Topic 04: Data Lifecycle and Data Quality

BDM3302: Data Management

What is Data Lifecycle Management?

- Data lifecycle management (DLM) is an approach to managing data throughout its lifecycle, from data entry to data destruction. Data is separated into phases based on different criteria, and it moves through these stages as it completes different tasks or meets certain requirements. (IBM)

Phases of Data Lifecycle Management

Phases of Data Lifecycle Management

- A data lifecycle consists of a series of phases over the course its useful life. Each phase is governed by a set of policies that maximizes the data's value during each stage of the lifecycle. DLM becomes increasingly important as the volume of data that is incorporated into business workstreams grows.

Phase 1: Data Creation

- Data creation occurs throughout organizations. This stage involves acquiring data from internal sources like customer data. Most data is created by business functions such as finance, marketing, sales, and human resources rather than the IT department whose typically responsible for managing the data once it's been created.
- Some data can be collected from external like market research companies. Once the data is collected, it must be cleansed and organized so that it can be used effectively. Data collection can be a time-consuming and costly process, but it is essential for businesses to have accurate and up-to-date data.

Phases of Data Lifecycle Management

Phase 2: Data Storage

- Once data has been created, it is typically stored on a computer hard drive or in a secure location. Maintaining data so it can be accessible and usable. Data must be backed up regularly to prevent loss in disaster like earthquake, fire, tsunami. Storage also involves backups that makes the data remains protected. Once data is stored, responsibility for its management typically falls to the IT or security team.
- Storage protections include access control around who can read and overwrite the data, device control such as data encryption, backups to protect the data from loss, plus security measures to protect the backups themselves.

Phases of Data Lifecycle Management

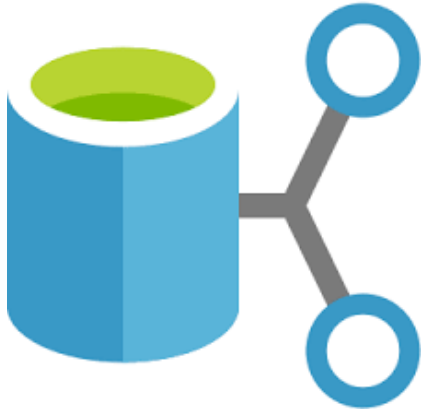
เอาข้อมูลออกมาใช้โดยทุกคนในองค์กร => ความเสี่ยงข้อมูลหาย

Phase 3: Data Usage

- This stage involves using the data to make decisions and take action. Data can be used for marketing, product development, or any other purpose. During this stage, data is accessed, viewed or processed. Protecting data during this phase will usually fall equally between the lines of business and the IT department.
- Protections during data usage include access control, encryption, data rights management for copyrighted information and data loss prevention, which involves software and business rules to prevent unauthorized access to sensitive information.

มาช่วยกู้เผาข้อมูล

Phases of Data Lifecycle Management



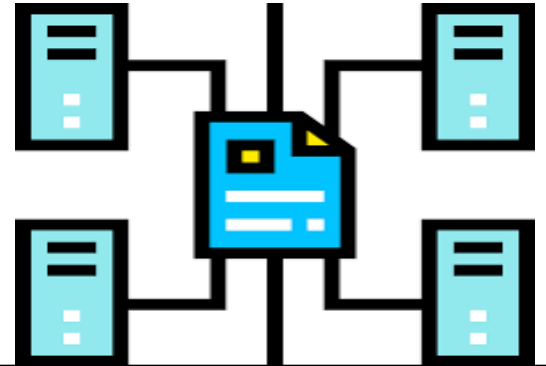
Phase 4: Data Sharing

พอเราใช้ → เรา save และ share ข้อมูลใหม่ๆ ให้คนอื่น และคนนอก.

- Data is often shared amongst internal employees and to corporate partners outside of the organization. Data sharing can occur through the network, via removable media, or across the internet via transfer sites (FTP) or email. When data is shared, it can be beneficial to business, but it also carries some risks.

Protect คนนอก สิ่งป้องกัน

- Data sharing safeguards involve access control, encryption, network security (firewalls/intrusion detection) and data loss prevention. When organizations are dealing with third-party vendors, they should have clear measures in place for data removal and verification after services have ceased.



Phases of Data Lifecycle Management

Phase 5: Data Archiving

เก็บระยะยาว ถาวร

- This phase involves storing data in a safe place for future use. Data can be stored on servers, in the cloud, or physical media. For short-term data protection, all data must be backed up regularly, either onsite or offsite. When an organization needs to retain data for the long term, it can be archived to tape or disk media and placed in remote, secure locations.
- An organization's operations team would usually take responsibility for archiving as opposed to IT or the lines of business. Protecting archived data include access control and encryption.

Phases of Data Lifecycle Management

Phase 6: Data Destroy

บางข้อมูลเราไม่ได้ใช้
↳ หายเลย เอากลับมาไม่ได้
* ทุกคนลบข้อมูลได้*

- When an organization's data reaches the end of its life, it must be permanently erased. Determining which data is erased, how it's erased and how that erasure is verified depends on several factors, such as content type, usage needs and regulatory requirements.
- When it's considered at all, the "Destroy" phase is most often addressed by the operations team. But when managed properly, end-of-life data destruction is truly the responsibility of all stakeholders, from IT to the lines of business.
- However, in practice, organization would like to have one more column to identify that a record (customer) data is active or inactive. Because when you deleted that record from database, it will be disappeared.

Reasons Why NOT To Delete Data

- Need to keep records for historical purposes
  - Accounting, reporting, legal, etc.
- Need to go back and look up information
- Legally required to keep information
  - Accounting data (Tax info, employee witholdings)
  - Regulatory authorities (stock trades, etc.)
  - Discovery for lawsuits

Data Quality Management

- Critical business decisions and allocation of resources are made based what is found in the data.
- Prices are changed, marketing campaigns created, customers are communicated with, and daily operations evolve around whatever data points are churned out by an organization's various systems.
- The data that serves as the foundation of these systems *must be good quality data*. Otherwise, we fail before we ever begin. It doesn't matter how pretty the screens are, how intuitive the interfaces are, how high the performance rockets, how automated the processes are, how innovative the methodology is, and how far-reaching the access to the system is, *if the data are bad—the systems fail*. And if the systems fail, or at the very least provide inaccurate information, every process, decision, resource allocation, communication, or interaction with the system will have a damaging, if not disastrous impact on the business itself.

Data Quality Management

- Data quality management (DQM) is a business principle that relies on the skillsets of people, processes, and technologies to ensure high-quality data throughout your organization.
- The goal of DQM extends beyond maintaining and improving data quality. In the long term, well managed data is the key to making educated business decisions and attaining your desired business outcomes. Teams must be able to trust that the data they're accessing is correct, current, and consistent across the board.

https://www.starfishetl.com/blog/data-quality-management-what-it-and-how-do-it

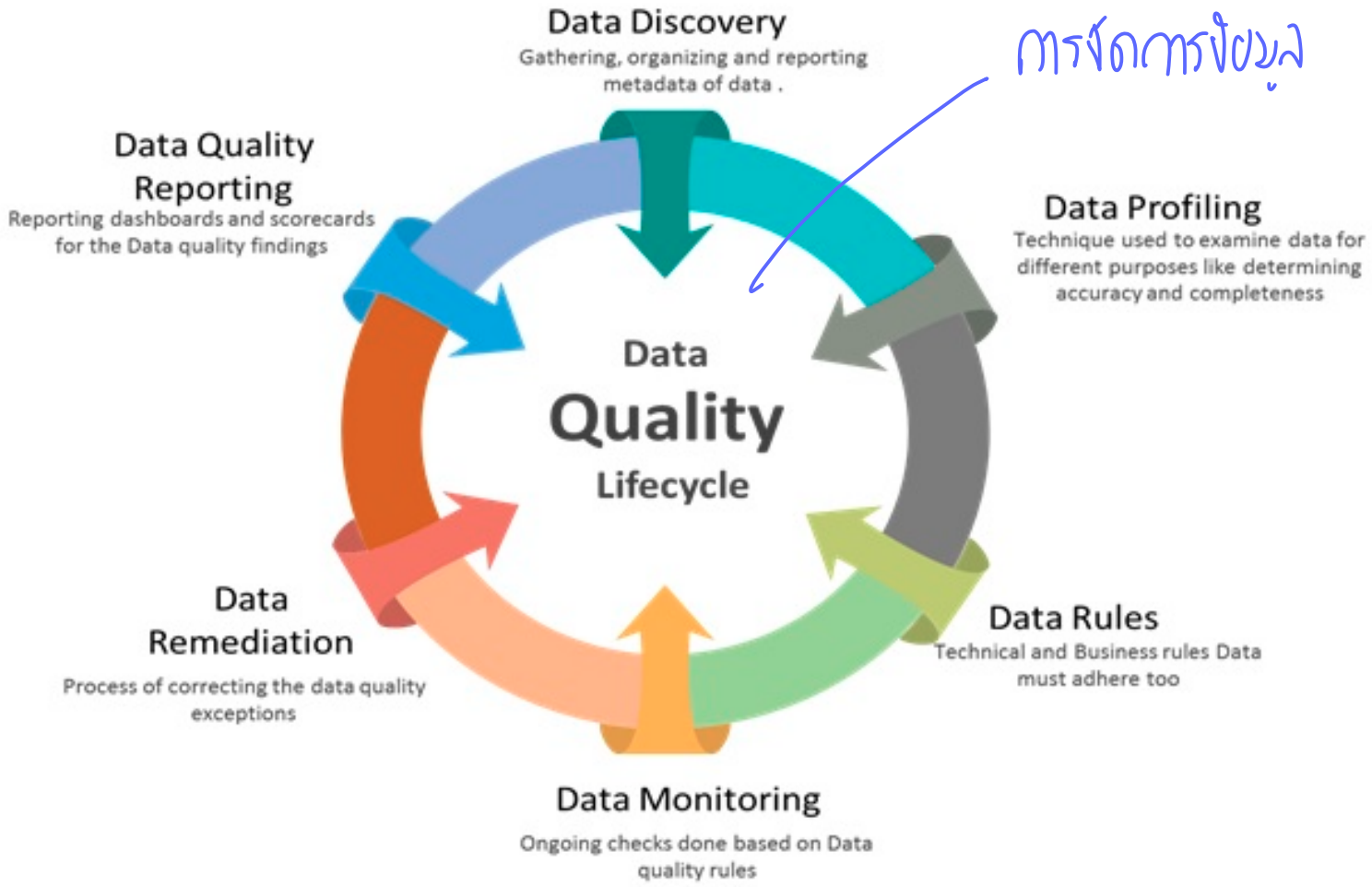
Data Quality Dimensions

- Data Quality Dimensions refers to an aspect or feature of information that can be assessed and used to determine quality of data.
- There are 6 keys for Data Quality Dimensions:
- **Accuracy** – Data accurately represent the "real-world" values. (incorrect of customer names or addresses) Data must be both accurate and precise enough for their intended use. Knowing sales accurately is important, but for many decisions, knowing sales only to the nearest $1000 per month for each product is sufficient.
- **Validity** – Data conforms to the syntax (format, type, range) of its definition. (incorrect customer type)
- **Timeliness** – means meeting the expectation for the time between when data are expected and when they are readily available for use. (changing customer address which is effective on June 1st is entered into the system on June 15th)
- **Completeness** – Data are complete in terms of required potential of data or Not Null. (customer address **missing** postal code or zip code)
- **Uniqueness** – Data are properly identified and recorded only once. (one customer is recorded twice in database, with different identifiers)
- **Consistency** - Data are represented consistently across the data set. (customer account is closed, but there is a new order associated to that account)

Why is Data Quality Management Important?

- The benefits of DQM ripple across every department in your organization. Reliable data reveals trends, informs proactive strategies, and boosts team performance and efficiency. Here are a few more benefits you can expect from data quality management:
  - Correct data creates more efficient data processes and more informed business decisions
  - Quality analytics offer a better view of customers, prospects, vendors, partners, etc.
  - Departmental alignment increases when analytical insights can be shared across teams
  - Businesses save money in the long term by avoiding initiatives that are unsupported by their data
  - Proper data quality helps streamline data governance procedures as well
  - Satisfaction and trust improve when a business can reference comprehensive customer data

Phases of Data Quality Management



การจัดการข้อมูล

1. Data Profiling

ข้อมูลเป็นยังไง

- The first step towards proper data quality, businesses review their data in detail to uncover issues as they relate to their quality goals. Data profiling looks at whether the **format and content of the data** matches its metadata and whether it is accurate, complete, and valid. Are there blank values? Duplicate data? Strange patterns? These types of questions can be answered with proper data profiling.
- To perform profiling, consider these three aspects:
- **Data structure discovery:** How is the data formatted? Is that consistent with the data standards the business is envisioning? Does the data pass mathematical checks such as the ability to accurately calculate sums?
- **Data content discovery:** Are there any fields or values that are incorrect or null? Do any specific rows in a table have issues? Are there any recurring patterns of concern? For example, are most phone numbers missing area codes?
- **Data relationship discovery:** How is the data interrelated? Does the metadata align?

2. Data Rules

มาตรฐาน จัดเก็บในรูปแบบไหน

- It’s no use going through the process of data profiling if you’re not going to maintain the new ideals going forward. That’s why businesses doing DQM must create a set of technical and business rules for standardizing how the data should be formatted and accessed going forward. These rules can lay out protocols for properly formatting numbers and dates (ex: DD/MM/YY vs. MM-DD-YY) or email addresses.
- One key piece of advice on data rules: don’t overdo it. Too many rules can be just as bad as no rules at all. Get together with stakeholders from different departments in the company to understand the data that is most valued to each of these departments. You can work together to build data rules that fulfill organization’s standards.

3. Data Monitoring

- Once your data rules are in place, data monitoring ensures data is periodically checked against those standards to maintain its integrity. Any deviations from the data rules should be reported through an automated process.
- Business intelligence software can often assist with this aspect of DQM by capturing the irregularities before they infiltrate your data sets and sending alerts to management. Monitoring the entire of data can run the risk of information overload, it's best practice to only monitor the data that drives business decisions. Any information considered vital or sensitive should be monitored through your automated processes.

4. Data Remediation

การแก้ไขข้อมูล

ข้อมูลที่ error

- Monitoring the data is necessary to reveal the deviations of your data away from your set rules. Data remediation initiates a process for resolving the data quality concerns that come from that monitoring. One of the most important aspects of data remediation is examining the root cause of the defective data. **Was it human error? or Processing issues?** เพราะจะรีบไปแก้ไขปัญหา
- If the data is found to be corrupt or inaccurate, your organization should decide whether that “bad data” should be modified, or more simply, deleted. Who will be in charge of this? Should stakeholders and team members be notified before the modifications/deletions take place? An agreed upon approach for how to handle data remediation is part of your long game for data quality management.
- Review data quality rules again to determine if anything should be adjusted or updated. If you uncover any data processes that are affected by the bad data, you’ll need to re-initiate them and align them with the new adjustments you’ve set in place.

5. Data Quality Reporting

- Data quality reporting helps teams easily keep track of the information gathered during the data quality management process. One of the easiest ways to keep data quality top-of-mind is to create and share data quality reports with your team. CRM dashboards can be utilized to show this data across departments, or quarterly scorecards can be created and emailed to the entire organization to keep everyone looped in.
- However, the way you choose to approach it, creating a standardized method for **cataloging data quality reports** is essential. It keeps everyone thinking about data quality and the role they play in maintaining it. It's also a tangible way to reflect on your data quality progress and share that progress with stakeholders in the organization.

6. Data Discovery

- The final step of the DQM lifecycle is the data discovery process. Data discovery deepens DQM by gathering, analyzing, and reporting on the metadata associated to your data.
- As a simple definition, metadata is just data about data. It adds context to your data to reveal deeper insights on where data is located, what it means, how it's being used, and more.
- Data discovery is a process of exploring data through visual tools that can help non-technical business leaders find new patterns and outliers to help an organization better understand the insights their data has to offer. (Oracle Netsuite)

Why do we need to manage our data?

- Impacted by government which require PDPA rules for every year
- Increasing pressure on departments or agencies to manage these data properly

Benefit of Good Data Management

- Benefits to Data Suppliers
  - Increased confidence and trust their data that will be used according to their agreed conditions of use
  - Clear understanding of the subsequent use of their data, documented in a formal Memorandum of Agreement signed by supplier and user
  - Fair return for the use of their data

Benefit of Good Data Management

- Benefits to Data Brokers/Intermediaries
  - Better quality
  - Better care of the data holdings
  - Better control over the data
  - Improved knowledge and understanding of data holdings
  - Improved business processes, including better and more efficient use and re-use of data
  - Increased confidence that the organization complies with statutory and non-statutory obligations
  - Better control over access to data
  - More sensible and consistent data charges and conditions of use
  - Increased trust and confidence of customers in the quality and reliability of outputs

Benefit of Good Data Management

- Benefits to users and customers
  - Improved awareness and understanding of what data are available for current and future use
  - Improved access to data, free from unnecessary obstacles, safeguarded from disclosure of personal information or infringement of legal and contractual obligations
  - Better quality and timely information i.e., access to the right information at the right time
  - Better value for money, resulting from clear, fair and consistent data charges and conditions of use
  - Better exploitation of data generally, enabled by easier data exchange and integration with other harmonized data
  - Efficiency gains resulting from easier use of better-quality data