빅데이터 분석 및 응용

L00: Orientation

Summer 2020

Kookmin University

Instructor

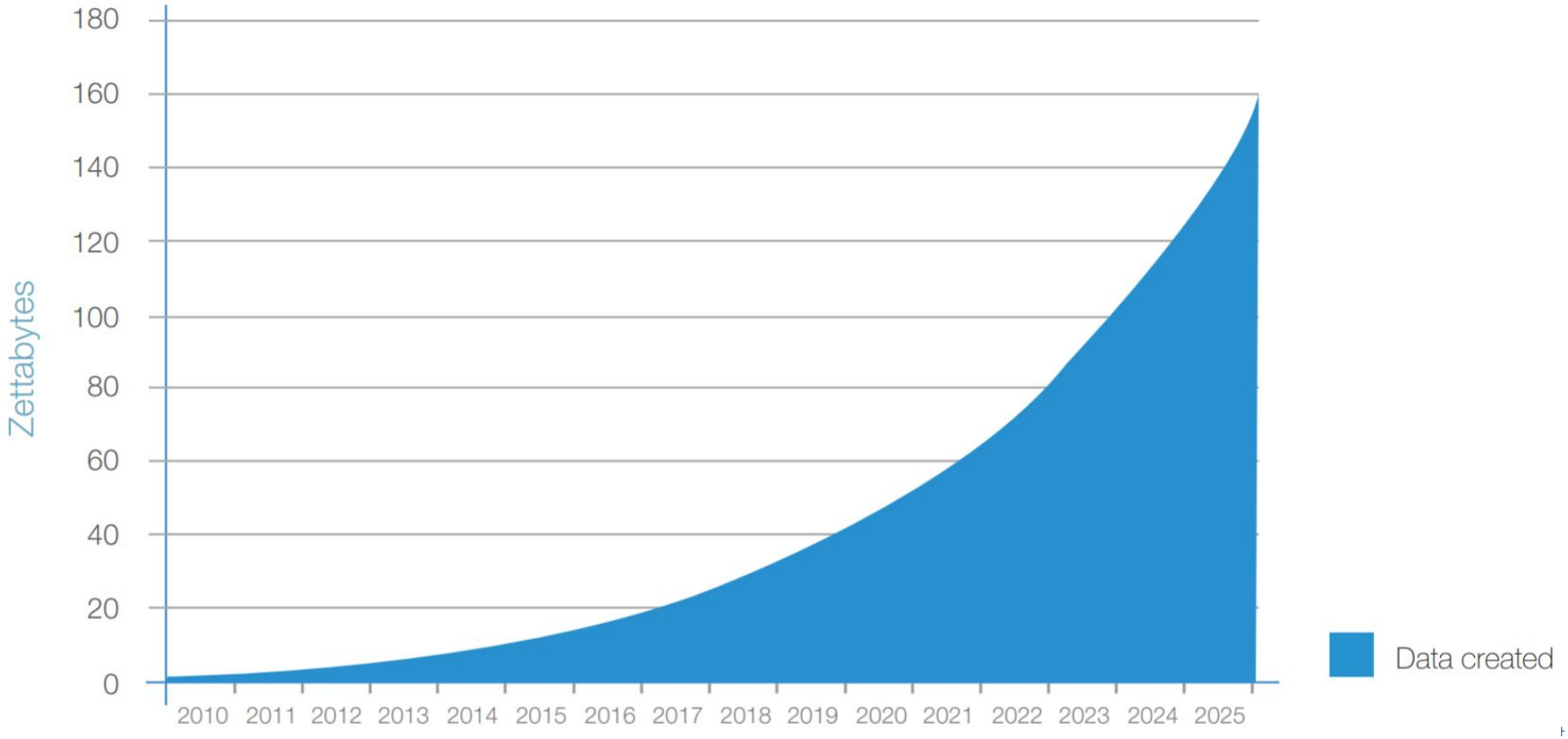
- **Name:** 박하명
- **Office:** 휘랑관 (E4) 814호
- **E-mail:** hmpark@kookmin.ac.kr
- **Homepage:** http://hmpark.me
- **Research Interests:**
  - Data mining
  - Big-data analysis & processing
  - Distributed computing (Hadoop, Spark etc.)
  - Graph algorithms

강의 개요

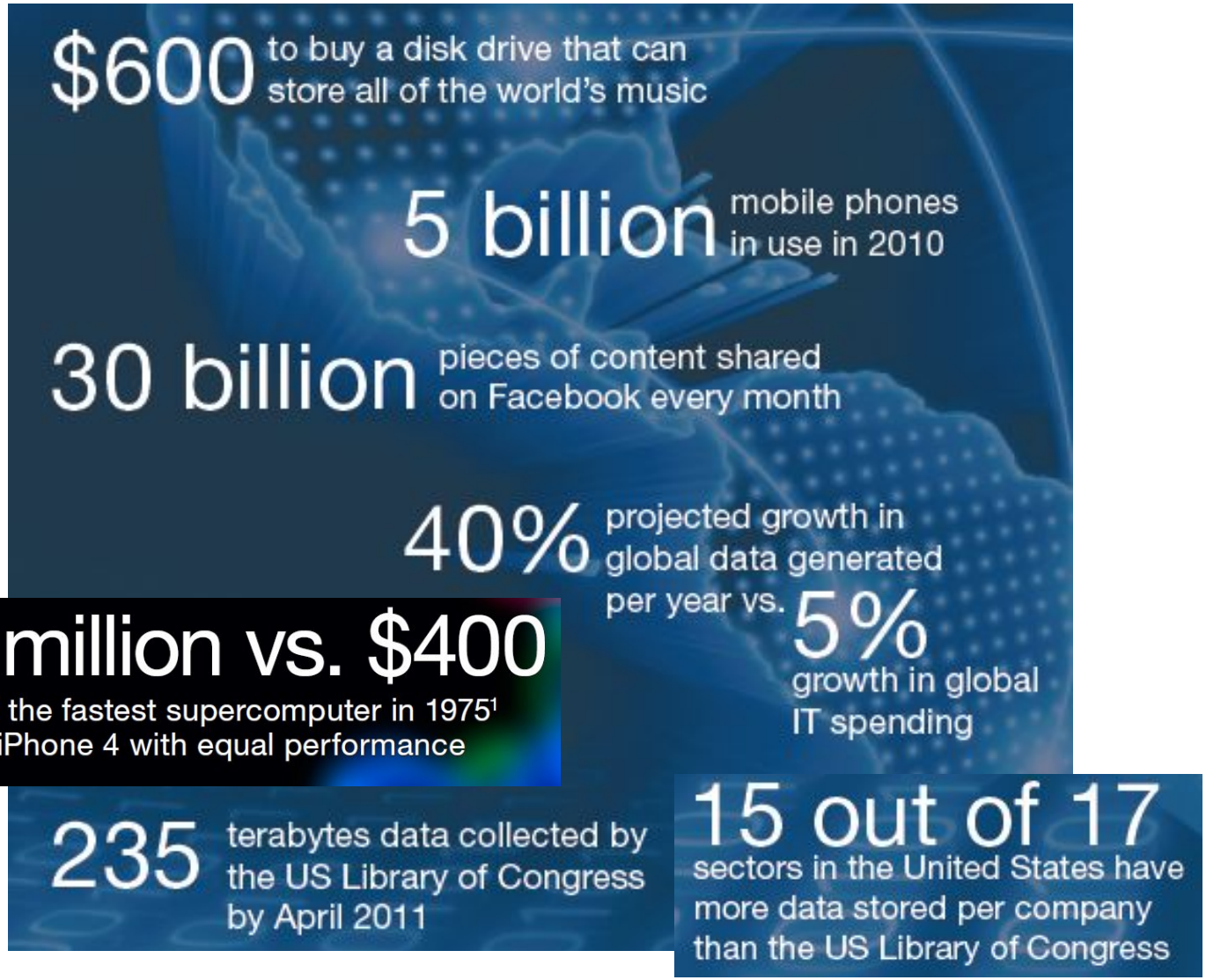
- **강의 시간**
  - 토요일 15:40~18:25
- **강의실**
  - 미래관 445호
- **강의 홈페이지**
  - ecampus 활용 (공지, 수업자료, 과제 등)
- **점수 배점**
  - 프로젝트 60% 과제 30% 출석 10%

Big Data

The sum of the world's data – the DataSphere — will grow from 33 zettabytes in 2018 to a mind-boggling 175ZB by 2025. (Data Age 2025, IDC)



Big Data


$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
$5 million vs. $400
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance
235 terabytes data collected by the US Library of Congress by April 2011
15 out of 17
sectors in the United States have more data stored per company than the US Library of Congress

Big Data


POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unspits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

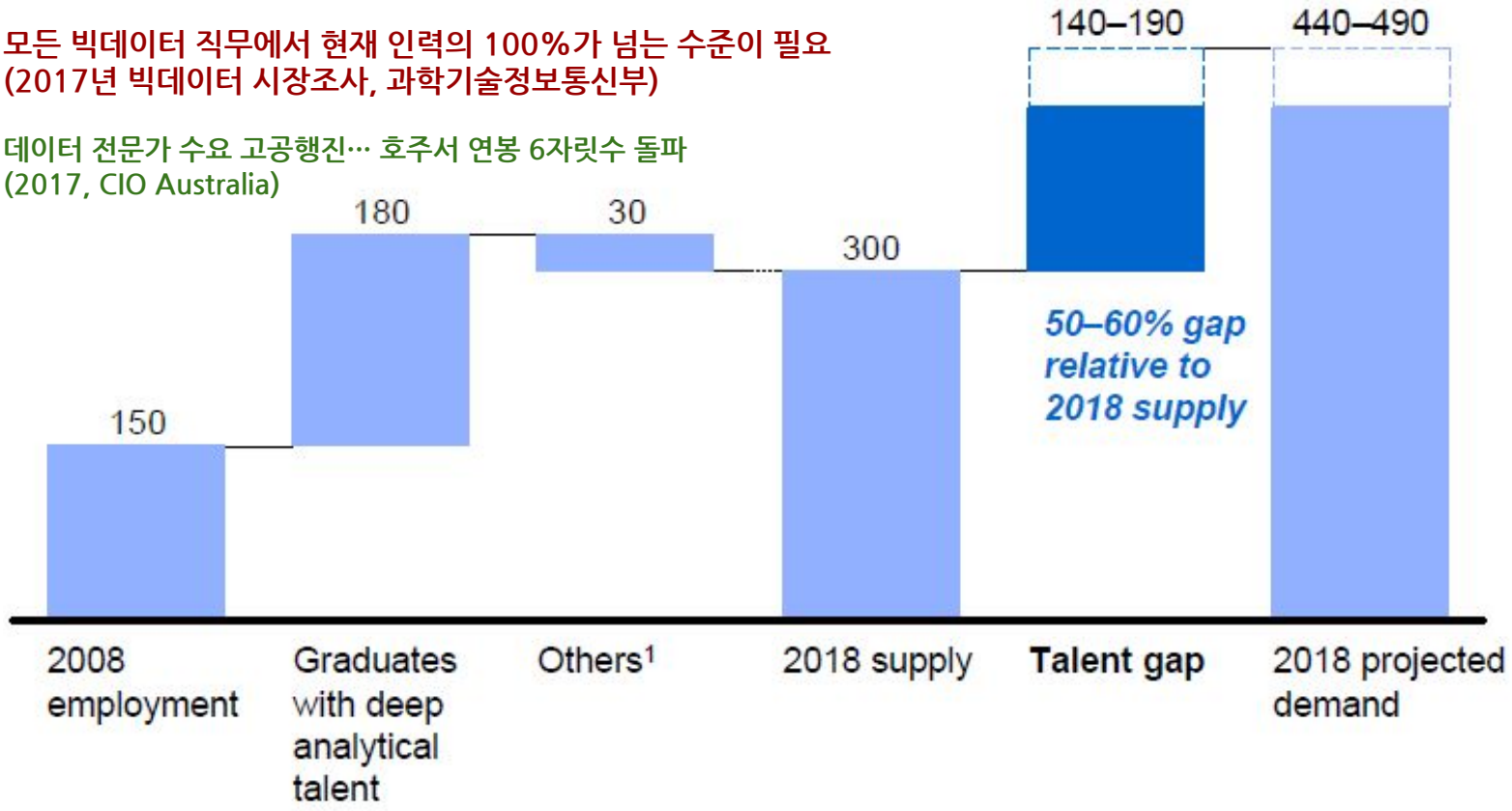
Good news: Demand for Big Data

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people

모든 빅데이터 직무에서 현재 인력의 100%가 넘는 수준이 필요
(2017년 빅데이터 시장조사, 과학기술정보통신부)

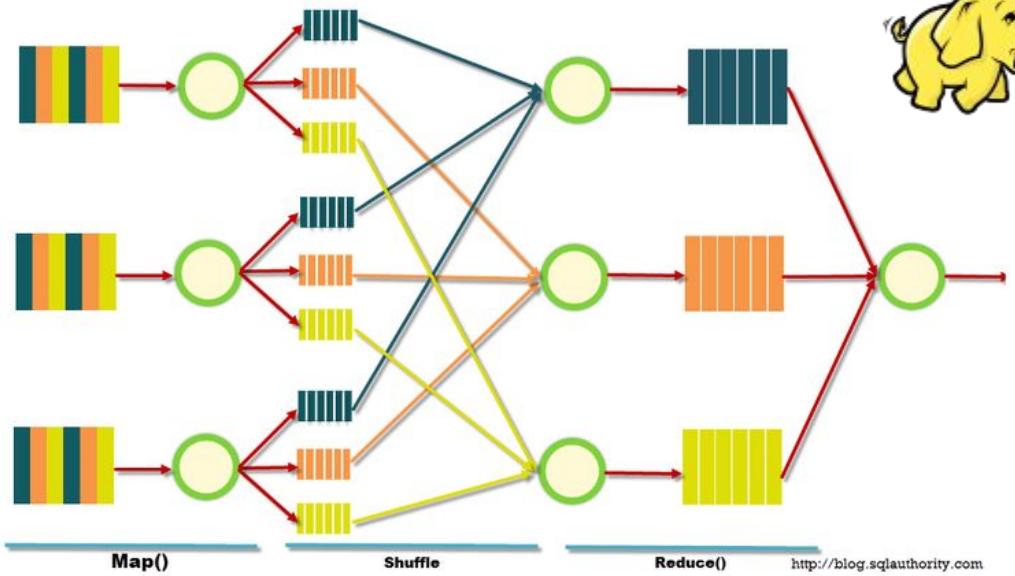
데이터 전문가 수요 고공행진… 호주서 연봉 6자릿수 돌파
(2017, CIO Australia)



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).
SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

MapReduce

- **MapReduce:** 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크
- **Hadoop:** Yahoo에서 제작한 MapReduce의 오픈소스 버전



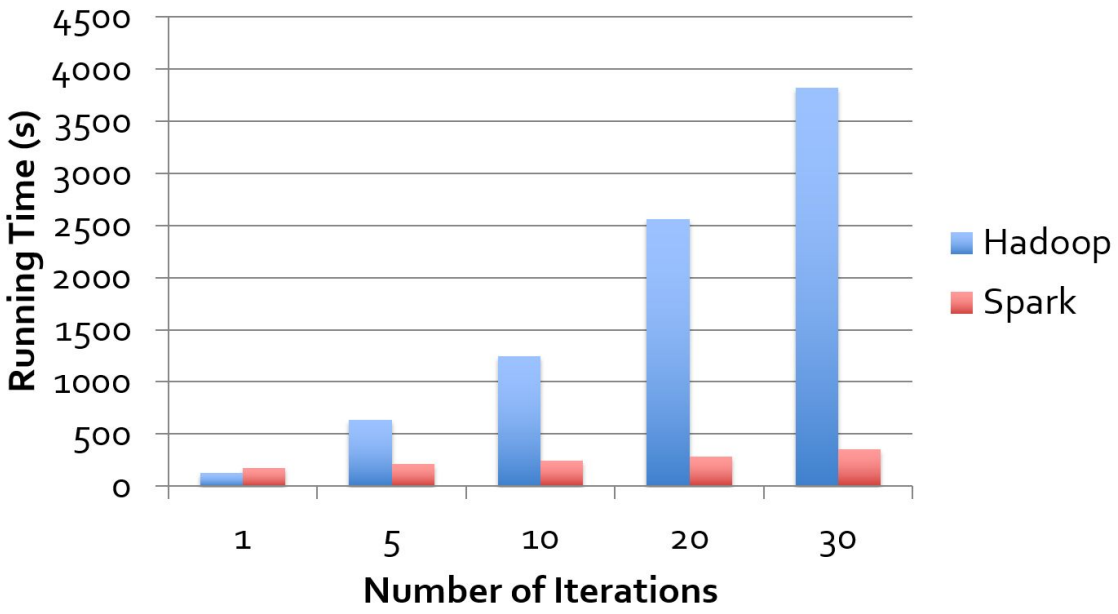
Why MapReduce?

- 아주 큰 데이터는 한 대의 컴퓨터에 담을 수 없거나 처리하는데 시간이 오래 걸린다
  - **컴퓨터 한 대**가 **평균 500GB의 하드디스크**를 가진다면,
  - **10TB**의 데이터를 처리하려면 **20대의 컴퓨터 필요**
  - 컴퓨터 한 대로는 10TB를 읽는데만 4일이나 걸림!

- 빅데이터 분석을 위해서는, 큰 규모의 분산 데이터 처리 및 프로그래밍 프레임워크가 필요

Why Spark?

- MapReduce는 fault tolerance를위해 disk를 적극 활용한다 → **느림**
- Spark: 확장성을 조금 포기하더라도, disk보다는 메모리를 적극 활용한다면 더 빠른 처리가 가능하지 않을까?



강의 일정 (변경될 수 있음)

주차	내용
1주차	Orientation
2주차	MapReduce 개요 및 분산파일시스템
3주차	MapReduce 프로그래밍 모델
4주차	MapReduce dataflow
5주차	Hadoop 실습
6주차	Cluster에서 Hadoop 실습
7주차	Spark
8주차	Spark 실습 (1) – text mining
9주차	Spark 실습 (2) – graph mining
10주차	Project Review

Questions?