to The Complete Data Management Masterclass

Presentation by George Smarts
https://www.udemy.com/course/data-management-masterclass-the-complete-course/

What will you learn in this course?

Course Modules

Module 1- The Basics of Data Management

Module 2 - Data Ethics

Module 3 - Data Governance

Module 4 - Data Architecture

Module 5 - Data Modeling and Design

Module 6 - Data Storage and Operations

Module 7 – Data Security

Module 8 – Data Integration & Interoperability

Module 9 – Document and Content Management

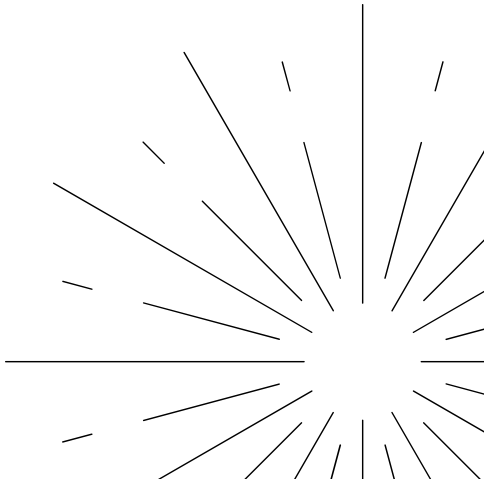
Module 10 – Master & Reference Data Management

Module 11 – Data Warehousing & Business Intelligence

Module 12 – Metadata Management

Module 13 – Data Quality

Module 14 – Big Data & Data Science



https://www.udemy.com/course/data-management-masterclass-the-complete-course/

Why take this course?

Reason 1 - We cover every Data Management subject area

Reason 2 - Real examples of use cases

Reason 3 – CDMP prep

Reason 4 - Tips and tricks from experience

Reason 5 - Answers most questions that you can expect in interviews

Reason 6 - Gain confidence in attending any data meeting

Reason 7 – Additional resources provided for further reading

Download the course resources

Main resource - PDF Presentation file with all lessons

Other resources - additional files/resources provided in various lessons

Module 1 - The Basics

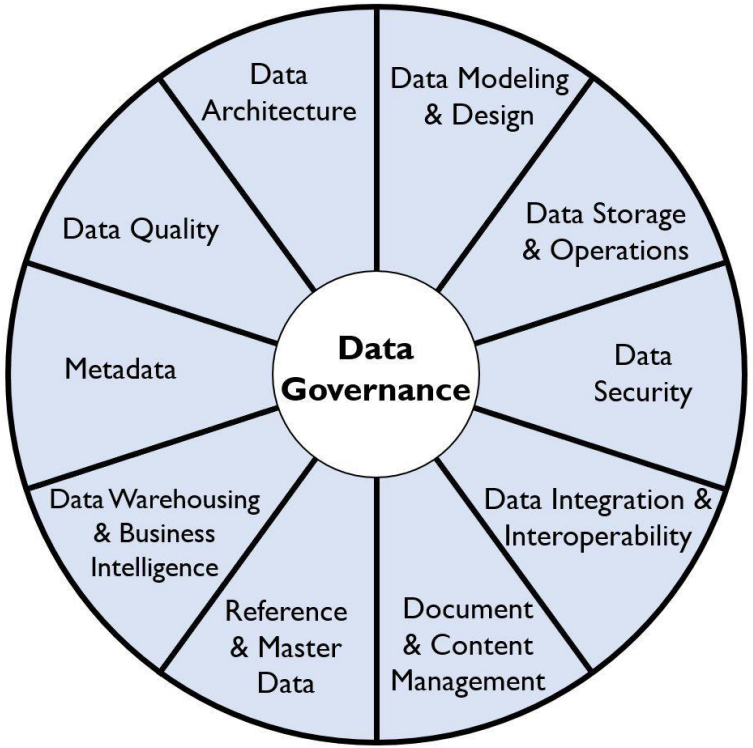
What is Data Management?

‘**Data management** is the practice of collecting, organizing, protecting and storing an organization’s data so it can be analyzed for business decisions.”

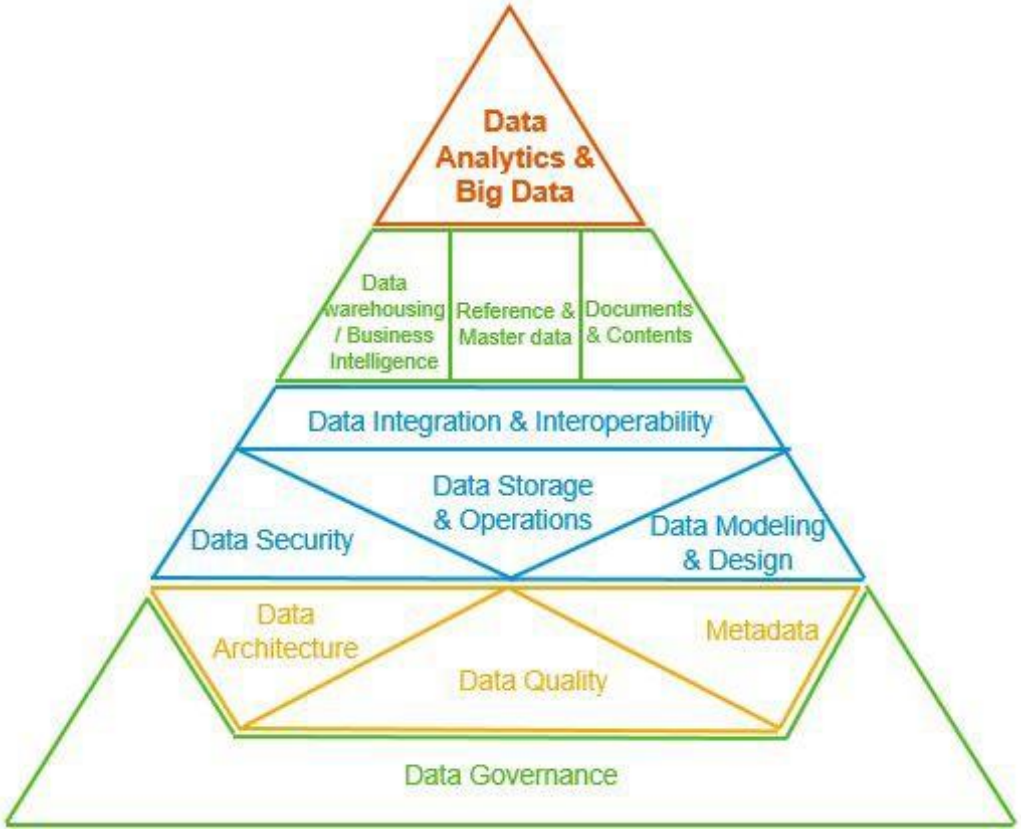
Presentation by George Smarts
Source of quote - Tableau.com

Data Management Subject Areas

1. Data Governance
2. Data Modeling & Design
3. Data Storage & Operations
4. Data Security
5. Data Integration & Interoperability
6. Document & Content Management
7. Reference & Master Data
8. Data Warehousing & Business Intelligence
9. Metadata
10. Data Quality
11. Data Architecture



DMBOK pyramid by Peter Aiken



Source: Peter Aiken developed the DMBOK pyramid

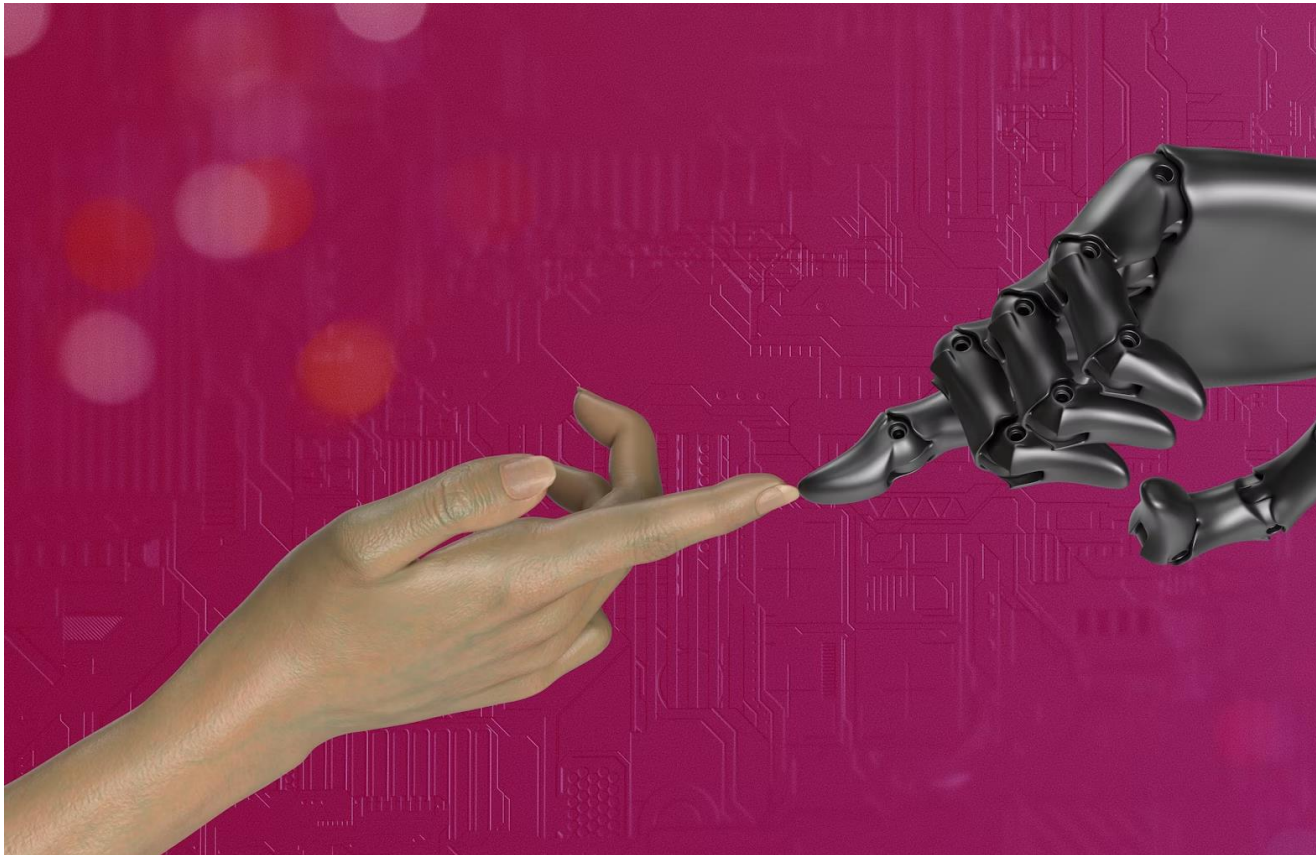
Module 2 - Data Ethics

Definition of Data Ethics

Data ethics refers to the set of moral principles that guide how data is collected, stored, analyzed, and shared.

It addresses various ethical considerations:

- Privacy
- Consent
- Transparency
- Fairness
- Accountability



Facebook and Cambridge Analytica

In 2018, the Cambridge Analytica scandal emerged as one of the most significant examples of poor data ethics. The UK-based political consulting firm gained unauthorized access to the personal data of approximately 87 million Facebook users. This data was initially collected for academic research but was misused for political profiling and targeted advertising during elections.

Consequences of the Scandal

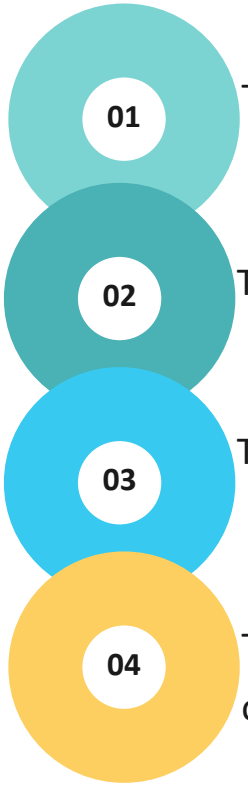
- Financial Impact: $36 billion in Facebook's market value. Additionally regulatory fines totaling nearly $6 billion from FTC.
- Reputational Damage
- Regulatory Scrutiny

Ethical Violations Highlighted

- Lack of Consent
- Data Misuse
- Transparency Issues

Source of image: https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

GOALS OF DATA ETHICS



To define ethical handling of data in the organization

To educate staff on the organization risks of improper data handling

To change/install preferred culture and behaviours on handling data

To monitor regulatory environment, measure, monitor and adjust organization approaches for ethics in data

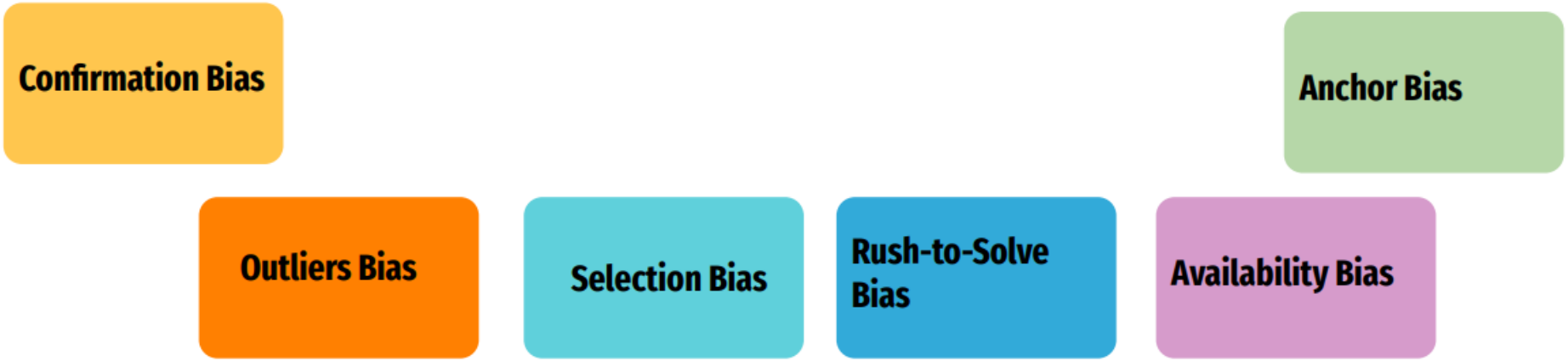
Source: DAMA-DMBOK: Data Management Body of Knowledge: 2nd Revised Edition - CHAPTER 2 Data Handling Ethics

Risk of Unethical Data Handling

What	What is it
Timing	It is possible to lie through omission or inclusion of certain data points in a report or activity based on timing.
Misleading Visualizations	Charts and graphs can be used to present data in a misleading manner.
Unclear Definitions or Invalid Comparisons	Statistical 'smoothing' of numbers over a period could completely change perception of the number.
Bias	In statistics, bias refers to deviations from expected values. These are often introduced through systematic errors in sampling or data selection.
Transforming and Integrating Data	Data integration presents ethical challenges because data is changed as it moves from system to system.
Obfuscation / Redaction of Data	Obfuscating or redacting data is the practice of making information anonymous, or removing sensitive information.

Source: DAMA-DMBOK: Data Management Body of Knowledge: 2nd Revised Edition - CHAPTER 2 Data Handling Ethics

Common types of bias when working with data



Presentation by George Smarts

Confirmation Bias

Example:

- Internet searching - “Why Schumacher is the greatest Formula 1 driver”

How to avoid Confirmation Bias:

1. Understand your assumptions and beliefs before you proceed with the data analysis
2. Don’t jump to conclusions due to your knowledge
3. Compare your beliefs to the findings

Outliers Bias

Example:

Averages of a KPI that do not represent the real picture

You have 5 Sales in October
Sale 1 = $5
Sale 2 = $7
Sale 3 = $6
Sale 4 = $52
Sale 5 = $7
Average = $15.40

How to avoid Confirmation Bias:

1. Do not fully trust averages
2. Understand the distribution of data
3. Look into the outliers and understand their impact
4. Decide what measurement will make most sense

Selection Bias

Example:

Leadership deciding on specific dataset to be analyzed.
Tom (General Manager UK) wants you to analyze the potential of expanding into a new sales territory in the UK by only analyzing a specific set of customers.

How to avoid Selection Bias:

1. User random selection of data if you cannot process all data

2. Make sure the data provides a good representation for the decisions that will need to be made

3. Do not base your data selection based on opinion of people

Rush-to-Solve Bias

Example:

Leadership wants to make a quick decision and decides to make conclusions based on limited set of data.
For example making Sales Targets for next year only based on finalized numbers for the first half of current year

How to avoid Rush-to-Solve Bias:

1. Think about the worst case scenario if wrong decision is made due to rush-to-solve attitude

2. What is the benefit of making a decision faster compared to the worst case scenario?

3. Choose what is more important

Availability Bias

Example:

News reporting multiple plane crashes over a period of a week. This leads to the audience start perceiving that plane crashes are more common than they actually are. Not based on actual statistical analysis.

A team makes business decisions based on easily available data, ignoring the data that will be harder or more time consuming to gather.

How to avoid Availability Bias

Three questions to ask yourself

1. Were you provided all the data needed?
2. Can you really afford to make business decision on limited data?
3. Can you wait to make a decision until you gather all the information?

Anchor Bias

Example:

Purchasing an item after doing research in 2 stores and believing you made a great deal since you bought it from the second store for $30 less.
Later you find out that you could have bought the item for $80 cheaper.

How to avoid Anchor Bias

1. Don't rush to conclusions based on the initial findings in a dataset
2. Take your time to understand the data
3. Do comprehensive analysis of the data before you make any final conclusions

What Data Ethics actually involves

6 Key Activities

1. Review Data-Handling Practices
2. Identify Principles, Practices, and Risk Factors
3. Create Ethical Data Handling Strategy
4. Address Practices Gaps
5. Communicate and Educate Staff
6. Monitor and Maintain Alignment

Source: DAMA-DMBOK: Data Management Body of Knowledge: 2nd Revised Edition - CHAPTER 2 Data Handling Ethics

Deliverables of Data Ethics

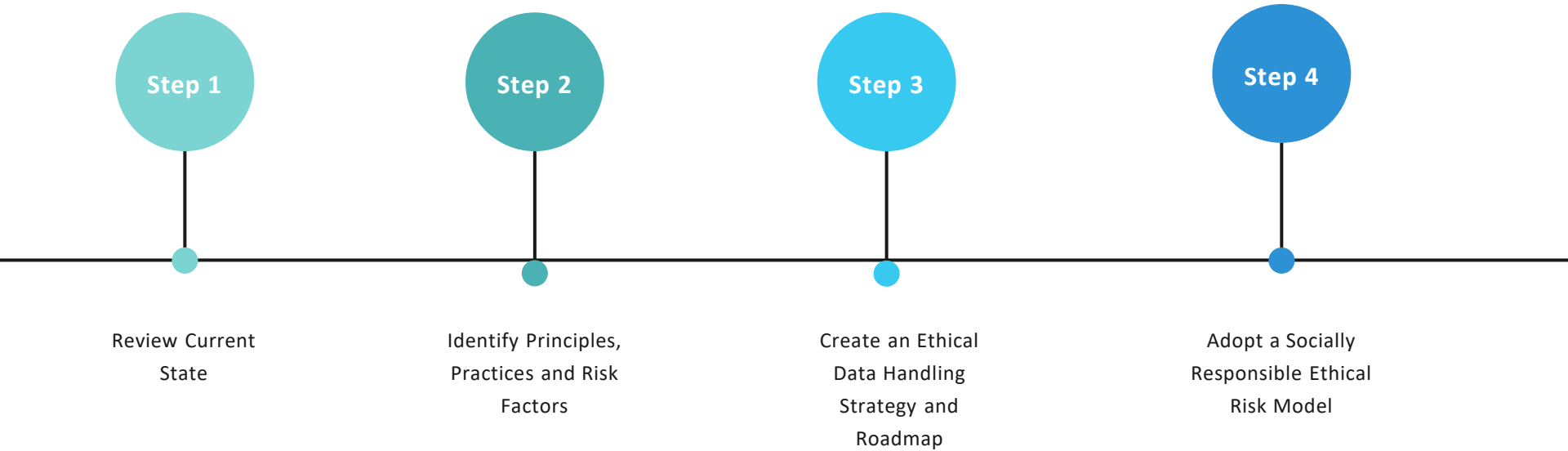
8 Key Deliverables

1. Current Practices and Gaps
2. Ethical Data Handling Strategy
3. Communication Plan
4. Ethics Training Program
5. Ethical Corporate Statements on Data
6. Aligned Incentives, KPIs and targets
7. Updated Data Policies
8. Ethical Data Handling Reporting

Source: DAMA-DMBOK: Data Management Body of Knowledge: 2nd Revised Edition - CHAPTER 2 Data Handling Ethics

Establishing Ethical Data Culture

Step by Step



Source: DAMA-DMBOK: Data Management Body of Knowledge: 2nd Revised Edition - CHAPTER 2 Data Handling Ethics

Module 3 - Data Governance

What is Data Governance?

- **Rules**, **Processes** and **Accountability** that allow the organization to better manage the availability, usability, security and integrity of the corporate data sources.

- **Tip - Think about it as bringing data under control and keeping it secure and consistent.**

7 reasons why you need Data Governance

1. Secure your data
2. Ensure compliance with regulations and data privacy laws
3. Improve the data quality
4. Avoid inconsistent data silos
5. Improve trust in the data
6. Better decision making
7. Improve efficiency

Regulations and Data

Amazon: $886 million

WhatsApp - $267 million

Home Depot: $200 million

Uber: $148 million

Yahoo: $85 million

Capital One: $80 million

*The U.S. relies on a “combination of legislation, regulation and self-regulation” rather than government intervention alone. There are approximately 20 industry- or sector-specific federal laws, and more than 100 privacy laws at the state level (in fact, there are 25 privacy-related laws in California alone).

Data Governance Core Principles

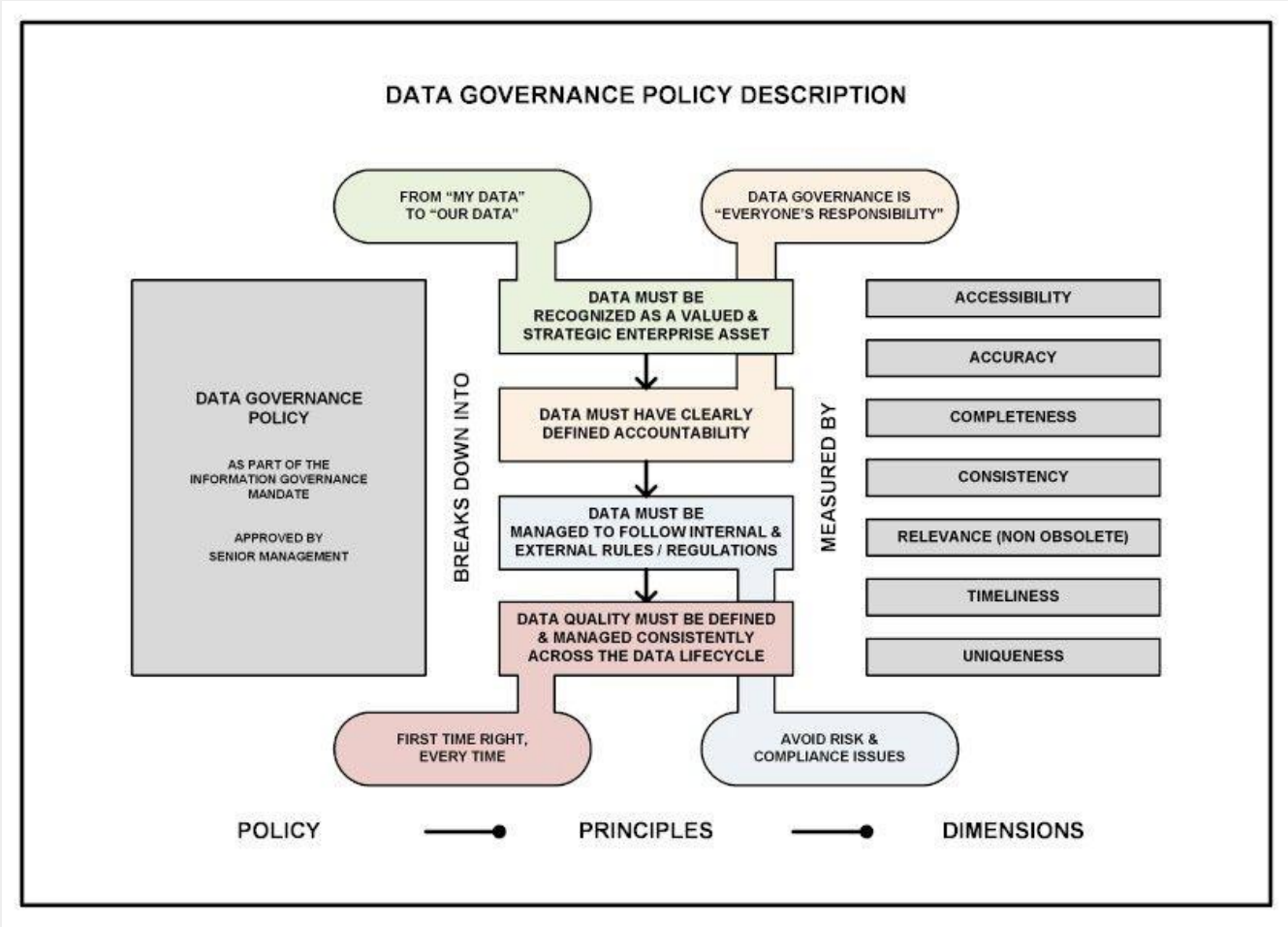


Image Credit: Robert S. Seiner, KIK Consulting

Governance Roles and Responsibilities

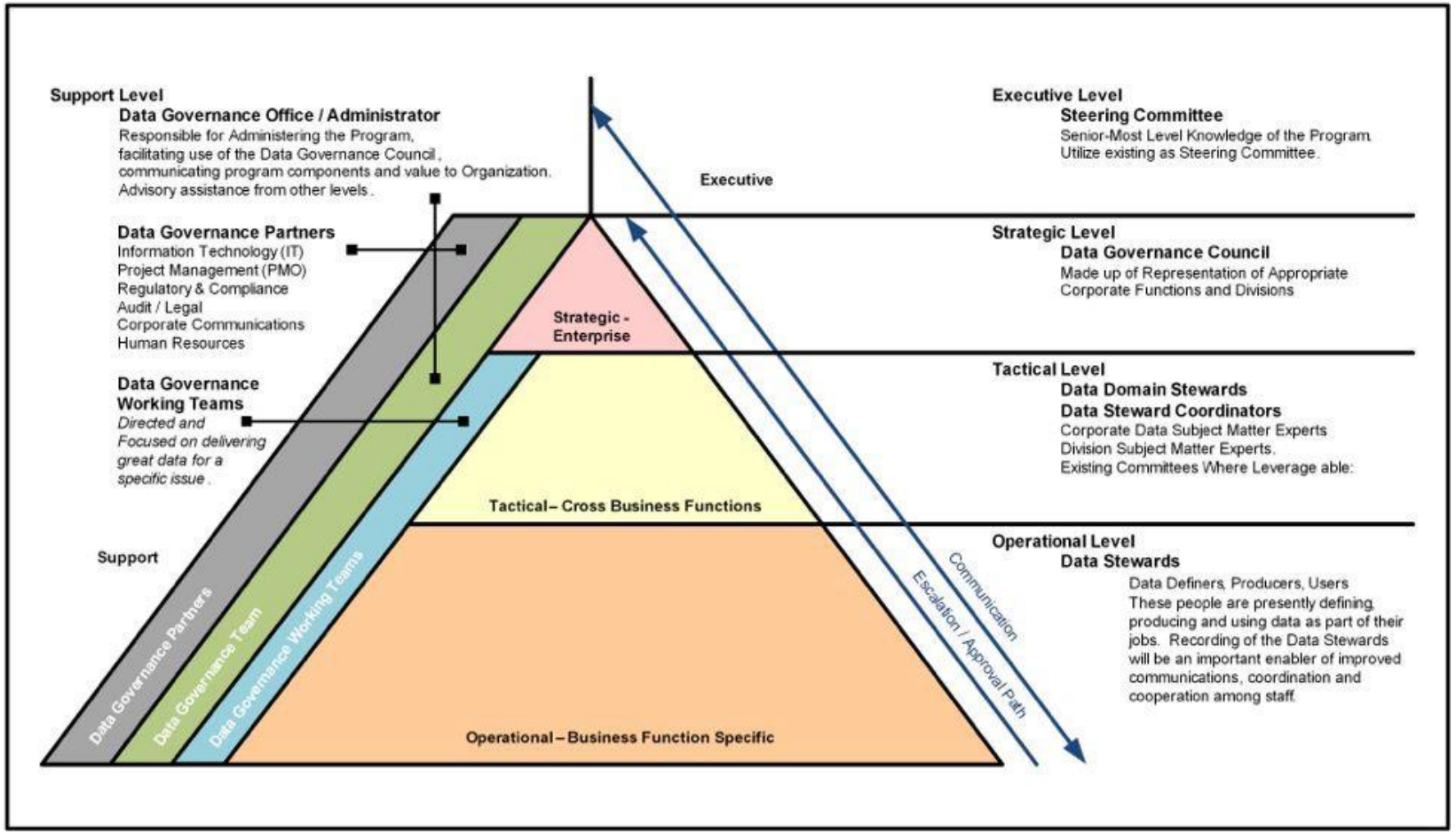


Image Credit: Robert S. Seiner, KIK Consulting

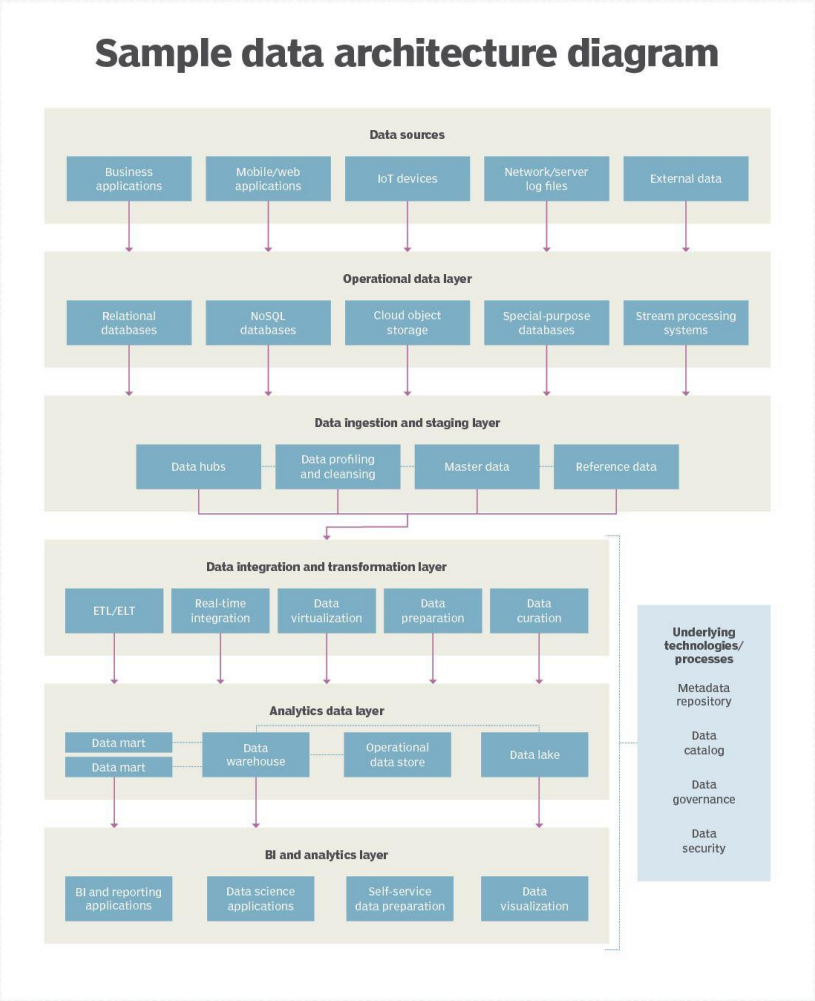
Module 4 - Data Architecture

What is Data Architecture

‘**Data architecture** is a set of rules, policies, standards and models that govern and define the type of data collected and how it is used, stored, managed and integrated within an organization and its database systems. It provides a formal approach to creating and managing the flow of data and how it is processed across an organization’s IT systems and applications”

Data Architecture Example

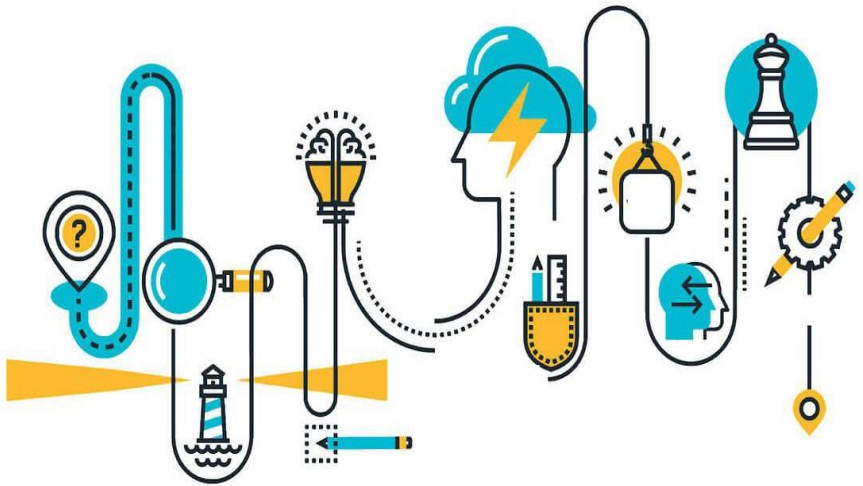
"**Data architecture** is a set of rules, policies, standards and models that govern and define the type of data collected and how it is used, stored, managed and integrated within an organization and its database systems. It provides a formal approach to creating and managing the flow of data and how it is processed across an organization's IT systems and applications"



Source of image:
https://www.techtarget.com/searchdatamanagement/definition/What-is-data-architecture-A-data-management-blueprint

Data Architecture Principles

1. *Data is a shared asset*
2. *Users require adequate access to data*
3. *Security is essential*
4. *Common vocabularies ensure common understanding*
5. *Data should be curated*
6. *Data flows should be optimized for agility*



Sources: https://www.atscale.com/blog/the-six-modern-principles-of-modern-data-architecture/

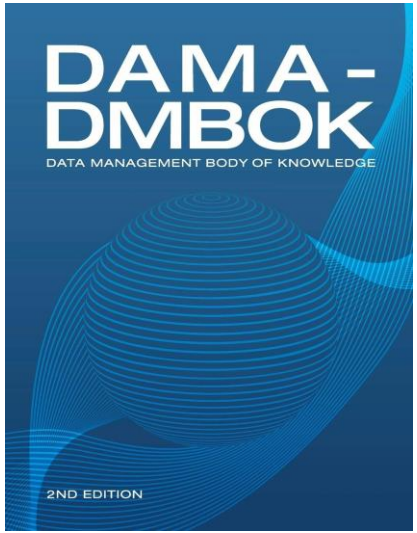
Data Architecture components

1. *Data pipelines*
2. *Cloud storage*
3. *Cloud computing*
4. *AI and ML models*
5. *Data streaming*
6. *Container orchestration*
7. *Real-time analytics*

Sources: https://www.cio.com/article/190941/what-is-data-architecture-a-framework-for-managing-data.html

Data Architecture frameworks

1. *DAMA-DMBOK-2*
2. *Zachman Framework for Enterprise Architecture*
3. *The Open Group Architecture Framework (TOGAF)*



ZACHMAN CERTIFIED™ ENTERPRISE ARCHITECT

TOGAF®

Data Architecture best practices

1. *Cloud-native*
2. *Robust and scalable data pipelines*
3. *Seamless data integration*
4. *Real-time data enablement*
5. *Decoupled and extensible*
6. *Domain-driven*
7. *Balanced*

Source: https://medium.com/@sutherl99/top-7-characteristics-of-a-modern-data-architecture-83b1e61f97f2

Data Architecture Roles

Data Architect!!!

Data Modeler

Data Integration Developer

Data Engineer

Module 5 - Data Modeling and Design

What is Data Modeling

“**Data modeling** is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structure.

The goal is to illustrate the types of data used and stored within the system, the relationships among these data types, the ways the data can be grouped and organized and its formats and attributes.”

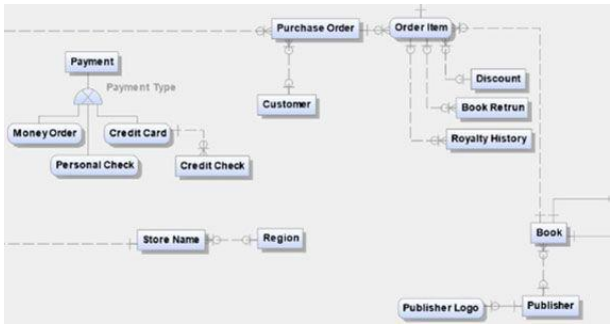
https://www.ibm.com/cloud/learn/data-modeling

Data Modeling vs Data Architecture

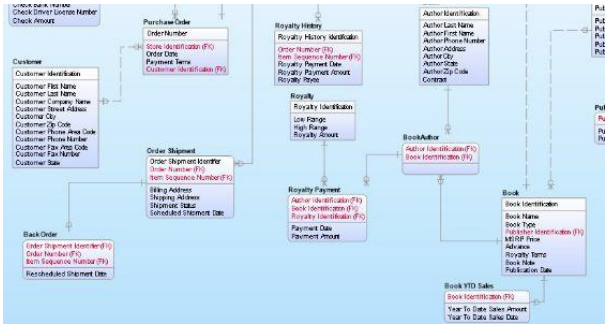
5 Key Differences	
Data Modeling	Data Architecture
Focuses on the representation of the data	Focuses on what tools and platforms to use for storing and analyzing data
Focus on accuracy of the data	Focus is on the infrastructure housing the data
Focus on reliability of the data	Focus on keeping the data safe
Representation of reality	Framework of systems and logistics
Represents a limited set of business concepts	Covers the data infrastructure of the entire organization

The 3 Levels of Data Models

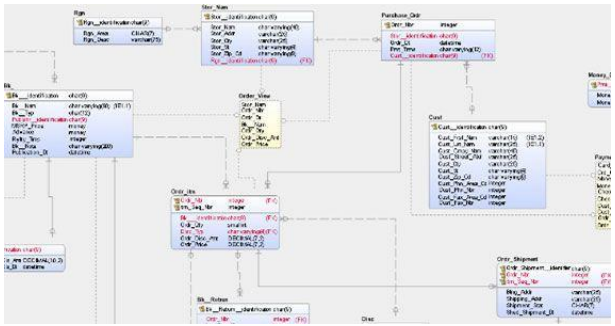
Conceptual data model



Logical data model



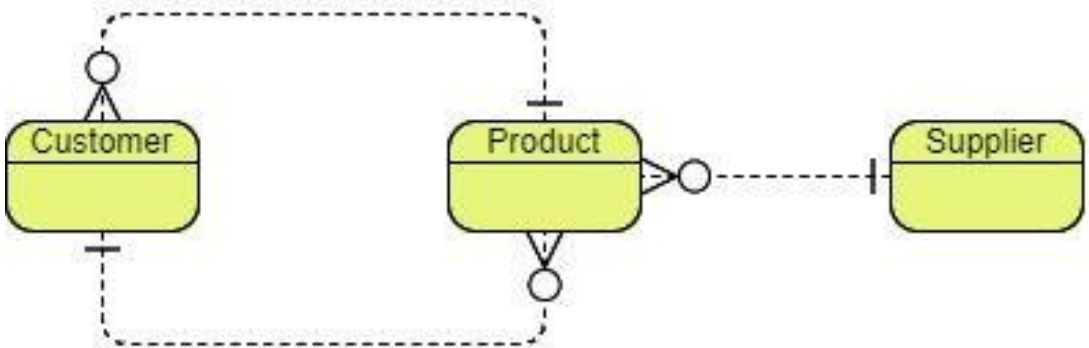
Physical data model



Conceptual data model

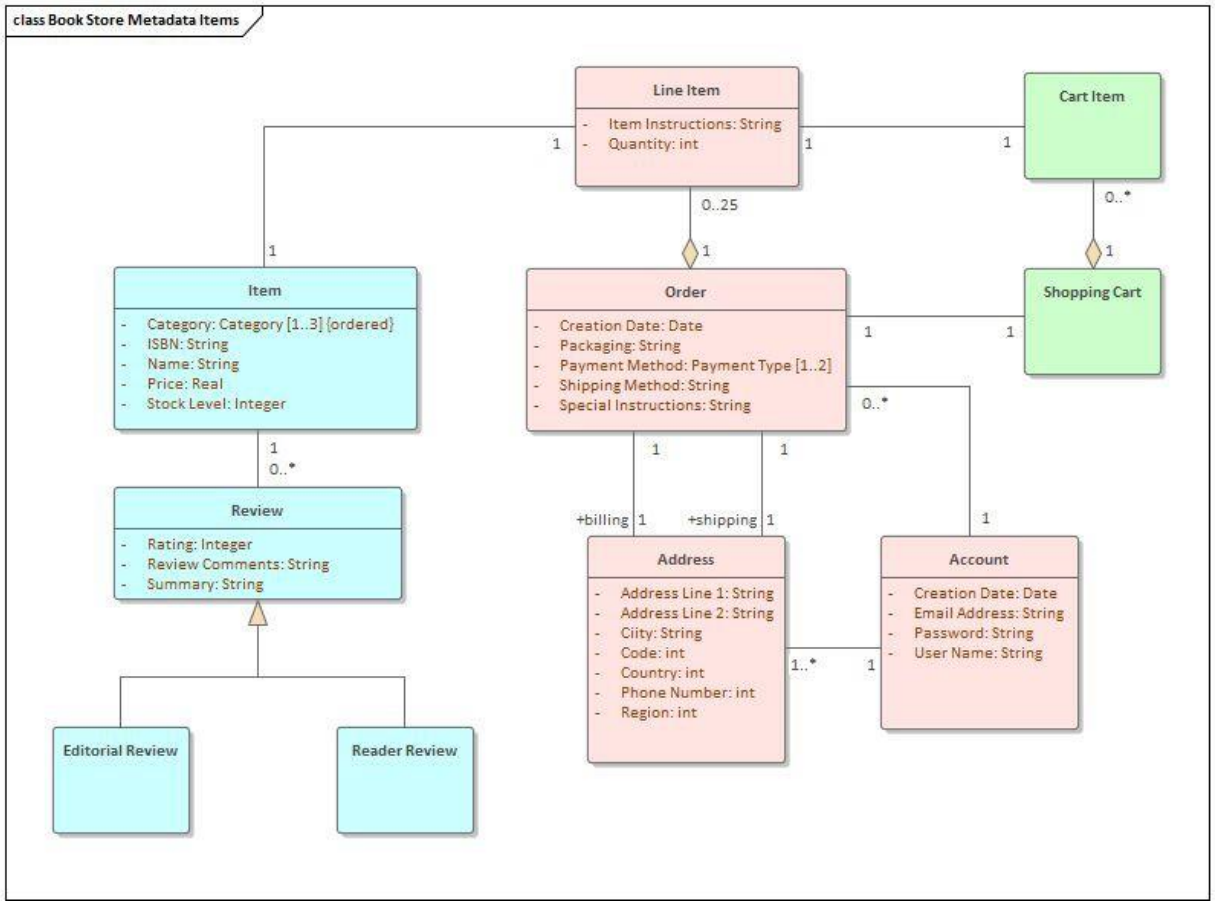
Elements of a Conceptual data model:

- Entity
- Attribute
- Relationship

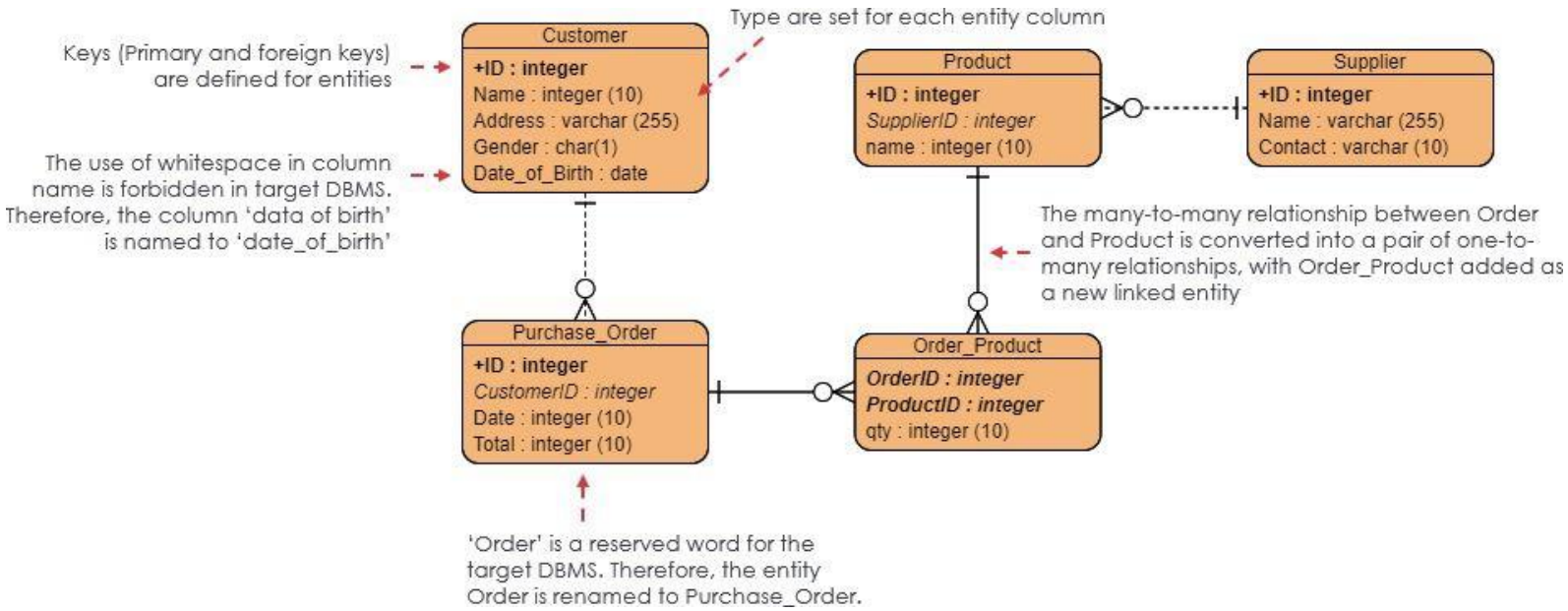


Source of Image: https://online.visual-paradigm.com/knowledge/visual-modeling/conceptual-vs-logical-vs-physical-data-model

Logical data model



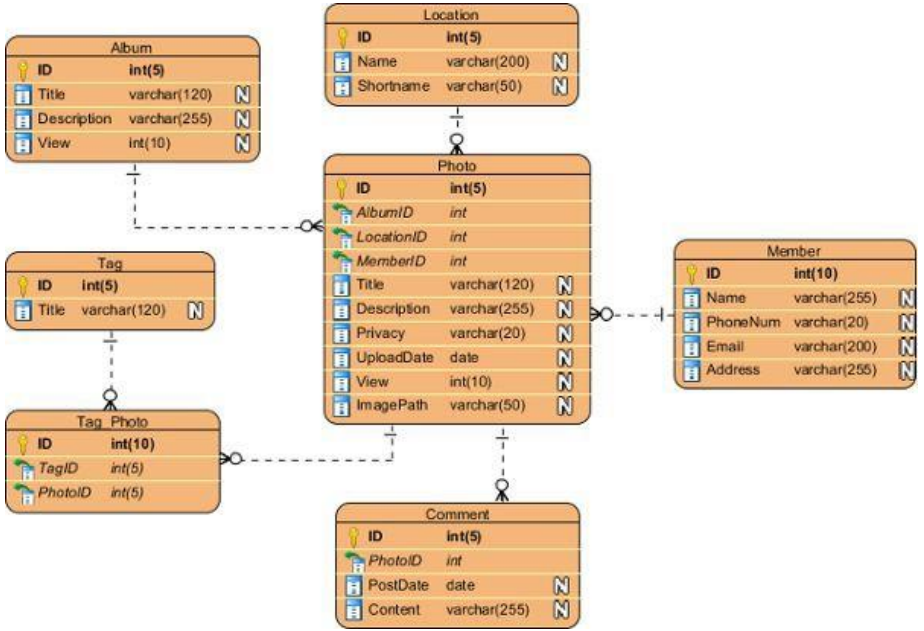
Physical data model



Source of Image: https://online.visual-paradigm.com/knowledge/visual-modeling/conceptual-vs-logical-vs-physical-data-model

Data Modeling Process

1. Identify the entities
2. Identify the attributes of each entity
3. Identify relationships among entities
4. Map attributes to entities completely
5. Assign keys, decide on degree of normalization
6. Finalize and validate the data model



Source: https://www.ibm.com/cloud/learn/data-modeling

Benefits of Data Modeling

- Reduce errors in software and database development
- Reduced cost
- Better documentation
- Improve database performance
- Improved communication between developers and BI teams
- Ease and speed the process of database design

Data Modeling Tools

- erwin Data Modeler
- ER/Studio
- DbSchema
- ERBuilder
- HeidiSQL
- Navicat Data Modeler
- Toad Data Modeler
- SQL Database Modeler

Source: https://spectralops.io

Module 6 - Data Storage and Operations

What is Data Storage and Operations

"**Data storage and Operations** includes the design, implementation, and support of stored data, to maximize its value throughout its lifecycle, from creation/acquisition to disposal"

Two sub-activities:

1. Database support
2. Database technology support

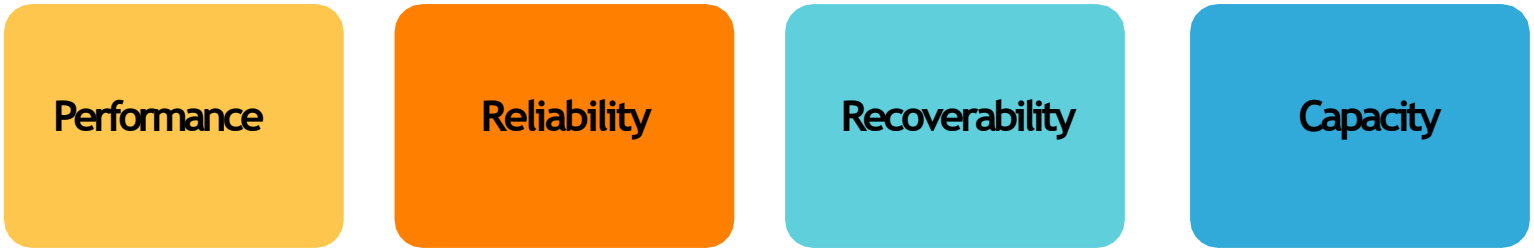
Source: DAMA-DMBOK2 Data Management

Benefits of good Data Storage Strategy

1. *Reduce capital expenses*
2. *Reduce operational expenses*
3. *Easier data management*
4. *Optimized resource utilization*
5. *Easier scalability*
6. *Better performance*
7. *Better user experience*

Data Storage activities

Data Storage Management key attributes



Data Storage Best Practices

Know your data needs

Tiered data storage

Disaster recovery plans

Consolidate Storage Pools

Utilize automated storage management tools

Module 7 - Data Security

What is Data Security

“**Data Security** is the process of protecting digital information from unauthorized access, corruption, or theft throughout its entire lifecycle.”

Why is Data Security important

- The average total cost of a ransomware breach is $4.62 million, slightly higher than the average data breach of $4.24 million (IBM)
- The average per record (per capita) cost of a data breach increased by 10.3 percent from 2020 to 2021 (IBM)
- The average cost of a breach with a lifecycle over 200 days is $4.87 million (IBM)
- 39 percent of costs are incurred more than a year after a data breach (IBM)
- In 2021, the United States was the country with the highest average total cost of a data breach was at $9.05 million (IBM)
- 34 percent of data breaches in 2018 involved internal actors (Verizon)
- It took an average of 287 days to identify a data breach (IBM).

Source: https://www.varonis.com/blog/data-breach-statistics

The Goals of Data Security

1. Enabling appropriate access and preventing unauthorized access
2. Enabling compliance with regulations
3. Creating appropriate policies for privacy, protection and confidentiality
4. Ensuring that stakeholders requirements for privacy and confidentiality are met

Source: DAMA International. (2024). *Data Management Body of Knowledge (DMBOK)*, 2nd Revised Edition. Chapter 7: Data Security

The Principles of Data Security

1. Collaboration
2. Enterprise approach
3. Proactive Management
4. Clear accountability
5. Metadata-driven
6. Reduce risks by reducing exposure

Source: DAMA International. (2024). *Data Management Body of Knowledge (DMBOK)*, 2nd Revised Edition. Chapter 7: Data Security

Types of Data Security

Sources: https://www.varonis.com/blog/data-security

Data Security Risks

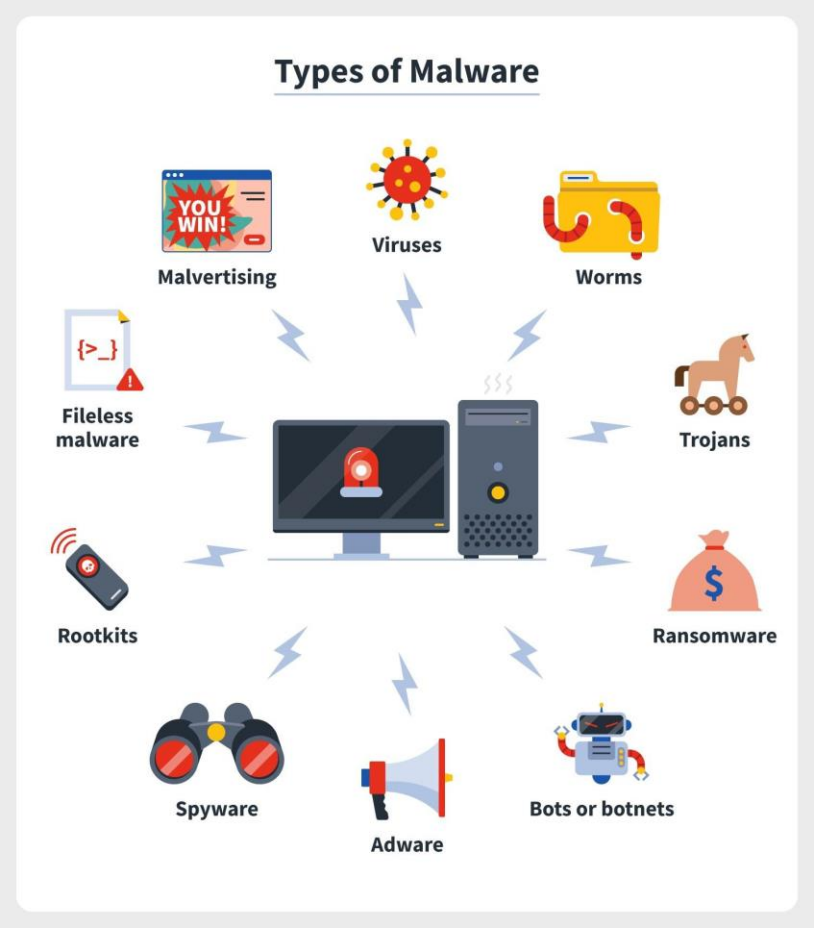
- Accidental Data Exposure
- Phishing
- Malware
- Insider Threats
- Password Attack
- Denial-of-Service (DOS)
- Man-in-the-Middle (MITM)
- SQL Injections
- Zero-day Exploit

Accidental Data Exposure

Phishing

Source: https://www.itgovernance.co.uk/phishing

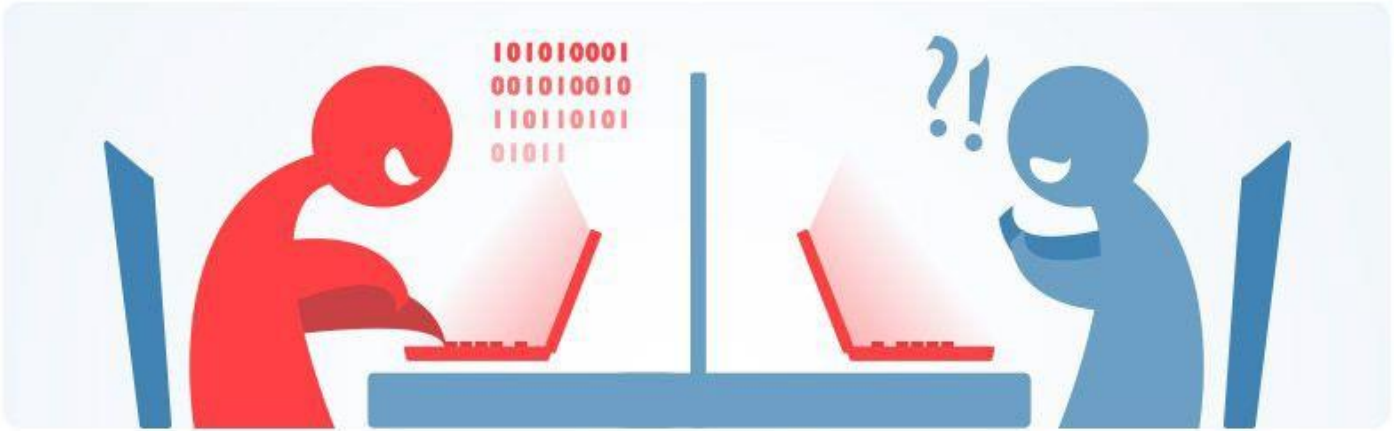
Malware



Source:
https://us.norton.com/internetsecurity-malware-types-of-malware.html

Insider Threats

Insider threat classification by CA Technologies



Malicious insiders

Intentionally use their access to sensitive data to harm the company

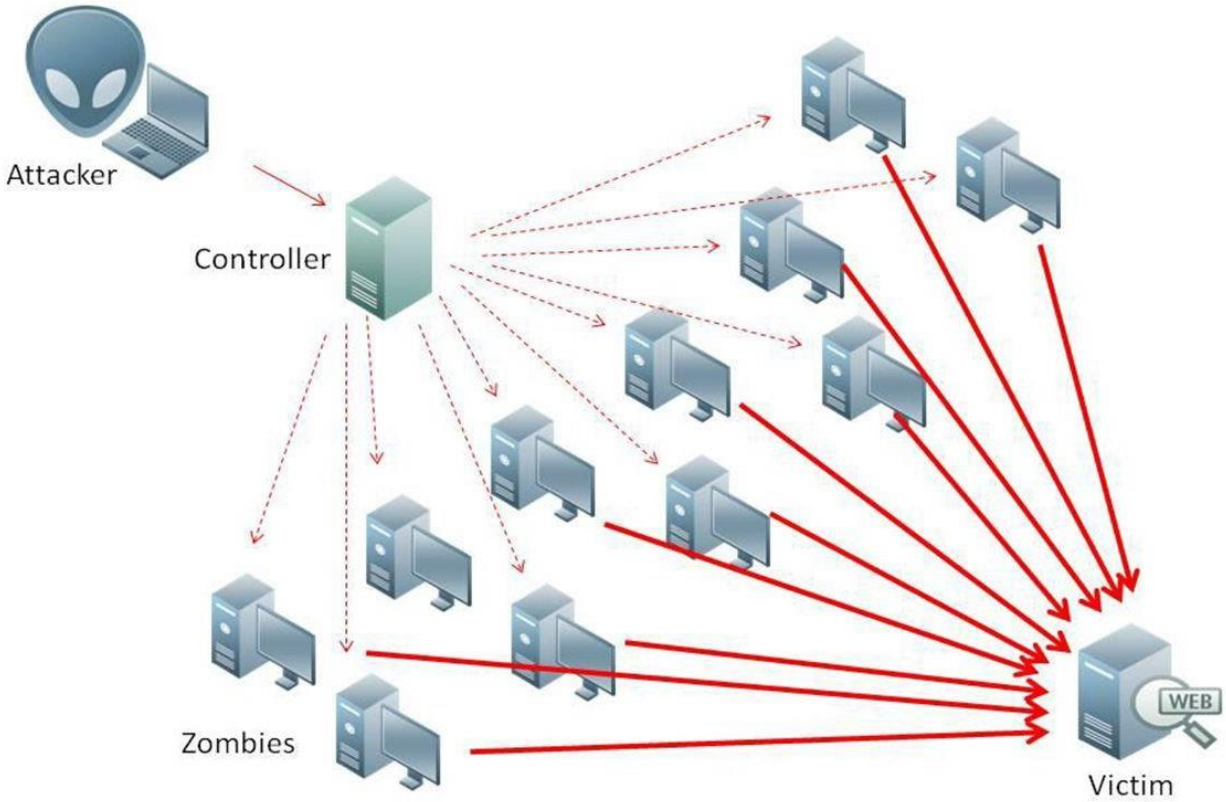
Inadvertent insiders

Cause damage to the company unintentionally

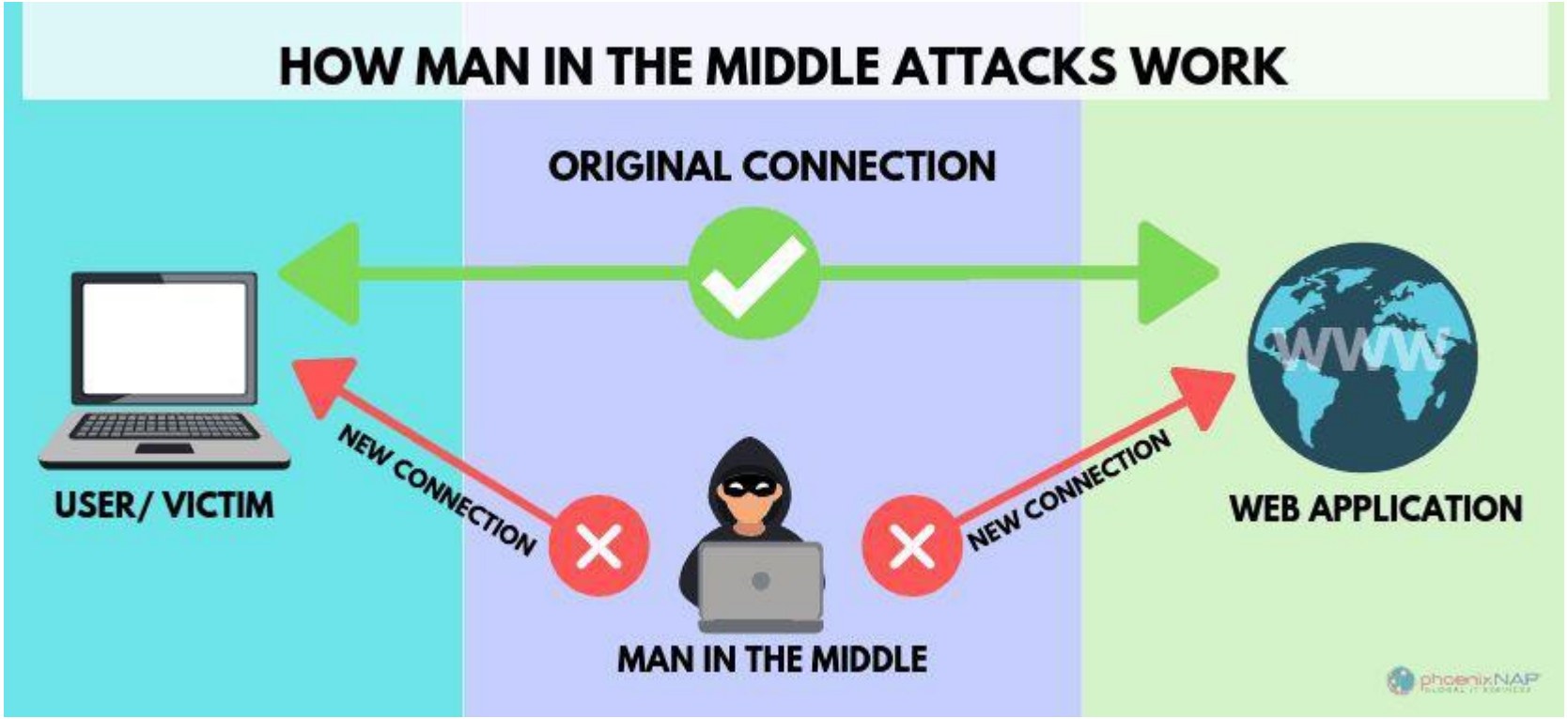
Source: https://www.ekransystem.com/en/blog/insider-threat-definition

Password Attacks

Denial-of-Service (DOS)



Man-In-The-Middle



SQL Injections

Image Source: https://www.avast.com/c-sql-injection

Zero-Day Exploit

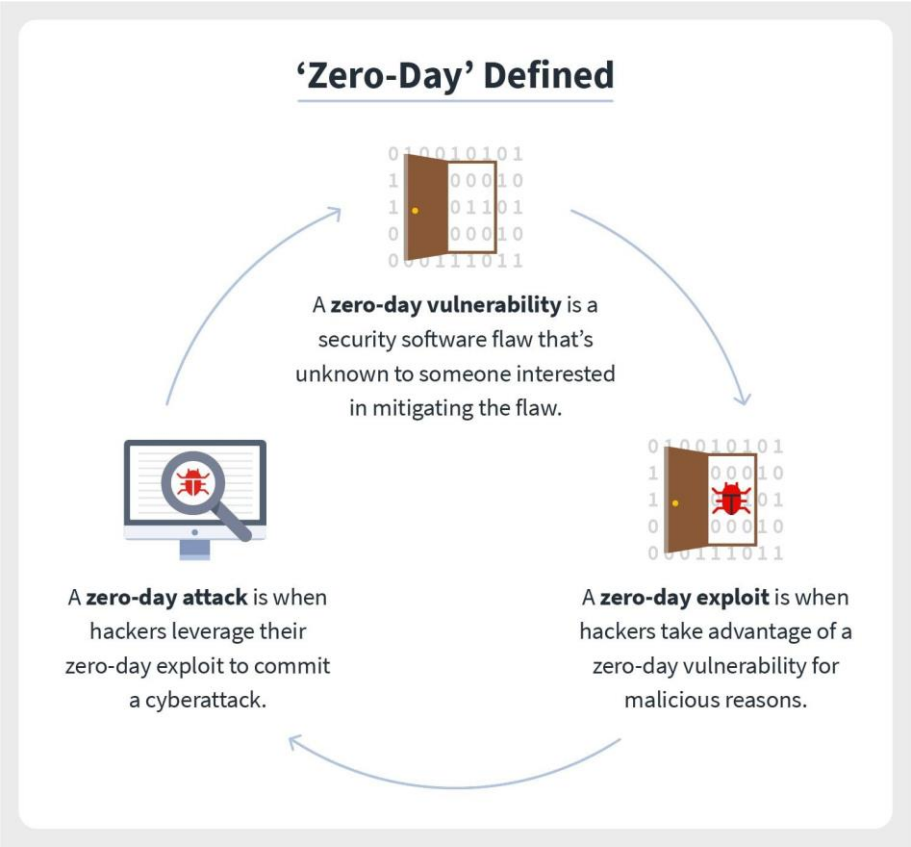


Image Source: https://us.norton.com/internetsecurity-emerging-threats-how-do-zero-day-vulnerabilities-work.html

Data Security Activities

1. Identify Data Security Requirements
2. Define Data Security Policy
3. Define Data Security Standards

Source: DAMA International. (2024). *Data Management Body of Knowledge (DMBOK)*, 2nd Revised Edition. Chapter 7: Data Security

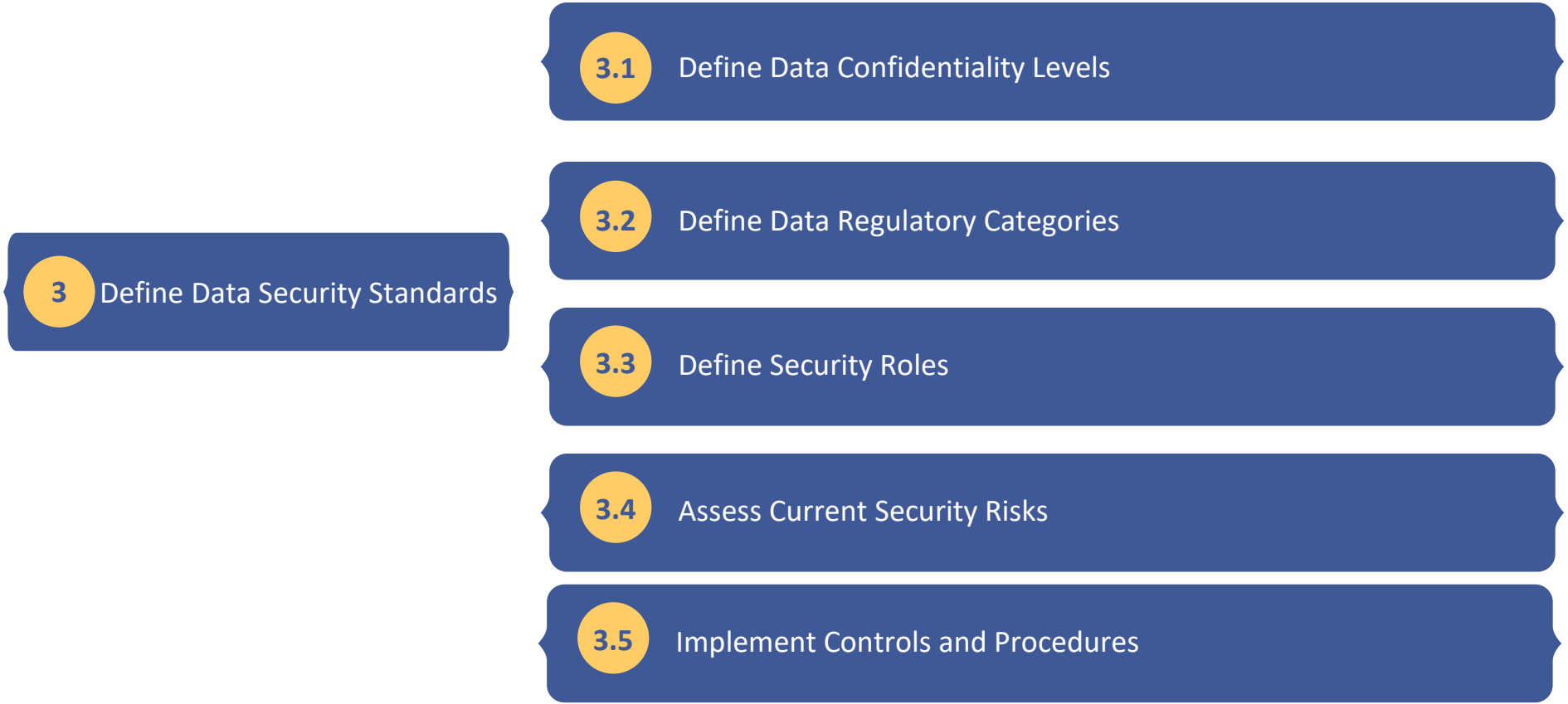
Data Security Activities

Source: DAMA International. (2024). *Data Management Body of Knowledge (DMBOK)*, 2nd Revised Edition. Chapter 7: Data Security

Data Security Activities

Source: DAMA International. (2024). *Data Management Body of Knowledge (DMBOK)*, 2nd Revised Edition. Chapter 7: Data Security

Data Security Activities



Source: DAMA International. (2024). *Data Management Body of Knowledge (DMBOK)*, 2nd Revised Edition. Chapter 7: Data Security

Data Security Tools

Source: DAMA International. (2024). *Data Management Body of Knowledge (DMBOK)*, 2nd Revised Edition. Chapter 7: Data Security

9 Best Practices to secure your data

1. Employees education
2. Create Insider Threat Policies
3. Phishing Simulations
4. Backup data
5. Update Systems and Software
6. Utilize HTTPS
7. Maintain Compliance
8. Use multi-factor authentication
9. Employe latest secure coding practices

Module 8 - Data Integration & Interoperability

What is Data Integration

‘**Data Integration** is the process of consolidating data from different sources into one, unified view for efficient data management”

Example of Data Integration

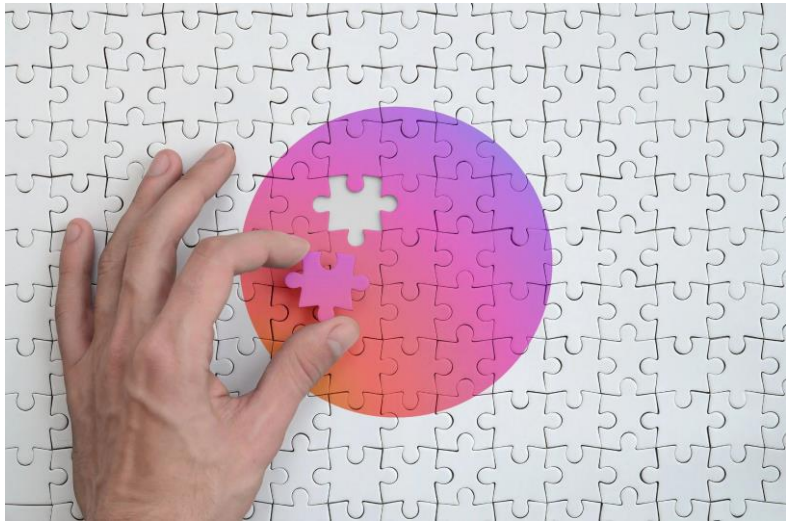
Company A uses the following:

- Salesforce for customer information and sales pipeline data
- External vendor for additional customer firmographics
- Internal database that tracks customer satisfaction ratings from surveys
- Internal financial system that tracks the sales revenue per customer
- Marketing department database on customer campaigns

Goal: Integration all the above sources into single source!

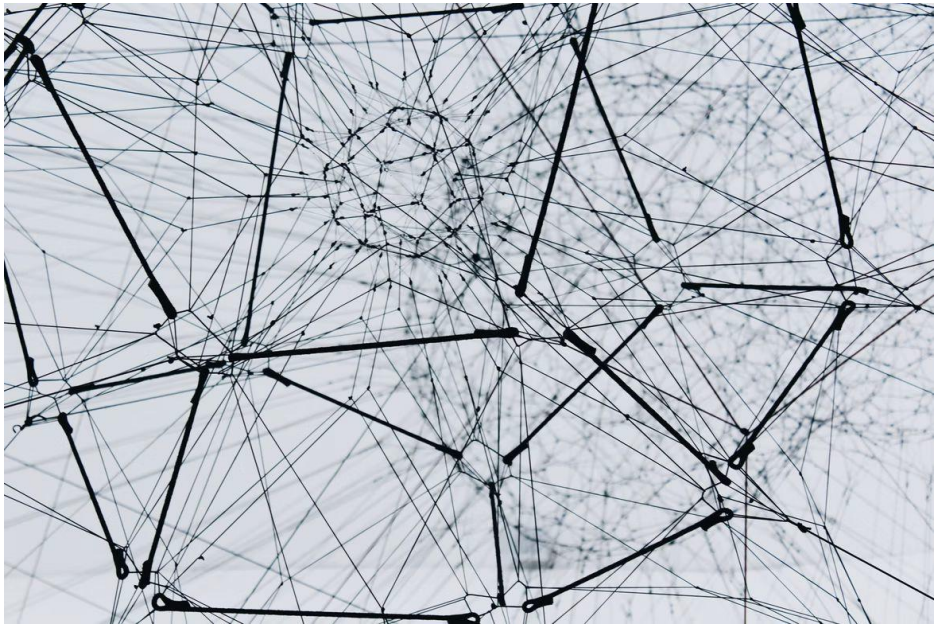
Importance of Data Integration

- Complete view of business intelligence, insights and analytics
- Increased efficiency and ROI - no need for manual data gathering
- Better employee, customer and partner experience
- Improves Collaboration
- Eliminates Data Silos
- Reduces errors
- Faster innovation, sales, time to market
- Improves Data Quality and Integrity



Techniques for Data Integration

- Manual Data Integration
- Middleware Data Integration
- Application Based Integration
- Uniform Access Integration
- Common Storage Integration (Data Warehousing)
- Data Virtualization



Manual Data Integration

Pros:

- Low cost
- Total control

Cons:

- Difficult to scale
- Human error

Source: https://www.talend.com/resources/data-integration-methods/

Middleware Data Integration

Pros:

- Better data streaming
- Easier access between systems

Cons:

- Maintenance
- Limited functionality

Source: https://www.talend.com/resources/data-integration-methods/

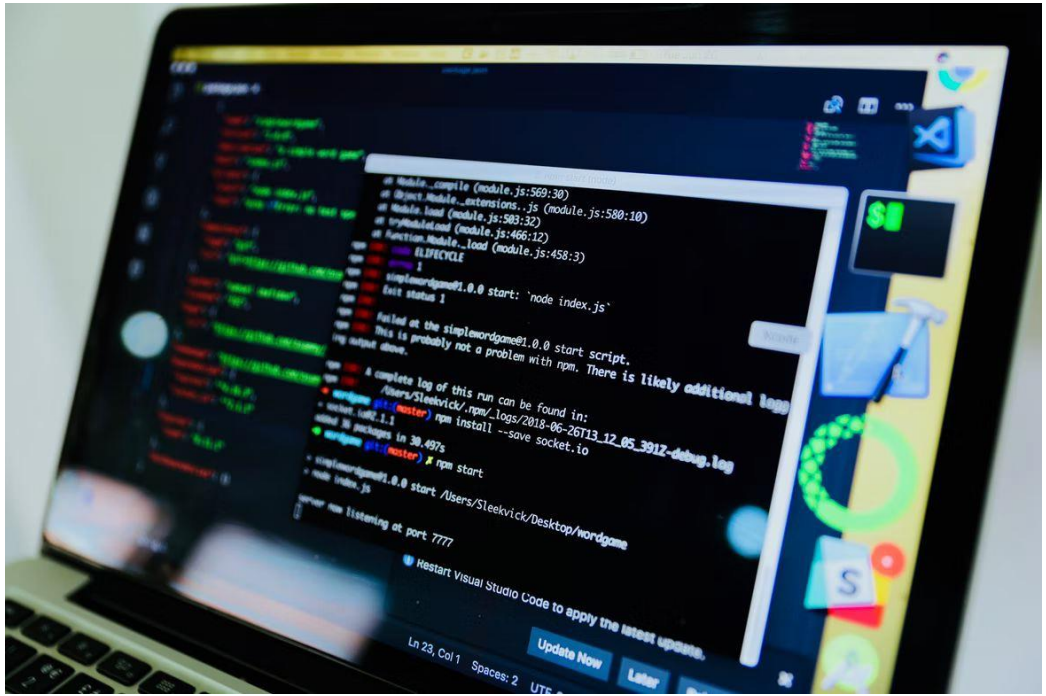
Application Based Integration

Pros:

- Simplified processes
- Easier information exchange
- Fewer resources used

Cons:

- Maintenance
- Inconsistent results
- Complicated setup
- Difficult data management



Source: https://www.talend.com/resources/data-integration-methods/

Uniform access Integration

Pros:

- Lower storage requirements
- Easier data access
- Simplified view of data

Cons:

- Data integrity challenges
- Strained systems

Source: https://www.talend.com/resources/data-integration-methods/

Common Storage Integration (Data Warehousing)

Pros:

- Reduced burden
- Increased data version management control
- Cleaner data
- Enhanced data analytics

Cons:

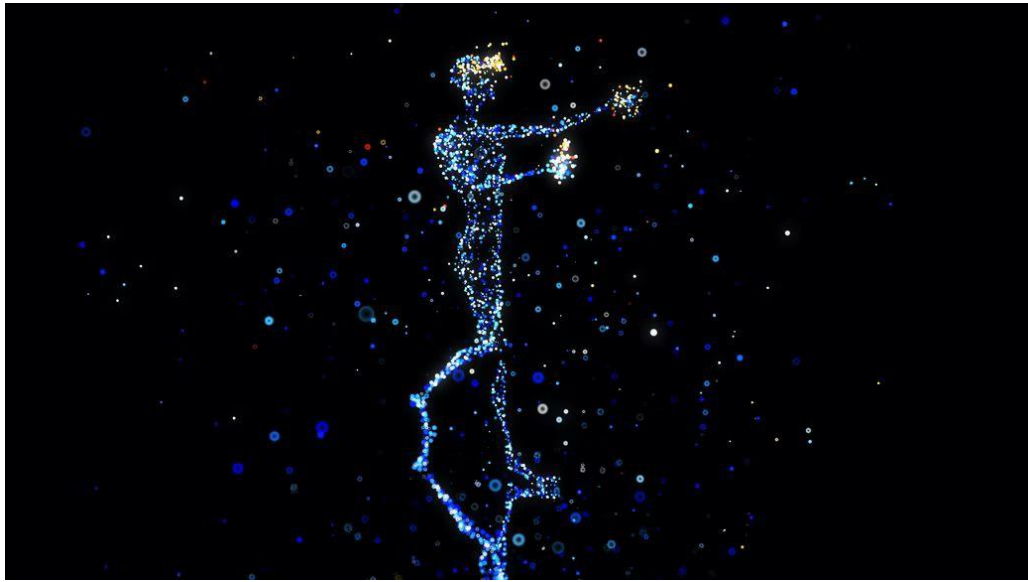
- Increased storage costs
- Higher maintenance costs

Source: https://www.talend.com/resources/data-integration-methods/

Data Virtualization

Pros:

- No need to move data
- Scalable
- No need to maintain data in multiple locations

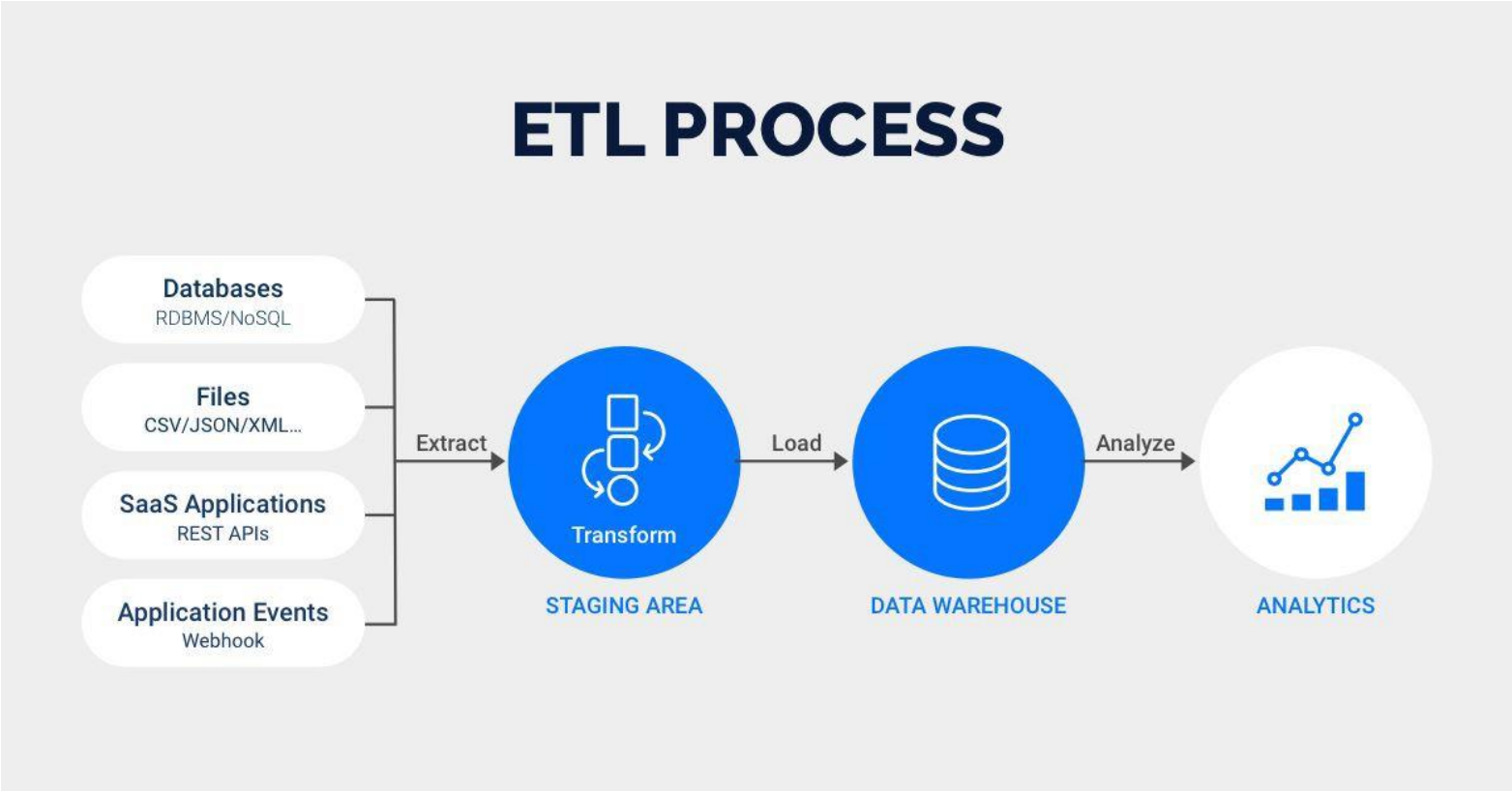


Cons:

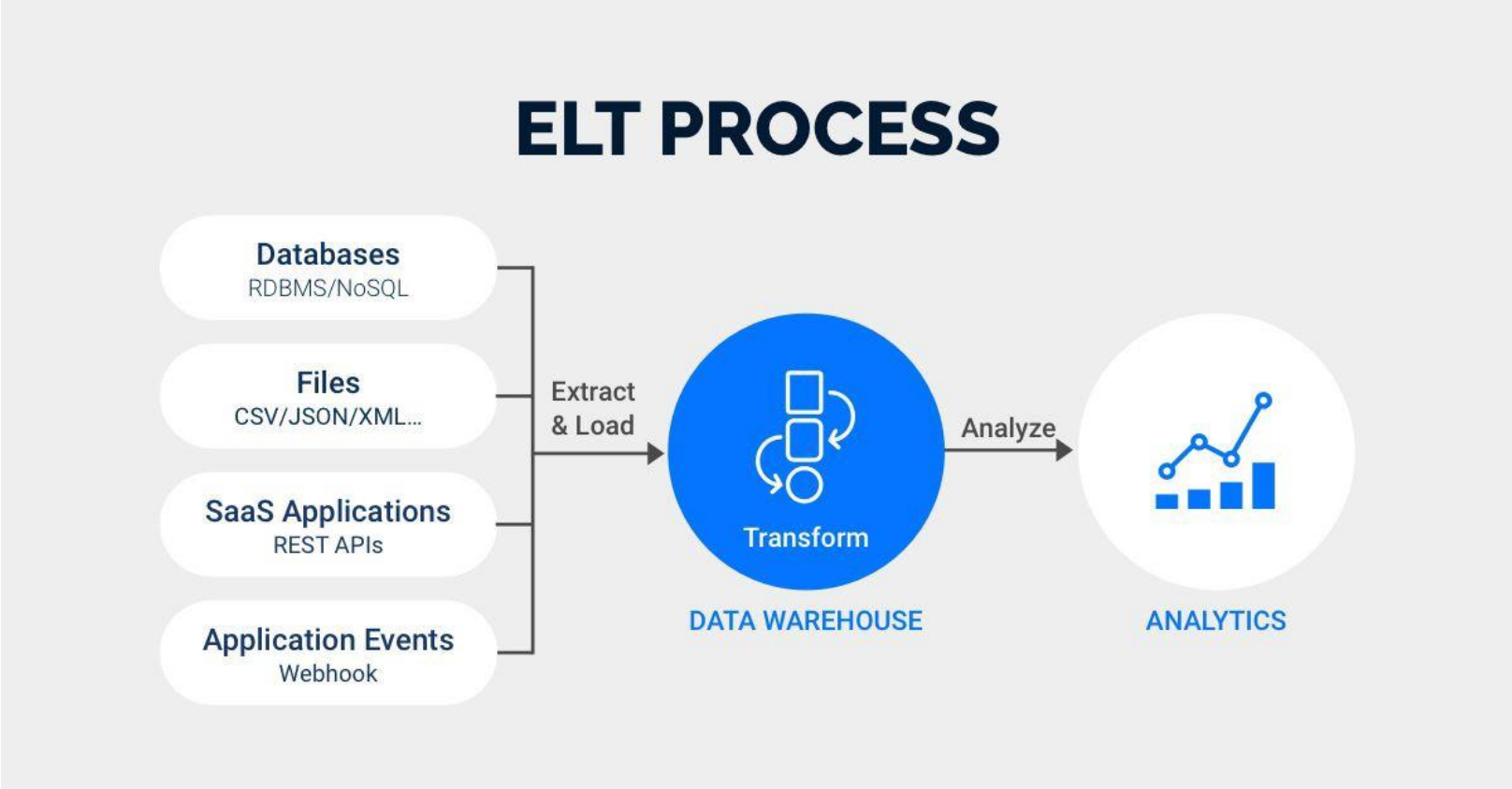
- Cost

Source: https://www.talend.com/resources/data-integration-methods/

ETL


ETL PROCESS
Databases
RDBMS/NoSQL
Files
CSV/JSON/XML...
SaaS Applications
REST APIs
Application Events
Webhook
Extract
Transform
STAGING AREA
Load
DATA WAREHOUSE
Analyze
ANALYTICS

ELT


ELT PROCESS
Databases
RDBMS/NoSQL
Files
CSV/JSON/XML...
SaaS Applications
REST APIs
Application Events
Webhook
Extract
& Load
Transform
DATA WAREHOUSE
Analyze
ANALYTICS

ETL vs ELT

Pros of ETL:

- **Compliance -** ETL is better for compliance with GDPR, HIPAA, CCPA and other standards
- **Implementation -** many ETL tools and experts
- **Maturity -** more documentation, tools and best practices

Pros of ELT:

- **Maintenance -** all data is always flowing with automated process
- **Speed -** data available in Data Warehouse faster since no transformation layer
- **Cost -** only cloud-based platforms needed at lower cost. On premise ETL processes require expensive hardware

Data Integration Tools

On-premise tools

- Oracle Data Service Integrator
- Informatica PowerCenter
- IBM InfoSphere Information Server

Cloud-based tools

- SnapLogic
- Talend Cloud Integration

Open-source tools

- Talend Open Studio
- Tibco Jaspersoft

Data Integration Best Practices

1. Identify your business needs first
2. Include internal business expert in the data integration team
3. Consider Long-Term Goals
4. Take into consideration the total cost of all methods
5. Avoid very complex data integration solutions
6. Choose a flexible solution

Module 9 - Document & Content Management

What is Document & Content Management

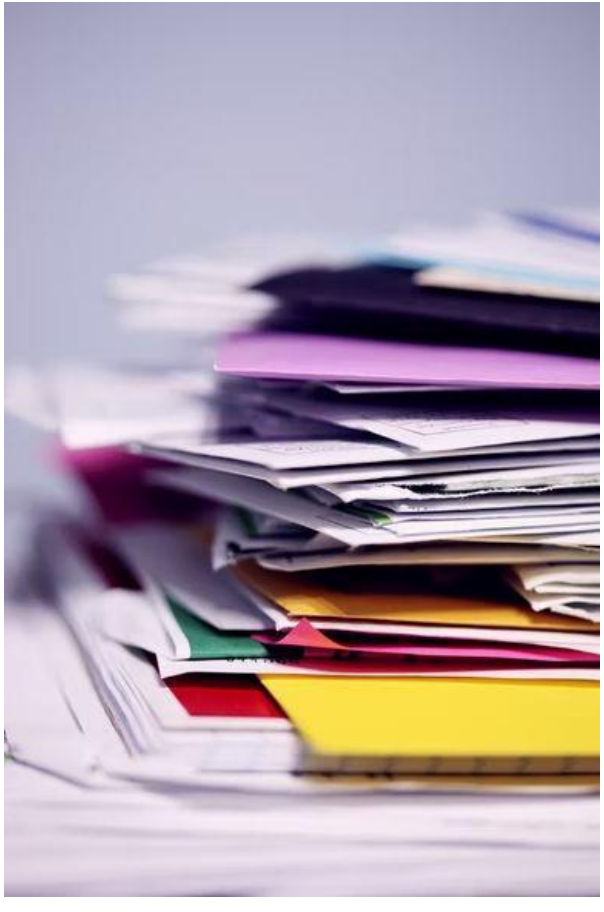
‘**Document & Content management** is the process of establishing planning, implementation and control activities for lifecycle management of data and information found in any form or medium - outside of relational databases”

Why do we need Document & Content Management?

- Comply with legal obligations
- Comply with customer expectations regarding records management
- Effective and efficient storage, retrieval and use of documents and content
- Integration between structured and unstructured content

What is a DMS?

- DMS stands for Document Management System
- What are the benefits?
- Types of DMS



What is a CMS?

- CMS stands for Content Management System
- What are the benefits?
- Types of CMS

ECMS

- ECMS stands for Enterprise Content Management System
- What are the benefits?
- CMS vs ECM

Figure 1: Magic Quadrant for Content Services Platforms

Source: Gartner (October 2021)

Document Management vs Enterprise Content Management

Comparison	Document Management System (DMS)	Enterprise Content Management System (ECMS)
Type of Data	Structured data in traditional formats (Word, PDF, PowerPoint, Excel, etc)	Structured + unstructured data such as images, audito, video files, HTML, etc
Main purpose	Workflow management and regulatory compliance	Storage, retrieval and publishing of content
Key difference	DMS is a software	ECM is a set of tools and processes. ECM is a broader version of DMS
Company size	DMS only solution can work well for small companies	ECM solution needed in bigger organizations

Module 10 - Master & Reference Data Management

What is Master Data

DAMA Guide to Data Management Body of Knowledge: “Master Data represents data about the business entities that provide context for business transactions”

Gartner: ‘Master Data is the consistent and uniform set of identifiers and extended attributes that describes the core entities of the enterprise including customers, prospects, citizens, suppliers, sites, hierarchies and chart of accounts”

What is Reference Data

DAMA Guide to Data Management Body of Knowledge: “Reference data is data used to classify or categorize other data”

Examples of Reference Data:

- Postal codes
- Language codes
- Customer segments
- Country codes
- Cost centers

Master Data vs Reference Data

Comparison	Master Data	Reference Data
Main purpose	Represents the business objects which contain the most valuable, agreed upon information shared across the organization	Data that defines the set of permissible values to be used by other data fields
More on usage	Master data is the data shared by multiple systems, applications, processes in the organization	Reference data is a type of master data that is used by other data fields
Examples	● Customer information - names, phone numbers and addresses ● Product information - product name and location ● Partner data - partner name and address	● Fixed conversion rates - weight, temperature, length, etc ● Currency codes ● Language codes ● Customer Segments ● Cost centers ● Postal codes ● Units of measurement

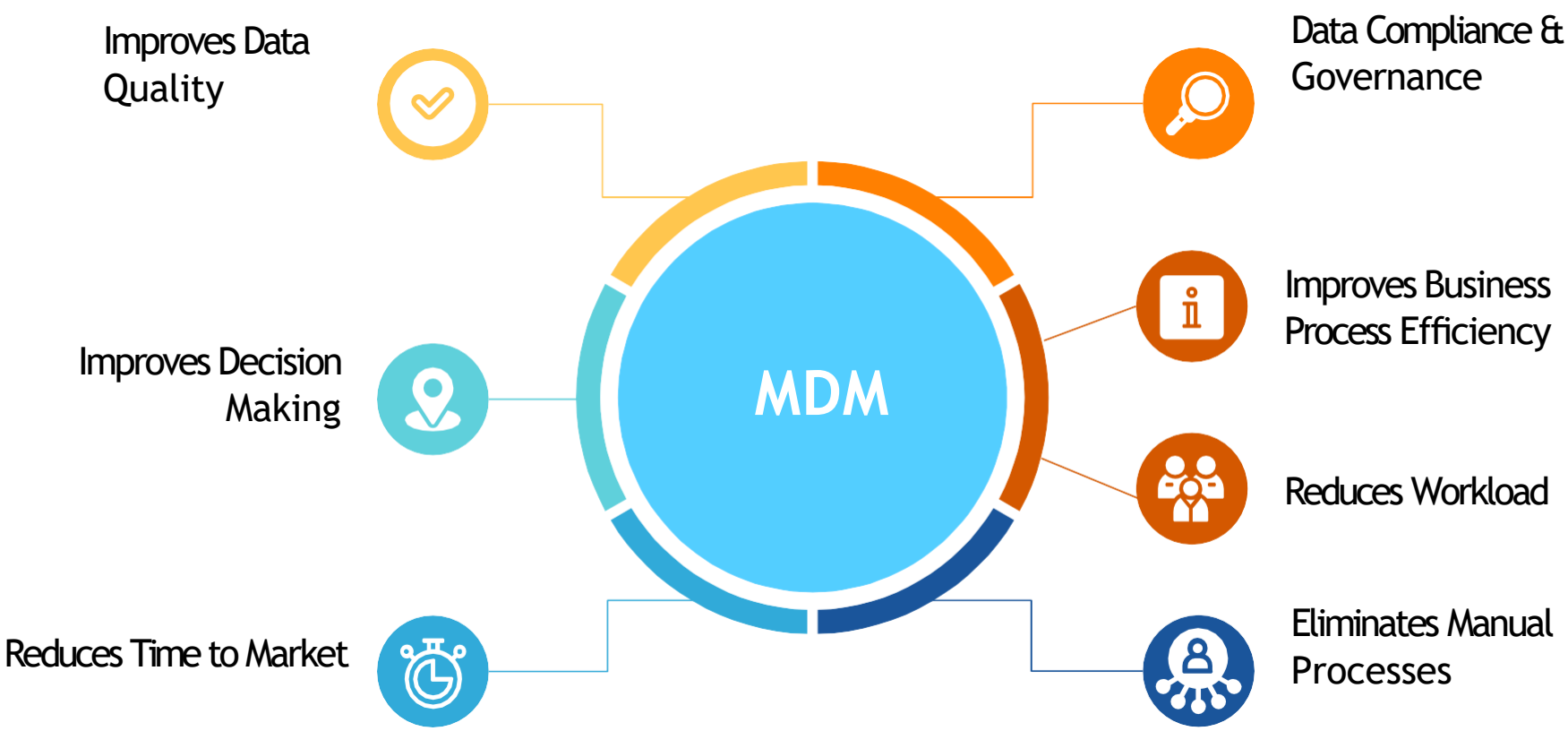
What is Master Data Management (MDM)

Master Data Management is the process of creating and maintaining a single master record - or single source of truth - for each person, place, and thing in a business.

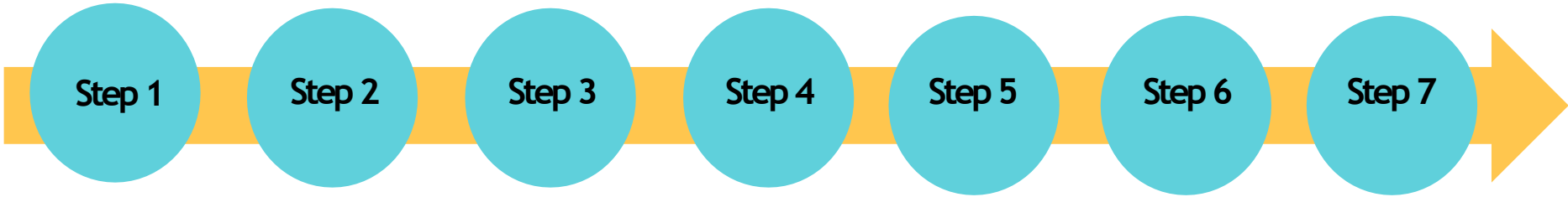
Through MDM, organizations gain a trusted, current view of key data that can be shared across the business and used for better reporting, decision-making, and process efficiency.

Source - definition SAP. www.sap.com/insights/what-is-master-data-management-mdm

Why MDM is important



Steps to implement MDM



Learn about MDM:

- Important Capabilities
- Implementation Styles
- Domain Categories

Analyze your current data environment

Define your business goals

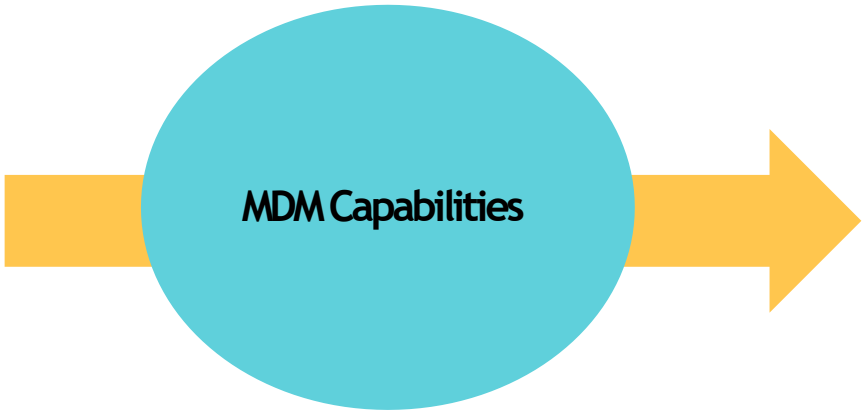
Secure the right sponsorship

Get the right team and tool

Implement a quick win

Repeat Step 6 and monitor

MDM Solution Capabilities



- Workflow/BPM
- Loading/Sync/Business Services
- Data Modeling
- Information Quality/Semantics
- Perform/Scale/Availability/Security
- Hierarchy Management
- Data Stewardship
- Data Governance
- Multiple Implementation Styles
- Multiple Usage Scenarios
- Multiple Domain and Multidomain
- Product Suite Internal Integration

Source: Gartner - Critical Capabilities for Master Data Management Solutions

MDM Implementation Styles

Consolidation

- **Ideal for reporting or analytics that reside in a BI/data warehouse**
- Nonintrusive to the business
- BI is the business platform
- Any industry
- Benefits dependent on success of BI strategy
- No attempt to clean up source data

Registry

- **Low control, autonomous environments**
- Nonintrusive to edge applications
- Emphasis is on remote data and application-to-application integration (lots of real-time network access)
- Distributed governance
- Faster to implement than coexistence and centralized

Coexistence

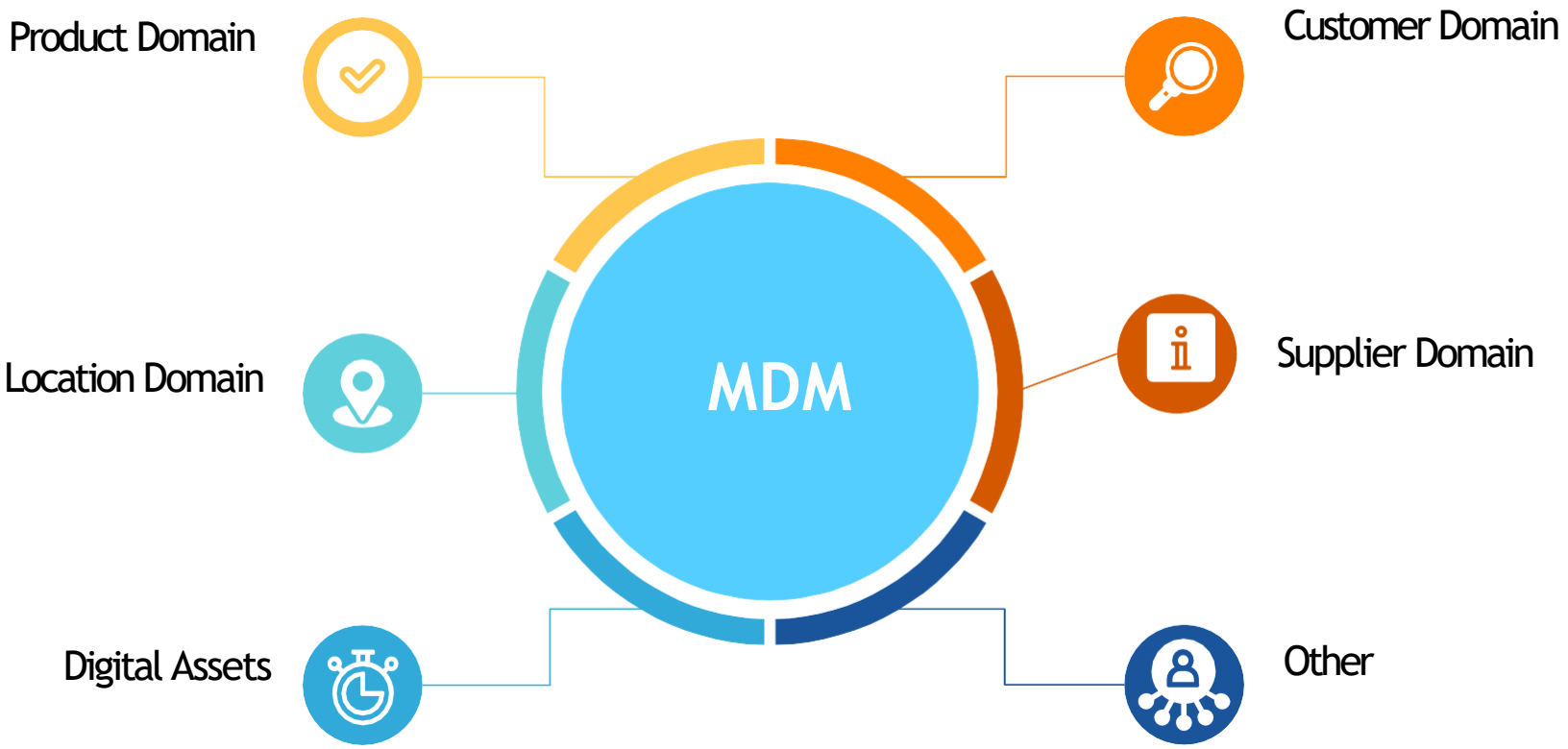
- **Large-scale distributed model**
- Largest change to information infrastructure
- Greatest need to mirror data
- Global and local governance
- Greatest risk over control, security
- Focused on shared services

Centralized

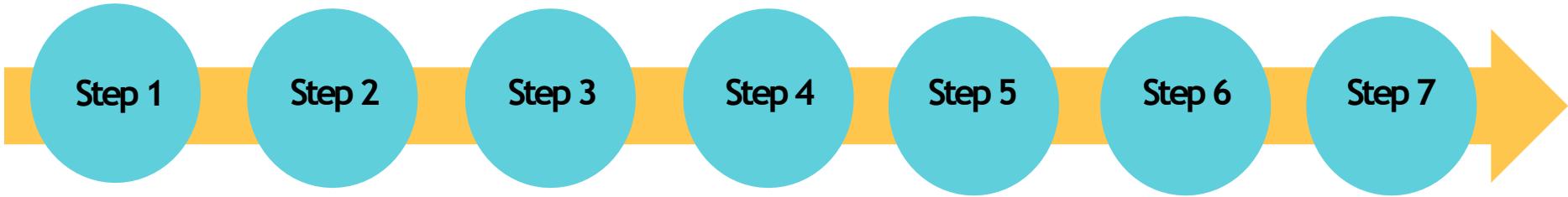
- **High-control, top-down environments**
- Largest change to application infrastructure
- Hugely invasive to the business
- Centralized governance
- Greatest control over access, security
- Focus on common services

ID: 351471

MDM Domains


Product Domain
Customer Domain
MDM
Location Domain
Supplier Domain
Digital Assets
Other

Steps to implement MDM



Learn about MDM:
- Important Capabilities
- Implementation Styles
- Domain Categories

Analyze your current data environment

Define your business goals

Secure the right sponsorship

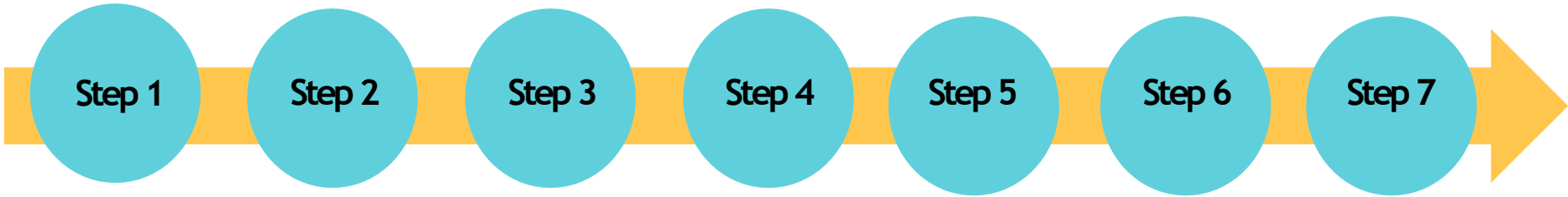
Get the right team and tool

Implement a quick win

Repeat Step 6 and monitor

1. What are the different company departments?
2. What kind of data do they use?
3. Where is the data coming from?
4. Is it well integrated with the company tools/apps and central repositories?
5. Any data silos?
6. What are the main data problems for the business?
7. Are there data quality controls in place?
8. How is the data being governed?
9. Who are the main data stewards in the department?
10. Any documentation that will help the MDM program?

Steps to implement MDM



Learn about MDM:
- Important Capabilities
- Implementation Styles
- Domain Categories

Analyze your current data environment

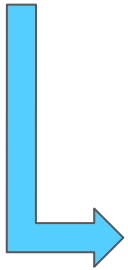
Define your business goals

Secure the right sponsorship

Get the right team and tool

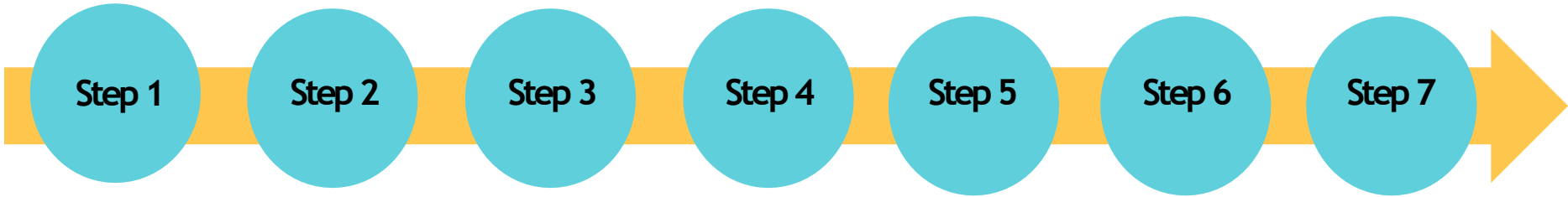
Implement a quick win

Repeat Step 6 and monitor



Some business drivers for MDM include:
- Up sell and cross sell opportunities
- Complete view of customers
- Improved Data Quality and business decisions
- Centralization of data
- Reduce costs of data maintenance and support
- Improved customer experience
- Other

Steps to implement MDM



Learn about MDM:
- Important Capabilities
- Implementation Styles
- Domain Categories

Analyze your current data environment

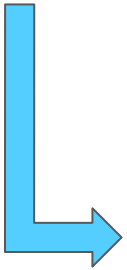
Define your business goals

Secure the right sponsorship

Get the right team and tool

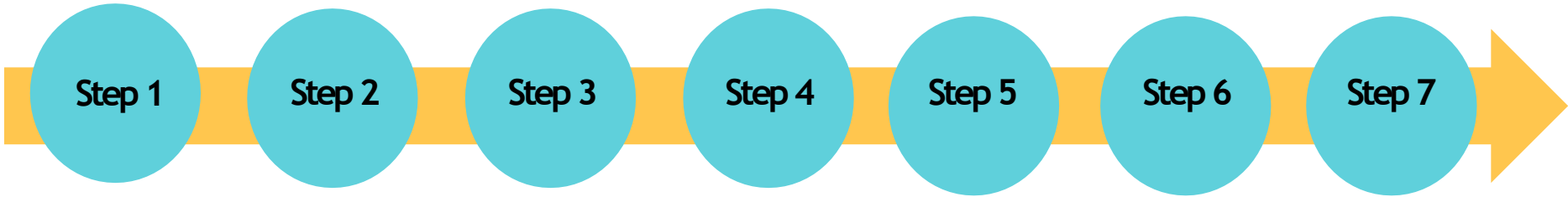
Implement a quick win

Repeat Step 6 and monitor



- MDM Programs can take years to implement
- Present your goals, strategy and success criteria
- Go to the top for long term budget commitment

Steps to implement MDM



Learn about MDM:
- Important Capabilities
- Implementation Styles
- Domain Categories

Analyze your current data environment

Define your business goals

Secure the right sponsorship

Get the right team and tool

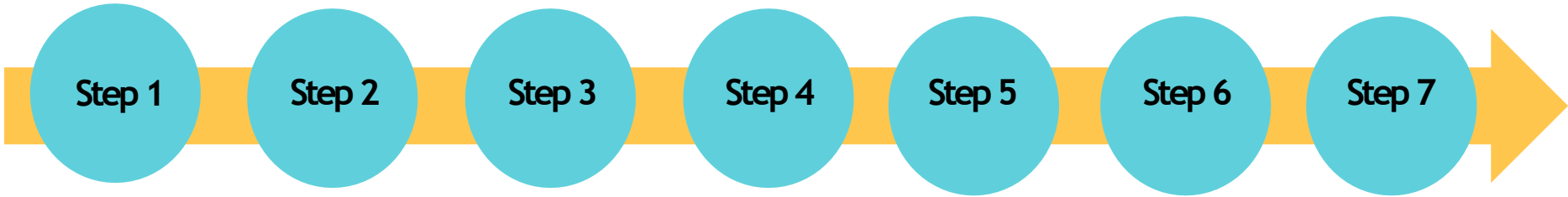
Implement a quick win

Repeat Step 6 and monitor

Get the Right Team:

- **Business Team -** *Sponsor (already covered), business analysts/SME, end users/data stewards
- **Program Team -** MDM Specialist, Data Architects, Program Manager
- **Tech Team -** DBA, Developers, Integration Experts, System Admins

Steps to implement MDM



Learn about MDM:
- Important Capabilities
- Implementation Styles
- Domain Categories

Analyze your current data environment

Define your business goals

Secure the right sponsorship

Get the right team and tool

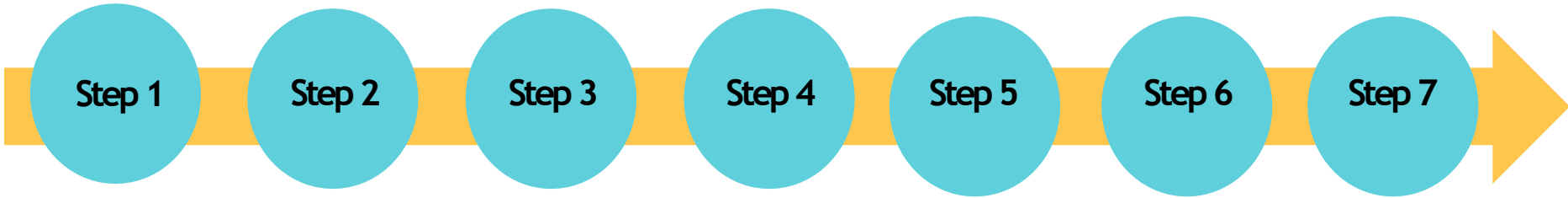
Implement a quick win

Repeat Step 6 and monitor

Get the Right MDM Solution (Master data management software that consists of policies, governance, standard tools and processes that facilitate defining and managing organization's data from a single point):

1. Understand your needs - features, functionality and business processes
2. Make a list of top 3-4 MDM solutions to further explore
3. Explore with a free trial

Steps to implement MDM



Learn about MDM:
- Important Capabilities
- Implementation Styles
- Domain Categories

Analyze your current data environment

Define your business goals

Secure the right sponsorship

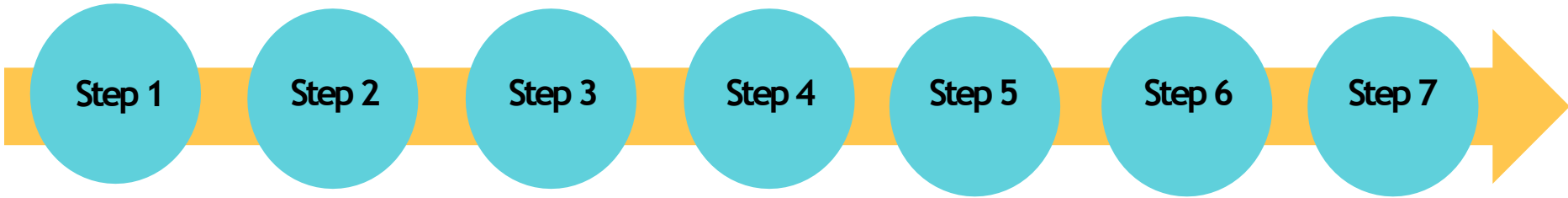
Get the right team and tool

Implement a quick win

Repeat Step 6 and monitor

- Start with a small pilot project to show the power of MDM (look for the stakeholders that were vocal about their data problems in Step 2)
- Leads to your first happy customers
- Secures your budget for the long term

Steps to implement MDM



Learn about MDM:
- Important Capabilities
- Implementation Styles
- Domain Categories

Analyze your current data environment

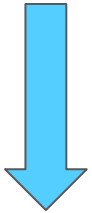
Define your business goals

Secure the right sponsorship

Get the right team and tool

Implement a quick win

Repeat Step 6 and monitor



- Maintain the long term vision and release “wins” at regular intervals
- Keep on tracking the KPIs
- Maintain leadership’s interest in the MDM Program

Module 11 - Data Warehousing & Business Intelligence

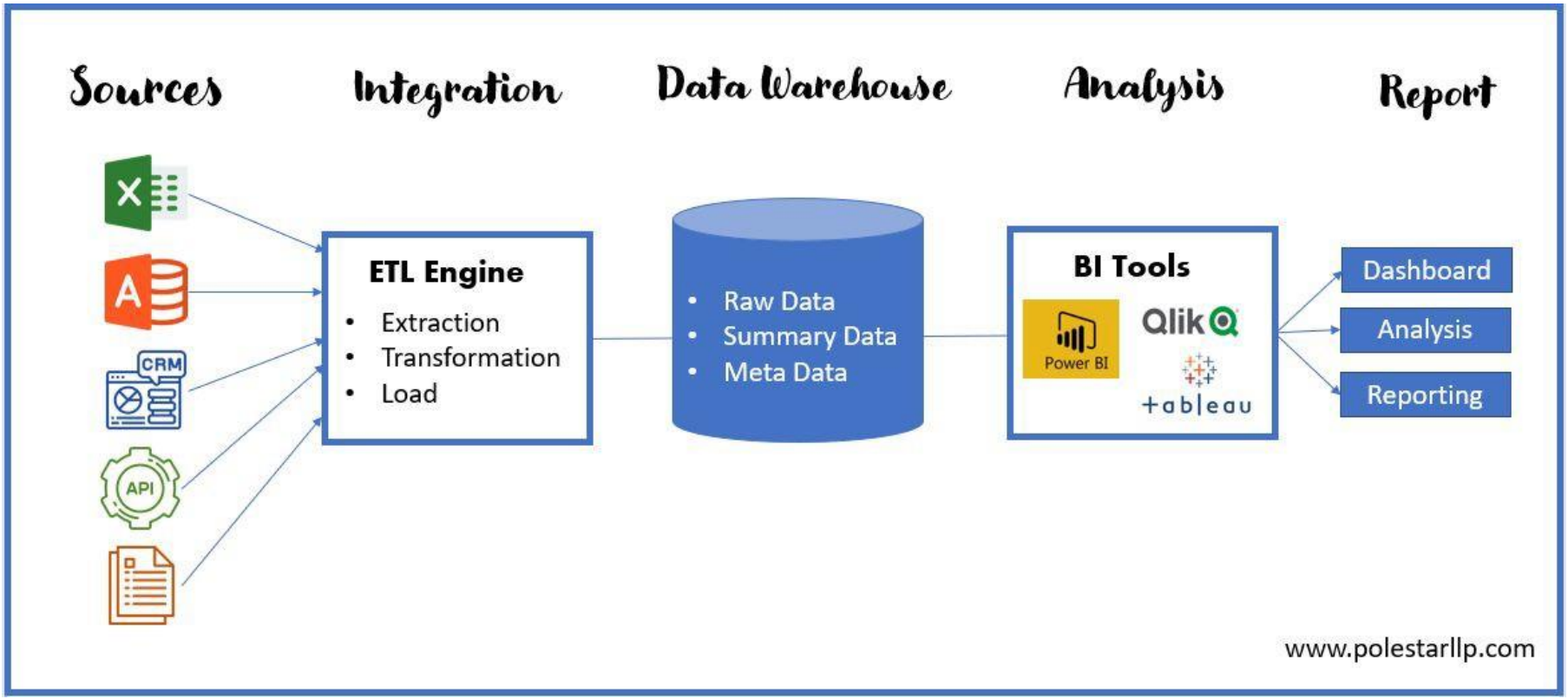
What is Data Warehousing and Business Intelligence

“Data warehousing is the secure electronic storage of information by a business or other organization. The goal of data warehousing is to create a collection of historical data that can be retrieved and analyzed to provide useful insight into the organization's
operations.
Data warehousing is a vital component of business intelligence. That wider term encompasses the information infrastructure that modern businesses use to track their past successes and failures and inform their decisions for the future.

”

Source: Jake Frankenfield - Investopedia

How the process works



What is a Data Warehouse (DW/DWH)

Data Warehouse (DW or DWH), and sometimes also referred to as Enterprise Data Warehouse (EDW) is a central repository of integrated data from one or more different sources.

Data Warehouses are used to store current and historical data in one single place.

Data Warehouse Components

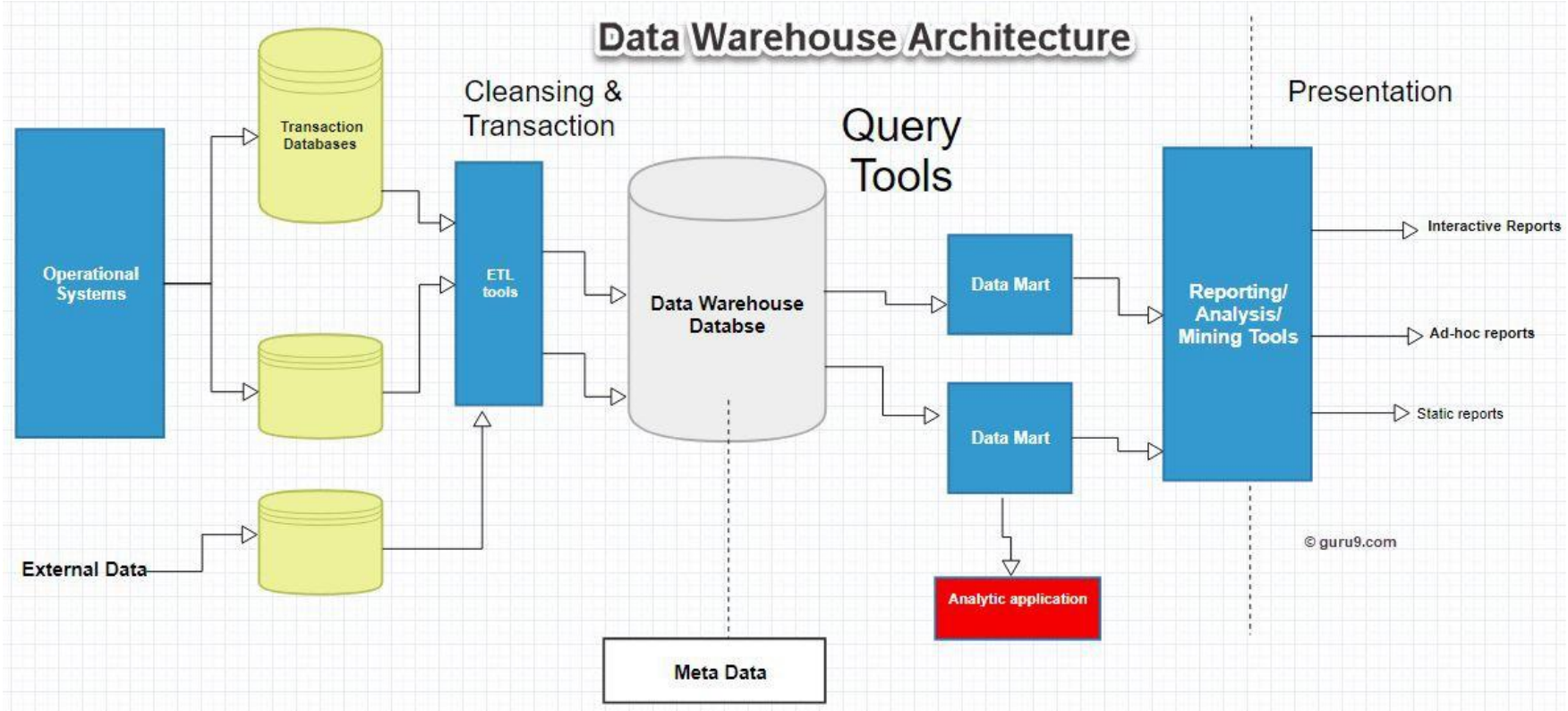


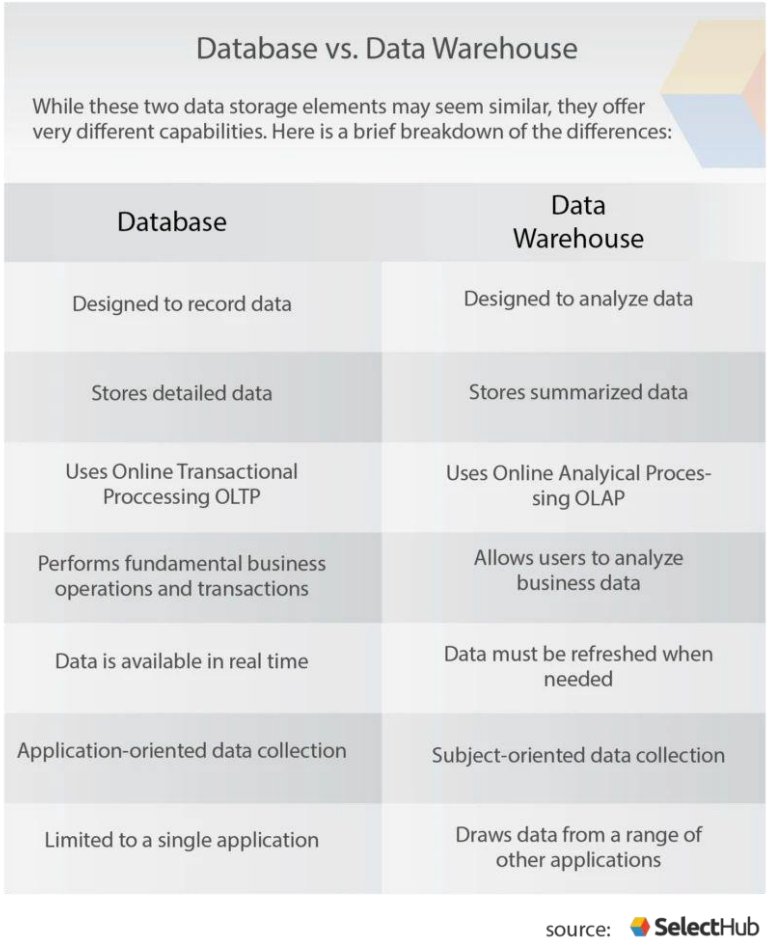
Image Source: guru99.com

What is a Data Warehouse (DW/DWH)

Data Warehouse (DW or DWH), and sometimes also referred to as Enterprise Data Warehouse (EDW) is a central repository of integrated data from one or more different sources.

Data Warehouses are used to store current and historical data in one single place.

Data Warehouse vs Database



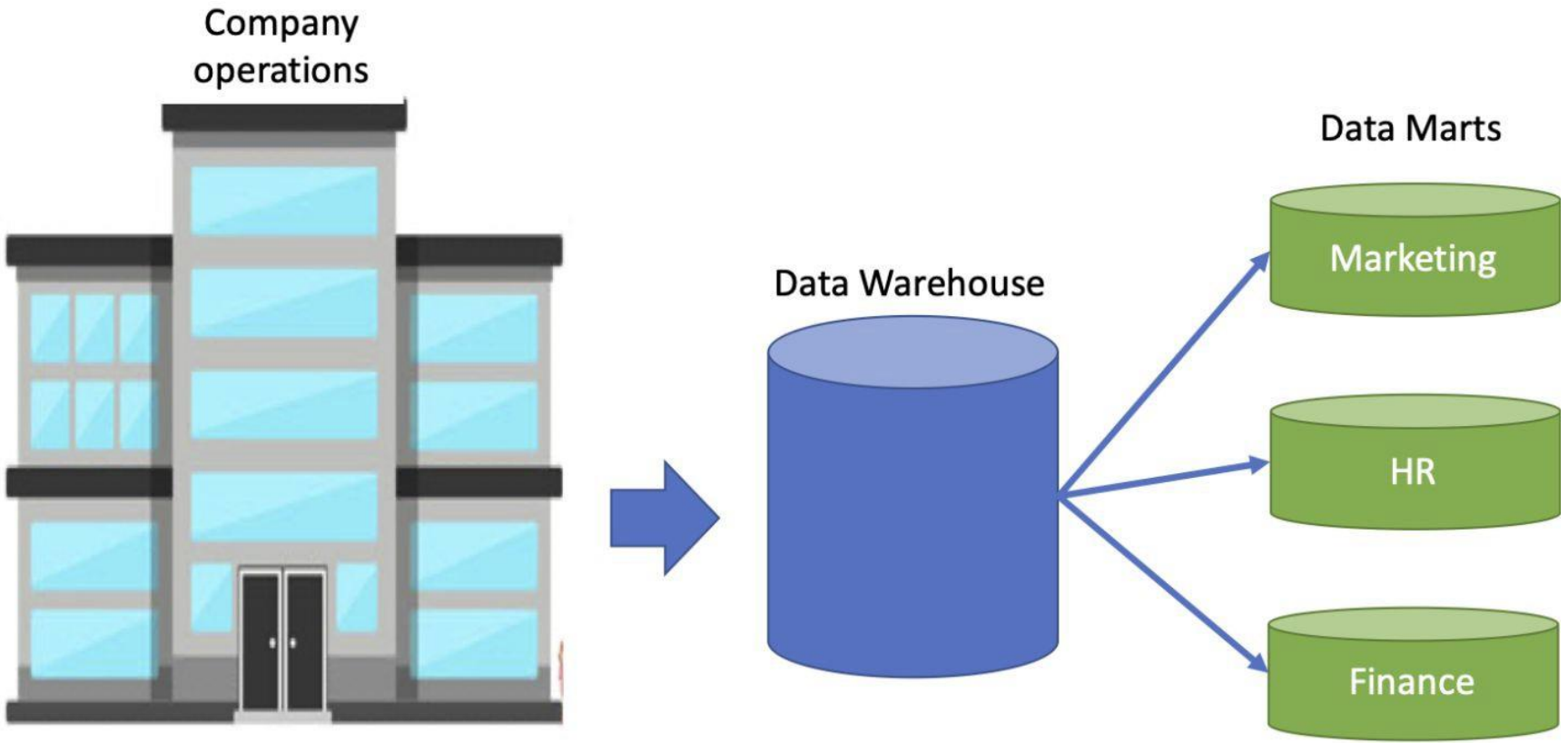
Database vs. Data Warehouse

While these two data storage elements may seem similar, they offer very different capabilities. Here is a brief breakdown of the differences:

Database	Data Warehouse
Designed to record data	Designed to analyze data
Stores detailed data	Stores summarized data
Uses Online Transactional Proccessing OLTP	Uses Online Analyical Proces-sing OLAP
Performs fundamental business operations and transactions	Allows users to analyze business data
Data is available in real time	Data must be refreshed when needed
Application-oriented data collection	Subject-oriented data collection
Limited to a single application	Draws data from a range of other applications

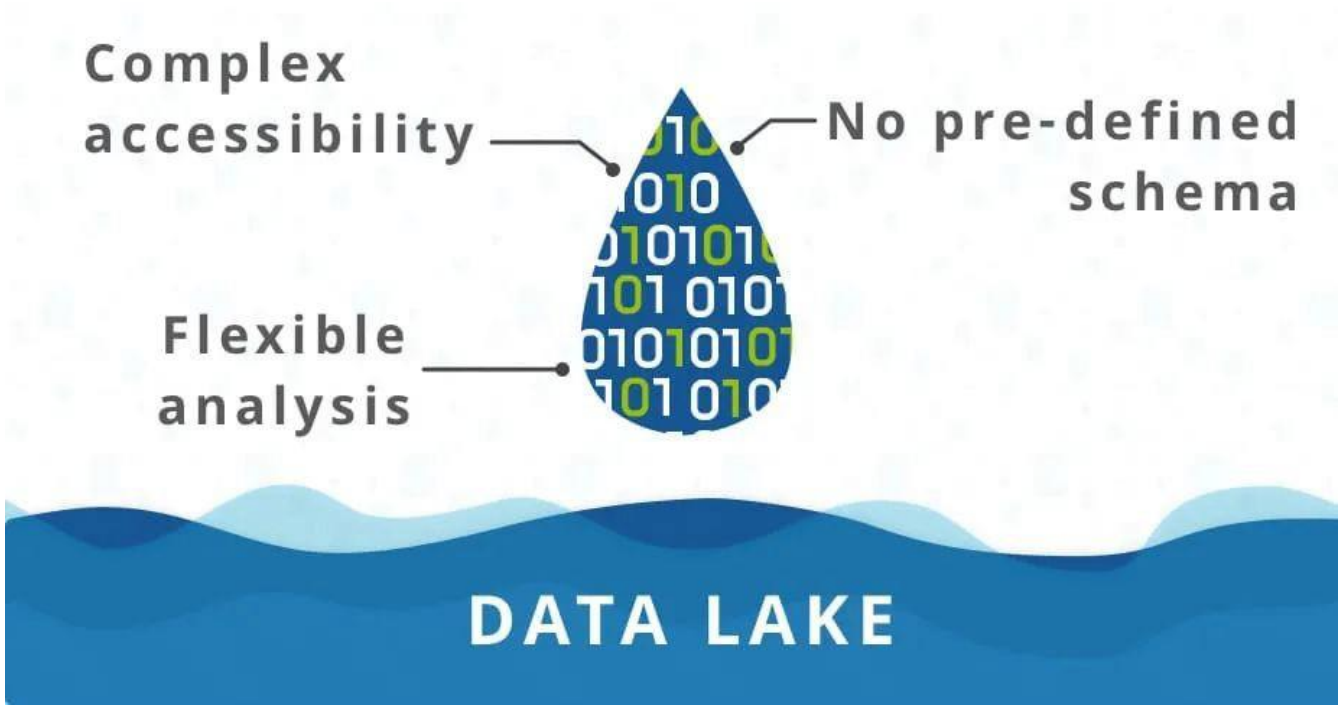
source: SelectHub

Data Warehouse vs Data Marts



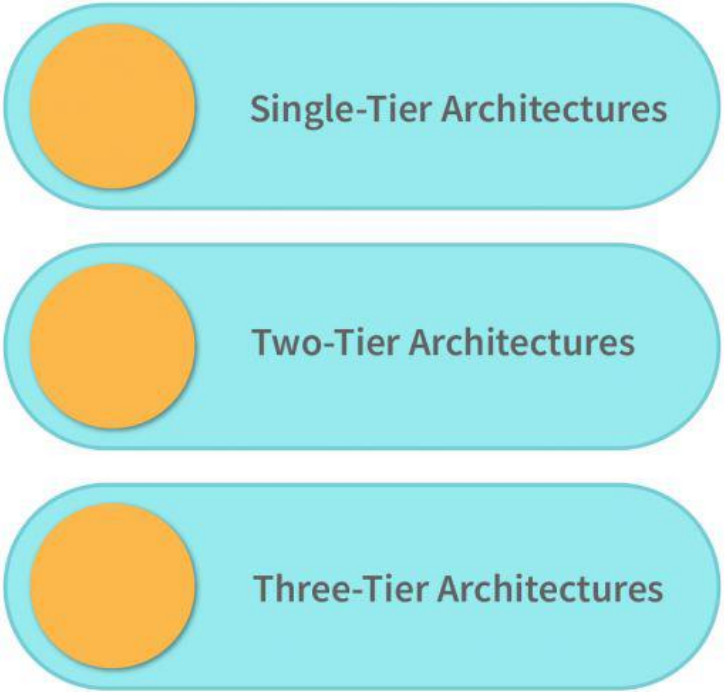
Data Warehouse vs Data Lake

A data lake is a central storage repository that holds big data from many sources in a raw, granular format.



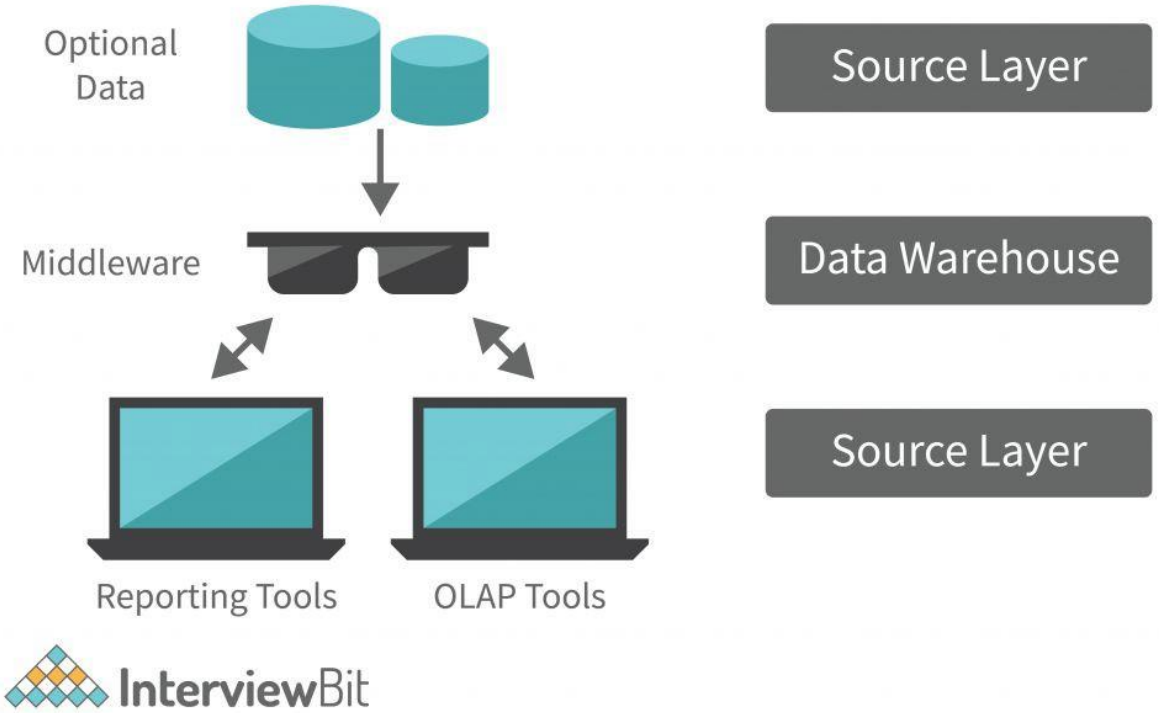
Source: https://www.talend.com/resources/what-is-data-lake/

Data Warehouse Architecture Types



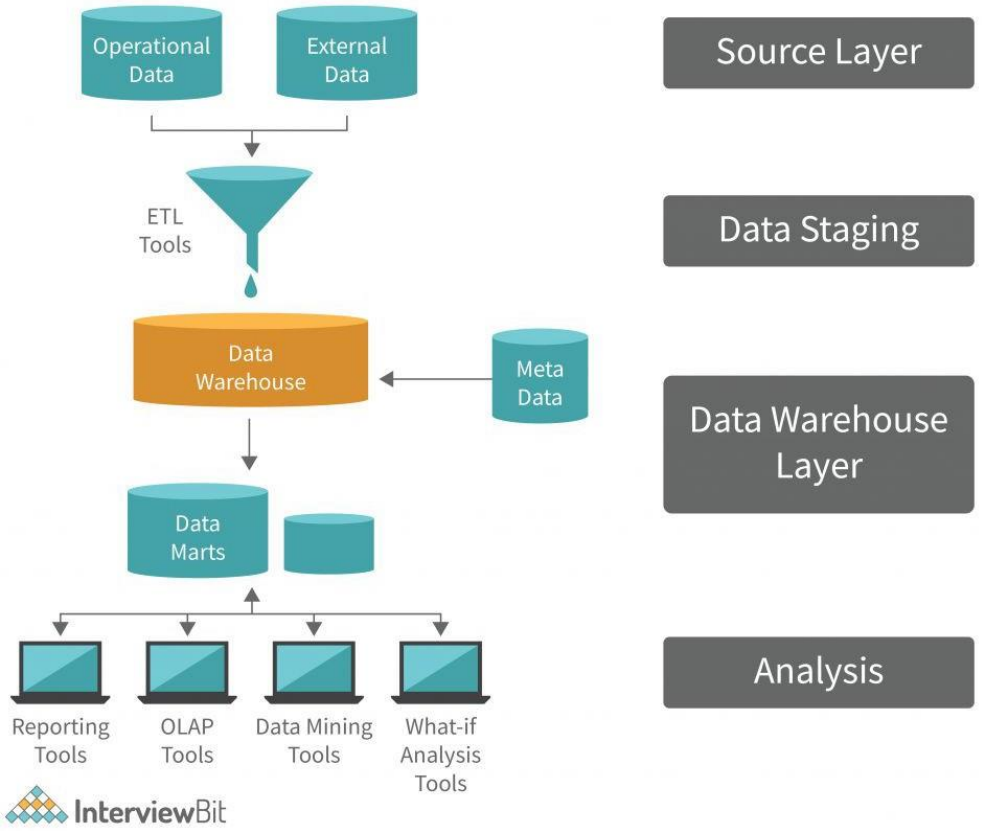
Source: https://www.interviewbit.com/blog/data-warehouse-architecture/

Single-Tier Data Warehouse Architecture



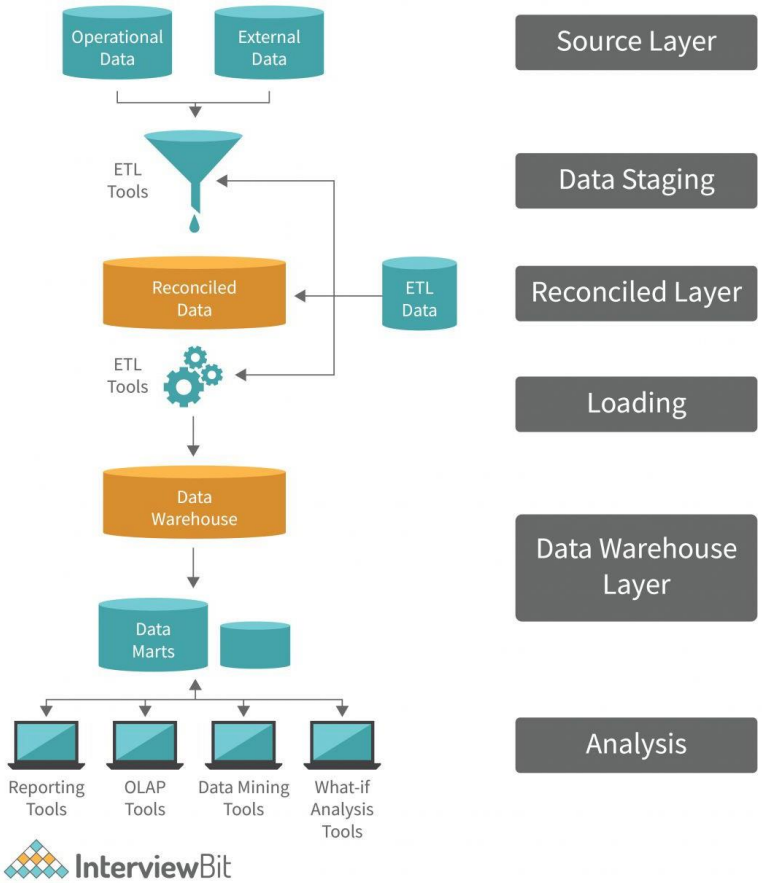
Source: https://www.interviewbit.com/blog/data-warehouse-architecture/

Two-Tier Data Warehouse Architecture



Source: https://www.interviewbit.com/blog/data-warehouse-architecture/

Three-Tier Architecture for a Data Warehouse System



Source: https://www.interviewbit.com/blog/data-warehouse-architecture/

What is Business Intelligence

"Business intelligence (BI) leverages software and services to transform data into actionable insights that inform an organization's strategic and tactical business decisions."

Source: www.cio.com

Business Intelligence vs Business Analytics

“The primary distinction between business intelligence and business analytics is the focus on when events occur. **Business intelligence is focused on current and past events that are captured in the data. Business analytics is focused on what's most likely to happen in the future**..**”**

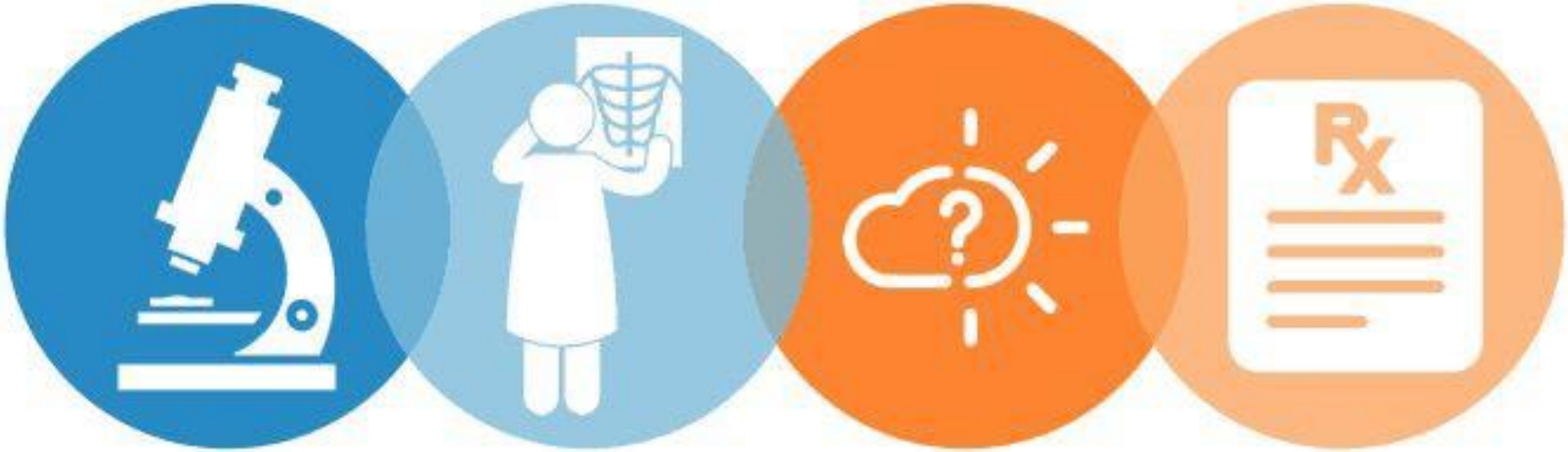
Source: www.mixpanel.com

Applications of Business Intelligence

1. Sales Intelligence
2. Visualization
3. Reporting
4. Performance Management

Source: https://www.selecthub.com/business-intelligence/4-key-bi-applications

Categories of BI analysis



Descriptive
Explains what happened.

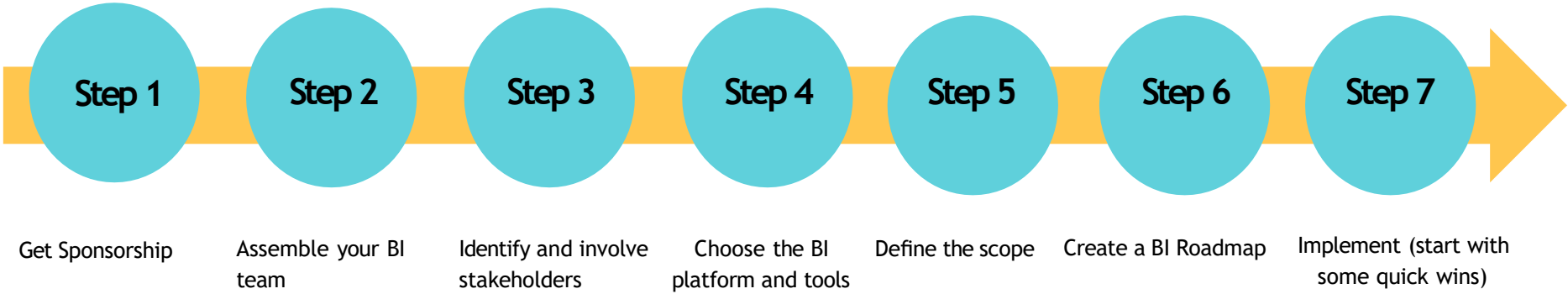
Diagnostic
Explains why it happened.

Predictive
Forecasts what might happen.

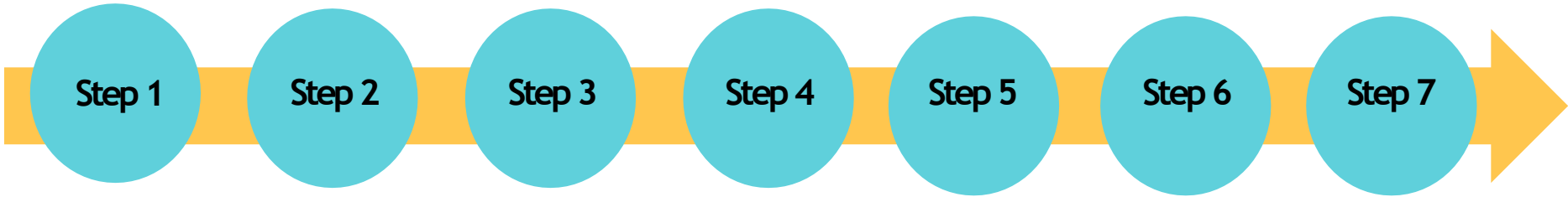
Prescriptive
Recommends an action based on the forecast.

Source: https://www.analyticsinsight.net/four-types-of-business-analytics-to-know/

Create a Business Intelligence Strategy



Create a Business Intelligence Strategy



Get Sponsorship

Assemble your BI team

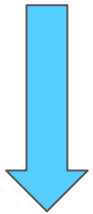
Identify and involve stakeholders

Choose the BI platform and tools

Define the scope

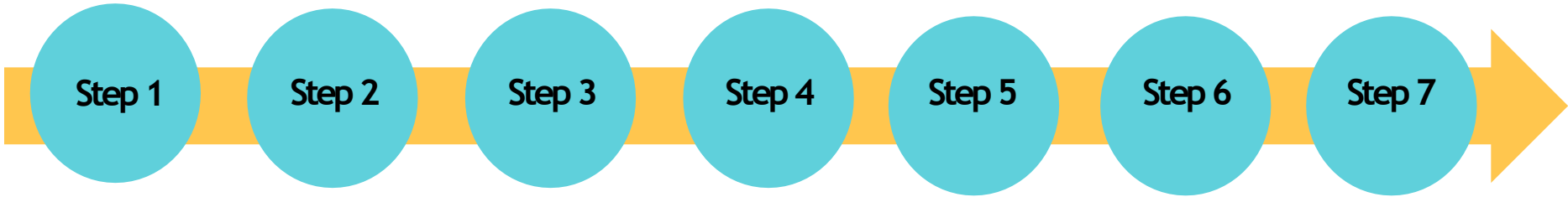
Create a BI Roadmap

Implement (start with some quick wins)



- Executive level sponsorship is key
- Update sponsor regularly
- Quick wins will help

Create a Business Intelligence Strategy



Get Sponsorship

Assemble your BI team

- Program Manager
- IT Owner
- Enterprise Architect
- Data Stewards
- etc

Identify and involve stakeholders

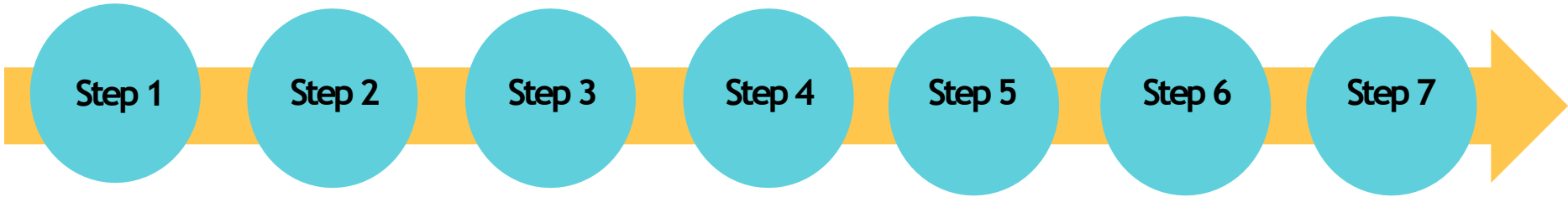
Choose the BI platform and tools

Define the scope

Create a BI Roadmap

Implement (start with some quick wins)

Create a Business Intelligence Strategy



Get Sponsorship

Assemble your BI team

Identify and involve stakeholders

Choose the BI platform and tools

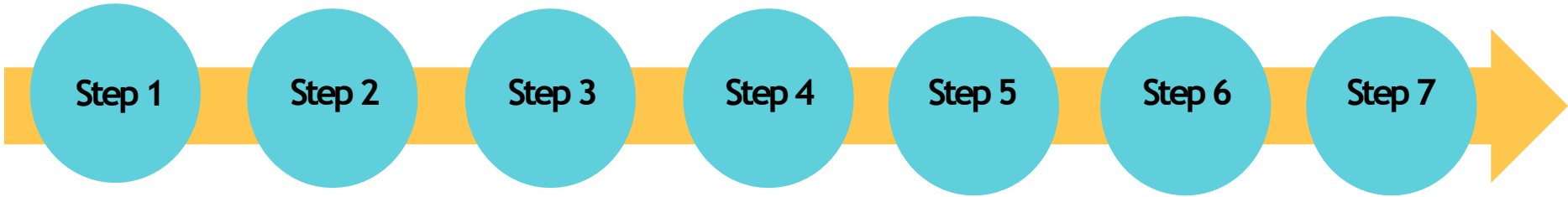
Define the scope

Create a BI Roadmap

Implement (start with some quick wins)

- Representative from every affected business group
- Involve early

Create a Business Intelligence Strategy



Get Sponsorship

Assemble your BI team

Identify and involve stakeholders

Choose the BI platform and tools

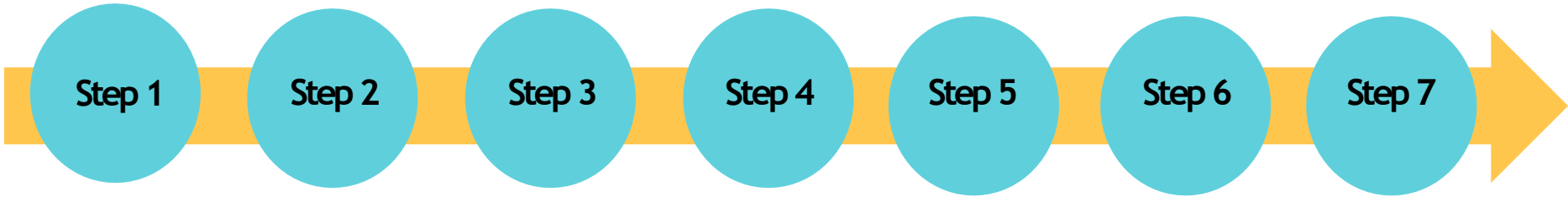
Define the scope

Create a BI Roadmap

Implement (start with some quick wins)

- Evaluate BI platforms against your needs
- Rank importance of different functionalities

Create a Business Intelligence Strategy



Get Sponsorship

Assemble your BI team

Identify and involve stakeholders

Choose the BI platform and tools

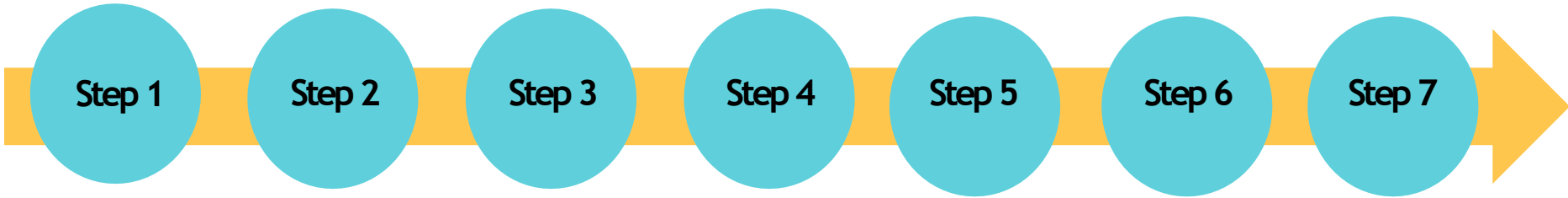
Define the scope

Create a BI Roadmap

Implement (start with some quick wins)

- What do you want to change?
- What do you envision the processes to look like?
- Do you start from a limited scope first?
- What will be the scope of analysis done?
- What is priority for the organization?

Create a Business Intelligence Strategy



Get Sponsorship

Assemble your BI team

Identify and involve stakeholders

Choose the BI platform and tools

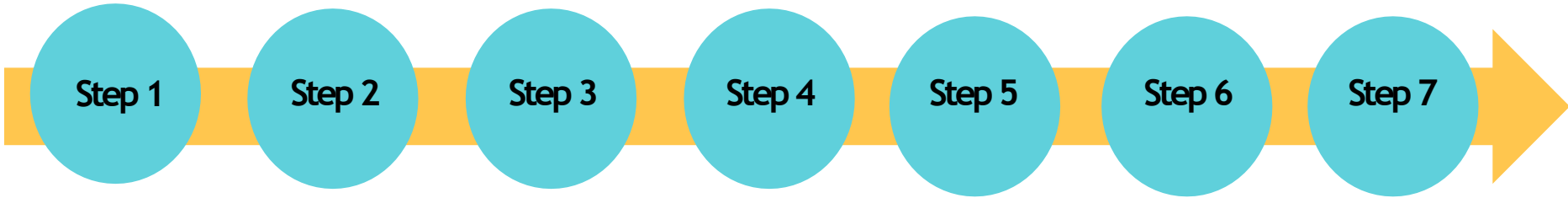
Define the scope

Create a BI Roadmap

Implement (start with some quick wins)

- Plan the scope in logical steps of release
- Keep in mind the dependencies!
- Adapt if needed
- Do not overcommit if not sure

Create a Business Intelligence Strategy



Get Sponsorship

Assemble your BI team

Identify and involve stakeholders

Choose the BI platform and tools

Define the scope

Create a BI Roadmap

Implement (start with some quick wins)

- Have strong partnership between stakeholders, IT and BI teams
- Do not forget about security and security profiles
- Break up implementation into multiple phases
- Do not forget about training and change management

Big Data and BI

- What is Big Data?
- Combining BI and Big Data

Self-Service Business Intelligence

- What is Self-Service BI
- Why is it important
- What are the challenges
- Best Practices

Magic Quadrant for Analytics and BI Platforms

Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms

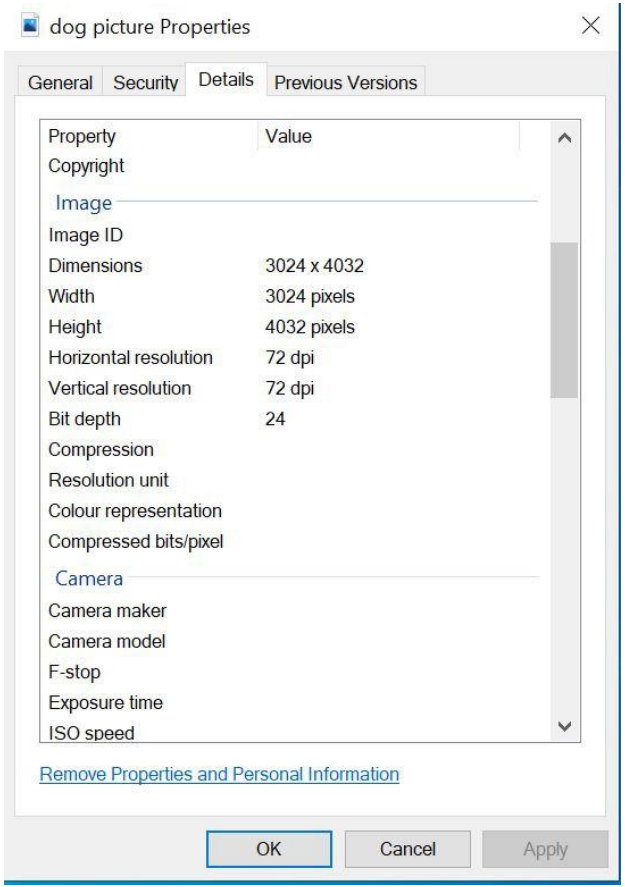
Source: Gartner (March 2022)

Module 12 - Metadata Management

What is Metadata

"**Metadata** is data that provides information about other data"

Example of Metadata



What is Metadata Management?

It is the portfolio of best-practice processes and technologies that allow businesses to manage this data about their data and derive insights for more effective data management. It allows users of all kinds – business, technical, and operational – to search for, understand, and securely access the data they need to do their jobs.

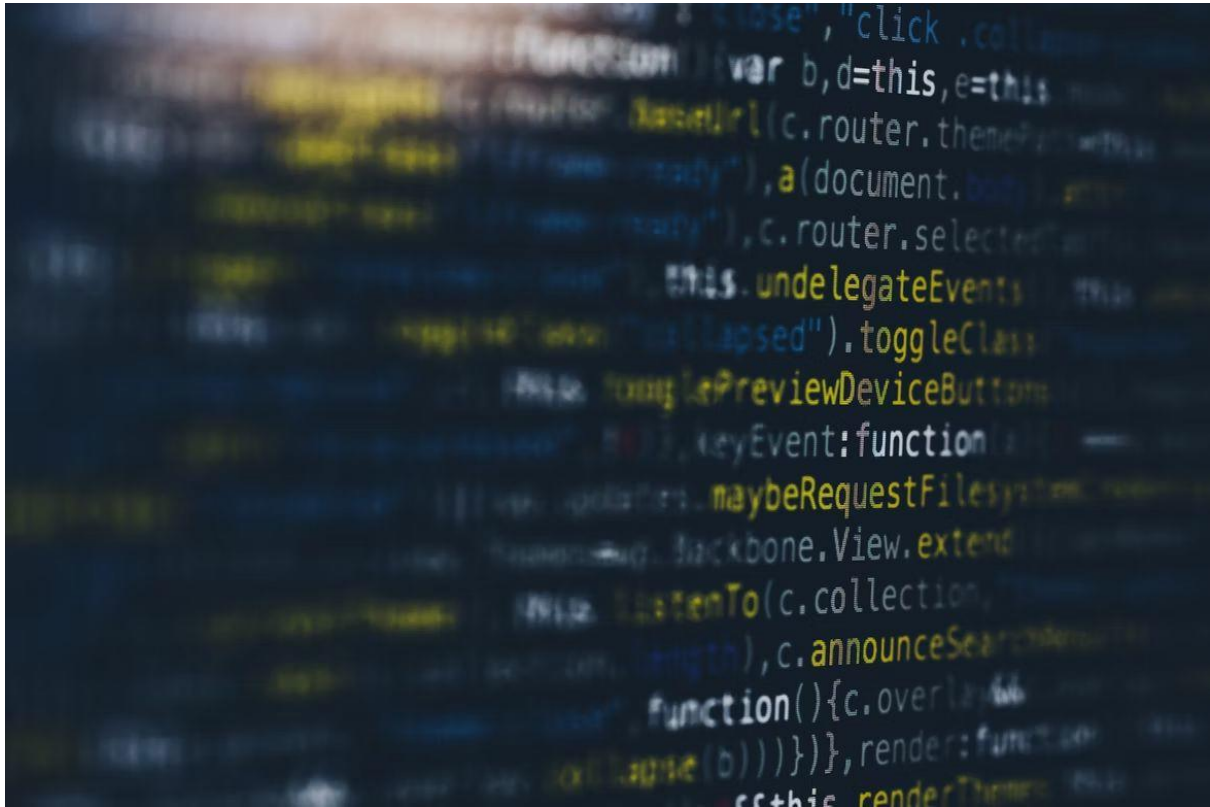
Source: https://www.informatica.com/resources/articles/what-is-metadata-management.html

Why do you need Metadata Management?

- Improved Consistency - establish a common business language
- Capture institutional knowledge
- Better data quality
- Faster access to insights
- Faster project delivery timelines
- Reduced costs
- Improved regulatory compliance

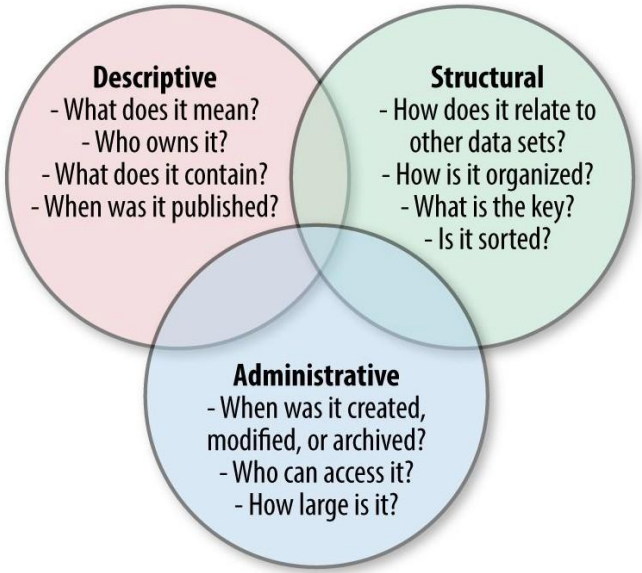
Types of Metadata

- Descriptive metadata
- Structural metadata
- Administrative metadata



Descriptive Metadata

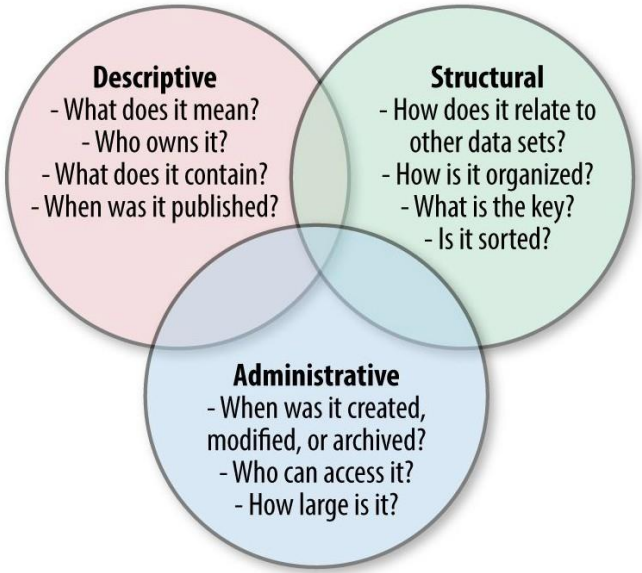
Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.



Source: The three types of metadata based on NISO's classification. Source: O'rielly

Structural Metadata

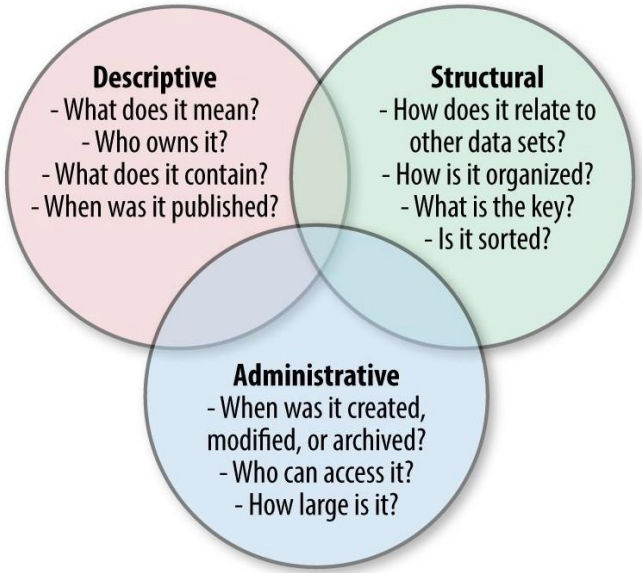
Structural metadata is used to specify the relationships between components of a digital object (internal structure) and between different digital objects (external structure)



Source: The three types of metadata based on NISO's classification. Source: O'rielly

Administrative Metadata

Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.



Source: The three types of metadata based on NISO's classification. Source: O'rielly

Metadata Activities

1. Understand Metadata Requirements
2. Define Metadata Strategy
3. Define Metadata Architecture
4. Create and Maintain Metadata
5. Query, Report and Analyze Metadata

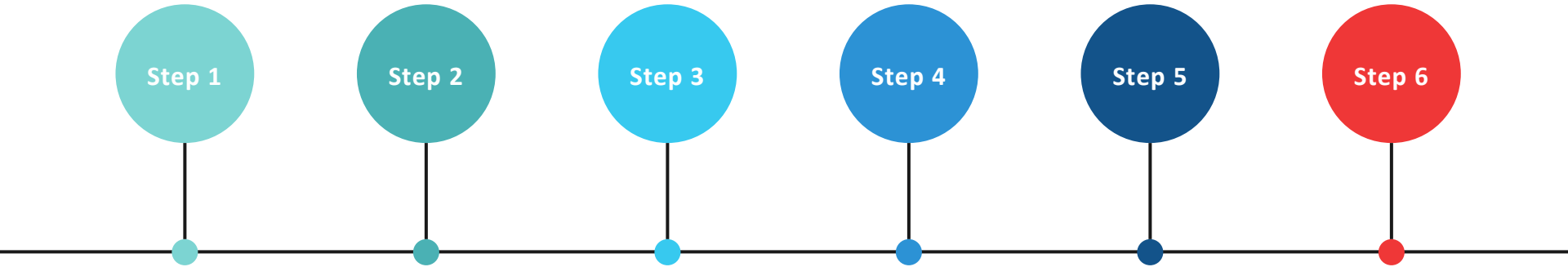
UNDERSTANDING METADATA REQUIREMENTS

Explanation: This involves determining what kind of metadata is necessary for the organization's data systems. It's about identifying the types of metadata that are essential for data processing, compliance, and business intelligence.

Example: An IT company developing a cloud storage service needs to understand metadata requirements for file storage. This could include file types, sizes, creation and modification dates, access permissions, and user activity logs. This metadata helps in managing storage systems, providing user access history, and optimizing data retrieval.

Understand Metadata Requirements

Step by Step



Identify Stakeholders

Define Data Objectives

Analyze Existing Data

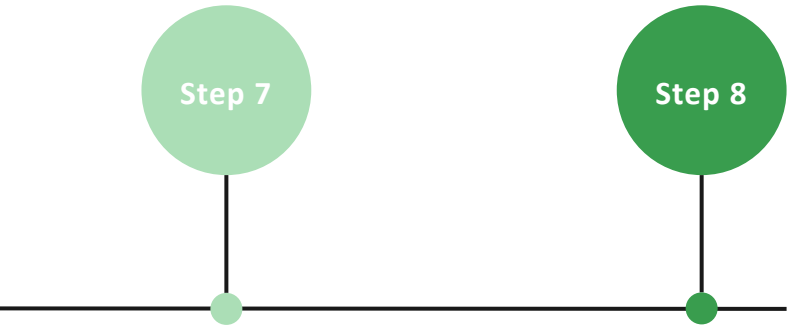
Determine Metadata Standards and Regulations

Assess Metadata Usage Scenarios

Gather Requirements

Understand Metadata Requirements

Step by Step



Prioritize Requirements

Document and Validate Requirements

DEFINE METADATA STRATEGY

Explanation: This step is about outlining a plan for how metadata will be handled within the organization. It includes the selection of metadata standards, the definition of processes for metadata creation, update, and deletion, and the establishment of roles for metadata governance.

Example: The IT department of a financial institution might define a metadata strategy that includes using the ISO 20022 standard for metadata concerning financial transactions. The strategy would detail how metadata is to be captured at each transaction point and the process for updating it to reflect transaction status changes.

DEFINE METADATA ARCHITECTURE

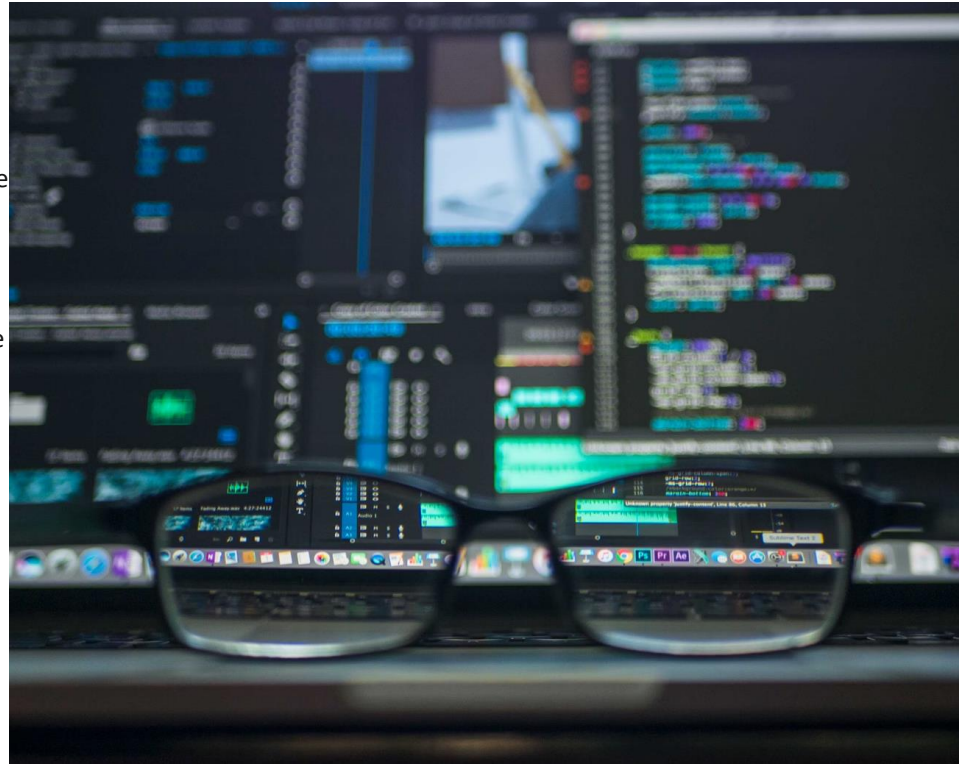
Explanation: This activity involves designing the technical framework that supports the collection, storage, and dissemination of metadata. It includes the selection of tools and technologies, the design of metadata models, and the integration of metadata with other data management systems.

Example: A software development company might design a metadata architecture that includes a centralized metadata repository, which integrates with their version control system. This would allow for tracking changes in code, dependencies, and deployment configurations, facilitating better software lifecycle management.

CREATE AND MAINTAIN METADATA

Explanation: This is the operational phase where metadata is generated according to the established requirements and strategy, and it is kept up-to-date to reflect the current state of the underlying data.

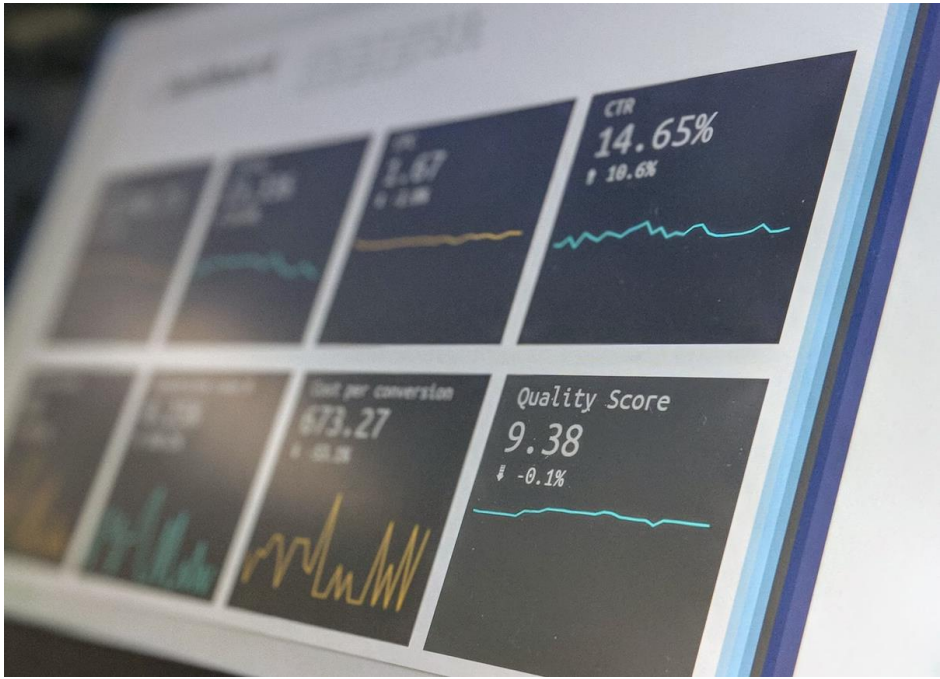
Example: An IT service provider could create metadata for each of its services, including service name, version, dependencies, configurations, and performance metrics. Maintenance would involve updating the metadata when services are upgraded, dependencies are changed, or new configurations are added.



QUERY, REPORT, AND ANALYZE METADATA

Explanation: This involves using metadata to gain insights into the IT infrastructure, improve data governance, and support decision-making. It includes building tools to search metadata, generate reports, and conduct analyses.

Example: An IT company might use metadata queries to monitor the performance of its network infrastructure. By analyzing metadata on network traffic, device status, and incident reports, the company can identify bottlenecks, predict hardware failures, and optimize network performance.



Magic Quadrant for Metadata Management Solutions

Figure 1. Magic Quadrant for Metadata Management Solutions

Source: Gartner (November 2020)

Implement Metadata Management

Step 1

Select the Metadata Program team

Step 2

Define the Metadata Strategy

Step 3

Adopt Metadata Standards

Step 4

Get the right Metadata Management tool

Step 5

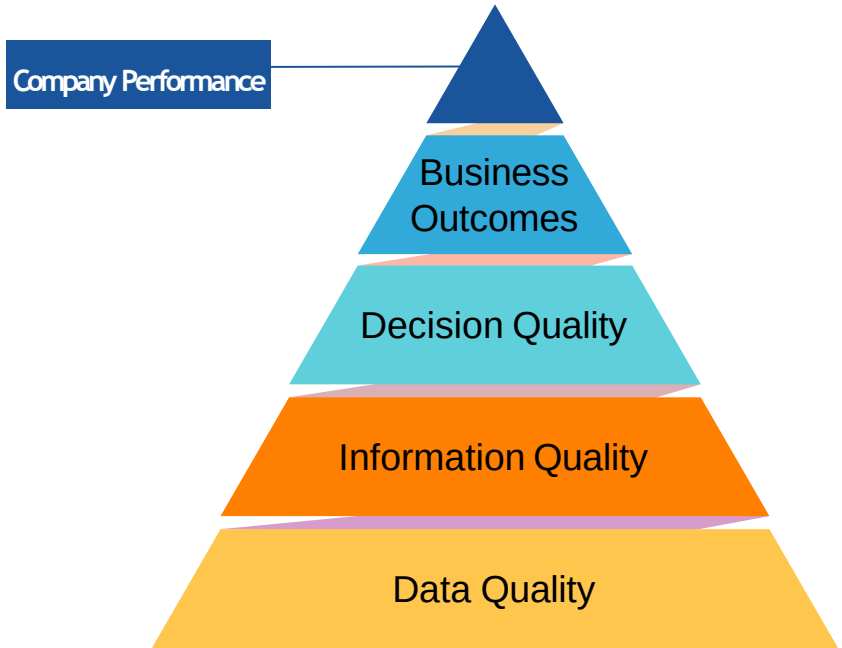
Deploy and expand

Module 13 - Data Quality Management

What is Data Quality?

Simple Definition of Data Quality:

"Data quality is defined by how well a given dataset meets a user's needs. Data quality is an important criteria for ensuring that data-driven decisions are made as accurately as possible"



Presentation by George Smarts

Data Quality Management

Definition of Data Quality Management:

"Set of practices that aim at improving and maintaining a high quality of information within the organization"

Pillars of Data Quality Management

- People
- Data Profiling
- Defining Data Quality
- Data reporting
- Data Repair

Presentation by George Smarts

Cost of Poor Data Quality

- According to IBM's estimate, the US lost $3.1 trillion yearly due to bad data.
- Gartner.com suggests that organizations lose between $10 to $14 Million USD annually due to poor data.
- Integrate reported that around 40% of all leads have inaccurate data.
- Cio.com identified that around 80% of companies believe they lost revenue due to data challenges.
- MIT Sloan reported that employees spend half of their time coping with managing data quality tasks.
- Pragmaticworks states 20 to 30 percent of operating expenses are due to bad data.
- Econsultancy.com reported that due to poor data, companies having mail delivery issues lost about 30% of their revenue, in addition to the 21% of businesses experienced reputation damages.
- Gartner also reported that data scientists spend around 80% of their time cleaning and organizing data.

Source: https://www.gartner.com/en
https://www.invensis.net/blog/cost-of-bad-data-quality/

Data Quality Dimensions

Accuracy
Data accurately represents “reality”

Validity
Data conforms to the expected format, type and range

Timeliness
Data is available when expected and needed

Completeness
All the data required for a particular use is present and can be used

Uniqueness
Data appears only once in a data set

Consistency
Data values do not conflict with other values across different data sets

Presentation by George Smarts

Data Quality Improvement Process

Module 14 - Big Data & Data Science

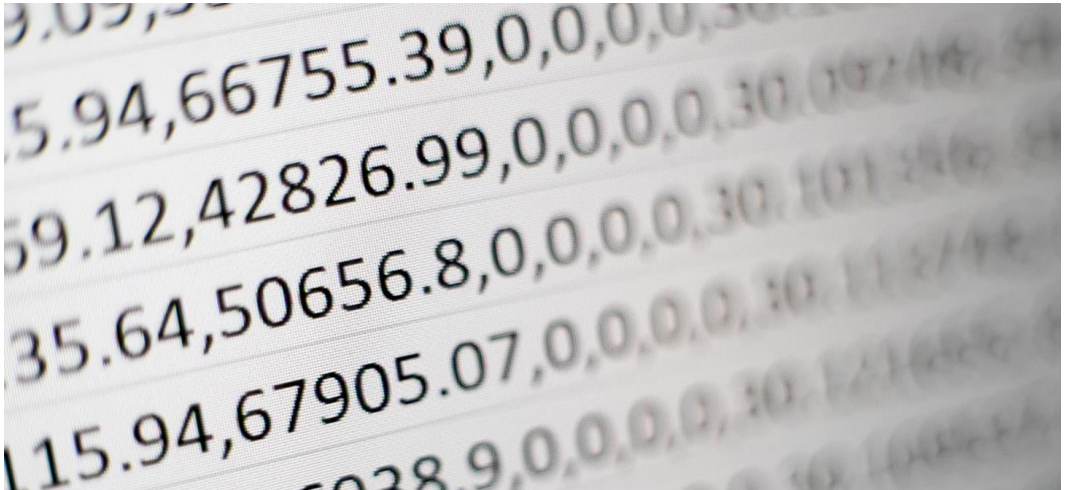
Definition of Big Data

Big data refers to extremely large datasets that are so complex and vast that traditional data processing software cannot handle efficiently.

Key Characteristics:

- **Large Volumes:** Think of the massive amounts of information generated every second, like social media posts, online transactions, and sensor data from devices.
- **Fast Processing:** The data is generated quickly, often in real time, requiring immediate analysis to make timely decisions.
- **Different Types:** Big data comes in various forms, including text, images, videos, and more, making it challenging to analyze.

In simple terms, big data is about managing and analyzing huge amounts of information to find patterns and insights that can help businesses and organizations make better decisions.



Uber and Big Data

Here are 5 primary applications of big data at Uber:

- Dynamic Pricing and Surge Pricing
- Driver-Rider Matching
- Predictive Analytics for Demand Forecasting
- Route Optimization
- Fraud Detection and Safety Measures

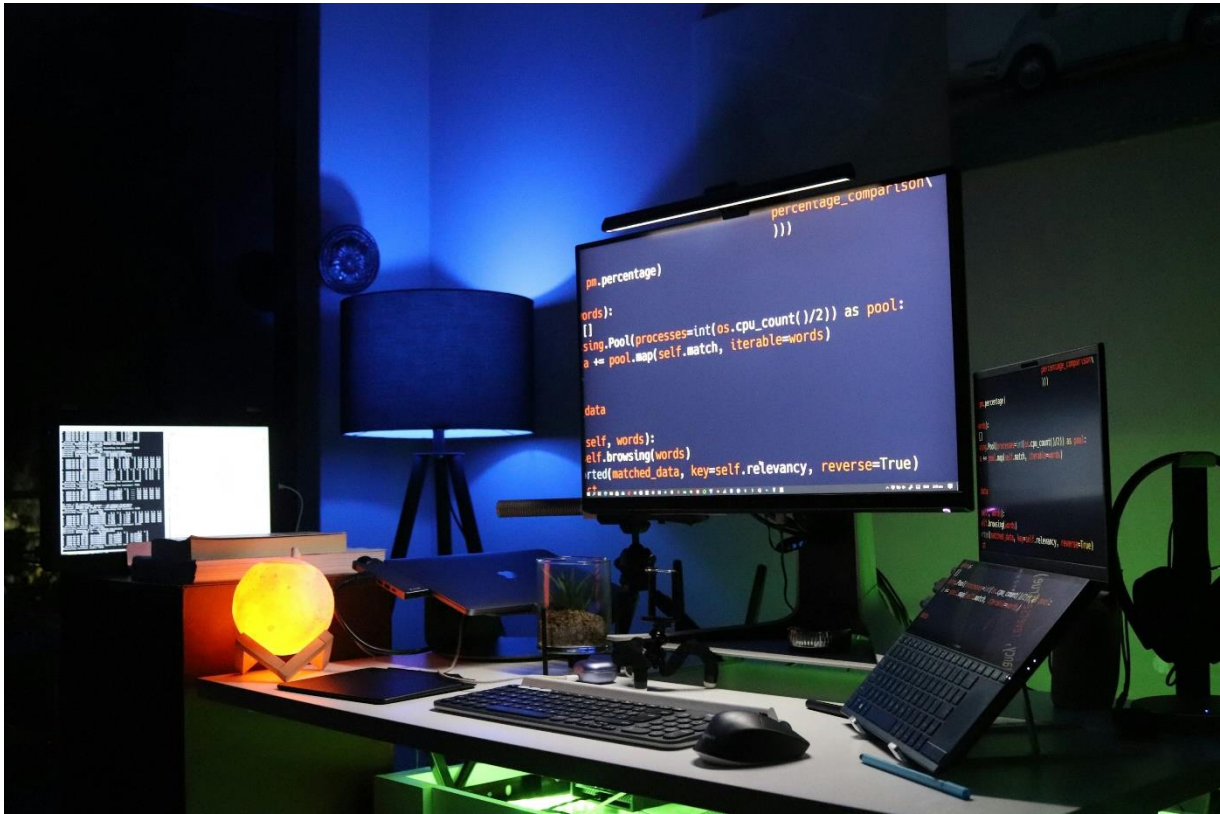
https://d3.harvard.edu/platform-digit/submission/uber-knows-you-how-data-optimizes-our-rides/

Definition of Data Science

Data science is a field that combines various techniques and theories from statistics, mathematics, computer science, and domain knowledge to extract meaningful insights from structured and unstructured data. It involves processes such as data collection, cleaning, analysis, and visualization to inform decision-making and drive business strategies.

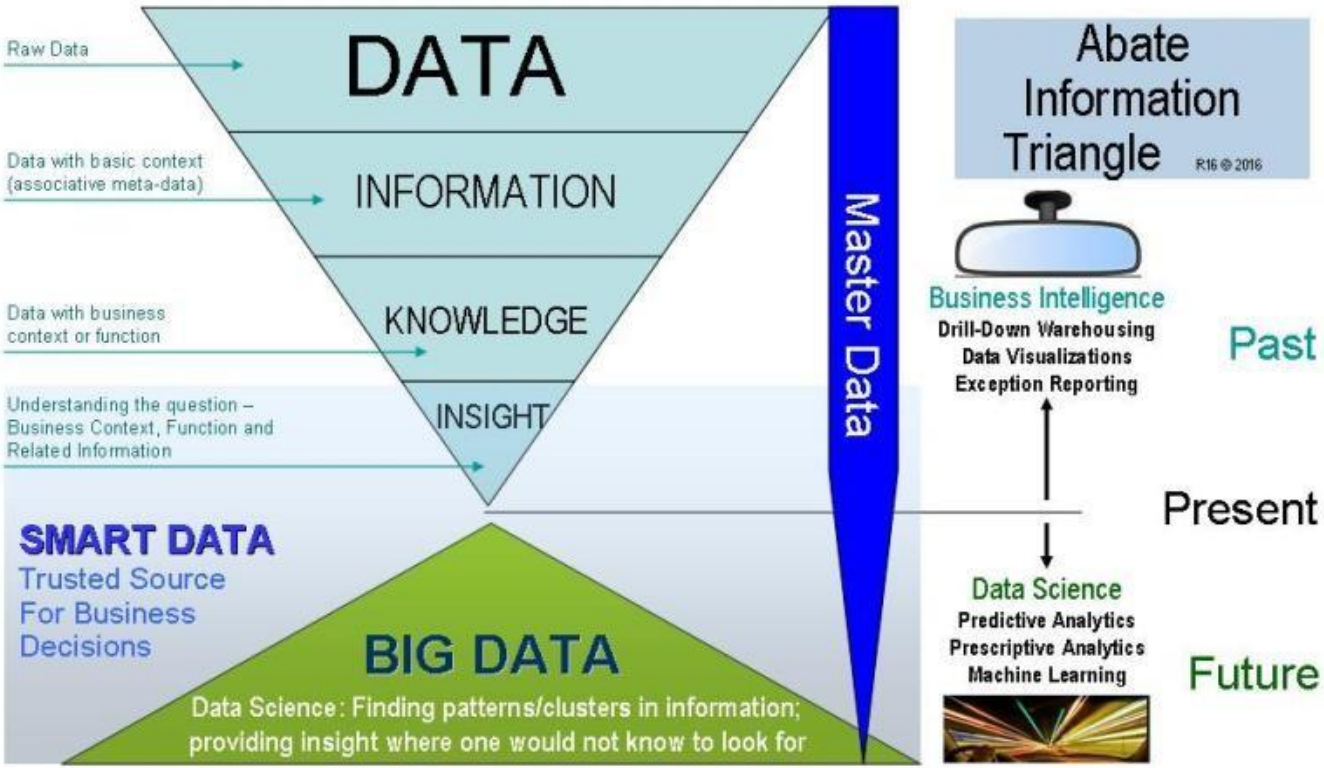
Key skills in Data Science:

1. Statistical Analysis
2. Programming
3. Machine Learning
4. Data Visualization



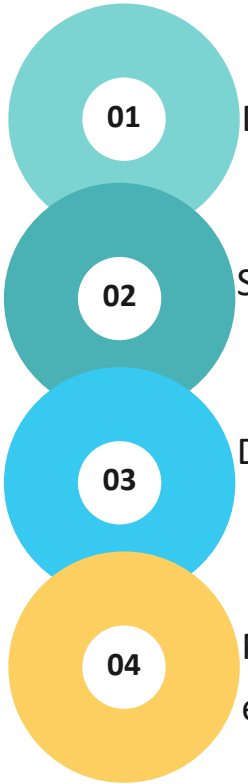
The Abate Information Triangle

The Abate Information Triangle is a foundational model developed by Robert J. Abate that illustrates the relationship between Data, Information, and Knowledge. This framework has gained particular relevance in the context of Big Data, highlighting how organizations can effectively manage and utilize vast amounts of information.



Source: https://www.linkedin.com/pulse/turning-information-actionable-insights-flipping-abate-cbip-cdmp/

GOALS OF BIG DATA AND DATA SCIENCE



Discover relationships between data and the business

Support the iterative integration of data sources(s) into the enterprise

DIscover and analyze new factors that might affect the business

Publish data using visualization techniques in an appropriate, trusted and ethical manner.

Source: DAMA-DMBOK: Data Management Body of Knowledge: 2nd Revised Edition - CHAPTER 14 Big Data and Data Science

What Big Data actually involves

7 Key Activities

1. Define Big Data Strategy and Business Needs
2. Choose Data Sources
3. Acquire and Ingest Data Sources
4. Develop Hypotheses and Methods
5. Integrate data for analysis
6. Explore Data Using Models
7. Monitor and Deploy

Source: DAMA-DMBOK: Data Management Body of Knowledge: 2nd Revised Edition - CHAPTER 14 Big Data and Data Science

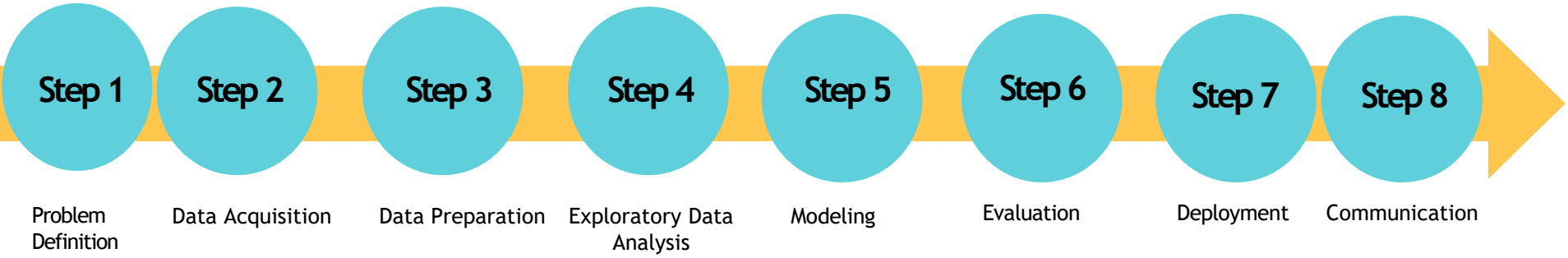
Deliverables of Big Data

6 Key Deliverables

1. Big Data Strategy and Standards
2. Data Sourcing Plan
3. Acquired Data Sources
4. Initial data analysis and hypotheses
5. Data insights and findings
6. Enhancement Plan

Source: DAMA-DMBOK: Data Management Body of Knowledge: 2nd Revised Edition - CHAPTER 14 Big Data and Data Science

The Data Science Lifecycle



Step 1 - Summarize the data

- We summarize to simplify the data
- We summarize to quickly identify “normal” and “odd”
- We summarize to provide more context to the data

Four key areas of data summarization

- Centrality
- Dispersion
- Replication (aggregation)
- Shape

Source:https://www.dataanalytics.org.uk/data-analytics-knowledge-base-tips-tricks-r-excel/statistics-guide/data-summary/

Data Centrality

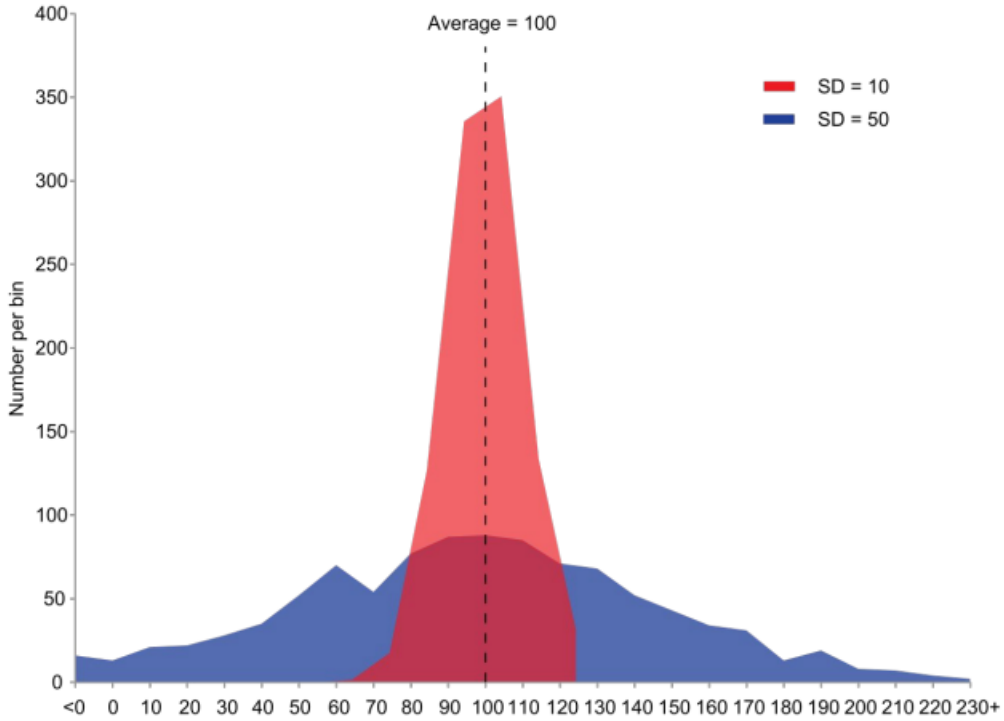
$$\text{Mean } \bar{x} = \frac{\sum x_i}{N}$$

$$\text{Median} = \begin{cases} \frac{(N+1)^{th}}{2} \text{ term; when N is odd} \\ \frac{\frac{N^{th}}{2} \text{ term} + \left(\frac{N}{2} + 1\right) \text{term}}{2} \text{ ; when N is even} \end{cases}$$

Mode = The value in the data set that occurs most frequently

Data Dispersion

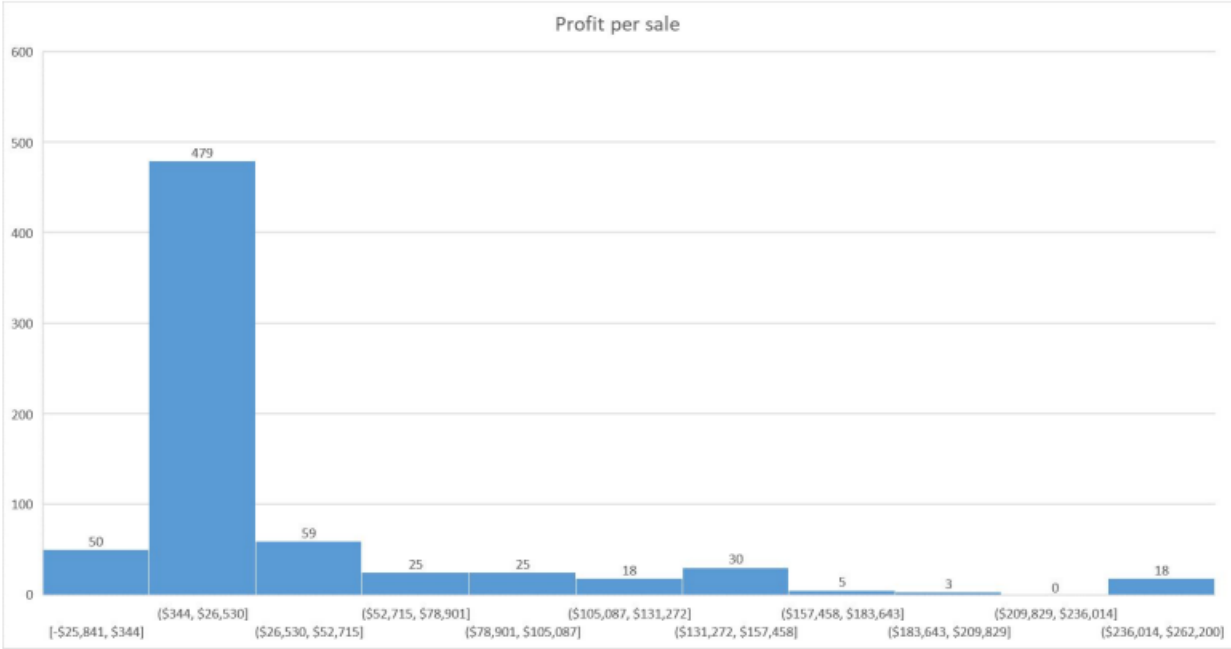
- Range
- Standard Deviation
- Inter-Quartile range



Data Repetition

	A	B	C	D	E	F	G	H
1	Year	2023						
2								
3	**Sum of Sales**	**Column Labels**						
4	**Row Labels**	**Hat**	**Jeans**	**Shirt**	**Shoes**	**Socks**	**T-Shirt**	**Grand Total**
5	France	$2,943,703	$1,005,746	$2,101,032	$528,537	$4,187,817	$582,703	$11,349,538
6	Germany	$2,158,598	$2,983,373	$6,498,114	$5,671,875	$6,434,146	$327,027	$24,073,134
7	India	$4,148,103	$1,279,961	$1,005,718	$3,593,731	$870,321	$115,116	$11,012,950
8	United Kingdom	$2,521,553	$1,162,950	$45,540	$1,697,265	$820,275	$554,514	$6,802,097
9	United States of America	$2,942,177	$5,306,425	$798,099	$1,040,857	$2,225,964	$136,749	$12,450,272
10	**Grand Total**	**$14,714,134**	**$11,738,455**	**$10,448,504**	**$12,532,265**	**$14,538,523**	**$1,716,110**	**$65,687,990**
11								

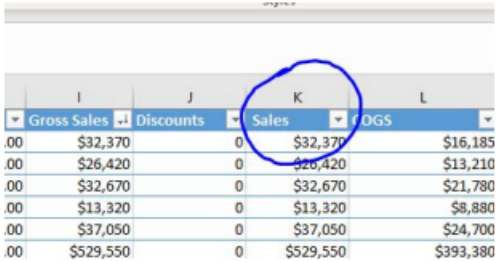
Data Shape



Presentation by George Smarts

Exercise - Histogram

1. Create a Histogram out of the Sales column
2. Copy the Histogram into a new worksheet called “Histogram”
3. Change the Title to Order Sales $ Frequency
4. Change the bin sizes to 100,000 width
5. Change the bars to green color
6. Add data labels



	I	J	K	L
	Gross Sales	Discounts	Sales	COGS
.00	$32,370	0	$32,370	$16,185
.00	$26,420	0	$26,420	$13,210
.00	$32,670	0	$32,670	$21,780
.00	$13,320	0	$13,320	$8,880
.00	$37,050	0	$37,050	$24,700
.00	$529,550	0	$529,550	$393,380

Exercise - Pivot Chart

1. Insert a Pivot Chart from the Cali Fashion data
2. Name the new worksheet “Pivot Chart”
3. The value field you will be using is **Sales**
4. Your axis will be **Country** followed by **Segment**
5. Add a title - “Sales per Country and Segment”
6. Remove the axis buttons from the bottom of the chart
7. Remove Sum of Sales button and Total label
8. Add label to the bars and format to currency with no decimals
9. Change the color of the bars
10. Play around with the drill-down functionality


Congratulations!