DataPipeline 최적화 과정(진행중)

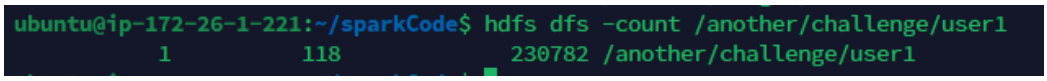
내용유무	☑
FE/BE/INFRA	data
작성자	승엽 종호
중요도	⭐⭐⭐⭐⭐

대략 1시간의 러닝 정보(사용자 정보를 1초마다 수집)를 조회 시 30초 소요

→ 짧은 시간(10분 내외)는 금방 가져왔었지만, 1시간의 데이터의 경우 오랜 시간이 소요되어 개선 필요

개선 시도

- 파일 압축
  - I/O 속도와 관련되어 있다 생각하여 하나의 파일이 받을 수 있는 Row의 갯수 증가시킴

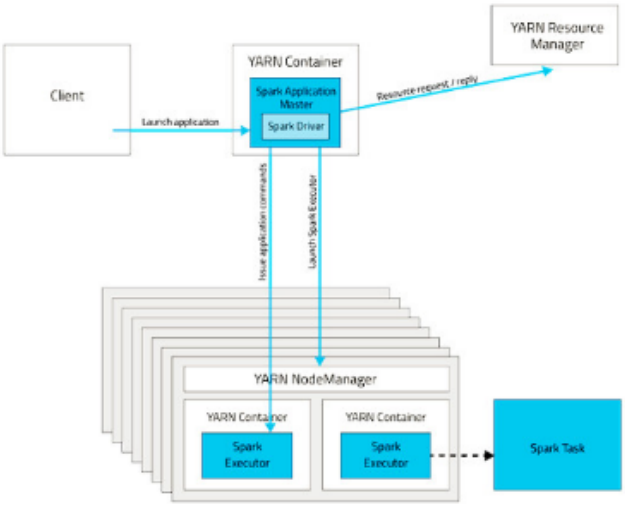


  - 파일 압축을 통해 갯수 줄임

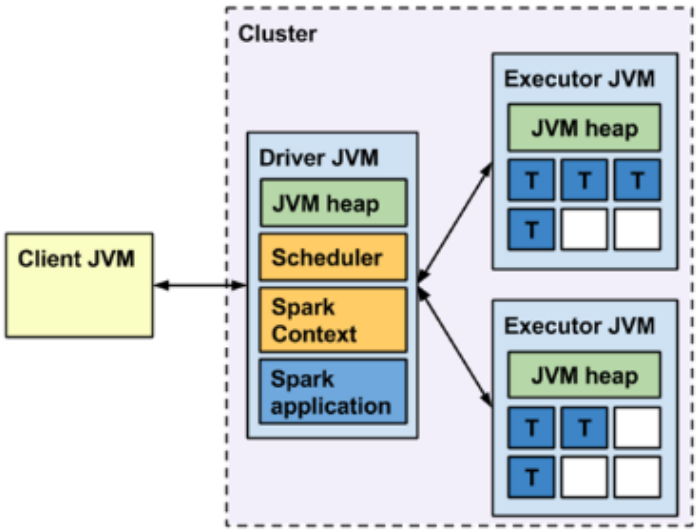
    → 10~15초 소요

    - 시간이 단축되긴 했지만, 개선이 필요함
- repartition을 통해 partition 갯수 조정
  - 한번에 여러 개를 받을 수 있도록 분산 병렬 처리 시도
    - 적정한 partition을 찾고있지만, 기술적 미숙함으로 인해 현재 진행 중
- rest api 서버에서 스파크를 이용하지 않는방법도 확인중
- spark 구조 변화(진행 예정)

- yarn-cluster 에서 standalone-cluster 시도 예정
  - yarn-cluster



    - standalone - cluster(Master, Slave)



- end 토픽을 추가해서 종료시에만 연산이 진행 될 수 있도록 하여 쓸데없는 비용을 줄여 성능을 높였다.