고성능, 고확장성 빅데이터 분석 플랫폼

박경석*, **, 유찬희*, **, 김유선*, **, 엄정호*
*한국과학기술정보연구원
**UST 빅데이터과학과
gspark@kisti.re.kr

High-performance and Highly Scalable Big Data Analysis Platform

Kyongseok Park*, **, Chan Hee Yu*, **, Yuseon Kim*, ** Jung-Ho Um*
*Korea Institute of Science and Technology Information
**Department of Big Data Science, UST

요 약

빅데이터를 활용한 기계학습 모델을 개발하기 위해서는 빅데이터 처리를 위한 플랫폼과 딥러닝 프레임워크 등 고급 분석을 수행할 수 있는 도구의 활용이 동시에 요구된다. 그러나 빅데이터 플랫폼과 딥러닝 프레임워크를 자유롭게 활용하기 위해서는 상당한 수준의 기술적 지식과 경험이 필요하다. 또한 빅데이터를 이용한 딥러닝 모델을 개발할 경우 분산처리와 병렬처리에 대한 지식과 추가적인 작업이 요구된다. 본 연구에서는 빅데이터를 활용한 기계학습 모형을 자유롭게 개발 및 공유하고 분산 딥러닝을 위한 시스템적 지원을 통해 분야별로 딥러닝 모형을 개발하는 응용 연구자들이 활용할 수 있는 플랫폼을 제시하였다. 본 연구를 통해 다양한 분야의 연구자들이 자신의 데이터를 이용하여 모형을 개발할 경우 분산처리와 병렬처리를 위한 기술적 제약을 극복하고 보다 빠르고 효율적인 방법으로 모형을 개발하고 현업에 활용할 수 있을 것으로 기대한다.

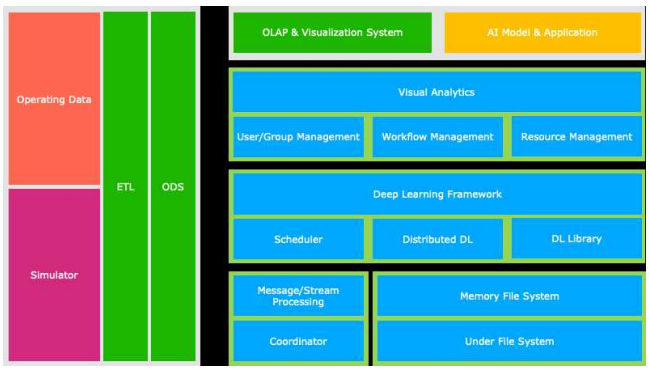
1. 서론

데이터의 증가에 따라 대규모 데이터를 빠르고 효율적으로 분석할 수 있는 환경에 대한 요구도 증가하고 있다. Hadoop이나 Spark 등 대표적인 빅데이터 플랫폼을 비롯한 다양한 플랫폼들이 많은 연구자와 개발자들을 통해 활용되고 있다. 또한 딥러닝을 비롯한 다양한 유형의 기계학습 모델이나 수치모델 등을 개발할 경우 추가적으로 딥러닝 프레임워크 등의 고급 분석 도구가 필요하다. 이러한 플랫폼과 도구들은 기본적으로 프로그래밍 언어를 이용하여 사용자가 분석 모델을 개발하는 것을 목적으로 하고 있다.

대규모 데이터를 분석할 경우 분산처리나 병렬처리를 위해 추가적인 지식과 작업을 필요로 한다. 다양한 유형의 빅데이터 플랫폼과 딥러닝 프레임워크를 활용하기 위해서는 각 도구들의 환경과 특징을 이해해야 하며 각 도구들의 장단점이 존재하기 때문에 활용 목적에 따라 적절한 도구를 선정하고 기능을 정확히 이해할 수 있어야 한다.

본 연구에서는 복잡한 딥러닝 모형을 보다 효율적으로 개발하고 공유할 수 있는 환경과 함께 대규모 데이터를 빠르게 처리할 수 있도록 분산처리와 병렬처리를 플랫폼의 사용자 환경 수준에서 지원하는 환경을 제공하여 모형 개발의 효율성을 높일 뿐만 아니라 분산처리와 병렬처리에 대한 전문 지식과 경험 없이도 대규모 데이터를 빠르게 분석할 수 있는 환경을 제공하고 있다. 본 연구에서 제안한 플랫폼은 발전설비 진단을 위한 실시간 센서 데이터 분석에 중점을 두고 구현되었다. 그러나 이에 국한되지 않고 제조 데이터, 위성영상을 비롯한 다양한 분야에서 활용이 가능하다.

2. 플랫폼 개요 및 구성 요소



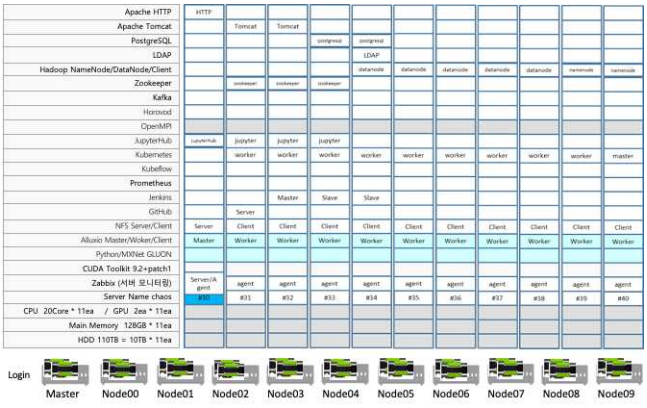
(그림 1) 플랫폼 아키텍처

본 연구에서 제안한 플랫폼은 빅데이터의 입수와 저장, 관리를 비롯하여 빅데이터 처리와 분석에 이르는 전 과정에 적용하기 위한 목적으로 구현되었다. 데이터는 실제 운영 상황에서 발생한 데이터와 수치 모델로 생성한 합성 데이터를 동시에 활용할 수 있다.

원본 데이터를 처리하는 첫 단계에서는 빅데이터 플랫폼을 이용하여 전통적인 DW에서 수행하는 ETL 절차를 수행하고 ODS에서 각 응용에 적합한 데이터 처리 과정을 거칠 수 있도록 구성하였다. 이후 ODS에 저장된 데이터를 분석에 적합한 환경으로 전송하여 HDFS에 데이터를 저장하고 메모리 기반 파일시스템인 Alluxio에 활용도가 높은 데이터를 적재하여 분석을 수행하게 된다[1, 2].

다음 단계에서는 분산 딥러닝을 수행하기 위한 프레임워크와 라이브러리를 기반으로 GUI 분석 환경을 제공하여 연구자들이 손쉽게 복잡한 딥러닝 모형 개발과 병렬처리를 지원할 수 있도록 지원하고 있다[3].

연구자들이 개발한 모형은 플랫폼을 사용하는 이용자 간 공유하고 재활용할 수 있으며 GUI 기반의 딥러닝 모형뿐만 아니라 Python 코드를 직접 공유하고 재활용할 수 있도록 지원한다[4]. 특히 코드의 재활용은 GUI 기반으로 개발한 모형을 정교하게 개선하고 보다 복잡한 처리 흐름과 모형의 구성요소를 갖는 수준 높은 모형을 개발할 수 있도록 지원하고 있다.



(그림 2) 플랫폼 구성 요소

빅데이터 처리 환경에서는 기본적인 데이터 집계, 기초적인 분석 및 가시화 등 현업 이용자 관점에서 직관적 작업이 가능하고 추가적인 분석이 필요한 경우 Spark을 이용한 분석도 가능하다. 모형 개발이 완성된 경우 서비스 시스템에 적용할 수 있도록 모형을 배포하고 실행할 수 있는 환경을 제공하고 있다.

개발한 플랫폼은 다수의 노드에 노드별로 다수의 GPU를 장착하여 모형을 개발할 때 CPU와 GPU 자원을 연구자의 요구에 맞춰 유연하게 선택하여 분석할 수 있다. 따라서 응용 사례별로 플랫폼에서 가용한 자원의 범위 내에서 빅데이터 처리에 적합한 수준으로 자원의 규모를 선택하고 실행할 수 있다.

3. 결론

본 연구에서는 발전 설비 등 다양한 설비에서 생산되는 실시간 센서 데이터와 제조·공정·서비스 데이터 및 위성영상 등 다양한 유형의 빅데이터를 분석할 수 있는 플랫폼을 제공하고 있다. 본 연구에서 제안한 플랫폼에서는 딥러닝 모형 개발의 효율성뿐만 아니라 모형의 공유와 배포 등 빅데이터 분석 전 프로세스를 지원한다. 따라서 다양한 분야의 응용 연구자들이 본 연구에서 제안한 플랫폼을 이용하여 연구자의 기술적 지식과 경험의 제약을 극복하고 보다 도전적으로 연구를 수행할 수 있을 것으로 기대한다.

사사

이 논문은 산업통상자원부의 재원으로 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구입니다. (No. 20181110100420)

이 논문은 중소벤처기업부의 재원으로 중소기업기술정보진흥원(TIPA)의 지원을 받아 수행한 연구입니다. (S3126610)

참고문헌

[1] Xu Chang and Li Zha, “The Performance Analysis of Cache Architecture Based on Alluxio over Virtualized Infrastructure”, IPDPSW, 2018, pp. 515-519

[2] Youngmoom Eom, Jinwoong Kim, Deukyeon Hwang, Jaewon Kwak, Minho Shin, Beomseok Nam, “Improving Multi-dimensional query processing with data migration in distributed cache infrastructure”, HiPC, 2014, pp. 1-10

[3] Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V., Mao, M. Z., Ranzato, M., Senior, A., Tucker, P., Yang, K., and Ng, A. Y., “Large scale distributed deep networks”, NIPS, 2012, pp. 1223-1231

[4] 박경석, 유찬희, Komal Sarda, 엄정호, “분산 딥러닝 모델 개발을 위한 고수준 분석 플랫폼”, 한국정보처리학회 추계학술대회, 2020, pp. 804-806