



WELCOME!

to The Complete Data Quality Masterclass

Presentation by George Smarts

Full course - <https://www.udemy.com/course/data-quality-management-masterclass-the-complete-course/>

What will you learn in this course?

Module 1 - The basics of Data Quality and Data Quality Management

Module 2 - What are the Data Quality Dimensions

Module 3 - What are Data Quality Rules and how to set them

Module 4 - Data Quality Techniques:

- > Data Profiling
- > Data Parsing
- > Data Standardization
- > Identity Resolution
- > Data Linkage
- > Data Cleansing
- > Data Enhancement
- > Data inspection and monitoring

Module 5 - The Data Quality Roles

Module 6 - The Data Quality Process

Module 7 - Data Quality Tools

Module 8 - Data Governance vs Data Quality Management

Module 9 - Data Quality best practices

Why take this course?

Reason 1 - The most comprehensive course out there that covers Data Quality topics from A to Z

Reason 2 - Properly structured course divided into 60+ lessons

Reason 3 - Practical! We look at real examples of how what we learn is applied

Reason 4 - I do not teach only on how to do something, but also why it is done

Reason 5 - Best practices from the industry

Reason 6 - Practice materials and resources included

Reason 7 - Course from someone with real experience



Download the course resources

Resource 1 - PDF version of the presentation

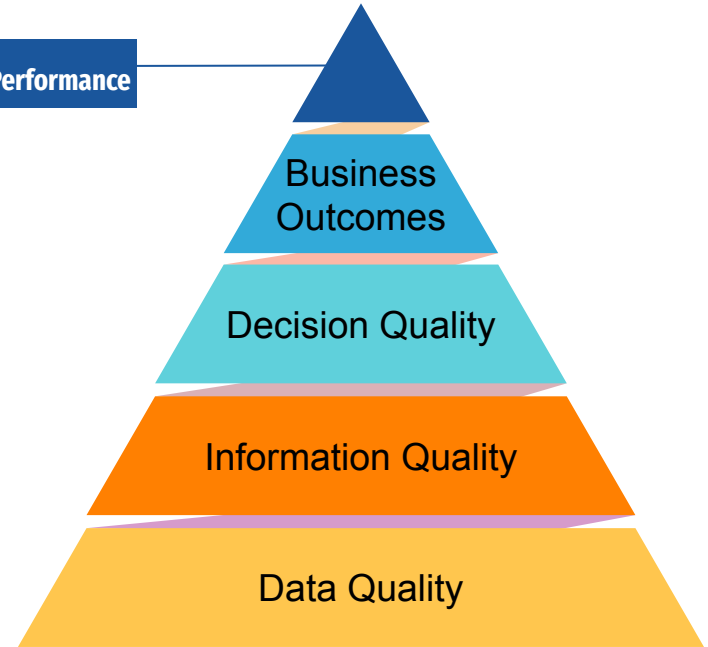
Resource 2 - Sample dataset for hands on exercises

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"

Company Performance



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

Impact of poor data quality



Decision Making

Data based decisions and policies are only as good as the data they are based on



Reputation risk

GDPR issues and negative media coverage



Missed opportunities

Failure to deliver service to customers
Failure to sale to relevant contacts

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.

Data error for \$125M

What Happened?

On September 23, 1999 NASA lost a \$125M Mars Climate Orbiter

Why it Happened?

Miscalculations due to use of imperial units instead of metric units sent the orbiter off course.

NASA used metric units in their software system

The spacecraft builder used imperial units

Note: According to NASA, the cost of the mission was \$327.6 million



Why do we have bad data?

- Intuition is more important to management than data
- Manual data entry errors
- Data Silos
- Data migration and conversion projects
- Scaling of the business and its datasets
- No Data Governance rules



Data Quality Dimensions



Data Quality Rules



- Also known as data validation rules
- Business rules for data
- Automated data quality checks
- Primary tool for determining data quality

Defining your Data Quality Rules

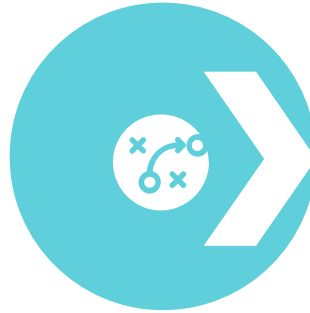
Step 1



Documentation of requirements

- > Technical description
- > Importance/Priority
- > Listing dependent data quality rules and data attributes

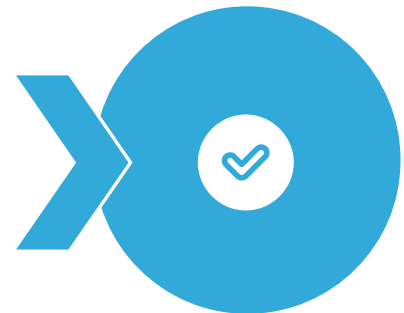
Step 2



Translation into machine-readable format

- > Create machine executable validation rule
- > Define test cases

Step 3



Implementation

- > Implement the automatic check in your data sources
- > Review

Data Profiling

Data Profiling is a technique for discovering and investigating data quality issues, such as duplication, lack of consistency, and lack of accuracy and completeness.

- Data Profiling involves analyzing one or multiple data sources and collecting metadata
- The data steward uses the results to investigate the origin of the data errors
- Previously done manually, now using Data Quality Tools
- The tools provide data statistics, such as degree of duplication, etc



Data Parsing

Data Parsing is the process of separating complex data entries into separate fields. It can also mean converting data into a different data format.

Example - converting full name or address into separate fields

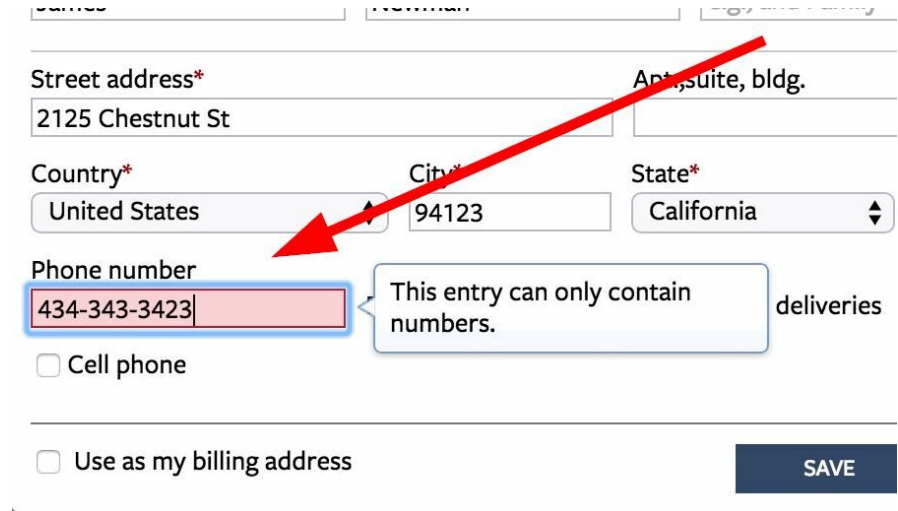
Example - converting HTML into plain text



Data Parsing Benefits

Extract pieces of data to validate if it follows a specific pattern!

- Use Cases for Data Parsing:
 - Check phone number
 - Check mail addresses
 - Check names
 - Check addresses



Form fields and validation:

- Street address*: 2125 Chestnut St
- City*: 94123
- State*: California
- Phone number: 434-343-3423
- Country*: United States
- Ant, suite, bldg.
- Cell phone: ☐
- Use as my billing address: ☐
- SAVE button

Validation message: This entry can only contain numbers.

deliveries

Data Standardization

"Data Standardization is the process of converting data to a common format"



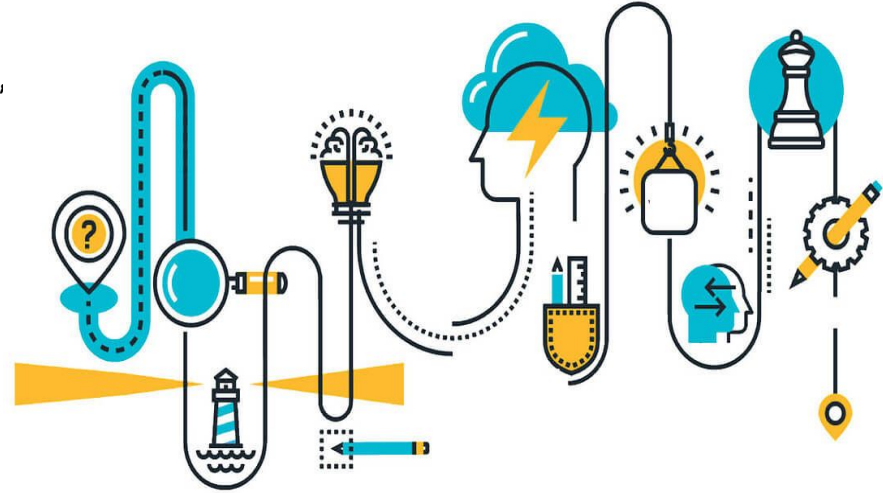
Identity Resolution

“Identity Resolution is a process that checks, validates and appends information across datasets to create a single, data-rich profile for a person, business or other entry”



Identity Resolution Process

1. **Identify** - channels, platforms and devices
2. **Connect** - connect the dots between the different channels, platforms, devices
3. **Match** - based on a defined set of attributes (same household, IP, wi-fi network, timing patterns, etc)
4. **Validate** - validate that it is the same identity
5. **Activate** - create a single, data-rich profile



7 Benefits of Identity Resolution

1. Single Customer View
2. Personalized Customer Experience
3. Happy Employees
4. Understand customer's network
5. Focused view for each function
6. Contextual Marketing
7. Governance



Data Linkage

“Data linkage, also known as record linkage, is the process of identifying, matching and merging records that correspond to the same person from several datasets or even within one dataset”



Data Cleansing

“Data cleansing is the process of resolving corrupt, inaccurate, incomplete or irrelevant data”



Data Enhancement

“Data enhancement is a data improvement process that adds information from third-party or internal datasets to increase the value that the organization derives from the data ”

Key Points:

- *Data enhancements builds on parsing, standardization and record linkage*
- *Sometimes called data enrichment*



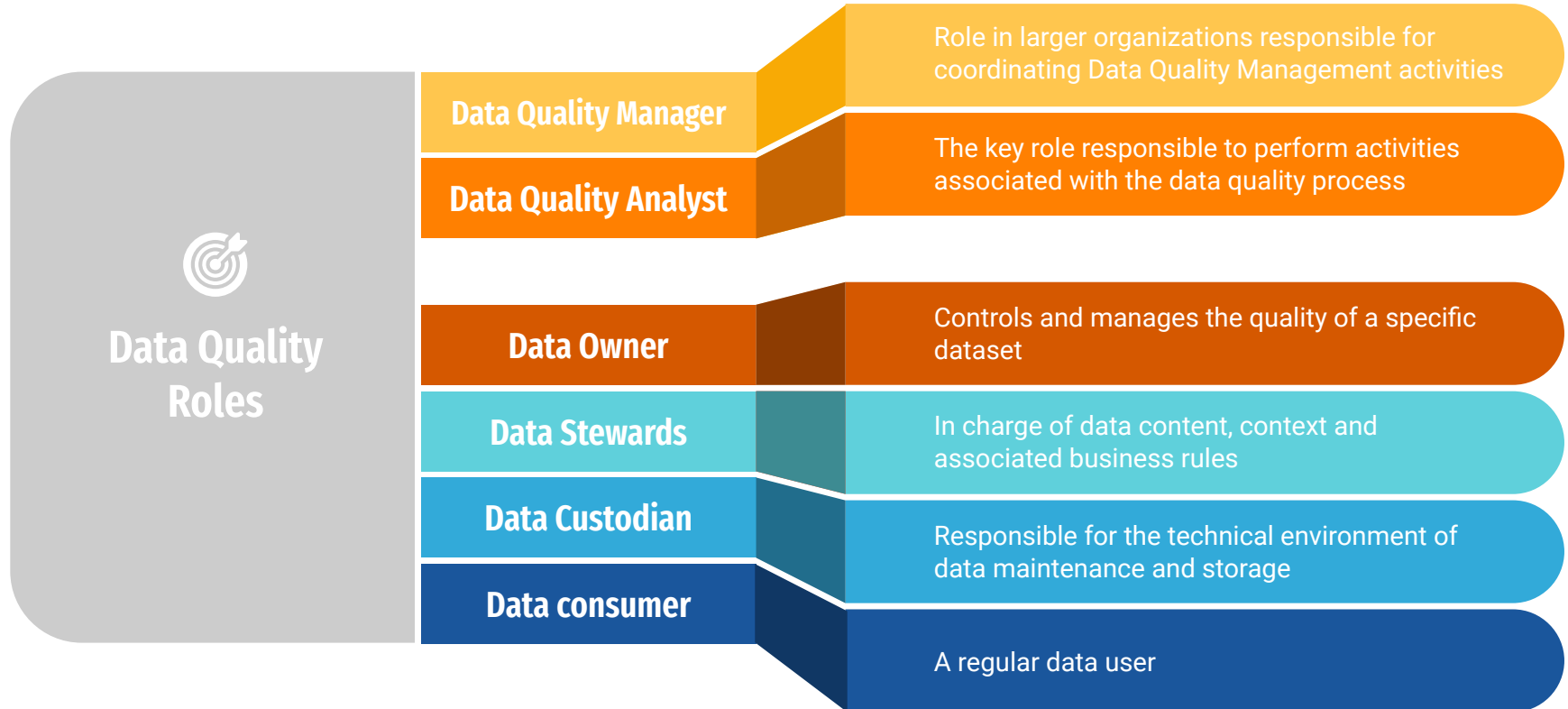
Data Inspection and Monitoring

The process of Data Inspection and Monitoring:

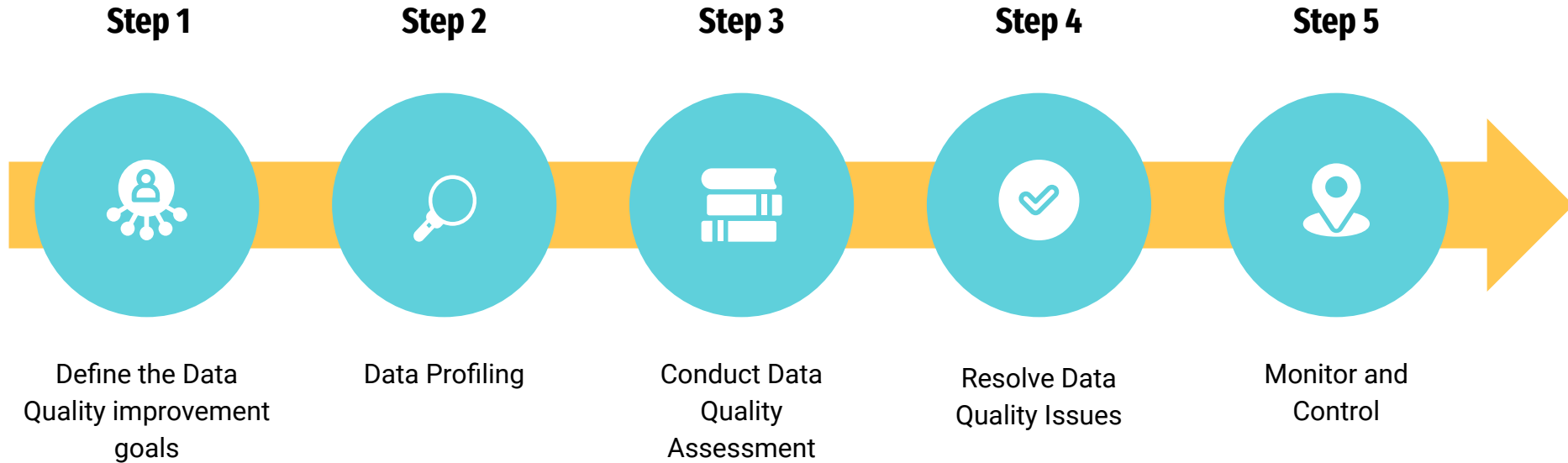
- *Data Profiling exposes potential business rules*
- *Data Quality Analyst documents the rules and confirms their criticality*
- *The rules describe the end-user data expectations*
- *The rules are used to measure and monitor the data*
- *Monitoring of the rules provide a proactive assessment of compliance*
- *The results are used to populate Data Quality Metric scorecard for Leadership*



Data Quality Roles



Data Quality Improvement Process



Data Root Cause Analysis - 5 Whys

1. Why do we have incorrect full name entries?

Due to manual data entry errors

2. Why do we have manual data entry errors?

The fields in the systems have no mandatory format

3. Why do the fields in the system do not have a mandatory format?

It was requested but not prioritized for delivery by the IT team

4. Why was it not prioritized for delivery by IT?

Limited resources for data team requests

5. Why are there limited resources for data team requests

Management strategy to dedicate 90% of IT resources to new software development

Importance of Data Quality Tools

- Data Quality tools help to make data more trustworthy and easier to manage
- Data Quality tools are becoming more critical due to the increased data complexity
- Data Quality tools can handle a variety of tasks:
 - Data consolidation
 - Data cleansing
 - Data validation reconciliation
 - Sample testing
 - Data mapping
 - Validating mailing addresses and contact info
 - Data analytics
 - and much much more!



What is important in a DQ Tool?

Profiling

Parsing

**Record Linkage
and Merging**

Data Cleansing

Standardization

**Reporting
functionality**

**Multi Data
Domain
Support**

Scalability

**Role-Based
Usability**

Performance

Magic Quadrant for Data Quality Solutions

Figure 1: Magic Quadrant for Data Quality Solutions



Source: Gartner (September 2021)

Steps to choose the correct DQ tool



Data Governance

Why spend time learning about Data Governance?

- You need both Data Quality Management and Data Governance
- A Data Governance Framework supports Data Quality initiatives
- Improves the long term success of data quality initiatives

Data Quality vs Data Governance

Data Quality

To what extent the data is accurate, complete, timely, valid, unique and consistent

Key Data Quality questions:

- How complete is the data?
- Is the data accurate without duplicates?
- How quickly can users access the data?

Data Governance

The process of managing the availability, usability, integrity and security of the data in the enterprise systems

Key Data Quality questions:

- How well does the data comply with the GDPR?
- Who is responsible for what parts of our data strategy?
- How do we train people to comply with the data policies?

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

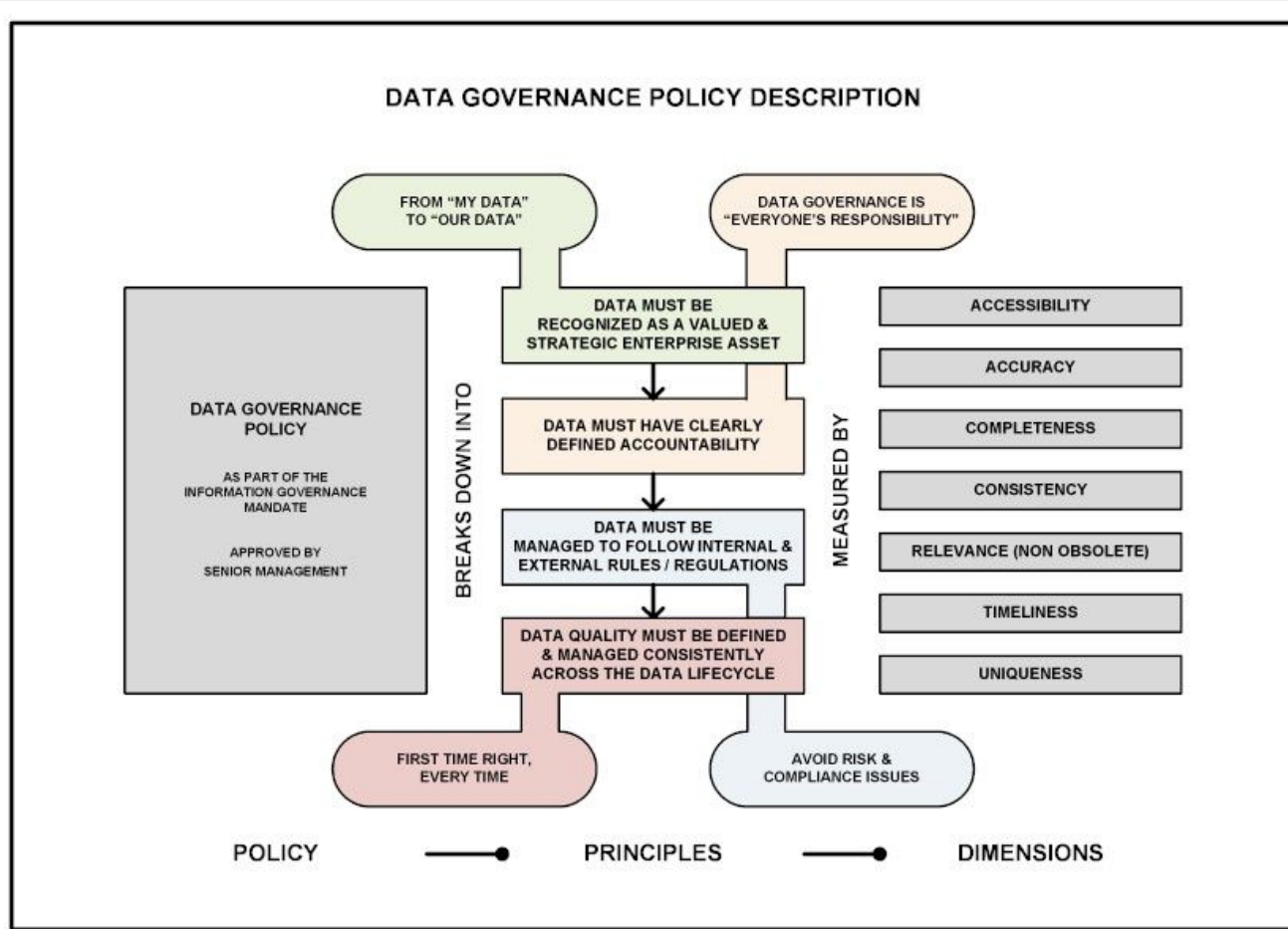
Yahoo: \$85 million

Capital One: \$80 million

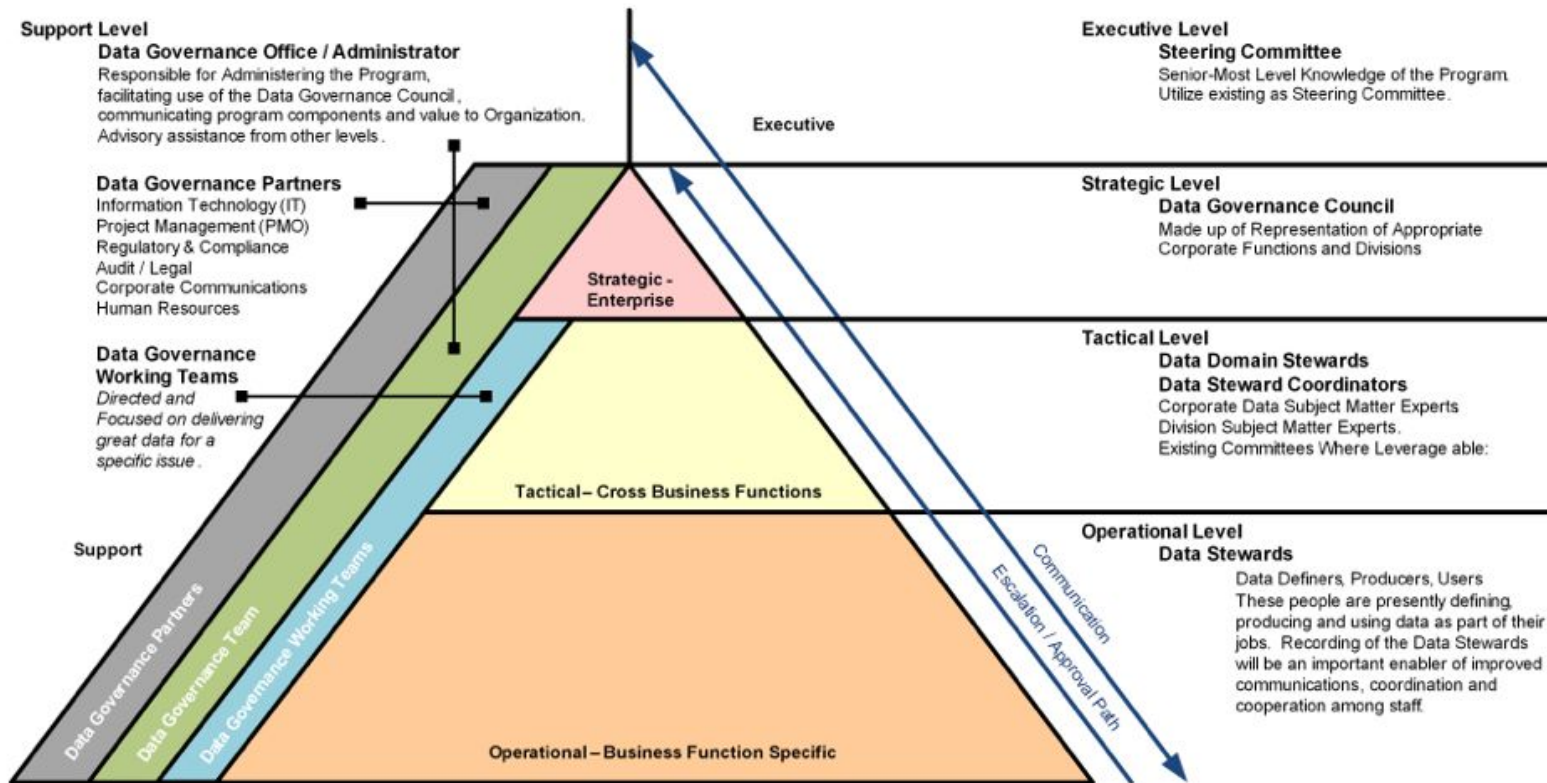


*The U.S. relies on a “[combination of legislation, regulation and self-regulation](#)” rather than government intervention alone. There are approximately 20 industry- or sector-specific federal laws, and more than 100 privacy laws at the state level (in fact, there are 25 privacy-related laws in California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

Data Governance Core Principles



Governance Roles and Responsibilities



Data Quality Best Practices





Congratulations!

Infographics

You can add and edit some **infographics** to your presentation to

- Choose your favourite infographic and insert it in your presentation using Ctrl C + Ctrl V or Cmd C + Cmd V in Mac.
- Select one of the parts and **ungroup** it by right-clicking and choosing “Ungroup”.
- **Change the color** by clicking on the paint bucket.
- Then **resize** the element by clicking and dragging one of the square-shaped points of its bounding box (the cursor should look like a double-headed arrow). Remember to hold Shift while dragging to keep the proportions.
- **Group** the elements again by selecting them, right-clicking and choosing “Group”.
- Repeat the steps above with the other parts and when you’re done editing, copy the end result and paste it into your presentation.
- Remember to choose the “**Keep source formatting**” option so that it keeps the design. For more info, please visit **Slidesgo School**.

