



# CSIE59830/CSIEM0410/ AIIA50050 Big Data Systems

Shiow-yang Wu (吳秀陽)  
CSIE, NDHU, Taiwan, ROC

## Happy Moon Festival !!



- Also known as **Mid-Autumn Festival** or **Mooncake Festival**
- One of the most important **cultural festival** in the Greater China region.



## What is Big Data?



- The **growth** of data in **Volume**, **Velocity**, **Variety** and **Veracity** (quality and accuracy) are in such an **unprecedented scale** that traditional data management systems can no longer handle it properly. (more Vs later)
- **New technologies** (AI/ML, IoT, ...) rely heavily on the processing of **huge data sets**.
- **Online services** (ChatGPT, YouTube, Meta, IG, ...) need to handle hundreds of millions of users issuing billions of request at the same time.
- We need **new frameworks, platforms, systems** and **tools** to deal with extremely large data sets and service requests.

## Examples of Big Data



- **Walmart**, the world's biggest retailer with over **10,900** stores across the globe serving more than **245 million** customers.
- Need to process **2.5 PB** of data every **hour**.
- **Data Café** – an **analytics hub** at Bentonville headquarters pulls in information from **200 sources**.
- **200 billion rows** of transactional data processing
- **Algorithms** are designed to blaze through them in **microseconds** to come up with **real-time solutions**.
- (<https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109>)(2023/07)

## Examples of Big Data



- Facebook(Meta) is another good example.
- 2.989 billion monthly active users, among them, 2.037 billion are daily active users (April 2023).
- Ad Revenues for Q1 2023: \$28.1 billion.
- 3<sup>rd</sup> most visited website outranked only by Google and YouTube (June 2023)
- Every 60 seconds, 136,000 photos are uploaded, 510,000 comments are posted, and 293,000 status updates are posted.
- (<https://www.simplilearn.com/how-facebook-is-using-big-data-article>)(2023/06)

## How Big? How Fast?



- About 1.7 MB of new data is created every second for every human.
- An average of 694,000 hours of video are streamed on YouTube every minute. (2023/02)
- Facebook generates 4 PB of data per day. (2023/03)
- 3.5(or 2.5) quintillion( $10^{18}$ ) bytes (or 3.5 million TB) of data are created each day. (2023/07)
- 97 zettabytes : the volume of data created worldwide in 2022. (Statista)
- 90% of the world's data has been created in the last two years.

## Activities in 60 Seconds



- 5.9 million Google searches take place.
- More than 231 million emails are sent.
- Around 500 hours of videos are uploaded on YouTube.
- Instagram users share around 66,000 photos.
- 1.7 million content pieces are shared on Facebook.
- 16 million text messages are sent.
- More than \$443,000 is spent by consumers on Amazon.
- Around 347,200 tweets are tweeted on Twitter.
- Over 120 professionals join LinkedIn.

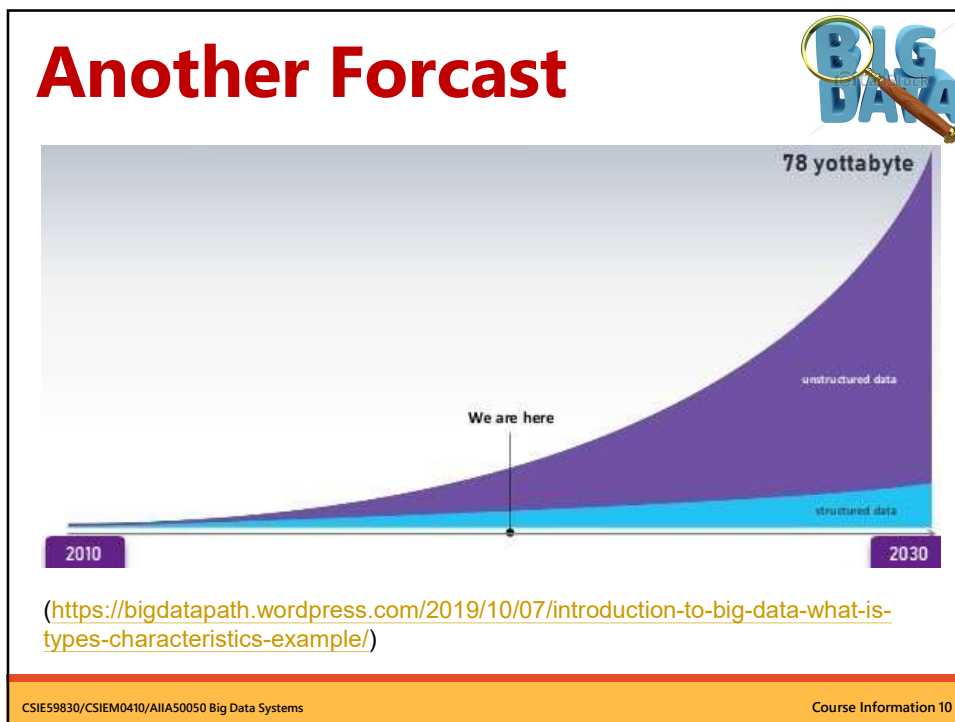
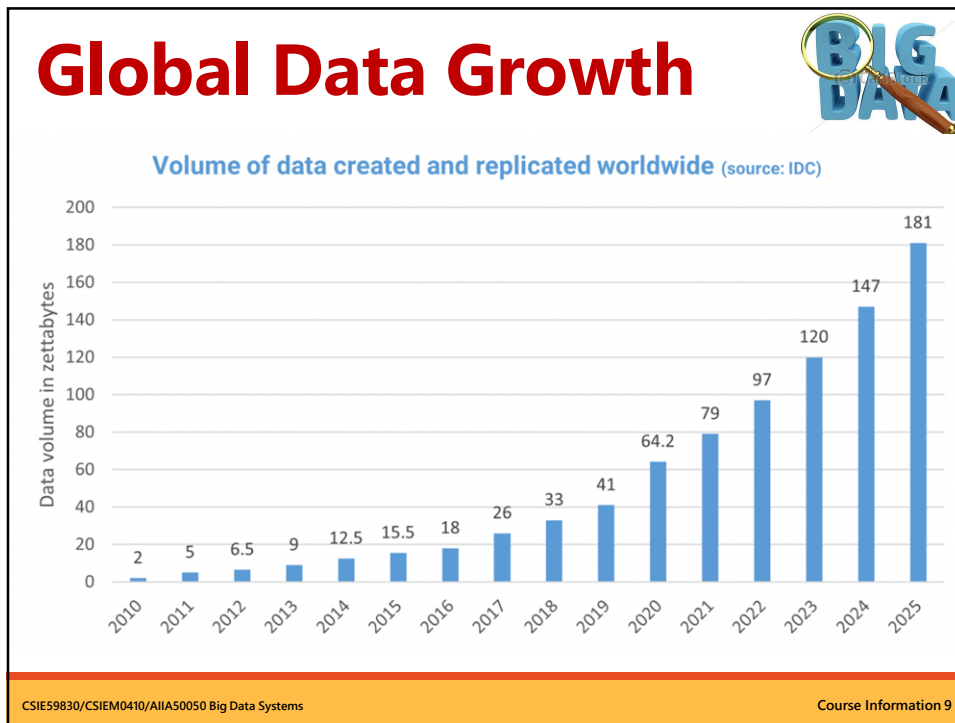
(<https://www.demandsage.com/big-data-statistics/>)

## Size of Data



- Exabyte, Zettabyte, Yottabyte, ... ?

Name	Symbol	Binary	Decimal
byte	B	$2^0=1$ byte	$10^0=1$
kilobyte	KB	$2^{10}=1.024$ byte (B)	$10^3=1.000$
megabyte	MB	$2^{20}=1.048.576$ B	$10^6=1.000.000$
gigabyte	GB	$2^{30}=1.073.741.824$ B	$10^9=1.000.000.000$
terabyte	TB	$2^{40}=1.099.511.627.776$ B	$10^{12}=1.000.000.000.000$
petabyte	PB	$2^{50}=1.125.899.906.842.624$ B	$10^{15}=1.000.000.000.000.000$
exabyte	EB	$2^{60}=1.152.921.504.606.846.976$ B	$10^{18}=1.000.000.000.000.000.000$
zettabyte	ZB	$2^{70}=1.180.591.620.717.411.303.424$ B	$10^{21}=1.000.000.000.000.000.000.000$
yottabyte	YB	$2^{80}=1.208.925.819.614.629.174.706.176$ B	$10^{24}=1.000.000.000.000.000.000.000.000$
brontobyte	BB	$2^{90}=1.237.940.039.285.380.274.899.124.224$ B	$10^{27}=1.000.000.000.000.000.000.000.000.000$
geopbyte	GeB	$2^{100}=1.267.650.600.228.229.401.496.703.205.376$ B	$10^{30}=1.000.000.000.000.000.000.000.000.000.000$



# IoT Trends in 2023

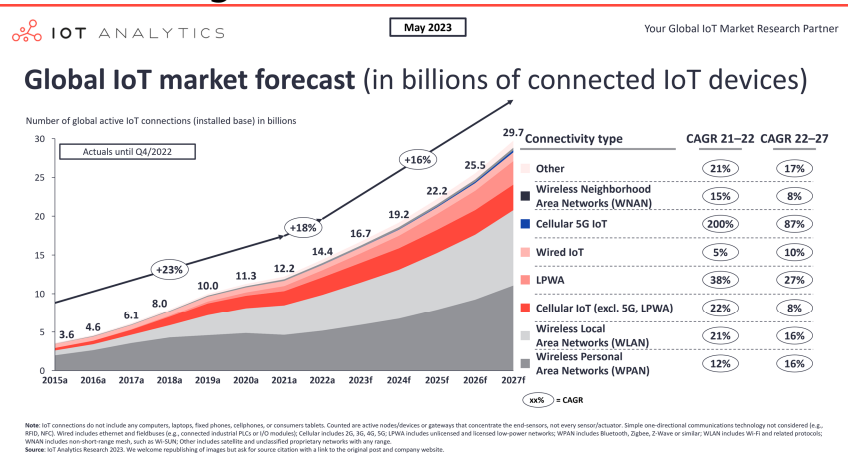


- **IoMT** (Internet of Medical Things), **IoRT** (Internet of Retail Things), **IoLT** (Internet of Logistics Things), **IoWT** (Internet of Workforce Things), .... (IoT everywhere)
- **16.7 billion** IoT devices in 2023, **>29 billion** by 2027. (IoT Analytics)
- IoT market size: from **\$201 billion**(2023) to **\$483 billion**(2027) (IoT Analytics)
- Most IoT applications are all built on top of **streaming big data analytics**.

# IoT Big Data



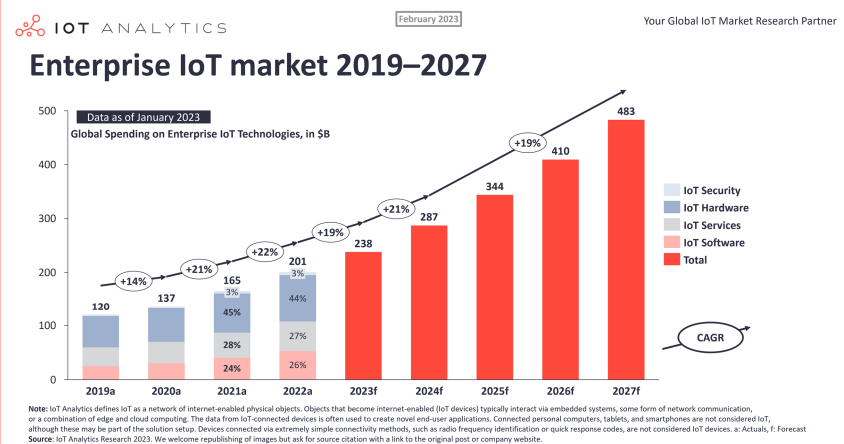
- **Internet of Things(IoT)** push even further the need for big data.



## IoT Market Size



- Global IoT market shows resilience despite economic downturn. (IoT Analytics)



## Big Data and AI



- The relationship between big data and AI has been described as **symbiotic**(共生) or **synergistic**(協同).
- Big data analytics leverages AI/ML for better data analysis.
- AI/ML rely on massive scale of data to learn and improve decision-making.
- The **true power** is unleashed when they **join forces**.



## Why Big Data?

From business point of view:

- Big data can unlock significant value by **making information transparent**.
- Big data can help organizations collect more accurate and detailed **operational information** to expose variability and boost performance.
- Big data allows **ever-narrower segmentation** of customers and therefore much more **precisely tailored** products or services.

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Course Information 16



## Why Big Data?



- Sophisticated analytics can substantially **improve decision-making**, **minimize risks**, and **unearth valuable insights** that would otherwise remain hidden.
- Big Data can be used to develop the **next generation** of **products** and **services**.
- Big Data is a “**Big Deal**”!
- Big Data means **Big Business**!

## Main Functions of Big Data

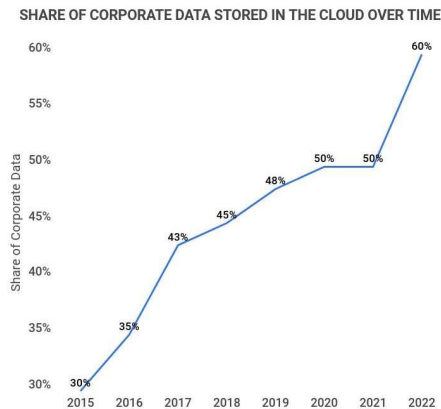


- **Descriptive Analytics**: figure out **what** happened in the **past**
- **Diagnostic Analytics**: understand **why** something happened in the **past**
- **Predictive Analytics**: assess the **probability of occurrence** in the **future**. Early warning, fraud detection, preventative action, forecasting, etc.
- **Prescriptive Analytics**: gives the user precise (prescriptive) **recommendations** (respond to the queries such as “What should I do if “x” occurs?”)

## Corporate Data Trend



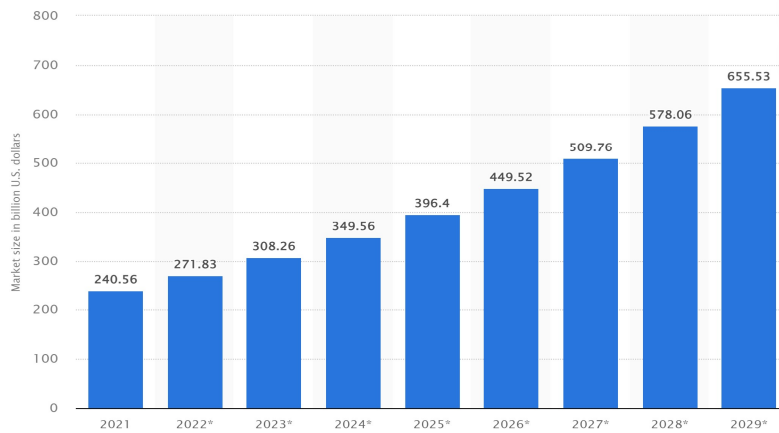
- As of 2022, 60% of corporate data worldwide is stored in the cloud. (Zippia)



## Global Big Data Market



- The global market size is expected to reach \$655.53 billion by 2029. (Statista)



## What about this course



- This is an **introductory course** on big data concepts, processing and systems.
- To **get hands-on experience** with popular **open source big data tools**: Hadoop, Spark, HPCC, HBase, Cassandra, MongoDB, Structured Streaming, Storm, SAMOA, Kafka, Flink, Neo4j, ...

## Systems vs Analytics



- “**Play**” with open-source big data tools.
- Learn **how to use** them in applications.
- Understand the underlying **principles** and **mechanisms**.
- More on big data **processing technologies and systems**.
- Limited cover on big data analytics (another course).
- Big data trends

# MAD(ML, AI, and Data) Landscape 2023

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE

Version 1.0 - Feb 2023 © Matt Turck (@mattturck), Kevin Zhang (@kzhang13) & FirstMark (@firstmarkapp) Blog post: mattturck.com/MAD2023 Interactive version: MAD.firstmarkapp.com Comments? Email: MAD2023@firstmarkapp.com **FIRSTMARK**

(https://www.ixahub.com/stories/key-takeaways-from-the-2023-ml-ai-data-landscape-report)

CSIE59830/CSIEEM0410/AIIA50050 Big Data Systems
Course Information 23

# Topics 1

- Introduction
  - What is big data?
  - Why big data?
  - Examples of big data
  - The challenges and opportunities of big data
- **General purpose** big data platforms
  - Distributed and cluster computing
  - Apache Hadoop and MapReduce
  - Cloudera(CDH, Cloudera Distribution for Hadoop)
  - Apache Spark and in-memory computation
  - HPCC(High Performance Computing Cluster), also referred to as DAS(Data Analytics Supercomputer)



## Topics 2



- Big data storage **architecture**
  - Distributed nodes
  - Scale-out NAS
  - All-solid-satae drive (SSD) arrays
  - Object-based storage
  - DNA storage
- Big data storage **systems**:
  - Distributed file systems and big data storage
  - Google GFS and Apache HDFS
  - Cloud Storage
- **Data lake**\*\*
- Big data storage **security**\*\*

\*\* : If time allows

## Topics 3



- Big data systems for **structured/semi-structured data**
  - SQL, NoSQL, NewSQL, Distributed SQL
  - Apache HBase, Cassandra, CouchDB, Drill, Impala, Hive
  - Spark SQL, DataFrames, Datasets
  - MongoDB
  - Google BigQuery, Spanner, F1
  - Presto

## Topics 4



- Big **graph** processing
  - The challenges of big graph processing
  - Pregel family of systems(BSP, Pregel, Giraph)
  - GraphLab family of systems(GraphLab, PowerGraph, GraphChi)
  - Spark GraphX, GraphFrames
  - Neo4j graph database
  - Titan distributed graph database
- **RDF** processing systems\*\*
  - NoSQL-based RDF systems
  - Hadoop-based RDF systems
  - Spark-based RDF systems

## Topics 5



- Big data **stream** processing
  - The challenges of **big data streaming**
  - Spark Streaming, Structured Streaming
  - Apache Storm, Samza, Flink
  - Apache SAMOA
- **IoT streaming** big data processing
- Streaming big data **applications**

## Topics 6



- Big data **pipelining** tools
  - Pig Latin
  - Tez
- Big data **ETL**(extract, transform, and load) and **integration** tools
  - Apache Airflow
  - Apache Kafka
  - Apache Camel
- Big data **orchestration** tools\*\*

## Topics 7



- Big data **analytics**, other systems and trends
  - Google Cloud Platform (GCP) vs Amazon Web Services (AWS)
  - RapidMiner
  - KNIME
  - Tableau
  - R Language and RStudio
- Open Data
- Beyond Hadoop and Spark
- Big data systems landscape

## Special Topic



- **Big data** and **AI** :
  - The relationship between big data and AI
  - How do big data and AI work together?
  - Powerful synergy of big data and AI
- How **ChatGPT** and other similar **Large Language Models**(LLMs) work?
- **LLaMA**, the open source LLM from Meta
- The future of big data and generative AI

## Special Topic



- **IoT** data stream processing:
  - ThingsBoard
  - DeviceHive
  - Kaa
  - DSA (Distributed Services Architecture)
  - SiteWhere
  - Node-RED
  - OpenRemote
  - Zetta
  - ThingSpeak



## Administrative Information



- Course Title: Big Data Systems
- Course Number: CSIE5983/CSIEM0410/AIIA50050
- Lecture Time: Tue 14:10~17:00
- Classroom: Engineering Building C305
- Office Hours: Tue 17:00~18:00
- Grading Policy:
  - Assignments 35%
  - Independent Study and Presentation 15%
  - Final exam 25%
  - Term project 25%

## Course Related Pages



- Course homepage:  
<http://web.csie.ndhu.edu.tw/showyang/BigDataSys2023f/index.html>
- Instructor's homepage:  
<https://web.csie.ndhu.edu.tw/showyang/>
- All lecture notes will be available online.

## Textbook & References



- No required textbook.
- Jawwad Ahmad Shamsi and Muhammad Khojaye. *Big Data Systems: A 360-degree Approach*. Chapman & Hall/CRC, 2021.
- Balamurugan Balusamy, Nandhini Abirami R, Seifedine Kadry, Amir H. Gandomi. *Big Data: Concepts, Technology, and Architecture*. Wiley, 2021.
- Sherif Sakr. *Big Data 2.0 Processing Systems: A Systems Overview, 2nd Edition*, Springer, 2020.
- S. Sasikala and D. Renuka Devi (Authors), Raghvendra Kumar (Editor). *Research Practitioner's Handbook on Big Data Analytics*. Apple Academic Press, 2023.
- Guido Dartmann, Houbing Song, et al. *Big Data Analytics for Cyber-Physical Systems: Machine Learning for the Internet of Things*. Elsevier Science Publishing Co Inc, 2019.
- Kai Hwang and Min Chen. *Big Data Analytics for Cloud, IoT and Cognitive Computing*. John Wiley & Sons Ltd., 2017.
- Tom White. *Hadoop: The Definitive Guide, 4th Edition*, O'reilly, 2015.
- Jure Leskovec, Anand Rajaraman, Jeff Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2010~2014.
- Mohammed J. Zaki and Wagner Meira JR. *Data Mining and Machine Learning - Fundamental Concepts and Algorithms, 2nd Edition*. Cambridge University Press, 2020.

## Assignments



- There will be several programming assignments for you to get hands-on experience on big data tools/systems.
- A lecture topic will start with its origin (mostly Google) and then its open source counterpart (mostly Apache).
- Learn to use **virtual machines** (VirtualBox, VMWare, Cloudera, ...) to build your Hadoop/Spark cluster on your desktop.
- You can then build your own cluster with lab machines for true parallel processing.

## Independent Study



- All students are to conduct independent study on self-selected topics (discussed with me first).
- Pick an **open source** big data system **not discussed in the class** as the study target.
- Prepare a presentation and a demonstration of the system in class.
- Every student must present and demo.

## Exam



- No midterm exam. Term **project proposal** and **tool demo** instead.
- A **final exam** at the final exam week.
- The exam questions will be on the basic concepts, systems and applications of big data techniques.
- Only the lecture part. Student presentation not included.
- Open book(notes, code, ...).
- No electronic devices allowed.

## Term Project



- There will be a modest scale term project for you to show your creativity.
- May use any big data tools for your project.
- Prepare a **project proposal** and a **tool(s) demo** on **VMs** by the end of the midterm week.
- Prepare a **project demo** at the end of the semester to explain your project to me.
- Turn in the **project** and **report** one week after the final exam.

## Recommendations



- Read the assigned readings before the class, participate in the discussion, **ask questions!**
- Learning by doing. Start early!
- You will need to learn **Linux** and **Python**.
- Grading policy revisited:
  - Assignments (35%)
  - Independent study and presentation (15%)
  - Final exam (25%)
  - Term project (25%)

## Resources




- Wikipedia, Big Data.  
([http://en.wikipedia.org/wiki/Big\\_data](http://en.wikipedia.org/wiki/Big_data))
- Wikibook, Data Science: An Introduction.  
([http://en.wikibooks.org/wiki/Data\\_Science:\\_A\\_n\\_Introduction](http://en.wikibooks.org/wiki/Data_Science:_A_n_Introduction))
- Apache Hadoop (<http://hadoop.apache.org/>)
- Apache Spark (<https://spark.apache.org/>)
- You can use any VM software (VirtualBox, VMWare, Cloudera, ...) for your virtual cluster.

## ChatGPT(AI in general) Policy



- With such a useful tool like ChatGPT, there is no reason to ban it.
- **ChatGPT coding** is becoming an important practice in the industry.
- However, with electronic calculator, you still need to learn arithmetic.
- With ChatGPT, you still need to learn big data technologies.
- It is OK to use it in assignments.
- You **CANNOT** use it in the **exams** !!


## Is Big Data Dying or Dead?



- Data never sleeps. It grows only faster.
- So how can big data be dying?
- **Big data processing** is at the **heart** of AI, ML, IoT & many new services.
- It is becoming a **norm**.
- **Every** data engineer need to know big data.

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 43

## Big Data is Alive and Kicking



- **DBMS** used to be considered only needed for **large** data.
- Now every company and organization use DBMS with no need to emphasize "large".
- **Data** is being generated **everywhere** at an **exponential rate**.
- Big data is becoming the **norm**.
- Every company and organization will use **big data systems** with no need to emphasize "big".

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Course Information 44