COMM1822

Term 2 2022

UNSW SYDNEY

Introduction to Databases for Business Analytics

Week 8 Big Data 1

Lecturer-in-Charge: Kam-Fung (Henry) Cheung
Email: kf.cheung@unsw.edu.au
Tutors: Theresa Tran
Liam Li Chen
Kathy Xu
PASS Leader: Srilekha Chandrashekara Kolaki

WARNING

This material has been reproduced and communicated to you by or on behalf of the University of New South Wales in accordance with section 113P(1) of the Copyright Act 1968 (Act). The material in this communication may be subject to copyright under the Act. Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

Do not remove this notice

Copyright

- There are some file-sharing websites that specialise in buying and selling academic work to and from university students.

- If you upload your original work to these websites, and if another student downloads and presents it as their own either wholly or partially, **you might be found guilty of collusion — even years after graduation**.

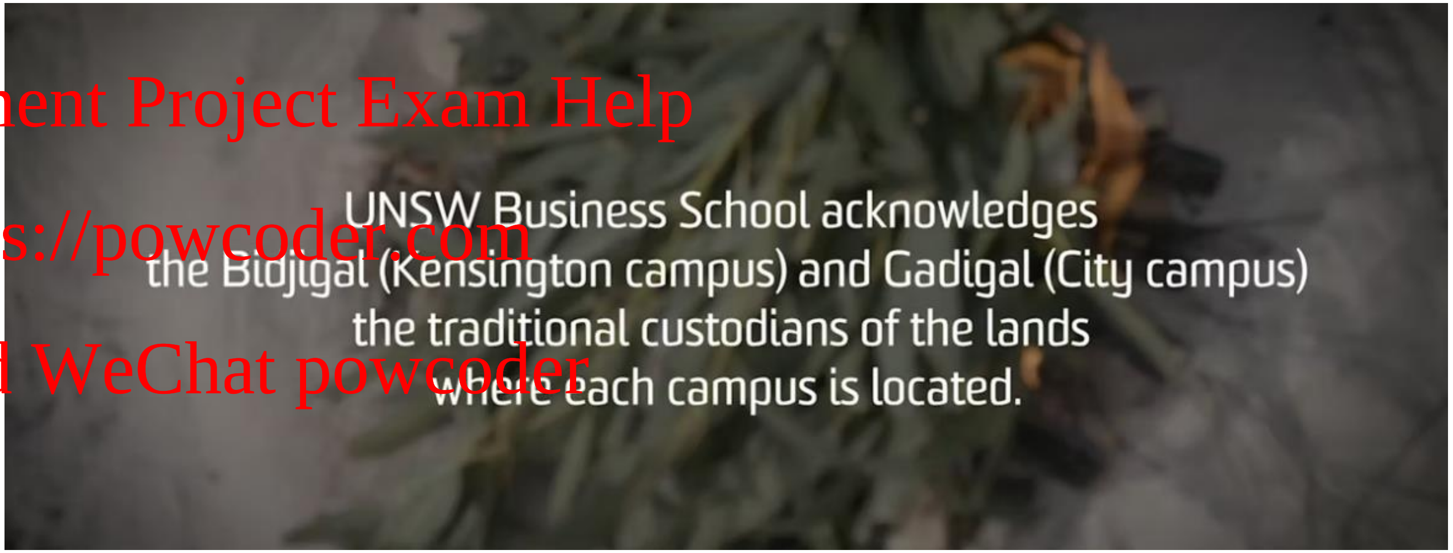
- These file-sharing websites may also accept purchase of course materials, **such as copies of <u>lecture slides</u> and <u>tutorial handouts</u>. By law, the copyright on course materials, developed by UNSW staff in the course of their employment, belongs to UNSW. It constitutes copyright infringement, if not academic misconduct, to trade these materials.**

Acknowledgement of Country

UNSW Business School acknowledges the Bidjigal (Kensington campus) and Gadigal (City campus) the traditional custodians of the lands where each campus is located.

We acknowledge all Aboriginal and Torres Strait Islander Elders, past and present and their communities who have shared and practiced their teachings over thousands of years including business practices.

We recognise Aboriginal and Torres Strait Islander people’s ongoing leadership and contributions, including to business, education and industry.



UNSW Business School acknowledges the Bidjigal (Kensington campus) and Gadigal (City campus) the traditional custodians of the lands where each campus is located.

UNSW Business School. (2022, May 7). *Acknowledgement of Country* [online video]. Retrieved from https://vimeo.com/369229957/d995d8087f

UNSW
SYDNEY

At UNSW you are free to...

- Respectfully disagree about anything
- Express different opinions
- Write your beliefs
- Show your beliefs
- Leave any club or organisation

It's not acceptable to...

- Attempt to censor opinions
- Use hate speech
- Make threats or instil fear
- Make false accusations
- Access or share others private information without consent

We are here to help...

- Tell a teacher
- Tell UNSW Psychology and Wellness
- Report to UNSW Complaints
- Report to UNSW Security
- Report a crime to police

Find out more

W8 Learning Outcomes

What is Big Data?

❑ Buzz Word!
❑ Cannot fit into a USB flash drive
❑ A large and complex dataset
❑ Social media
❑ IoT streaming of data
❑ Capturing of Media

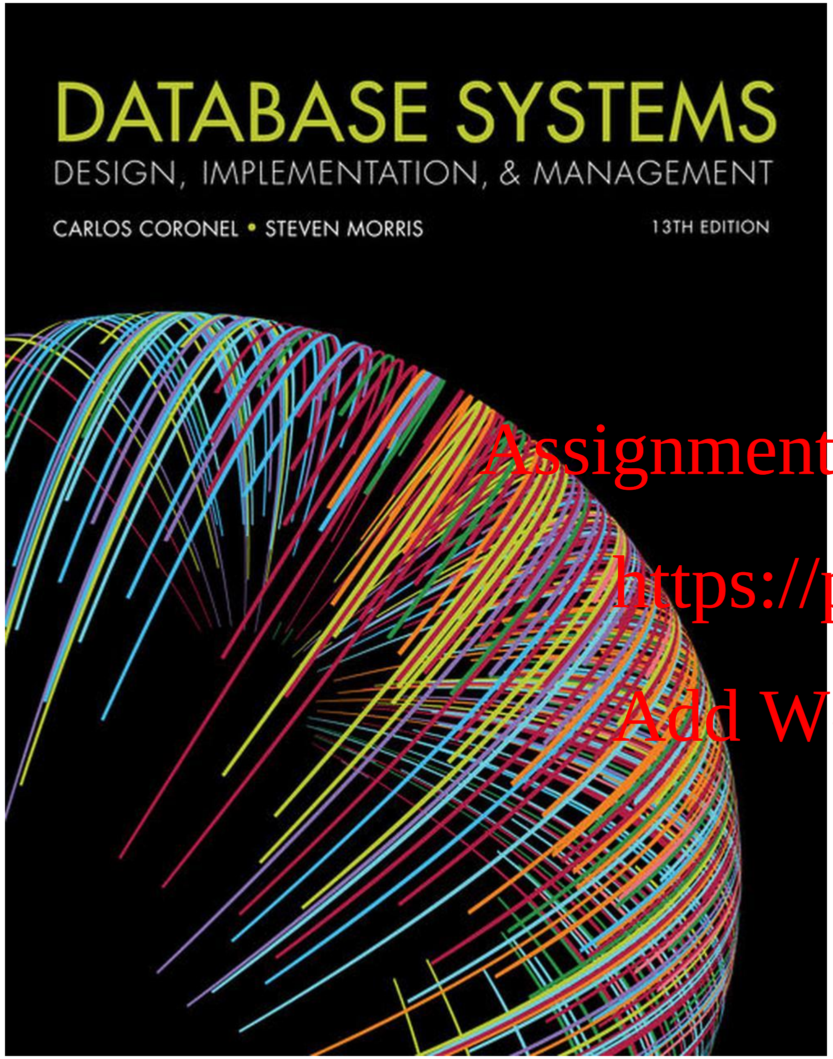
3Vs and more Vs

Big Data is classified into three types:

❑ Structured
❑ Unstructured
❑ Semi-Structured

Big Data

❑ Hadoop
❑ NoSQL


DATABASE SYSTEMS
DESIGN, IMPLEMENTATION, & MANAGEMENT
CARLOS CORONEL • STEVEN MORRIS
13TH EDITION

Chapter 14

Big Data and NoSQL

14-1 to 14-3

The Next Big Thing?

BIG

DATA

Big Data

- Refers to set of **data analysis and predictive analysis techniques for large and complex sets of raw data** (difficult or impossible to capture in ER models).
- Uses **machine learning** and **data mining** techniques on raw data (instead of organizing data upfront into neat structures) to make sense of the data.
- **Relational model: structure/schema on write**
- **Big Data model: structure/schema on read**

- Big data emerges because:
  - **much larger set of data sources** (e.g., Internet search/browsing, mobile devices)
  - **much cheaper costs to store data** (e.g., costs of hard disc drives reduced substantially)
  - **growing interest in identifying patterns** for business purposes (in all kinds of data)
  - **scaling out** instead of scaling up

Big Data

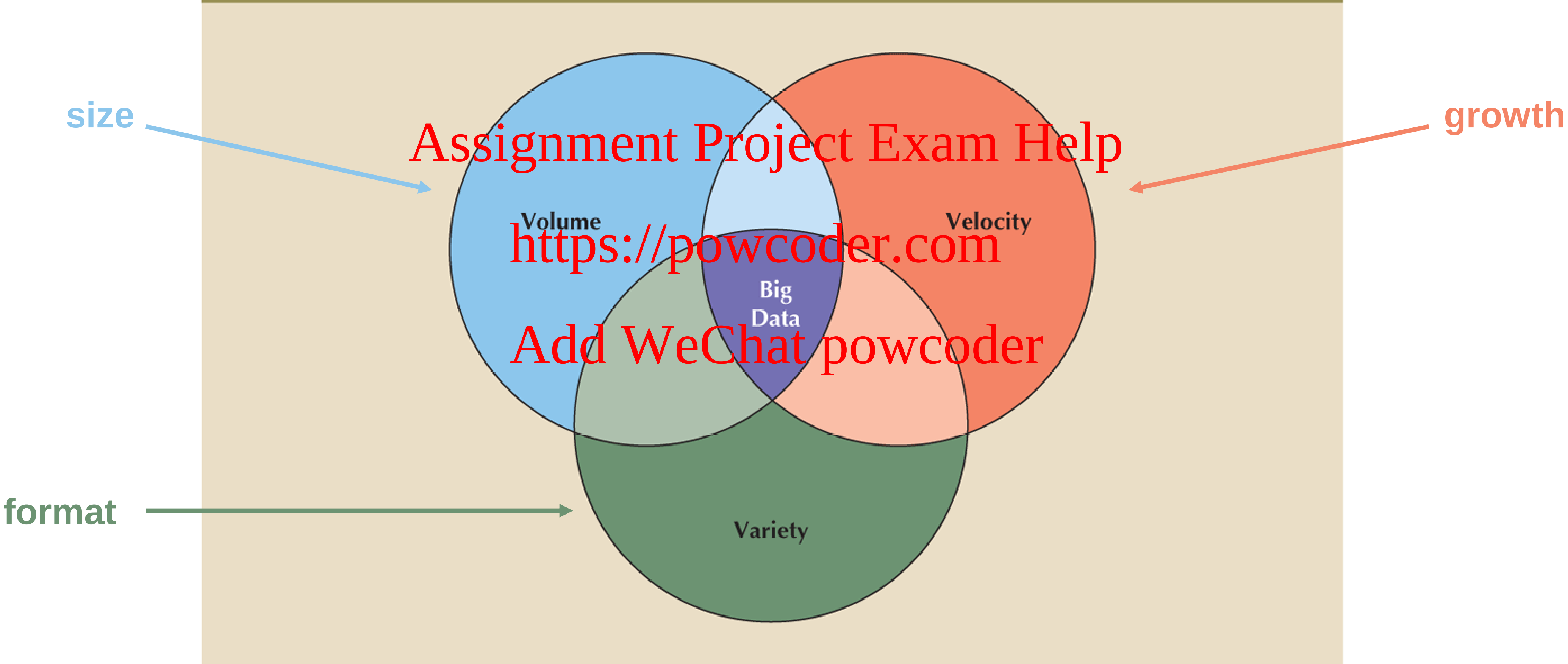
- Name: 7920 Disc Drive
- Product Number: 7920
- Introduced: 1977
- Division: Disc Memory
- Original Price: **$17000**
- Catalog Reference: 1979, page 641

http://hpmuseum.net/display_item.php?hw=272

3Vs and … more Vs

A Few Years Ago …

FIGURE 14.1 ORIGINAL VIEW OF BIG DATA



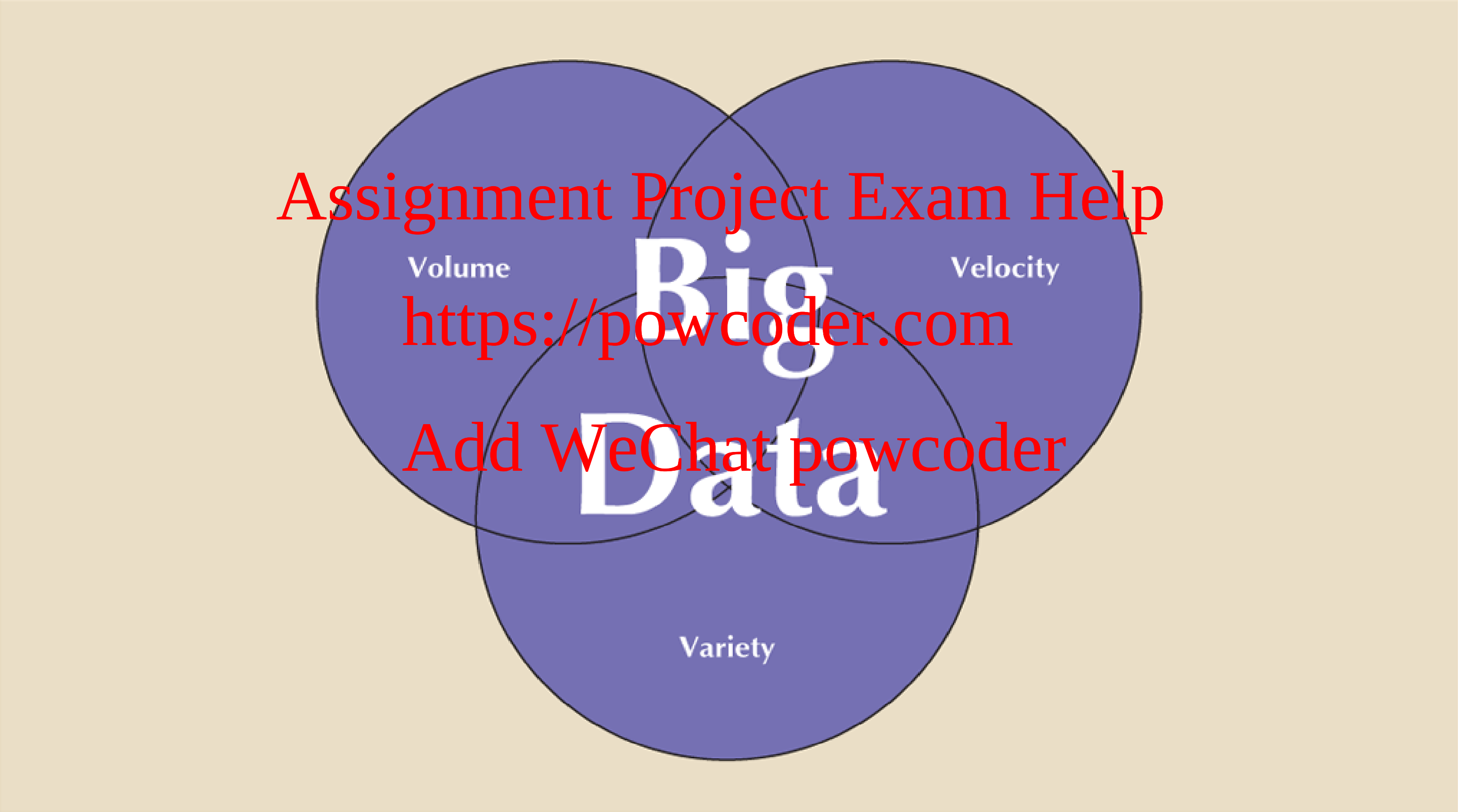
size

growth

format

Today

FIGURE 14.2 CURRENT VIEW OF BIG DATA



Big Data

1. **Volume:** Quantity of data to be stored storage issue
   - ❑ **Scaling up** is keeping the same number of systems but migrating each one to a larger system. e.g., 100 GB to 100 TB
   - ❑ **Scaling out** means when the workload exceeds server capacity, it is spread out across a number of servers.

2. **Velocity:** Speed at which data is entered into system and must be processed storage issue; data need to be processed rapidly
   - ❑ **Stream processing** focuses on input processing and requires analysis of data stream as it enters the system.
   - ❑ **Feedback loop processing** refers to the analysis of data to produce actionable results. (Details will be shown later.)

Big Data

3. **Variety:** Variations in the structure of data to be stored
   - ❑ **Structured data** fits into a predefined data model — relational DB
   - ❑ **Unstructured data** does not fit into a predefined data model

e.g., maps, images, emails, texts, tweets, videos, …

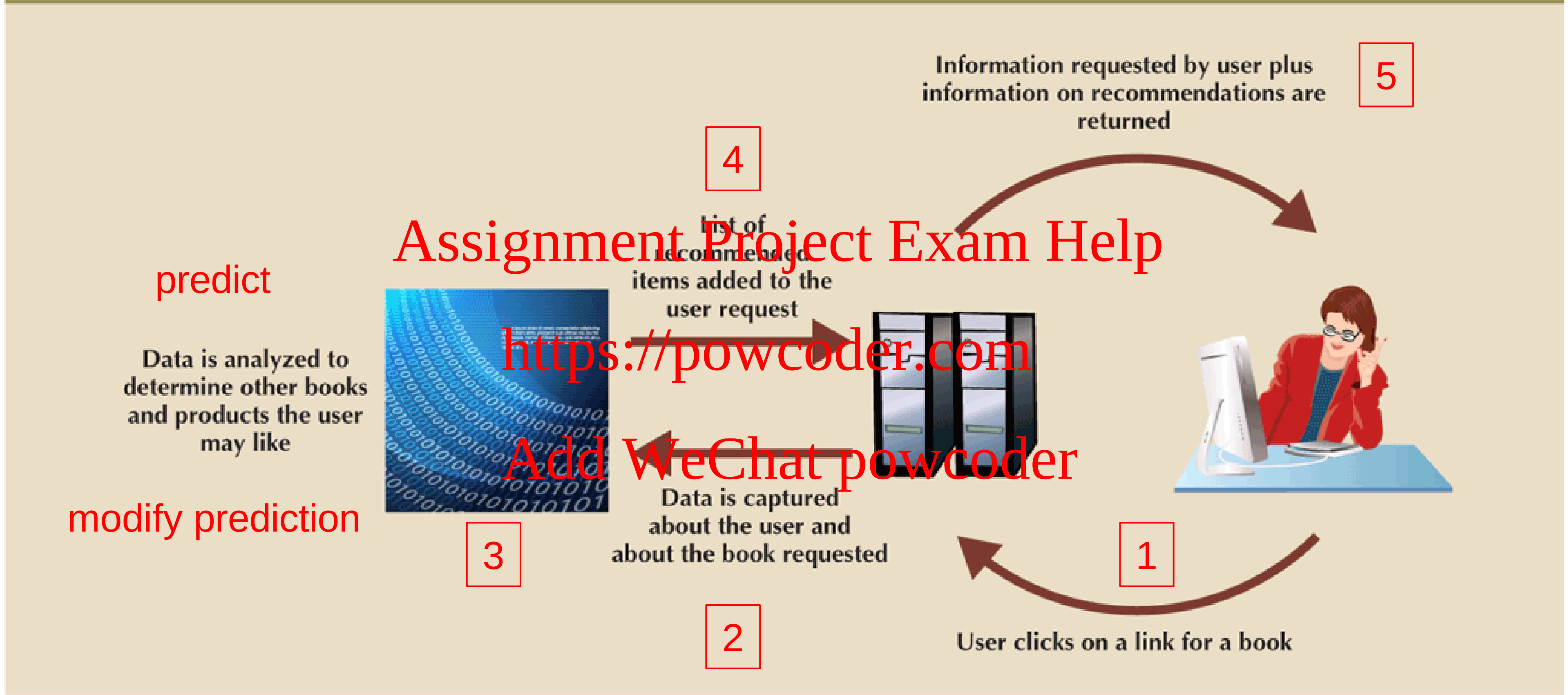
❖ **Other Characteristics**

- ❖ **Variability:** Changes in meaning of data based on context
  - ❖ **Sentiment analysis** attempts to determine attitude

Sarcasm (does ‘good’ really mean good?

- ❖ **Veracity:** Trustworthiness of data — accuracy
- ❖ **Value:** Degree of data can be analyzed for meaningful insight
- ❖ **Visualization:** Ability to graphically present data to make it understandable to users

FIGURE 14.3 FEEDBACK LOOP PROCESSING


Information requested by user plus information on recommendations are returned
5
4
List of recommended items added to the user request
predict
Data is analyzed to determine other books and products the user may like
modify prediction
3
Data is captured about the user and about the book requested
2
1
User clicks on a link for a book

Big Data Classification

Structured

Unstructured

Semi-Structured

Structured Data

Any data types that clearly defined be stored, accessed and processed in a fixed format can be defined a ***structured* data**.

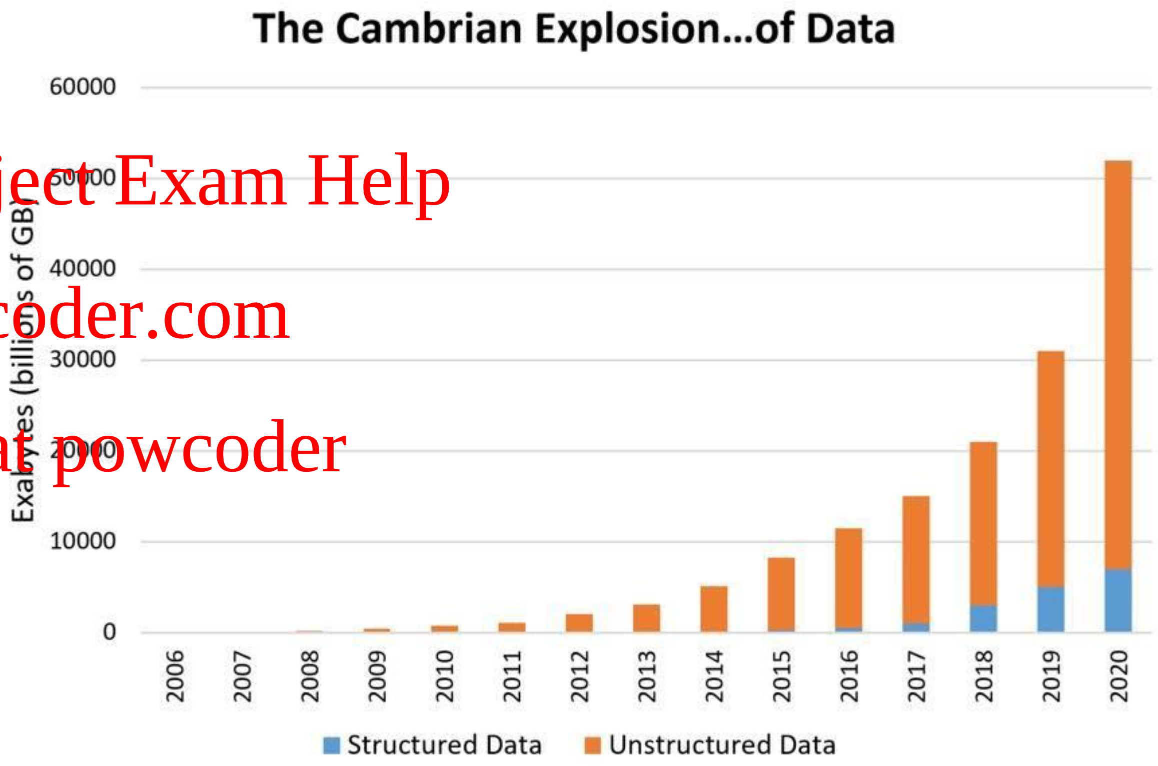
A good example is data stored in a table in a normalized database. You can easily search and retrieve the data from a table using SQL tools. For instance, in the Sales_Person table, we can find the Year of Hire for Sales_Person No. 101 is 1995, Cookie Biscuit.

Sales_Person_Num	**Sales_Person_Name**	**Year_of_Hire**	**Department_Num**
101	Cookie Biscuit	1995	10
102	Sweet Candy	1998	20
103	Chocolate Milk	2002	20

Unstructured Data

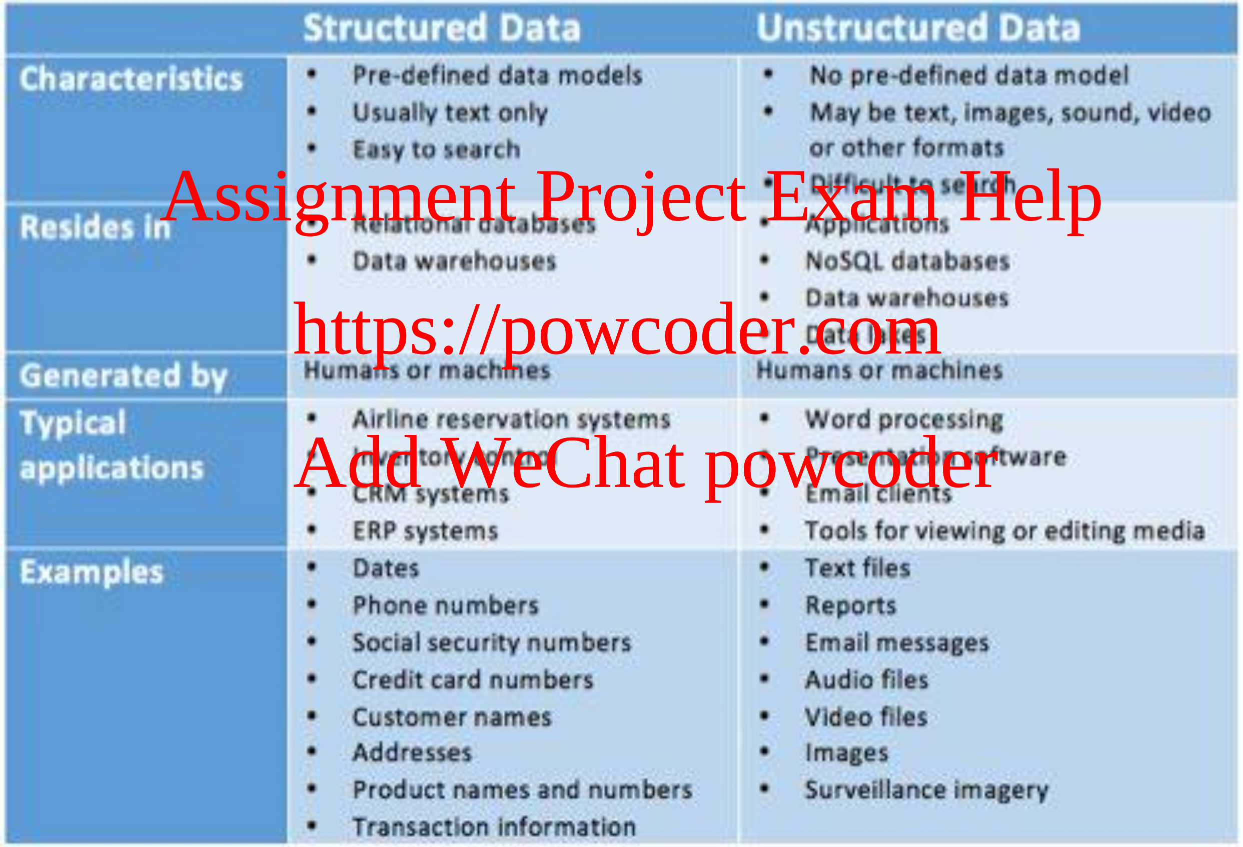
- ❑ ***Unstructured* data** can simply be described as not *structured data*; that is, anything that cannot be described as *structured data*.

- ❑ Examples of *unstructured* data include free text, videos, images, etc. The ability to analyze social media such as Facebook, Twitter, and WeChat, and images are among the key drives behind the growth of Big Data.



https://www.cprime.com/resources/blog/when-big-data-big/

Differences between Structured Data and Unstructured Data



	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

https://www.datamation.com/big-data/structured-vs-unstructured-data.html

Semi-Structured

- ***Semi-Structured data*** is crossed between Structured Data and Unstructured Data, i.e., it has both forms of data. Examples include Electronic Data Interchange (EDI), Markup Language XML, and Open Standard JSON (JavaScript Object Notation).
- For example, as shown below, XML document is organized in a hierarchy with “open” and “close” tags and encoded rules that defines a human- and machine-readable format.

```xml
<?xml version="1.0" standalone="yes"?>
<conversation>
  <greeting>Hello World! </greeting>
  <response>Hello, who are you? </response>
</conversation>
```

The Human Face of Big Data

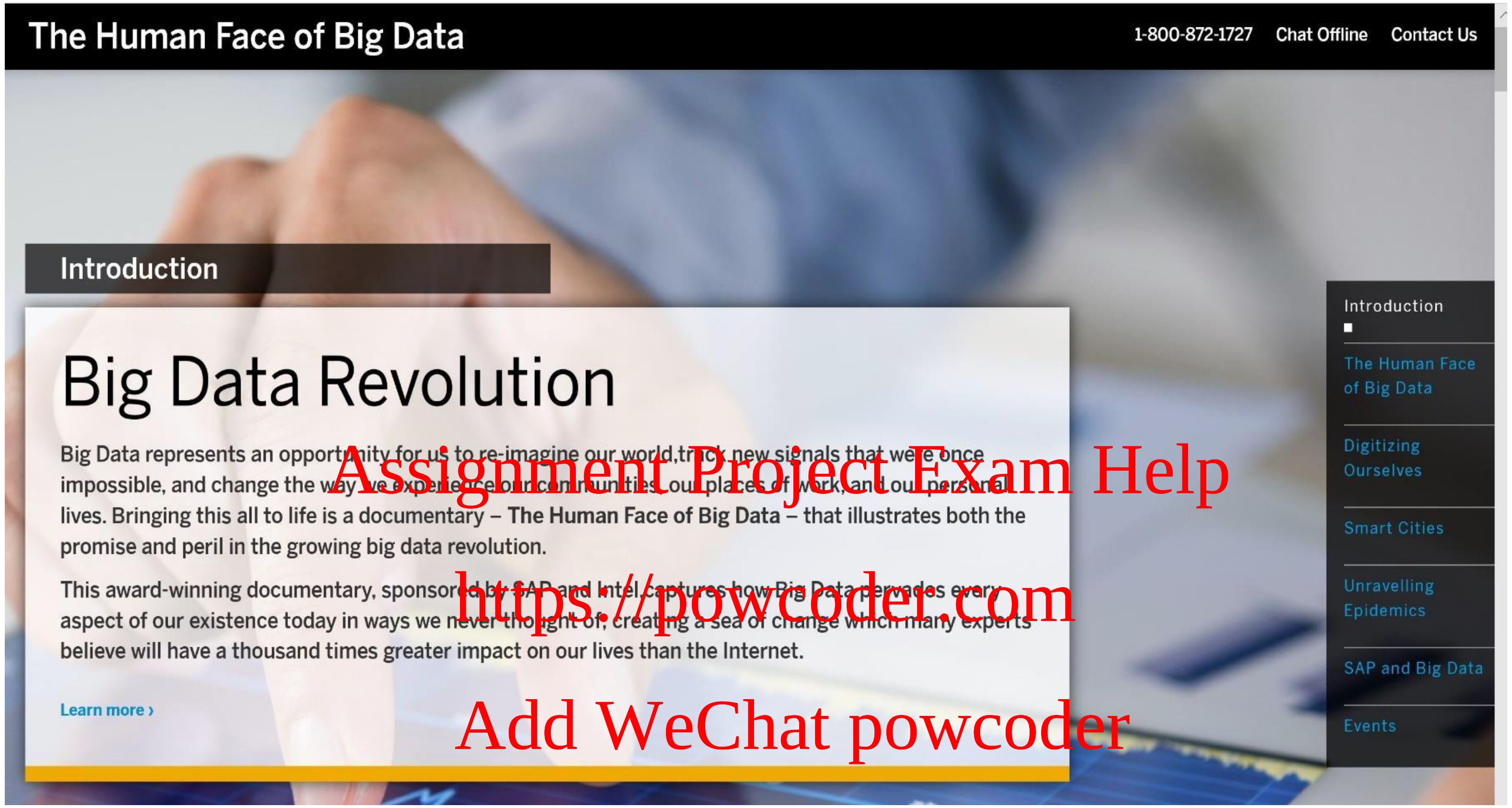
UNSW SYDNEY

The Human Face of Big Data

The impact of Big Data could be described the next major revolution since the Agricultural Revolution and Industrial Revolution. We can call it Digital Revolution or Big Data Revolution. Today, we have already seen large corporations, particularly the large Chinese companies, use Big Data, Artificial Intelligence, and Machine Learning extensively to drive their business strategies to gain competitiveness.

This award-winning documentary was created to explain how Big Data has evolved the way we work, shop, socialize, live, and benefit from Big Data as well as the rise of negative issues associated with Big Data. Big Data is collected, stored, and used across a wide range of products and services.

You will learn how Big Data can be used in various areas, and how Big Data influences.



Big Data Revolution

Big Data represents an opportunity for us to re-imagine our world, track new signals that were once impossible, and change the way we experience our communities, our places of work, and our personal lives. Bringing this all to life is a documentary – The Human Face of Big Data – that illustrates both the promise and peril in the growing big data revolution.

This award-winning documentary, sponsored by SAP and Intel, captures how Big Data pervades every aspect of our existence today in ways we never thought of, creating a sea of change which many experts believe will have a thousand times greater impact on our lives than the Internet.

The Human Face of Big Data https://www.youtube.com/watch?v=bIY3LUZ7i8Y

Warning: The music in the video is a bit loud in some sections, so you might want to test and control the volume.

Topic: Digitising Ourselves (17:36 to 23:55 of the video on previous slide)

❑ Collecting data about oneself!

❑ Pattern recognition algorithm - change the way as a society

- Personal devices, such as Apple Watch, Samsung Watch, and Fitbit, contain apps and sensors used to collect data about your health (as an example).
- If you have such personal devices, the question here is: can these devices influence on how you behave? Examples can include do you pay attention to the output (such as graph or numbers) from these apps, or do you have a goal of burning number of calories per day?

Topic: Building a Global Brain (23:55 to 25:55)
Topic: Creating Intelligence System (25:55 to 28:50)

Data is collected from you via devices. You react based on the data presented to you, and the action you have taken becomes another data point in this Big Data system. This becomes a cycle where the Big Data has an impact on you, and then your action becomes a data point in the Big Data.

In the video, it discusses about scheduling of buses. One of the suggestions is to be more proactive based on the needs of bus, i.e., instead of ten buses regularly travelling on one route. The bus can be diverted to another route if the demand for this particular route is reduced but a higher demand for the other route. Some would call this as building a smart city from Big Data. Thus, the city like Boston could be functioned more efficiently based on the data, i.e., “responsive to our needs”.

Topic: Targeting You (38:23 to 41:05) [1]

Target has used Big Data to identify pregnant women as part of their marketing strategy to target that segment of the consumers, provide better customer services, and improve their revenue. This practice is common among the retailers, hotel industry, airline industry and gambling industry, which offer loyalty programs to their customers as a way of rewarding them for being their customers.

The original intention of offering loyalty program is to build a customer relationship. However, in the case of Target, they use the customer information further with Big Data to create a profile of their customers who purchase products related to pregnancy and baby.

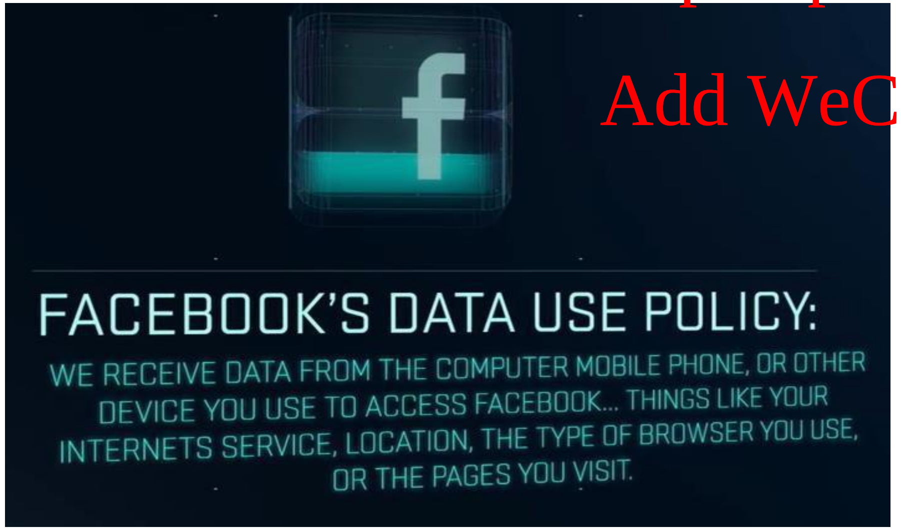
Topic: Targeting You (38:23 to 41:05)[2]

Another example is nearly all the search engines, such as Google, generate their revenue by producing advertisements based on what your searches.

Companies want to advertise their products on the Internet, and these search engine companies offer their services to the customers who search terms or phrases which meet the advertising criteria.

Topic: The Dark Side (41:06 to 45:59)

One of the criticisms on Facebook is they have been collecting data without fully reveal their intention, and how they would use your data once they collected. They can build a profile of you as an individual.



FACEBOOK'S DATA USE POLICY:

WE RECEIVE DATA FROM THE COMPUTER MOBILE PHONE, OR OTHER DEVICE YOU USE TO ACCESS FACEBOOK... THINGS LIKE YOUR INTERNETS SERVICE, LOCATION, THE TYPE OF BROWSER YOU USE, OR THE PAGES YOU VISIT.

Moreover, National Security Agency (NSA) has been collecting data for a number of years.

Big Data

Hadoop

NoSQL

Hadoop

- De facto standard for most Big Data storage and processing

- Java-based framework for distributing and processing very large data sets across clusters of computers
  1. Hadoop Distributed File System (HDFS): low-level distributed file processing system that can be used directly for data storage
  2. MapReduce: programming model that supports processing large data sets

Hadoop Distributed File System (HDFS)

Based on several key assumptions

- **High volume:** default block sizes is 64 MB and can be configured to even larger values
- **Write-once, read-many:** model simplifies concurrency issues and improves data throughput
- **Streaming access:** optimized for batch processing of entire files as a continuous stream of data
- **Fault tolerance:** designed to replicate data across many different devices so that when one fails, data is still available from another device

Why we need HDFS?

HDFS enables us to

- deal with very large datasets
- solve big data problems in a distributed manner
- use cheap hardware rather than expensive servers
- have a stable data storage which is fault tolerant
- store data in different platforms
- mange data using a set of Unix-style file system commands

Acknowledgement: The slide is provided by Xin Cao, UNSW CSE, ZZEN9313

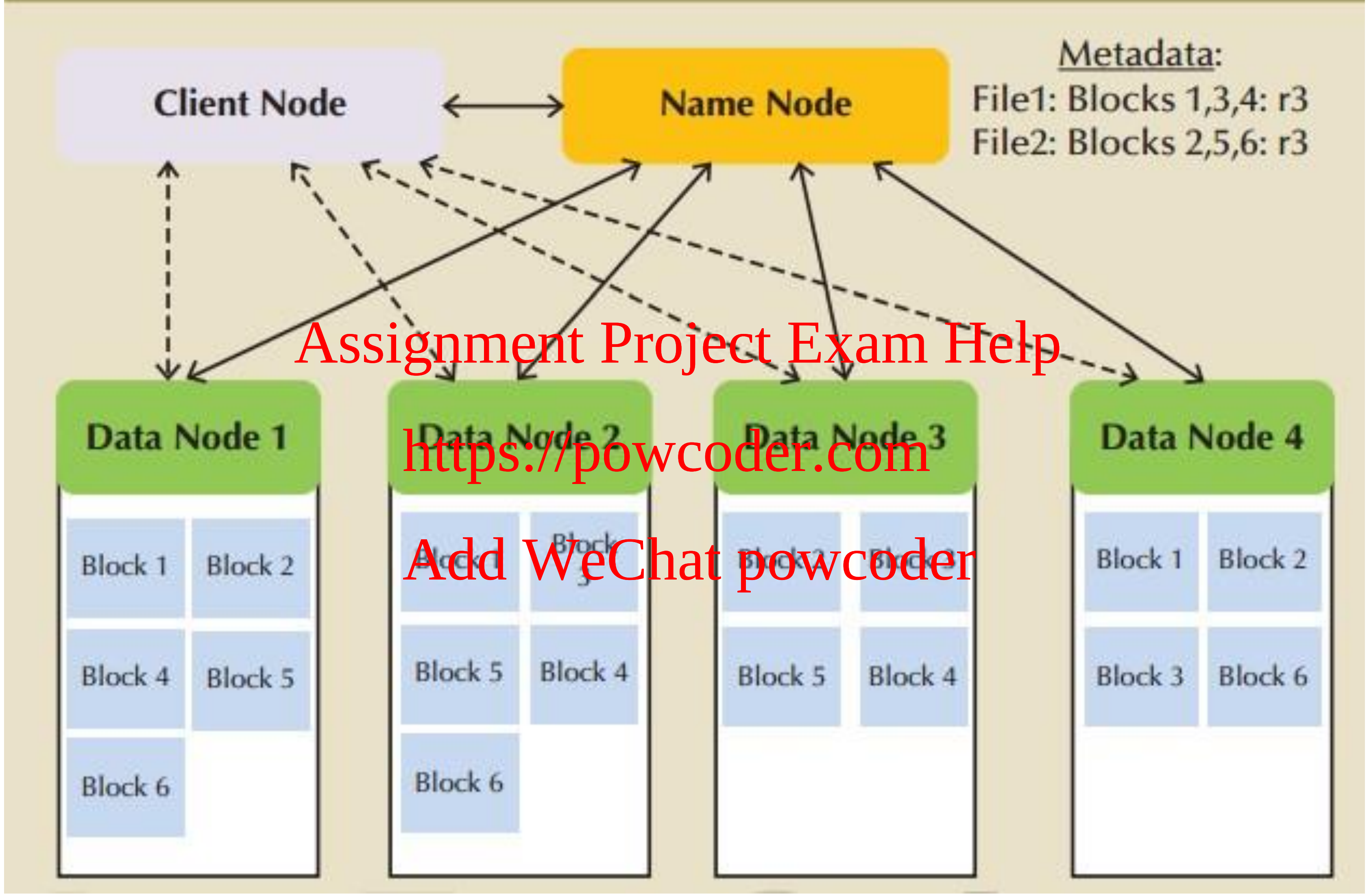
Nodes

More in next week

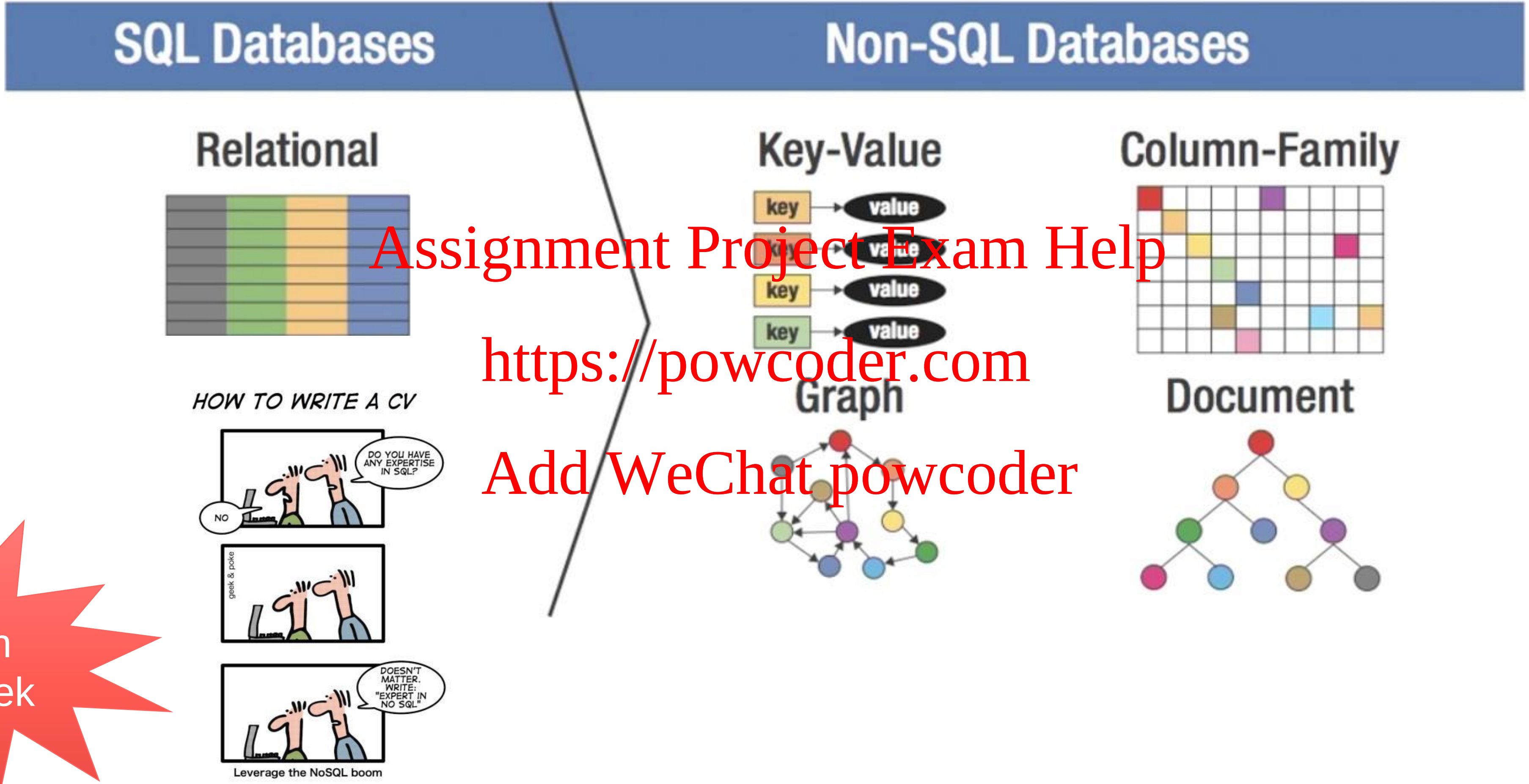
Hadoop uses several types of nodes:

- A **node** is just a computer that perform one or more types of tasks within the system
- **Data node** stores the actual file data
- **Name node** contains file system metadata
- **Client node** makes requests to the file system as needed to support user applications
- **Data node** communicates with **name node** and send back block reports and heartbeats

FIGURE 14.4 HADOOP DISTRIBUTED FILE SYSTEM (HDFS)



NoSQL



More in next week

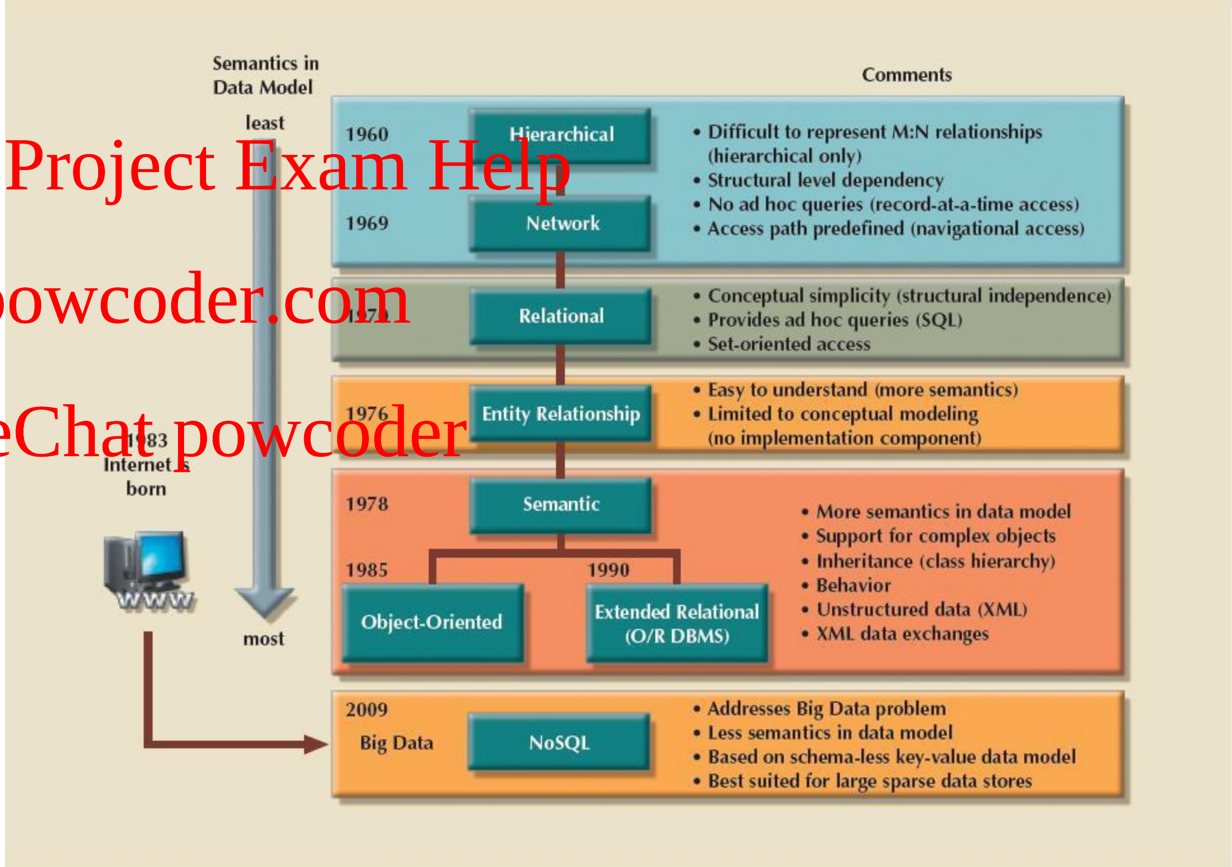
Source: Poulson/lynda.com

Discussion: Data Management Models

1. **File systems** models
2. **Relational** models
3. **Object-oriented** models
4. **Big Data** models

Polyglot persistence: The coexistence of a variety of data storage and data management technologies within an organization’s infrastructure.

FIGURE 2.6 THE EVOLUTION OF DATA MODELS



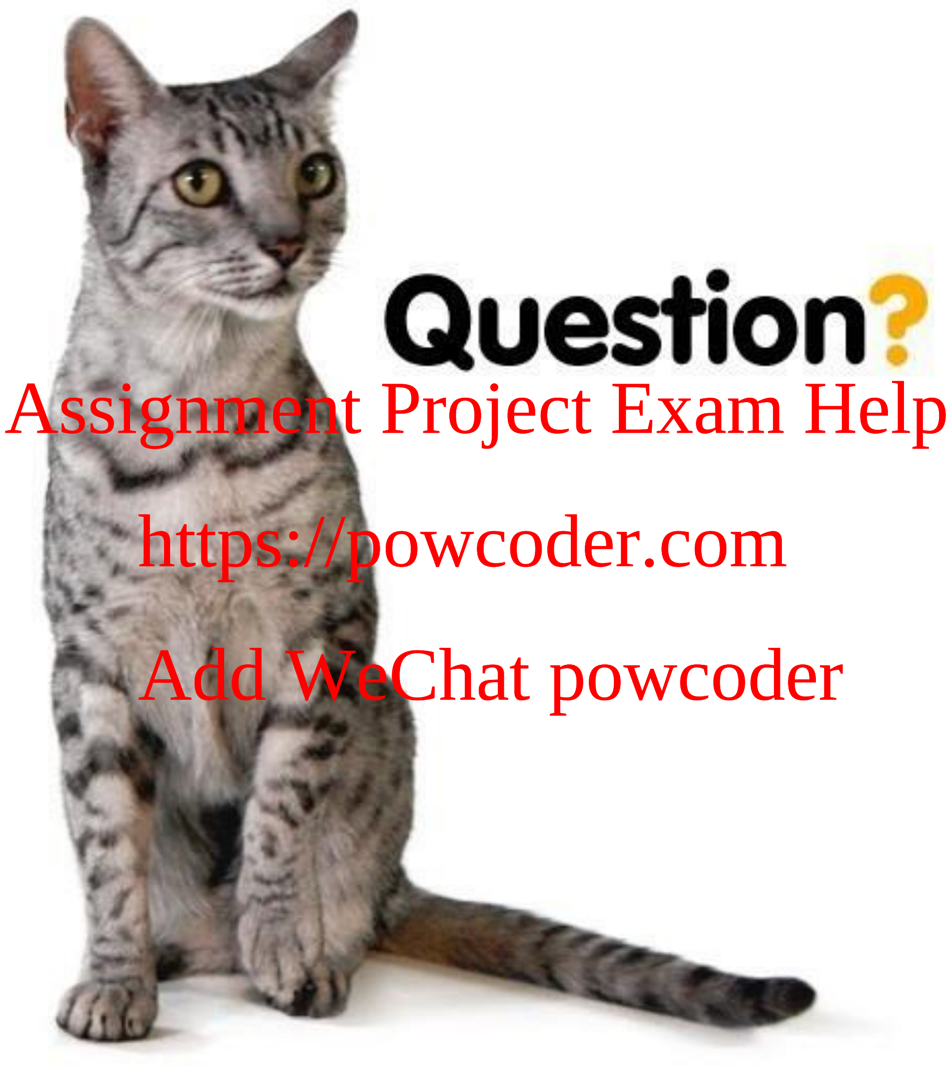
Hierarchical / Network:
- Difficult to represent M:N relationships (hierarchical only)
- Structural level dependency
- No ad hoc queries (record-at-a-time access)
- Access path predefined (navigational access)

Relational:
- Conceptual simplicity (structural independence)
- Provides ad hoc queries (SQL)
- Set-oriented access

Entity Relationship:
- Easy to understand (more semantics)
- Limited to conceptual modeling (no implementation component)

Semantic / Object-Oriented / Extended Relational:
- More semantics in data model
- Support for complex objects
- Inheritance (class hierarchy)
- Behavior
- Unstructured data (XML)
- XML data exchanges

NoSQL:
- Addresses Big Data problem
- Less semantics in data model
- Based on schema-less key-value data model
- Best suited for large sparse data stores



Source: patrickmahaney.com