ICT기반 환경 센서 데이터 분석을 위한 데이터베이스 적합성 비교 연구

문주현*, 박수용*, 우성주**, 신용태**
*숭실대학교 컴퓨터학과
**숭실대학교 컴퓨터학부
e-mail:hoopster@soongsil.ac.kr

A Study on the Database Conformance for the Analysis of ICT-Based Environmental Sensors

Ju-Hyeon Moon*, Soo-Yong Park, Seongju Woo, Yong-Tae Shin
*Dept of Computer Science, Soongsil University

요 약

환경 센서는 센서의 특징과 같은 변수에 따라 센서에서 발생하는 데이터가 일정하기 못하고, 광범위에서 실시간으로 발생하기 때문에 환경 센서 데이터 수집에 사용하는 데이터베이스 선정에 어려움이 있다. 본 논문에서는 각 데이터베이스의 특징을 실시간성과 확장성, 비용으로 비교하였다. ICT기반 환경 센서 데이터 수집에 적합한 데이터베이스는 MongoDB, OpenTSDB, MachBase DBMS이다.

1. 서론

최근 ICT(Information, Communication, Technology) 분야의 발전과 함께 스마트 센서 기술과 IoT(Internet of Things) 기술을 활용하여 넓은 범위의 환경 데이터 수집을 위한 환경 구축의 난이도가 낮아졌으며, 빅데이터 분석 기술과 인공지능 기술을 활용한 환경 데이터 분석이 활발히 연구 되고 있다. 그러나 환경 센서는 센서의 제조사, 센서의 특징 등 여러 변수에 따라 센서에서 발생 시키는 데이터가 일정하지 못하고, 광범위에서 실시간으로 데이터가 발생하기 때문에 환경 센서 데이터의 수집에 사용하는 데이터베이스 선정에 어려움이 있다. 따라서 본 논문에서는 광범위한 ICT기반 환경 센서 데이터의 수집에 알맞은 데이터베이스를 선정하고자 각기 다른 장점을 가진 데이터베이스의 특징과, ICT기반 환경 센서 구축 및 데이터 수집에 필요한 실시간성, 확장성, 비용을 평가하여 ICT기반 환경 센서 데이터 수집 및 분석에 적합한 데이터베이스를 제시한다.

본 논문의 구성은 다음과 같다. 2장 관련 연구에서는 RDBMS, MongoDB와 같은 각기 다른 데이터베이스를 살펴본다. 3장 비교 분석에서는 2장에서 살펴본 데이터베이스를 비교분석한다. 4장에서는 결론과 향후 연구방향을 제시한다.

2. 관련 연구

2.1 MariaDB

MariaDB는 MySQL과 같은 오픈소스 관계형 데이터베이스로 현재 사용되는 데이터베이스의 대부분은 관계형 데이터베이스를 기반으로 한다. 테이블과 스키마를 가지고 있으며, 트랜잭션을 지원하여 데이터의 무결성이 보장된다. 각 테이블 마다 고정적인 스키마를 가지고 있기 때문에 데이터를 입력할 시 테이블 형식에 맞는 입력 값이 필요하다.

2.2 MongoDB

MongoDB는 문서 지향적 NoSQL 데이터베이스로 RDBMS의 record와 유사한 개념의 document라는 JSON Object 형태의 key, value 쌍으로 이루어진 데이터 구조로 구성된다. MongoDB는 고정된 스키마를 가지고 있지 않아, document 각자의 고유한 스키마를 가지는 동적 스키마의 특성을 가지기 때문에 MongoDB는 많은 양의 데이터를 빠르게 저장할 수 있다.[1] MongoDB는 Sharding을 지원하여 여러 개의 데이터베이스로 데이터를 분할하는 클러스터 구축이 가능하며, Replication을 지원하여 데이터베이스의 장애가 발생 하였을 때 빠른 대응이 가능하다.

2.3 HDFS

HDFS(Hadoop Distributed File System)은 대용량 데이터를 분산 처리할 수 있는 오픈소스 빅데이터 분석 프레임워크인 Hadoop의 분산 파일 시스템이다. HDFS는 대용량의 파일을 일정 크기(128MB)의 블록으로 나누어 분산된 서버에 저장하여 저장된 대용량의 데이터를 빠르게 처리할 수 있다.[2] HDFS는 저사양의 서버 여러 대를 이용하여 클러스터 구축이 가능하고, Mapreduce 및 Hadoop Eco System에 다양한 프레임 워크를 사용할 수 있어 빅데이터 분석에 적합하다. 그러나 HDFS는 배치성으로 데이터를 저장하고 처리하기 때문에 실시간으로 발생하는 데이터 처리에 적합하지 않다.

2.4 Redis

Redis는 메모리 기반의 key, value 쌍으로 구성된 데

이터를 저장하는 고성능 데이터베이스이다. Redis의 모든 작업은 메모리에서 시행되기 때문에 디스크를 사용하는 다른 데이터베이스보다 더 빠른 읽기와 쓰기 작업이 가능하다. Redis는 디스크에 비해 메모리의 용량이 제한되기 때문에 많은 양의 데이터를 처리할 시 고용량의 메모리를 구축하기 위해 많은 비용이 필요하여 소량의 데이터를 고성능으로 처리할 때 합리적인 데이터베이스이다.[3] 그러나 Redis는 메모리기반 데이터베이스인 동시에 In-memory Cache로도 사용 가능하여 메모리 캐시를 기존의 사용 중인 데이터베이스에 적용하면 데이터베이스의 입출력 성능이 향상되기 때문에 대규모 데이터를 실시간으로 처리하기 위해 사용된다. Redis는 데이터베이스의 클러스터 구축과 MongoDB의 Replication과 같은 기능을 제공하여 실시간으로 데이터를 다른 서버에 복제하기 때문에 데이터 손실에 대한 위험이 매우 적다.

2.5 OpenTSDB(HBase)

OpenTSDB는 HBase를 기반으로 한 오픈소스 분산형 시계열 데이터베이스이다. OpenTSDB는 실시간으로 발생하는 시계열 데이터 수집에 특화되어 있으며, 많은 디바이스에서 짧은 시간에 발생하는 데이터의 수집이 가능하다. OpenTSDB에 저장되는 데이터는 HBase의 두 개의 테이블에 저장되며 tsdb 테이블은 시계열 데이터의 저장과 쿼리 기능을 제공하고, tsdb-uid 테이블은 모든 데이터에 대한 고유한 인덱스 값을 저장한다.[4]

OpenTSDB의 데이터 저장을 담당하는 HBase는 컬럼 기반의 데이터베이스이며 Hmaster 서버와 Hregion 서버로 구성되기 때문에 노드만 추가하면 데이터베이스 확장이 가능한 선형적인 확장 구조를 제공한다. HBase는 HDFS와 같이 Hadoop Eco System의 프로젝트 중 하나로 Mapreduce 및 Hadoop Eco System에 다양한 프레임워크를 사용할 수 있어 빅데이터 분석에 적합하다.

2.6 InfluxDB

InfluxDB는 GO lang으로 작성된 오픈소스 시계열 데이터베이스이다. InfluxDB는 고정된 스키마가 없는 동적 스키마 구조를 가지고 있으며, SQL-like language를 지원하여 다른 NoSQL 데이터베이스보다 사용하기 간편한 것이 장점이다.[5]

InfluxDB는 같은 오픈소스 시계열 데이터베이스인 OpenTSDB와 비교 했을 때 쓰기 성능에서 압도적으로 빠른 속도를 보여 데이터의 입력이 많은 센서 네트워크 환경에서 적합하다. InfluxDB는 노드만 추가하여 간편한 확장이 가능하게 설계되었지만, 확장 기능은 별도의 비용을 지불하여 사용하는 Enterprise 버전에서만 사용할 수 있어 다른 오픈소스 데이터베이스과 다르게 비용이 발생하는 것이 단점이다.

2.7 Machbase DBMS

Machbase DBMS는 로그성 데이터를 실시간으로 저장, 분석이 가능한 시계열 데이터베이스이다. Machbase DBMS는 한 번 입력된 데이터의 업데이트가 불가능하도록 설계되어있기 때문에 이미 입력된 데이터의 위변조가 불가능하다. Machbase DBMS는 초당 30만건에서 200만건까지 데이터 입력이 가능하여 매 초 발생하는 시계열 데이터의 입력에 좋은 성능을 기대할 수 있다.[6] Machbase DBMS는 실시간으로 입력되는 데이터의 크기를 압축하여 일정 크기의 데이터 블록을 생성하여 저장함으로써 데이터의 양을 줄임으로써 데이터의 입출력 크기가 줄어 낮은 비용으로 높은 성능을 기대할 수 있다.

3. 비교 분석

ICT기반 환경 센서 데이터는 넓은 범위에서 실시간 동시다발적으로 발생하기 때문에 환경 센서 데이터를 저장하는 데이터베이스는 실시간성을 띄어야하고 많은 양의 데이터가 수집되기 때문에 데이터베이스의 확장성 또한 용이해야한다. 2장에서 조사한 데이터베이스 중 HDFS는 Hadoop Eco System를 사용할 수 있기 때문에 빅데이터 분석에 좋은 성능을 보일 수 있지만, 배치형식의 데이터 저장방식으로 실시간으로 데이터를 저장해야하는 환경 센서 데이터 저장에는 적합하지 않다. 2장에서 조사한 데이터베이스들은 모두 Clustering, Sharding과 같은 확장 기술을 제공하고 있어 확장성에 관한 조건은 조사한 데이터베이스 전부 적합한 것으로 판단된다. 그러나 InfluxDB는 오픈소스로 비용이 들지 않지만, Cluster를 구축하기 위한 기술을 사용하기 위해선 일정의 비용을 지불하여

[표 1] 데이터베이스 비교

	MariaDB	MongoDB	HDFS	Redis	OpenTSDB	InfluxDB	Machbase DBMS
종류	RDBMS	NoSQL	NoSQL	NoSQL	NoSQL	NoSQL	NoSQL
저장 방식	record	document	file	memory, disk	column	memory, disk	log table
비용	free	free	free	free	free	free	free
실시간	O	O	X	O	O	O	O
확장	Clustering	Sharding	Clustering	Clustering	Clustering	Clustering (not free)	Clustering

Enterprise 버전을 구입해야하는 비용적인 문제가 있다. Redis는 데이터베이스로도 사용되지만 용량대비 비용이 disk를 사용하는 데이터베이스보다 memory가 월등히 높기 때문에 memory cache로써 다른 데이터베이스와 함께 사용하는 것이 비용측면에서 더 우수하다. 따라서 Redis와 InfluxDB는 ICT 환경 센서 데이터 수집에 적합하지 못한 데이터베이스이다. MariaDB는 실시간성, 확장성, 비용 세 가지 모두 충족하는 데이터베이스이지만 스키마를 동적으로 구성하지 못하기 때문에 특정 센서 마다 고유한 테이블을 구성해야하는 단점으로 데이터에 유연하지 못한 수집 프로세스를 구축해야한다. 따라서 MariaDB 또한 넓은 범위에서 실시간으로 발생하는 ICT기반 환경 센서 데이터 수집에는 적합하지 못한 데이터베이스이다.

4. 결론

본 논문에서는 ICT기반 환경 센서 데이터 분석을 위한 데이터베이스 적합성에 대한 비교 연구를 진행하였다. ICT기반 환경 센서 데이터의 수집을 위한 데이터베이스 비교는 각 데이터베이스의 특징과 데이터를 실시간으로 처리하는 실시간성, 많은 양의 데이터를 수집할 수 있는 확장성 그리고 비용 3가지 기준으로 진행하였다. MongoDB는 document기반의 빠른 데이터 처리와 Sharding을 활용한 확장성의 장점이 있었고, OpenTSDB와 Machbase DBMS는 센서 데이터와 같은 시계열 데이터에 특화된 시계열 데이터베이스로 빠른 쓰기 속도와 Clustering을 지원하여 많은 양의 데이터를 수집할 수 있는 장점이 있어 ICT기반 환경 센서 데이터 분석을 위한 데이터베이스는 MongoDB, OpenTSDB, Machbase DBMS이다. 향후 3개의 데이터베이스를 구축하고 실제 환경 데이터수집에 대한 성능평가가 필요하다.

Acknowledgement

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.IITP-2019-0-00135, ICT 기반 환경 모니터링 센서 신뢰성 검증 및 평가 플랫폼)

참고문헌

[1] Y. Kang, I. Park, J. Rhee and Y. Lee, "MongoDB-Based Repository Design for IoT-Generated RFID/Sensor Big Data," in IEEE Sensors Journal, vol. 16, no. 2, pp. 485-497, Jan.15, 2016.

[2] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, 2010, pp. 1-10.

[3] Jing Han, Haihong E, Guan Le and Jian Du, "Survey on NoSQL database," 2011 6th International Conference on Pervasive Computing and Applications, Port Elizabeth, 2011, pp. 363-366.

[4] B. Agrawal, A. Chakravorty, C. Rong and T. W. Wlodarczyk, "R2Time: A Framework to Analyse Open TSDB Time-Series Data in HBase," 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, Singapore, 2014, pp. 970-975.

[5] Influxdata, https://www.influxdata.com

[6] Machbase, https://www.machbase.com/machbase#overview