UOT 004.89

Qasımova R.T.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
rena.gasimova@science.az

"BIG DATA" ANALİTİKASI: MÖVCUD YANAŞMALAR, PROBLEMLƏR VƏ HƏLLƏR

Verilənlərin həcminin artması və real zamanda onların analizinə olan tələbat böyük verilənlərin analitikasının yaranmasına gətirib çıxarır. Məqalədə böyük verilənlərin analitikasının mövcud problemləri araşdırılır, böyük həcmli verilənlərin analizi üçün ən çox istifadə olunan bəzi metodlar təhlil olunur və bir sıra tövsiyələr verilir. Eyni zamanda Big data emalının texnoloji mərhələləri, böyük verilənlərin əsas xarakteristikaları və xüsusiyyətləri tədqiq olunur.

Açar sözlər: *verilənlər xəzinəsi, bulud, verilənlər bazasının idarəetmə sistemləri, data processing, big data, big data analytics, NoSQL, MapReduce, Hadoop, OLAP.*

Giriş

Verilənlər formal şəkildə təsvir olunan faktlardır, onları informasiyaya çevirmək lazımdır. Son dövrədək verilənlərin emalı (*data processing*) haqqında danışdıqda, nisbətən kiçik həcmli verilənlər üzərində alqoritmik, məntiqi və ya statistik əməliyyatların ayrılmaz birliyi nəzərdə tutulurdu. Lakin kompüter texnologiyalarının real dünyaya yaxınlaşması verilənlərin real dünyada informasiyaya çevrilməsi tələbatını artırır. Emal olunan verilənlər çoxalır, emal sürətinə tələbat artır, real-vaxt rejimində müxtəlif formatlı çox böyük həcmli verilənlər massivini emal edə bilən informasiya texnologiyaları (İT) həllərinin yaradılması zərurəti yaranır.

İT məntiqi olaraq material texnologiyalardan az fərqlənir, girişdə xam verilənlər, çıxışda isə insan qavraması üçün daha asan formada olan, intellektin gücü ilə informasiyanı faydalı biliyə çevirən strukturlaşmış informasiya olur. Yəni, kompüter xam verilənləri yenidən emal edib, faydalıları ayırıb istifadə üçün yararlı formada yazır. Müxtəlif mənbələrdən və müxtəlif formatlarda toplanan verilənlərin birgə analizi və onlardan yeni biliklərin və faydalı məlumatların əldə olunması adi texnoloji prosesdir. Bu səbəbdən informasiya texnologiyalarına ümumi qanunauyğunluqlar tətbiq edilməlidir, çünki buna əsasən digər texnologiyalar inkişaf edir. Bu isə hər şeydən əvvəl yenidən emal olunan xammalın həcminin çoxalması və emalın keyfiyyətinin yüksəlməsidir. Beləliklə, "Big Data" adı altında kompüter texnologiyalarında ciddi dəyişikliklərə səbəb olacaq keyfiyyət keçidi yarandı və təsadüfi deyil ki, onu sənaye inqilabı adlandırırlar.

Zaman keçdikcə kompüter proqramları özlərinin çoxcəhətlilikləri ilə real dünyaya daha da yaxınlaşırlar. Emal olunmamış verilənlərin həcminin artması onların real vaxtda analizinin zəruriliyi ilə birlikdə *Big Data Analytics* məsələsini effektiv həll etməyə imkan verən alətlərin yaradılması və tətbiqini zəruri edir. Bunun nəticəsidir ki, son zamanlar böyük həcmdə informasiya massivlərinin toplanması, eyni zamanda onların çox böyük sürətlə artması, həm akademik mühitdə, həm də İT sənayesində daha çox diqqət cəlb etməyə başlamışdır. *Miniwatts Marketing Group* analitik şirkətinin apardığı statistik hesabata əsasən, 2015-ci ilin birinci rübündə 3 milyarddan çox insan, yəni planetin əhalisinin 42,4%-i İnternetə qoşulmuş və mobil rabitə abunəçilərinin sayı 7,1 milyarda çatmışdır. 2020-ci ildə İnternetə qoşulan qurğuların sayının 50 milyard olacağı gözlənilir. 2012-ci ildə dünyada rəqəmsal informasiyanın həcmi 2,7 zetabayt olmuşdur. 2015-ci ildə bu həcmin üç dəfə və növbəti hər il üçün 40% artması proqnozlaşdırılır [1–5].

Lakin rəqəmsal informasiyanın belə sürətlə artımı, verilənlərin müxtəlifliyi, onların ötürülmə sürətinin yüksək artımı çoxsaylı problemlərin yaranmasına səbəb olur. Qeyd edildiyi kimi, artıq böyük verilənlərin saxlanılması, real-vaxt rejimində emalı, analizi və idarə edilməsi problemlər yaratmışdır. Bununla belə, böyük verilənlər (BV) problemi hələ ilkin araşdırmalar

səviyyəsindədir, yəni bu sahə hələ də tam olaraq təhlil olunmayıb. Aparılan tədqiqatlar "Big Data" anlayışını, onun mahiyyətini, müxtəlif xarakteristikalarını təsnifləndirməyə, BV-nin mənbələrini, bu texnologiyanın imkanlarını, problemlərini, təhlükəsizlik məsələlərini hərtərəfli tədqiq etməyə imkan verir. Tədqiqatlar göstərir ki, BV-nin emalı və analizi mükəmməl analitik texnologiyalar və alətlər tələb edir [6–8].

Son zamanlar nüfuzlu beynəlxalq təşkilatlar, elmi qurumlar, İT sahəsindəki nəhəng şirkətlərin iştirakı ilə Big Data probleminə həsr olunmuş çoxsaylı konfranslar, simpoziumlar, seminarlar, forumlar keçirilməkdədir. Burada çox böyük həcmdə verilənlərin emalının müxtəlif aspektləri müzakirə və tədqiq olunur. Bir sıra açıq suallar əsasında bu sahədə gələcək elmi-tədqiqat istiqamətləri və imkanları müəyyən edilir. Bu tədqiqat istiqamətləri BV-nin problemlərinin həlli üçün optimal metodların işlənməsinə zəmin yaradır. Ona görə də, bu texnologiyanın elmi nəzəri problemlərinin araşdırılması və elmi-tədqiqat obyekti kimi öyrənilməsi vacibdir, aktualdır.

Big Data xüsusiyyətləri

Bu gün Big Data İT sahəsindəki nəhəng vendorların və beynəlxalq analitik agentliklərin istifadə etdikləri anlayışdır. Qeyd edək ki, Big Data texnologiyasını və bu texnologiyanı digərlərindən fərqləndirən xüsusiyyətləri araşdırarkən müxtəlif fikirlərə rast gəlmək olur. Mütəxəssislərin fikrincə, Big Data texnologiyalar çoxluğudur və aşağıdakı xüsusiyyətlərə malikdir [9–16]:

– *Çox böyük həcm.* Hazırda BV-nin miqyası terabaytlardan zetabaytlara qədər həcm ilə xarakterizə olunur. Bu gün müəssisə səviyyəsində verilənlərin belə həcmi artıq heç kəsi təəccübləndirmir.

– *Çoxlu sayda informasiya mənbələri.* Məlumdur ki, biznes analitika (BA – *Business intelligence*) emal olunmayan informasiyanı anlamlı, rahat formaya çevirmək üçün lazım olan metod və alətlərdir. Ənənəvi BA-da bir neçə informasiya mənbələri var, Big Data isə onlarla, minlərlə xarici mənbələrə malikdir.

– *Strukturlaşdırılmamış informasiya.* Ənənəvi verilənlər xəzinəsi (VX) relyasion baza əsasında qurulmuşdur və strukturlaşdırılmış informasiyanın axtarışı üçün nəzərdə tutulmuş alətdir. Adi relyasion bazaların emal edə bilmədiyi yüzlərlə terabayt və ekzabayt həcmində müxtəlif tip strukturlaşdırılmamış informasiyanın toplanması və saxlanması, axtarışı, təhlükəsizliyi, analizi kimi məsələlərin həlli üçün yeni alətin yaradılmasına ehtiyac yaranır. Big Data mənbələrinin əsas hissəsini də strukturlaşdırılmamış və qismən strukturlaşdırılmış verilənlər təşkil edir. Bu halda Big Data yanaşması prinsipial olaraq daha mükəmməl həll olan şablonlar əsasında axtarışı təklif edir, bu isə öz növbəsində relyasion verilənlər bazasından fərqli olan informasiyanın saxlanması strukturunu nəzərdə tutur.

– *Dinamiki dəyişən informasiya.* BV çox böyük informasiya massivlərinin toplanması ilə xarakterizə olunur ki, bu da hazırda istənilən təşkilatın qarşılaşdığı problemdir. Böyük həcm ilk növbədə saxlama problemi yaradır ki, bu da genişmiqyaslı saxlama və paylanmış emal tələb edir. Hətta girişdə filtrlənmiş verilənləri də saxlamaq getdikcə daha baha başa gəlir. Məlumdur ki, saxlama texnologiyaları bahadır və onların uçuzlaşması yeni verilənlər mənbəyinin yaranmasına nisbətən daha yavaş baş verir. Bununla bağlı təşkilat səviyyəsində bu və ya digər verilənlərin hansı vaxta qədər saxlanması dəqiq müəyyən olunmalıdır. Məsələn, bəzi verilənlər təşkilatda uzun illər ərzində tələb oluna bilər, digərləri isə artıq bir neçə saatdan sonra (analitiklər onlardan lazım olanı götürdükdən sonra) faydasız ola bilərlər.

Son zamanlar verilənlərin emalı və yadda saxlanılmasına xidmət edən "grid" və "cloud computing" texnologiyalarının tətbiqi saxlama sahəsindəki problemləri, demək olar ki, aradan qaldıra bilmişdir. Beləliklə, Big Data BA-nın ənənəvi alətləri ilə həll olunan oxşar məsələləri həcm, mənbə, struktur və paylanma mərhələsində daha geniş kontekstdə həll edir. Nəticədə Big Data və BA texnoloji səviyyədə bir-birindən ciddi fərqlənirlər (cədvəl 1) [17–21].

Cədvəl 1

Big Data və biznes analitika arasındakı fərqli xüsusiyyətlər

Ənənəvi BA	Big Data
Vahid korporativ VX	Xəzinə paylanmış fayl sistemində yerləşə bilər
Verilənlərin formatı emal funksiyalarının tələbinə uyğunlaşdırılır	Emal funksiyaları müxtəlif verilənlərin formatına uyğunlaşdırılır
Verilənlərin formatı strukturlaşdırılmışdır	Həm strukturlaşdırılmış, həm də strukturlaşdırılmamış informasiya ilə işləmək nəzərdə tutulur
Tarixi informasiya	Ən yeni verilənlərlə iş
Verilənlərin əldə edilməsi, ötürülməsi və emalı əsasən ardıcıl, dəqiq müəyyən olunmuş prosedurlara uyğun baş verir	Verilənlərin kütləvi paralel emalı tətbiq olunur (Massively Parallel Processing, MPP)

Big Data emalının texnoloji mərhələləri

Hazırda aparılan elm-tədqiqat işlərinin əksəriyyəti BV-nin texnoloji problemlərinin həllinə yönəlmişdir. Big Data-nın əsas mənbələrini sensor və sosial şəbəkələr, müxtəlif sahələr haqqında məlumat ötürücüləri, bank əməliyyatları, coğrafi informasiya sistemləri (GİS), Qlobal Mövqetəyinetmə Sisteminin siqnalları (GPS), elmi eksperimentlər, elektron poçt, smartfonlar vasitəsilə alınmış rəqəmsal foto və videolar, böyük şirkətlərin bazarları, böyük satış mərkəzləri, domen adları serverləri (Domain Name Server, DNS) və s. təşkil edir. Bu mənbələr tərəfindən yaradılan böyük həcmli verilənlərin məhsuldar istifadəsi üçün uyğun texnologiyalar lazımdır. Bunlar verilənlərin toplanması, saxlanması və analitik emalı sistemlərindən ibarətdir. Onun müxtəlif hissələrində aşağıdakı texnologiya qruplarından istifadə edilə bilər [22]:

– analitik alqoritmlər;
– paralel proqramlaşdırma metodları;
– bulud hesablama resursları;
– fərdi kompüterdən strateji təyinatlı superkompüterlərədək hesablama sistemləri;
– saxlama sistemləri;
– səbəkələr;
– mürəkkəb teleskoplardan və tomoqraflardan sadə radiotezlikli identifikasıya (Radio Frequency Identification, RFID) texnologiyalarınadək müxtəlif növ daxiletmə qurğuları və s.

Sadalanan siyahıda səviyyələr yüksəldikcə, verilənlərin biliklərə yaxınlığı da artır. Analitika verilənlərin biliyə çevrilməsi prosesini başa çatdırır. Siyahıda aşağı səviyyələr yoxlanmış elmi və mühəndis həllərinə əsaslanır, yuxarı səviyyələr isə yeni olduqlarına görə zəif inkişaf etmişlər və kifayət qədər elmi təminata malik deyillər. Mütəxəssislər bilavasitə İT-dən Big Data üçün texnologiyalar kateqoriyasına verilənlərin kütləvi paralel emalı (*Massively Parallel Processing, MPP*) platformalarında analitik sistemləri, verilənlərin bulud servisini, *Hadoop* və *MapReduce* texnologiyalarını və NoSQL tipli paylanmış verilənlər bazasının idarəetmə sistemlərini (VBİS) və s. daxil edirlər. *Apache Software Fondation*-in layihəsi olan *Hadoop* daha geniş yayılmış texnologiyadır, paylanmış hesablama mühitində böyük həcmli verilənlərin (petabayt miqyasında) emalı və analizi üçün əsas platformadır, MapReduce modelinin açıq kodlu (open access) sistemidir. *Hadoop* iki əsas komponentdən təşkil

olunmuşdur: *Hadoop MapReduce və Hadoop Distributed File System (HDFS)*. Burada *MapReduce* paralel hesablamalara, *HDFS* paylanmış fayl sistemi isə verilənlərin idarə edilməsinə cavab verir [23–27].

BV-in növbəti dalğasını insanın kompüterlə qarşılıqlı əlaqəsini təmin edən yeni qurğular yarada bilər. Onlara elektron kağız, taktil əks əlaqə texnologiyaları (haptics), müxtəlif videoşlemlər (video visor), həmçinin xüsusi yaddaşlı məhsul və əşyalar, Vikipediyaya oxşar açıq kontekstlə işləmək üçün sistemlər və s. aiddir. Məlumdur ki, geniş mənada texnologiya əmtəə və xidmətlərin istehsalı üçün zəruri bilik kapitalı, tətbiqi mənada isə məhsulun hazırlanması prosesində maddənin, enerjinin, informasiyanın çevrilməsi, materialların emalı və yenidən emalı, hazır məhsulların yığılması, keyfiyyət və idarəetməyə nəzarət üçün lazım olan metodlardır. Texnoloji ardıcıllıqdakı maşın və mexanizmlər ilkin xammalı istehlaka hazır olan məhsula çevirir. Hazırkı inqilab amorf, qeyri-müəyyən "informasiya texnologiyaları"nı rədd edərək daha aydın "verilənlər texnologiyaları"nı təsdiq edir. Texnoloji ardıcıllığın girişində xam verilənlər, çıxışında isə insan tərəfindən istifadəyə hazır olan verilənlər olmalıdır. Bu aspektdən müasir texnologiyalar verilənlərdəki informasiyanın yeni biliklərə çevrilməsinə və ya biliklərin əldə edilməsinə imkan verməlidirlər [28, 29].

Verilənlərlə işləyən texnologiyaların əsas fərqləndirici xüsusiyyətlərindən biri ondadır ki, onlar insan üçün işləyirlər. Avtomatik quraşdırılmış sistemlər istisna olmaqla qalan kompüter sistemləri son nəticədə insanın istifadə etdiyi verilənlərin hazırlanmasına xidmət edir, yalnız bu sistemlər verilənləri informasiyaya və sonra biliyə çevirə bilərlər. Bu cür texnologiyaların paradoksallığı ondadır ki, texnoloji ardıcıllığın girişində verilənlərin həcmi daim artırsa, buradan da böyük verilənlərin problemi yaranır. Giriş və çıxışdakı verilənlərin həcmi arasındakı disproporsiya mahiyyətcə verilənlərlə işləyən texnologiyaların inkişafında əsas istiqaməti müəyyən edir: daxil edilən verilənləri itirməməklə hansısa bir yolla giriş axınını cilovlamaq, sonra bütün verilənlərdən daha əhəmiyyətlilərini seçmək və insanın qavraması üçün asan şəkildə təqdim etmək lazımdır. Hazırkı iqtisadi böhran exafloodla (exaflood, petabaytdan sonra gələn, 1018 bayta bərabər olan verilənlərin ölçü vahidi exabyte və flood sözlərindən yaranmışdır) bağlı sahələrdə inkişafı dayandırmamış, lazımi verilənləri o qədər də mühüm olmayanlardan ayırmaq üçün filtr rolunu oynayan mürəkkəb hadisələrin emalı texnologiyalarının (*Complex Event Processing, CEP*) inkişafını sürətləndirmişdir. Bu qrup texnologiyalar BV-lərlə bağlı əsas məsələlərdən birini həll edir, yəni xam verilənlərdən yeni, gizli biliklərin və faydalı məlumatların əldə olunmasına keçidi asanlaşdırır [30].

Bu gün böyük həcmli verilənlərdən faydalı informasiyanın aşkarlanması üçün ən qabaqcıl metodlar təmin edilir, belə məsələləri effektiv həll etməyə imkan verən texnoloji alətlər yaradılır. Mütəxəssislər, tətbiq sahəsindən asılı olmayaraq, BV-nin emalının texnoloji ardıcıllığını yeddi əsas mərhələyə bölürlər [31–33]:

– *Verilənlərin toplanması*. Xam verilənlər xəzinələrdən, qurğuların vericilərindən, şəbəkə mənbələrindən toplana bilər. Bu mərhələ mühəndis nöqteyi-nəzərindən daha ənənəvidir, lakin burada da verilənlərin tipini ayırmaq lazımdır, məsələn, mətn verilənlərini ədədi verilənlərlə səhv salmaq olmaz.

– *Verilənlərin sintaksis və qrammatik təhlili.* Burada verilənlərin strukturlaşması, onların kateqoriyalar üzrə paylanması həyata keçirilir, təhlil isə bir neçə səviyyədə yerinə yetirilir. Aşağı səviyyəyə fiziki siqnalların emalı, sıxılmış faylların açılması və deşifrələnməsi daxildir. Bit səviyyəsində mətn, media və digər faylların ayrılması, mətn səviyyəsində qramatik və struktur təhlili aparılır.

– *Filtrasiya.* Burada verilənlərin analizi metodlarına, məsələn, mürəkkəb hadisələrin emalı texnologiyaları vasitələrinə müraciət etmədən giriş axınını azaldaraq, yalnız faydalı verilənləri saxlamaq lazımdır.

– *Verilənlərin əldə edilməsi.* Verilənlərin əldə edilməsi verilənləri ayırmağa imkan verən statistik və digər metodlardan ibarətdir, verilənləri uyğun riyazi kontekstdə həll edir.

– *Təqdimat.* Verilənlər üçün daha yaxşı təqdimat formasını müəyyən etmək tələb olunur (diaqramlar, siyahılar, ağaclar və s.). Verilənlərin vizuallaşma üsulu sadə cədvəl və qrafiklərdən tutmuş mürəkkəb iki və üç ölçülü təsvirlərə qədər dəyişir.

– *Təqdimatın təkmilləşdirilməsi.* Burada verilənlərin təqdimatı formalarının redaktəsi aparılır.

– *Qarşılıqlı əlaqə.* Verilənlərlə fəal işləmək üçün verilənlərlə manipulyasiya fəndləri və daha əyani təqdimatı təmin edən üsullar işlənir.

Verilənlərin toplanması və onların təhlili ənənəvi texnologiyalar vasitəsilə reallaşır, filtrasiya və əldə edilmə verilənlər haqqında yeni elmin predmetidir, verilənlərin təqdimatı və dəqiqləşdirilməsi qrafiki dizayn sahəsinə daxildir.

Big Data analitika

Yüzlərlə terabayt və ekzabayt həcmində BV-nin mövcud metodologiyalarla və ya alətlərlə toplanması, idarə edilməsi, saxlanması və onlardan faydalı informasiyanın əldə edilməsi ciddi problemdir. Həm strukturlaşdırılmış, həm də strukturlaşdırılmamış informasiya ilə işləmək, daha dərin intellektual analiz aparmaq və analizin nəticələrini vizuallaşdırılmaq BV-nin analitikasının əsas məsələlərindəndir. Verilənlərin həcminin artması və real zamanda onların analizinə olan tələbat BV-nin ən əsas problemlərindən sayılan BV-nin analitikası *(Big Data Analytics)* istiqamətinin yaranmasına gətirib çıxarmışdır [28, 34–35]**.**

Big Data Analytics müxtəlif tipli verilənlərdən ibarət olan böyük verilənlər yığımının öyrənilməsi prosesidir. Yəni, BV-dən gizli qanunauyğunluqları, naməlum korrelyasiyaları və digər faydalı işgüzar informasiyanı aşkarlamaq üçündür. Analitik verilənlər daha səmərəli marketinqə, yeni gəlir almaq imkanlarına, müştərilərə xidmət keyfiyyətinin yaxşılaşmasına, işin effektivliyinin artmasına, təşkilatların rəqabət və digər biznes üstünlüklərinə gətirib çıxara bilər. Bu istiqaməti digər tətbiqlərdən fərqləndirən böyük həcm, sürət və mürəkkəblik kimi xarakteristikalar uyğun texnologiyalar tələb edir. Buna görə də, bu gün Big Data Analytics sahəsində əsas istehsalçılar xüsusi proqram-aparat sistemlərini təklif edirlər: *SAP HANA, Oracle Big Data Appliance, Oracle Exadata Database Machine, Oracle Exalytics Business Intelligence Machine, Teradata Extreme Performance Appliance, NetApp E-Series Storage Technology, IBM Netezza Data Appliance, EMC Greenplum, HP Converged Infrastructure əsasında Vertica Analytics Platform*. Bununla yanaşı, kiçik və yeni başlayan şirkətlərin də böyük həcmli verilənləri səmərəli emal edən proqram-aparat alətləri vardır. Onlara *Cloudera, DataStax, Northscale, Splunk, Palantir, Factual, Kognitio, Datameer, TellApart, Paraccel, Hortonworks* aiddir [36–38].

Verilənlər informasiya almaq üçün emal edilir, bu informasiyanın həcmi o qədər olmalıdır ki, insan onu biliyə çevirə bilsin. Həcm BV-nin ən əsas xarakteristikasıdır. Həcmindən asılı olaraq BV üç qrupa bölünür [22, 39–41]:

– 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.

Fast Data üçün emal yeni biliklərin alınmasını nəzərdə tutmur, onun nəticələri aprior biliklərlə əlaqələndirilir, bu və ya digər proseslərin necə keçməsinə nəzarət edir, eyni zamanda baş verənləri daha yaxşı və ətraflı şəkildə görməyə, hansısa hipotezləri təsdiq və ya inkar etməyə imkan verir. Yalnız mövcud texnologiyaların kiçik bir hissəsi *Fast Data* məsələlərinin həllinə yarayır. Bunlara xəzinələrlə işləyən bəzi texnologiyaları göstərmək olar, məsələn, *Greenplum, Netezza, Oracle Exadata, Teradata, Verica* tipli VBİS və s. Bu texnologiyaların iş sürəti verilənlər həcminin artımı ilə sinxron artır.

Big Analytics vasitələri ilə həll olunan məsələlər kəmiyyət və keyfiyyətcə çox fərqlənirlər. Uyğun texnologiyalar isə yeni biliklərin alınmasına və faydalı məlumatların əldə olunmasına kömək etməklə, verilənlərdə olan informasiyanı yeni biliyə çevirirlər. Başqa sözlə, qərarın seçilməsində süni intellekt texnologiyaları nəzərdə tutulmur, analitik sistem "müəllimlə təlim" prinsipi üzrə qurulur və onun bütün analitik potensialı təlim prosesində ona tətbiq edilir. Belə analitikanın klassik nümunələri *MATLAB, SAS, Revolution R, Apache Hive, SciPy Apache* və *Mahout* məhsullarıdır [42–44].

Yüksək səviyyə, *Deep Insight* müəllimsiz təlimi (*unsupervised learning*) və analitikanın müasir metodlarının istifadəsini, həmçinin müxtəlif vizuallaşma üsullarını nəzərdə tutur. Bu səviyyədə aprior bilik və qanunauyğunluqların aşkarlanması mümkündür.

Keyfiyyət baxımından Big Data Analytics proqramları nəinki yeni texnologiyalar, həm də yeni düşüncə tərzi tələb edir. Analitikaya ilkin verilənlərin hazırlanması vasitələrindən, vizuallaşmadan və nəticələri insana təqdim edən digər texnologiyalardan ayrıca baxılır. Hətta *The Data Warehousing Institute* kimi təşkilatın analitikaya baxışları başqadır. Təşkilatın məlumatına görə, hazırda müəssisələrin 38%-i idarəetmə praktikasında *Advanced Analytics* vasitələrindən istifadə imkanlarını tədqiq edir, 50%-i isə yaxın üç il ərzində bunu etməyi nəzərdə tutur. Qeyd etmək lazımdır ki, müəssisələrdə bu sahəyə belə maraq biznes sahəsindən çoxlu arqument gətirməklə əsaslandırılır. Belə ki, müəssisələrə yeni şəraitdə daha təkmil idarəetmə sistemi tələb olunur. Onun yaradılmasına əks əlaqə qurulmasından, yəni qərar qəbulunu dəstəkləyən sistemlərdən başlamaq tələb olunur ki, bunun da nəticəsində gələcəkdə qərar qəbulunu avtomatlaşdırmaq mümkün olacaqdır. Texnoloji obyektlərin avtomatlaşdırılmış idarəetmə sistemlərinin yaradılması 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 [45].

Analiz üçün yeni vasitələr verilənlər mənbələrinin çox olmasına, həmçinin müxtəlif formatlarda (strukturlaşdırılmış, strukturlaşdırılmamış, qismən strukturlaşdırılmış) olmasına, həm də müxtəlif indeksləşmə sxemlərindən (relyasion, çoxölçülü, noSQL) istifadə edilməsinə görə tələb olunur. Belə ki, böyük həcmli verilənləri toplamaq, birgə emal etmək və analiz üçün uyğun şəklə salmaq çox çətin olur. Ənənəvi üsullarla verilənlərlə işləmək artıq mümkün olmur. *Big Data Analytics* daha böyük və mürəkkəb massivlərə tətbiq edildiyindən, *Discovery Analytics* və *Exploratory Analytics* terminlərindən də istifadə edilir. Necə adlandırılmasından asılı olmayaraq, mahiyyət eynidir – qərar qəbul edən şəxsləri müxtəlif proseslər haqqında məlumatlarla təmin edən əks əlaqəni yaratmaq tələb olunur [41, 46–48].

Müasir İT faktorları: BV, analitika və bulud texnologiyalarını bu gün bir-birindən ayrı təsəvvür etmək mümkün deyildir. Çoxsəviyyəli saxlama sistemlərinə artan tələbat, bulud texnologiyalarının real olaraq mövcudluğu BV-nin analitikasına marağı artırmışdır. Bulud texnologiyaları böyük hesablamaların aparılmasında son dərəcə müvəffəqiyyətli yanaşmalardandır, buludlarda saxlamaq, buludlarda hesablamalarsız BV ilə işləmək mümkün deyildir. 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. İT sahəsindəki nəhəng şirkətlər yeni nəsil saxlama sistemlərində məhz miqyaslama aspektlərinə və verilənlərin çoxsəviyyəli saxlanmasına böyük diqqət ayırmışlar [50, 51].

Praktika göstərir ki, bu gün analitik məsələlərin yerinə yetirilməsi üçün sistemləri çox yükləmək tələb olunur. Lakin biznes tələb edir ki, bütün servis, əlavələr və verilənlər həmişə əlçatan olmalıdır. Bundan başqa, hazırda analitik tədqiqatların nəticələrinə tələbat çox yüksəkdir, çünki savadlı, düzgün və vaxtında keçirilən analitik proseslər bütövlükdə biznes işinin nəticələrini əhəmiyyətli artırmağa imkan verir.

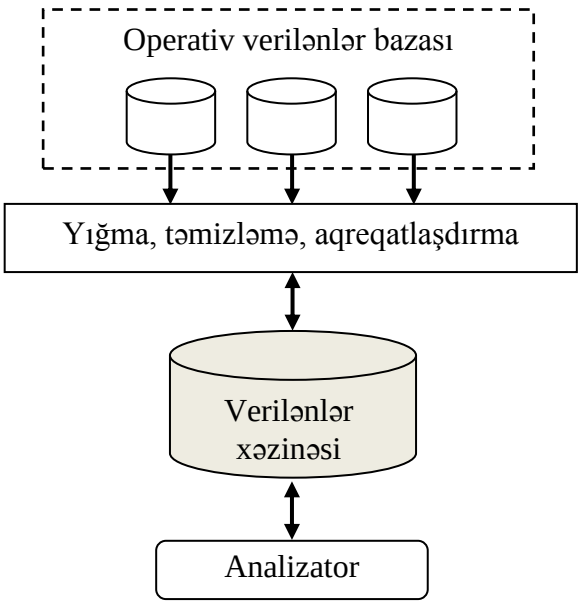
Böyük verilənlərin analizi problemləri. Bu gün BV-nin analizi üçün VX texnologiyalarının yaradılması zamanı istənilən yanaşma və metodlardan istifadə etməyə cəhd

göstərilir. Bununla yanaşı, ənənəvi əməliyyatların bəzi xüsusiyyətləri BV-nin emalı spesifikasına zidd ola bilər. Verilənlərin operativ və analitik emalı məsələlərinin əhəmiyyətli fərqi verilənlər bazaları texnologiyalarının inkişafının ilk vaxtlarında yaranmışdı. VX – *"Data Warehouse"* termini 70-ci illərdə Bill İnmon tərəfindən təklif edilmiş, lakin bu texnologiyalara maraq 20 il keçəndən sonra, belə sistemlərə real tələbat yarandıqda və lazımi hesablama gücləri əlçatan olduqda baş vermişdi [52].

VX-də verilənlərin emalı mərhələləri verilənlərin toplanmasından, təmizləmədən, yükləmədən, analizdən və nəhayət, analizin nəticələrinin təqdimatından ibarətdir. Bu mərhələlərin hər birində verilənlər üzərində xüsusi əməliyyatlar yerinə yetirilir. Qeyd etmək lazımdır ki, əgər BV-nin analizi üçün VX texnologiyalarının tətbiqinə cəhd olsa, onda yalnız alqoritmlərin analizinə deyil, həm də verilənlərlə işin bütün mərhələlərinə diqqət yetirmək lazımdır.

Verilənlərin toplanması. VX-nin yaradılmasında məqsəd ayrı-ayrı verilənlərin emalı sistemlərində yığılmış verilənlərin təmizlənməsi, razılaşdırılması, inteqrasiya edilməsi və analizdə istifadə üçün rahat formaya gətirilməsidir (şəkil 1). VX-də nəzərdə tutulur ki, informasiya operativ informasiya bazasından çıxarılır, lazımi şəklə çevrilir, yoxlanılır və yalnız bundan sonra sistemə yüklənilir. Yəni, VX-də verilənlərin hazırlanması texnologiyası bir-biri ilə əlaqəli üç mərhələdən ibarətdir [53]:

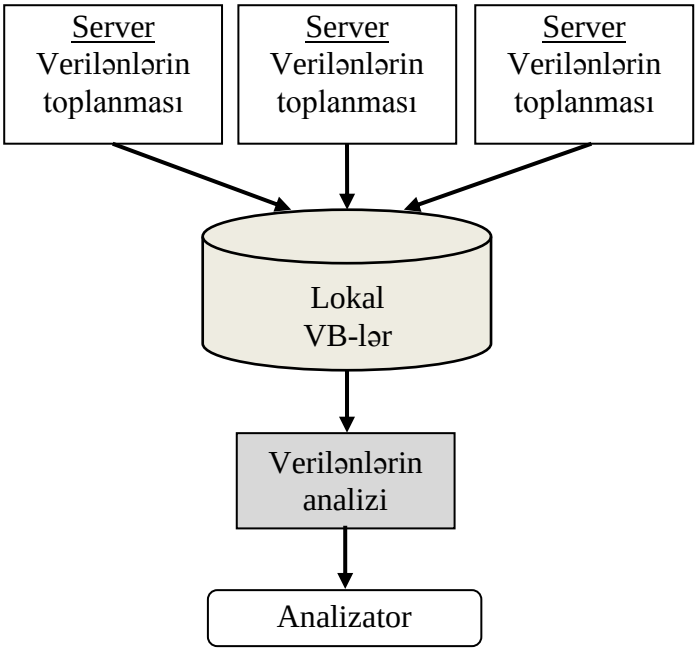
1. verilənlərin toplanması (*Data Acquisition*).
2. verilənlərin təmizlənməsi (*Data Cleaning*).
3. verilənlərin aqreqatlaşdırılması (*Data Consolidation*).



Şəkil 1. Ənənəvi analizdə verilənlərin xəzinəyə toplanması

Sadalanan əməliyyatlar dövriliklə yerinə yetirilir. Lakin BV ilə iş zamanı müxtəlif mənbələrdən girişə daxil olan verilənlərin analizini təmin etmək üçün belə dövrilük həmişə mümkün deyil. İnformasiyanın toplanması ilə onun analiz üçün əlçatanlığı arasındakı müddət VX-nin qurulması zamanı əməliyyatların yerinə yetirilməsinə lazım olan vaxtdan kiçik ola bilər. Belə məsələyə nümunə kimi sosial şəbəkələrdə tez yayılan informasiyanın və onun mənbələrinin müəyyən edilməsi, fəal istifadəçilərin təyin edilməsi, neqativ ifadələrin aşkarlanması və ya konfidensial informasiyanın sızması faktlarının aşkarlanması məqsədi ilə monitorinqi göstərilə bilər [54].

Məlumdur ki, bütün bu hadisələr daha tez aşkarlanmalı və neytrallaşdırılmalıdır. Lakin burada verilənlərin qeyri-formal təsviri movcuddur ki, onların da emalı üçün (ilkin olaraq yüksək sürəti ilə fərqlənməyən) mətnlərin intellektual analizi alqoritmləri tələb olunur. Məsələn, sosial şəbəkələrin monitorinqi məsələsində girişə istifadəçilərin ifadələri, şərhlər, qoyulmuş qiymətlər, fotoşəkillər və s. verilir. Aydındır ki, məşhur sosial şəbəkələrin monitorinqini çoxlu sayda istifadəçilərlə keçirmək olar. Həm də aydındır ki, istifadəçilərin çox olması və onların yüksək fəallığı səbəbindən bu məsələnin həlli qeyri-formal şəkildə təsvir edilən böyük həcmli verilənlərin yığım və emalını nəzərdə tutur [55]. VX-də verilənlərin ilkin emalı zamanı (məsələn, uyğunsuzluqların axtarışında) xəzinənin əvvəl yığılmış tərkibinin (verilənlərinin) istifadə edilməsi nəzərdə tutulur ki, bu da BV ilə iş zamanı çətin yerinə yetirilir. Problem ondadır ki, bu verilənlər həmişə paylanır, həm də təkcə analiz üçün deyil, yiğim üçün rahat olur. Məsələn, telekommunikasiya sistemlərində verilənlər regional serverlərdə toplanır (şəkil 2).



Şəkil 2. Böyük verilənlərin analizi

Beləliklə, aparılan təhlillər göstərir ki, ənənəvi VX-lərdə bütün verilənlər həmişə vahid məntiqi blokdan keçir, burada onlar konvertasiya olunur, yoxlanılır, təmizlənir, yüklənir və bu əməliyyatların yerinə yetirilməsi vaxtı nadir hallarda bütün qalan sistemlər üçün vacib olur. Lakin BV-nin emalı zamanı belə vahid blok ola bilməz. Qeyd etmək lazımdır ki, hələ ki, intensiv girişli verilənlər axınına malik məsələlər çox deyil və paylanmış, lakin məntiqi vahid sistem şəklində yığım, təmizləmə, çevirmə və yükləmə blokunu reallaşdırmaq mümkündür.

BV-nin analizi. Böyük həcmli verilənləri müxtəlif məqsədlər üçün operativ analiz etmək həmişə idarə edənlər üçün ciddi problem yaradır. Elmi-texniki ədəbiyyatın analizi göstərir ki, tətbiq olunan məhsullar, eləcə də digər analitik sistemlər hələ də tam olaraq bütün tələblərə cavab vermir. *Gartner, McKinsey Global* İnstitutu, *IDC (International Data Corporation)* və s. kimi kompaniyaların son tədqiqatlarında BV böyüyən, dinamik inkişaf edən sahə kimi təqdim edilir. *McKinsey Global* İnstitutunun analitik hesabatında BV-də tətbiq edilən analiz metodlarının adları verilmişdir [5]:

- *Data Mining sinfinin metodları* – assosiativ qaydaların öyrənilməsi (*association rule learning*), təsnifat, klaster analizi, reqressiya analizi və s.;
- *onlayn analitik emal (OLAP);*

- *kraudsorsinq (crowdsourcing) – əmək münasibətləri olmadan, ictimai müqavilə əsasında cəlb edilən bir qrup şəxsin gücü ilə verilənlərin emalı;*
- *verilənlərin birləşməsi və inteqrasiyası* (*data fusion and integration*) – dərinə analizi təmin etmək üçün müxtəlif mənbələrdən müxtəlif verilənləri inteqrasiya etməyə imkan verən texniki vasitələrdir. Buna nümunə olaraq, siqnalların rəqəmsal emalı və təbii dilin emalını (tonal analiz daxil olmaqla) göstərirlər;
- *maşin təlimi* – buraya müəllmlə və müəllimsiz təlim, həmçinin kollektiv təlim (*Ensemble learning*), baza modelləri əsasında kompleks proqnozlar almaq üçün statistik analiz və ya maşın təlimi əsasında qurulan modellərdən istifadə (*constituent models*, bir model daha boyuk modelin tərkib hissəsi olanda) daxildir;
- süni neyron şəbəkələri, şəbəkə analizi, optimallaşdırma, o cümlədən genetik alqoritmlər;
- obrazların tanınması*;*
- prediktiv analitika;
- imitasiya modelləşdirilməsi;
- fəza verilənlərinin analizi (*Spatial analysis*) - verilənlərdə topoloji, həndəsi və coğrafi informasiyadan istifadə edən metodlar;
- statistik analiz, buna nümunə kimi A/B testləşməsi və zaman sıralarının analizi göstərilir;
- analitik verilənlərin vizuallaşdırılması – nəticələrin alınması, sonrakı analiz, ilkin verilənlər kimi istifadə üçün informasiyanın şəkillər, diaqramlar, interaktiv imkanlar və animasiyadan istifadə etməklə təqdimatı.

BV üzərində çoxölçülü əməliyyat aparmaq üçün klassik B-ağaclardan başlamış bütün mövcud metodların tətbiqi yararlıdır. Bu gün ənənəvi VX-lər verilənlərin analizi alətlərinin təxminən eyni dəstini təqdim edirlər [56–58]. BV-lərdə sadalanan metodları işləməyə imkan verən, operativ yaddaşda terabaytlarla informasiyanı analitik emal edən məhsullar da yaranmışdır, məsələn, *SAP HANA, Greenplum Chorus, Oracle Exalytics, Oracle Exadata, Aster Data nCluster. Bundan başqa, Netezza, Teradata, Greenplum* və s. şirkətlərinin ənənəvi relyasion verilənlərinin idarə edilməsi sistemi əsasında terabaytlar və ekzabaytlarla verilənləri səmərəli emal edən proqram-aparat alətləri vardır. Belə həllərin potensial imkanlarını başa düşmək üçün həmin metodların əsasında olan alqoritmlərə baxmaq, həmçinin onların BV-nin emalına açar olan mümkün paralel hesablamaları üçün yolları analiz etmək lazımdır. Bu zaman verilənlərin konkret paylanmış emalı texnologiyalarına bağlanmamalı, yalnız BV üçün xarakterik olan əsas parametrləri (şəbəkə qarşılıqlı əlaqələrin intensivliyi, həcm, sürət və s.) nəzərə almaq lazımdır. İstənilən VBİS-in iş sürətinə təsir edən mühüm faktor giriş/çıxış əməliyyatlarının sayı və yaradılmış indekslərin effektivliyidir. Beləliklə, BV ilə işləyərkən məsələnin həllində ilk addım kimi emal olunacaq verilənlərin formatına, analizin tipinə, tətbiq olunan emal metodlarına, eyni zamanda məqsədli sistemin alacağı, yükləyəcəyi, emal və analiz edəcəyi, saxlayacağı verilənlərin mənbələrinə görə təsnifatlandırmaq tövsiyə olunur.

Onlayn analitik emal. OLAP (*On-line Analytical Processing*) – idarəedici qərarların qəbul edilməsini dəstəkləmək məqsədilə onlayn rejimdə verilənlərin operativ, analitik, çoxölçülü emalı deməkdir. Metodun mahiyyəti çoxölçülü kubun qurulması və onun müxtəlif kəsiklərinin alınmasındadır. Analitik verilənləri çoxölçülü kub şəklində təşkil etməklə onların sadə, anlamlı modeli alınır. Analizin nəticəsi, adətən, xanalarında aqreqat göstəricilər olan (say, orta, minimal və ya maksimal qiymət və s.) cədvəl olur. OLAP sisteminin fəaliyyəti ümumi halda dörd mərhələdən ibarətdir:

– verilənlərin toplanması;
– verilənlərin saxlanması;
– verilənlərin analitik altsistemlərin çoxölçülü kublarına yüklənməsi;
– verilənlərin təsviri.

Reallaşmasından, yəni analiz üçün istifadə olunan verilənlərin harada saxlanmasından asılı olaraq çoxölçülü analiz sistemləri aşağıdakı növlərə bölünürlər [53]:

- MOLAP – çoxölçülü OLAP (*Multidimensional OLAP*). Həm verilənlər, həm də aqreqatlar çoxölçülü VB-də saxlanılır.
- ROLAP – relyasion OLAP (*Relational OLAP*). Verilənlər öz əvvəlki relyasion VB-də saxlanılır, aqreqatlar isə həmin VB-də yaradılmış xüsusi işçi cədvəllərdə yığılır.
- HOLAP – hibrid OLAP (*Hybrid OLAP*). Verilənlər relyasion, aqreqatlar isə çoxölçülü VB-də saxlanılır.
- DOLAP – masaüstü OLAP (*Desktop OLAP*). Çoxistifadəçi rejimi dəstəkləməyən lokal, çoxölçülü analiz üçün nəzərdə tutulmuş məhsul olub, kiçik həcmli verilənlərin işlənməsində istifadə edilir.

Onların arasında geniş yayılmışı relyasion VBİS əsasında olan ROLAP sistemləridir və onlar daha şəffaf və öyrənilmişdir. MOLAP və HOLAP sistemlərinin daxili quruluşu, adətən, daha qapalıdır və konkret kommersiya məhsullarının nou-xau sahəsinə aiddir. MOLAP verilənləri çoxölçülü modeldə təsvir edir, lakin daxildə ROLAP-dan "ulduz" və "qardənəciyi" sxemini istifadə edir. VBİS nöqteyi-nəzərindən ROLAP verilənlər bazası adi relyasion bazadır və onun üçün bütün əməliyyatları dəstəkləyir. Lakin bir sıra çatışmazlıqları vardır. Məsələn, 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 yerləşdirilməsini optimallaşdırmaq mümkün deyil, aralıq aqreqat qiymətlərinin keşləşdiriməsinə imkan yoxdur. Nəhayət, analitik sorğuların yerinə yetirilməsi zamanı yüksək sürətliliyə tələbat səbəbindən, dərin statistik analiz aparmaq və həllin optimal yerinə yetirmə planı ilə işləmək mümkün deyildir. ROLAP-da verilənlər bazasının çoxölçülüyünü nəzərə almayan relyasion sorğu optimallaşdırmaları istifadə olunur. MOLAP texnologiyalarında bu çatışmazlıqlar yoxdur və buna görə yüksək analiz sürəti əldə etməyə imkan verir [59].

BV-nin analizi zamanı MOLAP, ROLAP, HOLAP texnologiyasının seçilməsi verilənlər bazasının yenilənmə tezliyindən asılıdır. Emalın paralel hesablanması nöqteyi-nəzərindən ilk baxışda istənilən çoxölçülü kub ölçülərdən birinin bölgüləri üzrə "kəsilə" və bir neçə server arasında paylana bilər. Məsələn, kubu zaman dövrlərinə (illər və aylara görə), ərazi əlamətinə görə (hər bir server öz regionuna cavabdehdir) və s. bölmək olar. Məlumdur ki, kubun bölünməsi zamanı çoxölçülü sorğunun yerinə yetirilməsi bir deyil, bir neçə serverin üzərinə qoyulur, bundan sonra nəticələr bütöv halda toplanır. Məsələn, əgər istifadəçi göstərilən müddətdə ölkə üzrə tələb olunan statistikanı soruşursa, verilənlər isə bir neçə regional OLAP-server üzrə paylanmışsa, onda hər bir server öz xüsusi cavabını qaytarır və nəticədə bütün verilənlər bir yerə toplanır. Əgər verilənlər zaman meyarı üzrə paylanarsa, onda baxılan sorğu nümunəsinin yerinə yetirilməsi zamanı bütün yük bir serverin üzərinə düşər. Belə vəziyyət aşağıdakı problemlərin yaranmasına səbəb olur [46]:

– verilənlərin serverlər üzrə optimal paylanmasını əvvəlcədən müəyyən etmək çətin olur;

– analitik sorğuların bir hissəsi üçün hansı verilənlərin, hansı serverlərdən lazım olacağı əvvəlcədən məlum olmur.

BV-də bu o deməkdir ki, çoxölçülü analiz üçün mövcud yanaşmalar yaxşı miqyaslana bilər və onlar paylanmış informasiyanın toplanmasına imkan verirlər (şəkil 2). Belə ki, hər bir server müstəqil informasiyanı toplaya, onu təmizləyə və lokal bazaya yükləyə bilər.

Klasterləşmə metodu. Məlumdur ki, klasterləşmə verilənlərin analizi və *Data mining* sahəsində əsaslı məsələlərdən biridir. Klasterləşmə çoxölçülü analizin riyazi modelidir, bir sıra obyektləri xarakterizə edən göstəricilər çoxluğuna əsaslanaraq, onları klasterlərə ayırır və hər bir klasterə daxil olan obyektlər digər klasterə daxil olan obyektlərlə müqayisədə daha oxşar və eyni olurlar. Klasterləşmə obyektlərin rəqəmlə ifadə olunan parametrlərinin arasındakı məsafəyə əsaslanır.

Klaster analizində oxşarlığın kəmiyyət qiymətləndirilməsi üçün metrika anlayışı daxil edilir. Belə ki, klasterləşmə analizində əsas mühüm şərtlərdən biri metrikanın seçilməsi sayılır (obyektlərin yaxınlığı). Təsnifat olunan obyektlər arasında oxşarlıq və ya fərq onlar arasında olan metrik məsafəyə görə müəyyən edilir (girişə iki vahid daxil olur, çıxışda onların oxşarlıq dərəcəsi tam identik vahidlər üçün sıfır olur). Klaster analizində obyektlər arasında müxtəlif məsafə ölçülərindən istifadə edilə bilər. Təsnifat olunan dəyişənlər üçün məsafə ölçüsünün və çəki əmsallarının tapılması klaster analizinin çox önəmli mərhələsidir. Müasir mərhələdə klasterləşdirmə verilənlərin analizində çox vaxt birinci addım kimi çıxış edir. Klasterlərə bölünmə üsuluna görə alqoritmləri dəqiq və qeyri-dəqiq, iyerarxiyalı və qeyri-iyerarxiyalı, iterativ və s. metodlara təsnifatlandırmaq olar [56, 60, 61].

BV-nin klasterləşmə problemi ondadır ki, mövcud alqoritmlər ilkin verilənlərdə istənilən informasiya vahidinə bilavasitə müraciət imkanını nəzərdə tutur, yəni alqoritmlərə hansı vahidlərin lazım olacağını əvvəldən tapmaq mümkün deyil. Öz növbəsində ilkin verilənlər müxtəlif serverlərə paylana bilər, bu halda hər bir klasterin mütləq olaraq bir serverdə saxlanması təmin edilmir. Əgər verilənlərin serverlər üzrə paylanması klasterləşmə alqoritmi üçün müəyyən edilsə (yəni, alqoritm hesab edir ki, verilənlər paylanmış virtual yaddaşda yerləşir), onda bu böyük həcmli verilənlərin bir serverdən digərinə köçürülməsinə gətirib çıxaracaqdır. Belə olan halda bu problemin həlli belə ola bilər. Hər bir serverdə yalnız həmin serverdəki verilənlərlə əməliyyat aparan alqoritm işə düşür, çıxışda tapılmış klasterlərin parametrləri və klaster daxilində elementlərin sayından asılı qiymətləndirilən çəkiləri verilir. Sonra alınan informasiya mərkəzi serverdə toplanır və metaklasterləşdirmə yerinə yetirilir, çəkiləri nəzərə alınmaqla yaxın yerləşən klaster qrupu seçilir. Bu metod universaldır, yaxşı paralel emalı təmin edir və ondan digər klasterləşmə alqoritmləri də istifadə edə bilər. Lakin o ciddi elmi tədqiqatlar aparılmasını, real verilənlərdə testləşməni və alınan nəticələrin digər lokal metodlarla müqayisəsini tələb edir [46].

Göründüyü kimi, vahid universal klasterləşmə alqoritmi mövcud deyildir. İstənilən alqoritmin istifadəsi zamanı onun üstünlükləri və nöqsanlarını başa düşmək, daha yaxşı işlədiyi verilənlərin təbiətini və miqyaslanma qabiliyyətini nəzərə almaq vacibdir. Araşdırmalar göstərir ki, BV-nin analizi üçün klasterləşdirmə metodlarının əksər hissəsi olduğu kimi yararlı deyildir və əlavə tədqiqatlar aparmaq zərurudir. Klasterləşmə alqoritmlərinin təkmilləşdirilməsi sahəsində tədqiqatlar daim davam edir [62].

Təsnifat metodu. Klassik metodlardan olan təsnifat məsələsi reqressiya məsələsinə oxşayır, bir dəyişənin digərlərindən asılılığının qurulması və istifadəsidir. Məsələn, yüksək səviyyəli domen zonalarının qiymətləri haqqında verilənlər bazası varsa, yeni alınacaq domen adın zonaya uyğun təqribi qiymətini əvvəlcədən söyləməyə imkan verən qaydalar sistemini qurmaq olar. Təsnifatın reqressiyadan fərqi ondadır ki, zaman sırası analiz edilmir, girişə verilən qiymətlər nizamlana bilmirlər. Son zamanlar çoxlu təsnifat metodları işlənmişdir (Bayes funksiyaları, həllər ağacı, neyron şəbəkələri və s.), bunlardan hər biri yaxşı işlənmiş elmi nəzəriyyəyə malikdir (özüöyrənən sistemlər, müəllimlə öyrənmə metodları) [63, 64].

Tədqiqatlar göstərir ki, təsnifat metodları eyni sxem üzrə qurulur. Əvvəlcə nisbətən kiçik seçimdən alqoritmin öyrənilməsi yerinə yetirilir, sonra alınan qaydalar qalan seçimlərə tətbiq edilir. Birinci mərhələdə klassik öyrənmə alqoritminin işi üçün paralel hesablama aparmadan verilənlər massivinin bir serverə köçürülməsi mümkündür. Lakin ikinci mərhələdə verilənlər müstəqil emal oluna bilərlər. Bu zaman özüöyrənmə nəticəsində alınmış qaydalar sistemi hər bir serverə köçürülür və bu serverdə saxlanan verilənlər massivi ondan ötürülür. Alınan nəticələr serverdə saxlana və ya sonrakı emal üçün göndərilə bilər. Beləliklə, təsnifatlandırıcıların öyrənmə mərhələsində BV üzərində iş aparılmır. Belə ki, sistemlərin öyrənilməsi üçün hazırlanan belə həcmin seçimi yoxdur, təsnifatlandırma mərhələsində verilənlərin ayrı-ayrı hissələri bir-birindən asılı olmadan emal edilir. Bu o deməkdir ki, mövcud təsnifat metodları BV-lə iş üçün yararlıdırlar [46].

Qanunauyğunluqların axtarışı. Təhlillər göstərir ki, böyük həcmli verilənlərdən avtomatik olaraq yeni qanunauyğunluqların aşkarlanmasında assosiativ qaydaların axtarışı alqoritmlərindən daha çox istifadə edilir. Assosiativ qaydaların axtarışı alqoritmlərinin tətbiq sahələri genişdir. Onları ticarətdə, təbabətdə, veb-səhifələrin seyr edilməsinin analizində (*Web Mining*), mətnlərin analizində (*Text Mining*), əhalının sayılması üzrə verilənlərin analizində, telekommunikasiya avadanlığının nasazlıqlarının analizi və proqnozlaşdırılmasında və s. hallarda uğurla tətbiq edirlər. Assosiativ qaydalar müxtəlif xüsusiyyətlər arasında korrelyasiyanı müəyyən edirlər. Yəni, assosiativ qaydalar əlaqəli hadisələr əsasında qanunauyğunluq tapmağa imkan verirlər. Assosiativ qaydada analizin məqsədi aşağıdakı asılılıqların qurulmasıdır: əgər tranzaksiyada bir neçə X elementlər toplusuna (yığınına) rast gəlinirsə, onda bunun əsasında nəticə çıxarmaq olar ki, digər Y elementlər toplusu də bu tranzaksiyada yaranmalıdır. Belə asılılıqların qurulması bizə çox sadə və intuitiv aydın olan qaydalar tapmağa imkan verir.

Assosiativ qaydalar axtarışı alqoritmləri bütün X və Y qaydalarının tapılması üçün təyin olunmuşdur, həm də bu qaydaların doğruluq və dəstəklənmə əmsalları uyğun olaraq minimal dəstək (minsupport) və minimal doğruluq (minconfidence) adlanan bəzi əvvəlcədən təyin olunan sərhəd qiymətlərindən böyük olmalıdır. Analitiki praktiki nöqteyi-nəzərdən bu qaydaların doğruluq və dəstəklənmə əmsalları maraqlandırır, lakin bu göstəricilər arasında hansı balansın olması konkret praktiki məsələnin sualıdır. Assosiativ qaydaların tapılması məsələsi iki altməsələyə bölünür [65]:

1. *Minsupport* həddini ödəyən bütün elementlər toplusunun tapılması. Belə elementlər toplusu tez-tez rast gəlinən toplu adlanırlar.

2. 1-ci bəndə əsasən *minconfidence* həddini ödəyən doğruluqla tapılan elementlər toplusundan qaydaların generasiyası.

Assosiativ qayda:

– (ilk növbədə) əhəmiyyətli olmalı, yəni tədqiqat aparılan verilənlərdə X və Y elementlər toplusu kifayət qədər tez-tez bir yerdə rastlaşmalıdırlar;

– dəqiq olmalı, yəni X toplusunu ehtiva edən tranzaksiyalarda Y toplusunu ehtiva edən tranzaksiyaların payı yüksək olmalıdır;

– maraqlı olmalıdır, yəni tranzaksiyada X toplusunun olması bu tranzaksiyada Y toplusunun olması ehtimalını artırmalıdır.

Qanunauyğunluqların axtarışı məsələsi Apriori alqoritminin köməyi ilə həll olunur. Aydındır ki, BV-nin emalı nöqteyi-nəzərindən əsas əməliyyat aqreqasiya funksiyasının hesablanmasıdır, bu da çoxölçülü analiz vasitəsilə yerinə yetirilir. BV-nin emalına mane olan Apriori alqoritminin digər mühüm anı informasiya əlamətləri toplusu ilə iş görə bilməsidir. Lakin bir şərti nəzərə almaq lazımdır ki, baxılan toplunun sayı informasiya əlamətlərinin sayından asılıdır, verilənlərin konseptual modelindən, onların həcmindən asılı deyil. BV-nin emalı üçün Apriori alqoritminin dəyişməsi aqreqasiya funksiyasının hesablanması üsuluna əsaslandığından, çoxölçülü analizin paralel hesablama alqoritmləri burada işlənə bilər.

Reqressiya analizi. Reqressiya analizi, bir və ya bir neçə asılı olmayan dəyişənin digər asılı dəyişənə təsirinin statistik analizi üsuludur. Məqsədi bir asılı olan, və bir və ya bir neçə asılı olmayan dəyişənlər arasındakı münasibəti araşdırmaqdır. Burada asılı dəyişənlərin sistematik təsiri nəticəsində asılı olmayan dəyişənlərin özlərini aparmasını izah etmək və onları təsadüfi təsirlədən ayırd etmək əsas məsələlərdəndir.

Reqressiya adı altında göstərilən ədədi kəmiyyətin verilən müddətdə dəyişməsini təsvir edən parametrik funksiyanın qurulması başa düşülür. Bu funksiya məlum verilənlər əsasında qurulur, sonra bu kəmiyyətin sonrakı qiymətlərinin əvvəlcədən tapılması üçün istifadə edilir. Metodun girişinə bu kəmiyyətin verilmiş şərtlərdə vəziyyətini təsvir edən "zaman-qiymət" şəklində cüt ardıcıllığı daxil olur, məsələn, konkret regionda konkret məhsul növünün satışının həcmi. Çıxışda tədqiq olunan kəmiyyətin vəziyyətini təsvir edən funksiyanın parametrləri olur.

İstifadə olunan parametrik funksiyanın növündən asılı olmayaraq onun parametrlərinin qiymətlərinin seçimi eyni üsulla həyata keçirilir. Kəmiyyətin müşahidə olunan qiymətləri ilə (yəni, metodun girişinə verilən) parametrlərinin cari qiymətlərində funksiyanın verdiyi qiymətlər arasında ümumi fərq hesablanır. Sonra ümumi fərqi azaltmaq üçün parametrlərin qiymətlərinin necə nizamlanması müəyyən olunur. Bu əməliyyatlar cəm fərqi lazımi minimuma çatana və ya onun sonrakı azalması mümkün olmayana qədər təkrar olunur.

Reqressiya analizində verilənlərin emalı nöqteyi-nəzərindən əsas əməliyyatlar mövcud ümumi fərqin hesablanması və parametrlərin qiymətlərinin təyin edilməsidir. Əgər birinci əməliyyat aydın şəkildə paralel yerinə yetirilirsə (cəm hissələrlə ayrı-ayrı serverlərdə hesablanır, sonra isə mərkəzi serverdə cəmlənir), ikinci isə çətin olur. Daha ümumi halda çəkilərin korreksiyası zamanı məlum riyazi faktdan istifadə olunur**:** bir neçə parametrin funksiyası qradiyent istiqamətində artır, qradiyentə əks istiqamətdə azalır. Öz növbəsində qradiyentin hesablanması hər bir parametrə görə funksiyanın xüsusi törəməsinin hesablanmasından ibarətdir ki, bu da çəkilərin cəminin hesablanmasına əsaslanan diskret diferensiallaşmaya gətirilir. Nəticədə parametrlərin qiymətlərinin təyini, həmçinin paralel yerinə yetirilə bilən cəmlənməyə gətirilir. Əgər reqressiya analizi çəkilərin cəminin hesablanmasına gətirilirsə, onda o çoxölçülü analiz kimi BV üzərində iş zamanı eyni tətbiq dərəcəsinə malikdir. Yəni, reqressiya analizi ilə işləyən sistemlər tamamilə miqyaslana və informasiyanın paylanmış yığımı şərtində işləyə bilərlər. Beləliklə, verilənlərin analizində mövcud alqoritmlər paralel hesablama apara bildikləri üçün potensial olaraq BV-nin analizi üçün istifadə oluna bilərlər [46, 66].

BV-ni təsvir etmək üçün mövcud olan ümumi xarakteristikalar, böyük həcm, yüksək sürət və verilənlərin müxtəlifliyi BV-də analizin aparılmasını və nəticədə yeni biliklərin aşkarlanmasını, faydalı məlumatların əldə edilməsini çətinləşdirir. Bu xarakteristikalar böyük verilənlərin əsas problemlərini özündə əks etdirməklə analitik arxitekturanın miqyaslanmasına böyük tələbatlar yaradır. Hal-hazırda bu tələblərin öhdəsindən gələn vahid alət mövcud deyildir. Nəticədə, bu gün çoxlu hibrid arxitektura yaranmışdır. Bunlardan genişlənmiş analitika platforması (*Advanced Analytics Platform*) praktiki eksperimentlər gedişində inkişaf etmişdir və BV ilə iş üzrə lazımi imkanları təmin edir.

Analiz nəticələrinin vizuallaşdırılması problemləri. Aparılan araşdırmalar göstərir ki, BV-nin analizinə həsr edilən elmi-tədqiqat işlərinin əksəriyyətində əsas diqqət bilavasitə analiz məsələlərinə yönəlib və alınan nəticələrin emalına fikir verilmir. Nəzərdə tutulur ki, mövcud metodlar hesabatların generasiyası, həmçinin müxtəliv növ diaqram və ya qrafiklərin qurulması şəklində tətbiq ediləcək. Lakin analizin nəticələrinə baxmaq üçün mövcud metodlar aşağıdakı səbəblərə görə tətbiq edilməyə bilər [46]:

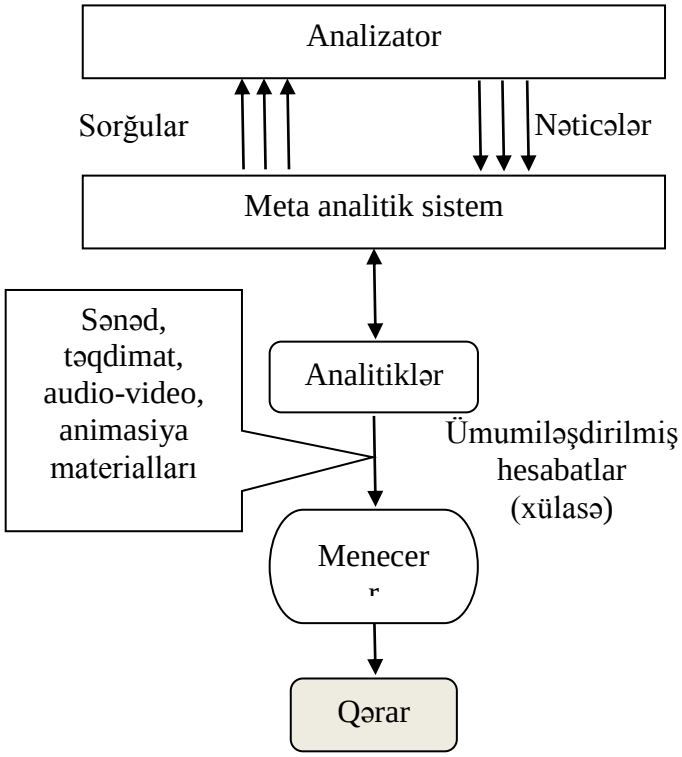
– girişdə böyük həcm verilənlər çıxışda böyük sayda analiz nəticələrini yaradır. Məlumdur ki, indi bir çox qanunauyğunluqlar statistik xətaları dəf edə bilirlər. Lakin qərar qəbulu üçün yalnız əsas qanunauyğunluqlarla kifayətlənmək olmaz. BV-nin analizində qəbul edilən qərarların maksimal səmərəliliyinə nail olmaq üçün çox az fərqlənən qanunauyğunluqlar və trendlər də nəzərə alınmalıdır. Əks halda ən müxtəlif məlumatların böyük axınlarının emalının, ümumiyyətlə, mənası olmaz.

– çıxış informasiyasının konseptual modeli olduqca çətinləşir. VX-nin tipik hesabatı ondan çox parametrlərə (məsələn, vaxt kəsiyi, region və s.) malik olmur, çünki, hesabat süni alınır və boş sətir və sıfırlardan ibarət olur. Lakin BV üçün bu belə deyil.

Verilənlərin analizi məsələlərində ilkin verilənlərin həcmi artdıqca, əvvəlcə sadə axtarış və verilənlərin baxış səviyyəsinə, sonra çoxölçülü və statistik analiz səviyyəsinə, sonra isə *Data Mining* səviyyəsinə keçərək, uyğun analiz metodlarından istifadə edilir. Lakin ilkin verilənlərin həcmi artdıqca, *Data Mining* səviyyəsində də həddən artıq çıxış informasiyası olur. Yəni, əgər əvvəllər qərar qəbulu üçün bir neçə hesabat vərəqinə baxmaq lazım gəlirdisə, BV-nin analizində bu belə deyil. Belə ki, qərar qəbul edən şəxs üstünə düşən çoxlu verilənlər yığınından daha

mühüm və əhəmiyyətli informasiyanın seçilməsi problemi ilə qarşılaşır. Bu problemi avtomatlaşdırılmış vasitələrin istifadəsi və işin yenidən təşkili üsulları ilə həll etmək olar [67].

Avtomatlaşdırılmış vasitələr daha mühüm hesabatlar seçməyə imkan verir, məsələn, satış dinamikasının monitorinqi zamanı elə hesabatlar seçilir ki, göstəricilərin əvvəlki dövrlə müqayisədə kəskin dəyişməsi müşahidə olunur. Lakin belə metod həmişə tətbiq olunmur. Belə ki, göstəricilərin dəyişməsi sadəcə sistemə məlum olmayan xarici amillərlə izah oluna bilər, məsələn, neftin qiymətinin aşağı düşməsi maşınların qiymətinin artması ilə izah oluna bilər və bu o demək deyil ki, növbəti dövrdə belə satış həcmlərini planlaşdırmaq lazımdır. Böyük sayda hesabatlardan mühüm və əhəmiyyətli informasiyanın seçilməsi işin yenidən təşkilinə əsaslanır. Bu zaman ayrıca insanlar seçilir ki, onların da vəzifələrinə hesabatların baxışı və qərar qəbul edən şəxslərə göndərilən xülasənin formalaşdırılması daxil olur (şəkil 3).



Şəkil 3. Böyük verilənlərin analizində işin yenidən təşkili

Belə xülasələrin formalaşdırılması əhəmiyyətli dərəcədə mövcud hesabatların generasiyası vasitələrindən aşağıdakı səbəblərə görə fərqlənir [46]:

– xülasəyə ən müxtəlif növlü informasiya daxil edilə bilər. Yəni, eyni sənəd özünə həm satış göstəricilərini, qiymətlərin artım dinamikasını, ştat cədvəlində dəyişiklikləri və həm də fotolara qədər hər şeyi daxil edə bilər;

– bütün xülasələr unukaldır və ümumi formatda olmaya bilər;

– xülasə həmişə konkret adam üçün konkret hala görə hazırlanır, nəticədə halın spesifikliyini və bu adamın xüsusiyyətlərini nəzərə alır. Bu nə vaxtsa çap edilmiş mətn sənədi, auditoriya qarşısında çıxış üçün prezentasiya, audioyazı və ya video ola bilər.

Bir sözlə, xülasə qərar qəbulu üçün bütün informasiyanı maksimal tez və əyani almağa imkan verməlidir. Sadalanan tələblərdən avtomatlaşdırılmış metaanalitik sistemlər konsepsiyası yaranır, bu da analiz nəticələrini vizuallaşdırmağa və onların əsasında sadəcə baxış üçün təyin olunan sənəd və təqdimatların yaranmasına, audioaxınların və videoroliklərin montajına, Flash animasiya yaratmağa imkan verir. Lakin bu vasitə və metodlar ilk baxışdan göründüyü kimi sadə

deyillər, müxtəlif informasiyanın seçilməsinə və maksimal əyani təqdimatına fokuslanmışlar. Bu məsələnin əl ilə həlli personala böyük yük yaradır, işin heç olmasa qismən avtomatlaşdırılması isə məlumatların xülasədə təqdimatının daha səmərəli üsullarının tapılması ilə bağlı fənlərarası elmi araşdırmaları tələb edir. Hazırda belə sistemlər lazımsız görünə bilər, lakin BV-nin analizi ilə bağlı praktiki məsələlərin sayı artdıqca, hesabatların tez və rahat hazırlanmasında onlara zərurət yaranacaqdır. Əks halda qərar qəbul edən personal sadəcə hesabat və analitik nəticələrin okeanında bata bilər.

Qeyd etmək lazımdır ki, BV ilə işləmək üçün bəzi hallarda VX-də istifadə olunan və işdə öz səmərəsini göstərən metodlar tətbiq olunandır və mövcud alqoritmlərdən bəziləri böyük paylanmış informasiya massivlərinin emalı üçün adaptasiya oluna bilər. Bununla yanaşı alınan nəticələrin əyani təqdimatında ciddi çətinliklər yarana bilər. Girişə daxil olan informasyanın böyük həcmi nəticəsində çıxışda müxtəlif növlü hesabatların sayı kəskin artır. Onların rahat təqdimatı üçün ənənəvi xəzinələr üçün istifadə edilən hesabatların generatorlarından prinsipial fərqlənən yeni proqram vasitələri lazımdır.

Nəticə

BV-də tətbiq edilən analiz metodları araşdırılarkən, onlardan hər birinin üstünlükləri və çatışmazlıqları təhlil edildi. Bu gün verilənlərin həcminin artması, onların real vaxtda analizinin zəruriliyi *Big Data Analytics* məsələsini effektiv həll etməyə imkan verən alətlərin yaradılması və tətbiqini tələb edir. *Big Data Analytics* vasitələrini və biznes-analitika texnologiyalarını, əsasən, İT nəhəngləri istifadə edirlər. Bu da biznes-analitikanın müəssisələrdə istifadəsinin aşağı mədəniyyətindən və biznes-istifadəçi tərəfindən mövcud analiz metodlarının qavrama çətinliyindən irəli gəlir. Bunu nəzərə alaraq, bəzi şirkətlərin (məsələn, *Information Builders*) analitikləri tərəfindən real vaxt rejimində istənilən mənbədən daxil olan verilənlərlə işləməyə imkan verən, istifadədə ən sadə qiymətləndirilən məhsul təklif edilir.

Big Data sadə düşüncə deyil, gələn texniki inqilabın simvoludur. Ona görə də, fəlsəfi olaraq demək olar ki, bu gün Big Data sivilizasiya üçün yeni bir idrakın əsasını qoyur. BV ilə analitik iş zəruriliyi İT-sənayesini tamam dəyişəcək yeni proqram və aparat platformalarının yaranmasını stimullaşdırır. Artıq bu gün böyük həcmli verilənlərin analizi üçün ən qabaqcıl metodlar təmin edilir: süni neyron şəbəkələri – bioloji neyron şəbəkələrinin təşkili və funksiyası prinsipi üzrə qurulmuş modellər, prediktiv analitika, statistika, *Natural Language Processing* metodları və s. Həmçinin, ekspertləri cəlbedən metodlar və ya kraudsorsinq, A/B testləşmə, sentiment analiz və s. vasitələrdən istifadə edilir.

Nəticələrin vizuallaşdırılması üçün məlum metodlar, məsələn, bulud teqləri və ən yeni *Clustergram, History Flow* və *Spatial Information Flow* istifadə edilir. BV texnologiyası tərəfindən *Google File System, Cassandra, HBase, Lustre* və *ZFS* paylanmış fayl sistemləri, *MapReduce* və *Hadoop* proqram konstruksiyaları və digər qərarlar dəstəklənir. Ekspertlərin qiymətlərinə görə, BV-nin təsiri altında ən çox transformasiyaya istehsalat, səhiyyə, ticarət, inzibati idarəetmə sahələri uğrayacaqdır.

Beləliklə, dünyada gedən proseslərə və aparıcı ölkələrin təcrübəsinə nəzər saldıqda aydın olur ki, müxtəlif mənbələrdən avtomatik və fasiləsiz olaraq generasiya olunan verilənlərin real-vaxt ərzində analitik emalı və analizində fərqli modellər, konseptual yanaşmalar təklif olunur. Bu mühüm faktı nəzərə alaraq, beynəlxalq təcrübədə BV-nin analitikası istiqamətində aparılan tədqiqatların araşdırılması olduqca zəruridir. Aydındır ki, çox böyük həcmdə rəqəmsal verilənlərin toplanması cəmiyyətin bütün sferalarının qarşılaşdığı problem olduğu üçün, onunla bağlı elmi-tədqiqat işləri də bir çox elm sahəsi üçün aktualdır. Çünki bu xam, strukturlaşdırılmamış verilənlər cəmiyyətin bütün sahələrini kökündən dəyişə biləcək təsirə malik bilik mənbəyidir. Bu baxımdan, Big Data texnologiyasının bütün aspektlərinin elmi-nəzəri əsaslarının işlənməsi böyük əhəmiyyət kəsb edir.

Ədəbiyyat

1. Miniwatts Marketing Group, Worldwide Internet Market Research, www.miniwatts.com
2. 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
3. Worldwide Big Data Technology and Services 2013-2017 Forecast, http://www.idc.com
4. Data Science Central, The online resource for Big Data practitioners, www.datasciencecentral.com
5. Big data: The next frontier for innovation, competition, and productivity. Analyst report, McKinsey Global Institute, May 2011. www.mckinsey.com
6. Madden S. From Databases to Big Data // IEEE Internet Computing, 2012, vol.16, no.3, pp.4–6.
7. What is big data? - Bringing big data to the enterprise, 2013. www-01.ibm.com
8. 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
9. Clifford L. Big data: How do your data grow? // Nature, 2008, vol.455, pp.28–29.
10. 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
11. Wei Fan, Albert Bifet. Mining big data: current status, and forecast to the future // ACM SIGKDD Explorations Newsletter, 2012, vol.14, no.2, pp.1–5.
12. Maté A., Llorens H., Gregorio E. An integrated multidimensional modeling approach to access big data in business intelligence platforms / Proceedings of the 2012 international conference on Advances in Conceptual Modeling (ER'12), Heidelberg, 2012, pp.111–120.
13. Szalay A., Gray J. 2020 Computing: Science in an exponential world // Nature, 2006, vol. 440, pp.413–414.
14. McAfee A., Brynjolfsson E. Big Data: The Management Revolution // Harvard Business Review, 2012, vol.90, no.10, pp.60–68.
15. Birke R., Björkqvist M., Chen L. Y., Smirni E., Engbersen T. (Big)data in a virtualized world: volume, velocity, and variety in cloud datacenters / Proceedings of the 12th USENIX conference on File and Storage Technologies (FAST'14), USENIX Association Berkeley, CA, USA, 2014, pp.177–189.
16. Richard Price. Volume, velocity and variety: key challenges for mining large volumes of multimedia information // Proceedings of the 7th Australasian Data Mining Conference (AusDM '08), Australia, 2008, vol.87, pp.17–23.
17. Chiang R.H.L., Goes P., Stohr E.A. Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline // ACM Transactions on Management Information Systems (TMIS), 2012, vol.3, no.3, Article 12 (pp.1–13).
18. Chen H., Chiang R.H. L., Storey V.C. Business intelligence and analytics: from big data to big impact // Journal MIS Quarterly, 2012, vol.36, no.4, pp.1165–1188.
19. Omar El-Gayar, Prem Timsina. Opportunities for Business Intelligence and Big Data Analytics in Evidence Based Medicine / HICSS '14 Proceedings of the 2014 47th Hawaii International Conference on System Sciences(HICSS '14), USA, 2014, pp.749–757.
20. Statchuk C., Iles M., Thomas F. Big data and analytics / Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research (CASCON '13), USA, 2013, pp.341–343.
21. Foster Y., Kesselman C., Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations // Intern. J. of High Performance Computing Applications, 2001, vol.15, no.3, pp.200–222.
22. Черняк Леонид. Большие Данные – новая теория и практика // Открытые системы, 2011, № 10, с.18–25.

23. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters // Communications of the ACM, 2008, vol.5, no.1, pp.107–113.
24. Lee K-H., Lee Y-J., Choi H., Chung Y.D., Moon B. Parallel data processing with MapReduce: a survey // ACM SIGMOD Record, 2011, vol.40, no.4, pp.11–20.
25. Brunozzi Simone. Big Data and NoSQL with Amazon DynamoDB / Proceedings of the 2012 workshop on Management of big data systems (MBDS '12), USA, 2012, pp.41–42.
26. Weiyi Shang, Zhen Ming Jiang, Hadi Hemmati, Bram Adams, Ahmed E. Hassan, Patrick Martin. Assisting developers of big data analytics applications when deploying on hadoop clouds / Proceedings of the 2013 International Conference on Software Engineering (ICSE '13), NJ, USA, 2013, pp.402–411.
27. Чак Лэм. Hadoop в действии. Издательство: ДМК Пресс 2012, 424 с.
28. Черняк Леонид. Вычисления с акцентом на данные // Открытые системы, 2008, № 8, с. 36–39.
29. Wu X., Zhu X., Wu G., Ding W. Data Mining with Big Data // Journal IEEE Transactions on Knowledge and Data Engineering, 2014, vol.26, no.1, pp. 97–107.
30. Wang Y.H., Cao K., Zhang X.M. Complex event processing over distributed probabilistic event streams // Computers & Mathematics with Applications, 2013, vol.66, no.10, pp.1808–1821.
31. Черняк Леонид. Смутное время СУБД // Открытые системы, 2012,№2, с. 16–21.
32. Черняк Леонид. Что делать с хаосом данных? // Открытые системы, 2013, № 9, с.16–20.
33. Дубова Наталья. Большие Данные крупным планом // Открытые системы, 2011, № 10, с.30–33.
34. InfoSphere Platform: Big Data Analytics, 2013, http://www-01.ibm.com/software
35. Jacobs A. The pathologies of big data // Communications of the ACM. 2009, vol.52. no.8, pp.36–44.
36. Вахрамеев Кирилл. СУБД для анализа Больших Данных // Открытые системы, 2011, № 10, с.26–29.
37. Babu S., Herodotou H. Massively Parallel Databases and MapReduce Systems // Foundations and Trends in Databases, 2013, vol.5, no.1, pp.1–104.
38. Vignesh Prajapati, Big Data Analytics with R and Hadoop, Publisher: Packt Publishing Ltd, 2013, pp.238.
39. Черняк Леонид. Свежий взгляд на Большие Данные // Открытые системы, 2013, № 7, с. 48–51.
40. Krish Krishnan. Data Warehousing in the Age of Big Data. 1st Edition, Morgan Kaufmann Publishers Inc. San Francisco, USA, 2013, 370 p.
41. Билл Фрэнкс. Укрощение больших данных. Как извлекать знания из массивов информации с помощью глубокой аналитики, пер. с англ. Андрея Баранова, М.: Манн, Иванов и Фербер, 2014, 352 с.
42. Big Data - What Is It? 2013, www.sas.com/big-data
43. MathWorks, www.mathworks.com/discovery/big-data-matlab.html
44. Hadoop Distributed File System. http://hadoop.apache.org/docs
45. Witt D., Gray J. Parallel Database Systems: The Future of High Performance Database Systems // Communications of the ACM, 1992, vol.35, no.6, pp. 85–98.
46. Селезнев К. Проблемы анализа больших данных // Открытые системы, 2012, №7, с.25–29.
47. Gudivada V.N., Rao D., Raghavan V.V. NoSQL Systems for Big Data Management / Proceedings of the 2014 IEEE World Congress on Services (SERVICES '14), USA, 2014, pp.190–197.
48. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим, пер. с англ. Инны Гайдюк, М.: Манн, Иванов и Фербер, 2013, 240 с.

49. Kenn Slagter, Ching-Hsien Hsu, Yeh-Ching Chung, Daqiang Zhang. An improved partitioning mechanism for optimizing massive data analysis using MapReduce // The Journal of Supercomputing, 2013, vol.66, no.1, pp.539–555.
50. Əliquliyev R.M, Hacırəhimova M.Ş. Big data fenomeni: problemlər və imkanlar // İnformasiya Texnologiyaları Problemləri, 2014, №2, s.3–16.
51. Marcos D. Assunção, Rodrigo N., Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya. Big Data computing and clouds // Journal of Parallel and Distributed Computing, 2015, vol.79, pp.3–15.
52. Inmon W. H. "Building the Data Warehouse," 3rd Edition, John Wiley & Sons, Inc., New York, 2002, 412 p.
53. Əliquliyev R.M., Qasımova R.T., Ələkbərova İ.Y. Qərarların qəbul edilməsini dəstəkləyən müasir konsepsiyalar haqqında // AMEA-nın Xəbərləri, fizika-riyaziyyat və texnika elmləri seriyası, 2005, №2, s.70–75.
54. Tonkin E.L., Pfeiffer H.D. Zombies Walk Among Us: Cross-Platform Data Mining for Event Monitoring / Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW '13), USA, 2013, pp.452–459.
55. Krishna Kumar K.P., Geethakumari G. A taxonomy for modelling and analysis of diffusion of (mis)information in social networks // International Journal of Communication Networks and Distributed Systems, Switzerland, 2014, vol.3, no.2, pp.119–143.
56. Alguliev R.M., Gasimova R.T. Identification of Categorical Registration Data of Domain Names in Data Warehouse Construction Task // Intelligent Control and Automation, 2013, vol.4, no.2, pp.227–234.
57. Alguliev R.M., Gasimova R.T. On a approach for intellectual analysis of registration data of domain names // International Journal of Ubiquitous Computing and Internationalization, 2011, vol.3, no.1, pp.27–30.
58. Ordonez C. Can we analyze big data inside a DBMS? / Proceedings of the sixteenth international workshop on Data warehousing and OLAP (DOLAP '13), USA, 2013, p. 85–92.
59. Алгулиев Р.М., Касумова Р.Т., Алекперова И.Я. Об одном подходе выполнения сложных запросов на основе технологии OLAP // Информационные технологии моделирования и управления, 2006, № 6, с.728–731.
60. Qasımova R.T. Milli domen adları ilə bağlı biliklər bazasının yaradılmasının konseptual əsasları haqqında // Bakı Universitetinin Xəbərləri. Fizika-Riyaziyyat Elmləri Seriyası, 2010, № 4, s.95–102.
61. Park H.S., Jun C.H. A simple and fast algorithm for K-medoids clustering // Expert Systems with Applications, 2009, vol.36, no.2, pp.3336–3341.
62. Нейский И.М., Филиппович А.Ю. Методика адаптивной кластеризации фактографических данных на основе интеграции алгоритмов MST и Fuzzy C-means // Известия высших учебных заведений. Проблемы полиграфии и издательского дела. М.: Изд-во МГУП, 2009, №3, с. 48–61.
63. Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan. Automatic Subspace Clustering of High Dimensional Data // Data Mining and Knowledge Discovery, 2005, vol.11, no.1, pp.5–33.
64. Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases / Proceedings of the ACM SIGMOD Conference on Management of Data, Washington D.C., May 1993, pp.207–216.
65. Tsai-Hung Fan, Dennis K. J. Lin, Kuang-Fu Cheng. Regression analysis for massive datasets // Journal Data & Knowledge Engineering, 2007, vol.61, no.3, pp.554–562.
66. Abousalh-Neto N.A., Kazgan S. Big data exploration through visual analytics / Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST '12), USA, 2012, pp.285–286.

67. Phil Simon. The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions. Publisher: Wiley, 2014, 240 p.

УДК 004.89
Касумова Рена Т.
Институт Информационных Технологий НАНА, Баку, Азербайджан
rena.gasimova@science.az
Аналитика «BIG DATA»: существующие подходы, проблемы и решения
Увеличение объема данных и потребность их анализа в реальном времени привели к необходимости создания одной из основных проблем аналитики больших данных. В статье рассматриваются существующие проблемы аналитики больших данных, расследуются наиболее часто используемые методы для их анализа и дан ряд рекомендаций. В то же время исследуются технологические этапы Big data технологии, основные характеристики и особенности данных.
Ключевые слова: *хранилища данных, облака, системы управления базами данных, Big data, Big data analytics, NoSQL, MapReduce, Hadoop, OLAP.*

Rena T. Gasimova
Institute of Information Technology of ANAS, Baku, Azerbaijan
rena.gasimova@science.az
"BIG DATA" Analytics: available approaches, problems and solutions
Increased volume of data and demand for ad hoc analysis of data leads to the rise of one of the biggest problems of Big Data called Big Data analysis. This article studies the current problems and most frequently used methods of big data analysis and gives some recommendations. The article also investigates the technological stages of Big data processing, and the basic characteristics and features of big data.
Keywords: *data warehouse, cloud, database management systems, data processing, big data, big data analytics, NoSQL, MapReduce, Hadoop, OLAP.*