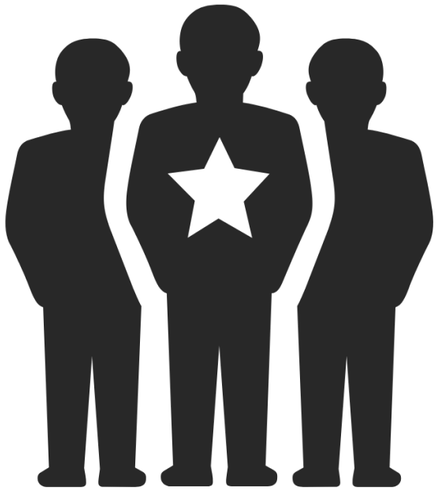
Data governance

UNDERSTANDING MODERN DATA ARCHITECTURE

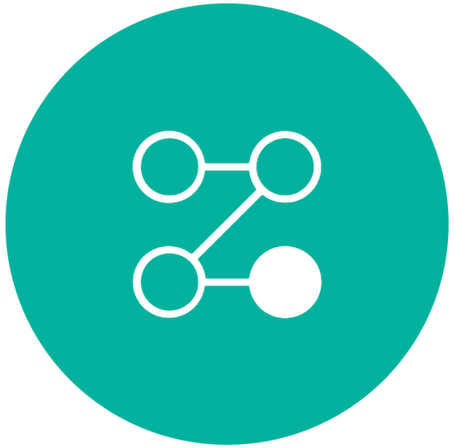
Miller Trujillo
Senior Software Engineer

datacamp

What is data governance?



People



Processes

Tools

- Know our data
  - What data do we have?
  - Quality
  - Origin
- Secure our data
  - Access management
  - Encryption
- Compliance & regulations

[1] https://www.oreilly.com/library/view/data-governance-the/9781492063483/

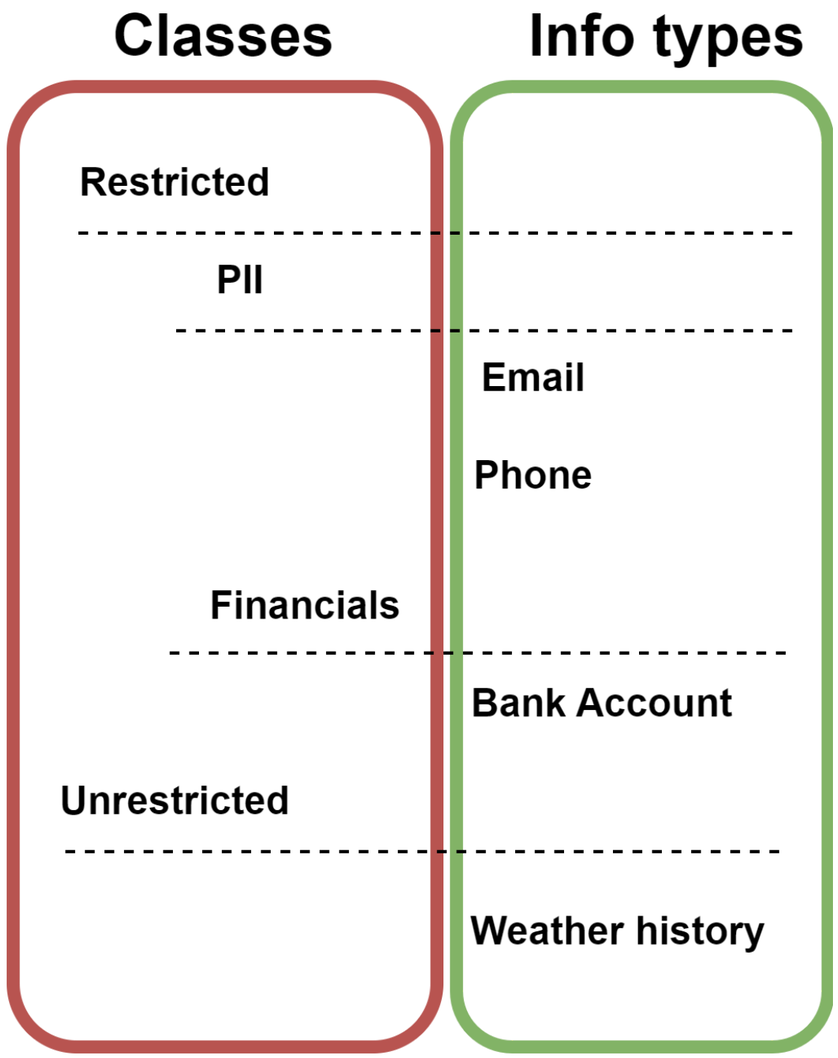
The people

Category	Roles	Responsibilities
Governor/Approver	Data owner/steward	Implement data governance strategy, classify data and, manage the data
User	Data analyst/scientist	Consume data to derive insights and make decisions
Ancillary/Additional	C-executive, legal	Support the overall data governance strategy

The processes

- Know our data
- Classification
- Data lineage - origin
- Data quality
  - What is good and bad data?
  - What to do with bad data?
- Governors normally own these processes

The tools



GCP Dataplex/Catalog

AWS Glue Data Catalog

Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE

Metadata Management

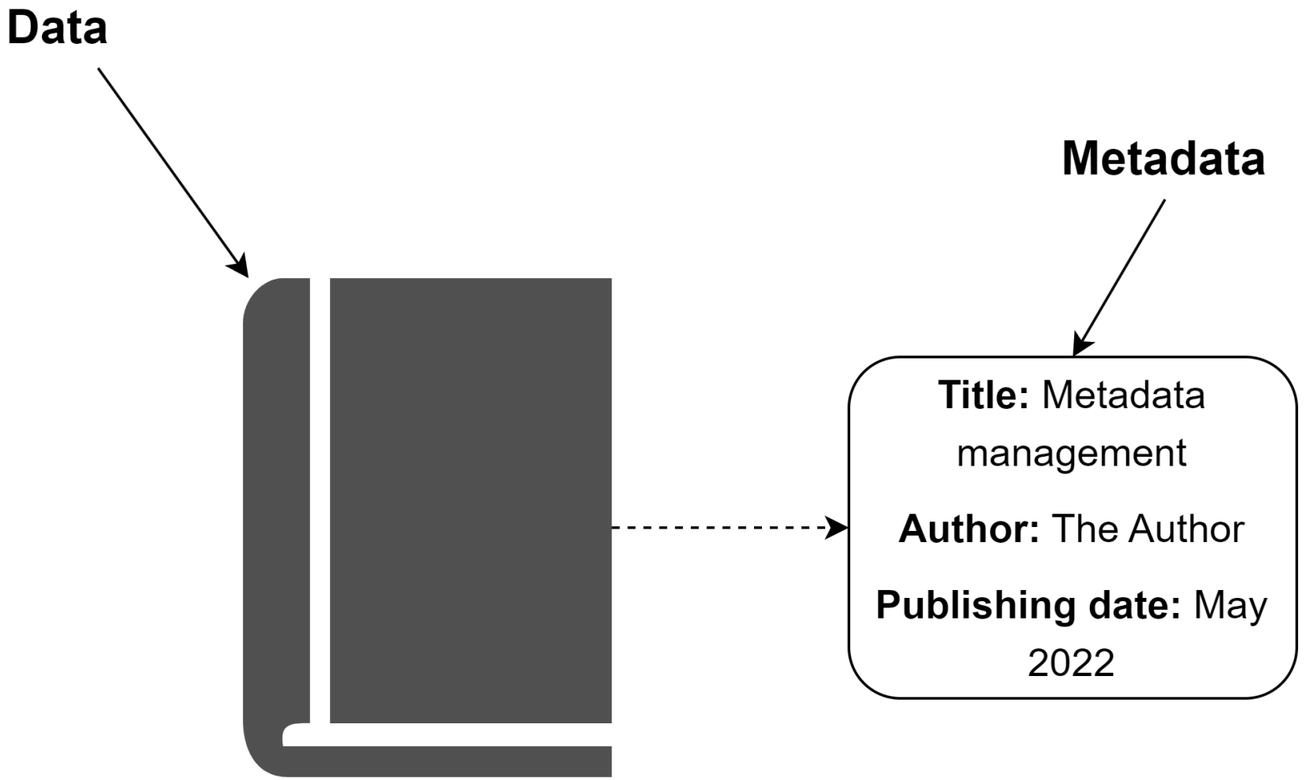
UNDERSTANDING MODERN DATA ARCHITECTURE

Miller Trujillo
Senior Software Engineer

datacamp

What is metadata?

- Data about data



Metadata types

Type of metadata	Data example	Book catalog example
Technical metadata	Data types, relationships, column names, data sources	Book's ISBN, number of pages
Business metadata	Business definitions, rules, data owner	Book's title, author, publisher, genre
Operational metadata	Timestamps, ETL job status, data quality metrics	Date of book acquisition, condition of the book
Usage metadata	Who accessed the data, when, and how it was used	Who checked out the book, when, and for how long

Where to store your metadata?

GCP Data Catalog

- Managed metadata service
- Integrates with GCP services
- Can register external metadata

AWS Glue Catalog

- Central metadata repository
- Integrates with AWS services
- Crawlers can catalog external data

Data catalogs

- Azure Data Catalog
- Apache Atlas
- CKAN
- Datahub
- Collibra
- ...

Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE

Data Security

UNDERSTANDING MODERN DATA ARCHITECTURE

Miller Trujillo
Senior Software Engineer

datacamp

Risks and consequences

Data Breach

- Unauthorized access
- Violates data confidentiality

Impact

- Financial: Fraud, penalties
- Reputation: Reduced trust, customer loss

Data protection measures

Access Control

- "Door to your data house"
- Defines who accesses what data

Encryption

- "Secret letter"
- Protects data at rest and in transit

Data Masking

- Uses data without revealing sensitive parts

Strengthening security in the cloud

IAM

- Permissions, role-based access
- Free to use

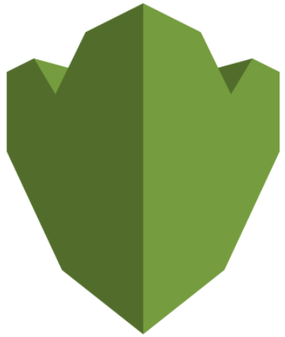
GCP IAM

AWS IAM

KMS

- Safe for cryptographic keys
- Handles key creation, rotation, deletion
- Customer managed key

GCP KMS



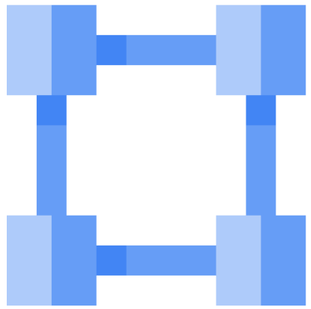
AWS KMS

Strengthening security in the cloud: The network

Firewalls, private networks

VPC (Virtual Private Cloud)

- Virtual, private, secure, isolated network within the cloud



GCP VPC

AWS VPC

VPC Service Controls

- Additional security measures
- Limit access based on the context

- More effort and cost to setup proper networking

Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE

Observability

UNDERSTANDING MODERN DATA ARCHITECTURE

Miller Trujillo
Senior Software Engineer

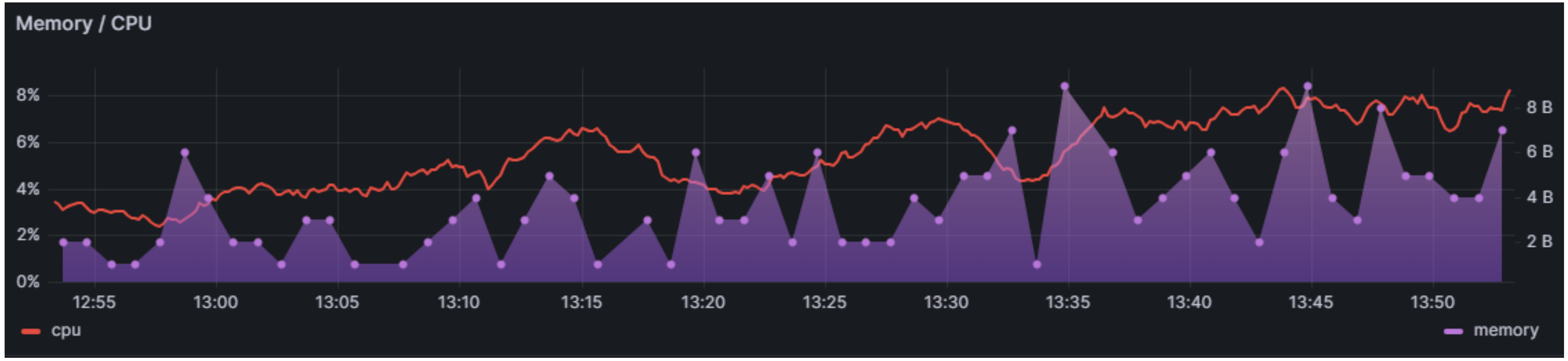
datacamp

What is observability?

- Understanding the insides from the outside.
- Data observability
  - Understand data as it moves
- Complex distributed systems
- Problem resolution
- Reliability

Key aspects of observability: Monitoring & metrics

- Monitoring, logging, and tracing.
- **Monitoring:** Continuously check of system's status
- **Metrics:** Numerical values emitted by the system's



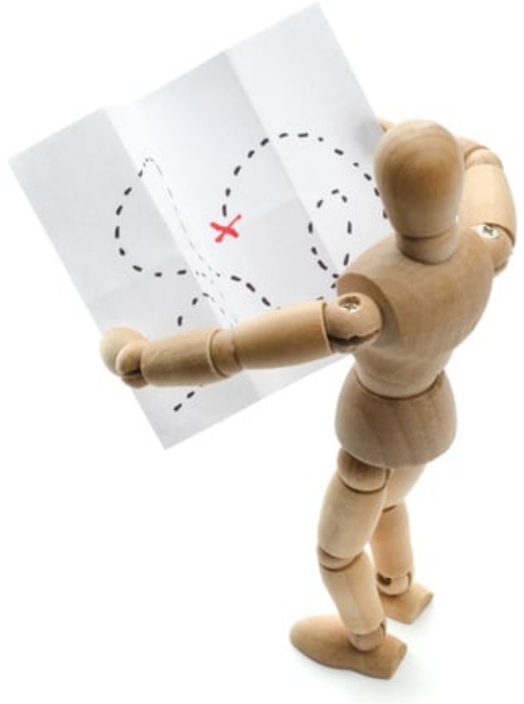
Key aspects of observability: Logging & tracing

logging

- Records of events
  - Examples:
    - Information or debug messages
    - Exception stack trace
- Audit logs
  - **Who** did **what** and **when**

Tracing

- Track an specific request or element through different stages



Observability platforms

AWS Cloudwatch

GCP Operations

Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE