빅데이터 분석 시스템을 활용한 온라인 검색 광고 플랫폼 설계 및 개발에 관한 연구

노선택*, 홍승형*, 김경수*, 송영기*, 김환철*
*㈜ 맥퍼트
e-mail : pearl@macpert.com

A Study on Design and Development for Online Search Advertisement Platform using Big Data Analysis System

Seon-Taek Noh*, Seung-Hyung Hong*, Kyung-Soo Kim*, Young-Ki Kim*, Hwan-Cheol Kim*
*Macpert Co., Ltd

요 약

온라인 검색 광고는 인터넷 사용자의 증가, 그리고 온라인 광고 수요의 규모가 커짐에 따라 광고 시장에서 보조적인 역할에서 벗어나 주도적인 위치로 변화하고 있다. 지속적인 규모성장과 수요 증가에도 불구하고 기존의 관계형 데이터베이스에 의존한 온라인 검색 광고 플랫폼은 구조적인 한계로 인해 유연한 자원 확장이나 분석속도의 보장성을 유지할 수 없다. 본 논문에서는 빅데이터 분석 시스템을 이용하여 온라인 검색 광고 플랫폼을 설계 및 구현함으로써, 데이터 저장 공간을 유연하게 확장할 수 있으며, 일정한 시간으로 수렴할 수 있는 안정적인 분석 속도를 유지하는 시스템을 제안한다.

1. 서론

온라인 광고는 글로벌화, 고효율 등의 특성을 가지고 있기 때문에 광고주들을 비롯한 온라인 마케팅 담당자들이 선호하는 매체 중 하나이다.[11,12] 특히 키워드 검색광고는 광고와 밀접한 관계가 있는 키워드를 검색한 사용자에게 노출됨으로써 높은 관심과 잠재적인 구매 고객으로 하여금 효과적인 광고를 할 수 있다는 장점을 가지고 있다.[5]

또한 인터넷 사용자의 증가, 그리고 온라인 광고 시장의 규모가 커짐에 따라 기존 광고의 보조적 역할에서 벗어나 광고산업의 주도적인 역할을 하고 있으므로 효율적인 온라인 검색 광고 플랫폼에 대한 수요는 늘어나고 있다.[1]

하지만 이러한 지속적인 성장에도 불구하고 기존의 관계형 데이터베이스(RDB)와 스토리지에 의존한 온라인 검색 광고 플랫폼은 기하 급수적으로 늘어나는 검색 키워드 정보 및 사용자의 기본 정보를 담고 있는 원천데이터(raw data)와 이를 분석하여 각종 리포트 결과를 산출하는 분석데이터(analysis data)의 양을 수용하기에는 구조적인 한계를 가지고 있다.[2] 따라서, 키워드 검색 광고에서 발생되는 비관계형 데이터를 관계형 데이터로 변환하여 저장해야 하는 시점에서 발생하는 비용적인 문제와 분석 속도가 데이터 량에 따라 비선형적으로 증가하게 되는 문제점이 발생 하게 된다.

본 논문에서는 빅데이터 분석 시스템을 이용하여 저비용으로 구축할 수 있으며, 증가하는 원천데이터 및 분석데이터의 양을 유연하게 수용하여 일정 시점에서 수렴할 수 있는 분석 속도를 산출 할 수 있는 온라인 검색 광고 플랫폼을 제안한다.

2. 관련 연구

2.1 온라인 검색 광고 플랫폼

온라인 광고란 웹을 기술적 기반으로 인터넷상에서 이루어지는 광고의 한 형태로 소리, 문자, 화상 등의 수단으로 인터넷을 매체로 게재, 방영하는 광고의 한 형태라 할 수 있다.[4,12]

온라인 광고의 유형은 크게 이미지, 동영상의 형태로 웹사이트 내에 노출되는 방식인 노출형 광고(Display Ad)와 인터넷 검색 창에 입력한 키워드 검색 결과에 따라 광고가 노출되는 검색 광고(Search Ad)로 구분된다. 특히 검색 광고는 2011 년 기준 광고 집행 추정치가 1 조 2,388 억원으로 같은 해 기준 노출형 광고 집행 추정치 6417 억원으로 두 배 가량의 규모를 나타내고 있다.[1]

검색 광고는 검색키워드를 입력하면 검색결과를 보여주면서 화면 상단에 관련된 배너광고가 나타나는 형태로써 1990 년대 후반부터 도입된 온라인 광고 형태 중 하나이다. 특히 국내 검색광고는 검색 트래픽의 성장과 함께 CPC(cost-per-click) 광고비 과금 체계로 전환함으로써 2003 년 140%의 성장률[1]을 시작으로 꾸준히 규모가 커져가고 있다. 특히, 검색의 특성상 소비자가 검색하는 특정 키워드는 불특정 다수에

게 노출되는 노출형 광고에 비해 소비자의 관심이나 구매의지가 비교적 높기 때문에 광고 효율성에서 큰 장점이 있다. 따라서 온라인 광고에서의 온라인 검색 광고 플랫폼에 대한 중요성은 더욱 높아져 가고 있는 실정이다.

3. 빅데이터 분석 시스템

일반적으로 빅데이터는 저장, 관리 및 분석에 대한 크기가 일반적인 데이터베이스 소프트웨어 도구의 능력을 넘은 데이터 셋을 말한다.[6] 단순히 수천 테라바이트나, 수십 페타 바이트등의 양이 많은 것이 아닌 해당 시스템에서의 활용 빈도에 대한 데이터 양에 대한 관점에서 접근하는 정의가 빅데이터이다.

검색광고에서 50% 가까이 점유율(단독집행 및 동시집행 포함)을 차지하고 있는 네이버는 방문자가 3200 만여명에 달하며, 쇼핑, 경매분야에서 상위 순위에 랭크된 옥션,G 마켓의 경우도 각각 1500 만명의 방문자 트래픽을 발생시키며 각 사이트별 접속후 검색률은 다음과 같다. [9,10]

<표 1> 주요 매체별 중복방문을 제외한 데이터량

	순방문자 (*1000)	검색전환율	검색건수 (*1000)	데이터량 (GB)
네이버	32,172	4.2%	135	1.35
다음	28,918	8.3%	240	2.4
옥션	15,285	12%	183	1.83
G 마켓	15,055	10.9%	164	1.64

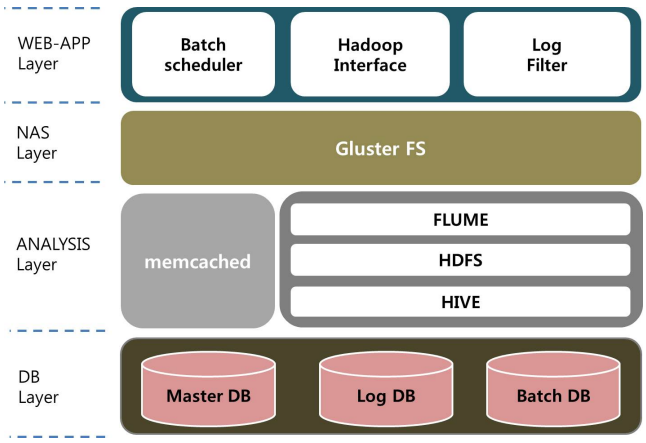
위의 표는 단일 사용자가 한 번 접속 후 한 번 키워드검색을 했을 때의 발생되는 데이터 량을 나타낸다. 이때 발생하는 데이터를 요청 데이터(request data)라고 하며 해당 키워드에 해당하는 광고가 노출되는 데이터를 노출 데이터(impression data)라고 한다. 1 개의 키워드 검색에서 발생되는 데이터 량을 1KB 라고 가정 했을 때, 일반적으로 매체에서 하나의 요청당 노출되는 키워드 광고는 5 개이므로 온라인 검색 광고 플랫폼에서 발생되는 원천데이터는 평균적으로 일일 10GB 가량의 양이 발생된다. 하지만 위의 데이터는 중복되지 않은 방문자 수 별 검색 전환율을 적용한 수치이므로 실제로는 더 많은 검색과 데이터 량이 발생한다. 다음(daum) 매체의 경우 2012 년 3 월 기준 월 검색 쿼리 수는 10 억건을 넘어서며, 월 페이지 방문수(Page Visit)는 137 억건을 넘어선다.[2] 이는 일 70TB 에 달하는 데이터 량이 발생하는 수치이다.

온라인 검색 광고 플랫폼에서 분석 및 광고노출에 필요한 원천데이터 및 분석데이터는 효율적인 분석과 광고 노출을 위해 시간 단위, 일 단위, 월 단위, 연 단위로 체계적인 관리가 필요하며, 유입되는 원천데이터의 양이 증가함에 따라 분석량과 결과 데이터가 기하급수적으로 증가하게 된다. 또한 분석량과 데이터의 양이 증가함에 따라 저장 및 분석에 필요한 자원(resource hardware)을 유연하게 확장할 수 있는 기술이 필요하게 된다.

본 논문에서는 온라인 검색 광고 플랫폼을 위해 저장, 산출, 관리 및 분석 데이터를 다루는 부분을 하둡 에코 시스템의 한 분야인 HDFS, 웹 로그를 실시간으로 HDFS 와 연동하는 FLUME, HDFS 상의 원천데이터를 분석하는 HIVE 기술을 활용하였다. [7,8,11]

4. 온라인 검색 광고 플랫폼 설계

제안하는 온라인 검색 광고 플랫폼은 4 개의 계층으로 구성되어 있으며, 아래의 그림은 검색 광고에 필요한 원천데이터 수집 및 분석에 대한 모듈에 대해 단순화 시킨 모델이다.



(그림 1) 수집/분석 모듈 레이아웃

4.1 웹 어플리케이션 계층

온라인 검색 광고 플랫폼에서 최상위의 계층으로써, 기존 매체 및 광고 사이트와의 API 연동 인터페이스 역할을 한다. 또한 요청 및 노출에 대한 원천 데이터를 필터링하여 NAS 계층에 적재하는 역할을 하며, WAS 단계에 존재하는 계층이다.

수집 및 분석에 대한 일감(job)을 관리하는 배치 스케줄러가 있으며, 외부 매체와 API 통신을 하고, 유효 로그를 필터링하여 HDFS 로 적재할 수 있도록 데이터를 구조화 시키는 로그 필터 모듈이 있다. 마지막으로 빅데이터 분석 시스템으로 구성된 Hadoop 과 연계하는 인터페이스가 이 계층에 속해있다.

4.2 NAS(Network Attached Storage) 계층

온라인 검색 광고 플랫폼에서 생성되는 모든 물리 파일이 저장되고 관리되는 계층이다. 플랫폼 내에 이루어지는 모든 로그 파일이 저장되며, FLUME 을 통해 HDFS 로 적재되는 원천데이터의 정합성을 위해 외부 매체로부터 유입되는 데이터를 LOG 파일 형태로 별도로 저장, 관리한다. NAS 계층은 오픈소스인 Gluster FS 를 사용하여 구성하였으며, 분산-복제모드로 설정되어 있다.

4.3 분석 계층

기존 RDB 에서 join, merge 등의 분석 수행을 했던 부분을 대신하여 분석, 수행하는 계층으로써, 웹 어플

리케이션 계층에서 전달받은 원천데이터를 저장하는 HDFS 모듈, 각 WAS 에서 필터링 된 원천데이터를 HDFS 로 전달하는 FLUME 모듈, HDFS 상에 적재된 원천데이터를 분석하여 통계 결과 및 산출 데이터를 생성하는 HIVE 모듈로 구성되어 있다. HDFS 상의 모든 데이터는 각 클러스터에 구성된 노드에 3 개로 복제되어 저장되며, HIVE 상에서 실행되는 분석 쿼리나 스크립트는 웹 어플리케이션 계층에서 관리되는 일감을 통해서만 구동된다.

또한, HIVE 의 특성상 쿼리 수행이 실시간으로 이루어질 수 없기 때문에 실시간으로 확인, 관리해야 하는 데이터는 memcached 를 이용하여 병합처리 해 관리함으로써 각 작업을 상호 보완적으로 수행할 수 있도록 구성하였다.

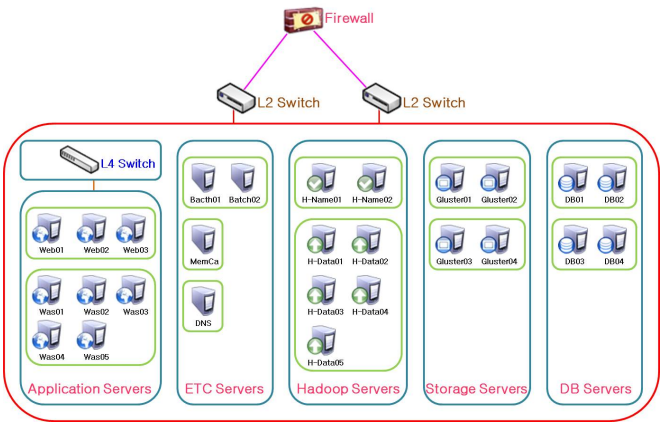
4.4 데이터베이스 계층

비관계형 원천데이터로부터 분석된 결과를 관계형으로 저장하여 관리하는 계층이다. 일반적으로 온라인 검색 광고 플랫폼상에서 가장 많이 차지하고 있는 데이터는 외부 매체에서 유입되는 요청/노출에 대한 수년 또는 십 수년간 적재되어 있는 데이터이다. 일별, 또는 월별, 년 별 등의 해당기간에 대한 분석 결과 데이터는 총합 및 중복제거 된 데이터이므로 상대적으로 양이 매우 적은 수준이다. 따라서 기존 관계형 데이터베이스에서 관리하는 것이 효율적이다. 특히, 각종 결과 리포트나 통계 결과 데이터는 웹이나 어플리케이션을 통해 실시간으로 접근하여 조회되어야 하기 때문에 데이터베이스 계층에서의 관리 필요성이 더욱 크다.

5. 구현 및 성능 평가

5.1 시스템 구현

온라인 검색 광고 플랫폼의 테스트 모듈을 구현하기 위해 27 대의 서버를 구성하였으며 각 모듈당 구성현황은 아래의 그림과 같다.



(그림 2) 온라인 검색 광고 네트워크 구성

어플리케이션 서버군은 웹서버와 WAS 로 구성되어 있으며 각 2.4GHz 쿼드코어 * 2 CPU 와 64GB 메모리로 구성된 서버 8 대를 배치하였으며, 각 WAS 에는 LOG FILTER 가 3 개의 인스턴스로 구동된다.

하둡 서버군은 2.4GHz 쿼드코어 * 2 CPU 와 16GB 메모리로 구성된 7 대를 배치하였으며 2 대는 Name-node 와 secondary-node 로 구성, 5 대는 data-node 로 구성하여 총 용량 120TB 를 구성하였다.

외부 매체에서 전달받은 원천데이터를 HDFS 로 전달하기 위한 FLUME 설정은 각 WAS 별로 해당 작업당 에이전트(agent)를 구별하여 클릭데이터 에이전트 5 개, 요청데이터 에이전트 5 개, 노출데이터 에이전트 5 개를 설정하였다. 컬렉터(collector)는 하둡 data-node 상에 설치하여 총 3 대의 서버에 설정함으로써 결함처리(failover)가 가능하도록 하였다.

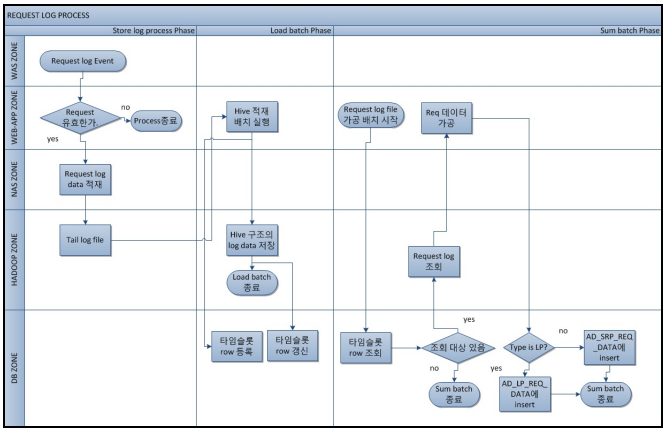
5.2 성능 평가

본 논문에서 제안한 온라인 검색 광고 플랫폼의 성능평가를 위해 두 가지 시나리오를 통해 기존 RDB 에 의존한 플랫폼과 빅데이터 분석 시스템을 도입한 플랫폼의 분석시간을 산출하여 비교하였다.

<표 2> 수행단계 별 비교 시나리오 표

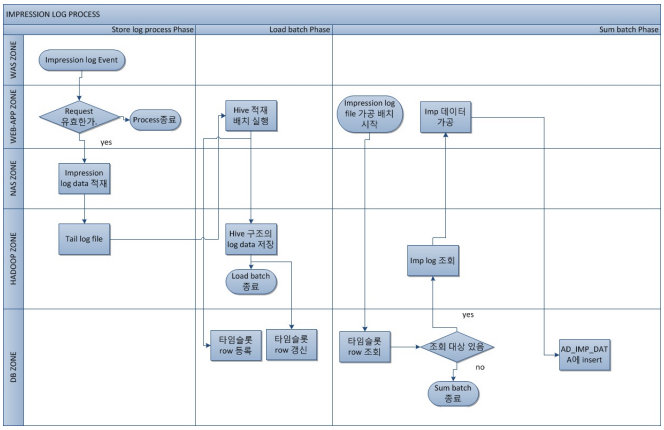
비교 플랫폼	비교 수행단계(A)	비교 수행 단계(B)	분석 시스템
기존 RDB 의존 플랫폼	요청(노출)데이터를 log4j 에 적재한뒤 RDB 로 인서트	rdb 쿼리를 통한 요청(노출)ID 로 중복제거 후 일별 통계 처리	oracle 11g
제안한 플랫폼	요청(노출)데이터를 flume 을 통해 hdfs 로 적재	hive 를 통한 요청(노출)ID 로 중복제거 후 일별 통계 처리	flume+ hive

첫 번째 시나리오는 외부 매체에서 들어오는 요청 데이터를 일별, 주별, 월별로 분석하는 시간을 측정하였다.



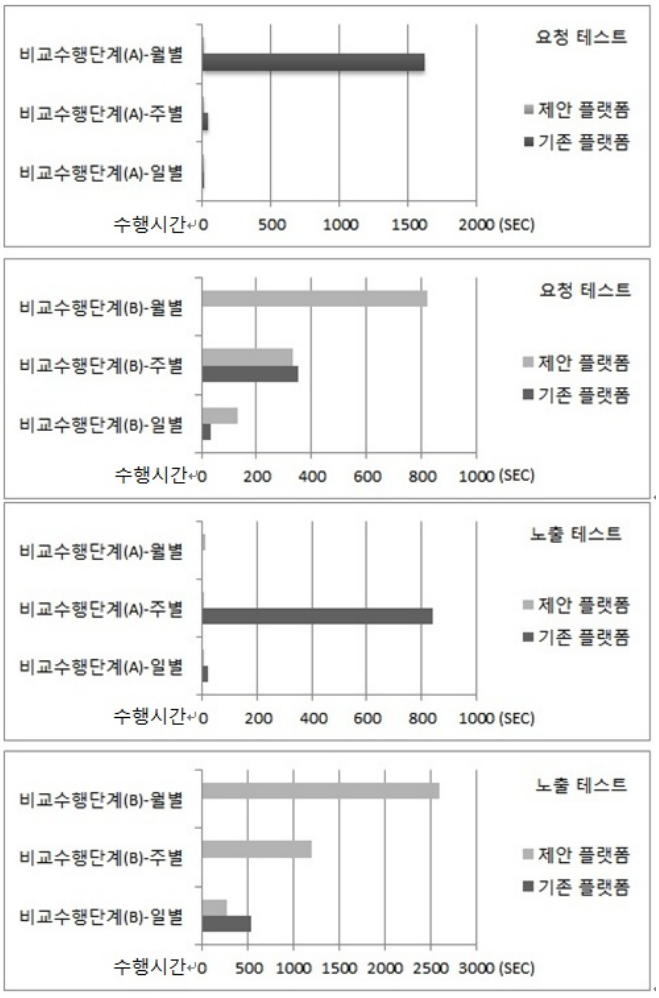
(그림 3) 요청 데이터 분석 엑티비티

두 번째 시나리오는 요청 키워드에 대한 해당 광고를 노출하는 노출 데이터를 일별, 월별, 주별로 분석하는 시간을 측정하였다.



(그림 4) 노출 데이터 분석 엑티비티

실험 데이터는 일일 평균 약 500 만건의 요청 데이터가 유입되며, 해당되는 광고의 노출 수는 요청당 평균 2.5 건으로 가정하였다. 데이터 량은 요청, 노출 1 건당 0.3KB~0.6KB 로 한정하였다. 실험 결과는 다음과 같다.



(그림 5) 요청 및 노출 분석 비교 수행 결과 그래프

위의 그래프에서 요청 및 노출에 대한 기존 플랫폼의 결과는 처리건수가 약 5 천만건 이상이 되면 응답없음의 결과가 나오게 된다. 하지만 HIVE 를 이용한 분석 시스템을 이용하면 1 억 5 천만건의 분석 데이터를 1 시간 이내에 처리된다. 또한 비교수행단계(A)의 경우에는 HIVE 에서 단순히 파일을 HDFS 로 이동하는 속도로 해당 테이블에 데이터를 넣는 효과를 나타내기 때문에 기존 RDB 에서의 인서트처리와는 큰 차이가 남을 알 수 있다.

6. 결론 및 향후 연구

기존 플랫폼은 일일 5 천만건 이하의 일 단위 분석 속도 및 효율에 효과적인 결과를 나타냈지만 5 천만건이상의 주 또는 월 단위의 대용량 데이터 분석에서는 제안한 플랫폼이 비교적 효과적인 분석 결과를 나타냈다. 또한 하둡 에코 시스템의 특성상 기존 컴퓨터(Commodity Hardware)급의 클러스터 구성만으로도 유연한 저장 공간 및 처리 자원을 조절할 수 있으므로 가변적인 온라인 검색 광고 시장에서 더욱 효과적이다.

HIVE 의 병합 및 조인(join) 속도는 기존 RDB 에 의존한 플랫폼에 비해 상당히 느리다는 단점이 존재한다. 일 단위가 아닌 시간 단위, 또는 분 단위의 분석 및 통계 처리 능력은 부적합할 가능성이 높다. 따라서 이를 대처할 수 있는 부분을 보완하여 실시간 분석에도 효과적인 결과를 얻을 수 있는 연구가 필요하다.

참고문헌

[1] "2011 년 온라인 광고 시장 규모", 온라인 광고협회, 2012
[2] "개인 정보 보호와 활용의 조화 방안", 유창하, 2012
[3] "인터넷 광고에 대한 효율성에 관한 연구 : 인터넷 광고 활성화를 위한 업체사례를 중심으로",김석균, 2007
[4] "온라인 광고에 대한 광고대행사와 광고주간 인식의 차이에 관한 연구", 조윤기, 김문정, 2009
[5] "키워드 검색 광고 운영 DB 데이터 분석을 통한 CPM 와 CPC 방식의 광고효과 연구", 김도연, 임규건, 이대철, 2011
[6] "Big data: The next frontier for innovation, competition, and productivity" , McKinsey Global Institute, 2011
[7] "http://hadoop.apache.org", HADOOP
[8] "http://hive.apache.org/", HIVE
[9] "http://koreanclick.com", 코리안클릭
[10] "http://webstatsdomain.com"
[11] "https://cwiki.apache.org/confluence/display/FLUME/Index", FLUME
[12]"Internet Marketing and e-Commerce," Hanson, Ward, and Kalyanam, Kirthi, 2007
[13] "Interactivity and its facets revised, " Johnson, G. J., Bruner, G. C., and Kumar, A., Journal of Advertising, Vol. 35, No. 4, pp. 35-52, 2006.
[14] "Strategic use of information technologies in the tourism industry, ", Buhalis, D., Tourism Management, Vol. 19, No. 5, pp. 409-421, 1998