E-səhiyyə üçün Konseptual Big Data Arxitekturası

Yadigar İmamverdiyev
AMEA İnformasiya Texnologiyaları İnstitutu
yadigar@lan.ab.az

***Xülasə*— Big Data texnologiyaları e-səhiyyə sistemlərinin qurulması üçün vacib əhəmiyyət daşıyan yanaşmalar və alətlər təqdim edir. Bu işdə e-səhiyyə sistemlərinə real zaman rejimində daxil olan böyük həcmli və müxtəlif formatlı tibbi verilənlərin paylanmış klaster sistemlərində saxlanması və bu verilənlərin dərin analitika və maşın təlimi metodları ilə analizi üçün nəzərdə tutulmuş hibrid bulud-Big data platforması üçün konseptual arxitektura təklif edilir. Həyat qabiliyyətli Big Data həllinin yaradılması üçün Hadoop ekosistemindən zəruri alətlərin seçilməsi məsələsinə də baxılır.**

Açar sözlər*— *e-səhiyyə; Big Data; Hadoop; Apache Spark; Big Data Analytics; MapReduce; Big Data arxitekturası.

I. GİRİŞ

Big Data texnolgiyaları böyük həcmli müxtəlif formatlı verilənləri real zaman rejimində toplamağa və həmin sürətlə də emal edərək yeni biliklər əldə etməyə imkan verir [1]. Hazırda bu texnologiyalar digər sahələrlə yanaşı, səhiyyədə də geniş tətbiqlər tapmağa başlayır. Pasiyentlərin keyfiyyətli tibbi xidmət tələbatını ödəmək üçün ölkələr öz səhiyyə sistemlərində getdikcə daha tez-tez bu texnologiyaya müraciət edirlər. Təsadüfi deyil ki, Ümumdünya Səhiyyə Təşkilatı üzv-ölkələrin e-səhiyyə profilini qiymətləndirərkən Big Data-nı ayrıca komponent kimi nəzərə alır [2]. Big Data texnologiyalarından istifadə nəticəsində səhiyyə sferasının digər sahələrlə müqayisədə təkcə inkişaf etmiş ölkələrdə deyil, orta və aşağı gəlirli ölkələrdə də daha çox fayda əldə edəcəyi gözlənilir [3].

Səhiyyə sistemi informasiya texnologiyalarına hələ 1950-ci illərdən müraciət edir. Həmin vaxtdan keçən onilliklər ərzində informasiya texnologiyaları böyük inkişaf yolu qət etmişdir; bu dövr ərzində dünya tibb təcrübəsində də olduqca böyük həcmdə strukturlaşdırılmış və strukturlaşdırılmamış verilənlər toplanmışdır. Əvvəlki dövrlərdə səhiyyə sistemində əsas məqsəd aparıcı biznes-proseslərin avtomatlaşdırılması idi, hazırda isə toplanmış informasiyanın analizi və əvvəllər aşkarlanması mümkün olmayan qanunauyğunluqların aşkara çıxarılması əsasında səhiyyə sisteminin işinin effektivliyini yüksəltmək əsas məqsəddir.

Big data tibbi məlumatların aqreqasiyası və intellektual analizi üçün böyük imkanlar açır. Məsələn, Big data analitikası xəstəliklərin yayılmasını proqnozlaşdırmağa, müxtəlif müalicə üsullarının effektivliyini müqayisə etməyə, risk qruplarını müəyyən etməyə, müalicəni fərdiləşdirməyə, səhiyyə sistemində zəif yerləri aşkara çıxarmağa, səhiyyə müəssisəsinin işini xarkterizə edən yüzlərlə göstəricini uçota almağa, müalicənin effektivliyini qiymətləndirməyə, səhiyyə xərclərini proqnzlaşdırmağa imkan verir. Vaxtında tibbi diaqnostika nəticəsində insanların və dövlətin səhiyyə xərcləri azalır, müalicənin effektivliyi əhəmiyyətli dərəcədə yüksəlir.

İnformasiya axınının sıçrayışla artması şəraitində tibb mütəxəssislərinin sürətli və keyfiyyətli qərarlar qəbul etməsi getdikcə çətinləşir, cəmiyyət isə daha yüksək keyfiyyətli tibbi xidmət və yeni nəsil tibbi texnologiyaların tətbiqini tələb edir. Belə şəraitdə Big Data e-səhiyyədə islahatların əsas elementinə çevrilir [4, 5].

Hazırda səhiyyə müəssisələrində tibbi verilənlərin mərkəzləşdirilmiş və paylanmış saxlanmasının təşkili olduqca aktualdır. Bu işin məqsədi e-səhiyyə üçün konseptual Big data arxitekturasının işlənməsidir. İşdə e-səhiyyə sistemlərindən daxil olan verilənlər üzərində dərin analitikanı və maşın təlimini dəstəkləyən verilənlər platforması üçün arxitektura təklif edilir. Bu platforma istənilən mənbədən daxil olan verilənləri səmərəli şəkildə qəbul etməli, saxlamalı və onları Hadoop və ya SQL 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.

II. E-SƏHİYYƏDƏ İNQİLAB: BULUD VƏ BIG DATA

Səhiyyə sferasında verilənlər xəstəliklərin və hər bir pasiyent üçün seçilən müalicələrin sənədləşdirilməsində vacib rol oynayır. Hazırda istənilən tibb müəssisəsi böyük həcmli müxtəlif informasiya ilə qarşılaşır: pasiyentlərin xəstəlik tarixçələri, elektron tibb kartları, müayinələrin nəticələri və s. ABŞ Milli Səhiyyə İnstitutlarının (National Institutes of Health) məlumatına görə, tibbi məlumatların həcmi hər iki ildə dörd dəfə artır. Verilənlərin həcminin belə sürətlə artması şəraitində ortalama xəstəxananın hazırda hər il təxminən 665 terabayt tibbi verilənlər generasiya edəcəyi gözlənilir. Verilənlərin ümumi həcminin sürətlə artması ilə yanaşı, problem həm də ondadır ki, səhiyyə sahəsində verilənlərin 70 %-dən çoxu strukturlaşdırılmamış verilənlərdir, onların artım sürəti isə strukturlaşdırılmış verilənlərin artım sürətini 10 dəfə üstələyir.

Tibbi sənədlərin ənənəvi formalardan elektron formalara keçirilməsi prosesləri də əksər ölkələrdə aktiv icra mərhələsindədir, bu böyük həcmli verilənlərin bir mənbəyini təşkil edir. Eyni zamanda, teletibb sistemləri geniş yayılmağa başlayır [6], bu sistemlərdən alınan fizioloji verilənlərin mürəkkəb, intensiv axınları böyük həcmli verilənlərin digər bir mənbəyidir. Lakin müxtəlif klinik bölmələrdən və teletibb sistemlərindən daxil olan verilənlər çox zaman pərakəndə və müxtəlif formatlarda generasiya olunurlar, bu bütün informasiyanın vahid infrastrukturda konsolidasiyasını əhəmiyyətli dərəcədə çətinləşdirir.

Hazırda səhiyyə müəssisələri qarşısında bir sıra mürəkkəb məsələlər durur, onların həlli zamanın təxirəsalınmaz tələbidir:

tibbi xidmətin keyfiyyətin daim yüksəldilməsi; verilənlərin uzun müddət ərzində harada və nə vaxt yaradılmasından və saxlanmasından asılı olmadan elektron sağlamlıq kartlarında olan informasiyanın vahid sistemdə təqdim olunması; müxtəlif bölmələr və müəssisələr arasında tibbi verilənlərin effektiv mübadiləsinin təmin edilməsi. Ənənəvi IT-infrastrukturlar, xüsusilə də müxtəlif tibbi istiqamətlər və ya bölmələr üçün əlaqələndirilməmiş şəkildə fəaliyyət göstərən verilənləri saxlama sistemlərinin infrastrukturları getdikcə şaxələnirlər, onlara xidmət mürəkkəbləşir və xərcləri artır, onların idarə edilməsi çətinləşir, bu sistemlərin məhdud imkanları isə müasir cəmiyyətin tibb müəssisələrindən tələb etdiyi effektivlik səviyyəsini və xidmət keyfiyyətini təmin etməyə imkan vermir. Bu hazırda Big Data texnologiyalarının istifadə edilməsi məsələsinin aktuallığını şərtləndirən əsas səbəblərdən biridir.

Belə şəraitdə unifikasiya edilmiş elə həllin yaradılması tələb edilir ki, informasiyanı saxlamaqla yanaşı, müxtəlif mənbələrdən toplanmış verilənləri ilk növbədə effektiv şəkildə emal etməyə imkan versin. Yaradılmış həll geniş yayılmış protokollarla uyarlılığı təmin etməli, verilənlərin müxtəlif formatlardan (HL7, DICOM və s.) vahid formata çevrilməsini nəzərdə tutmalı və daxil olan bütün informasiyanı vahid elektron arxiv şəklində mərkəzləşdirilmiş saxlamağa imkan verməlidir.

Böyük həcmli verilənlərin effektiv saxlanmasına və idarə edilməsinə yeni yanaşma bulud texnologiyalarına əsaslanır [7] və səhiyyə xidməti göstərən təşkilatların federasiyada birləşməsini nəzərdə tutur, bu federasiyada hər bir təşkilat pasiyentlər haqqında bu və ya digər verilənləri təqdim edir. Bu zaman xidmət provayderlərinin müxtəlif təşkilatlardan daxil olan elektron sağlamlıq kartlarını sinxronlaşdırmasına ehtiyac qalmır – lazımi verilənlər bulud prinsipi ilə saxlanılır. Federasiyaya daxil olan istənilən təşkilatın əməkdaşı öz səlahiyyətləri çərçivəsində pasiyentin sağlamlıq vəziyyəti haqqında verilənlərə müraciət edə bilər.

"Big Data" terminin ilk dəfə 2008-ci ildə daxil edilməsindən sonra keçən müddətdə bir sıra vacib texnoloji dəyişikliklər baş vermişdir, strukturlaşmamış informasiyanın saxlanması və emalı buludlara daşınıb, saxlama qurğularının həcmi artmaqla yanaşı, qiymətləri ucuzlaşıb, Hadoop/MapReduce texnologiyalar steki genişlənib və populyarlığı artıb. Bahalı olmayan biznes-analitika alətləri və prediktiv analitika sistemləri sifarişçilərə əlyetərdir, bazarda verilənlər saxlancının yeni kateqoriyası – analitik saxlanclar meydana çıxıb və s.

Nəticədə, böyük həcmli verilənlərin idarə edilməsi və analitik emalı üçün Big Data texnologiyaları və alətləri praktiki olaraq istənilən təşkilata əlyetəndir. Tibb işçiləri də açıq kodlu proqram texnologiyaları əsasında yaradılmış hibrid Big Data-bulud həllindən istifadə edərək tibb kartları, təcili tibbi yardım stansiyalarının jurnalları və sosial media daxil olmaqla bir çox mənbədən toplanmış strukturlu və struktursuz verilənləri müstəqil analiz edə və qısa zamanda keyfiyyətli qərarlar qəbul edə bilərlər.

III. HADOOP EKOSİSTEMİ HAQQINDA ÜMUMİ 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 [8]:

- **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 və Hive 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:

- **Hbase** – BigTable paradiqmasını reallaşdıran sütun verilənləri bazası;
- **Cassandra** – yüksək məhsuldarlıqlı paylanmış "key-value" verilənlər bazası;
- **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;
- **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 (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 (DataNode) və işçi (TaskTracker) işləyən işçi kompüterlər çoxluğundan ibarətdir. MapReduce iki mərhələdən ibarətdir [9, 10]:

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.

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

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 müşaiyət 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: Cloudera, Hortonworks, MapR birinin distributivlərinə daha çox müraciət edilir.

IV. BIG DATA ANALITIKA VASİTƏLƏRİ

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 [11].

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 [12].

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 [11].

Apache Hadoop texnologiyası çərçivəsində yaradılmış proqram təminatı 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 (join) əməliyyatından minimal istifadə etməklə sürətli axtarışı həyata keçirmək, böyük intensivlikli hadisələr axınını operativ emal etmək və s. olar.

V. E-SƏHİYYƏ ÜÇÜN KONSEPTUAL BIG DATA ARXİTEKTURASI

E-səhiyyə situasiya mərkəzinin analitik sistemi böyük həcmdə verilənləri emal etməli və həm daxil olan verilənlərin, həm də onların analizinin nəticələrini əks etdirmək üçün rahat interfeys təqdim etməlidir. E-səhiyyə üçün təklif edilən konseptual Big Data arxitekturası şəkil 1-də göstərilib. Konseptual arxitekturada seçilmiş Apache Hadoop, Apache Spark və Apache Spark Streaming texnologiyaları yüzlərlə terabaytdan yüzlərlə petabayta qədər həcmdə verilənləri emal etməyə imkan verir. Emal nəticələrinə sürətlə müraciət etmək üçün Solr server-indeksatoru istifadə edilir.

Servis şini. E-səhiyyə sisteminə verilənlər müxtəlif tibbi informasiya sistemlərindən, teletib sensorlarından daxil olur, burada onlar indeksləşdirilir, arxivləşdirilir, pasiyentin 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.

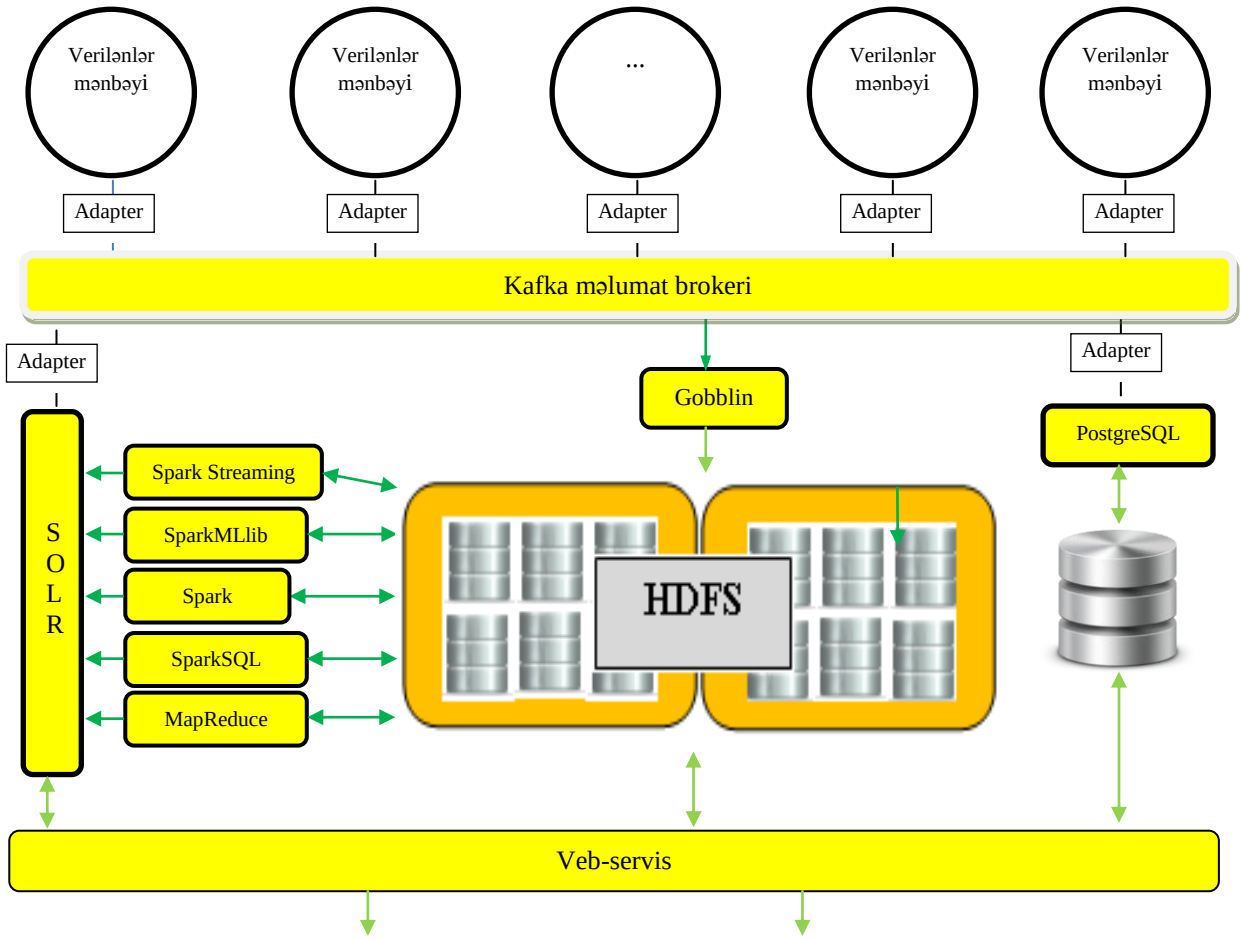
Apache Hadoop həllər stekinə daxil olan məlumat brokeri Apache Kafka axın məlumatlarının yüksək məhsuldarlıqlı paylanmış emalı üçün xüsusi olaraq yaradılıb və yuxarıda sadalanan nöqsanların əksəriyyətini aradan qaldırmağa imkan verir.

Gobblin – verilənləri Kafka-dan HDFS-ə yükləmək üçün proqram təminatıdır (LinkedIn tərəfindən işlənib). Əvvəlcə bu məqsədlə LinkedIn yenə özünün yartdığı 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. 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.

Verilənlərin analizi. E-səhiyyə üçün çoxpilləli kompleks emal tələb edən mürəkkəb analitika məsələlərinin həlli üçün konseptual arxitekturada Apache Spark texnologiyasından istifadəyə üstünlük verilmişdir, onun tətbiqi MapReduce ilə müqayisədə məhsuldarlığı əhəmiyyətli dərəcədə yüksəltməyə imkan verir, bu əməliyyatların bilavasitə operativ yaddaşda yerinə yetirilməsi hesabına baş verir.

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. Bu problemi Spark aradan qaldırır.

Spark da verilənlərin lokallığı ideyasından istifadə edir, lakin hesablamaların əksəriyyətini disk əvəzinə yaddaşa çıxarır. Spark-da əsas anlayış RDD (Resilient Distributed Dataset) – elastiki paylanmış verilənlər toplusudur. RDD üzərində əksər əməliyyatlar hesablamasız ötüşür, hesablamalar yalnız tələb edildikdə yerinə yetirilir. Apache Spark-a çoxluqlarla əməliyyatların baza toplusu, emal edilən verilənlərə SQL-formatında müraciət etmək üçün funksiyalar toplusu (ing. *Apache Spark DataFrames*), statistika funksiyaları və maşın təlimi funksiyaları toplusu (ing. *Apache Spark MLlib*), qrafların paylanmış emalı üçün funksiyalar toplusu (Apache Spark GraphX) daxildir. 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 atrmasını 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.



Şəkil 1. E-səhiyyə sistemi üçün konseptual Big Data arxitekturasının ümumi sxemi

Hadoop-SQL aləti. Hadoop infrastrukturunda SQL-yönümlü bir neçə tətbiqi proqram var: Hive, Spark SQL və s..

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, bu onu dəfələrlə sürətləndirərək interaktiv analitika üçün yararlı edib. Hive-də verilənlərin saxlanmasının optimallaşdırılmış sütun formatı ORC (ing. *Optimized Row Columnar*) istifadə edilir.

Zaman sıraları üçün verilənlər bazası. E-səhiyyə 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. Tibbi sensorlardan daxil olan verilənlər əksər hallarda zaman sıralarının xüsusiyyətlərini daşıyır. Zaman sıralarını relyasyon 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 [13].

OpenTSDB özlüyündə TSD (ing. *Time Series Daemon*) demonundan və komanda utilitləri toplusundan ibarətdir. Demonlar verilənləri saxlamaq üçün HBase bazas;ndan istifadə edirlər, həmçinin verilənlərə giriş üçün açıq protokolları dəstəkləyirlər.

Başqa sözlə, OpenTSDB – HBase-nin zaman sıralarını saxlamaq üçün istifadəsi 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.

NƏTİCƏ

Bulud texnologiyaları və Big Data analitikası gələcəyin səhiyyə infrastrukturunun əsasını təşkil edir və ölkədə səhiyyə işini dünya standartları səviyyəsində təşkil etməyə kömək edir, infrastrukturu və münasibətləri daha səmərəli idarə etməyə imkan verir, ən qiymətli dəyər olan insan sağlamlığı haqqında informasiyadan tam şəkildə istifadə etməyə şərait yaradır. Bu işdə təklif edilmiş konseptual Big Data arxitekturası elektron səhiyyə üzrə müxtəlif miqyaslı praktiki layihələrdə nəzərə alına bilər.

ƏDƏBİYYAT

[1] Y. N. Imamverdiyev “Big Data texnologiyalarının böyük perspektivləri və problemləri,” İnformasiya cəmiyyəti problemləri, №1, s.23–34, 2016.

[2] WHO Global Observatory for eHealth: Atlas of eHealth country profiles. World Health Organization 2016, 392 p.

[3] R. Wyber, S. Vaillancourt, W. Perry, P. Mannava, T. Folaranmi & L. A. Celi “Big data in global health: improving health in low- and middle-income countries,” Bulletin of the World Health Organization, vol. 93, no. 3, pp:203–208, 2015.

[4] M. Cottle, S. Kanwal, M. Kohn, T. Strome, N. Treister “Transforming health care through big data. Strategies for leveraging big data in the health care industry.” New York: Institute for Health Technology Transformation; 2013.

[5] P. Groves, B. Kayyali, D.Knott, S. van Kuiken, The ‘big data’ revolution in healthcare: Accelerating value and innovation. McKinsey & Company. 2013. 22 p.

[6] mHealth: New horizons for health through mobile technologies. Global Observatory for eHealth series - Volume 3. World Health Organization. 2011. 112 p.

[7] S. Basu, A. Karp, J. Li, J. Pruyne, J. Rolia, S. Singhal, J. Suermondt, R. Swaminathan, “Fusion: Managing healthcare records at cloud scale,” IEEE Computer Society, vol. 45, no. 11, pp.,42-49, 2012.

[8] T. White Hadoop: The definitive guide. O'Reilly Media, Inc., 2012.

[9] J. Dean, S. Ghemawat “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.

[10] K. H. Lee, Y. J. Lee, H. Choi, Y. D. Chung, B. Moon “Parallel data processing with MapReduce: a survey,” ACM SIGMOD Record, vol. 40, no. 4, pp. 11–20, 2012.

[11] K. Karthik, G. Kollias, V. Kumar, A. Grama, “Trends in Big Data analytics”, Journal of Parallel and Distributed Computing, vol. 74, no. 7, pp. 2561–2573, 2014.

[12] Sh. M. Weiss, N.Indurkhya, T.Zhang, F. Damerau Text Mining: Predictive methods for analyzing unstructured information. Springer; 2005, 260 p.

[13] S. Prasad, S. B. Avinash “Smart meter data analytics using OpenTSDB and Hadoop,” Innovative Smart Grid Technologies-Asia (ISGT Asia), 2013, pp. 1–6.