

# COMM1822

Term 2 2022

## Introduction to Databases for Business Analytics

Assignment Project Exam Help

Week 8 Big Data 1

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

Lecturer-in-Charge: Kam-Fung (Henry) Cheung

Email: [kf.cheung@unsw.edu.au](mailto:kf.cheung@unsw.edu.au)

Tutors: Theresa Tran

Liam Li Chen

Kathy Xu

PASS Leader: Srilekha Chandrashekara Kolaki



## Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Copyright

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

**Assignment Project Exam Help**

- If you upload your original work and presents it as their own either on your own or through other student downloads **https://eduassistpro.github.io/** **you might be found guilty of collusion — even years after graduation**

**Add WeChat edu\_assist\_pro**

- These file-sharing websites may also accept purchase of course materials, **such as copies of lecture slides and tutorial handouts**. 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.

Assignment Project Exam Help

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.

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

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

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

## Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# 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

## Big Data

- ☐ Hadoop
- ☐ NoSQL

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

## 3Vs and more Vs

## Big Data is classified into three types:

- ☐ Structured
- ☐ Unstructured
- ☐ Semi-Structured

Chapter 14  
Assignment Project Exam Help  
and NoSQL  
<https://eduassistpro.github.io/>  
4-3  
Add WeChat edu\_assist\_pro

# The Next Big Thing?

**BI**

Assignment Project Exam Help

<https://eduassistpro.github.io/>

**DA**

Add WeChat edu\_assist\_pro





# 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 i ke sense of the data).
- ❑ **Relational model: structure/schema on** <https://eduassistpro.github.io/>
- ❑ **Big Data model: structure/schema on** [Add WeChat edu\\_assist\\_pro](#)
- ❑ 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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

[http://hpmuseum.net/display\\_item.php?hw=272](http://hpmuseum.net/display_item.php?hw=272)

# 3Vs and ... more Vs

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro



**UNSW**  
SYDNEY

# A Few Years Ago ...

size → Assignment Project Exam Help ← growth

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

format →

# Today

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# 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. 100 GB to 100 TB
- ☐ **Scaling out** means increases server capacity, it is spread out across a <https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

## 2. **Velocity:** Speed at which data is added 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

Assignment Project Exam Help

images, emails, texts, tweets, videos, ...

<https://eduassistpro.github.io/>

## ❖ **Other Characteristics**

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

Add WeChat edu\_assist\_pro

Sarcasm (does 'good' really mean good?)

5

4

# Assignment Project Exam Help

predict

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

modify prediction

3

1

2



# Big Data Classification

Structured

Unstructure

Semi-Struct

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro



UNSW  
SYDNEY

# Structured Data

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

## Assignment Project Exam Help

A good example is data stored in a database. You can easily search and retrieve the data from a table. In the Sales\_Person table, we can find the Year of Hire for Cookie Biscuit.

Add WeChat edu\_assist\_pro

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*.

Assignment Project Exam Help

<https://eduassistpro.github.io/>

- ❑ 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 drivers behind the growth of Big Data.

Add WeChat edu\_assist\_pro

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

# Differences between Structured Data and Unstructured Data

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

<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 is organized in a hierarchy with “open” and “close” tags and encoded in a human- and machine-readable format.

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

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

# The Human Face of Big D

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro



# 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** extensively to drive their business strategies to gain **competitive advantage**.  
<https://eduassistpro.github.io/>

This award-winning documentary was created to show how Big Data has evolved the way we work, shop, socialize, live, and benefit from it, 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.

## Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

**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 smart watches, and Fitbit, contain apps and sensors used to track health (as an example).
  - If you have such personal devices, the question is: can these devices influence on how you behave? Examples: 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 system.

In the video, it discusses about how the city can be more proactive based on the needs of bus travellers. One of the suggestions is to have a fleet 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 retail, airline industry and gambling industry, which uses Big Data to target their customers as a way of rewarding them for being their customers.

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

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.

Assignment Project Exam Help

<https://eduassistpro.github.io/>

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

Add WeChat edu\_assist\_pro

# 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 you as an individual.

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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro



# Big Data

## Hadoop

## NoSQL

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro



# Hadoop

❑ De facto standard for most Big Data storage and processing

❑ Java-based framework for storing and processing very large data sets across multiple servers

1. **Hadoop Distributed File System (HDFS)**: a scalable distributed file processing system that can be used 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:** improves data throughput and reduces currency issues and
- ❑ **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 problem
- ☐ use cheap hardware
- ☐ have a stable data storage
- ☐ store data in different platforms
- ☐ manage data using a set of Unix-style file system commands

Assignment Project Exam Help

inner

<https://eduassistpro.github.io/>

servers

lerant

Add WeChat edu\_assist\_pro

# Nodes



Hadoop uses several types of nodes:

- ❑ A **node** is just a computer that perform one or more types of tasks within the sys
- ❑ **Data node** stores t
- ❑ **Name node** contains file syste
- ❑ **Client node** makes requests to
- ❑ **Data node** communicates with **name node** and send back block reports and heartbeats

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

stem as needed to

## Assignment Project Exam Help

<https://eduassistpro.github.io/>


Add WeChat edu\_assist\_pro

# NoSQL

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro



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

Assignment Project Exam Help

<https://eduassistpro.github.io/>

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

Add WeChat edu\_assist\_pro

# Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

Source: patrickmahaney.com