spark 구축

내용유무	☑
FE/BE/INFRA	data
작성자	승엽 종호
중요도	⭐⭐⭐⭐⭐

spark 설치

```
wget https://dlcdn.apache.org/spark/spark-3.4.0/spark-3.4.0-bin-hadoop3.tgz

tar -zxvf spark-3.4.0-bin-hadoop3.tgz
```

설정파일

- spark-env.sh

```
# Options read in any cluster manager using HDFS
# - HADOOP_CONF_DIR, to point Spark towards Hadoop configuration files
43 export HADOOP_CONF_DIR=$HADOOP_CONF_DIR

# Options read in YARN client/cluster mode
# - YARN_CONF_DIR, to point Spark towards YARN configuration files when you use YARN
47 export YARN_CONF_DIR=$HADOOP_CONF_DIR
```

- spark-defaults.conf

```
spark.master  yarn
```

트러블슈팅

1. master / workers 로 나누어서 작업을 하려 했고 직접 구축하였었지만, standalone방식은 보안과 관련된 이슈도 있을 뿐 아니라, 대화형 쿼리를 사용하지 않고, 실제 현업 환경과 비슷하게 구성하여 작업하기 위해 yarn 방식 채택
    a. standalone vs yarn
        i. standalone
            1. 간편하게 구축 가능
            2. 설정 및 관리가 쉬움
            3. 대화형 쿼리 지원

ii. yarn(Yet Another Resource Negotiator)

1. Hadoop ecosystem의 일부인 클러스터 관리자
2. 동일한 클러스터에서 실행되는 여러 애플리케이션에 걸쳐 리소스 관리 가능
    a. 대규모 환경에 적합

b. master, slave 구축 방법

i. master, slave에 각각 spark 설치

ii. master, slave에 각각 환경설정

- spark-env.sh

```
export SPARK_WORKER_INSTANCES=2
# Options read in any cluster manager using HDFS
# - HADOOP_CONF_DIR, to point Spark towards Hadoop configuration files
export HADOOP_CONF_DIR=$HADOOP_CONF_DIR
# Options read in YARN client/cluster mode
# - YARN_CONF_DIR, to point Spark towards YARN configuration files when you use YARN
export YARN_CONF_DIR=$HADOOP_CONF_DIR
# Options for the daemons used in the standalone deploy mode
# - SPARK_MASTER_HOST, to bind the master to a different IP address or hostname
export SPARK_MASTER_HOST=k8b308m.p.ssafy.io
# - SPARK_MASTER_PORT / SPARK_MASTER_WEBUI_PORT, to use non-default ports for the master
export SPARK_MASTER_PORT=9999
export SPARK_MASTER_WEBUI_PORT=10000
```

  master host, master port, 사용할 worker의 instance 개수를 설정해야 함

    - 부가적인 설정으로 cpu 코어수 등 가용할 자원 설정 가능

- workers

```
worker에 사용할 hostname
```

- spark-defaults.conf

```
spark.master  yarn
```

iii. master, slave 시작

- master

```
// spark/bin
start-master.sh
```

- worker

```
start-workers.sh
```

2. 방화벽 문제
    a. 클러스터 서버들과 마스터간에 통신을 할때 포트 문제로 통신이 되지않는 문제점 발생
        i. yarn에 사용되는 여러 포트를 추가하기도 하고 설정을 바꾸어 보며 수정을 해 보았지만 실패
        ii. 클러스터 서버들의 방화벽만 종료해보고 실행
            1. 정상 작동
            2. 추후 사용포트 확인 예정