Journal of the Korea Convergence Society
Vol. 7. No. 5, pp. 1-6, 2016 http://dx.doi.org/10.15207/JKCS.2016.7.5.001

하둡 분산 환경 기반의 데이터 수집 기법 연구

진고환
우송대학교 IT융합학부

A Study on the Data Collection Methods based Hadoop Distributed Environment

Go-Whan Jin
Division of IT Convergence, Woosong University

요 약 최근 빅데이터 활용과 분석기술의 발전을 위하여 많은 연구가 이루어지고 있고, 빅데이터를 분석하기 위하여 처리 플랫폼인 하둡을 도입하는 정부기관 및 기업이 점차 늘어가고 있는 추세이다. 이러한 빅데이터의 처리와 분석에 대한 관심이 고조되면서 그와 병행하여 데이터의 수집 기술이 주요한 이슈가 되고 있으나, 데이터 분석 기법의 연구에 비하여 수집 기술에 대한 연구는 미미한 상황이다. 이에 본 논문에서는 빅데이터 분석 플랫폼인 하둡을 클러스터로 구축하고 아파치 스쿱을 통하여 관계형 데이터베이스로부터 정형화된 데이터를 수집하고, 아파치 플룸을 통하여 센서 및 웹 애플리케이션의 데이터 파일, 로그 파일과 같은 비정형 데이터를 스트림 기반으로 수집하는 시스템을 제안한다. 이러한 융합을 통한 데이터 수집으로 빅데이터 분석의 기초적인 자료로 활용할 수 있을 것이다.

• 주제어 : 빅데이터, 하둡, 아파치 스쿱, 아파치 플룸, 융합

Abstract Many studies have been carried out for the development of big data utilization and analysis technology recently. There is a tendency that government agencies and companies to introduce a Hadoop of a processing platform for analyzing big data is increasing gradually. Increased interest with respect to the processing and analysis of these big data collection technology of data has become a major issue in parallel to it. However, study of the collection technology as compared to the study of data analysis techniques, it is insignificant situation. Therefore, in this paper, to build on the Hadoop cluster is a big data analysis platform, through the Apache sqoop, stylized from relational databases, to collect the data. In addition, to provide a sensor through the Apache flume, a system to collect on the basis of the data file of the Web application, the non-structured data such as log files to stream. The collection of data through these convergence would be able to utilize as a basic material of big data analysis.

• Key Words : Big Data, Hadoop, Apache Sqoop, Apache Flume, Convergence

1. 서론

최근 정보기술의 발전으로 데이터 수집 및 분석에 대한 다양한 서비스들이 출현하고 있다. 특히 RFID, USN 시스템의 발전으로 많은 데이터가 발생되어 수집되고 있고, 오프라인에서 동작하는 각종 기기들이 유무선 통신망을 통하여 각종 센서의 데이터를 실시간으로 전송하고 있어 데이터 수집 및 자료의 정확성이 높아지고 있다

*Corresponding Author : 진고환(gwjin@wsu.ac.kr)
Received September 11, 2016 Revised October 11, 2016
Accepted October 20, 2016 Published October 31, 2016

[1,2]. 스마트 폰의 확산과 개인 블로그, SNS 등의 서비스로 방대한 양의 데이터가 발생하는 빅데이터 환경이 만들어지고 있는 상황으로[3], 빅데이터는 사회, 과학 기술 등 모든 영역에서 가치가 있는 정보를 제공할 수 있는 가능성을 제시하며 그 중요성이 부각되고 있다[4]. 특히 데이터는 주로 데이터베이스에 보관되어 있는 정형데이터가 주를 이루고 있으나, 스마트 시대의 도래와 통신수단의 발달로 비정형 데이터가 전체 데이터의 80%를 상회하고 있는 것이 현실이다[5]. 이러한 많은 데이터들은 단지 보관만 한다면 그 가치는 사라질 것이며, 이를 처리하고 분석해야만 새로운 가치를 창출할 수 있고[6], 이를 위한 빅데이터 처리 및 분석기술이 비약적으로 발전하고 있다[7]. 또한 소셜네트워크 및 사물인터넷에서 자동으로 발생되고 있는 수많은 데이터의 양과 형태는 과거의 처리기술로는 불가능한 상태이며, 새로운 융합 기반을 요구하고 있다[8].

빅데이터의 활용과 분석 기술의 발전을 위한 연구중에서 대표적인 것은 아파치 소프트웨어 재단의 오픈소스인 하둡과 이를 효율적으로 사용 할 수 있는 서브프로젝트를 중심으로 하는 하둡에코시스템으로서, 빅데이터 분석을 위한 주요한 트렌드가 되고 있다. 특히 축적된 데이터를 해석하고 분석하는 업무에서 빅데이터 분석의 표준처럼 평가되고 있는 하둡은 빅데이터 프로젝트 수행을 진행한 응답 기업의 24%의 지지를 얻어 가장 빈번하게 사용되는 분석 툴로 조사 되었다[9]. 한편 빅데이터의 처리와 분석에 대한 관심도가 높아지면서 그와 병행하여 다양한 시스템에 분산된 정형/비정형의 데이터를 어떻게 단일 하둡파일시스템(HDFS : Hadoop Distributed File System)에 적재하는가 하는 데이터 수집 기술이 주요한 이슈가 되고 있다.

이러한 수집 기술은 조직 내부 및 외부에서 분산된 데이터 소스로 부터 필요한 데이터를 검색하여 자동 및 수동으로 수집하는 과정과 관련한 기술로서, 단순하게 데이터 확보하는 것이 아닌 검색/수집/변환을 통해 정제된 데이터를 확보하는 기술을 의미한다[10].

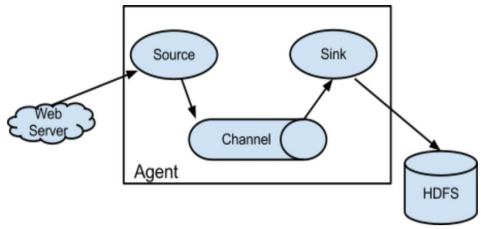
특히 IoT(Internet of Things) 환경에서 센서와 같은 다양한 데이터 소스로 부터 발생하는 데이터를 하둡 시스템으로 체계적으로 수집 및 저장하기 위해 아파치 플룸을 활용하는 연구[11]와, 웹서비스의 웹로그 데이터를 효율적으로 수집하는 기법에 관한 연구[12], 아파치 스쿱을 통하여 관계형 데이터베이스에 저장되어 있는 대용량의 정형 데이터를 빠르게 처리하기 위한 연구[13], 사회연결망 분석을 위하여 데이터 분석 도구를 이용한 연구[14]등이 진행되고 있으나, 데이터 분석에 대한 연구에 비하여 데이터 수집에 관한 연구는 상대적으로 미미한 실정이다.

또한 기존에 진행된 연구의 대부분은 정형 데이터 수집 혹은 비정형 데이터 수집에 제한적으로 연구되고 있어, 정형/비정형 데이터가 복합적으로 수집되는 시스템에 대한 연구가 필요한 실정이다. 이에 본 논문에서는 아파치 스쿱(Apache Sqoop)을 통해 관계형 데이터베이스로 부터 정형화된 데이터를 수집하고 아파치 플룸(Apache Flume)을 통해 IoT 및 웹 애플리케이션으로부터 데이터 파일이나 로그 파일과 같은 비정형 데이터를 스트림 기반으로 수집하는 시스템을 제안한다.

2. 관련 연구

2.1 플룸(Flume)

다음의 [Fig. 1]은 아파치 플룸의 아키텍처이다.



[Fig. 1] Apache Flume Architecture

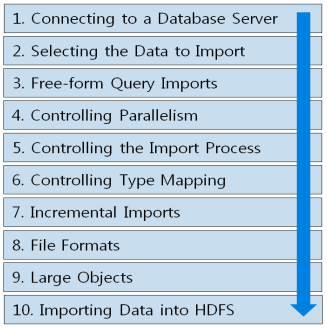
데이터가 발생해도 데이터의 손실 없이 전송 할 수 있을 것인가 하는 신뢰성(Reliability)의 문제와 내고장성(Fault tolerant), 분산된 시스템 환경에서 탄력 있게 대응할 수 있는 시스템의 확장성(Scalability)의 문제를 고민하게 되었고, 하둡에코시스템의 하나인 아파치 플룸은 훌륭한 해결책이 되고 있다. 플룸은 웹서버와 같은 외부의 소스를 통하여 데이터가 입력되고, 채널을 통하여 싱크로 나가는 구조로 되어 있다[15].

플룸은 분산된 대량의 데이터를 안정적이고 효율적으로 전달할 수 있는 시스템으로, 분산된 구조로 확장할 수 있어, 데이터의 양이 급증해도 기존 서버의 설정이나 구조 변경 없이 손쉽게 확장할 수 있는 장점을 가지고 있다. 이러한 플룸은 주로 로깅 시스템에 사용되며, 다양한 장

비에서 저장된 다양한 형태의 데이터를 하둡 등의 저장소에 저장할 수 있게 데이터를 이동시키는 역할을 한다. 특히, 플룸의 경우 다양한 데이터 플로우를 구성할 수 있으며 마스터 노드에서 통합적으로 관리할 수 있는 웹 페이지를 제공하고, 하둡과 통합이 잘되어 있어 실시간 데이터 수집에 매우 효율적으로 사용할 수 있다[16].

2.2 스쿱(Sqoop)

아파치 스쿱은 맵리듀스를 기반으로 구현된 데이터 적재 프로그램으로 아파치 소프트웨어 재단의 최상위 레벨 아파치 프로젝트이다. 특히 관계형 데이터베이스 및 하둡파일시스템 사이에 데이터 적재가 가능하기 때문에 많은 프로젝트에서 널리 사용하고 있다. 스쿱은 모든 적재 과정을 자동화하며 병렬처리 방식으로 작업하고, 좋은 내고장성(fault tolerance)을 지원한다[17,18,19]. 스쿱은 row-by-row 방식으로 관계형 데이터베이스에 저장되어있는 테이블을 읽어 하둡파일시스템에 저장하며, 테이블 하나를 파일셋으로 저장한다. 스쿱이 관계형 데이터베이스를 하둡파일시스템에 적재하는 절차는 다음의 [Fig. 2]와 같다[22].



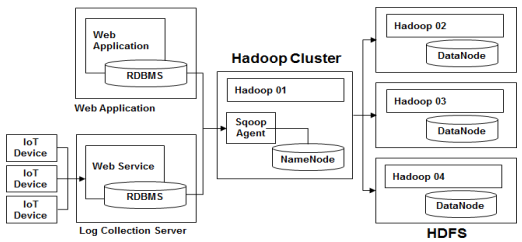
[Fig. 2] Data Loading Process of Apache Sqoop

병렬처리 방식으로 적재하기 때문에 적재한 후에 시스템에서 여러 개의 파일로 저장하게 된다. 스쿱을 사용하여 시스템에 저장되어 있는 파일셋을 읽고, 관계형 데이터베이스로 적재하는 Export 과정도 가능하다. 파일셋을 병렬처리 방식으로 읽고 레코드 형태로 변환하여 관계형 데이터베이스의 해당 테이블에 삽입한다[20,21].

3. 제안 시스템

3.1 스쿱을 이용한 정형 데이터 수집

웹 애플리케이션 환경에서 상당수의 데이터는 정형 데이터로서 각각의 데이터베이스에 적재되어 있다. 그러므로 하둡을 통한 빅데이터 분석에 앞서 관계형 데이터베이스 환경에 있는 데이터를 손실 없이 하둡파일시스템으로 수집하는 것이 중요하다. 다음의 [Fig. 3]은 분산된 시스템 환경에서 스쿱을 통해 하나 이상의 관계형 데이터베이스에 접속하여 데이터를 수집하고, 하둡파일시스템에 적재하는 구성도이다.



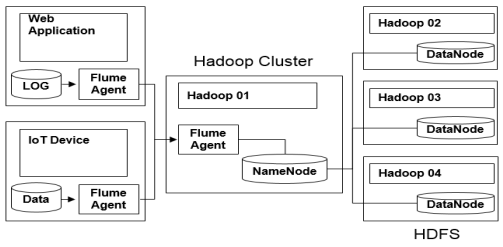
[Fig. 3] Data Acquisition System using the Scoop

스쿱 에이전트는 하둡 서버에 설치되어 동작한다. 스쿱 에이전트는 JDBC(Java Database Connectivity) API를 사용하여 하나 이상의 목적지 관계형 데이터베이스에 연결되어 단일 혹은 전체 테이블의 데이터를 맵리듀스 기반으로 데이터를 수집하여 적재한다.

3.2 플룸을 이용한 비정형 데이터 수집

웹서버의 로그파일은 서비스의 질을 높이고 보안이나 해킹을 탐지하는데 좋은 참고자료가 된다. 또한, 웹 서버에 이상이 생겼을 경우 발생한 오류를 조기에 발견하는 데에도 사용되는 중요한 자료이다[23].

다음의 [Fig. 4]는 플룸을 이용한 비정형 데이터 수집 시스템의 구조이다. 플룸은 하나 이상의 웹 애플리케이션에서 발생하는 로그 파일이나 데이터 파일과 같은 비정형 데이터를 Avro, Thrift, Syslog, Netcat과 같은 스트림 방식으로 수집할 수 있다. 수집된 데이터는 하둡파일시스템에 적재하여 하둡을 통한 데이터 분석이 가능하다.

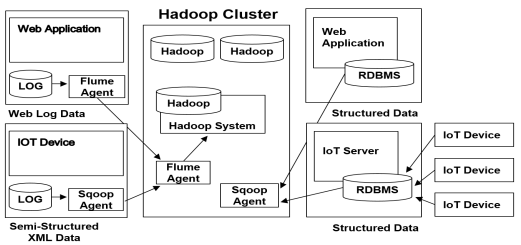


[Fig. 4] Data Acquisition System using the Flume

또한 분산 환경의 경우 별도의 설정 없이 에이전트의 추가도 가능하다.

4. 데이터 수집 실험

제안 시스템의 실험 환경 구축을 위하여 운영체제는 Win7(64bit)와 Ubuntu 14.04 버전을 기반으로 하였고, 데이터베이스는 MySQL-Server 5.6, 하둡 클러스터 구축은 Hadoop 1.2.1, 웹 애플리케이션은 Tomcat 7.0.62를 이용하였다. 실험을 위하여 4대의 컴퓨터로 하둡 클러스터를 구축한 후 데이터 생성을 위하여 별도의 자바 애플리케이션과 아파치 톰캣을 사용한 웹 애플리케이션을 구현하였다. 다음의 [Fig. 5]는 본 논문에서 제안하는 시스템의 실험 환경을 위한 구성도이다.

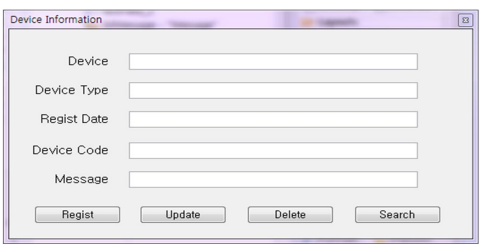


[Fig. 5] Test Environment Configuration

각각의 애플리케이션을 통하여 MySQL에 접속하여 데이터를 생성한 후 생성된 데이터를 스쿱을 통해 하둡 파일시스템에 저장한다. 또한 웹 애플리케이션에서 발생하는 로그 파일이나 데이터 파일과 같은 비정형 데이터 수집을 위하여 플럼을 통하여 수집한다.

4.1 정형 데이터 수집

다음의 [Fig. 6]은 자바 에이전트 입력 화면으로 각각의 데이터를 입력 받아 데이터베이스 서버의 디바이스 테이블에 입력 데이터를 삽입하는 기능을 수행한다.

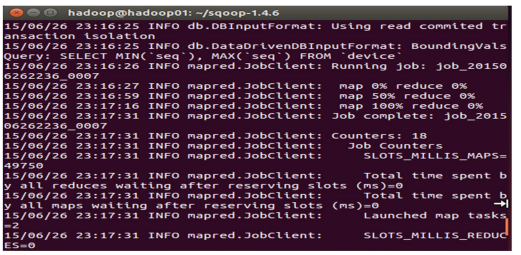


[Fig. 6] Java Agent UI

데이터를 입력 한 후 다음의 [Fig. 7]과 같은 스쿱 커맨드 명령을 실행함으로서 데이터베이스 및 대상 테이블에 접속한다. 스쿱 명령 실행 후 접속된 데이터베이스 및 대상 테이블에서 다음의 [Fig. 8]과 같이 하둡 맵리듀스 작업을 거쳐 하둡파일시스템에 적재된다.

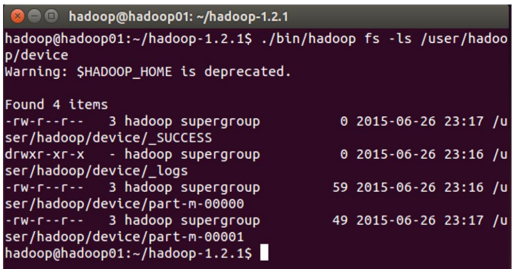
```
./bin/sqoop import --connect
jdbc:mysql://hadoop01/hadoop
--username root --password 0000 --table device
```

[Fig. 7] Sqoop Command



[Fig. 8] Data Acquisition(Sqoop)

다음의 [Fig. 9]는 하둡파일시스템에 적재된 디바이스 테이블의 데이터 결과로서, 적재된 파일은 하둡파일시스템 명령을 통해 확인할 수 있다.



[Fig. 9] Device Table Data

4.2 비정형 데이터 수집

다음의 [Fig. 10]은 플룸의 형상정보이다.

```
agent01.sources = avroGenSrc
agent01.channels = memoryChannel
agent01.sinks = hdfs-sink
agent01.sources.avroGenSrc.type = avro
agent01.sources.avroGenSrc.bind = 192.168.56.101
agent01.sources.avroGenSrc.port = 3333
agent01.sources.avroGenSrc.channels =
memoryChannel
agent01.sinks.hdfs-sink.type = hdfs
agent01.sinks.hdfs-sink.hdfs.path =
hdfs://hadoop01:9000/user/flume
agent01.sinks.hdfs-sink.hdfs.fileType = DataStream
agent01.sinks.hdfs-sink.hdfs.writeFormat = Text
agent01.sinks.hdfs-sink.channel = memoryChannel
agent01.channels.memoryChannel.type = memory
agent01.channels.memoryChannel.capacity = 100000
agent01.channels.memoryChannel.transactionCapacity
= 100000
```

[Fig. 10] Flume Configuration Information

플룸은 데이터를 전송해주는 에이전트 서버와 데이터를 입력 받는 수집 서버에 각각 설치된다. 형상정보는 데이터 수집 서버에 설치되어 반정형/비정형 데이터를 수집하기 위하여 각각의 전송 에이전트의 신호를 입력 받을 수 있다.

```
<HEALTHKIT-INFO>
        <HEALTH-DATE>0000-00-00</HEALTH-DATE>
        <HEALTH-ID>0000000</HEALTH-ID>
        <HEALTH-HEIGHT>000</HEALTH-HEIGHT>
        <HEALTH-WEIGHT>000</HEALTH-WEIGHT>
    <HEALTH-HEART_RATE>000</HEALTH-HEART_RATE>
    <HEALTH-CALORIES_BURNED>000</HEALTH-CALORIES_BURNED>
    <HEALTH-RUN>000</HEALTH-RUN>
    <HEALTH-STEP>000</HEALTH-STEP>
</HEALTHKIT-INFO>
```

[Fig. 11] Semi-Structured XML Data

플룸은 source,channels,sink 설정을 통해 전송 에이전트의 신호를 입력 받아 하둡파일시스템에 적재한다. 앞의 [Fig. 11]은 반정형 데이터로서 XML로 변환되어 플룸 에이전트간의 스트림 방식으로 하둡파일시스템에 저장된 데이터이다. 이때 수집된 데이터는 관계형 데이터베이스에 저장되지 않고, 다음의 [Fig. 12]와 같이 에이전트로부터 전송된 데이터가 하둡파일시스템에 적재된다.

```
hadoop@hadoop01: ~/flume-1.3.1
source avroGenSrc started.
2015-06-26 23:38:16,044 (pool-6-thread-1) [INFO - org.apache.av
ro.ipc.NettyServer$NettyServerAvroHandler.handleUpstream(NettyS
erver.java:171)] [id: 0x3cc425b2, /192.168.56.102:59422 => /192
.168.56.101:3333] OPEN
2015-06-26 23:38:16,052 (pool-7-thread-1) [INFO - org.apache.av
ro.ipc.NettyServer$NettyServerAvroHandler.handleUpstream(NettyS
erver.java:171)] [id: 0x3cc425b2, /192.168.56.102:59422 => /192
.168.56.101:3333] BOUND: /192.168.56.101:3333
2015-06-26 23:38:16,053 (pool-7-thread-1) [INFO - org.apache.av
ro.ipc.NettyServer$NettyServerAvroHandler.handleUpstream(NettyS
erver.java:171)] [id: 0x3cc425b2, /192.168.56.102:59422 => /192
.168.56.101:3333] CONNECTED: /192.168.56.102:59422
2015-06-26 23:38:23,996 (hdfs-hdfs-sink-call-runner-0) [INFO -
org.apache.flume.sink.hdfs.BucketWriter.doOpen(BucketWriter.jav
a:208)] Creating hdfs://hadoop01:9000/user/flume/FlumeData.1435
329500595.tmp
2015-06-26 23:38:54,488 (hdfs-hdfs-sink-roll-timer-0) [INFO - o
rg.apache.flume.sink.hdfs.BucketWriter.renameBucket(BucketWrite
r.java:427)] Renaming hdfs://hadoop01:9000/user/flume/FlumeData
.1435329500595.tmp to hdfs://hadoop01:9000/user/flume/FlumeData
.1435329500595
```

[Fig. 12] Loading Data into HDFS

4. 결론

빅데이터를 분석하기 위해 하둡을 도입하는 정부기관 및 기업이 점차 늘어나면서 다양한 데이터를 기존 관계형 데이터베이스나 데이터웨어하우스로부터 데이터의 손실 없이 하둡파일시스템에 전송하는 방법에 대한 수요가 점차 늘어나고 있다. 또한 하둡은 실시간 데이터 분석에 대한 제약이 따르기 때문에 이를 지원하기 위해 아파치 그룹에서는 하둡의 서브프로젝트인 스쿱과 플룸을 상위 레벨 프로젝트로 승격하여 데이터 수집에 대한 보다 활발한 연구 활동을 진행하고 있다. 이에 본 논문에서는 빅데이터 분석 플랫폼인 하둡을 클러스터로 구축하여 웹 애플리케이션 로그 및 IoT 장비에서 발생하는 정형 및 비정형 데이터를 각각 스쿱과 플룸을 통해 하둡파일시스템으로 수집하여 실시간 데이터 분석이 가능한 시스템을 제안한다. 이를 통해 정부기관이나 기업에서 보관 중인 대량의 데이터를 스트림 방식을 통해 수집할 수 있으며, 실시간으로 발생하는 IoT 장비의 로그 데이터를 수집하여 데이터 분석이 가능하다.

향후 연구에서는 대규모의 데이터 수집 및 병렬 데이터 수집에 대한 연구를 통하여 수집 서버 성능 향상에 대한 연구가 계속되어야 할 것이다.

REFERENCES

[1] O. B. Kwon, K. S. Kim, “The Design and Implementation of Location Information System using Wireless Fidelity in Indoors”, Journal of Digital Convergence, Vol. 11, No. 4, pp. 243-249, 2013.

[2] K. H. Lee, D. I. Kim, D. H. Kim, M. Y. Sung, Y. K. Lee, S. Y. Jung, “Implementation of Real-Time Video Transfer System on Android Environment”, Journal of th Korea Convergence Society, Vol. 3, No. 1, pp. 1-5, 2012.

[3] J. T. Kim, B. J. Oh and J. Y. Park, “Standard Trends for the Big Data Technologies”, Electronics and Telecommunications Trends 2013, ETRI, pp. 92-99, 2013.

[4] Y. S. Jeong, Y. T. Kim, G. C. Park, “Subnet Selection Scheme based on probability to enhance process speed of Big Data”, Journal of Digital Convergence, Vol. 13, No. 9, pp. 201-208, 2015.

[5] M. G. Song, S. B. Kim, “A Study of improving reliability on prediction model by analyzing method Big data”, Journal of Digital Convergence, Vol. 11, No. 6, pp. 103-112, 2013.

[6] M.J. Song, “Big Data is Creating Future Business

Map", Hansmedia, 2012.
[7] K. S. Noh, S. T. Park. K. H. Park, "Convergence Study on Big Data Competency Reference Model", Journal of Digital Convergence, Vol. 13, No. 3, pp. 55-63, 2015.
[8] S. H. Namn, K. S. Noh, "A Study on the Effective Approaches to Big Data Planning", Journal of Digital Convergence, Vol. 13, No. 1, pp. 227-235, 2015.
[9] BigData Monthly, "Big Data in the World," BigData World, Report, Vol. 8, 2015.
[10] S. A. Shin, K. E. Kim, "Classification and the Current State of Big Data Technology", National Information Society Agency, Korea Big Data Center, 2013.
[11] Y.H. Kang, "Design of a Framework of a System for Handling Streaming Data by Using Apache Flume", Journal of KIIT, Vol. 12, No. 11, pp. 127-132, 2014.
[12] U. G. Han, J. H. Ahn, "Load Balancing Method for Improving Performance of Apache Flume Log Aggregator", Proceeding of KIIT, pp. 314-317, 2014.
[13] Liu Chen, J.H. Ko, J.M. Yeo, "Analysis of the Influence Factors of Data Loading Performance Using Apache Sqoop", Journal of KIPS, Vol. 4, No. 2, pp. 77-82, 2015.
[14] K. C. Choi, J. A. Yoo, "A reviews on the social network analysis using R", Journal of the Korea Convergence Society, Vol. 6, No. 1, pp. 77-83, 2015.
[15] Apache Flume 1.4.0 User Guide, https://flume.apache.org/FlumeUserGuide.html.
[16] K. J. Park, "Big Data Eco System(Around the Platform)", Journal of KIIE, ie Magazine, Vol. 19, No. 3, pp. 41-47, 2012.
[17] Apache Sqoop, http://sqoop.apache.org
[18] Kathleen Ting, Jarek Jarcec Cecho, "Apache Sqoop Cookbook", O'Reilly, 2013.
[19] Rinusha Irudeen, Sanjeeva Samaraweera, "Big data solution for Sri Lankan development: A case study from travel and tourism", in Advances in ICT for Emerging Regions, International Conference on, 2013.
[20] Nodar Momtselidze, Alex Kuksin "Hadoop Integrating with Oracle Data Warehouse and Data Mining", in Journal of Technical Science and Technologies, Vol .2, No. 1, 2013.
[21] Ankit Jain, "Instant Apache Sqoop", Packt Publishing Ltd, 2013.
[22] Ognjen V. Jodzic, Dijana R. Vukovic, "The Impact of Cluster Characteristics on HiveQL Query Optimization", in Telecommunications Forum (TELFOR), 21st, 2013.
[23] K.B. Ryu, H.J. Park, "Mobile Web Server Log Analyzer", Proceeding of KSII, Vol. 5, No. 2, pp. 73-76, 2004.

저자소개

진 고 환(Go-Whan Jin) [정회원]

- 1990년 2월 : 한국과학기술원 산업공학과 (공학석사)
- 1999년 2월 : 한국과학기술원 테크노경영대학원 (공학박사)
- 2002년 3월 ~ 현재 : 우송대학교 IT융합학부 교수

<관심분야> : 빅데이터, 이동통신, 기술경영, 통신최적화