Big Data texnologiyalarının təhlükələri

Məkrufə Hacırəhimova

AMEA İnformasiya Texnologiyaları İnstitutu

makrufa@science.az

***Xülasə*— Məqalədə informasiyanın emalında yeni eranı əks etdirən "Big data" texnologiyalarının qısa xülasəsi verilir. Bu texnologiyanın bəzi təhlükəsizlik aspektləri tədqiq olunur.**

Açar sözlər— big data; big data analytics; security; anonymisation, privacy; encryption.

I. GİRİŞ

Dünya miqyasında informasiyanın daim artan sürəti və həcmi cəmiyyətdə "informasiya partlayışı" kimi xarakterizə olunur. Yeni əsrin əvvəllərindən başlayaraq rəqəmsal verilənlər hər il həndəsi silsilə ilə artmaqdadır [1]. Bəşəriyyətin mövcudluğundan 2003-cü ilə qədərki dövrdə dünyada cəmi 5 ekzabayt məlumat generasiya olunmuşdursa, 2012-ci ildə rəqəmsal informasiyanın həcmi 500 dəhə artaraq 2.7 zetabayt olmuş və 2015-ci ildə üç dəfə artması, növbəti hər il 40% artaraq, 2020-ci ildə dünyada informasiyanın həcminin 44 zetabayta çatacağı proqnozlaşdırılır [2]. Bu da Yer kürəsinin hər nəfərinə düşən 5200 qiqabayt informasiyaya bərabərdir. İnformasiyanın həcminin artmasında kompüter, mobil qurğular və İnternetin rolu böyükdür. Erikson şirkətinin məlumatına görə 6 milyard telefon (2012), Gartner şirkətinin məlumatına görə 2 milyard kompüter (2012), İKT sahəsinin əsas göstəriciləri üzrə Beynəlxalq Telekommunikasiya İttifaqının (2014) tədqiqatına görə dünya əhalisinin təxminən 40% İnternet istifadəçisidir. Dünyada hər gün 2.5 trilyon bayt məlumat hazırlanır, hər dəqiqədə 100 milyon email göndərilir, Google axtarış sistemində 2 milyon axtarış sorğusu, Facebook sosial şəbəkəsində 350 qiqabayt məlumat emal olunur və 570-dən çox veb-sayt yaradılır, hər dəqiqə 72 saatlıq yeni video YouTuba yüklənir və s.. Göründüyü kimi veb, sosial şəbəkələr, mobil qurğular, kredit kartları vasitəsilə edilən tranzaksiyalar və s. rəqəmsal verilənlər axınının artmasına gətirib çıxarmış, informasiya bolluğu yaranmış, dünya sanki informasiya ilə doldurulmuşdur. Bunun məntiqi nəticəsi olaraq verilənlərin emalı, saxlanılması və istifadəsində yeni eranı əks etdirən böyük verilənlər (*ing. big data*) termini meydana çıxmışdır [3]. "Big data" dedikdə xüsusi informasiya massivləri başa düşülür ki, ənənəvi verilənlər bazası və alətləri bunları idarə edə bilmir və ya kifayət qədər yaxşı etmir. Bu verilənlərin emalı üçün kifayət qədər saxlama tutumu və hesablama gücü tələb olunur. "Big data" artıq elmi ictimaiyyət, biznes-cəmiyyətləri, hökumət strukturları tərəfindən strateji resurs kimi dəyərləndirilir [4,5], 2013 və 2014-cü illərdə əsas texnoloji istiqamət adlandırırlar [2,6], bəzən də yeni neft (*new oil*) kimi qiymətləndirirlər [7]. "Verilənlər haqqında elm" (*data science)* 2013-cü ildən başlayaraq bir çox aparıcı universitetlərdə akademik fənn kimi tədris olunmaqdadır.

"Big data"nı təyin etməyə və digər verilənlərdən fərqləndirməyə kömək edən ilk model, çox böyük sürətlə (*velocity*) və müxtəlif mənbələrdən (*variety*) toplanan çox böyük həcmdə (*volume*) verilənləri daha səmərəli istifadə etmək, saxlamaq, analiz edərək ondan daha qiymətli informasiyanı əldə etmək ideyasını özündə əks etdirir [8]. İngilis dilli mənbələrdə bunu «3V»lər də adlandırırlar. Analitiklər bəzən "5V"lər kimi təsvir edilən dördüncü – həqiqilik (*veracity*) və beşinci – dəyər (*value*) [2] xarakteristikalarını da qeyd edirlər.

Hazırda böyük həcmli verilənlərin saxlanılması, idarə olunması, analizi və vizuallaşdırılması üçün IBM, Microsoft, SAS, HP, EMC kimi informasiya texnologiyaları nəhəngləri tərəfindən paralel emalı təmin edən müxtəlif proqram-aparat həlləri mövcuddur. Problemin həlli məqsədi ilə İT sahəsinin nəhənglərindən olan Google şirkəti tərəfindən 2004-cü ildə Google File System və MapReduce [9] proqram-aparat platforması yaradılmışdır ki, bu da böyük verilənlər üzərində paralel proqramlaşdırmanın əsasıdır. Bunun əsasında paylanmış hesablama mühitində böyük verilənlərin emalı və analizi üçün açıq kodlu Apache Hadoop və Hadoop File System [10] proqram təminatları işlənmiş və bununla da "big data" texnologiyalarının əsası qoyulmuşdur. Bu gün operativ yaddaşda terabaytlarla informasiyanın analitik emalı üçün SAP şirkətinin – HaNa (High-performance Analytic Appliance), Oracle şirkətinin Oracle Exalytics, Oracle Exadata məhsulları mövcuddur. Bundan başqa Netezza, Teradata, Greenplum və s. şirkətlərinin ənənəvi relyasiya verilənlərinin idarə edilməsi sistemi əsasında terabaytlar və ekzabaytlarla verilənləri səmərəli emal edən proqram-aparat vasitələri işlənmişdir.

"Big data" texnologiyalarının da digər texnologiyalar kimi iki tərəfi: faydaları və təhlükələri vardır. Belə ki, böyük verilənlər bir tərəfdən cəmiyyətin bütün sahələrini kökündən dəyişə biləcək təsirə malik bilik mənbəyidir, biznes və analitikanın bütün aspektlərində yeni üfüqlər açır. Digər tərəfdən, biz nə qədər informasiyanı rəqəmsallaşdırırıqsa və əlavə informasiya toplayırıqsa, informasiya əlçatan olur və ondan istifadə edən subyektlərin sayı da çoxalır, bir o qədər də bədxahlar üçün potensial imkanlar yaratmış oluruq. İnformasiyanın təhlükəsizliyi baxımından bədxahlar tərəfindən informasiyanın oğurlanması, təhrif olunması və şəbəkələrin sındırılması, fərdi informasiyaların asanlıqla ələ keçməsi və s. kimi təhlükələr yaradır. İnsanların xəbəri və razılığı olmadan onlara məxsus fərdi məlumatların analiz olunması etik və hüquqi cəhətdən yolverilməzdir, təhlükəsizlik və gizlilik baxımından ciddi problemdir.

Problemlərə müxtəlif prizmadan baxmaq olar: informasiya təhlükəsizliyi üçün "big data" texnologiyalarının tətbiqi və ya "big data" texnologiyalarında informasiya təhlükəsizliyi [19]. Təqdim olunan məqalədə əsas hədəf, məhz ikinci yanaşmadır.

II. BIG DATA VƏ İNFORMASİYA TƏHLÜKƏSİZLİYİ

Böyük verilənləırin (BV) mövcud olması ilə onların intellektual analizi yeni keyfiyyətdə tədqiqat sahəsi kimi izlənilməyə başlandı. Real vaxta maksimum yaxın rejimdə verilənlərin analizinə olan tələbat müxtəlif parametrlər, xarakteristikalar, hadisələr və s. arasındakı korrelyasiyanı tapmağa, klassifikasiya və analitik hesabatlar və bunun əsasında proqnozların verilməsinə imkan verən böyük verilənlərin analitikasının (*Big Data Analytics*) yaranmasına gətirib çıxardı [11,12]. Gartner analitik şirkətinin tədqiqatında qeyd edildiyi kimi, BV-nin analizi cinayətlərin və təhlükəsizliyin pozulması hallarını üzə çıxarmaqda əsas rol oynayacaq, 2016 -cı ilə kimi böyük şirkətlərin 25%-i kiber-təhlükəsizlik sistemlərində "big data" analitikanı istifadə edəcəklər. Gartner ekspertləri hesab edirlər ki, Big Data analitikası informasiya təhlükəsizliyi sistemlərini daha etibarlı etməyə imkan verəcəkdir [20].

Hər şeydən əvvəl BV korporativ maraqlar baxımından biznes-proseslərin səmərəliliyini artırmağa imkan verir. BV-nin toplanması və analizinin köməyi ilə gəlirləri və xərcləri optimal idarə etmək, maliyyə göstəricilərini yaxşılaşdırmaq və şəffaflığı yüksəltmək mümkündür. "*İnsan-maşın*" və "*maşın-maşın*" (*machine-to-machine – M2M*) kimi ikitərəfli qarşılıqlı əlaqə nəticəsində müxtəlif mənbələrdən və müxtəlif formatda (*strukturlaşdırılmış və strukturlaşdırılmamış*) fasiləsiz olaraq generasiya olunan verilənlərin birgə analizi və onlardan yeni biliklərin və faydalı məlumatların əldə olunması yeni elmi kəşflərin edilməsində, dövlət və özəl təşkilatlarda əsaslandırılmış düzgün qərarların qəbul edilməsində, hüquq qaydalarının qorunmasında, milli təhlükəsizlik, terrorizm, xəstəlik epidemiyalarını əvvəlcədən söyləməyə, marketinq işlərinin yaxşılaşdırılmasına, insanların gizli davranışlarını üzə çıxarmaq, məqsəd və niyyətlərini anlamaq, onların digər insanlarla, ətraf mühitlə qarşılıqlı əlaqəsini başa düşməkdə, maliyyə sektorunda milli səviyyədə iqtisadi riskləri daha yaxşı anlamaq, siyasətçiləri və tənzimləyici orqanları istiqamətləndirmək və risk sistemlərini daha yaxşı idarə etməkdə bu texnologiyadan istifadə önəmlidir.

"Big data" texnologiyalarının tətbiqi mövcud təhlükəsizlik modellərinin köhnəldiyini üzə çıxarmışdır. 15 il əvvəl istifadə olunan təhlükəsizlik yanaşması bu gün üçün adekvat deyildir. Eyni zamanda böyük verilənlərin həcm, müxtəliflik və sürət kimi xarakteristikaları isə təhlükəsizlik və gizlilik problemini daha da kəskinləşdirir [1]. Geniş miqyaslı "bulud" infrastrukturu, verilənlərin mənbələrinin müxtəlifliyi, axın şəklində informasiyanın toplanması və böyük həcdə informasiyaların "buludlarda" miqrasiyası təhlükəsizlik sistemlərinnin "zəif" cəhətlərini üzə çıxarmışdır. Belə ki, BV-nin genişlənməsi ilə ənənəvi təhlükəsizlik mexanizmləri kifayət deyildir. Eyni zamanda verilənlərin axını çox çevik və sürətli təhkükəsizlik həlləri tələb edir.

CSA (*Cloud Security Alliance)* tərəfindən Big data sistemlərinin on təhlükəsizlik problemləri verilmiş və onları dörd qrup üzrə aşağıdakı kimi təsnifatlandırılmışdır [13]:

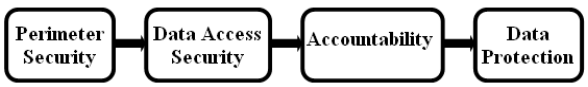
- infrastruktur təhlükəsizliyi (paylanmış proqram mühitinin təhlükəsizlik tədbirləri, qeyri-relyasiya tipli verilənlər xəzinəsi üşün ən yaxşı təhlükəsizlik təcrübəsi);
- verilənlərin gizliliyi (gizliliyi saxlayan miqyaslanan və kompozit data mining və analitika; kriptoqrafiya ilə həyata keçirilən girişə nəzarət və təhlükəsiz kommunikasiya; girişə qranular nəzarət)
- verilənlərin idarə edilməsi (verilənlərin saxlanması və ötürülməsi jurnalını möhkəmləndirmək; verilənlərin mənşəyi, qranular auditlər);
- tamlıq və reaktiv təhlükəsizlik (real-vaxt rejimində təhlükəsizliyin monitorinqi, girişin yoxlanılması/ filtirlənməsi).

Təsnifatdan da göründüyü kimi "big data" sitemlərinin təhlükəsizlik infrastruktutunu təmin etmək üçün paylanmış hesablamalar və verilənlər xəzinəsi mühafizə olunmalıdır. Hər şeydən əvvəl saxlanılan və gizli informasiyanın özünün təhlükəsizliyi üçün kriptoqrafiya və qranular giriş nəzarət vasitələrindən istifadə olunmalıdır. Böyük həcmi informasiyanın idarə olunması verilənlər xəzinəsi üçün verilənlərə effektiv nəzarət və mənşəyini müəyyən etmək üçün miqyaslanan və paylanmış həllər tələb edir. Nəhayət, müxtəlif nöqtələrdən axın şəklində daxil olan verilənlərin tamlığı yoxlanmalı və real-vaxt rejimində təhlükəsizlik insidentləri üzrə analitik təhlil aparılmalıdır.

Təhlükəsizlik və gizlilik problemlərinin həllində, adətən üç məsələnin yerinə yetirilməsi lazımdır:

1. *Modelləşdirmə*: kiber-hücum və ya verilənlərin sızmaları senarilərini əhatə edən təhlükəsizlik modelini formalizasiya etmək;
2. *Analiz*: təhlükəsizlik modeli əsasında mümkün həllər tapmaq;
3. *Realizasiya*: tapılmış həlli mövcud infrastrukturda tətbiq etmək.

Təhlükəsizliyi təmin etmək üçün bütün Hadoop məhsullarında dörd səviyyəli təhlükəsizlik (*4-Layer Security*) modeli tətbiq edilir [14] (şək.1).



Şəkil 1. Hadoop sistemində verilənlərin təhlükəsizlik modeli

Təhlükəsizlik perimetri (*Perimeter Security*): istifadəçinin autentifikasiyasına cavab verir və şəbəkə təhlükəsizliyi problemini Massaçusets Texnologiya Universiteti tərəfindən yaradılmış Kerberos şəbəkə protokolu vasitəsilə yerinə yetirir. Kerberos klient-server tətbiqlərində gizli-açar şifrlənməsi (*secret-key cryptography*) vasitəsi ilə etibarlı autentifikasiyanı təmin etmək üçün nəzərdə tutulmuşdur. Yəni, informasiya sistemlərinin təhlükəsizıiyini təmin etmək üçün şəbəkə üzrə autentifikasiya və dayanıqlı kriptoqrafiya alətləri təqdim edir. Həqiqiliyin yoxlanmasında de-fakto standart hesab olunan Kerberos, hər kəs üçün açıqdır əlyetərlidir [19].

Verilənlərə əlyetərlik (*Data Access Security*): istifadəçilərin ancaq verilənlərə çıxışa icazəsinin olduğunu,

xidmət və resurslara isə olmadığını təmin etmək üçün nəzərdə tutulur.

Hesabatlılıq *(Accountability):* bu təhlükəsizlik səviyyəsinin ümumi məqsədi hesabatlılığı təşviq etməkdir. Hadoop-da administratorlara verilənləri nəzarət və çıxışı audit etməkyə imkan verir. Əlavə olaraq verilənlərin mənşəyini, yəni verilənlərin hansı mənbədən daxil olduğunu müəyyən etməyə imkan verir. Bu səviyyənin təhlükəsizliyini dəstəkləmək üçün Navigator adlanan xüsusi məhsul vardır.

Verilənlərin qorunması *(Data Protection)*: sonuncu təhlükəsizlik aspektidir və verilənlərin şifrlənməsi, maskalanması və daha çox sahəni əhatə edir [21].

III. BIG DATA VƏ FƏRDİ MƏLUMATLARIN QORUNMASI

Təxminən 30 il bundan əvvəl bəlkə də insanların şəxsi həyatını mühafizə etmək və anonimliyi təmin etmək nisbətən asan idi. Çünki, şəxsi informasiyaları toplayan və saxlayan avtomatlaşdırılmış sistemlər az idi, İnternet çox primitiv idi, hətta onun haqqında bilənlərin sayı belə az idi. Son illər isə hər şey dəyişmiş, rəqəmsallaşma *(GPS siqnalları, mobil telefonlar, e-mail, elektron alqı-satqı, sosial şəbəkələrdəki yazışmalar, elektron tibbi yazılar və s.)* geniş miqyas almışdır. Marketing, bank, sığorta, tibb və s. sahələrdə müştərilər haqqında toplanan informasiya fərdi məlumatların gizliliyi və anonimliyini keçmişdə qoymuşdur.

"Big data" müstəsna faydalı informasiya versə də təhlükəsizlik və gizlilik ilə bağlı yeni etik problemlər yaradır. İstər şirkətlər, istərsə də dövlət təşkilatları tərəfindən insanlar haqqında toplanan informasiya artıq nəzarətdən çıxmış, açıq və gizli verilənlər arasındakı sərhəd silinməkdədir.

Fərdi məlumatlar açıq və konfidensial kateqoriyalara bölünür. Açıq fərdi məlumatlar kateqoriyasına müəyyən olunmuş qaydada adsızlaşdırılmış, subyekt tərəfindən açıq elan olunmuş və ya ümumi istifadə üçün yaradılmış informasiya sisteminə subyektin razılığı ilə onun barəsində daxil edilmiş məlumatlar aiddir. Şəxsin adı, soyadı və atasının adı daimi açıq fərdi məlumatdır. Konfidensial məlumatlardan fərqli olaraq, açıq kateqoriyalı fərdi məlumatların gizliliyinin təmin edilməsi tələb olunmur. Konfidensial fərdi məlumatlar qanunla müəyyən olunmuş hallar istisna olmaqla, üçüncü şəxslərə yalnız subyektin razılığı əsasında verilə bilər.

Verilənlər bilavasitə insan fəaliyyətini əks etdirir. İnsanlar müalicə götürürlər, mal və xidmət alırlar, veb saytlar axtarır, telefon zəngləri edirlər. Onların hansı coğrafi məkanda olduqları, ailə üzvləri ilə əlaqələri, siyasi fəaliyyəti, sosial dairələri və maraqları öz smart telefonları və istifadə etdikləri proqram əlavələri ilə daim izlənilir və şirkətlər tərəfindən toplanır. Bu gün, bu məlumatların böyük hissəsi fərdin razılığı olmadan toplanır və istifadə edlir. Belə ki, sığorta və banklar müştərilərinin borcu olduğunu öyrənməklə onlara kredit verməkdən imtina edir, marketlər valideynlərdən qabaq qızın hamilə olduğunu müəyyən edirlər, həkim öz pasientinə elektron-tibbi məlumatlar əsasında hətta xəstənin gizlətmək istədiyi informasiya malik olur [15].

Yarana biləcək təhlükələrdən biri də odur ki, insanlar təsadüf nəticəsində cinayətkar kimi məsuliyyətə cəlb olunur. Belə ki, böyük həcmdə multimedia (*audio,video*) verilənlərinin yayılması, sosial şəbəkələrə yükləməyə imkan verən onlayn proqramlar, istənilən insident zamanı təsadüfən kadra düşən günahsız yoldan keçənlərin hüquqları pozulur. Məsələn, Bostonda partlayışların təhqiqatı zamanı bir neçə nəfər terror aktı yerindəki fotoşəkillərin sosial şəbəkə saytlarında yerləşdirildiyinə görə şübhəlilər sırasına düşmüşdü.

Bu gün müasir informasiya texnologiyaları faktorları: böyük verilənlər, analitika və bulud texnologiyalarını bir-birindən ayrı təsəvvür etmək mümkün deyildir. Bulud texnologiyaları saxlanma, böyük hesablamaların aparılmasında son dərəcə müvəffəqiyyətli yanaşmalardandır [16]. Lakin, gizlilik problemi insanlara şəxsi məlumatlarını ümumiyyətlə, informasiyanı buludlarda saxlamağa mane olmaqda davam edir. Bu problem BV-nin intellektual analizi və analitikasının (*big data mining and analytics*) inkişafı ilə daha da ciddiləşmişdir [17]. Çünki, burada relevant nəticə almaq üçün fərdiləşdirilmiş və lokal-bazaya əsaslanan xidmətlər kimi şəxsi informasiya tələb edir. Şəxsə aid informasiya çox dəqiq yoxlamalara və profilləşmə narahatlığına, oğurlanma və itmə kimi şərtlərə məruz qalır [18].

Şəxsi verilənlərin təhlükəsizliyinə cavabdeh təşkilatlar adətən de-identifikasya üsullarından o cümlədən, anonimlik (*anonymization*), təxəllüs (*pseudonymization*), şifrləmə (*encryption*), açar-kodlaşdırma (*key-coding*) və s. istifadə edirlər. Anonimlik ad, ünvan və sosial təhlükəsizlik nömrələrini silməklə gizliliyi təmin edirsə, təxəllüs bu informasiyanı ləqəb və süni identifikasiya ilə əvəz edir. Açar-kodlaşdırma şəxsi informasiyanı kodlaşdırır və onun dekodlaşdırılması üçün açar yaradır.

İnsanların şəxsi verilənlərini "Big Data" sistemlərinə təqdim etməsi istəyinin müəyyən məqamları vardır. Bunun üçün aşağıdakı təhlükəsizlik həlləri təmin olunmalıdır:

- "Big Data"-nın təhlükəsiz və gizliliyin qorunması ilə toplanması və emalı;
- "Big Data" analizinin təhlükəsiz mühitdə və gizliliyin qorunduğu şəkildə həyata keçirilməsi;
- "Big Data" sistemləri üçün verilənlərin saxlanılması və saxlanma siyasətinin təhlükəsiz (*və gizlilik rejimində*) tətbiqi.

Əks təqdirdə istifadəçilər verilənlərini "Big Data" sistemlərinə təqdim etməkdə tərəddüd edirlər.

Konfidensial fərdi məlumatlar qanunvericilikdə nəzərdə tutulmuş tələblərə uyğun səviyyədə mülkiyyətçi, operator və bu məlumatlara giriş hüququ olan istifadəçilər tərəfindən mühafizə olunmalıdır. Fərdi məlumatların toplanılması, işlənilməsi və mühafizəsi dövlət tərəfindən hüquqi aktlarla tənzimlənməlidir.

NƏTİCƏ

Biz hazırda ekzabayt və zetabaytlarla böyük verilənlər axınının istehsalını təmin etmiş elm, texnika və texnologiyaların geniş yayıldığı erada yaşayırıq. "Big Data" texnologiyaları tətbiq etməklə bir çox elm sahəsində (*fizika, astronomiya, tibb, geologiya və s.*) tez bir zamanda böyük nailiyyətlər, idarəetmədə və biznes fəaliyyətində gəlir, rəqabətdə müəyyən üstünlüklər əldə etmək mümkündür. Bütün bunlara baxmayaraq, verilənlər çoxaldıqca, onlardan istifadə edən subyektlərin sayı da çoxalır. Verilənlərin əksəriyyəti fərdi məlumat olduğundan xüsusi mühafizə olunmalıdır. İnsanların xəbəri olmadan onlar haqqındakı verilənlərin analiz olunması yeni problemlər yaradır. Ona görə də bu texnologiyaların köməyi ilə emal və analiz olunan informasiyanın təhlükəsizlik məsələləri isə istər tədqiqatçıların, istərsə də istehsalçıların diqqət mərkəzində olmalıdır. "Big Data" texnologiyalarının müasir cəmiyyətə təsirini ümumiləşdirərək deyə bilərik ki, istənilən yeni texnologiya kimi o da, "bütün dərdlərə dərman deyil"dir. Kifayət qədər güclü bir alət olsa da müəyyən nöqsanları, məhdudiyyətləri və təhlükələri vardır. Bəşəriyyət üçün bir sıçrayış, elm və texnikadan başlayaraq biznesə qədər inkişafa doğru atılmış bir addımdır.

ƏDƏBİYYAT

[1] http://www.emc.com/leadership/digital-universe/iview/big-data-2020.htm

[2] Worldwide Big Data Technology and Services 2013–2017 Forecast. http://www.idc.com

[3] F. Diebold. "Big Data dynamic factor models for macroeconomic measurement and forecasting." Discussion Read to the Eighth World Congress of the Econometric Society, 2000.

[4] Big Data Research and Development Initiative.www.whitehouse.gov/

[5] The Australian Public Service Big Data Strategy. www.finance.gov.au/big-data

[6] Big data: The next frontier for innovation, competition, and productivity. Analyst report. McKinsey Global Institute, May 2011. http://www.mckinsey.com/.

[7] Data is the New Oil of the Digital Economy. 2014. http://www.wired.com

[8] D. Laney, "3D data management: Controlling data volume, velocity and variety," Technical report. META Group, Inc (now Gartner, Inc.), February 2001. http://blogs.gartner.com/

[9] J. Dean, S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Proceedings of the Sixth Symposium on Operating System Design and Implementation, vol.6, 2004, pp.137–150.

[10] Hadoop. http://hadoop.apache.org/.

[11] K. Kambatla, G. Kollias, V. Kumar, A. Grama, "Trends in Big Data Analytics," Parallel Distributed Computing, vol. 74, no.7, pp. 2561–2573, 2014.

[12] X. Wu., X. Zhu, G. Q. Wu, W. Ding, "Data mining with big data," IEEE Transactionson Knowledge and Data Engineering, vol.26, no. 1, pp.97-107, 2014.

[13] Cloud Security Alliance: "Expanded Top Ten Big Data Security and Privacy Challenges." 2013.

[14] T. Malbrecht, Z. Prekopcsak, "Big Data security on Hadoop." www.rapidminer.com

[15] C.Duhigg. "How companies learn your secrets." New York Times, 2012

[16] D. Agrawal, S. Das, Amr El Abbadi, "Big Data and Cloud Computing: Current State and Future Opportunities," Proc. of the 14th International Conference on Extending Database Technology, pp.530-533, 2011.

[17] D. Che, M. Safran, Z. Peng, "From big data to big data mining: challenges, issues, and opportunities," B. Hong, X. Meng, L.Chen, W. Winiwarter, W.Song (Eds.), Database Systems for Advanced Applications, Springer, Berlin Heidelberg, pp. 1–15, 2013.

[18] O. Tene, J. Polonetsky, "Big Data for all: Privacy and user control in the age of analytics," Northwestern Journal of Technology and Intellectual Property, vol.11, no.5, pp.239-273, 2013.

[19] Big data and privacy, MIT 2013. http://bigdata.csail.mit.edu

[20] http://cloudtimes.org/2014/02/12/gartner-report-big-data-will-revolutionize-the-cybersecurity-in-next-two-year/

[21] Big data and data protection. https://ico.org.uk/media/for-organisations/documents/1541/big-data-and-data-protection.pdf