멀티 클러스터 기반 국가연구데이터커먼즈 간 워크플로우 연계 방안 설계 및 구현

김다솔*, 이상백, 박성은, 조민희, 이미경, 송사광, 임형준
한국과학기술정보연구원 연구데이터공유센터
{dskim, sb_lee, pse3598, mini, jerryis, esmallj, hjyim}@kisti.re.kr

Design and Implementation of Workflow Federation Method for Multi-cluster Based Korea Research Data Commons

Dasol Kim, Sang-baek Lee, Seong-eun Park, Minhee Cho, Mikyoung Lee, Sa-kwang Song, Hyung-jun Yim
Research Data Sharing Center, Korea Institute of Science and Technology Information

요 약

최근 오픈 사이언스 문화가 확산됨에 따라 오픈 데이터, 오픈 소스 소프트웨어와 같은 공개된 리소스들을 효율적으로 공유 및 활용하기 위한 방법이 주목을 받고 있다. 본 논문에서는 연구 소프트웨어의 재현성을 향상시키기 위한 국가연구데이터커먼즈(KRDC)를 소개하고 다중 KRDC 클러스터 간 워크플로우 연계 방안을 제안한다. 국가연구데이터커먼즈는 연구 소프트웨어와 분석 환경인 인프라를 결합하여 함께 제공하는 서비스로, 멀티 노드 쿠버네티스(kubernetes) 클러스터를 기반으로 동작한다. 따라서, 서로 다른 KRDC 프레임워크에 존재하는 리소스들을 하나의 워크플로우로 연계하는 것은 복잡한 사용자 인증/인가 문제, 보안 상의 문제를 고려하여야 한다. 본 논문에서는 프록시(proxy) 앱을 사용하는 워크플로우 연계 기능을 제안하고, 이를 지원하기 위한 통합 인증, 인가 체계와 연계 방안을 구현한다. 제안하는 방법을 두 개의 KRDC 프레임워크를 대상으로 적용하여 제안 워크플로우 연계 방법의 유효함을 확인한다. 본 논문에서 제안하는 워크플로우 연계 방법과 시나리오는 실제 멀티 클러스터 연계 방안을 구현한 사례로, KRDC 프레임워크 뿐만 아니라 다양한 쿠버네티스 기반 리소스 연계에 활용할 수 있는 우수한 결과로 사료된다.

1. 서론

오픈 사이언스의 등장 이후 오픈 액세스, 오픈 데이터, 오픈 소스 등 오픈 사이언스 구성 요소들에 대한 관심이 증가하고 있다. 특히, 미국 정부에서 2023 년을 오픈 사이언스의 해(The year of Open Science)로 선언하고 NASA, CERN 을 비롯한 세계적인 선도 연구기관에서 이를 주도하면서 더욱 활발한 논의가 진행되고 있다[1]. 연구 분야에서는 이러한 문화가 연구에 활용되는 데이터, 소프트웨어를 비롯하여 분석 및 운영 환경까지 점차 확대되고 있다.

이에 따라, 과학기술 분야에서는 연구 데이터뿐만 아니라 소프트웨어, 인프라를 함께 제공하여 결과물의 공유 및 활용을 극대화하기 위한 연구 데이터 커먼즈(Research Data commons)가 주목을 받고 있다. 유럽, 호주 등 오픈 사이언스를 선도하는 국가에서는 이미 많은 논의를 통해 구현이 진행되고 있으며, 국내에서는 이에 발맞춰 필요성을 인식하고 관련 논의가 시작되고 있다[2]. 본 논문에서는 이를 지원하기 위한 국가연구데이터커먼즈(KRDC)를 소개하고 다중 KRDC 프레임워크 간 워크플로우 실행 연계 방안과 시나리오를 제안한다.

2. 연구 배경

오픈 사이언스는 연구, 교육, 사회, 정책, 기술 등 다양한 분야에서 연구의 전 과정을 개방적으로 전환하려는 움직임을 의미한다. 특히, 연구 데이터 및 소프트웨어가 인공지능 기반 연구의 핵심요소로 자리잡으면서 더욱 활발하게 오픈 사이언스 문화가 확산되고 있다. 이에 따라, 연구 데이터의 효율적인 관리 및 활용에 대한 요구 또한 증가하고 있으며, 많은 나라에서 국가 차원의 데이터 플랫폼을 운영하고 있다. 대표적으로 유럽 연합의 OpenAIRE[4], 일본의 NII ROCS[5], 호주의 ARDC[6]가 있으며, 국내에는 국가연구데이터플랫폼인 DataON[7]이 있다.

현재는 데이터뿐만 아니라 연구에 활용되는 리소스 전반의 통합 관리 필요성이 증가하고 있으며, 이러한 배경에서 연구 데이터 커먼즈 개념이 등장하였다. 연구 데이터 커먼즈는 연구 데이터와 컴퓨팅 리소스의

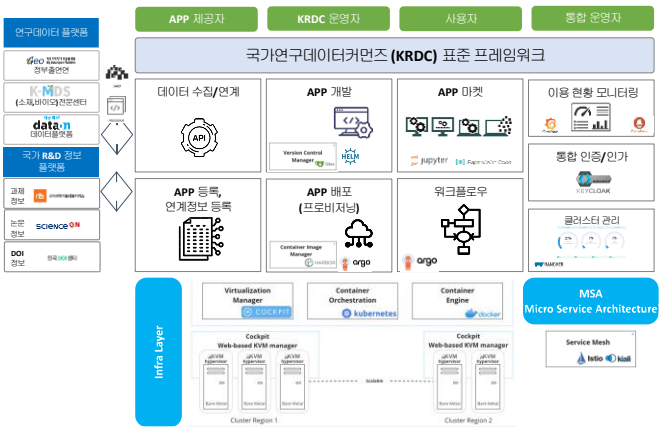
활용성을 극대화하여 경쟁력을 확보하는 것이 주된 목적이며, 국내에서도 이에 대한 요구가 증가하고 있다[3]. 따라서, 이미 많은 연구자들은 연구 소프트웨어나 분석도구, 컴퓨팅 자원의 공유 및 활용을 적극적으로 원하고 있으며 이에 대한 지원 방안이 필요하다.

이러한 필요에 대응하기 위해서 효율적인 컴퓨팅 자원의 수집, 공유, 활용을 기술적으로 지원하기 위한 통합 플랫폼 개발이 필요하다. 국가연구데이터커먼즈는 이를 구현하기 위해 설계되었으며, KRDC 프레임워크, KRDC 허브 서비스, KRDC 스토리지 서비스로 구성된다. 본 논문에서는 서로 다른 KRDC 프레임워크 간 워크플로우 연계 방법을 설계 및 구현하고, 이를 위한 통합 인증/인가 기능을 함께 제안한다.

3. 국가연구데이터커먼즈 개요

본 절에서는 국가연구데이터커먼즈의 구성요소와 기능에 대해 소개한다. 앞서 설명한 것과 같이, 국가연구데이터커먼즈는 다음의 세 가지 요소로 구성되어 있다. 첫째, KRDC 허브 서비스는 컴퓨팅 리소스의 공유 및 통합 인증/인가 기능을 담당한다. 둘째, KRDC 프레임워크는 핵심 구성요소로써 리소스 제작 및 배포를 지원한다. 셋째, 스토리지 서비스는 이미지, 소스코드 등 프레임워크 동작에 필요한 저장소부터 개인별, 공유 저장소를 포함한다.

그림 1 은 국가연구데이터커먼즈의 핵심을 담당하고 있는 KRDC 프레임워크의 구성도를 나타낸다. 그림을 보면, 데이터 수집부터 앱 개발과 배포, 사용자 관리를 위한 통합 인증/인가 기능까지 프레임워크에서 제공해야할 핵심 기능들이 포함되어 있는 것을 알 수 있다. 또한, 프레임워크의 각 기능들은 MSA (Micro Service Architecture)로 구현되며, 각각의 모듈들은 다수의 노드로 구성된 쿠버네티스 클러스터를 기반으로 동작한다. 본 논문에서는 KRDC 프레임워크 간 워크플로우 연계 방안에 대해 다룬다.

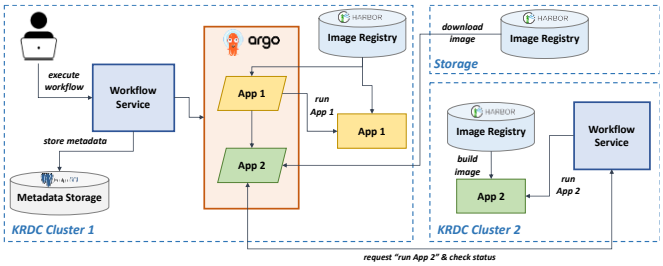


(그림 1) KRDC 프레임워크 구성도

4. 워크플로우 연계 방안 설계 및 구현

본 절에서는 서로 다른 클러스터에 KRDC 프레임워크가 구축된 환경에서, 각 프레임워크의 앱을 하나의 워크플로우로 실행하는 연계 방안을 설계한다. 제안하는 방법은 프록시 앱을 사용하는 방식으로, 프록시 앱은 로컬 클러스터에 타 클러스터에서 실행되는 앱의 상태정보를 동기화한다. 타 프레임워크로 앱 실행을 요청하는 경우, 로컬 프레임워크는 원격 앱을 프록시 형태로 생성하고 앱이 종료되면 반환된 결과를 사용하여 다음 동작을 수행한다.

제안 방법은 이종 쿠버네티스 클러스터 간 리소스 연계 기능을 구현하기 위해 고안한 방법이다. 일반적으로 사용되는 클러스터 연계 방안은 configuration 파일을 사용하여 리소스에 대한 접근을 허용하는 것이다. 하지만, 이는 클러스터 접근 정보가 하나의 설정 파일로 관리되기 때문에 외부로 노출될 경우 심각한 보안 문제로 이어진다. 따라서, 본 논문에서는 통합된 인증/인가 체계를 설계 및 적용하여 이러한 문제를 방지한다. 그림 2 는 설계한 KRDC 의 멀티 클러스터 간 워크플로우 연계 시나리오를 나타낸다.

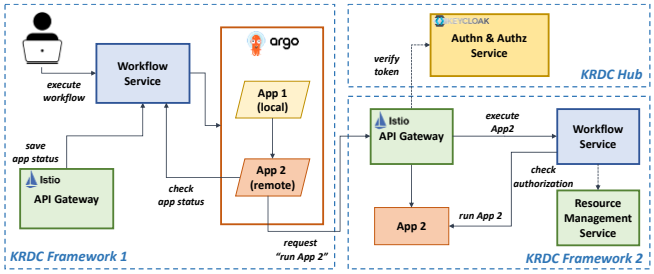


(그림 2) 멀티 클러스터 간 워크플로우 연계 시나리오

그림을 보면, KRDC 클러스터 1 내에 App 1 과 App 2 로 구성된 워크플로우를 확인할 수 있다. App 1 은 워크플로우 실행 주체와 동일한 로컬 클러스터에 존재하기 때문에 프레임워크 자체 이미지 저장소에서 컨테이너 이미지를 빌드하여 App 1 이 실행된다. 다음으로 App 2 는 외부에 위치한 KRDC 클러스터 2 에 존재하며, 이에 대한 실행 시나리오는 두 가지로 나뉜다. (1) 앱 이미지가 공용 스토리지에 존재하는 경우와 (2) 다른 KRDC 클러스터에 존재하는 앱을 실행 요청하는 경우이다. 시나리오 (1)은 그림 2 의 오른쪽 상단에서 볼 수 있듯이 이미지를 클러스터 1 로 다운받은 후 앱을 실행한다. 시나리오 (2)는 클러스터 2 에 App 2 실행 요청을 전달하고, 클러스터 2 내부에서 App 2 에 해당하는 이미지를 빌드하여 실행된 결과와 상태를 반환한다. 각 시나리오의 동작을 수행하기 위해서는 허브 또는 클러스터 간 인증, 인가 기능이 지원되어야 한다.

특히, 시나리오 (2)와 같이 다른 클러스터에 존재하

는 앱을 실행하려는 경우, 요청한 사용자의 접근 및 실행 권한에 대한 확인이 동시에 이루어져야 한다. 이를 위해, Keycloak[8]을 사용한 통합 인증/인가 체계를 구현하며, 허브 서비스와 프레임워크 모두에 이를 적용한다. 그림 3 은 시나리오 (2)에서 KRDC 허브를 통해 권한을 획득하고 워크플로우를 실행하는 과정을 나타낸다. 그림을 보면, 워크플로우를 실행하는 프레임워크 1 은 허브에 타 프레임워크의 앱 실행 권한을 사전에 요청하고, 실제 앱은 프레임워크 2 에서 실행된다. 따라서, 프레임워크 1 에는 프록시 앱이 생성되며, 이를 통해 App 2 실행 결과와 상태 정보를 전달받고 워크플로우에 반영한다.



(그림 3) 통합 인증/인가를 적용한 워크플로우 연계

그림 4 는 제안하는 워크플로우 연계 동작 과정을 나타낸 알고리즘이다. 이종 KRDC 프레임워크 간 워크플로우 동작 과정에 대해 상세히 알 수 있으며 이때, 앱은 python 으로 작성되고 argparse 방식으로 입력 파라미터를 받는다고 가정한다.

Algorithm 1 Multi-cluster workflow federation procedure in KRDC.

```
Input:
    workflow: a list of the workflow information.
    user: a list of user information.
Output:
    resultFile: the file of the workflow execution results.
procedure WORKFLOWFEDERATION(workflow, user)
    execInfo := parseWorkflow(workflow);
    execAuth := checkAuth(execInfo, user);
    if execAuth is not empty then
        for i := 1 to execInfo.length() do
            app := parseApp(execInfo[i]);
            if app.location is remote then
                appAuthz := validateAuthz(execAuth, app);
                // Receive parameters in the app through argparse
                proxyApp := createProxyApp(appAuthz, app);
                tempResult := requestExecApp(proxyApp);
                Append tempResult to resultFile;
            end if
            tempResult := executeApp(app);
            Append tempResult to resultFile;
        end for
    end if
    return result of workflow execution
end procedure
```

(그림 4) KRDC 의 워크플로우 연계 알고리즘

그림을 보면, 워크플로우 실행 정보와 사용자 정보를 받아 사용자 인증 절차를 수행한다. 사용자 인증이 끝나면, 워크플로우에 정의된 앱을 순차적으로 실행한다. 이때, 앱이 로컬 클러스터에 존재하면 그림 2 의 App1 과 같이 자체 이미지 레지스트리에서 이미지를 받아 빌드한다. 만약, 앱이 다른 클러스터에 존재한다면 사용자의 원격 앱 실행 권한을 검증하고 로컬 클러스터에 프록시 앱을 생성한다. 프록시 앱은 타 클러스터의 실제 앱을 쿠버네티스 서비스 형태로 실행하여 상태 정보와 실행 결과를 받는다. 원격 앱 실행이 종료되면 결과를 저장하고 프록시 앱은 종료된다. 제안하는 연계 방법은 KRDC 프로토타입에 적용하여 동작의 유효함을 확인하였으며, 향후 안정화를 통해 KRDC 1 차 구축 배포판에 포함될 예정이다.

5. 결론

본 논문에서는 국가연구데이터커먼즈 개념과 KRDC 프레임워크를 소개하고, 이종 KRDC 프레임워크 간 워크플로우 연계 기능을 설계 및 구현하였다. 기존 클러스터 연계 방식의 보안 이슈를 고려하여 프록시 앱을 사용한 연계 기능을 설계하였으며, 통합 인증/인가 기능을 통해 이를 효과적으로 지원하는 방안을 제안하였다. 제안하는 방법과 연계 시나리오는 KRDC 프로토타입에 적용하여 유효함을 확인하였으며, 향후 안정화를 통해 KRDC 1 차 구축 배포판에 포함될 예정이다. 본 논문의 제안 방법과 연계 시나리오는 다양한 쿠버네티스 기반의 멀티 클러스터 연계에 활용 가능하며, 확장성이 우수한 결과라 생각된다.

ACKNOWLEDGMENT

이 논문은 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다(과제번호: (KISTI)K-23-L01-C03-S01, (NTIS)1711198423).

참 고 문 헌

[1] C. Gentemann, “Why NASA and Federal Agencies are Declaring this the Year of Open Science,” *nature*, Jan. 2023.

[2] H. Yim, M. Lee, S. Song, D. Seo, and M. Cho, “Design and Application of Korea Research Data Commons for Data-driven Research and Development,” *Journal of Korean Institute of Intelligent Systems*, Vol. 32, No. 5, pp. 392-400, Oct. 2022.

[3] S. Song, and D. Seo, “Revitalization Plan for R&D Research Outcomes: focusing on research data”, In *Proc. of the Korea International Conference on Convergence Content*, Jeju, Korea, Dec. 2021.

[4] OpenAIRE, https://www.openaire.eu/.

[5] NII ROCS, https://rcos.nii.ac.jp/en/.

[6] ARDC, https://ardc.edu.au/.

[7] DataON, https://dataon.kisti.re.kr/.

[8] A. Chatterjee and A. Prinz, “Applying Spring Security Framework with Keycloak-based OAuth2 to Protect Microservice Architecture APIs: A Case Study,” *Sensors*, Vol. 22, Issue 5, 1-27, Feb. 2022.