Proqram Mühəndisliyində Big Data Problemləri

Tamilla Bayramova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
tamilla@iit.ab.az

***Xülasə* — Məqalədə "Big Data" konsepsiyasının rellaşdırılmasında istifadə edilən əsas texnologiyalar və onların yaradılması üçün istifadə edilən proqramlaşdırma dilləri araşdırılmışdır. Böyük verilənlərin emalı üçün proqram vasitələrini işləyərkən meydana gələn problemlər təhlil olunmuşdur.**

Açar sözlər — Big Data, Hadoop, MapReduce, SCOPE, JAQL, NoSQL

I. GİRİŞ

Müasir dünyada informasiyanın həcmi eksponensial şəkildə artır, onun saxlanılması və emalı çətinləşir. Bunlar sosial şəbəkələrdən daxil olan verilənlər, fərdi verilənlər, sensor qurğularının göstəriciləri, tranzaksiyalar, administrativ verilənlər və s. ola bilər. Bu informasiyaların böyük hissəsi strukturlaşdırılmamış verilənlərin payına düşür. Böyük verilənlərdən istifadə edərkən bir sıra çətinliklər və suallar yaranır (məsələn, verilənlərə əlyetərlilik, etik məsələlər, verilənlərin yadda saxlanılması və metodoloji problemlər). Bu verilənlər cəmiyyətə, siyasətə və iqtisadiyyata əhəmiyyətli dərəcədə xeyir gətirə bilər. İnsanlar haqqında ətraflı verilənlərin olması cinayətkarlığın azaldılması, tibbi xidmətin keyfiyyətinin artması, idarəetmənin səmərəliliyinin yüksəldilməsi və s. bu kimi sahələrdə istifadə edilə bilər. Böyük həcmdə informasiyanın emalı üçün 2000-ci illərdə bir bir qrup konsepsiya, metod və instrumental vasitələr yaradıldı və bunlar "Big Data" texnologiyaları adı altında ümumiləşdirildi [1].

Big Data termini çoxsaylı mübahisələr yaradır, əksəriyyət belə hesab edir ki, bu ənənəvi üsullarla emal edilə bilməyən, strukturlaşdırılmış və strukturlaşdırılmamış böyük informasiya yığınıdır. Ümumilikdə isə Big Data aşağıda göstərilən əlamətlərlə xarakterizə olunur [2]:

- ***Həcm*** (volume) – verilənlər bazası emalı və saxlanılması ənənəvi üsullarla çox çətin olan böyük həcmli informasiyadan ibarətdir və onlarla işləmək üçün yeni yanaşma və təkmilləşdirilmiş vasitələr tələb olunur;
- ***Sürət*** (velocity) – emalın nəticələrinin vaxtında alınması çox vacib amillərdəndir, həcm artdıqca emal üçün çox böyük sürət tələb olunur;
- ***Müxtəliflik*** (variety) – eyni zamanda müxtəlif formatlı strukturlaşdırılmış və strukturlaşdırılmamış informasiyanı emal etmək imkanı olmalıdır;
- ***Həqiqilik*** (veracity) – toplanan məlumatların düzgünlüyünə diqqət yetirilməlidir;
- ***Dəyər*** (value) – verilənlər şirkət üçün faydalı olmalıdır və ona xeyir gətirməlidir.

Məhdud zaman ərzində və ya verilən vaxt müddətində nəticələrin alınmasının vacibliyi böyük verilənlərin təqribi emalı metodlarının tətbiqinə gətirir. Big Data konsepsiyasının rellaşdırılmasında istifadə edilən əsas texnologiyalar və onların yaradılması üçün istifadə edilən proqramlaşdırma dilləri aşağıda araşdırılmışdır.

II. BİG DATA TEXNOLOGİYALARINDA İSTİFADƏ EDİLƏN PROQRAM VASİTƏLƏRİNİN BƏZİ PROBLEMLƏRİ

Böyük verilənlər (BV) müxtəlif mənbələrdən daxil olduğundan onları öz aralarında əlaqələndirmək, çevirmək və sənədləşdirmək lazımdır. Verilənlərin təhlili zamanı xətaların mənbəyinin və təbiətinin əvvəlcədən müəyyən olunması BV-də yaranan riskləri azaldır. Bunun üçün də verilənləri generasiya edən, massivlərin yaradılmasında və emalında istifadə edilən proqram vasitələrinə diqqət yetirilməlidir. İnformasiya sistemlərinin məhsuldarlığının və yaddaş qurğularının həcminin artması avadanlıqların imkanlarının qeyri məhdud olması haqda fikirlər yaradır. Hal-hazırda emal ediləcək verilənlər ilə onların saxlanılma imkanları arasındakı fərq sürətlə artır və verilənlərin səmərəli şəkildə idarə edilmə problemləri daha da aktuallaşır. Getdikcə artan verilənlərin təhlili və emali üçün keyfiyyətsiz proqram kodunun yazılması isə sistemlərin istismarı zamanı həlli mümkün olmayan problemlər yarada bilər [3].

Verilənləri idarə etmək üçün proqramları yazarkən düşünmədən obyekt-yönümlü metodologiyadan istifadə edilməsi çox pis nəticələrə gətirə bilər. Belə ki, obyektlər şəklində yaradılmış kiçik sorğuların sayı artdıqca onlar böyük informasiya axınına çevrilir və ən müasir çoxprosessorlu serverlər belə bu axını emal edə bilmir. Daxil olan sorğuların əksəriyyəti təkrarlandığı üçün onların və proqram kodunun optimallaşdırılması lazım gəlir. Verilənlərin analitik emalı məsələləri Big Data problemlərini həll etməyə yönəlib. Sorğuların cavablarını sürətləndirmək və istifadəçilərə daha keyfiyyətli xidmət göstərmək üçün müəssisələr Big Data texnologiyalarından istifadə etməli olurlar. Böyük həcmli verilənlərin emalı və analizini həyata keçirmək məqsədi ilə informasiya texnologiyaları (İT) sahəsinin nəhəngləri tərəfindən proqram-aparat platformaları yaradılmışdır.

Onlardan **Apache Hadoop** böyük həcmli (terabayt, petabayt) verilənləri yadda saxlamağa və BV-lə əlaqəli alqoritmləri paralel emal etməyə imkan verən paylanmış fayl sisteminin işlənilməsi və istifadə edilməsi üçün proqram platformasıdır (Software Framework) [4]. Hazırda Hadoop ən böyük İnternet-servis və İT şirkətlərində uğurla tətbiq olunur, bunlardan Yahoo, Facebook, Twitter, Amazon, Apple, Ebay və s. göstərmək olar. Hadoop platformasında əsasən 2 komponentdən istifadə edilirr:

- **Hadoop Distributed File System** (HDFS) verilənləri sürətli əlyetərliliyi təmin edən paylanmış fayl sistemidir;
- **MapReduce** hesablama klasterində böyük həcmli verilənlərin paylanmış emalı üçün proqram platformasıdır.

Bəzi hallarda əvvəl paralel sonra ardıcıl və yenidən paralel hesablamalar aparmaq lazım gəlir. MapReduce elə layihələndirilib ki, həm aralıq, həm də son nəticələr diskə yazılır. Ona görə də yuxarıda qeyd etdiyimiz tip məsələlərin həllində diskə yazma və oxuma vaxtlarını nəzərə alsaq hesablama müddəti bir neçə dəfə artacaq. Belə hallarda Spark proqram platforması tətbiq edilə bilər, burada əksər aralıq hesablamalar diskə deyil, yaddaşa göndərilir.

Cədvəl formalı verilənlərin təhlili daha asan olduğundan Hadoop infrastrukturunda SQL-yönümlü instrumental vasitələr də var [5]:

- **Hive** – HDFS fayl sistemində yadda saxlanılmış verilənlər üzərində kifayət qədər mürəkkəb sorğuları yerinə yetirməyə imkan verir.
- **Impala** – Cloudera şirkətinin məhsuludur. Sonuncudan fərqli olaraq MapReduce platformasından deyil, C++ dilində yazılmış öz platformasından istifadə edir.
- **Spark SQL** – Spark platformasında işləmək üçün yaradılmışdır.

Hadoop bazasında BV-nin emalı və analizi zamanı SQL – vasitələr bəzi hallarda kifayət etmir. Belə məsələlərin həllində NoSQL (Not only SQL) bazaları daha səmərəli olduğu üçün **HBase** proqramından istifadə edilir.

Geniş tətbiqinə baxmayaraq Hadoop proqram platforması bəzi məhdudiyyətlərə malikdir [6]:

- Hadoop klasterinin miqyaslanma imkanının məhduddur;
- Paylanmış hesablamaları yerinə yetirən alternativ proqram modeli yoxdur (Hadoop 1.0 ancaq Map Reduce hesablamalarını dəstəkləyir);
- Hadoop platformasında tək-tək imtina nöqtələrinin olduğuna görə etibarlılığa yüksək tələblərin qoyulduğu mühitlərdə istifadə edilə bilmir;
- Versiyalarındakı uyğunsuzluq: versiyasını yeniləndikdə bütün hesablama qovşaqlarında yenidən qurulmalıdır;
- Yeniləşən axın tipli verilənlərlə işi dəstəkləyə bilmir.

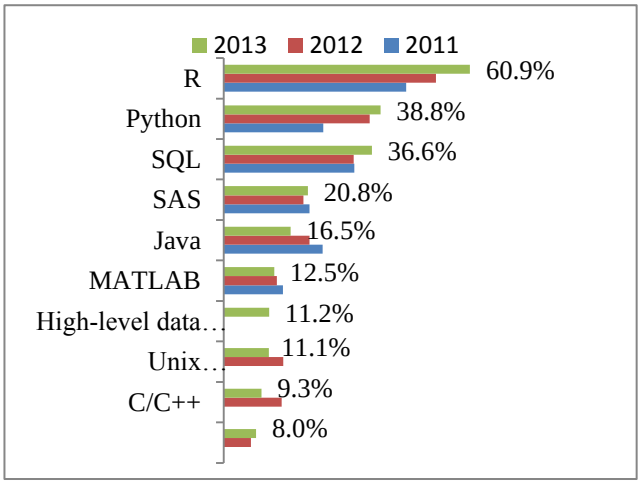
Çatışmazlıqlarına baxmayaraq özündə verilənlərin kütləvi-paralel emalı konsepsiya və texnologoyalarını birləşdirən Hadoop platforması BV-nin analizi üçün əsas platformadır.

Həm tədqiqat mərkəzlərinin, həm də böyük şirkətlərin diqqət mərkəzində olan belə maraqlı ideyalardan biri Hadoop və bulud texnologiyalarının birləşdirilməsidir. 2009-cu ildən başlayaraq IaaS/PaaS (İnfrastructure as a servis/Platform as a service) platformalarını işləyən iri şirkətlər bulud texnologiyaları ilə Hadoop platformasının birgə həllini təqdim etməyə başladılar.

Böyük verilənlərlə işləyən proqram təminatını yaradarkən proqram mühəndisliyinin əsas problemlərini nəzərə almaq lazımdır. Bunlar tələblərin təyin edilməsi, bu tələblərə uyğun proqram məhsulunun işlənilməsi və sınaq zamanı yaranan bir sıra problemlərdən ibarətdir. Belə proqramları işləyərkən aşağıdakı şərtlər nəzərə alınmalıdır:

- Miqyaslana bilmə: proqram getdikcə artan informasiyanı saxlamağa imkan verməlidir;
- Xətalara davamlı olmalıdır;
- İstənilən əməliyyat sistemində icra oluna bilməlidir;
- Məsələnin həllinin avtomatik olaraq paralelləşməsi;
- Çevik (strukturlaşdırılmış və strukturlaşdırılmamış verilənlərin saxlanması və təhlili) işləməlidir;
- Səmərəli olmalıdır;
- Risklər minimal olmalıdır.

Son illərdə Big Data texnologiyalarının yaradılmasında istifadə edilən proqramlaşdırma dillərinin populyarlığının artması bu sahəyə olan marağın artmasını göstərir. KDNuggets sorğu aparmışdır, 700 peşəkar proqramçının cavabları əsasında verilənlərin intellektual analizi, Data mining və elmi araşıdırmalarda ən çox istifadə edilən proqramlaşdırma dillərinin siyahısını vermişdir [7].



Şəkil 1. Proqramlaşdırma dillərinin reytinqi

Şəkil 1-dən göründüyü kimi proqramçıların verilənlərin statik emalı və qrafika ilə işləmək üçün yaradılmış R dilindən istifadəsi stabil olaraq artmaqda davam edir. Ümumilikdə Hadoop platforması ilə bu və ya digər səviyyədə əlaqəsi olan dillərdən istifadə 19% artmışdır. Bu proqramlaşdırma dilləri ilə yanaşı yeni proqramlaşdırma dillərindən də istifadə olunur, bunlara misal olaraq aşağıdakı dilləri göstərmək olar: SCOPE çevik yüksək səviyyəli proqramlaşdirma dilidir. İstifadəsi çox sadədir və analitik emalın müxtəlif klasterlərdə parallel yerinə yetirilməsi üçün nəzərdə tutulub [8].

Çox mürəkkəb və zəif strukturlu verilənlərin (məsələn, mətnlərin təhlili) emalı üçün JAQL (A JavaScript Object Notation Query Language) skript dili işlənilib [9].

NƏTİCƏ

Big Data texnologiyalarının işlənilməsində qazanılmış uğurlara baxmayaraq proqramçılar qarşısında bəzi açıq suallar qalır. Bunlar real zamanda prosesi dayandırmadan verilənlərin emal edən, xaker hücumlarının qarşısını alan məxfiliyi qorumaqla verilənləri emal edən və özü adaptiv proqram platformalarının işlənilməsidir.

ƏDƏBİYYAT

[1] Новиков Б.А., Графеева Н.Г., Михайлова Е.Г.Big Data: Новые задачи и современные подходы //Компьютерные инструменты в образовании, 2014, No4, с.10-18.
[2] Beyer Mark A., Laney Douglas. The Importance of «Big Data»: A Definition. Stamford, CT: Gartner, 2012.
[3] Tauheed Farhan, Nobari Sadegh, Biveinis Laurynas, Heinis Thomas, Ailamaki Anastasia. Computational Neuroscience Breakthroughs through Innovative Data Management, Proceedings of the 7th East-European Conference on Advances in Databases and Information Systems, Italy, 2013, p.14-27.
[4] http://www.codeinstinct.pro/2012/08/hadoop-overview.html
[5] http://www.zdnet.com/article/sql-and-hadoop-its-complicated
[6] http://www.codeinstinct.pro/2012/08/hadoop-design.html
[7] [http://www.kdnuggets.com/2013/08/languages-for-analytics-data-mining- data-science.html
[8] Chaiken Ronnie, Jenkins Bob, Larson Per-Ake, Ramsey Bill, Shakib Darren, Weaver Simon, Zhou Jingren. ˚ SCOPE: easy and efficient parallel processing of massive data sets, Journal proceedings of the VLDB Endowment, 2008. Vol 1, No. 2, p. 1265–1276.
[9] Beyer Kevin S., Ercegovac Vuk, Gemulla Rainer, Balmin Andrey, Eltabakh Mohamed Y., Kanne Carl-Christian, Ozcan Fatma, Shekita Eugene J.. ¨ Jaql: A Scripting Language for Large Scale Semistructured Data Analysis, Proceeding of the 37th International Conference on Very Large Data Bases, Seattle, WA, 2011, Vol. 4, No.2, p.1272-1283.