1. 빅데이터의 개요

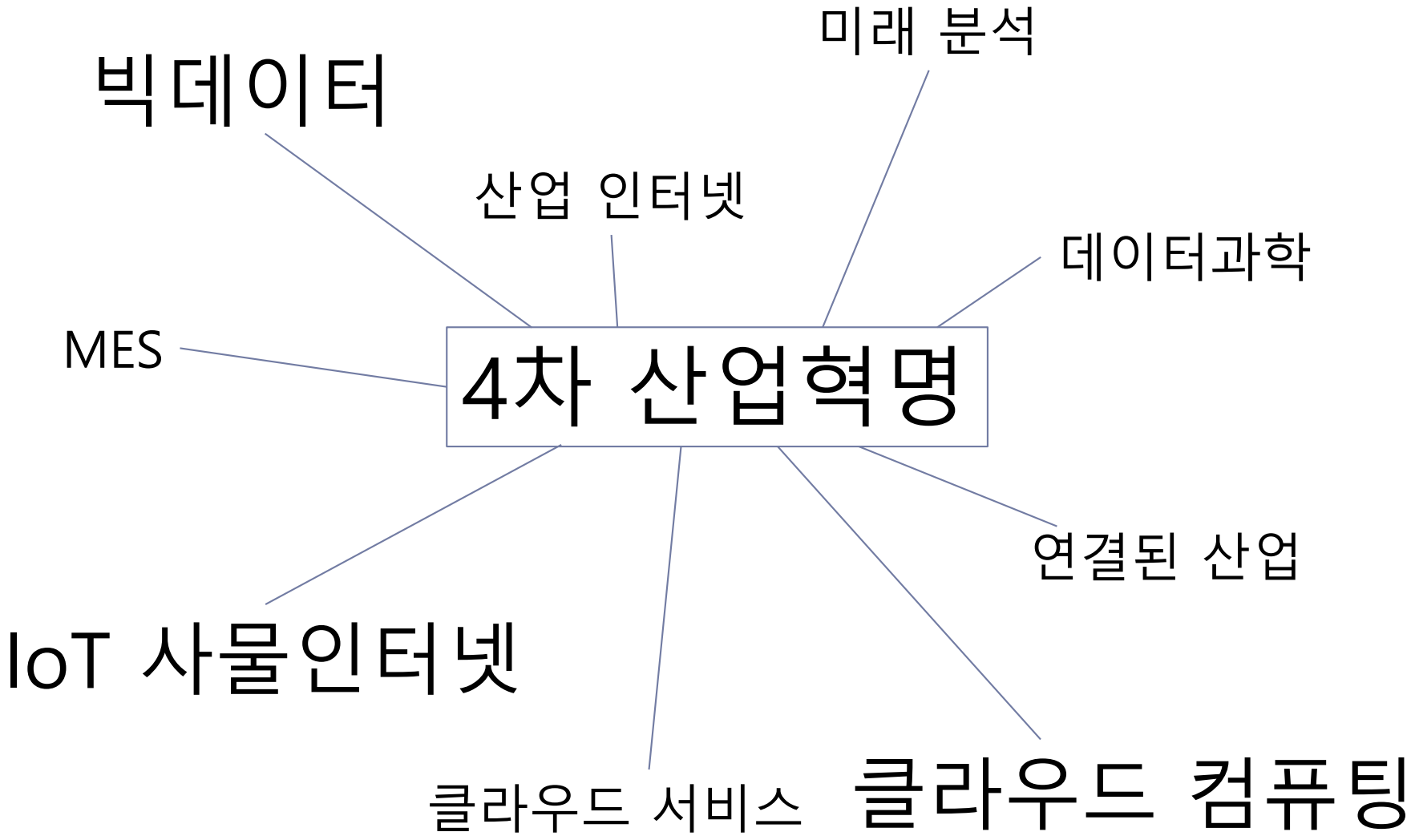
목차

- 4차 산업혁명
- 빅데이터 처리의 필요성
- 전반적인 빅데이터 처리 과정에 대한 이해
- 웹의 구성과 동작에 대한 이해
- 빅데이터 처리를 위한 도구

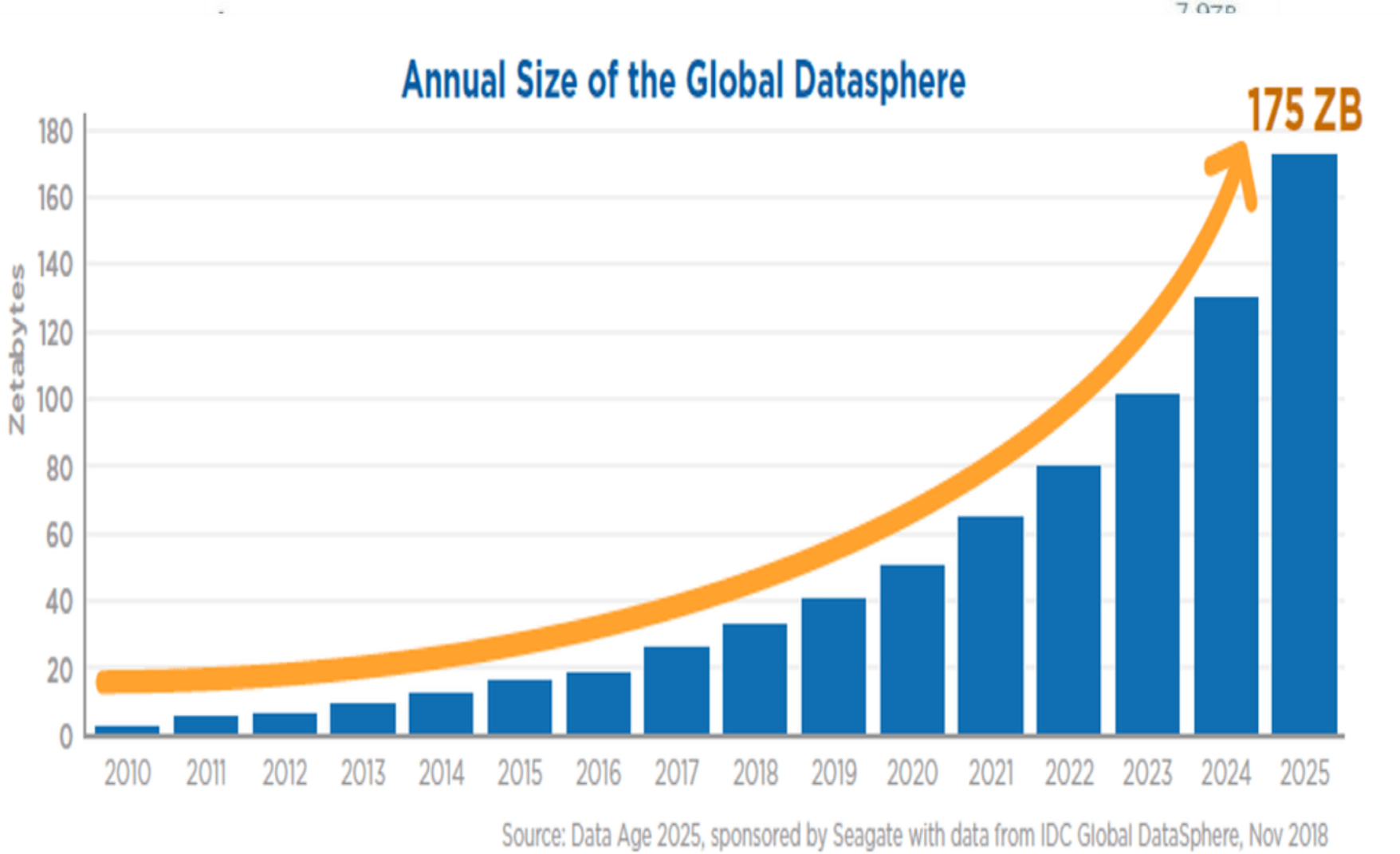
4차 산업혁명

- 기업들이 제조업과 정보통신기술을 융합해 작업 경쟁력을 제고하는 차세대 산업혁명
  - 클라우드 슈밥

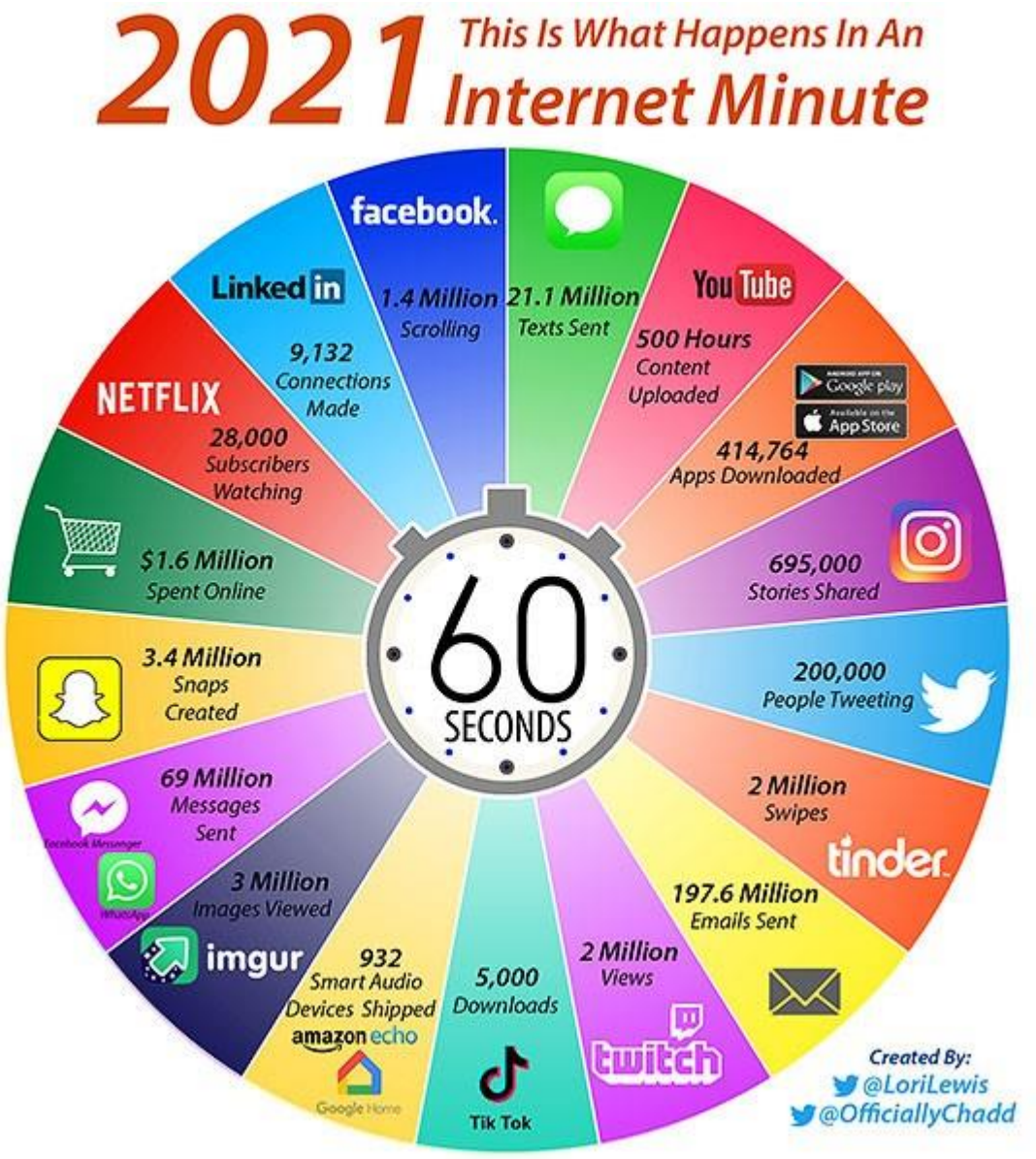
4차 산업혁명


빅데이터
미래 분석
산업 인터넷
데이터과학
MES
4차 산업혁명
연결된 산업
IoT 사물인터넷
클라우드 서비스
클라우드 컴퓨팅

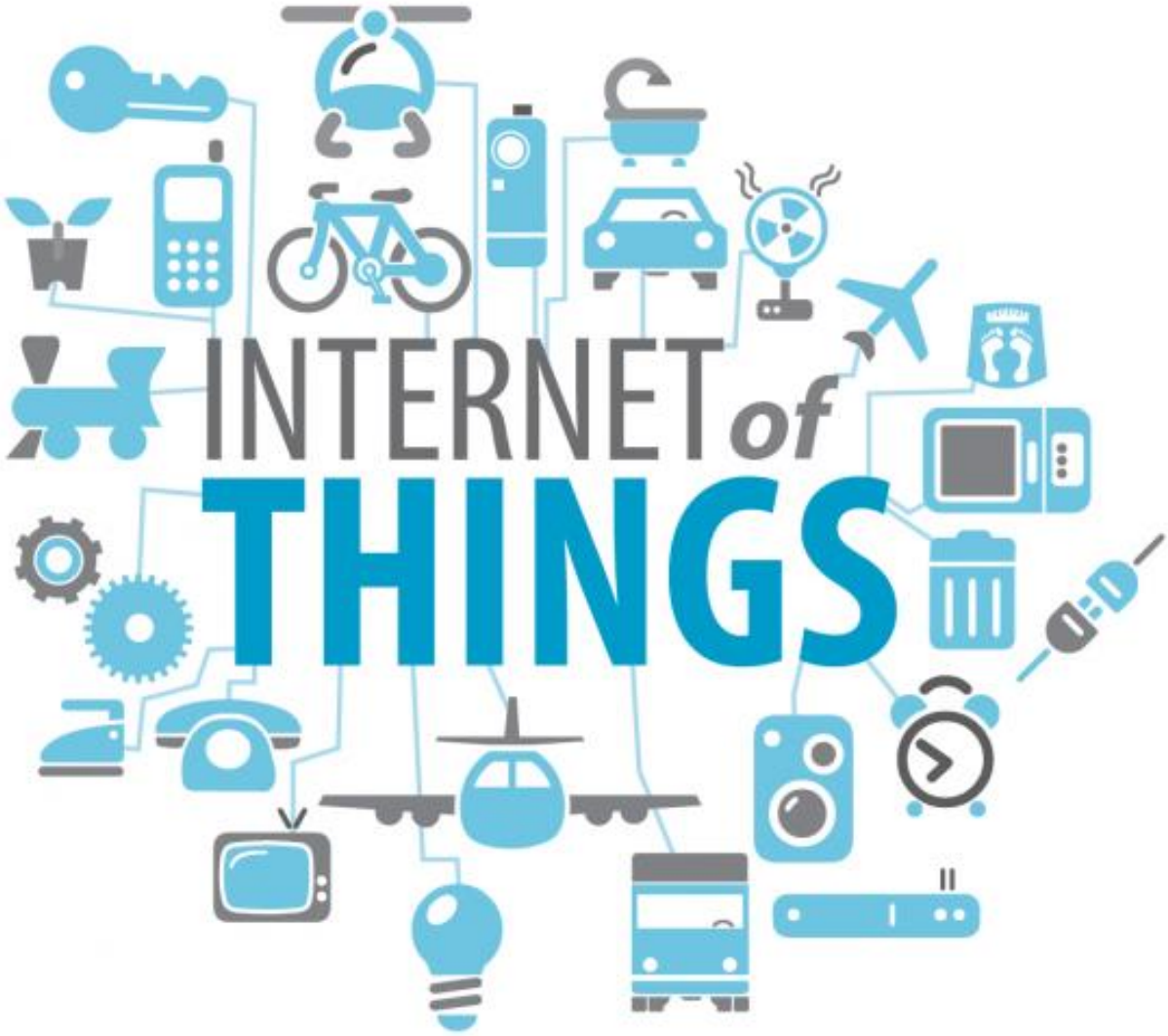
빅데이터 등장 배경


Annual Size of the Global Datasphere
175 ZB
Zetabytes
180
160
140
120
100
80
60
40
20
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

빅데이터 등장 배경


2021 This Is What Happens In An Internet Minute
60 SECONDS
1.4 Million Scrolling
facebook
21.1 Million Texts Sent
500 Hours Content Uploaded
YouTube
414,764 Apps Downloaded
Google play
App Store
695,000 Stories Shared
200,000 People Tweeting
2 Million Swipes
tinder
197.6 Million Emails Sent
2 Million Views
twitch
5,000 Downloads
Tik Tok
932 Smart Audio Devices Shipped
amazon echo
Google Home
3 Million Images Viewed
imgur
69 Million Messages Sent
Facebook Messenger
WhatsApp
3.4 Million Snaps Created
$1.6 Million Spent Online
28,000 Subscribers Watching
NETFLIX
9,132 Connections Made
LinkedIn
Created By:
@LoriLewis
@OfficiallyChadd

빅데이터 등장 배경

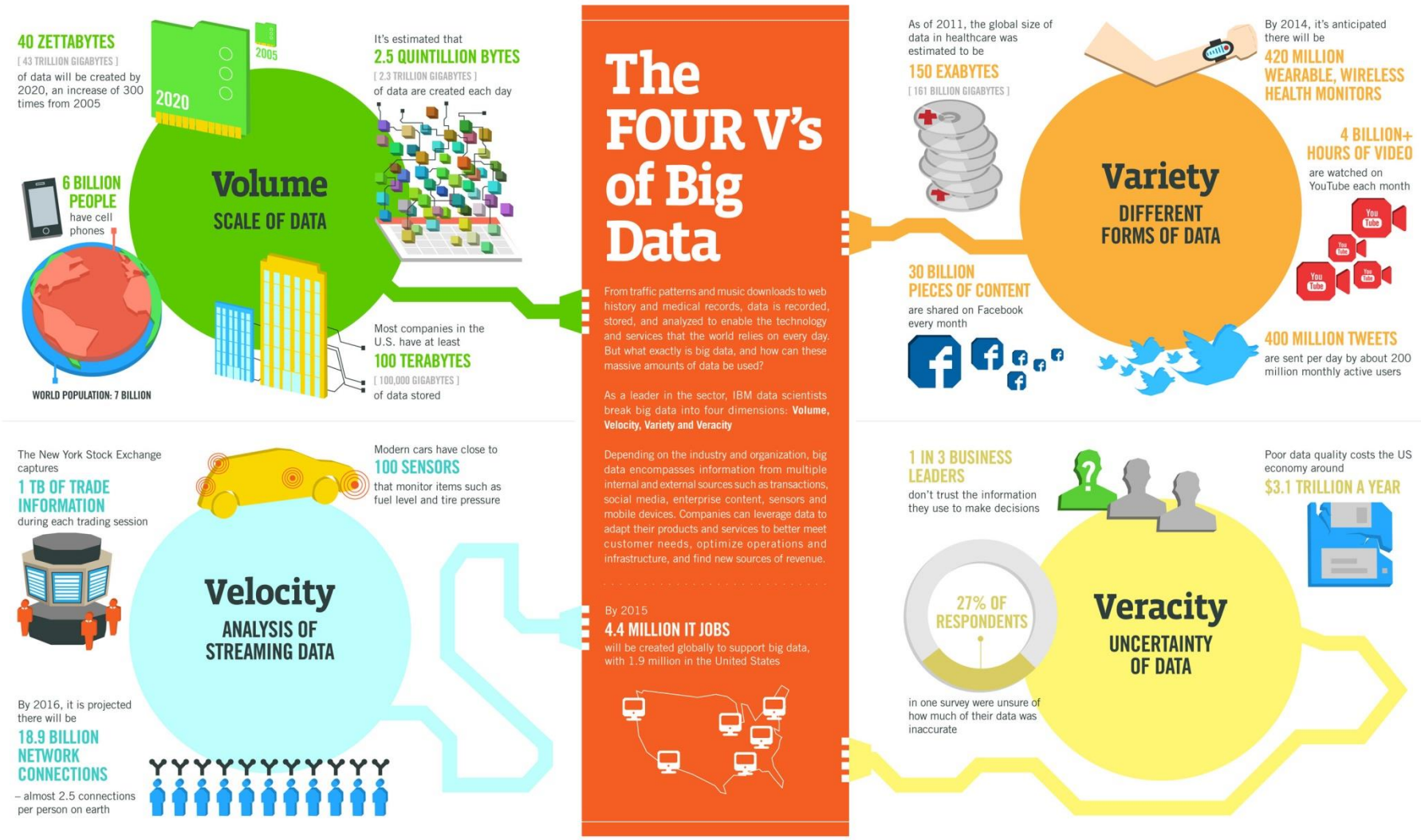


빅데이터란?

빅(Big) + 데이터(Data)

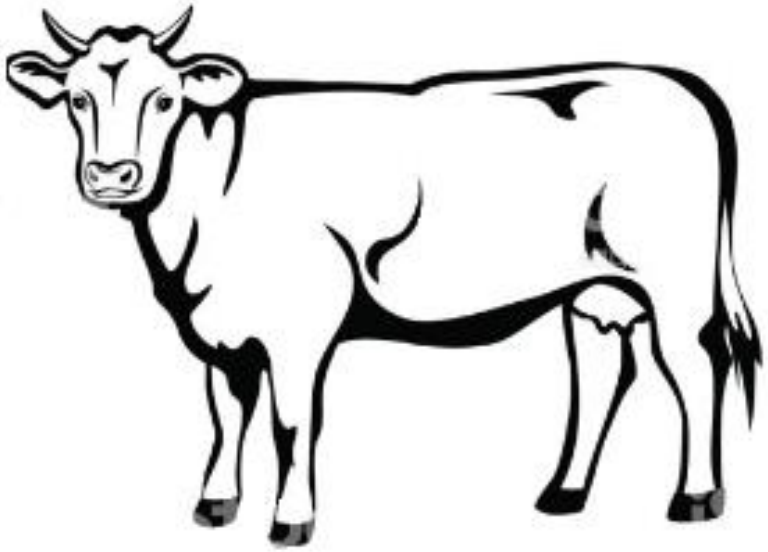
테라바이트 이상의 데이터?
어마어마하게 많은 데이터?

빅데이터의 특징



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM



813 kg

( )kg

빅데이터 처리의 필요성

- 기업과 기관의 요구
- 기업들은 새로운 데이터의 활용을 필요로 한다
  - Social data, Public data, Commercial data 등
- 현재 빅데이터 플랫폼 구축에 혈안이 되어 있다
- 빅데이터 플랫폼을 구성하기 위한 도전 과제는?
  - 통합의 어려움
  - 유지보수의 어려움
  - 신속한 가치 창출의 어려움
  - **사람/조직의 한계**

왜? 데이터 분석을 배워야 하나요?

기술➔ 가격 ➔ 가치

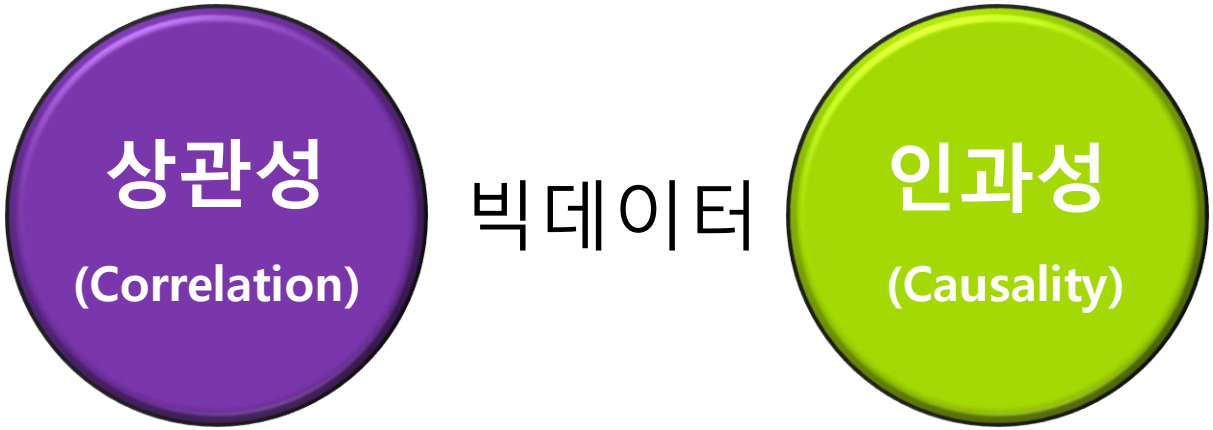
가치는 데이터로부터 나온다!!!

따라서 미래를 보려면 데이터를 보아야 한다

데이터 분석의 목적은 예측을 하기 위함이고

데이터 분석을 통해 얻어 내는 결과물은
미래에 내가 원하는 것을 얻기 위함이다.

주목 받는 데이터 과학



- 과거 -

데이터 저장, 접근법에 초점

언급량에 의한
데이터의 상관성을
중심으로 분석

- 최근 -

인과성에 의한,
왜('why')를 찾고자 하는
목적으로 분석
(패턴을 알아내어,
다음 패턴을 찾기 위한)

빅데이터 처리의 전반적인 과정

웹을 구성 요소

- 클라이언트와 서버
- 웹 문서의 구성 요소
- 웹 브라우저의 작동 방식

웹 구성요소

- 웹
  - World Wide Web(WWW)
  - 인터넷에 연결된 클라이언트들이 정보를 공유할 수 있는 공간
  - 웹 페이지 – 웹에서 보는 문서의 페이지 하나하나
  - 웹 사이트 – 웹 페이지의 모임
  - 인터넷 주소(URL과 IP)를 이용하여 접속
  - 하나의 웹 사이트는 여러 개의 웹 페이지로 구성.
  - HTML, JavaScript, CSS… 등으로 구성

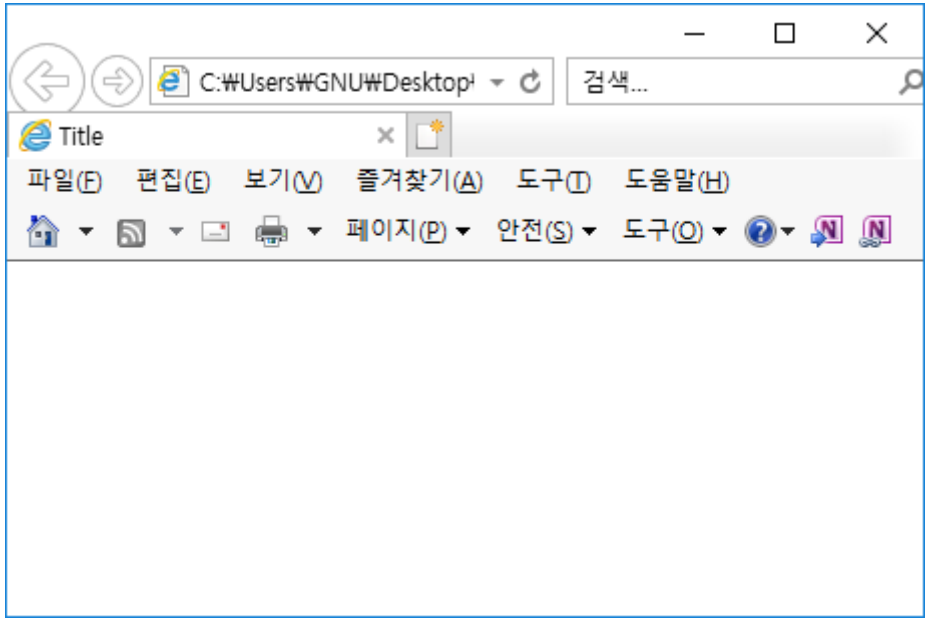
웹 구성요소

- HTML(HyperText Markup Language)
  - 웹 문서를 만들기 위한 프로그래밍 언어
  - 웹을 이루는 가장 기본적인 단위, *.html
  - '< >'(태그)로 감싸져 있음, 여는 태그 < >와 닫는 태그</ > 한 쌍으로 구성

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8">
    <title> Title </title>
   </head>
  <body>

  </body>
</html>
```

실행결과



웹 구성요소

- HTML(HyperText Markup Language)

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8">
    <title> Title </title>
  </head>

  <body>


  </body>
</html>
```

헤더 부분 – 웹 페이지의 정보를 담고…
<meta> <title> <style> <script> <link> …

바디 부분 – 웹 페이지 내용을 담고…
<p> <h>   <table> <input>
<button> <select> <a> <img>
  …

웹 구성요소

- HTML(HyperText Markup Language)
  - <a> 태그 – 다른 페이지로 이동
  - href 속성 – 이동되는 링크 정보

  - 주소 전체를 작성 하는 경우
  - 상대경로(상대주소)를 작성하는 경우
  - 절대경로(절대주소)를 작성하는 경우

 경상국립대학교

웹 구성요소

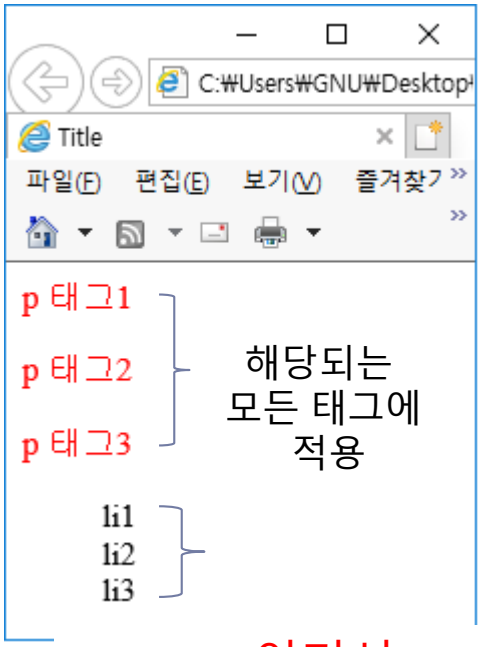
- CSS(Cascading Style Sheets)
  - 웹 사이트를 꾸며주는 역할

  - 데이터 수집을 하는데 왜? 웹 사이트를 꾸며주는 CSS를 알아야 할까?
    - CSS로 웹 사이트를 꾸며주기 위해 해당 태그에 접근하는 방식을
      크롤러에서도 똑같이 사용할 수 있으므로

    - 셀렉터(selector)
      : CSS를 이용하여 꾸미기 위해 특정 요소에 접근하는 것

웹 구성요소

- CSS
  - 셀렉터 없는 경우



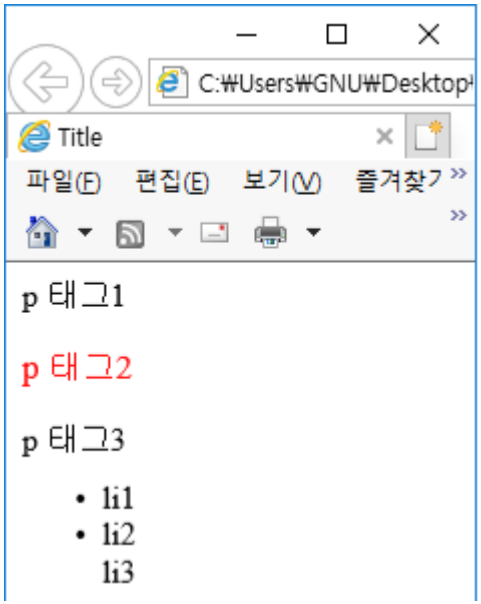
여기서
중요한 것은
태그를 이용해
접근할 수 있다는 것!!!

```
<html>
 <head>
  <meta charset="UTF-8">
  <title> Title </title>
  <style>
        p{
         font-size : 17px;
         color:red;
        }

        li{
         list-style-type: none;
        }
  </style>
 </head>
 <body>
        <p> p 태그1 </p>
        <p> p 태그2 </p>
        <p> p 태그3 </p>
        <ul>
           <li> li1 </li>
           <li> li2 </li>
           <li> li3 </li>
        </ul>
 </body>
</html>
```

웹 구성요소

- CSS
  - 셀렉터 사용



여기서
중요한 것은
원하는 요소에만 정확히
접근할 수 있다는 것!!!

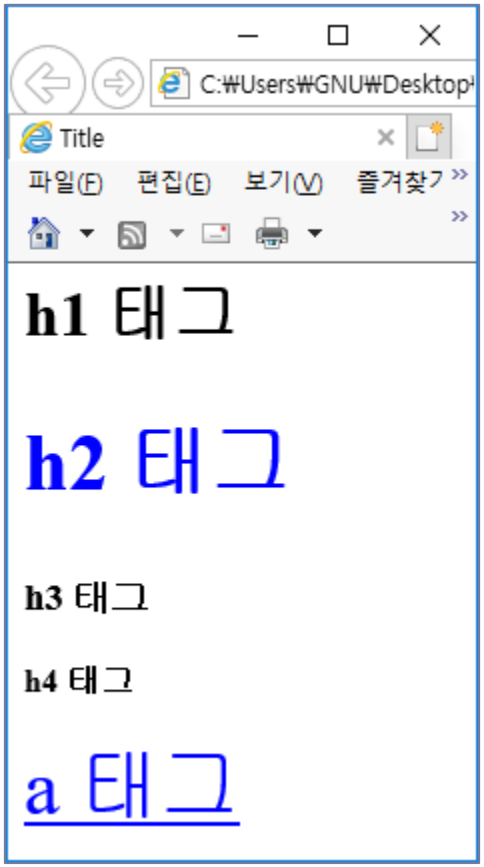
```
<html>
 <head>
  <meta charset="UTF-8">
  <title> Title </title>
  <style>
        p.p-target{
         font-size : 17px;
         color : red;
        }

        li.li-target {
         list-style-type: none;
        }
  </style>
</head>
<body>
        <p> p 태그1 </p>
        <p class="p-taget"> p 태그2 </p>
        <p> p 태그3 </p>
        <ul>
           <li> li1 </li>
           <li> li2 </li>
           <li class="li-taget"> li3 </li>
        </ul>
 </body>
</html>
```

웹 구성요소

class를 이용하여 셀렉터 만들기

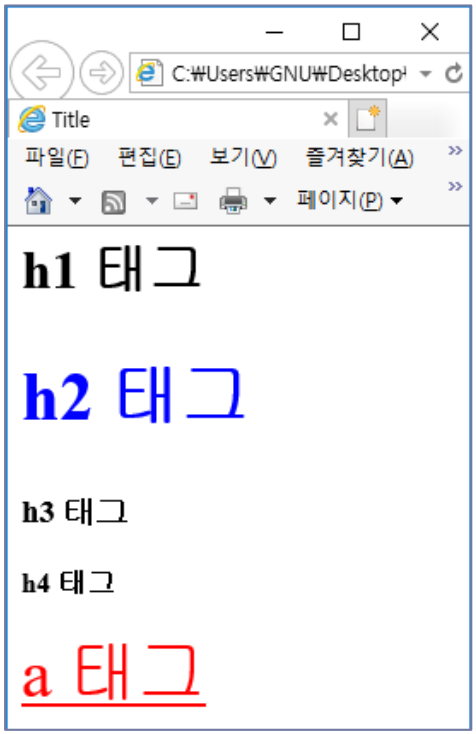
- CSS




```
<html>
  <head>
    <meta charset="UTF-8">
    <title> Title </title>
    <style>
            .target{
             font-size : 40px;
             color : blue;
            }
    </style>
  </head>
  <body>
            <h1> h1 태그 </h1>
            <h2 class="target"> h2 태그 </h2>
            <h3> h3 태그 </h3>
            <h4> h4 태그 </h4>
            <a href="/" class="target"> a 태그 </a>
  </body>
</html>
```

웹 구성요소

- CSS



id를 이용하여 셀렉터 만들기

```
<html>
  <head>
    <meta charset="UTF-8">
    <title> Title </title>
    <style>
            #target1{
             font-size : 40px;
             color : blue;
            }
            #target2{
             font-size : 40px;
             color : blue;
            }
    </style>
  </head>
  <body>
            <h1> h1 태그 </h1>
            <h2 id="target1"> h2 태그 </h2>
            <h3> h3 태그 </h3>
            <h4> h4 태그 </h4>
            <a href="/" id="target2"> a 태그 </a>
  </body>
</html>
```

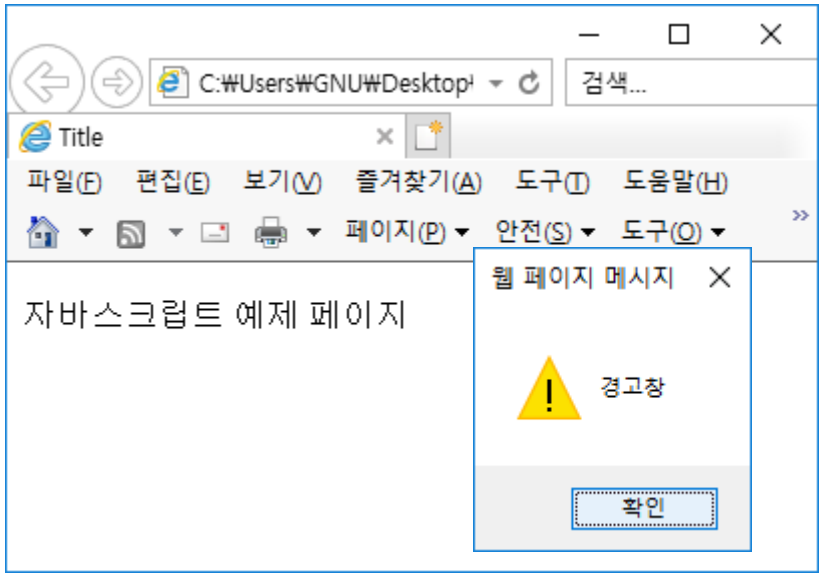
웹 구성요소

- JavaScript
  - 웹 사이트에 원하는 기능을 넣어 줄 수 있음
  - <script> 태그 안에 기능에 대한 코드를 넣는다.
  - <script> 태그는 <head>에 들어가거나, <body>의 가장 하단에 위치.
  - HTML 코드를 만들 수 있음

웹 구성요소

- JavaScript

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8">
    <title> Title </title>
    <script>
            alert('경고창')
    </script>
  </head>
  <body>
    <p>자바스크립트 예제 페이지</p>
  </body>
</html>
```



서버와 클라이언트

- 서버
  - 인터넷을 통해 연결된 클라이언트에게
    데이터 또는 서비스를 제공하는 프로그램
- 클라이언트
  - 데이터, 서비스를 요청하는 프로그램
  - 요청한 데이터를 사용자에게 보여주는 프로그램
- 서버를 어떤 구조로 만들었는지 파악하고 유추하는 것이 중요

서버와 클라이언트

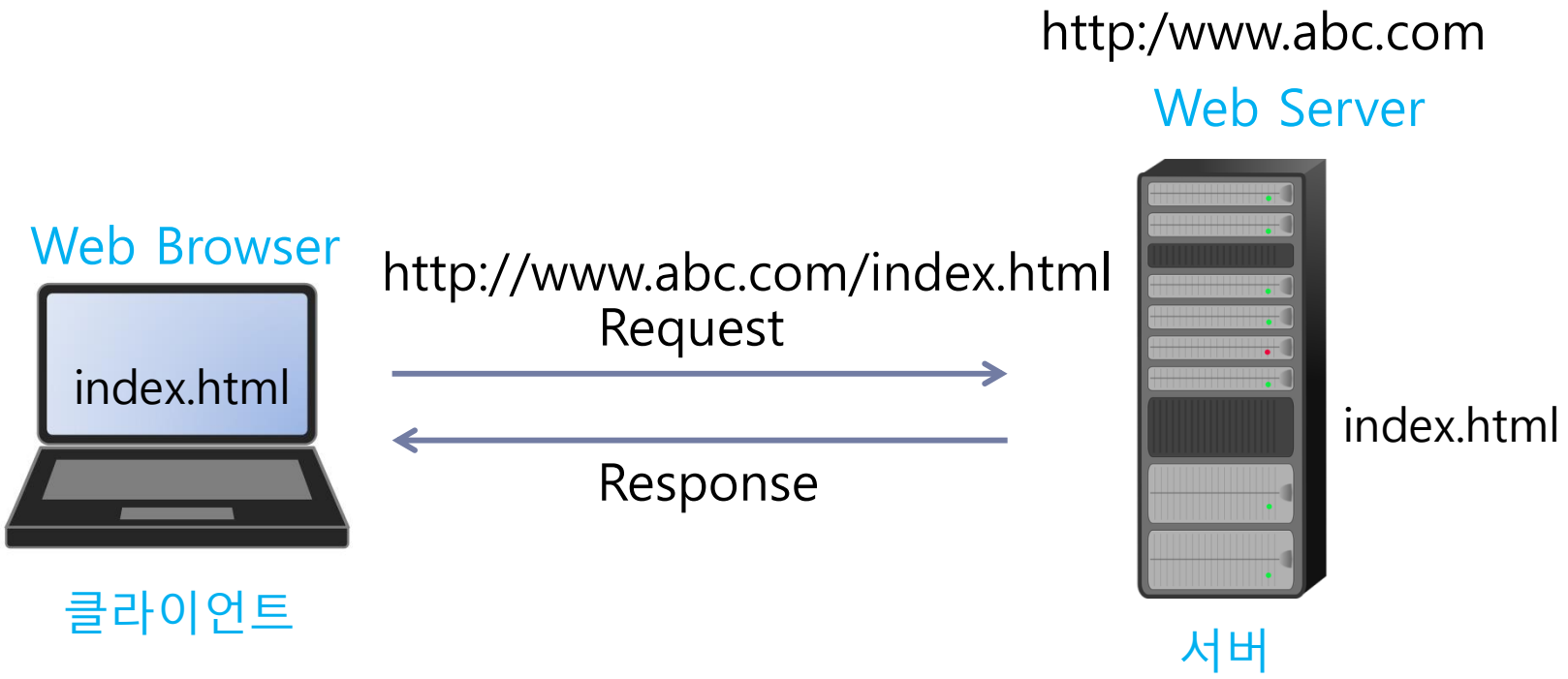
- URL(Uniform Resource Locator)
  - 네트워크 상에서 자원을 요청하는 규약
  - 데이터를 주고 받을 때 URL을 이용하여 데이터를 주고 받음

프로토콜://주소 또는 IP:포트번호/리소스경로?쿼리스트링

인터넷 : http HTTPS, FTP, SFTP, SSH…	www. abc.com 196.255.3.12	http : 80 HTTPS : 443 FTP : 20, 21 telnet : 23 SSH : 22	/score/a	name=kim&kor=98

http://www.abc.com/score?name=kim&kor=98

서버와 클라이언트

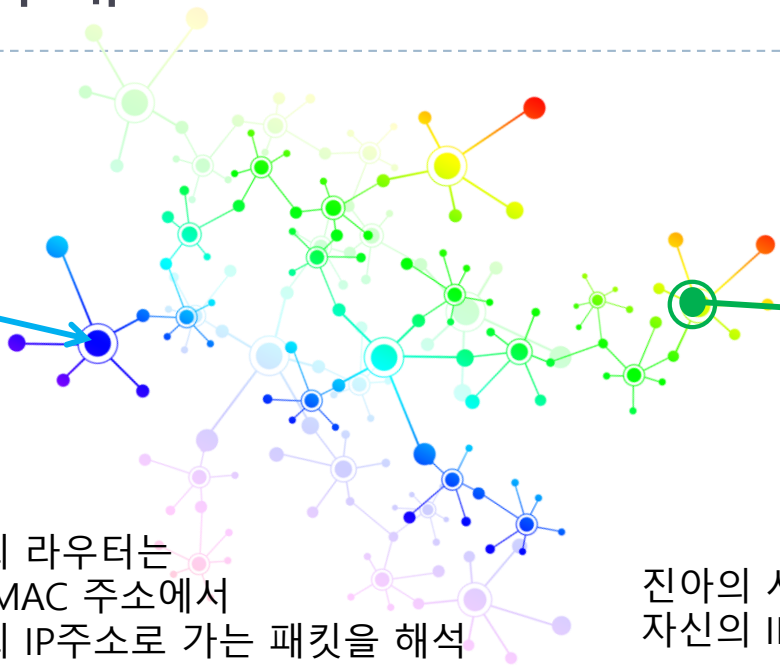

http:/www.abc.com
Web Server
Web Browser
http://www.abc.com/index.html
Request
index.html
Response
index.html
클라이언트
서버

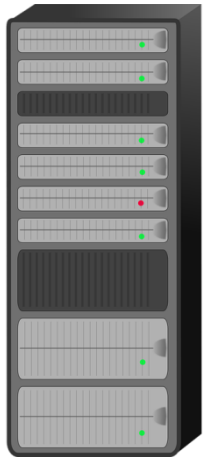
웹 동작의 이해

Web Server

Web Browser

서준





진아

서준의 라우터는
서준 MAC 주소에서
진아의 IP주소로 가는 패킷을 해석

서준의 라우터 고유 IP주소를
패킷에 발신자 주소로 기록 후
인터넷에 보냄

진아의 서버는
자신의 IP주소에서 그 패킷을 받음

패킷 헤더에서 포트 번호를 찾고
적절한 어플리케이션에 보냄.

웹 서버 어플리케이션은
서버 프로세서에서 데이터를 받음

데이터 : GET요청, index.html

html 파일을 찾고 새 패킷으로 묶어서
자신의 라우터를 통해 서준의 컴으로 전송.

서준의 라우터 주소 진아의 IP 주소
110100001010101 100000101001111 101101010000111

패킷(데이터)

서준의 라우터 주소 진아의 IP 주소 서준의 라우터 IP주소
11010000110101010 10000010100111110 1001010100001111

빅데이터 처리를 위한 도구

빅데이터 처리를 위한 도구