빅데이터 기술 개요 - 강의 노트

Slide 1:

안녕하세요!《빅데이터 기술 개요》 과정에 오신 것을 환영합니다.
오늘 우리가 들어가는 내용은 제1장 —— "빅데이터 기술 개요"입니다.

제 연락처와 홈페이지가 위에 있으니, 밤에 잠이 안 오시면 제 홈페이지를 들러보세요. 불면증 해결에 도움이 될지도 모릅니다 💤

Slide 2:

이번 장에서 주로 다룰 내용:

- 빅데이터란 무엇인가?
- 빅데이터가 가져온 영향은?
- 핵심 기술에는 어떤 것들이 있는가?
- 그것의 계산 방식과 대표적인 기술은 무엇인가?

간단히 말하면, 이번 장은 "빅데이터의 과거와 현재"입니다.

Slide 1:

大家好！欢迎来到我们的《大数据技术概述》课程。
我们今天进入的是第一章内容——"大数据技术概述"。

我的联系方式和主页都在上面，如果你晚上睡不着觉，欢迎去我主页看看，也许能帮你催眠(cuīmián) 💤。

Slide 2:

我们这章主要讲：

- 什么是大数据？
- 大数据带来了什么影响？
- 核心技术有哪些？
- 它的计算方式和代表性技术又是什么？

简单来说，这一章就是"大数据的前世今生"。

Slide 3:

이제 오늘의 핵심을 살펴봅시다:
빅데이터의 개념.
여러분 아시나요? "빅데이터"라는 단어는 "데이터 증량제"를 먹은 것처럼 들리지만 🤣, 사실 단순히 "데이터가 많다"는 것만은 아닙니다.
빅데이터에는 4개의 V가 있고, 이제 하나씩 명확히 설명하겠습니다!

Slide 4:

(노트 없음)

Slide 5:

여러분 아시나요?
"빅데이터"라는 단어는 "데이터 증량제"를 먹은 것처럼 들리지만 🤣,

사실 단순히 "데이터가 많다"는 것만은 아닙니다.
빅데이터에는 4개의 V가 있고, 이제 하나씩 명확히 설명하겠습니다!

Slide 3: (그냥 넘어갈 것)

现在，我们来看今天的重点(**zhòngdiǎn)**：
大数据的概念。
你们知道吗？"大数据"这个词听起来像是吃了"数据增肥药(zēngféi yào)"🤣，
其实它不仅仅是"数据很多"这么简单。

大数据有4个V，我们接下来会一一讲清楚！

Slide 4:

(슬라이드 보면서 대충 이야기 할 것)

Slide 5:

你们知道吗？
"大数据"这个词听起来像是吃了"数据增肥药zēngféi yào"🤣，

其实它不仅仅是"数据很多"这么简单。

大数据有4个V，我们接下来会一一讲清楚！

不仅仅" (bùjǐnjǐn): "~뿐만 아니라"

Slide 6: 4V 구체적 설명

자, 하나씩 살펴봅시다:

- **Volume**: 데이터 양이 엄청나게 크다
- **Velocity**: 데이터가 엄청나게 빠르게 생성된다
- **Variety**: 데이터 유형이 매우 다양하다
- **Value**: 대부분의 데이터는 사실 별로 가치가 없고, 가치 밀도가 매우 낮다

마치 우리가 사진을 찍는 것처럼 —— 매일 수십 장을 찍지만, 실제로 SNS에 올리는 건 한두 장뿐이고, 나머지는 모두 "데이터 무덤"으로 들어갑니다.

학생여러분 저를 따라서 함께 이 4가지 특성을 소리내어 읽어봅시다.

Slide 6: 4V具体解释

来，我们一个一个来看：

- Volume：数据量超级大。
- Velocity：数据产生得飞快。
- Variety：数据类型五花八门。
- Value：大部分数据其实没啥(**méi shá)**价值，价值密度很低。

就像我们拍照——每天拍几十张，但真正发朋友圈的只有一两张，剩下的都进了"数据坟墓(fénmù)"。

同学们，请跟我一起大声朗读这四个特点。

Slide 7: 빅데이터 목차

오늘 학습할 개요를 봅시다. 사실 방금 말한 네 개의 V입니다:

- 데이터 유형이 다양함 (Variety)
- 처리 속도가 빠름 (Velocity)
- 데이터 양이 많음 (Volume)
- 가치 밀도가 낮음 (Value)

학생A: 선생님, 이 순서가 4V의 순서를 거꾸로 한 거 아닌가요?
선생님: 맞아요, 저는 "거꾸로 하는 것"을 좋아해요. 그래야 여러분이 더 쉽게 기억하거든요 😂

그리고 오늘은 "데이터 유형의 다양화"부터 시작하겠습니다!

Slide 7:大数据的提纲

我们来看今天的学习提纲，其实就是刚才那四个V：

- 数据类型繁多（Variety）
- 处理速度快（Velocity）
- 数据量大（Volume）
- 价值密度低（Value）

学生A：老师，这顺序是不是背反了4V的顺序啊？
老师：对啊，我就喜欢"反着来"，这样你们更容易记住😂。

而且我们今天就从"数据类型多样化"开始说起！

Slide 8: VOLUME 데이터 양이 많음

여러분, 혹시 눈치채셨나요? 요즘 동영상 하나 찍으면 수백 메가바이트고, 쇼폼 플랫폼 좀 보면 어느새 핸드폰에서 "저장 공간 부족" 알림이 뜹니다!

이것이 바로: 데이터가 정말 점점 더 커지고 있다는 것입니다!

VOLUME은 "대량화", 데이터가 상상할 수 없을 만큼 많다는 뜻입니다!

상상할 수 있나요? 집에 있는 하드디스크, USB, 메모리 카드를 다 꺼내고, 할아버지 할머니 것까지 다 꺼내도, 여전히 저장 공간이 부족합니다 😱

Slide 9: IDC는 누구?

여러분 혹시 IDC를 들어보셨나요? "인도 대주방"이 아니에요 🤣. 미국의 매우 권위 있는 연구 기관으로, 정식 명칭은 "International Data Corporation"입니다.

Slide 10:

그들은 매년 IT 산업에 관한 연구 보고서를 발표하는데, 우리가 주목할 가치가 있습니다. 지금 우리가 이야기하는 빅데이터 성장 추세가 바로 그들의 연구에서 나온 것입니다.

Slide 8: VOLUME 数据量大

大家有没有注意到，现在我们拍个视频就几百兆**(zhào)**，刷个短视频平台，不知不觉手机就提示"内存不足"了！

这就说明：数据真的越来越大啦！

VOLUME 就是"大量化"，数据多到你无法想象！

你能想象吗？就算你把家里的硬盘、U盘、手机卡都拿出来，连同你爷爷奶奶的也拿出来，还是不够存 😱。

Slide 9: IDC是谁？

你们有没有听过 IDC？

不是"印度大厨 **(chú)**"哦🤣，是美国一个非常权威的研究机构，全名叫"International Data Corporation"。

Slide 10:

他们每年都会发布关于IT行业的研究报告，非常值得我们关注。我们现在讲的大数据增长趋势(qūshì)，就是从他们的研究中来的。

Slide 11:
(노트 없음)

Slide 12: 빅데이터의 무어의 법칙
IDC에 따르면, 인류 사회의 데이터는 매년 50%의 속도로 증가합니다. 즉 ——
📈 **2년마다 데이터 양이 두 배로 증가합니다!**

이건 마치 "데이터가 성장 호르몬을 먹은 것" 같지 않나요?
우리는 이 추세를 "빅데이터의 무어의 법칙"이라고 부릅니다.

학생B: 선생님, 그럼 몇 년 후에는 채팅 기록도 저장할 수 없게 되는 건가요?
선생님: 맞아요. 나중에는 이별 채팅 기록도 유료로 다운로드해서 백업해야 할지도 모릅니다 😂

Slide 11:
(노트 없음)
-
Slide 12:大数据的摩尔定律
IDC说了，人类社会每年数据以50%的速度增长，也就是说——
📈 每两年，数据量就翻一倍！

这是不是有点像"数据吃了激素(jīsù)"？

我们叫这个趋势为"大数据摩尔(**Mó ěr)**定律"。

学生B：老师，那是不是过几年，我们连聊天记录都存不起啦？
老师：是的，说不定你以后连分手聊天记录都要付费(fù fèi)下载备份😂。

Slide 13: 최근 2년 = 전체 역사?

충격적입니다! 인류가 최근 2년 동안 생성한 데이터 양이 그 이전 수천 년 동안의 전체 데이터 총량과 같습니다!

상상해보세요: 도서관에 들어갔는데, 사서가 말합니다: "죄송합니다, 저희가 이 2년 동안 추가한 책이 당신 할아버지 세대의 모든 책을 합친 것보다 많습니다."

이것이 바로 우리가 현재 살고 있는 빅데이터 시대입니다! 📚📚📚📚📚📚

Slide 13:近两年 = 全部历史？

震惊！人类在最近两年产生的数据量，等于之前几千年全部数据的总和！

想象一下：你走进一间图书馆，管理员跟你说："不好意思，我们这两年新增的书，比你爷爷那一辈所有书加起来还多。"

这，就是我们现在处在的大数据时代！📚📚📚📚📚📚

Slide 14: 2020년, 전 세계에는 얼마나 많은 데이터가?

IDC의 예측에 따르면, 2020년까지 전 세계적으로 35ZB(제타바이트)의 데이터가 있을 것입니다.

여러분이 물을지도 모릅니다: ZB가 뭔가요? 어떤 브랜드 과자 이름 같은데 🤣
급하게 생각하지 마세요. 지금 바로 "데이터 단위 환산"을 배워봅시다.

Slide 14:到了2020年，全球有多少数据？

根据 IDC 的预测，到 2020 年，全球会有 35ZB（泽字节） 的数据。

你可能会问：ZB 是什么东西？听起来像是某个品牌的辣条 **là tiáo** 🤣

别急，我们马上就来学习"数据单位换算"。

Slide 15: 2020년, 전 세계에는 얼마나 많은 데이터가?

IDC의 예측에 따르면, 2020년까지 전 세계적으로 35ZB(제타바이트)의 데이터가 있을 것입니다.

여러분이 물을지도 모릅니다: ZB가 뭔가요? 어떤 브랜드 과자 이름 같은데 🤔

급하게 생각하지 마세요. 지금 바로 "데이터 단위 환산"을 배워봅시다.

Slide 15: 그냥 슬라이더 보고 간단하게 말할 것

到了2020年，全球有多少数据？

根据 IDC 的预测，到 2020 年，全球会有 35ZB（泽字节） 的数据。

你可能会问：ZB 是什么东西？听起来像是某个品牌的辣条 🤔

别急，我们马上就来学习"数据单位换算"。

Slide 16: 데이터 단위 환산

자, 여러분, 복습해봅시다:

단위 환산 관계

- 1 KB = 1024 Byte
- 1 MB = 1024 KB
- 1 GB = 1024 MB
- 1 TB = 1024 GB
- 1 PB = 1024 TB
- 1 EB = 1024 PB
- 1 ZB = 1024 EB
- 1YB = 1024ZB

🌟 **이 구호를 기억하세요**: "Byte로 시작, YB로 끝, 작은 것부터 큰 것까지, 점점 더 비싸진다"

📦 예를 들어, 만약 여러분이 클라우드 저장소를 사용한다면, 무료 공간이 5GB 정도일 텐데, 1TB를 사면 스스로 "데이터 부자"라고 느낄 거예요. 그렇다면 35ZB는 어떤 개념일까요? 말 그대로 "우주 저장소 갑부"입니다!

Slide 16:数据单位换算

来，同学们，咱们复习一下：

单位换算关系

- 1 KB= 1024 Byte
- 1 MB= 1024 KB
- 1 GB= 1024 MB
- 1 TB= 1024 GB
- 1 PB= 1024 TB
- 1 EB= 1024 PB
- 1 ZB= 1024 EB
- 1YB = 1024ZB

🌟 记住这个口诀："Byte 开头，YB 收尾，从小到大，越来越贵"。

📦 举个例子，如果你用百度网盘，免费空间可能是 5GB，买到 1TB 都觉得自己"数据富豪"，那你想想 35ZB 是什么概念？简直是"宇宙存储富翁"！

Slide 17: VARIETY 다양화

방금 Volume을 설명했고, 이제 두 번째 V로 들어갑니다: Variety, 다양화!

우리 생활 속 다양한 유형의 데이터는 매우 다양하고, 더 이상 예전처럼 규칙적인 표 형태가 아닙니다.

마치 한 대학에 문과생도 있고, 이과생도 있고, 예술 특기생도 있는 것처럼 —— 구조가 완전히 다릅니다.

Slide 18:

(노트 없음)

Slide 17:VARIETY 多样化

我们刚讲完 Volume，现在进入第二个 V：Variety，多样化！

我们生活中各种类型的数据五花八门，不再是以前那种规规矩矩的表格啦。

就像一所大学，有的是文科生、有的是理科生，还有的是艺术特长生——结构完全不同。

Slide 18: 슬라이드 보면서 간단하게 말하고 넘어간다.

Slide 19: 구조화 데이터 vs 비구조화 데이터

이 그림이 알려주는 것:

- **구조화 데이터(Structured)**: 표 안에 정렬된 것들, 예를 들어 Excel, 데이터베이스
- **비구조화 데이터(Unstructured)**: 형식 없이 뒤죽박죽인 것들, 예를 들어 동영상, 이미지, 오디오, 트위터, SNS

Slide 19: 结构化 vs 非结构化

这张图告诉我们：

- 结构化数据（Structured）：表格里排好队的那种，比如 Excel，数据库。
- 非结构化数据（Unstructured）：杂七杂八没格式，比如视频、图片、音频、微博、朋友圈。

Slide 20: 구조화 데이터 vs 비구조화 데이터

통계에 따르면, 90%의 데이터가 비구조화 데이터입니다.
즉, 여러분의 셀카, 하소연, 팬질, 이모티콘... 전부 빅데이터의 일부입니다!

Slide 20: 结构化 vs 非结构化

根据统计，90% 的数据是非结构化的。
也就是说，你的自拍、吐槽 **(tǔ cáo)**、追星 **(zhuī xīng)**、表情包**(biǎo qíng bāo)**... 全部都属于大数据的一部分！

Slide 21: 데이터 유형의 분류도

다시 이 그림을 봅시다. 데이터 유형의 풍부함을 보여줍니다:

👨‍🔬 **과학 연구 분야:**
- 유전체(Genome)
- LHC 입자 가속기
- 우주 탐사

🏢 **기업 응용 분야:**
- 거래 기록
- 애플리케이션 로그
- 이메일, Word 파일 등

🌐 **네트워크 데이터:**
- 클릭스트림, 웨이보,트위터, 블로그, 위키피디아

😂

Slide 21: 数据类型的分类图

我们再来看这张图，展示了数据类型的丰富性：

👨‍🔬 科学研究领域：
- 基因组（Genome）
- LHC 粒子加速器
- 空间探测

🏢 企业应用领域：
- 交易记录
- 应用日志
- 邮件、Word 文件等

🌐 网络数据：
- 点击流、微博、推文、博客、维基百科

😂

Slide 22: 다양한 유형의 데이터

이 그림은 다양한 유형의 데이터를 나열합니다:

- 📄 **텍스트** (예: 이메일, 보고서, 웹페이지)
- 🖼️ **이미지** (사진, X-ray, 이모티콘 😆)
- 🎬 **비디오** (감시 영상, 라이브 방송, 숏폼)
- 💻 **로그** (애플리케이션 로그, 시스템 기록)
- 🛒 **거래 기록** (온라인 쇼핑, 금융 결제)

보세요, 지금의 데이터 종류는 더 이상 "한 냄비에 밥 짓기"의 개념이 아니라, "훠궈에 모든 재료가 다 들어있는" 것입니다!

Slide 22: 不同类型的数据

这张图列出了各种不同类型的数据：

- 📄 文本（例如：邮件、报告、网页）
- 🖼️ 图像（照片、X光片、表情包 😆）
- 🎬 视频（监控视频、直播、短视频）
- 💻 日志（应用日志、系统记录）
- 🛒 交易记录（电商购物、金融支付）

你看，现在的数据种类，已经不是"一个锅里煮饭"的概念了，是"火锅里什么料都有"！

Slide 23: 전통 알고리즘이 무용지물?

이렇게 많은 종류의 데이터를 마주하니, 전통 알고리즘은 마치 노인이 신기술을 마주한 것처럼 —— 조금 "역부족"입니다 😅

왜일까요? 예전 알고리즘은 주로 구조화 데이터를 처리했는데, 지금 동영상, 오디오 같은 비구조화 데이터를 만나면,
주판 들고 왕자 영광(모바일 게임) 하는 것과 같습니다 —— 현실적이지 않죠!

Slide 24: 새로운 알고리즘을 설계해야!

그렇다면 우리에게 필요한 것은?

💡 **새로운 알고리즘!**
이런 "새로운 말썽꾸러기 데이터"를 처리하기 위해, 우리는 이미지를 인식하고, 음성을 이해하고, 심지어 자동으로 글을 쓸 수 있는 알고리즘을 설계해야 합니다.

Slide 23: 传统算法失效了？

面对这么多种类的数据，传统的算法就像老年人面对新科技 —— 有点"力不从心" 😅

为什么呢？以前算法主要处理结构化数据，现在遇到视频、音频这些非结构化数据，就像拿算盘去打王者荣耀(**Wángzhě Róngyào**)——不现实！

Slide 24: 要设计新算法！

所以我们需要什么？

💡 新算法！
为了处理这些"新型顽皮(**wán pí**)数据"，我们要设计能识别(**shí bié**)图片、能理解语音、甚至能自动写文章的算法。

Slide 25: VELOCITY 빠른 처리

다음은 세 번째 V를 설명합니다: Velocity!
의미는 —— 처리 속도가 빨라야 합니다!

여러분 발견하지 못했나요? 지금 배달 주문하든, 영상 보든, 옷 사든, 1초 안에 로딩이 안 되면 "인터넷 느리다"고 느껴지죠?

그래서 우리는 "실시간 응답"이 필요하고, 이를 "1초 법칙"이라고도 합니다.

Slide 26: 1초 법칙

1초 법칙이란 무엇일까요?
예를 들어 쇼핑 앱을 열고 "검색"을 클릭하면 ——
⚡ **1초 이내에 시스템이 원하는 상품, 여러분이 좋아할 만한 브랜드, 심지어 어제 봤던 바지까지 추천해야 합니다!**

만약 5초 후에 나타난다면 ——
죄송하지만, 사용자는 이미 옆에 있는 다른 쇼핑 앱으로 갔습니다 🤣

이것이 바로 빅데이터 속도의 마법입니다!

Slide 25: VELOCITY 快速化

接下来我们讲第三个V：Velocity！
意思就是——处理速度要快！

你有没有发现？现在不管是点外卖、看视频、买衣服，一秒钟不加载就觉得"网卡了"？

所以我们需要"实时响应**(xiǎng yìng)**"，也叫"一秒定律"。

Slide 26: 一秒定律

什么是一秒定律？
比如你打开淘宝，点了个"搜索"——

⚡ 一秒之内，系统就要告诉你你想要的商品、猜**(cāi)**你喜欢 **cái ní xǐ huān** 的品牌、甚至你昨天看过的裤子还在推荐！

要是在五秒钟之后才出来——不好意思，用户已经切到隔壁**(gé bì)**的拼多多**(Pīn duō duō)**去了🤣

这就是大数据速度的魔力**(mó lì)**！

Slide 27: 관심 있는 콘텐츠를 1초 내에 제공

이런 경험 해보셨나요? 친구와 메신저로 "훠궈 먹고 싶다"고 말했는데, 어떤 앱을 열면 훠궈 광고가 즉시 튀어나옵니다!

학생들: 무서워요, 핸드폰이 저를 감시하나요?
선생님(웃음): 안심하세요, 사실 백그라운드의 빅데이터 시스템이 너무 똑똑해서 그래요. 여러분이 도청당한 게 아니라, 여러분이 "예측당한" 거예요!

Slide 28: 모든 마우스 클릭에는 데이터가

우리가 인터넷에서 하는 모든 클릭, 모든 스크롤은 기록됩니다 —— 이것이 바로 클릭스트림(Clickstream)입니다!

마치 마트에 가면, 모든 발걸음, 어떤 진열대를 보는지, 어떤 제품을 집는지, 모두 "투명 점원"이 조용히 기록하는 것과 같습니다 🕵️‍♀️

Slide 27: 您感兴趣的内容一秒内呈现

你有没有这种体验？你刚在微信跟朋友说完“想吃火锅”，打开某个App，火锅广告立刻弹出来！

同学们：吓死我了，手机在监视我吗？

老师（笑）：放心，其实是后台的大数据系统太聪明了，不是你被监听，而是你“被预测”了！

Slide 28: 每一次鼠标点击都有数据

我们在网上每一次点击、每一次滑动(**huá dòng**)，都会被记录下来——
这就是点击流（Clickstream）！

就像你去超市，每走一步、看哪个货架、拿起什么产品，全部都被“隐形店员(**yǐn xíng diàn yuan**)”悄悄记下来 🕵️‍♀️。

拖动 (tuō dòng) 드래그

Slide 29-30: 클릭스트림 데이터 흐름 생성

이러한 클릭 행동은 데이터 흐름을 생성하고, 백그라운드에서 실시간으로 분석합니다. 예를 들어:

- 언제 클릭했나요?
- 어떤 페이지에서 클릭했나요?
- 클릭 후에 구매했나요?

이것은 마치 우리가 인터넷에서의 "행동 발자취"와 같습니다. 여러분이 걸어간 모든 발걸음을 시스템이 조용히 기록합니다(부드럽게) 😎

Slide 30: 클릭스트림 데이터 흐름 생성

이러한 클릭 행동은 데이터 흐름을 생성하고, 백그라운드에서 실시간으로 분석합니다. 예를 들어:

- 언제 클릭했나요?
- 어떤 페이지에서 클릭했나요?
- 클릭 후에 구매했나요?

이것은 마치 우리가 인터넷에서의 "행동 발자취"와 같습니다. 여러분이 걸어간 모든 발걸음을 시스템이 조용히 기록합니다(부드럽게) 😎

Slide 29-30: 生成点击流数据流

这些点击行为会生成一条条数据流，后台会实时分析，比如：

- 你什么时候点的？
- 在哪个页面点的？
- 点完以后有没有买？

这就像是我们在互联网上的"行动足迹(**zú jì**)"，你走过的每一步，系统都悄悄(**qiāo qiāo**)记下来了（温柔地）😎。

Slide 30: 生成点(**diǎn jī**)击流数据流

这些点击行为会生成一条条数据流，后台会实时分析，比如：

- 你什么时候点的？
- 在哪个页面点的？
- 点完以后有没有买？

这就像是我们在互联网上的"行动足迹"，你走过的每一步，系统都悄悄记下来了（温柔地）😎。

Slide 31: 현재 접속 중인 콘텐츠 실시간 분석

백그라운드는 여러분이 현재 보고 있는 콘텐츠에 따라 실시간 분석을 수행합니다.

예를 들어 여러분이 "립스틱"을 보고 있으면, 시스템은 즉시 백그라운드에 알립니다: "빨리 아이섀도를 추천해! 그녀도 관심 있을 거야!"

이런 분석은 전혀 기다리지 않고, 1초 이내에 완료해야 합니다. 그렇지 않으면 사용자가 스크롤해서 넘어가버리면 "비즈니스 기회를 놓치게" 됩니다!

Slide 31: 实时分析正在访问的内容

后台会根据你当前正在看的内容，进行实时分析。

比如你在看"口红"，系统会立刻告诉后台："快推荐眼影(**yǎn yǐng**)！她可能也感兴趣！"

这种分析完全不等人，必须在1秒内完成。不然用户滑过去了，就"错过了商机"！

-

Slide 32: 사용자 특성 vs 그룹 특성 매칭

시스템은 어떻게 물건을 추천할까요?

📌 **여러분을 "여러분과 비슷한 사람들"과 매칭합니다:**

- 여러분이 사지 않은 것을, 다른 사람이 샀다면 → 여러분에게 추천!
- 다른 사람이 관심 있는 브랜드도 여러분이 관심 있을 수 있어요!

이것을 그룹 특성 기반 추천이라고 하며, "Collaborative Filtering"이라고도 합니다 —— "가십 알고리즘" 같지 않나요 😆?

Slide 32: 用户特征 vs 群体特征匹配

系统会怎么推荐东西？

📌 它会把你和"和你相似的人群"做匹配：

- 你没买过的东西，别人买过 → 推荐给你！
- 别人关注的品牌，你也可能感兴趣！

这就叫基于群体特征的推荐，也叫"Collaborative Filtering"——

听起来是不是像是"八卦算法"😆？

Slide 33:

(노트 없음)

Slide 34: 온라인 쇼핑의 클릭 행동

예시를 봅시다:

👆 온라인 쇼핑몰에서 상품을 탐색할 때, 한 번 클릭할 때마다 시스템 백그라운드는 "와, 또 하나의 데이터가 왔어!"라고 합니다.

학생E: 선생님, 저는 아직 주문하지도 않았는데 기록되나요?

선생님(웃음): 당연하죠! 여러분이 어떤 이미지를 클릭하는지, 몇 초 동안 망설이는지, 찜했는지 안 했는지... 모두 "가치 있는 정보"입니다!

Slide 35: 실시간 상품 추천

여러분이 샴푸를 탐색하면, 다음 초에 페이지가 자동으로 팝업합니다:

🧴 「**이것도 좋아하실 수 있어요: 컨디셔너, 세안제, 두피 마사지기!**」

이것이 바로 빅데이터의 추천 시스템입니다. 남자친구보다 여러분을 더 잘 알아요! (웃음)

학생F: 선생님, 그럼 제 남자친구를 빅데이터로 대체할 수 있나요?

선생님: 최소한 선물하는 면에서는... 가능할지도요 😆

Slide 33: (앞의 내용과 함께 설명 할 것)

Slide 34: 京东购物的点击行为

我们来看看京东这个例子：

👆 当你在京东上浏览商品，每点一次，系统后台就会"哇，又来一条数据啦！"

学生E：老师，我还没下单就被记录了？

老师（笑）：当然！你点哪一张图、犹豫(**yóu yù**)了几秒钟、有没有收藏……统统都是"有价值的信息"！

Slide 35: 实时推荐商品

你浏览一个洗发水，下一秒页面就自动跳出(**tiào chū**)：

🧴「你可能还喜欢：护发素、洗面奶、头皮按摩器！」

这就是大数据的推荐系统啦，它比你男朋友还懂你！（笑）

学生F：老师，那我男朋友可以被大数据取代吗？

老师：至少在送礼物这方面……可能可以 😆

Slide 36: 시간 지연 vs 비즈니스 가치

⚠️ **주의**: 만약 추천 정보가 5초, 10초 늦게 나타나면, 사용자는 이미 다른 곳으로 클릭해서 나가버렸고, 비즈니스 기회를 놓쳤습니다!
그래서 시스템은 반드시 —— **실시간 추천 + 초 단위 반응**을 해야 합니다!

📈 이렇게 해야만 "데이터"를 "비즈니스 가치"로 전환할 수 있습니다.

마치 아이스크림 판매처럼, 날씨가 더울 때 팔아야지, 눈이 내린 후에 배달하면... 누가 먹겠어요? ❄️🍦

Slide 37: VALUE (가치 밀도 낮음)

드디어 마지막 V를 설명합니다: Value (가치 밀도 낮음)
무슨 의미일까요?

📦 **데이터는 많지만, 유용한 것은 사실 많지 않습니다.**

마치 택배 상자 더미처럼, 99개를 열어봐도 별 쓸모가 없고, 단 1개만 여러분이 원하는 것입니다!
학생G: 그럼 왜 이렇게 많은 쓸모없는 데이터를 저장하나요?
선생님: 그 "1%"의 가치가 억만 단위의 비즈니스 기회일 수 있기 때문입니다! 💰

Slide 36: 时间延迟 vs(duì) 商业价值

⚠️ 注意：如果推荐信息晚了五秒十秒才出现，用户早就点出去了，商机就错过了！

所以系统必须做到——实时推荐 + 秒级反应！

📈 只有这样，才能让"数据"转化为"商业价值"。

就像卖冰淇淋(**bīng qí lín**)，必须在天气热的时候卖，等下雪了再送过来……你说谁还吃啊？❄️🍦

Slide 37: VALUE（价值密度低）

终于我们讲到最后一个V啦：Value（价值密度低(**dī**)）
什么意思呢？

📦 数据很多，但有用的其实不多。

就像一堆快递盒子(**yī duī kuài dì hé zi**)，打开99个没啥用，只有1个是你要的！

学生G：那我们为啥(**wèi shá**)还要存这么多垃圾数据？
老师：因为那"1%"的价值，可能是亿万级的商业机会！💰

Slide 38: 방대한 양 vs 가치 있는 데이터

우리가 마주하는 것은 많은 데이터입니다 —— 아마도:

- 🗑️ 지루한 기록
- 💤 아무 일도 일어나지 않은 영상
- 🤯 중복된 센서 정보

하지만 그 중에서 만약 한 정보가 예상치 못한 사건을 포착한다면 (예를 들어 지진 순간의 영상), 그 단일 지점의 가치는 매우 높습니다!

💎 마치 사막에서 한 방울의 물을 발견하거나, SNS에서 짝사랑하는 사람의 셀카를 찾는 것처럼 —— 모두 "희귀 자원"입니다!

Slide 39: 가치 밀도 낮음, 다시 강조

다시 한번 강조하면, "가치 밀도가 낮다"는 의미는:

🔹 많은 데이터가 "평범하고", 별 의미가 없음 🔹 하지만 특수한 사건이 하나 발생하면, 이 작은 데이터는 극히 귀중해짐!

예를 들어 24시간 감시 카메라로 촬영한 내용 중 99.99%는 "아무도 지나가지 않는" 화면이지만, 도둑이 들어온 그 몇 초 —— 그게 바로 가치가 있는 것입니다!

Slide 38: 海量 vs 有价值数据

我们面对的是一大堆数据

——可能是：

🗑️ 无聊的记录

💤 没发生任何事件的视频

🤯 冗余**rǒng yú**的传感器信息

但其中，如果有一条信息捕捉**bǔ zhuō**到意外事件（比如某地震瞬间的视频），它的单点价值就非常高！

💎 像是在沙漠(**shā mò**)里发现一滴水，或者在朋友圈里找到你暗恋(àn liàn)对象的自拍(**zì pāi**)——都是"稀(**xī**)有资源"！

-

Slide 39: 价值密度低，再强调一下

再强调一次，"价值密度低"意味着：

🔹 很多数据都是"普通"的、没什么意义的

🔹 但只要发生一个特殊(**tè shū**)事件，这一小段数据就变得极其宝贵！

比如你24小时监控摄像头拍摄(**pāi shè**)的内容，99.99% 都是"没人路过"的画面，但某个小偷进来的那几秒钟——就值钱了！

Slide 40: 데이터 vs 쓸모없는 데이터

이 그림을 보세요. 왼쪽은 "데이터처럼 보이는" 많은 내용이지만, 사실... 모두 쓸모없는 쓰레기 데이터입니다.

학생H: 그럼 제가 SNS에 올린 음식 사진은 쓰레기인가요?
선생님: 여러분에게는 당연히 아니죠! 하지만 빅데이터 분석에게는 "맛 특성 인식 실패"의 픽셀 더미일 뿐일지도 모릅니다 😂

Slide 41: 예상치 못한 사건의 데이터 가치

이 그림이 강조하는 것은:
📸 사고 발생 순간의 영상
🚨 비정상 심전도
🌪️ 기상 급변의 데이터 수집

이러한 데이터는 비록 "조금"일뿐이지만, 단위 가치가 매우 높습니다!
그래서 데이터는 많은 것이 중요한 게 아니라, "적절한 시기"가 중요합니다.

Slide 40: 数据 vs 垃圾数据

你看这张图，左边是一大堆"看起来像数据"的内容，但其实……都是没用的垃圾数据。

学生H：那我朋友圈晒的美食图算不算垃圾？
老师：对你来说当然不是！但对大数据分析来说，也许就是"味道特征识别失败"的一堆像素图 😂

Slide 41: 意外事件的数据价值

这张图强调的是：
📸 一段事故发生瞬间的视频
🚨 一个异常心电图
🌪️ 一次气象突变的数据采集

这些数据虽然只是"一点点"，但它的单位价值非常高！
所以说，数据不在多，而在于"恰到好处(**qià dào hǎo chù**)"。

Slide 42: 단일 지점 가치 vs 대량 데이터

이 비교 그림을 다시 봅시다:

- 왼쪽: 많은 "평범한" 데이터
- 오른쪽: 한 중요한 순간의 기록, "다이아몬드"처럼 반짝반짝 빛남 ✨

마치 수만 개의 "오늘 뭐 먹었어" 트윗 중에, 갑자기 "프러포즈 성공" 한 개를 보면 —— 이것이 바로 즉시 인기 검색어에 오를 수 있습니다! 🔥

Slide 43: 반복 강조 + 이미지 복습

이 페이지들은 주로 시각적 복습으로, 오늘 배운 내용을 다시 한번 이미지 형식으로 제시합니다:

📌 **여러분은 핸드폰으로 사진을 찍어 노트로 사용할 수 있고, 4V 구호 카드를 만들 수도 있습니다:**

- **Volume**: 데이터 양이 많음
- **Variety**: 데이터 유형이 다양함
- **Velocity**: 처리 속도가 빠름
- **Value**: 가치 밀도가 낮음

학생I: 선생님, 빅데이터 주제 랩을 만들어도 될까요? **선생님**: 당연하죠! 다음 수업에 프리스타일로 들려주기를 기대합니다! (웃음)

Slide 42: 单点价值 vs 大量数据

再看这个对比图：

左边：很多"平平无奇"的数据

右边：一个重要瞬间的记录，像"钻石(**zuàn shí**)"一样闪闪发光 ✨

就像在上万条"今天吃了啥"的微博中，突然看到一条"求婚成功"的——这条立刻能上热搜(**rè sōu, hot search**)！🔥

Slide 43: 重复强调 + 图像回顾

这些页面主要是视觉回顾，把我们今天讲的内容再一次以图像形式呈现：

📌 你们可以拿手机拍下来当笔记，也可以做个4V口诀卡片：

- Volume：数据量大
- Variety：数据类型多
- Velocity：处理速度快
- Value：价值密度低

学生I：老师，我能出一个大数据主题RAP吗？

老师：当然！我期待你下节课freestyle给大家听！（笑）

퀴즈 정답

Slide 44:

✅ **정답: C**

Slide 45:

✅ **정답: C**

Slide 46:

✅ **정답: D**

Slide 47:

제4문제

어떤 유형의 데이터가 비구조화 데이터에 속하나요?

- A. Excel 스프레드시트
- B. SQL 데이터베이스의 주문 기록
- C. 이미지와 비디오
- D. 은행 거래 내역서

Slide 48:

✅ **정답: B**

Slide 49:

✅ **정답: C**

Slide 50:

✅ **정답: C**

Slide 51:

✅ **정답: D**

퀴즈 정답

Slide 44:

✅ 正确答案：C

Slide 45:

✅ 正确答案：C

Slide 46:

✅ 正确答案：D

Slide 47:

第4题

哪种类型的数据属于非结构化数据？

A. Excel 表格

B. SQL 数据库中的订单记录

C. 图片和视频

D. 银行流水账单

Slide 48:

✅ 正确答案：B

Slide 49:

✅ 正确答案：C

Slide 50:

✅ 正确答案：C

Slide 51:

✅ 正确答案：D

Slide 52:

✅ **정답: D**

Slide 53:

✅ **정답: B**

Slide 54: 마무리

좋습니다, 오늘 수업은 여기까지입니다!
오늘 우리는 "빅데이터 시대"의 전말을 알아보았고, 정보화 물결, 3대 지원 기술, 데이터가 어떻게 생성되고 어떻게 변화하는지 배웠습니다.
다음 수업에서는 계속해서 "빅데이터의 핵심 기술"을 탐구하겠습니다.

마지막으로 질문 하나: 오늘 여러분은 자신이 매일 사실 "데이터를 생산"하고 있다는 것을 발견했나요? 😉

Slide 52:

✅ 正确答案：D

Slide 53:

✅ 正确答案：B

Slide 54:

好了，我们本节课到这里就结束啦！
今天我们认识了“大数据时代”的来龙去脉(**lái lóng qù mài)**，从信息化浪潮、三大支撑技术，到数据如何产生、如何演变。

下节课我们会继续探讨“大数据的关键技术”。

最后提问一下：你今天有没有发现，自己每天其实也在“生产数据”呢？😉

谢谢大家，下课！👏👏👏