Hadoop 포팅 매뉴얼

1. 배포 환경

- OS 베이스 이미지: ubuntu:20.04, python:3.8-slim
- JVM 버전: OpenJDK 11 (openjdk-11-jdk-headless)
- Hadoop 버전: 3.3.4
- Spark 버전: 3.5.5 (Hadoop3용)
- Python 버전: 3.8
- IDE 권장: VSCode (Docker, SSH Remote 지원 필요)

2. 빌드 시 사용되는 환경 변수

- 공통 환경 변수 (Dockerfile)

ENV HADOOP_HOME=/usr/local/hadoop

ENV SPARK_HOME=/usr/local/spark

ENV JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64 (마스터, 워커)

ENV JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64 (Flask)

ENV PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$SPARK_HOME/bin:$SPARK_HOME/sbin

- YARN 관련 설정
  - yarn-site.xml 및 mapred-site.xml에 정의되어 있음

3. 배포 시 특이사항

- `docker-compose -p hadoop-cluster -f ./hadoop/docker-compose.yaml up -d --build`를 통해 도커 실행
- 최초 실행시 `hdfs namenode -format`
- `init-hdfs.sh` 스크립트로 초기 디렉토리 자동 생성

- `flask-api`의 `init-hdfs.sh`를 통해 Spark 작업 관련 의존성 압축 및 HDFS 업로드

4. DB 접속 정보 등 프로젝트(ERD)에 활용되는 주요 계정 및 프로퍼티가 정의된 파일 목록

- .env : PostgreSQL 연결 정보

DB_NAME=cocoa

DB_USER=a507

DB_PASSWORD=

DB_HOST=

DB_PORT=5432

- core-site.xml : HDFS 기본 주소 설정
- hdfs-site.xml : Namenode 및 Datanode 저장 디렉토리, replication factor 설정
- yarn-site.xml : YARN 리소스 및 로그 디렉토리, RM 주소 설정
- mapred-site.xml : MapReduce → YARN 연동 및 환경 설정