

Hadoop 을 이용한 자동차 센서 데이터 분석 기법 연구

윤재열*, 임지연**, 김이준***, 김응모*
성균관대학교 전자전기 컴퓨터공학과
e-mail : vntlffl@naver.com

Data Analysis of Car Sensor System using Hadoop Framework

Jae-Yeol Yoon*, Ji-Yeon Lim**, Iee-Joon Kim***, Ung-Mo Kim*
School of Information and Communication Engineering
Sungkyunkwan University

요 약

최근 정보의 다양화와 대량화로 인해 대두된 빅데이터 환경은 여러 분야로의 연구 방향을 제시하고 있다. 이미 데이터 양이 페타바이트 규모를 넘어서고 있으며, 이를 처리하기 위한 방안이 연구 중이다. 본 논문에서는 무선 통신기기 및 센서 기술의 발달로 대규모 네트워크 구축이 가능해진 센서 데이터 중 차량에 사용되는 센서 데이터를 연구하는 방안을 제시하고자 한다. 빅데이터 개념이 대두되면서 이슈화되고 있는 Hadoop 시스템을 이용하여 자동차 센서 데이터(CAN Message)를 분석하고자 한다.

1. 서론

최근 정보의 다양화와 대량화로 인해 빅데이터 라는 개념이 대두되고 있다. 이미 현재 데이터의 양이 페타 바이트 규모를 넘어서고 있으며, 각 국가 나 기업에서는 급속도로 증가하는 데이터를 분석하여 활용하기 위한 방안을 연구 중이다.[11]

대용량 데이터 처리 분석을 위해 더그 커팅이 개발한 Hadoop 은 대규모 분산 컴퓨팅을 지원하는 아파치의 오픈 소스 프로젝트이다. 이는 높은 확장성과 성능을 제공하면서 결합 탐지와 자동 복구 기능이 우수하며, 특히 구글의 MapReduce 병렬 처리 모델을 포함하고 있어 과학 기술 분야에서도 점차적으로 도입 및 활용되고 있다.

그 중 무선 통신기기 및 센서 기술의 발달로 대규모의 USN 의 구축이 가능해지면서 여러 응용분야에서 이를 활용하는 방안이 연구되고 있다.[1][2][3]

최근 자동차 센서 네트워크에 대한 선진 자동차 국가 들의 기술 개발 이 빠르게 진행되고 있으며, 자동차의 기계, 전자, 통신과 제어 공학의 발전으로 인하여 안전성과 편의성을 추구하기 위한 요구가 까다로워지고 있다. 또한 자동차에서 발생하는 각종 정보를 따져보자면, 현대의 차량에서 수십, 수백 개의 센서 데이터를 가지고 있다. 따라서 수백, 수천 대의 자동차에서 발생하는 정보 또한 빅데이터가 아닐 수 없을 것이다. 자동차에서 발생하는 각종 정보를 수집하고 이를 가공하여 시스템 제어 장치에 제공하여 활용 할 수 있도록 하여 주기 위한 것이 차량 내 센서 네트워크 시스템이다. 자동차 센서 네트워크 기술은 차량의 무게를 줄여 주며, 차량의 기능을 향상 시키며, 자동차 각 부품의 고장 여부를 진단, 처방하여 마치 유기

적인 생명체처럼 작동하게 한다.[4]

최근 대규모의 센서 데이터를 저장 및 처리 하는 방안으로 클라우드 컴퓨팅을 활용하는 연구가 진행되고 있는데, [5]에서는 SOA 를 통한 인터넷과 센서 네트워크의 통합을 제안하였고, [6]에서는 센서 네트워크와 클라우드를 이용한 웹 기반의 응용 프로그램과의 결합에 대한 프레임 워크를 제시하였다. [7]에서는 클라우드 서버를 통한 u-의료 시스템을 제시하였다. 또한 센서 네트워크에서 수집된 데이터를 클라우드 게이트 웨이로 전송하여 빠른 응답처리가 가능한 규칙 기반의 이상상황 검출과 이벤트 처리에 관해 제안하였다.[8]

기존의 연구들은 센서 네트워크와 클라우드와의 결합에 치중하고 있으나 Hadoop 이라는 최근 빅데이터 환경에 맞게 개발된 프레임워크로의 적용에 대한 연구가 부족하다. 따라서 Hadoop 시스템을 이용하여 자동차 센서 데이터를 분석하기 위한 방안을 연구한다.

본 논문의 구성은 다음과 같다. 2 절에서는 관련연구로서, CAN 통신 데이터와 Hadoop 에 관해 살펴보고, 3 절에서는 본 연구를 수행하기 위한 필요 요구 사항에 대해 기술하며, 4 절에서는 제안 방향, 마지막은 결론 및 향후 연구 방향에 대해서 기술한다.

2. 관련연구

본 절에서는 본 논문에서 제안하기 위한 자동차 센서 데이터와 Hadoop 에 관해 기술한다.

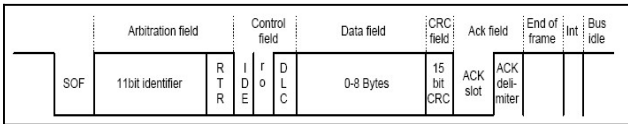
자동차 센서 네트워크에는 4 가지 표준을 따른다. LIN(Local Interconnect Network), CAN(Controller Area Network), MOST(Media Oriented Systems Transport), IEEE 1394. 이중 제어모듈, 센서, 액츄에이터 등을 연결하

기 위해 만들어진 시리얼 통신 프로토콜인 CAN 프로토콜만을 본 연구에 적용한다.

(1) CAN(Controller Area Network)

CAN은 각종 자동차내의 제어 장치들을 디지털 직렬 통신을 이용해 연결되어 사용하고 있으며, 저속으로는 125kbps 전송속도와 고속 전송속도로는 최대 1Mbps가 있다. 한 스테이션에서 고장이 발생하더라도 시스템의 일부는 사용할 수 있는 버스형(bus structure)과 스타형(star structure) 토폴로지를 지원하며 열악한 환경이나 고속, 충격이나 진동 노이즈가 많은 환경에서도 잘 견디기 때문에 차에 자동차에 적용되고 있다. [9]

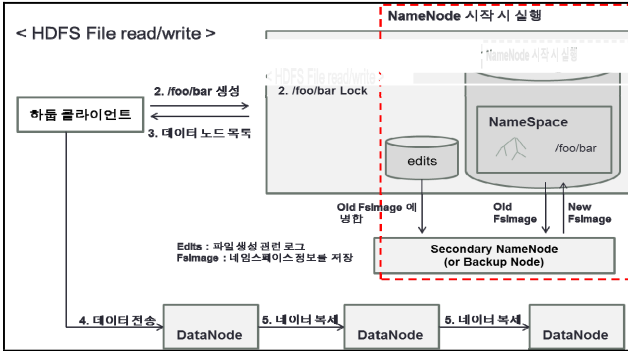
그림 1은 CAN 통신에 사용되는 메시지의 형태를 보여준다. CAN 메시지는 버스 시스템으로 구성된 네트워크의 중계를 위한 Arbitration field와 통신을 제어하기 위한 Control field, 실제 전송될 의미 있는 data가 포함된 Data field 등으로 구성되어 있다.



(그림 1) CAN Message frame format

(2) Hadoop의 데이터 처리 및 분석 구조

Hadoop은 단순한 프로그래밍 모델을 사용하여 컴퓨터클러스터 상에서 대규모 데이터를 분산 처리할 수 있도록 고안된 프레임워크이다. Hadoop은 크게 Hadoop 파일 시스템(HDFS)과 MapReduce로 구성된다.

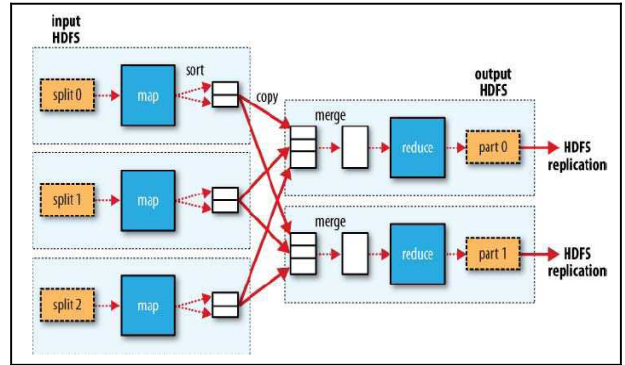


(그림 2) HDFS File read/write

Hadoop 파일 시스템은 테라 혹은 페타바이트 규모의 대규모 데이터를 저장할 수 있는 분산 파일 시스템으로 고성능의 파일 접근 및 입출력을 제공한다. Hadoop 파일 시스템에서 파일은 병렬처리 프로그램을 위한 고 가용성을 제공하고 결함에 효율적으로 대처할 수 있도록 기본 64MB의 블록 단위로 저장되는데, 이때 Replication 기능을 통해 둘 이상의 Datanode에 복제 및 분산되어 저장된다(그림 2). 따라서 시스템 오류로 인한 데이터 손실이 발생할 경우 다른 노드에 복제되어 있는 data를 통해 자동적으로 복구 과정을 수행하여 안정성을 높이게 된다. 그림 2는 Hadoop 파일 시스템의 기본 동작을 나타낸 것이다.

Hadoop MapReduce는 대규모의 데이터를 독립된 작업으로 나누어서 분산 병렬 처리할 수 있도록 설계된

프로그래밍 모델이다. MapReduce는 전체적인 MapReduce 작업 상태를 관리하는 Jobtracker와 맵 작업 혹은 리듀스 작업을 실행하는 Tasktracker로 나뉘게 된다. MapReduce는 그림 2와 같이 Map 단계와 Reduce 단계로 구성된다.



(그림 3) MapReduce 동작 과정

MapReduce는 그림 3과 같이 동작하지만 이는 개념적인 동작 일뿐 실질적인 동작이나 구성은 개발자에게 달려있다. 따라서 본 연구를 위해서 자동차 센서 데이터 맞게 프로그램을 개발한다.

Hadoop의 가장 큰 특징은 데이터 지역성(data locality)을 고려한 분산 병렬 데이터 처리가 가능하다는 점이다. Hadoop의 실행 노드는 Datanode와 Tasktracker를 함께 실행하고 있다. 전체 입력 데이터는 사용자가 전달한 속성 인자에 기반하여 논리적으로 분할된 데이터인 Input-Split을 계산한다. Hadoop은 InputSplit의 데이터 위치를 고려하여 각각의 데이터를 처리하기 위한 Tasktracker를 선택하고 맵 작업을 실행하게 된다. 즉, Hadoop은 데이터를 저장하고 있는 노드에서 데이터 처리 및 분석 작업이 이루어지게 된다. [13]

3. 필요 요구 사항 분석

자동차 센서 데이터를 분석하기 위해서 필요한 요구사항을 분석하여 연구 방향을 제시한다.

(1) CAN To Ethernet Gateway(C2E)

네트워크를 통해 Hadoop 시스템으로 CAN 통신 데이터를 수집하기 위해서는 C2E라는 장비가 필요하다. 이는 CAN 통신을 인터넷과 직접적으로 연결하기 위해서 다른 네트워크 간의 프로토콜 체계를 변환하여 주는 Gateway 장치의 필요성으로 인해 개발되었다.

본 논문에서는 PC와 CAN2Ethernet Gateway RS232나 TCP/IP 통신으로 데이터를 주고 받을 때의 2가지 형식 중 데이터의 흐름을 쉽게 확인하며, 제어할 수 있는 터미널 방식을 통해 접근한다. [10]

(2) CAN Message

본 논문은 실질적인 데이터 처리에 관하여 연구하기 위해 Data field만을 다룬다.

Byte 0	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6	Byte 7
추가 데이터 바이트 수	Mode	PID Code	0X55	0X55	0X55	0X55	0X55

그림 4 CAN Message-Data Field(Request)

그림 4 는 전달하는 데이터로서 전체 8 바이트 중 첫 번째 바이트(Byte0)는 몇 바이트의 정보가 오는지를 의미 하는 것으로 그 뒤로 Mode 와 PID Code 가 각각 1 바이트씩 2 바이트가 오기 때문에 2 가 되며, 세 번째 바이트(Byte2)는 PID 코드가 위치하고 그 이상은 사용되지 않으므로 0x55로 채워진다.

Byte 0	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6	Byte 7
추가 데이터 바이트 수	Custom Mode	PID Code	Value 0 (A)	Value 1 (B)	Value 2 (C)	Value 3 (D)	0X55

(그림 5) CAN Message-Data Field(Response)

그림 5 는 전달받은 데이터에 대한 응답 데이터로서 첫 번째 바이트는 질의 메시지와 동일하게 이후에 올 추가 데이터의 크기이다. 두 번째는 질의 메시지에서 사용된 Mode 값에 0x40 을 더한 값과 같다. 네 번째부터는 질의한 PID 에 대한 응답데이터가 온다. 데이터 크기는 1 바이트부터 4 바이트까지만 가질 수 있으며, 마지막 Byte7 은 사용되지 않아 0x55 로 채워진다. 응답데이터는 A, B, C, D 로 명명되고, 기본적으로 제공되는 PID 표를 통해 제공되며, Mode 는 현재 데이터 표시용 0X01 만 이용하며, 차량의 고장 진단 코드를 얻고자 할 경우 0X03 을 사용한다. [12]

(3) Hadoop 프레임 워크 H/W 구성

Hadoop 을 구축하기 위해서는 일반 CPU 보다는 서버 CPU 인 Xeon 을 사용하는 것이 적합하다. 또한 일반적으로 최소 3 번의 Rplication 을 수행하는 Hadoop 의 특성에 기반하여 최소 Slave PC 3~4 대로 구성한다. (그림 6)

<ul style="list-style-type: none"> 1G~10G 스위치 Num of Node <ul style="list-style-type: none"> Master : 최소 1 ~ 2 대 Slave : 3 ~ 4 대 Node 별 Spec. <ul style="list-style-type: none"> 2 * CPU(4 Core Per CPU) Xeons 2.5GHz 이상 or CPU(8 Core Per CPU) Xeons 2.5GHz 이상 4 * 1TB SATA HDD 16G RAM
--

(그림 6) Hadoop 구축을 위한 H/W Spec.

(4) Hadoop 프레임 워크 S/W 구성

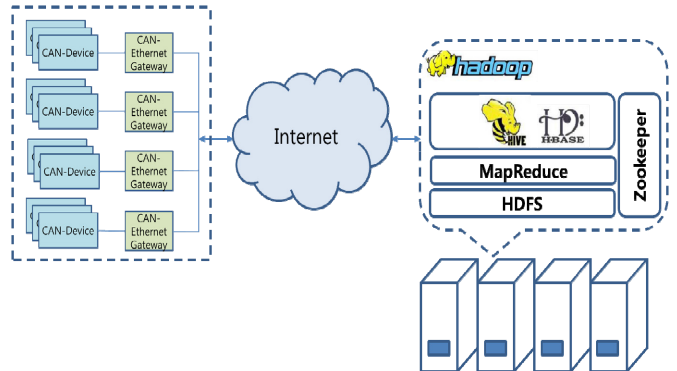
Hadoop 의 에코시스템으로 많은 S/W 가 있지만 최소한의 구성을 통해 시스템을 구축하고자 한다.

특징	명칭
대용량 분산 파일 시스템	Hadoop File System
대용량 분산 데이터 저장소	Hbase + Hive(integration)
배치 분산 병렬 처리	MapReduce(Hadoop)
분산 관리	ZooKeeper

(그림 7) Hadoop 구축을 위한 S/W 구성

Hadoop 시스템은 내부적으로 Linux 명령어들을 사용하기 때문에 환경설정이 간단하다. 특히 Ubuntu OS 에서 구성하기로 한다. Hadoop 시스템은 No-SQL 을 사용한다. No-SQL 의 종류는 다양하지만 Hadoop 에 최적화되어 있는 Hbase 를 통해 데이터를 관리 한다. No-SQL 은 기본적으로 join, union 과 같은 기능을 제공하지 않기 때문에 필요한 경우에 RDB 의 SQL 과 유사한 Hive QL 을 사용하여 Hbase 와 Native Hive 테이블간의 조합을 통해 Join, Union 을 수행한다.

4. 시스템 구성도



(그림 8) 전체 시스템 구성

그림 8 은 제안하고자 하는 시스템을 대략적으로 구성하여 나타낸 것이다. C2E 장비를 통해 네트워크로 연결하여 CAN 메시지를 제안하고자 하는 Hadoop 시스템에서 데이터를 전송 받는다.

5. Hadoop 을 이용한 데이터 처리

Hadoop 시스템으로 전송되는 CAN 메시지는 분산 컬럼 지향 데이터 베이스인 HBase 에 통해 저장되며, 여러 노드에 분산 및 중복 저장하여 확장성이 좋고 MapReduce 병렬 처리에 용이하게 한다.

실질적으로 의미 있는 데이터가 담긴 응답 DataField 가 Hadoop 시스템으로 전송되어 데이터 처리시 표준화된 메시지 형태로 변환하여 HDFS 에 업로드 한다. 이는 Mode, PID, 응답 메시지 정보가 들어있는 Byte3~6 값으로 구성된다. Hbase 를 통해 저장되는 데이터의 형태는 그림 9 와 같다.

Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6
Custom Mode	PID Code	Value 0 (A)	Value 1 (B)	Value 2 (C)	Value 3 (D)

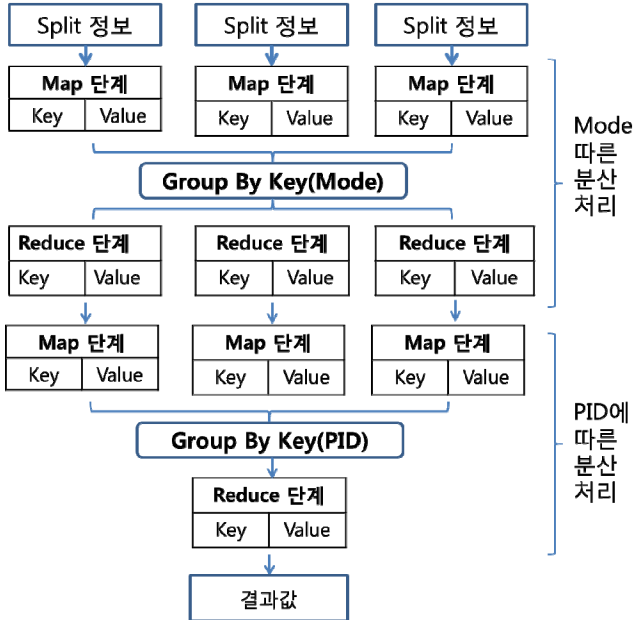
Row (Byte 1 ~2)		Column(byte 3~6)			
Custom Mode	PID Code	Value 0 (A)	Value 1 (B)	Value 2 (C)	Value 3 (D)

(그림 9) CAN Message Format 변환

MapReduce 데이터 처리를 위해서 HDFS 에 업로드된 정보를 입력 data 에 관한 Split 형태로 변환할 필요가 있다. 그림 9 와 같이 변환된 메시지는 다음과 같

이 분산 처리 된다.

변환된 Split 정보를 2 단계의 MapReduce 작업을 통해 Mode 별, PID 별로 데이터를 분류 및 병합하여 그룹화 시킨다. 최종적으로 Reduce 단계를 거친 후 분산 처리 된 데이터의 결과값은 다시 HDFS 를 통해 저장한다.



(그림 10) MapReduce 작업을 통한 데이터 분류

6. 결론

본 논문에서는 자동차 센서 데이터를 분석하기 위해 Hadoop 시스템을 적용하는 방향을 제시한다. 최근 이슈화 되고 있는 빅데이터를 분석하기 위한 프레임워크로서 Hadoop 시스템으로 통한 자동차 센서 데이터를 분석하기 위한 방안을 연구한다. 수백 수천대의 자동차의 탑재 되어 있는 센서 데이터를 실시간으로 분석하기 위해서는 클라우드 컴퓨팅 환경을 적용하는 것이 가장 효율적이라고 생각한다. 이를 통해 보다 확장성 있게 차량 상태를 모니터링 하는 데 사용된다면 좀더 빠른 응답 처리가 가능하지 않을까 예상된다. 본 논문은 아직 기획 단계이기 때문에 실험이 절대적으로 요구된다. CAN 데이터의 실험 및 분석 결과가 성공적으로 수행된다면, 자동차 센서 데이터 전체에 따른 Hadoop 시스템 적용도 기대해 볼만 하다.

참고문헌

[1] Aberer K. Alonso G. and Kossmann D. "Data Management for a Smart Earth – The Swiss NCCR-MICS initiative", SIGMOD Record, Vol. 35, No. 4, pp. 40-45, 2006.
 [2] Aberer K. Hauswirth M. and Salehi A. "Infrastructure for Data Processing in Large-Scale Interconnected Sensor Networks", 2007 International Conference on Mobile Data Management, pp. 198-205, May 2007.
 [3] Jardak C. Riihijärvi J. and Mähönen P. "Extremely Large-scale Sensing Applications for Planetary WSNs", In

Proceedings of the 2nd ACM international Workshop on Hot Topics in Planet-Scale Measurement, pp. 1-6, June 2010.

[4] 심영일, “새롭게 발전하는 자동차 센서 네트워크기술” Machinery Industry, pp. 66-73, 2008
 [5] V. Rajesh, J. M. Gnanasekar, R. S. Ponmagal, and P. Anbalagan, "Integration of Wireless Sensor Network with Cloud", 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 321-323, 2010.
 [6] M. Hassan, B. Song, and E. Huh, "A Framework of Sensor-cloud Integration Opportunities and Challenges", Conference On Ubiquitous Information Management And Communication, pp. 618-626, 2009.
 [7] X. Lee, S. Lee, P. T. True, L. T. Vinh, A. M. Khattak, M. Han, D. V. Hung, M. M. Hassan, M. Kim, K. Koo, Y. Lee, and E. Huh, "Secured WSN-integrated cloud computing for u-life care", In Proceedings of the 7th IEEE Conference on Consumer Communications and Networking Conference, pp. 702-703, Jan. 2010.
 [8] 클라우드 기반 센서 데이터 관리 시스템 설계 및 구현
 [9] Gye-Do P. Won-Kyo S. Jang-Myung L. “Moving robot distributed control system of CAN”
 [10] SYS TEC electronic GmbH. “Controller Area Network(CAN) - Ethernet Gateway System Manual”, 2007
 [11] 류성일, “Big Data 시대가 가져올 비즈니스 패러다임의 변화”, KT 경제 경영 연구소, 2011
 [12] 마을 소프트웨어, “Multi-CAN Analyzer 를 이용한 차량 ECU 모니터”, 2011
 [13] 김형준, 조준호, 안성화, 김병준, “클라우드 컴퓨팅 구현 기술”, 에이콘, 2011