UOT 004.9

Alıquliyev R.M.[1], İmamverdiyev Y.N.[2]
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]r.aliguliyev@gmail.com, [2]yadigar@lan.ab.az

NEFT-QAZ SƏNAYESİ ÜÇÜN KONSEPTUAL BIG DATA ARXİTEKTURASI

Big Data texnologiyaları neft-qaz sənayesi sistemlərinin qurulması üçün vacib əhəmiyyət daşıyan yanaşmalar və alətlər təqdim edir. Bu işdə neft-qaz sənayesi sistemlərinə real zaman rejimində daxil olan böyük həcmli və müxtəlif formatlı verilənlərin paylanmış klaster sistemlərində saxlanması və onların dərin analitika və maşın təlimi metodları ilə analizi üçün nəzərdə tutulmuş hibrid Big Data platforması üçün konseptual arxitektura təklif edilir. Rəqabət qabiliyyətli Big Data həllinin yaradılması üçün Hadoop ekosistemindən zəruri alətlərin seçilməsi məsələsinə baxılır.

Açar sözlər: *neft-qaz sənayesi, Big Data, Hadoop, Apache Spark, Big Data Analitika, MapReduce, Big Data arxitekturası.*

Giriş

Hazırda neft-qaz sənayesi yüngül neft (həm sıxlığına, həm də çıxarılmasının asanlığına görə) erasından ağır neft mərhələsinə keçməkdədir [1]. Yer təkində neft ehtiyatları tükənmir, daha uzaq və çətin şəraitə malik ərazilərə və daha dərin qatlara gedir, onu çıxarmaq getdikcə çətinləşir. Bu keçid dövrünün əsas xarakterik cəhətlərindən biri istehsalat tsiklinin bütün zəncirində informasiya texnologiyalarının geniş tətbiq edilməsidir; bunu bəzən neft-qaz sənayesində elmi-texniki inqilab kimi də xarakterizə edirlər [2].

Bu tendensiyada aparıcı xətt "rəqəmsal mədən" (və ya "ağıllı mədən") texnologiyalarının tətbiq edilməsidir [3–5]. Rəqəmsal yatağın fəlsəfəsi "ölç-modelləşdir-qərar qəbul et-yerinə yetir-nəzarət et" prinsipidir. Rəqəmsal mədənlərin instrumental əsasını optik lifli sensorlar təşkil edir. Neft quyusunda quraşdırılmış bu sensorlar temperaturun, təzyiqin və digər parametrlərin paylanmış ölçülməsinə imkan verir. İnformasiya optik lifli sensorlardan rəqəmsal mədənlərin əsas nöqtəsinə – istismarın real vaxt rejimində idarə edilməsi mərkəzinə və ya monitorinq mərkəzinə ötürülür. Mədənlərdən geoloji informasiyanın real zaman rejimində toplanması və ötürülməsi üçün optik lifli sistemlərin yaradılması, istehsal və texnoloji verilənlərin monitorinq mərkəzlərində emalı, proseslərin və proses verilənlərinin real rejimdə 3D-vizuallaşdırılması, robototexnikanın tətbiqi əməliyyatların məsafədən idarə edilməsinə, distant xidmətlərə imkan yaradır [6].

Monitorinq mərkəzlərində məsələlərin həlli üçün standart yanaşma hesablama gücü yüksək və qiyməti baha hesablama resurslarından istifadə edilməsidir, lakin onlar böyük həcmli informasiyanın arzulanan sürətlə emalına tam zəmanət vermir. Alternativ variant Big Data emalı üçün Apache Hadoop proqram texnologiyaları steki əsasında "yüksək məhsuldar verilənlər analitikası" (*ing. High Performance Data Analytics*) sinfindən olan həllərdən istifadə edilməsidir [7]. Big Data texnologiyaları verilənlər axınının paylanmış operativ emalını aparmağa və monitorinq mərkəzlərində istifadə edilən real zaman analitik sistemləri qurmağa imkan verir.

"Big Data" terminin ilk dəfə 2008-ci ildə daxil edilməsindən [8] keçən müddət ərzində bir sıra vacib texnoloji dəyişikliklər baş vermişdir: strukturlaşdırılmamış verilənlərin saxlanması və emalı buludlara daşınıb, saxlama qurğularının həcmi artmaqla yanaşı, qiymətləri ucuzlaşıb, Hadoop ekosistemi formalaşıb. Bahalı olmayan biznes-analitika alətləri və prediktiv analitika sistemləri artıq sifarişçilərə əlyetərdir, bazarda verilənlər saxlancının yeni kateqoriyası – analitik saxlanclar meydana çıxıb və s. [9, 10].

Big Data texnologiyaları böyük həcmli müxtəlif formatlı verilənləri real zaman rejimində toplamağa və həmin sürətlə də emal edərək yeni biliklər əldə etməyə imkan verir. Hazırda bu texnologiyaları digər sahələrlə yanaşı, neft-qaz sənayesində də tətbiq etməyə başlayırlar [11, 12].

Big Data texnologiyalarının köməyi ilə verilənləri dəyərə necə çevirmək barədə ümumi

bəyanatlar olduqca çoxdur. Neft-qaz sənayesi də istisna deyil. Bir çox şirkətlərin marketinq bülletenlərində neft-qaz sənayesi verilənlərinin aqreqasiyası və intellektual analizi üçün Big Data-nın böyük imkanları vurğulanır. Lakin Big Data analitikasının neft-qaz sənayesində tətbiqi hələ də eksperimental səviyyədədir [13]. Yalnız bir neçə şirkət Big Data texnologiyasını tətbiq etməyə cəhdlər edir [14]. Məsələn, Chevron şirkəti seysmik verilənlərin emalı üçün Hadoop (*ing. IBM BigInsights*) istifadə edir. Shell isə seysmik sensor verilənləri üçün Amazon Virtual Private Cloud (*ing. Amazon VPC*) Hadoop-un istifadəsi üzrə pilot layihə həyata keçirmişdir. Big Data konsepsiyasının neft-qaz sənayesində tətbiqini isbatlayan daha bir neçə nümunə də göstərmək mümkündür.

Verilənlərin həcminin eksponensial artması ilə yanaşı, mövcud şirkətlərin əksəriyyətində bu verilənlərə müraciət intensivliyi o qədər də yüksək deyil. Buna görə çox tələb edilməyən verilənlərin onlara operativ giriş imkanı istisna edilmədən saxlanmasını iqtisadi baxımdan necə əsaslandırmaq məsələsi meydana çıxır [15]. Bu baxımdan hazırda neft-qaz sənayesi müəssisələrində verilənlərin mərkəzləşdirilmiş saxlanmasının və paylanmış emalının təşkili olduqca aktualdır.

Bu işin məqsədi neft-qaz sənayesi üçün konseptual Big Data arxitekturasının işlənməsidir. İşdə neft-qaz sənayesi sistemlərindən daxil olan verilənlərin dərin analitika və maşın təlimi metodları ilə analizini dəstəkləyən Big Data platforması üçün arxitektura təklif edilir. Bu platforma istənilən mənbədən daxil olan verilənləri səmərəli şəkildə qəbul etməli, saxlamalı və onları Big Data analitikası alətlərinə əlyetər etməlidir. İlkin məqsəd böyük həcmdə verilənləri paylanmış klaster sisteminə yükləməkdən ibarətdir. Saxlanan verilənlərin həcmi ilə sorğu müddəti arasında balans saxlamaq üçün hibrid yanaşmanın işlənməsi zəruridir. Həyat qabiliyyətli Big Data həllinin yaradılması üçün Hadoop ekosistemindən zəruri alətlərin seçilməsi məsələsinə də baxılır.

Neft-qaz sənayesində böyük həcmli verilənlər

Neft-qaz sənayesində bütün verilənlər vacib əhəmiyyət daşıyır və bütün fəaliyyət sahələrində həmişə böyük həcmdə verilənlər generasiya edilirdi, lakin bu gün ekstremal həcmdə verilənlər inanılmaz böyük sürətlə generasiya edilir. Bu, həm də son dövrlər sensor qurğularının yaradılması texnologiyalarında qazanılmış uğurlarla əlaqəlidir (seysmik kəşfiyyatda 4C (4-komponentli) sensorları, quyularda, neftin və qazın toplanması, hazırlanması və nəqli sistemlərində optik lifli sensorlar). Geofizika (4D seysmika – 4-cü ölçü zamandır, mədəndə müəyyən zaman intervalı ilə götürülmüş seysmik ölçmələr ardıcıllığıdır), geologiya və neft-qaz yataqlarının işlənməsi (4D monitorinq) böyük həcmli verilənlərin əsas mənbəyidir [16].

Neft-qaz sənayesində verilənləri iki sinfə ayırmaq olar: texnoloji proseslərin monitorinqi və idarə edilməsində əldə edilən verilənlər; şirkətin və əməliyyatların idarə edilməsi proseslərində emal edilən verilənlər. Bu siniflərdən olan verilənlərin bəzi xarakteristikalarına qısa nəzər salaq.

Hazırda neft və qaz yataqlarının axtarışında əsasən seysmik kəşfiyyat metodu tətbiq edilir. Bu metodun mahiyyəti yerin (və ya dənizin) səthində süni yolla, məsələn, partlayışla elastik dalğalar yaratmaq və sonra onları yer səthində xüsusi seysmik qəbuledicilərlə qeydiyyata almaqdan ibarətdir. Bir yatağın seysmik kəşfiyyatı zamanı əldə edilən verilənlərin həcmi onlarla terabayta çata bilər [16].

Geoloji-seysmik verilənlərin həcminin sürətlə artması səbəbləri aydındır. Bu bir çox cəhətdən iki- və üçölçülü seysmik analizdən dördölçülü seysmik analizə keçilməsi ilə bağlıdır, bu verilənlərin həcminin on dəfələrlə artmasını şərtləndirir. Bundan başqa, çətin şəraitdə olan yataqların işlənməsinə daha tez-tez cəhdlər edilir ki, bu daha təfsilatlı informasiya alınmasını təmin edən seysmik kəşfiyyat aparılmasını tələb edir [17].

Son illər quyuların qazılmasında nəzarət üçün karotaj, ölçmə və 4D seysmik monitorinq kimi innovativ texnologiyalar (*ing. LWD (logging while drilling)*, MWD (ing. *measurement while drilling*) və SWD (*ing. seismic while drilling*)) istifadə edilir. Şaquli və üfüqi quyuların qazılmasında, layların hidravlik açılmasında da passiv seysmika istifadə edilə bilər [17].

Optik lifli sensorların tətbiqi temperaturun, təzyiqin və digər parametrlərin quyuda hər 100 m, 10 m, 1 m və hətta 10 sm-dən bir ölçülməsinə imkan verir [18]. Boruda sensorların istifadəsi yolu ilə real vaxt rejimində quyunun işinə və onun vəziyyətinə nəzarəti təmin etmək mümkündür. Praktikada ən çox tələb edilən alət boru vizualizatoru ola bilər, onun tətbiqi ilə mədəndə neftçilərin çoxdankı arzusu – yatağın bütün həyat dövrü ərzində borunun hər bir nöqtəsində mexaniki və fiziki xarakteristikaların vəziyyətini gözlə görmək arzusu həyata keçə bilər. Tək bir quyuda belə sensorlardan alınan verilənlərin gündəlik həcmi bir neçə terabayta çatır.

Neft və qaz kəmərlərinə nəzarət etmək, müxtəlif növ defektlərin yerini və ölçülərini müəyyən etmək üçün bu kəmərlərdə çox böyük sayda maqnit axını sensorları (*ing. Magnetic Flux Leakage, MFL*) istifadə edilir. Sensorlar kəmərin çevrəsi boyunca bərabər məsafədə yerləşdirilir və bu sensorlar hər 3 mm-də MFL siqnalları ölçür. Bunun nəticəsində toplanan verilənlərin həcmi olduqca böyük olur [19].

Böyük həcmli informasiyanın periferiyadan mərkəzə ötürülməsi üçün peyklərdən istifadə edilə bilər (məsələn, Qazprom "Yamal" peyklər qrupundan istifadə edir). Lakin real vaxt rejimində işə keçdikdə, peyk rabitəsi kanallarının böyük həcmdə verilənlərin ötürülməsinə gücü yetməyə bilər. Bu halda çox böyük həcmdə verilənlərin ötürülməsi üçün optik lifli rabitə kanallarından istifadə edilir. Məsələn, 2010-cu ilin sonunda BP şirkəti Meksika körfəzində bütün dəniz platformalarını birləşdirən, uzunluğu 1200 km olan optik-lifli rabitə kanalının qurulmasını başa çatdırmışdır. Bu layihənin həyata keçirilməsi 80 milyon dollara başa gəlmişdir. Norveçdə də oxşar təcrübə var, lakin sahil xətti uzun olduğuna görə BP-dən fərqli olaraq qapalı sistem deyil, optik lifli rabitə kanalından dəniz platformalarına birbaşa çıxışlar edilib. Verilənlər isə sahildə emal edilir. Rusiyada da Stokman yatağı, Xəzər dənizində Lukoil dəniz platformaları üçün oxşar layihələr vardır [20].

Neft-qaz yataqlarının 3D geoloji və hidrodinamika modellərində də böyük həcmdə sintetik verilənlər generasiya edilir. Neft-qaz yataqlarının geoloji modelləşdirilməsi zamanı layın 3D modeli qurulur və onun əsasında laydakı karbohidrogen ehtiyatı qiymətləndirilir. Geoloji modelin əsasında qurulan 3D hidrodinamika modeli yatağın işlənməsi prosesində layın xassələrinin və ehtiyatların həcminin dəyişməsini, quyular üzrə neftin (qazın) çıxarılması tempini göstərir. Yataqların geoloji və hidrodinamiki modelləşdirilməsi üçün Landmark, Roxar, Schlumberger və TimeZYX kimi böyük şirkətlərin xüsusi proqram təminatından istifadə edilir. Quyuda quraşdırılmış sensorlardan alınan informasiya əsasında geoloji və hidrodinamiki modellər adaptasiya olunur. Neft-qaz yataqlarının geoloji və hidrodinamiki modelləri əsasında müxtəlif mətn, cədvəl, qrafiki hesabatlar alınır. Beləliklə, neft-qaz yataqlarının geoloji və hidrodinamiki modelləşdirilməsində terabaytlarla, hətta petabaytlarla seysmik, geofizika, mədən verilənlərinin emalı rutin məsələyə çevrilir [21].

Verilənlərin daha bir mənbəyi şirkətdə aparılmış geoloji və geofiziki tədqiqatlar üzrə toplanmış tarixi məlumatların rəqəmsallaşdırılmasıdır. Şirkətin bütün mədənləri və quyuları üzrə toplanmış məlumatlar (qazma və hasilat da daxil olmaqla), obyektlərin texniki pasportları sahə standartları üzrə rəqəmsallaşdırılır və mərkəzləşdirilmiş arxivə daxil edilir.

Neft-qaz şirkətlərində istehsalat proseslərinin kompleks avtomatlaşdırılması ilə yanaşı, idarəetmə proseslərinin və biznes-proseslərin informasiyalaşdırılması da geniş miqyasda həyata keçirilir (bir sıra şirkətlərdə SAP sisteminin tətbiqi üzrə layihələr). İdarəetmə prosesləri üzrə verilənlər – büdcələşdirmə, idarəetmə hesabatları, faydalı iş əmsalı sistemi, maliyyə hesabatlarının konsolidasiyası, kontraktların, xəzinənin, risklərin, insan resurslarının idarə edilməsi sistemlərində emal edilir. Əməliyyat prosesləri üzrə verilənlər – layihələrin, təmir işlərinin, emal əməliyyatlarının, tədarük və sifarişlərin idarə edilməsi, istehsalın planlaşdırılması, ətraf mühitin mühafizəsi, nəqlin idarə edilməsi və loqistik planlaşdırma sistemlərində toplanır.

Neft-qaz şirkətləri məlum səbəblərə görə mətbuatın, kommersiya strukturlarının, həm də hakimiyyət orqanlarının diqqətində olurlar və buna görə də kənar mənbələrdən də böyük həcmdə strukturlaşdırılmamış informasiyanın toplanması və analizi olduqca aktualdır. Bu mənbələrə

kütləvi informasiya vasitələri, veb-saytlar, sosial media, elektron poçt, müxtəlif hesabatlar, şəkillər və multimedia aid olunur. Qeyd etmək lazımdır ki, bu verilənlər strukturlaşdırılmamış və ya yarım-strukturlaşdırılmış şəkildədir. Bu səbəbdən onları ənənəvi verilənlər anbarında saxlamaq, müntəzəm müraciət və analiz etmək olduqca mürəkkəbdir.

Təşkilati strukturun mürəkkəbliyi, geniş əraziyə səpələnmiş mürəkkəb istehsalat prosesləri və fəaliyyət sahələrinin müxtəlifliyi neft-qaz şirkətlərindən müəssisənin effektiv idarə edilməsi üçün zəruri olan bütün verilənləri vahid informasiya fəzasında birləşdirən həllər tələb edir. Ənənəvi IT-infrastrukturlar, xüsusilə də müxtəlif fəaliyyət istiqamətləri və ya bölmələr üçün əlaqələndirilməmiş şəkildə fəaliyyət göstərən verilənləri saxlama sistemlərinin infrastrukturları getdikcə şaxələnir, onlara xidmət mürəkkəbləşir və xərcləri artır, onların idarə edilməsi çətinləşir. Bu sistemlərin məhdud imkanları isə müasir neft-qaz sənayesi müəssisələrindən tələb edilən iqtisadi effektivlik səviyyəsini təmin etməyə imkan vermir. Bu hazırda neft-qaz sənayesində Big Data texnologiyalarının istifadəsi məsələsinin aktuallığını şərtləndirən əsas səbəblərdən biridir.

Hadoop ekosistemi haqqında ümumi məlumat

Hadoop ekosistemi Big Data texnologiyalarının sinonimi hesab edilir. Başlanğıcda Hadoop verilənlərin klasterlərdə saxlanması və MapReduce metodu ilə paralel emalı üçün bir alət idi, hazırda isə Hadoop böyük həcmli verilənlərin emalı ilə bu və ya digər şəkildə əlaqəli olan texnologiyaların (təkcə MapReduce vasitəsilə deyil) böyük bir stekidir.

Hadoop nüvəsinə (*ing. core*) aşağıdakılar daxildir [22]:

- **Hadoop Distributed File System (HDFS)** – paylanmış fayl sistemidir, praktiki olaraq qeyri-məhdud həcmdə verilənləri saxlamağa imkan verir.
- **Hadoop YARN** (*ing. Yet Another Resource Negotiator* – «daha bir resurs vasitəçisi») – klasterin resurslarının və məsələlərin idarə edilməsi üçün platformadır.
- **Hadoop MapReduce** – paylanmış MapReduce-hesablamaların proqramlaşdırılması və yerinə yetirilməsi platformasıdır.
- **Hadoop Common** – Hadoop ekosistemində digər modulların istifadə etdikləri utilitlər və kitabxanalar toplusudur. Məsələn, HBase modulları HDFS-ə müraciət etmək üçün Hadoop Common-da saxlanan Java arxivlərindən (JAR fayllarından) istifadə edirlər.

Hadoop-la bilavasitə əlaqəli olan, lakin Hadoop nüvəsinə daxil olmayan çox sayda Apache layihələri mövcuddur:

- **Hive** – böyük həcmli verilənlər üzərində SQL sorğuları üçün proqram təminatı (SQL-sorğuları MapReduce-məsələləri ardıcıllığına çevirir);
- **Pig** – verilənlərin yüksək səviyyəli analizi üçün proqramlaşdırma dilidir. Bu dildə bir sətirlik proqram kodu MapReduce-məsələlər ardıcıllığına çevrilə bilər;
- **Hbase** – BigTable paradiqmasını reallaşdıran sütun verilənləri bazası;
- **Cassandra** – yüksək məhsuldarlıqlı paylanmış "key-value" verilənlər bazası;
- **ZooKeeper** – konfiqurasiyanın paylanmış saxlanması və konfiqurasiyaya dəyişikliklərin sinxronlaşdırılması üçün servis;
- **Mahout** – böyük həcmli verilənlər üzərində maşın təlimi proqram kitabxanası.

Hadoop haqqında danışdıqda, ilk növbədə onun fayl sistemi – HDFS nəzərdə tutulur. İlkin yanaşmada adi fayl sistemi fayl deskriptorları cədvəlindən və verilənlər sahəsindən ibarətdir. HDFS-də cədvəl əvəzinə xüsusi server – adlar serveri (*ing. NameNode*) istifadə edilir, verilənlər isə çox sayda DataNode-lar üzrə paylanır. Verilənlər bloklara bölünür (adətən 64 Mb və ya 128 Mb), hər bir fayl üçün server onun yolunu, blokların və blok replikatorlarının siyahısını yadda saxlayır. HDFS sistemi UNIX-in klassik ağacşəkilli direktoriyalar strukturuna malikdir, istifadəçilərin hüquqlar üçlüyü və hətta konsol komandaları da oxşardır.

HDFS-in əsas xüsusiyyəti olduqca etibarlı olmasıdır. Hadoop klasterinin klassik konfiqurasiyası bir ad serverindən, bir MapReduce masterindən (JobTracker adlanır) və hər birində verilənlər serveri

(*ing. DataNode*) və işçi (*ing. TaskTracker*) işləyən işçi kompüterlər çoxluğundan ibarətdir. MapReduce iki mərhələdən ibarətdir [23, 24]:

1. *Map* – hər bir verilənlər bloku üzərində paralel və (mümkün olduqca) lokal yerinə yetirilir. Böyük həcmdə verilənləri proqramın olduğu yerə daşımamaq üçün proqram verilənlərin olduğu serverə göndərilir və verilənləri emal edir;

2. Reduce – əsas qovşaq ilkin emal edilmiş verilənləri işçi qovşaqlardan toplayır, onları birləşdirir və məsələnin həllini formalaşdırır.

Hadoop MapReduce verilənlərin paket emalı üçün nəzərdə tutulub, məsələlər növbə ilə həll edilir, bu isə sistemin işini yavaşıdır. Buna görə çox zaman Hadoop verilənlərin emalı üçün deyil, onların saxlanılması üçün istifadə edilir.

Hadoop distributivləri

Hadoop layihəsi Apache Software Foundation təşkilatının yuxarı səviyyəli layihəsidir, buna görə əsas distributiv və bütün digər işləmələr üçün mərkəzi repozitari məhz Apache Hadoop hesab edilir. Lakin bu distributivin praktikada tətbiqi bir sıra çətinliklərlə müşayiət olunur: Hadoop-u klasterdə quraşdırmaq üçün kompüterləri əvvəlcədən sazlamaq, paketləri quraşdırmaq, bir çox konfiqurasiya fayllarına düzəlişlər etmək və digər əməliyyatlar tələb edilir. Bu işlər insan tərəfindən yerinə yetirilir və bu zaman köməkçi sənədlər çox vaxt natamam olur və ya heç olmur. Buna görə praktikada bu işləri avtomatlaşdıran distributivlər istifadə edilir, üç şirkətdən birinin: Cloudera, Hortonworks, MapR distributivlərinə daha çox müraciət edilir.

CDH (*ing. Cloudera Distribution including Apache Hadoop*) distributivi – Cloudera Manager-in idarəsi altında Hadoop ekosisteminin ən populyar alətlərini birləşdirir. Cloudera Manager klasterin qurulmasını, bütün komponentlərin quraşdırılmasını və onların sonrakı monitorinqini öz üzərinə götürür. CDH ilə yanaşı, şirkət özünün digər məhsullarını da, məsələn, Impalanı (verilənlər bazası) inkişaf etdirir.

HDP (*ing. Hortonworks Data Platform*) distributivi – fərqləndirici xüsusiyyəti ondan ibarətdir ki, öz məhsullarından daha çox Apache məhsullarının inkişaf etdirilməsinə səy göstərir (cədvəl 1). Məsələn, Cloudera Manager əvəzinə Apache Ambari, Impala əvəzinə Hive istifadə edilir. Nəticədə HDP daha açıqdır və klasterləri idarəetmə sistemi Ambari həm Linux, həm də Windows üçün həllər qurmağa və Azure HDInsight bulud mühitinə miqrasiya etməyə imkan verir. Buna görə, monitorinq mərkəzinin qurulması üçün baza distributivi kimi HDP seçilə bilər, o, həlləri maksimal sayda platformaya tirajlamağa imkan verir.

Cədvəl 1

Verilənlərin emal mərhələləri üzrə HDP 2.3 distributivinə daxil olan proqram təminatı

Emal mərhələsi	Proqram təminatı
Giriş verilənləri axınının qəbul edilməsi	Kafka məlumat brokeri
İndeksləmə, axtarış, qruplaşdırma, klasterləşdirmə, sadə statistik emal	Solr indeksatoru
İntensiv axın verilənlərinin onlayn emalı	Apache Spark Streaming, Storm
Böyük həcmli xətti verilənlərin paket emalı	MapReduce, Apache Spark, Apache Spark DataFrame
Böyük qraf verilənlərinin paket emalı	GraphX, Apache Giraph
Statistik emal, maşın təlimi, prediktiv analiz	Apache Spark ML

MapR distributivi – gəlirlərinin əsas mənbəyi konsaltinq və tərəfdaşlıq proqramları olan əvvəlki iki şirkətdən fərqli olaraq MapR öz məhsullarının bilavasitə satışı ilə məşğul olur. Üstün cəhətləri çox sayda optimallaşdırmalar və Amazon ilə tərəfdaşlıq proqramıdır. Mənfi

cəhətlərindən biri pulsuz versiyada (M3) funksionallığın məhdudlaşdırılmasıdır. Bundan başqa, Apache Drill-in əsas ideoloqu və əsas yaradıcısı da MapR şirkətidir.

Big Data analitikası

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

Big Data ekosisteminə 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, maşın təlimi, mətnlərin intellektual analizi əsasında Big Data analitik alətləri qurulur [25].

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 emalının və analizinin necə təşkil edilməsi mürəkkəb elmi-tədqiqat məsələlərindən biridir. Strukturlaşdırılmamış verilənlərin intellektual analizi elmi tədqiqatların nisbətən cavan sahəsidir, mətn verilənlərin intellektual analizi – Text Mining sahəsində daha çox tədqiqatlar aparılıb [26–29].

3. Big Data analizi − Big Data analizi üçün statistik analiz, verilənlərin intellektual analizi, maşın təlimi, imitasiya modelləri, optimallaşdırma üsulları, verilənlərin vizuallaşdırılması, verilənlərin aqreqasiyası və inteqrasiyası və s. üsulları istifadə edilir. Prediktiv analitika ayrıca istiqamət kimi fərqləndirilir [30].

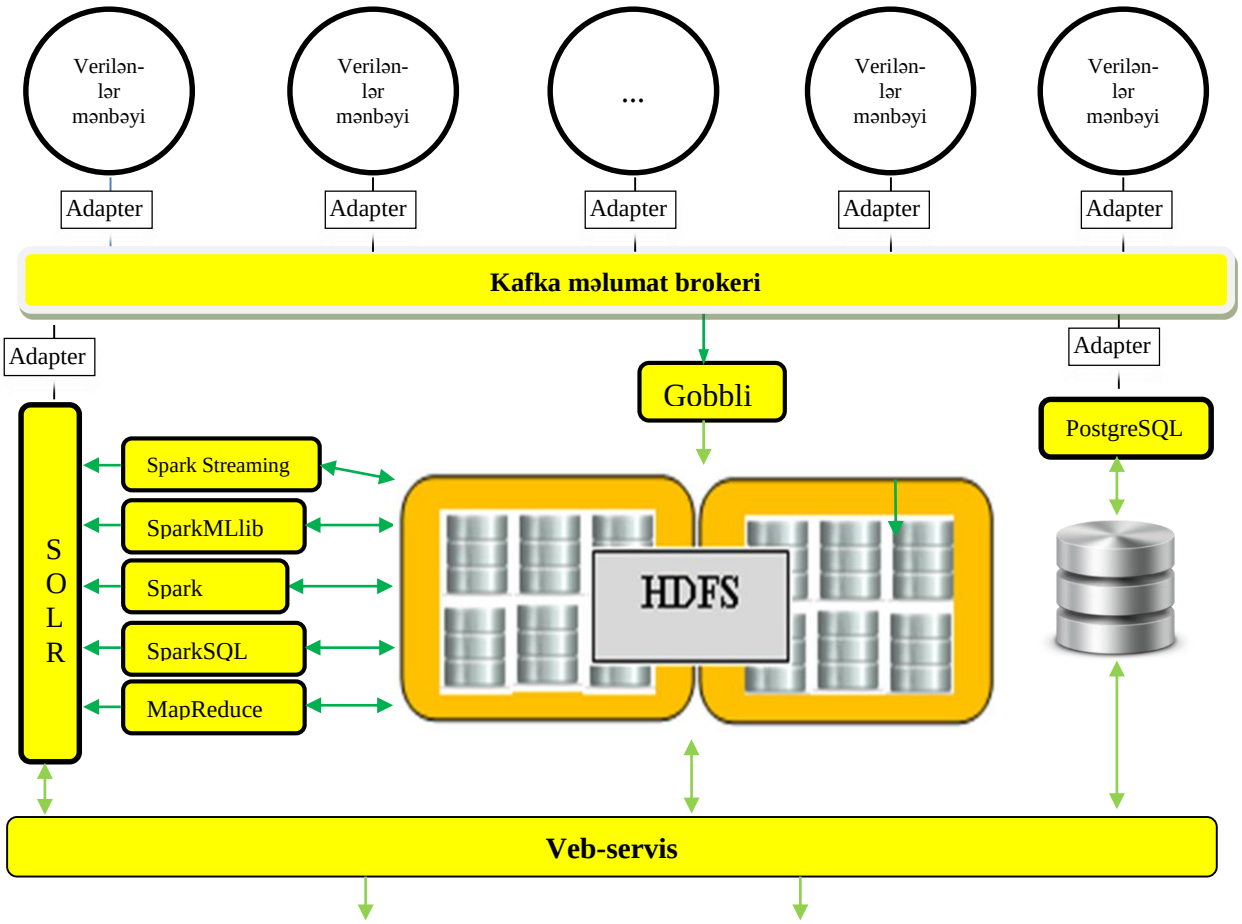
İdarəetmə qərarlarının operativ qəbul edilməsi üçün zəruri olan bütün verilənlərin asan başa düşülən şəkildə əks etdirilməsi üçün vizuallaşdırma alətləri: rəng identifikatorları, yarımşəffaflıq, verilənlərin laylarla yerləşdirilməsi, izoxətlərin, diaqramların qurulması və s. geniş istifadə edilməlidir [31].

Apache Hadoop texnologiyası çərçivəsində yaradılmış proqram təminatları bu sistemlərin əsas üstünlüklərindən birindən – üfüqi miqyaslanmadan istifadə etməklə verilənlərin analizin bütün mərhələlərində paylanmış emalını təmin etməyə imkan verir. Mövcud alqoritmlər hətta bir serverdən ibarət olan klasterdə də effektiv işləyə və emal edilən verilənlərin həcmi yüz dəfələrlə artdıqda serverləri təcili rejimdə qoşmaqla miqyaslana bilər .

Big Data texnologiyalarının köməyi ilə kvazi-strukturlaşdırılmış böyük həcmli verilənləri emal etmək, məhsuldarlığı və emal edilən verilənlərin həcmini mütənasib artırmaqla avtomatik üfüqi miqyaslanmanı yerinə yetirmək, birləşdirmə (*ing. join*) əməliyyatından minimal istifadə etməklə sürətli axtarışı həyata keçirmək, böyük intensivlikli hadisələr axınını operativ emal etmək və s. olar.

Neft-qaz sənayesi üçün konseptual Big Data arxitekturası

Neft-qaz sənayesi situasiya mərkəzinin analitik sistemi böyük həcmdə verilənləri emal etməli və həm daxil olan verilənlərin, həm də onların analizlərinin nəticələrini əks etdirmək üçün rahat interfeys təqdim etməlidir. Neft-qaz sənayesi üçün təklif edilən konseptual Big Data arxitekturası şəkil 1-də göstərilib. Konseptual arxitekturada seçilmiş Apache Hadoop, Apache Kafka, Apache Spark və Apache Spark Streaming texnologiyaları yüzlərlə terabaytdan yüzlərlə petabayta qədər həcmdə verilənləri emal etməyə imkan verir [32, 33]. Emal nəticələrinə sürətlə müraciət etmək üçün Solr indeksator-serveri istifadə edilir.



Şəkil 1. Neft-qaz sənayesi üçün konseptual Big Data arxitekturasının ümumi sxemi

Servis şini. Neft-qaz sənayesi sisteminə verilənlər müxtəlif informasiya sistemlərindən, rəqəmsal mədən sensorlarından daxil olur, burada onlar indeksləşdirilir, arxivləşdirilir, obyektin vəziyyəti qiymətləndirilir və s. Bütün verilənlər axınlarının koordinasiyasını təmin etmək və verilənlərin ötürülməsinə çəkilən xərcləri azaltmaq üçün servis şinləri istifadə edilir. Servis şini müxtəlif mənbələrdən məlumatların qəbul edilməsini, saxlanmasını və istehlakçılara paylanmasını mərkəzləşdirilmiş şəkildə yerinə yetirməyə imkan verir. Servis şininin əsas elementi kimi relyasion VBİS (verilənlər bazasını idarəetmə sistemi) istifadə edilə bilər, o, ötürülən verilənləri və hər bir mənbə və istehlakçı üzrə ötürmənin vəziyyəti barəsində xidməti məlumatların saxlanmasını təmin etməlidir. Klaster rejimində paylanmış emal zamanı böyük yüklənmələr və məlumatların zəmanətli çatdırılmasını təmin edən mexanizmlərin mürəkkəbliyi səbəbindən, servis şinlərində səhvlər (imtinalar), həmçinin verilənlərin sinxronlaşdırılması və ötürülməsi ilə bağlı problemlər baş verə bilər. Axın məlumatlarının yüksək məhsuldarlıqlı paylanmış emalı üçün xüsusi olaraq yaradılmış bir sıra həllər mövcuddur (ActiveMQ, RabbitMQ, Apache Kafka və s.). Apache Hadoop ekosisteminə daxil olan məlumat brokeri Kafka yuxarıda sadalanan nöqsanların əksəriyyətini aradan qaldırmağa imkan verir.

Məlumat sistemləri əsasən iki modelə əməl edirlər – növbə və nəşr/abunə. Kafka bu iki modeli ümumiləşdirən konsepsiyadan istifadə edir. Apache Kafka paylanmış nəşr/abunə növbəsidir. Müxtəlif kateqoriyalar üçün növbələr mövzular (Kafka terminologiyasında topic) adlanır. Kafka məlumat nəşr edənləri istehsalçı (*ing. producer*), məlumatı oxuyanları isə istehlakçı adlandırır. Məlumat növbələri müəyyən bufer təmin edirlər, Kafka mövzularına misal axınlardır. Brokerlər – Kafka klasterini təşkil edən serverlər istehsalçılar və istehlakçılar arasında verilənlərin nəqli kanalı kimi çıxış edirlər.

Mövzular bölmələrə (*ing. partition*) bölünür. Hər bir bölmə nizamlanmış məlumatlar ardıcıllığıdır. Məlumatlar bölməyə arasıkəsilmədən əlavə edilir. Bölmədə hər bir məlumata unikal sıra nömrəsi (ofset adlanır) verilir. Bölmədə istənilən məlumata nömrəsi ilə müraciət etmək olar.

Hər bölmədə bir server lider kimi çıxış edir. Lider bölmədə bütün oxuma və yazma sorğularını idarə edir. Liderin ardıcılları isə onu replikasiya edirlər. Bölmələrdən biri sıradan çıxdıqda, istehsalçılar və istehlakçılar hiss olunmadan digər serverə qoşulurlar.

Bölmələr iki məqsədə – miqyaslama və paralelliyə xidmət edirlər. Mövzu bir neçə bölməyə bölünə bilər, hər bölmə müxtəlif serverlərdə ola bilər (miqyaslama). İstehlakçı eyni anda müxtəlif bölmələrdən oxuya bilər (paralellik). Axın məlumatlarının ötürülməsi üzrə Kafka məhsuldarlığı bir serverdə saniyədə on minlərlə məlumata çata bilər. Xətti miqyaslama hesabına, məsələn, LinkedIn sistemində olduğu kimi on milyonlarla məlumat emal oluna bilər.

Gobblin – verilənləri Kafka-dan HDFS-ə yükləmək üçün proqram təminatıdır (LinkedIn tərəfindən işlənib). Əvvəlcə bu məqsədlə LinkedIn yenə özünün yaratdığı Camus-dan istifadə edirdi (gündə milyardlarla məlumat yüklənirdi). Gobblin müxtəlif mənbələrdən böyük həcmli verilənlərin Hadoop–da "həzmi" (*çıxarılması, çevrilməsi və yüklənməsi – ing. extracting, transforming, and loading, ETL*) üçün universal platformadır.

Verilənlərin indeksləşdirilməsi. Solr serveri axın verilənlərinin indeksləşdirilməsini praktiki olaraq real vaxt rejimində yerinə yetirir (indeksdə hadisələrin fiksasiyası bir neçə saniyədən bir aparılır) və müxtəlif nüvələrdə (cədvəllərdə) yazıların müxtəlif növləri, bulud rejimi istifadə edildikdə üfüqi miqyaslama hesabına indekslənən verilənlərin qeyri-məhdud həcmi, axtarış sorğularının zəngin dili təmin edilir [34]. Bundan başqa, sorğuların əksəriyyəti üzrə sürətli axtarış (bir neçə millisaniyə), fasetləmə, sorğulanan verilənlərin qruplaşdırılması və sadə statistik analizi, həmçinin saxlanan verilənlərin klasterizasiyası təmin edilir. Solr-da əsas xüsusiyyət onun üfüqi miqyaslanması, böyük həcmli (yüz milyonlarla yazı) verilənlərdə informasiyanın yüksək sürətlə axtarışı və emalıdır. Əsas verilənlər axını Solr girişinə adapterlərdən daxil olur, adapterlər indekslənən axınlar (mövzular) üzrə verilənləri Kafka brokerindən alırlar. Verilənlər real vaxtda və ya avtonom rejimdə analiz yerinə yetirən proseslərdən də daxil ola bilər. Verilənlərin bir hissəsi üzərində düzəliş aparmaq, keyfiyyətin yaxşılaşdırılması prosesində əlavə etmək və ya silmək olar.

Böyük həcmli verilənlərin analizi. Böyük həcmli verilənlərin analizi üçün Hadoop əsasında müxtəlif platformalar mövcuddur – Spark, Storm, GraphMap, H2O və s.

GraphMap – maşın təlimi alqoritmlərinin paralel yerinə yetirilməsinə yönəlib. Map mərhələsində bir-birindən asılı olmadan (müxtəlif qovşaqlarda) yerinə yetirilə bilən hesablamalar müəyyən edilir, Reduce mərhələsində isə nəticələr birləşdirilir.

Storm Twitter tərəfindən açıq kodlu layihə kimi 2011-ci ildə açıqlanıb. "Real zaman verilənlərinin emalı üçün Hadoop" kimi təqdim olunur, axın emalına və kəsilməz hesablamalara yönəlib (nəticələr alındıqca axın üsulu ilə ötürülür).

Təklif edilən konseptual arxitekturada böyük həcmli verilənlərin analizi üçün Apache Spark seçilmişdir. Bu seçimi aşağıdakı kimi əsaslandırmaq olar. MapReduce-un əsas üstünlükləri – miqyaslanma, istifadənin sadəliyi və imtinalara dayanıqlığıdır. Lakin Hadoop-da MapReduce realizasiyasının bir sıra nöqsanları da vardır. Əsas nöqsan iterativ alqoritmlərin (*ing. Machine Learning*) yerinə yetirilməsi zamanı məhsuldarlığın aşağı olmasıdır.

Standart MapReduce elə layihələndirilib ki, bütün nəticələr – həm son, həm də aralıq nəticələr diskə yazılır. Nəticədə diskə yazma və oxuma əməliyyatlarının müddəti hesablamaların öz müddətlərindən bir neçə dəfə böyük ola bilər. Spark bu problemi aradan qaldırır. Spark da verilənlərin lokallığı ideyasından istifadə edir, lakin hesablamaların əksəriyyətini disk əvəzinə bilavasitə operativ yaddaşa çıxarır. Bu, MapReduce ilə müqayisədə məhsuldarlığı əhəmiyyətli dərəcədə yüksəltməyə və iterativ alqoritmləri effektiv yerinə yetirməyə imkan verir.

Bunu nəzərə alaraq, neft-qaz sənayesi üçün çoxpilləli kompleks emal tələb edən mürəkkəb analitika məsələlərinin həlli üçün konseptual arxitekturada Apache Spark texnologiyasından istifadəyə üstünlük verilmişdir [36]. Spark-ın əsas üstünlüyü axın verilənləri ilə sürətli işləməsidir. Spark məsələlərin paylanmasını dəstəkləmir, bunun əvəzinə Mesos klaster menecerindən istifadə edir, o resursların izolyasiyasını və onların şəbəkə üzrə paylanmasını təmin edir.

Apache Spark-da əsas anlayış RDD (*ing. Resilient Distributed Dataset*) – imtinalara

dayanıqlı paylanmış verilənlər kolleksiyasıdır. RDD anlayışına əsaslanan paylanmış hesablamalar modeli Kaliforniya Universitetində (Berkli) işlənmişdir. RDD – qovşaqlar çoxluğu üzrə paylanmış dəyişməyən obyektlər kolleksiyasıdır [37]. RDD üzərində əksər əməliyyatlar hesablamasız ötüşür, hesablamalar yalnız tələb edildikdə yerinə yetirilir. Verilənlər toplusunun bir hissəsi itdikdə onları bərpa etmək olar, buna görə kolleksiya dayanıqlı adlanır.

RDD modeli əsasında Spark-dan əlavə, xüsusi paylanmış sistemlər – verilənlər saxlancı (*ing. Shark*), qraf verilənlərinin emalı sistemi (*ing. GraphX*), axın verilənlərinin emalı sistemi (*ing. Spark Streaming*), maşın təlimi alqoritmləri kitabxanası (*ing. MLlib*), emal edilən verilənlərə SQL-formatında müraciət etmək üçün funksiyalar toplusu (*ing. Spark DataFrames*) da işlənmişdir [37].

Spark-ın maraqlı xüsusiyyətlərindən biri odur ki, o, əməliyyatların yerinə yetirilməsi üçün Hadoop YARN-dan istifadə etmir, özünün məxsusi axın API-si və müxtəlif qısamüddətli zaman intervallarında müstəqil paket emalı prosesləri vardır. Spark-da verilənlərin paylanmış saxlama sisteminin olmaması onun nöqsanlarından biridir. Buna görə də Spark-ın işə salınması üçün HDFS-in verilənləri saxlama imkanlarından istifadə edilir. Faktiki olaraq, Spark Hadoop-u tamamlayır və Hadoop fayl sistemi ilə birgə işləyə bilər.

Apache Spark klasterin serverləri üzrə paylanmış verilənləri emal etməyə imkan verən əməliyyatların geniş sinfini dəstəkləyir, bu klasterin prosessor gücünü maksimal effektiv istifadə etməyə imkan verir. Bölmələrə bölgü düzgün aparılsa, aralıq verilənlərin şəbəkə ilə ötürülməsini minimuma salmaq olar, bu serverlərin sayını artırdıqda klasterin məhsuldarlığının praktiki olaraq xətti artmasını təmin edir.

Operativ xarakterli məsələlərin həlli üçün daxil olan verilənlərin Apache Spark Streaming modulu əsasında axın emalı istifadə edilir. Hər bir operativ məsələ üçün proses işə salınır və prosesdə Kafka məlumatlar brokerindən müvafiq axının oxunması və qoyulmuş məsələdən asılı olaraq onun analizi həyata keçirilir.

Hadoop-SQL aləti. Hadoop infrastrukturunda SQL-yönümlü bir neçə tətbiqi proqram var [38]: Hive, Spark SQL və s. Hadoop-un SQL ilə uyarlığı üçün Hive tətbiq edilməlidir [39]. Hive – Hadoop platformasında tarixən ilk və hazırda ən populyar verilənlər bazasını idarəetmə sistemidir. Sorğu dili kimi HiveQL istifadə edilir, SQL-in qısaldılmış dialekti olsa da, HDFS-də saxlanan verilənlər üzərində kifayət qədər mürəkkəb sorğular yerinə yetirməyə imkan verir. Hive son versiyalarda klassik MapReduce-dən Tez platformasına keçib [40], bu onu dəfələrlə sürətləndirərək interaktiv analitika üçün yararlı edib. Hive-də verilənlərin saxlanmasının optimallaşdırılmış sütun formatı ORC (*ing. Optimized Row Columnar*) istifadə edilir [41].

Zaman sıraları üçün verilənlər bazası. Neft-qaz sənayesi sistemlərində verilənlərin emalı iki axına bölünür – daxil olan verilənlərin real zamanda emalı və paket emalı. Daxil olan verilənlərin böyük əksəriyyəti məhz real zamanda emal olunur. Sensorlardan daxil olan verilənlər əksər hallarda zaman sıralarının xüsusiyyətlərini daşıyırlar. Zaman sıralarını relyasion verilənlər bazalarında reallaşdırmaq olar. Lakin verilənlərin daxilolma sürəti çox böyükdürsə, onda NoSQL həll tələb edilir. Zaman sıraları üçün NoSQL sistem konseptual səviyyədə vahid verilənlər modelindən imtina edir və verilənlər modelini məsələdən asılı olaraq seçməyi təklif edir. Hazırda zaman sıraları ilə işləmək üçün NoSQL sistemlərdə tez-tez istifadə edilən həllərdən biri OpenTSDB-dir [42].

OpenTSDB özlüyündə TSD (*ing. Time Series Daemon*) demonundan və komanda utilitləri toplusundan ibarətdir. Demonlar verilənləri saxlamaq üçün HBase bazasından istifadə edirlər, həmçinin verilənlərə giriş üçün açıq protokolları dəstəkləyirlər. Başqa sözlə, OpenTSDB – HBase-nin zaman sıralarını saxlamaq üçün istifadəçi sxemidir (arxitekturadır). Bu sxemə interfeys elementləri (*ing.TSD*) və HBase-də verilənlərin təsviri modeli daxildir. Demonlar bir-birindən asılı deyil, bu verilənlər axınlarının sayı artdıqda üfüqi miqyaslamanı təmin etmək üçündür. Zaman sıraları verilənlərinə giriş və vizuallaşdırma üçün istifadəçi interfeysi kimi Graphana istifadə edilir.

SQL operatorlarının zaman sıralarına tətbiqində müəyyən özəlliklər var, məsələn, oxuma əməliyyatı (*ing. SELECT*) emal metodları ilə birbaşa əlaqəlidir. Fasiləsiz daxil olan verilənlərin emalı həmişə müəyyən pəncərə ilə – verilənlərin müəyyən qısa zaman intervalına aid hissəsi ilə əlaqədardır.

Bunun nəticəsində zaman sıraları olan relyasion verilənlər bazasında mürəkkəb şərtlər olan oxuma sorğuları deyil, verilənlərin ardıcıl porsiyalarla oxunması sorğuları üstünlük təşkil edir. Zaman sıralarının emalının qeyd edilən bu və digər xüsusiyyətləri onların saxlanması və emalı üçün xüsusi sistemlərdən istifadə edilməsini tələb edir. Böyük həcmli zaman sıraları verilənlərinin emalı üçün Hadoop SQL alətləri istifadə etmək lazım gəlir. Lakin OpenTSDB-də istifadə edilən "wide table" və "blob" formatları bu verilənlərə SQL-əsasında müraciəti çətinləşdirir. Bu halda Spark SQL vasitəsilə zaman sıraları verilənləri ilə işləməyin bəzi üstünlükləri vardır.

Zaman sıraları ilə işləyən NoSQL sistemlər üçün vacib cəhət MQTT (*ing. Message Queueing Telemetry Transport*) protokolunun dəstəklənməsidir [43]. MQTT maşınlararası kommunikasiyaların (*ing. M2M*) və Əşyaların İnternetinin (*ing. IoT*) dəstəklənməsi üçün populyar kommunikasiya protokoludur. MQTT nəşr/abunə modelinin reallaşdırılması üçün nəqliyyat protokolu kimi istifadə edilir. Sensorlar (ölçmə cihazları) öz verilənlərini nəşr edirlər, verilənlər bazaları abunəçi kimi çıxış edir və nəşr olunmuş verilənləri saxlamaq üçün oxuyurlar. Zaman sıraları ilə əlaqədar MQTT şininin dəstəklənməsinə ehtiyac yarana bilər.

Nəticə

Bulud texnologiyaları və Big Data analitikası gələcəyin neft-qaz sənayesi infrastrukturunun əsasını təşkil edir və bu sahəni dünya standartları səviyyəsində təşkil etməyə kömək edir, infrastrukturu və münasibətləri daha səmərəli idarə etməyə imkan verir, ən qiymətli dəyər olan verilənlərdən tam şəkildə istifadə etməyə şərait yaradır. Bu işdə təklif edilmiş konseptual Big Data arxitekturası neft-qaz sənayesi üzrə müxtəlif miqyaslı praktiki layihələrdə nəzərə alına bilər.

Ədəbiyyat

1. Campbell C.J., and Laherrère J.H. The end of cheap oil // Scientific American, 1998, vol.278, no.3, pp.78–83.
2. Cross L.R. The technology revolution in oil and gas, 2014. www.worldservicesgroup.com/publications.asp?action=article&artid=6496
3. Saputelli L. A., Bravo C., Moricca G., Cramer R., Nikolaou M., Lopez C., Mochizuki S. Best practices and lessons learned after 10 years of Digital Oilfield (DOF) implementations // SPE Paper 167269, SPE Kuwait Oil and Gas Show and Conference, 2013, p.1. http://dx.doi.org/10.2118/167269-MS.
4. Dickens J., Feineman D., & Roberts S. Choices, changes and challenges: lessons for the future development of the Digital Oilfield // Society of Petroleum Engineers. 2012. http://dx.doi.org/10.2118/150173-MS.
5. Feineman D. R. Digital Oilfield Implementation: Learning From the Ghostbusters // Society of Petroleum Engineers, 2014. http://dx.doi.org/10.2118/167831-MS.
6. Holland D. Exploiting the Digital Oilfield: 15 Requirements for Business Value. Xlibris, 2012.
7. İmamverdiyev Y.N. Big Data texnologiyalarının böyük perspektivləri və problemləri // İnformasiya cəmiyyəti problemləri, 2016, №1, s.23–34.
8. Editorial: Community cleverness required // Nature, 4 September 2008, vol.455, no.7209, p.1.
9. Əliquliyev R.M., Hacırəhimova M.Ş. "Big Data" fenomeni: problemlər və imkanlar // İnformasiya texnologiyaları problemləri, 2014, №2, s.3–16.
10. Qasımova R.T., Big Data analitikası: mövcud yanaşmalar, problemlər və həllər // İnformasiya texnologiyaları problemləri, 2016, №1, s.75–93.
11. Hajirahimova M.S. Opportunities and challenges Big Data in oil and gas industry // Материалы Национального Суперкомпьютерного Форума (НСКФ-2015), Россия, Переславль-Залесский, 24-27 ноября, 2015.
12. Alıquliyev R.M., İmamverdiyev Y.N., Abdullayeva F.C. Neft-qaz sənayesi üçün Big Data analitikanın cloud computing platformasında analytics-as-a-service kimi reallaşdırılması imkanlarının tədqiqi // İnformasiya texnologiyaları problemləri, 2016, №1, s.11–26.
13. Feblowitz J. The Big Deal about Big Data in upstream oil and gas. IDC Energy Insights.

October 2012.
14. 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.
15. Sangvai P. Impact of Big Data in oil and gas industry // Proc. of the 10th Biennial International Conference & Exposition, 2013, pp.439–440.
16. Onajite E. Seismic Data Analysis Techniques in Hydrocarbon Exploration. Elsevier Inc., 2014.
17. Hyne N. Dictionary of Petroleum Exploration, Drilling & Production. 2nd Edition. 2014.
18. Zhang M., Ma X., Wang L., Lai Sh., Hongpu Zhou H., Zhao H., Liao Y. Progress of optical fiber sensors and its application in harsh environment // Photonic Sensors, 2011, vol.1, no.1, pp.84–89.
19. Shi Y., Zhang C., Li R., Cai M., Jia G. Theory and application of magnetic flux leakage pipeline detection // Sensors, 2015, vol.15, pp.31036–31055.
20. Bravo C.E., Saputelli L., Rivas F., Perez A. G., Nickolaou M., Zangl G., De Guzman N., Mohaghegh S., Nunez G. State of the art of artificial intelligence and predictive analytics in the E&P Industry: a technology survey // Society of Petroleum Engineers, 2013. http://dx.doi.org/10.2118/150314-PA.
21. Kamal S. Z., Williams J., Liddle J. Continuous improvement of assets through existing and new digital oilfield technology // Society of Petroleum Engineers, 2014. http://dx.doi.org/10.2118/167908-MS.
22. White T. Hadoop: the definitive guide. O'Reilly Media, Inc., 2012.
23. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters // Proc. of the 6th Conference on Symposium on Opearting Systems Design & Implementation, 2004, vol.6, pp.137–150.
24. 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.
25. 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.
26. Fan W., Bifet A. Mining Big Data: current status, and forecast to the future //ACM SIGKDD Explorations Newsletter, 2013, vol.14, no.2, pp.1–5.
27. Weiss Sh.M., Indurkhya N., Zhang T., Damerau F. Text mining: predictive methods for analyzing unstructured information. Springer; 2005, 260 p.
28. Aliguliyev R.M. A new sentence similarity measure and sentence based extractive technique for automatic text summarization // Expert Systems with Applications, 2009, vol.36, no.4, pp.7764–7772.
29. 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.
30. Siegel E. Predictive Analytics: The power to predict who will click, buy, lie, or die. Wiley; 1st edition, 2013, 320 p.
31. Mittelstadt S., Behrisch M., Weber S., Schreck T. et al. Visual analytics for the Big Data era - a comparative review of state-of-the-art commercial systems // Proc. of the IEEE Conference on Visual Analytics Science and Technology, 2012, pp.173–182.
32. Chardonnens T. et al. Big Data analytics on high velocity streams: a case study // IEEE International Conference on Big Data, 2013, pp.784–787.
33. Jones M.T. Spark, an alternative for fast data analytics. IBM developerWorks, November 2011.
34. Wang H-M., Wang H-W., Liu Y., Yang F. Design and implementation of SOLR-based information retrieval system for value-added service // The Journal of China Universities of Posts and Telecommunications, 2008, vol.15, pp.51–54.
35. Douglas K., Douglas S. PostgreSQL: a comprehensive guide to building, programming, and administering PostgreSQL databases, SAMS publishing, 2003.
36. Fazelat R. A Comprehensive analysis - data processing part Deux: Apache Spark vs Apache

Storm, January 2016. www.linkedin.com/pulse/comprehensive-analysis-data-processing-part-deux-apache-fazelat
37. Tian X., Lu G., Zhou X., Li J. Evolution from Shark to Spark SQL: preliminary analysis and qualitative evaluation. Big Data Benchmarks, Performance Optimization, and Emerging Hardware, 2015, pp.67–80.
38. Abadi D., Babu S., Özcan F., Pandis I. SQL-on-hadoop systems: tutorial // Proc. of the VLDB Endowment, 2015, vol.8, no.12, pp.2050–2051.
39. Thusoo A., Sarma J.S., Jain N., Shao Z., Chakka P., Anthony S., Liu H., Wyckoff P., Murthy R. Hive A Warehousing Solution Over a MapReduce Framework // Proc. of the VLDB Endowment, 2009, vol.2, no.2, pp.1626–1629.
40. Saha B., Shah H., Seth S., Vijayaraghavan G., Murthy A., Curino C. Apache Tez: a unifying framework for modeling and building data processing applications // Proc. of the ACM SIGMOD International Conference on Management of Data, 2015, pp.1357–1369.
41. Huai Y., Chauhan A., Gates A., Hagleitner G., Hanson E.N., O'Malley O., Zhang X. Major technical advancements in Apache Hive // Proc. of the ACM SIGMOD International Conference on Management of Data, 2014, pp.1235–1246.
42. Prasad S., Avinash S.B. Smart meter data analytics using OpenTSDB and Hadoop // Innovative Smart Grid Technologies-Asia, 2013, pp.1–6.
43. Hunkeler U., Truong H.L., Stanford-Clark A. MQTT-S – a publish/subscribe protocol for Wireless Sensor Networks // Proc. of the 3rd IEEE international Conference on Communication Systems Software and Middleware and Workshops, 2008, pp.791–798.

UOT 004.9

Алыгулиев Рамиз М.[1], Имамвердиев Ядигар Н.[2]
Институт Информационных Технологий НАНА, Баку, Азербайджан
[1]a.ramiz@science.az, [2]yadigar@lan.ab.az

Концептуальная архитектура Big Data для нефтегазовой промышленности

Технологии Больших Данных предоставляют важные подходы и инструменты для создания систем управления данными для нефтегазовой промышленности. В статье предлагается концептуальная архитектура для гибридной платформы Больших Данных, предназначенной для хранения и анализа данных большого объема методами глубокой аналитики и машинного обучения в распределенных кластерных системах, поступающих из систем нефтегазовой промышленности в режиме реального времени в различных форматах. Рассматривается также вопрос выбора необходимых инструментов из экосистемы Hadoop для создания жизнеспособного решения Больших Данных.

Ключевые слова: *нефтегазовая промышленность, Большие Данные, Hadoop, Apache Spark, MapReduce, анализ Больших Данных, архитектура Больших Данных*.

Ramiz M. Aliguliyev[1], Yadigar N. Imamverdiyev[2]
Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]a.ramiz@science.az, [2]yadigar@lan.ab.az

Conceptual Big Data architecture for the oil and gas industry

Big Data technologies provide important approaches and tools for the creation of data management systems for the oil and gas industry. The paper proposes a conceptual architecture for a hybrid Big Data platform for storing and analyzing large volumes of data gathered from oil and gas industry systems in real-time by deep analytics and machine learning methods in distributed cluster systems. We also consider the question of choice of necessary tools from Hadoop ecosystem for the building of a viable Big Data solution.

Keywords: *oil and gas industry, Big Data; Hadoop; Apache Spark; MapReduce; Big Data analytics; Big Data architecture.*