ETL 상에서 처리속도 향상을 위한 빅데이터 처리 시스템 제안

이정빈*, 박석천** , 길기범*** , 천승태***
*가천대학교 일반대학원 모바일소프트웨어학과
**컴퓨터공학과 정교수(교신저자)
***데이터스트림즈 기술연구소 주임, 책임 연구원
e-mail:leego1622@gmail.com

Suggestion of BigData Processing System for Enhanced Data Processing on ETL

Jung-Been Lee*, Seok-Cheon Park**, Gi-Beom Kil***, Seung-Tea Chun***
*Dept. of Mobile Software, Gachon University
**Dept. of Computer Engineering, Gachon University(Corresponding Author)
***Research Engineer, DATASTREAMS co., ltd

요 약

최근 디지털 정보량의 기하급수적인 증가에 따라 대규모 데이터인 빅데이터가 등장하였다. 빅데이터는 데이터가 실시간으로 매우 빠르게 생성되며 다양한 형태의 데이터를 가지며 이 데이터를 수집, 처리, 분석을 통해 새로운 지식을 창출한다. 그러나 기존의 ETL(Exact/Transform/Load) 연구에서 이러한 빅데이터를 처리 하는데 성능 저하가 발생되고 있으며 비정형 데이터를 관리할 수 없다. 따라서 본 논문에서는 기존의 ETL 처리의 한계를 극복하기 위해서 하둡을 이용하여 ETL 상에서 처리 속도를 높이고 비정형 데이터를 처리할 수 있는 빅데이터 처리 시스템을 제안하고자 한다.

1. 서론

디지털 정보량의 기하급수적인 증가에 따라 대규모 데이터가 중대 이슈로 부각되며 빅데이터(Big Data)라는 용어가 등장하였다. 빅데이터의 정의는 데이터가 실시간으로 매우 빠르게 생성되며 다양한 형태의 데이터를 가지며 이 데이터를 수집, 처리, 분석을 통해 새로운 지식을 창출한다.

기존의 ETL(Extract/Transform/Load) 처리에는 DB를 이용한 방법과 File을 이용한 방법이 존재한다. 하지만, 빅데이터의 비정형 데이터를 처리할 수 없고, 처리 속도면에서 부족한 문제점이 있다[1].

논문에서는 기존의 ETL 처리의 한계를 극복하기 위해서 하둡을 이용하여 ETL 상에서 처리 속도를 높이고 비정형 데이터를 처리할 수 있는 빅데이터 처리 시스템을 제안하고자 한다.

본 논문의 구성은 1장 서론에 이어 2장에서는 이와 관련된 연구들에 대해 살펴보고, 3장에서 하둡을 이용한 빅데이터 처리 시스템을 제안하고, 4장에서 결론을 맺는다.

2. 관련 기술

2.1 데이터 웨어하우스

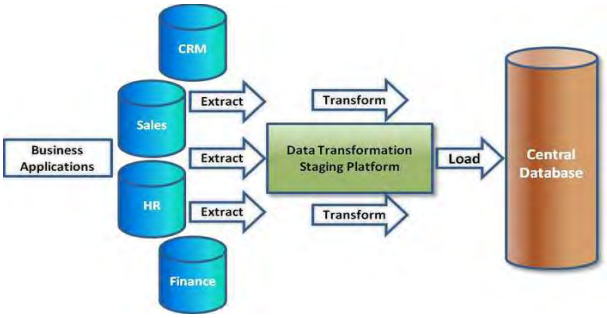
데이터웨어하우스는 운영업무 지원시스템인 OLTP(Online Transaction Process) 시스템에서 생성된 데이터로부터 다양한 분석 정보를 추출하여 의사결정 지원시스템인 OLAP(Online Analysis Process)에 사용하기 위한 데이터 저장고이다[2].

2.2 ETL

ETL은 데이터 웨어하우스(DW: Data Warehouse) 구축 시 데이터를 운영 시스템에서 추출하여 가공(변환, 정제)한 후 데이터 웨어하우스에 적재하는 모든 과정이다. ETL은 데이터 추출(Extraction), 변환(Transformation), 적재(Loading)의 약자이다.

ETL은 소스 시스템으로부터 데이터를 추출(extract)하고, 목적 데이터로 변환(transform)하고, 마지막으로 목표 데이터 마트로 데이터를 적재(load)한다[3].

ETL의 처리 프로세스는 (그림 1)과 같다.



(그림 1) ETL 프로세스

2.3 빅데이터

빅데이터는 데이터가 실시간으로 매우 빠르게 생성되며 다양한 형태의 데이터를 가지며 이 데이터를 수집, 처리, 분석을 통해 새로운 의미 있는 지식을 창출할 수 있는 데이터로 정의할 수 있다[4].

2.4 하둡

하둡(Hadoop)은 대용량 데이터 처리를 위해 거대한 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 오픈 소스 프레임워크이다. 구글파일시스템(GFS)을 벤치마킹하여 하둡분산파일시스템(HDFS:Hadoop Distributed File System)과 맵리듀스(MapReduce)를 구현한 것이다[5].

3. 향상된 빅데이터 처리 시스템 제안

3.1 HDFS의 특징

HDFS(Hadoop Distributed File System)는 하나의 마스터 노드와 여러 개의 슬레이브 노드로 구성된다. 마스터 노드는 파일 시스템 네임스페이스를 관리하고 클러이언트에 의한 파일 접근을 통제하는 단일 네임노드로 구성된다. 각 슬레이브 노드에는 노드에 붙은 스토리지를 관리하는 데이터 노드가 있다. 하둡 파일 시스템은 사용자의 데이터를 파일로 저장한다. 내부적으로 파일은 하나 또는 그 이상의 블록으로 쪼개지며 이러한 블록들은 데이터 노드의 집합 안에 저장된다[6].

3.2 MapReduce의 특징

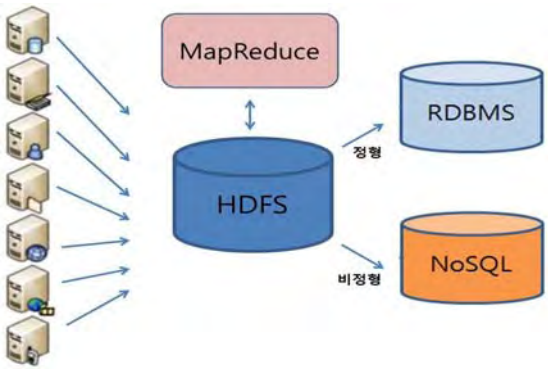
Google에서 개발한 MapReduce는 대용량 데이터를 다수의 서버로 구성된 클러스터에서 병렬 처리하는 연산 프로그래밍 모델이다. 이용자로부터 MapReduce에서 다루도록 할당 된 일을 잡(Job)이라 하며 이를 여러개의 태스크(Task)로 나누어 분산 처리한다. MapReduce는 데이터를 전체 문서를 작은 단위로 나누고, 나누어진 단위에서 데이터를 처리 하고, 처리 된 데이터를 각각 계산하여 합치는 작업이라 한다. 단순히 데이터를 나누어 처리하는 Map과 데이터를 모아주는 Reduce라는 두 개의 기본 데이터 변환 연산에서 나오고 각각의 함수를 구현함으로써 병렬처리하는 환경을 만들 수 있다[7].

3.3 ETL상에서 향상된 빅데이터 처리 시스템

기존의 ETL 시스템들은 정형적인 데이터를 처리하는 것에 중점을 두고 있지만 이후 등장한 페이스북, 트위터 등 SNS의 인기로 비정형 데이터가 생성되며 빅데이터 시대가 도래되었다.

따라서 ETL상에서 하둡을 이용한 빅데이터 처리 시스템은 하둡기반으로 빅데이터를 처리하고 NoSQL DB를 두어 비정형 데이터를 적재하는 시스템을 제안한다.

본 논문에서 제안하는 빅데이터 처리 시스템은 (그림 2)와 같이 각 여러 서버로 들어오는 빅데이터들을 HDFS에서 추출하여 분산 저장하고 MapReduce를 통한 병렬처리로 변환을 한다. 마지막으로 데이터의 형태 정형, 비정형에 따라 RDBMS 및 NoSQL로 분산 적재한다.



(그림 2) 제안하는 빅데이터 처리 시스템

4. 결론 및 향후 연구 방향

본 연구의 핵심은 하둡의 장점인 분산 저장, 병렬 처리를 통한 처리 성능 향상과 NoSQL을 이용한 비정형 데이터의 적재에 목적이 있다. 이를 위해 본 논문에서는 하둡의 HDFS와 MapReduce를 이용하여 기존의 ETL의 처리 성능을 향상시키고 NoSQL를 이용한 비정형 데이터를 적재하는 시스템을 제안하였다.

향후 제안하는 빅데이터 처리 시스템의의 구체적인 설계와 하둡의 HDFS와 MapReduce를 이용한 빅데이터 처리 시스템을 구현하고자 한다.

사사의 글

본 연구는 2015년도 지식 경제부의 SW전문인력양성사업의 재원으로 정보통신산업진흥원의 고용계약형 SW석사과정 지원사업(H0116-15-1003)으로부터 지원받아 수행되었습니다.

참고문헌

[1] 정윤철, “ETL상에서 파일 시스템을 이용한 대용량 ”, 2008

[2] 박종모, “데이터웨어하우스의 개발생상성 향상을 위한 측정지표의 설계 및 분석”, 2007

[3] 이승하, “보안 로그/이벤트 수집을 위한 BigData ETL 모델 설계”, 2014

[4] 김남중, “데이터 품질 기반의 빅데이터 성숙도 모델에 관한 연구”, 2007

[5] 이현종, “빅데이터 하둡 플랫폼의 활용”, 2012

[6] 조성환, “HDFS 기반 동적 데이터 관리를 위한 파일 관리자 설계”, 2013

[7] 김영길, “하둡 기반 빅 데이터 처리 플랫폼 통합 연결 인터페이스”, 2013