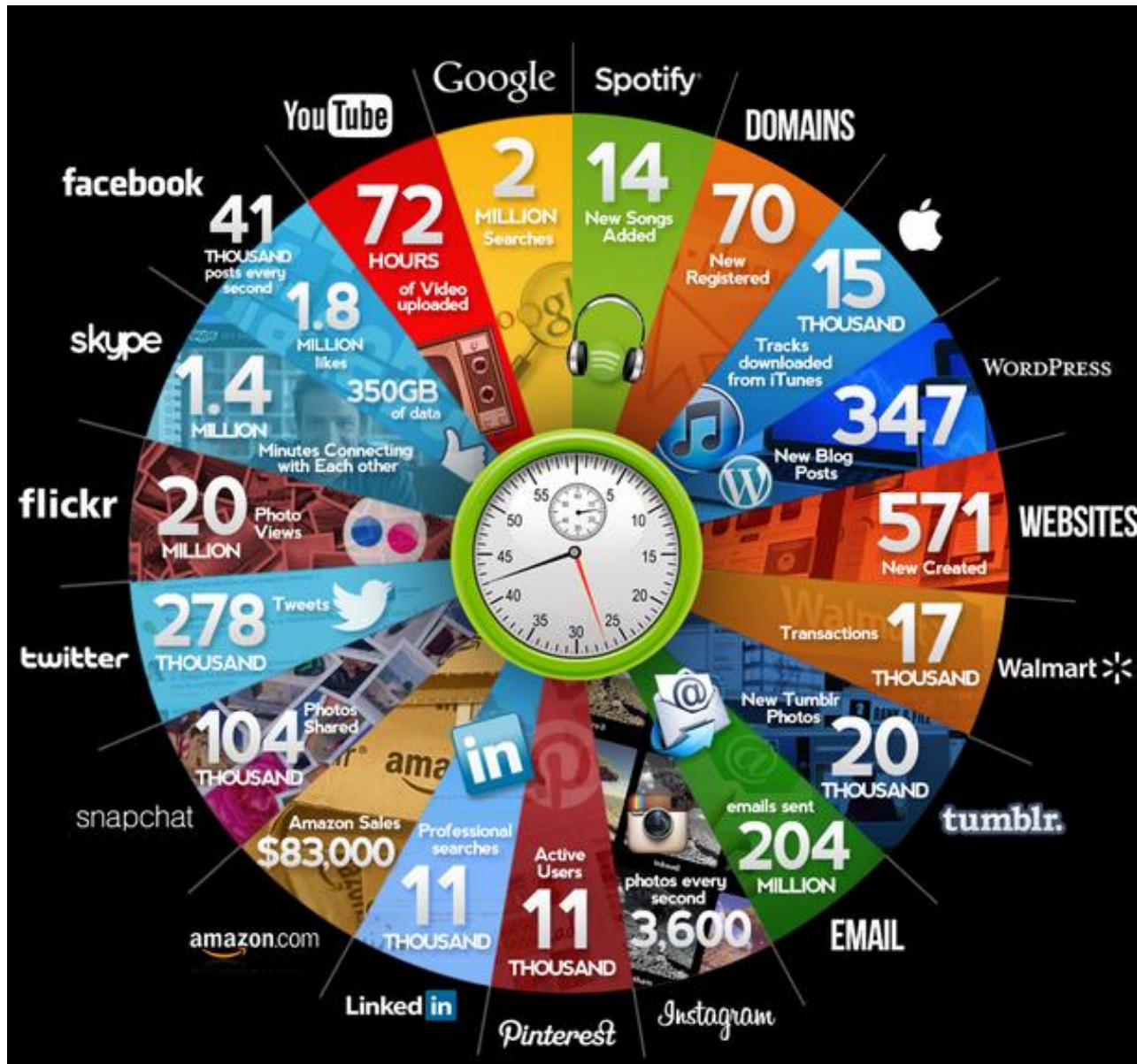


빅데이터 개요



2016년도 2학기 2학년 방과후학교

60초 사이에 일어나는 일들..



빅데이터 정의

그 전에는 분석이 불가능했지만, IT가 발전하면서 다룰 수 있게 된 방대한 양의 데이터

예) 휴대폰 통화 내역, 기상 정보, 실시간 교통량, 인터넷 검색 내역, 소셜 네트워크 서비스 메시지 등

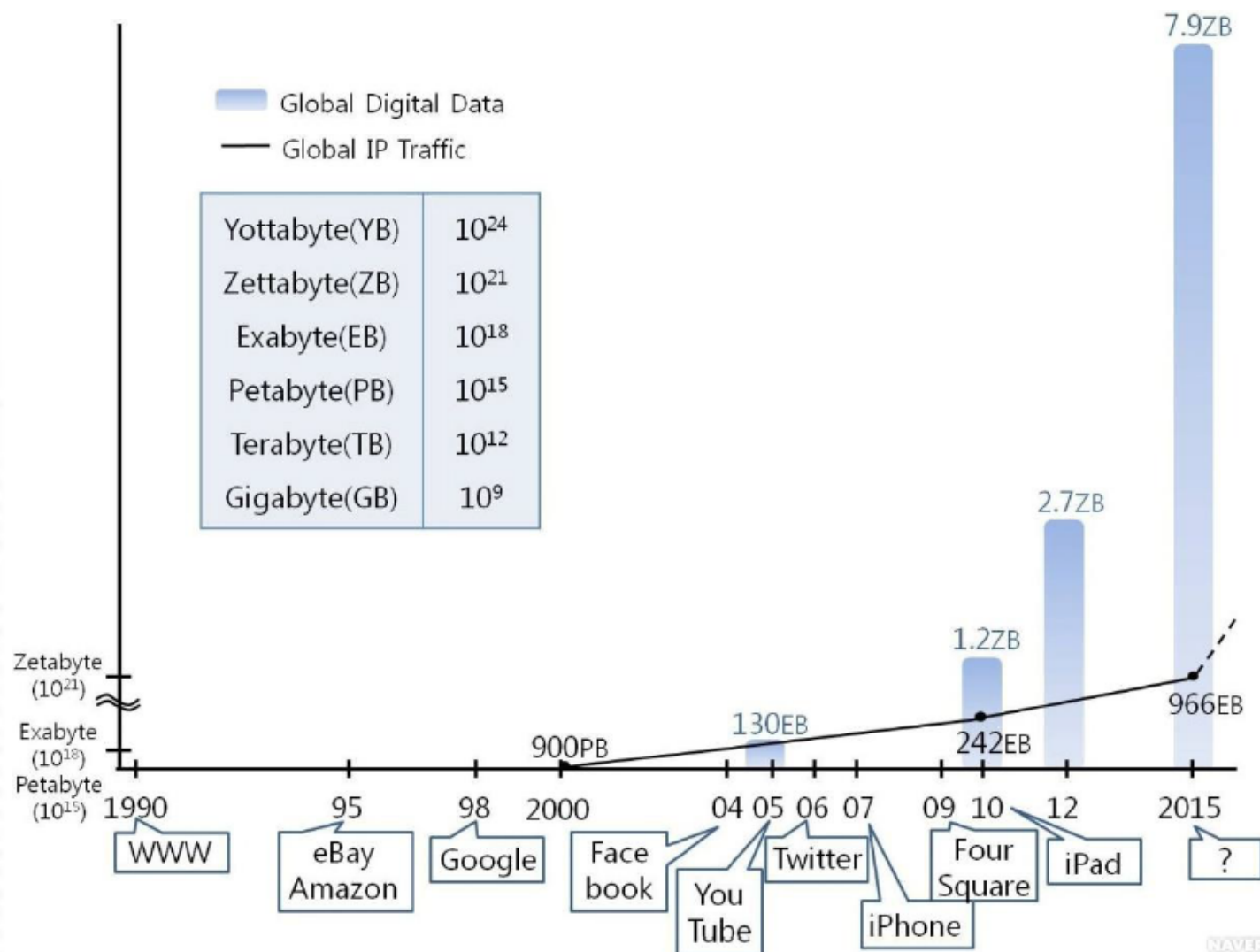


하루 평균 1
억 5,500만 건 생성



하루 평균
동영상 재생 건수 40억회

인터넷 기업의 등장과 글로벌 디지털 데이터 규모



빅데이터 특징

- 데이터의 양 (Volume)
- 데이터 생성 속도 (Velocity)
- 형태의 다양성 (Variety)

빅데이터 속에서는 이름이나 주민등록번호와 같은 개인식별정보, 성별·나이·직업과 같은 정형적인 특성보다는 ‘당신이 어디에서 무엇을 하며, 어떤 행동방식을 선호하는가’가 의미 있는 자료가 됨

빅데이터 특징

구분	기존	빅데이터 환경
데이터	<ul style="list-style-type: none">정형화된 수치자료 중심	<ul style="list-style-type: none">비정형의 다양한 데이터문자 데이터(SMS, 검색어)영상 데이터(CCTV, 동영상)위치 데이터
하드웨어	<ul style="list-style-type: none">고가의 저장장치데이터베이스데이터웨어하우스(Data-warehouse)	<ul style="list-style-type: none">클라우드 컴퓨팅 (비용효율적인 장비) 활용 가능
소프트웨어 /분석 방법	<ul style="list-style-type: none">관계형 데이터베이스(RDBMS)통계패키지(SAS, SPSS)데이터 마이닝(data mining)machine learning, knowledge discovery	<ul style="list-style-type: none">오픈소스 형태의 무료 소프트웨어Hadoop, NoSQL오픈 소스 통계솔루션(R)텍스트 마이닝(text mining)온라인 버즈 분석(opinion mining)감성 분석(sentiment analysis)

빅데이터 사례 1

■ 미국 국세청

- ❖ 정부기관 사기 방지 솔루션 : 방대한 자료로부터 이상 징후를 찾아내고 예측 모델링을 통해 과정의 행동 정보를 분석하여 사기 패턴과 유사한 행동 검출
 - ❖ 소셜 네트워크 분석을 통한 범죄 네트워크 발굴 : 계좌, 주소, 전화번호, 납세자 간의 연관 관계 분석, 범죄자와 관련된 소셜 네트워크를 분석하여 범죄자 집단에 대한 감시 시스템 마련
 - ❖ 다양한 데이터 분석을 통한 지능형 감시 시스템 구축, 탈세 및 사기 범죄 예방 시스템 구축
- 통합형 탈세 및 정부 사기 방지 시스템을 통해 연간 3,450억 달러에 달하는 세금 누락 및 불필요한 세금 환급 절감, 탈세자 수의 감축

빅데이터 사례 2

■ 일본 정부 지능형 교통 정보 시스템

- ❖ 다양한 사용자에게 의해 취득된 정보를 바탕으로 한 **실시간 교통 정보**를 공유함으로써 최적의 교통 안내 서비스 제공 및 교통 체증으로 인한 불필요한 에너지 낭비 방지
 - ❖ UTIS(Ubiqlink Traffic Information System)를 통한 독자적 도로교통정보망 구축 : **차량 주행속도 기반으로 도로 교통정보 예측**한 후, 최적 경로 안내 및 재난 상황 시 통행 가능 도로 정보 제공
- 자동차 주행 속도 계산하여 교통 정보 수집 (일본 전역 지정도시 택시 약 11,000대와 데이터 제공 동의한 사용자로부터 실시간 정보 수집)

빅데이터 사례 3

■ 구글

- ❖ 독감과 관련된 **검색어 빈도를 분석**해 독감 환자 수와 유행 지역을 예측하는 독감 동향 서비스 개발(google.org/flutrends)
 - 미국 질병통제본부(CDC)보다 예측력이 뛰어남
(CDC보다 1~2주 정도 확산 경로를 더 빠르게 예측함)
- ❖ 자동번역 시스템 - 구글은 수천만 권의 도서 정보와 유엔과 유럽의회, 웹사이트 자료를 활용해 64개 언어 간 자동번역 시스템 개발 (IBM도 캐나다 의회의 문서를 활용해 영어·불어 자동번역 시스템 개발을 시도했으나 실패)
- ❖ Microsoft가 장기간 대규모 투자로 만들어낸 스펠링 교정보다 우수한 프로그램을 개발하고 번역시스템에 적용 (**매일 3억건씩 발생하는 검색창의 오타 입력과 수정 정보를 활용**)

빅데이터 사례 4

■ 서울시 심야버스

- ❖ 심야버스 노선을 계획할 당시 심야 택시 승·하차데이터 500만 건과 이동통신사 KT가 보유한 통화기록 30만 건을 결합해 분석

→ 홍대, 동대문, 강남, 종로 등 주요 지역의 시간대별 인구 밀집도와 이동 경로를 파악해 노선도의 효율성을 높임 (하룻밤에 6,000명 이용)

