Data Classification

Data

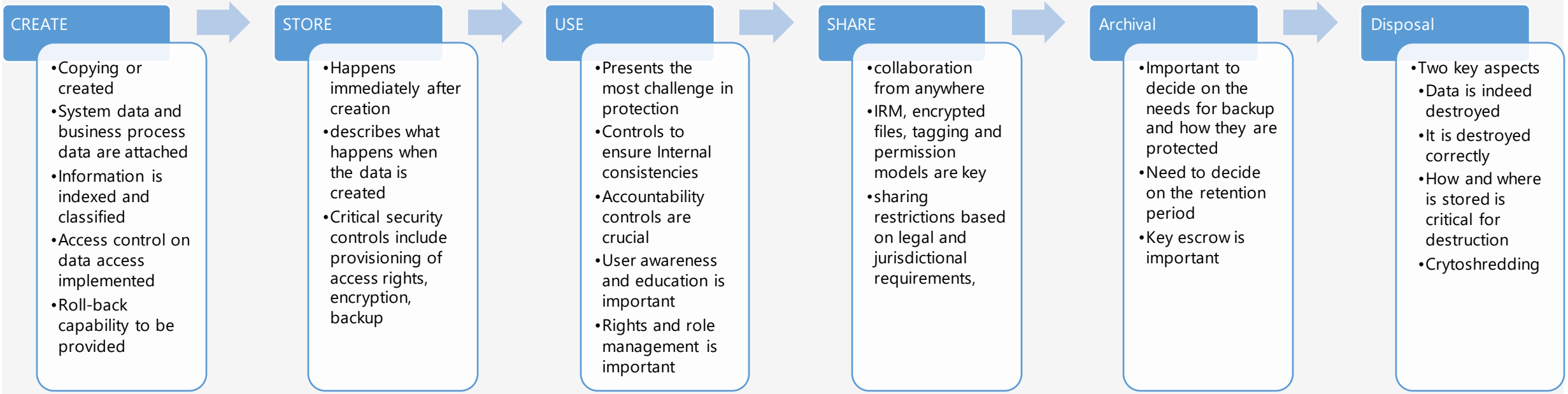
- Information : Data that is combined to form meaning
- Information has worth to the organization

Data Backup	Data Archive
• Copy of current data set that is used as backup if loss of the original data set • It becomes less useful over time	• Copy of data set that is no longer in use, but retained for use later • Data from original location is destroyed

Data Types

- **Sensitive Data:**
  - Any information that is not public or unclassified
  - Any type of data that an organization has value upon and shall protect or comply with law and regulations
- **Personally Identifiable Information:**
  - Any information that can identify an individual
  - Race, name, SSN, date, place of birth, biometric, medical, financial, employment information
- **Protected Health Information:**
  - A health related information that can be related to an individual
  - Oral or written information created or received by health care related entities
  - Relates to past, present or future medical information of an individual
- **Proprietary Data:**
  - Any data that helps an organization to maintain a competitive edge
  - If lost, it can seriously affect the primary mission of an organization

Data Life Cycle



- Cloud data lifecycle is not generally iterative
- Data may exist in one or more phases simultaneously

Data Owner

- Key aspect of good data management involves identification of information owner
- Individual or group that created, acquired or purchased information that supports the mission of the organization
- Has legal rights over the data
- Ownership implies the right to exploit the data as well as the right to destroy it
- Also referred as **Data Controller**

Data Owner Responsibilities

- Determine the impact the information has on the organization
- Understand the replacement cost of the information
- Establish the rules of appropriate use and protection of information
- Decide who has access to the information and what privilege
- Know when the information is inaccurate or no longer needed and should be destroyed
- Provide input to system owners regarding security requirements and controls for the information system that hold the data
- Assist in identification and assessment of common security controls
- Delegates day-to-day maintenance to the data custodian

Data Owner

- Data Owner shall **establish and document** the following
  - The ownership, IP rights and copyrights for their data
  - The statutory and non-statutory requirements relevant to their business to ensure the data is compliant
  - The policies for data security, disclosure, pricing and dissemination
  - Contracts with users and customers on conditions of use, before the data is released

Data custodian

- Data custodian ensures important data sets are developed, maintained and are accessible within their defined specifications.
- Best handled by entity that is most familiar with a datasets content and its management criteria
- Also referred to as **Data Processor**
- Responsibilities include
  - Adherence to data owner guidelines
  - Ensure access to appropriate users and maintaining appropriate level of security
  - Dataset maintenance, including data storage and archival
  - Dataset documentation, including changes to documentation
  - Quality Assurance and validation to assure ongoing data integrity

System Owner

- A person who owns the system processing sensitive information
- One system may have multiple information owners
- Responsibilities
  - Develop a system security plan in coordination with Information owners
  - Maintain the plan and ensure it operates according to the agreed security requirements
  - Ensure system users and support personal get security training
  - Update the plan whenever major change happens
  - Assist in identification, implementation and assessment of common security controls

Other Roles

- **Data Analyst / Data Steward**
  - Ensures data is stored in a way that makes more sense to the company
  - Responsible for architecting a new system that will hold company information or advice in purchase of a product
  - Works with data owners to help ensure that the structures setup support business objectives
- **Security Administrator**:
  - Responsible for maintaining specific security devices
  - Creating new user accounts, implementing new security software, testing security patches
  - Has the focus of keeping the network secure; network administrator has main focus on keep the IT running
- **Supervisor**
  - Ultimately responsible for all actions of the users under them
  - Responsible for making sure access changes are done for user accounts as and when there is change in user role

Data Classification

- Refers to the practice of differentiating between different types of information assets and providing some guidance as to how they must be protected
- It is an ongoing process and not one-time effort
- Important metadata item that should be attached to all information is ~ classification level
- The classification level should be always attached throughout the lifecycle of the information
- Primary Purpose:
  - Helps indicate the level of confidentiality, integrity and availability protection that is needed for each type of data
  - Helps ensure data is protected in a most cost-effective manner
- Each classification should have separate handling requirements and procedures

Classification & Categorization

- **Classification**
  - Identifies the value of the data to the organization
  - It also identifies how data owners can determine the proper classification, and personnel should protect data based on classification
  - Classification authority is the one who applies the original classification to the sensitive data
- **Categorization**
  - Process of determining the impact due to the loss of CIA of information to an organization
- Classification and categorization help to set baselines for information systems

Data Classification

- Data is classified by Sensitivity, criticality, Jurisdiction
- **Sensitivity**:
  - Loss to an organization if the information is released to unauthorized entities
  - Organizations can lose trust and spend expensive response efforts in remediation

- **Criticality**
  - Indicator of how the loss will impact the fundamental business process of the organization
  - It is that which is required for the organization to continue business

- **Jurisdiction**
  - The geophysical location of the source / storage point of the data

Classification Guidelines

- When classifying data, take into consideration
  - Who has access to data
  - How the data is secured
  - How long the data is retained
  - What methods used to dispose the data
  - Whether the data needs to be encrypted
  - What use of the data is appropriate
- Keep the classification small
- Classification should not be restrictive and detail oriented (either)
- Each classification should be unique and separate from others; no overlap effects
- Should outline how information is controlled and handled through its life cycle

Data Policy Definition Considerations

Cost

- Cost of providing access to data vs cost of providing the data

Ownership & Custodianship

- Who owns the data and who maintains the data

Privacy

- What data is private, what data is made public

Liability

- How protected the organization is from legal recourse

Sensitivity

- What type of data is in question; what is the impact, type and level of threat, vulnerability for the data

Existing Law and Policy Requirements

- May have impact on enterprise data policy

Policy & Process

- Consideration should be given to legal request for data and policies that may need to be put in places

Data discovery Methods

- **Label based Discovery**:
  - Accurate and sufficient label, helps organization determine what data it controls
  - Labels created by data owners will greatly help in discovery efforts
  - During discovery process, it is easy to collect and disclose relevant and only appropriate data using labels
- **Metadata based discovery:**
  - Set of data that gives information about other data
  - Often automatically created at the same time as the data
- **Content based Discovery:**
  - Pattern-matching discovery that looks for content of datasets
  - Prone to more false positives and generally slow

Data Types

- **Structured Data**
  - Data that is stored according to meaningful structures, discrete types and attributes
  - eg: relational database
- **Unstructured Data**
  - Unsorted data containing all types of content
  - eg: video, audio, emails content
- **Semi Structured Data:**
  - Uses tags or other elements to create fields and records within data without requiring rigid structures
  - Eg: XML, JSON, MongoDB

Storage

Cloud Storage Architectures

- **Long term Storage**
  - Storage designed for use for long term retention
  - Primarily used for backup and regulatory requirements
  - Cheaper to store, can turn out costlier for data retrieval
- **Ephemeral Storage**
  - Storage that is available only as long as the instance does
  - Data will be removed when an instance is terminated
- **Raw Storage**
  - Storage that the customer has direct access to
  - It's a form of virtualization that allows a particular VM to access storage logical unit number (LUN). LUN provides a dedicated portion of the overall storage for the use of the VM

Volume Storage:

- With Volume storage, customer is allocated a storage space within the cloud:
- **File Storage** (File-level Storage or File-based Storage):
  - Data is stored and displayed just as a file structure
  - Big Data analytical tools and processes use File Storage model
- **Block Storage:**
  - Blank volume a customer can you use to put anything
  - Provides flexibility and higher performance
  - Requires greater amount of administration and might entail requiring installation of OS and other app to use
  - Better suited for multiple types of data and kinds (OLTP databases)

Data Dispersion

- Refers to a technique used in cloud computing of breaking data into smaller chunks and storing them across different physical storage devices
- It also allows for **erasure coding** to allow for reconstruction of data is some segments are lost
- **Advantage**: Availability of data
- **Disadvantage**: Data gets dispersed creating legal and regulatory impact; Also, latency can also be an issue, due to the additional overhead required to perform the erasure coding and reconstruct data.

Object Based Storage:

- Data is stored as objects, not as files or blocks
- Objects include not only the production content, but also the associated meta data
- This architecture allows for significant level of description, including, marking, labeling, classification and categorization
- This enhances the capabilities of Indexing capabilities, Data policy enforcement and DLP capabilities
- Object storage is typically associated with IaaS
- VM Snapshots are stored as objects

Storage - SAN

- Provides secure storage among multiple computers
- SAN appears like a single disk to the customer, while storage is spread across multiple locations
- SAN uses **block-level storage** – data is broken into blocks of uniform size.

Storage - NAS

- This network storage solution uses TCP/IP and allows file-level access
- NAS appears to a customer as a single file system
- Most OS offer native support for NAS

Cloud Storage Threats

- For Long term storage, threats include credential theft, compromise, privilege escalation
- Risks to integrity of data
- DoS attack and service outages
- Cryptographic malware-style attacks
- Side channel attacks
- Data corruption or destruction
- Malware and Ransomware
- Improper disposal

Cloud Storage Controls

- Encryption
- Data obfuscation
- Hashing
- Tokenization
- DLP
- Keys , Secrets and certificate management

Encryption in Cloud

Cloud Service	Description
Storage-level Encryption	Encryption of data as it is written to storage. Keys are controlled by the CSP
Volume-level Encryption	Encryption of data written to volumes connected to specific VM Keys are controlled by the customer
Object-level Encryption	Encryption on all objects as they are written to storage Keys are controlled by the CSP
File-level Encryption	Implemented in Customer application Keys can be manually managed or through IRM ~ customer controlled
Application-level Encryption	Implemented in application typically using object storage Data is encrypted by the app prior to storage
Database-level Encryption	File-level by encrypting database files or utilizing transparent encryption Keys are controlled by the customer

Key Management Considerations

- Level of Protection must be higher or same as the data that it is protected
- Key recovery and escrow mechanisms
- Key revocation
- Key Lifetime
- Outsourcing Key Management

Hashing

- A Hash is used to guarantee the integrity of data, a MAC guarantees integrity AND authentication
- A Hash take a single input – a message and produces a message digest
- A MAC algorithm takes two inputs -- a message and a secret key -- and produces a MAC
- Hash can be applied to any size data block
- Hash produces fixed-length output
- One-way hash function is never used in reverse

Hashing Characteristics

- Hash should be computed over the entire message
- Hash should be a **one-way function**
- Given a message and Hash value, computing another message with the same Hash value should be impossible
- Resistant to Birthday attacks

Obfuscation

- Obfuscation refers to application of any technique in order to make the sensitive data less meaningful to unauthorized entities
- Obfuscation can be done in either static or dynamic configurations.
- **Static Obfuscation**: New dataset is created as a copy from the original dataset and only the obscured data is used
- **Dynamic Obfuscation**: data is obscured as it is accessed.

Obfuscation

Technique	Description
Randomization	Replacement of data using random information. Useful when real data needs to be removed but the attributes need to be maintained (length of the string, character set etc) so that testing will be done with equivalent data sets
Anonymization	Removing identifiable data
Pseudo-Anonymization	Removing identifiable data but leaves some elements that could be used to de-anonymize the data
Hashing	Hash are sometimes used to mask or anonymize the data
Shuffling	Using different entries within the data set to represent the data Helps create more realistic test data
Masking	Hiding data with useless characters without removing the data

Obfuscation

Technique	Description
Null	Deleting the raw data or displaying null entries. Some functionalities of the dataset will be drastically reduced.
Tokenization	Replacing sensitive data with a replacement value called token. Tokenization adds significant overhead to the process Two DB are maintained: 1 with the actual live sensitive data The other with nonrepresentational tokens mapped to each piece of the data in the first table **Process flow** User / Process calls the data -> it is authenticated by the token server -> token server retrieves the appropriate token from the token database -> calls the real data mapped to the retrieved token -> passes it to the calling user / process

Obfuscation

Technique	Description
Substitution	Swapping out some information for other data. Can be random or can follow integrity rules.
Value Variance	Applies mathematical changes to Numerical data This can be helpful for creating high realistic data for testing
Homomorphic Encryption	Can be used for Obfuscation The encrypted data is processed without first decrypting it

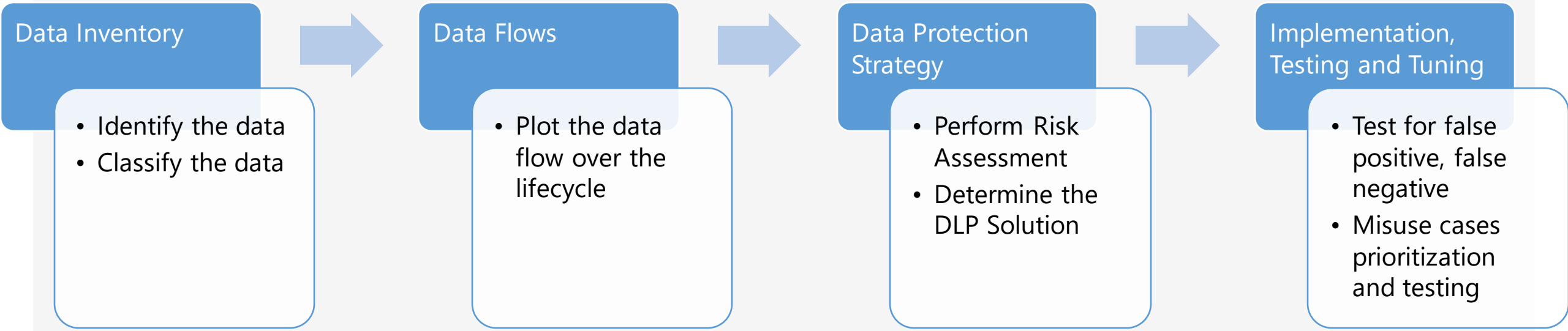
Data Loss Prevention

- Comprises actions that organizations take to prevent unauthorized external parties from gaining access to sensitive data
- DLP is concerned with external parties
- DLP should be integrated as part of Risk Management Approach
- DLP technology determination aspects
  - Sensitive data awareness
  - Policy engine
  - Interoperability
  - Accuracy (most critical)

DLP Goals

- Major Goals
  - Security Control
  - Policy Enforcement
  - Enhanced Monitoring
  - Regulatory Compliance

DLP Approach



Data Analytics Methods

- **Data Lake**:
  - Unstructured data storage mechanism with data in the form of files or blobs
- **Data Warehouse:**
    - Structured data storage in which data is normalized to fit in a defined data model
- **Data Mart:**
    - Data that has been warehoused, analyed and made available for specific usecases.

Data Analytics Methods

- **Data Mining**:
  - Running queries across various fields of data streams to detect and analyze previously unknown trends and pattern
- **Real-time Analytics:**
  - Providing data mining functionality concurrently with data creation and use.
  - It requires automation and require efficiency to perform properly

Data Analytics Methods

- **Online Analytic Processing (OLAP)**:
  - Provides users with analytic processing capabilities for a data source
  - Consists of consolidation, drill-down and slice/dice functions
  - Consolidation gathers multidimensional datasets into cubes
  - Drill-down and slice/dice allows users to analyze subset of the data cube.
  - Forensic analysis often makes use of OLAP to extract relevant information from log files

Data Retention Policy Considerations

- Retention Timeline
- Legal and Regulatory requirements
- Retention
- Data classification
- Data Deletion
- Archival and Retrieval
- Monitoring, Maintenance and Enforcement

Information Rights Management

- 3 Key traits of IRM
- **Data Rights**
  - Describe the actions authorized users can taken on a given asset and how these rights are set, applied, modified and removed
- **Provisioning**
  - Provisioning rights without ensuring business functionality is disturbed is critical
- **Access Models**
  - The access model is a critical part of design and implementation
  - How the data will be accessed is critical to determine the controls that will be applied

IRM Functions

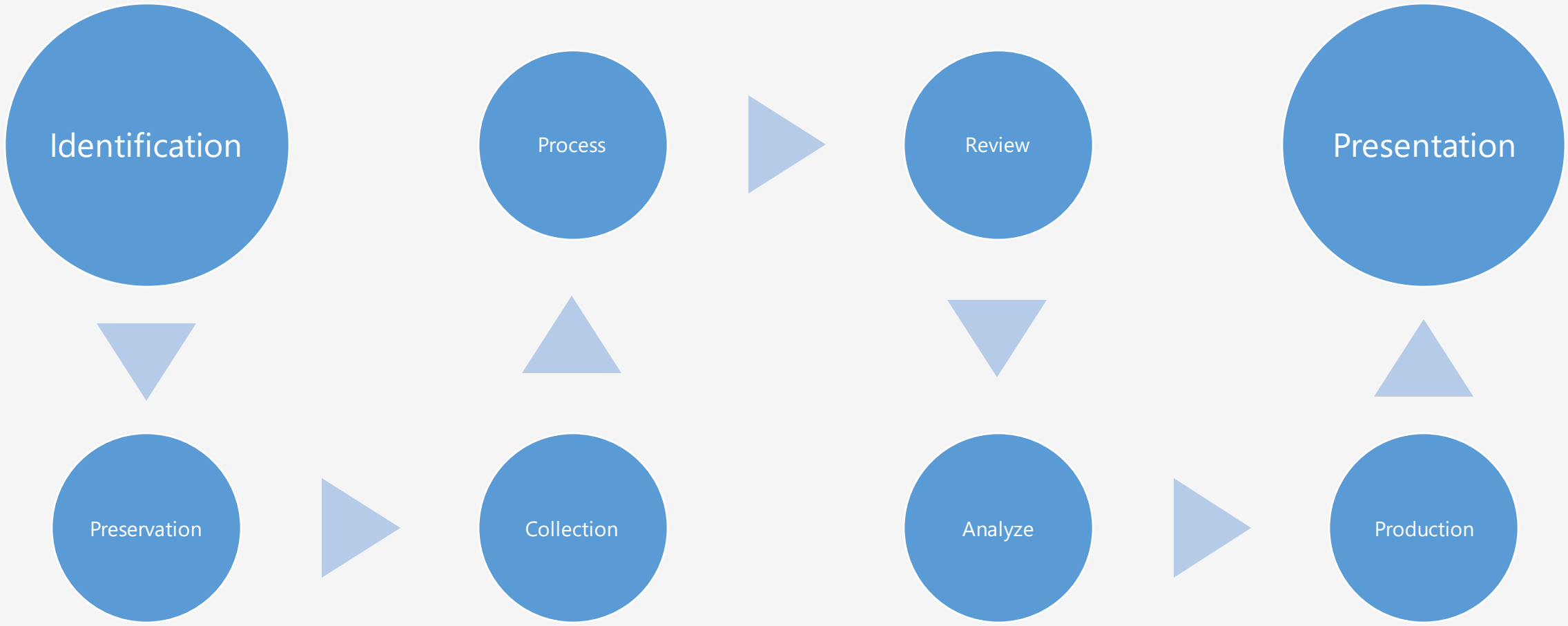
Function	Description
Persistent Protection	IRM should follow the content it protects. This protection should not be easy to circumvent
Dynamic Policy control	Allow data owners to modify ACLs and permissions for the protected data under their control
Automatic Expiration	Access and permissions to protected content should automatically expire, no matter where the content exists
Continuous Auditing	IRM should allow comprehensive auditing of the contents use and access
Replication Restriction	Restrict illegal and unauthorized duplication of protected content
Remote Rights Revocation	The data owner should have the ability to revoke the rights at any time
Interoperability	IRM solutions must offer support for users across different system types

Certificates and Licenses

- Licenses and Certificates are most common methods for identifying the users and systems in an IRM System
- Certificates help validate the identity of the user or computer
- License describes the access rights the user or computer have to the content they are attached to
- IRM attached to files require local clients or web applications to decrypt the file

eDiscovery

- Legal Holds are intended to ensure that data required for a case is collected and preserved



All the best