**Defining Data**

* Data refers to collections of digitally stored units, in other words, stuff that is kept on a computing device
* Single units of data are traditionally referred to as **datum**

Categories of Data:

* **Structured:** Data that has been formatted to a set structure; each data unit fits nicely into a table in a database. It’s ready for analysis. Examples include first name, last name, and phone number.
* **Unstructured:** Data that are stored in a native format must be processed to be used. Further work is required to enable analysis. Examples include email content and social media posts.
* **Semi-structured:** Data that contains additional information to enable the native format to be searched and analysed.

The Zettabyte Era:

|  |  |
| --- | --- |
| 8 bits | 1 byte |
| 1024 bytes | 1 kilobyte |
| 1024 kilobytes | 1 megabyte |
| 1024 megabytes | 1 gigabyte |
| 1024 gigabytes | 1 terabyte |
| 1024 terabytes | 1 petabyte |
| 1024 petabytes | 1 exabyte |
| 1024 exabytes | 1 zettabyte |
| 1024 zettabytes | 1 yottabyte |
| 1024 yottabytes | 1 brontobyte |

Data Volumes:

| **Unit** | **Value** | **Example** |
| --- | --- | --- |
|  |  |  |
| Kilobytes (KB) | 1,000 bytes | a paragraph of a text document |
| Megabytes (MB) | 1,000 Kilobytes | a small novel |
| Gigabytes (GB) | 1,000 Megabytes | Beethoven’s 5th Symphony |
| Terabytes (TB) | 1,000 Gigabytes | all the X-rays in a large hospital |
| Petabytes (PB) | 1,000 Terabytes | half the contents of all US academic research libraries |
| Exabytes (EB) | 1,000 Petabytes | about one fifth of the words people have ever spoken |
| Zettabytes (ZB) | 1,000 Exabytes | as much information as there are grains of sand on all the world’s beaches |
| Yottabytes (YB) | 1,000 Zettabytes | as much information as there are atoms in 7,000 human bodies |

From Data to Insight:

| ***Data*** | ***Information*** |
| --- | --- |
| Raw | Processed |
| Items such as characters, words, pictures, and numbers that have no meaning in isolation | Data that is organized and given context to have meaning |
| No analysis dependency | Dependent on the analysis of data |
| A diagram of a diagram  Description automatically generated  Unorganized and not dependent on context | Organized and dependent on context |
| Not typically useful alone | Useful alone |

The Role of Data in the 21st Century:

Data-Driven Decision-Making:

* data must be organized and analysed to understand patterns, make decisions, identify problems, and feed other systems.

**Data ownership** describes the rights a person, team, or organization has over one or more data sets.

Data Architecture:

* **data architecture** is the manner in which data design and management decisions are being made to align with Enterprise Architecture (EA) and in turn, with the business.
* data architecture is the agreed blueprint for how data supports an organization’s functions and technologies.
* high-quality enterprises and data architectures both exist, organizations run more smoothly, and they can transform as conditions (either internally or in the marketplace) dictate. The absence or poor implementation of both can stifle digital transformation efforts, create high levels of complexity, and increase the possibility of failure.

**data architecture** considers and typically supports:

* Ensuring data is available to those who need it and are approved to use it.
* Reducing the complexity of accessing and utilizing data
* Creating and enforcing data protections to support organizational policies and obligations.
* Adopting and agreeing to data standards
* Optimizing the flow and efficient use of data to eliminate bottlenecks and duplication.

The Lifecycle of Data:

1. **Creation:** This is the stage at which data comes into being. It may be manual or automated and get created internally or externally. Data is created all the time by a vast number of activities that include system inputs and outputs.
2. **Storage:** Once data is created and assuming you want it available for later use, it must be stored. It most likely will be contained and managed in a database. The database needs a home, too as a local hard drive, server, or cloud service.
3. **Usage:** Hopefully you’re capturing and storing data because you want to use it. Maybe not immediately, but at some point, perhaps for analysis. Data may need to be processed to be useful. That could include cleansing it of errors, transforming it to another format, and securing access rights.
4. **Archival:** In this stage, you identify data that is not currently being used and move it to a long-term storage system out of your production environment. If it's needed at some point in the future, it can be retrieved and utilized.
5. **Destruction:** Despite a desire by some to keep everything forever, there is a logical point where destruction makes sense or is required by regulation or policy. Data destruction involves making data inaccessible and unreadable. It can include the physical destruction of a device such as a hard drive.

**Defining Big Data**

One way to define and characterize big data is through these five Vs:

1. **Volume:** The sheer scale of data being produced is unprecedented and requires new tools, skills, and processes.
2. **Variety:** There are already a lot of legacy file formats, such as CSV and MP3, and with new innovations, new formats are emerging all the time. This requires different methods of handling, from analysis to security.
3. **Velocity:** With so many collection points, digital interfaces, and ubiquitous connectivity, data is being created and moved at increasing speed. Consider that in 2021, Instagram users created, uploaded, and share 65,000 pictures a minute.
4. **Variability:** The fact that the creation and flow of data are unpredictable.
5. **Veracity:** The quality, including accuracy and truthfulness, of large volume of disparate sets of data, can differ considerably, causing challenges to data management.

Consequences of Big Data: