SQL on Hadoop 기술 동향 및 보안 위협

윤한중 [1], 석상기 [1*]
[1] 서울과학기술대학교 컴퓨터공학과
[1]e-mail : {christ26y, sksuk}@seoultech.ac.kr

Security Threats and Review for SQL on Hadoop

Han Jung Youn[1], Sang Kee Suk[1*]
Dept. of Computer Science and Engineering,
Seoul National University of Science and Technology, Korea

요 약

SQL on Hadoop 기술은 하둡 분산 파일 시스템에 저장된 데이터를 대상으로 SQL 을 이용하여 사용자의 질의를 처리하는 기술이다. 기존의 Hadoop 시스템이 맵리듀스의 한계와 기존 시스템의 호환성으로 인해 RDBMS 와 병행사용이 불가피하다는 단점을 SQL 을 이용해 극복하고자 하는 것이다. 본 논문에서는 SQL on Hadoop 의 대표적 프레임워크인 Hive 와 Impala 의 특징과, 연구동향에 대해 살펴보고 예상되는 보안 위협에 대해 고찰한다.

1. 서론

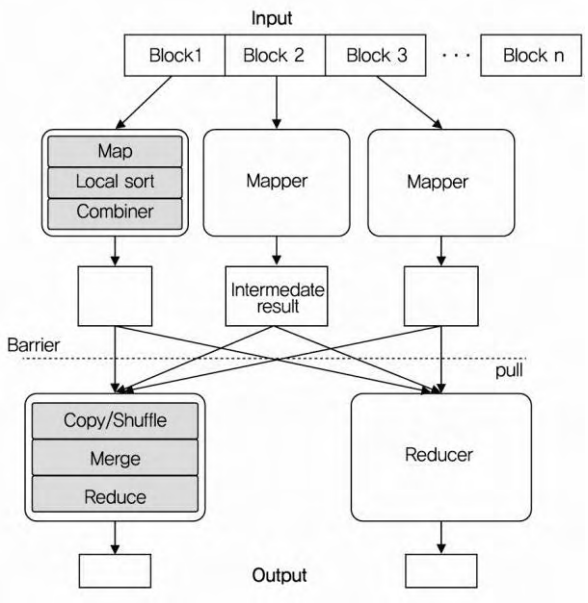
최근 하둡 분산 파일 시스템(HDFS)을 이용한 데이터 처리 기법이 빠르게 성장하고 있다. HDFS 는 수백, 수천의 컴퓨터를 이용하여 데이터를 분산 처리함으로써 기존의 관계형 데이터베이스 시스템(RDBMS)에 비해 대용량의 데이터를 처리할 수 있으며, 데이터의 양에 따라 유연성 있게 자원을 활용할 수 있다는 장점이 있다. 그러나 맵리듀스 기법이 실시간 데이터 처리에 있어서 RDBMS 에 비해 상대적으로 낮은 성능을 보이고 여러 한계를 보임으로 인해 기업들을 중심으로 다양한 연구들이 진행되고 있다. 그 중 SQL on Hadoop 기술은 HDFS 에 저장된 대량의 데이터를 대상으로 사용자에게 익숙한 SQL 을 이용하여 질의를 실시간으로 처리하고자 하는 기술이다. SQL on Hadoop(SoH) 기술은 이미 데이터분야의 표준 언어라고 할 수 있는 SQL 언어를 Hadoop 환경에서도 사용할 수 있다면 이미 축적되어있는 대량의 데이터환경을 포함할 수 있을 것이라는 생각을 기반으로, SQL 을 처리하는 속도를 향상시켜 Hadoop 의 성능적인 만족도 높이고자 하는 최신 기술의 총칭이다.

본 논문에서는 SoH 관련 기술인 MapReduce 와 SoH 의 대표적인 시스템인 Hive, Impala 에 대해 살펴보고 SoH 환경에서의 보안위협요소 및 정보보안의 중요성에 대해 고찰한다.

2. SQL on Hadoop 기술 및 배경 지식

2.1 Map Reduce

맵리듀스(MapReduce)는 Map 과 Reduce 함수를 직접 작성해 대량의 데이터를 처리하는 분산프로그래밍 모델이다. 맵리듀스의 입출력에는 블록 기반 분산 파일 시스템을 이용한다. 파일들은 기본적으로 64MB 의 블록 단위로 로컬 환경에 관리되고, 각 블록들은 내고장성의 지원을 위해 2 개의 복사본을 갖는다.



(그림 1) MapReduce 의 시스템 구조[1]

맵리듀스는 분산형 RDBMS 와 비교하여 모델이 단순하고 개발이 편리한 장점이 있다. 또한 데이터모델이나 스키마의 정의를 요구하지 않으므로 Text 형식과 같은 데이터도 쉽게 다룰 수 있다. 내고장성과 확장성은 맵리듀스의 가장 큰 특징으로, 맵리듀스 작업 중 장애가 발생해도 작업이 종료되지 않고 지속적으

*교신저자: 석상기(서울과학기술대학교)

로 수행, 완료된다.

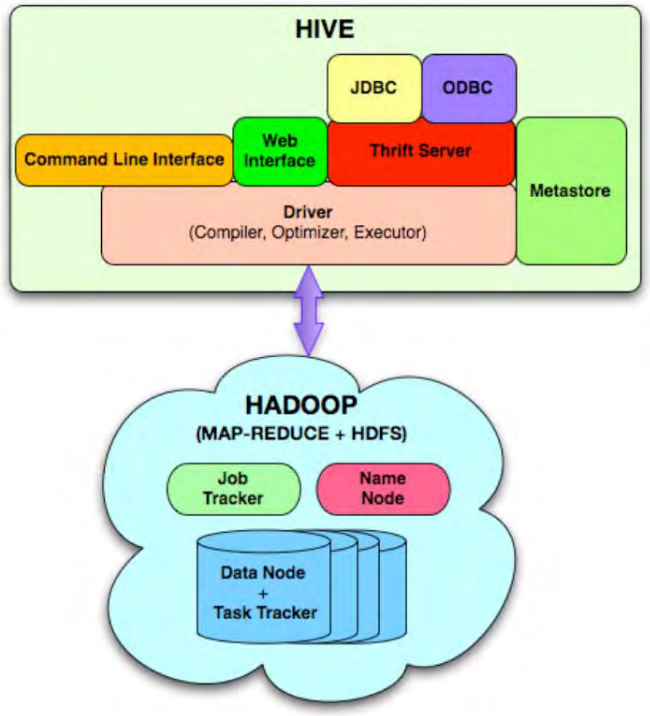
그러나 맵리듀스는 SQL 과 같은 고차원 언어와 스키마를 지원하지 않기 때문에 Join 이나 Loop 문과 같은 복잡한 알고리즘을 구현하기 어렵다. 또한 내고장성 확보를 위해 지속적으로 많은 디스크와 네트워크 입출력을 발생시키고, 데이터의 블록화로 인해 파이프라인 병렬화가 불가능해져 비효율적인 측면이 존재한다[2].

맵리듀스가 구현된 대표적인 시스템이 하둡(Hadoop)으로, 현재 아파치 재단에서 개발중인 공개 소프트웨어이다.

2.2 Apache Hive

하이브(Hive)는 페이스북에서 개발된 SQL on Hadoop 솔루션이다. Hadoop 상에서 동작하는 데이터 웨어하우스 도구로서, 데이터 요약, 비정형질의 및 분석기능을 제공하며 맵리듀스의 모든 기능을 지원한다. 하이브는 HiveQL 이라는 유사 SQL 을 이용해 질의를 지원하는데, 사용자가 작성한 HiveQL 질의는 Hive Driver 에 의해 맵리듀스 Job 으로 변환되어 HDFS 상에서 실행된다.

하이브는 기본적으로 명령어 기반 인터페이스이지만 ODBC 와 JDBC 를 지원하며 웹 GUI 또한지원하는 사용자 친화적인 시스템이다. 테이블 스키마의 정보와 같은 메타데이터는 Metastore 라는 이름으로 RDBMS 상에 메타데이터를 저장해 데이터의 조회가 수월하다[4].

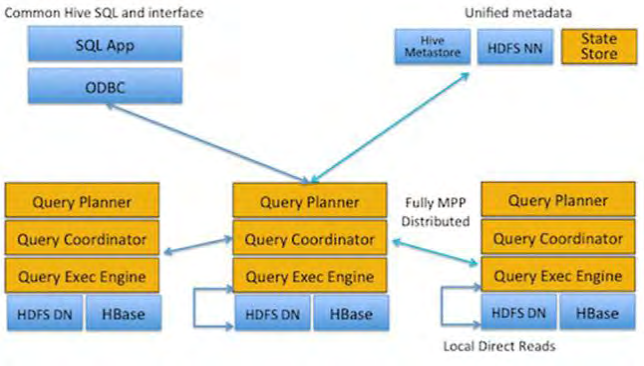


(그림 2) Hive 의 구조도[5]

2.3 Cloudera Impala

임팔라(Impala)는 HDFS 에 기반한 하나의 분산처리 도구이다. HiveQL 을 인터페이스로 사용하지만, 질의 응답시간을 줄이기 위해서 맵리듀스 프레임워크를 이용하지 않는다. 각각의 데이터 노드에 Impalad 라는 프로세스가 실행되며, Impalad 는 질의에 대한 계획을 세우고 실행하며 노드 사이의 값을 공유한다[6].

임팔라는 모든 질의를 파이프라인으로 실행하기 때문에 HDFS 상의 데이터 입/출력에 있어서 좋은 성능을 보인다. 일반적으로 임팔라는 하이브에 비해 2 배 이상 빠른 반응속도를 보이며 이는 하이브가 기본적으로 맵리듀스를 이용하며 오버헤드가 발생하는 설계적인 한계 때문이다.



(그림 3) Impala 의 구조도[7]

3. 기존 기술과 SQL on Hadoop 기술의 비교

SQL on Hadoop 기술은 하둡, 또는 HDFS 에 기반하여 맵리듀스 기술의 단점을 보완하고자 개발된 기술이다. <표 1>은 맵리듀스와 하이브, 임팔라의 특징을 비교한 것이다.

<표 1> MapReduce, Hive, Impala 의 비교

	MapReduce (Hadoop)	Hive	Impala
High-level language support	No	Yes (HiveQL)	Yes (HiveQL)
Data modeling	None	Table	Table
Scalability	Yes	Yes	Yes
I/O Efficiency	Low	Low	Better
Fault tolerance	Yes	Yes	No
Base Language	Java	Java	C++
Low Latency	No	Midium	Yes
Batch	Yes	Yes	Realtime
Purpose	Data Warehouse	Data Warehouse	Query Engine

4. SQL on Hadoop 정보보호 고려사항

SoH 환경에서는 대량의 데이터를 수집하고 분산저장, 분석 및 2 차 데이터 생성 등의 과정에 따라 다양한 보안 이슈가 존재한다. 본 장에서는 데이터를 저장하고 운영할 때 필요한 보안 문제에 대해 고려한다.

4.1 사용자 접근제어

대량의 데이터가 인증되지 않은 공격자에게 노출될 경우 치명적인 정보 유출위험이 존재하기 때문에 사용자의 신원 확인과 권한에 따른 접근 제어가 반드시 필요하다. 기업 내에서 인트라넷으로 제공되는 경우 직급에 따른 권한으로 사용자의 접근제어를 할 수 있겠으나, 서비스가 불특정 다수를 대상으로 제공될 경우 의도적인 공격에 쉽게 데이터가 노출될 수 있다.

4.2 데이터 무결성

SoH 환경에서 데이터는 여러 개의 노드에 분산 처리되고 병렬 저장되므로, 분산된 데이터의 무결성이 보장되어야 한다. SoH 환경은 2 차 데이터의 생성과 분석이 주된 목적이기 때문에 2 차 데이터에 대해 신뢰성을 제공하기 위해서는 분산된 데이터의 무결성이 요구된다.

4.3 데이터 가용성 및 복구

SoH 환경은 데이터 노드마다 파일이 분산되어 저장되므로, 단일 장애점(Sing point of Failure, SPOF) 문제가 취약하다. 이런 문제점을 해결하기 위해 HDFS 상에서 중요한 데이터를 저장한 데이터 노드는 백업 노드를 가지게 하고, 저장장치의 여유공간을 충분히 확보해야 한다.

4.4 데이터 기밀성

데이터의 기밀성 확보를 위해 중요 데이터는 암호화가 필수적이다. SQL on Hadoop 환경은 대량의 데이터를 실시간으로 처리하는 것이 주 목적이므로 데이터의 가용성이 떨어지지 않도록 경량화한 AES, DES 등의 적절한 암호화 기법을 적용해야 한다[8].

4.5 네트워크 및 웹 보안

SoH 환경은 기본적으로 네트워크에 기반하기 때문에 네트워크 환경의 보안 취약성을 가지고 있다. 악의적인 공격을 방지하기 위해 IDS(Intrusion Detection System), IPS(Intrusion Prevention System), 방화벽, https 등의 보안 기술을 분산처리 환경에 맞게 적용해야 한다[9].

5. 결론 및 고찰

본 논문에서는 현재 가장 대표적인 SQL on Hadoop 기술과 그에 대한 보안 고려사항에 대해 논의하였다.

SoH 기술은 앞으로 데이터 웨어하우징 및 빅데이터 시장에서 주류로 성장할 가능성이 막대한 시스템이지만 보안적으로 취약한 부분이 존재한다. 이는 SoH 기술의 적극적인 상용화에 큰 장애물이다. 사용자에게 쾌적한 데이터 분산처리 시스템을 제공하기 위해 지속적으로 SoH 보안 기술의 연구가 요구된다.

참고문헌

[1] Lee, Kyong-Ha, et al. "Parallel data processing with MapReduce: a survey." AcM sIGMoD Record 40.4 (2012): 11-20.

[2] 이경하, et al. “대규모 데이터 분석을 위한 MapReduce 기술의 연구 동향” 전자통신동향분석 제 28 권 제 6 호 (2013)

[3] 김종욱. "" SQL on Hadoop" 기술 동향." 한국멀티미디어학회지 18.1 (2014): 1-7.

[4] Kumar, Rakesh, et al. "Comparison of SQL with HiveQL." International Journal for Research in Technological Studies 1.9 (2014): 2348-1439.

[5] Thusoo, Ashish, et al. "Hive-a petabyte scale data warehouse using hadoop." Data Engineering (ICDE), 2010 IEEE 26th International Conference on. IEEE, (2010).

[6] Floratou, Avrilia, Umar Farooq Minhas, and Fatma Ozcan. "Sql-on-hadoop: Full circle back to shared-nothing database architectures." Proceedings of the VLDB Endowment 7.12 (2014).

[7] http://blog.cloudera.com/blog/2012/10/cloudera-impala-real-time-queries-in-apache-hadoop-for-real/, “Cloudera Impala: Real-Time Queries in Apache Hadoop, For Real”

[8] Lin, Hsiao-Ying, et al. "Toward data confidentiality via integrating hybrid encryption schemes and Hadoop distributed file system." Advanced Information Networking and Applications (AINA), 2012 IEEE 26th International Conference on. IEEE, (2012).

[9] 정교일, et al. "빅데이터와 정보보안." 한국정보기술학회지 10.3 (2012): 17-22.