KI Cloud: 슈퍼컴퓨터를 통한 빅데이터 분석 및 머신 러닝 서비스 구축 방안

박주원, 이승민, 정기문, 홍태영
한국과학기술정보연구원
{juwon.park, smlee76, kmjeong, tyhong}@kisti.re.kr

KI Cloud: Design and Implementation of BigData Analysis and Machine Learning Applications on Supercomputer

Ju-Won Park, Seungmin Lee, Kimoon Jeong, and TaeYoung Hong
Korea Institute of Science and Technology Information

요 약

전통적으로 기초 과학 분야의 대규모 워크로드 작업들은 슈퍼컴퓨터와 같은 대용량 클러스터 시스템을 이용하여 수행해왔다. 그러나 최근 빅데이터 및 머신 러닝과 같은 새로운 분야에서의 컴퓨팅 자원 요구가 증가하고 기존 사용자의 요구 사항도 다양해짐에 따라 기존의 클러스터 시스템 운영 환경에서는 많은 어려움이 나타나고 있다. 이러한 문제를 해결하기 위해 한국과학기술정보연구원(KISTI)에서는 지난 3 월부터 KI (KISTI Intelligent) Cloud 서비스를 개발하여 서비스를 제공하고 있다. KI Cloud 서비스는 다음과 같은 특징이 있다. 첫째, Jupyter 과 RStudio 와 같은 대화형 개발 환경을 웹을 통해 제공함으로써 사용자는 언제, 어디서나 손쉽게 서비스를 활용할 수 있다. 둘째, 컨테이너 기술을 활용하여 사용자가 요구하는 개발 및 실행 환경을 실시간으로 구성하여 제공한다. 셋째, 사용자의 서비스 환경을 동적으로 구성하여 제공함으로써 컴퓨팅 자원의 효율성을 높일 수 있다.

1. 서론

전통적으로 기상, 화학, 고에너지 물리와 같은 기초 과학 분야의 대규모 워크로드 작업들은 슈퍼컴퓨터와 같은 대용량 클러스터 시스템을 이용하여 수행해왔다 [1]-[3]. 그러나 최근 빅데이터 분석 및 머신러닝과 같은 새로운 분야에서의 컴퓨팅 자원의 요구가 증가하고 있고 기존의 기초 과학 분야의 작업 형태도 다양해짐에 따라 기존 클러스터 시스템을 통한 서비스 제공에 많은 어려움에 직면해 있다. 이러한 문제점을 해결하고 사용자 맞춤형 서비스를 제공하기 위해 한국과학기술정보연구원(KISTI)에서는 2020 년 3 월부터 'KI (KISTI Intelligent) Cloud' 서비스를 개발하여 제공해 오고 있다.

KI Cloud 서비스는 슈퍼컴퓨터와 같은 대용량 컴퓨팅 자원을 통해 Infrastructure-as-a-Service (IaaS)부터 Platform-as-a-Service (PaaS)까지 다양한 형태의 서비스를 제공하기 위해 개발된 클라우드 플랫폼으로 OpenStack[4], Kubernetes[5]와 같은 오픈 소스 기반으로 구축되어 운영되고 있다. 본 논문에서는 KI Cloud 서비스 중 PaaS 를 중심으로 알아보도록 한다. KI Cloud 는 다음과 같은 특징이 있다. 첫째, Jupyter, RStudio 와 같은 웹 기반 대화형 개발 환경을 제공한다. Jupyter 는 Julia, Python 과 같은 다양한 프로그램 개발 환경을 웹 브라우저를 통해 제공하는 툴킷으로 최근 머신러닝 튜토리얼과 같은 간단한 프로그램을 웹을 통해 쉽게 실행할 수 있도록 다양한 예제 파일이 공개되고 있다. RStudio 는 통계 및 데이터 분석에 많이 활용되고 있는 R 언어에 특화되어 데이터 관리, 그래픽 처리, 디버깅을 GUI 를 통해 쉽게 할 수 있도록 개발된 툴킷이다. 둘째, 사용자가 요구하는 개발 및 실행 환경을 실시간으로 제공할 수 있다. KI Cloud 의 경우 컨테이너 기술을 활용하여 다양한 사용자 요구 사항을 맞춤형으로 제공할 수 있다. 즉, 사용자는 자신의 개발 환경 및 실행 환경에 적합한 Docker 이미지를 검색/선택하거나 직접 이미지를 제작하여 KI Cloud 플랫폼을 통해 배포하고 웹을 통해 실행할 수 있다. 마지막으로 컴퓨팅 자원의 효율성을 높일 수 있다. KI Cloud 의 경우 제공되는 서비스는 사용자 요청을 기반으로 실시간으로 서비스를 배포하고 배포된 서비스를 이용하는 방식이다. 즉, 사용자의 요청에 따

라 동적으로 서비스를 구성하여 프로세스를 최소화함으로써 컴퓨팅 자원의 효율성을 향상시킬 수 있다.

2. KI Cloud 아키텍쳐 및 구현 결과

2.1 아키텍쳐

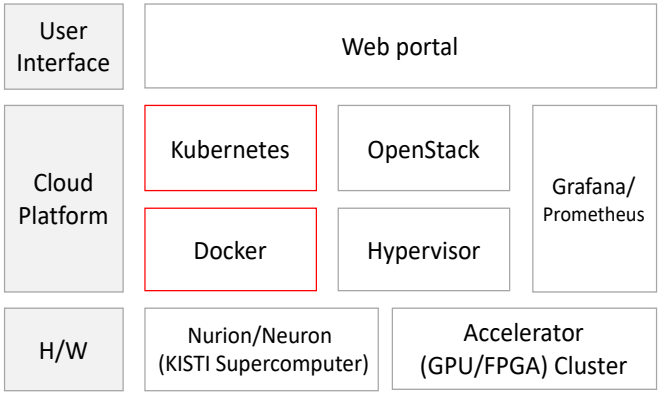

Fig. 1. Architecture of KI Cloud.

KI Cloud 는 슈퍼컴퓨터와 같은 대용량 클러스터 시스템을 통해 클라우드 서비스를 제공하기 위한 플랫폼으로 크게 가상 머신, 오브젝트 스토리지와 같은 컴퓨팅 인프라 자원을 제공하는 IaaS 와 Jupyter, RStudio 와 같이 개발 플랫폼을 제공하는 PaaS 로 구성된다. 각 서비스는 그림 1 에서 보는 바와 같이 OpenStack, Kubernetes 와 같은 오픈 소스를 기반으로 구축되어 운영되고 있다.

KI Cloud 에서 제공하는 PaaS 는 사용자 요구에 최적화된 서비스 제공을 위해 컨테이너 기반으로 설계되어 개발되었으며 Container runtime 으로는 Docker 를 사용하고 container orchestration tool 로는 kubernetes 를 활용하였다. 이와 같이 컨테이너 기술을 기반으로 PaaS 를 제공함에 다음 2 가지 이점이 있다. 첫째는 하이퍼바이져 기반 가상화 기술 대비 성능이 우수하다. 일반적으로 클라우드 서비스에서 많이 활용되는 Xen, KVM 과 같은 하이퍼바이져 기반의 가상화 기술은 하이퍼바이져에서 발생하는 overhead 로 인하여 성능저하가 발생한다. 그러나 컨테이너 기반의 가상화 기술은 Guest OS 가 없이 Host 의 커널을 공유하기 때문에 기존 하이퍼바이져 기반 방식 대비 성능이 우수하다[6]-[8]. 둘째, 다양한 사용자 요구에 맞는 서비스 환경을 실시간으로 제공한다. 하이퍼바이져 기술을 이용하여 가상 환경을 생성할 경우 수 분 ~ 수 십분의 시간이 소요된다. 그러나 컨테이너의 경우 수초 이내로 가상 환경을 생성할 수 있기 때문에 실시간 배포 및 서비스 이용이 가능하다[9].

2.2 구현 결과

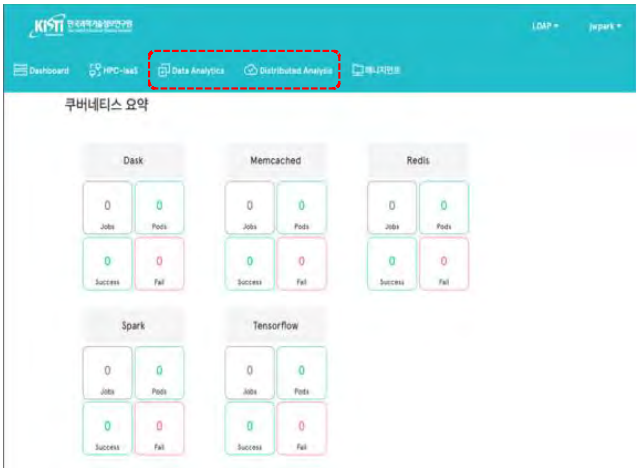

Fig. 2. KI Cloud Dashboard.

그림 2 는 KI Cloud 서비스 대시보드를 보여준다. 그림에서 보는 바와 같이 제공 되는 메뉴는 HPC-IaaS, Data Analytics, Distributed Analysis 가 있으며 PaaS 에 해당 서비스는 크게 3 가지로 분류할 수 있다. 첫째, Jupyter, RStudio 와 대화형 개발 환경이다. Jupyter 의 경우 그림 3 에서 보는 바와 같이 Jupyter Notebook 에 접속한 후 필요한 커널을 선택하여 이용할 수 있다. 현재 제공되고 있는 커널은 Python, Python with tensorflow, Python with tensorflow in GPU node, Scala, Spark 5 개의 커널이 제공되고 있으며 사용자의 요청이 있을 경우 관리자가 커널을 생성하고 기존 서비스에 추가하여 사용자에게 제공하고 있다. RStudio 의 경우 Docker 이미지 기반으로 배포/실행함으로써 사용자가 원하는 환경을 직접 제작하거나 Docker hub 에서 선택하여 이용할 수 있다.

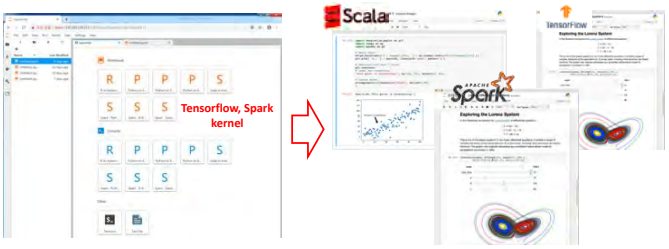

Fig. 3. Jupyter Service.

둘째, key-value 기반의 인메모리 데이터 베이스 서비스를 제공한다. 최근 NOSQL 방식의 데이터베이스에 대한 요구가 증가하고 있으며 특히, 실시간 데이터 분석이 가능한 인메모리 형태의 데이터 베이스에 대한 요구가 증가하고 있다. 이러한 요구 사항을 반영하여 KI Cloud 에서는 2 가지 형태 (Redis,

Memcached)의 인메모리 데이터 베이스 서비스를 제공한다.

마지막으로 Spark, 분산 Tensorflow 와 같은 대규모의 데이터 분석 환경을 제공한다. Spark 과 Tensorflow 의 경우 대화형 작업 환경이 아닌 배치 작업 형태로 실행되며 실행하고자 하는 Docker 이미지 및 자원의 규모를 선택하고 작업을 제출하면 kubernetes 에서는 가용 자원과 매칭하여 자원을 할당하고 Docker 이미지를 다운로드 받아 분산 환경에서 작업을 실행한다.

3. 결론

최근 빅데이터 분석 및 머신 러닝과 같은 새로운 분야의 서비스 제공 및 다양한 형태의 사용자 요구사항을 만족시키기 위해 KISTI 에서는 지난 3 월부터 KI (KISTI Intelligent) Cloud 서비스를 개발하여 제공해 오고 있다. KI Cloud 는 다음 3 가지 특징이 있다. 먼저, Jupyter, RStudio 와 같은 웹 기반 대화형 개발 환경을 제공한다. 둘째, 사용자가 요구하는 개발 및 실행 환경을 실시간으로 제공한다. 마지막으로 서비스의 동적 구성을 통해 컴퓨팅 자원의 효율성을 향상 시킨다. 특히, KI Cloud 는 서비스 환경을 컨테이너 기반 가상화 기술을 활용함으로써 성능 저하 없이 실시간으로 서비스를 배포하여 사용자에게 제공하고 있다.

참고문헌

[1] E. Deelman, D. Gannon, M. Shields, and I. Taylor, “Workflows and e-science: An overview of workflow system features and capabilities,” *Future Generation Computer Systems*, vol. 25, no. 5, pp. 528–540, 2009.

[2] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, “Examining the chal- lenges of scientific workflows,” *IEEE Computer*, vol. 40, no. 12, pp. 24–32, Dec 2007.

[3] Schlagkamp, Stephan, et al. "Understanding user behavior: from HPC to HTC." *Procedia Computer Science* 80 (2016): 2241-2245.

[4] “OpenStack.” [Online]. Available: https://www.openstack.org/ Oct. 2020.

[5] “Kubernetes.” [Online]. Available: https://kubernetes.io/. Oct. 2020.

[6] Li, Zheng, et al. "Performance overhead comparison between hypervisor and container based virtualization." *IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, 2017.

[7] Zhang, Jie, Xiaoyi Lu, and Dhabaleswar K. Panda. "Performance characterization of hypervisor-and container-based virtualization for HPC on SR-IOV enabled InfiniBand clusters." *IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2016.

[8] Babu, Anish, et al. "System performance evaluation of para virtualization, container virtualization, and full virtualization using xen, openvz, and xenserver." in *Proc. of 2014 Fourth International Conference on Advances in Computing and Communications*, 2014.

[9] Yamato, Yoji. "OpenStack hypervisor, container and baremetal servers performance comparison." *IEICE Communications Express* 4.7 (2015): 228-232.