1. 서론

(Alluxio는 Big Data를 위한 빠른 가상 스토리지 이다.) Tachyon으로도 알려진 Alluxio는 메모리 중심적 가상 분산 스토리지 오픈 소스 시스템으로, 데이터 접근을 메모리 속도로 가능하도록 해주는 안정적인 스토리지이다. Alluxio는 파일 API와 커맨드라인 toolkit을 사용해 host의 RAM 공간에 접근할 수 있도록 한다. 분산 프레임워크인 Apache Spark는 Alluxio를 사용해서 데이터 접근 속도를 증가시킬 수 있다. 분산 프레임워크인 Apache Spark의 경우 많은 양의 데이터를 처리하는 데에 성능이 매우 중요하다. 때문에, Spark의 성능을 도와주는 Alluxio를 데이터 prefetching을 통해서 더 좋은 성능을 내도록 하는 작업이 매우 중요하다.

Alluxio는 계층적 스토리지를 제공하는데, 높은 계층일수록 데이터를 읽어오는 속도가 빠르나, 용량이 적다. 많은 양의 데이터를 처리하는 데에 데이터를 읽어오는 속도는 전체 성능에 큰 영향을 끼치기 때문에 Alluxio내에서 사용할 데이터를 미리 높은 계층으로 prefetch하는 것은 전체 어플리케이션의 성능 향상에 큰 도움이 될 것이다. Data를 읽는 시간이 오래 걸리는 SSD나 HDD에 저장된 데이터를 어플리케이션과 비동기적으로 MEM으로 읽어오도록 코드를 구성하였다. Spark에서 application을 실행할 때 각 job에 필요한 data들을 Alluxio에 미리 prefetching을 해 놓음으로써, Spark의 성능을 최적화한다. Data prefetching 없이 Alluxio를 사용했을 때와의 성능 비교를 통해 Data prefetching이 Spark의 성능을 얼마나 증가시키는 지를 실험을 통해 알아낼 것이다.

기존의 문제점 보완할 점.