



Apache Spark

- 빅데이터 워크로드에 사용
- 오픈 소스 고속 통합 분석 엔진
- 데이터 처리 분야에서 가장 규모 큰 오픈소스 project
- 인메모리 기반의 데이터 처리
- 다양한 데이터 저장소 & 생태계 잘 구축
- 배치 및 실시간 처리

- Spark Core API(일반 실행)
 - Spark 플랫폼의 기본 일반 실행 엔진
- Spark SQL + DataFrame (구조화된 데이터)
 - 구조적 데이터 처리를 위한 Spark 모듈
- Stream (Stream분석)
 - Spark의 사용 편의성 & 내고장성을 그대로 활용하면서도 Stream 데이터와 과거 데이터에 강력한 인터랙티브 분석 app을 지원



내고장성이란?

: 시스템의 일부가 고장이 나도 전체에는 영향을 주지 않고, 항상 시스템의 정상 작동을 유지하는 능력

- MLlib (머신러닝)
 - 확장가능한 머신러닝 라이브러리
 - 고급 알고리즘 & 빠른 속도 제공

- GraphX (그래프 계산)
 - Spark를 기반으로 한 그래프 계산 엔진
 - 사용자가 대규모의 구조화된 그래프 데이터를 상호작용 방식으로 구축, 변환, 추론 하도록 지원
- 장점
 1. 속도
 - 여러 병렬 작업으로 데이터를 메모리에 캐시 → 빠른 속도
 2. 실시간 스트림 처리
 - 다른 프레임워크와 통합
 3. 통합 엔진 (여러 워크로드 지원)
 4. 사용 편리성 증가

등장배경

- MapReduce 형태의 클러스터 컴퓨팅 패러다임의 한계를 극복하고자 등장
- 이는 성능이 좋지 x
 - ⇒ In-memory 연산 통해 처리 성능을 향상시키기 위해 Spark 등장



MapReduce란?

-Disk로부터 데이터를 읽은 후,
 Map을 통해 흩어져 있는 데이터를 Key-Value 형태로 연관성 있는 데이터끼리 묶은
 후에,
 Reduce를 하여 중복된 데이터를 제거하고, 원하는 데이터로 가공하여 다시 Disk에
 저장