„Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ“

Názov projektu :

Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológii a systémov

ITMS : 26240120039

Názov výstupu :

Záverečná vedecká správa k pracovnému balíku č. 5 - Výskum ochrany súkromia

Obsah

Zoznam skratiek

FTP	File Transfer Protocol
HDFS	Hadoop File System
HTTP	Hypertext Transfer Protocol
IMAP	Internet Message Access Protocol
IP	Internet Protocol
ISV	Independent Software Vendor
KNN	k-Nearest Neighbours
NAT	Network Address Translation
OS	Operačný systém
SFTP	Secure File Transfer Protocol
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
SSH	Secure SHell
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VNC	Virtual Network Computing
VPN	Virtual Private Network
WFP	Windows Filtering Platform
WMI	Windows Management Instrumentation

1 Úvod

Náplňou balika je analýzu dát sieťovej prevádzky s cieľom sledovanie anomálií a zachytenia sieťových útokov. K splneniu tohto cieľa je potrebné pochopiť problematiku a možnosti zberu dát zo sieťovej prevádzky a ich analýzy. V prvom rade je potrebné zabezpečiť zachytenie a filtrovanie sieťovej komunikácie. Následne je potrebné vykonavať ich analýzu s využitím datamining techník. Tieto dve oblasti tvorili základ výskumu pre tento balíček, kde bol doraz kladený na analýzu problému aby sme výskum nadviazali na existujúce poznatky z tejto oblasti.

2 Analýza riešenia

Pre filtrovanie sieťovej premávky existuje niekoľko možnosti. Môžeme vykonávať filtrovanie z prostredia OS, kde sledujeme priamo komunikáciu samostatných zariadení a klientov, alebo sledovanie komunikácie na sieťových zariadeniach kde sledujeme súhrnu komunikáciu v rámci celej siete. Následne je potrebné získané dáta ukladať na vhodný úložný priestor a analyzovať.

2.1 Zachytenie a filtrovanie sieťovej komunikácie z prostredia OS

Cele filtrovanie je možne vykonať v režime používateľa, bez potreby písania ovládačov. Tento spôsob ale má množstvo obmedzení a pre bezpečnostné aplikácie je úplne nepoužiteľný. Druhou možnosťou je vykonanie filtrovania a zberu v režime jadra. Na systémoch windows k tomu možeme využiť windows filtering platform (WFP). WFP je bohatá a rozšíriteľná platforma pre monitorovanie, zachytávanie a spracovávanie sieťovej premávky na všetkých úrovniach sieťového zásobníka. Nahrádza rozhrania pre filtrovanie sieťovej premávky vo Windows XP a Windows Server 2003. WFP pozostáva z množiny spojov so sieťovým zásobníkom (network stack) a filtračného jadra riadiaceho spoluprácu so sieťovým zásobníkom. WFP poskytuje filtračnú infraštruktúru, do ktorej môžu ISV zapájať špecializované filtračné moduly. Niektoré sieťové služby OS Windows rozširujú vlastnosti štandardného systémového protokolového ovládača TCP/IP (protokol s riadením vysielanie/Internetový protokol) pomocou WFP. Jednou z možností ako pracovať s WFP je využitie balíka Windivert. WinDivert je balík pre Windows Vista, Windows Server 2008, Windows 7 a Windows 8. WinDivert môže byť použitý v užívateľskom móde pomocou prístupu k ovládaču ktorý pracuje v režime jadra na zachytávanie paketov, filtrovanie paketov, sniffing, firewall, NAT, VPN atď., bez nutnosti písania vlastného kódu v režime jadra.

2.2 Zachytenie a filtrovanie sieťovej komunikácie z sieťových zariadení

Druhou spomínanou metódou je sledovanie komunikácie na sieťových zariadeniach a priamo z prostredia siete. Pre analýzu tohto prístupu sme vykonali prieskum možností a dostupných riešení čo zahrnovalo:
- Štúdium možností a dostupných spôsobov, nástrojov pre zber sieťovej prevádzky [1],[2],[3],[4].

- Skúmanie nasledovných nástrojov na sledovanie prevádzky siete a monitorovanie prenášaného obsahu:

1. Microsoft Network Monitor
2. Nagios
3. Network Miner
4. Pandora FMS
5. BandwidthD
6. EasyNetMonitor

7. Zenoss Core
8. PRTG Network Monitor Freeware
9. Angry IP Scanner
10. ntopng
11. Total Network Monitor
12. NetXMS
13. xymon
14. WirelessNetView
15. Xirrus Wi-Fi Inspector
16. WireShark

Následne sa vykonalo vyhodnotenie informácií získaných o jednotlivých nástrojoch na sledovanie prevádzky siete a monitorovanie prenášaného obsahu, v rámci čoho bolo vykonané porovnanie protokolov SNMP, Netflow a WMI.

Pri tejto oblasti bolo potrebné sa tiež venovať:

- Určeniu hardvérovej špecifikácie pre zber údajov zo sieťovej prevádzky.
- Analýza dostatočnosti hardvérovej špecifikácie pre zber údajov zo sieťovej prevádzky.
- Štúdium formátu výstupných súborov sieťovej prevádzky pcap, jeho špecifikácie a binárnych komponentov, časť 1.
- Štúdium formátu výstupných súborov sieťovej prevádzky pcap, jeho špecifikácie a binárnych komponentov, časť 2.
- Simulované testovanie navrhnutej hardvérovej zostavy.
- Praktické testy s dátami na úrovni protokolu UDP/IP.
- Praktické testy s dátami na úrovni protokolu TCP/IP.
- Spracovanie výstupov testov pre jednotlivé protokoly (TCP/IP a UDP/IP), porovnávanie výsledkov.

2.3 Analýza dát využitím datamining techník

V rámci tejto časti sa vykonalo štúdium datamining techník so zameraním na vstupné dáta zo sieťovej prevádzky, čo zahŕňa:

- štúdium techník pre predspracovanie dát so zameraním na rýchlosť spracovania,
- návrh vhodnej štruktúry ukladania a správy veľkých objemov dát,
- prehľad analytických metód vhodných na spracovanie veľkých objemov dát v prostredí siete,
- štúdium možností sledovania prevádzky siete a jej monitorovania,
- analýzu spôsobu implementácie metód spracovania veľkých dát, a analýzu možností ich využitia pre účely sledovania sieťovej prevádzky.

2.3.1 Štúdium techník pre predspracovanie dát so zameraním na rýchlosť spracovania

Pre splnenie tejto oblasti bolo vykonané štúdium materiálov zameraných na:
dataminig techniky “support vector regression” [5],[6],[7],[8]
dataminig techniky “neurónové siete” [9],[10],[11],[12]
dataminig techniky “KNN model” [13],[14],[15],[16]
dataminig techniky “rozhodovacie stromy” [17],[18],[19],[20]

Ďalej bolo vykonané štúdium:
zhlukovacích algoritmov bez použitia trénovacích príkladov [21],[22] a multiscale metod [21].

2.3.2 Prehľad analytických metód vhodných na spracovanie veľkých objemov dát v prostredí siete

V tejto časti sme sa zamerali na tvorba prehľadu analytických metód vhodných na spracovanie veľkých objemov dát zo siete.

To zahrňalo:

- Štúdium pomocných metód (štatistické momenty vyšších rádov) pre analýzu údajov zo sieťovej prevádzky.
- Návrh parametrov štatistických metód vyšších rádov pre účely analýzy nami získaných údajov.
- Analýza získaných údajov zo sieťovej prevádzky podľa protokolov (http, smtp, ftp, sftp).
- Analýza získaných údajov zo sieťovej prevádzky podľa protokolov (ssh, imap pop3, vnc, rdesktop).
- Analýza získaných údajov zo sieťovej prevádzky podľa typu spojenia (spojovo orientované).
- Analýza získaných údajov zo sieťovej prevádzky podľa typu spojenia (nespojovo orientované).
- Analýza známych vektorov sieťových útokov

Podrobnejšie štúdium metód pre analýzu dát sieťovej prevádzky z odbornej literatúry: Analýzu protokolov [23],[24], analýzu typov spojenia [25],[26] a tiež analýzu známych vektorov sieťových útokov [27],[28],[29].

2.3.3 Štúdium možností sledovania prevádzky siete a jej monitorovania

Táto oblasť zahrňovala štúdium štatistických techník predspracovania údajov: štandardizácia (škálovanie rozptylu a odstránenie posunu strednej hodnoty), normalizácia, binarizácia. Ďalej

štúdium štatistických techník predspracovania údajov: kódovanie vlastností, dopĺňanie chýbajúcich hodnôt, redukcia dát. Tiež pre potreby monitorovania bolo nutné vykonať analýzu vhodnosti určenia štatistík vyšších rádov pri predspracovaní údajov: šikmosť, špicatosť. Hľadanie závislostí medzi atribútmi: závislosť symbolických atribútov, kontingenčná tabuľka, závislosť numerických atribútov.

Pri sledovaní a monitorovaní sieťovej prevádzky je potrebné zabezpečiť tiež proces predspracovania sledovaných dát pre potreby algoritmov a nástrojov pre sledovanie a monitorovanie sieťovej prevádzky. Zároveň štúdium predspracovania dát využijeme tiež pri procese analýzy sieťovej komunikácie, kde je taktiež potrebné data upraviť pre potreby algoritmov určených na analýzu.

Príprava hlavných procesov predspracovania pozostáva z návrhu nasledovných algoritmov na vyššej úrovni: 1. čistenie údajov, 2. integrácia údajov, 3. transformácia údajov, 4. redukcia (výber) údajov

V ďalšej časti sme sa zamerali na

- spresnenie algoritmov čistenia údajov: vypĺňanie chýbajúcich hodnôt, odstraňovanie výrazne odchýlených hodnôt, odstraňovanie nekonzistentností
- spresnenie algoritmov integrácie a sčasti aj transformácie údajov: identifikácia rovnakých objektov v rámci rozličných zdrojov údajov, detekcia a odstraňovanie rôznych konfliktov medzi údajmi, vylučovanie redundantných záznamov, agregácia údajov
- spresnenie algoritmov transformácie a redukcie (výberu) údajov: využitie hierarchie konceptov, min-max algoritmus, decimácia, prevod numerických znakov na nenumerické, redukcia dimenzie (výber sledovaných atribútov), redukcia množstva údajov (výber reprezentantov tried)

- predspracovanie a analýzu textu pre techniky klasifikácie a zhlukovania [30]

Po štútiu teoretických princípov a možností bolo vykonané:

- Zhrnutie výsledkov z analýzy testovacích údajov pre základné sieťové protokoly.
- Hľadanie možností optimalizácie nástrojov pre zber údajov zo sieťovej prevádzky na základe predošlých testov a ich výsledkov
- Vykonanie optimalizácie nástrojov pre zber údajov zo sieťovej prevádzky.
- Opätovný zber testovacích údajov pre sieťový protokol UDP/IP, s optimalizovanými parametrami nástrojov pre zber údajov zo sieťovej prevádzky.
- Opätovný zber testovacích údajov pre sieťový protokol TCP/IP, s optimalizovanými parametrami nástrojov pre zber údajov zo sieťovej prevádzky.
- Spracovanie získaných údajov v predošlých testoch.

- Vyhodnotenie optimalizačného procesu porovnaním s výsledkami prvotných testov, doladenie parametrov nástrojov pre zber údajov zo sieťovej prevádzky.
- Štúdium možností a dostupných spôsobov, nástrojov pre zber dát z operačného systému [31],[32],[33]
- Analýza možností zberu údajov z operačného systému Linux.
- Analýza možností zberu údajov z operačného systému Windows.
- Štúdium možností optimalizacie, nastrojov pre zber sieťovej prevádzky [34]

2.3.4 Návrh vhodnej štruktúry ukladania a správy veľkých objemov dát

Po zachytení informácií o sieťovej komunikácii je nutné ich ukladať. Predpokladá sa, že systém bude zbierať veľké množstvo dát preto je nutné využiť technológie a možnosti analýzy ktoré sú navrhnuté pre big data model. Pri týchto technológiach však je potrebné dať dôraz na zabezpečenie ochrany dát. Prvý problém je zlyhanie hardvéru. Ako náhle začneme používať veľa kusov hardvéru, šanca, že jeden zlyhá je pomerne vysoká. Štandardným spôsobom, ako sa dá vyhnúť strate údajov je cez replikácie. Sú to redundantne kópie dát, aby systém v prípade poruchy mal k dispozícii ďalšie kópie. Na obdobnom princípe pracuje Hadoop Distributed File System, teda súborový systém Hadoop.

Ďalší problém je, že pri väčšine analyzačných úlohách je nutne čítať dáta nielen z jedného disku, ale aj z niektorých z 99 ďalších diskov, čo môže byt dosť náročné. MapReduce, jedná z predností Hadoop-u poskytuje programovací model, ktorý rieši problém zápisu a čítania z viacerých diskov, transformuje uloženie dát do sady kľúčov a hodnôt. MapReduce pozostáva z krokov mapovania a redukovania, niekedy sa pridáva medzi tieto dva kroky ešte jeden krok a to je premiešavanie. Hadoop stanovuje spoľahlivé zdieľanie úložiska a systém na analýzu dát. Ukladanie je realizované pomocou HDFS a analýza je zabezpečená MapReducom. Ide o základný stavebný kameň Hadoop-u, ktorý však zahŕňa aj iné časti, preto si ich predstavme v nasledujúcej, osobitnej časti. Pre zber dát je potrebné vyvinúť ešte dalšie úsile. Môžeme vyvinúť vlastný systém pre zber, alebo použiť po vhodnej úprave už vytvorené systémy ako napr. chukwa.

Chukwa je open-source systém zberu dát pre sledovanie rozsiahlych distribuovaných systémov. Chukwa je nadstavbou Hadoop-u (respektíve jeho modulu Distributed File System (HDFS) a MapReduce) a dedí škálovateľnosť a robustnosť Hadoop-u. Chukwa tiež obsahuje flexibilnú a užitočnú sadu nástrojov pre zobrazovanie, sledovanie a analyzovanie výsledkov pre čo najlepšie využitie zozbieraných údajov.

Pri tejto oblasti bolo potrebne venovať pozornosť aj nasledujúcim otázkam:

- Analýza spôsobu implementácie metód spracovania veľkých dát ktorá zahŕňala štúdium článku merania kvality v technikách získavania údajov z veľkého objemu dát [35],[36]:
- Návrh multidimenzionálneho merania kvality údajov.
- Analýza nasledovných kritérií multidimenzionálneho merania kvality údajov:
    - presnosť,

- úplnosť,
- konzistencia,
- časová náväznosť.

- Analýza nasledovných kritérií multidimenzionálneho merania kvality údajov:
  - dôveryhodnosť,
  - pridaná hodnota informácie,
  - zmysluplnosť,
  - dostupnosť.

- Analýzu spôsobu paralelnej implementácie metód spracovania veľkých dát.
- Oboznámenie sa so špecifickou technológiou v distribuovanom prostredí MapReduce [37].
- Analýza implementácie metód paralelného spracovania veľkých dát v distribuovanom prostredí MapReduce [38], [39].
- Oboznámenie sa so špecifickou technológiou v distribuovanom prostredí HIVE [40].
- Analýza implementácie metód paralelného spracovania veľkých dát v distribuovanom prostredí HIVE [41].
- Oboznámenie sa so špecifickou technológiou v distribuovanom prostredí Pig a analýza implementácie metód paralelného spracovania veľkých dát v tomto distribuovanom prostredí [42].

2.3.5 Analýzu spôsobu implementácie metód spracovania veľkých dát, a analýzu možností ich využitia pre účely sledovania sieťovej prevádzky.

Pre analýzu spôsobu implementácie metód spracovania veľkých objemov dát sme riešili pomocou vytvorenia experimentov s rôznymi scenarmi. Ktomu bolo potrebné realizovať nasledujúce úlohy:

- Príprava údajov na experimenty. Spúšťanie experimentov zameraných na spracovanie veľkých objemov údajov zo siete.
- Vyhodnocovanie údajov získaných z experimentov zameraných na spracovanie veľkých objemov údajov zo siete.
- Optimalizácia vybraných atribútov na základe experimentov zameraných na spracovanie veľkých objemov údajov zo siete. Opakované spúštanie a vyhodnocovanie experimentov na základe optimalizácie.
- Opakované spúšťanie a vyhodnocovanie experimentov, a na základe finálnej optimalizácie výber vhodných atribútov pre analýzu údajov získavaných zo spracovania veľkých objemov údajov zo siete.

Tiež sme sa venovali štúdiu zhlukovacích algoritmov bez použitia trénovacích príkladov [21], [22] štúdiu multiscale methods [21] a analýze Data processing [22], a Data warehouse and OLAP technology [22] čo bolo potrebné pre hľadanie spôsobu implementácie týchto metód pre potreby analýzy sieťovej prevádzky pri veľkom objeme dát.

V rámci tejto oblasti sme je ale hlavná pozornosť venovaná štúdiu základnej teórie modelovania veľkých dát, návrh modelu a sémantike modelovacieho jazyka Web Ontology Language.
Tiež je venovaná pozornosť problematike generovania údajov pomocou nástrojov:

- Giraph.
- Storm.

Následne je potrebné vykonať:

- Aplikáciu modelu na vygenerované údaje.
- Hľadanie vzorov a predikcia správania sieťovej prevádzky.
- Definovanie normálneho stavu sieťovej prevádzky.
- Spustenie nástroja na analýzu sieťovej prevádzky za účelom hľadania anomálií.
- Sledovanie anomálií
- Pokračovanie v analýze sieťovej prevádzky za účelom hľadania anomálií.
- Analýza a experimentovanie s metódami a nástrojmi na modelovanie veľkých dát
- Vytváranie umelého testovacieho prostredia pre sledovanie anomálií jednotlivých protokolov pre sieťovú prevádzku prenosu veľkých objemov dát.
- Generovanie anomálií v testovacom prostredí pripravenom pre sledovanie anomálií protokolov.
- Vytváranie špeciálneho testovacieho prostredia pre sledovanie anomálií jednotlivých typov sieťových spojení pri prenose veľkého objemu údajov.
- Generovanie anomálií v testovacom prostredí pripravenom pre sledovanie anomálií typov spojení.
- Spracovanie údajov získaných z generovania anomálií pre obe vytvorené testovacie prostredia (vzhľadom na protokoly a vzhľadom na typy dátových spojení).
- Príprava príkladov aplikácie modelov pre sledovanie správania protokolov pre sprístupňovanie webového obsahu.
- Príprava príkladov aplikácie modelov pre sledovanie správania protokolov pre sprístupňovanie elektronickej komunikácie.
- Príprava príkladov aplikácie modelov pre sledovanie správania riadiacich protokolov toku údajov v sieti.
- Príprava príkladov aplikácie modelov pre zachytenie sieťových útokov vzhľadom na protokoly pre sprístupňovanie webového obsahu.
- Príprava príkladov aplikácie modelov pre zachytenie sieťových útokov vzhľadom na protokoly pre sprístupňovanie elektronickej komunikácie.

- Príprava príkladov aplikácie modelov pre zachytenie sieťových útokov vzhľadom na riadiace protokoly toku údajov v sieti.
- Testovanie modelov na sledovanie správania a predikciu prípadných útokov

3 Záver

Návrh a riešenie nami navrhovaného systému pozostáva z agentov ktorý sú inštalovaný na jednotlivých zariadeniach pripojených do sieťovej komunikácie, a z kolektorov ktoré zabezpečujú zber získaných dát od agentov. Následne sa dáta ukladajú na hadoop file systém (HDFS). Po ich uložení sa pomocou nástrojov strojového učenia vykonáva analýza týchto dát. Tak je možné zachytiť anomálie a potenciálne útoky ktoré sa prejavujú a je možné ich identifikovať v sieťovej prevádzke.
V September 2015 bolo zahájené experimentovanie s metódami a nástrojmi na modelovanie veľkých dát s cieľom analýzy týchto dát. Výsledkom analýzy je sledovanie anomálií, ich predikcia a zachytenie závislostí.

Bola vykonaná príprava príkladov aplikácie modelov pre sledovanie správania a zachytenia sieťových útokov a čiastočné testovanie a vyhodnotenie výsledkov testov. V tomto štádiu, ale ešte nie je možné vykonať relevantné závery a je potrebné v testovaní pokračovať, čo je úloha do ďalšieho obdobia.

Literatúra:

1. Software Tool for Passive Real-Time Measurement of QoS Parameters http://users.utcluj.ro/~dobrota/pdf/Software_Tool_for_Passive_Real_Time_Measurement_of_QoS_Parameters.pdf

2. Active measurement tool for the EuQoS project http://www.netit.upc.edu/images/pdf/2005/MOME2005.pdf

3. Annex of Monitoring quality of Internet access services in the context of net neutrality http://berec.europa.eu/files/document_register_store/2014/3/BoR%20%2814%29%2024%20Draft%20BEREC%20NN%20QoS%20Monitoring%20Report%20ANNEX.pdf

4. QUALITY OF SERVICE IN TELECOMMUNICATION NETWORKS
5. http://www.eolss.net/sample-chapters/c05/e6-108-14-00.pdf

6. www.cs.ucf.edu/courses/cap6412/fall2009/papers/Berwick2003.pdf

7. 2.pdf.aminer.org/000/337/560/uncertainty_support_vector_method_for_ordinal_regression.pdf

8. http://alex.smola.org/papers/2003/SmoSch03b.pdf

9. http://www.svms.org/regression/SmSc98.pdf

10. http://www.cs.sjsu.edu/faculty/lee/cs157b/Data_Mining_and_Neural_Networks_danny_leung.ppt

11. http://www70.homepage.villanova.edu/matthew.liberatore/Mgt2206/turban_online_ch06.pdf

12. http://www.jatit.org/volumes/research-papers/Vol5No1/1Vol5No6.pdf

13. http://www.cs.iastate.edu/~honavar/nn7.pdf?origin=publication_detail

14. http://chem-eng.utoronto.ca/~datamining/Presentations/KNN.pdf

15. http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf

16. http://cis.poly.edu/~mleung/FRE7851/f07/k-NearestNeighbor.pdf

17. http://www.ce.sharif.ir/courses/84-85/2/ce324/resources/root/Supplementary Materials for Final Exam/Data Mining (Classification)pdf.pdf

18. http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

19. http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec3.pdf

20. http://biolab.uspceu.com/datamining/pdf/DecisionTrees.pdf

21. http://www users.cs.umn.edu/~kumar/dmbook/dmslides/chap4_basic_classification.pdf

22. David L. Banks.,Classification, clustering, and data mining applications. http://books.google.sk/books?id=hQW_j7NPy98C&printsec=frontcover&dq=clustering&hl=sk&sa=X&ei=DKJfT_WRMZHOsgamz827CQ&ved=0CEAQ6AEwAg#v=onepage&q=clustering&f=false

23. Jiawei Han, Micheline Kamber., Data mining http://books.google.sk/books?id=AfL0t-YzOrEC&printsec=frontcover&hl=sk&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

24. Formal Analysis of Network Protocols http://repository.upenn.edu/cgi/viewcontent.cgi?article=1970&context=cis_reports

25. Security Analysis of Network Protocols https://www.usenix.org/legacy/event/sec08/tech/slides/mitchell_slides.pdf

26. A Summary of Network Traffic Monitoring and Analysis Techniques http://www.cse.wustl.edu/~jain/cse567-06/ftp/net_monitoring.pdf

27. Network Basics Companion Guide https://books.google.sk/books?id=sTrSAQAAQBAJ&pg=PA104&lpg=PA104&dq=Network+Basics+Companion+Guide&source=bl&ots=bsYUk0m7-7&sig=N_16KPpkcc3AtvnqxVDh--l8IA8&hl=en&sa=X&ved=0CEIQ6AEwBmoVChMIjcvxg-

3AyAIVxtYUCh18HAFP#v=onepage&q=Network%20Basics%20Companion%20Guide&f=false

28. DCE RPC Vulnerabilities New Attack Vectors Analysis http://www.coresecurity.com/content/dce-rpc-vulnerabilities-new-attack-vectors-analysis#sthash.927QL6Hl.dpuf

29. NBA of Obfuscated Network Vulnerabilities' Exploitation Hidden into HTTPS Traffic http://dx.doi.org/10.1109/ICITST.2014.7038827

30. Attack Methodology Analysis: Emerging Trends in Computer- Based Attack Methodologies and Their Applicability to Control System Networks https://inldigitallibrary.inl.gov/sti/3494179.pdf

31. Data, informace, znalosti a Internet http://books.google.sk/books?id=UJh-gLdTH8IC&pg=PA499&dq=clustering+dokumentov&hl=sk&sa=X&ei=F6ZfT4L0FszNsgbMtdSxBg&redir_esc=y#v=onepage&q=clustering%20dokumentov&f=false

32. Acquiring Volatile Operating System Data Tools and Techniques http://dl.acm.org/citation.cfm?id=1368516

33. Techniques and Tools for Recovering and Analyzing Data from Volatile Memory http://www.sans.org/reading-room/whitepapers/forensics/techniques-tools-recovering-analyzing-data-volatile-memory-33049

34. Emerging Digital Forensics Applications for Crime Detection, Prevention, and Security https://books.google.hr/books?hl=en&lr=&id=U8m6YLhoLesC&oi=fnd&pg=PR1&dq=Emerging+Digital+Forensics+Applications+for+Crime+Detection,+Prevention,+and&ots=EDNbHLG0TE&sig=Pi6n9h6tdyWZt3qnh2mGS2AO-po#v=onepage&q=Emerging%20Digital%20Forensics%20Applications%20for%20Crime%20Detection%2C

35. Software Tool for Passive Real-Time Measurement of QoS Parameters http://users.utcluj.ro/~dobrota/pdf/Software_Tool_for_Passive_Real_Time_Measurement_of_QoS_Parameters.pdf

36. Laure Berti-Équille: Measuring and Modelling Data Quality for Quality-Awareness in Data Mining; in Quality Measures in Data Mining, Studies in Computational Intelligence, Vol. 43, 2007, XIV, 314 p. Strany 101 až 126.

37. Christen P., Goiser K.: Quality and Complexity Measures for Data Linkage and Deduplication; in Quality Measures in Data Mining, Studies in Computational Intelligence, Vol. 43, 2007, XIV, 314 p. Strany 127 až 152.

38. MapReduce
http://chimera.labs.oreilly.com/books/1234000001811/apb.html

39. MapReduce
http://books.google.sk/books?id=drbI_aro20oC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

40. MapReduce
Jeffrey Dean , Sanjay Ghemawat, MapReduce: simplified data processing on large clusters, Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation, p.10-10, December 06-08, 2004, San Francisco, CA

41. Programming Hive By Edward Capriolo, Dean Wampler
http://books.google.sk/books?id=n8FffirFaQIC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

42. Hadoop: The Definitive Guide By Tom White
http://books.google.sk/books?id=drbI_aro20oC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

43. Pig Latin, a Parallel Dataflow Language
http://chimera.labs.oreilly.com/books/1234000001811/index.html