komunikace mezi serverem a obětí tak, aby jejich komunikace byla buď zcela nemožná, nebo alespoň velmi pomalá.

United States Computer Emergency Readiness Team definuje příznaky DDoS útoku takto:

- Neobvyklé zpomalení služby (při otevírání souborů nebo prostém přístupu).
- Celková nedostupnost části nebo celých stránek.
- Nemožnost se ke stránkám připojit.
- Extrémní nárůst obdrženého spamu.

Splnění některých podmínek ale ještě neznamená DDoS útok, může jít o prostý výpadek zaviněný hardwarem nebo softwarem samotného serveru bez cizího zavinění.

DoS útoky se dělí podle způsobu útoku do dvou skupin:

- První skupina útoků využívá známých bezpečnostních chyb pro shození nebo alespoň zpomalení služby nebo celého serveru.
- Druhá skupina využívá k útoku tzv. zaplavování - flood. To spočívá v odesílání uměle generovaných dat na cílový počítač. To způsobí vyčerpání

přenosového pásma nebo systémových prostředků serveru, což může mít za následek zpomalení nebo úplné odstavení služby nebo serveru [19].

Níže jsou uvedeny vybrané typy DOS útoků, které se pokusíme odchytit pomocí NetFlow záznamů.

1. **Útok PING FLOOD**

   Úspěch tohoto útoku stojí na předpokladu, že útočník disponuje rychlejší konektivitou než cílový počítač. Provedení útoku spočívá v odesílání dat na adresu napadeného počítače, čímž dojde k vyčerpání přenosové kapacity směrem k oběti. To způsobí nedostupnost tohoto stroje pro ostatní uživatele. Pro realizaci není třeba žádného speciálního softwaru, stačí využít běžných systémových nástrojů jako je ping.

2. **Útok SYN FLOOD**

   Jedná se zaplavení cílového stroje TCP pakety se SYN příznakem a upravenou hlavičkou odesílatele. Takový paket je přijat jako požadavek o spojení. Oběť odešle TCP paket s příznaky SYN a ACK a čeká na odpověď TCP procedury (TCP paket s ACK příznakem). Tato odpověď nedorazí, protože upravená hlavička neukazuje na počítač, který by požadoval navázání spojení. Z důvodu omezeného počtu možných připojení takto polootevřené spojení po určitou dobu blokuje jiné legitimní žádosti o spojení, nebo vyčerpá velké množství systémových prostředků na napadeném serveru.

PROLOMENÍ HESLA SSH

Je anomálie způsobená pokusem o prolomení SSH hesla hrubou silou. Tato anomálie se vyznačuje velkým počtem neúspěšných pokusů o autentizaci na

cílovém stroji. Služba SSH je typicky přístupná na portu 22 a po několikátém neúspěšném pokusu o autentizaci dojde většinou k ukončení spojení ze strany SSH serveru. Z toho plyne, že pokus o prolomení SSH hesla se na provozu ukáže jako velké množství po sobě následujících připojení na port 22, které jsou ukončené z cílové adresy [20].

6.3 Příprava NetFlow dat ke zpracování

Zpracování velkých a stále narůstajících objemů dat řeší dnes každý podnik. Objem NetFlow záznamů je dán velikostí sítě, na které je sonda připojena.
Sledovaný provoz se týkal 45 ekonomických jednotek. Sběr dat probíhal od 17.12.2010. Přibližná velikost dat exportovaných za jeden den činí kolem 50 MB dat. Exportovaná data za měsíc obsahovala cca 1,09 GB dat. Zpracovávat takové množství dat je hardwarově i časově velice náročné.

6.3.1 Zpracování získaných NetFlow záznamů

Získaná NetFlow data jsou plnohodnotná tzn.není provedeno vzorkování. Vzorkování paketů vybere do sběru záznamů každý n-tý paket podle nastavení v konfiguraci. Vzorkování se běžně používá pro snížení objemu a provedení statistik nad těmito daty. Tímto krokem dochází ke snížení objemu analyzovaných dat, ale za cenu ztráty cenných informací, které potřebuji použít pro detekci anomálií. Použita jsou tedy data kompletní.

V této tabulce je uvedená výpočetní kapacita dostupná k vypracování této diplomové práce.

Procesor	RAM
Intel(R) Core (TM)Core 2CPU 2,8GHz	4 GB

Původní záměr byl analyzovat a porovnávat sumární data za měsíc a následně pak měsíce mezi sebou. Vzhledem však k dostupné výpočetní kapacitě a rozsahu diplomové práce bylo nezbytné v tomto směru změnit rozsah analyzovaných dat. K analýze bude tedy použito několik náhodně vybraných jednodenních záznamů za měsíc leden.

Přesto se pořád potýkáme s velkým objemem dat. Tento objem dat by mohl být snížen použitím vhodného filtrování. Toto filtrování umožňuje provést už syntaxe pro nástroj nfdump. Pomocí znalostí získaných v kapitole 6.2.1 (Typy bezpečnostních incidentů) budou sestaveny jednotlivé filtry.

6.3.2 Syntaxe filtru pro NetFlow

Tato kapitola shrnuje význam výrazů používaných pro filtry na NetFlow sondách. Informace jsou čerpány z uživatelské příručky vyhotovené firmou
INVEA-TECH a.s.

Popis filtru není co do své délky omezen. Všechna klíčová slova mohou
být psána libovolně malými i velkými písmeny, pokud není řečeno jinak.
Každý filtr sestává z jednoho nebo více výrazů **expr**. Výrazy lze spojovat následujícím způsobem:

```
Filter = expr, expr and expr, expr or expr, not expr, ( expr ),
not ( expr )
```

Výraz **expr** může být tvořen některým z těchto primitiv:

```
Any
any Používá se jako dummy filtr. Pomocí 'not any' blokujte všechny toky.

Verze protokolu
inet nebo ipv4 pro IPv4 a inet6 nebo ipv6 pro Ipv6

Protokol
proto <protocol>, kde protocol může být některý ze známých protokolů TCP,
UDP, ICMP, GRE, AH atd., nebo proto num, kde num udává číslo protokolu.
```

IP adresa
[SourceDestination] **IP a.b.c.d** nebo **HOST a.b.c.d**, kde a.b.c.d je libovolná platná IPv4 nebo IPv6 IP adresa.
[SourceDestination]
určuje, zda se jedná o cílovou, nebo zdrojovou IP adresu. Může nabývat hodnot **SRC, DST** nebo libovolné kombinace **SRC and|or DST**. Vynechání volby SourceDestination odpovídá volbě **SRC or DST**.

Adresa sítě
[SourceDestination] **NET a.b.c.d m.n.r.s** pro dvojici IPv6 síť a maska sítě
[SourceDestination] **NET net/num**, kde net je platná adresa IPv4 nebo IPv6 sítě a num je počet bitů masky. Adresy sítí mohou být zkracovány (např. 172.16/16) v případech, že nedojde k víceznačnosti.

Port
[SourceDestination] **PORT** [comp] **num**, kde num je platné číslo portu. Je-li vynechán operátor comp, pak se předpokládá operátor '='.

Síťové rozhraní
[inout] **IF** num, kde num udává číslo síťového rozhraní.
[inout] udává směr rozhraní a může nabývat hodnot **IN** nebo **OUT**.

TCP příznaky
flags tcpflags, kde tcpflags je kombinace hodnot:
A ACK.
S SYN.
F FIN.
R Reset.
P Push.
U Urgent.
X Všechny příznaky zapnuty.
Na pořadí příznaků nezáleží. Příznaky, které nejsou zmíněny, se ve filtru neuvažují (don't care). Tzn. pokud si přejete zobrazit toky, které mají nastavený pouze příznak SYN, použijte syntaxi '**flags S and not flags AFRPU**'.

Type of Service
tos value, kde value je číselná hodnota 0..255.

Počet paketů
packets [comp] **num**
Udává počet paketů v NetFlow záznamu.

Počet bytů
bytes [comp] **num**
Udává počet bytů v NetFlow záznamu.

Počet paketů za sekundu
pps [comp] **num** [scale] udává počet paketů za sekundu v toku.
[scale] je násobek měrné jednotky. Může nabývat hodnot (Kilo) **k**, (Mega) **m**,(giga) **g**, (Terra) **t**. Násobek je roven 1024.

```
[comp] je komparátor.
Podporovány jsou následující komparátory:
=, ==, >, <, EQ, LT, GT. Je-li komparátor vynechán, předpokládá se
operátor'='.

Trvání toku
duration [comp] num udává dobu trvání toku v milisekundách.
Počet bitů za sekundu
bps [comp] num [scale] udává počet bitů za sekundu v toku.

Počet bajtů na paket
bpp [comp] num [scale] udává počet bajtů na paket v toku. [21]
```

6.4 Analýza získaných dat

Pro základní přehled jsou zde uvedeny sumární statistiky za jednotlivé vybrané dny, které budou dále analyzovány. Počet flows odpovídá jednotlivých řádkům v souborech, které byly zpracovány. Znamená to tedy, že získané statistiky vycházely ze základního souboru (resp. souborů za vybrané dny), které byly tvořeny 2 069 287 řádky.

Date	Time window	Total flows	Total bytes (G)	Total packets (M)	Avg bps (M)	Avg pps	Avg bpp
5.1.2011	2011-01-04 23:59:22 - 2011-01-05 23:59:52	462 807	42,2	92,2	4,0	1 118	468
7.1.2011	2011-01-10 23:59:25 - 2011-01-07 23:59:00	423 220	57,4	113,0	5,4	1 371	519
11.1.2011	2011-01-10 23:59:39 - 2011-01-11 23:59:26	433 621	51,6	104,5	4,9	1 268	505
17.1.2011	2011-01-16 23:59:28 - 2011-01-17 23:59:00	395 987	39	84,1	3,7	1 020	474
27.1.2011	2011-01-26 23:59:08 - 2011-01-27 23:59:56	353 652	37	80,1	3,5	972	472
Total	-	2 069 287	227,2	473,9	-	-	-
Avg	-	413 857	45	95	4,3	1 149,8	487,6

Tab. č.3 – Přehled základních statistik vybraných dat

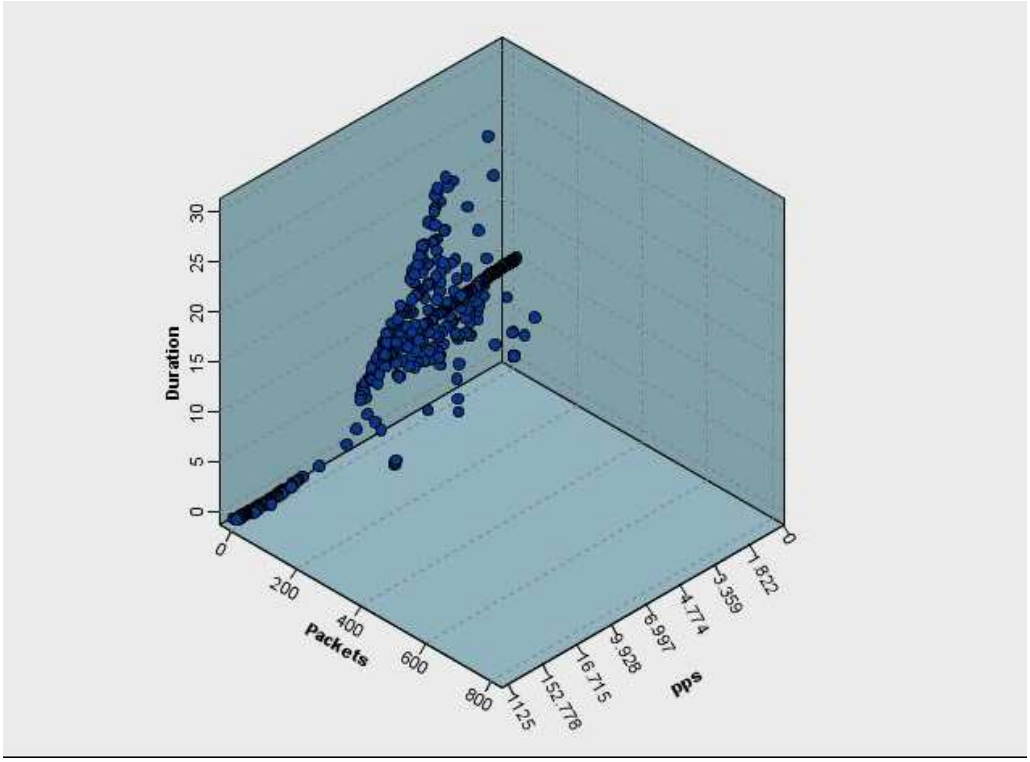
V následující tabulce je znázorněno průměrné množství unikátních cílových a zdrojových IP adres za vybrané dny. Nepříliš vysoké počty unikátních IP adres v datech jsou způsobeny především tím, že systémy jsou přístupné určité skupině

potencionálních uživatelů. Tudíž sledovaná síť není otevřena do celého internetu, ale pouze určitému adresnímu IP prostoru.

	5.1.2011	7.1.2011	11.1.2011	17.1.2011	27.1.2011
SRC IP	1099	1114	1249	902	913
DST IP	682	624	633	523	510

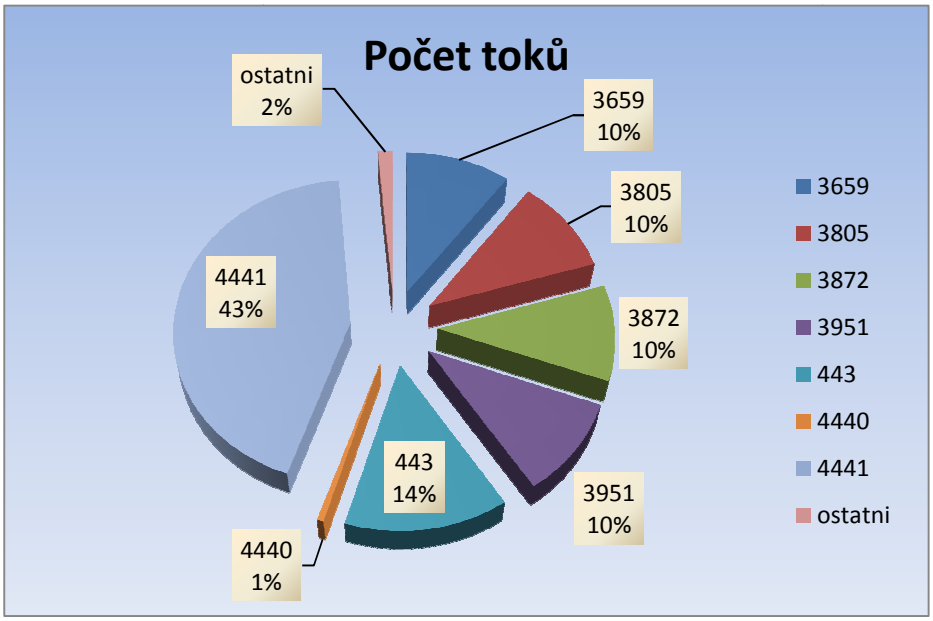
Tab.č.4 – Počty unikátních IP adres za vybrané dny

Graf zobrazuje závislost trvání toku, počtu paketů a přenesených paketů za senkundu.



Obr.č.13 – Graf závislosti délky toku, počtu paketů, přenesených paketů

Při pohledu na toky z hlediska portů, nalezneme následující použití služeb.



Obr.č.14 – Graf využití portů

Cílový port	Packets_Sum	Bytes_Sum	Počet toků
3659	2 754	139 204	2 707
3805	2 870	166 978	2 701
3872	2 761	140 074	2 700
3951	2 730	136 195	2 698
443	211 334	34 917 312	3 734
4440	2 384	200 000	146
4441	20 538	2 288 474	11 509
ostatni	26 555	7 974 101	364
celkem	316 599	92 058 147	4 578

Tab.č.5 – Poměr toků, přenesených paketů, bajtů vzhledem k portům

6.5 Implementace získaných znalostí

6.5.1 Detekce bezpečnostních incidentů v NetFlow

V této kapitole je popsána tvorba a konkrétní syntaxe filtrů pro NetFlow záznamy. Tyto filtry mají za účel podchytit vybrané síťové anomálie uvedené v kapitole 6.2.1 (Typy bezpečnostních incidentů). Při přepisu do syntaxe NetFlow bude čerpáno z technik uvedených v kapitole viz výše.

SKENOVÁNÍ

VERTIKÁLNÍ SKENOVÁNÍ

TCP

1. **Skenování SYN**
2. **Skenování CONNECT**

U těchto dvou typů skenování je postup stejný, rozdíl je v tom, že v případě skenování CONNECT dochází k úplnému navázání spojení. Oba skeny reagují stejným způsobem v případě uzavřených portů.

Syntaxe:

Identifikace potencionálních skenů - fáze požadavku:

```
proto TCP and flags S and not flags FPU and packets < 4
```

Dále je třeba indentifikovat stejné dvojice zdrojových a cílových IP adres. U těchto toků použijeme funkci agregace pro zjištění počtu toků a unikátních cílových portů mezi těmito IP adresami. Zjistíme nejvyšší počet dosažených toků za sekundu.
Čím vyšší bude hodnota dosažených toků, tím pravděpodobněji se jedná o sken. Stanovení této hodnoty, tedy určení jde-li o sken či nikoliv není snadné. Budeme-li uvažovat vyšši hodnoty, identifikujeme takto nižší počty skenů, ale zároveň i nižší

počty falešných poplachů. Proto je na místě ověřit u těchto toků i komunikaci v opačném směru.

Identifikace potencionálních skenů - fáze odpovědi:
`proto TCP and flags AR and not flags SFPU`
v tomto případě se jedná o skeny uzavřených portů. Dále ještě filtr
`proto TCP and flags AS and not flags FRPU`

3. Skenování FIN

Zde je podle techniky v kapitole 6.2.1 (Typy bezpečnostních incidentů) syntaxe filtru nejjednodušší. Hledáme v toku pouze jeden paket s příznakem FIN. Takto zjištěné toky jsou natolik nestandardní, že je lze považovat za skenování.

Syntaxe:
`proto TCP and flags F and not flags ASRPU and packets < 2`

4. Skenování NULL

Rovněž tento typ skenování je v běžném provozu dobře identifikovatelný, jelikož nemá nastavený žádný příznak. Výskyt takových toků můžeme považovat za skenování.

Syntaxe:
`proto TCP and not flags ASFRPU`

5. Skenování Xmas

V tomto případě je identifikace také jednoduchá, jelikož hledáme toky, které mají jeden paket a nastaveny nestandardní kombinaci příznaků FIN, PUSH a URGENT. Zjištěné toky uvažujeme jako skenování.

Syntaxe:

```
proto TCP and flags FPU and not flags ASR
```

UDP

Pro zjištění toků, které mohou být potencionální skenování použijeme tuto syntaxi.

Syntaxe:

Identifikace potencionálních skenů - fáze požadavku:

```
proto UDP and packets < 2
```

Opět je třeba indentifikovat stejné dvojice zdrojových a cílových IP adres. U těchto toků znovu použijeme funkci agregace pro zjištění počtu toků a unikátních cílových portů mezi těmito IP adresami. Zjistíme nejvyšší počet dosažených toků za sekundu.

A opakovaně platí, čím vyšší bude hodnota dosažených toků, tím pravděpodobněji se jedná o sken. Stanovení této hodnoty, tedy určení jde-li o sken či nikoliv není snadné. Budeme-li uvažovat vyšši hodnoty, identifikujeme takto nižší počty skenů, ale zároveň i nižší počty falešných poplachů. Proto je na místě ověřit u těchto toků i komunikaci v opačném směru.

Identifikace potencionálních skenů - fáze odpovědi:

```
proto ICMP and ICMP-TYPE 3 and ICMP-CODE 3
```

a v případě otevřeného portu

```
proto UDP
```

HORIZONTÁLNÍ SKENOVÁNÍ

Vycházíme ze situace, že horizontální skenování skenuje alepoň 2 IP adresy.

1. **Skenování PING**

Pro zjištění toků, které mohou být potencionální skenování použijeme tuto syntaxi.

Syntaxe:

```
Proto ICMP and ICMP-TYPE 8 and packets < 2
```

Dále je třeba indentifikovat toky se stejnou zdrojovou IP adresou. U těchto toků potřebujeme najít toky, které se liší v cílové IP adrese, aby jsme potrvdili předpoklad o horizontálním skenování. Zjistíme nejvyšší počet dosažených toků za sekundu. Čím vyšší bude hodnota dosažených toků s jedním paketem, tím pravděpodobněji se jedná o sken. Velký počet toků s jedním paketem na různé cílové adresy s vysokým počtem toků za sekundu poukazuje na skenování.

2. **Skenování TCP SYN**

Pro zjištění toků, které mohou být potencionální skenování použijeme tuto syntaxi.

Syntaxe:

```
proto TCP and flags S and not flags FPU and packets < 3 and
bpp < 100
```

Dále hledáme toky se stejnou zdrojovou adresou. Toky se musí lišit alespoň ve 2 cílových IP adresách. Určíme počet toků, počet unikátních cílových portů a IP adres. Zjistíme nejvyšší počet toků za sekundu. Opět platí vyšší počet toků za sekundu, pravděpodobněji se jedná o skenování.

DOS ÚTOKY

1. **Útok PING FLOOD**

Pro nalezení toků, jež by mohly identifikovat tento typ útoku je následující syntaxe.

Syntaxe:
Identifikace potencionálních skenů - fáze požadavku:
`proto ICMP and ICMP-TYPE 8 and ICMP-CODE 0`

Dále hledáme toky se stejnou dvojicí zdrojové a cílové adresy. Pro ověření, zda se nejedná o odpověď na ICMP, najdeme toky splňující následující filtr.

Identifikace potencionálních skenů - fáze odpovědi:
`proto ICMP and ICMP-TYPE 8 and packets < 2`

Je-li tato množina prázdná a počet toků za sekundu přesáhl měrnou hranici v daných jednotkách (k, M,G), je pravděpodobné, že se jedná o DOS útok typu PING FLOOD.

2. Útok SYN FLOOD

Pro nalezení toků, které by mohly identifikovat tento typ útoku je následující syntaxe.

Syntaxe:
`proto TCP and flags S and not flags AFRPU`

Hledáme takové toky, které se shodují ve zdrojové a cílové adrese a cílovém portu. Zjistíme počet odpovídajících toků a nejvyšší počet toků za sekundu. Obsahují-li nalezené záznamy velké množství toků a počet toků přesáhl měrnou hranici v daných jednotkách (k, M, G, T), je pravděpodobné, že se jedná o útok typu SYN FOOD.

PROLOMENÍ HESLA SSH

Pro nalezení toků, jež by mohly identifikovat pokusy o prolomení hesla SSH je syntaxe následující.

Syntaxe:

Identifikace potencionálních pokusů o prolomení hesla - fáze požadavku:

```
proto TCP and packets < 20 and dst port 22 and flags F
```

Hledáme takové toky, které mají stejné dvojice zdrojové a cílové IP adresy nad stanovenou hranici pokusů a připojení do SSH.

Identifikace potencionálních pokusů o prolomení hesla - fáze odpovědi:

```
proto TCP and packets < 20 and src port 22 and flags R
```

Tento filtr zobrazí pokusy o spojení ukončené serverem.

6.5.2 Aplikace dataminingových metod

Ve sledovaném provozu byla detekována anomálie. Tato anomálie se po vyfiltrování záznamů projevuje na jedné ekonomické jednotce. Více v kapitole 6.6.1 (Detekce anomálie).

Protože zpracování celkových dat, byť i za jeden den, je na tvorbu statistik a aplikaci dataminingových metod s dostupnou výpočetní kapacitou velmi časově náročné, budou se dále zpracovávat s nově exportovái data. Jedná se o data, ve kterých se vyskytla anomálie týkající se nadměrného provozu. Tento provoz byl zachycen na jedné ekonomické jednotce. Z tohoto důvodu jsou nově exportovaná data omezená adresovým prostorem přiřazeným tomuto subjektu. U těchto dat již nebude opakováno zpracování základním statistických informací, ale zaměříme se rovnou na použití některých modelovacích technik uvedených v kapitole 3.3. (Metody pro datamining). Pomocí dataminingového nástroje PASW Modeler je

aplikována metoda rozhodovacího stromu pro zjištění podstatných atributů pro nastavení systémových varování. Zpracována jsou data za 10 pracovních dní. Tato data byla exportována v podobě 2 souborů. Jeden obsahuje data za období, kdy se anomálie nevyskytla, druhý soubor zachycuje výskyt anomálie. Výstupy dat obou souborů obsahují 30 773 toků. Pro účely zpracování a doby odezev při zpracování takového objemu není nezbytně nutné data dále oddělovat.

Podle grafů uvedených v kapitole 6.6.1 se pokusíme pomocí rozhodovacího stromu zjistit, jaké atributy jsou pro detekci nejpodstatnější. Vzhledem k tomu, že anomálie je nejlépe viditelná na grafu zobrazujícím počet toků za sekundu, bude tento případ zvolen jako první. Dále, podle vizualizace Obr.č.15 – Detekovaná anomálie síťového provozu, budou provedeny stejné kroky pro počet paketů za sekundu a bity za sekundu. Na základě zjištěných informací provedeme nastavení alertů pro případ výskytu obdobné situace. Je nezbytné vycházet z údajů, které jsou obsaženy v konfiguračním rozhraní pro alerty. Proto je analýza zaměřena na tyto atributy. Total flows, Total packtes, Total bytes, Flows/s, Packets/s, Bits/s porovnané absolutní hodnotou nebo n-hodinovými průměry.

6.6 Výsledky analýzy

Na základě syntaxe uvedené v kapitole 6.5 Implementace získaných znalostí, možností filtrování v nástroji nfsen a dalšímu zpracování v PASW Modeleru byly zjištěny tyto výsledky.

Skenování	5.1.2011	7.1.2011	11.1.2011	17.1.2011	27.1.2011
SYN a CONNECT	7	3	0	1	0
FIN	0	0	0	0	0
NULL	10	15	2	0	0
Xmas	0	0	0	0	0
UDP	2	0	1	0	1
DOS útok					
PING FLOOD	0	0	0	0	0
SYN FLOOD	0	0	0	0	0
SSH prolomení hesla	0	4	0	4	0

Problém však vznikl s tím, že v prostředí NetFlow sondy není možné úplné zpracování informací o tocích, potřebné pro zachycení výskytu anomálií. Jedná se především o analýzu komunikace opačným směrem než byl prvotní požadavek. Zachycení podmínek filtru pro toky jedním směrem a logickým ‘and’ spojením s komunikací v opačném směru zároveň a takto získaná data vizualizovat, není v prostředí sondy možné. Toto se týká skenování SYN a CONNECT, UDP, DoS útoků PING FLOOD, SYN FLOOD. Proto je další analýza těchto údajů pro cíle této diplomové práce bezpředmětná. Detekce skenování FIN, NULL a Xmas a pokusů o prolomení hesla do SSH bude provedena v kapitole 6.6.2 (Nastavení alerts).

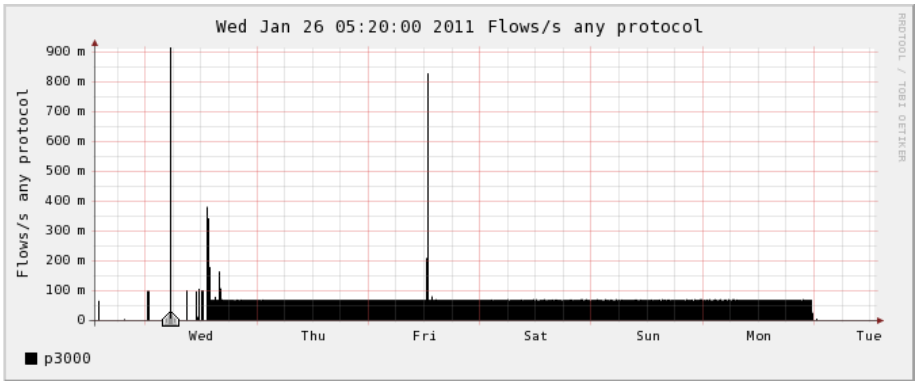
Je třeba se proto zaměřit na stejný problém z trochu jiného úhlu. Ze sledovaných dat byla zjištěna anomálie (více v kapitole 6.6.1 Detekce anomálie). Analyzovat se budou tato data a následně podle nich budou nastaveny alerty v rozhraní sondy. Tento způsob realizace by naplnil cíle diplomové práce. V případě výskytu

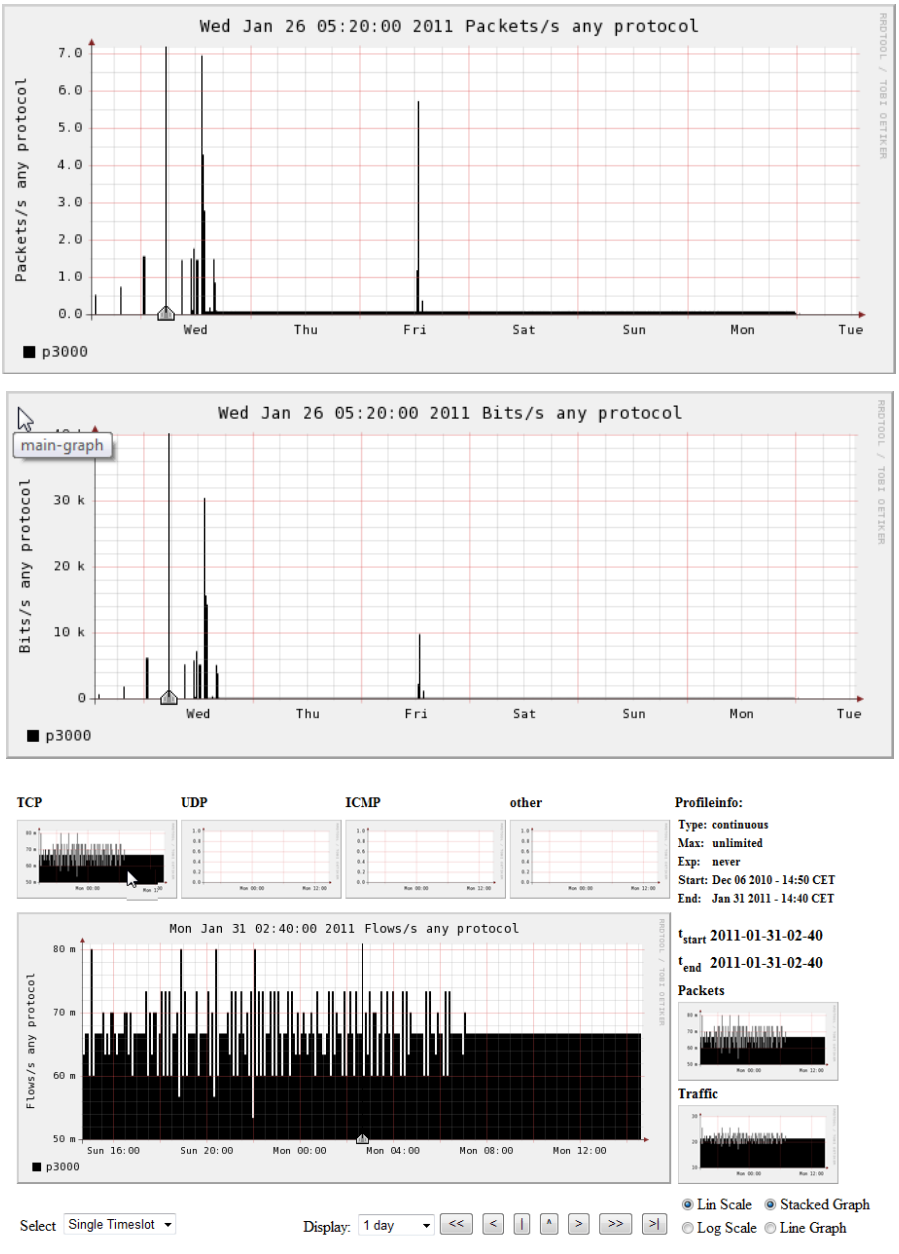
neobvyklé události, v podstatě empiricky nastavené odchylky od běžného provozu, upozorní systém varování správce pomocí mailové zprávy.

6.6.1 Detekce anomálie

Zpracována byla data za celý provoz všech 45 ekonomických jednotek za 5 pracovních dnů jednoho měsíce. Takto zúžený výběr dat k analýze byl zvolen z toho důvodu, že zpracování dat v rozdělení na všechny subjekty přesahuje rozsah diplomové práce. Vlivem toho nebyla ve zpracování celkových dat zachycena anomálie na jedné z 45 ekonomických jednotek. Jelikož však probíhá i vizuální analýza toků pomocí nástroje nfsen, byla anomálie identifikována. Další možnosti vizualizace anomálie jsou uvedeny v příloze B,C. Anomálie je způsobena nezvyklým nárůstem a trváním provozu po dobu 6 dní, viz Obr.č15 – Detekovaná anomálie síťového provozu.

Je nezbytné sledovat provoz odděleně za jednotlivé ekonomické jednotky. Sledovat hodnoty běžného provozu a tím budou odchylky na takto zvolených menších segmentech lépe idenfitikovatelné.



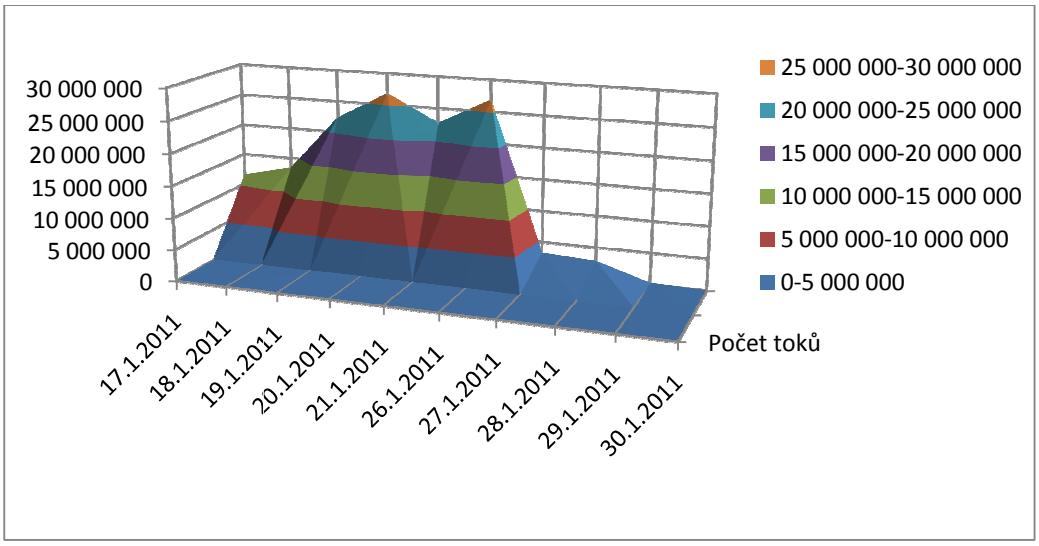


Obr.č.15 – Detekovaná anomálie síťového provozu

Následným krokem je prověřit možnosti nastavení alertu a přizpůsobit tomu analýzu, která má za cíl zjištění odchylek na nastavitelných atributech v alertu. Díky výskytu anomálie je možné provést nastavení systémových varování.

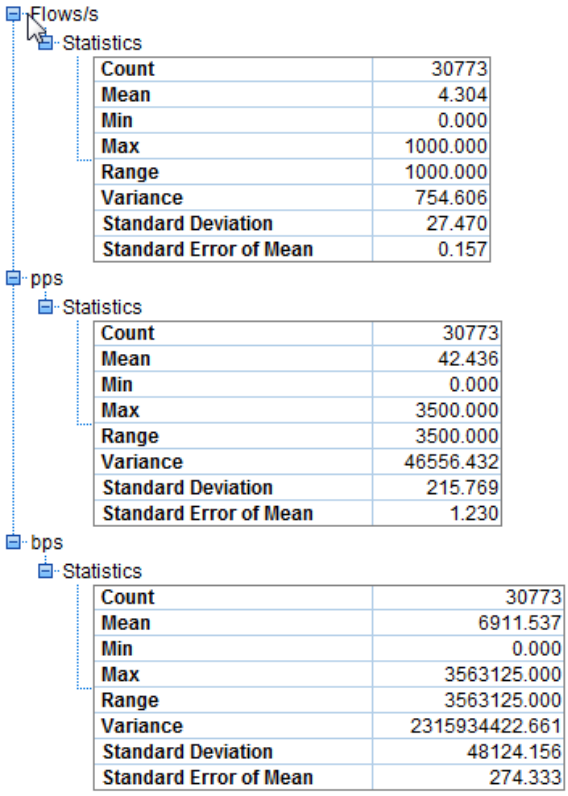
Na základě zjištěných výstupů analýzy bude pak provedeno nastavení alertů pro případ výskytu obdobné situace.

U skenování zkouší útočník různé porty na zařízení, které skenuje. Podle analyzovaných dat však sledovaná IP adresa komunikovala pouze s jedním cílovým portem. Na tomto portu běží jedna z aplikací sledované sítě. Ovšem zaznamenaný způsob provozu neodpovídá běžné komunikaci uživatele se serverem. Vzhledem k těmto skutečnostem vyřazuji z analýzy tyto atributy protokol, zdrojová a cílová adresa, cílový port. Tyto údaje se ve vybraných datech v podstatě neliší a nenabývají rozdílných hodnot rozhodných pro analýzu.



Obr.č.16 – Detekovaná anomálie počty toků

Zjištěné statistické hodnoty vybraných atributů, které pomohly identifikovat anomálii, vypočtené nástrojem PASW Modeler jsou uvedeny níže.



Obr.č.17 – Základní statistiky vybraných atributů

Aby bylo možné nastavit systém varování, alertů, je nezbytné vycházet z údajů, které jsou obsaženy v konfiguračním rozhraní pro alerty určené.

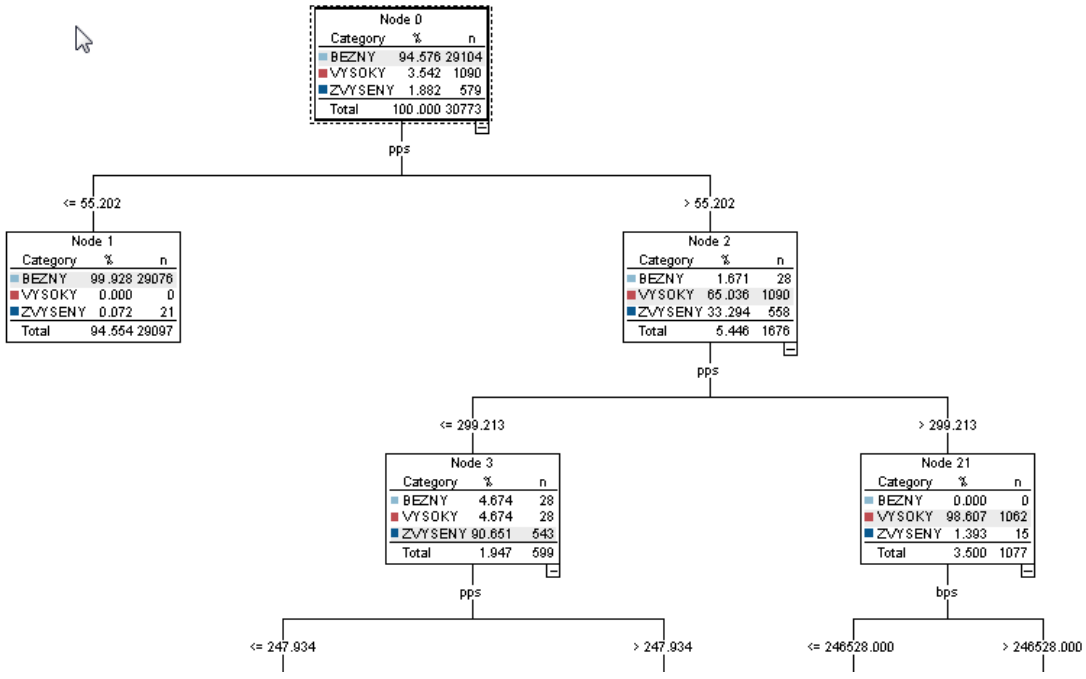
Kromě stejné možnosti použití filtru, jaká byla použita pro detekci skenování, DoS útoků a prolomení SSH hesla, jsou zde podmínky pro nastavení členěny do dvou, pro tento případ použitelných, skupin. Jedna se týká souhrnných statistik o tocích, druhá je založena na nastavení Top 1 statiky. Z toho důvodu byla analýza zaměřena na tyto atributy:

Total flows, Total packtes, Total bytes, Flows/s, Packets/s, Bits/s porovnané absolutní hodnotou nebo n-hodinovými průměry.

Pro vyhotovení stromu pomocí rozhodovacího algoritmu C.5 byla trénovací data doplněna o následující proměnné Práh toku, Práh pps (pakety za sekundu), Práh bps (bajty za sekundu). Pomocí doplňkových proměnných typu „Práh“ byla rozčleněna původní hodnota (toky za sekundu, pakety za sekundu a bajty za sekundu) na 3 kategorie:

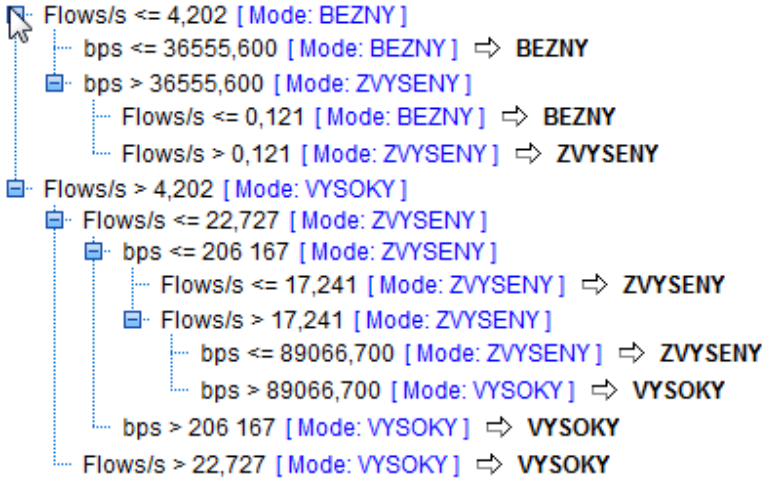
- BĚŽNÝ do výše průměru celého souboru (zahrnuje tedy i anomálii)
- ZVÝŠENÝ hodnoty mezi průměrem a horní hranicí odchylky
- VYSOKÝ od horní hranice odchylky dále

Zde je ukázka části rozhodovacího stromu Práh toků.

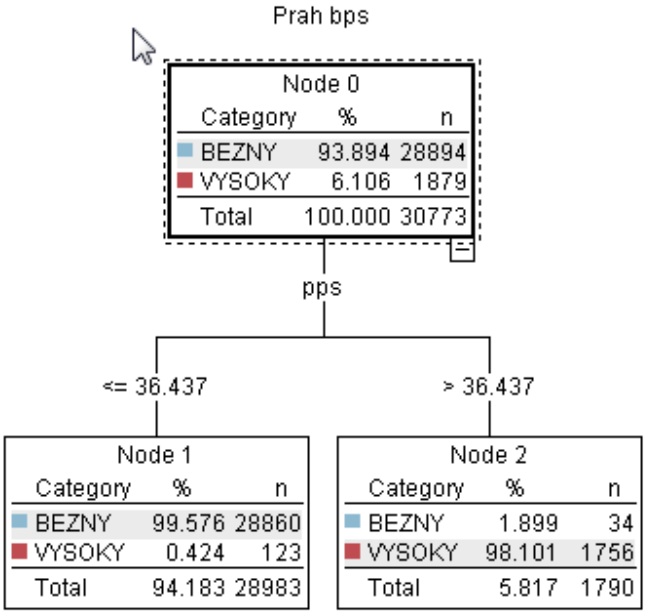


Obr.č.18 – Ukázka části rozhodovacího stromu pro atribut Práh toků

Obr.č.19 – Rozhodovací strom pro atribut Práh toků



Obr.č.20 – Rozhodovací strom pro atribut Práh pps (pakety za sekundu)



Obr.č.21 – Rozhodovací strom pro atribut Práh bps (bajty za sekundu)

Na základě hodnot v tabulce bude provedeno systémové opatření, které v případě dalšího výskytu takové anomálie upozorní správce zařízení.

	1. proměnná	%	2.proměnná	%
Prah toku	pps	98	flags	2
Prah pps	flows/s	95	bps	5
Prah bps	pps	100		

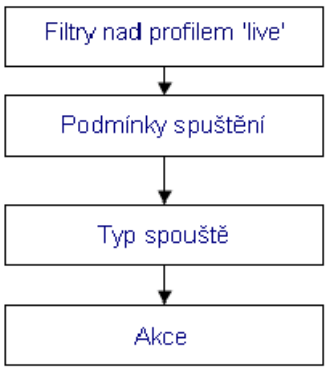
Tab.č.6 – Přehled vybraných atributů dle významnosti

Po vytvoření rozhodovacího stromu bylo zjištěno, že z hlediska očekávaného zisku vychází nejlépe atribut `pps` (pakety za sekundu), dále pak atribut `flows/s` (počet toků za sekundu).

6.6.2 Nastavení varování - alerts

V předchozí kapitole byl popsán postup tvorby rozhodovacího stromu na základě rozhodovacího algoritmu C.5 pro zpracování NetFlow dat s výskytem anomálie. Nyní zbývá provést nastavení alertu. Pro úplnost je třeba uvést možnosti nastavení reakcí na konkrétní události v rozhraní sondy.

Alert je možné definovat pomocí filtrů, podmínek spuštění, typu spouště a akce, která má být provedena.



Obr.č.22 –Schéma při nastavení varování

Syntaxe pro použití filtru je blíže popsána v kapitole 6.3.2 (Syntaxe filtru pro NetFlow). Podmínky spuštění lze definovat buď nad listem toků (Total flow summary) nebo nad statistikami toků (Top 1 statistics). Podmínky mohou být mezi sebou vzájemně zřetězeny (maximálně 6 podmínek). Spojení podmínek je uskutečněno pomocí logické spojky ‘or’ nebo ‘and’.

V případě splnění spouštěcích podmínek dochází k aktivaci spouště (trigger). Trigger lze nastavit na spuštění pokaždé je-li splněna podmínka. Pouze jedno spuštění za dobu platnosti podmínky, spuštění až po n-tém splnění podmínky. Dále je možnost, aby trigger zůstal blokován po dobu několika cyklů. Po spuštění

triggeru dochází ke uskutečnění akce, je-li nadefinována. Obvykle se jedná o odeslání emailu.[21]

Nastavení alertu vychází z výsledků rozhodovacích stromů uvedených v předchozí kapitole. Jelikož by ale výstup vytvořeného rozhodovacího stromu ukazoval jako vysoké i ty hodnoty, které se jednorázově několikanásobně zvýší (lze pozorovat v grafech sondy NetFlow obr.č.15 - Detekovaná anomálie síťového provozu)
je důležité zahrnout i faktor přetrvávání vysokých hodnot po určitý časový úsek. Tato situace je řešitelná pomocí spouště (triggeru) v konfiguračním rozhraní sondy.

Na základě analýzy časového trvání vysokých hodnot v tocích za sekundu a paketech za sekundu (peaks), pomocí grafického rozhraní sondy, bylo prověřeno, že na dané ekonomické jednotce nepřetrvávají takové výkyvy déle než 30 minut. Přesto, že extrémní výkyvy jsou až o řád vyšší než průměrné hodnoty během výskytu anomálie, nepřetrvávají ani zdaleka tak dlouhou dobu jako zachycená anomálie. Proto je časové hledisko podstatným atributem pro nastavení. Z tohoto důvodu musela být provedena další analýza, nyní už jen podstatných atributů zjištěných v tabulce tab.č.6 – Přehled vybraných atributů dle významnosti. Tato analýza je zaměřena na průměrné hodnoty atributů `flows/s` a `pps` v úhrnech jednotlivě za celé dny a zároveň po celých hodinách.

Date flow	Flows	Duration_Mean	Flows/s_Sum	Flows/s_Mean	pps_Sum	pps_Mean
Průměr z 17.1.2011	98,1111	10,6029	634,2557	6,9258	4 145,2200	49,8948
Průměr z 18.1.2011	90,8333	10,1733	241,1233	4,2078	2 850,6813	46,6734
Průměr z 19.1.2011	117,0000	11,1708	343,9675	3,1795	2 501,6914	26,9455
Průměr z 20.1.2011	116,1538	10,8161	335,5886	5,1973	4 145,9057	58,6145
Průměr z 21.1.2011	123,5556	11,0781	954,8362	8,1457	7 686,5536	73,4291
Průměr z 26.1.2011	268,3333	2,2933	5 626,2497	15,6922	57 155,4512	163,6733
Průměr z 27.1.2011	255,6800	0,7462	101,6841	0,3304	1 099,3150	3,5528
Průměr z 28.1.2011	250,5600	0,6600	145,2923	0,4728	1 199,5779	3,9031
Průměr z 29.1.2011	230,4000	0,0000	0,0000	0,0000	0,0000	0,0000
Průměr z 30.1.2011	220,2500	0,0000	0,0000	0,0000	0,0000	0,0000
Celkový průměr	199,8247	4,1945	860,0295	3,8184	8 479,7290	37,8828

Tab.č.7 – Úhrny vybraných atributů za jednotlivé dny dle C.5 rozhodovacího algoritmu

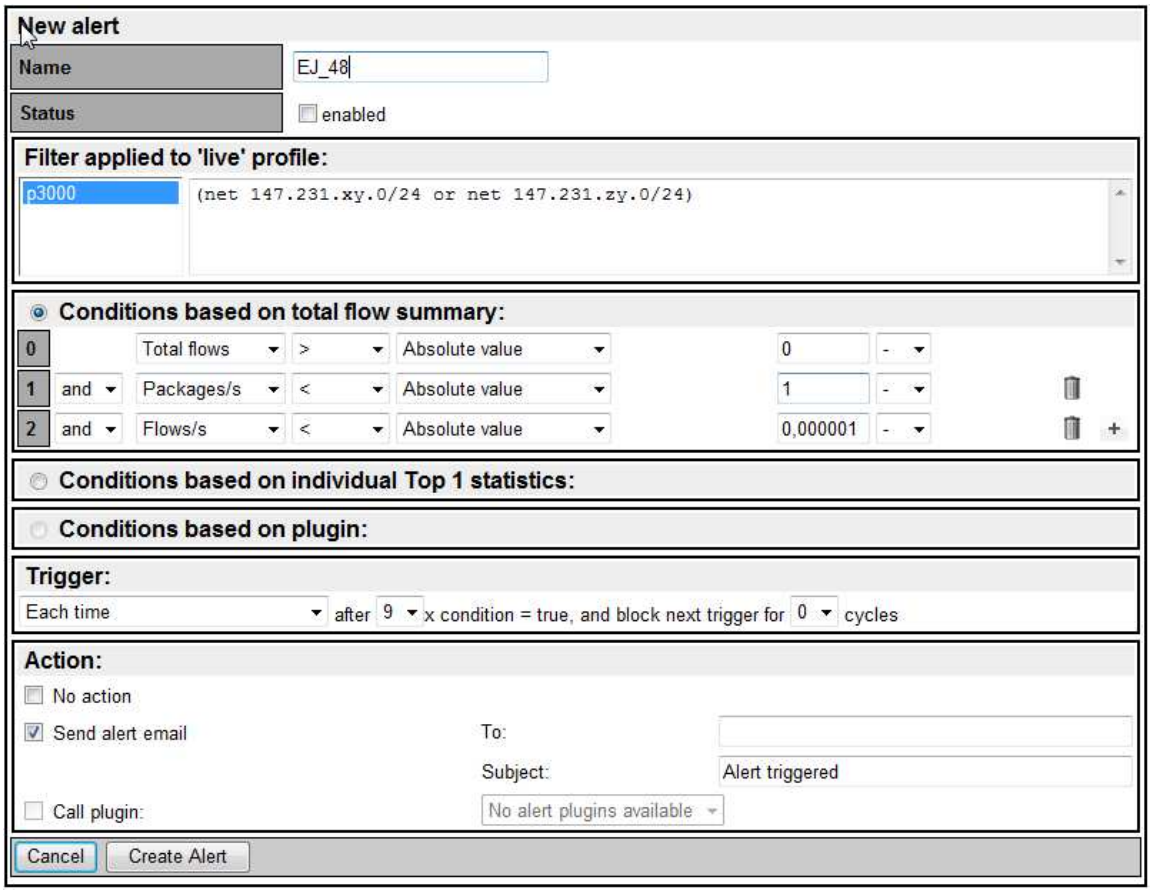
Tabulku úhrnů za jednotlivé hodiny zde z důvodu objemných dat není uvedena. Ovšem hodiné úhrny uvedených atributů, jak vyšlo z analýzy, jsou velmi podstatné. Pro úplnost uvádím rodíl hodnot v hodinových úhrnech během náběhu anomálie.

	Flows/s_Sum	Flows/s_Mean	pps_Sum	pps_Mean
1.hodina	77,838	0,432	2 594,115	14,412
2.hodina	6,935,733	22,519	71 201,690	231,174
3.hodina	15 436,623	49,476	156 404,447	501,296
4.hodina	38 277,776	65,459	380 968,183	652,343

Tab.č.8 – Vybrané atributy náběh anomálie

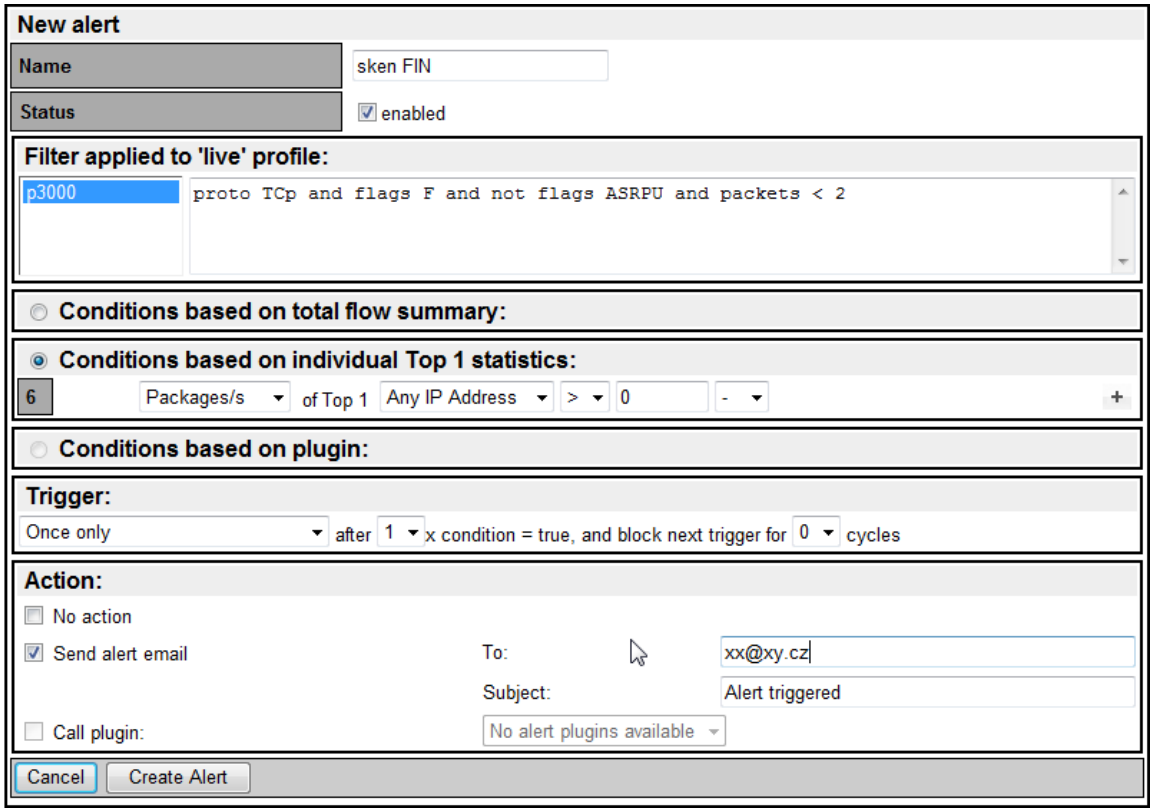
Podle výše uvedených údajů můžeme přistoupit k vlastnímu nastavení podmínek alertu.

Ve většině dat po výskytu tohoto náběhu se v poli Duration (délka trvání toku) objevuje nula. Je to způsobeno tím, že v tomto poli mělo být zobrazeno příliš malé číslo a ve výstupech se zobrazuje jako 0, protože počet toků stále neklesl pod 240 v hodinovém úhrnu. Po prověření situace se zobrazením 0 v rozhraní sondy a opravdu se tak zobrazují pouze data, která výpočtem souvisejí s polem Duration (délka trvání toku). Od této doby až do konce výskytu anomálie však přetrvává komunikace po celé dny, tedy 0-24hod. Minimální počet toků v hodinovém úhrnu je 120 a nikdy neklesl na nižší hodnotu. Není možné tuto komunikaci považovat za běžnou ve vztahu uživatel – server. Je tedy nutné vzít v potaz i nastavení atributu počtu toků a hodnot, které se zobrazují jako nulové. Situace tak poukazuje na vysoký počet velmi krátkých toků, což lze už svou podstatou považovat za anomálii provozu. Nastavení alertu je zobrazeno na obr.č.23 – Ukázka nastavení alertu pro anomálie konkrétní ekonomické jednotky.



Obr.č.23 – Ukázka nastavení alertu pro anomálie konkrétní ekonomické jednotky

Rovněž bylo provedeno nastavení pro detekci skenů FIN, NULL, Xmas a pokusů o prolomení hesla do SSH. Ukázka je k dispozici na obr.č.24 – Alert pro FIN sken



Obr.č.24 – Ukázka nastavení alertu pro sken FIN

7. Závěr

Cílem diplomové práce byla analýza síťového provozu s využitím technik dataminingu pomocí protokolu NetFlow a návrh opatření a jeho realizace na automatickou detekci anomálií vyskytujících se na počítačové síti (internet).

V kapitolách č. 2, 3 a 5 byla popsána teoretická východiska. Diplomová práce se dále zabývá analýzou síťového provozu konkrétního informačního systému, pomocí NetFlow protokolu. Získaná data byla podrobena analýze pomocí softarových nástrojů, zejména pak byl využit nástroj PASW Modeler.

Výsledkem analýzy je objevení anomálie na jedné ekonomické jednotce. Jednalo se o nestandardní chování ve smyslu nezvykle nadměrného síťového provozu. Dále byly také detekovány pokusy o skenování sítě a pokusy o prolomení služby SSH.
Po zjištění výše uvedených skutečností bylo přikročeno v kapitole č. 6.6.2 k návrhu automatických varování v případě výskytu anomálií. Samotná implementace těchto varování byla realizována na zařízení NetFlow sonda od firmy InveaTech a.s., přičemž je nutné podotknout, že obdobně by se dala tato automatická varování realizovat na jakémkoliv jiném zařízení s funkční instalací nástrojů nfdump a nfsen. Tímto lze považovat hlavní cíle této práce za splněné.

Při tvorbě této diplomové práce bylo nutné řešit některé technické a logické problémy spojené s dosažením výše uvedených cílů. Mezi hlavní komplikace patří tyto:

- Velký rozměr získaných vstupních dat pro analýzu (řádové GB), který neumožňoval analyzovat dostupnými hardwarovými prostředky celé

původně plánované období (měsíc). Bylo tedy přikročeno k výběru náhodných dat (pracovních dnů), jejichž rozměr byl již zpracovatelný.

- Je nutné zmínit a zopakovat, že navržená opatření byla určena pro konkrétní informační systém obsahující 45 ekonomických jednotek. Jde tedy o systém dosti rozsáhlý. Pokud bychom analýzu a následný návrh prováděli pro jiný typ, či jinou velikost informačního systému, mohla by být s velkou pravděpodobností realizovaná opatření zcela jiná. I když analýza dat a samotná implementace opatření by se příliš nelišila. Z toho vyplývá, že velmi důležitým aspektem je individuální přístup a správné vyhodnocení výsledků analýzy síťového provozu.
- Další cestou plnohodnotného řešení detekce anomálií přímo v rozhraní sondy by mohlo být zakoupení nebo vytvoření vlastních pluginů zaměřených na detekci anomálií.

Lze předpokládat, že do budoucna budou organizace stále zvyšovat investice a důraz na bezpečnost svých informačních systémů, které se často stávají jejich hlavní konkureční výhodou. Analýza síťového provozu sice nepatří mezi obranné mechanismy používané v první linii síťové bezpečnosti, mezi takové prostředky patří například firewally atd. Nicméně analýza prováděna na historyckých datech může odhalit závažné nedostatny v zabezpečení nebo konfikuraci systému. Analýza síťového provozu je tedy velmi účinným nástrojem stále častějších bezpečnostních auditů.

8. Seznam použitých zdrojů

[1] KANTARDZIC, M. Data Mining: Concepts, Models, Methods and Algorithms. IEEE Press & John Wiley, 2002.

[2] PARR, O. Data mining. COMPUTER PRESS, 2000. 330s. ISBN 80-7226-577-6

[3] PETR, P.DataMining Díl I.Univerzita Pardubice, 2008. ISBN 978-80-7395-098-9

[4] SPENCE, R. Information Visualization, Addison Wesley, Harlow: England, 2001.

[5] DONÁT, J. E-business pro manažery; 1 vyd.Praha: Grada Publishing, 2000, ISBN 80-247-9001-7

[6] LUKASOVA, ŠARMANOVA J.: Metody shlukové analýzy. SNTL, Praha 1985.

[7] WITTEN.I.H., FRANK.E., 1999].: Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufman, 1999, ISBN 1-55860-552-5.

[8] QUINLAN,J.R.: C4.5: Programs for Machine Learning. Morgan Kaufman, 1993, ISBN 1-55860-238-0

[9] UTGOFFf,P.E.,BERKMAN,N.C., Clouse,J.A. (1997): Decision tree induction based on efficient tree restructuring. *Machine Learning,* 29, 5-44.

[10] CLAISE.B.E,: Cisco Systems NetFlow Services Export Version 9.RFC 3954 (Standard), [online] Říjen 2004.
URL http://ietf.org/rfc/rfc3954.txt

[11]Cisco Systems, I.:Introducing to Cisco IOS NetFlow. Technická zpráva, Cisco, Říjen 2007

[12] NFDUMP. [online], Srpen 2005, [cit.2011-03-22].
URL http://nfdum.sourceforge.net

[13] NFSEN.[online],[cit.2011-03-22]
URL http://sourceforge.net

[14] HALLER. M.: Skenování portů:teorie, [online] Říjen 2006 [cit. 2011-03-25].
URL http://www.lupa.cz/clanky/skenovani-portu-teorie/

[15] HALLER. M.: Skenování portů:techniky, [online] Říjen 2006 [cit. 2011-03-25].
URL http://www.lupa.cz/clanky/skenovani-portu-techniky/

[16] POSTEL. J.: Transmission Control Protocol. RFC 793 (Standard), Září 1981, updated by RFC 3168.
URL https:// ietf.org/rfc/rfc793.txt

[17] MESSER.J.,: Secrets of Network Cartography: Comprehensive Guide to NMap, Professor Messre, LLC,[online] 2007, [cit.2011-03-25]
URL http://www.networkuptime.com/nmap/page3-8.shtml

[18] GONG, J.: Identifying P2P users using traffic analysis, SecurityFocus,[online] 20.5.2009, [cit.2011-03-29]
URL http://www.symantec.com/connect/articles/identifying-p2p-users-using-traffic-analysis

[19] GLENN, M.: A Summary of DoS/DDoS Prevention, Monitoring and Mitigation Techniques in Provider Environment, The SANS Institute, 2007,
URL
http://www.sans.org/reading_room/whitepapers/intrusion/summary-dos-ddos-prevention-monitoring-mitigation-techniques-service-provider-enviro_1212

[20] KRAWETZ, N.: Introducting to Network Security, Charles River Media, 2007, 1-58450-464-1.

[21] INVEA-TECH a.s.: FlowMon sonda, Uživatelská příručka (v4.00), 22.března 2010

Seznam obrázků

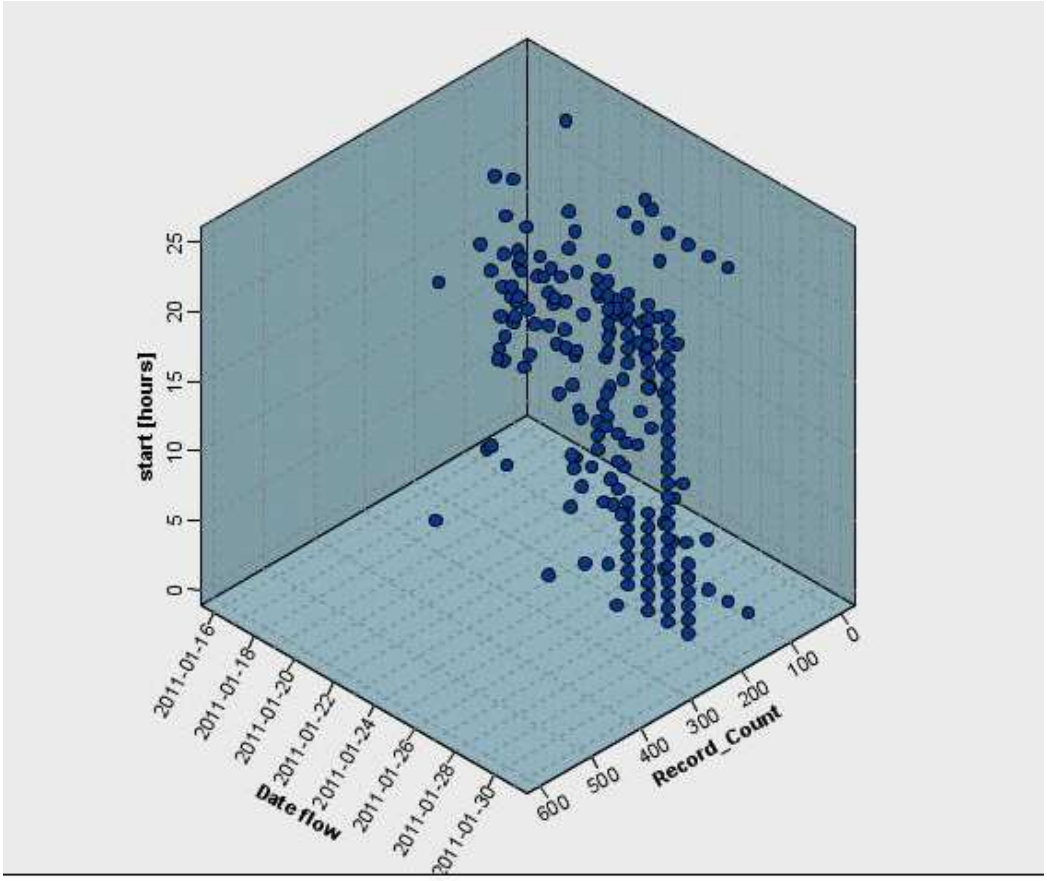
Seznam tabulek

9. Přílohy

Příloha A

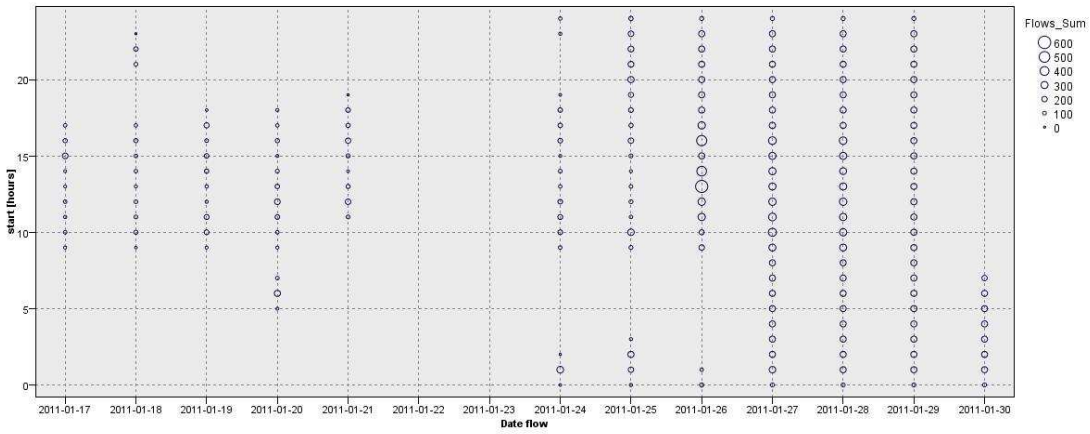
Date first seen	Duration	Proto	IP Addr	Flows	Packets	Bytes
2011-04-06 14:04:30.053	319.438	any	147.231.1.55	2397	106487	24.9 M
2011-04-06 14:06:45.110	171.185	any	147.231.61.198	274	2443	648975
2011-04-06 14:04:27.046	330.365	any	147.231.1.44	199	14782	1.3 M
2011-04-06 14:05:56.670	4.725	any	147.231.36.128	180	1611	391454
2011-04-06 14:04:42.086	254.575	any	147.231.9.20	141	1149	221385
2011-04-06 14:04:43.745	312.361	any	147.231.1.54	136	5001	1.6 M
2011-04-06 14:04:20.940	333.191	any	147.231.1.45	133	10674	766251

Příloha B



Scatter plot - vizualizace anomálie – počty toků v průběhu celých dnů neklesly pod 120

Příloha C



Bublinkový graf - vizualizace anomálie – počty toků v průběhu celých dnů neklesly pod 120