



Data Quality and Metrics

IS465: Data Management and Governance

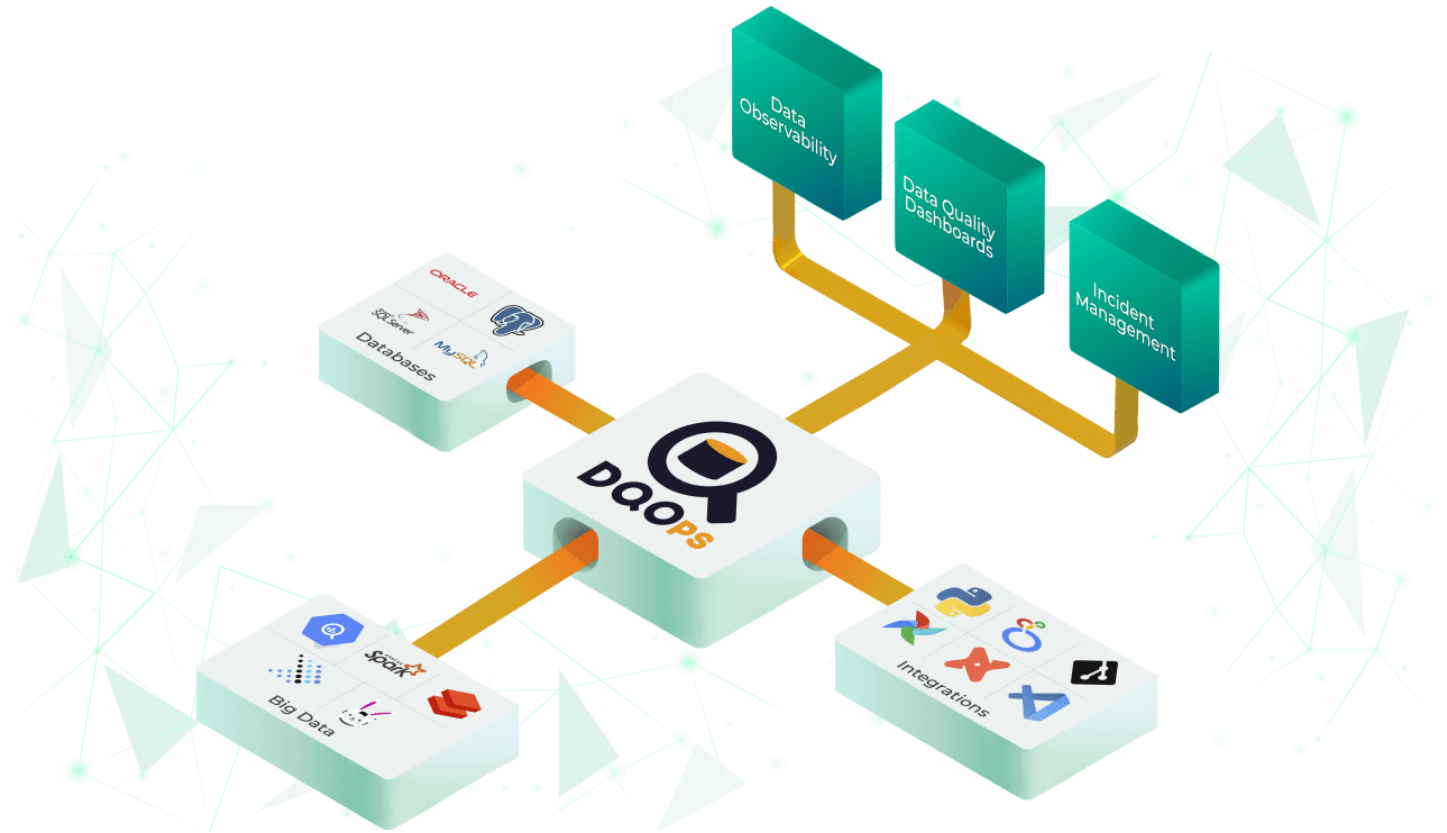
Outline

- Introduction to Data Quality
- Data Quality Dimensions
- Data Quality Metrics
- Data Quality Assessment
- Data Quality Improvement
- Data Cleansing and Normalization Techniques

Introduction to Data Quality

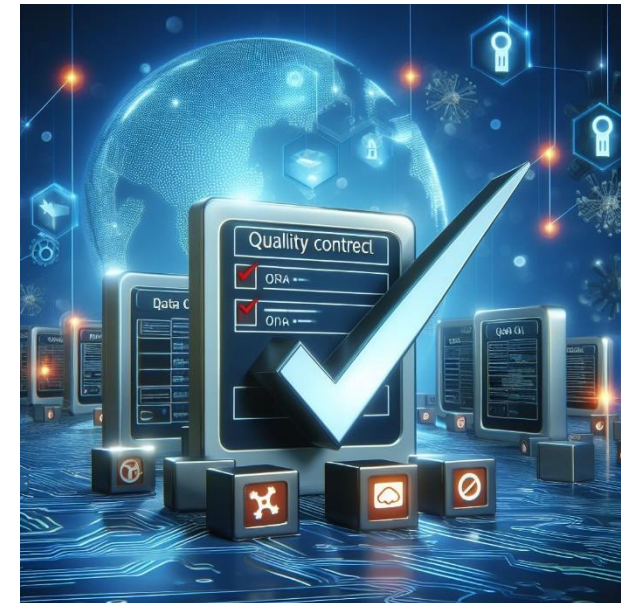
Ensuring Trustworthy Data for Informed Decision Making

- Data quality is a critical aspect of any organization's data management strategy.



What is Data Quality?

- Data quality refers to the degree to which data is accurate, complete, consistent, and reliable, and meets the requirements of its intended use
- Data quality is a multifaceted concept that encompasses various aspects of data, including accuracy, completeness, consistency, and reliability.



Why is Data Quality Important?

- Informed decision making
- Operational efficiency
- Customer satisfaction
- Competitive advantage
- Regulatory compliance

The Cost of Poor Data Quality

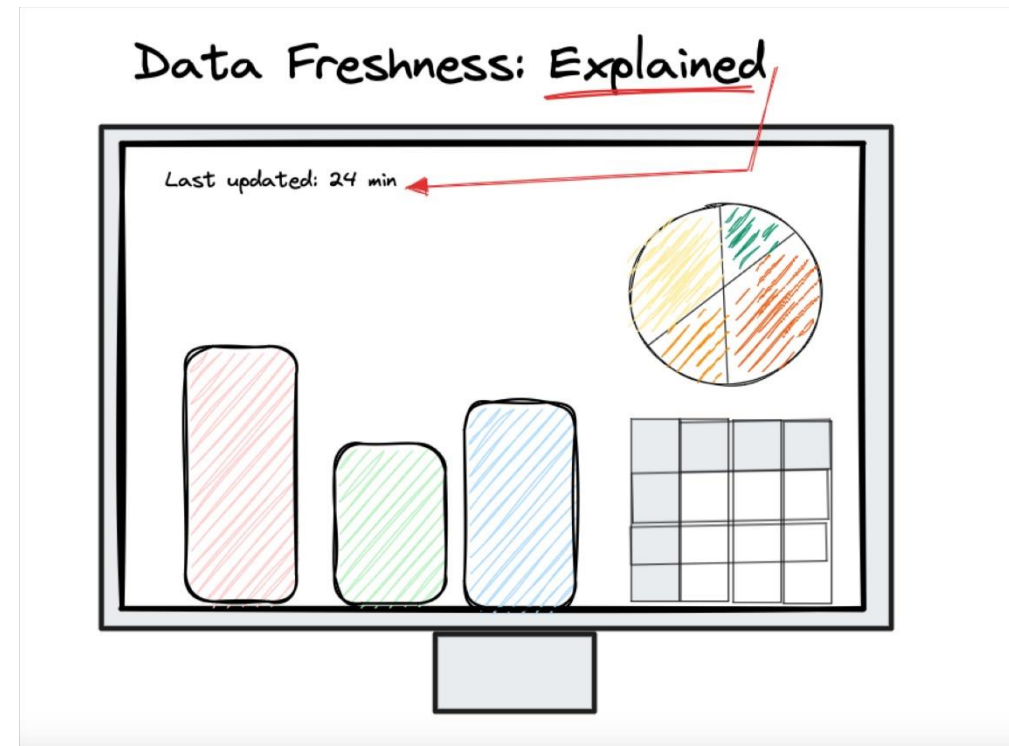
- Inaccurate decisions
- Financial losses
- Reputational damage
- Decreased customer trust
- Compliance issues

Data Quality Dimensions



Data Quality Metrics

- Error rate
- Data freshness
- Data coverage
- Data consistency ratio
- Data accuracy ratio



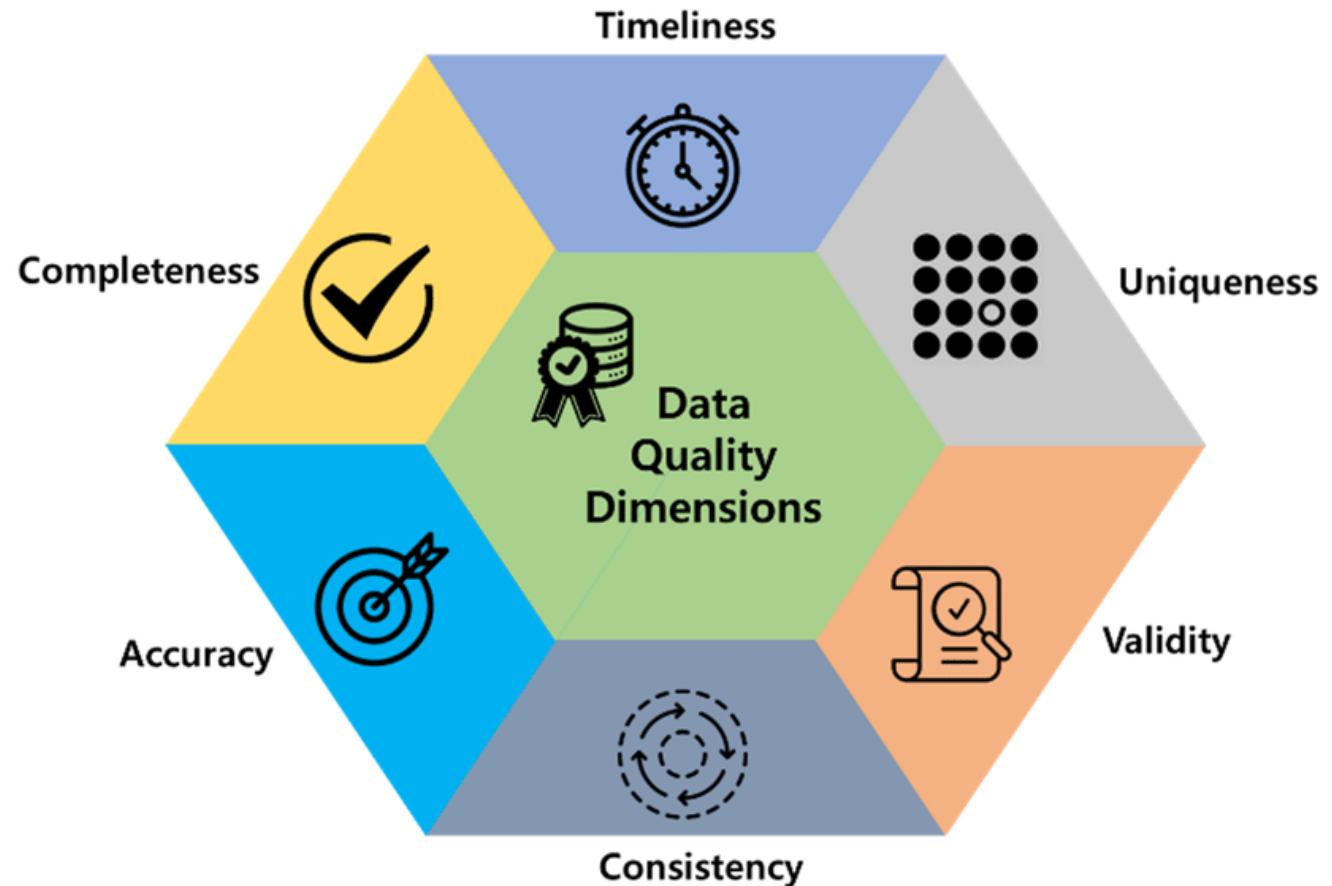
Data Quality in a Nutshell

- Data quality is critical for informed decision making
- Poor data quality has severe consequences
- Data quality can be measured across various dimensions
- Data quality metrics provide a way to quantify data quality

Data Quality Dimensions

Measuring the Quality of Your Data

- Data quality dimensions provide a framework for measuring and improving the quality of your data.



Accuracy

- The correctness of data values, ensuring that they reflect the real-world values they represent.
- Example: "Correct addresses, phone numbers, and dates of birth"
- Accuracy is critical in ensuring that data is reliable and trustworthy.



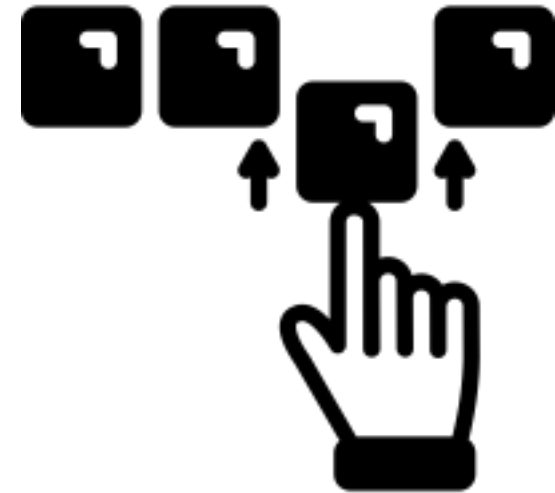
Accuracy Metrics

- Accuracy metrics provide a way to measure and quantify the correctness of data values.
 - Accuracy rate: percentage of correct data values
 - Error rate: percentage of incorrect data values
 - Precision: degree of accuracy in data values



Completeness

- The presence of all required data values, ensuring that no essential information is missing.
- Example: "All fields filled in a form, complete customer information"
- Completeness is essential in ensuring that data is comprehensive and usable.



Completeness Metrics

- Completeness metrics provide a way to measure and quantify the presence of all required data values.
 - Completeness rate: percentage of complete data records
 - Missing value rate: percentage of missing data values

Consistency

- The uniformity of data values across different systems or sources, ensuring that data is presented in a consistent format.
- Example: "Consistent formatting of dates, consistent use of abbreviations"
- Consistency is critical in ensuring that data is easily interpretable and comparable.

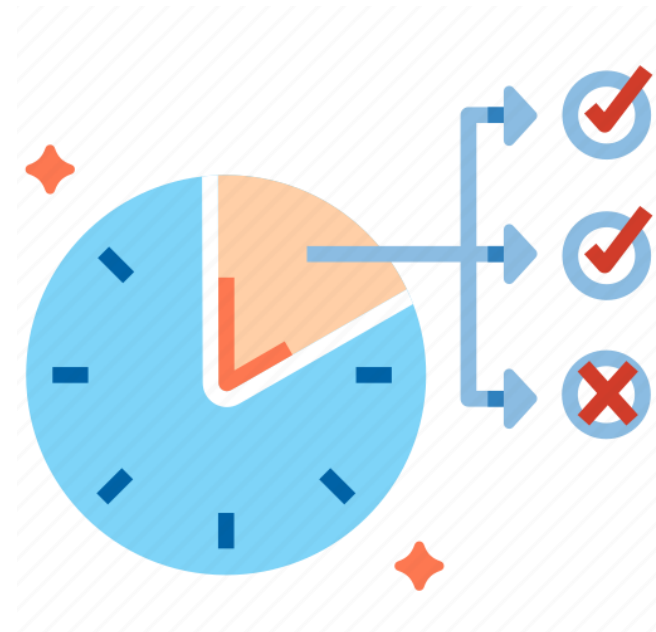


Consistency Metrics

- Consistency metrics provide a way to measure and quantify the uniformity of data values.
 - Consistency rate: percentage of consistent data values
 - Inconsistency rate: percentage of inconsistent data values

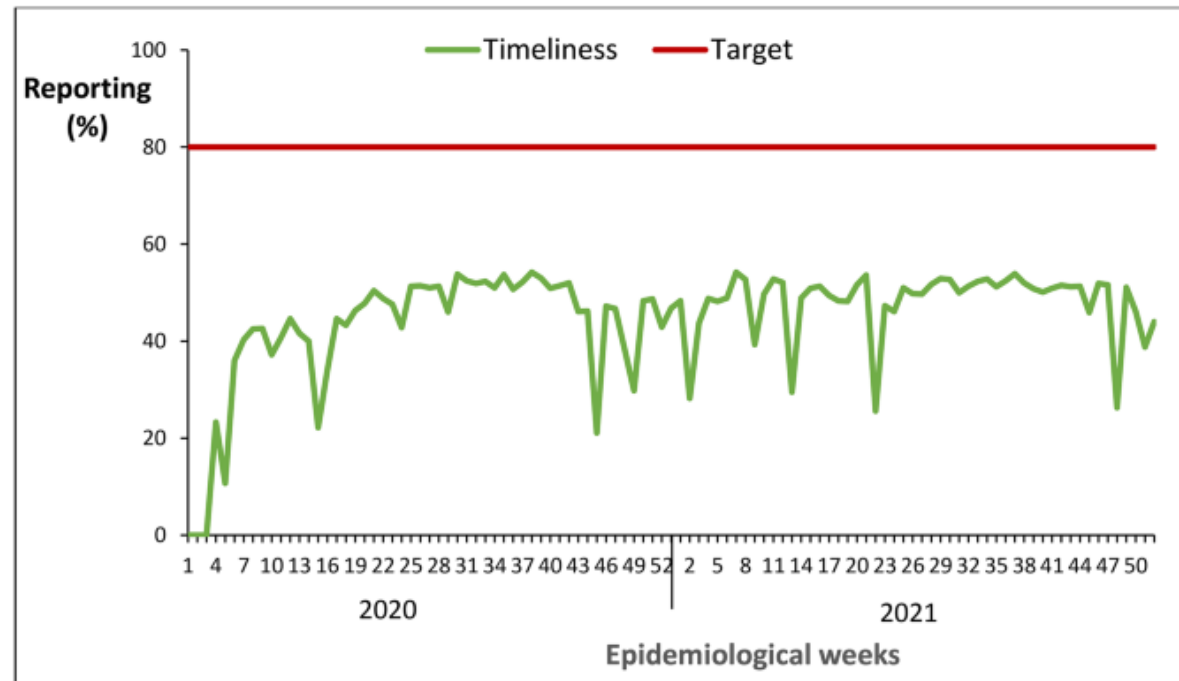
Timeliness

- The currency of data values, ensuring that data is up-to-date and reflects the current state of the real world.
- Example: "Up-to-date customer information, current product prices"
- Timeliness is essential in ensuring that data is relevant and actionable.



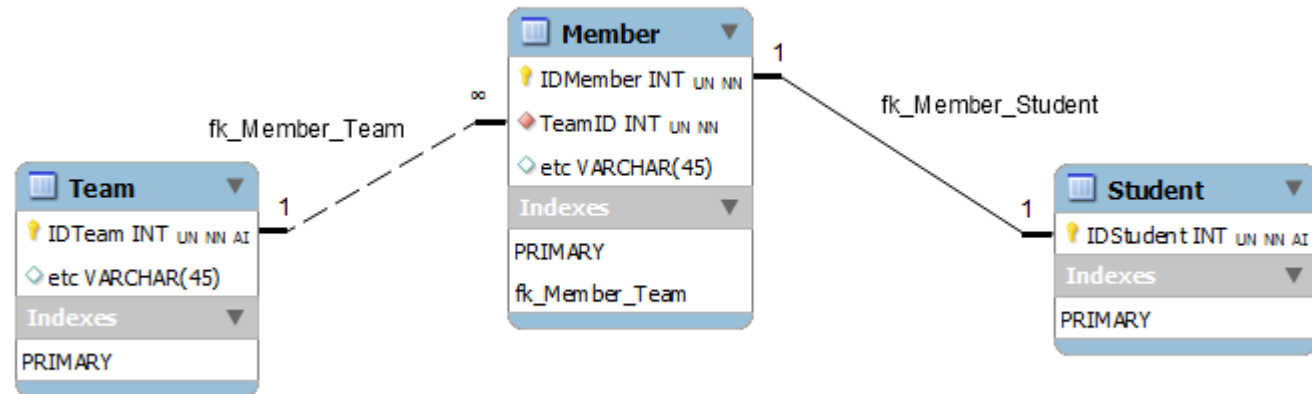
Timeliness Metrics

- Timeliness metrics provide a way to measure and quantify the currency of data values.
 - Timeliness rate: percentage of up-to-date data values
 - Staleness rate: percentage of outdated data values



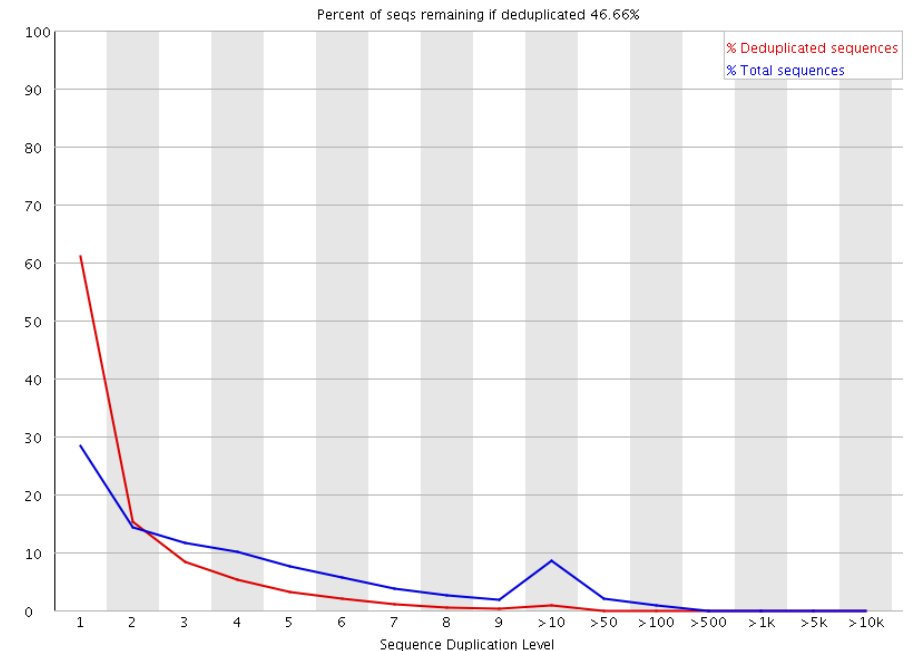
Uniqueness

- The absence of duplicate data values, ensuring that each data record is unique and distinct.
- Example: "No duplicate customer records, unique product IDs"
- Uniqueness is critical in ensuring that data is accurate and reliable.



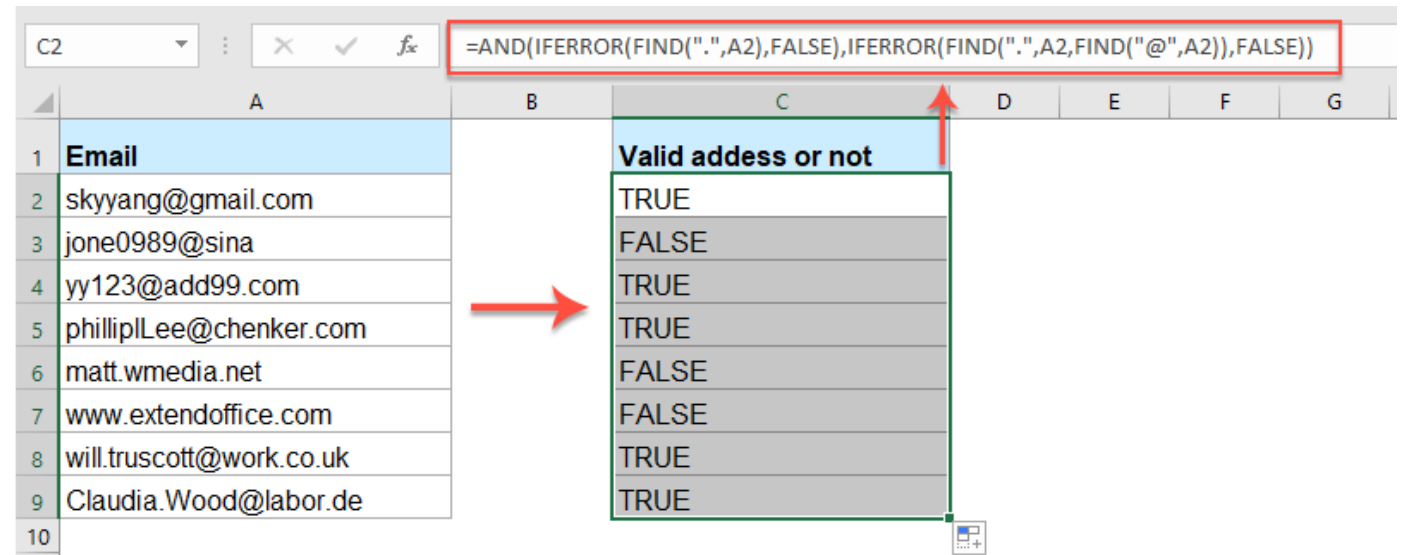
Uniqueness Metrics

- Uniqueness metrics provide a way to measure and quantify the absence of duplicate data values.
 - Uniqueness rate: percentage of unique data records
 - Duplication rate: percentage of duplicate data records



Validity

- The adherence of data values to defined rules and constraints, ensuring that data is correct and meaningful.
- Example: "Valid email addresses, correct formatting of phone numbers"
- Validity is essential in ensuring that data is reliable and trustworthy.



The screenshot shows an Excel spreadsheet with a formula bar at the top displaying the formula: `=AND(IFERROR(FIND(".",A2),FALSE),IFERROR(FIND(".",A2,FIND("@",A2)),FALSE))`. The formula is highlighted with a red box. Below the formula bar, the spreadsheet has columns A, B, C, D, E, F, and G. Column A is labeled "Email" and contains a list of email addresses. Column B is empty. Column C is labeled "Valid address or not" and contains the results of the formula. A red arrow points from the formula bar to the formula in cell C2. Another red arrow points from the "Valid address or not" header in cell C1 to the results in column C. The results in column C are: TRUE, FALSE, TRUE, TRUE, FALSE, FALSE, TRUE, TRUE.

	A	B	C	D	E	F	G
1	Email		Valid address or not				
2	skyyang@gmail.com		TRUE				
3	jone0989@sina		FALSE				
4	yy123@add99.com		TRUE				
5	philliplLee@chenker.com		TRUE				
6	matt.wmedia.net		FALSE				
7	www.extendoffice.com		FALSE				
8	will.truscott@work.co.uk		TRUE				
9	Claudia.Wood@labor.de		TRUE				
10							

Validity Metrics

- Validity metrics provide a way to measure and quantify the adherence of data values to defined rules and constraints.
 - Validity rate: percentage of valid data values
 - Invalidity rate: percentage of invalid data values

Data Quality Dimensions Summary

- Accuracy: correctness of data values
- Completeness: presence of all required data values
- Consistency: uniformity of data values
- Timeliness: currency of data values
- Uniqueness: absence of duplicate data values
- Validity: adherence to defined rules and constraints

Data Quality Metrics

Measuring the Quality of Your Data

- Data quality metrics provide a way to measure and quantify the quality of your data
- Metrics:
 - Quantitative
 - Qualitative, and
 - Composite metrics

Quantitative Metrics

- Quantitative metrics provide a clear and objective measure of data quality, allowing for easy comparison and tracking over time.
- Numerical measures of data quality, providing a precise and objective assessment of data quality.
- Examples:
 - Accuracy rate
 - Completeness rate
 - Error rate
 - Data freshness

QUANTITATIVE METRICS

Word Count



Time



Readability



Grade Level



Support Costs



Web Analytics









Quantitative Metrics Examples

- Accuracy rate: percentage of correct data values
- Completeness rate: percentage of complete data records
- Error rate: percentage of incorrect data values
- Data freshness: percentage of up-to-date data values

Qualitative Metrics

- Descriptive measures of data quality, providing a subjective assessment of data quality.
- Examples:
 - Data relevance
 - Data consistency
 - Data usability
 - Data freshness

QUALITATIVE METRICS

Quality		
Usability		
Customer Satisfaction		

Qualitative Metrics Examples

- Data relevance: how well the data meets the needs of its intended use
- Data consistency: how well the data conforms to a standard or format
- Data usability: how easy the data is to use and understand
- Data freshness: how up-to-date the data is

Composite Metrics

- Combinations of multiple metrics to provide a comprehensive view of data quality.
- Composite metrics provide a comprehensive view of data quality, taking into account multiple aspects of data quality
- Examples:
 - Data quality score
 - Data health index
 - Data fitness score

Composite Metrics Examples

- Data quality score: a weighted average of multiple metrics
- Data health index: a composite metric that takes into account multiple aspects of data quality
- Data fitness score: a metric that assesses the overall fitness of the data for its intended use

Benefits of Data Quality Metrics

- Improved data quality
- Increased confidence in data-driven decisions
- Better data management and governance
- Enhanced data analytics and insights

Challenges of Data Quality Metrics

- Defining and selecting relevant metrics
- Collecting and analyzing data
- Interpreting and communicating results
- Maintaining and updating metrics over time

Data Quality Assessment

Evaluating the Quality of Your Data

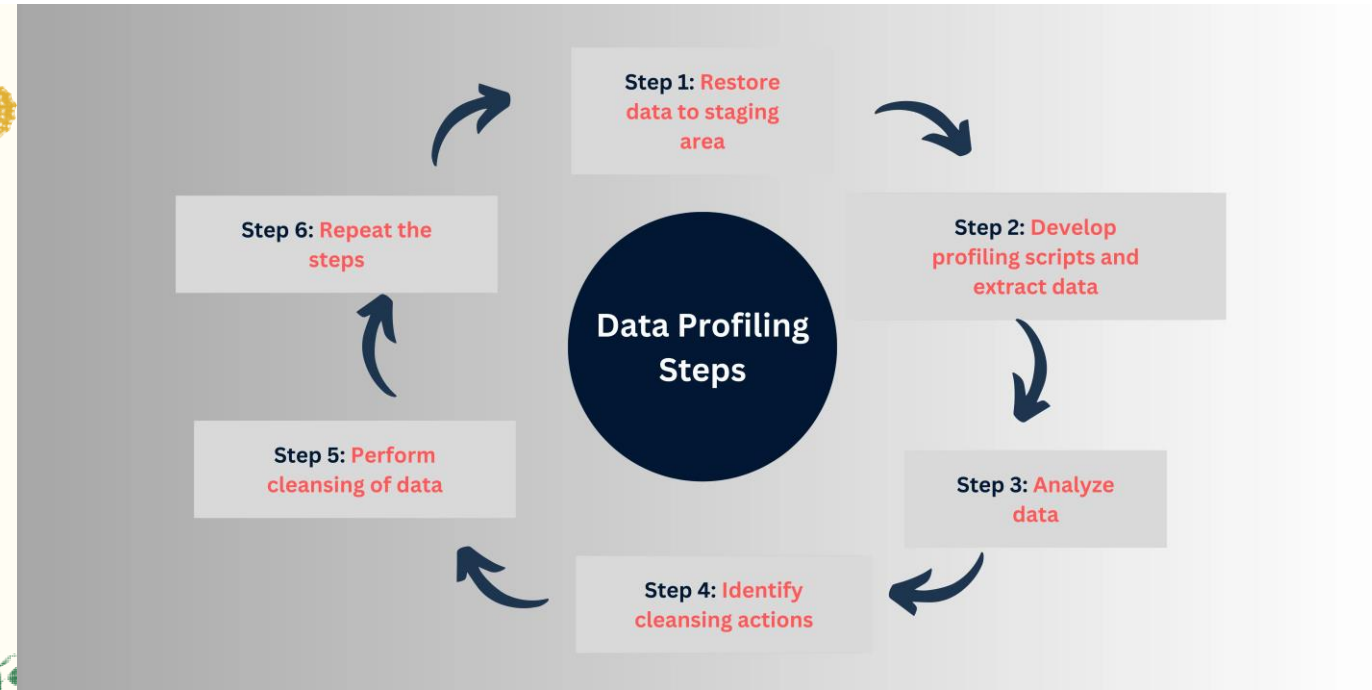
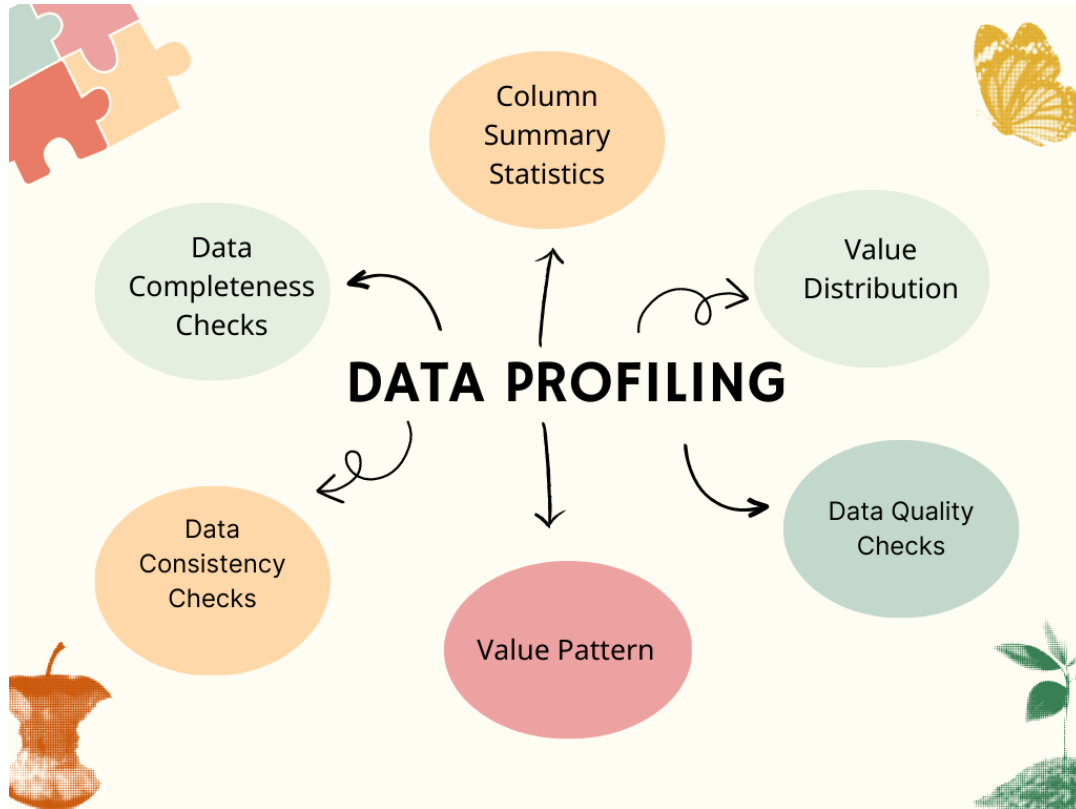
- Data quality assessment is the process of evaluating the quality of your data to identify issues and opportunities for improvement.
 - Data profiling
 - Data auditing
 - Data validation
 - Data quality dashboards

Data Profiling

- Analysis of data distribution, patterns, and relationships to identify quality issues.
- Data profiling provides a detailed understanding of the data, helping to identify quality issues and opportunities for improvement.
- Examples:
 - Analyzing data distribution to identify outliers
 - Identifying patterns in data to detect anomalies
 - Examining relationships between data elements to identify inconsistencies

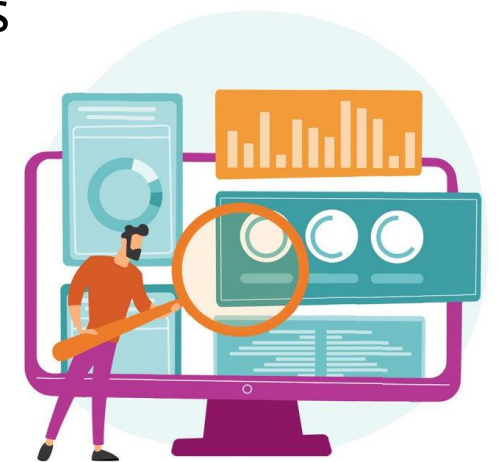


Data Profiling

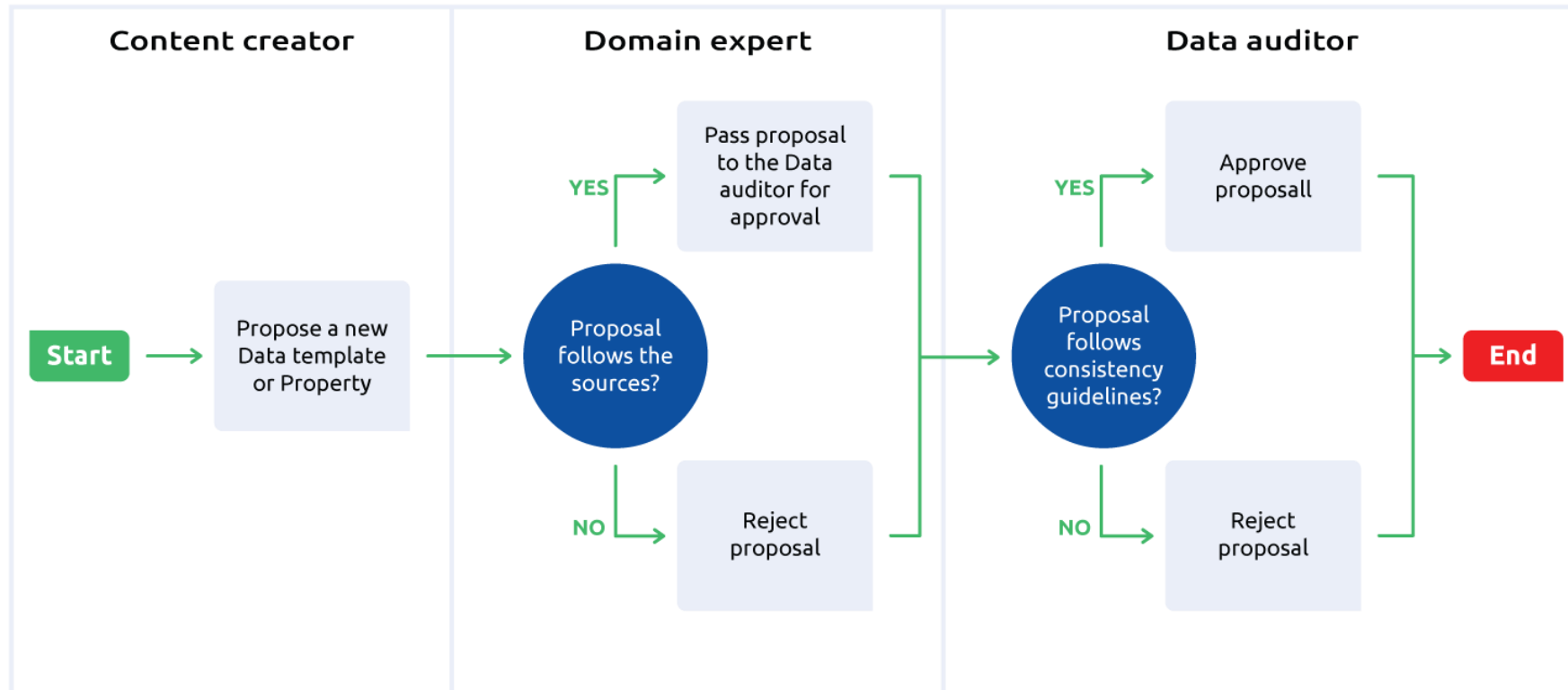


Data Auditing

- Verification of data values against a set of rules or constraints.
- Data auditing ensures that data values conform to expected rules and constraints, helping to identify and correct errors.
- Examples:
 - Checking data values against a set of predefined rules
 - Verifying data values against external sources
 - Identifying data values that do not conform to expected formats



Data Auditing



Data Validation

- Checking of data values against a set of predefined criteria.
- Data validation ensures that data values meet the required criteria, helping to identify and correct errors.
- Examples:
 - Checking data values against a set of predefined formats
 - Verifying data values against a set of business rules
 - Identifying data values that do not conform to expected criteria



Data Validation

Data Validation Process



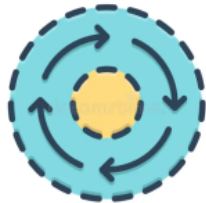
Data Type



Data Range



Data Constraint



Data Consistency



Code Structure



Code Validation

Data Validation

Benefits of Data Validation

Improves the
Efficiency of Data

1

Identifies
Inaccuracies

2

Reveals New
Data Insights

3

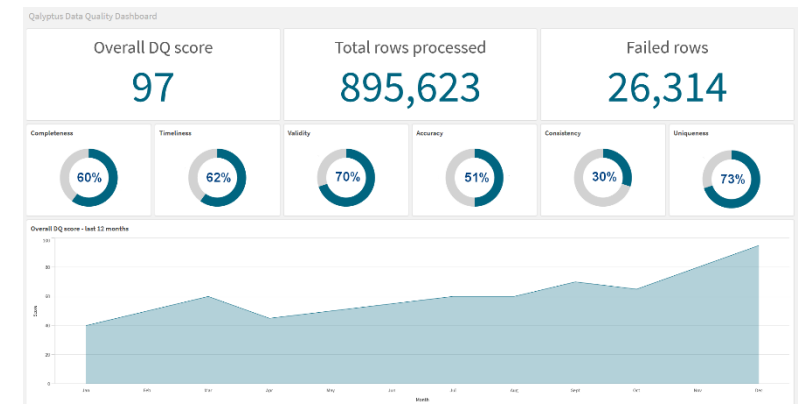
Enhances Customer
Satisfaction

4

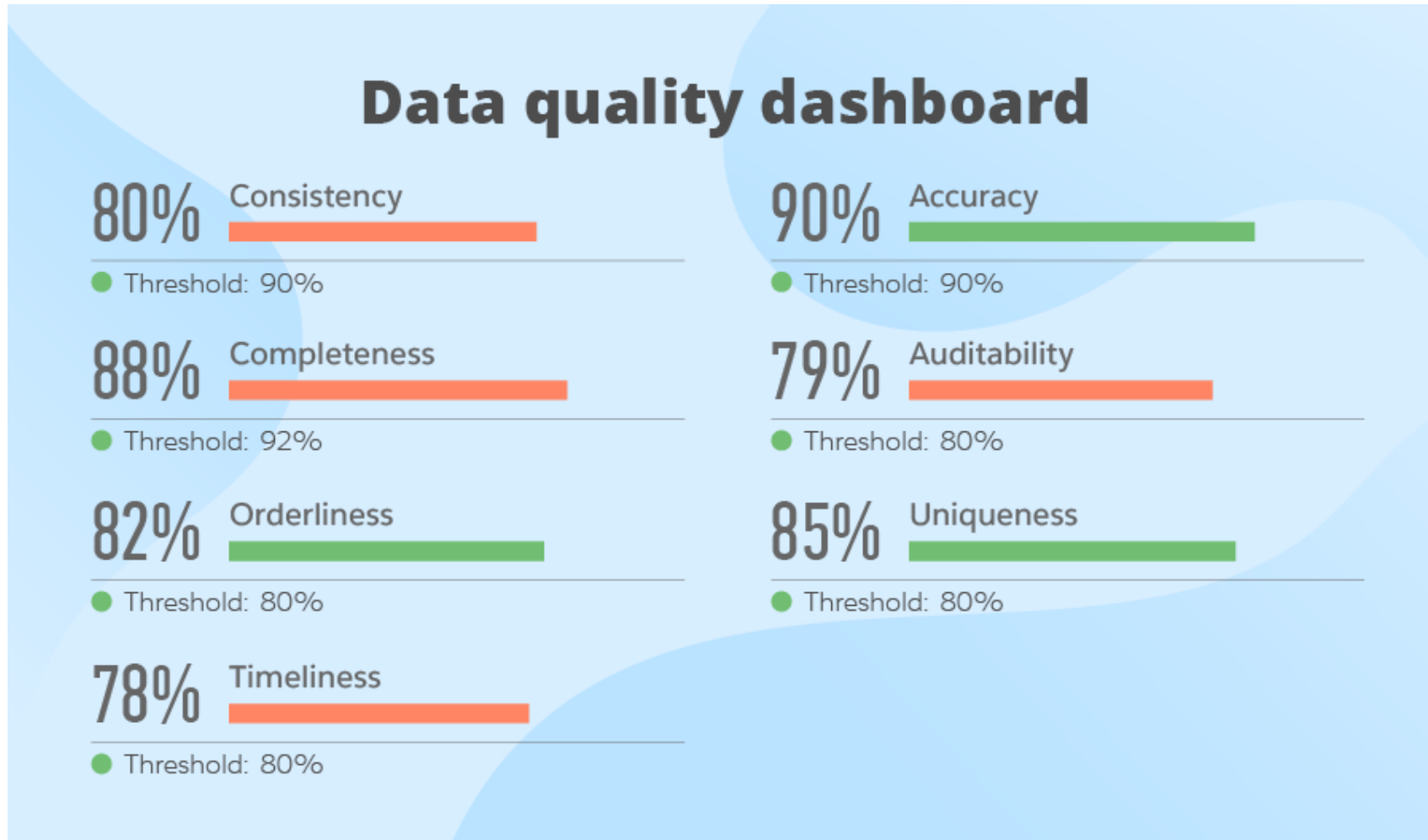


Data Quality Dashboards

- Visual representations of data quality metrics and trends.
- Data quality dashboards provide a visual representation of data quality, helping to identify issues and opportunities for improvement.
- Examples:
 - Dashboards showing data quality metrics such as accuracy and completeness
 - Dashboards showing data quality trends over time
 - Dashboards providing alerts and notifications for data quality issues



Data Quality Dashboards



Data Quality Assessment

- Benefits of Data Quality Assessment
 - Improved data quality
 - Increased confidence in data-driven decisions
 - Better data management and governance
 - Enhanced data analytics and insights
- Challenges of Data Quality Assessment
 - Defining and selecting relevant metrics
 - Collecting and analyzing data
 - Interpreting and communicating results
 - Maintaining and updating assessments over time

Data Quality Assessment Tools

- Examples:
 - Data profiling tools such as IBM InfoSphere QualityStage
 - Data auditing tools such as SAP Data Services
 - Data validation tools such as Talend Data Quality
 - Data quality dashboards such as Tableau



Best Practices for Data Quality Assessment

- Define clear goals and objectives
- Select relevant metrics and tools
- Collect and analyze data regularly
- Communicate results effectively
- Continuously monitor and improve

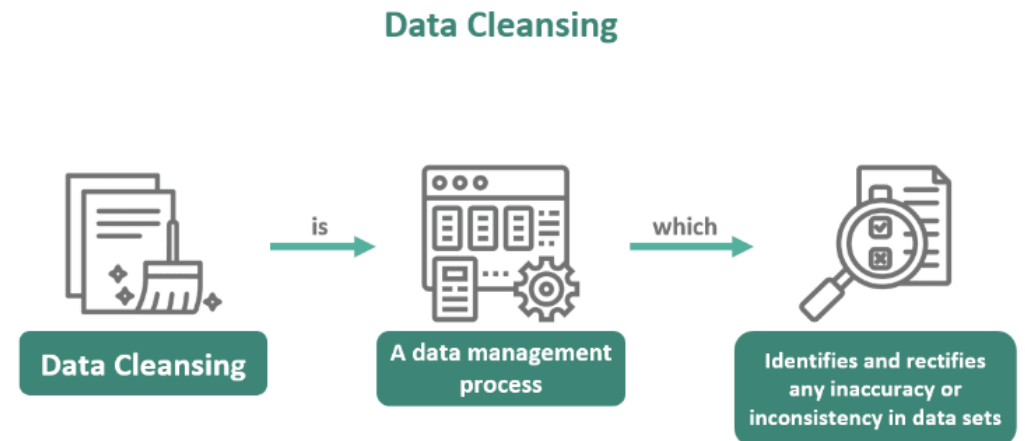
Data Quality Improvement

Enhancing the Quality of Your Data

- Data quality improvement is the process of enhancing the quality of your data to make it more accurate, complete, and reliable.
 - Data cleansing
 - Data normalization
 - Data quality rules
 - Data quality monitoring

Data Cleansing

- Identification and correction of errors, inconsistencies, and inaccuracies in data.
- Data cleansing is an essential step in data quality improvement, helping to identify and correct errors and inconsistencies in data.
- Examples:
 - Correcting spelling errors in customer names
 - Removing duplicate records
 - Filling in missing values



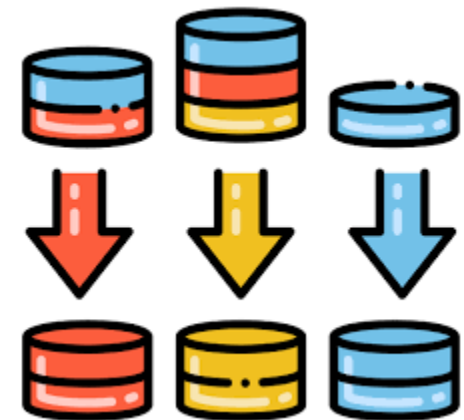
Data Cleansing Techniques

- Data scrubbing: removing unnecessary characters or values
- Data parsing: breaking down data into smaller components
- Data transformation: converting data into a consistent format



Data Normalization

- Transformation of data into a consistent format to improve quality and reduce redundancy.
- Data normalization helps to improve data quality by transforming data into a consistent format and reducing redundancy.
- Examples:
 - Standardizing date formats
 - Aggregating data into a single record
 - Summarizing data into a concise format

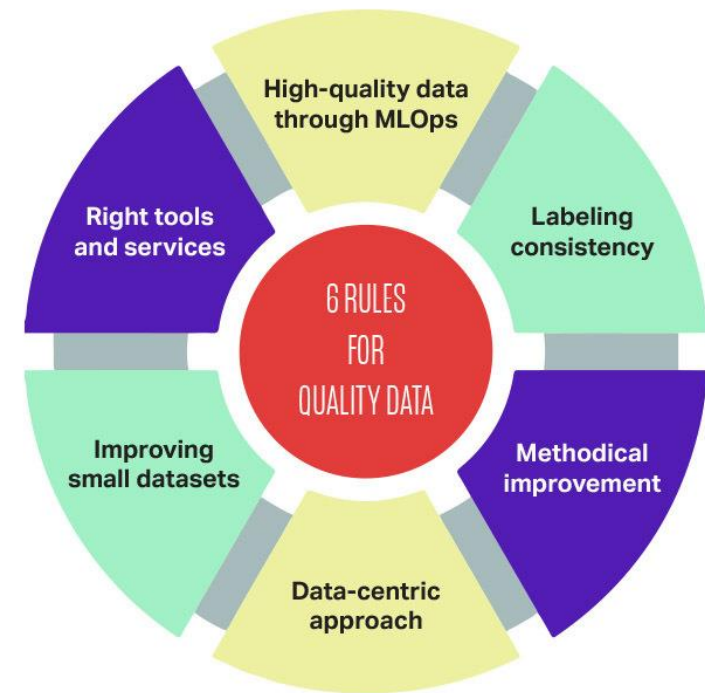


Data Normalization Techniques

- Data standardization: applying a consistent format to data
- Data aggregation: combining data into a single record
- Data summarization: condensing data into a concise format

Data Quality Rules

- Definition and implementation of rules to prevent data quality issues.
- Data quality rules help to prevent data quality issues by defining and implementing rules for data entry, validation, and metrics.
- Examples:
 - Defining rules for data entry
 - Implementing data validation checks
 - Establishing data quality metrics



Data Quality Monitoring

- Ongoing surveillance of data quality to detect and respond to issues.
- Data quality monitoring is an essential step in data quality improvement, helping to detect and respond to data quality issues in real-time.
- Examples:
 - Monitoring data quality metrics
 - Detecting data quality issues
 - Responding to data quality issues

Data Quality Improvement

- Benefits of Data Quality Improvement

- Improved data accuracy
- Increased confidence in data-driven decisions
- Better data management and governance
- Enhanced data analytics and insights

- Challenges of Data Quality Improvement

- Defining and implementing data quality rules
- Collecting and analyzing data quality metrics
- Communicating data quality issues to stakeholders
- Maintaining and updating data quality improvement processes

Data Cleansing and Normalization Techniques

Improving Data Quality through Data Cleansing and Normalization

- Data cleansing and normalization are essential steps in data quality improvement, helping to remove errors, inconsistencies, and inaccuracies from data and transform it into a consistent format.

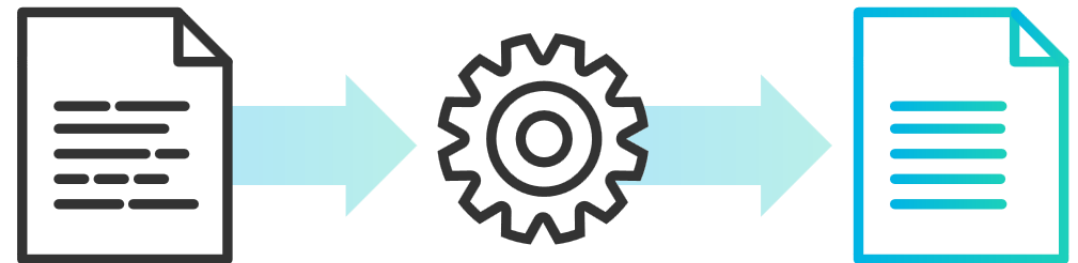
Data Scrubbing

- Removal of incorrect or invalid data values.
- Data scrubbing is an essential step in data cleansing, helping to remove errors and inconsistencies from data.
- Examples:
 - Removing invalid email addresses
 - Correcting spelling errors in customer names
 - Deleting duplicate records



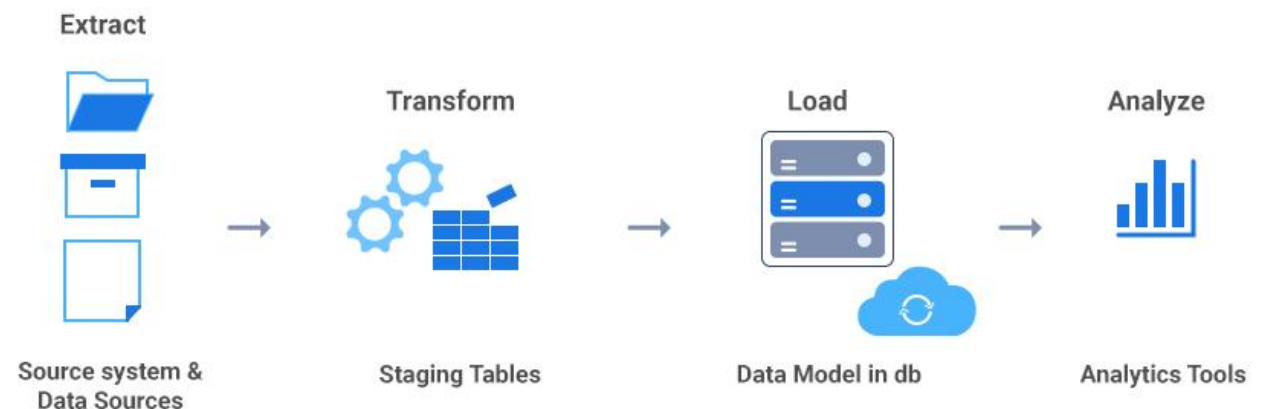
Data Parsing

- Extraction of relevant data values from unstructured or semi-structured data.
- Data parsing is a technique used to extract relevant data values from unstructured or semi-structured data, helping to improve data quality and reduce errors.
- Examples:
 - Extracting names and addresses from unstructured text
 - Parsing data from log files
 - Extracting data from social media posts



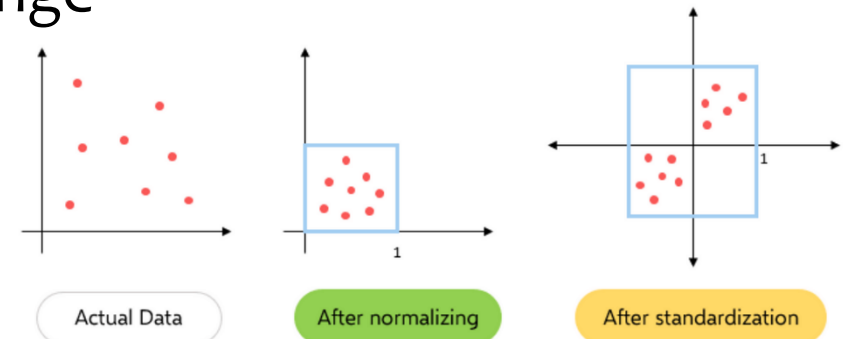
Data Transformation

- Conversion of data values from one format to another.
- Data transformation is a technique used to convert data values from one format to another, helping to improve data quality and consistency."
- Examples:
 - Converting date formats from MM/DD/YYYY to YYYY-MM-DD
 - Transforming data from one unit of measurement to another
 - Changing data types from string to integer



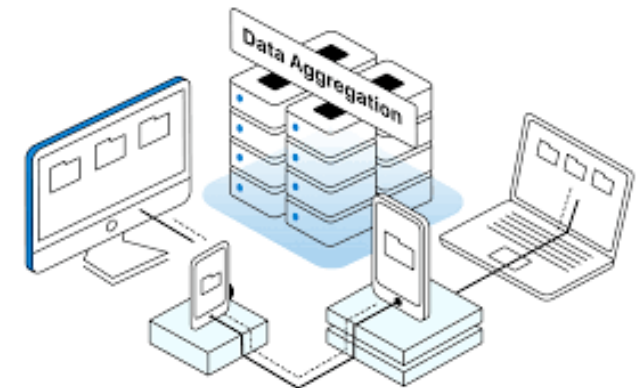
Data Standardization

- Application of consistent formatting and coding conventions.
- Data standardization is a technique used to apply consistent formatting and coding conventions, helping to improve data quality and reduce errors.
- Examples:
 - Standardizing date formats across a database
 - Applying consistent coding conventions for data values
 - Using standardized data formats for data exchange



Data Aggregation

- Combination of data values to reduce redundancy and improve quality.
- Data aggregation is a technique used to combine data values to reduce redundancy and improve quality, helping to improve data management and analysis.
- Examples:
 - Combining data from multiple sources into a single record
 - Aggregating data to reduce redundancy and improve quality
 - Creating summary tables to reduce data volume



Data Summarization

- Reduction of data values to a more compact form.
- Data summarization is a technique used to reduce data values to a more compact form, helping to improve data analysis and decision-making.
- Examples:
 - Creating summary reports to reduce data volume
 - Summarizing data to improve analysis and decision-making
 - Reducing data complexity through summarization

