3 주차 Homework			
학번	32151671	이름	박민혁

※ 문제를 해결하여 제출하시오. 필요시 실행 결과를 캡쳐하여 붙여넣는다.

[Python 복습] 주피터를 활용해 결과를 구하시오.

1. 문자열 'Big data processing'를 변수에 저장하고 그것을 한 문자씩 출력하는 프로그램을 작성하시오(1 점).

Answer:

```
Font: consolas / Font size: 10

String = ‘Big data processing’
for x in String:
    print(x)
```

```
In [12]: String = 'Big data processing'
         for x in String:
             print(x)

         B
         i
         g

         d
         a
         t
         a

         p
         r
         o
         c
         e
         s
         s
         i
         n
         g
```

2. 1 부터 50 까지 자연수의 합을 구하고 결과를 출력하는 프로그램을 작성하시오(2 점).

Answer:

```
Font: consolas / Font size: 10

result = 0
for x in range(1, 51):
    result += x
print(result)
```

```
In [13]: result = 0
         for x in range(1, 51):
             result += x
         print(result)

         1275
```

3. 피보나치 수열을 구하는 함수를 정의하고, 그 함수를 이용해 수열의 40 번째 값을 구하시오. 피보나치 수열은 첫째 항과 둘째 항이 1 이고 세번째 항부터는 바로 앞의 두 항의 합인 수열이다(1, 1, 2, 3, 5, 8, 13, ...). (3 점)

Answer:

```
Font: consolas / Font size: 10

// 재귀 구현
def fibo(x):
    if x == 1 or x == 2:
        return 1
    return fibo(x-1) + fibo(x-2)

print(fibo(40))

// dp 테이블 이용
def fibo():
    for i in range(3, 41):
        dp[i] = dp[i-1] + dp[i-2]

dp = [0] * 41
dp[1] = 1
dp[2] = 1
fibo()

print(dp[40])
```

```
In [40]: def fibo():
             for i in range(3, 41):
                 dp[i] = dp[i-1] + dp[i-2]

         dp = [0] * 41
         dp[1] = 1
         dp[2] = 1
         fibo()

         print(dp[40])

         102334155
```

[3 주차 확인 학습]

1. 하둡과 하둡 에코시스템에 대해 중학생 동생에게 설명한다고 가정하고, 자신의 언어로 설명해 보시오. (4 점)

Answer:

Font: 맑은 고딕(한글) / Font size: 10

하둡을 배우기 전에, Big Data 시대가 어떻게 됐는지에 대해서 알아야한다.
데이터가 폭발적으로 증가하였고, 데이터 유형이 구조적 데이터에서 비정형 즉 반구조적 데이터가 많아졌고, 데이터 특성이 다양해지고 복합적이며 실시간으로 변화한다. 그래서 이것들을 처리하기 위해 하둡이 필요하게 되었으며 하둡이 탄생하게 된 이유이다.

한명의 사용자가 하나의 DB 시스템에 요청을 하면 속도는 빠르다. 하지만 여러명의 사용자가 요청을 하게 되면 부하가 걸릴 것이고 느려질 것이다. 그래서 사용자가 직접 DB 에 요청을 하거나, 하둡을 이용해 요청을 하면 빨라질 것이다.
하둡이란 간단한 프로그래밍 모델을 사용하여 여러대의 컴퓨터 클러스터에 대규모 데이터 셋을 분산 처리 할 수 있게 해주는 프레임 워크이다. 단일 서버에서 수천대의 머신으로 확장 할 수 있도록 설계되었다. 일반적으로 하둡파일시스템과 맵리듀스 프레임워크로 시작 되었으나, 여러 데이터 저장, 실행 엔진, 프로그래밍 및 데이터 처리 같은 하둡 생태계 전반을 포함하는 의미로 확장 발전 되었다.

하둡의 프로세싱 과정은 다음과 같다.
데이터 수집 -> 데이터 저장 -> 데이터 처리 -> 데이터 분석 -> 데이터 시각화이다. 하둡의 에코 시스템에 대해 설명하겠다.

1. 분산코디네이터

Zookeeper	분산환경에서 서버간의 상호 조정이 필요한 다양한 서비스를 제공하는 시스템

2. 분산 리소스관리

YARN	작업 스케줄링 및 클러스터 리소스 관리를 위한 프레임워크로 맵리듀스, 하이브, 임팔라, 스파크등 애플리케이션들은 얀에서 작업을 실행

3. 데이터 수집

Flume	비정형 데이터 수집
Sqoop	정형 데이터 수집

4. 데이터 저장

HBase	구글 빅데이터 기반으로 개발된 비관계형 데이터베이스이며 하둡 및 하둡파일시스템 위에 빅테이블과 같은 기능을 제공
HDFS	애플리케이션 데이터에 대한 높은 처리량의 엑세스를 제공하는 분산 파일 시스템
Kafka	버퍼 기능이 있어 데이터를 저장

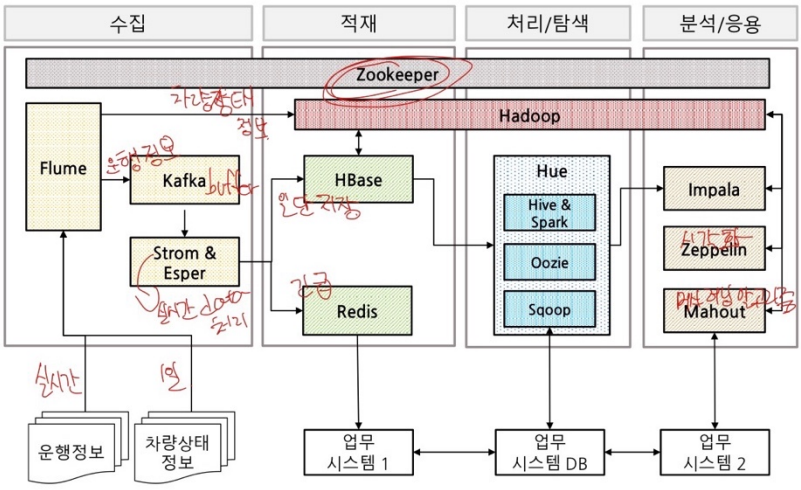
5. 데이터 처리

PIG	구조화, 비구조적 데이터 사용 가능
HIVE	구조화 된 데이터에서 사용 가능
Spark	메모리 기반 데이터 처리, 실시간으로 데이터를 처리(주식시장 그래프)
Mahout	분석 기계학습에 필요한 알고리즘을 구축하기 위한 오픈소스 프레임 워이며, 클러스터링, 추천시스템, 맵리듀스 등 머신러닝 알고리즘을 지원
STORM	실시간으로 데이터를 처리

6. 그 외

Apache Ambari	하둡을 모니터링 하여 상태를 확인
Oozie	하둡 관리

아래 사진은 스마트카 서비스 하둡 에코시스템 예시이다.



54

.