빅데이터 도구 트렌드 및 긍·부정적 인식 결정 요소 조사

이명진, 구자환, 김응모
성균관대학교 소프트웨어대학
zj1081923@skku.edu, jhkoo@skku.edu, ukim@skku.edu

A Survey on Trend and Factor Determining Positive and Negative Recognition for Big Data Tools

Myungjin Lee, Jahwan Koo, Ung-Mo Kim
College of Software, Sungkyunkwan University

요 약

디지털 기술의 발전으로 데이터의 규모와 형태의 다양성이 기하급수적으로 증가하고 있다. 많은 업계에서 빅데이터를 비즈니스와 사용자의 서비스 제공에 사용하고 있으며, 데이터의 중요성 또한 커지고 있다. 본 연구에서는 빅데이터를 처리하기 위한 단계를 수집, 저장, 그리고 처리 및 분석 단계로 나눈 후, 단계별로 가장 높은 관심도를 가진 도구를 선정하고, 소프트웨어 리뷰 분석을 통해 긍부정 인식을 판단하며 인식 결정 요인을 조사한다. 이를 통해 다양한 빅데이터 생태계 속에서 사용자들이 관심을 많이 두고 있는 빅데이터 도구의 트렌드를 쉽게 파악하고 관련 빅데이터 도구를 선택하는 데에 도움을 줄 수 있다.

1. 서 론

4차 산업 혁명으로 인해 다양한 디지털 기기들이 일반 대중들에게 보급되고 있다. 데이터의 양과 그 종류는 PC, 스마트폰, SNS, IoT 기기, 센서 등 다양한 전자 기기의 대중화에 따라 기하급수적인 속도로 증가하고, 다양화되어가고 있다. 최근의 플랫폼과 서비스는 데이터의 축적을 통해 양질의 경험을 이용자에게 준다. 과거부터 현재까지 페이스북에 업로드된 사진은 약 400억 개이며 전체 데이터의 규모는 하루에만 500테라바이트에 달한다[1].

이와 같은 흐름 속에서 새로운 데이터의 출현과 함께 방대한 양의 데이터를 처리하고 분석할 수 있는 기술 또한 함께 등장했다. 비정형 데이터들을 저장하고 분류하는 다양한 수집·저장 소프트웨어, 저장된 데이터들을 분석할 수 있는 분석기법 등 여러 가지 기술과 플랫폼이 개발되고 있다. 빅데이터 관련 시장 규모가 연간 1,000억 달러를 넘어서며 해마다 10%씩 성장할 것으로 전망된다. 이 수치는 소프트웨어 산업 전체의 거의 2배에 가깝다.

빅데이터의 양은 굉장히 빠른 속도로 증가하며, 과거에는 없었던 다양한 형태의 데이터로 수집된다. 이 방대한 양의 데이터로부터 유의미한 정보를 제공해야 한다. 이 모든 과정은 다양한 빅데이터 도구들이 처리하고, 분석하고, 관리한다. 빅데이터를 가공하는 프로세스는 수집, 저장, 처리, 분석, 시각화의 단계를 거친다. 각각의 단계에서는 그 단계에 특화된 소프트웨어 도구가 존재한다. 예를 들어 데이터 수집 단계에는 Flume, Sqoop, Chukwa, Nutch 등, 데이터를 처리하는 데에는 Kafka, Storm, Spark 등이 있다.

다양한 도구들이 존재하는 만큼 각각이 가진 특성들도 다르고, 그로 인해 사용자들이 선호하는, 혹은 선호하지 않는 도구들도 있을 것이다. 각 단계에서 소프트웨어끼리의 비교에 대한 분석은 NoSQL 데이터베이스 성능 평가에 관한 연구[2]와 빅데이터 처리를 위한 도구를 비교한 연구[3]와 같이 이전에 관련된 연구가 존재한다. 첫 번째 연구에서는 Hbase, Cassandra, MongoDB, Redis 네 개의 NoSQL 데이터베이스의 성능을 비교했다. 두 번째 연구에서는 Computing tools인 Hadoop, Cloudera Impala RTQ, IBM Netezza, Apache Giraph를 다양한 항목에서 비교 분석했다. 또한, Storage tools인 Hbase,

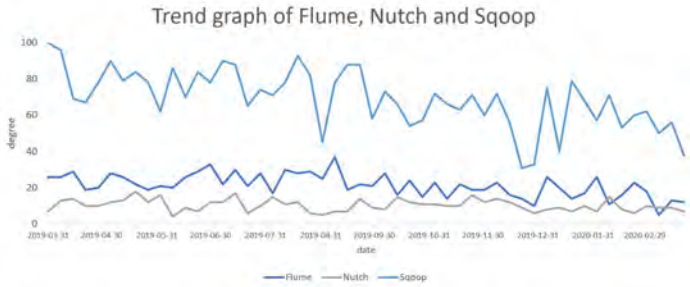
Apache Hive, Cassandra, Neo4j도 마찬가지로 다양한 항목에서 비교 및 분석했다.

그러나 소프트웨어 이용자들의 특정 소프트웨어 관심도와 트렌드에 대해서는 아직 구체적으로 연구된 바가 없다. 따라서 본 연구에서는 빅데이터 가공 단계를 수집, 저장, 처리 및 분석 이렇게 세 단계로 나누고, 각 단계에서의 트렌드를 분석하며 선정된 소프트웨어의 특징 및 장단점을 분석해보고자 한다.

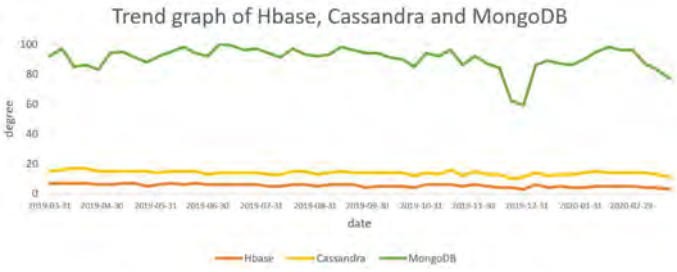
2. 사용자 관심도 분석

수집 단계에서는 Apache Flume, Nutch, Sqoop를, 저장 단계에서는 HBase, Cassandra, MongoDB를, 처리 및 분석 단계에서는 Apache Kafka, Spark, Hadoop 총 9개의 소프트웨어를 선정했다. 관심도는 구글 트렌드를 통해 확인했다. 기간은 모두 동일하게 2019년 3월 31일부터 2020년 3월 22일로 고정했다. 그래프의 y축 degree는 검색량이 가장 높은 지점의 검색 관심도를 100으로 해 나머지 값들을 계산한 결과이다.

그림 1의 그래프에서 볼 수 있듯이, 조사 기간 동안 Sqoop이 나머지 두 도구들보다 우위를 차지하고 있다.



(그림 1) Trend graph of Flume, Nutch and Sqoop

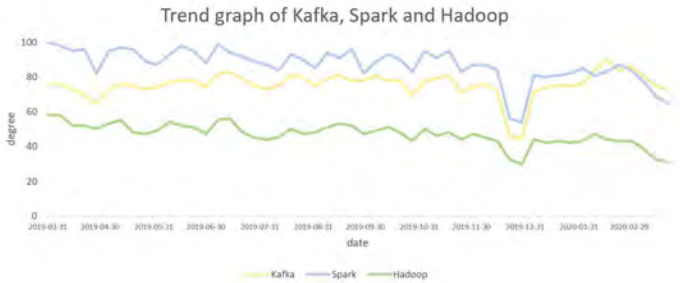


(그림 2) Trend graph of Hbase, Cassandra and MongoDB

그림 2의 그래프는 Hbase, Cassandra, MongoDB 중 MongoDB가 압도적인 관심도를 가지고 있는 것을 보여준다.

그림 3에서는 세 도구들 중 Spark의 관심도가 가장 높고 Kafka, Hadoop이 그 뒤를 잇고 있다.

이와 같은 결과로 말미암아 각 단계에서 Apache Sqoop, MongoDB, Apache Spark 세 도구를 정해 리뷰를 분석한다.



(그림 3) Trend graph of Kafka, Spark and Hadoop

3. 소프트웨어 리뷰 수집 및 분석

Apache Sqoop은 소프트웨어 리뷰 사이트인 'G2 Crowd', 'TrustRadis'에서, MongoDB는 'G2 Crowd', 'Capterra'에서, Apache Spark는 'G2 Crowd', 'Capterra', 'TrustRadius' 사이트에서 크롤링하여 데이터를 수집했다.

유효한 단어란 positive, negative, 혹은 neutral word로 분류될 수 있는 단어이다. 감성 분석을 통해 해당 소프트웨어에 대한 사용자의 반응이 긍정인지, 혹은 부정인지 판단했다. 분석을 위한 감성사전으로는 SentiWordNet[4]을 이용했다.

<표 1> 리뷰 데이터에 대한 감성분석 결과

Apache Sqoop	총 단어 수	2368
	유효한 단어 수	904
	positive score	0.60
	negative score	0.40
MongoDB	총 단어 수	70458
	유효한 단어 수	26360
	positive score	0.61
	negative score	0.39
Apache Spark	총 단어 수	7369
	유효한 단어 수	2890
	positive score	0.57
	negative score	0.43

표 1의 positive score과 negative score은 감성분석 결과 총 리뷰의 긍정정도, 부정정도를 점수화하여 백분율로 나타낸 값이다. 중립성을 띠는 단어는 제외했다.

표 1에서 Apache Sqoop은 긍·부정도가 각각 약 60%, 40%로 긍정적인 반응이 훨씬 크다는 것을 알 수 있다. 마찬가지로 MongoDB는 긍·부정이 각각 61%, 39%, 그리고 Apache Spark 역시 부정적인 반응보다 긍정적인 반응이 더 크다는 것을 알 수 있다.

이 결과를 통해 사용자들이 세 소프트웨어에 대해 만족감을 느끼고 있음을 알 수 있다. 그러나 긍·부정도가 극적인 차이를 보이지 않기 때문에 각 소프트웨어에 대한 특징과 강점뿐만 아니라 취약점 또한

조사했다.

3.1 Apache Sqoop 분석

Apache Sqoop은 Hadoop과 같은 관계형 데이터베이스처럼 구조화된 데이터 저장소 간에, 대량 데이터를 효율적으로 전송하도록 설계된 도구이다.[5] 맵리듀스를 기반으로 구현된 데이터 적재 프로그램이며, MySQL, Oracle 등의 관계형 데이터베이스와 Hadoop 파일 시스템 간 손쉬운 데이터 적재가 가능하기 때문에 널리 사용된다. 또한 데이터 전송을 병렬화해 빠른 성능을 보장한다. 그러나 Apache Sqoop은 command line 인터페이스로, GUI를 제공하지 않는다.[6-7]

RDBMS를 기반으로 하는 많은 어플리케이션 시스템을 운영하거나, 빠른 성능이 필요할 때 데이터 전송을 분할하고 병렬화할 수 있는 Sqoop은 좋은 솔루션이 될 수 있다. 그러나 event driven data를 다뤄야 하거나, 원천이 되는 데이터 저장소에 큰 부담이 가해져서는 안 되는 경우, 대량의 데이터 전송에 이용되는 Sqoop의 사용은 바람직하지 않다.[8]

3.2 MongoDB 분석

MongoDB는 높은 성능과 확장성을 가지고 있는 문서 기반 방식의 NoSQL DBMS이다. 가장 큰 특징으로는 RDBMS와는 다르게 고정된 스키마가 존재하지 않으며 JSON 형태의 문서로 저장하는데, 이를 BSON 형식이라고 부른다. key-value를 통해 복잡한 쿼리가 가능하고, 관계형 데이터베이스보다 응답속도가 빠르며, 인덱스 추가를 통해 처리 속도를 더 빠르게 할 수 있다. 그러나 복잡한 join이나 트랜잭션 처리에 제약이 있다는 단점이 있다. 또한 디스크에 쓰기가 비동기식으로 이루어져 데이터가 소실될 위험이 있다.

스키마가 존재하지 않기 때문에 데이터 모델의 변경, 추가, 확장이 비교적 쉽다. BSON 구조를 사용하므로 데이터를 직관적으로 파악할 수 있어 가독성이 높다. 그리고 native sharding을 지원하기 때문에 확장이 RDBMS보다 간단하다.[9]

IDC의 추정으로는, 데이터의 90%는 사전 정의된 데이터 모델이 없는 비정형 데이터이다. 이러한 종류의 데이터를 저장하기 위해서 사용자들은 다른 데이터베이스 저장 도구보다 MongoDB를 편리하게 선택한다. 방대한 양의 빠른 데이터 처리를 원하는 사용자라면 MongoDB를 사용하는 게 효율적이지만 높은 transactional application을 사용한다면, 혹은 은행과 같이 데이터 소실이 치명적인 결과를 초래하는 시스템에서 MongoDB는 좋은 선택이 아닐 수 있다.

3.3 Apache Spark 분석

Apache Spark는 대규모 데이터 처리를 위한 통합 분석 엔진이다. DAG 스케줄러, 쿼리 최적화 프로그램 및 물리적 실행 엔진을 사용해 배치·스트리밍 데이터 모두에 높은 성능을 보여준다. 예를 들어, Logistic regression에서 Spark는 Hadoop보다 running time이 약 100배 빠르다. 또한 SQL, DataFrames, 머신 러닝을 위한 MLlib, Graph X, Spark Streaming 등을 포함한 라이브러리를 제공한다. Spark API는 개발자 친화적이며, 분산 처리 엔진의 복잡성을 간단한 메소드 호출로 가린다. 인메모리 컴퓨팅을 지원해 Hadoop과 같은 디스크 기반 엔진과 비교해 훨씬 빠르게 데이터를 쿼리할 수 있다. 그러나 Apache Spark는 현재의 모든 작업을 메모리 내에서 유지하는데, 이는 메모리 리소스의 부족을 야기한다.[10-11]

Spark는 Hadoop의 MapReduce 코드보다 훨씬 간결하고, 다양한 언어를 위한 API로 쉽게 작성할 수 있다. 또한 배치 프로그래밍과 인메모리 컴퓨팅에서 최소 10배, 최대 100배의 성능을 보이는 Spark는 사용자 입장에서 매우 매력적으로 보인다.

큰 빅데이터를 빠른 속도로 처리하려는 사용자는 Spark가 적절한 도구가 될 것이다. 그러나 비교적 작은 크기의 데이터를 처리하거나, 혹은 사용 가능한 메모리가 제한적이라면 Spark보단 Hadoop이 좋은 선택이 될 수 있다.

4. 결론

빅데이터 처리 과정을 수집, 저장, 처리 및 분석 세 단계로 나누어 각각의 단계에 특화된 소프트웨어를 선정했다. 각각의 단계에서 Apache Sqoop, MongoDB, Apache Spark가 사용자들에게 가장 높은 관심도를 가지고 있었다. 소프트웨어 리뷰 사이트에서 리뷰 데이터를 분석했고, 세 도구 모두 부정적인 반응보다 긍정적인 반응이 많았다. 따라서 Apache Sqoop, MongoDB, Apache Spark 모두 사용자들에게 편의성을 제공하고, 사용자에게 매력적인 소프트웨어라고 결론 내릴 수 있다. 그러나 긍정도와 부정도의 차이가 크지 않기 때문에 강점뿐만 아니라 취약점 또한 조사했다. 이를 통해 각 도구의 특징 및 장단점을 분석하고 소프트웨어 특성에 따른 적합한 사용자의 유형 또한 제안했다.

본 연구를 통해 언급된 빅데이터 도구들의 현재 트렌드가 어떠한지 살펴보고, 빅데이터 처리 각 단계에서의 도구 선택에 도움을 줄 수 있을 것이다.

참 고 문 헌

[1] 이긍희 외 4인, “빅데이터의 이해”, 한국방송통신대학교 출판문화원, 2016.

[2] 박홍진, “다양한 NoSQL 데이터베이스의 성능 평가 연구”, 한국정보전자통신기술학회논문지, Vol.9, No.3, pp.298-305, 2016.

[3] Bakshi Rohit Prasad and Sonali Agarwal, “Comparative Study of Big Data Computing and Storage Tools: A Review”, International

Journal of Database Theory and Application, Vol.9, No.1, pp.45-66, 2016.
[4] Text Learning Group, http://sentiwordnet.isti.cnr.it/
[5] Apache Sqoop, https://sqoop.apache.org/
[6] 진고환, “하둡 분산 환경 기반의 데이터 수집 기법 연구”, Journal of the Korea Convergence Society, Vol.7, No.5, pp.1-6, 2016.
[7] Varsha B.Bobade, “Survey Paper on Big Data and Hadoop”, International Research Journal of Engineering and Technology(IRJET), Vol.3, No.1, pp.861-863, 2016.
[8] Tomcy John, Pankaj Misra, “Data Lake for Enterprises”, Packt, 2017.
[9] MongoDB, https://www.mongodb.com/
[10] Apache Spark, https://spark.apache.org/
[11] Abdul Ghaffar Shoro, Tariq Rahim Soomro, “Big Data Analysis: Ap Spark Perspective”, Global Journal of Computer Science and Technology (C), Vol.15, No.1, pp.7-14, 2015.