**Data**

Data is a collection of raw facts and figures that we need to process to extract meaning or information. Data is nothing unless it is processed or aligned in some context, when data is structured and organized, results in Information.

There are the four (4) data formats commonly seen in data processing, they include structured, semi structures, quasi-structured and unstructured data format.

String, Integers, float and boolean are some common examples of data type.

**Big Data**

Big Data defines a situation in which data sets have grown to such enormous sizes that conventional information technologies can no longer effectively handle the size of the data set. Big data grows exponentially (becoming more rapid) and is boundless (volume), has varied formats (variety), and has high complexity (voracity and velocity). The four Vs of big data: volume, variety, voracity and velocity.

Big data emerges from a variety of sources comprising of Internet of things (IoTs), factories, enterprise resource planning and customer relations management systems according to industry experts.

Many big data tools used by enterprises to identify business opportunities, improve performance, and drive decision-making are open source and not designed with security in mind. This huge increase in data consumption leads to many data security concerns, which include Information theft, distributed denial-of-service (DDoS), ransomware etc.

It is difficult to secure big data because of constant access by different users, the presence of open source tools and the multiple feeds of data from sources with different protection needs. Big data security challenges can be experienced both on-premises and in the cloud. It is prevalent in non-relational database, endpoint vulnerabilities, data mining solutions, access distributed data processing and storage task.

Data security threats can he handled using security techniques such as, user access centralized key management, encryption, and intrusion detection and prevention (IDPS). One way that organizations can protect data is through encryption, which applies algorithms to scramble data so that it is readable only by someone who holds the key to decrypt it. Encryption takes a piece of data, commonly called the plaintext, combines it with a cryptographic key. This produces a scrambled version of the data called the cipher text. It is possible to decrypt the data to recover the plaintext using the key, but without the key, the cipher text hides all information about the original data, other than its length.

Using user access control to protect data, we turn to cryptographic techniques to secure data in storage. The primary goal of this technique is to enforce access control to data stored in potentially untrusted repositories. That is, we give authorized parties access to the data they need while ensuring that unauthorized parties, either outsiders trying to gain access or malicious insiders in the organization managing the repository, cannot access sensitive data.

An intrusion detection and prevention system (IDPS) is a solution that monitors network for simpler threats and then takes flag complex threats to alert security teams to stop any threat detected.

Internet of things (IoTs) are physical objects such as our phones, appliances, lighting systems, irrigation systems, security cameras, vehicles and cities equipped with sensors and software, which enable them to interact with each other by collecting and exchanging data via wireless network, with little human intervention. The Internet of things (IoTs) simplifies and automates complicated tasks by uniting the objects under one common infrastructure, transforming them into smart objects and remotely controls them.

**Data Wrangling**

When data scientists identify useful data sources for solving the business problem they then proceed to extract, clean, and format the necessary data from those sources. Data wrangling is the process that ensures that the data is in a format that is clean and accurate to yield a data set that is suitable for exploration and analysis. Data wrangling extracts the most valuable information from the dataset.

The process of data wrangling includes first finding the appropriate data that is necessary for the analysis. This data can be from one or multiple sources, such as tweets, bank transaction statements in a relational database, sensor data, and so on. Next, the data is cleaned, where there is missing data, we will either delete or substitute it with the help of several techniques. If there are outliers, we need to first detect them and then handle them appropriately. If data is from multiple sources, we will have to perform join operations to combine it.