UOT 004.9:351

İmamverdiyev Y.N.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
yadigar@lan.ab.az

BİG DATA TEXNOLOGİYALARININ BÖYÜK PERSPEKTİVLƏRİ VƏ PROBLEMLƏRİ

Big Data müxtəlif mənbələrdən fərqli formatlarda yüksək sürətlə daxil olan böyük həcmli verilənlərin emalı, analizi və onlardan faydalı biliklərin çıxarılması texnologiyalarını və alətlərini özündə birləşdirir. Son dövrlər elmi və kütləvi ədəbiyyatda Big Data texnologiyaları elektron dövlət, biznes, səhiyyə, elm, istehsal və digər fəaliyyət sahələrində yeni perspektivlər açan və inqilabi dəyişikliklər edə biləcək texnologiyalar kimi təqdim edilir. Bu işdə həmin iddiaların söykəndiyi arqumentlərin potensialını müəyyən etmək və Big Data sahəsində düzgün strategiya seçmək üçün həmin texnologiyaların mahiyyəti, xüsusiyyətləri və təşəkkül tarixi, texnoloji komponentləri və analitik imkanları kritik analiz edilir, üstünlükləri, perspektivləri və mövcud problemləri göstərilir.

Açar sözlər: *Big Data, Big Data analitikası, verilənlərin intellektual analizi, Hadoop, rediktiv model.*

Giriş

Son onillikdə informasiya-kommunikasiya texnologiyalarının (İKT) inkişafı və cəmiyyətin müxtəlif sahələrində tətbiqi nəticəsində çoxsaylı mənbələrdən müxtəlif formatlı böyük həcmdə verilənlər toplanır. İnformasiya texnologiyaları bazarının tətbiqi ilə məşğul olan *IDC* analitik şirkətinin qiymətləndirməsinə görə, dünyada saxlanılan verilənlərin həcmi ildə 40 % sürətlə artır. 2010-cu il müəyyən mənada sərhəd xətti hesab edilə bilər – verilənlərin həcmi 1 Zettabayt (Zb) həddini keçmişdi (1 Zb təxminən 1 milyard qiqabayta bərabərdir), 2012-ci ildə qlobal verilənlər 2.7 Zb həcmində olmuşdur. Proqnozlara görə, 2020-ci ilə qədər bu rəqəm 40 Zb-a çatacaq.

E-dövlət, elm, iqtisadiyyat, nəqliyyat, rabitə, ticarət, turizm, səhiyyə və digər sahələrdə böyük həcmdə verilənlər generasiya edilir və bu sahələrin fəaliyyəti getdikcə daha çox verilənlərin toplanması, emalı və analizi sistemlərinin effektivliyindən asılı olur.

Müxtəlif mənbələrdən böyük sürətlə daxil olan fərqli formatlı böyük həcmli verilənlərin ənənəvi verilənlər bazası texnologiyaları ilə emalı mümkün deyil. Bu problem bu cür verilənlərin toplanması, emalı və saxlanması sahəsində innovasiyalara təkan vermişdir. Nəticədə İKT sənayesində böyük həcmdə verilənlərin toplanması, saxlanması, emalı və intellektual analizi üçün müvafiq modellər və proqram təminatları yaradılmışdır. Kompüterlərin hesablama gücü, verilənləri saxlama qurğularının həcmi, hər cür mümkün sensorlar şəklində aparat bazası və yüksək sürətli İnternet də belə məsələlərin həlli üçün artıq yetkin səviyyəyə çatmışdır.

Son bir neçə ildə emal edilən verilənlərin həcminin və müxtəlifliyinin sıçrayışla artması hadisəsi və onu müşayiət edən bir sıra texnoloji həllər nəticəsində "kəmiyyətdən keyfiyyətə keçid" baş verir – bu fenomeni Böyük Verilənlər (ing. *Big Data*) adlandırırlar. Ekspertlər *Big Data* trendinin İKT sənayesinin aşkar aparıcı gücü olduğunu qeyd edirlər [1]. Bu gün artıq aydın görünür ki, *Big Data* – kompüter elmlərinin müxtəlif təbiətli böyük həcmli verilənlərin emalı, analizi və onlardan faydalı biliklərin çıxarılması sahəsində tədqiqatlar üçün yeni perspektivlər açan aparıcı istiqamətlərindən biridir.

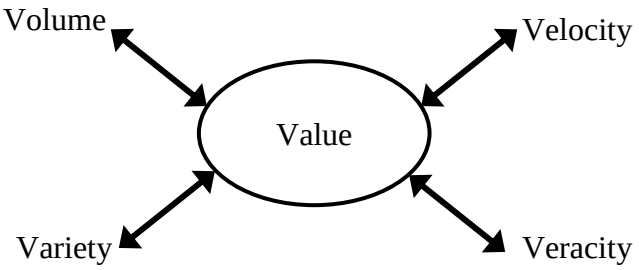
Bu gün İKT sənayesi həm dövlət sektorunun, həm də biznesin faydalanması üçün *Big Data* ilə işləyən müxtəlif yanaşmalar təklif etməyə hazırdır. Böyük müəssisələr (banklar, telekommunikasiya operatorları, pərakəndə satıcılar) müştəri bazalarında saxlanılan verilənlərin analizi əsasında müştəriləri haqqında, demək olar ki, hər şeyi öyrənə bilərlər. *Big Data* və bulud texnologiyalarının inteqrasiyası sayəsində *Big Data* kiçik müəssisələr üçün də böyük imkanlar

açır. Lakin *Big Data* texnologiyalarının təmin etdiyi bütün üstünlüklərə baxmayaraq, dünyada toplanan rəqəmsal verilənlərin yalnız 0,5%-i araşdırılır.

Big Data anlayışı

Big Data anlayışı üçün müxtəlif təriflər mövcuddur. İlk tərif *Meta Group* (indi *Gartner* şirkətinə daxildir) tərəfindən bu verilənlərin "*3V*" adlanan xarakteristikalarını təsvir etməklə verilmişdir: *Volume* (Həcm), *Velocity* (Sürət) və *Variety* (Müxtəliflik). Verilənlərin keyfiyyəti əsasında *IBM* dördüncü *V*-ni əlavə etmişdir: *Veracity* (Doğruluq). *Oracle Big Data*-nın gətirdiyi dəyəri nəzərdə tutaraq daha bir *V* əlavə etmişdir: *Value* (Dəyər) [2].

Bu üç tərifi birləşdirməklə "*5V*" adlanan daha əhatəli tərif alınır: *Volume*, *Velocity*, *Variety*, *Value* və *Veracity* (şəkil 1).



Şəki1 1. Big Data – 5V [2]

Volume (Həcm) – *Big Data* həcminin böyük olmasına görə ənənəvi üsullarla emalı mümkün olmayan verilənlərdir.

Velocity (Sürət) – həm yeni verilənlərin yaranması sürəti, həm də verilənləri emal sürəti nəzərdə tutulur. *Big Data* verilənlərin toplanmasının fenomenal sürətlənməsi və onların emalının mürəkkəbləşməsidir.

Variety (Müxtəliflik) – emal edilə bilən verilənlərin tiplərinin müxtəlifliyi nəzərdə tutulur: strukturlaşdırılmış, yarım-strukturlaşdırılmış, strukturlaşdırılmamış. *Big Data* növündən və həcmindən asılı olmadan verilənlərlə işləməyə imkan verən alətlər toplusuna deyilir.

Veracity (Doğruluq) – verilənlərin keyfiyyəti və mənbəyi nəzərdə tutulur: yaxşı, pis, qeyri-müəyyən, ziddiyyətli, natamam və s.

Value (Dəyər) – kiçik dəyər sıxlığına malikdir, yəni zəruri, qiymətli informasiyanın tapılması üçün çox böyük həcmdə verilənləri emal etmək lazım gəlir.

Big Data texnologiyaları böyük dəyər əldə etmək üçün olduqca müxtəlif mənbələrdən çox böyük həcmdə verilənləri yüksək sürətlə toplamağa, analiz etməyə və bu zaman avtomatik keyfiyyət nəzarəti ilə onların doğruluğunu təmin etməyə imkan verən texnologiyaların və arxitekturların yeni nəslidir [3].

Big Data həcmi. Adətən, *Big Data* anlayışı həcmi bir neçə terabaytdan xeyli çox olan verilənlər kimi təsvir edilir. Xüsusi halda bəzi verilənlər anbarı min terabayta, yəni bir petabayta qədər böyüyə bilər (1000 terabayt = 1 petabayt). Həcmi petabaytdan böyük verilənlər ekzabayt (Eb) ilə ölçülür, məsələn, 2009-cu ildə İnternetdə toplanan verilənlərin 500 Eb olduğu təxmin edilir. Verilənlərin ölçü vahidləri və uyğun həcmlər Cədvəl 1-də verilir.

Qeyd edək ki, verilənlərin həcminin aşağıdakı təsnifatı da məlumdur:

- *Böyük həcmli verilənlər*: 1000 meqabaytdan (1 qiqabaytdan) bir neçə 100 qiqabayta qədər;
- *Çox böyük həcmli verilənlər*: 1000 qiqabaytdan (1 terabayt) bir neçə terabayta qədər;
- *Big Data*: bir neçə terabaytdan yüzlərlə terabayta qədər;
- *Extremely Big Data*: 1000 terabaytdan 10000 terabayta qədər (1 petabaytdan 10 petabayta qədər).

Cədvəl 1. Verilənlərin ölçü vahidləri

Ölçü vahidi	**Həcm, bayt**	**Ölçü vahidi**	**Həcm, bayt**
Kilobayt (Kb)	10^3	Ekzabayt (Eb)	10^{18}
Meqabayt (Mb)	10^6	Zettabayt (Zb)	10^{21}
Qiqabayt (Qb)	10^9	Yottabayt (Yb)	10^{24}
Terabayt (Tb)	10^{12}	Brontobayt (Bb)	10^{27}
Petabayt (Pb)	10^{15}	Geopbayt (Gb)	10^{30}

Big Data mənbələri. Sosial şəbəkələr, veb-saytların loq-faylları, elmi verilənlər (astronomiya, fizika, insan genomu, meteorologiya, biokimya, biologiya) yaxşı məlum olan *Big Data* mənbələridir.

Verilənlərin 15-20%-i "Əşyaların İnterneti", o cümlədən çoxsaylı telefonlar, planşetlər və digər qurğular tərəfindən generasiya edilir. Proqnozlara görə, "Əşyaların İnterneti" tərəfindən generasiya edilən verilənlərin payı 2020-ci ildə 40%-ə çatacaq.

Müasir tibb texnologiyaları tibbi yardımın göstərilməsi ilə bağlı böyük həcmdə verilənlər generasiya edir (şəkillər, video, real vaxt rejimində monitorinq).

Elektrik stansiyaları kimi istehsal sahələrində bəzən on minlərlə parametr üçün hər dəqiqədə və hətta hər saniyədə verilənlərin fasiləsiz axını generasiya edilir. Bir neçə ildir ki, tətbiq edilən "*Smart grid*" texnologiyası ailələrin elektrik enerjisi sərfini hər dəqiqə və ya hər saniyə ölçməyə imkan verir.

Big Data sürəti. Verilənlərin həcmi və müxtəlifliyi dəyişdikcə, verilənlərin yaranma sürəti də dəyişir. Bu gün verilənlərin generasiya edildiyi sürət onların ənənəvi sistemlərdə emalını mümkünsüz edir. Hər gün təxminən 7 min petabayt yeni verilən generasiya edilir, onların yalnız 10%-i strukturlaşdırılmışdır – eyni zamanda bu nisbət daim azalır.

Böyük sürətlə verilən generasiya edən bir çox təşkilatlar vardır. *Twitter* hər dəqiqədə təxminən 5 Qb və ya gündə 7 Tb, *Facebook* dəqiqədə 7 Qb və ya gündə 10 Tb verilən generasiya edir. *Youtube* iddia edir ki, hər dəqiqədə 24 saatlıq video yüklənir.

Big Data-nın tarixi

Big Data müzakirələrinin əksəriyyəti biznes tətbiqləri ilə bağlı olsa da, əslində, bu termin korporativ mühitdə yaranmayıb, elmi məqalədə irəli sürülüb və yaranma tarixi dəqiq məlum olan çox az terminlərdən biridir [4]. 2008-ci ilin sentyabr ayının 3-də "*Nature*" jurnalının "Böyük həcmdə verilənlərlə işləmə imkanı açan texnologiyalar elmin gələcəyinə necə təsir göstərə bilər?" sualına həsr olunmuş xüsusi nömrəsi çapdan çıxmışdı. Bu xüsusi nömrə verilənlərin ümumiyyətlə elmdə və xüsusi halda elektron elmdə (*e-science*) roluna dair diskusiyaları yekunlaşdırırdı. *Big Data* terminini jurnalın redaktoru Klifford Linç işgüzar ingilis dili mühitində məşhur olan "böyük neft" və "böyük filiz" metaforalarına analoji olaraq təklif etmişdi.

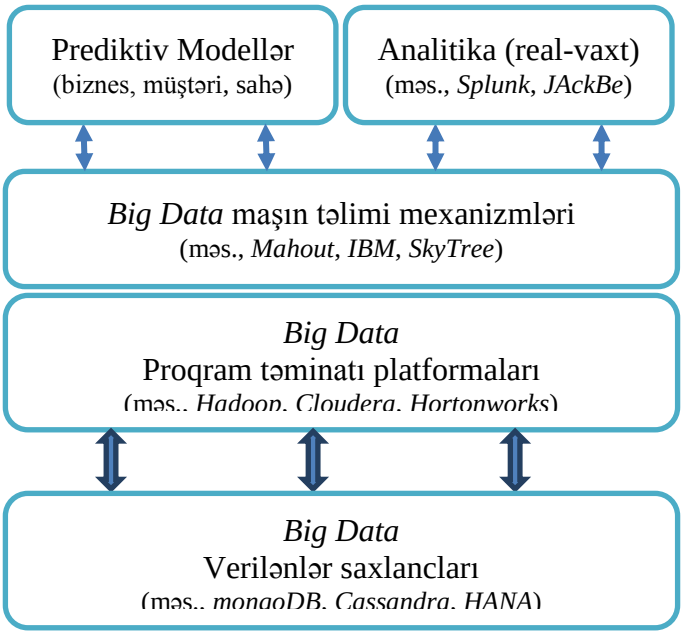
Lakin *Big Data* konsepsiyası yeni deyil, o, meynfreymlər və onlarla bağlı elmi hesablamalar dövründə meydana çıxmışdı. Məlumdur ki, elmtutumlu hesablamalar həmişə mürəkkəbliyi ilə fərqlənir və adətən, böyük həcmdə verilənlərin emalı zəruriliyi ilə sıx bağlı olur.

Big Data sənayesi bir çox şirkətin çox böyük həcmdə verilənləri emal etmək ehtiyacından yaranmışdır, çünki ənənəvi metodlar artıq bu işə yaramırdı. Məsələn, bəzi məlumatlara görə, *Google* bir gündə 24 Pb (24 milyon qiqabayt) informasiya emal edir. Bu həcmdə axınla bacaran superkompüterlərin qiyməti şirkətlərin əksəriyyəti üçün çox bahadır və onlar əvəz axtarmağa başladılar. İdeyalardan biri çox böyük sayda adi kompüteri şəbəkədə birləşdirməkdən və hesablamaları onlar arasında bölüşdürməkdən ibarət idi. Problem belə sistemlərdə daim qəzaların baş verməsi idi. Problemin həlli hesablamaları böyük şəbəkənin müxtəlif sahələrində təkrarlayan proqram oldu, bunun sayəsində elementlərdən birinin sıradan çıxması son nəticəyə təsir etmirdi [5].

Big Data ətrafında böyük canlanma *McKinsey* konsaltinq şirkətinin 2011-ci ilin iyununda açıqladığı "*Big Data*: innovasiyalarda, rəqabətdə və məhsuldarlıqda yeni sərhəd xətti" adlı hesabatından sonra yarandı. Hesabatda potensial *Big Data* bazarı milyardlarla dollarla qiymətləndirilirdi [1]. Hazırda *Big Data* sahəsinin potensialına görə İKT sənayesinin ən azı ikinci sektoru olması fikri hamılıqla qəbul olunur.

Big Data analitika vasitələri

Böyük həcmli verilənlərin toplanması, idarə edilməsi, analizi və vizuallaşdırılması üçün alətlər və texnologiyalar bir neçə sahəyə aiddir: statistik analiz, kompüter texnologiyaları, tətbiqi riyaziyyat və iqtisadiyyat. Onlardan bəziləri əvvəllər böyük olmayan verilənlərlə işləmək üçün istifadə edilirdilər, sonra isə böyük həcmli verilənlərə uğurla adaptasiya edilmişlər; digərləri isə elmi məsələlərdən meydana çıxmışdılar və əvvəldən böyük həcmdə verilənlərlə işləməyə yönəlmiş şirkətlər (ilk növbədə, *Google*, *Amazon*, *Yahoo*, *Facebook* və s.) tərəfindən idarə edilmişdilər.



Şəkil 2. *Big Data* ekosistemi

Big Data ekosistemini Şəkil 2-dəki kimi göstərmək olar. Ekosistemə daxil olan *Big Data* verilənlər saxlancları və proqram təminatı platformaları *Big Data* texnoloji bazasını təşkil edir, onlar müxtəlif mənbələrdən verilənlərin toplanmasını, saxlanmasını və idarə edilməsini təmin edirlər. Verilənlərin intellektual analizi (ing. *Data Mining*), maşın təlimi (ing. *Machine Learning*), mətnlərin intellektual analizi əsasında *Big Data* analitik alətləri qurulur [6-8].

Big Data ekosistemində problemləri üç istiqamətə ayırmaq olar:

1. Verilənlərin saxlanması və idarə edilməsi − həcmin yüzlərlə terabayt və ya petabayt olması verilənləri ənənəvi relyasion verilənlər bazalarının köməyi ilə saxlamağa və idarə etməyə imkan vermir.

2. Strukturlaşdırılmamış verilənlərin emalı − *Big Data* verilənlərinin əksəriyyəti strukturlaşdırılmamış verilənlərdir: mətn, video, audio, təsvirlər, multimedia və s. Strukturlaşdırılmamış verilənlərin emalını və analizini necə təşkil etməli?

3. *Big Data* analizi − *Big Data* analizi üçün statistik analiz, verilənlərin intellektual analizi, maşın təlimi, imitasiya modelləri, optimallaşdırma üsulları, verilənlərin vizuallaşdırılması, aqreqasiyası, inteqrasiyası və s. üsulları istifadə edilir. Prediktiv analitika ayrıca istiqamət kimi fərqləndirilir [9, 10].

Strukturlaşdırılmamış verilənlərin emalı. Strukturlaşdırılmamış verilənlər onların standart analitika alətləri ilə emalını çətinləşdirən və eyni zamanda, yeni biliklərin çıxarılması üçün unikal potensial təşkil edən bir sıra əlamətlərlə xarakterizə edilir. Birincisi, bu verilənlər olduqca *müxtəlif*dir. İkincisi, onlar *birmənalı* deyillər – verilənlərin eyni toplusu kontekstdən, dil və mədəniyyət xüsusiyyətlərindən asılı olaraq müxtəlif məna daşıya bilər. Üçüncüsü, onlar *dinamik*dirlər – zaman keçdikcə verilənlərin strukturu, qiymətləri dəyişir. Bundan başqa, strukturlaşdırılmamış verilənlər çox zaman subyektiv və emosional çalar xarakteri daşıyırlar.

Predmet sahəsinin *ontologiyasının* (strukturunun) müəyyən edilməsi strukturlaşdırılmamış verilənlərin strukturlaşdırılmış şəklə gətirilməsində ilk addımdır. Ontologiya – predmet sahəsinin təsviri sxemi və verilənlərin bu predmet sahəsinə aid edilməsi qaydalarından ibarətdir. Sxem kimi ona konseptlər daxil olur – mahiyyətlər, mahiyyətlərin atributları və əlaqələr. Əlaqələrə xidməti informasiyanı: münasibətin emosional çalarını, əlaqə predmetini, əlaqə üsulunu və s. əks etdirməyə imkan verən atributlar da daxil olmalıdır. Konseptlər, atributlar və əlaqələr üçün meyarlar – strukturlaşdırılmamış verilənlər axınından verilənləri bu və ya digər predmet sahəsinə aid edilməsi qaydaları müəyyən edilir.

Predmet sahəsinin ontologiyasının müəyyən edilməsindən sonra formalaşdırılmış strukturlara axtarış, klassifikasiya, vizuallaşdırma, analiz, proqnozlaşdırma, qanunauyğunluqların aşkarlanması, emosional çaların müəyyən edilməsi və faktların çıxarılması alətlərini tətbiq etmək olar.

Strukturlaşdırılmamış verilənlərin intellektual analizi elmi tədqiqatların nisbətən cavan sahəsidir, mətn verilənlərin intellektual analizi – *Text Mining* sahəsində daha çox tədqiqatlar aparılıb [11, 12]. *Text Mining* sahəsində tədqiqatların əsas sahələrinə mətnlərin klassifikasiyası, klasterləşdirilməsi, referatlaşdırılması, faktların, anlayışların çıxarılması (*feature extraction*), suallara cavabların axtarışı (*question answering*), tematik indeksləmə (*thematic indexing*), açar sözlər üzrə axtarış (*keyword searching*), mühakimələrin tonunun analizi (*Sentiment analysis*), rəylərin analizi (*Opinion Mining*) aid etmək olar [8, 10, 12].

Hazırda bir çox aparıcı proqram təminatı istehsalçısı *Text Mining* sahəsində məhsullar təklif edir, belə sistemlərə bəzi misallar aşağıda verilir:

- *Intelligent Miner for Text* (*IBM*);
- *TextAnalyst*, *PolyAnalyst* (*Megaputer*);
- *Text Miner* (*SAS*);
- *SemioMap* (*Semio Corp.*);
- *Oracle Text* (*Oracle*);
- *Knowledge Server* (*Autonomy*).

Prediktiv analitika. Prediktiv analitika (ing. *predictive analytics*) – verilənlərin və ya hadisələrin gələcəkdə proqnozlaşdırılması üçün indiki və keçmiş verilənlərin və ya hadisələrin analizində istifadə edilən statistika, verilənlərin analizi və oyunlar nəzəriyyəsinin metodları çoxluğudur.

Prediktiv analitikaya yaxın anlayış *Data Mining*-dir, çünki prediktiv analitika qismən oxşar metodları istifadə edir. Prediktiv analitikanın əsas mahiyyəti prediktorun və ya prediktorların (proqnozlaşdırılan hadisəyə təsir edən parametrlərin) müəyyən edilməsidir. Məsələn, sığorta şirkətləri sığorta ödənişinin müəyyən edilməsi zamanı yaş, sürücülük təcrübəsi kimi prediktorlara baxırlar. Prediktorlar çoxluğu prediktiv analitika modelini təşkil edir və bu model baxılan hadisəni gələcəkdə müəyyən ehtimalla öncədən xəbər verir.

Prediktiv analitikanın istifadəsinə ən məşhur misal – bankda kredit verilərkən müştərinin ödəmə qabiliyyətinin qiymətləndirilməsi üçün skorinq modellərinin tətbiqidir. Lakin prediktiv analitikanın tətbiq sahələri olduqca genişdir, ona ən böyük ehtiyac son istehlakçılarla işləyən bank və maliyyə xidmətləri, sığorta, əczaçılıq, dövlət sektoru, telekommunikasiya və informasiya texnologiyaları, pərakəndə satış kimi sahələrdədir.

Erik Siqel "*Predictive Analytics*" adlı kitabında prediktiv analitikanın ən geniş yayılmış on tətbiq sahəsini göstərir [13]: birbaşa marketinq; reklamın prediktiv istiqamətlənməsi; dələduzluq sxemlərinin aşkarlanması; investisiya risklərinin idarə edilməsi; müştərilərin saxlanması; tövsiyə servisləri; təhsil; siyasi kampaniyalar; tibbdə qərar qəbuletmə sistemləri; sığorta və ipoteka krediti.

Big Data modellərinin qurulması. Çox zaman məsələ *Big Data* verilənləri üçün dəqiq modellər qurmaqdan ibarət olur. Müxtəlif *Data Mining, Machine Learning* alqoritmlərinin böyük həcmli verilənlərin paralel emalı üçün *Map-Reduce* realizələri vardır. Lakin böyük həcmdə verilənlərin emalından alınan yekun modelin, həqiqətən də, dəqiq olmasını söyləmək çətindir.

Əslində, böyük olmayan verilənlərin modellərini qurmaq daha əlverişlidir. *Big Data* analizinə yanaşmalardan biri verilənlərin bütün həcmini seqmentləmə və klasterləşdirmə üçün istifadə etməkdən, alınan böyük olmayan seqmentlər və klasterlər üçün çox sayda modellər qurmaqdan və sonra uyğun model üzrə proqnoz verməkdən ibarətdir. Limit halında, gələcək alış-verişi proqnozlaşdırmaq üçün müştərilərin böyük verilənlər anbarında hər bir şəxsin ayrıca modeli qurula bilər.

Beləliklə, *Big Data*-nı dəstəkləyən analitika platforması yüzlərlə, hətta minlərlə modeli idarə etmək gücündə olmalı və lazım olduqda, onları yenidən kökləmək imkanına malik olmalıdır [14].

Big Data texnologiyaları

Paylanmış fayl sistemləri. *Big Data* (terabaytlar, petabaytlar) paylanmış fayl sistemlərində saxlana və sistemləşdirilə bilər. Paylanmış fayl sisteminin idarə edilməsi üçün standart texniki avadanlıq və açıq kodlu proqram təminatı (məsələn, *Hadoop*) istifadə edilməklə verilənlər anbarını etibarlı şəkildə nisbətən asanlıqla reallaşdırmaq olar [15].

Big Data paylanmış fayl sistemləri *Google File System*, *Lustre* (*Linux* və *Clusters* və ya *Lustre File System*, *LFS*), *IBM*-in fayl sistemi *General Parallel File System* (*GPFS*), *Hadoop Distributed File System* (*HDFS*) platformaları və bir çox digər həllərlə dəstəklənir.

Google File System daha çox axtarış məsələlərinə hesablanıb, *HDFS* isə analitik proqramlar üçün daha çox yarayır [16]. Lakin elə tətbiq sahələri vardır ki, sadalanan vasitələrdən heç biri tam şəkildə tələbləri təmin etmir və onlarda verilənlərə paylanmış müraciət tələb edilir, verilənlər müxtəlif anbarlar arasında, o cümlədən coğrafi paylana bilər. Bu zaman xərcləri minimallaşdırmaq üçün verilənlərin izafi miqrasiyasından qaçmaq və çoxsəviyyəli sistemdə verilənləri istifadə edildikləri yerə yaxın saxlamaq lazımdır.

Verilənlər bazalarının yeni növləri. Verilənlərin həcminin kəskin artmasının təsiri altında stabil sahə hesab edilən verilənlər bazalarını idarəetmə sistemləri (VBİS) sahəsində müəyyən fəallıq hiss olunmağa başlayır ki, bu da özünü– *NoSQL* və *NewSQL* kimi iki hərəkatın meydana çıxmasında özünü göstərir [17, 18].

NoSQL (*not only SQL* və ya *no SQL*) − ənənəvi relyasion verilənlər bazalarında istifadə olunan *SQL* vasitəsilə verilənlərə müraciətdən əhəmiyyətli dərəcədə fərqlənən verilənlər bazası modellərinin reallaşdırılmasına yönəlmiş bir sıra layihələri, yanaşmaları bildirən termindir (2009-cu ildə meydana çıxıb).

NoSQL bazaları - yeni növ verilənlər bazasıdır: qeyri-relyasion, paylanmış, açıq kodlu və üfüqi miqyaslanandır. *NoSQL*-həllərin tətbiqi zamanı verilənlər sxeminin təsviri üçün heş-cədvəllər, ağaclar və digər verilənlər strukturları istifadə edilə bilər.

NoSQL konsepsiyasının tərəfdarları qeyd edirlər ki, bu konsepsiya heç də relyasion modelləri və *SQL* dilini tam inkar etmir. Layihə o faktdan çıxış edir ki, SQL mühüm vasitədir, lakin o, universal ola bilməz. Relyasion verilənlər bazaları üçün göstərilən problemlərdən biri böyük həcmdə verilənlər ilə pis işləməsidir. Layihənin məqsədi *SQL*-in çevik olmadığı yerlərdə verilənlər bazalarının imkanlarını genişləndirməkdir.

Effektiv klaster həlləri. Hazırda paralel verilənlər bazaları texologiyaları geniş yayılıb. Bu texnologiya prosessorlar çoxluğunun vahid verilənlər bazasına müraciətini təmin edir ki, bu da tranzaksiyaların daha yüksək səviyyədə buraxılış qabiliyyətinə çatmasına, çox sayda istifadəçinin eyni zamanda işləməsini dəstəkləməyə və mürəkkəb sorğuların yerinə yetirilməsini sürətləndirməyə imkan verir.

SNA (*Shared Nothing Architecture*) – resursları paylaşılmayan arxitektur daha yaxşı miqyaslanır və getdikcə daha populyar olur. *SNA* – paylanmış müstəqil hesablama arxitekturudur, burada hər bir qovşağın öz yaddaşı, disk massivi və giriş-çıxış qurğuları vardır. Belə arxitekturda hər bir qovşaq özlüyündə müstəqildir və şəbəkənin digər qovşaqları ilə heç nə ilə bölüşmür. Hər bir *SNA* qovşağı digər qovşaqlarla xüsusi protokolla qarşılıqlı əlaqədə olaraq, öz məsələsini yerinə yetirir. Belə sistemlərin məhsuldarlığını hər bir qovşaqda prosessorlar, operativ yaddaş, disk yaddaşı əlavə etməklə və ya belə qovşaqların sayını artırmaqla yüksəltmək olar.

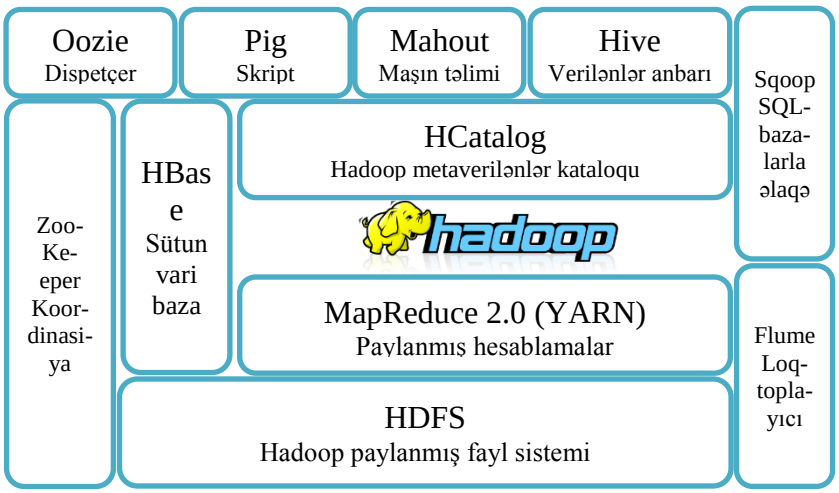
Big Data və bulud texnologiyaları. Bulud texnologiyaları, hər şeydən əvvəl, *Big Data* analizi üçün səmərəliliyi, miqyaslanmanı, miqrasiyanı və genişlənməni təmin edən çevik yanaşmadır [19]. Bulud mühiti verilənlərə müraciətin səmərəliliyini yüksəltməyə kömək edir və böyük həcmdə verilənləri emal etmək üçün resursların çevik çoxluğunu təklif edir. Bu zaman nəhəng həcmdə verilənlərin saxlanması və onların emalı üçün yetərli miqdarda hesablama resurslarının təmin edilməsi problemləri həll edilir. Buludda verilənlər bir neçə sahədə yerləşdirilir, bu onları istifadəçiyə yaxın yerləşdirməyə, müraciət vaxtını azaltmağa və məhsuldarlığı artırmağa imkan verir.

Big Data və bulud texnologiyalarının inteqrasiyanı məqsədyönlü dəstəkləmək üçün *Pivotal Initiative* adlı virtual şirkət yaradılıb, ona *Pivotal Labs*, *Greenplum*, *vFabric*, *Cloud Foundry*, *Spring* və *Cetas* kimi şirkətlər daxildir. Söhbət *PaaS* və *Big Data* analitikası həllərinin vahid strukturda birləşdirilməsindən gedir. Bu alyansda *VMware* məhsulları infrastruktur və *PaaS*, *Greenplum* sistemləri – analitika, *Pivotal* – məhsul istiqamətlərinin bir tamda birləşdirilməsi və ümumi kommersiya həllinin yaradılması üçün cavabdehdir.

Hadoop **ekosistemi**

Hazırda *Hadoop* ekosistemi (Şəkil 3) *Big Data* üçün sinonim hesab edilir. *Hadoop*-da verilənlərin avtomatik paralelləşdirilməsini və onların klasterlərdə emalını təmin edən *MapReduce* texnologiyası reallaşdırılıb (2005-ci ildə *Doug Cutting* və *Mike Cafarella* tərəfindən yaradılıb, *Hadoop* adı *Cutting*-in kiçikyaşlı oğlunun oyuncaq filinin adından götürülüb). *Hadoop*-un komponentlərinin çoxu müxtəlif *Apache* layihələrində yaradılmış açıq kodlu proqram təminatıdır [16, 20].

Aşağıda *Hadoop* ekosisteminə daxil olan bəzi komponentlərin qısa təsviri verilir:



Şəkil 3. Hadoop ekosistemi

HDFS (*Hadoop Distributed File System*) **–** *Hadoop*-un nüvəsini təşkil edir, həcmi bir neçə terabaytdan petabayta qədər olan verilənlər anbarlarının saxlanması və idarə edilməsi üçün paylanmış fayl sistemidir. *HDFS* giriş verilənlərini bloklara bölür və blokların hər biri serverlər çoxluğunda onlara ayrılmış yerlərdə yerləşdirilir. Kommunikasiya üçün *TCP/IP* səviyyəsi istifadə edilir. *HDFS* imtinalara dayanıqlıdır, komponentlərdən hər hansı biri sıradan çıxsa, bu sistemin ümumi işinə təsir etmir. *HDFS* tətbiqi proqramlara minlərlə qovşaq və petabaytlarla verilənlər miqyasında işləməyə imkan verir. Məsələn, *HDFS* sistemi *Yahoo!*-da təxminən 40 Petabayt veriləni idarə edir.

MapReduce – kompüter klasterlərində böyük həcmlərdə (petabaytlarla) verilənlərin paralel emalı üçün *Google*-nin 2004-cü ildə təklif etdiyi paylanmış hesablama modelini reallaşdırır (*Java*-da). *MapReduce*-un işi iki addımdan ibarətdir: *Map* və *Reduce*. *Map*-addımda giriş verilənləri ilkin emal edilir. Bunun üçün kompüterlərdən biri (*master node* – əsas qovşaq) məsələnin ilkin verilənlərini alır, onu hissələrə bölür və ilkin emal üçün paylanmış fayl sisteminin işçi qovşaqlarına paylayır. *Reduce*-addımda əsas qovşaq ilkin emal edilmiş verilənləri işçi qovşaqlardan toplayır, onları birləşdirir və məsələnin həllini formalaşdırır [21].

Pig – böyük həcmdə verilənləri qiymətləndirmək üçün yüksək səviyyəli dildir. *Apache Pig* komponenti *Big Data* çoxluqları üzərində komandaların yaradılması və yerinə yetirilməsi ideyası ilə yaradılmışdı. *Pig* proqramlarının əsas xüsusiyyəti paralleləşdirmədir, bu böyük verilənlər çoxluqlarını idarə etməyə kömək edir. *Pig* komponenti *MapReduce* proqramlar ardıcıllığını generasiya edən kompilyatordan və '*Pig Latin'* dilindən ibarətdir, *Hadoop* paylanmış verilənlər bazalarında *SQL*-ə oxşar sorğuların yerinə yetirilməsinə dəstək verir.

Hive – verilənlər anbarı infrastrukturudur, *Hadoop* fayl sistemində yerləşən böyük həcmdə verilənlərə *SQL* vasitəsilə müraciət etmək üçün tətbiq edilir, *MapReduce* tam dəstəklənir. *Hive*-in digər özəlliyi sorğuları sürətləndirmək üçün bit-xəritə indeksləri kimi indeksləri dəstəkləməsidir. *Apache Hive Facebook* tərəfindən yaradılmışdı, hazırda digər şirkətlər tərəfindən də istifadə edilir və inkişaf etdirilir.

HCatalog – *Hadoop*-da yaradılmış verilənlər üçün cədvəllərin və saxlancların idarə edilməsi servisi təmin edir. *HCatalog* ortaq sxem və verilənlərin tipi mexanizmi, cədvəl abstraksiyası kimi özəlliklər təklif edir, bunlar *Hadoop*-un *Pig*, *MapReduce*, *Streaming* və *Hive* kimi komponentlərinin dayanıqlı işləməsini dəstəkləyir.

HBase (*Hadoop DataBase*) – paylanmış, sütunvari verilənlər bazasıdır (*Google*-un *BigTable*-dən qaynaqlanır), saxlanma məqsədləri üçün *HDFS*-dən istifadə edir. Bir tərəfdən, *MapReduce* istifadə etməklə paket rejimində hesablamaları, digər tərəfdən isə, nöqtə sorğularını (ing. *random reads*) idarə edir.

Zookeeper – *Hadoop* ekosisteminin başqa bir əhəmiyyətli komponentidir. Onun əsas funksiyası koordinasiya məlumatlarını saxlamaq, adlandırma, paylanmış sinxronlaşdırmanı və qrup servislərini təmin etməkdir, bunlar müxtəlif paylanmış sistemlər üçün olduqca vacibdir. Əslində, *HBase*-in işləməsi *ZooKeeper*-dən asılıdır.

Mahout – maşın təlimi üçün proqram təminatıdır, klassifikasiya, klasterizasiya, tövsiyə və paketli kolloborativ süzgəc kimi əsas alqoritmlər daxildir. Əsas alqoritmlər *Hadoop*-un yuxarı səviyyəsində *Map/Reduce* paradiqması ilə reallaşdırılıb, lakin onu *Hadoop*-dan kənarda da xətti cəbr və statistikaya hədəflənmiş proqram kitabxanası kimi istifadə etmək olar.

Ekosistemə daxil olan ***Sqoop*** və ***Flume*** verilənləri *Hadoop*-klasterlərə və əksinə köçürmək üçün istifadə edilir.

Çox vaxt *Hadoop* verilənlərin saxlanması və emalının standart texnologiyaları ilə birlikdə istifadə edilir, bəzən isə *Storm, Dremel, Drill* kimi innovativ həllər də əlavə edilir. Bundan başqa, biznes analitikası məhsullarının praktiki olaraq bütün əsas istehsalçıları məhsullarına *Hadoop*-klasterlərdə daim saxlanan verilənlərə müraciət üçün funksional imkanlar əlavə edirlər. Komponentlərin bu siyahısını dəfələrlə genişləndirmək olar, çünki getdikcə daha çox şirkət bu və ya digər cəhətdən *Hadoop*-la əlaqəsi olan məhsullarla bazara daxil olur.

Big Data-nın bəzi problemləri

Big Data tətbiqlərinin əsas problemlərdən biri *Big Data* layihələrinin effektivliyinin qiymətləndirilməsi ilə bağlıdır. Bu cür layihələrin effektivliyinin iki mənbəyini göstərmək olar: 1) Bu böyük həcmdə verilənlərin analizinə çəkilən xərcləri və zamanı azaldır, operativ qərar qəbul edilməsi üçün informasiyanın tez hazırlanmasına imkan verir; 2) *Big Data* texnologiyalarının tətbiqi e-xidmətlərin fərdiləşdirilməsini təmin edir. Bununla yanaşı, elan edilən iqtisadi effektivlik və məqsədəuyğunluq vədlərinə baxmayaraq, *Big Data* layihələrinin iqtisadi effektivliyini hesablamaq olduqca çətindir.

İkinci problem *Big Data* texnologiyalarını müxtəlif sahələrdə tətbiq edə bilən mütəxəssislərin hazırlığı ilə bağlıdır. Belə mütəxəssislər bir tərəfdən riyazi statistika, verilənlərin analizi, maşın təlimi sahəsində hazırlıqlı olmalı, proqramlaşdırma vərdişlərinə malik olmalı, *İBM*, *Oracle*, *HP*, *SAS*, *SAP* və başqa şirkətlər tərəfindən təklif edilən aparat-proqram komplekslərində işləməyi bacarmalıdırlar. Digər tərəfdən, onlar konkret sahədə *Big Data* texnologiyalarının tətbiq edildiyi məsələlərin qoyuluşu vərdişlərinə malik olmalıdırlar. Onlar baxılan sahədə fəaliyyətin metod, ssenari və alqoritmlərini bilməli, *Big Data* texnologiyalarını realizə edən aparat-proqram sistemlərinin funksional xarakteristikalarına dair tələbləri formalaşdırmağı bacarmalıdırlar.

Eyni zamanda, yeni texnologiyalara aludəçilik məyusluğa da gətirib çıxara bilər. Məsələn, *Big Data* analizi nəticəsində mənasız korrelyasiyalar aşkarlana bilər – harvard Universitetinin professoru David Leynueber sübut edir ki, *S&P* 500-ə daxil olan şirkətlərin gəlirlərini Banqladeşdə kərə yağı istehsalının həcmi əsasında yüksək dəqiqliklə proqnozlaşdırmaq olar [22]. Maraq kəsb edən bir çox problem böyük olmayan, seyrək verilənlər (ing. *Sparse data*) toplusundan istifadə edilməklə də öyrənilə və həll edilə bilər.

"*Small Data*" hərəkatı. *Big Data* sahəsində uğurların arxasında başqa bir vacib inqilab da gizlənir: kiçik verilənlər (ing. *Small Data*). Verilənlərin saxlanma qiymətlərinin sürətlə düşməsi sayəsində onların toplanmasında və emalında kütləvi demokratikləşmə baş verir. Böyük həcmdə verilənlər üzərində mərkəzləşmə və nəzarət tendensiyalarının (*Big Data*) əksinə, texnologiyalar əsrində ən böyük potensial məhz mərkəzləşməmiş, paylanmış verilənlər və biliklər ekosistemindədir:

- Hətta *Microsoft* və *Yahoo!* kimi təşkilatlarda da işlərin çoxunu tək bir kompüterdə yerinə yetirmək olur. Məsələn, *Microsoft*-da orta iş ölçüsü 14 Qb və işlərin 80%-i 1 Tb-dan azdır. *Yahoo!*-da təxmini orta iş ölçüsü 12 Qb-dır [23].
- Tədqiqatlar göstərir ki, *Facebook*-da işlər üstlü paylanma qanununa tabedir, kiçik işlər üstünlük təşkil edir [24]. İşlərin ən azı 90%-də giriş verilənlərinin həcmi 100 Qb-dan kiçikdir. *Facebook*-da *Hadoop* yüklənməsinin diqqətlə öyrənilməsi nəticəsində aşkarlanmışdır ki, işlərin çox kiçik azlığı terabayt və daha böyük miqyasa çatır, işlərin əksəriyyətində giriş və çıxış verilənləri meqabayt-qiqabayt diapazonunda olur [25].

Fərdi məlumatların qorunması. *Big Data* texnologiyaları şəxsi həyatın toxunulmazlığı baxımından ciddi problemlər yaradır. Müxtəlif sosial-iqtisadi aktorlar istifadəçilər haqqında olduqca çox verilənlər toplayırlar – istifadəçinin veb-saytdakı sosial şəbəkələrdəki davranışı, istifadəçi ilə əlaqəsi olan şəxslərin davranışları və əlaqələri, onların alış-veriş davranışları və s. Bundan başqa, kommunikasiyaların emosional çalarları da analiz edilir. İstifadəçinin sosial

mediada nə yazdığı ilə yanaşı, necə yazdığını da analiz edirlər. Bir sözlə, potensial istifadəçi barəsində vacib və ya lazımsız görünən hər şey toplanır və onun 360 dərəcəli profili yaradılır.

Beləliklə, müasir izləmə texnologiyaları və *Big Data* analizi əsasında insanın "beşikdən qəbrədək izlənilməsi" imkanın yaranması onun şəxsi həyatının toxunulmazlığını qorumaq cəhdləri ilə cəmiyyətin məlumat tələbatı arasında olduqca kritik dilemma yaradır [26].

Nəticə

Hazırda *Big Data* sahəsinə ictimai maraq pik həddindədir. Hələlik *Big Data* texnologiyalarının yaxın gələcəkdə insanların həyatında fərdi kompüter və İnternet kimi zərurətə çevriləcəyini söyləmək çətindir, lakin onun həyatımızı və biznes mühitini dəyişdirməsi göz önündədir və bu, heç kimdə şübhə yaratmır.

Big Data texnologiyalarının böyük potensialı var və müxtəlif fəaliyyət sahələrinə ciddi təsir edəcəkdir. Təkcə korporativ informasiya arxitekturunu dəyişməklə iş bitmir. Demək olar ki, təşkilatın bütün struktur bölmələrinin işində dəyişikliklər etmək tələb olunur. Verilənlərin analizi prosesi dəyişərək təcrid olunmuş funksional sahədən biznes-proseslərin tərkib hissəsinə çevriləcək. *Big Data* texnologiyalarının böyük potensialından səmərəli faydalanmaq üçün böyük həcmli verilənlərin idarə edilməsi və analizinə məqsədyönlü və yaxşı düşünülmüş *Big Data* strategiyasının işlənilməsi xüsusi əhəmiyyət daşıyır.

Ədəbiyyat

1. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A.H. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. 2011.
2. Baaziz A., Quoniam L. How to use Big Data technologies to optimize operations in Upstream Petroleum Industry / International Journal of Innovation, 2013, vol. 1, no. 1, pp. 19-29.
3. Feblowitz J. The Big Deal about Big Data in upstream oil and gas. IDC Energy Insights. October 2012.
4. Editorial: Community cleverness required // Nature, 4 September 2008, vol. 455, no. 7209, pp. 1-1. doi:10.1038/455001a
5. Dean J., Ghemawat S. MapReduce: Simplified data processing on large clusters / Proc. of the 6th Conference on Symposium on Opearting Systems Design & Implementation (OSDI'04), 2004, vol. 6, pp. 137-150.
6. Han J., Kamber M., Jian P. Data mining: concepts and techniques. Morgan Kaufmann, 2006.
7. Bishop C.M. Pattern recognition and machine learning. Springer. 2006.
8. Feldman R., Sanger J. The Text Mining Handbook: Advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.
9. Junqué de Fortuny E., Martens D., Provost F. Predictive modelling with Big Data: Is bigger really better? // Big Data, 2013, vol. 1, no. 4, pp. 215-226.
10. Weiss Sh. M., Indurkhya N., Zhang T., Damerau F. Text Mining: Predictive methods for analyzing unstructured information. Springer; 2005, 260 p.
11. Aliguliyev R.M. A new sentence similarity measure and sentence based extractive technique for automatic text summarization // Expert Systems with Applications, vol. 36, no. 4, 2009, pp. 7764–7772.
12. Alguliev R.M., Aliguliyev R.M., Isazade N.R. Multiple documents summarization based on evolutionary optimization algorithm // Expert Systems with Applications, 2013, vol. 40, no. 5, pp. 1675-1689.
13. Siegel E. Predictive Analytics: The power to predict who will click, buy, lie, or die. Wiley; 1st edition. 2013. 320 p.

14. Karthik K., Kollias G., Kumar V., Grama A. Trends in Big Data analytics / Journal of Parallel and Distributed Computing, 2014, vol. 74, no. 7, pp. 2561-2573.
15. White T. Hadoop: The definitive guide. O'Reilly Media, Inc., 2012.
16. Ghemawat S., Gobioff H., Leung S. The Google file system / Proc. of the 19th ACM Symposium on Operating Systems Principles, 2003, pp. 29-43.
17. Anglade T. noSQL Tapes. http://www.nosqltapes.com.
18. Stonebraker M., Madden S., Abadi D. J., Harizopoulos S., Hachem N., Helland P. End of an Architectural Era (It's Time for a Complete Rewrite) / Proc. of the 33rd International Conference on Very Large Data Bases (VLDB '07), 2007, pp. 1150-1160.
19. Agrawal D., Das S., El Abbadi A. Big data and cloud computing: current state and future opportunities / Proc. of the 14th International Conference on Extending Database Technology, 2011, pp. 530-533.
20. Shvachko K., Kuang H., Radia S., Chansler R. The Hadoop distributed file system / IEEE 26th Symposium on Mass Storage Systems and Technologies, 2010, pp. 1-10.
21. Lee K.H., Lee Y.J., Choi H., Chung Y.D., Moon B. Parallel data processing with MapReduce: a survey // ACM SIGMOD Record, 2012, vol. 40, no. 4, pp. 11-20.
22. Leinweber D., Stupid Data Miner tricks: Overfitting the S&P 500 // The Journal of Investing, 2007, vol. 16, no. 1, pp. 15-22.
23. Rowstron A., Narayanan D., Donnelly A., O'Shea G., Douglas A., Nobody ever got fired for using Hadoop on a cluster / Proc. of the Workshop on Hot Topics in Cloud Data Processing (HotCDP), 2012, Article No. 2. doi:10.1145/2169090.2169092
24. Ananthanarayanan G., Ghodsi A., Wang A., Borthakur D., Kandula S., Shenker S., Stoica I. PACMan: Coordinated memory caching for parallel jobs / Proc. of the 9th USENIX Conference on Networked Systems Design and Implementation, 2012, pp. 20.
25. Chen Y., Alspaugh S., Katz R.H. Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads // Proc. of the VLDB Endowment (PVLDB), 2012, vol. 5, no. 12, pp. 1802–1813.
26. Tene O., Polonetsky J. Privacy in the age of big data: A time for big decisions // Stanford Law Review Online, 2012. http://www.stanfordlawreview.org/online/privacy-paradox/big-data

УДК 004.9:351

Имамвердиев Ядигар Н.
Институт Информационных Технологий НАНА, Баку, Азербайджан
yadigar@lan.ab.az

Большие перспективы и проблемы технологии Больших Данных

Большие Данные охватывают технологии и инструменты для сбора, обработки, анализа и извлечения полезных знаний из структурированных и неструктурированных данных большего объема, генерируемых с высокой скоростью разными источниками. В последнее время научная и популярная литература представляет Большие Данные как технологию, открывающую новые перспективы и революционные изменения в электронном государстве, бизнесе, здравоохранении, науке, производстве и других областях деятельности. С целью определения истинного потенциала аргументов, поддерживающих эти утверждения, и выбора правильной стратегии для больших данных в этой работе критически анализируются сущность, характеристики и история становления, основные строительные компоненты и аналитические возможности этих технологий, а также указываются преимущества, перспективы и существующие проблемы.

Ключевые слова: *Big Data, Большие Данные, аналитика больших данных, интеллектуальный анализ данных, Hadoop, предиктивная модель.*

Yadigar N. Imamverdiyev
Institute of Information Technology of ANAS, Baku, Azerbaijan
yadigar@lan.ab.az

Big prospects and problems of Big Data technology

Big Data covers technologies and tools for collecting, processing, analyzing and extracting useful knowledge from structured and unstructured data of large volumes generated at high speed by different sources. Recently, scientific and popular literature promotes Big Data as technology which opens new perspectives and revolutionary changes in e-government, business, health, science, industry and other fields. In order to determine the true potential of arguments supporting these assertions and to choose the right strategy for Big Data this paper critically examines essentials, characteristics, basic building components and analytical capabilities of Big Data, and identifies advantages, prospects and existing problems.

Keywords: *Big Data; Big Data analytics; Data Mining; Hadoop; predictive model.*