COSC2083

Introduction to IT

Lecture 4: Data and Storage

RMIT UNIVERSITY

Overview

- Introduction
- Data Storage
- Assessments
- Questions?

Short Exercise!

Communication + Topics / Interests in IT

- Form a pair with a person nearby
- Compare your answers to the question:

  “Do you backup your data?”

  “How do you backup your data?”

  “What are some older storage technologies?”

You have a few minutes

Session 1

Hardware

RMIT UNIVERSITY

Data and Storage

Three interrelated aspects:

- Technology
- Organisation and Encoding
- Scale

RAM - Random Access Memory

Random access means any cell can be accessed at any time (and in any order!)

- **Volatile** – contents cleared when machine is switched off
- Very fast compared to other forms of memory
- **DRAM**: dynamic RAM (replenishes charges constantly)
- **SDRAM**: synchronous DRAM – faster still
- Often have small very fast **caches** and **registers.**

Magnetic Disk (1)

- Thin spinning metal disk with magnetic coating
- Each disk contains a number of circular tracks
- Often several disks stacked on top of each other
- Cylinders made up of tracks made up of sectors
- Can have very large storage this way
- Slow access time!

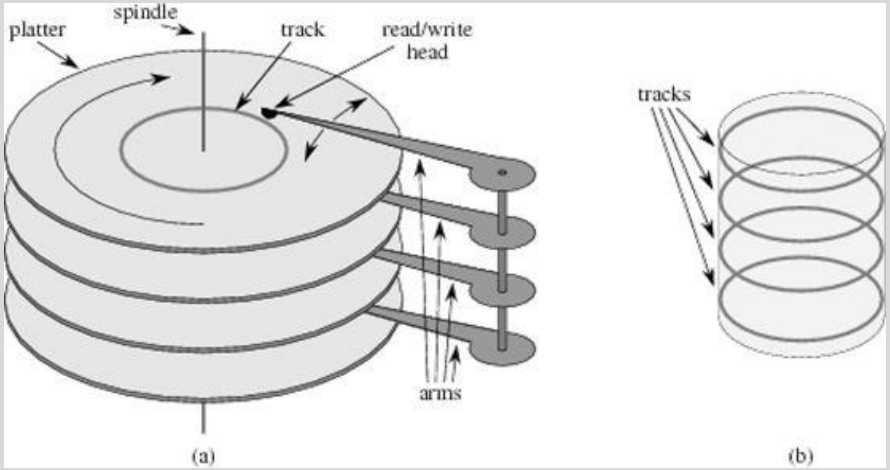
Magnetic Disk (2)

Hard Disk – HDD

- **Seek time:** move heads from one track to another
- **Latency time:** half time for complete disk rotation
- **Access time:** seek time + latency time
- **Transfer rate:** rate data can be read from disk

Magnetic Disk (3)

‘Typical’ Hard disk



- Seek time: 2ms to 15ms
- Latency time: 8ms to 20ms
- Transfer rate: 0.5 GB per second
- Sounds fast, but is actually quite slow …

Video Resource

How a Hard Disk Drive Works

How Do Hard Drives Work?

Solid State Drives (SSD) (1)

SSD (Solid State Drive)

- don’t have moving parts
- consume **less power**, less heat,
- generate **no vibrations,** less harm
- are based on flash technology

Solid State Drives (SSD) (2)

- Flash technology is semiconductor-based memory that preserves its information when powered off
- Has data access speeds much faster than mechanical disks
  - Microseconds vs. milliseconds

Solid State Drives (SSD) (3)

- Most storage vendors now offer storage systems using only SSDs
- Main disadvantage: price per gigabyte
- Higher than mechanical but dropping fast

Flash Drives

- Disks of all sorts are slow compared to other circuits
- Flash drives ‘write’ small electronic circuits
- Eventually decay after many changes of data
- Suitable for slow-changing data, not main memory
- Portable and much more resilient than disks

Optical Discs (CDs, DVDs, Blu-ray)

- Laser readers rather than magnetic ones
- Disks more error-tolerant than magnetic one

Type	Features	Date	Storage
CD	Compact Disc	1928	700MB
DVD	Multiple Layers	1995	4.7GB
Blu-ray	Blue Laser	2006	25GB

Note: There are other capacities - Single Sided, Double Sided, Multi-layer etc.

RAID

Redundant Array of Independent Disks

Provide a solution with:

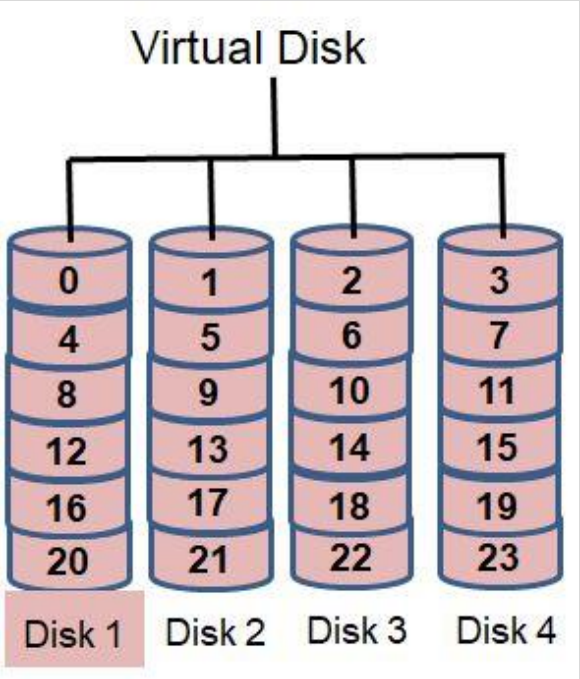
- High availability of data
- Improvements of performance

Uses multiple redundant disks

RAID 0 – Striping (2)

RAID 0 is also known as **striping**

- Provides an easy and cheap way to increase performance
- Uses multiple disks, each with a part of the data on it



RAID 0 – Striping (2)

- RAID 0 actually lowers availability

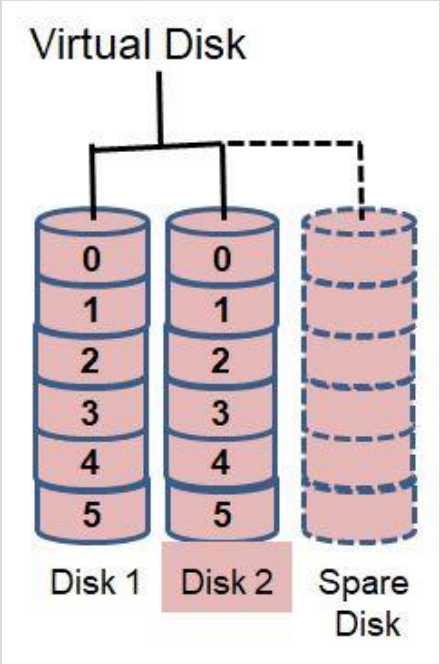
  If one of the disks in a RAID 0 set fails, all data is lost

- Only acceptable if losing all data on the RAID set is no problem (for instance for temporary data)

RAID 1 – Mirroring (1)

RAID 1 is also known as **mirroring**

- A high availability solution that uses two disks that contain the same data
- If one disk fails, data is not lost as it is still available on the mirror disk



RAID 1 – Mirroring (2)

- The most reliable RAID level
- High price
- 50% of the disks are used for redundancy only
- A spare physical disk can be configured to automatically take over the
- Task of a failed disk

Session 2

Encoding

RMIT UNIVERSITY

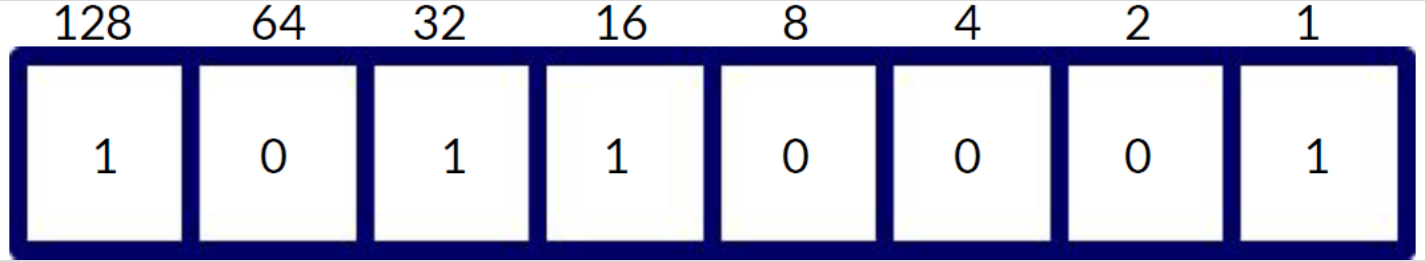
Binary Codes (1)

- Base 2
- Binary: 1 or 0, on off

Binary Codes (2)

Simple [base 10 / Decimal] numbers in binary:

- 0000 = 0
- 0001 = 1
- 0010 = 2
- 0011 = 3
- 0100 = 4
- 0101 = 5



ASCII

American **S**tandard **C**ode for **I**nformation **I**nterchange

- 7-bit patterns to represent
  - Letters (Uppercase and Lowercase)
  - Numbers
  - , . , ; “ $ % @ * & ! ? < > …
- Total of 128 different characters
  - 01001000 H
  - 01100101 e
  - 01101100 l
  - 01101100 l
  - 01101111 o
  - 00101110 .

Unicode (1)

- Computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems.

Unicode (2)

How?

- Uses 16 bits [not 7 - like ASCII]
- Can do Chinese, Japanese & Hebrew characters

So, Unicode can be used to represent 137,439 characters.

Quite a significantly larger count than ASCII...

Two’s Complement

Alright, but how do we store **negative integers**?

To get the Two's Complement negative notation of an integer:

- Write out the number in binary
- Invert the digits
- Add one to the result.

Bit Pattern	Value
011	3
010	2
001	1
000	0
111	-1
110	-2
101	-3
100	-4

Session 3

Big Data

RMIT UNIVERSITY

Big Data

Define Big Data

Diverse, high-volume, high-velocity information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.

Ted-Ed: Big Data - Tim Smith

Big Data - 3 Vs

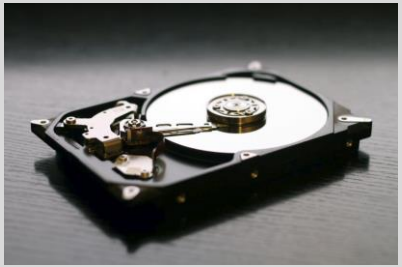
- **Volume** - its sheer size makes it a big data. Consider machine-generated data, which are generated in much larger quantities than non-traditional data.
- **Velocity** - The rate at which data flow into an organization is rapidly increasing. Velocity is critical because it increases the speed of the feedback loop between a company and its customers.
- **Variety** - Big Data formats can change very rapidly.

What are the three V's?

- **Volume**
  - When dealing with Big Data, we are talking about **petabytes** of data (FYI **trillion** of bytes)
  - It was estimated recently that Facebook stored about **20 petabytes** of 260 billion photos
  - That's 1,500 CDs or 220 DVDs

Volume

- The term “Big” is relative anyways
- Subject to various factors


Where do data come from?

- **Internal sources**
  - Sensors
  - Databases
- **External sources**
  - Social media
  - Customer's interactions (clickstream)
  - Suppliers/Partners

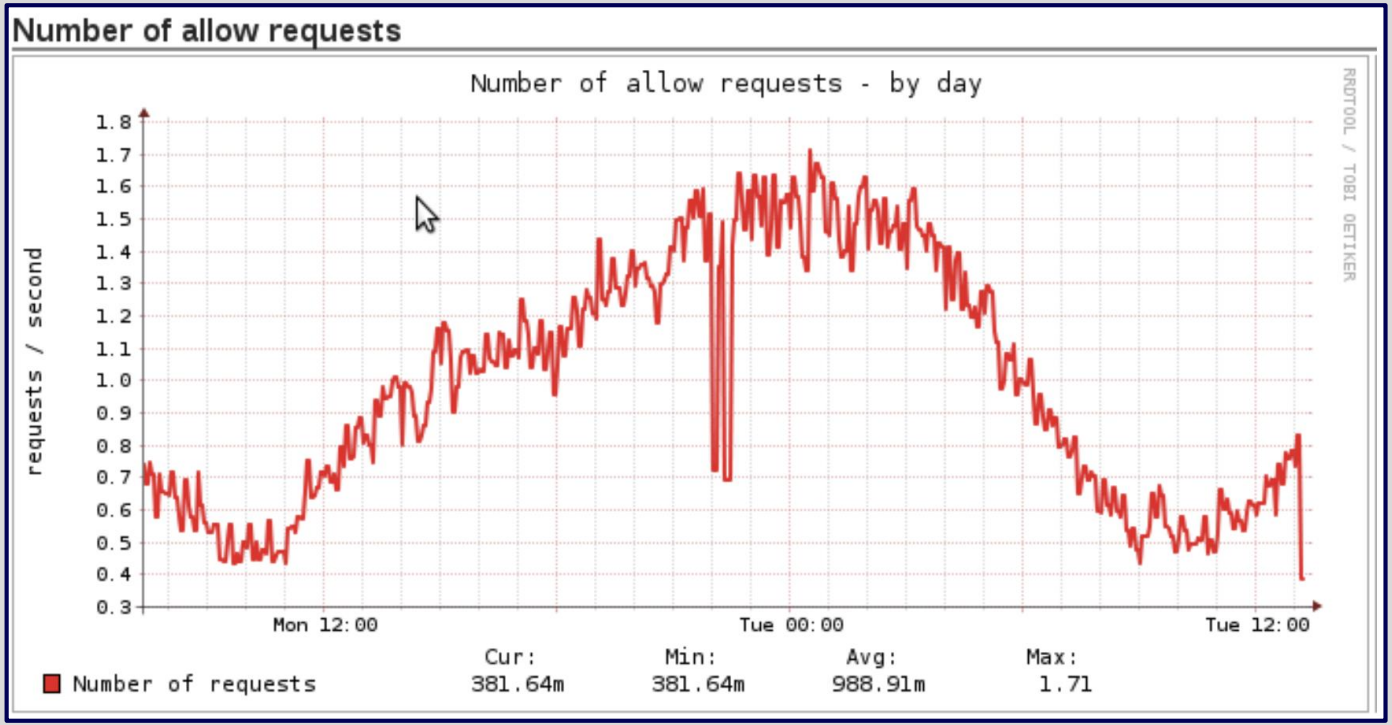
Where do data come from?

- Sensors in a single jet engine can generate 10 terabytes of data in 30 minutes.
- With more than 25,000 airline flights per day, the daily volume of data from just this single source is incredible.
- **Smart electrical meters**, sensors in heavy industrial equipment, and **telemetry** from automobiles increase the volume of Big Data

The Second V of Big Data

- **Velocity**
  - The rate at which data are generated and the speed at which it should be analysed
  - Wal-Mart, for instance, processes more than one million transactions per hour
  - It also reflects the speed of the feedback loop between your company and the customers

Daily requests at Spotify



https://labs.spotify.com/2013/03/15/backend-infrastructure-at-spotify/

The Third V of Big Data

- **Variety**
  - Unstructured vs. structured data
  - Tabular data constitute **only 5%** of all existing data
  - How to store & analyse (written) text, voice, facial features, images, video etc.?

Even more V's ??

- **Veracity**
  - Coined by IBM, which refers to the unreliability of data coming from different sources
- **Variability**
  - Coined by SAS, which refers to the variation in velocity
- **Value**
  - Coined by Oracle. Data increase value as they are processed

Data Mining

- Data can tell more than the obvious things. When properly analysed big data can reveal valuable patterns and information - helping predict trend and behaviour.
  - “People who bought X also bought Y …”
- Detecting credit card fraud - habits of card owner formed. If your card is stolen and used fraudulently, the usage often varies noticeably from your established pattern

Why Big Data

- Personalisation marketing
- Better pricing (e.g., auto adjust prices)
- Cost reduction (e.g., optimisation)
- Improved customer service
- Improved quality of life (for smart cities)

Leveraging Big Data (1)

- Organisations must do more than simply manage Big Data; they must also gain value from it.
- In general, there are six broadly applicable ways to leverage BD to gain value

Leveraging Big Data (2)

- Creating Transparency
- Enabling Experimentation
- Segmenting Population to Customise Actions
- Replacing/Supporting Human Decision Making with Automated Algorithms
- Innovating New Business Models, Products, and Services
- Organisations Can Analyse Far More Data

Now hold your fire…

- Big Data **is only valuable** when it contributes to better decision making!
  - When your boss suggests Big Data, be careful…
- This boils down to **two main processes**:
  - Data management; and
  - Data analytics (to be covered later)

Session 4

Organising Data

RMIT UNIVERSITY

Organising Data (1)

So how do we do it…

… Sometimes not very well!

- Forms which request repeated information
- Addresses can vary in format
- Data increases exponentially with time
- …

Organising Data (2)

- Multiple data sources
  - Data Rot + Degradation
  - Data Security + Quality + Integrity + Regulation

Data Rot (1)

- Data Rot refers primarily to problems with the media on which the data are stored. Over **time**, **temperature**, **humidity**, and exposure to **light** can cause physical problems with storage media and thus make it difficult to access the data.

Data Rot (2)

- The second aspect of data rot is that finding the machines needed to access the data can be difficult.
- ...We've covered how data could be stored, and mentioned past and present technologies.

Data Governance (1)

- To address the numerous problems associated with managing data, organizations are turning to data governance.
- Data governance is an approach to managing information across an entire organization.
  - It involves a formal set of business processes and policies that are designed to ensure that data are handled in a certain, well-defined fashion.

Data Governance (2)

- So an organisation would follow unambiguous rules for **creating**, **collecting**, **handling**, and **protecting** its information.

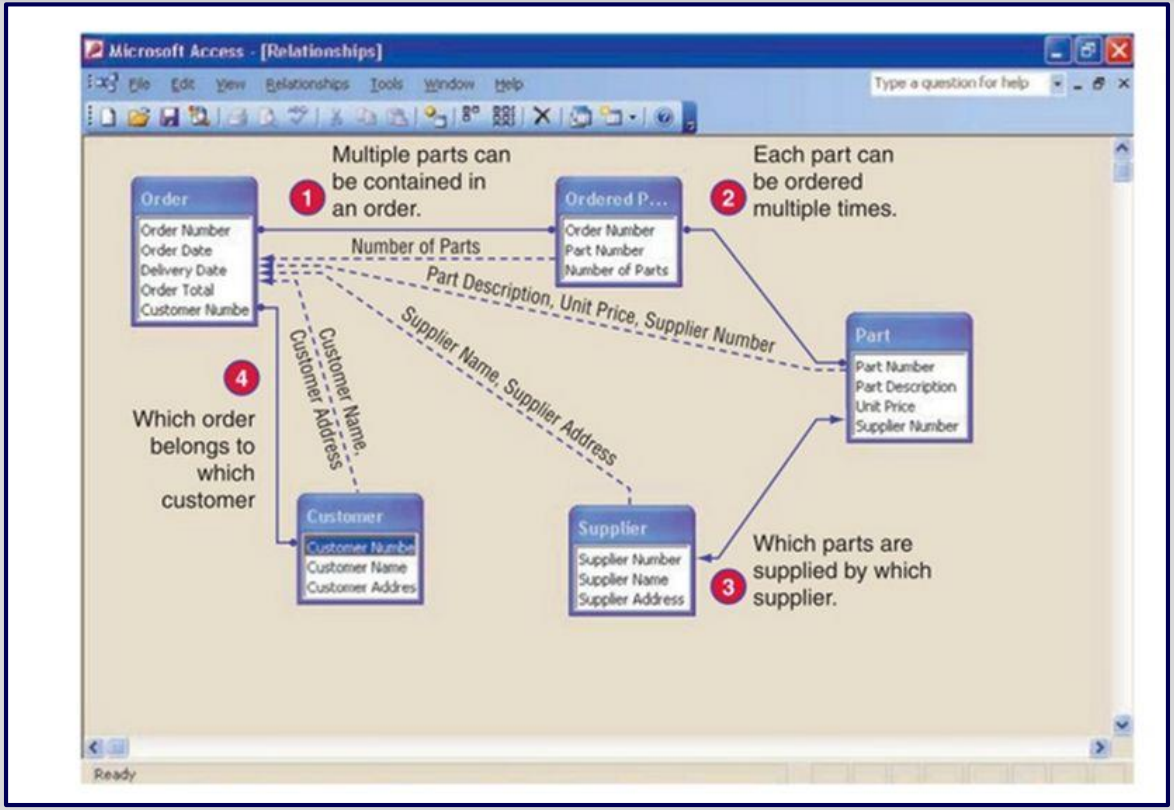
Relational Databases (1)

- A relational database generally is not one big table—that contains all of the records and attributes.
- Such a design would entail far too much data redundancy. Instead, a relational database is usually designed with a number of related tables.
- Each of these tables contains records (listed in rows) and attributes (listed in columns).

Relational Databases (2)

- The most commonly performed database operation is requesting information.
- Structured query language (SQL) is the most popular query language used for this operation.

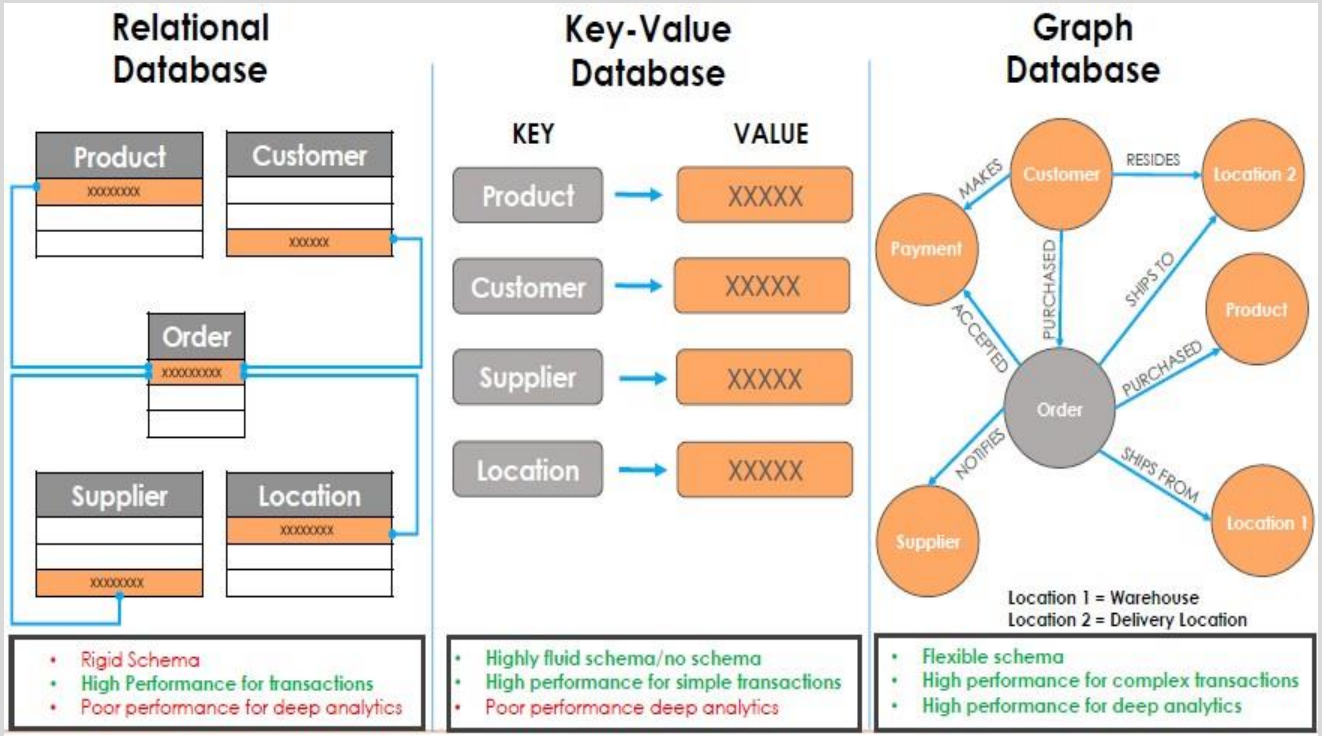
Relational Databases (3)



NoSQL Databases

- In addition, many organizations are turning to NoSQL databases (think of them as “not only SQL” databases) to process Big Data.
- These databases provide an alternative for firms that have more and different kinds of data (Big Data) in addition to the traditional, structured data that fit neatly into the rows and columns of relational databases.

Different types of databases



NoSQL database examples

- Let’s watch a video about MongoDB
  - https://www.youtube.com/watch?v=EE8ZTQxa0AM

NoSQL database examples

- Graph database Neo4J
- https://www.youtube.com/watch?v=3RARmkXtp30

Data security, quality, integrity

- Data security, quality, and integrity are critical, yet they can be easily jeopardised.
- Information sharing between firms
- Data intensive operations = less security?
- Missing or unreliable data
  - Imagine the impacts in healthcare context!

Regulations

- Data localisation
- Privacy concerns
- Data ownership
  - aka mining data responsibly
- Legal regulations have made it a top priority for companies to better account for how they are managing information.

Goals for this week

- Finish off Assignment 1

  GitHub Pages Website

- Work on a ‘Pitch for a Group’ to be ready for next week

- Start thinking about Assignment 2

Copyright © 2017 RMIT University Vietnam