쿠버네티스 환경에서 Hadoop/Spark 클러스터를 프로비저닝하고 공동 관리하는 빅데이터 플랫폼 개발

김도희*, 백혜원*, 이도윤*, 허현진*, 고석주*, 이동욱**

Big data platform to provision and co-manage Hadoop/Spark clusters in Kubernetes environments

Dohee Kim*, Hyewon Baek*, Doyoon Lee*, Hyeonjin Heo*, Seokjoo Koh*, Dongwook Lee**

요 약

컨테이너 기반 가상환경과 빅데이터 플랫폼이 주목받는 상황에서 대량의 데이터를 효율적으로 관리하는 것은 필수적이다. 하지만 빅데이터 분석을 위해 인프라를 구축하려면 많은 시간과 비용이 소요된다. 이러한 문제점을 해결하기 위해 본 논문에서는 전문지식이 없는 사용자도 웹 기반 인터페이스를 통해 컨테이너 오케스트레이션 환경에서 Hadoop/Spark 클러스터를 간편하게 구축하고 공유할 수 있는 플랫폼을 제안한다.

Ⅰ. 서 론

미국 시장 조사 기관 IDC의 조사 결과에 따르면 2020년 COVID-19의 확산으로 인한 언택트 시대의 도래로 데이터양이 대폭 증가했다[1]. 이러한 데이터 폭증과 맞물려 Apache Hadoop과 Spark 등 분산 데이터 처리를 지원하는 빅데이터 플랫폼은 SW산업 분야에서 중요한 영역이 되었다.

하지만 인프라 엔지니어 및 데이터 엔지니어가 부재하는 일반 기업에서 데이터 처리를 위해 Hadoop과 Spark를 직접 설치하고 활용하기엔 큰 어려움이 있고, 직접 빅데이터 플랫폼을 다룬다고 해도 전통적인 온프레미스 환경에 서버를 구축하는 것은 하드웨어부터 소프트웨어까지 관리해야 하는 번거로움이 있다. 또한, 온프레미스 방식 사용 시 서버 사양이 실제로 필요한 사양보다 과하게 산정된다면, 전체 컴퓨팅 자원 중 일부밖에 활용되지 않아 자원 사용의 효율성 문제를 낳고 이는 예산 낭비로 직결된다. 반면에 클라우드 방식을 채택한다면 손쉽고 빠르게 플랫폼을 구축할 수 있고, 빠르게 증가하는 데이터에 대한 확장성과 끊임없이 변화하는 요구사항에 대한 유연성을 확보할 수 있다[2].

빅데이터 플랫폼 구축의 어려움과 온프레미스 환경의 단점을 보완하여 본고에서는 클라우드 네이티브 환경에서 컨테이너와 컨테이너 오케스트레이션 기술을 기반으로 Hadoop과 Spark 클러스터를 구축하는 웹 서비스를 제안한다. Hadoop은 대용량 파일을 저장할 수 있는 분산 파일 시스템을 제공하고, 대규모 데이터 처리 OSS 중 가장 범용성이 있다[3]. 하지만, Hadoop은 디스크에서 데이터를 처리하여

* Kyungpook National Univ. doheedev@gmail.com, qorgp13@gmail.com, doyun7433@naver.com, hyeonjin4870@naver.com, sjkoh@knu.ac.kr
** DataStreams Corp, dwlee@datastreams.co.kr

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2021-0-01082)

속도가 다소 느리다. 이 단점을 보완하여 인메모리 환경에서 데이터를 빠르게 분석하는 Spark와 함께 사용하였다.

기존 시스템과의 차별성은 웹페이지의 대시보드에서 Hadoop/Spark 클러스터에 대해 사용자 간 초대 기능을 제공하여 클러스터를 공동 관리하고 협업을 가능케 하는 것이다.

이를 위해 쿠버네티스를 채택하고 웹 프론트엔드, 웹 백엔드, 쿠버네티스 제어 엔진 파트로 나누어서 해당 시스템을 개발한다.

II. 시스템 설계

1) 시스템 구성도

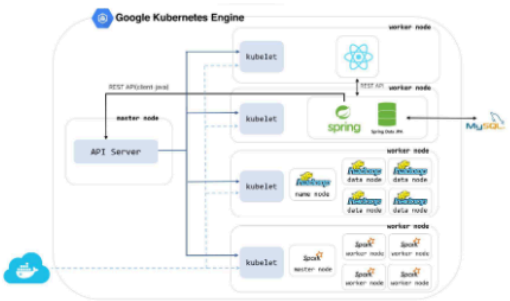

그림 1. 시스템 구성도
Fig. 1. System Architecture

본고에서 구현한 빅데이터 플랫폼은 관리형 쿠버네티스 엔진인 GKE(Google Kubernetes Engine)를 활용하여 클라우드 환경에서 구축하였다. 사용자는 React로 개발된 웹 기반 인터페이스를 통해 빅데이터 클러스터를 생성할 수 있다. 웹 프론트엔드의 요청은 웹 백엔드가 받아서 쿠버네티스 제어 엔진으로 전달한다. 이때, 웹 백엔드와 쿠버네티스 제어 엔진은 Spring Framework를 활용하여 개발되었고, RestTemplate를 통해 서로 통신한다. 쿠버네티스 제어 엔진은 쿠버네티스(GKE) master node의 API Server로 접근하여, 사용자의 요청사항에 맞게 Hadoop/Spark 클러스터를 구축하는 역할을 한다[4].

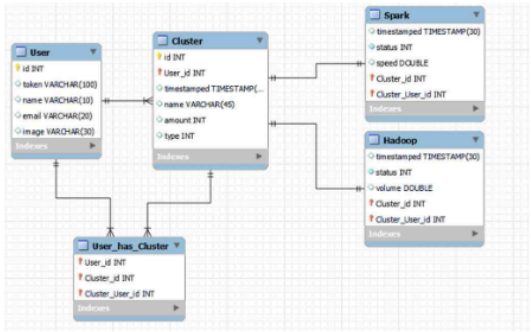

그림 2. 개체-관계 다이어그램
Fig. 2. Entity-Relationship Diagram

2) 데이터베이스 설계

클러스터 생성 시 Hadoop 또는 Spark 중 원하는 클러스터를 선택하고 개수를 입력하면 해당 정보가 데이터베이스에 저장된다. 이때, Hadoop 또는 Spark 클러스터를 구성하는 노드 개수를 3개로 지정하면 master node 1개, slave node 3개가 생성된다.

Spark 클러스터를 생성했을 때는 처리 속도를, Hadoop 클러스터를 생성했을 때는 데이터양을 각각 데이터베이스에 저장해서 시각화를 돕는다.

또한, 클러스터별로 사용자 초대가 가능하여 사용자들은 하나의 클러스터를 공동 관리할 수 있다. 따라서 데이터베이스 상에서 User와 Cluster 필드는 다대다 관계로 구성되어 데이터베이스에서 관리된다.

3) Hadoop/Spark 이미지 생성 과정

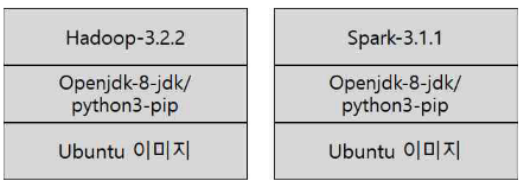

그림 3. Hadoop/Spark 설치 구조
Fig. 3. Hadoop/Spark Installation Structure

먼저 docker Hub로부터 ubuntu 이미지를 받은 후 쿠버네티스 노드에 컨테이너를 띄우고 bash로 접속한다.

컨테이너에 접속했다면 필요한 패키지 및 라이브러리를 설치한다. 위 그림에서는 openjdk-8과 python3-pip를 설치했다. 그리고 Hadoop 및 Spark 바

이너리 파일을 다운로드해서 환경 변수를 설정해준다.

모든 설정이 완료되면 컨테이너를 이미지로 만들어서 사용한다.

III. 구 현

본 장에서는 웹을 통해 클러스터의 정보를 입력받고, 쿠버네티스를 통해 빅데이터 클러스터가 생성되는 과정을 순서대로 열거한다.

우선, 사용자는 빅데이터 프레임워크의 종류를 선택한 후, 클러스터의 이름과 개수를 지정한다. 입력한 정보와 사용자에 대한 정보는 REST API를 통해 웹 서버로 전송된다. 이때, 서버는 사용자 정보를 Session을 통해 획득한다.

그림 5-1. 생성된 worker node들
Fig. 5-1. Created worker nodes

그림 5-2. Deployment 정보
Fig. 5-2. Deployment Information

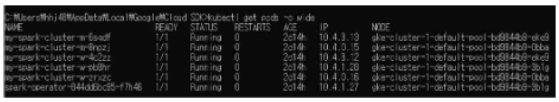

그림 5-3. Pod 정보
Fig. 5-3. Pod Information

Google에서 제공하는 google cloud SDK shell을 이용하여 생성한 worker node 및 오브젝트들을 보여주겠다.

그림 5-1은 GKE를 이용하여 생성한 쿠버네티스 클러스터 내의 모든 노드들을 보여준다.

앞서 웹을 통해 입력받은 Hadoop 및 Spark 클러스터의 개수 정보를 이용해서 YAML 파일을 생성하여 deployment 해두었다. 이때 spark라는 이름의 namespace를 두어 독립적인 공간을 할당해 주었다.

그림 5-2는 spark namespace 내의 deployment 정보를 알려준다. Spark-operator라는 이름으로 하나의 컨테이너가 생성된 것을 알 수 있다.

그림 5-3은 생성된 pod들의 정보를 보여준다. 각 pod들이 그림 5-1에서 보여주었던 3개의 노드에 적절히 분배되어있는 것을 확인할 수 있다.

IV. 결 론

본 논문에서는 쿠버네티스를 기반으로 master node의 API Server와 통신하는 제어 엔진을 구현하고 Hadoop/Spark 클러스터를 구성 및 공동 관리하는 서비스를 개발했다. 이 시스템을 통해 온프레미스 환경에서 빅데이터 분석 환경 구축 시 발생하는 문제점을 해소하고 나아가 전문지식이 없는 사용자도 간편하게 빅데이터 클러스터를 생성하며 여러 사용자가 공동으로 관리하는 방법을 제안한다.

이 시스템은 편리한 빅데이터 관리환경을 제공하기 때문에 '유기견 입양 플랫폼', '전기차 충전소 위치 알리미' 등과 같은 빅데이터를 사용하는 프로그램에 활용될 것으로 기대된다.

참 고 문 헌

[1] John Rydning, Michael Shirer. "Data Creation and Replication Will Grow at a Faster Rate than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts." IDC. March 24, 2021. https://www.idc.com/getdoc.jsp?containerId=prUS47560321

[2] 이강표. (2018). 빅데이터 분석 플랫폼, 어디에 구축할 것인가?. 연구방법논총, 3(2),

[3] 이현종. (2012). 빅데이터 하둡 플랫폼의 활용

[4] 서동우, 김명일, 박상진, 김재성, 정석찬. (2019). 엔지니어링 서비스 지원을 위한 클라우드 기반 빅데이터 플랫폼 개발 연구