What is big data and what are its characteristics?

The phrase "big data" has become trendy nowadays. It refers to sets of data which are far too massive to be handled with traditional hardware and software for data analysis. Today, many data gathering capabilities have experienced explosive growth and this leads to challenging issues in storing and analyzing the resulting data.

We see data come from everywhere, including finance, marketing, weather forecasting, medicinal science and e-commerce. Even the search engines like Google and posts to social media sites such as Facebook and Twitter are contributing to the availability of big data. These developments have led to the evolution of an entirely new profession: *the data scientist*, who can combine the fields of statistics, mathematics, computer science and engineering with knowledge of a specific application.

The characteristics of big data include three factors which distinguish it from other types of data: *volume, velocity* and variety.

Volume is easy to understand. With big data, the volume is certainly massive. In fact, new terminology has been used to describe the size of these datasets. For example, in here we do not talk about megabyte or gigabyte of data anymore but petabyte which refers to 1.0×10^{15} bytes of data.

Remember a byte is a single unit of storage in a computer's memory and it is used to represent a single number, character, or symbol. A byte however, consists of eight bits, each consisting of either a 0 or 1. Hence a petabyte means 1,000 trillion bytes!

Velocity refers to the speed at which data is gathered. Big datasets are continuously gathered at very high speed, generally received through streaming the data or by a complex event processing.

When we download and watch a movie from an internet source such as Youtube, the data is being downloaded at an extremely high speed while the movie is playing. We usually get very annoyed if our computer encounters a slow internet connection because we will face with annoying interruptions or glitches as the data downloads online.

Streaming is useful when we need to make decisions in real time. For example the geologist in an earthquake-prone region has to receive quick information on unusual earth plate movement, and the traders have to make split-second decisions as new market information becomes available.

Complex event processing (CEP) refers to the use of data to predict the occurrence of events based on a specific set of factors. With this type of processing, data is examined for patterns that could not be found with more traditional approaches, so that better decisions may be made in real time. A good example is our GPS device's ability to guide us to reroute our car driving based on the traffic and accident data available.

Variety refers to the fact that the contents of a big dataset may consists of a number of different formats, including spreadsheets, word processing documents, videos, photos, music clips, email/text messages and so on. Storing a huge quantity of these incompatible types is one of the major challenges of big data.

We have to find ways to extract useful information from multiple types of disparate files. In addition to the traditional database management systems including the usual relational, hierarchical and network models, we can consider: (1) distributed storage to spread the data out over multiple storage devices for quicker access, and (2) parallel processing the data by different processors. The human genome project is wholly dependent on having a server farm to sort out the seemingly infinitive number of possibilities.

In conclusion, if we can only process the big data available to us by traditional approaches whilst the big data is continuously streaming in, we will not able to leverage ALL the data available to make informed and strategic decisions, and will surely miss out many untapped information sources and opportunities.