



Data Management & Security Playbook for Social Impact Organizations

*A guide for ICT4D practitioners to develop secure data strategies for today,
tomorrow, and 20 years from now*

Foreword & Acknowledgements

This data playbook equips social impact organizations with essential strategies to elevate data maturity and drive meaningful impact. By adhering to these guidelines, organizations can determine how to improve their data practices at the level that they are, while also charting a path to improve their data practices if they so choose.

This playbook was written by several team members from the technology social enterprise [Dimagi](#), leveraging our many years of experience advising leading social impact organizations in their data and security strategies. These authors include [Rishabh Rath](#) (Director of Sales Engineering), [Erin Quinn](#) (Senior Director of Customer Success), [Kai Cowger](#) (Director of Technology), and [Clayton Sims](#) (Chief Technology Officer), with reviews from [Amy Smith](#) (Senior Director of Partnerships), [Gillian Javetski](#) (Managing Director) and [Sarah Strauss](#) (Director of Revenue Marketing).

Our team at Dimagi would like to also thank Matthew Konda, the Founder of the security firm [Jemurai](#) (now [Crux Security](#)) for his review of this Playbook, as well as for the thoughtful insights over the years in advancing our own security work at Dimagi.

We also would like to provide a special thank you to the team at [Digital Square](#) for providing financial support in building out this playbook and for their continued focus on educating others about the importance of [securing](#) Global Goods.

Finally, thank you as well to all of our CommCare partners and users who shared their insights and feedback, contributing to the use cases and overall development of this playbook.

We hope that this playbook provides value to social impact organizations navigating the sometimes messy but essential worlds of data as well as security. For any questions about the contents of the playbook, please feel free to email info@dimagi.com.

Overview

Data, when managed effectively, can be a powerful asset in the hands of global health and development organizations. It's central to Monitoring and Evaluation (M&E) and Program Management, boosting program visibility, ensuring accountability, and providing insights into operational health. Data can reveal trends in growth and decline, and provide justification for future program funding. And most importantly, it's critical in ensuring that the right services are getting to the right people. This playbook is designed to guide organizations through structured data strategies, making it easier to enhance decision-making and improve service delivery.

For non-profit organizations, social enterprises, and frontline programs, collecting and organizing data is essential for delivering critical services. In low-resource settings, where much of the data collection takes place, organizations encounter unique challenges that shape their data strategies. These can include a lack of basic resources like internet access, varying levels of digital literacy, diverse reporting requirements, and ongoing security threats such as hacking attempts. Understanding and addressing these complexities is crucial for developing effective data roadmaps.

By following the guidelines in this playbook, organizations can be empowered to:

- Pragmatically identify their data maturity, organized across three levels
- Learn about the different stages involved in the [data lifecycle](#)
- Understand the essential data and data security requirements for each lifecycle stage in the context of both social impact organizations and their organizations' maturity
- Develop a plan to fill in any gaps, and grow their data maturity.

This framework is tailored for data managers, organizational leaders, field teams, and analysts seeking robust data management and actionable insights. It is designed to support digital strategies across organizations of all sizes and resource levels, offering guidance on data protection, efficient collection, thorough cleaning, effective analytics, and secure data pipelines.

Foreword & Acknowledgements	1
Overview	2
Mapping Data Lifecycle Stages	3
Adapting the Framework for Social Impact Organizations	4
Data Maturity Levels: Social Impact Organizations	5
Level 1: Essential Data Practices	6
Level 2: Handling More Complex Needs	7
Level 3: Sophisticated, Enterprise Data Needs	7
Mapping Social Impact Organizations' Data Journeys	9
All Stages: Data Security as a Foundation	9
Stage 1: Data Generation and Collection	12
Stage 2: Data Processing	16
Stage 3: Data Storage and Management	18
Stage 4: Data Analysis, Visualization, & Interpretation	20
Appendix: CommCare Features for Your Data Journey	24

Mapping Data Lifecycle Stages

The concept of [Data Lifecycle Stages](#) refers to all of the steps required to collect, manage, and leverage data, from data generation through to deletion.

There are numerous data lifecycle stage frameworks¹ to refer to, all which are slightly different. For this project, the team at Dimagi selected [Harvard Business School's Data Lifecycle Model](#), a widely recognized framework. This model includes eight essential stages: generation, collection, processing, storage, management, analysis, visualization, and interpretation.

- **Generation:** Data is created through interactions and processes.
- **Collection:** Data is gathered using methods like forms, surveys, and direct observation.
- **Processing:** Data is cleaned and transformed for use.
- **Storage:** Data is stored in databases.
- **Management:** Data is organized and secured.
- **Analysis:** Data is analyzed for meaningful insights.
- **Visualization:** Data is presented graphically.
- **Interpretation:** Makes sense of the findings to inform future projects.

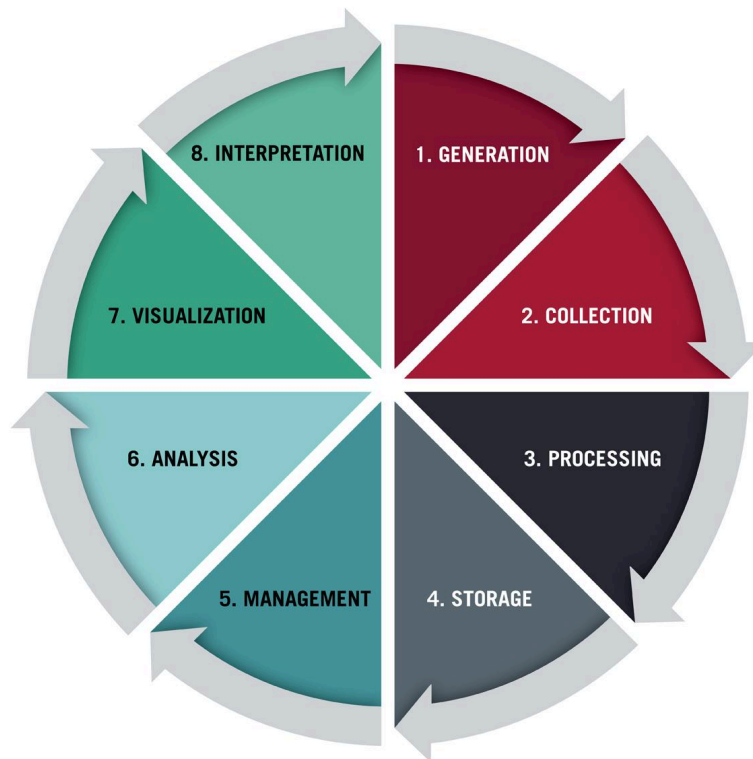


Figure 1: Harvard Business School's [Data Lifecycle Model](#)

¹ Popular examples of data lifecycle frameworks include Microsoft's [Team Data Science Process](#) (TDSP), The [Knowledge Discovery in Databases](#) (KDD) process, and The [Data Management Body of Knowledge](#) (DAMA-DMBOK).

Robust data management is crucial at all maturity stages, ensuring data integrity, confidentiality, and availability. Each component enhances data collection, processing, and analysis. This framework emphasizes the iterative nature of data projects, where insights from one cycle inform the next, highlighting the importance of continuous improvement and strategic data management. Understanding these stages facilitates effective communication with data teams and improves project planning and execution.

Adapting the Framework for Social Impact Organizations

In putting together this framework for social impact organizations, the [Harvard Business Model Framework](#) has been adapted in the following ways for social impact organizations:

Adding Security At Every Stage

Incorporating security at every level of the data lifecycle is crucial, especially when dealing with sensitive data commonly found in social impact work, such as personal health information and demographic details. Ensuring data security throughout data generation, collection, processing, storage, and analysis helps protect against breaches and unauthorized access. Each section of the framework includes specific security measures to safeguard data integrity and confidentiality.

Adapting Data Generation and Collection to Focus on Data Modeling

The data model should be at the forefront of how data is generated and collected. This involves not only collecting data but structuring and modeling it in ways that align with the organization's mission and objectives. Effective data modeling ensures that the data collected is relevant, accurately reflects the reality on the ground, and can be used to inform decisions and measure impact.

Combining Storage and Management

Organizations, particularly smaller ones like community health programs, often rely on cloud software due to limited access to comprehensive storage infrastructures. Combining data storage and management into one stage makes the framework more pragmatic and aligned with the realities of these organizations. This combined stage emphasizes using scalable, secure cloud solutions to store and manage data efficiently.

Combining Analysis, Visualization and Interpretation

Combining data analysis, visualization, and interpretation into a single stage streamlines the process of deriving actionable insights from data. Presenting data in a clear, understandable format is crucial for stakeholders who may not have technical expertise. This combined stage ensures that data is not only analyzed but also effectively communicated to support strategic decision-making and demonstrate impact.



Figure 2: Adapted Data Lifecycle Stages Framework

Data Maturity Levels: Social Impact Organizations

Each organization will begin their data journey somewhere within the matrix below. As long as an organization’s needs are fully met at the level at which they are currently operating, it is not necessary for them to continue to progress to the next level. In fact, there may be many good reasons (resources, capacity, funding) to intentionally stay at a lower maturity level. Organizations sit in one of these three levels, and in some cases, they may also mature their way up to the next maturity level. For example, one community health program may never move from Level 1, while a second community health program may scale in size and complexity, and need to move from Level 1 to 2.

Maturity	Example Organization	Challenges	Goals
Level 1	Community Health Program <i>Small team (5-10 members)</i>	Limited technical expertise, budget constraints, basic data management needs	Establish secure data collection and storage, ensure compliance with basic regulations
Level 2	Regional NGO <i>Medium-sized team (20-50 members)</i>	Managing increased data volumes, ensuring data accuracy, maintaining compliance	Implement advanced data management practices, enhance data security, support complex workflows
Level 3	International Corporation <i>Large team (100+ members)</i>	Managing extensive data sets, ensuring robust data security, integrating multiple systems	Utilize advanced data management tools, implement comprehensive security measures, support large-scale workflows

Figure 3: Data Maturity Levels

By mapping the four data lifecycle stages by the three different data maturity levels, you can get a sense of the state that organizations find themselves in and how they may want to grow:






Data Lifecycle Stages →	 Data Security			
Data Maturity Levels ↓	 1. Generation & Collection	 2. Processing	 3. Storage & Management	 4. Analysis, Visualization, & Interpretation
Level 1	Basic data collection tools	Simple data cleaning with Excel	Encryption Solutions (Bitlocker)	Excel, Google Sheets, In-product reporting
Level 2	Longitudinal user and data hierarchies	Automated data cleaning and deduplication tools	Multi-Factor Authentication	Interactive Dashboards (Microsoft Power BI, Tableau, Google Data Studio)
Level 3	Advanced tools and frameworks Multi-layered Case Management	Advanced automated cleaning and APIs Comprehensive data audits	Single sign-on (SSO) Biometric authentication	Python, R, PostgreSQL, MySQL, and Microsoft SQL Server

Figure 4: Dimagi-Adapted Data Lifecycle Stages x Data Maturity Levels

Level 1: Essential Data Practices

Organizations at this level are just beginning their digital transformation journey. They typically focus on foundational steps such as basic data collection and foundational security measures. These organizations often have limited resources and expertise in data management and are looking to establish a straightforward, secure system for collecting and managing data. The primary goal at this level is to build a solid foundation that ensures data is collected efficiently and stored securely, with minimal complexity. An early maturity level can be representative of small organizations or larger organizations that are earlier on in their technological journey.

Organizations in this phase might use basic features offered by mobile data collection tools, opting for low-code and no-code form builders for creating simple digital forms and fundamental workflows to track individual records. The emphasis is often on ensuring that data is encrypted during transmission and storage, implementing basic role-based access controls to limit data access to authorized personnel, and maintaining compliance with fundamental regulations like GDPR and HIPAA. These organizations are generally focused on building initial capabilities that can be scaled up as they grow and their needs become more complex.

Level 2: Handling More Complex Needs

At this level, organizations have typically expanded their operations and need to handle more complex data and security requirements. They might implement advanced data management practices and enhanced security protocols to support increased data volumes and more sophisticated workflows. This level usually involves the introduction of more sophisticated tools and processes to manage data more effectively and securely. Organizations running several projects across regions or verticals often fall into this category.

Organizations at this level often start to develop advanced forms with conditional logic and validation rules, enabling more accurate and relevant data collection. Workflows are customized to handle more intricate scenarios and larger data sets, allowing for detailed tracking and management of records. Data dictionaries may be created to include detailed descriptions, labels, and categorizations for all data properties, facilitating better collaboration and understanding among users. Enhanced encryption protocols are often implemented to provide stronger protection for data at rest and during transmission. Detailed role-based access controls and two-factor authentication are typically introduced to enhance data security, ensuring that only authorized users can access sensitive information. Compliance with additional industry-specific regulations and certifications is also usually ensured to meet the specific needs of their sector.

Level 3: Sophisticated, Enterprise Data Needs

Organizations at this level generally manage extensive data sets and require highly sophisticated data management and security practices. They typically utilize advanced tools and frameworks to ensure robust data protection and efficient data handling across multiple projects and locations. These organizations operate at a large scale, managing detailed and complex data models that require advanced integration and customization capabilities. While large organizations with programs running at scale across multiple regions and sectors often fall into this category, smaller organizations with very complex workflows can be similarly placed into the complex category.

At this level, organizations often utilize the full capabilities of their data collection tools to create highly customized forms, supporting complex data models and workflows. Multi-layered workflows are usually designed to adapt to various programmatic needs, providing detailed tracking and management of records across multiple projects and locations. Complex management might include nested relationships between different entities being tracked, referral workflows, or other ancillary workflows like one-way and two-way messaging.

Robust encryption techniques are generally employed to ensure that data remains secure at all times, both in transit and at rest. Granular role-based access controls are implemented, allowing for precise management of user permissions and ensuring that sensitive data is only accessible to those with the appropriate authorization. Advanced authentication methods, such as single sign-on, are used to further enhance security. Compliance with the highest levels of industry standards and certifications is usually maintained, ensuring that the organization's security practices are aligned with global best practices. These comprehensive security measures provide

the necessary protection for the organization's complex and large-scale operations, safeguarding their data and supporting their continued growth and success.

By understanding these maturity levels and the specific needs and goals of different types of organizations, this playbook provides a clear pathway for enhancing data management capabilities, ensuring robust data protection, and driving successful outcomes at every stage of an organization's development.

Mapping Social Impact Organizations' Data Journeys

By mapping the data journey, organizations can identify current capabilities, pinpoint areas for improvement, and implement targeted strategies to advance through the data maturity levels. This approach ensures that organizations can effectively leverage data as a strategic asset, drive innovation, improve decision-making, and maintain robust data security at every stage. This section provides a comprehensive framework for assessing and enhancing data management practices.

All Stages: Data Security as a Foundation

Data security should be the foundational cornerstone for any organization utilizing global good technology, regardless of scale. Ensuring the protection of sensitive information is crucial for maintaining trust, safeguarding privacy, and ensuring the integrity of services provided to all clients - especially those in vulnerable communities.

As organizations grow and scale, the complexity of their security needs and solutions may increase, but the *fundamental principle* remains the same: **protecting data is essential for protecting people.** In essence, data security is vital for empowering organizations to deliver on their mission while ensuring the safety and privacy of those they serve.

Map Out and Plan for a Responsible Digital Footprint

Every piece of digital infrastructure adopted brings both new risks and obligations for stewardship. It is a key starting point for organizations to understand the demands on their resources for maintaining different systems, and ensure that they can account for how they will be managed responsibly.

When considering choices for new infrastructure, organizations should seek to establish not just base costs (ie: licensing), but a holistic cost of adoption and maintenance against a fixed standard of security controls and practices. If the personnel and capacity required for any required maintenance (applying patches), modernization (required direct updates to implementations), and compliance (auditing to ensure maintenance occurs) are not considered upfront, such activities are likely to be deferred when new costs appear.

This can pose a significant sleeping danger. As organizations grow, so do the consequences for “weak links” in their infrastructure which they may not have the resources to oversee. This can lead to being taken by surprise with risks like “Side Channel” attacks in which adversaries take advantage of vulnerabilities in seemingly unimportant systems as a backdoor into the data of critical sensitive systems that sit right next to them.

It's important for organizations to recognize when the conditions for security aren't realistic and when to say 'no.' The data that's hardest to lose is data you don't have in the first place.

Prevent the Most Common Vector for Security Breaches

Despite the popular image of ‘dark hat’ hackers, the most common factor in data breaches isn’t sophisticated, or even middling, technical wizardry. By an overwhelming margin most breaches (nearly half) simply involve an attacker “taking over” an existing user’s account, generally through trying a username and password that the user reused on another site which was breached. Working with this knowledge helps establish a few priorities for organizations.

The first is that modern standards-based best practices are supported and adopted for new data systems, rather than bespoke guidelines. For example, the most recent NIST SP 800-53 standards provide evidence that Multi-factor Authentication (MFA), which requires users to provide additional authentication based on a physical asset, for accounts with significant privilege is one of the most crucial controls in the current security environment, while periodic password rotation (which is commonly found in organizational policies) is actually associated with an increased risk.

Organizations should seek a posture that allows them to minimize the surface area of authentication across their digital footprint. With the myriad identities and systems that are encountered on a day to day basis, it’s impractical and unrealistic to expect users to reliably create and remember hundreds of complex, unique passwords for authentication. Adopting Single Sign-On (SSO) and/or a robust enterprise password manager is an increasingly important step for organizations with significant footprints, as is ensuring that the tools the organization adopts are compatible with these tools.

Finally, understanding that most data is breached not through a failure of hard controls, but by authenticated accounts, it is important to account for the scope and visibility of user data access. One key approach is the Principle of Least Privilege, granting users only the minimum access rights necessary to perform their tasks. Following this principle is important to choose technologies that have established role-based access control (RBAC) features, which allow flexible building and assigning of privileges that grant just enough access for all users to do their work, but not too much to increase the risk of exposure. Additionally, it is important to be able to account for the actions taken by a given user in the event of a breach with the scope of access they do hold. Ensuring that tools have sufficient logging of user activity helps ensure that the full consequences of one of these failures can be sufficiently understood.

Consider Human Factors and Limitations

Preventing re-used credentials being leveraged by attackers (so called “password stuffing”) is just the first example of the common attack vectors which are based on human, rather than technology, behaviors. The majority of breaches include some form of human error or failure to have occurred, either due to unclear or unrealistic expectations or due to malicious intentional manipulation. It is essential that the security of digital data can’t be undermined by an individual point of failure, and that tools provide support for users in maintaining a secure approach rather than encumbering them. Enterprise Password Managers are a great realization of this ideal, providing users with an improved day-to-day experience while increasing organization control.

Even robust technical defenses can be overcome by manipulating people into revealing information or performing actions that compromise security. These attacks can take many forms, from

deceptive emails to elaborate schemes, all by exploiting trust, curiosity, or fear. The consequences of a successful social engineering attack can be severe, potentially leading to data breaches and reputational damage. Given that a single employee's mistake can compromise an entire organization's security, comprehensive and ongoing training to recognize and resist these tactics is essential for all staff members. This human-focused defense is, quite often, the best defense against an ever-evolving landscape of cyber threats.

Secure Handling of Data, Encryption, and Anonymization

Once data is collected and stored within a secure tool, it is paramount to ensure any usage of the data is handled in a safe, standardized way. Data should not be stored on local devices unless necessary, and data handling actions (data cleaning, for example) should be performed within secure channels such as through a tool's user interface instead of downloading data files onto a device to perform the same actions. It is especially valuable to identify specific tools in your data ecosystem which can be used reliably to share data in safe ways which minimize the reach of data and can revoke access to it, in addition to encrypting it, and ensure that teams have socialized the use of these tools in their regular roles.

Equally important to the channels of distribution is ensuring that data can be handled with the minimal subset required for a given task. Especially when dealing with datasets that include personally identifiable information, it is important to always extract minimum data about any subject needed by stripping away as much identifiable information as possible while retaining the value needed for analysis and reporting. This not only protects the privacy of individuals but also helps follow data protection regulations in various countries around the world. Tools chosen for data collection should make these practices easy, such as through built in anonymization tools and customizable reports.

Case Study: Dimagi's SOC 2 Journey

Dimagi is a social enterprise that delivers open-source digital solutions for social impact organizations. Their flagship product, CommCare, is used worldwide for data collection and service delivery.

As a leading Global Good provider, security plays a major role in how Dimagi thinks about data, engineering, project design, and more. Because a strong security stance is critical to Dimagi's operations, Dimagi is proud to support CommCare as the first and only SOC 2 certified [Global Good](#) software.

[SOC 2](#) (Service Organization Control 2) is an auditing standard for assessing an organization's controls related to the Trust Services Criteria: security, availability, processing integrity, confidentiality, and privacy of data. It ensures that a service provider securely manages data to protect the privacy and interests of its clients.

In 2020, Dimagi partnered with [Jemurai](#) (now [Crux Security](#)) a cybersecurity firm, to undergo a rigorous audit by an independent third party auditing firm to verify its adherence to the Trust

Services Criteria. The process involves documenting and implementing comprehensive policies and procedures, ensuring robust data protection measures, and demonstrating that these controls are effectively designed and adhered to at a specific point in time. Dimagi attained SOC 2 Type I certification in 2021.

Dimagi has since achieved SOC 2 Type II certification in 2022 and has maintained it ever since. Maintenance for SOC 2 Type II certification requires continuous monitoring of security controls, yearly independent audits of those controls, and updates to policies and procedures as new security standards are released. Each year, multiple staff members spanning various teams at Dimagi come together to undergo a rigorous audit of all of their security controls and through this process ensure they are able to provide comprehensive evidence showing that they adhere to those controls. They also perform penetration and vulnerability tests and a complete backup restore / disaster recovery test with written reports on each. Ensuring SOC 2 compliance has ensured security remains at the forefront of all Dimagi's policies and procedures.

Collectively, these practices ensure Dimagi's software products are the safest and most trustworthy products for frontline workers everywhere.

Questions about the report can be sent to support@dimagi.com.

Stage 1: Data Generation and Collection

Establishing a **robust data model** and **efficient data collection** processes is crucial for effective data management and security. Tailoring data generation and collection methods to meet organizational needs enhances **data consistency** and **reliability**, providing a solid foundation for all data-related activities. This section outlines practices for setting up a data model and collecting data at three levels of data maturity. Each level includes selected practices to enhance data generation, collection, management, and security.

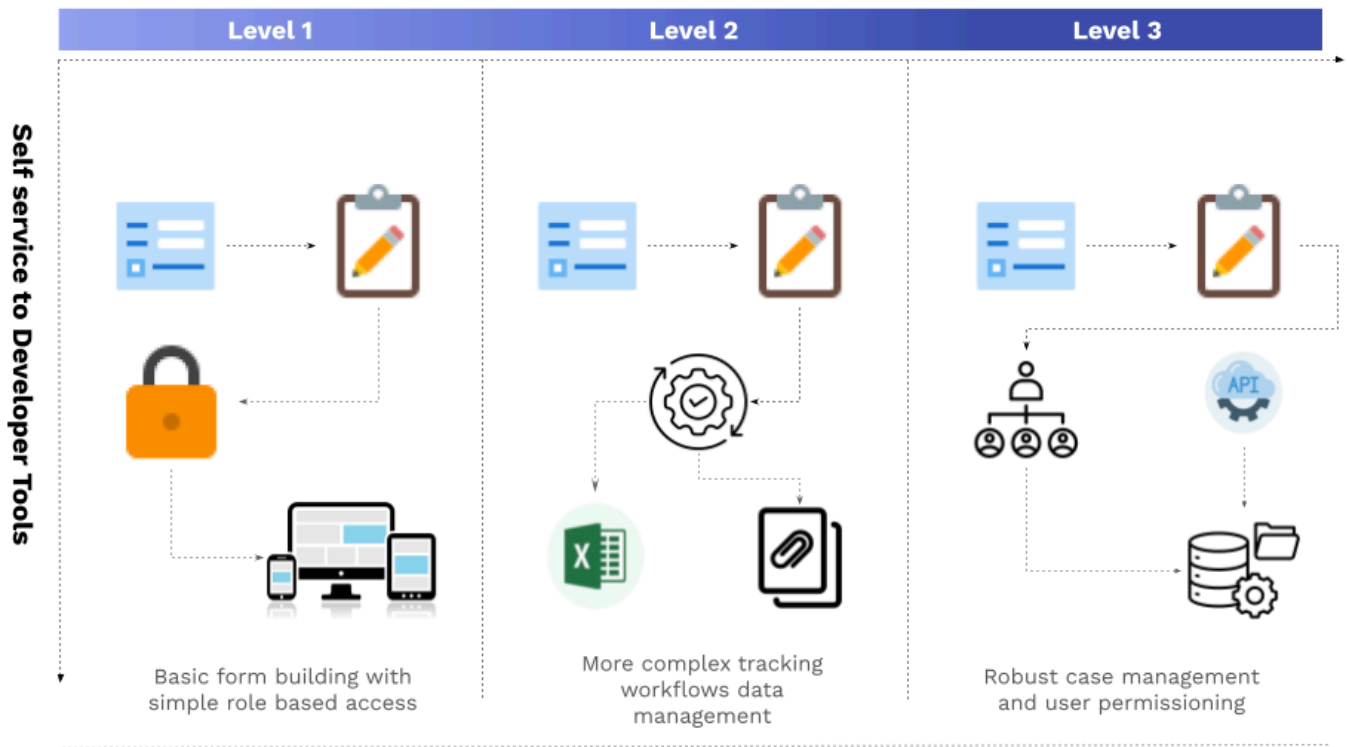


Figure 5: Data Generation & Collection Methods Across Maturity Levels

Level 1 Organizations

Organizations at this level are beginning their **digital transformation journey**, focusing on foundational steps such as **basic data collection** and **security measures**. Building a simple yet efficient data model is essential for organizations with limited resources. Platforms that allow for **mapping key indicators** and creating **linkages between data** are recommended to minimize manual work.

Organizations should use **intuitive form-building tools** with drag-and-drop interfaces to create straightforward digital forms for basic data collection. Forms should be **user-friendly** and capable of functioning **offline** if necessary, supporting data collection in **low-resource settings**. Implementing simple **case management workflows** to securely track individual cases over time is essential, establishing key indicators for entities and documenting them in a basic data management guide to maintain consistency.

Data should be collected through **mobile and web applications**. In low-resource settings, these applications should allow users to work offline and store data locally on their devices until it can be sent to the server. Secure data transmission must be ensured by encrypting data during submission and storage, providing reliable data collection capabilities even in low-connectivity environments. Mobile applications should be intuitive for end users and not require any technical expertise, ensuring quick adoption at the field level.

Organizations should implement **role-based access controls** to ensure data security and compliance. This helps in protecting sensitive information by allowing only authorized personnel to access or modify data, thereby maintaining data integrity and confidentiality.

Level 2 Organizations

As workflows get more complex, there is a need for solutions that can offer **user and data hierarchies**. For example, organizations tracking more than one related entity can leverage data collection platforms that offer features for it. Often, multiple users may collect data for the same set of clients. To ensure that data is linked to the right users, **user groups** can be created for accessing and updating the same records.

Organizations should develop forms with **conditional logic, validation rules, and multi-language support** to enhance data accuracy and relevance. These forms should handle more complex data structures and workflows, improving data collection and service delivery. Tailoring tracking entities to address complex scenarios and larger data volumes, incorporating elements such as referrals, hierarchical supervision, and nested relationships, is important. **Monitoring entities over time** helps evaluate progress and outcomes accurately.

At this level, organizations should implement more comprehensive data management practices, including the use of **advanced metadata and data cataloging techniques** to enhance data organization and accessibility. This helps maintain data quality and facilitates efficient data retrieval and analysis. This is beneficial for programs that are scaling to maintain data hygiene and coherence.

Along with regular form submissions from mobile or web applications, the system should allow **bulk data updates**. These updates can represent the initial import of data into the system at the beginning of a project or updates to existing records. Implement **no-code solutions** that use tools like formatted **Excel files**, allowing even non-technical staff to complete data imports. This capability enables the updating of multiple records simultaneously, making data management more accessible and efficient for scaling organizations.

Level 3 Organizations

Organizations at the complex level manage extensive data sets and require sophisticated data management and security practices, utilizing **advanced tools and frameworks** to ensure robust data protection and efficient data handling across multiple projects and locations. **Application release management** can be a driving factor at this level. For example, organizations looking to collect global indicator data for programs implemented across different teams or geographies should invest in solutions that allow building **template applications** or question libraries. These should be easily replicated or pushed to multiple users with the option of locking content if needed.

Organizations should utilize advanced **form-building capabilities** to create highly customized forms that support complex data models and workflows. Integrating advanced **data capture methods** (e.g., barcode scans, GPS tagging) and appearance attributes to meet diverse data collection needs is essential. Designing **multi-layered case management** workflows that adapt to various

programmatic needs supports detailed tracking and management of cases across large-scale operations. Complex release management practices should be followed to meet programmatic needs.

Organizations should implement **hierarchical data sharing** to manage user permissions and data access effectively. This allows controlled data sharing within defined hierarchies to ensure data isolation and privacy while facilitating collaboration. Employing **advanced data security measures**, including robust encryption techniques, **multi-factor authentication**, and **single sign-on**, is crucial to protecting sensitive information. Ensuring compliance with industry standards and regulations helps maintain data integrity and confidentiality.

Organizations often ingest data from other external sources. To achieve this, they typically explore robust automated solutions with the help of **APIs**. Using API integration can be beneficial for programmatic data collection and incorporating multiple sources of data. This is usually a requirement for larger implementations where data is collected in multiple formats and through different systems.

By following these tailored practices for data modeling and collection, organizations can ensure data integrity, accuracy, and usability at each stage of their data maturity. This approach facilitates efficient data management, enhances data security, and enables organizations to make informed decisions.

Security Best Practices: Data Generation and Collection

When it comes to generating and collecting data, organizations of all sizes need to be committed to protecting sensitive information and ensure data integrity. Below are some best practices for how:

1. **Ensure Secure Collection Methods:** Utilize secure methods for data collection, such as HTTPS for web applications and secure channels for mobile data transmission. Ensure all data is encrypted both in transit and at rest to protect sensitive information from unauthorized access during collection and storage. For mobile and web applications, ensure encryption during transmission to prevent data interception and unauthorized access.
2. **Create Access Controls:** Implement stringent access controls to restrict data access to authorized personnel only. Role-based access controls (RBAC) should be tailored to organizational needs, ensuring that users have the minimum necessary permissions. Use secure login methods for mobile devices, such as passwords or passkeys, and consider mobile device management tools to lock and wipe lost or stolen devices.
3. **Conduct Security Audits and Data Integrity Checks:** Regularly conduct security audits to identify potential vulnerabilities and ensure adherence to security policies. This involves examining access controls, encryption protocols, and data management practices. Implement integrity verification measures to ensure the collected data remains unaltered, using checksums and validation rules during data entry.

4. **Run Regular Data Security Trainings:** Provide regular training for data collectors on data security best practices. Ensure they understand the importance of data protection and how to handle sensitive information securely.

Stage 2: Data Processing

Effective data processing is essential for maintaining data quality and ensuring accurate, actionable insights. Regular data processing activities help organizations transform raw data into usable formats, rectify discrepancies, and remove duplicate records. Data wrangling and data cleaning are key components of processing data. While data cleaning focuses on removing inaccurate and inconsistent data, data wrangling encompasses the entire process of transforming raw data into a more usable form.

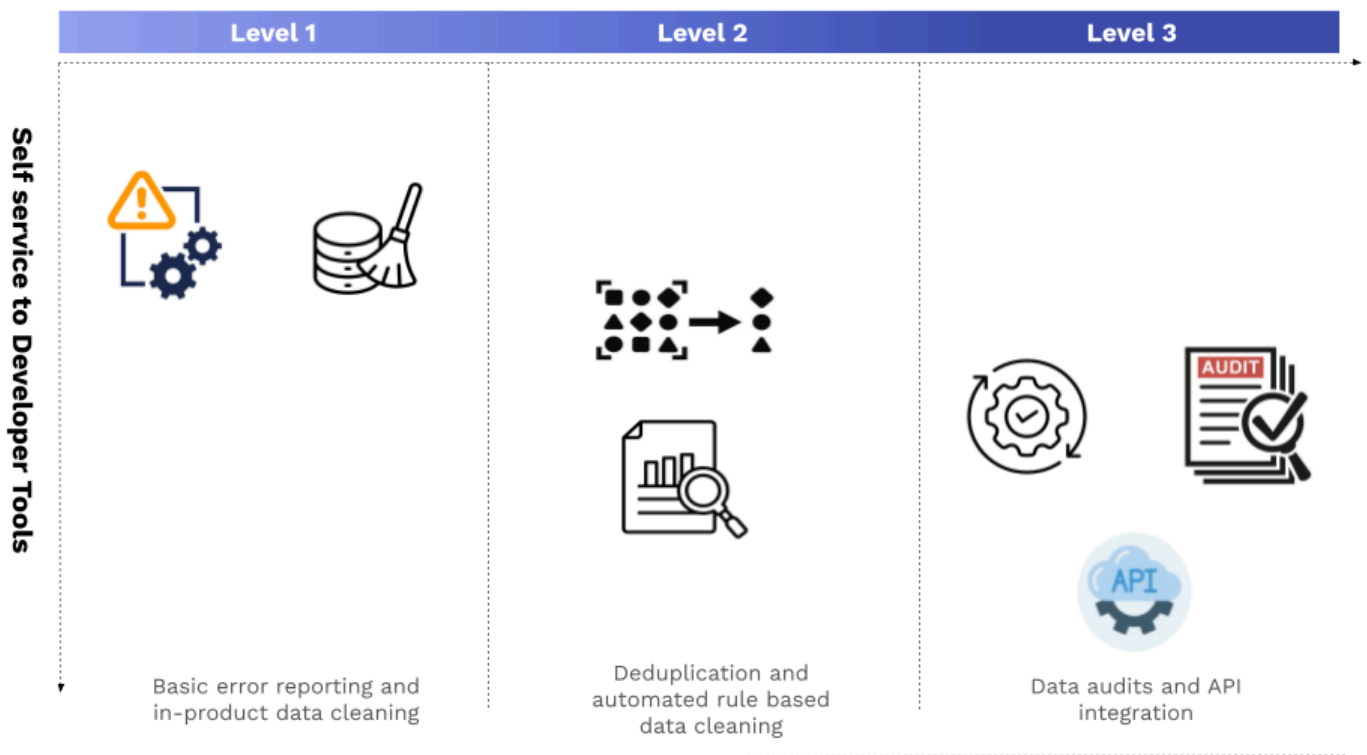


Figure 6: Data Processing Methods Across Maturity Levels

Level 1 Organizations

Organizations at this level are beginning their digital transformation journey, focusing on foundational tools and practices to clean their data. Easy to use tools like **Excel** are good options for organizations to process data.

At this stage, organizations can begin to consider **simple data cleaning tools and processes** to ensure data accuracy, consistency, and reliability. These tools help identify and correct errors in data entries, providing a solid foundation for further analysis. For example, using **aggregate reports** in Excel that allow access to individual entity-level data for cleanup after submission from the field can be an efficient way to keep data clean.

In addition, organizations should **establish basic error reporting mechanisms** to track and correct data issues as they arise and implement processes for manual review of data to catch discrepancies and ensure accuracy. Regularly auditing the data using tools like **Google Sheets** or **in-product reports** from **data collection platforms** helps identify and address errors. Most data collection platforms offer web-based reports that can help organizations identify and resolve data discrepancies. This approach does not require additional investment in tools and expertise.

Level 2 Organizations

At this level, organizations require more advanced tools and practices to handle increased data volumes and complexity in their cleaning processes. With more resources available, these organizations can invest in automated tools and systems that enhance data processing efficiency and accuracy. The focus shifts to managing larger datasets and implementing robust error detection and correction mechanisms.

Organizations should utilize **automated data cleaning and deduplication tools** to manage larger datasets efficiently. These tools can automatically detect and correct common data issues, such as duplicates and formatting errors. Implement robust deduplication processes to ensure the dataset remains clean and reliable, which is crucial for maintaining the accuracy of large datasets. Some data collection platforms offer deduplication features which can be utilized to **clean data in-product**. Organizations should take a proactive approach in deciding their data processing ecosystem as it will define how efficiently they can implement scaling programs.

Enhanced error reporting and monitoring systems should be developed to track data quality metrics. Tools like **Power BI** or **Tableau** can create dashboards that regularly review these reports to proactively identify and address data quality issues, ensuring ongoing data integrity and reliability.

Level 3 Organizations

Organizations at this level require highly sophisticated tools and methods to manage extensive datasets and perform detailed data cleaning. It is imperative that automated solutions are deployed to ensure data integrity is maintained.

Organizations should utilize **advanced automated cleaning and API integration** features to manage and clean data continuously. Defining criteria for identifying and correcting data issues and ensuring these processes run regularly is crucial. Employing robust API strategies to integrate real-time data cleaning and synchronization across multiple systems enhances the overall data quality and consistency.

Comprehensive data audits should be implemented for all data transactions to preserve them for compliance and best practices. Data collection platforms with robust audit mechanisms should be chosen at this level so that organizations can securely track large amounts of data. Regularly reviewing and analyzing these logs helps ensure data accuracy and integrity. **Advanced analytical techniques** are preferred to detect anomalies and trends, providing deeper insights into data quality issues.

By adopting these tailored practices for data processing, organizations can ensure data integrity, accuracy, and usability at each level of their data maturity.

Security Best Practices: Data Processing

Security considerations during data processing are crucial to maintain data quality and protect sensitive information.

1. **Work Only In Secure Processing Environments:** Ensure that the technology you use has secure environments for data processing activities. This includes secure servers and cloud environments with strong access controls and encryption. Ask the technology's data security policy to validate this.
2. **Anonymize Sensitive Data:** For sensitive datasets, implement data anonymization techniques to protect personal information. This includes removing or obfuscating personally identifiable information (PII) and using pseudonymization techniques.
3. **Implement Access Controls:** Implement strict access controls to ensure that only authorized personnel can access and process data. Use multi-factor authentication (MFA) to enhance security.
4. **Validate Processed Data:** Implement validation rules and checks during data processing, including detecting and correcting errors and inconsistencies. Ensure that data isn't tampered with by maintaining a detailed audit logs of data processing activities.

Stage 3: Data Storage and Management

Efficient and secure data storage preserves data integrity and accessibility. Implementing robust storage solutions with backups and encryption ensures data safety and availability. Advanced data management practices and tools help monitor performance, identify trends, and uncover insights, leading to improved operational efficiency and strategic planning.

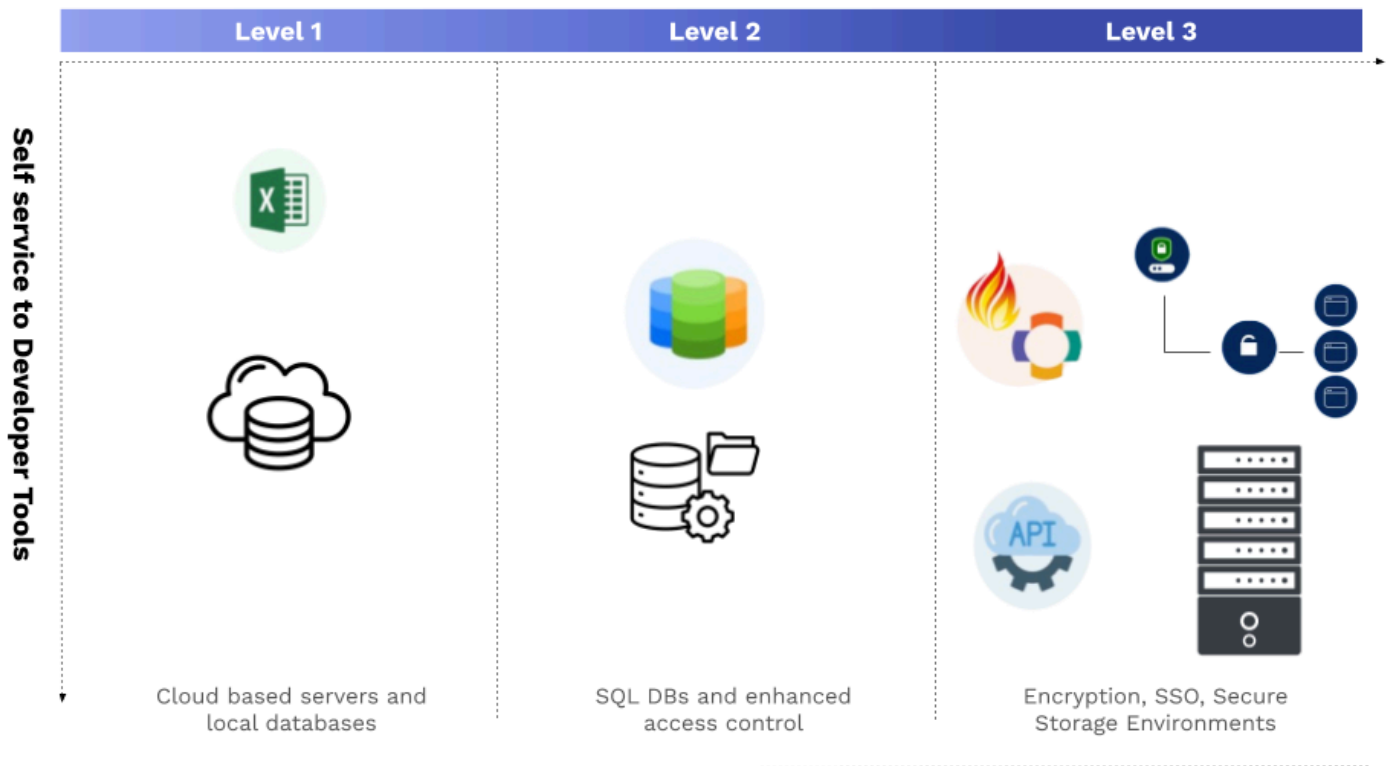


Figure 6: Data Storage & Management Methods Across Maturity Levels

Level 1 Organizations

At this level, **basic storage solutions** such as **Google Drive** or **Microsoft OneDrive** with basic encryption are used for storing data. Regular backups should be implemented to prevent data loss, and basic access controls should limit data access to authorized personnel.

Additionally, utilizing local storage with encryption solutions like **BitLocker** ensures data remains secure even when stored on local devices.

Level 2 Organizations

At this level, organizations handle increased data volumes and need more advanced storage solutions to maintain data integrity and security. Sophisticated storage solutions such as **AWS S3**, **Microsoft Azure Blob Storage**, or **Google Cloud Storage** are utilized. These solutions offer advanced encryption, automated backups, and scalable storage capacities to handle large data volumes. Features supporting data redundancy and high availability are crucial for organizations scaling their programs to ensure reduced downtimes and stable storage systems.

Enhanced security is maintained through **multi-factor authentication (MFA)** and detailed **role-based access control (RBAC)**. Regularly reviewing and updating access permissions with tools like **Entra ID** or **Okta** ensures that only necessary personnel have access to sensitive data.

Level 3 Organizations

Organizations at this level manage extensive data sets and require sophisticated storage solutions with comprehensive security measures. **Highly secure storage environments** such as **Amazon Redshift**, **Google BigQuery**, or **Microsoft Azure SQL Database** with robust encryption, automated disaster recovery, and comprehensive access logs are used. These solutions support high availability and seamless scalability across multiple locations, benefiting organizations handling large amounts of data across different geographic areas.

Granular access controls are implemented to precisely manage data access and ensure compliance with industry regulations. Advanced authentication methods, such as **single sign-on (SSO)** with **Okta** or **biometric authentication**, further secure data access.

Security Best Practices: Data Storage & Management

Security considerations for data storage are vital to protect data from unauthorized access and ensure its integrity and availability.

1. **Ensure All Data Is Encrypted:** Ensure that all stored data is encrypted at rest. Use strong encryption algorithms to protect data from unauthorized access.
2. **Control Data Access:** Use role-based access controls (RBAC) to ensure that only authorized personnel have access to data. In addition, ensure the physical security of data storage locations such as server rooms or data centers - particularly if you are self-hosting your data.
3. **Run Regular Backups:** Perform regular backups of stored data to prevent data loss. Ensure backups are also encrypted and stored securely.
4. **Create a Data Retention Policy:** Implement data retention policies to manage the lifecycle of stored data. Ensure that data is retained for the necessary period and securely disposed of when no longer needed.

Stage 4: Data Analysis, Visualization, & Interpretation

Establishing efficient and secure data analysis and visualization practices is crucial for interpreting data and presenting it in an understandable manner. Data analysis and visualization ensure that data insights are communicated effectively, supporting real-time decision-making and strategic planning. By leveraging data visualization tools, organizations can enhance data interpretation and support comprehensive data analysis.

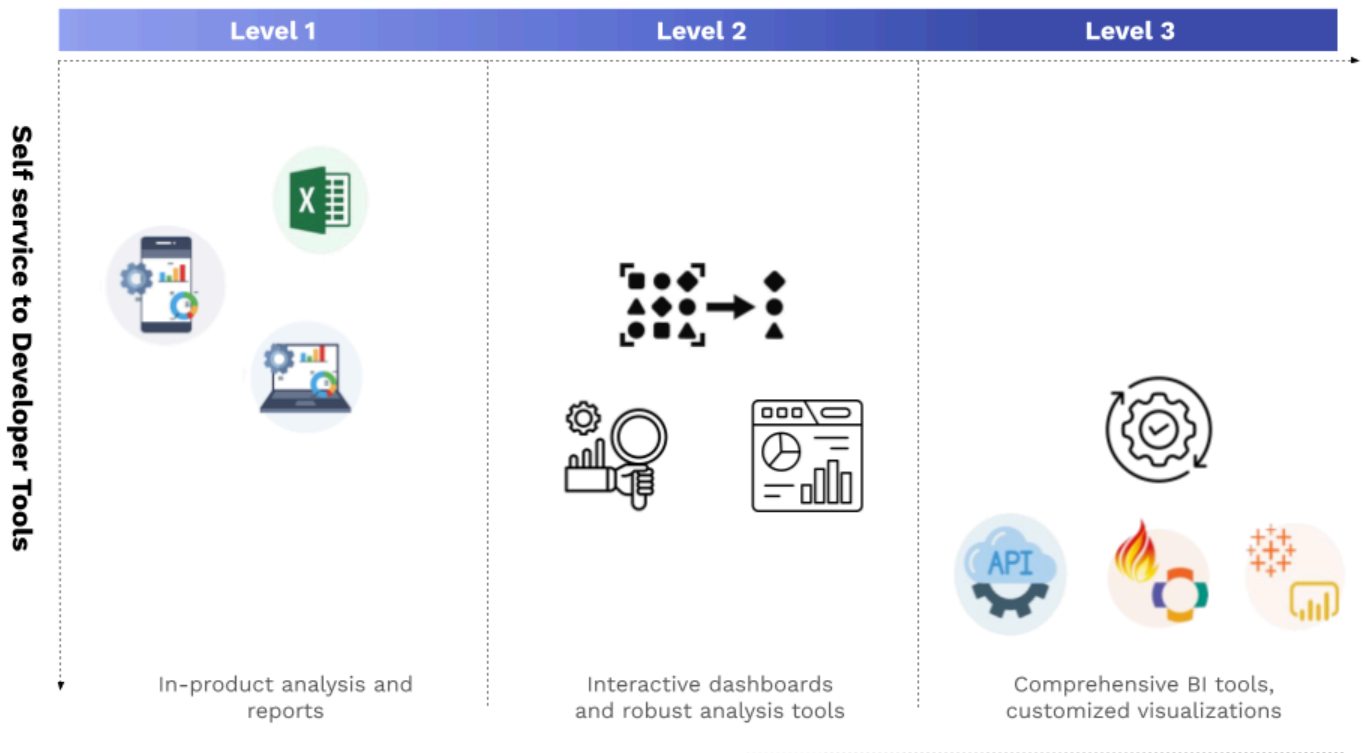


Figure 8: Data Analysis, Visualization, and Interpretation Methods Across Maturity Levels

Level 1 Organizations

At this stage in their journey, organizations may only have the capacity, resources, or even need to invest in **basic analysis tools such as Excel** for data analysis and reporting. Beyond being cost effective, Microsoft Excel has the added benefit of being an industry standard tool-- almost anyone who has worked with data before in their life is comfortable working within Excel. It is essential to ensure any tools used support secure data handling and provide basic encryption for data at rest and in transit.

Similarly, organizations at this phase may only need **straightforward data visualization techniques (such as Excel or Google Sheets)** to visualize their data. An organization focused on tracking, for example, ten primary indicators may only need simple bar graphs to meaningfully show and draw insights from their data. Because visualization needs at Level 1 are typically much lower than at levels 2 and 3, organizations may choose to select a data collection system with stronger **in-product reporting** for ease of use.

Though the number of indicators and visualizations may be small and less complex, even at this maturity level, it is important to ensure that data gets shared securely. Organizations should ensure that any system they choose has access controls to limit viewing to authorized personnel. For instance, remember that an Excel file with raw data and visualizations in it that lives unencrypted on a local computer or is sent unencrypted via email is *not secure*.

Level 2 Organizations

Organizations at this level have progressed beyond basic data analysis and are adopting more advanced tools and methods.

Tools like **Microsoft Power BI**, **Tableau**, or **Google Data Studio** are commonly used for data analysis and visualization. These tools are capable of handling larger datasets and creating **interactive dashboards** that allow for deeper insights and trend analysis. Organizations can visualize data using more **complex charts**, such as scatter plots and heat maps, which provide a more comprehensive view of the data.

Despite their advanced capabilities, these organizations must still prioritize data security. Advanced **encryption methods** and **access control mechanisms** are essential to ensure data privacy and compliance with regulations. Implementing automated reporting systems can also enhance efficiency and accuracy by generating regular reports with minimal manual intervention.

Level 3 Organizations

Organizations at Level 3 are highly data-driven and utilize advanced analytics to drive decision-making.

These organizations utilize advanced analytics, including machine learning and AI, to perform predictive analytics, anomaly detection, and advanced statistical analyses. Tools like **Python**, **R**, and specialized **machine learning platforms** are commonly used. Visualization techniques like **geospatial mapping**, **real-time data visualization**, and **complex network diagrams** are employed to uncover hidden patterns and insights.

Data security is paramount, with stringent security protocols such as **multi-factor authentication**, **role-based access control**, and continuous monitoring implemented to protect sensitive data. Seamless integration of various data sources into a unified data platform allows for comprehensive analysis and reporting. Development of **custom analytics solutions** tailored to specific organizational needs provides a competitive edge and supports strategic decision-making.

Organizations at this level also require sophisticated methods to manage extensive data sets and perform detailed analysis and integration. Using APIs to export data to SQL databases like **PostgreSQL**, **MySQL**, and **Microsoft SQL Server** ensures efficient data transfer. Concurrent data synchronization tools can run exports from multiple sources into a SQL database, ideal for consolidating data from different teams or locations. Custom integrations with other tools, such as **Salesforce** for CRM, **AWS S3** for storage, and **Google BigQuery** for data platforms, enhance data management capabilities. These robust integrations, supported by APIs and data forwarding, ensure an efficient data pipeline tailored to business requirements.

By adopting these tailored practices for setting up a data pipeline, organizations can ensure efficient and secure data flow at each stage of their maturity, leveraging versatile export capabilities to meet their data management and reporting requirements.

Security Best Practices: Data Analysis, Visualization, & Interpretation

Security considerations in data analysis and visualization are crucial to protect sensitive information and ensure the accuracy and confidentiality of insights derived from data.

1. **Anonymize Sensitive Data:** Anonymize sensitive data before analysis to protect personal information. This includes techniques like masking and pseudonymization.
2. **Implement Strict Access Controls:** Implement strict access controls to manage who can access data analysis tools and visualizations. Use role-based access controls (RBAC) to ensure only authorized personnel have access. Maintain audit logs of data analysis and visualization activities. This helps in tracking usage, identifying potential security incidents, and ensuring accountability.
3. **Use Only Secure Visualization Tools:** Use secure tools and platforms for data analysis and visualization. Ensure these tools support data encryption and secure access controls.
4. **Complete Data Integrity Checks:** Implement checks to ensure the integrity of data used in analysis and visualizations. This includes verifying data sources and ensuring data has not been tampered with.

Appendix: CommCare Features for Your Data Journey

For organizations using CommCare, below are some potentially useful features that you can leverage at different points of the data journey, organized by where they are in their data maturity.

Data Security			
CommCare Feature	Level 1	Level 2	Level 3
Industry-Recognized Certifications and Compliance	Adhere to basic regulations, such as GDPR and HIPAA	Maintain compliance with highest industry standards, including SOC 2, NIST 800-53, GDPR, and HIPAA	
Encryption and Data Security	Encrypt data transmissions using TLS/SSL and HTTPS protocols	Use AES-256 encryption for data stored on servers, ensuring robust protection against unauthorized access	Employ advanced encryption techniques for all data transmissions and storage to ensure maximum security
Access Controls and User Authentication	Implement basic role-based access controls (RBAC) and two-factor authentication (2FA)	Enhance RBAC and 2FA to ensure that only authorized users can access sensitive data and functionalities	Use single sign-on (SSO) via identity management providers (e.g., Entra ID , OneLogin, Okta)
Audit Logging	Maintain basic audit logs for user actions, including authentication and data access.	Preserve audit logs for at least 6 years, ensuring adherence to regulations like HIPAA and supporting best practices.	Implement comprehensive audit logging for all user actions, ensuring full traceability and compliance with industry standards and regulations.
Data Generation & Collection			
CommCare Feature	Level 1	Level 2	Level 3
Form Builder	Create digital forms tailored to basic data collection needs	Develop advanced forms with conditional logic and validation rules	Create highly customized forms supporting complex data models, workflows, and data capture.
Case Management	Track and manage individual cases through straightforward workflows.	Customize case management workflows for intricate scenarios, including referral and supervision workflows. Set up case sharing groups to allow multiple users to manage the same cases.	Design multi-layered case management workflows for detailed tracking and management
Data Dictionary	Organize and explain key case properties with basic descriptions and labels	Expand to include detailed descriptions, labels, and categorizations for all case properties.	Develop a comprehensive data dictionary with extensive documentation and examples for all case properties
User Management	Set basic roles and permissions in CommCare	With organizations , manage and organize users, permissions, and data within the platform, supporting scenarios where multiple groups or entities use the system simultaneously.	

Submitting Data	Submit data via CommCare mobile apps, suitable for low resource settings with limited or no internet connectivity.		Submit forms via APIs
Web Apps	Use CommCare applications as Web Apps for facility-based workflows		
Case Imports	Create new cases or update case data in bulk by uploading formatted Excel files		Leverage bulk case update APIs
Data Processing			
CommCare Feature	Level 1	Level 2	Level 3
Data De-Identification	Ensure that personally sensitive data is de-identified		
Data Removal	Understand basic data removal practices in CommCare.		
Data Cleaning	With data cleaning tools , clean case data and form submissions, using reports to identify discrepancies	Identify and remove duplicate cases using user-defined rules	Set automatic update rules to ensure current data and reduce manual workload
Data Storage and Management			
CommCare Feature	Level 1	Level 2	Level 3
Performance Testing	Setup caseload and mobile device testing for scaling projects		
Storage	Save data to excel manually or programmatically through data exports	Save data to BI tools and SQL databases through OData feeds and DET	Robust APIs for integrating CommCare with CRM systems, storage solutions, and other data platforms
Data Analysis, Visualization and Interpretation			
CommCare Feature	Level 1	Level 2	Level 3
Analysis	Direct Excel/CSV exports configurable from the Data tab	Data Exports support enhanced export capabilities with configurable options	With the Data Export Tool , export data to SQL databases using a command line wrapper for interfacing with APIs.
Visualization	With the Case List Explorer , visualize, filter, and interact with cases using a user-friendly interface Simple Excel dashboard integrations for data visualization	In-product report builder for creating summaries, bar graphs, pie charts and mapping reports. OData feeds for direct integration with BI tools like Power BI and Tableau.	Custom third party integrations through robust APIs, integration platforms and more.