AIG150- Week 7

Data governance: Tools, People, and Processes

Reading Text: Ch 01-03

Data Governance: The definitive guide

Agenda

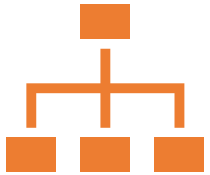
—What is data governance ?

—Why is it becoming more important ?

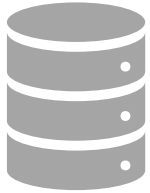
—Data governance in action

—Ingredients of data governance

Data Governance



Data governance is about identifying important data around an organization and to ensure the quality, integrity, security, availability, and usability of the collected data is maintained



It needs to be in place from the time a factoid of data is collected or generated until the point in time in which the data is destroyed or achieved

Data governance needs to be in place to ensure:

1. Data is accessed only by permitted users in permitted ways
2. Data is auditable, meaning all accesses, including changes, are logged
3. Data is compliant with regulations

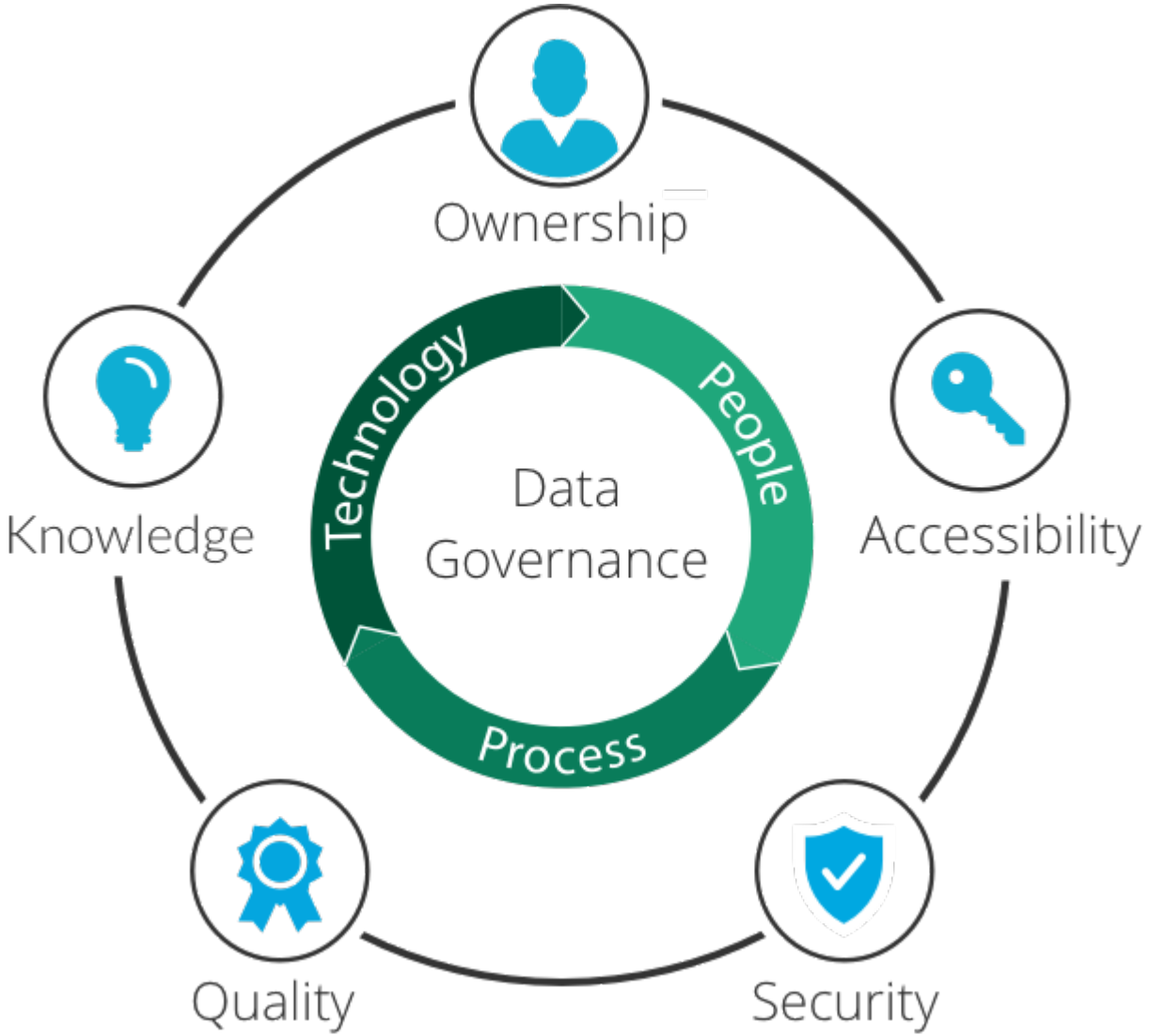
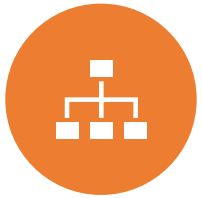


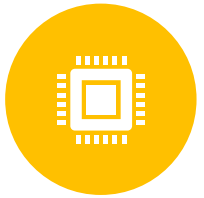
Image taken from https://www.imperva.com/learn/data-security/data-governance/

Questions To Ask



Does your organization collect high quality data?

Is the data stored securely?



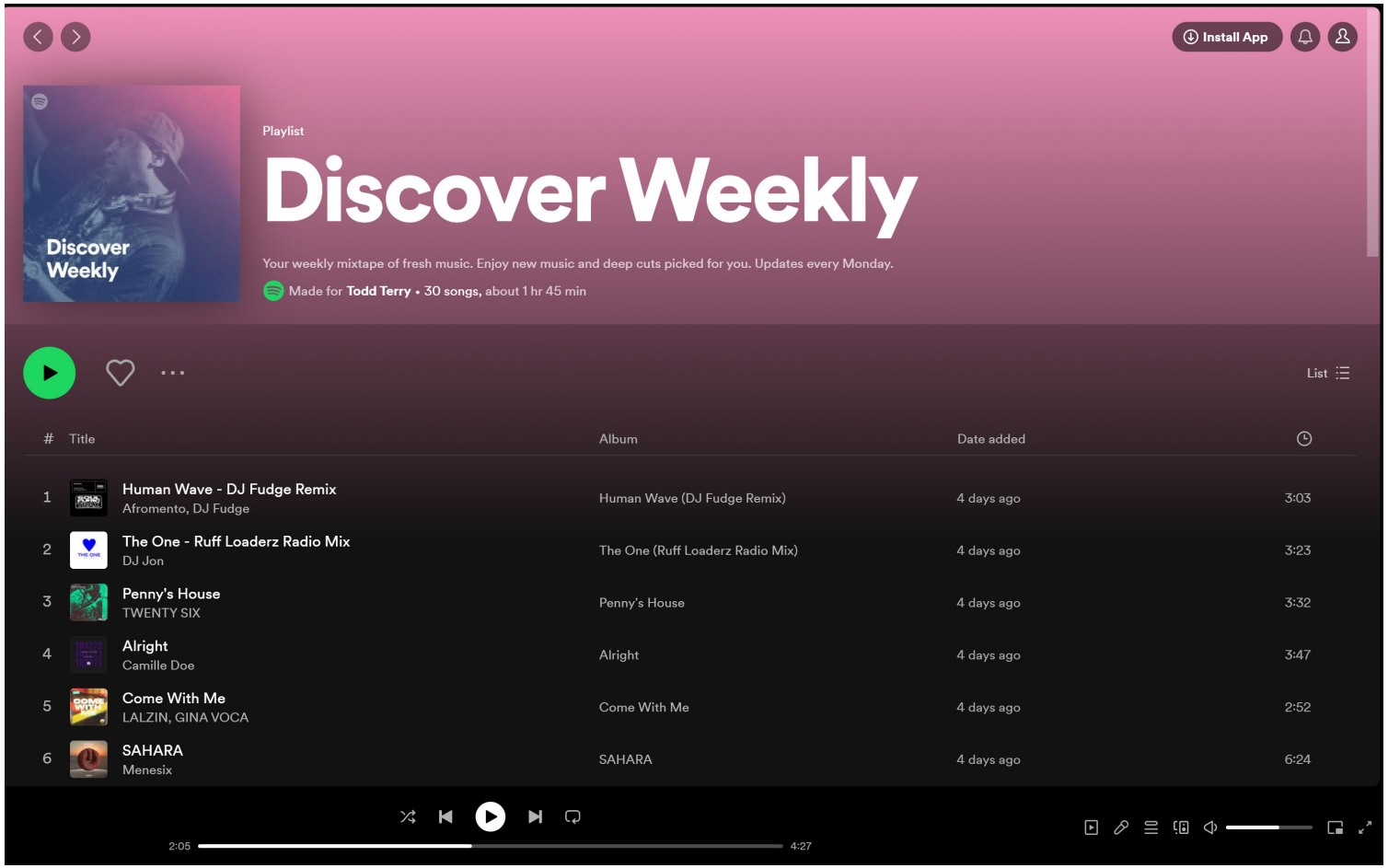
Can employees access only the data they should have access to ?

Does your organization have the mechanism to assess new data analytical techniques ?

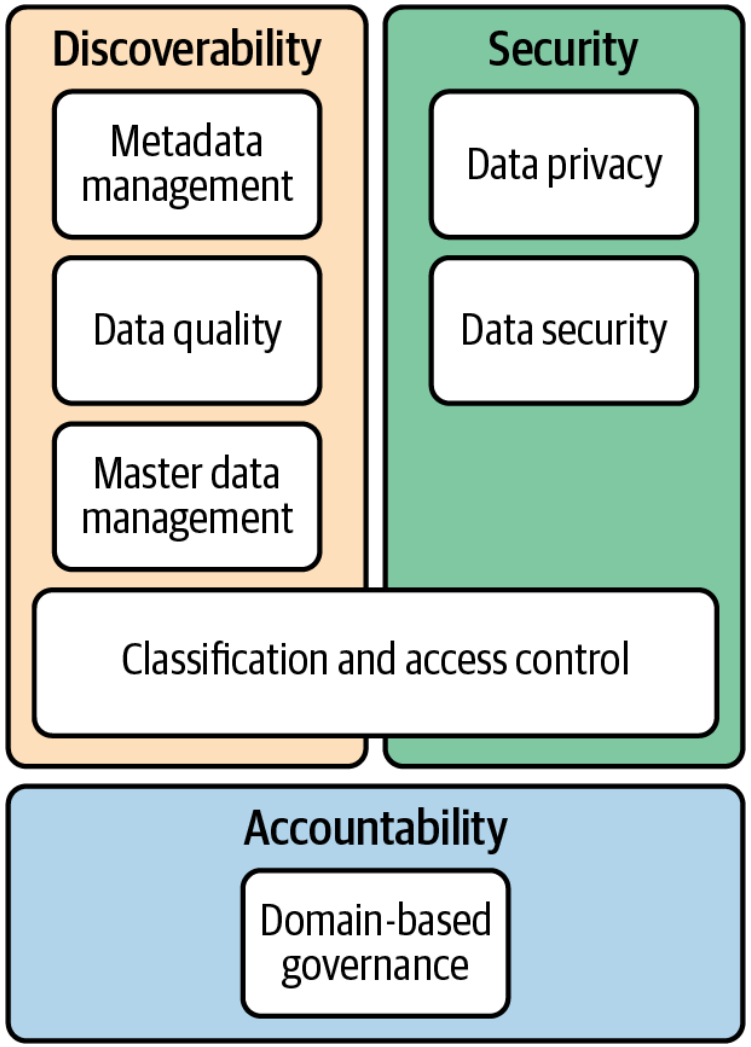
Are analytics performed beneficial to the organization ?

Is data privacy and security well taken care of?

With Data Governance In Place

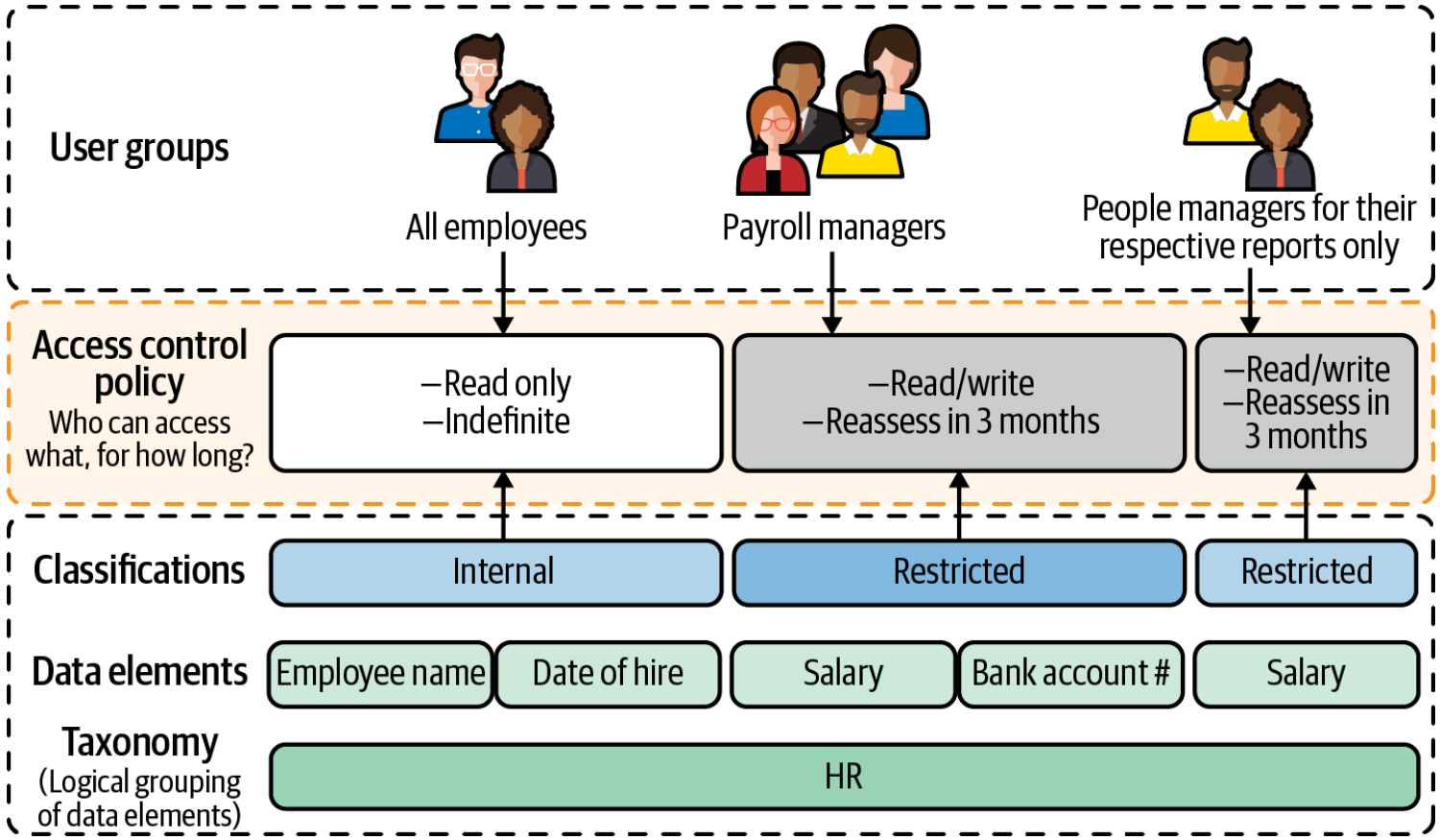


Enhancing Trust In Data



Taken from the reference text
Figure 1-2. The three key aspects of data governance that must be addressed to enhance trust in data

Classification & Access Control

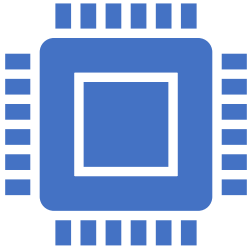


Taken from the reference text Figure 1-3. Protecting the human resources information of employees

Data Governance Versus Enablement Versus Security

Data governance is mostly focused on making data accessible, reachable, and indexed for searching across the relevant constituents, usually the entire organization's knowledge-worker population

Data enablement goes further than making data accessible and discoverable; it extends into tooling that allows rapid analysis and processing of the data to answer business-related questions: "how much is the business spending on this topic," "can we optimize this supply chain," and so on



Data security, which intersects with both data enablement and data governance, is normally thought about as a set of mechanics put in place to prevent and block unauthorized access

Why Data Governance Is becoming More Important ??

- The size of data is growing, IDC predicts that global datasphere will balloon to 175 ZB by 2025.
- The number of people viewing and working with the data has grown exponentially. IDC reports that there are over 5 billion people in the world interacting with data, projected to increase to 6 billion (75% of the world's population) in 2025.
- Variety of ways to produce and collect data.
- More kinds of data including sensitive data are now being collected.
- Laws and Regulations such as GDPR, PIPEDA, HIPPA.
- Ethical concerns around the use of data.

Business Value Of Data Governance

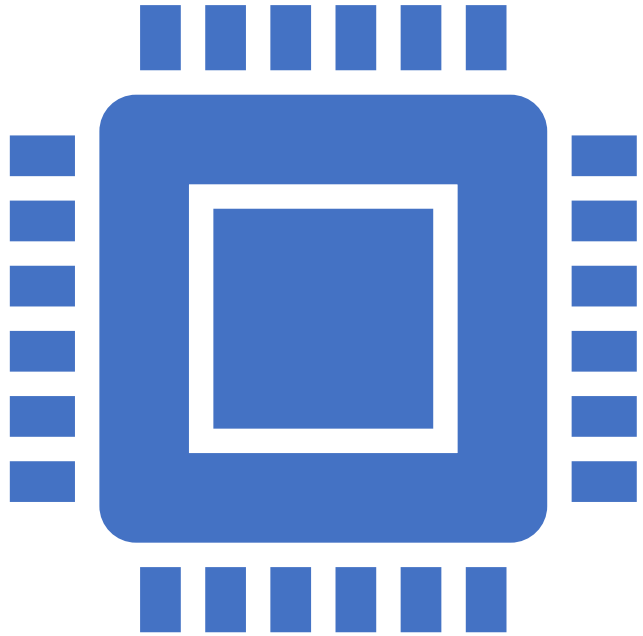
- Data governance ideally will allow all employees in the organization to access all data (subject to a governance process) under a set of governance rules (defined in greater detail below), while preserving the organization's risk posture (i.e., no additional exposure or risks are introduced due to making data accessible under a governance strategy).

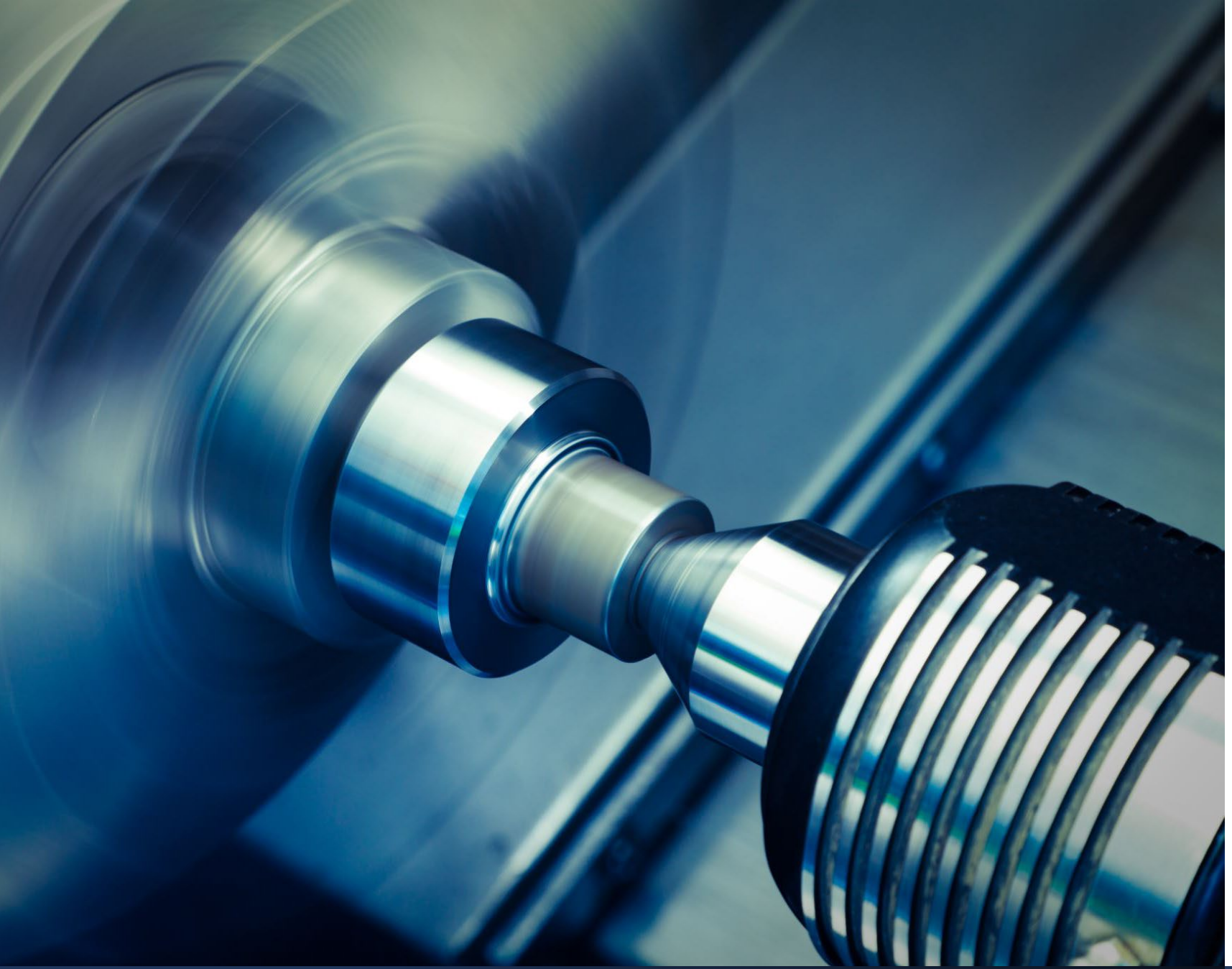
- It should lead to better decision making, better opportunity discovery, and a more productive organization.

- Help organization to achieve goals such as improve operations, find additional sources of revenue, or even monetize data directly.

Things To Consider

- Changing regulations and compliance needs
- Data accumulation and organization growth
- Moving data to the cloud
  - Data governance is easier in cloud
- Data infrastructure expertise

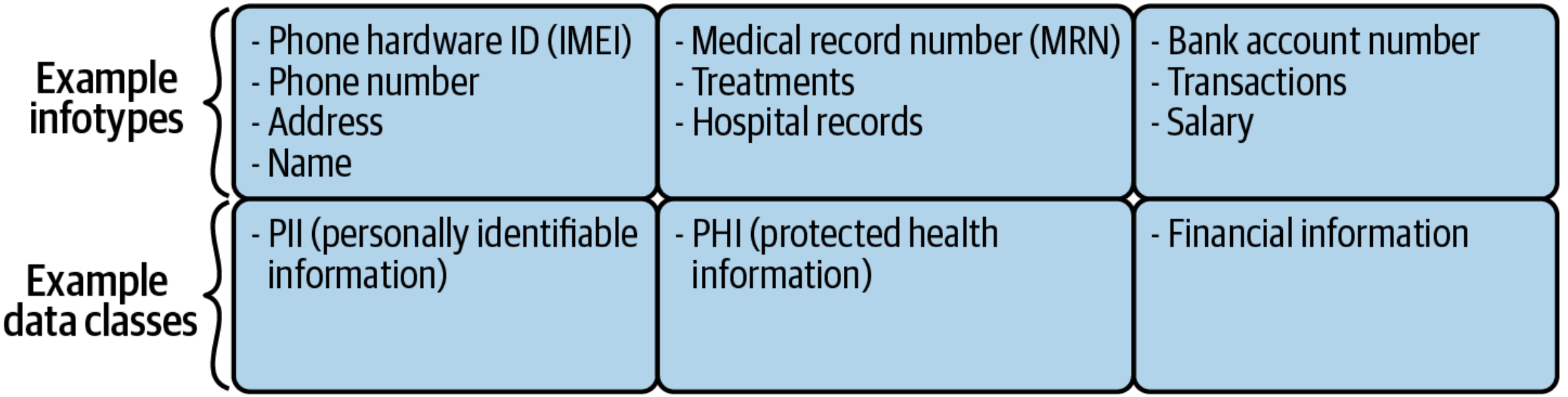




Ingredients of Data Governance: Tools

- Automation
- Machine learning
- Automatic policy applications

Enterprise Dictionary



Taken from reference text
Figure 2-1. Data classes and infotypes

Data Class

A data class references a set of *policies*: the same retention rules and access rules are required on this data.

A data class references a set of individual infotype.

Once the data the organization handles is defined in an enterprise dictionary, policies that govern the data classes can be assigned to data across multiple containers.

The desired relationship is between a policy (e.g., *access control*, *retention*) and a data class rather than a policy and an individual container.

Example:Analysts cannot access PII” versus “Analysts cannot access column #3 on Table B”) and supports scaling to larger amounts of data.

Data Class Hierarchy

Policy tags

Policy tags are tags with access control policies that can be applied to sub-resources, for example, BigQuery columns.

Name	ID	Description
Restricted	3247623653529953690	Highly Restricted Data
PHI	4081878655865131464	Patient Health Information
Drug_Details	348889402753783706	Details about a drug perscribed
NHS_Number	4099447459463431825	Patient ID
Treatment_Details	6587645476172403944	Details about a treatment or condition
PII	1690556303680165819	Personally identifiable data
Email	5606010836299662298	Email address
IMEI	7077445421065241870	Cellphone hardware ID
IP_Addr	2449414728069309088	IP Address of a session/connection
Personal_Car_VIN	7187828684927708308	Vehicle Identifier
Phone_Num	8401384437536803987	Phone number
SSN	9118232350617909155	US Social Security Number
Sensitive	5013925770628759512	Sensitive Data
Financials	358397642325435489	Financial Data
Bank_Account	8370833355300570	International Bank account ID
Credit Card Num	6313828804358283165	Credit Card number
Unrestricted Data	8097084282273622955	Unrestricted Data, broad access
Car_Details	4696597770432605648	Generic Details about a vehicle

Taken from reference text
Fig 2-2.data class hierarchy

Data Access and Policy Book

- Who (inside or outside the organization) can access a dataclass?
- The retention policy for the data class (how long data is preserved)
- Data residency/locality rules, if applicable
- How the data can be processed (OK/Not OK for analytics, machine learning, etc)?
- Other considerations by the organization

Data Retention

To limit liability, risk management, and exposure to legal action, an organization will usually define a maximum (and a minimum) retention rate for data.

In the case of financial institutions, for example, it is common to find requirements for holding certain kinds of data (e.g., transactions) for a minimum of seven years.

Policies

01

For compliance, the organization needs to be able to prove to a regulator that it has the right policies in place around the handling of data.

02

A regulator will require the organization to submit the policy book, as well as proof (usually from audit logs) of compliance with the policies.

03

The regulator will require evidence of procedures to ensure that the policy book is enforced and may even comment on the policies themselves.

Data Policies And Use Cases

Data Classification, Cataloging & Metadata Management

Automate data classification by tagging the data as “belonging to a class”.

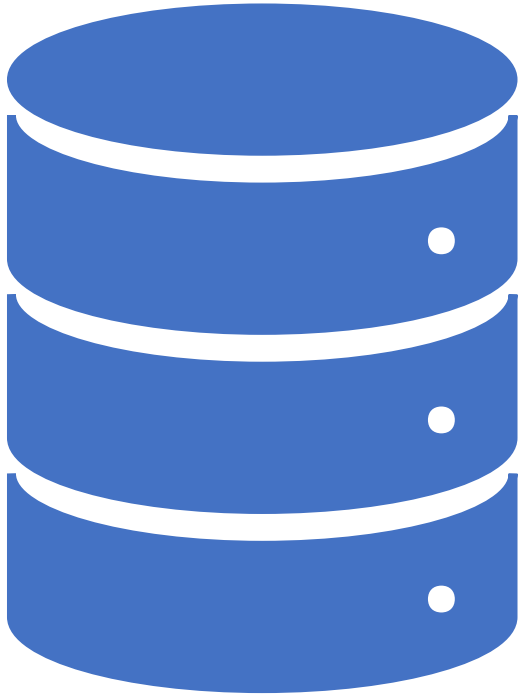
Automatically apply policies that control access to and retention of the data according to the definition of the data class.

Should apply controls and policies on metadata.

Data catalog manages the metadata. It should be able to efficiently index all the information and must be able to present it to users whose permissions allow it, using high-performing search and discovery tools.

More Tasks For Automation

- Data assessment and profiling
- Lineage tracking
- Key management and encryption
- Data retention and deletion
- Identity and access management
- User authorization and access management



Ingredients of Data Governance: People and Processes

People: Roles, Responsibilities, & Hats

Table 3-1. Different hats with their respective categories and the tasks associated with them

Hat	Category	Key tasks
Legal	Ancillary	Knows of and communicates legal requirements for compliance
Privacy tsar	Governor	Ensures compliance and oversees company's governance strategy/process
Data owner	Approver (can also be governor)	Physically implements company's governance strategy (e.g., data architecture, tooling, data pipelining, etc.)
Data steward	Governor	Performs categorization and classification of data
Data analyst/data scientist	User	Runs complex data analytics/queries
Business analyst	User	Runs simple data analyses
Customer support specialist	Ancillary (can also be a user)	Views customer data (but does not use this data for any kind of analytical purpose)
C-suite	Ancillary	Funds company's governance strategy
External auditor	Ancillary	Audits a company's compliance with legal regulations

Table taken from reference text

Know Your Company

Legacy companies
Cloud natives/ digital only
Retail
Highly regulated
Small companies
Large companies

People & Process Together (1)

Issues and consideration

- Hats versus roles and company structure
- Tribal knowledge and SMEs
- Definition of data
- Old access methods
- Regulation compliance

People & Process Together (2)

Strategies for success

- Data segregation within storage system
- Data segregation and ownership by line of business
- Creation of “views” of datasets
- A culture of privacy and security