의료 데이터기반 빅데이터 분석 서비스 구현

김성수°, 정문영, 이태휘, 원종호
한국전자통신연구원
{**sungsoo, mchung, taewhi, jhwon**}**@etri.re.kr**

Implementation of Medical Data-Based Big Data Analytics Service

Sung-Soo Kim°, Moonyoung Chung, Taewhi Lee, Jongho Won
ETRI

요 약

의료 데이터는 병원이 구축하는 전자건강기록 (EHR) 데이터와 보편화되고 있는 개인용 건강정보 측정장치들을 통해 축적되는 데이터를 중심으로 지속적으로 폭증하고 있다. 이러한 의료 빅데이터 분석을 통해, 의료비 절감과 공공의료 서비스 개선에 필요한 연구활동들이 활발히 진행되고 있다. 이러한 대규모 의료 빅데이터에 대한 효과적인 분석 처리를 위해서, 하둡 분산파일 시스템 (HDFS)에 저장된 데이터에 대해 SQL 질의처리를 지원하며, 웹을 통해 인터랙티브하게 의료 빅데이터 분석을 수행할 수 있는 시스템이 요구된다. 본 논문에서는 DAG 기반 분산처리 엔진을 이용한 하둡 데이터에 대한 SQL 처리 엔진을 이용한 의료 데이터기반 빅데이터 분석 서비스를 제공할 수 있는 시스템을 구현하였다. 제안한 시스템은 빅데이터 분석을 위해 27만건의 실제 병원 건강검진 데이터를 활용하였다. 본 시스템은 에드혹 (ad-hoc) 질의처리를 통해 초기 데이터 분석에 유용한 기능을 제공할 뿐만 아니라, 주요 집계를 통한 효과적인 의료 관련 트렌드 분석을 제공할 수 있음을 보여주었다.

1 서론

개인과 병원이 획득하는 의료 및 헬스 데이터는 매우 많은 양 (*volume*)과 텍스트, 의학영상 등과 같은 다양성 (*variety*)을 보유하고 있으며, 그 발생 속도 (*velocity*) 또한 중환자실의 환자 모니터링 장치에서 발생하는 데이터는 초당 1,000 레코드가 넘을 정도로 빠르게 축적된다 [1]. 의료 데이터 수집 장치들로부터 생성되는 데이터도 큰 시장을 형성해 나가고 있다. 맥킨지는 2025년 사물 인터넷 (IoT; Internet of Things)의 경제적 파급력을 약 3조 달러로 보며, 산업별 비중은 헬스케어가 15%를 차지하여 제조 15%와 공동 1순위를 차지할 것으로 전망하고 있다 [2].

국내의 경우, 건강보험심사평가원은 년간 4,600만 명, 200억 건의 5년치 의료 빅데이터를 보유하고 있다. 국가적으로 공공의료 서비스 개선을 위해 빅데이터를 활용하여 분석하는 것은 아주 중요하다. 데이터 분석가 및 과학자들이 이러한 의료 빅데이터를 활용한 의미있는 분석을 수행하기 위해서는 빅데이터기반 질의처리 엔진과 대화형 (interactive) 분석 도구가 필수적이다.

본 논문의 의료 빅데이터 분석 서비스 측면에서 기여한 바를 요약하면 다음과 같다.

- *의료 데이터 트렌드 분석:* 의료 빅데이터의 각 항목별(진단질병별, 성별, 연령별, 지역별) 통계 정보를 대쉬보드로 제공함으로써 전체적인 의료데이터 경향을 직관적으로 분석할 수 있는 기능을 제공한다.
- *에드혹 (ad-hoc) 질의처리:* 데이터 과학자들이 초기 의료 빅데이터에 대한 데이터 탐색 시에 유용한 인터랙티브한 에드혹 (ad-hoc) 질의를 웹기반 분석 인터페이스를 통해 제공한다.

이와 같이, 본 논문은 의료 빅데이터를 하둡기반 분산파일시스템에 저장하고, SQL 질의를 통해 의료 빅데이터 통계 분석과 에드혹 질의 처리를 수행할 수 있는 웹 기반 분석 시스템을 소개한다.

2 관련 연구

본 절에서는 관련 연구를 크게 *의료 빅데이터 서비스* 분야, 하둡기반의 *대규모 병렬 처리* 분야, 하둡상의 질의처리를 위한 *SQL온하둡* 분야로 나누어 기술한다.

의료 빅데이터 서비스 의료 빅데이터 서비스는 데이터 수집 단계, 데이터 분석 단계, 분석 결과를 토대로 의사결정에 반영하는 데이터 활용 단계로 이루어진다. IBM은 슈퍼컴퓨터 왓슨 (Watson)의 빅데이터 분석 능력을 극대화하여 의료진의 데이터 활용도를 향상시켜주는 서비스를 제공하고 있다. 특히, 전자건강기록 (EHR; Electronic Health Record) 데이터와 사물인터넷 기술을 접목하여, 개선된 의료 서비스들 (환자 모니터링, 고령자들의 홈케어, 만성질환 치료 및 관리 등)을 소개하고 있다. Explorys 사 (미국)는 의료 빅데이터 분석 서비스를 제공을 목표로 클라우드 기반 플랫폼인 'Explorys DataGrid'를 개발하였다 [3]. 국내의 경우, 병원을 중심으로 각기 보유한 전자의료기록 (EMR)과 전자건강기록 (EHR) 데이터를 체계화한 통합 임상데이터웨어하우스 (CDW; Clinical Data Warehouse)를 구축하고 있다.

대규모 병렬 처리 (Massively Parallel Processing) 하둡 (Hadoop)은 저가형 클러스터에서 빅데이터에 대한 분산 저장과 분산 처리를 제공하는 기술이다.분산저장을 위해 하둡 분산파일시스템 (HDFS; Hadoop Distributed File System)을 이용하고 초기 분산처리는 맵리듀스 (MapReduce) 프레임워크를 이용해서 처리했으나,

현재는 테즈 (Tez), 스파크 (Spark) 등과 같은 DAG (Directed Acyclic Graph) 기반의 성능 향상된 계산 프레임워크 (computational framework)들을 사용한다. 본 논문에서도 DAG 기반 계산 프레임워크를 구현한 엔진을 사용하였다 [4].

SQL온하둡 (SQL-on-Hadoop) 하둡 상에서 분석을 위한 질의처리를 수행하는 SQL온하둡 시스템들이 인기를 얻고 있다. 이전에 데이터 분석가나 데이터 과학자는 하둡상의 데이터 분석을 위해 복잡한 맵리듀스 프로그램을 작성하는 데 어려움을 겪었다. 이러한 문제를 극복하기 위해, 초기 페이스북 (facebook)에서 SQL과 유사한 질의 (HiveQL)를 맵리듀스로 변환하여 수행할 수 있는 하이브 (Hive)를 소개하였다. 이와 같이, SQL온하둡이란 SQL 질의문을 통해 HDFS에 저장된 데이터를 분석할 수 시스템을 말한다 [5]. SQL온하둡 시스템으로 타조 (Tajo), 임팔라 (Impala), 드릴 (Drill), 프레스토 (Presto), SparkSQL 등이 있다.

본 논문에서는 의료 데이터 분석가들이 웹을 통해 의료 빅데이터를 분석할 수 있는 시스템을 소개한다. 제안하는 웹기반 의료 빅데이터 분석 시스템은 다음과 같은 2가지 컴포넌트를 포함하고 있다.

- *의료 빅데이터 분석용 웹 어플리케이션:* 주요한 통계정보를 가시화하여 보여주는 대시보드와 에드혹 질의 실행도구를 제공하는 웹 응용프로그램이다.
- *SQL온하둡 시스템:* 의료 빅데이터 분석용 웹 어플리케이션에서 요청받은 SQL 질의를 처리하는 시스템이다.

3 시스템 구조

본 논문에서 제안하는 시스템은 하둡기반 빅데이터 질의처리 엔진인 키위 (KIWI[1]) 클러스터를 기반으로 의료 빅데이터에 대한 웹 기반 분석 서비스를 제공하는 시스템이다.

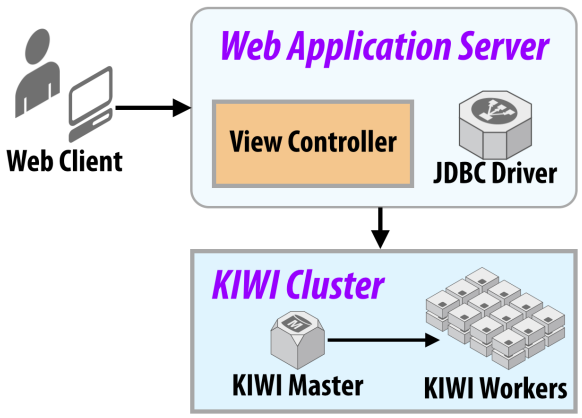

그림 1: 의료 빅데이터 분석 시스템 구조

1) 본 연구에서 개발한 SQL-on-Hadoop 엔진 일종으로 'Key Impact on data Warehouse Infrastructure'에 대한 약어임

그림 1은 제안하는 시스템 구조도를 보여 주고 있다. 여기서, 사용자는 웹 어플리케이션 서버 (WAS)를 통해 제공되는 웹 UI를 이용해서 빅데이터 분석을 수행한다. 사용자용 분석 웹 UI는 웹 어플리케이션 서버상의 웹 어플리케이션으로 의료 빅데이터를 접근하기 위해, 키위 클러스터와 JDBC 인터페이스를 통해 연결한다. 빅데이터 관리 및 처리를 담당하는 키위 클러스터 마스터 (Master)는 웹 어플리케이션 서버로부터 전달받은 요청질의를 처리한다. 키위 마스터는 요청 받은 질의에 대한 분산실행계획을 수립하고, 각 워커 (Worker) 노드에게 태스크를 할당하는 역할을 맡고 있다.

시스템 구조에서 사용자가 웹브라우저를 통해 통계 데이터 조회에 대한 요청 처리과정을 살펴보면 다음과 같다.

1. 사용자는 분석 웹 UI를 통해 통계 데이터 조회를 요청한다.
2. WAS는 통계 데이터 조회 요청을 JDBC를 통해 KIWI 클러스터의 마스터에게 전달한다.
3. 키위 마스터는 SQL 질의 처리를 수행하여 구성된 각 워커노드에게 태스크를 할당한다.
4. 키위 마스터는 SQL 질의 처리 결과를 WAS에게 전달하고, 분석 웹 어플리케이션은 결과를 가시화 한다.

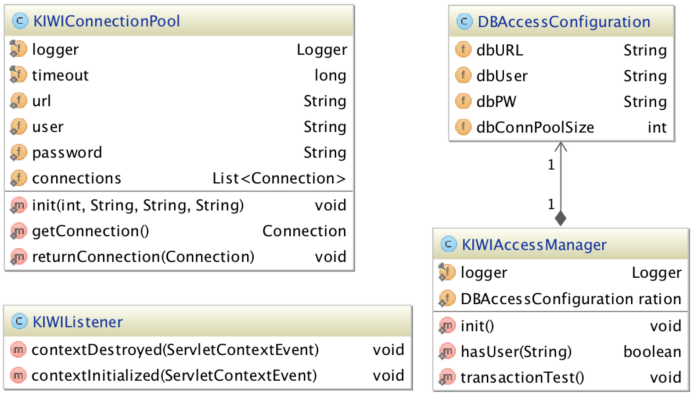

그림 2: 의료 데이터 접근 관련 클래스 다이어그램

4 시스템 구현

본 논문의 의료 빅데이터 분석용 웹 어플리케이션은 스프링 프레임워크, 키위 (KIWI) JDBC 라이브러리를 이용하여 자바 언어로 개발하였다. 의료 데이터를 저장, 관리하는 SQL온하둡 시스템은 아파치 타조 [4]를 코드베이스 (codebase)로 사용하고 추가 기능들 (질의 이력관리, SQL 인라인 맵리듀스 프로그램 수행 기능 등)을 확장하여 키위 시스템을 구현하였다. 그림 2는 SQL온하둡 시스템인 키위를 통해 데이터를 접근하기 위해 필요한 관련 클래스에 대한 다이어그램을 보여주고 있다. `KIWIConnectionPool` 클래스는 데이터 컨넥션 성능 향상을 위해, 일정의 컨넥션을 연결하고 있다가, 요청이 들어오면 컨넥션을 할당해주는 역할을 담당한다.

키위 데이터 접근을 관리하는 `KIWIAccessManager` 클래스는 데이터 접근을 위한 설정을 관리한다. `KIWIListenser` 클래스는 어플리케이션이 키위에서 비동기로 요청한 질의등에 대한 결과가 왔을 때 이벤트를 처리하기 위한 리스너 클래스다.

표 1은 시스템 구현 및 실험에 사용한 데이터 구성을 보여 주고 있다. 클러스터는 리눅스 KVM (Kernel-based Virtual Machine)로 생성된 16개의 가상머신 노드로 구성하였다. 실험은 우분투 서버 (14.04 LTS) 운영체제에서 JDK7, 하둡 2.5.1 버전을 설치하여 수행하였다. 실험에 사용한 의료데이터는 실제 병원에서 6개월간의 익명화된 건강검진 데이터를 사용하였다.

표 1: 클러스터 및 실험 데이터 구성

항목	구성
총노드 수	16 노드 (1 Master, 15 Workers)
노드당 구성 하드웨어	2 VCores, 6Gb RAM
의료 데이터 건수	약 27만건
데이터 수집기간	6개월
데이터 크기	3Gb

본 논문의 웹기반 의료 빅데이터 분석 서비스는 전체 의료데이터에 대한 집계연산을 통한 통계정보 서비스와 에드혹 질의처리 서비스로 구분할 수 있다. 그림 3와 같이 의료 데이터 웨어하우스에 대한 속성별 통계 정보를 가시화해주는 대시보드와 ANSI-SQL 기반 에드혹 질의를 편집하고 실행할 수 있는 도구를 포함하고 있다. 대시보드를 통해 보여주는 통계정보로는 가장 잦은 진단 종합통계, 성별에 따른 내원 진단통계, 지역별 내원환자통계, 연령별 내원현황, 지역별 내원환자통계가 있다. 또한, 키위 클러스터의 수행성능을 모니터링 하기 위해 질의처리 성능 측정 모듈을 구현하였고, 에드혹 질의 처리에 대한 처리 소요시간을 제공한다.

5 결론

본 논문은 의료 빅데이터 분석 서비스를 제공하기 위한 시스템을 제시했다. 제시한 시스템에서 하둡 상에서 SQL기반 질의를 처리하는 시스템 (SQL-on-Hadoop)인 DAG 기반 실행 엔진인 키위 엔진을 구현하였고, 키위 엔진을 이용한 웹 사용자 인터페이스 (Web UI)를 구현하였다.

제안한 의료 데이터기반 빅데이터 분석 서비스를 이용하면, 데이터 분석가와 데이터 과학자 같은 사용자는 주요 집계 연산을 통해 의료 빅데이터 트렌드를 직관적으로 분석할 수 있다. 또한, 빅데이터 초기 분석을 수행하기 위해 필요한 ANSI-SQL 에드혹 질의들도 키위 시스템을 통해 효과적으로 수행할 수 있다.

향후 연구주제로는 주요한 집계연산 (SUM, AVG, COUNT 등)의 결과로 근사결과와 오차정보를 함께 제공함으로써, 응답시간을 단축할 수 있는 근사 질의 처리 (approximate query processing) 기법 [6]에 관한 것이다.

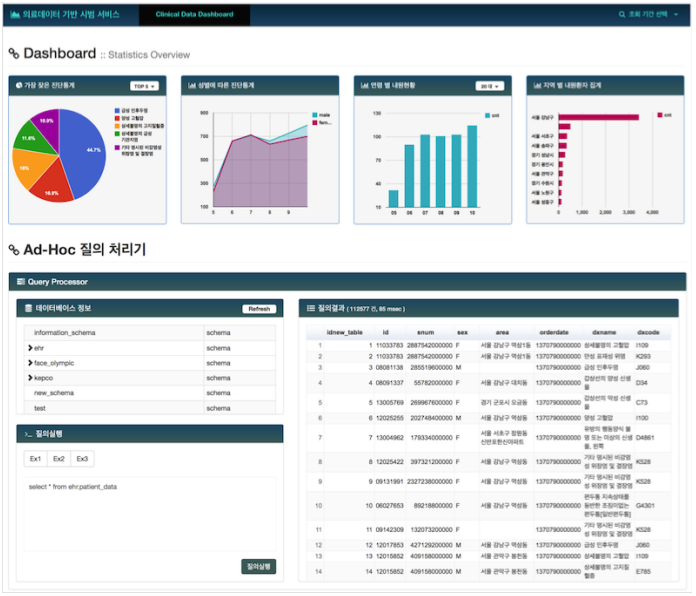

그림 3: 의료 빅데이터 분석 서비스 웹 UI

감사의 글

본 연구는 ETRI R&D 프로그램 (‘듀얼모드 배치처리 분석을 제공하는 빅데이터 플랫폼 핵심 기술 개발, 15ZS1400’)의 일환으로 수행하였습니다.

참고 문헌

[1] 정현학, “‘바이오헬스 빅데이터 플랫폼’ 만들자,” *보건산업동향*, pp. 10–15, 2015.

[2] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, “The ‘big data’ revolution in healthcare,” *McKinsey&Company*, 2013.

[3] 한정수, “클라우드 환경에서 의료 빅데이터 활용 및 전망,” *Journal of Digital Convergence*, pp. 341–347, 2014.

[4] H. Choi, Y. D. Chung, J. Son, H. Yang, B. Lim, S. Kim, and H. Ryu, “Tajo: A Distributed Data Warehouse System on Large Clusters,” in *Proceedings of the 2013 IEEE International Conference on Data Engineering (ICDE 2013)*, ICDE ’13, pp. 1320–1323, 2013.

[5] A. Floratou, U. F. Minhas, and F. Özcan, “SQL-on-Hadoop: Full Circle Back to Shared-nothing Database Architectures,” *Proc. VLDB Endow.*, vol. 7, pp. 1295–1306, Aug. 2014.

[6] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica, “BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data,” in *Proceedings of the 8th ACM European Conference on Computer Systems*, EuroSys ’13, pp. 29–42, 2013.