빅 데이터 기술을 이용한 에너지 효율화 분석 시스템 제안

육찬심[01] 임효진[2] 이주경[3]
[01]경기대학교 컴퓨터과학과, 차세대융합기술연구원 공공데이터연구센터
[2]한양대학교 정보시스템학과, 차세대융합기술연구원 공공데이터연구센터
[3]아주대학교 컴퓨터공학부, 차세대융합기술연구원 공공데이터연구센터
chansim1008@kyonggi.ac.kr, rskf415@hanyang.ac.kr, rudolph0724@ajou.ac.kr

A proposal of an energy efficiency management system using Big data technology

Chansim Youk[O1] Hyojin Lim[2] Jukyong Lee[3]
[01]School of Science, Kyonggi University & Public Data Research Center, Advanced Institutes of Convergence Technology
[2]Department of Information System, Hanyang University & Public Data Research Center, Advanced Institutes of Convergence Technology
[3]School of Computing, Ajou University & Public Data Research Center, Advanced Institutes of Convergence Technology

요 약

전력에 대한 가치의 비중이 낮아지면서 전력 소비의 극대화를 위하여 많은 연구들이 진행되었고 빅 데이터의 등장으로 빅 데이터를 이용한 연구가 다방면으로 진행되고 있다. 본 논문은 전력과 비 전력 데이터를 이용하여 전력 소비의 효율성을 극대화하기 위하여 빅 데이터 분석시스템 서버를 구성하였고, 서버에 설치되어있는 Hadoop을 통해 데이터의 저장과 분석이 이루어진다. 데이터는 Data Portal Server에 게이트웨이를 통하여 수집되며, 수집된 데이터를 Sqoop이 HDFS 위의 Hadoop Database(HBase)로 적재한다. 본 연구는 향후 저장이 완료된 데이터에 한하여 패턴을 찾고 데이터마이닝(Data Mining)기술을 적용하여 전력효율의 극대화를 실현한다.

1. 서 론

현대사회가 발전하면서 사람의 생활은 많이 편리해졌다. 이러한 편리함이 익숙해지면서 사람들은 당연하다고 생각하는 것에 대하여 가치의 비중이 많아 낮아졌다. 대표적으로 전력을 예로들 수 있는데 전력사용의 익숙함으로 사람들로 하여금 전력에 대한 가치의 비중을 축소시켰으며 이는 전력 소비의 효율성을 감소시키는 계기가 되었다.

전력 소비의 효율성 감소는 전기누출사고 등 사람에게 직접적인 영향을 미칠 뿐만 아니라 지구온난화 등 환경에도 많은 영향을 미친다. 전력 소비의 효율성이 감소하는 문제점을 해결한다면 전기를 생성하며 만들어지는 온실가스의 양을 절약하여 지구온난화를 예방할 수 있다. 또한 국가적으로 백 억 원에 이르는 전기 절감 효과를 볼 수 있으며 블랙아웃(Black out)사태를 대비한 예비전력을 충분히 저장하는데 있다.

본 연구는 전력과 비 전력 데이터 분석 시스템 서버를 구성하여 전력 소비 효율성의 극대화가 이뤄질 수 있도록 그 토대를 마련한다.

* 본 연구는 2013년도 산업통상자원부의 재원으로 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. 20132010101800) 또한 차세대융합기술연구원 스마트 그리드 센터의 최중인 센터장님, 이인규 박사님, 그리고 군산대 통계컴퓨터과학과의 온병원 교수님의 지도하에 작성된 논문입니다.

본 논문의 구성은 2장에서 빅 데이터 분석 시스템 서버의 구성에 대하여 설명하고 3장에서는 Sqoop를 이용한 데이터 이동에 대하여 설명한다. 4장에서는 샘플 데이터를 이용한 웹 기반의 API에 대하여 설명하고 5장에서는 전력 사용의 효율화를 위한 플랫폼의 구축에 대하여 설명한다. 마지막으로 6장에서는 결론을 내린다.

2. 빅 데이터 분석 시스템 서버 구성

본 연구에서는 분산 처리 시스템 환경 구축을 위해 그림 1과 같이 12개의 Node로 구성 된 서버를 이용하였다.

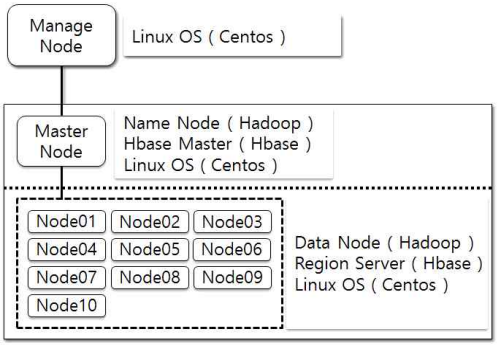

그림 1. 시스템 서버 구성

각 Node는 CentOS release 6.4(Final)와 Hadoop(Ver. 2.2.0), HBase(Ver. 0.98.2), Sqoop(Ver. 1.4.4)을 설치하였으며 하드웨어 구성은 다음과 같다.

- CPU : Intel(R) Xeon(R) CPU E5-2430 0 @ 2.20GHz
- Hard Disk : DELL PERC H310 8TB
- Main Memory :
  Manage Node=16G, Master Node=32GB, Node=24GB

Master Node는 Hadoop의 Name Node와 HBase의 HBase Master로 설정하였으며, Node01~10은 Hadoop의 Data Node와 HBase의 Region Server로 설정하였다. 또한 Node10의 경우 Standby Namenode로 설정하였다.

본 서버는 Hadoop과 HBase를 이용하여 수집해온 전력과 비 전력 데이터를 분산저장하고 패턴을 찾아 데이터 마이닝 기술을 적용하여 분석한다. Client는 분석이 완료된 데이터를 제공되는 API을 이용하여 요청할 수 있으며, 서버에서 JSON 형식으로 제공받을 수 있다.

2.1 분산처리 시스템 Hadoop

구글의 분산파일 시스템(GFS) 논문이 공개된 후, 그 구조에 대응하는 체계로 만들어진 Hadoop은 스케일아웃(Scale Out)을 지원하며 빅 데이터를 처리하기 위해 많은 기업들이 선호하고 있는 분산처리 시스템이다.[1,2,6]

빅 데이터라고 불리기 위해서는 3V(Variety, Velocity, Volume) 1C(Complexity) 중 2가지 이상을 만족해야 한다. 본 연구를 위해 수집되는 전력과 비 전력 데이터는 빅 데이터의 조건에 충분히 만족하고 있다.[3]

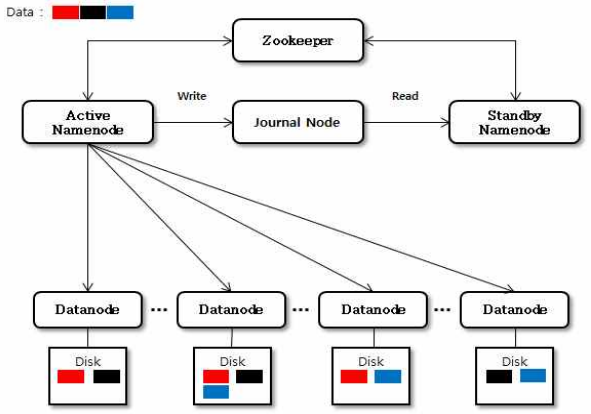

그림 2. Hadoop2의 HDFS 저장 구조

그림 2에서 Active Name Node는 Journal Node에 Edit Log를 기록하고, Standby Namenode는 이를 읽고 저장한다. 또한 Active Namenode는 각각의 Datanode에 데이터를 분산시켜 저장하고 저장이 완료된 데이터에 대한 메타데이터를 자신의 디스크에 저장한다. 하나의 데이터는 데이터의 안정성을 보장하기 위해서 기본적으로 3개의 Node에 저장된다.(설정에 따라 저장되는 Node의 개수는 바뀔 수 있다.)[1]

Hadoop1과 Hadoop2(Hadoop 1.x.x와 2.x.x는 통상 Hadoop1, Hadoop2라 칭한다.)의 경우 Name Node의 개수가 다르다. Hadoop1의 경우 Namenode가 1개 존재하는 반면, Hadoop2의 경우 Namenode가 그림 2처럼 2개 존재한다. 2개의 Namenode중 Standby Namenode가 Active Namenode에 문제 발생시 Zookeeper를 통하여 이를 감지하고 Active Namenode로 Automatic Failover가 이루어진다.

본 연구에서 두개의 Namenode를 사용하여 시스템의 신뢰성과 안정성을 보장하고 새로운 프레임워크(Frame Work)인 YARN을 적용한 Hadoop2를 선택하였다.

3. 데이터 수집

데이터 수집의 단계는 크게 DP서버(Data Portal Server) 수집, 이동, 적재로 나눌 수 있다. DP서버에는 각 게이트웨이에서 계량된 전력 데이터가 저장된다. 이렇게 수집된 데이터를 추후 그림 4에서 보는바와 같이 Hadoop에서 빅 데이터를 분석하기 위해 Hadoop 시스템의 HBase로 옮겨 적재한다. 데이터 이동에는 Sqoop을 사용하며 데이터 수집 단계의 목표는 각 게이트웨이로부터 사용전력에 대한 데이터를 효과적으로 수집하여 미리 구축된 Hadoop 시스템에 안정적으로 전송하고, 분석을 위한 준비를 마치는 것이다.

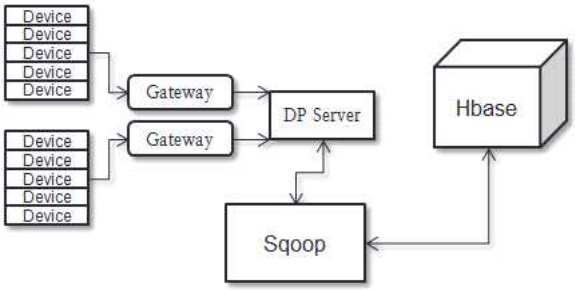

그림 4. 전력 데이터의 흐름

3.1 Data Portal Server

DP서버는 전력을 계량하는 게이트웨이를 모아서 관리하는 서버이며 실시간으로 계량되는 정보와 타임스탬프(Time Stamp)를 확인할 수 있다. 각 게이트웨이에는 하나 이상의 디바이스가 연결될 수 있으며 본 연구에 사용된 게이트웨이 장치는 2개이고 각 게이트웨이마다 5개의 디바이스가 연결되어있다.

DP서버 상에서는 수집된 데이터의 RDBMS로 PostgreSQL가 사용되었다. 계량 값 별로 타임스탬프와 함께 다수의 테이블에 실시간으로 수집 및 저장된다.

3.2 Sqoop을 이용한 데이터 이동

RDBMS에서 Hadoop으로 데이터를 이동하는 도구로는 Sqoop이나 Flume을 사용할 수 있다. Sqoop은 그림 5와 같이 Hadoop과 관계형 데이터베이스 사이에 데이터를 전송하기 위한 도구이다. 관계형 데이터베이스가 제공하는 JDBC를 이용하여 데이터의 양방향 전송이 가능하므로 RDBMS와 HDFS 또는 HBase는 데이터의 양방향 전송이 가능하다. Flume은 다수의 서버에 있는 테이블이나 파일을 읽어올 수 있으며 데이터를 받아오기 위해서는 송신측 서버에 Flume Client를 설치할 필요가 있다.[1]

본 연구에서는 PostgerSQL에서 HBase로 데이터를 전송하는데 DP 서버가 많지 않고, 송신측 서버에 대한 개입을 최소화하기 위해서 Sqoop(Ver. 1.4.4)을 사용하였다.

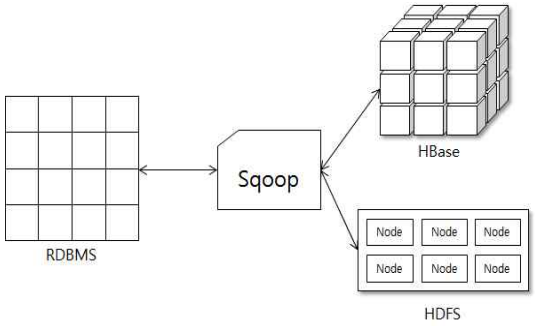

그림 5. Sqoop System Flow

3.3 Hadoop 시스템에 데이터 적재

구성된 Hadoop 시스템에서 분석하기 위한 데이터를 적재하는 방법은 3가지가 있다. 첫 번째로 HDFS에 직접 적재하는 경우, 저장된 데이터는 Mapper와 Reducer를 디자인하여 분석할 수 있다.[2] 두 번째로 데이터웨어하우스인 Hive를 이용하는 경우, Hive SQL언어를 이용해 질의하면 자동적으로 Map과 Reduce 작업을 수행하여 분석할 수 있다.[4] 세 번째로 HBase로 적재하는 경우, Row-key와 Column-key를 기반으로 데이터를 저장하고, 데이터의 확장에 따라 Region을 자동적으로 분할한다. 또한 데이터분석을 위해 Mapper나 Reducer를 직접 설계하지 않아도 된다.[5]

본 연구에서는 HBase를 이용하여 데이터 적재와 분석을 빠르고 용이하게 할 수 있는 환경을 구축하였다.

4. 샘플데이터를 이용한 웹 기반 데모 API

서버에 저장된 전력 데이터를 Client에게 제공을 위하여 웹 기반의 데모 API를 HBase의 라이브러리를 이용하여 자바언어와 자바스크립트언어로 개발하였다.

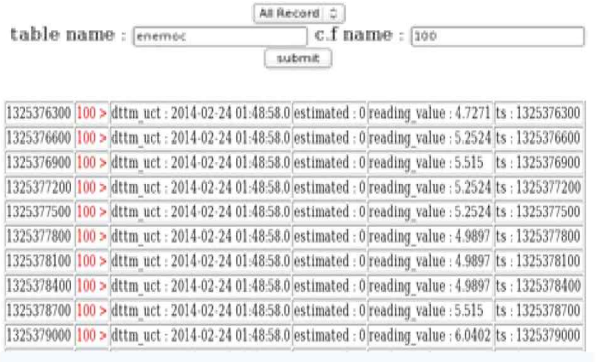

그림 6. 웹 기반의 데모 API

그림 6은 HBase에 저장되어있는 테이블형식의 데이터를 제공해주는 웹 기반의 데모 API이다. 그림6에서 볼 수 있듯이 table name과 c.f name(column family name)을 입력하게 되면 해당 c.f의 모든 데이터를 볼 수 있다.

본 데모 API에서는 전체 테이블의 데이터 또는 c.f name의 데이터, row-key에 해당하는 데이터를 볼 수 있다.

5. 전력 효율성 극대화를 위한 플랫폼

본 연구에서 플랫폼을 구축하기 위해서 자바언어를 사용하였으며 플랫폼의 계층적 구조는 그림 6과 같다.

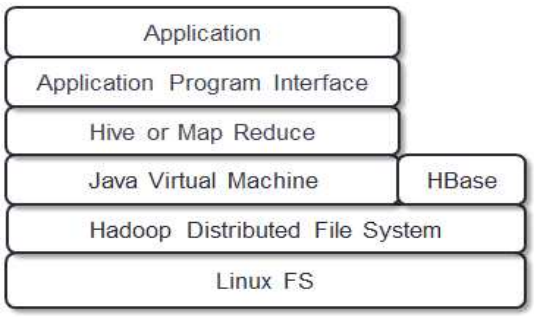

그림 7. 플랫폼의 계층적 구조

리눅스를 기반으로 Hadoop을 설치하고 Hadoop의 HDFS위에 HBase를 설치하였다. Sqoop을 통하여 DP 서버로부터 데이터를 수집하며, 수집된 데이터는 HBase의 테이블 형태로 저장하고 분석한다. JVM을 통해 Hive or Map Reduce와 HDFS을 연동할 것이고 Client는 자바언어로 만들어진 API를 통하여 Hive or Map Reduce에 의해 데이터를 제공받을 수 있다. 이때 제공되는 데이터는 JSON 형식이다.

5. 결론 및 향후 연구

본 연구에서는 빅 데이터 시스템을 이용한 전력 효율화 분석 시스템 구성에 관하여 연구하였다. 전력과 비 전력 데이터를 수집 및 저장하고 분석한 결과를 API를 통하여 Client의 요청 하에 제공한다. 데이터를 Client에게 제공함에 따라 전력 소비 효율성의 극대화를 실현한다.

향후 연구로는 웹 기반의 API를 제공할 뿐 만 아니라 다른 Application을 통해서 데이터를 제공받을 수 있도록 API 개발을 할 계획이다. 또한 Client가 수집된 데이터를 JSON 형태로 제공받을 수 있으며 데이터의 패턴을 분석하기 위하여 데이터 마이닝 기술을 적용할 것이다.

참고문헌

[1] Tom, W. “Hadoop The Definitive Guide, 3rd Edition”, O'Reilly Media, July 2009

[2] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Communications of the ACM, 2008

[3] 온병원, 빅 데이터의 이해와 활용, 한국에너지기술평가원 2012년 12월

[4] Edward, C. “Programming Hive”, O'Reilly Media, November 2012

[5] Lars, G. “HBase: The Definitive Guide”, O'Reilly Media, September 2011

[6] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung, "The Google File System", SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principle Pages 29-43