




Big Data: Concepts, Challenges & Opportunities

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

Lecture material is mostly home-grown, partly
taken with permission and courtesy
from Professor Shih-Wei Liao of NTU.



Outline

- What is Big Data? (The Big Data Phenomena)
- Big data examples
- Big data concepts
- Challenges and opportunities
- Summary
- Next: Big Data Computing & Systems

Size of Data



- What is the maximum file size you have dealt with so far?
 - Movies/Files/Streaming video that you have used?
 - What have you observed?
- What is the maximum download speed you get?
- Simple questions:
 - What's the capacity of your HD?
 - You know GB, TB, right?
 - How about PB, EB, ZB, or YB?

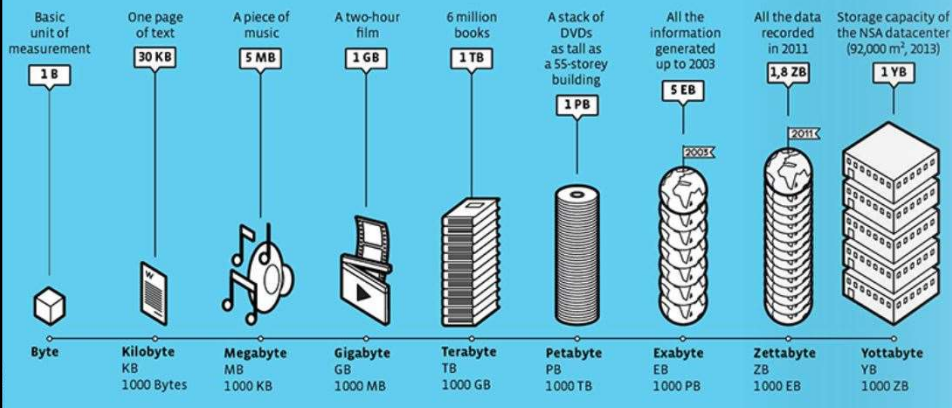
Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Still no idea?

Scale of Data in Bytes



COMPARATIVE SCALE OF BYTES

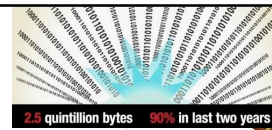


Data in Bytes



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	$10^{30}=1.000.000.000.000.000.000.000.000.000.000$

What is Big Data?



- 1.7MB of data is created per **second** per **person**.
- 90% of world's data has been created in **past 2 years**.
- 328.77 million **terabytes** is created **every day**.
- 463 **exabytes** (10^{18}) of data each **day** by 2025.
- 95 million **photos** and **videos** shared every **day** on IG.
- 120 **zettabytes** data worldwide in 2023. Will reach 181 **zettabytes** by 2025.
- 15.14 **billion** IoT devices are connected worldwide.
- To download **all the data** on the internet, an internet user will need approximately **180 million years**.

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

Data Everywhere



- Lots of data is being collected and warehoused
 - Web data, e-commerce
 - Purchases at department/grocery stores
 - Bank/Credit Card transactions
 - Social Networks
 - Photos/Videos sharing
 - Smart home IoT sensors
 - Robots, UAV/UGV
 - ...



Huge Amount of Data



- There are huge volumes of data in the world:
 - From the **beginning of recorded time** until **2003**, we created **5 exabytes** of data.
 - In **2011**, the same amount was created every **two days**.
 - In **2013**, the same amount of data is created every **10 minutes**.
 - In **2020**, **every person** generates **1.7 MB** in just a **second (146.88 GB a day)!!**
 - In **2023**, the data volume worldwide is **120 ZB**.

Every Minute on the Internet



- 2.4b(2014), 3.4b(2016), 3.8b(2017), 4.4b(2019), 4.66b(2020), 5.18b(2023) **internet users worldwide.**
- **Every minute** on the Internet:
 - **Netflix** users stream 452K hours of video
 - **Instagram** users share 66K photos and video
 - **Youtube** users upload 500 hours of videos
 - **Twitter** users send 575K tweets
 - **Facebook** : 2.1M active users
 - **Spotify** add 28 tracks
 - **Email** users send 231.4M messages
 - **Zoom** hosts 104.6k hours of meetings
 - **Google** users conducts 5.9M searches
 - **Consumers** : 6M shopping online

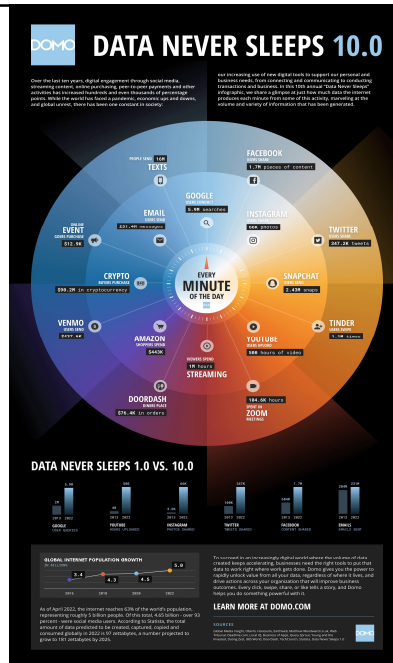
640K ought to be enough for anybody.



(<https://www.domo.com/data-never-sleeps>)

(<https://localiq.com/blog/what-happens-in-an-internet-minute/>)

Data Never Sleeps



© Can Stock Photo - 109120457

Activities in 60 Seconds



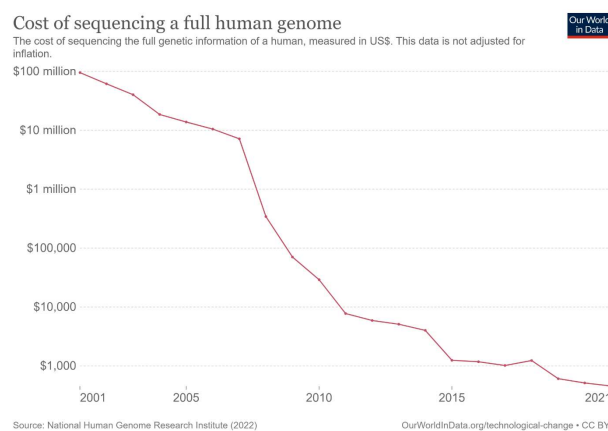
- 5.9 million Google searches take place.
- More than 231 million emails are sent.
- Around 500 hours of videos are uploaded on YouTube.
- Instagram users share around 66,000 photos.
- 1.7 million content pieces are shared on Facebook.
- 16 million text messages are sent.
- More than \$443,000 is spent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

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

Fallen Cost of Processing



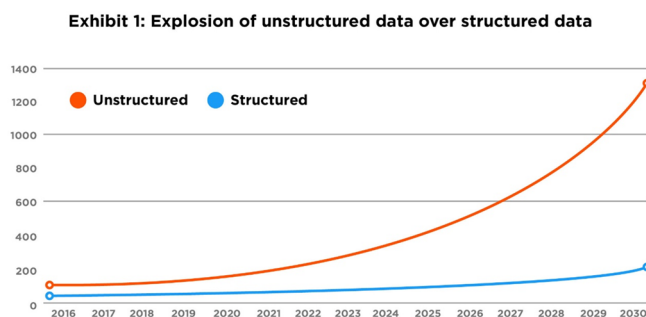
- The cost of processing continues to fall (eg. cost of sequencing a full human genome):



Semi-/Unstructured Data is Growing Much Faster



- Traditional data is **structured** (text, tables, DB, ...).
- New data is mostly **semi-** or **unstructured** (XML, audio/video, graphs, ...) which is **growing much faster**.



The Big Data Phenomena



- “**Big Data**” is used to characterize the **phenomena**:
 - Huge data, growing faster than ever!
 - Reduced processing cost
 - But don’t know how (semi-/unstructured data)
- **Jim Gray** described the big data phenomena as the **Fourth Paradigm** and called for a **paradigm shift** in the computing architecture and large-scale data processing mechanisms.
- Need to develop a **new generation** of **computing systems** and **tools** to **manage, analyze, and visualize** the data flood.

How Big is Big?



- The size to be considered “big” is changing.
- Big Data was referred to one gigabyte or 1 GB in 1999.
- Today, the term symbolizes **peta bytes** (1024 terabytes), **exabytes** (1024 petabytes) or **zettabytes** (1024 exabytes).
- Stu Feldman (Chief Scientist of Schmidt Futures, ex Google VP) says at least **10TB** in terms of **data rate**.

Is Big Data a real discipline standing on its own?



- Some heavyweights said “Big Data is not new. Database and data mining have been around for more than 30 years.”
- Big data already disrupted the field of data models and relational database and demanded new ways of building systems. (量變造成質變)
- In the case of data mining, see the free book by Professor Ullman: “*Mining of Massive Datasets*” (<http://www.mmds.org/>)

Is Big Data just a small part of Cloud Computing?




- Some said “Big Data is just a small part of Cloud Computing. Don’t make a big deal out of it.”
- The truth is:
 - They are technologies with different focuses.
 - Cloud focuses more on elastic computing and warehouse computing.
 - Big Data focuses more on managing huge data sets for enterprise cloud and possible-time/real-time analytics.
- It’s a big deal:
 - Many impossible business model → possible now.

Big Data Implication



- Big Data Everywhere :
 - E.g., Google indexed World Wide Web pages
 - Which demanded the creation of MapReduce and NoSQL.
- Big Data is not just an isolated discipline: Big Data is not just **red hot** in one discipline.
- When data explodes, **new data properties** appear, **new business applications** emerge, which solidify the discipline.

Big Data Applications




Applications Everywhere: Domain knowledge, Business models

Analytics: Ad-hoc analytics, statistics, AI & machine learning

Systems: Tools, Infrastructure

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 19

Big Data from CSO's Perspective



- New revenue vs. cost reduction
 - Top-line vs. Bottom-line
 - CSO (Chief Strategy Officer) vs. CIO (Chief Information Officer)
- First, here we talk about Big Data from CSO's perspective
- IBM says, **Big Data = Big Business**
- Explore new business models and services.

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 20

Big Data from CIO's Perspective

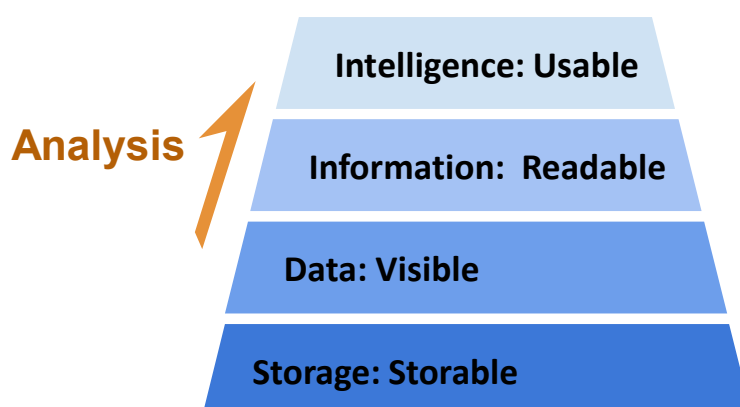


- CIO = Chief Information Officer
- To CIOs, Big Data means:
 - No more expensive **machines**: use commodity systems
 - No more expensive **DB** software: use open-source, NoSQL/NewSQL/Distributed SQL systems
 - No more expensive **storage**: No RAID. Just common hard drives
- To CIOs, Big Data means **large scale distributed computing** with **commodity systems** on **open-source software**.


Back to Basics: What is Data?



Texts, Records, Statistics...




Digital World



JAN 2023 **ESSENTIAL DIGITAL HEADLINES**
OVERVIEW OF THE ADOPTION AND USE OF CONNECTED DEVICES AND SERVICES


TOTAL POPULATION



8.01 BILLION

URBANISATION **57.2%**


UNIQUE MOBILE PHONE USERS



5.44 BILLION

vs. POPULATION **68.0%**

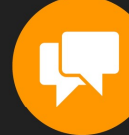
INTERNET USERS



5.16 BILLION

vs. POPULATION **64.4%**

ACTIVE SOCIAL MEDIA USERS



4.76 BILLION

vs. POPULATION **59.4%**


SOURCES: UNITED NATIONS, GOVERNMENT BODIES, GAMA INTELLIGENCE, ITC, WORLD BANK, EUROSTAT, CHINIC, APRI, HAMA & KANTAR, CIA WORLD FACTBOOK, COMPANY ADVERTISING RESOURCES AND PUBLISHED REPORTS, CQMM, BETA RESEARCH CENTER, KEPRO ANALYTIC. **ADVISORY:** SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. **COMPARABILITY:** SIGNIFICANT REVISIONS TO SOURCE DATA INCLUDING SCHEMATIC REVISIONS TO POPULATION DATA, RESULTS ARE **NOT COMPATIBLE** WITH PREVIOUS REPORTS. ALL FIGURES USE THE LATEST AVAILABLE DATA, BUT SOME SOURCE DATA MAY NOT HAVE BEEN UPDATED IN THE PAST YEAR. SEE NOTES ON DATA FOR FULL DETAILS.

we are social | Meltwater

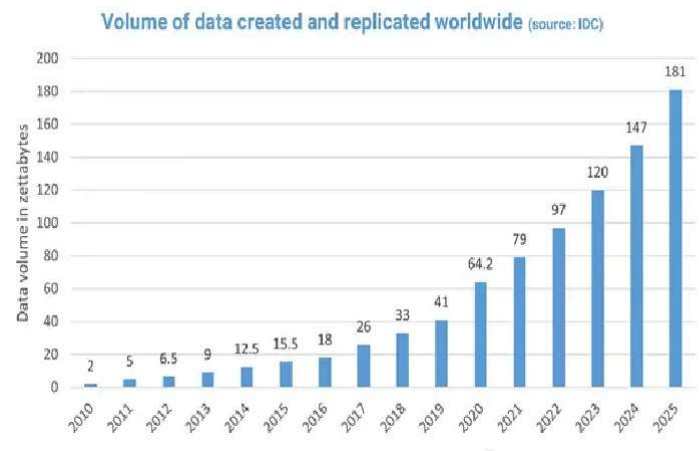
(<https://datareportal.com/reports/digital-2023-global-overview-report>)


CSIE59830/CSIEM0410/AIIA50050 Big Data Systems
Introduction 23

Data is Growing Exponentially



Volume of data created and replicated worldwide (source: IDC)





CSIE59830/CSIEM0410/AIIA50050 Big Data Systems
Introduction 24

Big Data Challenges



- Storing and managing of data
- Preparing for scalability
- Timely analytics and actionable insights
- Integration of data from different sources
- Data quality and security
- Selecting and exploring the right tools
- Need talent and skilled people
- Cost management

Where is Big Data?

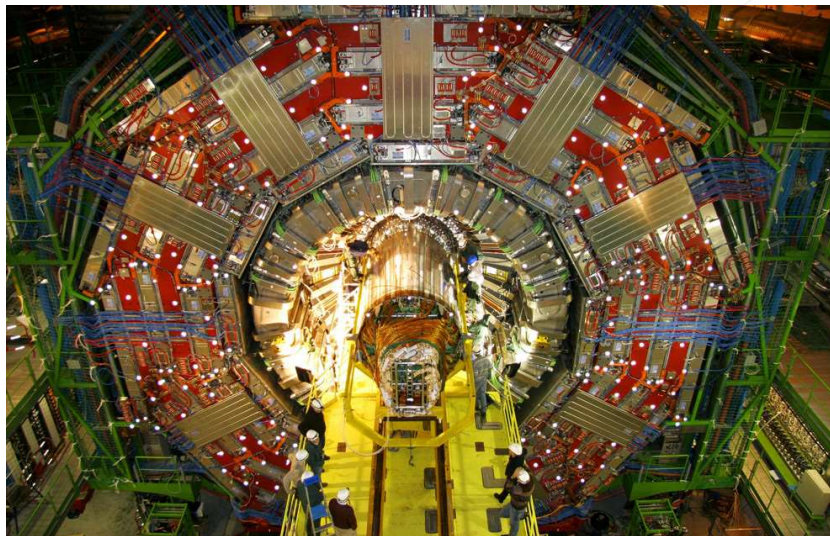


© Can Stock Photo - 129720457



Big Data Examples

Large Hydron Collider (LHC)



LHC Aerial Photo



CSIE59830/CSIEM0410/AIIA50050 Big Data Systems

Maximilien Brice, © CERN

Introduction 29

Large Hadron Collider (LHC)



- 150 millions sensors
- 1 billion collisions per second
- Recording 30+ petabytes of data per year
- 100+ petabytes of data are permanently archived
- The data are distributed on the Worldwide LHC Computing Grid (WLCG) for analysis.
- Shut down at the end of 2018 for major upgrades.
- Became operational again on 22 April 2022.

(<https://home.cern/resources/faqs/facts-and-figures-about-lhc>)

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems

Introduction 30

The Earthscope (地球鏡)



- The **Earthscope** was one of the world's largest science project(2003-2018) to track North America's geological evolution.
- Records data over 3.8 million square miles with 4,000+ connected instruments generating 67 terabytes of data.
- It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more.



(<https://www.earthscope.org/>)

WalMart



- World's biggest retailer
- **10,900** stores across the globe(2023)
- More than **245 million** customers
- Walmart **Data Café** (private cloud) on **250 node Hadoop** cluster
- Process **2.5 PB** of data every **hour**.
- **200 billion rows** of transactional data
- Information from **200 streaming sources**
- Algorithms to blaze through them in **micro seconds** to come up with **real-time solutions**.

(<https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109>)

Facebook(Meta)



- Over **2.95 billion** monthly active users(MAU), among them, **1.62 billion** users visit every day.
- Every **minute**:
 - 400 users sign up
 - 510,000+ comments are made
 - 293,000 status updates
 - 136,000 photos are uploaded
 - 4 million posts are liked
- Generates **4 petabytes** of data per day (stored in **Hive** containing about **300 petabytes** of data)

(<https://thesocialshepherd.com/blog/facebook-statistics>)


(<https://kinsta.com/blog/facebook-statistics/>)

More Big Data Examples




- 25 Big Data Examples and Applications
(<https://builtin.com/big-data/big-data-examples-applications>)
- 14 Big Data Examples & Applications Across Industries
(<https://www.simplilearn.com/tutorials/big-data-tutorial/big-data-applications>)
- Top 10 Big Data Applications in Real Life
(<https://intellipaat.com/blog/10-big-data-examples-application-of-big-data-in-real-life/>)


Who's Generating Big Data




Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)



Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)




- Progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely and scalable way

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems
Introduction 35



The Model Has Changed...

- The model of generating/consuming data has changed

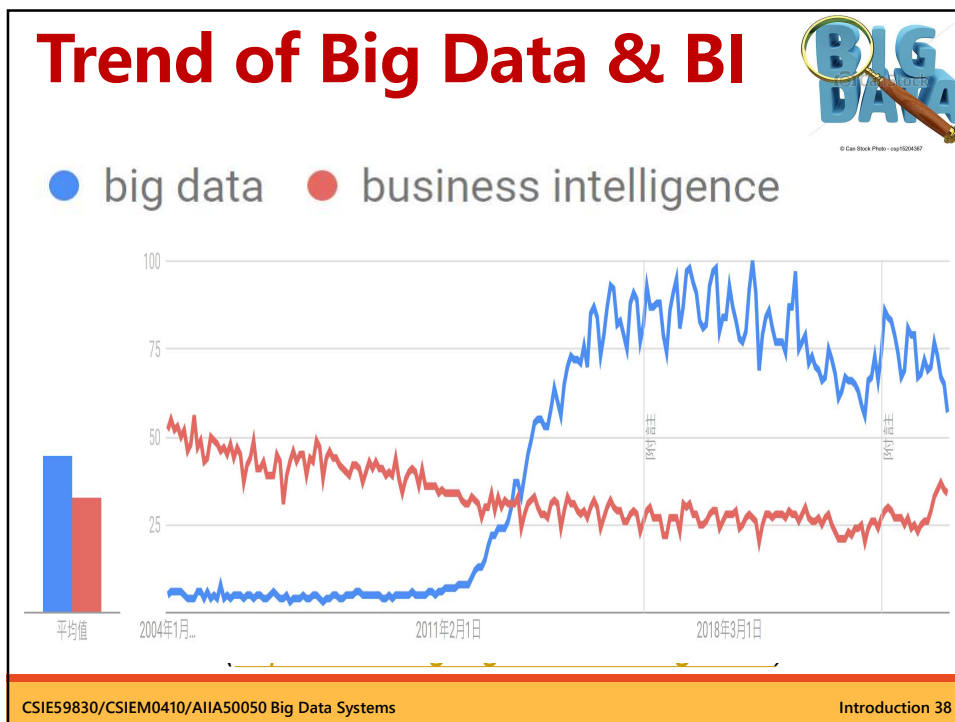
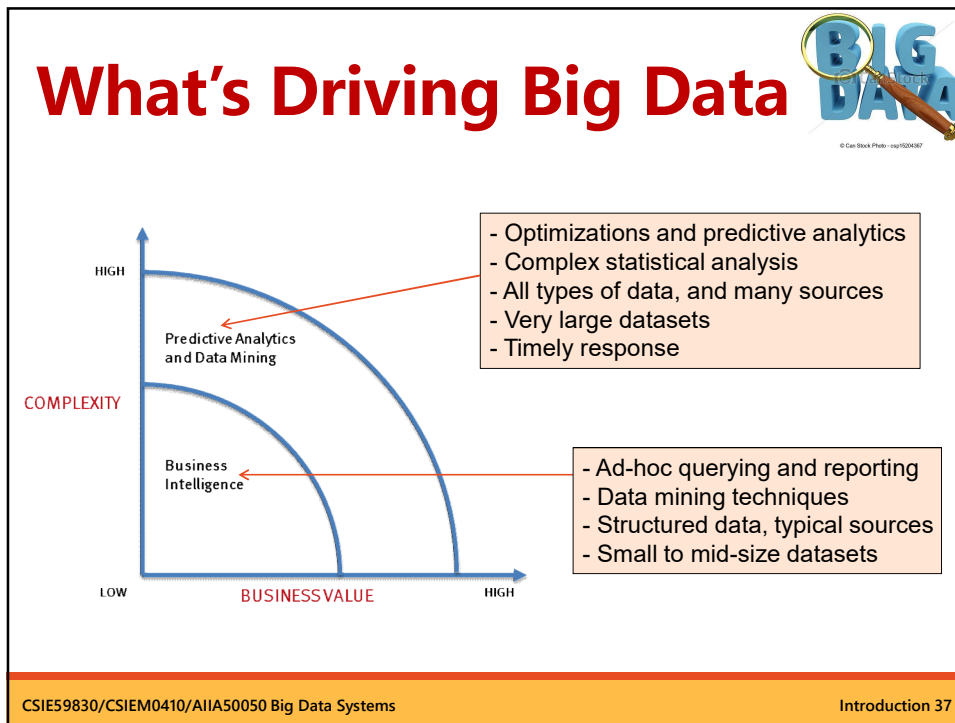
Old Model: Few companies are generating data, all others are consuming data

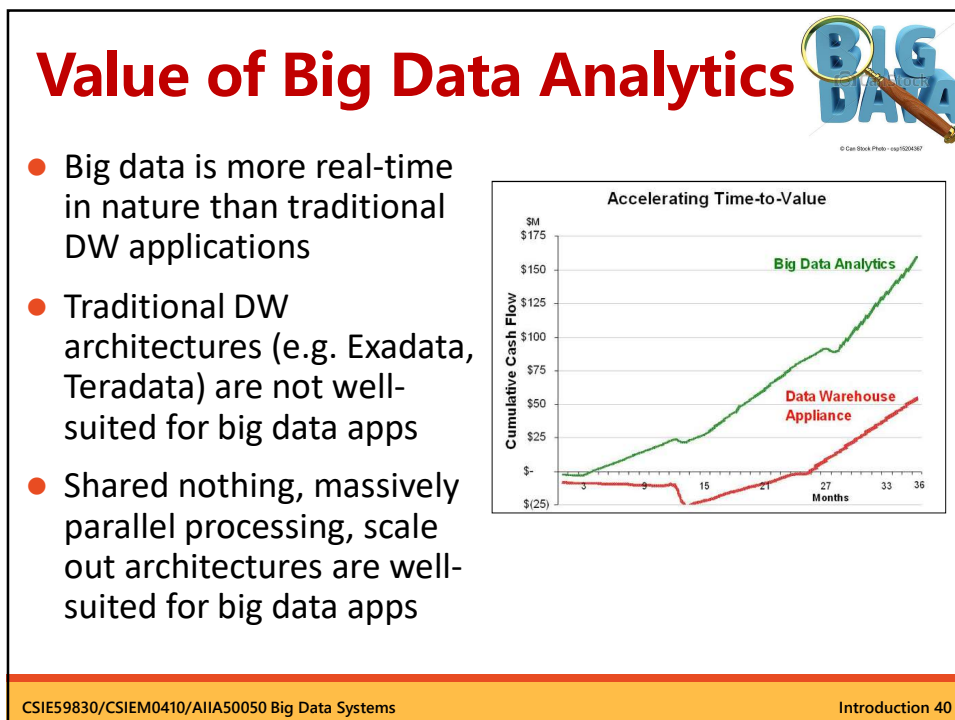
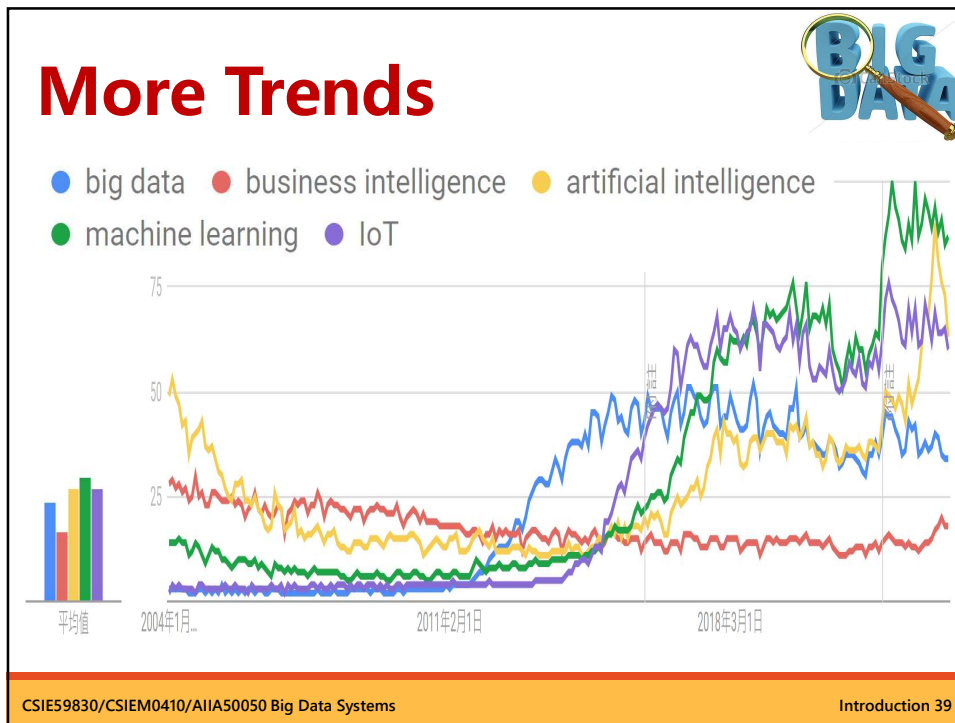


➔


New Model: all of us are generating data, and all of us are consuming data


➔


CSIE59830/CSIEM0410/AIIA50050 Big Data Systems
Introduction 36





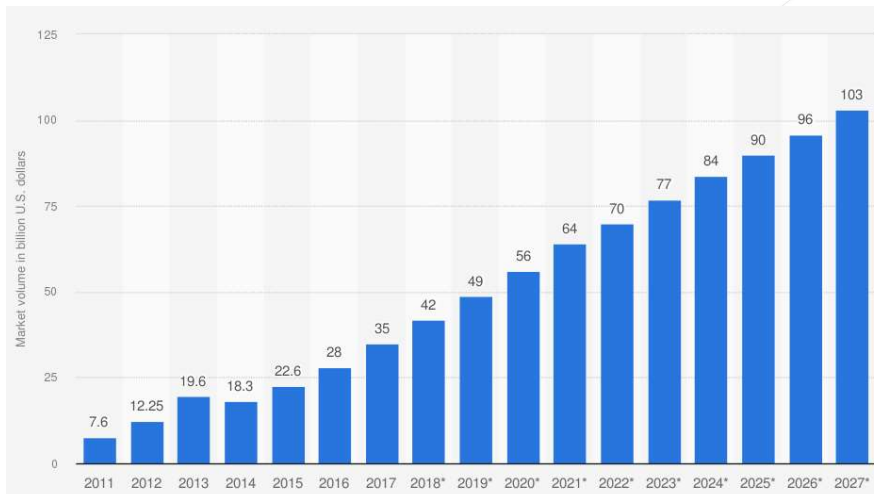
Data Business



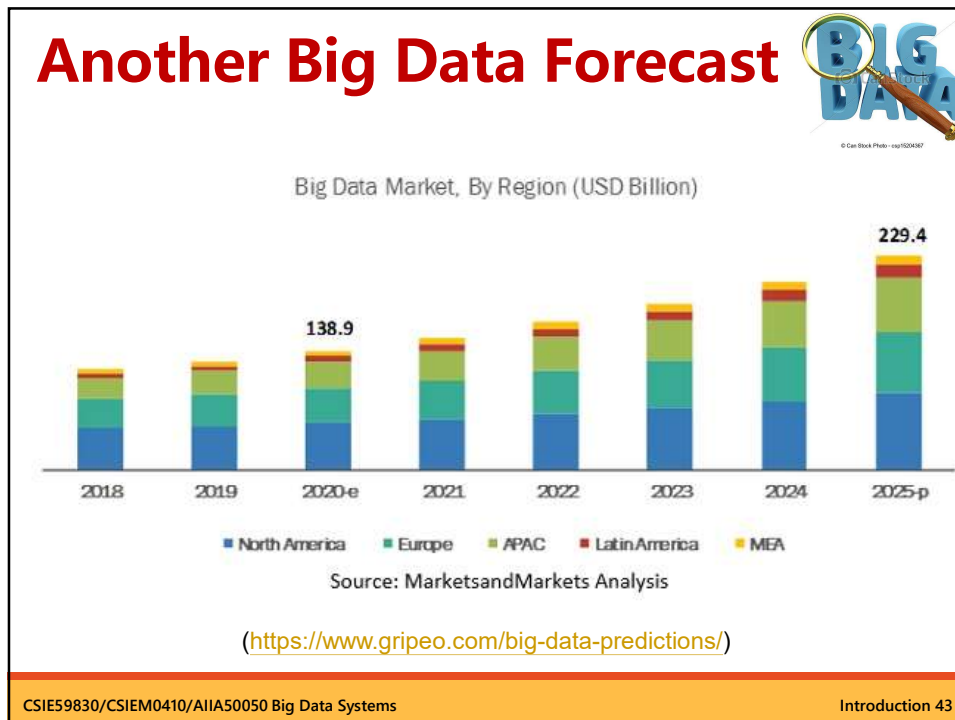
- Retails(零售)
- Traffic(交通)
- Health(健康)
- Education(教育學習)
- Manufacturing(製造)
- Contents(數位內容)
- Agriculture(農業)
- Advertising(廣告)
- Telecommunication(電信)
- Finance(金融)
- Smart grid(智慧電網)
- ...



Global Big Data Market



(<https://www.statista.com/statistics/254266/global-big-data-market-forecast/>)



Types of Data

- Structured data (Tables/Transaction/Legacy Data)
- Text data (Web)
- Semi-structured data (Email, XML, markup language code, ...)
- Unstructured data
- Multimedia data (images, audio, video, VR/AR)
- Graph data
 - Social Network, Semantic Web (RDF), ...
- Streaming data (sensors, IoT, ...)
 - Real-time, can only scan it once

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 44

Value of Big Data



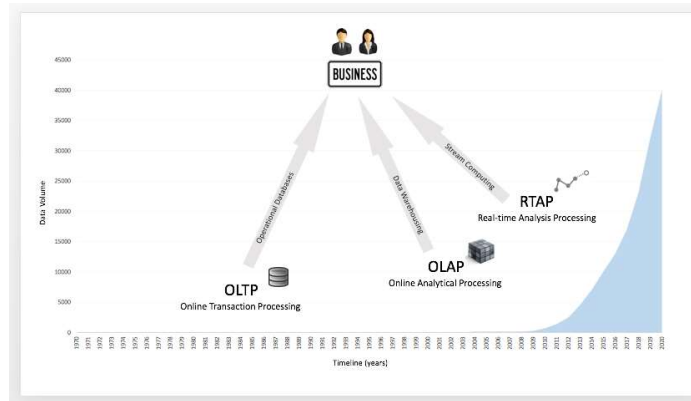
- When the data size grows, **intrinsic properties** of data emerges.
- “A kilo of data is worth more than a gram of algorithm”. (一斤資料勝過一兩演算法)
- A small improvement of algorithm is no longer important.
- The key is what the **data** tells us.
- The focus is on designing algorithms that **scale** !!

What to do with these data?



- Aggregation and Statistics
 - Data warehouse and OLAP
- Indexing, Searching, and Querying
 - Keyword based search
 - Pattern matching (XML/RDF)
- Knowledge discovery
 - Data Mining
 - Statistical Modeling
- Machine learning

Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data warehousing)
- **RTAP:** Real-Time Analytics Processing (Big data technology)

Myths About Big Data



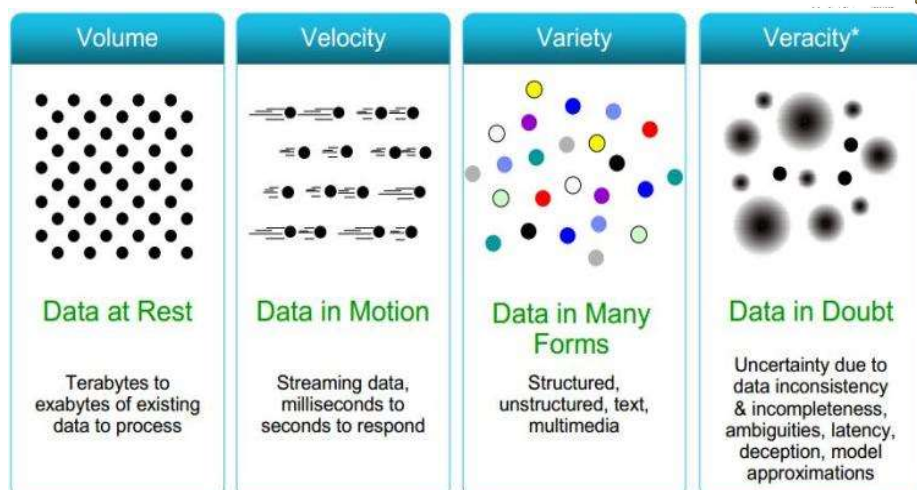
- **“Big Data is Only About Massive Data Volume”**
 - Volume is just an element of Big Data
- **“Big Data is all-powerful”**
 - Can get All Of The Data
 - Big Data Yields Certainty
 - Can answer WHY

“Big Data is Only About Massive Data Volume”?



- **4 Vs**
 - **Volume** : The starting point of Big Data, but the least important of 4 elements.
 - **Variety** : Traditional data management processes can't cope with the heterogeneity of big data.
 - **Velocity** : Data is generated in real time, with demands for usable information to be served up immediately.
 - **Veracity** : Refers to the biases, noise and abnormality in data. How to make data to be trusted for the organization to make crucial decision?

4 Vs of Big Data



4 Vs of Big Data

Big data Expands on 4 fronts

<http://whatis.techtarget.com/definition/3Vs>

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 51

4 Vs of Big Data

The FOUR V's of Big Data

Volume: SCALE OF DATA

- 40 ZETTABYTES (43 TRILLION TERABYTES) of data will be created by 2020, an increase of 300 times from 2009.
- 6 BILLION PEOPLE have cell phones.
- WORLD POPULATION: 7 BILLION
- It's estimated that 2.5 QUINTILLION BYTES (2.5 TRILLION QUADRANTS) of data are created each day.
- Most companies in the U.S. have at least 100 TERABYTES (100,000 GIGABYTES) of data stored.

Velocity: ANALYSIS OF STREAMING DATA

- The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session.
- Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure.
- By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS - almost 2.5 connections per person on earth.

Variety: DIFFERENT FORMS OF DATA

- As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES (150 BILLION GIGABYTES).
- By 2014, it's anticipated there will be 420 MILLION WEARABLE WIRELESS HEALTH MONITORS.
- 4 BILLION+ HOURS OF VIDEO are watched on YouTube each month.
- 30 BILLION PIECES OF CONTENT are shared on Facebook every month.
- 400 MILLION TWEETS are sent per day by about 200 million monthly active users.

Veracity: UNCERTAINTY OF DATA

- Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors, and mobile devices. Companies can leverage data to predict their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
- By 2015, 4.4 MILLION IT JOBS will be created globally to support big data, with 1.2 million in the United States.
- 1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions.
- 27% OF RESPONDENTS
- in one survey were unsure of how much of their data was inaccurate.
- Poor data quality costs the US economy around \$3.1 TRILLION A YEAR.

Source: McKinsey Global Institute, TechCrunch, Gartner, EMC, SAS, IBM, NESTEC, Q&Q

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 52

The 5 Vs of BigData

The diagram illustrates the 5 Vs of Big Data as five horizontal arrows pointing outwards from a central vertical axis. Each arrow is labeled with a 'V' and has a corresponding definition to its right. The background is purple with a magnifying glass icon in the top right corner.

- Volume** (blue arrow): The size of the data
- Velocity** (green arrow): The speed at which the data is generated
- Variety** (orange arrow): The different type of data
- Veracity** (red arrow): The trustworthiness of the data in terms of accuracy
- Value** (pink arrow): Just having BigData is of no use unless we can turn it into value

Powered by StackDataLabs

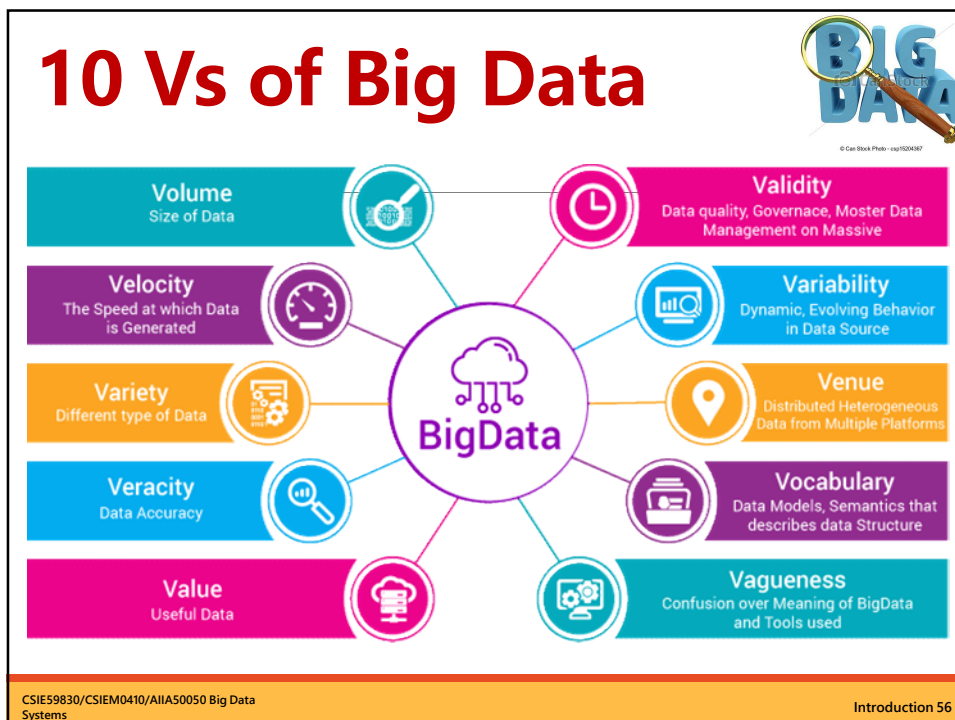
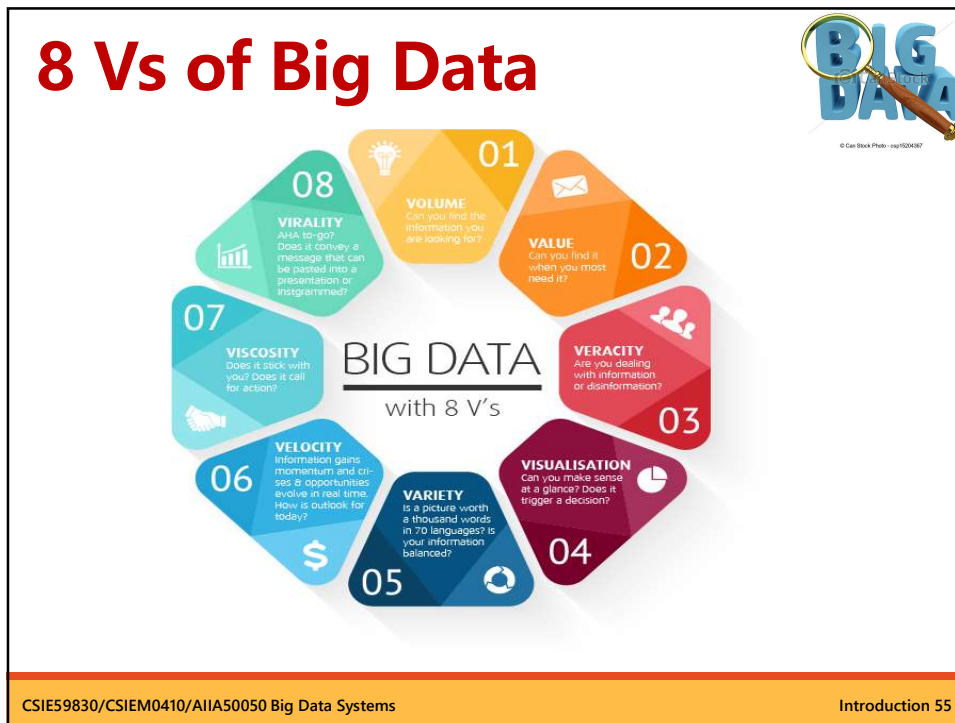
CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Course Information 53

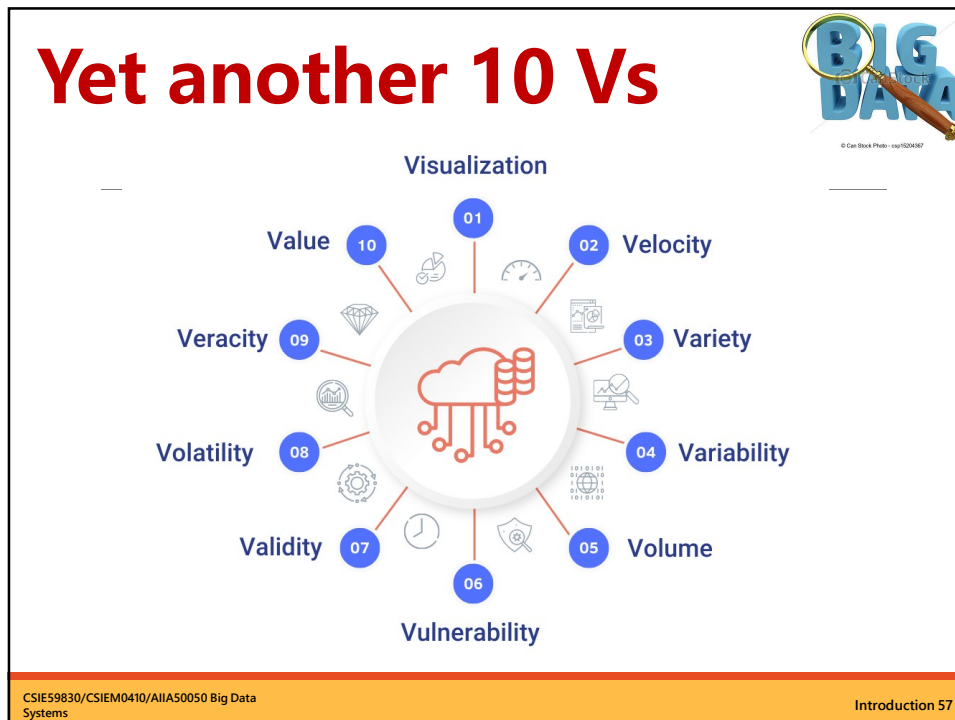
5 Vs of Big Data

The diagram shows the 5 Vs of Big Data arranged in a circle around a central pentagon labeled '5 Vs of Big Data'. Each 'V' is in a box and has a list of characteristics below it. The background is white with a magnifying glass icon in the top right corner.

- Volume**
 - Terabytes
 - Records/Arch
 - Transactions
 - Tables, Files
- Velocity**
 - Batch
 - Real/near-time
 - Processes
 - Streams
- Value**
 - Statistical
 - Events
 - Correlations
 - Hypothetical
- Veracity**
 - Trustworthiness
 - Authenticity
 - Origin, Reputation
 - Availability
 - Accountability
- Variety**
 - Structured
 - Unstructured
 - Multi-factor
 - Probabilistic

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 54











Meanings of Some Vs

- **Validity:** The issue of collecting data which is correct and accurate for the intended use.
- **Volatility:** How long is data valid and how long should it be stored.
- **Variability:** Big data is variable, i.e. variance in meaning, changing of meaning (rapidly).
- **Visualization:** Making data comprehensible, easy to understand and read.
- The list keeps growing to **42 Vs** (you cannot be serious!?)

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems

Introduction 58

Big Data is not just Hadoop

Understand and navigate federated big data sources		Federated Discovery and Navigation
Manage & store huge volume of any data		Hadoop File System MapReduce
Structure and control data		Data Warehousing
Manage streaming data		Stream Computing
Analyze unstructured data		Text Analytics Engine
Integrate and govern all data sources		Integration, Data Quality, Security, Lifecycle Management, MDM

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 59

Problems when data is BIG

- How to store ?
- How to retrieve ?
- How to process ?
- How to analyze ?
- How to handle streaming data in real-time?

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 60

Emerging Technologies for Managing Big Data



- Architecture
- Storage
- Computing
- Graph
- Database/Data warehousing
- Stream processing
- Real-time Analytics & Business knowledge
- Big data as a service

How to store?



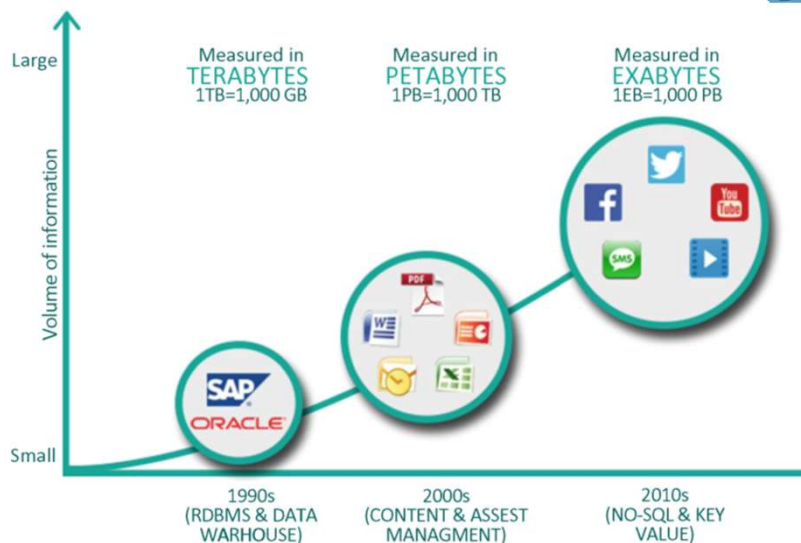
- It's not likely to store it on a single machine
 - Facebook generates TBs of data every day
 - **> 500 hours** of content are uploaded to YouTube **every minute**. That's **82.2 years** of new video every day.
- Distributed File System
 - Google File System (GFS)
 - Hadoop Distributed File System (HDFS)
- Big data storage systems

Example: How to retrieve ?



- You may want to use a traditional database system to organize data
 - MySQL, PostgreSQL,
- Unfortunately, they don't scale well to big data level...
 - One naive reason is that they usually run on only 1 machine.

Evolution of Data/Tech



Structure of Big Data



- The structure of data can be classified into:
 - **Structured data**: Data with a defined format and structure (RDB, spreadsheets, CSV, ...)
 - **Semi-structured data**: Textual data files with a flexible structure that can be parsed (XML, ...)
 - **Quasi-structured data**: Textual data with erratic data formats (Web click stream data, ...)
 - **Unstructured data**: Data that have no inherent structure (text docs, PDF files, images, videos, ...)
- Use different tools for different cases.

Latency Requirements

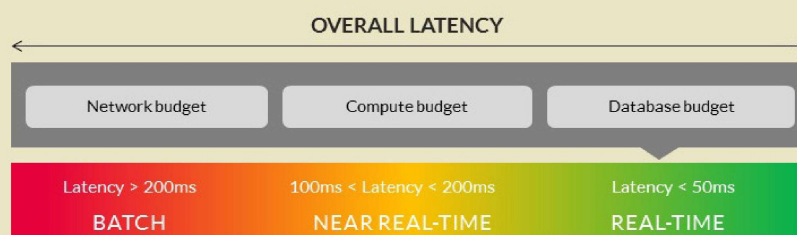
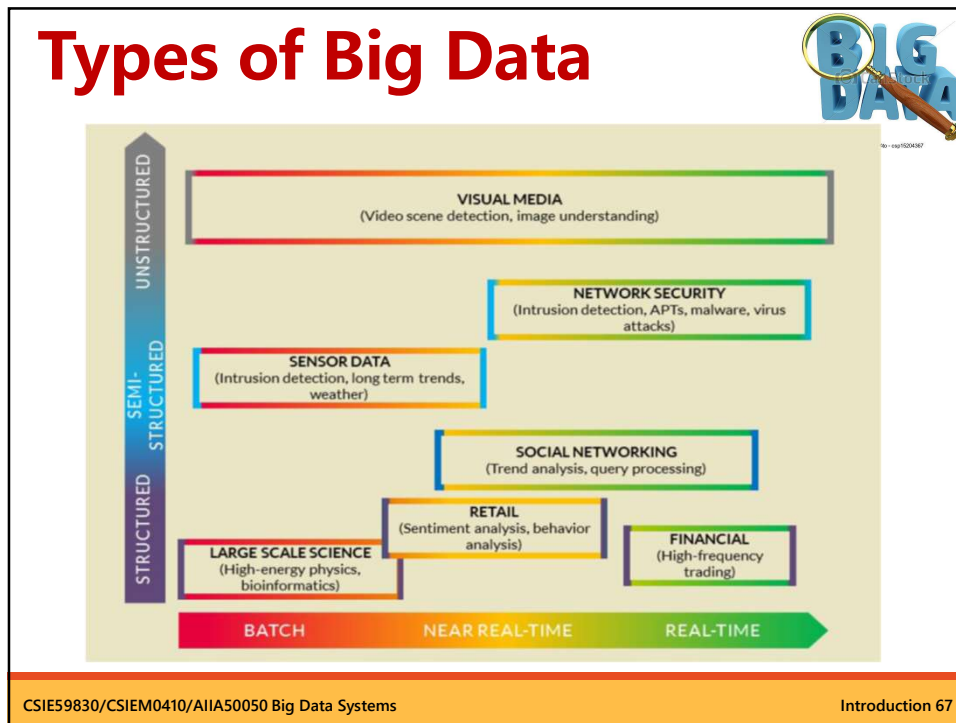



Figure 2: Characterization of latency requirements


- Low latency requirements generally imply that the data must be processed as it comes in.




Storage & Warehousing


- BigTable / HBase
- Cassandra
- MongoDB
- Hive / Spark SQL

APACHE HBASE 

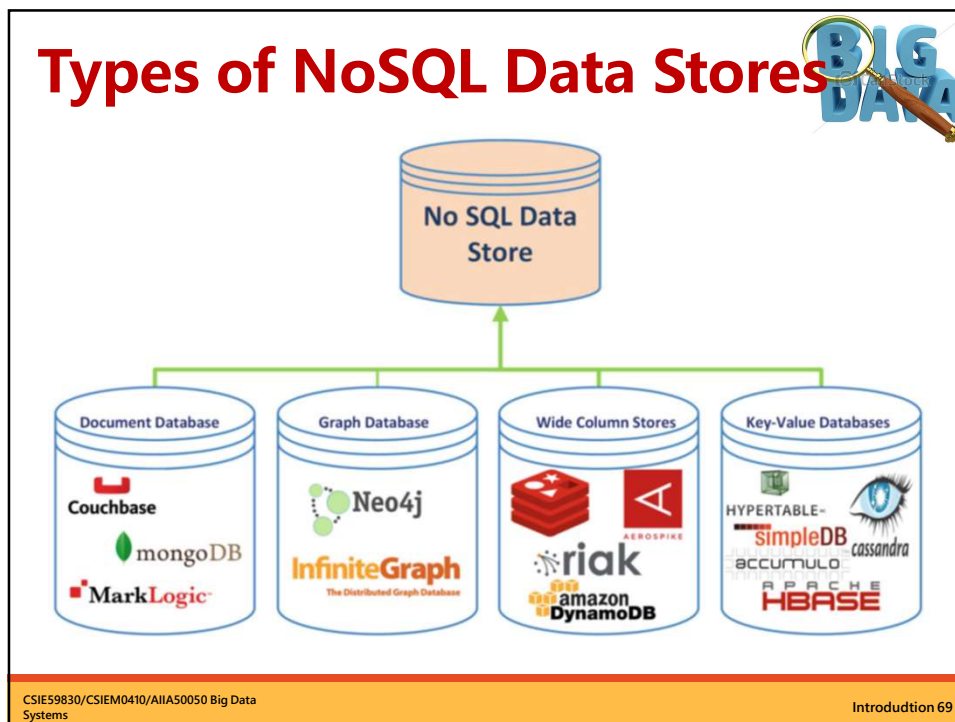
Cassandra 

Spark SQL 

HIVE 

mongoDB 

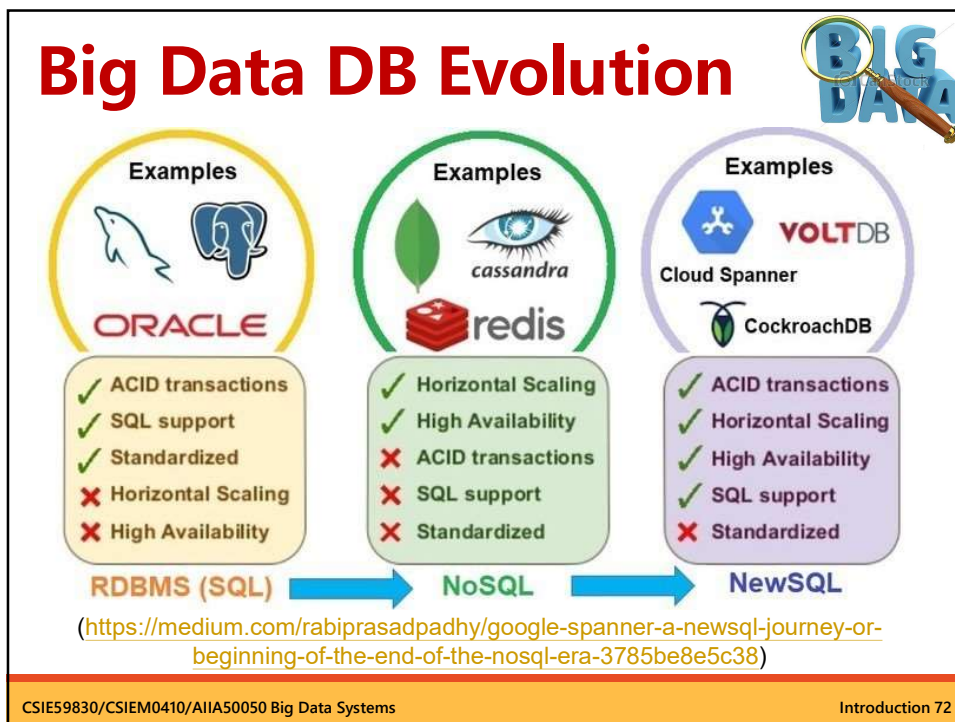
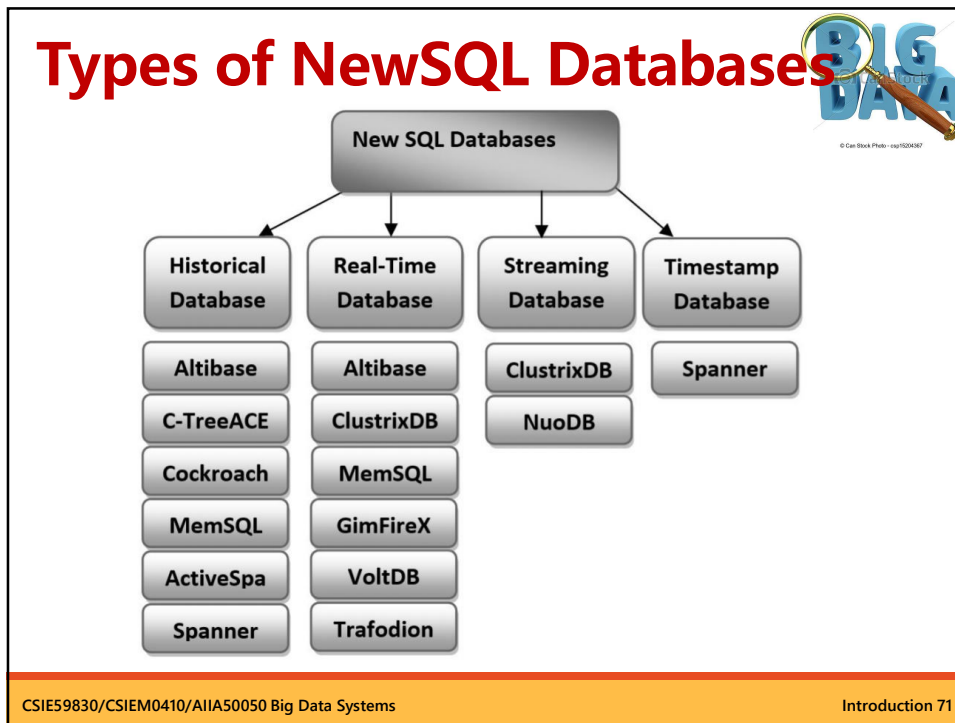
CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 68

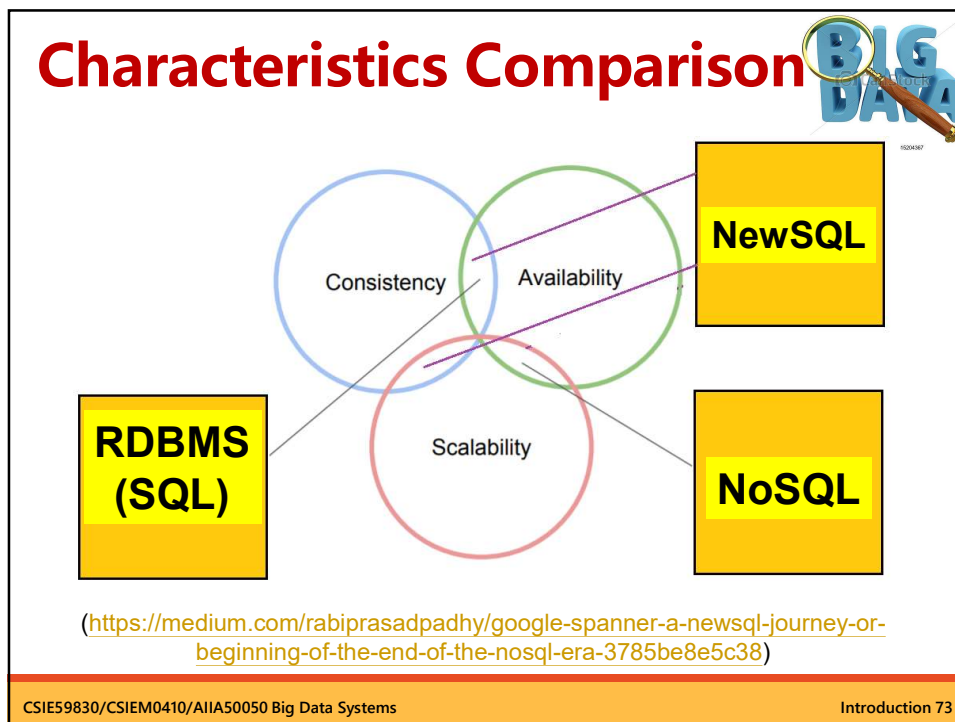


NewSQL Databases

- ACID guarantee of traditional SQL databases is good but not scalable.
- NoSQL databases scale out well but do not support ACID.
- **NewSQL databases** come to solve the problem.
 - Support ACID
 - Distributed
 - Can be scale-out
 - Handle large volume of data with great performances

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 70








How to process ?



- Finally, we can get the data we need efficiently from the monster-like data set.
- And it's time to do something cool now
 - Retrieval, mining, learning, ...
- But you'll soon face some trouble...
 - Data can't fit in memory / disk on a single machine
 - Not powerful enough with a single machine

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 74

Computing




- MapReduce 
- Massive Parallel Processing
- Spark: in-memory computing 
- STORM: real-time streaming
- ...

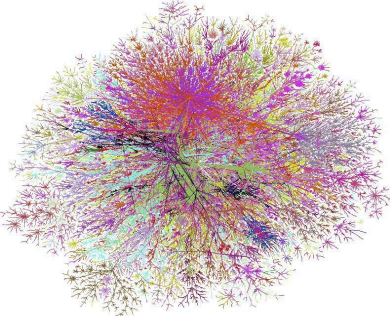



CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 75

Big Graph Computing




- **Graphs** abstract entities and interactions using **vertices** and **edges**
- **Graph algorithms** handle generic problems on graphs and can be adapted to real problems
- Many applications call for the processing of **large graphs**.



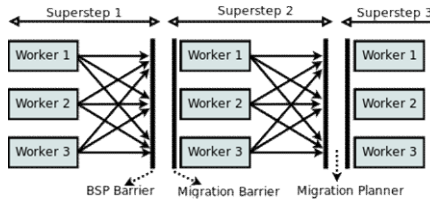
A hairball graph depicting the internet in 2004.

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 76


Big Graph Processing





- Pregel, BSP
- Giraph
- Neo4j Graph DB
- ArangoDB
- Apache TinkerPop
- ...



BSP Barrier Migration Barrier Migration Planner






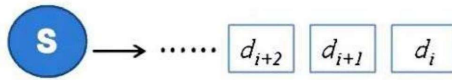


CSIE59830/CSIEM0410/AIIA50050 Big Data Systems
Introduction 77


Streaming Big Data



- A **streaming data source** is a potentially unbounded data source that keeps generating **data stream** over time.
- A **data stream processing node** is a node that accepts one or more input data streams, processes the data in some way, and generates one or more output data streams.



Data Source



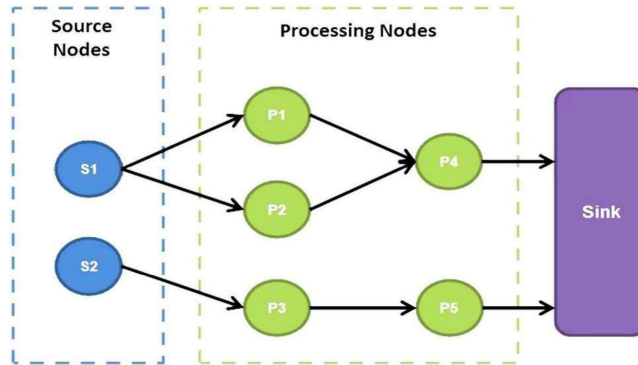
Processing Node

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems
Introduction 78

Stream Processing Platform



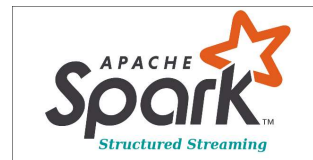
- A **stream processing platform** provide constructs and libraries for modeling, implementing and executing applications over streaming data flows.

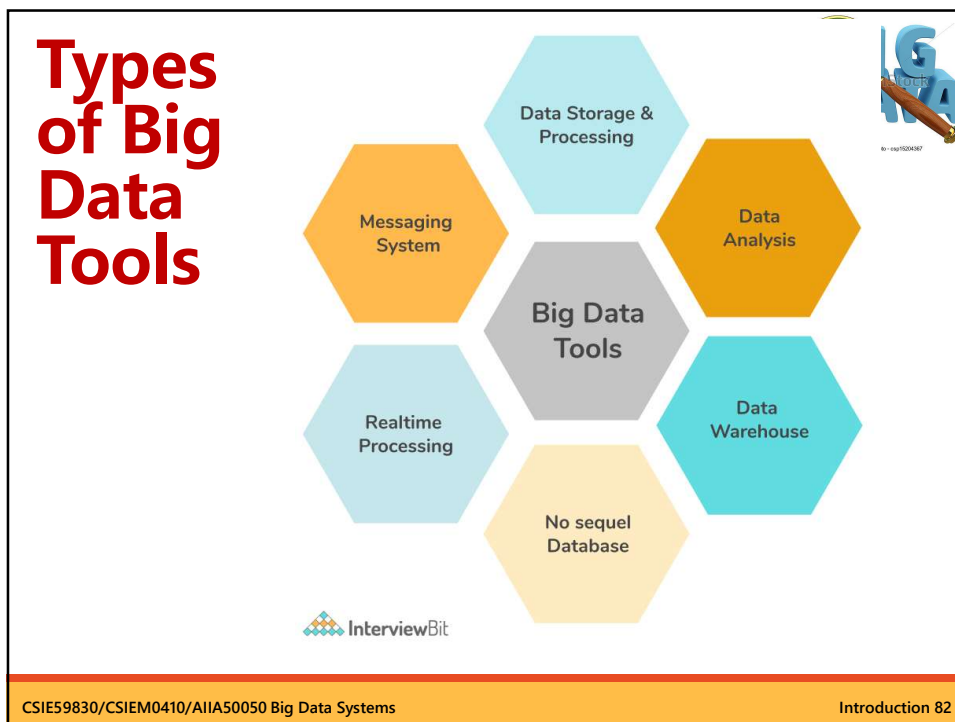
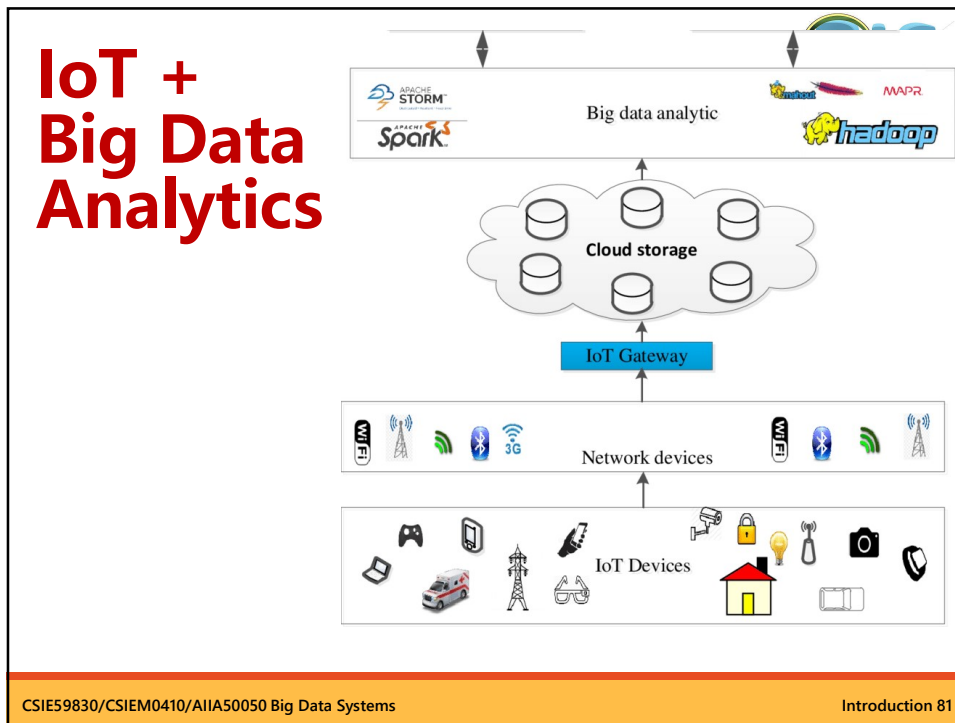


Stream Analytics Platforms



- Apache Kafka
- Spark/Structured Streaming
- Apache Storm
- Apache Flink
- Apache Samza
- ...





Big Data Ecosystem

DATA SOURCES

Internal data sources such as data from CRM system, ERP system, sales reports, etc.

External data sources such as government statistics and media channels

DATA STORAGE

Big data storage software tools store, manage and retrieve massive amounts of data.

DATA MINING

Data mining tools allow businesses to extract usable data from a huge set of raw data to find relationships, patterns, and anomalies.

DATA ANALYTICS

Although data mining tools incorporate data analysis, there are software designed specifically with advanced analytical capabilities.

DATA VISUALIZATION

Data visualization software is also a type of data analytics tool. However, they are specifically designed to take the raw data and presenting it with beautiful and easy digestible visuals like graphs and charts.

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 83

Evolving Big Data Tools

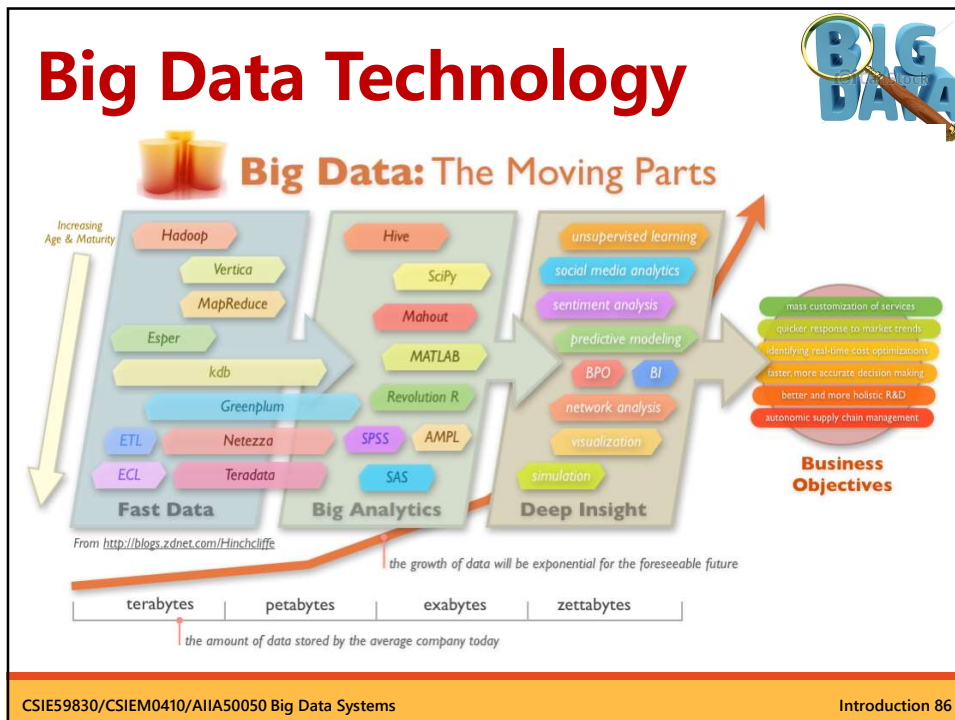
ANALYTICS			
DATA ANALYST PLATFORMS		DATA SCIENCE PLATFORMS	
Microsoft, pentaho, alteryx, Digital Reasoning, guavus, AYASDI, ATTIVO, Datameer, Quid, incorta, interana, ClearStory, Origami, ENDOR, MODE, Bottlenose, switchboard		IBM, KNIME, data iku, DOMINO, rapidminer, CONTINUUM ANALYTICS, ALGORITHMIA, DATAWATCH, ANGOSS, SAS	
BI PLATFORMS	VISUALIZATION	MACHINE LEARNING	
Microsoft, aws, Domo, Wave Analytics, looker, THOUGHTSPRINT, ATSCALE, ARCADIA DATA, Information Builders, GoodData, MicroStrategy, birst	tableau, SAP, Google Cloud, celonis, Qlik, Periscope Data, ZEPL, GOMDATA, plotly, CHARTIO, YOU CAN TOGO	Azure Machine Learning, aws, Google Cloud, H2O, DataRobot, gamalon, ELEMENT AI, VISENZE, deepsense.io, bonsai	
COMPUTER VISION	HORIZONTAL AI	SPEECH & NLP	
Microsoft Azure, Amazon Rekognition, clarifai, Cloud Vision API, EVER AI, deepomatic, twentybn, neurals	IBM Watson, Cortana, Face++, sentient, Voyager, Affectiva, Numenta, PETUUM, nqalogics, CURIOUS AI, OSARO, BLUE VISION	Google Cloud, twilio, amazon alexa, narrative science, semantic machines, Mobvoi, SoundHound Inc., PRIMER, volocera, NIJUNCE, MindMeld, nuance, snips, yspop	
SEARCH	LOG ANALYTICS	SOCIAL ANALYTICS	WEB / MOBILE / COMMERCE ANALYTICS
ORACLE, elasticsearch, EXALEAD, COVEO, Lucidworks, ATTIVO, swiftype, algalia, alphasense, MAANA, omni:us, SINEQUA	splunk, sumologic, LOGGLY, TIMBER, kibana, Logz.io	Hootsuite, sprinklr, NETBASE, synthesio, simplereach, bitly, predata, SimilarWeb	Google Analytics, mixpanel, AMPUTUDE, sumall, Airtable, RESCI, SIGOPT, granify, custora

CSIE59830/CSIEM0410/AIIA50050 Big roduction 84

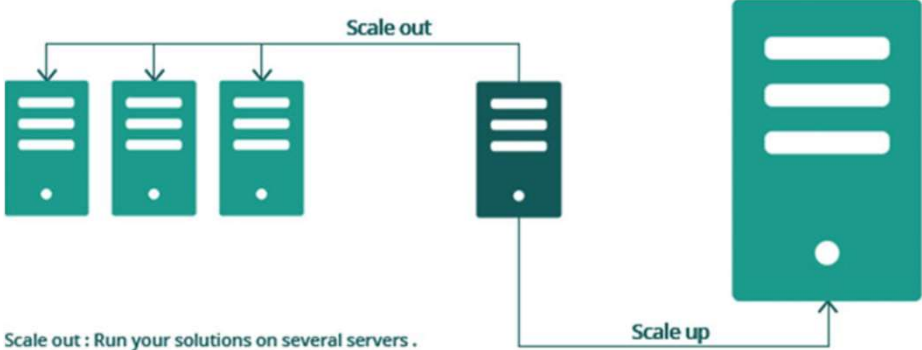
MAD(ML, AI, and Data) Landscape 2023

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

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Course Information 85



Scale Out vs Scale Up




- Big data is about scaling out instead of scaling up.

Scale out : Run your solutions on several servers .
Scale up : Run your solutions on bigger server.

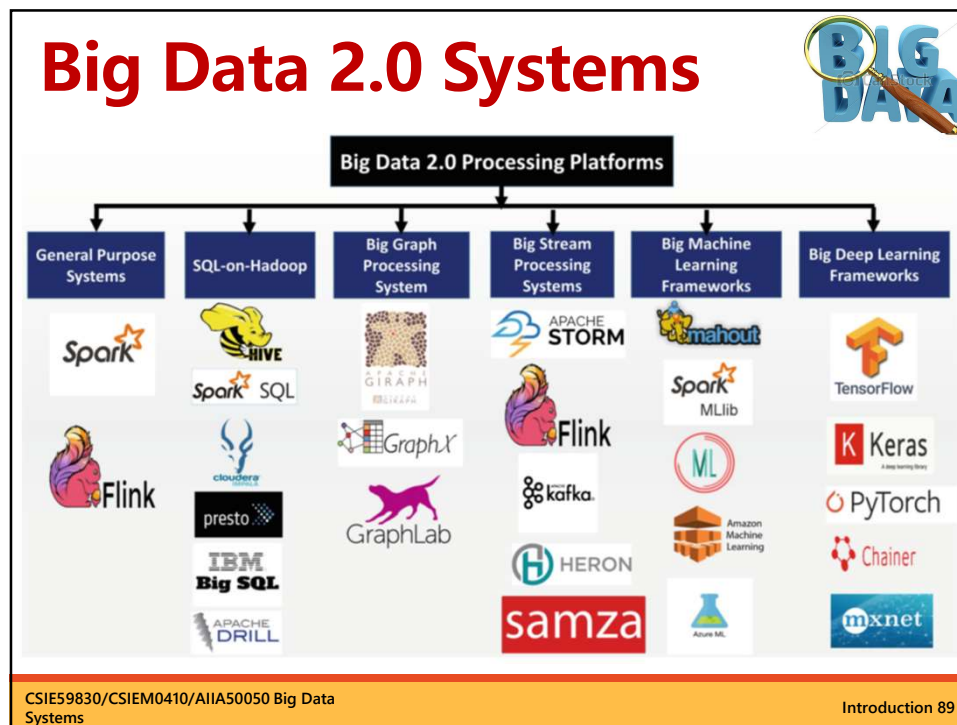
CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 87

Summary



- Big data era is **already** here!
- Calls for **advanced models** of storing, managing, processing, and analyzing data.
- You may come across articles claiming “Big data is dead”, “Big data era is coming to an end” ...
- The truth is that **everything** will be big data. It is becoming the **norm**.
- Some authors call it “**Big Data 2.0**”!!

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 88



Coming Lectures

- Big data processing architecture
 - Hadoop
- General purpose big data processing system
 - MapReduce
 - Spark
 - HPCC
- Data mining algorithms based on MapReduce and Spark

CSIE59830/CSIEM0410/AIIA50050 Big Data Systems Introduction 90

Coming Lectures



- Storage systems for big data processing
 - Google File System
 - Hadoop Distributed File System
 - Google Cloud Storage/Datastore/BigTable
- NoSQL/NewSQL database systems
 - Hbase, Cassandra
 - MongoDB
 - VoltDB
- Data warehousing systems
 - Google BigQuery
 - Apache Hive
 - Spark SQL

Coming Lectures



- Systems for big graph processing
 - Google Pregel, BSP, Giraph
 - Neo4j
 - Apache TinkerPop
- Systems for stream processing
 - Spark Streaming, Structured Streaming
 - Apache Storm, Samza, Flink
 - Apache SAMOA (distributed streaming ML framework)
- ETL and API integration tools
 - Apache Kafka
 - Apache Camel
 - Apache Airflow

Coming Lectures



- Big data analytics**
 - Google Dremel, Apache Drill and Apache Impala
 - Google Cloud Platform vs Amazon Web Services
 - Beyond Hadoop
- Big machine learning framework**
 - Apache Mahout
 - Spark MLlib
- Big data analytics and ML are covered in different classes.

Assignment 0



- Select the **virtual machine (VM) software** (VirtualBox, VMWare, Cloudera, ...) to build your virtual Hadoop/Spark cluster.
- Install the VM software and construct several virtual host machines (eg. Ubuntu) for the cluster.
- Familiarize yourself with the Linux VMs to prepare for subsequent assignments.
- No need to turn in anything.