Big Data

So far in this course you have learned about the Internet of Things and Cloud Computing. Besides the benefits and applications that we already discussed, there is another product of modern computing that is worth looking into; data, lots and lots of data. In this module we will take a close look at the huge collection of information that results from all the computing that is going on in the world today. In this module, you will learn about Big Data.

Objectives:

05_Obj01: Define Big Data
05_Obj02: Identify the key enablers of Big Data
05_Obj03: Identify the categories of Big Data
05_Obj04: Identify the main sources of Big Data
05_Obj05: Identify the V's of Big Data

What is Big Data?

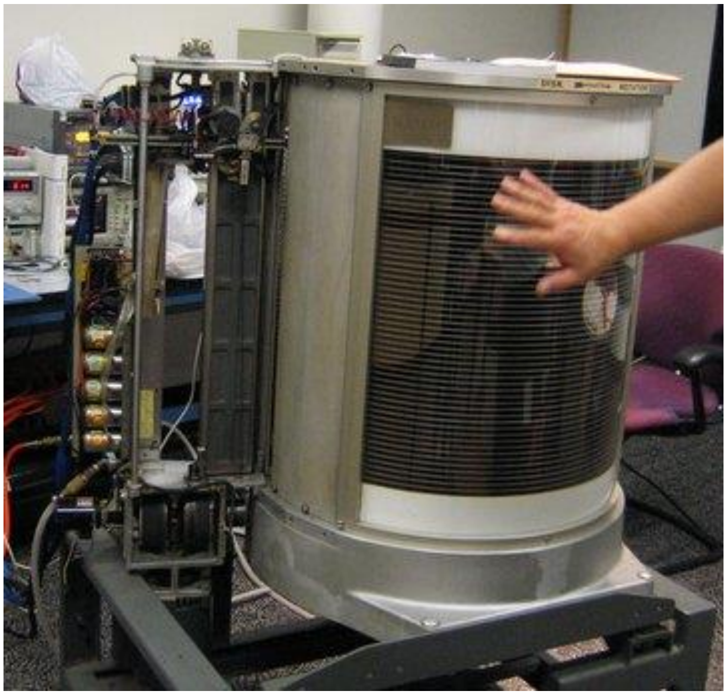
According to the Merriam-Webster dictionary, Big Data is "*an accumulation of data that is too large and complex for processing by traditional database management tools*". Big data is the massive collection of structured and unstructured digital information that is generated by digital computing. We have come to a point where there are so much data that traditional techniques and applications are no longer enough to process all of them.

How is Big Data possible?

Just like Cloud Computing and the Internet of Things, Big Data is not a single technology. It is a concept; a single term that represents the combination of different factors that enable the collection, management, and analysis of vast collections of data. Understanding these enabling factors or enablers is necessary to understand how Big Data works. Here are the three key enablers Big Data:

Data Storage

Data storage or the process of recording information is the main factor that makes big data possible. Technologies that enable people to record electronic information have been around for decades but it is only recently that data storage technologies were able to support the scale that is required by Big Data.



IBM 350 RAMAC Retrieved August 11, 2017, from **https://commons.wikimedia.org/wiki/File:IBM_350_RAMAC.jpg**

In the early of computing, data storage was limited and expensive. To put it in perspective, the **IBM 350 RAMAC**, the first hard drive which was released in 1956 could only store 5 megabytes and was as big as a fridge. Today, you can buy a 1 terabyte (1 million megabytes) external hard drive that can fit in your pocket for about 60 US dollars or 3000 Philippine pesos

Processing Capacity

Processing capacity or processing power refers to how many operations a processor can perform in a given amount of time. Stored data is pretty much useless unless it can be accessed and manipulated. This is why processing capacity is a key enabler of Big Data. We need powerful computers to process all the information that we collect.

Data Availability

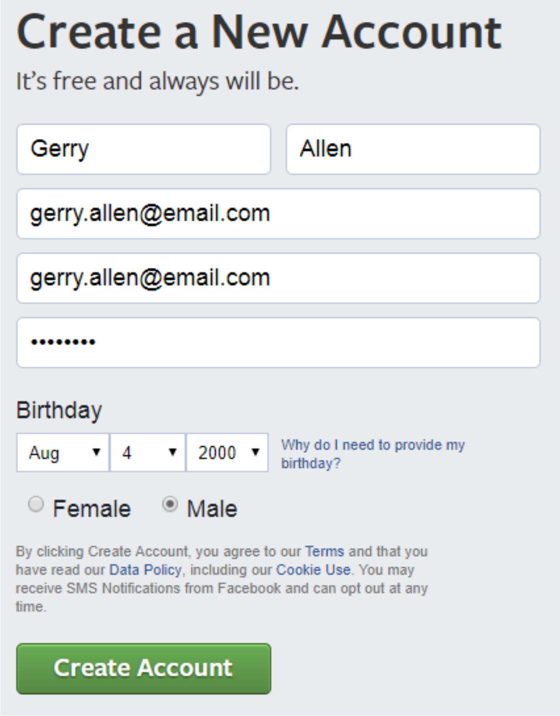
Data availability refers to the process of making data accessible to users and applications whenever it is needed. Information needs to be easy to reach. Even if you have a lot of stored data and computers powerful enough to process them, it would not matter unless you have an efficient way for your computers to access the data.

Categories of Big Data

In order to differentiate the categories of Big Data, first you have to know about metadata. Metadata is data that provides information that other data. In short, metadata is *data about data*. To know what category a piece of information belongs to, you have to determine if it has metadata attached to it. There are two main categories:

Structured Data

The first one is structured data which has metadata and a fixed length and format. These characteristics make structured data easy to locate, store, and process.



Create a New Account

It's free and always will be.

Gerry

Allen

gerry.allen@email.com

gerry.allen@email.com

••••••••

Birthday

Aug ▾ 4 ▾ 2000 ▾

Why do I need to provide my birthday?

○ Female ● Male

By clicking Create Account, you agree to our Terms and that you have read our Data Policy, including our Cookie Use. You may receive SMS Notifications from Facebook and can opt out at any time.

Signup page. Retrieved August 3, 2017, from https://facebook.com

For example, when you fill out an online form, (e.g. Facebook signup) you are entering information in fixed locations or fields. Each of these fields has metadata attached to it so when submit the form; the system will know which piece of information is your first name, last name, birth date, and so on.

Unstructured Data

The second type is *unstructured data* which is not organized in a readily recognizable way. Structured data doesn't have metadata.

> My name is Gerry Allen. My friends call me Gerry. I am 17 years old. I am male. I was born in August 4, 2000. My email address is gerry.allen@email.com. I won't tell you my password because that's private.

Here's an example, this text contains the same information as the Facebook signup form. However, the information in it is not presented in a recognizable format. It does not have metadata that points out which are names, dates, emails, etc.

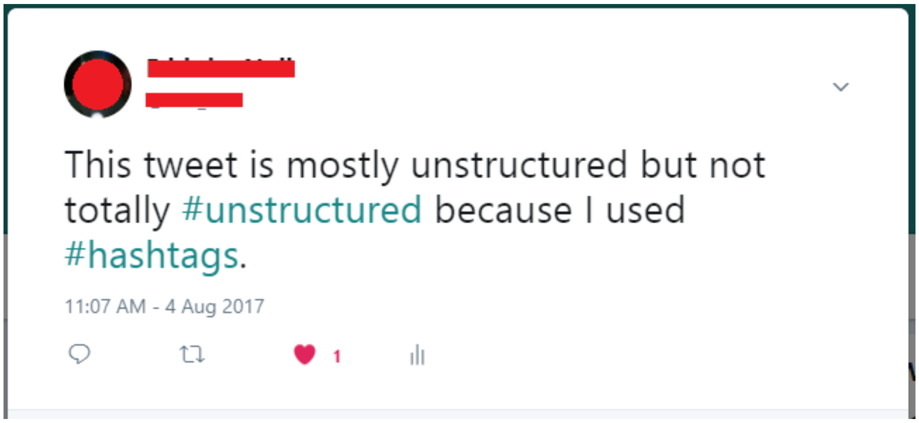
It is estimated that 80% of all data are unstructured. This includes online social interactions (to be discussed later) such as messages, pictures, sounds, and videos. Unstructured data is useless in its present form. It needs to be processed to yield structured data.

Semi-structured data

You may come across the term "semi-structured data" As the name implies, semi-structured data is a cross between structured and unstructured data. Semi-structured data is not completely organized but it has some recognizable parts.

> My name is Gerry Allen. My friends call me Gerry. I am 17 years old. I am male. I was born on the August 4, 2000. My email address is gerry.allen@email.com. I won't tell you my password because that's private.

Here's an example: There are programs that can automatically recognize certain data types within text documents such as dates, URL's and email addresses.



[redacted]

This tweet is mostly unstructured but not totally #unstructured because I used #hashtags.

11:07 AM - 4 Aug 2017

You are probably familiar with hashtags. A hashtag is a *metadata tag* that is used in social media to identify messages related to a specific topic. In the above example, if you click on either of the two hashtags, you will be redirected to the collection of tweets that share the same topic. Hashtags enable social media sites to categorize otherwise unstructured data.

Where does Big Data come from?

Many people say that Big Data is everywhere but in order to properly utilize big data, it is important to know where it specifically comes from. The following are the three main sources of Big Data:

Social Data

Social data comes from people’s online social interactions. The most common example of this is social media. Status updates, preferences, photos, videos, etc. are recorded by social media sites. They stay in the servers unless we choose to delete them or if we delete our account. In Facebook alone, it is estimated that 300 million photos are uploaded every day.

Transactional Data

It is a common practice to keep digital records of business transactions, even the ones that are done offline. For example, when you buy groceries, your transaction is recorded by the store. These records can be analyzed to gain valuable information. Therefore, companies have to keep records of all their transactions. It is estimated that by 2020, there will be 450 billion business transactions created in the Internet per day.

Machine Data

People are not the only ones that create data. In the first module, you learned about smart objects and the sensors that enable them to detect real-world input such as temperature, pressure, and motion.

These inputs are transmitted through the wireless network and saved in servers.

The V's of Big Data

To better understand Big Data, we have to take the 3V's into consideration. The Three V's are the main defining characteristics of Big Data. They are:

Volume

The main characteristic of Big Data is its size, hence the name. It is estimated that 2.5 quintillion bytes or 2.3 trillion gigabytes of data are created each day.

Variety

Big data is complex. Variety refers to the different types of data. Earlier, we discussed structured data which is composed of many known data types. On top of that, there's unstructured data which has the potential to yield more data types. Finally we have different sources of big data, each with their own unique data types.

Velocity

Big Data is also fast. Velocity is the frequency of incoming data. In 2016, the average Global IP data traffic was more than 88,719 petabytes per month. 1 petabyte is equal to 1 million gigabytes. In 2016, it is estimated that there were 18.9 billion network connections in the world. That means there is an average of 2.5 connections per person. As of July 2015, 400 hours of video is uploaded to Youtube **every minute**.

Other V's

While Volume, Variety and Velocity are generally recognized as the main defining attributes of Big Data, other V's can also be considered to further define it. Here are some of them:

- **Veracity** - the trustworthiness of data. It is important to differentiate between correct and incorrect data. For example, some people create fake or multiple accounts. Information from these accounts can be redundant or incorrect.

- **Value** - the usefulness of data. Being correct does not automatically make data useful. Some of the information collected from the Internet may not be useful to anyone.

- **Variability** - the constant change in the meaning of data. A piece of information can have varying meanings. In many cases, automated systems cannot recognize the contexts that apply to certain data.

- **Visualization -** condensing and converting data into understandable format. After the data has been processed, it needs to be presented in a form that can be understood by people. For example, after analysing information from millions of users, the most important results are presented in chart or diagram form.

Glossary of Terms

BIG DATA – is an accumulation of data that is too large and complex for processing by traditional database management tools

CLOUD COMPUTING – is sometimes referred to as *the cloud*. It is the practice of storing, accessing, and processing data in remote locations through the Internet

DATA STORAGE – or the process of recording information is the main factor that makes big data possible

PROCESSING CAPACITY – or processing power refers to how many operations a processor can perform in a given amount of time

DATA AVAILABILITY – refers to the process of making data accessible to users and applications whenever it is needed

Sources:

Ibmbigdatahub.com. (n.d.). Where does big data come from? Retrieved September 17, 2017, from http://www.ibmbigdatahub.com/infographic/where-does-big-data-come

Datascience.berkeley.edu. (2014, September 03). What Is Big Data? Retrieved September 17, 2017, from https://datascience.berkeley.edu/what-is-big-data/

Sas.com. (n.d.). Big Data Insights. Retrieved September 17, 2017, from https://www.sas.com/en_us/insights/big-data.html

Marr, B. (2015, November 19). Big Data: 20 Mind-Boggling Facts Everyone Must Read. Retrieved September 17, 2017, from https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#70af93d717b1

Internetsociety.org. (n.d.). Global Internet Report 2016. Retrieved September 17, 2017, from http://www.internetsociety.org/globalinternetreport/2016/

Hadoopadmin.co.in. (n.d.). Sources of BigData. Retrieved September 17, 2017, from http://www.hadoopadmin.co.in/sources-of-bigdata/

Karr, D. (2017, January 14). Big Data Brings Marketing Big Numbers. Retrieved September 17, 2017, from https://martech.zone/ibm-big-data-marketing/

Merriam-webster.com. (n.d.). Big Data. Retrieved September 17, 2017, from https://www.merriam-webster.com/dictionary/big%20data

Ibmbigdatahub.com. (n.d.). The Four V's of Big Data. Retrieved September 17, 2017, from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Dummies.com. (n.d.). The 4 V’s of Big Data. Retrieved September 17, 2017, from http://www.dummies.com/careers/find-a-job/the-4-vs-of-big-data/

Ibm.com. (2017, May 16). The biggest data challenges that you might not even know you have. Retrieved September 17, 2017, from https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

Internetlivestats.com. (n.d.). Internet Users. Retrieved September 17, 2017, from http://www.internetlivestats.com/internet-users/

Zephoria.com. (2017, August 29). Top 20 Facebook Statistics - Updated August 2017. Retrieved September 17, 2017, from https://zephoria.com/top-15-valuable-facebook-statistics/

Granville, V. (n.d.). A Comprehensive List of Big Data Statistics. Retrieved September 17, 2017, from http://www.bigdatanews.datasciencecentral.com/profiles/blogs/a-comprehensive-list-of-big-data-statistics

Statista.com. (n.d.). YouTube: hours of video uploaded every minute 2015 | Statistic. Retrieved September 17, 2017, from https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

Computerhistory.org. (n.d.). 1956: First commercial hard disk drive shipped. Retrieved September 17, 2017, from http://www.computerhistory.org/storageengine/first-commercial-hard-disk-drive-shipped/

McNulty, E., McNulty, T. A., Szekely, B., Salazar, J., & White, I. (2017, May 08). Understanding Big Data: The Seven V's. Retrieved September 17, 2017, from http://dataconomy.com/2014/05/seven-vs-big-data/

Links to Videos and Readings

Video: Big Data – Tim Smith

Smith, T. (2013, May 03). Big Data - Tim Smith. Retrieved September 17, 2017, from https://youtu.be/j-0cUmUyb-Y

PDF:

Chakraborty, G., Dr. (n.d.). Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining. Retrieved September 17, 2017, from https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf