What is Big Data?

Learn about applications of big data and how we can describe it with the 3 Vs

Data is Everywhere

We generate data from all kinds of activities, whether a transaction at a local store, a website we visit, or even the location of our cell phone. At the time of this writing, an average of 500 million tweets are written on Twitter each day. Imagine being the person at Twitter who has to analyze this data! Data of this size is often referred to as *big data*.

What exactly is big data? In general, big data is any data that is too big for a typical modern computer to process and analyze. This means, however, that the definition of big data is relative to the amount of computing power we have available. For example:

> Most current personal computers have somewhere between 8-32 GB of random access memory (RAM) available for data processing. That means, from the perspective of a personal computer, any dataset larger than 10-20 GB might be too large to process.
>
> A large enterprise can take advantage of larger computing resources (i.e., a warehouse of servers or the cloud), so 100+ GB might be the upper limit for the size of data the enterprise can handle.
>
> Data measured in terabytes is perhaps the largest amount of data being worked with at this time (1 TB = 1000 GB).

In the following applet, try adjusting the slider to see what qualifies as big data as we increase our modern computing power

Big data hasn’t always been a concept with respect to data analysis. For most of history, we have been able to handle the amount of data we collect. Before computers, scientists would perform calculations on handwritten data for a research sample. With the invention of computers, we were able to process data more quickly and were generally able to keep up with the amount of data we had available.

In more recent history, however, sources of data have continued to grow and are outpacing the growth of computing power. In the mid-2000s, after the massive growth of the internet, many analysts in the industry were struggling to handle their own data. Roger Mougalas coined the term “big data” when referring to a dataset that was unmanageable with current business intelligence tools.

Multiple choice

How big does data have to be for it to be considered big data?

Greater than 1 terabyte (TB)

Greater than our available computing power can handle

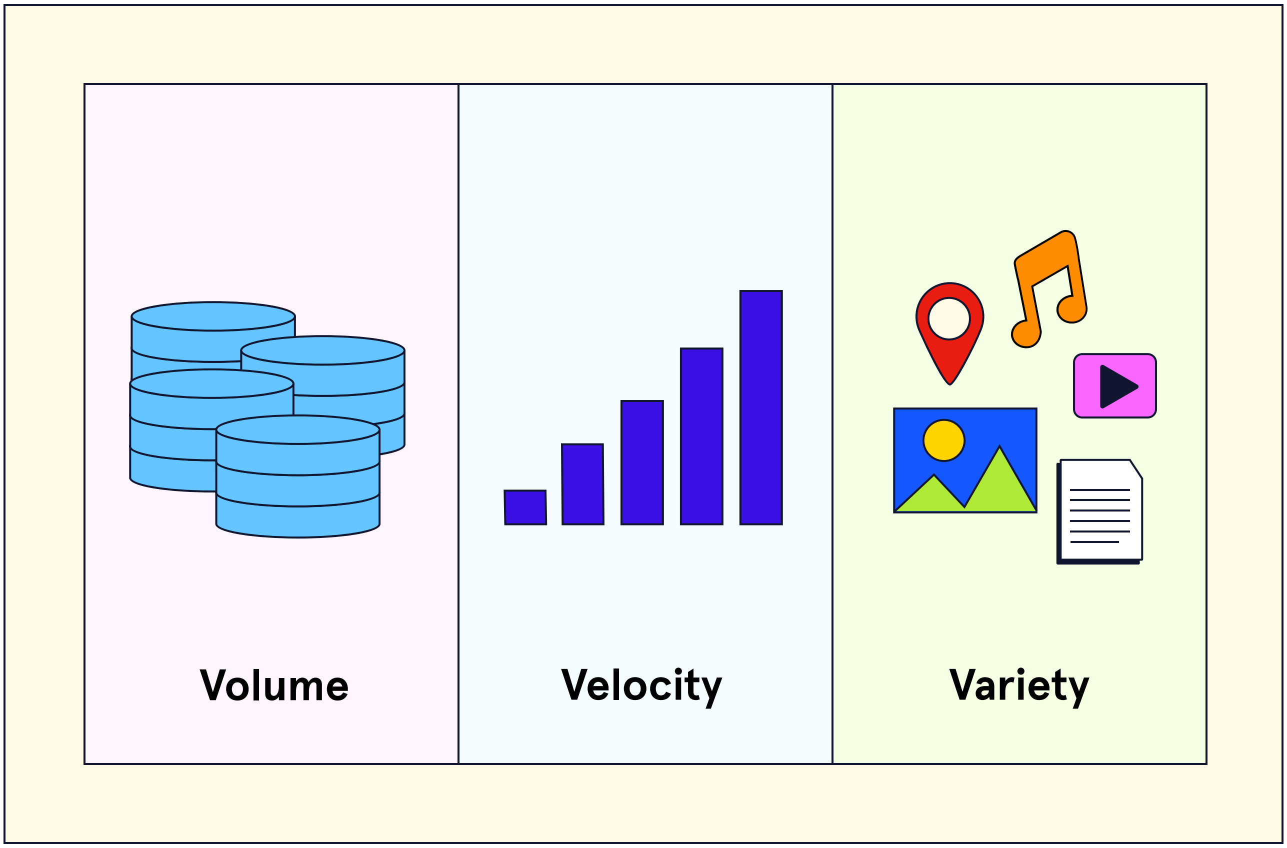
Greater than 100 gigabytes (GB)

Greater than 1 petabyte (PB)

👏 Correct! Big data is defined relative to modern the computing power available to process the data. There is no specific number to define the size of big data.

The 3 Vs

Big data is a relative concept that can be difficult to grasp. It may be easier to define big data using the features that make it hard to handle in the first place. We can generally categorize big data by what are known as the three Vs: volume, velocity, and variety. Depending on where you look (or who you ask), there may be a different number of Vs. Some will say that there are 4, 5, 10, or even 17 Vs of big data! The three Vs we talk about here are the core of most definitions and give a complete picture of big data's features.



Volume

Big data is “big”. While this may seem obvious, it’s an important concept to cover. As previously mentioned, the definition of “big” is that the data is bigger than the amount of available computing power. Currently, zettabytes of data are created every year (for reference, a zettabyte is 1 billion terabytes).

Velocity

Big data has velocity, meaning that it is growing quickly. If data were simply large, but slow-changing, then over time our computing power would eventually catch up to the size of the data. Through means like apps and sensors, data becomes faster, cheaper, and easier to collect automatically and continuously.

Variety

Big data also has variety, meaning that it comes in different, and sometimes complex, forms. In today’s data ecosystem, data comes in many more formats than the data tables of old. Data can be categorized as structured (data tables with rows and columns), semi-structured (think JSON files with nested data), and unstructured (audio, image, and video data). Each of these data formats presents different challenges in processing.

Multiple choice

What are the 3 Vs that describe big data?

Vertical, Veracity, Variety

Volume, Vertical, Variety

Volume, Velocity, Value

Volume, Velocity, Variety

👏 Correct! The 3 Vs of big data describe its size, speed, and complexity.

Big Data Applications

What can we do with big data? When do we run into big data in the real world? Let’s explore a few examples of big data applications across various industries.

Social Media

With an average of 500 million tweets per day, Twitter data would definitely qualify as big data. Despite the massive amount of data at its disposal, Twitter provides analytics for each user with the ability to dive into historical Tweet activity and identify trends. In order to provide this for each Twitter user, they must be using some kind of big data toolkit to store and analyze the data.

Healthcare

If we look at the healthcare industry, a rising trend that many providers want to enable is known as evidence-based medicine. Healthcare providers want to combine data from several sources, including cell phone apps, diagnostic tests, and previous medical records, to give recommendations for each patient. Providers hope this will avoid expensive and unnecessary tests and improve patient outcomes. In this case, there is a lot of data from many different sources and formats, but the efforts could provide a huge impact for their patients.

Finance

In the financial industry, credit card companies aim to reduce the amount of fraudulent transactions, as these cost money and cause hardship for their customers. Using different tools, credit card companies are able to analyze every single credit card transaction and use machine learning models to identify transactions that could be fraudulent. This saves a massive amount of money for themselves and for their customers!

Final Thoughts

No matter which industry we look at, we will find numerous examples of big data everywhere, and there are more every day. However, we need to be aware that big data often comes with both big challenges and big effects. With the right tools and techniques, we can begin to extract value from big data while being mindful of both its limitations and impacts.