한국컴퓨터정보학회 동계학술대회 논문집 제23권 제1호 (2015. 1)

하둡기반 공간 빅데이터 저장 관리 시스템 구조

이강우*, 조은선°
*한국전자통신연구소 공간정보기술연구실
°충남대학교 컴퓨터공학과
e-mail:kwlee@etri.re.kr*, eschough@cnu.ac.kr°

An Architecture for a Spatial Big-Data Management System on Hadoop

Kang-Woo Lee*, Eun-Sun Cho°
*Spatial Information Technology Research Section, ETRI
°Dept. of Computer Sci. and Eng., Chungnam National University

● 요 약 ●

본 논문에서는 하둡 환경상에서 개발 중인 공간 빅데이터 저장 관리 시스템의 구조를 설명한다. 본 시스템은 공간 센서 및 IoT의 등장으로 대용량화된 공간 데이터로 인한 기존 공간 정보 처리 시스템의 성능적 한계를 극복하기 위한 목적으로 개발 중이다. 본 시스템은 효과적인 대용량 데이터 처리를 위해 현재 활발히 연구되고 있는 빅데이터 처리 기술과 공간 정보 처리 기술을 접목하여, 대용량의 공간 정보를 수집, 저장 관리하는 기능을 제공한다. 또한 효과적인 공간 데이터의 접근을 위해 스크립트 언어 기반의 공간 정보 처리 언어를 제공하고, SQL 형식의 선언적 공간 정보 질의 처리 기능도 제공하기 위해 개발 중에 있다.

키워드: 하둡(Hadoop), 공간 정보(geo-spatial information), 빅데이터(bigdata)

I. 서 론

기존의 공간 정보는 전통적으로 데이터베이스 관리 시스템(DBMS)를 통해 저장되고 관리되어 왔다. 공간 정보 활용 응용들은 수행에 필요한 정보를 DBMS를 통해 획득하여 수행된다. 공간 정보 응용의 활용 분야의 폭이 넓어짐에 따라 보다 다양한 형태의 데이터도 DBMS에 포함되어 관리되기 시작했다.

특히, 최근 사회관계망을 통한 데이터 생산과, IoT 분야의 등장으로, 보다 다양한 공간 정보를 포함하는 데이터가 급격히 증가하게 되었고, 사용자들은 이러한 데이터를 기존의 공간 정보 시스템을 활용하여 분석하여 유용한 정보를 도출하고자 하였다. 그러나 기존 공간 정보 시스템으로는 급격히 증가하는 공간 정보를 원활히 처리하지 못해 만족스러운 성능을 보이지 못하게 되었다.

이 문제를 해결하기 위해, 최근 활발히 연구가 진행되고 있는 빅데이터 처리 기술을 활용하고자 하였으나, 이런 기술들은 빅데이터에 포함된 공간 정보를 효과적으로 처리하지 못하여 만족할 만한 처리 성능을 보이지 못하였다.

본 논문에서는 최근에 각광을 받은 빅데이터 처리 기술을 활용하여 공간 정보의 저장 관리 및 접근 기능을 개발하여 대규모의 공간 빅데이터를 효율적으로 수집, 저장, 관리 할 수 있는 기능을 제공하기 위해 개발 중인 "공간 빅데이터 저장 관리 시스템"을 소개한다.

II. 관련 연구

Hadoop의 빅데이터 처리 기술과 기존의 공간 정보 처리 기술을 융합하여, 대용량 공간 빅데이터를 효과적으로 처리하고자 하는 연구는 일부 진행되어 왔다. [2]는 Hadoop 상에 공간 정보 처리 계층을 올린 SpatialHadoop 에 대해 기술한다. 특히 SpatialHadoop에서는 Hadoop의 데이터 처리 언어인 Pig를 공간 정보 처리 기능을 부가한 Pigeon을 제공하여 비교적 용이하게 공간 분석을 수행할 수 있도록 한다.

애모리 대학에서도 Hadoop 프레임워크에 공간 정보 처리 기능을 올린 시스템에 대한 연구를 진행해 왔다[1]. 이 대학에서는 공간 정보 인덱싱을 통해 공간 정보 접근 속도를 높이는 분야에 연구를 집중하고 있다. [4]는 Shared-Nothing 방식의 Hadoop 환경하에서의 대용량 공간 정보를 병렬적으로 처리하는 연구를 진행해 왔다. Hadoop 환경하에서의 공간 빅데이터 처리에 관한 연구 개발은 학계 뿐만 아니라 공간 정보 처리 업계에서도 진행되고 있다. ESRI사는 단순 HDFS뿐만 아니라, Hive를 UDF를 통해 확장하여 공간 정보를 질의할 수 있도록 한다.

III. 본 론

그림 1은 본 연구를 통해 개발 중인 공간 빅데이터 관리 시스템의 구조를 보여준다.

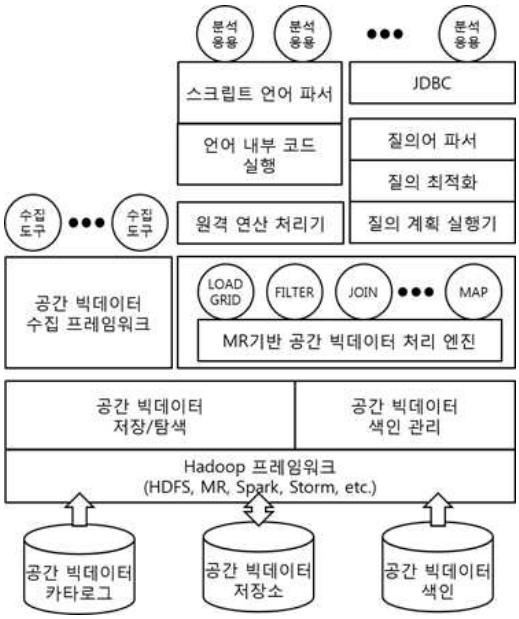

그림 1. 시스템 구조
Fig 1. System Architecture

Hadoop 프레임워크

본 시스템은 Hadoop 프레임워크를 사용한다. 본 시스템에 의해 관리되는 모든 공간 정보는 HDFS를 통해 저장되고 입출력된다.

공간 빅데이터 저장/탐색

HDFS에 저장된 공간 빅데이터의 저장 포맷과 이에 따른 데이터 입출력을 담당하는 모듈이다. 공간 빅데이터의 주요 데이터 형태인 레코드 형식과 그리드 형식의 데이터를 HDFS에 효과적으로 저장하기 위해 저장 형식과 데이터 분할 정책을 담당한다. 특히 그리드 형식의 데이터는 순차 읽기/쓰기 특성을 고려하여 열(row)단위의 분할 정책을 사용한다.

공간 빅데이터 색인 관리

공간 색인은 HDFS에 저장된 공간 정보 중 특정한 조건을 만족하는 부분만을 신속히 접근하기 위한 방법을 제공한다. 기본적으로 기존 다차원 색인 기법을 HDFS의 파일 분할 방식을 활용하여, 같은 조건/지역의 색인 정보를 같은 클러스터에 위치시키는 spatial-aware 인덱싱 기법을 사용한다.

공간 빅데이터 연산자

공간 관리 연산자는 시스템에서 제공하는 기본적인 공간 데이터 접근/처리 인터페이스를 정의한다. 일반적으로 하나 이상의 공간 정보(파일)을 입력으로 받아 이를 처리하여 다시 1개의 공간 정보(파일)을 생성하는 기능을 갖는다.

제공되는 각 연산자들의 자세한 구현 알고리즘은 사용자에게 알려질 필요는 없고, 연산자가 제공하는 입출력 정보만을 통해 사용할 수 있기 때문에, 구현 투명성을 제공한다. 뿐만 아니라, 서로 다른 방식으로 구현된 복수의 연산자가 동일 인터페이스를 갖을 수도 있기 때문에, 활용 상황에 따른 연산자 교체가 가능하여 성능 최적화의 기본 방식을 제공한다.

공간 빅데이터 처리 언어

공간 빅데이터 연산자는 Java API 기반의 인터페이스로 이를 사용하기 위해서는 Java 함수 호출을 통해 구동된다. 그러므로 Java 프로그래밍과 맵리듀스 기반의 프로그래밍에 익숙하지 않은 분석 응용 개발자의 경우 사용하기 어려운 한계를 갖는다.

이 문제를 해결하기 위한 방법으로 공간 정보를 접근하고 처리할 수 있는 스크립트 언어 기반의 공간 빅데이터 처리 언어를 제공한다.

응용 개발자가 본 언어를 통해 작성된 프로그램은 그림 3과 같이 스크립트 언어 파서를 통해 내부 중간 코드로 변환한다. 중간 코드은 여러 공간 연산자들로 구성된 실행 절차 계획으로 정의된다. 이렇게 정의된 계획은 비용 기반의 최적화 기법을 통해 보다 효과적으로 수행될 수 있는 계획으로 변환된다.

최적화를 거친 최종 중간 코드는 '중간 코드 실행기'를 통해 기술된 공간 연산자들을 계획된 순서대로 호출하는 방식으로 실행된다.

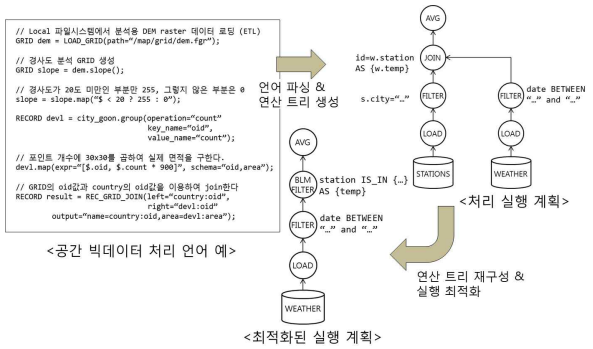

그림 2. 공간 빅데이터 언어 처리 과정
Fig 2. Spatial Big-Data Language Process

공간 빅데이터 질의 처리

본 시스템은 스크립트 기반의 공간 정보 접근 뿐만 아니라, 기존 DBMS와 같이 SQL에 기반한 공간 정보 접근 방식을 제공한다.

사용자에 의해 작성된 SQL은 스크립트 언어에서의 처리 방법과 유사하게 공간 연산자 실행 계획으로 변환되고, 비용 기반의 최적화를 거쳐 질의 계획 실행기에 의해 처리된다.

공간 빅데이터 수집 프레임워크

공간 빅데이터 수집 프레임워크는 다양한 형태의 공간 정보를 생성하는 데이터 소스와 연동하여 원시 데이터를 수집하여, 이로부터 필요한 공간 정보를 추출하여 본 시스템이 지원하는 데이터 형식으로 변환하는 적재하는 기능을 수행한다.

IV. 결 론

본 논문에서는 최근에 각광을 받은 빅데이터 처리 기술을 활용하여 공간 정보의 저장 관리 및 접근 기능을 개발하여 대규모의 공간 빅데이터를 효율적으로 수집, 저장, 관리 할 수 있는 기능을 제공하기 위해 개발 중인 "공간 빅데이터 저장 관리 시스템"의 구조를 설명하였다.

본 시스템은 다양한 형태의 대용량 공간 정보를 수집하여 저장 모듈을 통해 적재하고, 공간 연산자를 통해 접근할 수 있는 기능을 제공한다. 또한 편리한 공간 데이터의 접근을 위해 스크립트 언어 기반의 공간 정보 처리 언어와, SQL 기반의 공간 정보 질의 처리 기능도 제공하기 위해 개발 중에 있다.

감사의 글

본 연구는 '국토교통부 국토공간정보연구사업 국토공간정보의 빅데이터 관리, 분석 및 서비스 플랫폼 기술개발(14NSIP-B091011-01) 과제'의 연구비 지원에 의해 연구되었음.

참고문헌

[1] Ablimit Aji and et.al., "Hadoop GIS: a high performance spatial data warehousing system over mapreduce," Proceedings VLDB Endowment. Aug 2013; 6(11).

[2] Ahmed Eldawy,, "SpatialHadoop: Towards Flexible and Scalable Spatial Processing using MapReduce," In the SIGMOD/PODS Ph.D. Symposium, June, 2014.

[3] Jason Long, "GIS Tools for Hadoop: Big Data Spatial Analytics for the Hadoop Framework", Esri blog, http://esri.github.io/gis-tools-for-hadoop/

[4] Abhishek Sagar, "Large Spatial Data Computation on Shared-Nothing Spatial DBMS Cluster via MapReduce," Msc Thesis, June 2012.