구글 클라우드 FHIR 객체의 Big Query 수행

김소연, 김민채, 진주은, 김나연, 이정훈
제주대학교 데이터사이언스학과
{carol7378, brianna0324, jooenujin365, nayeooon21}@naver.com, jhlee@jejunu.ac.kr

Big Query execution for FHIR objects on Google Cloud

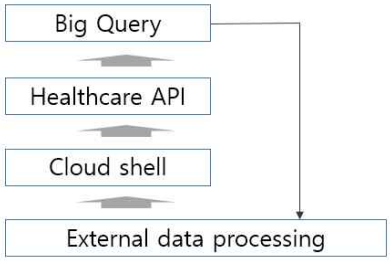
Soyeon Kim, Minchae Kim, Jooeun Jin, Nayeon Kim, Junghoon Lee
Dept. of Data Science, Jeju National University

요 약

본 논문에서는 구글 클라우드에 1차적으로 저장된 Healthcare API 서비스의 FHIR 객체들을 Big Query 서비스로 전환하고 질의를 작성하여 결과를 확인하는 과정을 설명한다. 이 과정에서 IAM을 위한 Big Query 테이블로의 입력 권한 부여 과정과 중첩된 필드들을 포함하고 있는 FHIR 객체의 명세과정이 핵심적인 단계가 되고 있으며 위 서비스들의 연계에 의해 대용량의 의료정보들이 구글 클라우드 상에 저장되고 사전분석되어 추가적인 정밀 분석을 위한 기저 자료를 제공할 수 있다.

1. 서론

구글 클라우드는 대용량의 데이터를 저장함은 물론 Big Query나 Machine Learning과 같은 기능을 제공한다[1]. 더욱이 HL7에 의해 제정 보완되고 있는 의료정보 표준인 FHIR (Fast Healthcare Interoperability Resources) 모델[2]을 위한 Healthcare API까지 확장하고 있는데 이들을 적절하게 활용하면 대용량의 정보를 저장하는데 그치지 않고 예비분석이나 고속의 기계학습 기능을 제공한다. <그림 1>에서 보는 바와 같이 예비 분석에 의해 산출된 결과는 다른 시스템에 출력되어 추가적인 정밀 분석도 가능하다. 의료정보는 그 내용이 복잡하여 모델링 하기 어려워서 관련 표준들이 널리 사용되지 못하고 있지만 최근에는 구글 클라우드에서도 지원할 정도가 되고 있다.



<그림 1> 구글 클라우드 서비스 연계

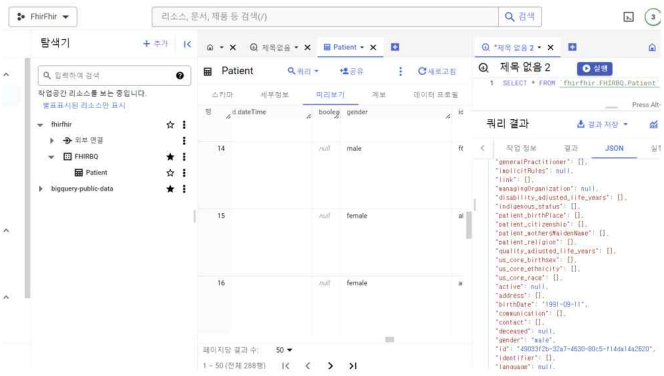
FHIR 자원들 중에서 다른 자원에 대한 참조를 갖고 있지 않은 Patient, Practitioner 객체들의 먼저 입력되어야 하는데 각 객체는 구글에서 부여한 고유한 id를 갖는다. 이후 병원 방문에 해당하는 Encounter 객체는 관련된 환자, 의사 등에 대한 참조를 속성으로 포함한다. 이때 존재하지 않는 참조 id를 포함한다면 입력이 차단된다. 이후 Observation은 X-Ray, 심전도, 혈액 검사 등 다양한 환자의 기록들을 저장하고 있는데 위의 자원들에 대한 참조를 정확하게 포함하여야 한다. 특히 원격 의료를 위한 개인용 측정기기들도 이 표준에 따라 저장하여야 한다.

2. FHIR 자원의 Big Query 전송

구글 클라우드에서 의료정보 분석 기능을 사용하려면 Big Query, Healthcare API, IAM (Identity and Access Management) 등 3 가지 서비스를 설정하여야 한다. 먼저 Big Query에서 데이터셋을 생성하여 FHIR 자원을 읽어올 수 있는 저장소를 준비하한다. 다음으로 IAM에서 이 데이터셋에 다른 서비스들이 접근할 수 있는 권한을 설정한다. 이 과정에서 IAM 관리자는 전체 프로젝트 내에서 기본적으로 Healthcare API를 사용자의 하나로 간주하고 있는데 이 API에는 기본적으로 서비스 에이전트의 역할이 주어져 있다. 이에 부가하여 Big Query 작업사용자와 데이터 편집자 역할을 부여하여야 한다. 다

음에는 Healthcare API로 가서 데이터셋을 선택하고 내보내기 기능을 수행하는데 그 대상을 BigQuery로 선택한다.

내보내기가 성공하면 Big Query 서비스에서 그 테이블이 보이는데 이 테이블은 HL7에서 정의한 리소스의 모든 속성을 스키마에 포함하고 있다. 그러나 <그림 2>에서 보는 바와 같이 업로드되는 필드들은 이들 중 일부만 포함하고 있으므로 대부분의 속성들이 Null 필드를 갖고 있다. 일반적으로 FHIR 자원을 Healthcare API에 올릴 때 강한 검사를 하고 특히 ID 부분에서 정해진 포맷을 준수하는 경우만 허용하기 때문에 필드 값에서 오류가 발생하지는 않으며 권한 문제에서 오류가 발생할 수도 있다. Big Query 테이블로 변환하는 과정에서 일부 속성은 중첩될 수 있어서 한 필드 내에 다른 필드들을 포함할 수 있다. 예를 들어 이름의 경우도 family name과 given name을 부속성으로 갖는다.



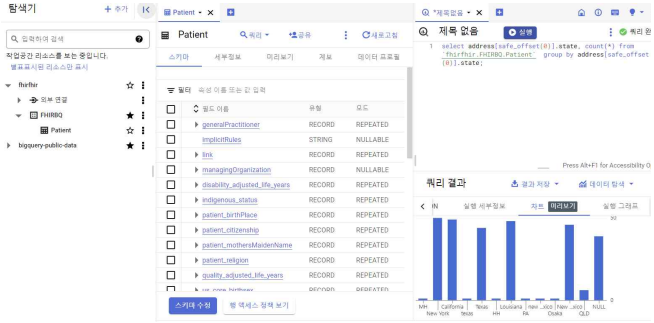
<그림 2> Big Query 데이터 셋 생성

Big Query의 질의에서 중첩된 필드에 대한 질의를 하는 경우에는 점으로 구분하여 계층적으로 필드를 명시하는데 해당 필드가 어떤 레코드에서는 null인 경우가 있다. 이런 경우 데이터 질의의 안전성을 유지하기 위하여 <표 1>에서 보는 바와 같이 safe_offset을 사용한다. 이 질의에서 fhirfhir는 프로젝트 이름, FHIRBQ는 Big Query에서 생성한 데이터 세트의 이름, Patient는 Healthcare API에서 입력된 테이블의 이름이다.

<표 1> Big Query 질의문의 예

```
select address[safe_offset(0)].state, count*)
from 'fhirfhir.FHIRBQ.Patient'
group by address[safe_offset(0)].state;
```

<그림 3>은 <표 1>의 질의를 수행한 결과를 보이고 있다. 현재 시험적으로 입력된 환자 정보 레코드들은 주소 필드에 state를 포함하고 있는데 이 서브필드를 키로 그루핑하고 그 개수를 구한 결과이다. 이는 먼저 state와 개수를 속성으로 갖는 테이블 형태의 출력을 생성하기도 하면 Json 형태의 파일로 변환할 수도 있다. 또 Big Query에서는 자동으로 막대그래프를 생성해준다. 이와 같이 다양한 형태의 데이터를 구글 코랩, CSV 파일 등으로 출력하여 추가적인 분석을 가능하게 한다.



<그림 3 > 질의문 수행 결과

3. 결론 및 추후 과제

구글 클라우드는 Computing Engine, Big Query, Machine Learning 등 다양한 기능을 제공하고 있으며 의료정보 표준인 FHIR 모델까지도 지원하고 있어서 막대한 양의 의료정보들이 클라우드에 저장되고 분석될 수 있는 환경을 제공하고 있다. 더욱이, 병원방문, 환자, 의사 등의 객체들의 상호 참조에 관련된 강한 무결성 검사를 수행하고 있어서 고도의 정확성이 요구되는 의료정보 처리에 더욱더 활용될 것으로 전망된다.

Acknowlegment

이 성과는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(IITP)(2021-0-00146)의 지원을 받아 수행된 연구임.

참고문헌

[1] R. Botez, et al., "Deploying a Dockerized Application With Kubernetes on Google Cloud Platform," COMM, pp. 471-476, 2020.

[2] https://www.hl7.org/fhir/