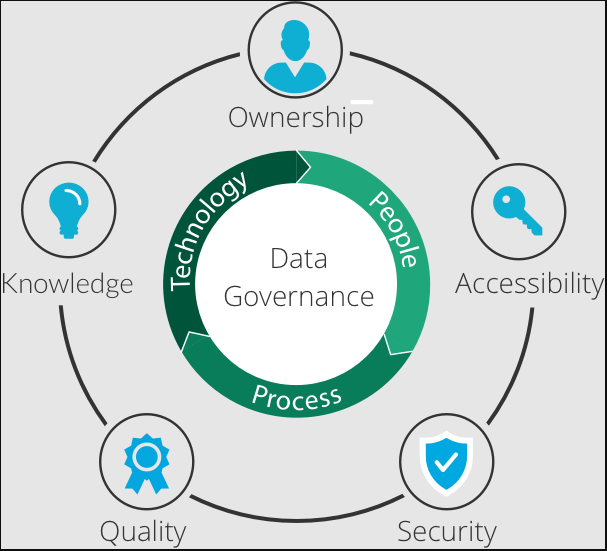
**Data Governance**

**What is data governance?**

Data governance is a comprehensive approach that comprises the principles, practices and tools to manage an organization’s data assets throughout their lifecycle. By aligning data-related requirements with business strategy, data governance provides superior data management, quality, visibility, security and compliance capabilities across the organization. Implementing an effective data governance strategy allows companies to make data easily available for data-driven decision-making while safeguarding their data from unauthorized access, and ensuring compliance with regulatory requirements.



* **What are the business benefits of data governance?**

Data governance is essential for unlocking the value of data, which is a critical asset for organizations. By implementing a robust data governance approach, businesses can leverage their data assets, gain a competitive edge, and earn and maintain customer trust by ensuring sound data and privacy practices.

1. **Increased operational efficiency and reduced costs**

Effective data governance allows organizations to create a single source of truth for their data estate, preventing data sprawl and silos, and reducing duplication. This leads to increased efficiency, reduced costs and easier management of security and governance concepts across the data estate.

1. **Improved productivity and faster decision-making**

Data governance promotes data democratization by ensuring data accuracy, consistency, and trustworthiness. It helps data users find high-quality data quickly, promoting a better understanding of the data’s meaning and context, leading to increased productivity and faster decision-making.

1. **Enhanced collaboration and value realization**

A strong data governance program lays the foundation for enhanced data collaboration and sharing across teams, business units, and partners. This helps organizations promote knowledge sharing and build a better data culture, leading to increased innovation, better decision-making, and maximizing the value of their data.

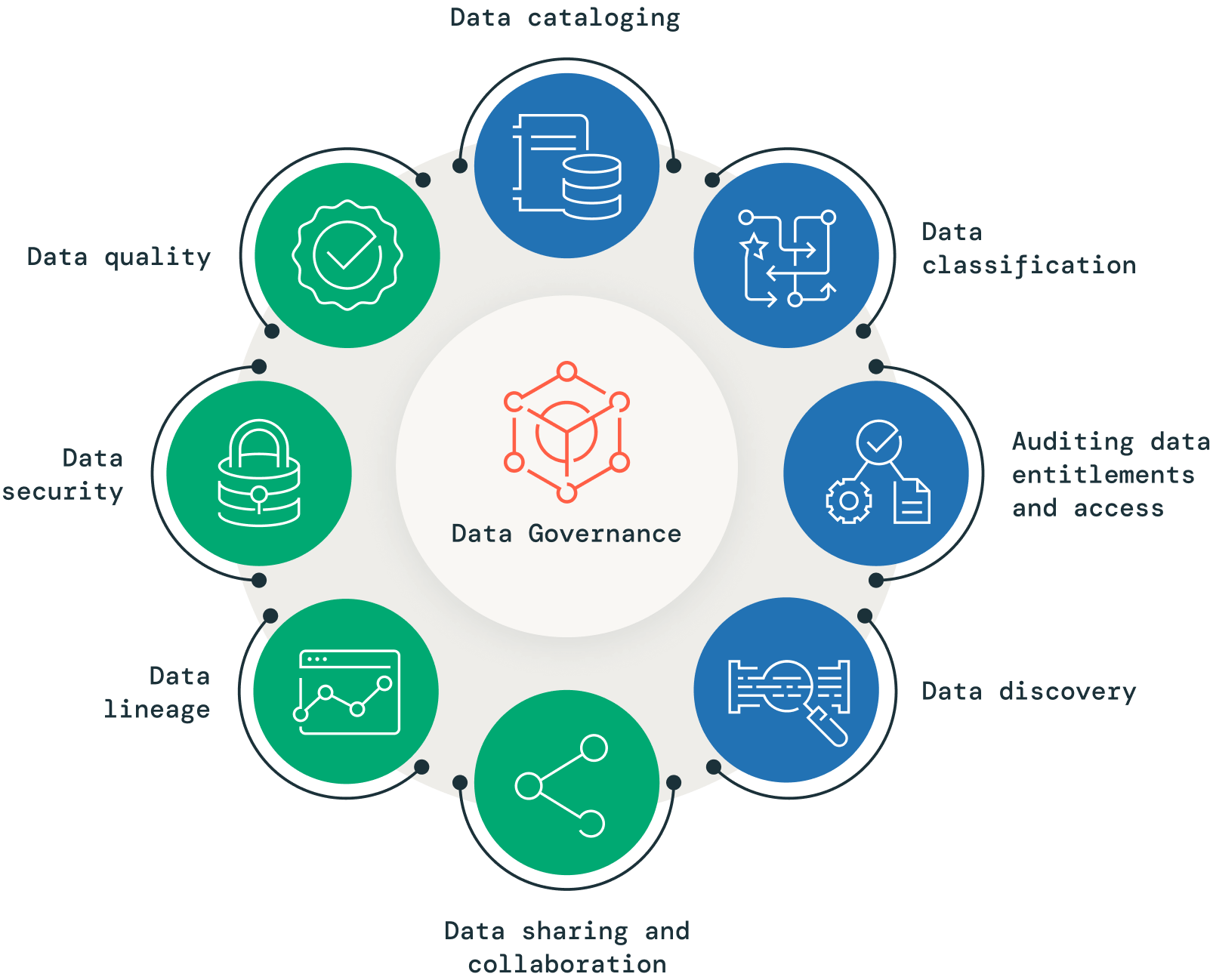
1. **Enhanced security and privacy**

Data governance mitigates security and privacy risks by implementing controls and processes to prevent unauthorized access and misuse of sensitive data. It promotes a culture of trust and transparency with stakeholders.

1. **Better compliance with regulations and standards**

Effective data governance results in better compliance with regulatory requirements, such as **HIPAA, FedRAMP, GDPR or CCPA**. This protects the organization’s reputation, avoids potential financial and legal consequences, and increases stakeholder trust.

* **Key elements of data governance -**

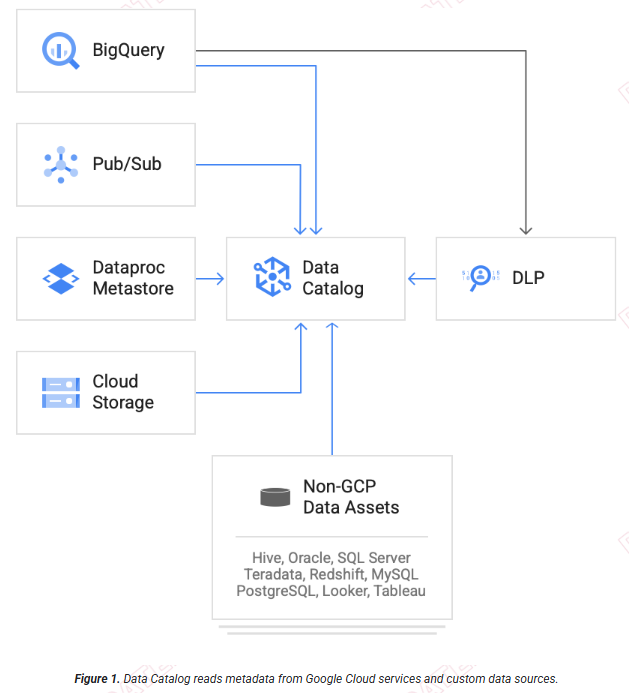
****



**Data Cataloging –**

Effective data governance requires knowledge of the data that exists within an organization. This is where a data catalog comes in, as it provides a centralized metadata repository for an organization’s data assets. A data catalog allows stakeholders to quickly discover, understand and access the data they need, improving data-related activities such as discovery, governance and analytics. It acts as a searchable index of all the data available, including information about its format, structure, location and usage, providing semantic value to an otherwise unidentifiable sea of information. Incorporating a data catalog into a governance program can help organizations improve their data management, enhance collaboration, reduce redundancy and ensure proper access controls and audit information retrieval.

* **Need for Data Catalog –** 
  + ***Searching for insightful Data:*** 
    - Data consumers don't know the location and origin of data. They have to navigate data "swamps".
    - Data consumers don't know what data to use to get insights because most data isn't well documented and, even if documented, isn't well maintained.
    - Data can't be found and is often lost when it resides only in people's min
  + ***Understanding data:***
    - Is the data fresh, clean, validated, approved for use in production?
    - Which dataset out of several duplicate sets is relevant and up-to-date?
    - How does one dataset relate to another?
    - Who is using the data and who is the owner?
    - Who and what processes are transforming the data?
  + ***Making data useful:***
    - Data producers don't have an efficient way to put forward their data for consumers. If there's no self-service, consumers may overwhelm producers. Several data engineers can't manually provide data to thousands of data analysts.
    - Valuable time is lost if data consumers have to find out how to request data access, wait without a defined response time, escalate, and wait again.
* **How Data Catalog Works –**

****



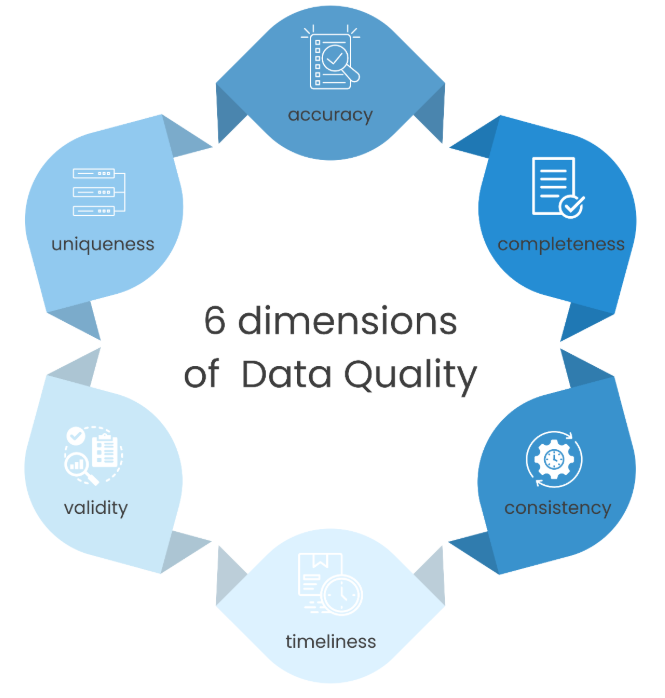
**Data Quality –**

In today’s data-driven world, ensuring high data quality is crucial for accurate analytics, informed decision-making and cost-effectiveness. Data quality directly impacts the reliability of data-driven decisions and is a key aspect of data governance. To maintain effective data governance, organizations must prioritize the evaluation of key data quality attributes such as accuracy, completeness, freshness and compliance with data-quality rules. Therefore, a strong focus on data quality is essential in any data governance strategy, as it helps trace data lineage, enforce data quality rules, and track changes. Don’t let poor data quality compromise your business decisions and resource allocation — prioritize data quality as a critical part of your data governance efforts for better outcomes.

* **Data Quality Process –**



* **Data Quality main 6 Diminutions –**

****

**1. Accuracy**

Accuracy refers to the correctness of the data. It is considered accurate when it reflects real-world values or attributes it is supposed to represent. Inaccurate data can lead to incorrect conclusions and decisions.

* **Example:** In a sales dataset, accuracy means that every value in the “amount” column equals the actual amount charged to the customer.
* **Improvement:** Verify data against authentic references or actual entities to ensure it maintains accuracy throughout its lifecycle.

**2. Completeness**

Completeness assesses the presence of all required data elements within the dataset, without missing or omitted values. Data should be comprehensive and contain all necessary information for its intended use.

* **Example:** A salesperson requires customer contact information to follow up. Missing phone numbers make the data incomplete.
* **Improvement:** Set standards for the minimum information required and conduct regular checks to ensure completeness.

**3. Consistency**

Consistency is about the absence of conflicts or differentiations between different data elements, either between different datasets or within a dataset. Data should be consistent both within itself and with other related data.

* **Example:** If an employee’s termination date differs between HR and payroll systems, the data is inconsistent.
* **Improvement:** Regular audits and reconciliation processes can help maintain consistency across datasets.

**4. Timeliness**

Timeliness refers to the relevance and currency of data, ensuring it is up-to-date and suitable for decision-making. Timely data is updated and important for specific use cases. Outdated data can be misleading and result in poor decision-making.

* **Example:** Real-time fraud detection systems need timely data updates to function effectively.
* **Improvement:** Implement real-time or scheduled data update mechanisms to ensure the data is timely.

**5. Validity**

Validity ensures that data adheres to rules or standards, confirming its appropriateness for the intended purpose. Valid data conforms to predefined formats, values, and ranges essential for ensuring its utility and reliability.

* **Example:** Postcodes must adhere to a specific numeric format to be considered valid.
* **Improvement:** Regular validation checks and adherence to industry standards are essential for maintaining validity.

**6. Uniqueness**

Uniqueness assesses whether a dataset contains duplicate records or not. It involves identifying and eliminating repeated or redundant information, making sure that each data entry is distinct.

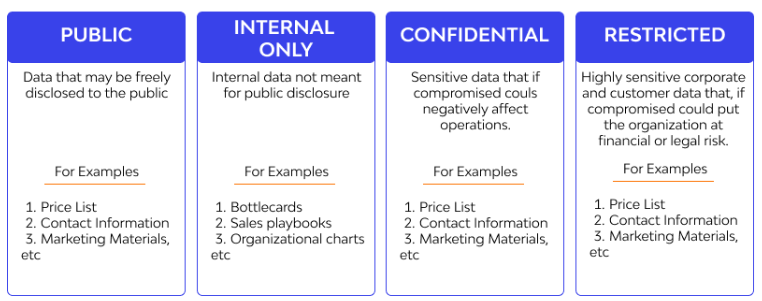
* **Example:** Duplicate customer records can lead to incorrect conclusions in analysis.

**Improvement:** Identifying overlaps and using data cleansing and deduplication techniques to maintain uniqueness.

** Data Classification –**

Data classification is a crucial part of data governance that involves organizing and categorizing data based on its sensitivity, value and criticality. With the exponential growth of data, businesses are increasingly concerned about protecting sensitive data, mitigating risks and ensuring data quality. Classification allows organizations to identify and classify data based on its risk level and importance, allowing them to apply appropriate security measures and policies. A robust data classification system enhances data governance, reduces risks and ensures data quality and protection at scale.

* **Data classification categories**

****

1. **Public**— Information that is freely available and accessible to the public without any restrictions or adverse consequences, such as marketing material, contact information, customer service contracts, and price lists.
2. **Internal**— Data with low security requirements, but not meant for public disclosure, such as client communications, sales playbooks, and organizational charts. Unauthorized disclosure of such information can lead to short-term embarrassment and loss of competitive advantage.
3. **Confidential**— Sensitive data that if compromised could negatively impact operations, including harming the company, its customers, partners, or employees. Examples include vendor contracts, employee reviews and salaries, and customer information.
4. **Restricted**— Highly sensitive corporate data that if compromised could put the organization at financial, legal, regulatory, and reputational risk. Examples include customers’ PII, PHI, and credit card information.

* **Data Classification Process**

****

* **Best Practices for Data Classification**



 **Data Security –**

Organizations understand the significance of granting high-quality data access to their teams to drive insights and business value, while prioritizing sensitive data protection against unauthorized access. Effective data access management is crucial for data security and governance, and a good data security governance program should include access controls that define which groups or individuals can access what data. These controls can be highly specific, down to the individual record or file. As data breaches and regulations such as GDPR and CCPA pose increased risks, businesses must establish clear governance policies that define who can access sensitive data sets and how to track any misuse. Unauthorized access to private or sensitive information should not occur, and implementing effective access management strategies is essential to safeguard data and maintain customer trust.

* **Types of Data Security**

****

### **Encryption**

Data encryption is the use of algorithms to scramble data and hide its true meaning. Encrypting data ensures messages can only be read by recipients with the appropriate decryption key. This is crucial, especially in the event of a data breach, because even if an attacker manages to gain access to the data, they will not be able to read it without the decryption key.

[**Data encryption**](https://www.fortinet.com/resources/cyberglossary/encryption) also involves the use of solutions like tokenization, which protects data as it moves through an organization’s entire IT infrastructure.

### **Data erasure**

There will be occasions in which organizations no longer require data and need it permanently removed from their systems. Data erasure is an effective data security management technique that removes liability and the chance of a data breach occurring.

### **Data masking**

Data masking enables an organization to hide data by obscuring and replacing specific letters or numbers. This process is a form of encryption that renders the data useless should a hacker intercept it. The original message can only be uncovered by someone who has the code to decrypt or replace the masked characters.

### **Data resiliency**

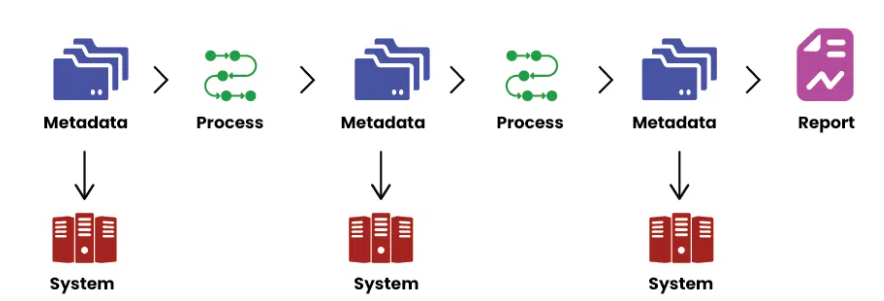
Organizations can mitigate the risk of accidental destruction or loss of data by creating backups or copies of their data. Data backups are vital to protecting information and ensuring it is always available. This is particularly important during a data breach or ransomware attack, ensuring the organization can restore a previous backup.

 **Auditing data entitlements and Process –**

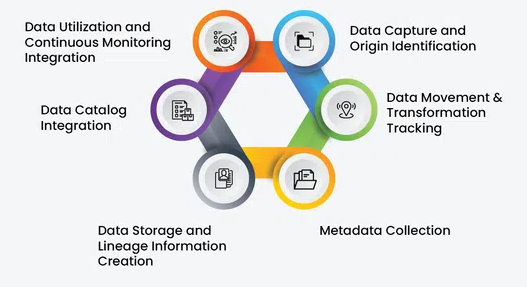
Effective data access auditing is a critical aspect of data governance and security governance programs, particularly in regulated industries. By understanding who has access to what data and tracking recent access, organizations can proactively identify overentitled users or groups and adjust their access accordingly, minimizing the risk of data misuse. Without proper audit mechanisms in place, an organization may not be fully aware of their risk surface area, leaving them vulnerable to data breaches and regulatory noncompliance. Therefore, a well-designed audit team within a data governance or security governance organization plays a key role in ensuring data security and compliance with regulations such as GDPR and CCPA. By implementing effective data access auditing strategies, organizations can maintain the trust of their customers and protect their data from unauthorized access or misuse.

 **Data Lineage –**

Data lineage is a powerful tool that helps organizations ensure data quality and trustworthiness by providing a better understanding of data sources and consumption. It captures relevant metadata and events throughout the data’s lifecycle, providing an end-to-end view of how data flows across an organization’s data estate. As an essential pillar of a pragmatic data governance strategy, data lineage enables organizations to become compliant and audit-ready while reducing the operational overhead of creating audit trails manually, and providing trusted sources for audit reports. Additionally, data lineage empowers data consumers to perform better analyses, and helps data teams perform root cause analysis of any errors, significantly reducing debugging time.

****

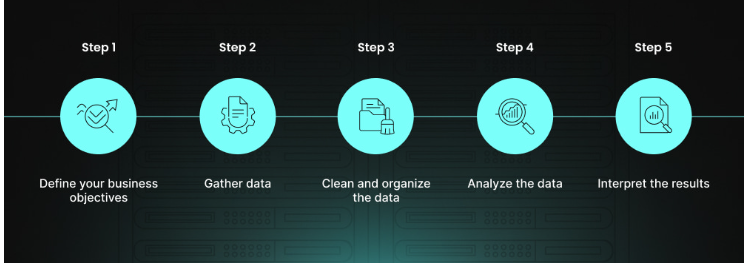
* **How Data Lineage Works?**
  1. **Data Capture and Origin Identification:** The process begins with identifying the starting point or the resource where the data comes from. This can be a customer's sign-up form on a webpage, a sensor in a production facility, or any other component that operations into this nature.
  2. **Data Movement and Transformation Tracking:** Data is covered by these data lineage tools via process and the system thus the movement. This processing implies the kinds of transformations that may include cleaning, filtering and enrichment (combining it with additional data).

****

* 1. **Metadata Collection**: Metadata, which we can define as the "data about data", performs a very important function in the process of lineage. Data will leave the metadata behind that will be informative about its traits, for example, its format, creation date, and the executed transformations.
  2. **Data Storage and Lineage Information Creation:**The center of the repository the smart meter captures these data and sends through the smart grid t move calculations down. By data lineage tools, this data is used to create automatically a lineage map that illustrate and track the data path between data system.
  3. **Data Catalog Integration**: In most situations data-lineage is in the data catalogue with the data catalog. The central library of data lineage is accessible to the users, which means that they can find any information about say raw data used to convey a given message.
  4. **Data Utilization and Continuous Monitoring:**Business analysts, data scientists, and other stake holders can use the data catalog to investigate the data provenance and this information helps them to understand the uses of the data. This transparency means that they have use of the most accurate, high-quality data possible to conduct in-depth, balanced analysis and decision-making.

** Data Discovery –**

As organizations continue to gather massive amounts of data from various sources, it’s becoming increasingly important to make this data easily discoverable for analytics, AI or ML use cases. This is critical to accelerate data democratization and unlock the true value of the data. Furthermore, with the emergence of modern data assets like dashboards, machine learning models, queries, libraries and notebooks, data discovery has become a key pillar of a robust data governance strategy. Organizations should view data discovery as a fundamental aspect of their data governance strategy. It enables data teams to easily locate data assets across the organization, collaborate on various projects, and innovate quickly and efficiently. This helps to prevent data duplication, which can be problematic as it costs money to persist them, and may lead to governance challenges at different security levels.



### **Define your business objectives**

The process begins with setting clear goals that provide direction and focus for the data discovery efforts.

This step also helps guarantee that the subsequent data analysis fully complies with the specific needs and challenges of the business.

### **Gather data**

Then you should proceed with collecting data from a variety of sources, both internal and external in order to assure a wide-ranging scrutiny.

Remember that it is mandatory to gather a diverse range of data to gain a holistic view of the business landscape.

### **Clean and organize the data**

Here, the emphasis is on removing inaccuracies and inconsistencies from the data to ensure its reliability.

Properly organized data facilitates more efficient and accurate analysis and leads to more reliable insights.

### **Analyze the data**

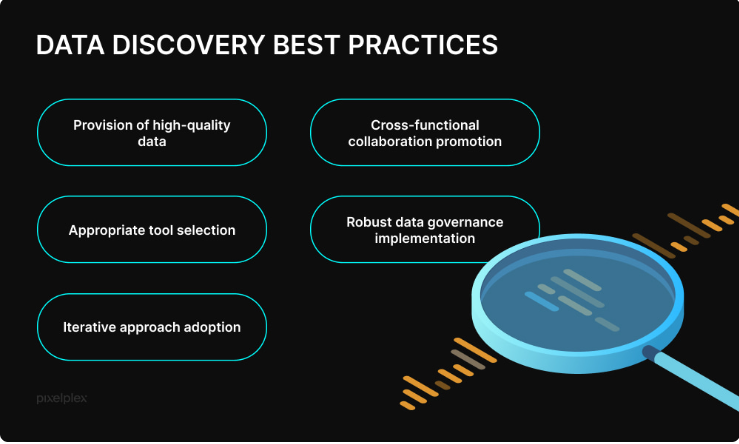
In this stage, the concealed value and potential of data begin to emerge. The emphasis is on utilizing sophisticated analytical tools to comb through the data and identify patterns and insights that may not be immediately apparent.

### **Interpret the results**

The final step in the data discovery process is paramount for translating analysis into practical outcomes. It involves two key actions:

* Making informed decisions based on the performed analysis
* Aligning these insights with the business’s strategic objectives

### **Best Data Discovery Practices 🡪**



1. **Provision of high-quality data**

Quality data serves as the foundation of effective analysis. Thus, prioritizing the accuracy, completeness, and dependability of data is essential, as these elements directly influence the validity of the insights derived.

1. **Appropriate tool selection**

Selecting data discovery tools must be in harmony with the unique requirements of the business. The key considerations include the tool’s functionality, user-friendliness, and its compatibility with current systems.

1. **Cross-functional collaboration promotion**

When different departments like IT, data science, and business units work together, the quality and relevance of insights improve. This joint strategy harnesses the strengths of diverse skill sets, yielding more thorough and practical interpretations of data.

1. **Robust data governance implementation**

A comprehensive data governance framework is imperative to guarantee ethical usage, adherence to compliance standards, and robust security, thus providing effective protection and diligent management of data.

1. **Iterative approach adoption**

Treating data discovery as an enduring and adjustable process facilitates the continuous refinement of techniques and secures alignment with new data and insights, thereby sustaining its relevance.

 **Data Sharing and Collaboration –**

Data sharing and collaboration are vital components in today’s business environment, with organizations exchanging data with internal teams, external partners, and customers across multiple clouds, data platforms and regions. As the demand for external data continues to grow, it is critical for organizations to securely exchange data while maintaining control and visibility over how their sensitive information is used. Data cleanrooms play a critical role in secure and controlled data collaboration, ensuring that data privacy regulations are upheld. It is essential for organizations to invest in open format, interoperable and multicloud data sharing technologies to meet their data-driven innovation needs. Moreover, data marketplaces serve as a bridge between data providers and consumers, facilitating the discovery and distribution of data sets. Therefore, it is crucial to recast data sharing as a business necessity and a crucial pillar of a robust data governance strategy.