

# 1-4. Spark란? 왜 Python인가?

· 행보의 어큰 라이터는?

1. SQL Database 이용 행 대신 하드드라이브.

2. 분산시스템. 여러 컴퓨터와 컴퓨터로 데이터 분배.

여기서! 스파크가 쓰인다.

· 분산컴퓨터는 확장이 쉽다.

→ 단순히 컴퓨터를 추가하면 된다.

· 네고장성.

→ 분산시스템은 한 컴퓨터가 고장나도 네트워크를 통해 다른 컴퓨터를 구동할 수 있지만 로컬컴퓨터는 개선화된 기에서 오류가 생기면 별다른 방법이 없다.

분산시스템에선 한 컴퓨터가 고장나도 데이터의 처리와 분산이 문제가 없다.

## Hadoop.

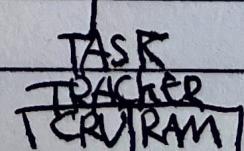
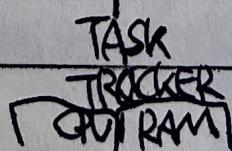
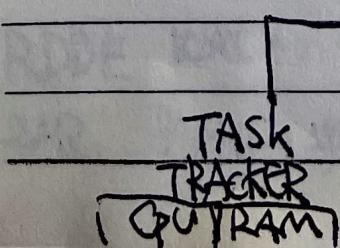
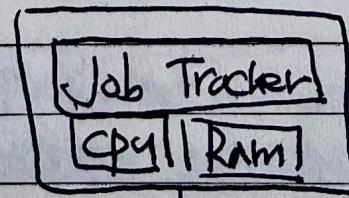
· 방대한 파일을 여러 컴퓨터에 분산시키는 풍선.

HDFS 사용 (대용량 데이터를 분산 시스템으로 처리)

1 데이터 물체를 복제해 네고장성을 높여, 한 컴퓨터가 고장이나도 다른 컴퓨터에 복제된 데이터 물체로로 전송이 가능하라.

1 MapReduce 활용. MapReduce는 분산된 작업이 연산될 수 있게 한다.

→ HDFS는 방대한 양의 데이터를 분산하고 MapReduce는 분산된 Dataset 처리를 가능케한다.



SPARK UCH 2013. 2. 오픈소스

스파크는 MAPREDUCE의 대안

MAPREDUCE - SPARK의 관계.

VS

· 다양한 기능이 구현된 API를 사용 가능하다.

e.g. 가상드라, AWS S3, HDFS (민파일시스템).

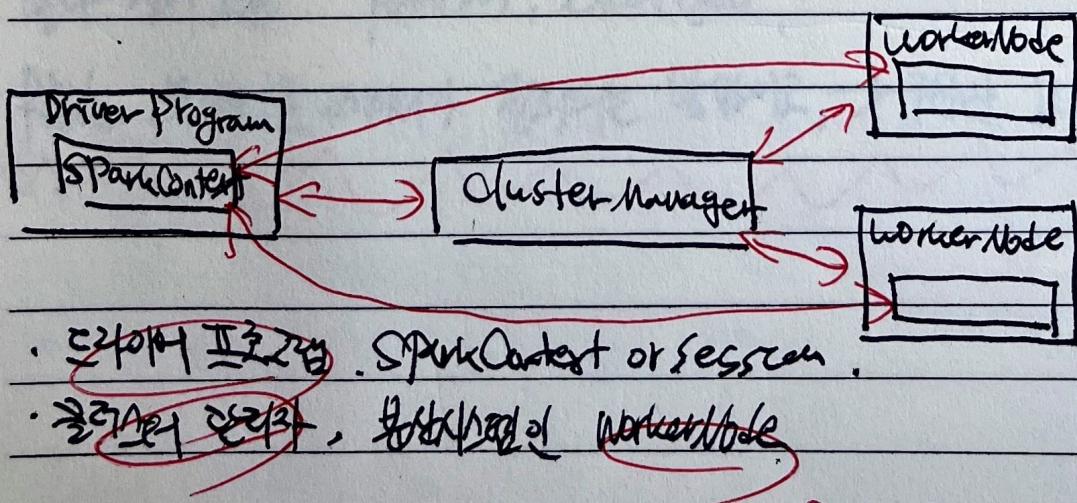
! 스파크는 MAPREDUCE보다 100배 더 빠르게 운용할 수 있다.

어떻게?

(1) MAPREDUCE는 Map과 Reduce의 단계가 따로 이루어지지만 스파크는 메모리에 데이터를 기록한다.  
따라서 병렬 처리로 작동하여 메모리가 가득 찬 경우에는 데이터를 디스크에 기록한다.  
Spark의 방식이 다르다. (맵기록작업)

(2) RDD (회복성 분산 데이터)

· 분산된 데이터의 핵심, 내구성, 데이터의 복원(데이터를 복구할 수 있는 능력)  
· HDFS 이외에도 많은 데이터 소스를 쓸 수 있다.



RDD는 변환X, 자\_인턴운산, 개성 가능?

그리고, 변환과 같은 2가지 행동이 있다.

번화: 우리가 알아야 할 대사피.

액션: 그 대사피를 실행으로 옮기는 행동 그 자체!

비서도 출시 PySpark을 활용해 데이터를 불러오는데...  
액션을 추적하기 실제 경우 어떤 결과도 없다.

BigData에서 동작됨. Why?

큰 용량의 파일을 처리하기에 불이익한거요. 여기서 이유에는 번화작업은 모두 연산할 필요성이 놓여야.

→ 번화작업의 흥미는 데이터의 평균값, 혹은 특정 데이터의 값의 개수를 세는 것에 있다.

→ 이런 풀이 그의 숫자보다 큰 값을 가지고 있는지 알아보는 작업일 수 있다.

예 필요한 데이터만을 원하면 되기 때문에 이 모든 번화작업이 속도 연산될 필요는 없다. 아주 방대한 힘.

1. 어떤 값을 학습할 때와 같이 학습률로 연산하면 우리가 간다.

2. 번화와 액션을 나누어 호출이 보면 처리된다.

1과 2. 0이 나온다. 데이터와 함께 기반의 연산을 처리하는 제로 간다.

예외 케이스는... pandas, tensorflow.

Pandas는 물리적으로 데이터가 처리되는 방식이고, → 구현은 DataFrame!

## \* RDD

· 아파치 스파크의 데이터 처리를 위한 추상화된 데이터 구조.

분산화된 데이터의 모음.

인라인ais 쓰기, 병렬성, 병렬 가능성, 자연 계산 ..

### 1. RDD로 데이터 처리 :

스파크에서 데이터 처리를 위해 RDD를 사용한다.

RDD를 통해 데이터를 로드하고 변환하는 작업을 수행한다.

log data를 RDD로 로드하고 필요한 정보를 추출, 변환.

### 2. 변환 및 작업 수행.

RDD를 이용하여 Transformation을 수행.

Map, Filter, Reduce 등의 변환 연산을 사용하여 데이터 가공. 합계.

### 3. 액션 연산 수행.

Action 연산은 실제로 결과를 반환하는 연산이다.

### 4. HDFS에 저장 : 빛전 연산을 통해 처리한 데이터나 결과는 파일이나 HDFS에 같은 분산 파일 시스템에 저장가능.

\* PySpark에서 데이터를 불러오는 작업은 transformation이 포함된다.

데이터를 불러오는 작업 자체는 스파크가 실행되는 중에는 실제로 데이터를 불러오는

것이 아닌, 데이터 처리 계획을 구현하기 위한 단계다.

Action을 실행하지 않으면 어떤 결과도 나타나지 않는다.

data.show() → 데이터 빛전 연산 수행.