2/5/2024

Data Creation, Data Capture, Data Classification
Ethical Considerations, Data Validation, Data Privacy

Data created any time information is recorded/measured about the world:

1. Human generated (answering surveys, twitter posts, video, etc.)
2. Machine generated data (satellites, sensors, log files, etc.) Internet of Things
3. Organization generated (sales records, government records, etc.)

Data creation was hard.

Steps to capture data

1. Decide what data to collect
2. What data capture tools to use?
    a. Organization and structure of data/files
    b. Data validation components
    c. Enable open and flexible formats, proprietary formats should be well documented
    d. Allow the data to be moved with high quality
3. Collection process – documented, transparent, reproducible
4. Compliance with privacy regulations

How is data classified?

Privacy level – public, internal-only, confidential, restricted
Context-based (sensitive information)
Content-based
User-based

Factors to consider:

Confidentiality
Integrity of the data – tends to require more storage space, variable accessibility
Availability
Data Type – some types of data require more storage
#<text<images<videos

Methods of Validation:

- Manual Intervals – visually inspecting all data "by hand", with people/lots of human hours
- Defined intervals – ex. Temperature data
- Equal Intervals
- Quantiles
- Standard deviation intervals (identification of outliers)
- Natural Breaks
- Geometric intervals
- Custom ranges

Text-analysis would have other kinds of checks.

Validation can focus on several types of factors:

- Data Type
- Ranges
- Uniqueness
- Consistency
- Non-null

Ethical Considerations

Institutional Review Boards – research data collection

Consent

Transparency

Accountability

Anonymity

Bias

Resources:

1. http://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%202/1.%2001-05.pdf
2. https://ardc.edu.au
3. https://www.imperva.com/learn/data-security/data-classification
4. https://digitalguardian.com/blog/what-data-classification-data-classification-definition
5. https://www.techtarget.com/searchdatamanagement/definition/data-classification
6. https://towardsdatascience.com/the-ethics-of-data-collection-9573dc0ae240
7. https://www.ama.org/marketing-news/the-murky-ethics-of-data-gathering-in-a-post-cambridge-analytica-world/
8. https://www.safe.com/what-is/data-validation/
9. https://www.varonis.com/blog/data-privacy
10. https://dataprivacymanager.net/5-things-you-need-to-know-about-data-privacy/

Extended commentary:

Data Creation refers to the process of generating new data. This could include manual data entry, the creation of new files or databases, or the automatic generation of data through sensors or other devices.

Data Capture refers to the process of collecting and acquiring data. This could involve downloading data from external sources or copying data from existing databases.

Data Classification refers to the process of organizing and categorizing data. This is an important step in data management as it allows for easier search and retrieval of data. Data classification may involve assigning metadata or tags to data to describe its content, format, and other characteristics. It may also involve identifying sensitive or confidential data and applying appropriate security measures.

Ethical considerations related to data validation and privacy are essential in ensuring that data is collected, processed, and used in a responsible and ethical manner. Here are some points to consider:

Data validation: When validating data, it is important to ensure that the data is accurate, reliable, and complete. Ethical considerations related to data validation include ensuring that the data is

not biased, that it is collected and processed in an objective manner, and that any errors or omissions are corrected promptly.

Data privacy: Protecting the privacy of individuals is a crucial ethical consideration in data analysis. This includes ensuring that the data is collected and used only for the intended purpose, that it is stored securely, and that individuals' personal information is protected. It is important to comply with applicable data privacy laws and regulations, such as GDPR, CCPA, and HIPAA.

Informed consent: Informed consent is an essential ethical consideration when collecting data from individuals. It involves ensuring that individuals are fully informed about the purpose of data collection, how their data will be used, who will have access to their data, and their rights with respect to their data. Individuals must give their consent voluntarily and without coercion.

Bias and discrimination: Ethical considerations related to bias and discrimination involve ensuring that data analysis is conducted in an objective and unbiased manner. This includes identifying and mitigating any biases in the data or data analysis process that could lead to discriminatory outcomes.

Transparency: It is important to be transparent about the data analysis process, including how data is collected, processed, and used. This includes being clear about any assumptions made, any limitations of the data, and any potential biases or conflicts of interest.

Overall, ethical considerations are essential in ensuring that data analysis is conducted in a responsible and ethical manner. By considering these ethical principles, data analysts can help to ensure that their work benefits individuals and society as a whole, while avoiding any negative impacts or unintended consequences.

Data capture refers to the process of obtaining and collecting data from various sources, including electronic devices, sensors, social media, surveys, and other sources. There are several methods used for capturing data, including:

Manual Data Entry: This involves entering data into a computer system manually, either from paper forms or other sources such as emails or other electronic documents.

Automated Data Capture: This involves using technology to capture data automatically, such as sensors, barcodes, RFID tags, or other automated systems.

Web Data Capture: This involves collecting data from websites, either through web scraping or by using APIs.

Social Media Data Capture: This involves capturing data from social media platforms such as Facebook, Twitter, and Instagram.

Mobile Data Capture: This involves capturing data from mobile devices such as smartphones or tablets, either through apps or other mobile technologies.

Overall, the method of data capture depends on the type of data being captured and the specific requirements of the data analysis project. It is important to ensure that the data is captured accurately and ethically, and that appropriate measures are in place to protect the privacy and security of the data.

Data classification is the process of organizing data into different categories or groups based on the level of sensitivity, value, or criticality to an organization. It involves identifying and assigning a specific label or tag to each data type to determine its level of confidentiality, integrity, and availability. The classification process usually includes the following steps:

Identify data types: Determine the different types of data an organization collects, such as personal data, financial data, confidential data, public data, etc.

Define data categories: Create categories or groups for data types based on their sensitivity or value to the organization, such as classified, sensitive, confidential, public, etc.

Assign labels or tags: Apply specific labels or tags to each data type based on its category, such as a classification level, color code, metadata, etc.

Define handling procedures: Develop guidelines for handling each data type based on its classification level, including access controls, storage, sharing, retention, and disposal.

Monitor and update: Regularly monitor and review the data classification system to ensure it remains up-to-date and effective, and update the classification of data as needed.

Proper data classification is essential for data protection, compliance with regulatory requirements, and effective data management. It helps organizations to identify and prioritize data security controls and allocate resources accordingly.

Data validation is the process of ensuring that the data collected is accurate, complete, and consistent. It involves a series of checks and tests to ensure that the data is of high quality and can be used for analysis.

The validation process involves several steps, including:

Checking for completeness: This involves ensuring that all required fields have been filled out and that there are no missing values.

Data formatting: This step involves ensuring that the data is in the correct format, such as date format or numerical format, so that it can be analyzed accurately.

Range checks: This involves checking that the data falls within a specific range or threshold, such as a temperature reading that falls within a normal range.

Cross-field validation: This involves checking that the data entered in one field matches the data entered in another field. For example, ensuring that a customer's age matches their birthdate.

Check for duplicates: This involves identifying and removing any duplicate data.

Data consistency: This involves ensuring that the data is internally consistent, such as checking that the sum of a set of values matches the total.

Overall, data validation is an essential step in ensuring that the data collected is accurate and of high quality.