UNIT 1

Data for Cyber Security and Forensic

-What is data science in computer engineering?

-Type of data encountered in CSF,

-Synthetic v/s real-world data,

-Data discovery and classification , outliers, incident reporting

-Assessing quality of data sets, Big Data , Cloud Computing,

---Business Intelligence,

-Cleansing unstructured data ,

-Data security strategies

what is data science in computer engineering?

Data science in computer engineering refers to the application of various techniques, processes, and systems to extract meaningful insights and knowledge from data.

It involves the use of computational and statistical methods to analyze and interpret complex datasets, with the ultimate goal of informing decision-making, solving problems, and creating value.

In the context of computer engineering, data science often involves working with large volumes of data generated by computer systems, networks, and other digital sources.

Some components of data science in computer engineering include:

1. **Data Collection**: Gathering relevant data from various sources, which can include databases, log files, sensors, and other data repositories.

2**. Data Cleaning and Preprocessing**: Preparing the data for analysis by addressing missing values, outliers, and other issues. This step ensures that the data is suitable for modeling and analysis.

3. **Exploratory Data Analysis (EDA):** Examining the data through statistical and visual methods to identify patterns, trends, and relationships. EDA helps in gaining a better understanding of the dataset.

4. **Feature Engineering**: Selecting, transforming, or creating new features from the existing data to improve the performance of machine learning models.

5. **Machine Learning**: Building predictive models using algorithms and statistical models to make sense of the data and make informed predictions or classifications.

6. **Data Visualization**: Creating visual representations of the data to communicate findings effectively. This can include charts, graphs, and dashboards.

7. **Big Data Technologies:** Dealing with large-scale datasets often requires the use of specialized tools and technologies, such as Apache Hadoop, Spark, and other distributed computing frameworks.

8. **Data Security and Privacy**: Ensuring that data handling practices adhere to ethical and legal standards, especially when dealing with sensitive information.

9. **Optimization**: Fine-tuning algorithms and models to improve efficiency and performance in terms of speed, accuracy, and resource utilization.

Type of data encountered in CSF

In cybersecurity and forensic investigations, various types of data are encountered, and analyzing these data sources is crucial for understanding and responding to security incidents. Here are some common types of data encountered in cybersecurity and forensic investigations:

1. **Logs**:

   - ***System Logs***: Records events and activities on a computer or network.

   - ***Application Logs***: Capture information related to specific applications or software.

   - ***Security Logs***: Contain data on security-related events, such as login attempts, firewall activity, and intrusion detection system alerts.

2. **Network Traffic Data**:

   - Packet Captures: Record the raw data of network packets, enabling the analysis of network traffic.

   - ***NetFlow Data***: Summarizes network traffic flows, providing insights into communication patterns.

3. **File System Data**:

   - File Metadata: Information about files, such as creation and modification dates, file size, and permissions.

   - File Content: Actual data within files, including documents, images, and executables.

4. **Memory Dump**:

   - ***Memory Forensics***: Analyzing the contents of a computer's volatile memory (RAM) to identify running processes, open network connections, and other system information.

5. **Registry Data**:

   - ***Windows Registry***: Contains system configuration settings, user profiles, and information about installed software.

6. **Endpoint Data**:

   - ***Endpoint Logs***: Logs generated by individual devices, such as workstations or servers.

   - ***Endpoint Artifacts***: Information left behind by system activities, including temporary files, caches, and artifacts of user interactions.

7. **Cloud-based Data**:

   - ***Cloud Logs***: Logs generated by cloud services, such as AWS CloudTrail or Azure Activity Logs.

   - ***Cloud Configurations:*** Information about cloud infrastructure settings and configurations.

8**. Email Data:**

   - ***Email Headers***: Information about the sender, recipient, and route of an email.

   - ***Email Content:*** The actual text or attachments within emails.

9. **Database Data:**

   - ***Database Logs***: Record database activities, including queries, modifications, and user access.

   - ***Data Records***: Actual content stored within databases.

10. **Incident Response Data**:
    - ***Incident Logs***: Documentation of actions taken during incident response activities.
    - ***Forensic Images***: Bit-for-bit copies of storage media for analysis without altering the original.

11. **User and Authentication Data**:
    - ***User Activity Logs***: Record user actions, login/logout times, and account changes.
    - ***Authentication Records:*** Details about user authentication attempts, successful and failed logins.

Effectively analyzing and correlating these types of data is essential for identifying security incidents, understanding their scope, and developing effective response strategies in cybersecurity and forensic investigations.

Synthetic v/s real-world data

In the field of Data Science, both synthetic and real-world data play important roles, and each has its advantages and disadvantages. Let's explore the differences between synthetic and real-world data:

1. Synthetic Data:

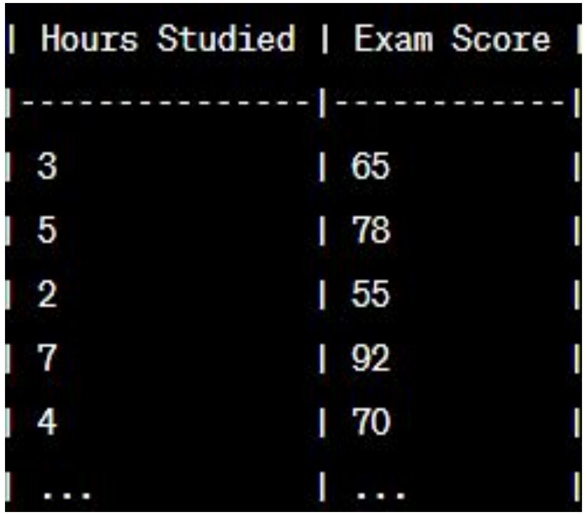
- *Definition*: Synthetic data is artificially generated data that mimics the characteristics of real-world data but is not collected from actual observations or measurements.
- *Advantages*:
    - Privacy: Since synthetic data is not derived from real individuals or entities, it can be used for testing and development without privacy concerns.
    - Scenario Testing: Useful for simulating various scenarios, including extreme or rare events, to evaluate how models perform in different situations.
    - Data Augmentation: Can be used to augment real-world datasets, especially when the available data is limited.
- *Disadvantages*:
    - Lack of Realism: Synthetic data may not fully capture the complexity and nuances present in real-world data.
    - Potential Biases: The process of generating synthetic data may introduce biases if the underlying assumptions or algorithms are flawed.

2. Real-world Data:

- *Definition*: Real-world data is collected from actual observations, measurements, or events in the natural environment.
- *Advantages*:
    - Authenticity: Represents the actual characteristics and patterns present in the real world.
    - Contextual Relevance: Useful for building models that are directly applicable to the specific domain or problem at hand.
    - Variability: Reflects the diversity and variability inherent in real-world situations.
- *Disadvantages*:
    - Privacy Concerns: Real-world data may contain sensitive information, leading to privacy issues.
    - Data Limitations: Availability, quality, and quantity of real-world data can be limiting factors, especially in certain domains.

Example For Synthetic data:

Let's consider a dataset representing the relationship between the number of hours students spend studying and their exam scores. A synthetic dataset for this scenario might look like this:



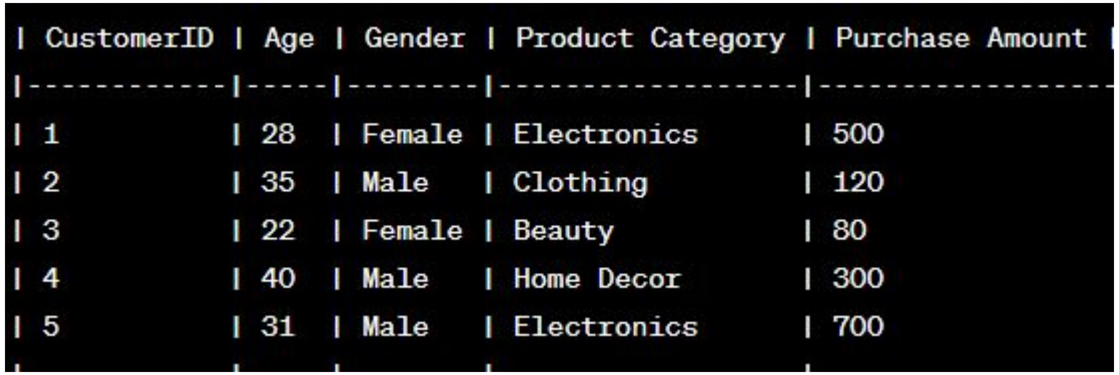
```
| Hours Studied | Exam Score |
|---------------|------------|
| 3             | 65         |
| 5             | 78         |
| 2             | 55         |
| 7             | 92         |
| 4             | 70         |
| ...           | ...        |
```

In this example:

- **Hours Studied:** The independent variable representing the number of hours a student spent studying.
- **Exam Score:** The dependent variable representing the corresponding exam score achieved by the student.

Example For Real-world data:

Consider a dataset from an e-commerce company that tracks customer behavior on their online platform. The dataset might include various features such as:



```
| CustomerID | Age | Gender | Product Category | Purchase Amount |
|------------|-----|--------|------------------|-----------------
| 1          | 28  | Female | Electronics      | 500
| 2          | 35  | Male   | Clothing         | 120
| 3          | 22  | Female | Beauty           | 80
| 4          | 40  | Male   | Home Decor       | 300
| 5          | 31  | Male   | Electronics      | 700
```

Choosing Between *Synthetic* and *Real-world* Data:

- *Use Case*: The choice between synthetic and real-world data often depends on the specific use case and goals of the data science project.
- *Data Availability*: If real-world data is scarce or unavailable, synthetic data can be a valuable resource for model development and testing.
- *Privacy and Ethical Considerations*: When dealing with sensitive information, synthetic data may be preferred to address privacy concerns.
- *Model Evaluation*: Real-world data is crucial for evaluating model performance in the actual conditions it will encounter.

In practice, a combination of both synthetic and real-world data is often used to leverage the strengths of each approach and mitigate their respective limitations.

Synthetic, real-world data in Detailed

What is Synthetic data in Data Science?:

It refers to artificially generated data that mimics the characteristics and patterns of real-world data without being directly obtained from actual observations or measurements. The primary purpose of creating synthetic data is to provide a substitute for or complement real-world datasets, especially in situations where obtaining or using real data is challenging due to privacy concerns, data scarcity, or other limitations.

Qualities of synthetic data in Data Science include:

1. **Generation Methods**:

   - Simulation: Creating data through simulation involves modeling the underlying processes or systems to generate data that approximates the behavior of real-world phenomena.

   - Statistical Methods: Using statistical techniques to generate synthetic data that follows the distribution and statistical properties observed in real datasets.

   - Machine Learning Models: Employing machine learning models, such as generative models like GANs (Generative Adversarial Networks) or autoencoders, to generate data that resembles real-world samples.

2. **Applications**:
   - Privacy Preservation: Synthetic data is often used when dealing with sensitive information to protect individual privacy. It allows researchers and data scientists to perform analyses and model development without exposing real individuals' personal details.
   - Data Augmentation: Synthetic data can be used to augment real-world datasets, especially when the available data is limited. This helps improve the diversity and size of the training dataset for machine learning models.
   - Scenario Testing: Synthetic data is valuable for testing models in various scenarios, including rare or extreme events, to assess their performance under different conditions.

Synthetic Challenges:
- Realism: While synthetic data aims to capture the characteristics of real data, there may be challenges in replicating the full complexity and nuances present in real-world datasets.
- Biases: The generation process may introduce biases if the assumptions or algorithms used in creating synthetic data do not accurately reflect the underlying data distribution.

Synthetic Use Cases:
- Healthcare: Synthetic data can be generated to represent patient records for research and development purposes while ensuring patient privacy.
- Finance: Synthetic financial data can be used for testing and validating algorithms and models without using real financial transactions.
- Cybersecurity: Simulating cyber threats and attacks using synthetic data helps enhance security measures without exposing real vulnerabilities.

What is Real-world data in Data Science?

It refers to data that is collected from actual observations, measurements, or events that occur in the natural environment. It is representative of the characteristics, patterns, and complexities present in the real world. Real-world data is contrasted with synthetic data, which is artificially generated to mimic the characteristics of real data but is not obtained through direct observation or measurement.

- Characteristics of real-world data in Data Science include:

1. Authenticity: Real-world data is genuine and reflects the actual behavior, attributes, and variability inherent in the environment from which it is collected.

2. Contextual Relevance: It is directly related to the specific domain or problem at hand. Real-world data provides the context necessary for building models that can be applied to practical scenarios.

3. Variability: Real-world data often exhibits variability, capturing the diverse patterns and outcomes that may be encountered in different situations.

4. Imperfections: Real-world data may contain noise, outliers, missing values, and other imperfections that are common in actual data collection processes. Dealing with these imperfections is a key challenge in Data Science.

5. Application to Decision-Making: Real-world data is used to make informed decisions, develop predictive models, and gain insights into the underlying patterns and trends in a given domain.

- Real-world data can come from various sources, including:

- Surveys: Data collected through surveys conducted in real-world settings.

- Sensors: Information gathered from sensors, such as those in Internet of Things (IoT) devices, that monitor physical phenomena in the real world.

- Transaction Records: Data from business transactions, financial records, and other real-world interactions.

- Social Media: Information derived from user interactions on social media platforms.

- Health Records: Patient data, medical records, and other healthcare-related information.

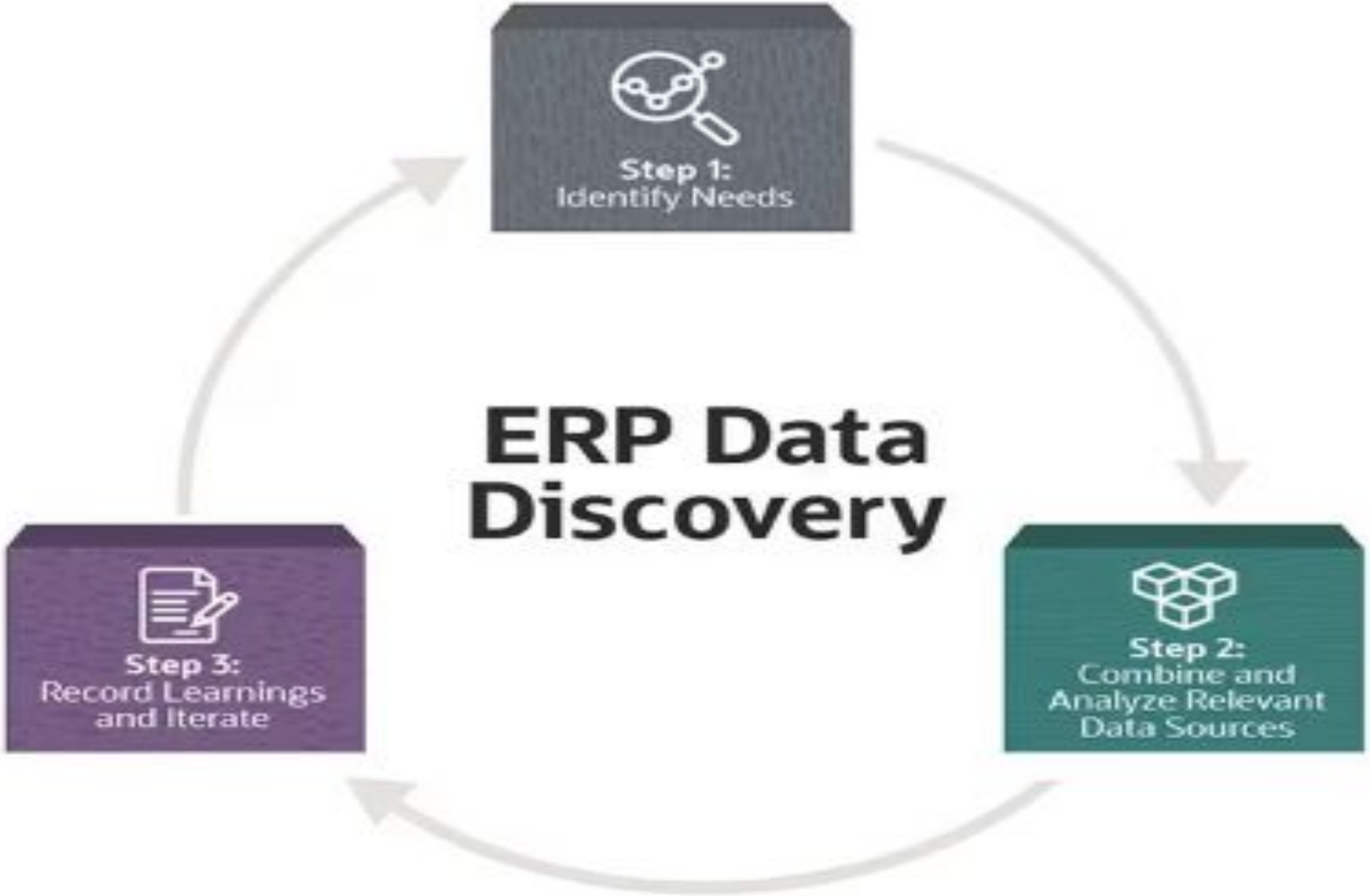
When working with real-world data, data scientists often need to preprocess and clean the data to handle missing values, outliers, and other issues. Additionally, they use statistical and machine learning techniques to extract meaningful insights and build models that can generalize well to new, unseen real-world data.

Data discovery and classification

What Is Data Discovery?

- *Data discovery is a crucial step in the data science process, involving exploring and understanding the dataset before performing more advanced analyses.*
- Data discovery is the process of navigating or applying advanced analytics to data to detect informative patterns that could not have been discovered otherwise. Like a golfer stepping back from the ball to assess the terrain before a putt, data discovery lets businesses take a step back from individual data points, combine data from multiple sources — including external third-party data — and see the big picture, which in turn leads to better decision-making and business strategy. So, when performing data discovery, you may not always know exactly what you’re looking for — you may simply be seeking patterns and outliers to better understand your data.
- Crucially, data discovery does not require business users to build elaborate models. Most companies that use data discovery do so as part of their business intelligence (BI) software, which provides them with a complete view of their organizations in a simple dashboard or visual format.

How Is Data Discovered?



ERP Data Discovery
Step 1: Identify Needs
Step 2: Combine and Analyze Relevant Data Sources
Step 3: Record Learnings and Iterate

How Is Data Discovered?

Data discovery is a five-step process. It is also an iterative process, which means companies can continue to collect, analyze and refine their data discovery approach over time by drawing on their results and feedback from business stakeholders.

- **Step 1: Identify needs.** Effective data discovery begins with a clear purpose, such as the resolution of a pain point. This means considering what kinds of data would be helpful to know, while remaining open to the unexpected insight along the way. For instance, a distributor of fast-moving consumer goods (FMCG) might decide to re-examine its logistics data in an effort to reduce food waste during shipment by 10%. Or a retail bank might analyze its web data with the aim of reducing bounce rates for new prospects.
- **Step 2: Combine data from relevant sources.** For data discovery to be effective, it is important to combine and integrate data from multiple sources because no single data stream tells the complete story. This process is sometimes referred to as data crunching.
- **Step 3: Cleanse and prep the data.** This is the heavy lifting part of data discovery — and a key part of its value. Cleaning the data and preparing it for analysis helps organizations reduce the "noise" in their data and get clearer direction from their data analyses.
- **Step 4: Analyze the data.** With information combined from multiple departments, integrated with external data and cleansed for analysis, business leaders can gain a complete view of their operations and solve the operational riddles that stand in the way of efficiency.
- **Step 5: Record learnings and iterate.** Data discovery is not a one-off process; it is a commitment to continuous improvement.

Author Malcolm Gladwell said it takes people 10,000 hours of practice to master a particular skill — and the same is true of businesses learning to master their data. They must treat data discovery as a way of life with the aim of improving and running more efficiently over time.

Data discovery methods

Similar to processes such as data normalization and competitive analysis, the data discovery process has been greatly improved by the rise of artificial intelligence and automated tools.

-> The two methods of data discovery processes and the reason for the transition from manual to automated discovery.

Manual data discovery

- Before the invention of automated data discovery tools, data specialists were required to spend countless hours manually preparing, mapping, and analyzing data.
- Today, automated data discovery tools and artificial intelligence work together to speed up this process. Manual data discovery involved the monitoring of metadata and data lineage to unpack trends within datasets.
- Extensive knowledge about data categorization and data lineage was required to manually map and organize data during this time.

Automated data discovery

- As mentioned above, the rise of automated data discovery due to the technological advancements in automation and AI has greatly influenced the rise of intelligent data discovery as a necessary practice for long-term business success. Automated data discovery is also often known as performing smart data discovery.
- Intelligent data discovery consists of data mapping specifications, data flow diagrams, data matrices, and other factors that make up a strategic data approach.
- ***Today, AI*** is able to visualize and map data using machine learning algorithms in ways that were not possible before. The AI analyzes data relationships and detects patterns that can provide valuable data-driven insight and accelerate business processes in the company.
- This advancement has also increased the readability of the discovery process making it a more business user-oriented process, not only suitable for data professionals.
- Automation allows sales teams, acquisition experts, and other business users to find relevant data insights.

What is data classification?

- Data classification is the process of separating and organizing data into relevant groups ("classes") based on their shared characteristics, such as their level of sensitivity, the risks they present, and the compliance regulations that protect them. To protect sensitive data, it must be located, classified according to its level of sensitivity, and accurately tagged. Then, enterprises must handle each group of data in ways that ensure only authorized people can gain access, both internally and externally, and that the data is always handled in full compliance with all relevant regulations.
- When done correctly, data classification makes using and protecting data easier and more efficient. However, this process is often overlooked, especially when organizations don't understand its full purpose, scope, and capabilities.

Data classification is the process of organizing data into categories that make it easy to retrieve, sort and store for future use.

- *A well-planned data classification system makes essential data easy to find and retrieve. This can be of particular importance for risk management, legal discovery and regulatory compliance.*

Why classify your data?

Data security and privacy suffer if organizations don’t know their data, including where it lives and how it needs to be protected.

To “know your data” means having understanding where all “sensitive” data is located across an enterprise.

According to Forrester, data privacy professionals, such as Data Privacy Officers (DPO), cannot effectively protect customer, employee, and corporate information if they don’t know the following:

- What data exists across their enterprise
- Where it resides exactly
- Its value and risk to the organization
- Compliance regulations governing the data
- Who is allowed to access and use the data

Data classification delivers this insight by providing a consistent process that identifies and tags all sensitive information wherever it resides across an enterprise — such as in networks, sharing platforms, endpoints, and cloud files. It works by enabling the creation of attributes for data that prescribe how to handle and secure each group according to corporate and regulatory requirements. Because the data is easy to find, organizations can apply protections that lower data exposure risks, reduce the data footprint, eliminate data protection redundancies, and focus security resources on the right actions. In this way, classification both streamlines and strengthens organizations’ data privacy and security protection programs.

Benefits of data classification

- Only 54% of companies know where their sensitive data is stored. This dark data is a serious problem in the fight to keep sensitive data secure, private, and in compliance. By launching comprehensive, well-planned data classification programs, organizations gain a wide range of benefits.

Improve data security

- Data classification enables organizations to safeguard sensitive corporate and customer data by answering the following critical questions:
- What sensitive data do we have (IP, PHI, PII, credit card, etc.)?
- Where does this sensitive data reside?
- Who can access, modify, and delete it?

Knowing the answers to these questions delivers several benefits, including:

- Decrease the sensitive data footprints, thereby, making data security more effective.
- Reduces access to sensitive data to only approved users.
- Understand the criticality of different types of data, so they can be better protected.
- Install the right data protection technologies, such as encryption, data loss prevention (DLP), and identity loss and protection (ILP).
- Optimize costs without wasting resources on non- or less-critical data.

Data classification can be expensive and cumbersome: Few organizations are equipped to handle data classification by traditional (manual) methods. This creates several challenges, including:

- Sensitive data has the potential to become lost in data silos where it is undiscoverable and unprotected.
- Mishandling of sensitive information can result in embarrassment for clients and loss of future revenue.
- Organizations can be fined and penalized for mishandling regulated data.
- Breached client information can spawn lawsuits, tarnish an organization's reputation, and lower goodwill.

Lack understanding of data classification best practices: Poor execution of data classification can result in a cascading series of data security and privacy failures, resulting in these challenges:

- Leadership has an "it won't happen to us" mindset.
- Data and privacy concerns get put in line behind other pressing priorities, such as sales, marketing, expansion, and product expenses.
- Companies don't know how to locate or identify their data.
- Organizations are out of sync with ever-evolving compliance regulations.
- Companies make data classification overly complex, thereby, failing to produce practical results.

Lack of enforcement of data privacy policies: Many organizations have data classification policies that are theoretical rather than operational. In other words, the corporate policy is not enforced, or it's left to business users and data owners to implement.

- The challenge stems from overlooking answering critical questions such as:
- Are inappropriate data privacy discussions happening at the top levels in an organization?
- Who is ultimately responsible for data privacy, and do they have the powers to implement and control solutions?
- Is sensitive and confidential information being shared with other entities?
- Are privacy and compliance policies being circumvented, either deliberately or inadvertently?

Example: NLP data, Image Processing data, Biomedical Image data,

Three(3) types of data classification systems

- There are three options for creating data classification programs:

1. **Manual** — Traditional data classification methods require human intervention and enforcement.
2. **Automated** — Technology-driven solutions eliminate the risks of human intervention, including excessive time and errors, while adding persistence (around-the-clock classification of all data).
3. **Hybrid** — Human intervention provides context for data classification, while tools enable efficiency and policy enforcement.

EXAMPLE: Data classification roles in the enterprise

- Data classification is not one person’s job — it’s everyone’s job. To optimize data classification programs, organizations should designate individuals who will be responsible for carrying out specific duties. For example, Forrester defines data classification roles and responsibilities in six ways.

1. **Data Champions:** A person is responsible for the organization’s use of data for business purposes. They have an incentive to ensure that data is protected and used appropriately. This role can come in different forms, such as a Chief Privacy Office (DLP) who is responsible for data strategy, including quality, governance, and monetization. What’s important is to ensure that an identified business stakeholder will support and drive data classification efforts as a part of the organization’s overall data strategy.
2. **Data Owners:** These are the people ultimately responsible for the data and information collected and maintained by his or her department or division. They are usually a member of senior management, and can also be line-of-business managers, division heads, or the equivalent. If the data resides and is primarily in use within their group, they own it. The aim is for data owners to provide an additional layer of context for classification, such as third-party agreements, which some of today’s automated tools can’t do yet.
3. **Data Creators**: Unless an organization has an automated data classification system, the responsibility of identifying new, freshly created pieces of data (including copies of existing data) as sensitive or not rests with its creator. Anyone within an organization can be a data creator. Data creators can ask themselves one simple question to determine sensitivity: Would it be acceptable for this data to find its way into the public domain or a competitor’s hands? If not, it’s sensitive data and should be appropriately classified.

4. **Data Users:** Anyone who has access to this data is a data user. Data users must use data in a manner consistent with the purpose intended, and comply with this policy, and all policies applicable to data use. Those who have authorization to handle and use the data are in the best position to provide feedback or answer questions about the data classification tags. For example, they can answer questions such as:

-“Is the classification appropriate and based on how the data is used?”

-“Are there circumstances or situations where the data could be handled differently from what’s allowed under the current classification?”

5. Data Auditors

This may be a risk and compliance manager, a privacy officer, a data officer, or an equivalent role. The data auditor reviews the data owner’s assessment of the classification and determines if it’s in line with business partner, regulatory, and other corporate requirements. The data auditor also reviews feedback from data users and assesses alignment between actual or desired data use and current data-handling policies and procedures.

6. Data custodian

IT technicians or information security officers are responsible for maintaining and backing up the systems, databases, and servers that store the organization’s data. In addition, this role is responsible for the technical deployment of all of the rules established by data owners and for ensuring that the rules applied within systems are working.

What is data discovery and classification?

Data discovery and data classification are two separate processes, but they go hand-in-hand to give your organization complete visibility into the data across your entire environment. Before we get into how they work together, let's look at a quick definition of each.

- **Data discovery**

Data discovery is the process of scanning your entire environment to find and identify where both structured and unstructured data resides across your business. This means looking across your entire network, including file servers and hardware, to determine where sensitive and regulated data lives.

- In essence, data discovery allows businesses to identify, classify and track sensitive data so that they have complete visibility into where their data lives. This helps companies to better protect their data, as well as ensure they are meeting regulatory compliance requirements.

- **Data classification**

Data classification is the process of identifying the types of data that a business has discovered, and then tagging that data to organize it into categories based on file type, content and other metadata.

- The process of data classification makes it easier to locate and retrieve sensitive data, as well as eliminate multiple duplications of data. This helps reduce storage and backup costs, increases visibility into where data lives and allows businesses to classify data by the type of regulation it is governed by – making compliance goals easier to achieve.

- **Data discovery and classification go hand-in-hand**

Data discovery and classification work together to give businesses complete visibility into what data they have, where it lives and what policies need to be put in place to protect the data and ensure it complies with data protection regulations. In short, data discovery and classification dramatically improves your data protection and enables your business to implement the controls required to achieve compliance.

Basically:
Data discovery involves identifying and locating data within an organization.
Data classification involves categorizing data based on its sensitivity and importance.

- Purpose: Helps organizations understand and manage their data, particularly in terms of compliance, security, and access controls.
- Tools: Automated discovery tools, metadata management systems.

Outlier

What is an outlier?

- In data analytics, outliers are values within a dataset that vary greatly from the others—they’re either much larger, or significantly smaller. Outliers may indicate variabilities in a measurement, experimental errors, or a novelty.
- In a real-world example, the average height of a giraffe is about 16 feet tall. However, there have been recent discoveries of two giraffes that stand at 9 feet and 8.5 feet, respectively. These two giraffes would be considered outliers in comparison to the general giraffe population.
- When going through the process of data analysis, outliers can cause anomalies in the results obtained. This means that they require some special attention and, in some cases, will need to be removed in order to analyze data effectively.

There are two main reasons why giving outliers special attention is a necessary aspect of the data analytics process:

1. Outliers may have a negative effect on the result of an analysis
2. Outliers—or their behavior—may be the information that a data analyst requires from the analysis

Types of outlier: There are two kinds of outliers:

- A **univariate outlier** is an extreme value that relates to just one variable. For example, Sultan Kösen is currently the tallest man alive, with a height of 8ft, 2.8 inches (251cm). *This case would be considered a univariate outlier as it's an extreme case of just one factor: height.*
- A **multivariate outlier** is a combination of unusual or extreme values for at least two variables. For example, if you're looking at both the height and weight of a group of adults, you might observe that one person in your dataset is 5ft 9 inches tall—a measurement that would fall within the normal range for this particular variable. You may also observe that this person weighs 110lbs. Again, this observation alone falls within the normal range for the variable of interest: weight. However, when you consider these two observations in conjunction, you have an adult who is 5ft 9 inches and weighs 110lbs—a surprising combination. That's a multivariate outlier.

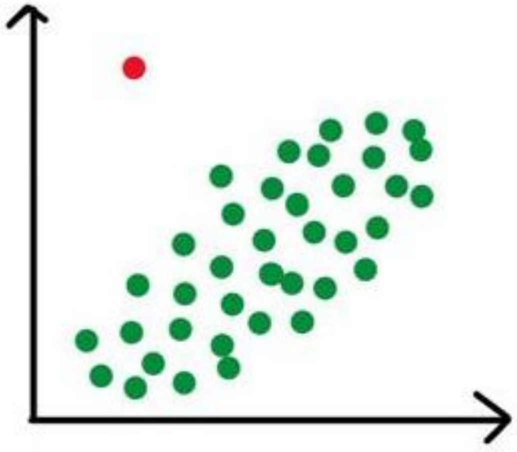
Besides the distinction between univariate and multivariate outliers, you'll see outliers categorized as any of the following:

- **Global outliers (otherwise known as point outliers)** are *single data points that lay far* from the rest of the data distribution.
- **Collective outliers** are seen as a subset of data points that are completely different with respect to the entire dataset.
- **Contextual outliers (otherwise known as conditional outliers)** are *values that significantly deviate from the rest of the data points* in the same context, meaning that the same value may not be considered an outlier if it occurred in a different context. Outliers in this category are commonly found in time series data.

Now we know what an outlier is, let's take a look at how they end up in datasets in the first place.

Global Outliers

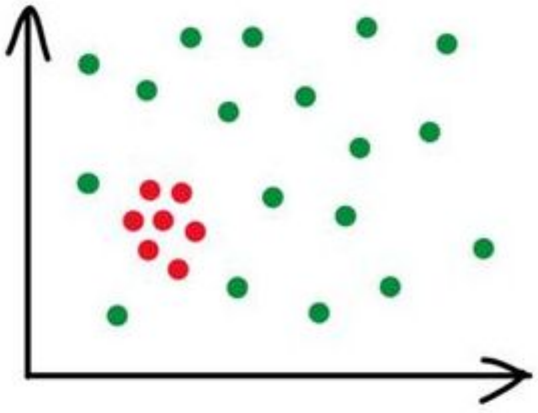
1. Definition: Global outliers are data points that deviate significantly from the overall distribution of a dataset.
2. Causes: Errors in data collection, measurement errors, or truly unusual events can result in global outliers.
3. Impact: Global outliers can distort data analysis results and affect machine learning model performa
4. Detection: Techniques include statistical methods (e.g., z-score, Mahalanobis distance), machine learning algorithms (e.g., isolation forest, one-class SVM), and data visualization techniques.
5. Handling: Options may include removing or correcting outliers, transforming data, or using robust methods.
6. Considerations: Carefully considering the impact of global outliers is crucial for accurate data analy and machine learning model outcomes.



The red data point is a global outlier.

Collective Outliers

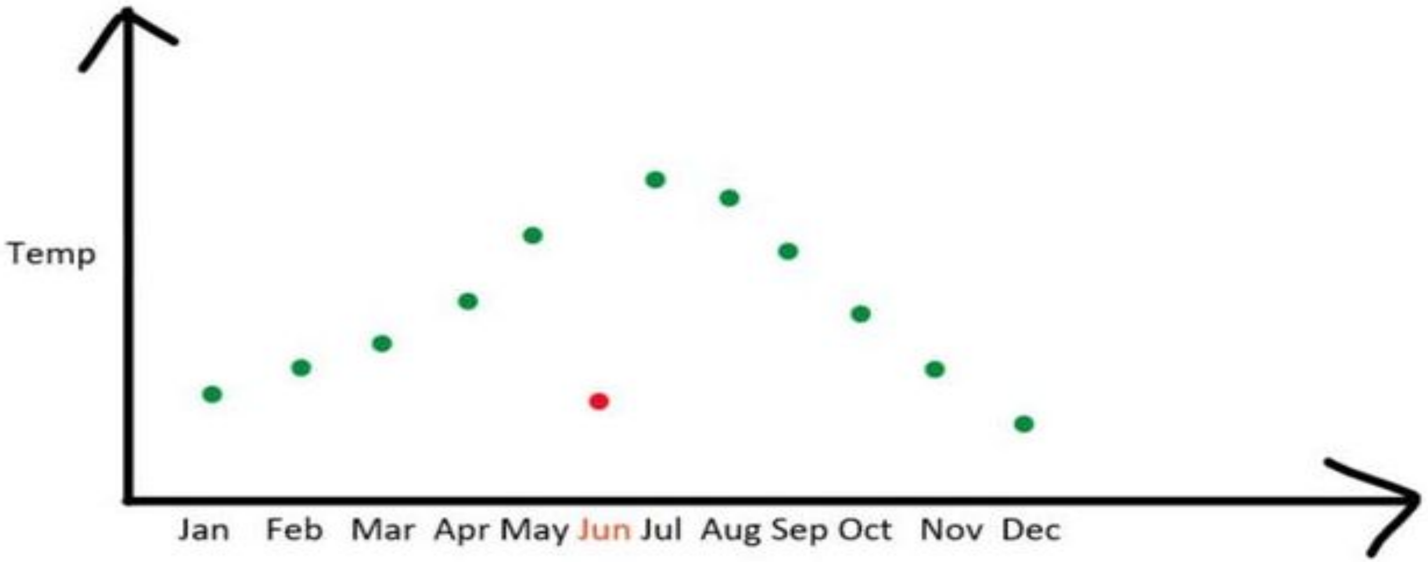
1. Definition: Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.
2. Characteristics: Collective outliers may not be outliers when considered individually, but as a group, they exhibit unusual behavior.
3. Detection: Techniques for detecting collective outliers include clustering algorithms, density-based methods, and subspace-based approaches.
4 Impact: Collective outliers can represent interesting patterns or anomalies in data that may require spec attention or further investigation.
5. Handling: Handling collective outliers depends on the specific use case and may involve further analys of the group behavior, identification of contributing factors, or considering contextual information.
6. Considerations: Detecting and interpreting collective outliers can be more complex than individual outliers, as the focus is on group behavior rather than individual data points. Proper understanding of the data context and domain knowledge is crucial for effective handling of collective outliers.



The red data points as a whole are collective outliers.

Contextual Outliers

- **1. Definition:** Contextual outliers are data points that deviate significantly from the expected behavior within a specific context or subgroup.
  2. Characteristics: Contextual outliers may not be outliers when considered in the entire dataset, but they exhibit unusual behavior within a specific context or subgroup.
  3. Detection: Techniques for detecting contextual outliers include contextual clustering, contextual anomaly detection, and context-aware machine learning approaches.
  4. Contextual Information: Contextual information such as time, location, or other relevant factors are crucial in identifying contextual outliers.
  5. Impact: Contextual outliers can represent unusual or anomalous behavior within a specific context, which may require further investigation or attention.
  6. Handling: Handling contextual outliers may involve considering the contextual information, contextual normalization or transformation of data, or using context-specific models or algorithms.
  7. Considerations: Proper understanding of the context and domain-specific knowledge is crucial for accurate detection and interpretation of contextual outliers, as they may vary based on the specific context or subgroup being considered.



A low temperature value in June is a contextual outlier because the same value in December is not an outlier.

What is the Difference Between Incident Response and Computer Forensics?

- Incident response and computer or cyber forensics both deal with the same issue; they are responses to a compromise, breach, or attack. Incident response is focused on the containment of a threat or attack. Forensics involves a thorough examination of the data in order to gain a complete understanding of the breach in order to remediate the attack and prevent a recurrence.
- **Incident Response**
- Incident response consists of actions taken immediately following a security compromise, attack, or breach. In addition to containing the attack, responders must also preserve all relevant evidence for later examination. This requires a team of experienced professionals who understand how to respond to the incident while carefully preserving evidence.
- Attempting to restore or recover information from a compromised computer or network could cause irreparable damage to files or the system. A dedicated, professional incident response team can handle even the most complex breach events with precision and speed, placing your organization in the best position to mitigate loss and keep the business operational.
- **Digital Forensics**
- Following an attack, there are two important questions to answer: “how did it happen?” and “how to prevent it from happening again?” Digital forensics is the process by which experts collect, examine, and analyze data from compromised computer systems and storage devices in order to answer these questions. This is done carefully, following professional best practices, to ensure that the evidence could be admissible in a court of law if necessary.
- Evidence collection includes identifying and securing infected devices and all data, including latent data, from the systems. Latent or ambient data is data that is not easily accessible (it could be hidden or deleted) and requires an expert to uncover. Once the evidence is collected and evaluated, it undergoes a detailed analysis to determine root cause, scope of breach, and what data may have been impacted. Each step of this process is carefully documented.

How do outliers end up in datasets?

Now that we've learned about what outliers are and how to identify them, it's worthwhile asking: how do outliers end up in datasets in the first place?

Here are some of the more common causes of outliers in datasets:

- ***Human error*** while manually entering data, such as a typo
- ***Intentional errors***, such as dummy outliers included in a dataset to test detection methods
- ***Sampling errors*** that arise from extracting or mixing data from inaccurate or various sources
- ***Data processing errors*** that arise from data manipulation, or unintended mutations of a dataset
- ***Measurement errors*** as a result of instrumental error
- ***Experimental errors***, from the data extraction process or experiment planning or execution

Natural outliers which occur "naturally" in the dataset, as opposed to being the result of an error otherwise listed. These naturally-occurring errors are known as novelties

How can you identify outliers?

- With small datasets, it can be easy to spot outliers manually (for example, with a set of data being 28, 26, 21, 24, 78, you can see that 78 is the outlier) but when it comes to large datasets or big data, other tools are required.
- Usually these outliers can be identify with visualizations or statistical methods, but there are many others available for implementation into your data analytics process. The method that you end up using will depend on the type of dataset you're working with, as well as the tools you're working with.

Incident Reporting

- Incident reporting involves the systematic capture, documentation, and management of those incidents that may affect the operations, safety, or security of a business. *Beyond simply acknowledging and addressing the immediate consequences of an incident*, this approach aims to gather comprehensive data, including information about the circumstances leading up to the event, the individuals involved, the impact on the organization, and any follow-up actions that were taken. By diligently reporting and recording incidents, businesses can learn from past experiences, identify patterns, and implement preventive measures to minimize future risks.
- Incident response is an organized, strategic approach to detecting and managing cyberattacks in ways that minimize damage, recovery time and total costs.

How is incident reporting important in businesses?

*An **incident** can cause business disruption, loss or destruction of sensitive data, litigation, and a tarnished reputation. Resolution without understanding is a great way to end up back where you started. We identify the breach's root cause while simultaneously eliminating unauthorized access and minimizing business interruption.*

Maintaining a productive work environment depends on the business' ability to prepare for and mitigate disruptive incidents. Incident reporting plays a crucial role in making this possible, creating a safe and secure working environment while also helping to foster a culture of continuous improvement within the organization.

This is because incident reporting empowers businesses by:

- **Creating a reliable reporting process**

Incident reporting provides employees with a structured process to report incidents as they occur. *It offers a clear channel for individuals to communicate any **safety** concerns, **accidents**, near **misses**, or **security** breaches they have witnessed or experienced, or that they have reason to anticipate.* By encouraging prompt reporting, businesses can quickly address emerging issues and mitigate potential risks before they escalate into more significant problems.

- **Identifying root causes and establishing preventive measures**

Managing the impact of incidents is not enough; companies need to *understand the origin of the incident* so that it may be addressed. Effective incident reporting helps organizations uncover the underlying factors contributing to disruptive events. *By capturing relevant details—such as the **circumstances**, individuals **involved**, and **potential** contributing factors*—businesses gain valuable insights into root causes. *This information enables them to establish necessary checks, procedures, and risk controls to better address and prevent similar incidents from occurring.*

- **Raising awareness and promoting a proactive workplace**

Communicating *and* ***reviewing*** *incidents through the reporting process raises awareness among employees about potential dangers and threats*. When incidents are shared and discussed within the organization, it fosters a culture of vigilance and responsibility, encouraging individuals to be more proactive in reporting potential threats or other issues they might encounter*. This heightened awareness contributes to a safer and more secure workplace for everyone.*

- **Ensuring regulatory compliance**

Incident reporting supports ***regulatory compliance by providing documentation*** *to ensure that the organization is operating within established legal requirements and industry standards*. It also facilitates the development of improved policies and regulations designed to address specific risks and mitigate potential hazards.

- **Improving transparency and risk assessment**

Incident reporting provides a clear picture of current conditions within an organization. ***By documenting and analyzing incidents, businesses can identify trends, patterns, and areas of concern.*** This data-driven approach to incident reporting enables companies to make informed decisions, improve transparency, and conduct more thorough risk assessments. *Any processes that need to be changed, improved, or eliminated are exposed, enhancing the organization’s overall safety and security.*

Type of incident that can be reported:

Health and safety threats in the workplace should always be major considerations, but they are only one part of the equation; incident reporting must be capable of accounting for a wide range of disruptive occurrences. The various events that organizations are encouraged to capture and document can be classified into the following categories:

- Near misses

Near misses refer to incidents where a potentially harmful event almost occurs but does not result in injury or damage. These close calls serve as valuable learning opportunities, providing insights into potential hazards and vulnerabilities in the workplace. Reporting near misses allows businesses to investigate the underlying causes, implement preventive measures, and avoid future accidents or injuries.

- Exposure incidents

Exposure incidents involve situations where employees *are exposed to hazardous substances, materials, or conditions that could affect their health*. These incidents include ***chemical spills, toxic fumes, biological agents, or other dangerous exposure***s. Reporting exposure incidents is crucial for ensuring timely medical intervention, identifying the source of exposure, and implementing measures to prevent similar incidents.

- Injuries and lost time

Incidents resulting in injuries or lost work time must be fully documented. This category encompasses accidents, physical injuries, and illnesses that affect employees' ability to perform their job duties. Reporting such incidents allows for more immediate medical attention, investigation into the root causes and liability, and implementation of corrective measures to prevent future injuries and maintain productivity.

- Sentinel events

Sentinel events are severe ***incidents that result in significant harm*** (or the potential for serious harm) to individuals or organizations. These events are typically difficult to anticipate and may include major security breaches, catastrophic equipment failures, life-threatening injuries, etc. Reporting sentinel events gives organizations more data to assist in thorough investigation, identify systemic issues, and implement necessary improvements to mitigate future risks.

Proper incident reporting means following an established documentation process to ensure all relevant data is recorded clearly and accurately. When creating an incident report, include the following key elements:

- **Location**
  Clearly specify the location where the incident occurred. This provides context and helps identify any site-specific factors that may have contributed to the incident.
- **Date and time**
  Include the date and time of the incident. This information is crucial for tracking incident trends, analyzing patterns, and identifying potential time-related factors.
- **List of people involved**
  Record the names and roles of individuals directly involved in the incident, including employees, contractors, customers, and any other relevant parties. This helps in subsequent investigations and communication with those affected.
- **Incident description**
  Provide a detailed and objective description of the incident, highlighting the sequence of events leading up to and following the occurrence. Include information such as the nature of the incident, equipment or processes involved, and any other relevant observations.
- **Damage/injuries**
  Document any damage caused to property, equipment, or the environment, and any injuries sustained by the individuals involved. Specify the extent of the damage or injuries to accurately assess the potential impact and associated costs.
- **Contributing factors**
- Identify and document the underlying factors that contributed to the incident. This may include equipment failure, human error, environmental conditions, or other elements. Understanding the root causes will help when it comes time to implement preventive measures.
- **Photos, videos, and other evidence**
  Whenever possible, include visual evidence such as photographs, videos, or diagrams that further clarify the incident report. Visual documentation can provide additional context and assist in investigations or future training initiatives.
- **Immediate action taken**
  Describe the immediate actions taken to address the incident, including any emergency response measures, first aid administered, or initial containment procedures. Prompt action is crucial for mitigating further damage or injuries, and correct documentation of these actions creates a clearer picture of the situation.
- **Recommendations**
  Offer recommendations on how to prevent or respond to similar incidents in the future. These suggestions may include changes to policies, procedures, or training programs, or the introduction of new safety measures. These recommendations should be practical and actionable.

Incident reporting is a critical process, requiring adherence to certain best practices to ensure its effectiveness and reliability. To maximize the value of incident reports, organizations must take steps to ensure that the information included in the report is:

- **Accurate**
  Accurate reporting is crucial for capturing the true nature of the incident. It is important to provide precise information without speculation or assumptions. Accuracy ensures that the incident report reflects the actual events and helps in analysis and decision-making.
- **Clear**
  Visual clarity and easy comprehension make complex incidents more accessible. Utilize clear and concise language, headings, bullet points, and paragraphs to organize the information effectively. Visual aids such as diagrams or flowcharts can also help improve accessibility.
- **Complete**
  A comprehensive incident report should present all the most relevant information. This includes the incident details, individuals involved, contributing factors, actions taken, and recommendations. A complete report promotes a thorough understanding of the incident and facilitates effective next steps.
- **Factual**
  Incident reports should be based on facts rather than personal opinions or subjective interpretations. It is important to present information objectively, relying on tangible evidence, eyewitness accounts, and documented observations to ensure the integrity and reliability of the report.
- **Valid**
  Validity in incident reporting refers to the authenticity and reliability of the information presented. Ensure that the data included in the report is verified, credible, and obtained from reliable sources. Validity instills confidence in the report's findings and recommendations, leading to informed decision-making.

While incident reporting plays an essential role in maintaining and promoting a safe, secure work environment, it is not always as simple as making a few notes and snapping some pictures. Certain challenges can hinder the effectiveness of incident reporting, and it is essential to be aware of these challenges and how to circumvent them. ***Here are some common hurdles to be aware of:***

- Incidents that go unreported, either for fear of blame or because the employee believes the issue is too minor.
- Inconsistency in reporting standards between departments, teams, or individuals.
- Lack of awareness about what kinds of incidents should be reported or what incident reporting processes entail.
- Poor or incomplete training on essential incident reporting processes or standards.
- Lack of an easy-to-use reporting system.
- Overly time-consuming reporting processes.
- Lack of appropriate follow-up after an incident is reported.
- In many of these circumstances, the difficulty comes from insufficient incident reporting tools and support. Having the right systems in place, backed by proper training and clearly established expectations, can help ensure that incident reporting solutions are in place and working properly when emergent situations occur.

.**Summary Incident Reporting:**

- **Definition:** Incident reporting involves documenting and analyzing events that deviate from expected behavior or could impact the security or functionality of a system.
- **Process**: Incident detection, reporting, response, and resolution.
- **Importance**: Enables organizations to respond effectively to security incidents, ensuring continuous improvement in security measures.

Assessing Quality of Data Sets

In data science, data quality assessment (DQA) is the process of applying business-approved requirements to a selected data set. The purpose of DQA is to understand the condition of data in relation to expectations or particular purposes.

Data quality can be assessed based on:

- Accuracy, Completeness, Consistency, Timeliness, Believability, Interpretability.
- Researchers typically assess data quality at both the group level and the individual level. They look for evidence that the data are: Consistent, Correct, Complete, Credible.

Some techniques for data profiling include:

- SQL queries
- Python libraries
- Data visualization tools
- DQA can help identify potential causes of poor data quality. These causes can link to your objectives to improve data quality.

Some other data science approaches to data quality include:

- Data analysis
- Examines data schemas and performs interviews to reach a complete understanding of data and related architecture and management rules
- Requirements analysis
- Surveys the opinion of data users and administrators to identify quality issues and set new quality targets

- Data quality measures the condition of data, relying on factors such as *how useful it is to the specific purpose*, *completeness*, *accuracy*, *timeliness* (e.g., is it up to date?), *consistency*, *validity*, and *uniqueness*.
- Data quality analysts are responsible for conducting data quality assessments, which involve assessing and interpreting every quality data metric. Then, the analyst creates an aggregate score reflecting the data's overall quality and gives the organization a percentage rating that shows how accurate the data is.

- **Accuracy**

The data must conform to actual, real-world scenarios and reflect real-world objects and events. Analysts should use verifiable sources to confirm the measure of accuracy, determined by how close the values jibe with the verified correct information sources.

- **Completeness**

Completeness measures the data's ability to deliver all the mandatory values that are available successfully.

- **Consistency**

Data consistency describes the data’s uniformity as it moves across applications and networks and when it comes from multiple sources. Consistency also means that the same datasets stored in different locations should be the same and not conflict. Note that consistent data can still be wrong.

- **Timeliness**

Timely data is information that is readily available whenever it’s needed. This dimension also covers keeping the data current; data should undergo real-time updates to ensure that it is always available and accessible.

- **Uniqueness**

Uniqueness means that no duplications or redundant information are overlapping across all the datasets. No record in the dataset exists multiple times. Analysts use data cleansing and deduplication to help address a low uniqueness score.

- **Validity**

Data must be collected according to the organization’s defined business rules and parameters. The information should also conform to the correct, accepted formats, and all dataset values should fall within the proper range.

How Do You Improve Data Quality?

People looking for ideas on how to improve data quality turn to data quality management for answers. Data quality management aims to leverage a balanced set of solutions to prevent future data quality issues and clean (and ideally eventually remove) data that fails to meet data quality KPIs (Key Performance Indicators). These actions help businesses meet their current and future objectives.

There is more to data quality than just data cleaning. With that in mind, here are the eight mandatory disciplines used to prevent data quality problems and improve data quality by cleansing the information of all bad data:

- **Data Governance**: Data governance spells out the data policies and standards that determine the required data quality KPIs and which data elements should be focused on. These standards also include what business rules must be followed to ensure data quality.
- **Data Profiling**: Data profiling is a methodology employed to understand all data assets that are part of data quality management. Data profiling is crucial because many of the assets in question have been populated by many different people over the years, adhering to different standards.
- **Data Matching**: Data matching technology is based on match codes used to determine if two or more bits of data describe the same real-world thing. For instance, say there’s a man named Michael Jones. A customer dataset may have separate entries for Mike Jones, Mickey Jones, Jonesy, Big Mike Jones, and Michael Jones, but they’re all describing one individual.
- **Data Quality Reporting**: Information gathered from data profiling, and data matching can be used to measure data quality KPIs. Reporting also involves operating a quality issue log, which documents known data issues and any follow-up data cleansing and prevention efforts.
- **Master Data Management (MDM):** Master Data Management frameworks are great resources for preventing data quality issues. MDM frameworks deal with product master data, location master data, and party master data.
- **Customer Data Integration (CDI):** CDI involves compiling customer master data gathered via CRM applications, self-service registration sites. This information must be compiled into one source of truth.
- **Product Information Management (PIM)**: Manufacturers and sellers of goods need to align their data quality KPIs with each other so that when customers order a product, it will be the same item at all stages of the supply chain. Thus, much of PIM involves creating a standardized way to receive and present product data.
- **Digital Asset Management (DAM):** Digital assets cover items like videos, text documents, images, and similar files, used alongside product data. This discipline involves ensuring that all tags are relevant and the quality of the digital assets.

Data Quality Best Practices

- Data analysts who strive to improve data quality need to follow best practices to meet their objectives. Here are ten critical best practices to follow:
- Make sure that top-level management is involved. Data analysts can resolve many data quality issues through cross-departmental participation.
- Include data quality activity management as part of your data governance framework. The framework sets data policies and data standards, the required roles and offers a business glossary.
- Each data quality issue raised must begin with a root cause analysis. If you don’t address the root cause of a data issue, the problem will inevitably appear again. Don’t just address the symptoms of the disease; you need to cure the disease itself.
- Maintain a data quality issue log. Each issue needs an entry, complete with information regarding the assigned data owner, the involved data steward, the issue’s impact, the final resolution, and the timing of any necessary proceedings.
- Fill data owner and data steward roles from your company’s business side and fill data custodian roles from either business or IT whenever possible and makes the most sense.
- Use examples of data quality disasters to raise awareness about the importance of data quality. However, while anecdotes are great for illustrative purposes, you should rely on fact-based impact and risk analysis to justify your solutions and their required funding.

Big data

Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time. These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.

The amount and availability of data is growing rapidly, spurred on by digital technology advancements, such as connectivity, mobility, the Internet of Things (IoT), and artificial intelligence (AI). As data continues to expand and proliferate, new big data tools are emerging to help companies collect, process, and analyze data at the speed needed to gain the most value from it.

Big data definitions may vary slightly, but it will always be described in terms of volume, velocity, and variety. These big data characteristics are often referred to as the “3 Vs of big data” and were first defined by Gartner in 2001.

In addition to these three original Vs, three others that are often mentioned in relation to harnessing the power of big data: **veracity**, **variability**, and **value**.

- **Veracity**: Big data can be messy, noisy, and error-prone, which makes it difficult to control the quality and accuracy of the data. Large datasets can be unwieldy and confusing, while smaller datasets could present an incomplete picture. The higher the veracity of the data, the more trustworthy it is.
- **Variability:** The meaning of collected data is constantly changing, which can lead to inconsistency over time. These shifts include not only changes in context and interpretation but also data collection methods based on the information that companies want to capture and analyze.
- **Value:** It’s essential to determine the business value of the data you collect. Big data must contain the right data and then be effectively analyzed in order to yield insights that can help drive decision-making.

basis.

- **Velocity**

Big data velocity refers to the speed at which data is generated. Today, data is often produced in real time or near real time, and therefore, it must also be processed, accessed, and analyzed at the same rate to have any meaningful impact.

- **Variety**

Data is heterogeneous, meaning it can come from many different sources and can be structured, unstructured, or semi-structured. More traditional structured data (such as data in spreadsheets or relational databases) is now supplemented by unstructured text, images, audio, video files, or semi-structured formats like sensor data that can’t be organized in a fixed data schema.

How does big data work?

The central concept of big data is that the more visibility you have into anything, the more effectively you can gain insights to make better decisions, uncover growth opportunities, and improve your business model.

Making big data work requires three main actions:

- **Integration:** Big data collects terabytes, and sometimes even petabytes, of raw data from many sources that must be received, processed, and transformed into the format that business users and analysts need to start analyzing it.
- **Management:** Big data needs big storage, whether in the cloud, on-premises, or both. Data must also be stored in whatever form required. It also needs to be processed and made available in real time. Increasingly, companies are turning to cloud solutions to take advantage of the unlimited compute and scalability.
- **Analysis:** The final step is analyzing and acting on big data—otherwise, the investment won't be worth it. Beyond exploring the data itself, it's also critical to communicate and share insights across the business in a way that everyone can understand. This includes using tools to create data visualizations like charts, graphs, and dashboards.

Big data benefits

- Improved decision-making

Big data is the key element to becoming a data-driven organization. When you can manage and analyze your big data, you can discover patterns and unlock insights that improve and drive better operational and strategic decisions.

- Increased agility and innovation

Big data allows you to collect and process real-time data points and analyze them to adapt quickly and gain a competitive advantage. These insights can guide and accelerate the planning, production, and launch of new products, features, and updates.

- Better customer experiences

Combining and analyzing structured data sources together with unstructured ones provides you with more useful insights for consumer understanding, personalization, and ways to optimize experience to better meet consumer needs and expectations.

- Continuous intelligence

Big data allows you to integrate automated, real-time data streaming with advanced data analytics to continuously collect data, find new insights, and discover new opportunities for growth and value.

- More efficient operations

Using big data analytics tools and capabilities allows you to process data faster and generate insights that can help you determine areas where you can reduce costs, save time, and increase your overall efficiency.

- Improved risk management

Analyzing vast amounts of data helps companies evaluate risk better—making it easier to identify and monitor all potential threats and report insights that lead to more robust control and mitigation strategies.

Challenges of implementing big data analytics

While big data has many advantages, it does present some challenges that organizations must be ready to tackle when collecting, managing, and taking action on such an enormous amount of data.

The most commonly reported big data challenges include:

- **Lack of data talent and skills.** Data scientists, data analysts, and data engineers are in short supply—and are some of the most highly sought after (and highly paid) professionals in the IT industry. Lack of big data skills and experience with advanced data tools is one of the primary barriers to realizing value from big data environments.
- **Speed of data growth.** Big data, by nature, is always rapidly changing and increasing. Without a solid infrastructure in place that can handle your processing, storage, network, and security needs, it can become extremely difficult to manage.
- **Problems with data quality.** Data quality directly impacts the quality of decision-making, data analytics, and planning strategies. Raw data is messy and can be difficult to curate. Having big data doesn’t guarantee results unless the data is accurate, relevant, and properly organized for analysis. This can slow down reporting, but if not addressed, you can end up with misleading results and worthless insights.
- **Compliance violations.** Big data contains a lot of sensitive data and information, making it a tricky task to continuously ensure data processing and storage meet data privacy and regulatory requirements, such as data localization and data residency laws.
- **Integration complexity.** Most companies work with data siloed across various systems and applications across the organization. Integrating disparate data sources and making data accessible for business users is complex, but vital, if you hope to realize any value from your big data.
- **Security concerns.** Big data contains valuable business and customer information, making big data stores high-value targets for attackers. Since these datasets are varied and complex, it can be harder to implement comprehensive strategies and policies to protect them.

Big data strategies and solutions

- Developing a solid data strategy starts with understanding what you want to achieve, identifying specific use cases, and the data you currently have available to use. You will also need to evaluate what additional data might be needed to meet your business goals and the new systems or tools you will need to support those.
- Unlike traditional data management solutions, big data technologies and tools are made to help you deal with large and complex datasets to extract value from them. Tools for big data can help with the volume of the data collected, the speed at which that data becomes available to an organization for analysis, and the complexity or varieties of that data.
- For example, data lakes ingest, process, and store structured, unstructured, and semi-structured data at any scale in its native format. Data lakes act as a foundation to run different types of smart analytics, including visualizations, real-time analytics, and machine learning.
- It's important to keep in mind that when it comes to big data—there is no one-size-fits-all strategy. What works for one company may not be the right approach for your organization's specific needs.

Big Data:

- **Definition**: Big Data involves large, complex datasets that traditional data processing methods struggle to handle.
- **Characteristics**: Volume, Velocity, Variety, Veracity, Value (5 Vs).
- **Technologies**: Hadoop, Apache Spark, NoSQL databases.

As example: four key concepts that our Google Cloud customers have taught us about shaping a winning approach to big data:

Open
Today, organizations need the freedom to build what they want using the tools and solutions they want. As data sources continue to grow and new technology innovations become available, the reality of big data is one that contains multiple interfaces, open source technology stacks, and clouds. Big data environments will need to be architected to be both open and adaptable to allow for companies to build the solutions and get the data it needs to win.

Intelligent
Big data requires data capabilities that will allow them to leverage smart analytics and AI and ML technologies to save time and effort delivering insights that improve business decisions and managing your overall big data infrastructure. For example, you should consider automating processes or enabling self-service analytics so that people can work with data on their own, with minimal support from other teams.

Flexible
Big data analytics need to support innovation, not hinder it. This requires building a data foundation that will offer on-demand access to compute and storage resources and unify data so that it can be easily discovered and accessed. It’s also important to be able to choose technologies and solutions that can be easily combined and used in tandem to create the perfect data toolsets that fit the workload and use case.

Trusted
For big data to be useful, it must be trusted. That means it’s imperative to build trust into your data—trust that it’s accurate, relevant, and protected. No matter where data comes from, it should be secure by default and your strategy will also need to consider what security capabilities will be necessary to ensure compliance, redundancy, and reliability

Cloud Computing

Cloud computing refers to the delivery of computing services, including storage, processing power, and applications, over the internet. Instead of relying on local servers or personal devices to handle data and applications, cloud computing allows users to access and use these resources through remote servers hosted on the internet.

Some characteristics of cloud computing include:
1. ***On-Demand Service****:* Users can access computing resources as needed, paying only for the resources they use. This eliminates the need for significant upfront investments in hardware and infrastructure.
2. ***Broad Network Access***: Cloud services are accessible over the internet from a variety of devices, such as laptops, smartphones, and tablets.
3***. Resource Pooling:*** Cloud providers pool and manage computing resources to serve multiple customers, enabling efficient resource utilization and cost savings.
4. ***Rapid Elasticity***: Cloud resources can be quickly scaled up or down based on demand. This allows organizations to adapt to changing workloads and requirements.
5. ***Measured Service***: Cloud computing resources are metered, and users are billed based on their usage. This pay-as-you-go model is cost-effective and allows organizations to optimize their spending.

Cloud computing is often categorized into three service models:
1. ***Infrastructure as a Service (IaaS)***: Provides virtualized computing resources over the internet. Users can rent virtual machines, storage, and networking components.
2***. Platform as a Service (PaaS):*** Offers a platform that includes tools and services for application development, such as databases, development frameworks, and hosting.
3. ***Software as a Service (SaaS):*** Delivers fully functional software applications over the internet. Users can access these applications without worrying about the underlying infrastructure.

Popular cloud service providers include Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and many others, each offering a range of services to meet various computing needs.

Types of cloud computing

- **Public Cloud**

Private cloud is a cloud environment in which all cloud infrastructure and computing resources are dedicated to, and accessible by, one customer only. Private cloud combines many of the benefits of cloud computing—including elasticity, scalability, and ease of service delivery—with the access control, security, and resource customization of on-premises infrastructure.

- **Private cloud**

Private cloud is a cloud environment in which all cloud infrastructure and computing resources are dedicated to, and accessible by, one customer only. Private cloud combines many of the benefits of cloud computing—including elasticity, scalability, and ease of service delivery—with the access control, security, and resource customization of on-premises infrastructure.

- **Hybrid cloud**

Hybrid cloud is just what it sounds like—a combination of public and private cloud environments. Specifically, and ideally, a hybrid cloud connects an organization's private cloud services and public clouds into a single, flexible infrastructure for running the organization’s applications and workloads.

Summary Cloud Computing:

- **Definition**: Cloud Computing delivers computing services over the internet, providing on-demand access to resources.
- **Benefits**: Scalability, Cost-efficiency, Accessibility.
- **Services**: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS).

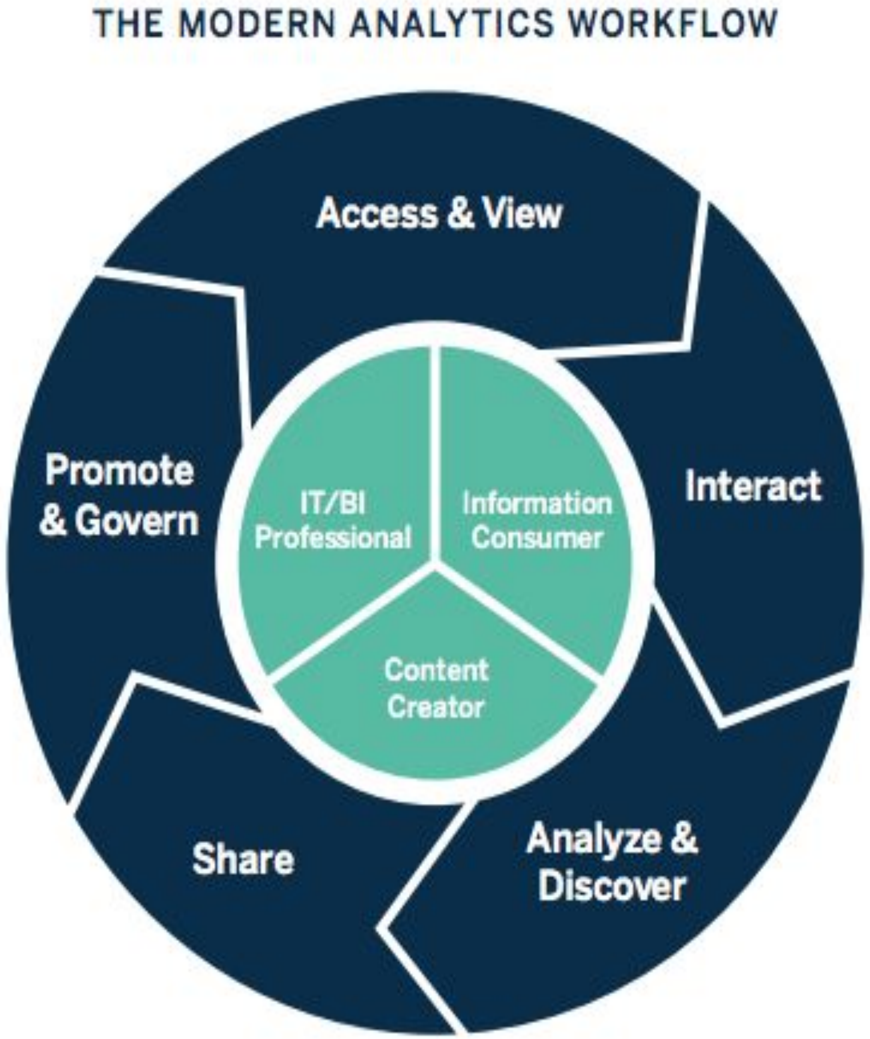
Business Intelligence

- Business Intelligence (BI) refers to technologies, processes, and tools that help organizations collect, analyze, and present business data to support decision-making. The goal of business intelligence is to provide insights into business operations, trends, and performance, enabling organizations to make informed and strategic decisions.
- Business intelligence combines business analytics, data mining, data visualization, data tools and infrastructure, and best practices to help organizations make more data-driven decisions.
- The term Business Intelligence was coined in 1989, alongside computer models for decision making. These programs developed further, turning data into insights before becoming a specific offering from BI teams with IT-reliant service solutions.

How business intelligence works

Businesses and organizations have ***questions*** and goals. To ***answer*** these questions and ***track performance against these goals***, they gather the necessary data, analyze it, and determine which actions to take to reach their goals.

- On the technical side, raw data is collected from business systems.
- Data is processed and then stored in data warehouses, the cloud, applications, and files.
- Once it’s stored, users can access the data, starting the analysis process to answer business questions.
- BI platforms also offer data visualization tools, which convert data into charts or graphs, as well as presenting to any key stakeholders or decision-makers.



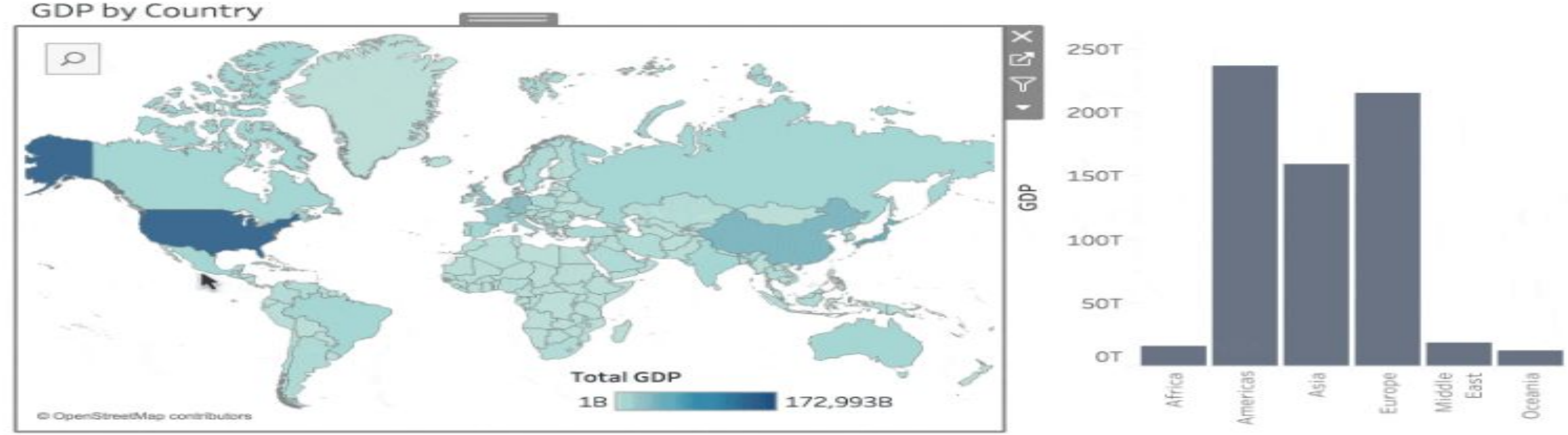
BI methods

Much more than a specific “thing,” business intelligence is an umbrella term that covers the processes and methods of collecting, storing, and analyzing data from business operations or activities to optimize performance. All of these things come together to create a comprehensive view of a business to help people make better, actionable decisions. Over the past few years, business intelligence has evolved to include more processes and activities to help improve performance. These processes include:

- **Data mining:** Using databases, statistics, and machine learning (ML) to uncover trends in large datasets
- **Reporting:** Sharing data analysis to stakeholders so they can draw conclusions and make decisions
- **Performance metrics and benchmarking:** Comparing current performance data to historical data to track performance against goals, typically using customized dashboards
- **Descriptive analytics:** Using preliminary data analysis to find out what happened
- **Querying:** Asking the data-specific questions, BI pulling the answers from the data sets
- **Statistical analysis:** Taking the results from descriptive analytics and further exploring the data using statistics such as how this trend happened and why
- **Data visualization:** Turning data analysis into visual representations such as charts, graphs, and histograms to more easily consume data
- **Visual analysis:** Exploring data through visual storytelling to communicate insights on the fly and stay in the flow of analysis
- **Data preparation:** Compiling multiple data sources, identifying the dimensions and measurements, and preparing it for data analysis

The difference between traditional BI and modern BI

- Historically, business intelligence tools were based on a traditional business intelligence model. This was a top-down approach where business intelligence was driven by the IT organization and most, if not all, analytics questions were answered through static reports.
- Traditional business intelligence is still a common approach for regular reporting and answering static queries. However, modern business intelligence is interactive and approachable. While IT departments are still an important part of managing access to data, multiple levels of users can customize dashboards and create reports on little notice. With the proper software, users are empowered to visualize data and answer their own questions.



Modern BI prioritizes self-service analytics and speed to insight.

Some of the top business intelligence benefits include:

- Data clarity
- Increased efficiency
- Better customer experience
- Improved employee satisfaction

Business Intelligence (BI) Summary:

- **Definition**: Business Intelligence involves technologies, processes, and tools for transforming raw data into meaningful insights.
- **Components**: Data warehousing, reporting, dashboards, data mining.
- **Purpose**: Supporting business decision-making processes.

Data security strategies

What Is Data Security?

- Data security is the process of assessing and implementing controls to protect digital assets and reduce risk.
- Digital assets may include databases, files, accounts, and other information that is sensitive or critical to operations.

Data security refers to the practices and measures implemented to protect data from unauthorized access, disclosure, alteration, destruction, and other threats. It is a critical aspect of information security and is essential for safeguarding sensitive and confidential information. Data security involves various strategies, technologies, and policies to ensure the confidentiality, integrity, and availability of data.

What Is a Data Protection Strategy?

- A data protection strategy is an organized effort that includes all the measures implemented for the purpose of protecting data in the organization. A data protection strategy can help organizations standardize the security of sensitive data and corporate information, ensuring the privacy of customers and employees and the security of trade secrets.
- Data protection strategies typically involve multi-step processes that define how security measures are implemented and maintained. The goal is to minimize the footprint of sensitive data and secure business-critical and regulated data. In particular, organizations use data protection strategies to prevent threat actors from gaining unauthorized access to data.

A successful data protection strategy will typically include the following components.

1. Access Control:
   - Strategy: Implement strong access controls, including authentication mechanisms like multi-factor authentication (MFA) and role-based access control (RBAC). Limit access to data to only those who need it for their roles, and regularly review and update access permissions.
2. Encryption:
   - Strategy: Use encryption for data both in transit (during communication) and at rest (when stored). This helps ensure that even if unauthorized access occurs, the data remains unreadable without the proper decryption keys.
3. Regular Software Updates and Patch Management:
   - Strategy: Keep software, operating systems, and applications up to date by applying security patches promptly. Regularly assess and update systems to address vulnerabilities and reduce the risk of exploitation.
4. Firewalls and Network Security:
   - Strategy: Deploy firewalls to monitor and control network traffic. Use intrusion detection and prevention systems to detect and respond to potential security threats. Segment networks to limit the potential impact of a security breach.

5. Data Backups:
   - Strategy: Implement a robust backup and recovery strategy. Regularly back up critical data and test the restoration process to ensure that data can be recovered in the event of data loss or a security incident.
6. Security Awareness Training:
   - Strategy: Educate employees about security best practices, the importance of protecting sensitive information, and how to recognize and report potential security threats. Foster a security-aware culture within the organization.
7. Data Loss Prevention (DLP):
   - Strategy: Deploy DLP solutions to monitor and control the flow of sensitive data within the organization. Implement policies to prevent unauthorized access, sharing, or leakage of sensitive information.
8. Incident Response Plan:
   - Strategy: Develop and regularly test an incident response plan that outlines the steps to be taken in the event of a security incident. This helps ensure a coordinated and effective response to mitigate the impact of a breach.
9. Physical Security:
   - Strategy: Secure physical access to data centers, servers, and other hardware that store sensitive information. Implement measures such as access control systems, surveillance, and environmental controls to protect physical infrastructure.
10. Vendor Risk Management:
    - Strategy: Assess and manage the security risks associated with third-party vendors and service providers. Ensure that they adhere to security best practices and have adequate measures in place to protect the data they handle.
11. Data Classification:
    - Strategy: Classify data based on its sensitivity and importance. Apply appropriate security controls and protection measures based on the classification. This helps prioritize security efforts and resources.

12. Regulatory Compliance:
- Strategy: Stay informed about relevant data protection regulations and compliance requirements. Ensure that data security measures align with legal and industry-specific standards to avoid legal consequences and reputational damage.

13. Employee Offboarding Procedures:
- Strategy: Develop and implement procedures for securely managing access to data when employees leave the organization. This includes revoking access credentials and ensuring the return or deletion of company devices.

14. Endpoint Security:
- Strategy: Implement endpoint protection measures, including antivirus software, endpoint detection and response (EDR) solutions, and device encryption, to secure devices such as laptops, desktops, and mobile devices.

15. Continuous Monitoring:
- Strategy: Implement continuous monitoring of network and system activities. Use security information and event management (SIEM) tools to detect and respond to security incidents in real-time.

By adopting a comprehensive and multi-layered approach to data security, organizations can significantly reduce the risk of data breaches and unauthorized access. Regularly reassessing and updating these strategies in response to evolving threats is essential for maintaining robust data security.

Data Security Strategies:
- **Encryption**: Protecting data by converting it into a secure format.
- **Access Controls**: Managing who has access to data and what actions they can perform.
- **Authentication** and Authorization: Verifying user identities and determining their permissions.
- **Data Masking:** Protecting sensitive information by replacing, encrypting, or scrambling original data.
- **Monitoring and Auditing**: Keeping track of data access and changes for security auditing.

Clean unstructured data

Why clean unstructured data?

- Unstructured data can contain a lot of noise, errors, inconsistencies, duplicates, missing values, or irrelevant information that can affect the quality and reliability of your analysis and models.
- Cleaning unstructured data can ***help you reduce the*** size, complexity, and ambiguity of your data, as well as improve its accuracy, completeness, and usability.
- Cleaning unstructured data can also ***help you extract useful*** features, insights, and patterns from your data, as well as prepare it for further processing, such as tokenization, normalization, encoding, or embedding.
- Cleansing unstructured data ***involves*** preparing and transforming raw, unorganized data into a structured and usable format. Unstructured data can include text, images, audio, video, and other forms of information that lack a predefined data model.

Cleansing unstructured data

Here are steps to help you cleanse unstructured data:

1. Define Data Quality Goals:
   - Clearly define your data quality goals and requirements. Identify the specific aspects of data quality that are important for your analysis or application, such as accuracy, completeness, consistency, and relevance.
2. Data Profiling:
   - Conduct data profiling to understand the characteristics of your unstructured data. This involves analyzing the content, structure, and patterns within the data to identify potential issues and areas for improvement.
3. Text Parsing and Tokenization:
   - For textual data, use natural language processing (NLP) techniques to parse and tokenize the text into meaningful units, such as words or phrases. This step is crucial for extracting relevant information and identifying patterns.
4. Remove Redundant or Irrelevant Information:
   - Eliminate unnecessary or redundant information from the unstructured data. This could include removing duplicate records, irrelevant sections, or noise that does not contribute to the analysis.
5. Normalization:
   - Normalize data to ensure consistency. This involves standardizing formats, units, and representations of data. For example, converting dates to a consistent format or standardizing text capitalization.
6. Entity Recognition:
   - Use entity recognition tools or techniques to identify and categorize entities within the unstructured data. This could include recognizing names, locations, dates, and other relevant information.
7. Remove Special Characters and Formatting:
   - Cleanse the data by removing special characters, formatting inconsistencies, and any other artifacts that may hinder analysis. This step helps ensure that the data is consistent and easy to work with.

8. Data Enrichment:
- Enhance the data by adding additional information from external sources. This could involve linking unstructured data to structured datasets, providing context and additional details for analysis.

9. Handle Missing Data:
- Address missing or incomplete data. Depending on the nature of the missing data, you may choose to impute values, remove incomplete records, or take other appropriate actions to ensure data completeness.

10. Quality Assurance:
- Implement quality assurance checks to validate the cleansed data. This may involve running validation scripts, performing statistical analyses, or comparing the data against predefined quality metrics.

11. Document the Cleansing Process:
- Document the steps taken during the cleansing process. This documentation is essential for transparency, reproducibility, and collaboration, especially if multiple individuals are involved in the data cleansing process.

12. Iterative Process:
- Data cleansing is often an iterative process. Continuously review and refine your cleansing techniques based on the results and feedback. Regularly revisit your data quality goals and make adjustments as needed.

13. Data Governance:
- Establish data governance practices to ensure that data quality is maintained over time. Implement policies, procedures, and responsibilities for managing and monitoring data quality.

14. Use Data Quality Tools:
- Leverage data quality tools and software that automate and streamline the cleansing process. These tools may include features for data profiling, cleansing, enrichment, and monitoring.

Cleansing unstructured data can be ***challenging***, but by following these steps and leveraging appropriate tools and techniques, you can improve the quality and usability of your data for analysis, reporting, and decision-making.

How to clean unstructured data from text?

- Text data can be sourced from various sources, such as web pages, documents, emails, tweets, reviews, or comments. It can have different formats, languages, styles, and tones; and may contain spelling mistakes, grammar errors, slang, abbreviations, emoticons, or hashtags.
- To clean unstructured data from text, you can use techniques such as removing unwanted characters with regular expressions or string methods; converting text to lowercase or uppercase; removing stop words with predefined lists or libraries; performing stemming or lemmatization to reduce words to their root form; correcting spelling or grammar errors using dictionaries or libraries; detecting and translating languages with APIs or libraries; and identifying and extracting entities with named entity recognition (NER) techniques or libraries.

How to clean unstructured data from images?

- Image data can come from various sources, such as cameras, scanners, websites, or social media platforms, and can have different formats, resolutions, orientations, colors, or qualities. It may also contain noise, artifacts, watermarks, or text.
- To clean unstructured data from images, you can use techniques such as resizing or cropping to adjust their dimensions and aspect ratios; rotating or flipping to correct their orientations and perspectives; converting to grayscale or color; enhancing or adjusting to improve contrast, brightness, sharpness, or saturation; filtering or smoothing to remove noise, blur, or edges; segmenting or masking to isolate or remove specific regions or objects; and detecting and extracting text using optical character recognition (OCR) techniques. Cleaning unstructured data from text or images can be a challenging and time-consuming task but is worth the effort as it can transform your data into a clean and structured form that can enhance your data science project and portfolio.

Cleansing Unstructured Data Summary:

- Challenges: Unstructured data (text, images, videos) lacks a predefined data model.
- Techniques: Natural Language Processing (NLP), Text Mining, Image Processing.
- Tools: Open-source tools like NLTK (Natural Language Toolkit) for NLP

Cleansing unstructured data in Python typically involves a combination of techniques for cleaning and pre-processing text data. Here's a general approach you can use:

1. **Tokenization**: Splitting the text into individual words or tokens.
2. **Lowercasing**: Converting all words to lowercase to ensure uniformity.
3. **Removing special characters**: Stripping out punctuation, symbols, and other non-alphanumeric characters.
4. **Removing stop words**: Removing common words (e.g., "the", "is", "and") that may not be relevant to the analysis.
5. **Lemmatization or stemming**: Reducing words to their base form to normalize variations (e.g., "running" to "run").
6. **Handling missing data**: Dealing with any missing or null values in the text data.

Question and Discussion

Thank You