클라우드 컴퓨팅 환경에서의 학습용 빅 데이터 플랫폼 설계

김준헌†
† 성균관대학교 전자전기컴퓨터공학과

Big Data Platform for Learning in Cloud Computing Environment

Jun Heon Kim†
† Dept. of Electrical and Computer Engineering, Sungkyunkwan University

요 약

정보 기술의 끊임없는 발전에 따라 광범위한 분야에서 방대한 양의 데이터가 발생하게 되면서 이를 처리하기 위한 빅 데이터에 대한 연구 및 교육이 활발히 진행되고 있다. 이를 위하여 데이터 분석 및 처리를 위한 고성능의 서버 및 분산 처리를 위한 다수의 컴퓨터가 필요하며 이는, 개인 혹은 저사양의 수업 환경에서 빅 데이터를 학습하는 데에 어려움을 겪게 한다. 때문에 가상 환경에서 원활한 빅 데이터 학습을 위한 클라우드 기반의 시스템이 필요하다. 이에 본 논문에서는, 빅 데이터 처리 기술의 하나인 Spark를 이용한 빅 데이터 플랫폼 구축에 대하여 기술한다.

1. 서 론

최근 모바일 시장의 규모가 커지고 전 세계적으로 소셜 네트워크가 활성화됨에 따라 기업 혹은 기관에서 처리해야 하는 데이터의 양이 급속도로 증가하게 되었다. 이로 인하여 각 조직에서는 쏟아지는 데이터를 분석 및 관리하여 가치를 창출해내는 빅 데이터에 대한 수요가 높아지고 있다[1].

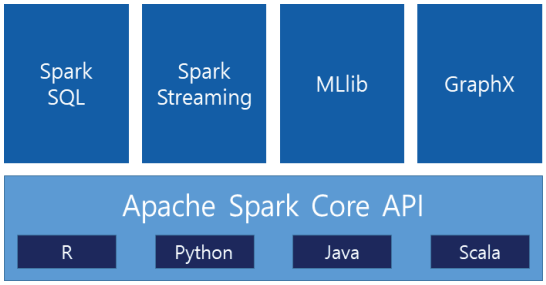
빅 데이터란 기존의 데이터 관리 도구로 데이터를 처리할 수 없는 대량의 비정형 데이터를 관리하는 기술로써, 실제로 이를 수행하기 위하여 다수의 고성능 컴퓨팅 환경이 필요하다[2]. 하지만 실제 교육 환경에서 모든 빅 데이터 처리 환경을 구축하여 사용하는 것은 많은 어려움이 따른다.

이에 본 논문에서는, 교육 환경에서의 클라우드 기반 빅 데이터 플랫폼에 대하여 서술한다. 단일 서버에서 용도에 맞는 빅 데이터 플랫폼을 구축한 후 이를 다수의 사용자가 사용할 수 있도록 작업을 분할하며, 사용자는 가상 환경에서 서버에 원격 접속함으로써 원활한 빅 데이터 분석이 가능하도록 설계한다.

2. 이론적 배경

2.1 Apache Spark

Apache Spark는 빅 데이터 분석을 위한 오픈 소스 데이터 처리 프레임워크이다. 이는 일괄 처리, 실시간 처리, 스트리밍 분석, 기계 학습 및 대화형 SQL과 같은 다양한 대형 데이터 작업 부하를 포괄하도록 설계되었으며, In-Memory 기반의 스토리지를 사용함으로써 Hadoop에 비하여 100배 빠른 데이터 처리가 가능하다. Spark는 원래 Scala 언어로 작성된 도구이지만 이외에도 Java, Python, R과 같이 응용 프로그램 개발을 위한 여러 프로그래밍 언어를 지원한다[3].

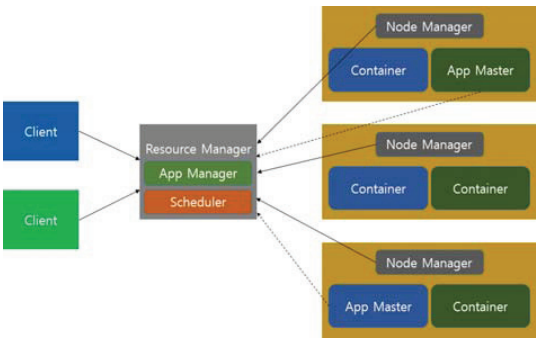


[그림 1] Apache Spark Architecture

2.2 Hadoop Yarn

Yarn은 기존의 Hadoop Map Reduce를 수행할 때 클러스터 자원 배분 시 병목 현상이 발생하는 문제점을 해결하기 위하여 개발된 클러스터 매니지먼트 프로그램이다. Yarn은 기존의 Map Reduce 시스템이 수행

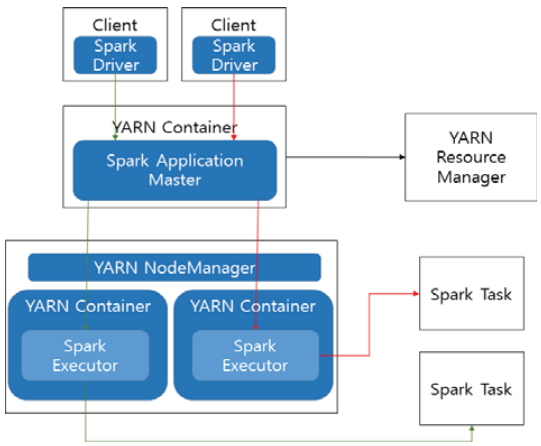
하는 자원 관리, Job 상태 관리를 Resource Manager와 Application Master로 분리하여 시스템 성능을 높인다. 또한, 기존의 Hadoop이 Map Reduce만을 실행할 수 있는 것과 달리 다양한 Application에 CPU 및 Memory를 할당함으로써 범용성 있는 컴퓨팅 클러스터를 수행할 수 있다[4].



[그림 2] Hadoop Yarn Architecture

3. 빅 데이터 학습 시스템 구성

본 논문의 시스템 구조는 [그림 3]과 같다. 우선, 서버에 Spark Task를 수행하기 위한 시스템을 구축한 후, YARN을 이용하여 개별 사용자의 업무를 클라우드 환경에서 제어한다. 이 때, YARN은 각 사용자의 업무를 클러스터링하여 자원을 적절히 배분하는 역할을 수행한다. 그 후, YARN Container로 전달된 사용자 프로그램을 Task로 변환하여 수행한다.



[그림 3] Proposed System Architecture

빅 데이터 학습 환경에서는, 일반적으로 가상 컴퓨팅 환경에서 데이터 분석을 수행하므로 한 대의 PC에 다수의 머신을 구성하는 것이 가능하다. 이를 이용해 사용자가 자신의 PC에서 다수의 클라이언트를 생성하여 서버에 접속한다. 서버에는 필요에 따라 학습에 관련된 여러 컴포넌트를 미리 설치하고, 사용자는 서버에 설치된 빅 데이터 분석 컴포넌트를 클라우드 환경에서 활용할 수 있다.

또한, 사용자는 Spark에서 제공하는 SparkR, Spark Web UI 등을 통하여 데이터 분석 결과, 수행중인 노드의 정보 등을 확인할 수 있다. 이를 통하여 사용자는 빅 데이터 처리 과정을 수월하게 이해하고 응용할 수 있다.

4. 결론 및 논의

본 논문에서는 저사양 환경에서 클라우드 컴퓨팅 환경을 구성하여 빅 데이터를 용이하게 학습할 수 있는 시스템 구축 방안에 대하여 서술하였다. 이를 통하여 단일 서버에서 다수의 사용자가 클라우드 컴퓨팅 환경에서 빅 데이터 플랫폼을 사용하여 데이터를 분석할 수 있다. 추후 연구로는, 다수의 서버 환경이 제공되었을 때 다수의 서버에 각기 다른 빅 데이터 플랫폼 환경을 구성하여 사용자의 요구에 따른 환경을 제공할 수 있는 커스터마이징을 진행할 예정이다.

Acknowledgement

"본 연구는 미래창조과학부 및 정보통신기술진흥센터의 SW중심대학지원사업(2015-0-00914)의 연구결과로 수행되었음"

참 고 문 헌

[1] Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data.* McGraw-Hill Osborne Media.

[2] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications, 19*(2), 171-209.

[3] Shoro, A. G., & Soomro, T. R. (2015). Big data analysis: Apache spark perspective. *Global Journal of Computer Science and Technology, 15*(1).

[4] Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., ... & Saha, B. (2013). Apache hadoop yarn: Yet another resource negotiator. In *Proceedings of the 4th annual Symposium on Cloud Computing* (p. 5). ACM.