**UOT 004.02**

***Əliquliyev R.M.*[1], *Hacırəhimova M. Ş.*[2]**
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]rasim@science.az, [2]makrufa@iit.ab.az

# "BIG DATA" FENOMENİ: PROBLEMLƏR VƏ İMKANLAR

*Məqalə "big data" fenomeninə həsr olunur. Məqalədə "big data" anlayışının mahiyyəti və böyük verilənlərin mənbələri, bu texnologiyanın imkanları, problemləri və mövcud yanaşmalar tədqiq olunur. 3V konsepsiyası, böyük verilənlərin analizi məsələləri araşdırılır. Bu konsepsiyanın reallaşdırılmasında mövcud proqram-aparat məhsulları analiz olunur.*

***Acar sözlər:*** *big data, data science, big data analytics, NoSQL, MapReduce, Hadoop, OLAP.*

## Giriş

İnternet bütün dünyada köklü dəyişiklikliyi – sənaye cəmiyyətindən informasiya cəmiyyətinə (İC) keçidi təmin etmişdir. Mütəxəssislər baş verən dəyişiklikləri haqlı olaraq daş dövrünün bitməsi ilə müqayisə edirlər. O dövrdə insanlar mədənlərdən metal çıxarmağı öyrənmişdilərsə, indi isə verilənlərdən informasiya əldə edirlər. Bu gün İC-nin əlamətləri hər tərəfdə hiss edilməkdədir. Hər kəs mobil telefon istifadəçisidir, hər kəsin evində kompüter var, hər müəssisədə böyük informasiya sistemləri və s. mövcuddur. Kompüterlər insan fəaliyyətinin bütün sferalarına nüfuz etdikcə, insanların İnternetə çıxış imkanları genişləndikcə, bu şəbəkə vasitəsilə göstərilən xidmətlərin sayı artdıqca, bulud hesablamaları, sensor texnologiyaları, saysız-hesabsız media qurğuları, ümumiyyətlə, rəqəmsal texnologiyaların səhiyyə, astronomiya, bioinformatika, nəqliyyat, dövlət idarəçiliyi və s. kimi sahələrdə geniş tətbiqi verilənlər axınının artmasına gətirib çıxarmış, dünya sanki informasiya ilə doldurulmuşdur. 2000-ci illərdən "big data" fenomeni meydana çıxmışdır."Big data" ingilis dilindən Azərbaycan dilinə tərcümədə "böyük verilənlər" (BV) kimi başa düşülür. Bu yeni termin həcm və mürəkkəblik baxımından mövcud idarəetmə metodları və intellektual analiz vasitələri ilə emal oluna bilməyən verilənləri təyin etmək üçün istifadə edilir.

İnkişaf etməkdə olan İC böyük elmi-texniki inqilab çağındadır və bu inqilab, biznesdən başlayaraq elmə qədər həyatımızın bütün sahələrinə nüfuz etməkdədir. Bu eyni zamanda sosial-iqtisadi, təhlükəsizlik, elmi baxımdan mürəkkəb bir problemdir. Problem böyük verilənlər ideyası ilə daha da dərinləşməkdədir. Belə ki, BV-nin saxlanması, idarə edilməsi, onlardan dəyər yaradılması ciddi problem yaratmışdır. Problem müxtəlif mənbələrdən avtomatik və fasiləsiz olaraq generasiya olunan verilənlərin real-vaxt ərzində emalı və analizində mövcud İT (*İnformasiya Texnologiyaları*) həllərin səmərəsiz olmasındadır. Mövcud vəziyyət "verilənləri yaratmaq son dərəcə asan, emal etmək isə son dərəcə çətin olur" fikrini söyləməyə əsas verir. Ona görə də bu texnologiyanın elmi-tədqiqat obyekti kimi öyrənilməsi vacibdir, aktualdır.

BV Google, Amazon, Facebook kimi nəhəng kompaniyalar, CERN (*European Organization for Nuclear Research*), NASA (*National Aeronautics and Space Administration*) tərəfindən həyata keçirilən elmi layihələrdə analiz edilməkdədir. IDC (International Data Corporation), McKinsey Global İnstitutu, Gartner və s. kimi kompaniyaların son tədqiqatlarında BV böyüyən, dinamik inkişaf edən sahə kimi təqdim edilir. Onlar "big data"nı 2013-cü ildə əsas texnoloji istiqamət və 2014-cü ildə informasiya kommunikasiya texnologiyaları (İKT) sahəsinin lokomotivi adlandırırlar [1–4].

## "Big data" anlayışı və mahiyyəti

Mürəkkəb informasiya sistemləri və verilənlərin analitik təhlilinə artan tələbat nəticəsində zamanın yeni çağırışı – "*big data"* termini son illərdə daha fəal surətdə leksikonumuza daxil olmaqdadır. Əslində "Big data" yeni fikir deyildir. Belə ki, kompüterlər yarandıqları ilk illərdə

ancaq hesablama məsələlərinin həllinə xidmət edirdilərsə, keçən əsrin 70-ci illərindən sonra onlar adi hesablayıcı qurğulardan verilənləri emal edən universal texnologiyaya çevrilmişlər. O vaxtdan "*data product"*, "*data tool"*, "*data application"*, "*data science"*, "*data scientist"* və hətta verilənlərləri geniş oxucu kütləsinə çatdıran "*data journalist"* kimi yeni terminlər meydana gəlməyə başlamışdır. "Big data" termini isə ilk dəfə 1998-ci ildə Silicon Graphics kompaniyasının kompüter elmləri üzrə mütəxəssisi Con Meşi ("*Big Data and the Next Wave of InfraStress*" adlı seminarda məruzə) tərəfindən istifadə edilmişdir [5, 6]. Bu terminə bir qədər sonra, 2000-ci ildə, F.Dieboldun akademik mühitdə dərc olunan tədqiqatında rast gəlinir [7]. Lakin termin, 100 ildən çox tarixə malik (1869-cu ildən nəşr olunan) həftəlik "Nature" jurnalının 2008-ci il sentyabr ayının "big data" mövzusuna həsr olunmuş xüsusi nömrəsində professor Klifford Linçin (*Clifford Lynch*) "Böyük verilənlər. Sizin verilənlər necə böyüyür?" ("*Big Data: How do your data grow*?") adlı məqaləsi ilə populyarlıq qazanmışdır [8]. Məqalədə böyük verilənlərlə işləmək imkanları verən texnologiyaların elmin gələcəyinə necə təsir edə bilər?" sualına cavab tapmağa cəhd olunur, verilənlərin elmdə, xüsusilə elektron elmdə (e-elm) rolu göstərilir. Məqalənin nəşrindən ötən illər ərzində termin biznes cəmiyyətləri nümayəndələri tərəfindən və akademik mühitdə intensiv istifadə olunmağa başlanmışdır. Bu sahəyə maraq isə 2011-ci ildən artmağa başlamışdır [9]. Mütəxəssislər BV-ni bəzən mineral resurslarla – böyük filiz mədəni (*largeore*), yeni neft (*new oil*), gizli biliklərin mənbəyi kimi "data mining", təbiət kataklizmləri (*data tornado, data deluge*), təbii fəlakət (*sunami*) və s. ilə müqayisə edirlər. Onlar qeyd edirlər ki, verilənlərin dəyərini başa düşənlər üçün XXI əsrdə verilənlər XVIII əsrdə istifadə olunmamış neft kimi çox qiymətli aktiv hesab olunur [10, 11]. 2012-ci ildə Davosda Beynəlxalq İqtisadi Forumda BV valyuta və ya qızıl kimi yeni sinif iqtisadi aktiv kimi bəyan edilmişdir [12]. Bəzən bu problemi Mur qanunu kimi də interpretasiya edirlər [13].

Verilənlərin mənimsənilməsi və istifadəsində yeni eranı əks etdirən bu anlayış nisbi anlayışdır və zaman-zaman dəyişə bilər. Yəni, bu gün üçün "böyük" hesab olunan informasiya, sabah üçün normal sayıla bilər [14]. Bu nisbilik həm də hesablama gücü ilə təyin olunur. Aydındır ki, hər bir kompüterin disk həcmi müxtəlifdir və onların yaddaş tutumu, məhsuldarlığı durmadan artmaqdadır. Bununla bərabər, verilənlərin həcmi də eksponensial olaraq artır. Hazırda BV həcmi 1 terabaytdan (*1tb=1024geqabayt*) başlayaraq daha böyük həcmli verilənlər (1*zetabayt=1024ekzabayt*) hesab olunur.

Vahid unifikasiya olunmuş tərif olmadığından elm, sənaye və kütləvi informasiya vasitələri kimi maraqlı tərəflər çox vaxt bir-birini təkzib edən müxtəlif təriflər təqdim edirlər [2,3,5,15–17]. Razılaşdırılmış tərifin olmaması "big data" fenomeninin anlaşılmasında qeyri-müəyyənlik yaradır. BV dedikdə, çox böyük, çox sürətli və emal üçün çox mürəkkəb verilənlər başa düşülür. Forrester konsaltinq kompaniyası butexnologiyanın mahiyyətini aşağıdakı kimi formalizə edir: " "Big data" özündə çox böyük həcmdə verilənlərdən məzmun çıxaran texnika və texnologiyaları birləşdirir". Kaliforniyanın Berkli Universitetinin kompüter elmləri üzrə professoru M.Franklinə görə: "böyük verilənlər – onlarla işləmək üçün böyük xərc tələb edən və onlardan çətinliklə informasiya əldə edilən verilənlərdir" [15]. McKinsey Global İnstitutunun "Böyük verilənlər: innovasiyalar, rəqabət və məhsuldarlıq üçün növbəti hüdud" adlı analitik hesabatında isə göstərilir ki, bu termin o verilənlər dəstinə aid edilir ki, onun həcminə informasiyanın daxil edilməsi, saxlanılması, idarə edilməsi və analizi üçün tipik verilənlər bazasının imkanları çatmır [2]. Tərifdən görünür ki, burada meyar kimi ancaq verilənlərin həcmi nəzərdə tutulur. Belə olan halda həcm problemi heç də yeni deyildir. Bu verilənlər bazası sahəsindəki çoxdan mövcud olan mövzudur ki, paylanmış verilənlər bazası, resursların birgə istifadəsi arxitekturunun yaranması da bu məsələnin həllinə xidmət etmişdir [18]. Oxşar tərif IDC-nin tədqiqatında da verilir. Tərifdə göstərilir ki, "Big Data" texnologiyaları yeni nəsil texnologiya, arxitektura olmaqla, böyük həcmli verilənlərdən çox kiçik zamanda böyük sürətlə analiz aparmaqla bilik əldə etmək və analizin nəticələrini təqdim etmək üçündür. Böyük verilənləri verilənlərin özü, verilənlərin analitikası və analitikanın nəticələri kimi xarakterizə

edirlər [3]. Bu tərif növbəti bölmədə şərh edəcəyimiz 3V modelinə [19] əsaslanır. Amerika Birləşmiş Ştatları Milli Standartlaşma və Texnologiyalar İnstitutu (NIST - ***National Institute of Standards and Technology***) isə iddia edir ki, BV dedikdə "həcmi mövcud üsul və sistemlərin imkanlarını ötüb keçən verilənlər" təsəvvür edilir. "Böyük" məfhumu mövcud hesablama səviyyəsinə görə nisbidir [16]. Massaçusets Texnologiya İnstitutunun professoru S. Madden "böyük verilənlər"i "çox böyük", "çox sürətli", "cox çətin" kimi təsvir edir [17]. Burada "çox çətin" ifadəsi mövcud alətlərlə emal oluna bilməyən verilənlərə aiddir. Stuart Ward və Adam Barker bu ideyaları ümumiləşdirərək böyük verilənləri böyük və ya mürəkkəb verilənlər dəstinin saxlanması və analizi NoSQL, MapReduce, maşın təlimi ilə məhdudlaşmayan bir sıra üsulların köməyi ilə təsvir edilən termin kimi təyin etmişlər [20]. Göründüyü kimi, BV-nin anlaşılmasında 3V modelini daha çox istifadə olunan və geniş yayılmış tərif kimi qəbul etmək olar. Çünki bu model uyğun texnologiya və məhsullara olan tələbləri əldə etmək üçün BV-ni daha yaxşı xarakterizə edir.

**"Big data" mənbələri**

Son onillikdə informasiyaya əlyetərliyin asanlaşması və kommunikasiya vasitələrinin çoxalması nəticəsində hər il rəqəmsal informasiyanın həcmi həndəsi silsilə ilə artır. Statistik rəqəmlərə nəzər salaq. Bəşəriyyətin mövcudluğundan 2003-cü ilə qədərki dövrdə dünyada cəmi 5 ekzabayt məlumat generasiya olunduğu halda, 2012-ci ildə rəqəmsal informasiyanın həcmi 500 dəfə artaraq 2.7 zetabayt olmuşdur. Dünyada informasiyanın həcminin 2015-ci ildə üç dəfə artması, növbəti hər il 40% artaraq, 2020-ci ildə 44 zetabayta çatacağı proqnozlaşdırılır [3]. Bu da Yer kürəsində hər nəfərə düşən 5200 geqabayt informasiya deməkdir. IBM (**International Business Machines)** kompaniyasının tədqiqatında isə informasiyanın 90%-nin son iki ildə yaradıldığı bildirilir [21].

İnternetə çıxış imkanlarının genişlənməsi son on il ərzində yaradılan və toplanan informasiyanın həcminin həqiqətən də çox böyük sürətlə artmasında bir canlanma yaratmışdır. Birləşmiş Millətlər Təşkilatının (BMT) 2014-cü ilin əvvəlinə olan məlumatına görə, Yer kürəsinin əhalisi 7.138 milyarda çatmışdır. 2013-cü ilin sonunda dünyada İnternet istifadəçilərinin sayı 2.7 milyard təşkil etmişdir. Beynəlxalq Telekommunikasiya İttifaqının proqnozlarına görə, 2014-cü ilin sonuna 3 milyard insan və ya planetin əhalisinin 42,3%-i İnternet istifadəçisi olacaqdır [14]. IBM-in məlumatına görə, dünyada hər gün 2.5 trilyon bayt məlumat hazırlanır, hər dəqiqədə 100 milyon e-mail göndərilir, Google axtarış sistemində 2 milyon axtarış sorğusu, Facebook sosial şəbəkəsində 350 gb məlumat emal olunur və 570-dən çox veb sayt yaradılır, hər dəqiqə 72 saatlıq yeni video YouTube internet-servisinə yüklənir və s. Bu rəqəmlər həqiqətən də verilənlərin həcminin çox böyük olduğunu bir daha sübut edir [21].

BV-nin əsas mənbələrini elmi eksperimentlər, sensor və sosial şəbəkələr, dövlət qurumlarının agentlikləri və portalları, iqlim haqqında məlumat ötürücüləri, nəqliyyatın intellektual idarə edilməsi sistemləri, GPS (*Global Positioning System*) siqnalları, GIS (*Geographic Information System*) sistemlər, böyük kompaniyaların bazaları, elektron poçt, smartfonlar vasitəsilə alınmış rəqəmsal foto və videolar, böyük satış mərkəzləri, bank əməliyyatları və s. təşkil edir. Bir sözlə, BV-nin mənbələri hədsiz dərəcədə çoxdur. Verilənlərin potensial mənbələrinin müxtəlifliyi artdıqca, qismən strukturlaşdırılmış və strukturlaşdırılmamış verilənlərin həcmi də artır. BV-nin mənbələri əsasən beş kateqoriyaya bölünür: veb verilənlər və sosial media; maşınlar vasitəsilə yaradılan verilənlər (machine-to-machine –M2M); böyük tranzaksiya verilənləri; elmi verilənlər; insanların yaratdığı verilənlər [22, 23].

***Veb verilənlər və sosial media.*** Veb, analitik təhlillər üçün çox zəngin, verilənlərin mənbələrinə görə isə çox fərqlidir. Bu mənbələrə veb səhifələr, onlayn məqalələr, bloqlar və s. daxildir [23]. Bu mənbələrin əsas hissəsini strukturlaşdırılmamış verilənlər (*mətn, video, təsvirlər*) təşkil edir və onların əksəriyyəti bir-biri ilə hipermətnlər vasitəsilə əlaqələnir. Veb verilənlərin bir qismini isə strukturlaşdırılmış metaverilənlər təşkil edir və adətən onlar qraf-

formadadır [23]. Digər veb-mənbə isə log fayllardır ki, bu informasiyalar da qismən strukturlaşdırılmış tipdədirlər. Bu informasiyaları istifadə etməklə şirkətlər istifadəçilərin davranışları əsasında müəyyən işləri həyata keçirə bilərlər. Sonuncu tip veb mənbə isə sosial şəbəkələrdir (*Facebook, YouTube, Twitter, WhatsAppvə s.*). Burada insanlar özləri haqqında doğru və ya yalan informasiyalar yazırlar.

***Maşınlar vasitəsilə yaradılan verilənlər.*** BV əsasən insanlar və maşınlar tərəfindən yaradılsa da, bunlardan sonuncusu əsas mənbə sayılır. Maşınlar həm informasiya istehsalçısı, həm də istehlakçısı olduğundan, verilənlərlə interaktiv işi deyil, müşahidələrin emalı zamanı böyük zəhmət tələb edən işlərin avtomatlaşdırılmasını, verilənlərin axtarışını və verilənlərə əlyetərliliyi təmin edən proqram təminatı tələb olunur ki, maşın maşınla (M2M) işləyə bilsin. Bu tip verilənlərə onlayn rejimdə istifadəçilərin davranışlarını izləyən sensorlar, video kameralar, texniki qurğuların və s. göstəricilərindən toplanmış informasiya daxildir. Əşyaların interneti (*Internet of Things*) ideyası M2M kommunikasiyasına ən parlaq nümunədir [22].

***Böyük tranzaksiya verilənlərini*** əsasən banklar, iri internet-mağazalar vasitəsilə edilən alqı-satqı əməliyyatları (*məsələn,Walmart supermağazalar şəbəkəsi saatda 2.5 petabayt klient əməliyyatları emal edir*)**,** onlayn ödənişlər zamanı yaranan strukturlaşdırılmış və qismən strukturlaşdırılmış verilənlər təşkil edir.

***Elmi verilənlərə*** LHS (Large Hydron Collider), LSST (Large Synoptic Survey Telescope), əl izləri, genetik analizlər və s. zamanı əldə olunmuş verilənlər və s. daxildir.

Dünyada insanlar özlərinin gündəlik fəaliyyətləri nəticəsində çoxlu sayda verilənlər generasiya edirlər. ***İnsanların yaratdıqları verilənlərə*** elektron məktublar, sənədlər, icmallar, qeydlər və s.daxildir. Soares insanlar tərəfindən yaradılan verilənləri iki kateqoriyaya: veb və tranzaksiya verilənlər qrupuna aid edir [22]. Bu da onu göstərir ki, verilənlər bir və ya bir neçə kateqoriyada təsnif oluna bilər.

**“Big Data” texnologiyaları**

**“Big Data” texnologiyalarının yaranmasını şərtləndirən əsas amillər.** Hər şeydən əvvəl “Big Data” texnologiyaları verilənlərin həcminin çox böyük sürətilə artması–eksponensial inkişafla ifadə olunan “*informasiya partlayışı*” fazasının yaranması ilə əlaqədardır. İnformasiyanın həddindən artıq çoxalması informasiya yükü problemi yaradır. Bu gün “*informasiya partlayışı*”nın xarakterik nümunələri çoxdur. Bu nümunələr haqqında bir qədər sonra danışacağıq.

İkincisi, biznes strukturlarda biznes-proseslərin informasiyalaşdırılması elmi təşkilatlarda ölçmələrin aparılmasında yeni imkanların yaranması, dövlət qurumlarında, həmçinin ictimai kommunikasiya şəbəkələrində xidmətlərin sayının və funksionallığının artması ilə bağlıdır. Yəni, bu o vaxta təsadüf edir ki, real-vaxt rejimində müxtəlif formatlı çox böyük ölçülü verilənlər massivini emal edə bilən İT həllər yaradılmış [24–28] və korporativ istifadəçilər üçün əlçatan olmuşdur. Sürətlə artan verilənlərin öhdəsindən gəlmək məqsədi ilə İT sahəsinin nəhənglərindən olan Google şirkəti tərəfindən File System Google [24] və MapReduce [25] proqram-aparat platforma yaradılmışdır. Bunun əsasında açıq kodlu Apache Hadoop və Hadoop File System [26–28] proqram təminatları işlənmiş və bununla da BV texnologiyalarının əsası qoyulmuşdur.

Üçüncü, BV artıq Amerika və bir sıra qərb dövlətlərində elmi ictimaiyyət, biznes-cəmiyyətləri, hökumət strukturları tərəfindən neft qədər strateji resurs kimi dəyərləndirilir, bu sahədəki problemlərə çox böyük önəm verilir. Onu demək kifayətdir ki, ABŞ prezident administrasiyası 2012-ci ilin martında BV sahəsində “Böyük verilənlərin tədqiqi və inkişafı” (*Big data Research and Development İnitiative*) təşəbbüsünü elan etmişdir [29]. Təşəbbüs çərçivəsində BV texnologiyalarının ABŞ dövlət siyasətinin aparıcı istiqamətlərində istifadəsi üçün kompleks tədbirlərin keçirilməsi *(konfransların, forumların və s.*), layihələrin işlənməsi nəzərdə tutulmuş, böyük həcmli rəqəmsal verilənlərin təşkili və analizi üşün aidiyyatı dövlət agentliklərinə (N*SF-National Science Foundation, NIH-National Institutes of Health, DoD -*

*Department of Defense, DoE- Department of Energy, DARPA- Defense Advanced Research Projects Agency və USGS- US Geological Survey*) 200 milyon dollar həcmində vəsait ayrılmışdır. 2013-cü ildə Yaponiya hökuməti BV üzrə milli proqramın işlənməsini dərc etmiş, Avstraliya "*Australian Public Service Big Data Strategy"* strategiyasını qəbul etmişdir. Böyük Britaniya, Almaniya, Çin və s. kimi dövlətlərdə də BV strateji əhəmiyyətli vacib resurs kimi dəyərləndirilməkdədir.

IDC mütəxəssisləri isə əsas amilləri "hard" disklərin qiymətinin ucuzlaşması, sensor texnologiyaların genişlənməsi, bulud (*cloud*) texnologiyaların və verilənlərin saxlanmasının virtuallaşdırılması və infrastrukturlarının köməyi ilə informasiya resursarına əlyetərliyin mümkünlüyü, eyni zamanda innovativ tətbiqi əlavələr və analitik alətlərin mövcud olması ilə əlaqələndirirlər [3].

Hazırda böyük verilənlər mövzusu istər biznes sahəsində, istərsə də elmi mühitdə son dərəcə populyardır. Nüfuzlu beynəlxalq təşkilatlar, elmi qurumlar tərəfindən çox böyük həcmdə informasiyanın emalının müxtəlif aspektlərinə həsr olunmuş çoxsaylı konfranslar, simpoziumlar, seminarlar, forumlar keçirilməkdədir. Müzakirə və tədqiq olunan əsas mövzular: böyük verilənlərin arxitekturu (Big Data Architecture), idarə edilməsi (big data management), modelləşdirilməsi (Big Data Modeling), analitikası (Big Data Analytics), alətləri (Big Data Toolkits), açıq platformalar (Big Data Open Platforms), "Big Data"xidmət kimi (Big Data As a Service), biznesin səmərəli idarə edilməsi (Big Data in Business Performance Management), e-dövlətdə və cəmiyyətdə böyük verilənlərin analitikası (Big Data Analytics in e-Government and Society), vizuallaşdırma (Visualization),təhlükəsizlik (security), böyük verilənlər üçün alqoritmlərdir və s.

Elmi və populyar jurnalların: Nature (2008), Science (2011), Computer (2013) və s. xüsusi nömrələri bu mövzuya həsr olunmuşdur. Son illərdə işıq üzü görmüş böyük verilənlərin elmi-nəzəri problemlərini işıqlandıran "Journal of Big Data", "International Journal of Big Data" (IJBD), International Journal of Big Data Intelligence (IJBDI), Big Data Research, Big Data & Society və s. kimi akademik jurnalların nəşri isə bir daha bu sahədəki problemlərin nə qədər aktual olmasından xəbər verir.

**"Big data" alətləri və texnoloji həllər.** Böyük verilənlərin emalı onlardan faydalı informasiyanı əldə etmək üçün son dərəcə səmərəli hesablama gücü və mükəmməl analitik imkanlara malik texnologiyalar tələb ediir. Hazırda böyük həcmli verilənlərin saxlanılması, idarə olunması, analizi və vizuallaşdırılması üçün IBM [30], Oracle [31, Microsoft [32], SAS [33], SAP [34], HP [35], Teradata, EMC və s., kimi informasiya texnologiyaları nəhəngləri tərəfindən paralel emalı təmin edən müxtəlif proqram-aparat həlləri mövcuddur. Dünyada çoxlu sayda informasiya sistemlərinin (*axtarış, GMail, Google Maps, Google Earth, Big Query və s.*) yaradıcısı Google şirkəti tərəfindən 2004-cü ildə ***MapReduce*** paylanmış hesablama modeli təqdim olunmuşdur. Bu model BV üzərində paralel proqramlaşdırmanın əsasıdır [24,25]. Onun əsas ideyası, böyük və mürəkkəb verilənləri kiçik hissələrə bölməklə emal etməkdir. MapReduce-un işi iki mərhələdən (Map və Reduce) ibarətdir. "Map" mərhələsində giriş verilənləri ilkin emal üçün əsas qovşağa (*master node*) göndərilir və orada digər kömpüterlər (*worker node*) arasında paylanılır. "Reduce" mərhələsində əsas qovşaq emal olunmuş verilənləri işçi qovşaqlardan toplayır və onun əsasında məsələnin həllinin nəticəsi formalaşdırılır.

Böyük verilənlərin de-fakto standartı hesab olunan Apache Software Fondation-un layihəsi ***Hadoop*** daha geniş yayılmış texnologiyadır, paylanmış hesablama mühitində böyük verilənlərin emalı və analizi üçün əsas platformadır, MapReduce modelinin açıq kodlu (*open access*) sistemidir və 1000 qovşaqdan çox miqyaslaşmanı təmin edir [26]. Hadoop iki əsas komponentdən ibarətdir: Hadoop MapReduce [27] и Hadoop Distributed File System (HDFS) [28]. Burada MapReduce paralel hesablamalara, HDFS paylanmış fayl sistemi isə verilənlərin idarə edilməsinə cavab verir.

***NoSQL (****Not Only SQL)* bu gün BV aləminin əsası hesab olunur və verilənlərin idarə edilməsinin miqyaslılıq (*scalability*), əlyetərlilik (*availability*) və verilənlərin uyğunlaşdırılması (*consistency*) kimi problemlərinin həllində tətbiq olunur [36]. Ədəbiyyatlarda paylanmış sistemlərin bu üç xüsusiyyəti Berkli universitetinin professoru Eric Brewer tərəfindən təklif olunmuş CAP (*Consistency, Availability u Partition Tolerance*) teoremi kimi də tanınır.

Ənənəvi informasiya xəzinələri çoxölçülü analiz (OLAP), klassifikasiya, klasterizasiya və s. alətlər dəstini təqdim edir, bu gün isə operativ yaddaşda terabaytlarla informasiyanın analitik emalı üçün SAP kompaniyasının HaNa (*High-performance Analytic Appliance*), Oracle kompaniyasının Oracle Exalytics, Oracle Exadata məhsulları mövcuddur. Bundan başqa, Netezza, Teradata, Greenplum və s. kompaniyalarının ənənəvi relyasiya verilənlərinin idarə edilməsi sistemi əsasında terabaytlar və ekzabaytlarla verilənləri səmərəli emal edən proqram-aparat alətləri vardır.

Müasir İT faktorları: böyük verilənlər, analitika və bulud texnologiyalarını bu gün bir-birindən ayrı təsəvvür etmək mümkün deyildir. Buludlarda saxlamaq, buludlarda hesablamalarsız böyük verilənlərlə işləmək mümkün deyildir. Çünki geniş miqyaslı və çoxsəviyyəli saxlama sistemlərinə artan diqqət və tələbat, bulud texnologiyalarının real olaraq mövcudluğu, həm də BV-nin analitikasına marağı artırmışdır. Qeyd etmək lazımdır ki, bulud texnologiyaları böyük hesablamaların aparılmasında son dərəcə müvəffəqiyyətli yanaşmalardandır. Burada böyük həcmli rəqəmsal informasiya IaaS *(Infrastructure as a service*), PaaS (*Platform as a service*), SaaS (*Software as a service*) "bulud" xidmətləri vasitəsi ilə mərkəzləşdirilmiş qaydada idarə olunur və saxlanılır [37].

**"Big data" texnologiyalarının faydaları.** Böyük verilənlərin də digər texnologiyalar kimi iki tərəfi: zərərləri və faydaları vardır. Birinci ilə mübarizə edərkən, ikincini yaddan çıxarmaq olmaz. "İnformasiya sunamisi" adlandırılan heterogen xam verilənlər cəmiyyətin bütün sahələrini kökündən dəyişə biləcək təsirə malik bilik mənbəyidir. Ona görə də yeni-yeni elmi kəşflərə imza atmaq, iqtisadi inkişafa nail olmaq, innovasiyaları stimullaşdırmaq məqsədi ilə bu verilənlər biliyə çevrilməlidir. BV texnologiyalarının faydasını göstərmək üçün 2009-cu ildə BMT-nin "Global Puls" təşəbbüsünü qeyd etmək lazımdır [38]. Məqsəd böyük verilənlərin imkanlarından istifadə etməklə BMT və onun partnyorlarına dayanıqlı inkişaf üçün yeni yanaşmaların axtarılmasına kömək etməkdir. Təşəbbüs çərçivəsində bir çox layihələr işlənilməkdədir.

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. Telemetrik qurğular vasitəsi ilə "*insan-maşın*" və "*maşın-maşın*" kimi ikitərəfli qarşılıqlı əlaqə nəticəsində müxtəlif mənbələrdən və müxtəlif formatlarda (*strukturlaşdırılmamış, zəif strukturlaşdırılmış və strukturlaşdırılmamış)* toplanan verilənlərin birgə analizi və onlardan yeni biliklərin və faydalı məlumatların əldə olunması yeni elmi kəşflərin edilməsində, dövlət, hökumət təşkilatları və özəl kompaniyalarda mühüm qərarların qəbul edilməsində, hüquq qaydalarının qorunmasında, sosial təminat, milli təhlükəsizlik, səhiyyə məsələlərində çox önəmlidir. Böyük verilənlərin analizi istehlakçıların alıcılıq qabiliyyətini öyrənməklə marketinq işlərinin yaxşılaşdırılmasına, insanların gizli davranışlarını üzə çıxarmaq, məqsəd və niyyətlərini anlamaq, onların digər insanlarla, ətraf mühitlə qarşılıqlı əlaqəsini başa düşməkdə kömək edə bilər.

BV-dən savadlı şəkildə istifadə edildikdə, hadisələrə operativ və dəqiq reaksiya göstərmək, düzgün qərar vermək mümkündür. Vətəndaşlarına xidmət etməkdə, milli problemlərin (səhiyyə, terrorizm, iş yerlərinin yaradılması, təbii fəlakətlər və s.) həllində və təhlükələri əvvəlcədən aşkarlamaqda hökümətin bu texnologiyadan istifadəsi vacibdir.

Böyük verilənlər maliyyə sektorunda milli səviyyədə iqtisadi riskləri daha yaxşı anlamaq, siyasətçiləri və tənzimləyici orqanları istiqamətləndirmək və risk sistemlərini daha yaxşı idarə etməkdə kömək ola bilər.

Tibbi aspektdən BV-in analitikası səhiyyə sistemində inqilabi dəyişikliklərə imkan verir. Əməliyyatların səmərəliliyinin yaxşılaşdırması, xəstəlik epidemiyalarını əvvəlcədən söyləməyə, səhiyyə sahəsinə xərcləri optimallaşdırmağa, kliniki sınaqların manitorinqinin keyfiyyətini yaxşılaşdırmağa kömək edə bilər.

Qeyd olunan üstünlüklərə baxmayaraq, verilənlər çoxaldıqca, onlardan istifadə edən subyektlərin sayı da çoxalır. Verilənlərin əksəriyyəti fərdi məlumat olduğundan xüsusi mühafizə olumalıdır. İnsanların xəbəri olmadan onların haqqındakı verilənlərin analiz olunması etik və hüquqi cəhətdən yolverilməzdir.

**“Big data” problemləri**

Böyük verilənlərin problemləri əsasən çox sürətlə böyüyən böyük həcmli informasiyanın real vaxt rejimində emalı, axtarışı, təsnifatlandırılması, analizi ilə bağlıdır. Tətbiq olunduğu sahələrdən asılı olmayaraq böyük verilənləri təsvir etmək üçün ümumi xarakteristikalar mövcuddur. Bu xarakteristikalar böyük verilənlərin əsas problemlərini özündə əks etdirməklə üç əsas qrupa bölünür: həcm (*volume*), sürət (*velocity*) və müxtəliflik (*variety*). İngilis dilli mənbələrdə bunu «3V» də adlandırırlar. Bu parametrlərin konvergensiyası böyük verilənləri təyin etməyə və digər verilənlərdən fərqləndirməyə kömək edir. Bu model ilk dəfə 2001-ci ildə D.Laney tərəfindən verilmişdir [19]. O, “böyük verilənlər” terminini istifadə etməsə də, elektron kommersiyada bir tendensiyanı: verilənlərin idarə edilməsinin daha vacib, daha çətin olacağını əvvəlcədən xəbər vermiş və daha sonra verilənlərin idarə edilməsində verilənlərin ölçüsünü, ötürülmə sürətini və müxtəlifliyini əsas problem kimi təyin etmişdir. Bu xarakteristikalar isə ümumilikdə “big data” texnologiyalarının əsas konsepsiyasını təşkil edir. Bu konsepsiya çox böyük sürətlə və müxtəlif mənbələrdən toplanan çox böyük həcmdə verilənləri daha səmərəli istifadə etmək, saxlamaq, analiz edərək ondan daha qiymətli informasiyanı əldə etmək ideyasını özündə əks etdirir. Qeyd etmək lazımdır ki, analitiklər bəzən “5V” (şək. 1) kimi təsvir edilən dördüncü – həqiqlilik (*veracity*) [21] və beşinci – dəyər (*value*) [1, 2] xarakteristikalarını da qeyd edirlər.


Şəkil 1. Böyük verilənlərin əsas xarakteristikaları

***Həcm.*** Həcm böyük verilənlərin ən əsas xarakteristikasıdır. Aydındır ki, böyük verilənlər çox böyük informasiya massivlərinin toplanması ilə xarakterizə olunur ki, bu da hazırda istənilən təşkilatlatın qarşılaşdığı problemdir. Hazırda BV-nin miqyası terabaytlardan zetabaytlara qədər həcm ilə xarakterizə olunur (*on il əvvəl informasiyanın həcmi terabaytlarla, daha sonra peta və*

*ekzabaytlar, hazırda isə zetabaytlarla ölçülməkdədir)* [3,39]. Həcm problemi ilk növbədə saxlama problemi yaradır ki, bu da genişmiqyaslı saxlama və paylanmış emal tələb edir. Bu gün saxlama məsələsinin həllində informasiyanın qurğular arasında miqrasıyasını həyata keçirən bir sıra texnologiyalar: DAS (*Direct-Attach-Storage*), NAS (*Network Attached Storage*), SAN (*Storage Area Networks*), HSM (*Hierarchical Storage Management*), ILM (*Information Life-cycle Management*)) mövcuddur [40,42]. Son zamanlar isə saxlama qurğularının yaddaş tutumunun artması, çoxsaylı kompüterlərin (server, kompüter və s.) hesablama və yaddaş resurslarının klasterləşməsi və virtualllaşdırılmasını həyata keçirməklə, verilənlərin emalı və yadda saxlanılmasına xidmət edən "grid" və "*cloud computing"* texnologiyalarının tətbiqi saxlama sahəsindəki problemləri demək olar ki, aradan qaldıra bilmişdir [37,43]. Həcmindən asılı olaraq BV üç qrupa bölünür [39,44–46]:

- Tez (sürətli) verilənlər (Fast Data) – onların həcmi terabaytlarla ölçülür;
- Böyük analitika (Big Analytics) - onların həcmi petabaytlarla ölçülür;
- Dərinə nüfuzetmə (Deep Insight) - onların həcmi ekzabaytlarla və zetabaytlarla ölçülür.

Qruplar bir-birindən yalnız verilənlərin həcminə görə deyil, həm də onların keyfiyyətli emalına görə fərqlənirlər. Statistik rəqəmlərdə əks olunan verilənlərin həcmi bir daha mütəxəssisləri bu sahədə yeni metod və alətlərin işlənməsinə sövq edir.

***Sürət.*** Həcm artdıqca, emal üçün də çox böyük sürət tələb olunur. Burada iki hal nəzərdə tutulur. Birinci, yeni verilənlər böyük sürətlə generasiya olunur, mövcudlar yenilənir və toplanır. İkincisi, sürət zaman problemi kimi dəyərləndirilir və mövcud emal texnologiyalarının verilənləri real vaxtda analiz etmək imkanına malik olması ilə izah olunur [23]. Bu işdə verilənlərin relyasiya idarəetmə sistemləri kifayət deyildir. Bu məsələdə şəbəkənin ötürücülük qabiliyyəti də xüsusi əhəmiyyət daşıyır.

***Müxtəliflik.*** Müxtəliflik BV-nin təbii özəlliklərindəndir. BV ilə işləmək ancaq verilənləri saxlamaq üçün böyük həcmli qurğular deyil, eyni zamanda böyük hesablama gücü də tələb edir. Əsas problem ondan ibarətdir ki, məlumatların əksəriyyəti çox vaxt müxtəlif mənbələrdən (*e-poçt, sosial şəbəkələr, GPS – koordinatları, müxtəlif sensorlardan, veb-saytlar və s.*) müxtəlif formatlarda daxil olur və müxtəlif indeksləşmə sxemi istifadə olunur. Ənənəvi relyasiya verilənlər bazasının sətir və sütunlarında ifadə olunmuş strukturlaşdırılmış (məsələn, maliyyə verilənləri) verilənlərlə bərabər, informasiyalar strukturlaşdırılmamış – mətn, video-audio *fayllar, təsvirlər və s.* tiplərdə olur. Bu tip verilənlər isə dünyada bütün informasiyanın 80-90%-ni təşkil edir. Bunları sadəcə olaraq, bir araya yığmaq və birgə emal etmək və analiz üçün uyğun şəklə salmaq çox çətin olur [47,48].

Toplanan məlumatların ***həqiqiliyi*** də önəmlidir. Çünki, əsasında qərar qəbul edəcəyimiz verilənlər nə qədər dəqiq və ya şübhəlidir? Məsələn, ötürücülər vasitəsilə alınan verilənlər daha etibarlıdır, nəinki sosial media verilənləri.

Digər bir faktor isə BV-nin ***dəyər*** yaratması xüsusiyyətinin olub-olmamasıdır. BV əlavə dəyər yaratmırsa, "məlumat zibilliyi"nə çevrilir. Məhz "big data"-nın xüsusilə biznes qurumları tərəfindən diqqətdə olması əlavə dəyər yaratma xüsusiyyətinə görədir. Buna görə də bu faktor marketinq xarakteristikası kimi qiymətləndirilir [1,2]. Çünki, informasiyanın qiyməti ancaq bizim onu necə istifadə etməyimizlə təyin olunur.

Özündə son dərəcə faydalı informasiyanı daşıyan, adi relyasiya bazalarının emal edə bilmədiyi yüzlərlə terabayt və ekzabayt həcmində mətn, təsvirlər, audio-video və s. tip strukturlaşdırılmamış informasiyanın toplanması və idarə edilməsi, saxlanması, təhlükəsizliyi, axtarışı, analizi (*analitik hesabatların generasiyası və vizuallaşdırılması, proqnozlaşdırma*) və s. kimi məsələlərin həllində yeni texnologiyalar, yanaşmalar, daha mükəmməl analiz üsulları tələb edir.

**“Big data” analitikası**

***Mövcud yanaşmalar.*** Biz hazırda ekzabayt və zetabaytlarla böyük verilənlər axınının istehsalını təmin etmiş elm, texnika və texnologiyaların geniş yayıldığı erada yaşayırıq. Elmi sahədə böyük verilənlər artır, çünki, indi elmi tədqiqatlar nəzəri düşüncələrdən daha çox elmi eksperimentlərə köklənmişdir [47–49]. Bu eksperimentlərin nəticəsində (*əsasən fizika, astronomiya və tibbkimi elmi sahələrdə*) isə çox böyük həcmdə rəqəmsal verilənlər yaranır. Söhbət, LHS (Large Hydron Collider), LSST (Large Synoptic Survey Telescope), Hubble teleskopu və s. kimi petabaytlarla informasiya verən layihələrdən gedir. Biznes sahəsində BV massivi yaranır, ona görə ki, hazırda insan fəaliyyətinin böyük hissəsi İnternetdədir, onlayn rejimdədir [47]. Zaman keçdikcə, verilənlərin həcminin artması və real zamanda onların analizinə olan tələbat BV-nin ən əsas problemlərindən sayılan böyük verilənlərin analitikasının (*Big Data Analytics*) yaranmasına gətirib çıxardı [30,45,51]. “Big Data Analytics” daha böyük və mürəkkəb massivlərə tətbiq edildiyində,n kəşf edən analitika (*Discovery Analytics*) və izah edən analitika (*Exploratory Analytics*) terminlərindən də istifadə edilir [45]. Necə adlandırılmasından asılı olmayaraq, mahiyyət eynidir – qərar qəbul edən şəxsləri müxtəlif proseslər haqqında məlumatlarla təmin edən əks əlaqəni yaratmaq.

Məlumdur ki, analiz müxtəlif parametrlər, xarakteristikalar, hadisələr və s. arasındakı korrelyasiyanı tapmağa, təsnifat və analitik hesabatlar və bunun əsasında proqnozların verilməsinə imkan verir [23]. Bu aspektdən müasir texnologiyalar verilənlərdəki informasiyanın yeni biliklərə çevrilməsinə və ya biliklərin əldə edilməsinə imkan verməlidirlər. BV-in saxlanması, emalı və analizi üçün böyük hesablama gücü, miqyaslılığı təmin edən arxitektur yanaşmalar, vahid infrastruktur tələb olunur [50].

BV mövzusu yeni olduğundan, onun ətrafında çoxlu mübahisələr mövcuddur. Məsələn, biznes analitika və BV analitikası arasındakı fərq. Biznes analitika ilə BV-in analitikası eyni məqsədə xidmət etsələr də, onlar üç aspektə görə bir-birindən fərqlənirlər [39, 46]:

- BV, onun tərifinə uyğun olaraq, daha böyük həcmdə informasiyanın emalı üçün nəzərdə tutulur.
- BV daha sürətlə alınan (*bəzən nəticələrin veb səhifənin yüklənməsindən də sürətli formalaşdırılması*) və dinamik dəyişən məlumatların emalı üçün nəzərdə tutulur.
- BV strukturlaşdırılmamış məlumatların emalı üçün nəzərdə tutulur.

Biznes analitika müəyyən zaman çərçivəsində biznes tərəfindən əldə olunan nəticələrin analizinin təsvir olunma prosesidir. Halbuki, böyük verilənlərin sürətli emalı və analizi gələcəkdən xəbər verməyə imkan verir, gələcək üçün biznesə tövsiyələr təklif etməyə qadirdir.

Mütəxəssislər böyük verilənlərin analizində iki yanaşmanı qeyd edirlər: saxlamaq və analiz etmək (*store and analyze*); analiz etmək və saxlamaq (*analyze and store*) [45]. Birinci halda verilənlərə analitik alətləri tətbiq etməzdən əvvəl tranzaksiya sistemlərindən (*OLTP- Online Transaction Processing*) alınan xam verilənlər emal olunur və xəzinəyə (*datawarehouse*) yüklənir. Məhz bu prinsip əsasında yaradılmış ənənəvi analitik həllər (*OLAP - Online Analytical Processing*) böyük verilənlərin analizi üçün nəzərdə tutulmadığından, uyğunsuzluq meydana çıxır.

Məlumdur ki, *superkompüterlər* saniyədə böyük sayda (*flops*) hesab əməliyyatlarını yerinə yetirmək imkanına malikdirlər və ancaq strukturlaşdırılmış verilənlərlə işləyərkən faydalıdırlar. Bu da superkompüterlərin tətbiqləri üzərinə məhdudiyyətlər qoyur. Onlar daha çox elmin bitib-tükənmədiyi modelləşdirmədə səmərəlidir. Con Hopkins universitetinin professoru, astronom A.Şali elmdə kompüterlərin tətbiqini ancaq hesablamaların sürətləndirilməsində deyil, böyük verilənlərin üzərində analitik işlərin təkmilləşdirilməsində görür. Onun fikrincə elmdə və texnikada “scop” sonluğu ilə bitən (*məsələn, mikroskop, teleskop, periskop, intraskop və s.*) çoxlu sayda qurğular vardır. İndi də verilənlər axınının analizi üçün bir hipotetik qurğu lazımdır ki, o verilənləri xüsusi saxlanc yerlərində toplamadan, yığım yerinə daha yaxın yerdə dərhal

emal etsin. Şali bu qurğunu *"DataScope"* adlandırır. Superkompüterlərdən fərqli olaraq, DataScope eksperimental verilənlərlə işləmək məqsədi daşıyır. [48]-də Şali və onun həmkarı Cim Qrey verilənlərin müasir elmdə rolu və eksperimental verilənlərin həcmi ilə elmi üsulların qarşılıqlı münasibətinin təkamülünə baxmışlar. Onlar elmdə istifadə olunan hesablama sistemlərində çoxnüvəli prosessorların hesablama gücü ilə saxlama sistemlərinin potensialı arasında disbalans olduğunu qeyd edirlər. Məqalədə onlar həmçinin elmdə verilənlərin rolunun üç postulatını verirlər: 1) elmi tədqiqatlar daha çox eksperimental verilənlərlə işləməyə yönəlir; 2) verilənlərin emalı problemləri miqyaslaşmanın tətbiqi ilə şərtləndirilir; 3) verilənləri hesablamaya deyil, hesablamaları verilənlərə yaxınlaşdırmaq lazımdır. Ümumiyyətlə, petabytlarla verilənlər massivlərinin emalının zəruriliyi *"Data – İntensive Computing"* [52] yanaşmasını ortaya çıxarmışdır ki, bu "hesablamalar deyil, verilənlər daha əsas sərvətdir" anlamına gəlir. Postulatların realizasiyası yolunda emal olunan verilənlərin həcmi operativ yaddaşın ölçüsünü keçərsə, onda ən məhsuldar saxlama sisteminin qoşulması kömək edə bilər. Problemin həlli kimi onlar həm də disk, prosessor və şəbəkə ilə işləyən miqyaslılıq arxitekturuna malik infrastruktur təklif edirlər.

Böyük verilənlərin analizi sahəsində ROLAP (*Relational On-Line Analytical Processing*), MOLAP (*Multi-Dimensional On-Line Analytical Processing*), HOLAP (*Hybrid Online Analytical Processing*) kimi ən müasir texnologiyalar mövcuddur [45]. Bunlardan hansısa birinin seçilməsi verilənlərin yenilənməsindən asılıdır.

ROLAP relyasiya bazası əsasında yaradılmışdır və bütün əməliyyatları dəstəkləyir. ROLAP sistemlər daha şəffaf və öyrənilmiş sistemlərdir və onların aşağıdakı çatışmazlıqları vardır:

- verilənlərin daxil edilməsi mərhələsinə nəzarət etmək mümkün deyil;
- statistika yığmaq və indekslərin saxlanması üçün optimal struktur seçmək olmur;
- yüksək giriş/çıxış sürətini təmin etmək üçün verilənlərin diskdə optimal yerləşdirilməsi mümkün olmur;
- aralıq aqreqatlaşdırılmış qiymətləri keşləşdirmək olmur;
- analitik sorğuların yerinə yetirilməsi zamanı yüksək sürətlilik tələb olunduğuna görə, dərin statistik analiz və həllin optimal planını seçmək mümkün deyildir.

MOLAP verilənləri həqiqətən də çoxölçülü modeldə təsvir edir, ancaq daxildə ROLAP-dan "ulduz" və "qardənəciyi" sxemini istifadə edir və ROLAP-ın çatışmazlıqları aradan götürlmüşdür. Bu da çox yüksək sürətli analiz etməyə imkan verir. Böyük verilənlər üzərində klassik metodların, "B-three" və daha mürəkkəb strukturların tətbiqi mümkündür. MOLAP və HOLAP qapalı sistemlərdir və kommersiya məhsullarının nou-xau sahəsinə aid edilir.

Böyük həcm, sürət və mürəkkəblik kimi xarakteristikalarla təyin olunan böyük verilənləri mövcud metodologiyalarla və ya alətlərlə idarə etmək və onlardan faydalı informasiyanı əldə etmək böyük verilənlərin analizində ciddi problemdir. Strukturlaşdırılmamış verilənlərin (mətn, aidio, video və s.) daha dərin intellektul analizi (*mining*) [46] və analizin nəticələrinin vizuallaşdırılması [53, 54] BV analitikanın əsas məsələlərindəndir. Problemin həllində data mining texnologiyaları sinfindən olan klassifikasiya, klasterləşdirmə, neyron şəbəkələr və s. kimi üsullar tətbiq olunur [55].

**"Big Data" və təhsilin yeni məsələsi.** Purdue Universitetinin professoru, statistika, verilənlərin vizuallaşdırılması, maşın təlimi sahəsində tanınmış mütəxəssis Uilyam Klivlend tərəfindən təklif edilmiş "data science" terminin yaranması (2001) ilə "data scientist" adlanan mütəxəssislərə tələbat bu kateqoriyadan olan kadrların hazırlanmasına marağı stimullaşdırmışdır. Təsadüfi deyildir ki, təxminən həmin vaxtdan da Elm və Texnika üçün Məlumat Komitəsi (CODATA – International Council for science: Committee on Data for Science and Technology) – elm və texnika üçün ədədi verilənlərin toplanması, qiymətləndirilməsi və saxlanılması üzrə beynəlxalq şuranın "Data Science Journal"ı (2002-ci ildən) nəşr olunmaqdadır.

"O´Reilly Radar" jurnalında nəşr olunan "Verilənlər haqqında elm nədir?" (What is Data Science?) məruzəsinin müəllifi Mayk Lukidis yazırdı: "Gələcək, verilənləri məhsula çevirə biləcək insan və kompaniyalara məxsus olacaq". Bu deyim, məşhur bir kəlamı yada salır: "Kim informasiyaya malikdirsə, o da dünyaya sahibdir". Bu gün bu aforizmə bir az düzəliş vermək olar: "Dünyanı verilənlərə və onların analizi texnologiyalarına malik olanlar idarə edir". İnformasiyanın alınması üsullarına sahib olmaq verilənlər haqqında elm (*data science*) adlanır. "*Data science"* termini hərfi mənada "*verilənlər haqqında elm" kimi tərcümə olunmamalıdır. Çünki, ingilis dilində "science"sözü* ancaq "*elm*" deyil, həm də "bacarıq", "məharət", "qabiliyyət" deməkdir. Demək ki, bu bilik və bacarıqlara əsaslanan elmdir. Verilənlər haqqında elm ənənəvi informatikadan başlayaraq riyaziyyata qədər müxtəlif sahələr üzrə bacarıq və vərdiş tələb edir. Bu sahə ilə məşğul olmaq üçün "*data scientist*" adlanan daha təkmilləşmiş mütəxəssislərə ehtiyac vardır. 2013-cü ildən başlayaraq Dandi Universitetində (Şotlandiya), Oklend Universitetində (Yeni Zelandiya), London İmperial Kollecində, Cənubi Kaliforniya Universitetində, Vaşinqton, Berkli, Nyu-York Universitetlərində verilənlər haqqında magistr proqramları tədris olunur. Bu problem təhsil məsələsidir.

**"Big data" perspektivləri**

Qeyd olunduğu kimi, verilənlər hər il həndəsi silsilə ilə artmaqdadır. Proqnozlar isə bu prosesin hələ davam edəcəyindən xəbər verir [1–3]. Bu baxımdan BV-nin saxlanılması, idarə edilməsi və analitikası sahəsində tədqiqatçıların və praktiklərin yaxın illərdə məşğul olacaqları problemlər çoxdur:

*Arxitektura.* Etibarlı, dayanıqlı, genişmiqyaslı, minimal texniki tələbləri olan və s. əsas xüsusiyyətlər tələb olunur.

*Analitika və mining.* Həddindən artıq böyük verilənlərlə işləyərkən səhv etməmək və çoxsaylı sorğuları vaxtında cavablandırmaq üçün statistik və daha dərin analitik təhlillər üçün yeni-yeni metod və alqoritmlərin işlənməsi gərəkdir. Əlbəttə verilənlərin çox da sadə olmayan analizi üsulları mövcuddur, lakin mətn, video və audio informasiyaların analizi ciddi problemdir. Problemin həllinə nail olmaq üçün daha mükəmməl nəzəri və praktik üsullara ehtiyac vardır.

*Vizuallaşdırma*. Vizuallaşdırma nəticələrin təqdim olunması üçün BV-nin analizində əsas məsələlərdəndir. Verilənlər həddindən çox olduqda, daha münasib və uyğun şəkildə vizuallaşdırmaq çətin olur. Yeni üsul və mexanizmlər lazımdır.

*Faydalı verilənlərin aşkarlanması*. İnformasiyanın hamısı faydalı olmadığından, böyük verilənlərin seçilməsi çox vacibdir. IDC-nin tədqiqatında [3] 2012-ci ildə saxlanılan bütün informasiyanın (898 ekzabayt) təxminən 18%-nin (158 ekzabayt) böyük verilənlər kateqoriyasına düşdüyü göstərilir və bu informasiyadan cəmi 3% faydalı verilənlər kimi analiz olunur.

Maddi (*xammal*) və qeyri-maddi (*verilənlər və ya informasiya*) resursların emalının inteqrasiya olunduğu "informasiya iqtisadiyyatı və ya rəqəmsal iqtisadiyyat" dövründə iqtisadiyyat (*bütün tarixi dövrlərdə olduğu kimi*) bazar vasitəsilə reallaşdırılan əks-əlaqə olmadan mövcud ola bilməz.

Əlbəttə, bazar əks əlaqə üçün yeganə vasitə deyildir. Əks-əlaqə prinsipi əsasında yaradılan sistemlər ayrılıqda götürülmüş təşkilatı deyil, həm də bütövlükdə milli iqtisadiyyatı kökündən dəyişə bilər. "Big data" texnologiyaları bazarı istehsalçıların proqram, aparat təminatı və xidmət üzrə satışdan əldə etdikləri gəlirlərlə ölçülür. IDC və IIA (*International Institute of Analytics*), Wikibon cəmiyyətinin tədqiqatlarında dünyada "big data" texnologiyaları bazarının həcminin (*2012-ci ildə proqram, aparat təminatı, xidmət üzrə*) 5 milyard dollardan bir qədər artıq təşkil etdiyi, 2014-cü ildə 16,1 milyard dollar olaçağı göstərilir. Bunlardan proqram təminatı 24%, xidmətlər 29%, saxlama sistemləri isə 45% təşkil edir. 2012-2017-ci illər ərzində gəlirlərin 50.1 milyard dollar olacağı proqnozlaşdırılır [1-4]. Wikibonun 2013-cü ildəki tədqiqatında [56] (70 kompaniya üzrə) dünyada "Big data" texnologiyaları nəhənglərinin gəlirləri göstərilmişdir.

Tədqiqatdan göründüyü kimi, böyük verilənlər bazarına IBM (*$1,368 milyon*), HP *($869 milyon*) vəDell *($652 milyon*) kompaniyaları liderlik edir. Həmin tədqiqatda həmçinin göstərilir ki, 2015-ci ildə bu sahədə 4,4 milyon IT iş yerlərinin yaradılması gözlənilir, onlardan 1,9 milyonu ABŞ-da olacaqdır.

**Nəticə**

BV böyük informasiya massivlərini istifadə etməyə imkan verən yeni nəsil texnologiyadır. Böyük həcm, sürət və müxtəliflik kimi xüsusiyyətlərlə xarakterizə olunan verilənlərin emalı və analizi məqsədilə yaradılmış *MapReduce, Hadoop, HDFS, NoSQL və s.* proqram-aparat platformaları mövcuddur, onlar tətbiq edilməkdə və inkişaf etdirilməkdədir. Böyük həcmli verilənlərlə işləməkdə superkompüterlər, qrid və bulud texnologiyaların köməyi ilə müəyyən problemlər aradan qaldırılmış olsa da, bu sahədə hələ də ciddi texniki və elmi-nəzəri problemlər möcuddur. Problem ancaq böyük həcmdə verilənlərin saxlanılması və idarə olunmasında deyil, həm də strukturlaşdırılmamış verilənlərin analizi və nəticələrin interpretasiyasındadır. BV-i mükəmməl analiz etməklə bir çox elmi sahələrdə tez bir zamanda yüksək nailiyyətlər, idarəetmədə və biznes fəaliyyətində gəlir, rəqabətdə müəyyən üstünlüklər əldə etmək mümkündür.

**Ədəbiyyat**


1. Worldwide Big Data Technology and Services 2013–2017 Forecast, (http://www.idc.com)
2. Big data: The next frontier for innovation, competition, and productivity. Analyst report, McKinsey Global Institute, May 2011. http://www.mckinsey.com/
3. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe/
4. Beyer M. A. and Laney D. The importance of big data: A definition. Stamford, CT: Gartner, 2012.
5. Diebold F. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.
6. Lohr S. The Origins of ‘Big Data’: An Etymological Detective Story . http://bits.blogs.nytimes.com/2013/
7. Diebold F. Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting. / Discussion Read to the Eighth World Congress of the Econometric Society, 2000.
8. Clifford L. Big data: How do your data grow? // Nature, 2008, vol.455, pp.28–29.
9. Google Trends for Big Data, 2013.
10. Is Data The New Oil? http://www.forbes.com/sites/perryrotella/2012/04/02/
11. Data Is the New Oil of the Digital Economy. http://www.wired.com/2014/07/
12. Big Data, Big Impact: New Possibilities for International Development, 2012. *www.weforum.org*
13. Moore's law applied to big data. http://www.datasciencecentral.com/forum/
14. Big Data: Big today, normal tomorrow, ITU-T Technology Watch Report, November 2013.
15. https://amplab.cs.berkeley.edu/
16. NIST Big Data Working Group (NBD-WG). http://bigdatawg.nist.gov/home.php.
17. Madden S. From Databases to Big Data // IEEE Internet Computing, 2012, vol.16, issue 3, pp.4–6.
18. Witt D., Gray J. Parallel Database Systems: The Future of High Performance Database Systems // Communications of the ACM, 1992, 35(6), pp. 85–98.
19. Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical report, META Group, Inc (now Gartner, Inc.), February 2001. http://blogs.gartner.com/
20. Ward J.S. and Barker A. Undefined By Data: A Survey of Big Data Definitions. http://arxiv.org/pdf/1309.5821.pdf

21. What is big data? - Bringing big data to the enterprise, 2013. http://www-01.ibm.com/
22. Soares S. Big Data Governance - An Emerging Imperative. MC Press Online, LLC, 1st edition, 2012.
23. Chen J., Chen Y., Xiaoyong D., et.all. Big data challenge: a data management perspective // Frontiers of Computer Science in China, 2013, 7(2), pp.157–164.
24. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters/ Proceedings of the Sixth Symposium on Operating System Design and Implementation, volume 6 of OSDI ’04, Berkeley, CA, USA, 2004, pp.137–150.
25. Ghemawat S., Gobioff H. and Leung S.T. The Google File System / Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, SOSP ’03, New York, USA, October 2003, pp.29–43.
26. Hadoop, http://hadoop.apache.org/
27. Hadoop MapReduce. http://hadoop.apache.org/docs/stable/mapred_tutorial.html
28. Hadoop Distributed File System. http://hadoop.apache.org/docs/
29. Big Data Research and Development Initiative. www.whitehouse.gov/
30. InfoSphere Platform: Big Data Analytics, 2013, http://www-01.ibm.com/software/
31. Oracle and Big Data: Big Data for the Enterprise, 2013, http://www.oracle.com/
32. Big Data, 2013, http://www.microsoft.com/.
33. Big Data - What Is It? 2013, http://www.sas.com/big-data/
34. SAP HANA integrates predictive analytics, text and big data in a single package, 2013, http://www54.sap.com/
35. Big Data Solutions, 2013, http://www8.hp.com/
36. Stonebraker M. Errors in Database Systems, Eventual Consistency, and the CAP Theorem // Communications of the ACM, April, 2010.
37. Agrawal D., Das S., Amr El Abbadi. Big Data and Cloud Computing: Current State and Future Opportunities / EDBT, march 22–24, 2011, Uppsala, Sweden.
38. UN Global Pulse. http://www.unglobalpulse.org.
39. Черняк Л. Большие Данные – новая теория и практика. М.: Открытые системы, 2011, №10.
40. Алгулиев Р.М., Фаталиев Т.Х., Гаджирагимова М.Ш. К созданию корпоративной распределенной архивной системы // Известия НАНА, 2003, №3, с.143–147.
41. Menon J., Treiber K. Daisy: A Virtual-disk Hierarchical storage Manager, Performance Evaluation Review, 25(3), December 1997, pp.37–44.
42. Chen Y. “Information Valuation for Information Lifecycle Management” / Proceedings of Autonomic Computing, June 2005, pp.135–146.
43. Foster Y., Kesselman C., Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations // Intern. J. of High Performance Computing Applications, 2001, 15(3), 200–222, www.globus.org
44. McAfee A. and Brynjolfsson E. Big Data: The Management Revolution. Harvard Business Review, 2012, vol.90, no.10, pp.60–68.
45. Селезнев К. Проблемы анализа больших данных // Открытые системы, 2012, №7, с.25–29.
46. Fan W., Bifet A. Mining Big Data: Current Status, and Forecast to the Future / SIGKDD, vol.14, issue 2, pp.1–5.
47. Mayer-Schönberger V. and Cukier K. Big Data - A Revolution That Will Transform How We Live, Work and Think. John Murray (Publishers), 2013.
48. Szala A., Gray J. 2020 Computing: Science in an exponential world // Nature, 2006, vol.440, pp.413–414.
49. Anderson C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete // Wired Magazine, July 2008. http://www.wired.com/science/

50. Bakshi, K.Considerations for big data: Architecture and approach / Proceedings of the IEEE Aerospace Conference, 3-10 march, 2012, pp.1–7.
51. Fayyad U. Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling, 2012. http://big-data-mining.org/keynotes/
52. Черняк Л. Вычисления с акцентом на данные // Открытые системы, 2008, №8, с.36–39.
53. Zhang J., Huang M. L. 5Ws Model for Big Data Analysis and Visualization / Proceedings of the IEEE 16th International Conference on Computational Science and Engineering (CSE), 2013, pp.1021–1028.
54. Siba F.N., Mohammad S., Kidwai H.K., Qamar B., Awwad F. Parallel Implementation and Performance Analysis of a 3D Oil Reservoir Data Visualization Tool on the Cell Broadband Engine and CUDA GPU / Proceedings of the 14th International Conference on High Performance Computing and Communication & 9th International Conference on Embedded Software and Systems (HPCC-ICESS), 2012, pp.970–975.
55. Wu X., Zhu X., Wu G.Q., Ding W. Data mining with bigdata // IEEE Transactionson Knowledge and Data Engineering, 2014, vol.26, issue 1, pp.97–107.
56. Big Data Market Size and Vendor Revenues. http://wikibon.org/wiki/

**УДК 004.02**

***Алгулиев Расим М.[1], Гаджирагимова Макруфа Ш.[2]***
Институт Информационных Технологий НАНА, Баку, Азербайджан
[1]rasim@science.az; [2] makrufa@science.az

**Феномен big data: проблемы и возможности**

Эта статья посвящена феномену big data. В статье исследуются термин big data, возможности, проблемы этой технологии. Анализируются 3V-концепции и задачи анализа больших данных. Рассматриваются существующие программные и аппаратные продукты в реализации этой концепции.

***Ключевые слова:*** *big data, data science, big data analytics, NoSQL, MapReduce, Hadoop, OLAP.*

***Rasim M. Alguliyev [1], Makrufa S. Hajirahimova [2]***
Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]rasim@science.az, [2]makrufa@science.az

**"Big Data" phenomenon: Challenges and Opportunities**

This paper is devoted to “Big Data” phenomenon. It explores the term of "Big Data", opportunities, challenges and existing approaches of this technology. 3V conception and the tasks of big data mining are analyzed. We also analyzed the existing software and hardware products in the implementation of this conception.

***Keywords:*** *big data, data science, big data analytics, NoSQL, MapReduce, Hadoop, OLAP.*