Proqram Mühəndisliyi və Data Science

Məkrufə Hacırəhimova[1], Hicran Gözəlova[2]
[1,2]İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*makrufa@science.az*, [2]*gozalova@yandex.ru*

***Xülasə*— "Big data" konsepsiyasının populyarlaşması ilə «data science» multidissiplinar elmi sahəyə çevrilmişdir. Məqalədə proqram mühəndisliyi üçün"Data Science" və "Data Science" üçün proqram mühəndisliyi məsələləri təhlil edilmişdir.**

Açar sözlər— big data, proqram təminatı, proqram mühəndisliyi, proqram mühəndisləri, Data Science, verilənlər alimləri

I. GİRİŞ

Big Data (BD) konepsiyasının əsasını müxtəlif mənbələrdən fərqli formatlarda yüksək sürətlə daxil olan böyük həcmli verilənlərin emalı, analizi və onlardan faydalı biliklərin çıxarılması təşkil edir. Verilənlərdən biliklərin əldə edilməsində yeni intellektual analiz modellərinin işlənməsi və proqram təminatlarının yaradılması kimi məsələlər aktuallaşır [1-3]. Bu konsepsiyanın inkişafı Verilənlər haqqında elm (Data science – DS) bəzən də datalogiya adlandırılan (datalogy) yeni elm sahəsinin yaranmasına gətirib çıxarmışdır [4-6]. Bu akademik sahə informatikanın bir bölməsi olmaqla rəqəmsal verilənlərin emalı, analizi və təqdim olunması problemlərini öyrənir [5-7]. O, riyaziyyat, statistika, obrazların tanınması, biliklər bazası, maşın təlimi və s. kimi multidissiplinar yanaşmaları birləşdirir. Yeni nəzəriyyələr, yeni üsullar, analitik alətlər alimlərə və biznes nümayəndələrinə BD-də gizli şəkildə olan biliklərin aşkarlanmasında kömək edə bilər. Bu da DS-in əsasını təşkil edir [7, 8]. BD erasında DS proqram mühəndislərinin yiyələndikləri xüsusi əhəmiyyətli bacarıqlar yerinə yetirilmiş layihələr əsasında yeni layihələri proqnozlaşdırmağa imkan verir [4]. DS ilə müqayisədə proqram mühəndisliyi (PM) proqram təminatının sistemli şəkildə layihələndirilməsi, işlənməsi, fəaliyyəti və müşayiət olunmasını təmin edən, həmçinin bu yanaşmaları tədqiq edən elmi nəzəri və pratik sahə olmaqla keçən əsrin 60-cı illərindən mövcuddur. Tədqiqatlar göstərir ki, son illər DS və PM sahələri bir-biri ilə inteqrasiyaolunmaqdadır. Ona görə də bu məsələyə iki cür yanaşmaq olar: DS-də PM və əksinə [8–12].

II. DATA SCİENCE ÜÇÜN PROQRAM MÜHƏNDİSLİYİ

BD-nin analizi bir çox layihələrin proqram təminatlarının işlənib hazırlanmasının əsasını təşkil edir. DS inkişaf etməkdə olan sahə olduğundan proqram mühəndislərinə təcrübəni yaxşılaşdırmağa kömək edir.

Proqram təminatı verilənlərin analizinin konkret aspektinin yekunudur. Verilənlərin analizi yerinə yetirilərkən, proqram təminatı, təkrar tətbiq oluna biləcək alətlər toplusunun təyin edilmiş bir modula tətbiqinitəmin edir, həmçininproseduru sistemləşdirməyə və standartlaşdırmağa imkan verir [4, 7].

Proqram təminatı analiz üçün dəqiq interfeys işləyib hazırlamaqla, prosedur və ya alətlərin funksionallığını formalaşdırır. Proqram təminatının sadədən mürəkkəbə üç səviyyəsi var:

- birinci səviyyədə, başlanğıc üçün prosedurun qısa təsvir formasından ibarət xüsusi kod yazılır;
- ikinci səviyyə istənilən proqramlaşdırma dillərindən birində yazılmış hər hansı bir funksiya ola bilər. Seçilmiş proqramlaşdırma dilinin qrafik qurmaq imkanı olarsa istifiadəçiyə giriş verilənlərini və çıxış qiymətlərini bilmək yetərlidir. Bu səviyyədə funksiyanın interfeysini təyin etmək vacibdir;
- üçüncü səviyyə funksiyalar toplusundan ibarəttam proqram təminatı paketidir. O, özünün xüsusi interfeysi ilə seçilir və bir neçə məsələyə tətbiq edilə bilir [4,6,11].

Son illərdə əməliyyatların optimallaşdırılması və qərarların qəbul edilməsi kimi məsələlər DS-dən istifadə etməklə yerinə yetirilir. Verilənlərin biliyə çevrilməsi üçün DS və verilənlər alimlərinə (Data Scientist) ehtiyac vardır. Verilənlər strukturlaşmış və ya strukturlaşmamış, yəni müxtəlif formatlı – mətn, təsvir, video və s. olur. Verilənlər alimləri, biznes-qərarların qəbul edilməsinə kömək məqsədilə verilənlərdən biznes məsələlərin həllində və ya verilənlərin analizində istifadə edirlər [2, 5].

III. PROQRAM MÜHƏNDİSLİYİ ÜÇÜN DATA SCİENCE

PM sahəsində təcrübələri təkmilləşdirmək məqsədilə tədqiqatçılar DS-dən geniş istifadə edirlər. Bu, özünü bir neçə aspektdən daha qabarıq büruzə verir:

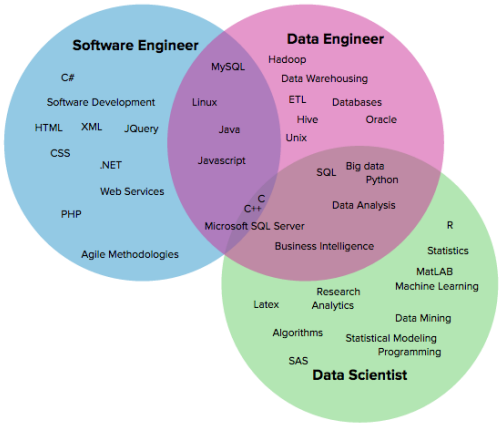
- məhsuldarlıq baxımından böyük həcmli verilənlərin analizi üçün verilənlərin təmizlənməsi və yeni, genişləndirilmiş analizdən istifadə edilməsi;
- verilənlərin mənasının və mənşəyinin qorunub saxlanması üçün müvafiq analizin aparılması;
- kommunikasiya və koordinasiya cəhətdən qeyri-müəyyənlikləri və riskləlri müzakirə etmək və verilənlər əsasında qərar qəbul etmək üçün xüsusi üsulların tətbiq edilməsi və s. [10].

DS, proqram təminatının işlənib hazırlanmasında, onun həyat tsiklinin daha effektiv idarə edilməsində proqram mühəndislərinə çox böyük dəstək verir.

Son illər verilənlərin artması ilə onların idarə edilməsində əsas rol oynayan mühəndis proqramçılar, verilənlər mühəndisləri və verilənlər alimləri arasında fərqlər müşahidə olunur. Bu fərqlilik şəkil 1-də təqdim olunmuş diaqramda öz əksini tapmışdır [10, 11]. Verilənlər mühəndisliyi (Data Engeneering) və

DOI: 10.25045/NCSoftEng.2017.20

DS daha gənc sahələrdəndir (qeyd etmək lazımdır ki, verilənlər mühəndisliyi əvvəllər PM-in tərkibinə daxil olmuşdur).



Şəkil 1. Proqram mühəndislərin verilənlər mühəndisi və verilənlər alimləri ilə müqayisəsi

Proqram mühəndisləri testetmə və analiz də daxil olmaqla, layihələndirmədən başlamış proqramın tərtib edilməsinə qədər bütün mərhələlərdə iştirak edərək sistem və proqram əlavələrinin yaradılmasını həyata keçirirlər. Proqram mühəndislərinin əsas işlərinə aşağıdakıları daxil etmək olar:

- interfeysin yaradılması;
- veb və mobil əlavələrin hazırlanması;
- əmliyyat sisteminin və proqram təminatının işlənib hazırlanması və s. [12,13].

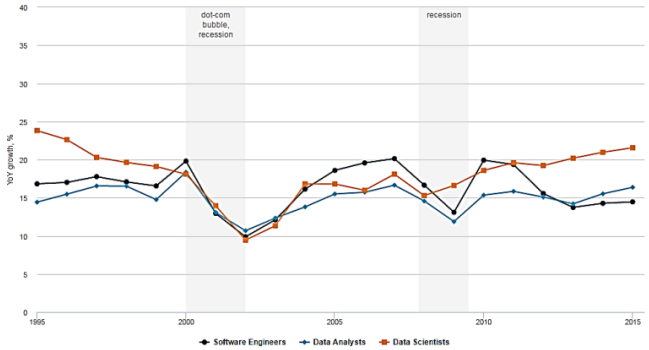
Verilənlər mühəndisi sistem yaradır, proqram təminatı mütəxəssisləri tərəfindən yaradılmış sistem və əlavələrdən verilənlərin əldə edilməsini və saxlanmasını təmin edir. Qeyd etmək lazımdır ki, verilənlər mühəndislərinin 40%-i əvvəllər mühəndis-proqramçı kimi fəaliyyət göstərmişlər. Onların işlərinə daxildir:

- verilənlərin genişləndirilmiş strukturu;
- paylanmış hesablamalar;
- paralel proqramlaşdırma;
- yeni biliklər və yeni alətlər (Hadoop, Spark, Kafka və s.);
- verilənlər bazasında ETL (Extract, Transform, Load) prosesinin qurulması.

Verilənlər alimləri verilənlər üzərində analiz aparırlar. Bu əməliyyat intellektual analiz və ya maşın təlimi alqoritmi ola bilər. Bu alqoritmlər sonradan proqram mühəndisləri və verilənlər mühəndisləri tərəfindən koda çevrilir və reallaşdırılır. Verilənlərin modelləşdirilməsi, maşın təlimi, alqoritmlər, biznes analitika və s. bu sahənin əsas işləridir [11-13].

Yaşadığımız Big data erasında daha dərin analitik təhlillərin aparılması və alətlərin işlənməsində həm verilənlər alimlərinə, həm də proqram mühəndislərinə böyük ehtiyac yaranmışdır. Verilənlərin paylanmış və paralel hesablama texnologiyaları, mobil və sensor texnologiyalar, Əşyaların İnterneti, sosial şəbəkələr və s. bunu şərtləndirən amillərdəndir. Dünyanın aparıcı universitetlərində verilənlər elmi və PM ixtisasları üzrə fənlərin müxtəlif təhsil pillələrində tədrisi buna əyani sübutdur.

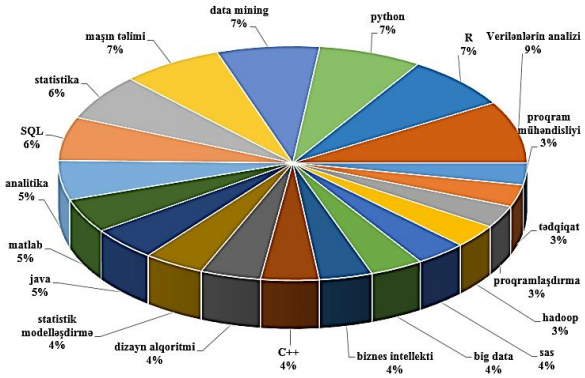
Müqayisə üçün PM, verilənlər alimivə verilənlər mühəndisliyi mütəxəssislərinin illər üzrə dinamikasına nəzər salaq (şəkil 2.):



Şəkil 2. PM və DS mütəxəssislərinin sayının illər üzrə artım tempi

Şəkil 2-dən göründüyü kimi, müxtəlif dövrlərdə hər üç sahə üzrə mütəxəssislərin sayında artma və azalmalar müşahidə olunur, lakin 2012-ci ildən başlayaraq, verilənlər alimlərinin sayında ardıcıl olaraq 50%-ə qədər artım diqqəti cəlb edir [11, 13].

Hazırda BD-nın toplanması, təmizlənməsi, analizi, proqram təminatının yaradılması və s. məsələləri yerinə yetirmək üçün verilənlər elmindən istifadə olunur. 2013-cü ildən başlayaraq dünyanın bir çox aparıcı universitetlərində “data science” akademik fənn kimi bakalavr, magistr və doktorant pilləsində tədris olunur [14, 15]. DS-in tədrisində ən çox öyrədilən bacarıqlar aşağıdakı şəkildə təsvir olunmuşdur (şəkil 3.):



Şəkil 3. DS-in tədrisində öyrədilən bacarıqlar

IV. VERİLƏNLƏR ALİMLƏRİNƏ ARTAN TƏLƏBAT

Proqram təminatının yaradılması və *istifadəyə verilməsi* (işə salınması) işlənib hazırlanma prosesi haqqında böyük həcmdə xam verilənləri əmələ gətirir [16]. Bu verilənləri emal etmək üçün yüksəkixtisaslı verilənlər alimlərinə ehtiyac var. Son zamanlar, proqram vasitələrindən istifadə etməklə verilənlərin analizinə aparan proqram təminatı işləyib hazırlayan təşkilatlar inkişaf etməyə başlamışlar [17,18]. Belə böyük həcmli verilənləri analiz etmək üçün Microsoft şirkətianaltik bacarıqlara və proqram təminatı işləmək qabilliyyətinə malik olan mütəxəssilərin iş stillərini araşdırmaq məqsədilə sorğu aparmışdır. Sorğunun nəticəsi olaraq, verilənlər alimlərini beş qrupa ayırırlar:

DOI: 10.25045/NCSoftEng.2017.20

1. Qərarların qəbulu üçün vacib olan verilənləri toplayan mühəndislərlə işləyən daxili provayderlər;

2. İntellektual modellərin yaradılması üçün maşın təlimi təcrübəsini istifadə edən modelləşdirmə mütəxəssisləri;

3. Mühəndis analizi və verilənlər analizi arasında sualları balanslaşdıran verilənlər platformasını yaradanlar;

4. Verilənlər sahəsində tam fəaliyyət göstərən mütəxəssislər (polymath);

5. Verilənlər alimlərinə rəhbərlik edən və mükəmməl təcrübəni yayan qrup rəhbərləri [16,19].

Verilənlər alimləri mütəxəssislər qrupunun tərkibinə təyin edilə bilər, yaxud ixtisaslaşmış DS qrupunun rəhbəri ola bilər. Onlar həm fərdi, həm də komanda şəklində işlədikdə, məhsuldarlıq daha yüksək olar. Belə olan halda çoxlu sayda müxtəlif problemləri olan təşkilatlarda bir-birini dəstəkləyən, təşkilatı DS haqqında məlumatlandıran qrup şəklində işləyən verilənlər alimləri daha yüksək nəticəyə nail ola bilərlər [14,16].

Bu gün böyük həcmli verilənlərdən faydalı informasiyanın aşkarlanması üçün ən qabaqcıl metodlar tətbiq edilir və effektiv texnoloji alətlər yaradılır.

Verilənlər alimləri daha geniş yayılmış Python və R proqramlaşdırma dillərindən istifadə edirlər. Bu açıq mənbə dilləri (open-source languages) mütəxəssislər tərəfindən dəstəklənir. Bundan əlavə DS-in tətbiqi üçün NumPy və SciPy kimi bir çox əlavə paketlər hazırlanmışdır. Bu dillər kompilyasiya edilmir, verilənlər alimlərinin dilin nüanslarına deyil, problemlərə istiqamətlənməsinə imkan yaradır. Verilənlər alimləri arasında populyar olan dillərdən biri də kompilyasiya edilmiş Scala dilidir [18, 19]. Bu dillərin MatLab, Stata, və SPSS kommersiya tipli dillərdən üstünlüyü onların istifadə üçün açıq giriş imkanına malik olmasıdır.

Proqram kodunun test edilməsi riyazi cəhətdən çox mürəkkəbdir. Bununla belə, verilənlər alimləri öz kod və alqoritmlərini test edirlər. Bu testlər o qədər etibarlı olmamasa da, bəzi DS-lər onlardan istifadə etməklə yoxlamalar aparır, testerlər isə onlarla işləyə bilirlər. Test edilmənin səmərəliliyi üçün verilənlər alimlərinin standart üsullara riayət etməklə yaxşı sənədləşdirilmiş və təkrarlanan kod yazmaları məqsədəuyğun hesab edilir [20-23].

NƏTİCƏ

Böyük verilənlərin təhlili bir çox təşkilatlar üçün həssas mövzu olaraq qalmaqdadır. Böyük verilənlərin proqram təminatının yaradılmasında proqram mühəndislərinin rolu böyükdür. Onlarla yanaşı, müəssisəyə daha çox gəlir gətirə biləcək mütəxəssislərin -verilənlər alimlərinin də bu sahədə rolu böyükdür.

BD-ni idarə etmək üçün “verilənlər alimləri”nin əsas bacarığı kod yazmaqdan başlayaraq, verilənləri analiz etmək və hamının anladığı tərzdə verilənləri vizuallaşdırmaqdır.

Verilənlər alimləri həm təşkilatda işləyənlərin təhsili ilə, həm də statistik alqoritmlərin realizasiyası ilə məşğul ola bilərlər. Həmçinin təşkilatlarda bir-birini dəstəkləyən, təşkilatı DS haqqında məlumatlandıran və qrup şəklində işləyən verilənlər alimləri daha yüksək nəticəyə nail ola bilərlər.

ƏDƏBİYYAT

[1] R.M. Əliquliyev, M. Ş. Hacırəhimova, “Big data” fenomeni: problemlər və imkanlar, İnformasiya texnologiyaları problemləri, 2014, №2, s. 3-16

[2] Y.N. İmamverdiyev, “Big data texnologiyalarinin böyük perspektivləri və problemləri”, İnformasiya cəmiyyəti problemləri, 2016, №1, s. 23–34

[3] R.M. Əliquliyev, M. Ş. Hacırəhimova, A.S Əliyeva, "Big data-nin aktual elmi-nəzəri problemləri”, İnformasiya cəmiyyəti problemləri, 2016, №2, s. 37–49

[4] M. Kim, T. Zimmermann, R. DeLine, Andrew Begel “The Emerging Role of Data Scientists on Software Development Teams”, ICSE '16 38th International Conference on Software Engineering, Austin, TX, USA — May 14 - 22, 2016

[5] Naur P. Datalogy, the science of data and of data processes and its place in education, Proc. IFIP Congress, Edinburgh, Scotland,. Amsterdam: North-Holland, 1968, pp. 48-52

[6] F. Jack Data science as an academic discipline Data Science Journal, Volume 5, 19 October 2006,pp. 163-164

[7] S.Madden, From Databases to Big Data, IEEE Internet Computing, 2012, vol.16, no 3, pp.4–6.

[8] 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/

[9] J. Fan, F.Han, H. Liu, Challenges of Big Data analysis, National Science Review, 2014, vol. 1, no 2, pp. 293–314.

[10] R. DeLine “Research Opportunities for the Big Data Era of Software Engineering”, 2015 IEEE/ACM 1st International Workshop on Big Data Software Engineering, pp. 26-29, 2015

[11] T. Menzies, E. Kocaguneli, F. Peters, B. Turhan “Data science for software engineering”, ICSE '13 35th International Conference on Software Engineering, San Francisco, CA, USA — May 18 - 26, 2013, pp.96-107

[12] Gorton, A. Basar Bener, A. Mockus“Software Engineering for Big Data Systems”, March/April 2016 | IEEE SOFTWARE, https://www.computer.org/csdl/mags/so/2016/02/mso2016020032.pdf

[13] https://www.stitchdata.com/resources/reports/the-state-of-data-science/

[14] M.Ş. Hacırəhimova, H.Y. Gözəlova, “Böyük Verilənlərdən Verilənlər Haqqında Elmə: Fənlərarası Perspektiv”, “Big data: imkanları, multidissiplinar problemləri və perspektivləri” I respublika elmi-praktiki konfransı, Bakı şəhəri, 25 fevral 2016, s. 184-187

[15] M.Ş. Hacırəhimova, H.Y Gözəlova, “Data Science və Onun Tibbi İxtisaslar üzrə Proyeksiyaları”,“Elektron tibbin multidissiplinar problemləri” I respublika elmi-praktiki konfransı, Bakı, 24 may 2016

[16] M. Kim, T. Zimmermann, R. DeLine, A. Begel “The Emerging Role of Data Scientists on Software Development Teams”, https://sfu-db.github.io/cmpt884-fall16/Papers/kim-icse-2016.pdf

[17] K. M. Anderson “Embrace the Challenges: Software Engineering in a Big Data World”, Big Data Software Engineering (BIGDSE), 2015 IEEE/ACM 1st International Workshop on Big Data Software Engineering, pp. 19-25, 2015

[18] NESSI – Software Engineering White Paper: Software engineering for and with Big Data, 2014, http://www.nessi-europe.eu/Files/Private/NESSI_SE_WhitePaper-FINAL.pdf

[19] M. Chambers, C. Doig, I. Stokes-Rees “Breaking Data Science Open”,O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA

[20] http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html

[21] http://www.seagate.com/files/www-content/ti-dm/_shared/images/r-and-python-pv0026-1-1409us.pdf

[22] B. Christian, Madsen et al., “Python for Analytics and The Role of R”, http://www.seagate.com/files/www-content/ti-dm/_shared/images/r-and-python-pv0026-1-1409us.pdf

[23] https://datafloq.com/read/5-best-python-libraries-for-data-science/994

DOI: 10.25045/NCSoftEng.2017.20