자율주행 데이터의 효과적인 처리를 위한 분산 데이터베이스 설계

백민석, 정현석, 공은빈, 오상윤*
아주대학교

white0825@ajou.ac.kr, ingu627@ajou.ac.kr, eb0904@ajou.ac.kr, *syoh@ajou.ac.kr

Design and Implementation of Distributed Database for Autonomous Driving

Minseok Baek, Hyunseok Jung, Eunbin Gong, Sangyoon Oh*
Ajou Univ.

요 약

자율주행 기술 고도화를 위해서는 관련 데이터의 효과적인 관리를 지원하는 시스템이 반드시 필요하다. 본 논문에서는, 비정형 대용량의 자율주행 데이터를 처리하기 위한 HDFS 와 HBase 기반의 분산 데이터베이스의 설계를 소개하며, 공개 자율주행 데이터의 ETL 과정을 통해 실증적인 효과를 분석한다.

Ⅰ. 서 론

높은 주행 안정성 보장 증 자율주행 기술 고도화를 위해서 주행 시 얻어진 센싱 데이터, 주행 알고리즘 검증을 위한 가상의 시나리오 데이터, 실 도로 및 주변 환경 데이터 등 다양한 관련 데이터를 확보하는 것은 중요한 문제이며 이를 효과적으로 관리할 수 있는 시스템이 반드시 필요하다. 특별히 최근에 소개된 A2d2[1] 공개 데이터 셋과 같이 카메라 등의 센서 기술 발전에 따라 자율주행 데이터의 용량 또한 커지는 추세이며, 이에 따라 대용량 데이터, 시계열 데이터 등 기존의 단일 노드 기반의 관계형 데이터베이스로는 효과적인 관리가 어려운 데이터들이 많아지고 있어 이에 대한 새로운 분산 기반의 데이터베이스의 설계와 구현이 필요하다.

이와 같이 자율주행에서 빅데이터 관리는 필수적이지만 현재까지 국제적으로 협의된 데이터 표준은 없는 상태이다. 따라서 본 논문에서는 비정형 데이터 보관 및 처리에 적합하여 목표하는 자율주행 데이터 처리에 적합하다고 판단되는 NoSQL 기반 분산 데이터베이스 적재 시스템 설계를 제시하고, 이에 대한 활용 방안 또한 제시하도록 한다.

Ⅱ. 자율 주행 데이터 관리를 위한 분산 데이터 베이스

(1) 분산데이터 베이스 설계

자율주행 데이터를 다루는데 있어, 기존 관계형 데이터베이스보다 분산을 통해서 확장하는 것이 용이한 NoSQL 기반의 데이터베이스를 선택하는 것이 적합하다. 세부 이유로는 자율 주행 관련 데이터들은 다양한 형태의 비정형 데이터가 다수 이며, 특별히 시계열 데이터를 자주 다루기 때문이며, 일반적으로 분산 파일 시스템의 사용은 단일 대용량 파일 시스템을 사용하는 것보다 효율적이기 때문이다.

본 논문에서는 MapReduce[2] 패러다임을 사용하여 크기가 큰 데이터 셋을 분석하고 변환할 수 있는 Hadoop Distributed File System(이하 HDFS)과 HDFS 위에서 작동할 수 있는 NoSQL 분산 데이터베이스인 HBase[3]를 선택하여 설계를 진행하였다. HBase 에 적재되는 데이터는 테이블과 row-key, column-family, timestamp 로 구성되며 수백만 개에 달하는 row 와 column 을 분산하여 저장할 수 있다. 또한 row-key 기반 정렬을 지원하고 column-family 에 임의의 하위 column 을 추가할 수 있어 유연한 데이터 처리가 가능하다. Timestamp 는 각 셀에 추가되어 이전 버전을 유지할 수 있도록 한다. 분산 환경에서 HBase 는 master 가 region 서버들을 관리하며 region 서버는 각 테이블의 데이터를 관리한다. 이 특성은 다른 NoSQL 대안으로 꼽을 수 있는 Cassandra[4]보다 처리 속도가 느릴 수 있으나 내결함성에서 강점이 있는 결과를 가지게 된다.

(2) 분산 데이터 베이스 구현 및 활용

실제 자율주행 데이터인 공개 데이터 셋 NGSIM US-101[5]의 ETL 처리 과정을 통해 분산 데이터베이스를 활용하여 데이터를 적재하는 과정을 설명한다. 해당 데이터는 고속도로에 인접한 36 층 건물 옥상에 설치된 CCTV 에서 특정 지역을 통과하는 차량들을 촬영한 영상을 바탕으로 차량의 정보와 움직임 등을 기록한 시계열 데이터이다. 해당 데이터를 CSV 형식으로 HBase 에 탑재하기 위해 Fig. 1 과 같은 ETL 과정을 거치며, 과정은 1 개의 master 서버와 여러 개의 region 서버로 구성된 HBase 환경에서 수행된다. 각 region 은 테이블의 부분 집합으로 region 서버에 나누어 분배되며,

이때 master 는 각 region 서버의 로드 밸런싱을 위해 region 을 재할당하거나 복구하는 등의 일을 한다. Row-key 와 region 의 매핑 정보는 메타데이터 형식으로 유지되며 이는 Zookeeper[6]에 위치한다. Zookeeper 조정자의 역할(Coordinator)로 메타데이터를 유지하고 각 서버의 상태를 관리한다.

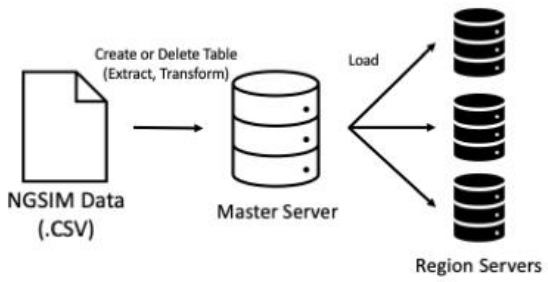

Fig. 1 ETL Process

클라이언트는 최초 데이터 적재 시 생성하는 테이블의 row-key 와 column-family 를 정의할 수 있으며, 이 column-family 의 정의 과정은 ETL 의 추출과정과 밀접한 관련이 있다. 이 과정을 통해 클라이언트가 정의하는 필요한 요소를 다른 요소들과 분리할 수 있으며, 이렇게 정의된 테이블에 의해 데이터가 변환될 것이기 때문에 이는 ETL 과정 중 추출(Extract)에 해당한다고 볼 수 있다. (본 논문에서는 이 과정을 추출로 정의한다.)

시계열 데이터를 시간 순으로 정렬하고자 하는 경우 예시 데이터에서는 'Frame_ID'를 row key 로 하여 시간 순으로 정렬할 수 있으며 필요에 따라 row-key 를 달리함으로써 분석에 활용할 수 있다. Fig.2 는 CSV 형식의 예시 데이터가 row key 와 column-family 를 갖는 HBase 데이터 모델로 변환된 예시이다. 클라이언트의 column-family 정의에 따라 column 이 묶여 매핑된 것을 확인할 수 있다. 또한 row key 를 기준으로 알파벳 내림차순 정렬된다. column-family 는 적재 이후에도 추가 및 삭제가 가능하여 확장성이 있으며 column 이 다시 여러 개의 column map 으로 구성된 형태이다.

CSV

Vehicle_ID	Frame_ID	V_length	...	Location
515	6548	13	...	us-101
2224	2230	14.3	...	us-101

HBase Data Model

Row Key	Vehicle		...	Location
	Column Family			
Frame_ID	ID	length	...	Location
2230	2224	14.3	...	us-101
6548	515	13	...	us-101

Fig. 2 ETL Example

적재(Load) 전 데이터는 생성된 HBase 테이블에 맞게 변환(Transform)된다. 예시 데이터인 NGSIM CSV 형식 데이터의 경우 HBase 에 내장된 MapReduce 라이브러리인 *ImportTsv* 를 통해 각 region 서버에 분산 적재할 수 있으며 병렬 처리된다. 분산 적재된 데이터에 대한 접근 또한 MapReduce 작업으로 처리된다. 만약 동일한 테이블에 다시 데이터를 적재할 경우 테이블의 timestamp 를 통해 구분할 수 있다.

III. 결론

본 논문에서는 대용량 자율주행 데이터에 효과적인 분산 시스템을 제시하고 이에 따라 얻을 수 있는 효과를 기술하였다. 데이터를 분산하여 병렬로 적재할 수 있는 것은 대용량 데이터를 데이터베이스에 적재하는 시간을 단축시킬 수 있으며 비용적 측면에서도 유리하다. 뿐만 아니라 본 연구에서 사용한 HBase 와 HDFS 를 기반으로 Spark 등의 데이터 처리 프레임워크를 연동하여 작업의 확장성을 높일 수 있다.

추후 연구에서는 다양한 실제 자율주행 시계열 데이터를 분산 데이터베이스에 탑재 및 처리하고 대안으로 채용될 수 있는 Cassandra 및 다른 분산 데이터베이스의 성능을 비교한다. 또한 데이터 적재 과정에서의 성능 최적화 문제, ETL 자동화 작업 등의 주제에 관심을 가지고 연구할 계획이다.

ACKNOWLEDGMENT

본 연구는 2023 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업(2022-0-01077), 2022 년도 한국연구재단 기본연구사업 (NRF-2021R1F1A1062779), 및 산업통상자원부 자율주행기술개발혁신사업 (2001 8248-주변 상황 인식 센서 성능 및 판단 기능 부족으로 인한 사고 위험 대응 기술(SOTIF) 개발)을 지원받아 수행한 연구임.

참 고 문 헌

[1] Geyer, Jakob, et al. "A2d2: Audi autonomous driving dataset.", arXiv preprint arXiv, 2020

[2] Dean, J. "MapReduce: simplified data processing on large clusters", Communications of the ACM, pp. 107-113, 2008

[3] Vora, M. N. "Hadoop-HBase for large-scale data.", Proceedings of 2011 International Conference on Computer Science and Network Technology, pp. 601-605, 2011

[4] Jogi, V. D. "Performance evaluation of MySQL, Cassandra and HBase for heavy write operation", 2016 3rd International Conference on Recent Advances in Information Technology (RAIT). IEEE, pp. 586-590, 2016

[5] U.S. Federal Highway Administration. (2005) US Highway 101 dataset.

[6] Hunt, P. "{ZooKeeper}: Wait-free Coordination for Internet-scale Systems", 2010 USENIX Annual Technical Conference, 2010